CSENET: COMPLEX SQUEEZE-AND-EXCITATION NETWORK FOR SPEECH DEPRESSION LEVEL PREDICTION

Cunhang Fan[1]*, Zhao Lv*[1]*, Shengbing Pei*[1]*, Mingyue Niu*[2]

[1] Anhui Province Key Laboratory of Multimodal Cognitive Computation,
School of Computer Science and Technology, Anhui University, Hefei 230601, China
[2] School of Computer and Information Engineering, Tianjin Normal University, Tianjin 300387, China

ABSTRACT

Automatic speech depression level prediction (SDLP) is a very challenging problem in affective computing. There are many studies that have acquired quite good performances for SDLP. However, most of the input speech features of these studies are based on the amplitude spectrogram, which loses the phase spectrogram information. Therefore, these speech features may lose some important information related to depression. In order to make full use of speech information, this paper proposes a complex squeeze-and-excitation network (CSENet) for SDLP. The complex spectrogram is used as the input speech feature, which contains both amplitude and phase spectrogram. In addition, to acquire a discriminative feature, the squeeze-and-excitation residual network is employed to extract deep speech feature. Finally, the attentive temporal pooling is utilized to dynamically select more important information according to the attention mechanisms. Experimental results on the AVEC 2013 and AVEC 2014 datasets prove the effectiveness of our proposed method. As for the mean absolute error (MAE) evaluation metric on AVEC 2013, our proposed method acquires state-of-the-art performance.

Index Terms: speech depression level prediction, complex spectrogram, SENet, attentive temporal pooling

1. INTRODUCTION

Depression is a common mental disorder, which makes people feel of sadness for a long time and deprives them of the joy of life. More seriously, depression endangers human health with self-mutilation and suicide [1]. Early diagnosis and treatment are very helpful for the cure of depression. Therefore, automatic depression detection is necessary to assist doctors in diagnosis.

Many works have shown that speech signals can reflect people's emotion and stress [2, 3]. In addition, physiological studies have also shown that speech signals are different between the depressed and normal individuals [4]. Based on the above research results, many works have been proposed to predict the depression level based on the speech signals [5, 6, 7, 8, 9, 10].

In order to improve the performance of speech depression level prediction (SDLP), there are so many speech features. For example, in [6], authors find that the Mel-frequency cepstral coefficient (MFCC) is a discriminative acoustic feature. Therefore, they apply the MFCC for SDLP and acquire a good performance. Besides the MFCC, in [10], authors use the acoustic low-level descriptors (LLDs) and 3D log Mel spectrograms for SDLP. In addition, in [7], authors utilize the pre-trained deep learning network to extract deep speech features, which are used for SDLP. Experimental results show that these features can improve the prediction performance. However, these above features are based on the amplitude spectrogram discard phase information. The phase spectrogram may contains much important information for SDLP. And many studies have shown that the phase spectrogram is very important for the speech quality and intelligibility [11, 12]. Therefore, it is unreasonable to abandon this information.

To address the above problem, this paper proposes a complex squeeze-and-excitation network (CSENet) for SDLP, which uses the complex spectrogram as the input feature. The complex spectrogram contains all of the speech information both amplitude and phase spectrogram. Therefore, we apply this feature to make full use of the speech information so that it can improve the prediction accuracy. In addition, to acquire a discriminative features for SDLP, the squeeze-and-excitation residual network (SE-ResNet) [13, 14] is applied to extract the deep speech features. Finally, motivated by [15, 16], we utilize the attentive temporal pooling to acquire long-term dependencies for low-dimensional speech representations and dynamically select more important information according to the attention mechanisms. The visualization results show that the attentive temporal pooling is very effective for discriminate different depression levels. The main contributions of this study can be summarized as follows:

- We propose a novel CSENet with attentive temporal pooling for SDLP.
- To be our best knowledge, this is the first work to apply the complex spectrogram for SDLP.
- The experimental results on the Audio/Video Emotion Challenge (AVEC) 2013 [17] and AVEC2014 [18] in-

