Are You Speaking with a Mask? An Investigation on Attention Based Deep Temporal Convolutional Neural Networks for Mask Detection Task

Yu Qiao[1], Kun Qian[2(✉)], Ziping Zhao[1(✉)], and Xiaojing Zhao[1]

[1] Tianjin Normal University, Tianjin, China
[2] The University of Tokyo, Tokyo, Japan
qian@p.u-tokyo.ac.jp

Abstract. When writing this article, COVID-19 as a global epidemic, has affected more than 200 countries and territories globally and lead to more than 694,000 deaths. Wearing a mask is one of most convenient, cheap, and efficient precautions. Moreover, guaranteeing a quality of the speech under the condition of wearing a mask is crucial in real-world telecommunication technologies. To this line, the goal of the ComParE 2020 Mask condition recognition of speakers subchallenge is to recognize the states of speakers with or without facial masks worn. In this work, we present three modeling methods under the deep neural network framework, namely Convolutional Recurrent Neural Network(CRNN), Convolutional Temporal Convolutional Network(CTCNs) and CTCNs combined with utterance level features, respectively. Furthermore, we use cycle mode to fill the samples to further enhance the system performance. In the CTCNs model, we tried different network depths. Finally, the experimental results demonstrate the effectiveness of the CTCNs network structure, which can reach an unweighted average recall (UAR) at 66.4% on the development set. This is higher than the result of baseline, which is 64.4% in S2SAE+SVM nerwork(a significance level at $p < 0.001$ by one-tailed z-test). It demonstrates the good performance of our proposed network.

Keywords: Computational paralinguistics · Deep learning framework · Mask condition recognition · Speech recognition

1 Introduction

COVID-19, as a pandemic, has more than 20 million confirmed patients (causing more than 748 000 deaths), and is still affecting more than 200 countries and territories globally at the time of writing this paper[1]. Computer audition (CA), a multidisciplinary field that leverages the advanced acoustic/audio signal processing and machine learning technologies to enable the machines having or even

[1] https://coronavirus.jhu.edu/map.html.

X. Shao et al. (Eds.): CSMT 2020, LNEE 761, pp. 163–174, 2021.
https://doi.org/10.1007/978-981-16-1649-5_14

outperforming the human hearing capacities, has been increasingly applied to the healthcare domain [12]. More recently, CA has been thought to have promising potential for fighting the COVID-19 pandemic due to its non-invasive and ubiquitous characteristic by nature [9,15].

In this paper, we aim to develop a speech-driven deep learning framework to recognize people with or without facial masks worn. The task is proposed as part of the INTERSPEECH 2020 Computational Paralinguistics ChallengE (ComParE) [14]. The data offered in this challenge is called the MASC (the Mask Augsburg Speech Corpus) dataset, which is the first to give access to recordings of speech from individuals wearing an operation mask. The labels of the data are their condition states while communicating, including Masking and Clear. Many existing works have been performed on the speech recognition research. Some acoustic features, such as the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [7], ComParE acoustic feature set and Bag of-Audio-Words (BoAW) feature set [13], combined with traditional machine learning methods, have been proved to be effective for recognizing the speech signals.

With the development of deep learning, neural network has made substantial achievements in the computational paralinguistics field. Neural network have been widely used due to the superior performance, such as speaker identification [11,17,19,21], language identification [1–3] and speech emotion recognition [5,20]. Therefore, various neural network frameworks such as convolutional neural network (CNN) and recursive neural network (RNN) have emerged. CNN is used to extract spatial features and generate feature maps. The extensive application from AlexNet to VGG model reflects the superior performance of CNN. The pre-trained AlexNet network was used to extract deep features, and then the features of the full connection layer were input into Support Vector Machine (SVM) for classification, which achieved good performance on the data set FAU-AIBO [6]. Two different convolution nuclei were used to extract time-domain and frequency-domain features respectively, and then the features were classified by CNN after fusion. Finally, the UAR of the four categories of emotions of IEMOCAP reached 68% [10]. Gated Recurrent Unit (GRU) and Long-Short Term Memory (LSTM) are also widely used, GRU is a variant of LSTM, they can solve the gradient vanishing problem in the RNN optimization process. Greff et al. benchmarked eight LSTM variants on speech recognition [8]. The combination of CNN and RNN is widely used. Mingyi et al. added LSTM after CNN, and found that the five convolution layers had the best performance on EmoDB [4]. However, with the deepening of network layers, some information will be lost because CNN has no memory function, and the operation time of RNN is relatively long.

Main contributions of this work can be summarised as follows: First, we compare the performance of two different network topologies on this classification problem and find the good effect of TCN on this classification problem. Second, we have introduced attention mechanism across all network structures to allow the network to focus on key features during training. Third, we integrate utterance level features into the network structure with good performance, realized the fusion of deep learning representation and utterance level features.