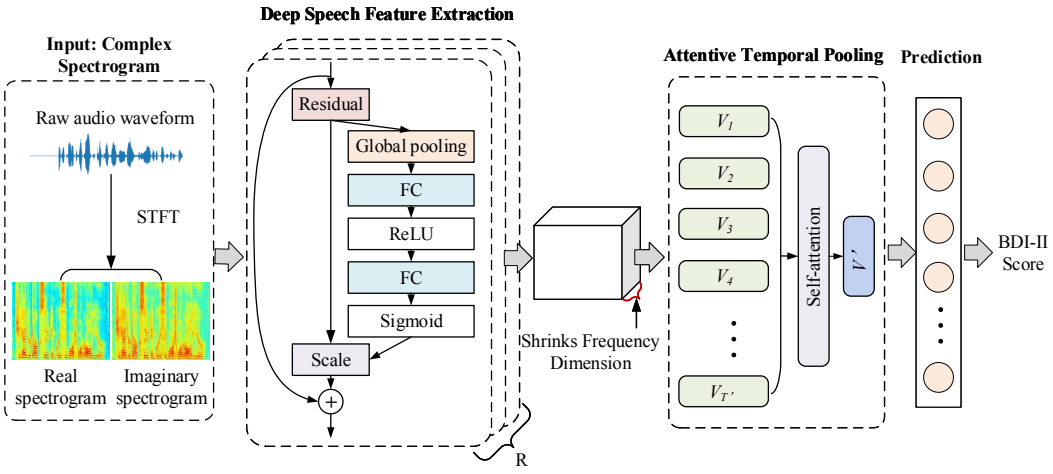


Fig. 1. The schematic diagram of our proposed CSENet method for SDLP. The CSENet consists of complex spectrogram, deep speech feature extraction and attentive temporal pooling.

dicate the superiority of our proposed method. In addition, as for mean absolute error (MAE) evaluation metric on AVEC2013, our proposed method acquires state-of-the-art performance.

2. OUR PROPOSED CSENET METHOD

Fig. 1 shows the schematic diagram of our proposed CSENet method for SDLP. From Fig. 1 we can find that the complex spectrogram is used as the input feature of CSENet. Then the SE-ResNet is applied to extract discriminative deep speech features. After SE-ResNet, frequency is shrunk to a lower dimension. Then we use another two convolution neural network (CNN) layers to further compress the frequency channel to one dimension. Finally, the attentive temporal pooling is utilized to acquire long-term dependencies.

2.1. Complex Spectrogram

In order to make full use of the speech information both amplitude and phase spectrogram, we explore the complex spectrogram as the input feature of SDLP. Firstly, the time domain speech signals $x[k]$ is converted into time-frequency domain by the short-time Fourier transformation (STFT):

$$Real[t,f] + i * Imag[t,f] = STFT(x[k]) \tag{1}$$

where $STFT$ denotes the function of STFT, $Real$ and $Imag$ are the corresponding real and imaginary part, respectively. t is the index of time frame and f is the index of frequency bin. k is the time index of speech signals.

Then the real and imaginary part of STFT are stacked together as the complex spectrogram $\mathbf{X}$:

$$\mathbf{X} = stack(Real[t,f], Imag[t,f]) \in \mathbb{R}^{2\times T\times F} \tag{2}$$

where the $stack(*)$ means the stack operation.

2.2. Deep Speech Features Extraction

Because the SE-ResNet [13] can explicitly model the interdependencies between the channels of its convolutional features so that it can improve the quality of deep speech representations. In addition, it can automatically obtain the importance of each feature channel, and then improve the useful features according to this importance and suppress the features that are not useful to the current task. Therefore, the SE-ResNet is used as the extractor of deep speech features to acquire discriminative acoustic features.

The SE block is a computational unit and shown in Fig. 1. It firstly uses the global pooling (GP) to squeeze the input vector so that acquires the global information embedding.

$$z_c = \mathbf{F}_{GP}(\mathbf{u}_c) \tag{3}$$

where $\mathbf{u}_c$ is the c-th channel input vector $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_C]$, which is generated by the convolution operation from the complex spectrogram. The $\mathbf{F}_{GP}$ means the GP operation. $\mathbf{z} = [z_1, z_2, ..., z_C] \in \mathbb{R}^C$ and z_c is the c-th element of $\mathbf{z}$.

Then two fully-connected (FC) layers are applied to make full use of the aggregated information by the squeeze operation and fully capture channel-wise dependencies.

$$\mathbf{s} = \sigma(\mathbf{W}_2\delta(\mathbf{W}_1\mathbf{z})) \tag{4}$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are the weight matrix of these two FC layers. In addition, the δ and σ denote the ReLU and sigmoid function, respectively.

Finally, the scale operation is used to acquire the final output of the SE block:

$$\tilde{x}_c = \mathbf{F}_{scale}(\mathbf{u}_c, s_c) = s_c\mathbf{u}_c \tag{5}$$

where $\tilde{\mathbf{X}} = [\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_C]$ and $\mathbf{F}_{scale}(\mathbf{u}_c, s_c)$ means the channel-wise multiplication.

After SE-ResNet, the frequency of input vector is shrunk to a lower dimension. In order to acquire the deep speech feature with only the time dimension, we use another two convolution neural network (CNN) layers to further compress the frequency channel to one dimension.

$$\mathbf{V} = \delta(\mathcal{G}(\tilde{\mathbf{X}})) \in \mathbb{R}^{D \times T'} \tag{6}$$

where $\mathcal{G}$ means the CNN function, $\mathbf{V} = [\mathbf{V}_1, \mathbf{V}_2, ..., \mathbf{V}_{T'}]$ denotes the deep speech features, and T' is the time dimension after SE-ResNet, D is the dimension of $\mathbf{V}_i$.

2.3. Attentive Temporal Pooling

Motivated by [15, 16], we utilize the attentive temporal pooling in order to not only make our model can process varying length inputs, but also to dynamically select more important information according to the attention mechanisms. The attentive temporal pooling mainly aims to covert a set of local descriptors $\mathbf{V}_i$ into a single global descriptor $\mathbf{V}'$.

$$d_i = \mathbf{V}_i^{Trans} \mathbf{V}_i \tag{7}$$

$$\alpha_i = \frac{\exp(d_i)}{\sum_i \exp(d_i)} \tag{8}$$

$$\mathbf{V}' = \mathbf{W}_3[concate\{\mu(\alpha_i \mathbf{V}_i); \zeta(\alpha_i \mathbf{V}_i)\}] \tag{9}$$

where μ and ζ denote the mean and variance. Finally, global descriptor $\mathbf{V}'$ can be acquired by a linear projection $\mathbf{W}_3$.

3. EXPERIMENTS AND RESULTS

3.1. Dataset

We conduct our experiments on two public datasets i.e., AVEC 2013 and AVEC 2014. AVEC 2013 dataset consists of 150 videos from 84 subjects. Each subject is required to perform 14 different tasks according to the instructions on the computer screen. The duration of each video ranges from 20 to 50 minutes. The average length of each recording is 25 minutes. There are three parts in AVEC 2013: training, development and test sets, each with 50 samples.

AVEC 2014 is a subset of AVEC 2013, which consists of two different tasks: Northwind and FreeForm. There are 150 videos for each task that is equally divided into three parts: training, development and test sets. In this paper, we merge these two tasks as a new database. Therefore, there are 100 videos for training, development and test sets, respectively.

As for AVEC 2013 and AVEC 2014, the Beck Depression Inventory-II (BDI-II) is used as the ground truth, which indicates the depression level.

3.2. Evaluation Metrics

In this work, in order to quantitatively evaluate SDLP results, root mean square error (RMSE) and mean absolute error (MAE) are used as evaluation metrics.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{10}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{11}$$

where y_i and $\hat{y}_i$ denote the true and predicted BDI-II score of the i-th sample, respectively.

3.3. Experimental setup

Firstly, we divide each audio sample into several segments, which has 3s and the overlap of two adjacent segments is 50%. The sampling rate of all generated speech waveform is 8000 Hz. As for the STFT, the hamming window is 50ms, window shift is 12.5ms and the fast Fourier transform (FFT) points are 512. Therefore, the dimension of the spectrogram is 257. The architecture of deep speech feature extractor is based on the ResNet18 [14]. We utilize 256 for the dimension of deep speech features, which means that D is 256. In addition, we use the Adam as our optimizer, and the learning rate is 0.002 for AVEC 2013. The learning rate of AVEC 2014 is 0.0006.

3.4. Experimental results

Table 1 and Table 2 show the experimental results for different SDLP methods on the AVEC 2013 and AVEC 2014 datasets. CResNet and CSENet mean that we apply the ResNet [14] and SE-ResNet [13] as our deep speech feature extractor, respectively.

3.4.1. Experimental performance of CResNet and CSENet

From Table 1 we can find that when we use the SE-ResNet replace the ResNet as the deep feature extractor, the performance can be improved. More specifically, compared with the CResNet method, our proposed CSENet method can reduce the RMSE and MAE from 9.42, 7.19 to 9.28, 6.79, respectively. In addition, from Table 2 we can find that although CSENet gets a worse performance for RMSE than CResNet, the result of MAE is still better than CResNet. These results indicate that the deep speech features extracted by SE-ResNet is more suitable for the SDLP than ResNet. The reason is that the SE-ResNet with squeeze and excitation operation can automatically obtain the importance of each feature channel. In addition, it can also improve the useful features according to this importance and suppress the features that are not useful to current task. Therefore, the CSENet can acquire a better performance than CResNet.

3.4.2. Comparison with other SDLP methods

Table 1 and Table 2 also show the results of other SDLP methods on the AVEC 2013 and AVEC 2014 datasets. Where

Table 1. Experimental results for different SDLP methods on the AVEC 2013 dataset. abs-SENet means that the amplitude is used as the input feature without the phase information.

Methods	RMSE	MAE
AVEC 2013 Audio Baseline [17]	14.12	10.35
PLS regression [8]	11.19	9.14
DCNN [9]	10.00	8.20
CNN-LSTM-SVR [6]	9.79	7.48
SAN-CNN [10]	9.65	7.38
STA-EEP [19]	9.50	7.14
abs-SENet	9.72	7.45
CResNet (ours)	9.42	7.19
CSENet (ours)	**9.28**	**6.79**

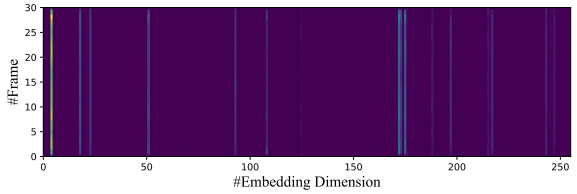
Table 2. Experimental results for different SDLP methods on the AVEC 2014 dataset.

Methods	RMSE	MAE
AVEC 2014 Audio Baseline [18]	12.56	10.03
Fisher Vector Encoding [20]	11.51	9.74
PCA+Linear Regression [21]	10.28	8.07
DCNN [9]	9.99	8.19
CNN-LSTM-SVR [6]	9.66	8.02
SAN-CNN [10]	9.57	7.94
STA-EEP [19]	**9.13**	7.65
abs-SENet	10.04	8.02
CResNet (ours)	9.24	7.18
CSENet (ours)	9.61	**7.13**

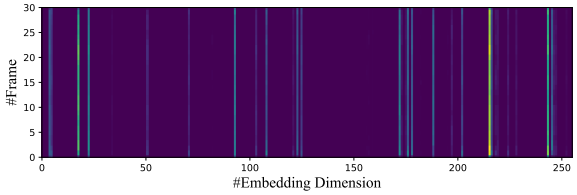
the [8, 9, 10, 17, 18, 20, 21] methods apply the LLDs and MFCC as the input speech features. The [6] only uses MFCC as the input speech features. And the amplitude spectrogram is utilized as the input speech features for [19]. These above speech features are lost the phase spectrogram information, which may loses some very important information for SDLP. In this paper, we apply the complex spectrogram as our speech features, which contains both the amplitude and phase spectrograms. From Table 1 we can find that compared with other SDLP methods, our proposed complex spectrogram-based SDLP method can get the best performance no matter RMSE or MAE. In addition, as for the AVEC 2014 dataset, although our proposed method get worse performance than [19] for RMSE, it acquires a better result for MAE evaluation metric. To be our best knowledge, as for the AVEC 2013 dataset, our proposed CSENet method gets the state-of-the-art performance for MAE evaluation metric. These results prove the effectiveness of our proposed method. In addition, these results also demonstrate that the complex spectorgram is a suitable feature for SDLP task.

3.4.3. Effectiveness of attentive temporal pooling

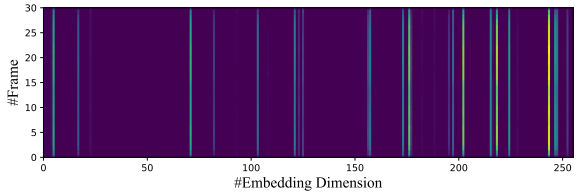
Fig. 2 shows the visualization results of attentive temporal pooling embedding for different individuals. The horizontal axis means the dimension of embedding, and vertical axis is the time frame after SE-ResNet. (a) is the healthy subject of No. 319-2. (b) is the mild depression subject of No. 220-3. (c) is the severe depression subject of No. 237-1. From Fig. 1 we can find that the more severe the level of depression, the more bright bars in the visualization figure, which indicates that this embedding is a very discriminative representation for SDLP. This result prove that the attentive temporal pooling can effectively learn depression information and it is very effective for SDLP.



(a) Healthy subject of No. 319-2



(b) Mild depression subject of No. 220-3



(c) Severe depression subject of No. 237-1

Fig. 2. The visualization results of attentive temporal pooling embedding for different individuals. (a) is the healthy subject of No. 319-2. (b) is the mild depression subject of No. 220-3. (c) is the severe depression subject of No. 237-1.

4. CONCLUSIONS

In order to make full use of speech information, this paper proposes a CSENet for speech depression level prediction. The proposed CSENet uses the complex spectrogram as the input feature, which contains both amplitude and phase spectrogram information. In addition, the SE-ResNet is applied to extract deep speech features. Finally, the attentive temporal pooling is used to dynamically select more important information according to the attention mechanisms. Experiments on AVEC 2013 and AVEC 2014 demonstrate that our proposed method is effective for SDLP and the complex spectorgram is a suitable feature for speech depression level prediction task. In future, we will explore different architecture of deep speech feature extractor to pay attention to different frequency bands.

5. ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (NSFC) (No.61972437), the Open Research Projects of Zhejiang Lab (NO. 2021KH0AB06) and the Open Projects Program of National Laboratory of Pattern Recognition (NO. 202200014, 202200012).

6. REFERENCES

[1] Maurizio Fava and Kenneth S Kendler, "Major depressive disorder," *Neuron*, vol. 28, no. 2, pp. 335–341, 2000.

[2] Jes Olesen, Anders Gustavsson, Mikael Svensson, H-U Wittchen, Bengt Jönsson, CDBE2010 Study Group, and European Brain Council, "The economic cost of brain disorders in europe," *European journal of neurology*, vol. 19, no. 1, pp. 155–162, 2012.

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

[4] 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.

[5] Cong Cai, Mingyue Niu, Bin Liu, Jianhua Tao, and Xuefei Liu, "Tdca-net: Time-domain channel attention network for depression detection," *Proc. Interspeech 2021*, pp. 2511–2515, 2021.

[6] 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.

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

[8] 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.

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

[10] 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.

[11] Kuldip Paliwal, Kamil Wójcicki, and Benjamin Shannon, "The importance of phase in speech enhancement," *speech communication*, vol. 53, no. 4, pp. 465–494, 2011.

[12] Cunhang Fan, Jianhua Tao, Bin Liu, Jiangyan Yi, Zhengqi Wen, and Xuefei 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.

[13] 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.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[15] Joao Monteiro, Jahangir Alam, and Tiago H Falk, "Generalized end-to-end detection of spoofing attacks to automatic speaker recognizers," *Computer Speech & Language*, vol. 63, pp. 101096, 2020.

[16] João Monteiro, Jahangir Alam, and Tiago H Falk, "Residual convolutional neural network with attentive feature pooling for end-to-end language identification from short-duration speech," *Computer Speech & Language*, vol. 58, pp. 364–376, 2019.

[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] Mingyue Niu, Jianhua Tao, Bin Liu, Jian Huang, and Zheng Lian, "Multimodal spatiotemporal representation for automatic depression level detection," *IEEE Transactions on Affective Computing*, 2020.

[20] 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.

[21] Asim Jan, Hongying Meng, Yona Falinie Binti A Gaus, and Fan 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.