In this article, We investigate and compare three topologies, i.e., Convolutional Recurrent NeuralNetwork (CRNN), Convolutional Temporal Convolutional Networ (CTCNs) and CTCNs with utterance level features. In addition, CNN and attention are added to both models to improve the network performance.

This paper is organized as follows: Firstly, we introduce the methods used in Sect. 2. Section 3 introduces experimental design, including data preprocessing, experimental setting, and experimental results. And the discussion will be given in Sect. 4. Finally, we conclude this study in Sect. 5.

2 Methods

2.1 BLSTM

BLSTM is composed of forward LSTM and backward LSTM. In the LSTM, there are three kinds of gates: forgetting gate, input gate and output gate. The forgetting gate can selectively forget some information, and the input gate new information selectively recorded, and in the output gate for output. In BLSTM, forward LSTM is used to help the network learn sequence characteristics forward and backward LSTM learns sequence information later. This design can help the network form sequence memory. When we input the extracted mask audio sequence, we can not only accumulate the information of the input moment, but also remember the information of the previous moment, which has a good effect on dealing with the time series problem.

The network structure diagram of BLSTM is shown in Fig. 1, from which we can see that the output layers results are jointly controlled by forward layers and backword layers, and the final output results can be expressed as follows by mathematical expressions:

$$h_t = f(w_1 x_t + w_2 h_{t-1}) \tag{1}$$

$$h_t^{'} = f(w_4 x_t + w_5 h_{t+1}^{'}) \tag{2}$$

$$O_t = g(w_3 h_t + w_6 h_t^{'}) \tag{3}$$

where, Eq. (1) represents the result of forward propagation, Eq. (2) represents the result of back propagation, and Eq. (3) represents the expression of the output result after BLSTM.

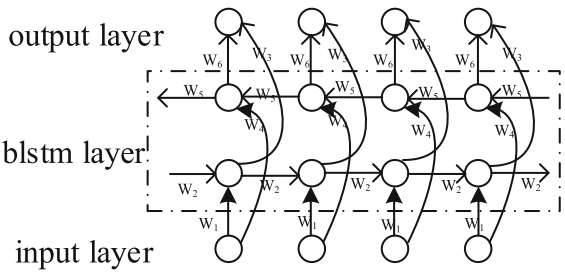
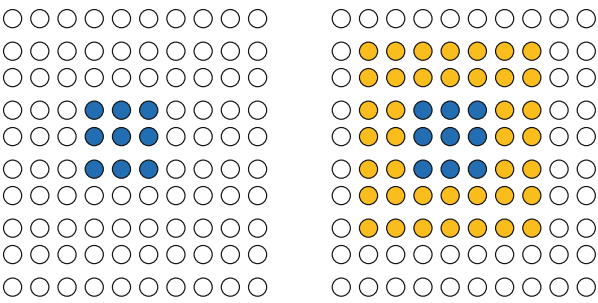


Fig. 1. BLSTM network structure.

2.2 TCN

Similar to BLSTM, TCN can also be used to handle time series problems. TCN network is all convolution operation, which means that TCN neural network can carry out large-scale parallel processing, which is shorter than BLSTM to some extent, which involves the skip layer connection of dilated convolution, causal convolution, and residual convolution.



(a) General Convolution (b) Dilated Convolution

Fig. 2. Contrast diagram of convolution receptive field.

Dilation rate parameter is involved in the part of dilation convolution, which is used to represent the size of the dilation, so that the convolution process has a larger receptive field. As shown in Fig. 2(a) represents the receptive field of dilated convolution, and (b) represents the receptive field of general convolution. From the figure, the advantages of the receptive field of dilated convolution can be clearly seen. Where, the size of the convolution kernel in (a) is 3, and after dilation rate, the size of the convolution kernel becomes 5, and finally the receptive field of (b) is obtained.

Where, the calculation of the size of the dilated convolution kernel follows: dilated filter $= d * (k - 1) + 1$, where d stands for dilation rate and k stands for the size of the convolution kernel.

By referring the dilative convolution to the causal convolution, the prediction at time t can take into account the sequence before time$t$, thus achieving a time memory effect similar to BLSTM. The skip layer of residual convolution is realized by 1D fully-convolutional network (FCN) [16], which equals the length of the output sequence to that of the input sequence [22].

2.3 Attention Mechanism

To ensure the reliability of model training, we added the attention layer to the network structure and the Attention mechanism after the weight causal layers in TCN, as shown in Fig. 3.

In this paper, the attention layer in the network structure is a sequence coding layer, which is a series of weight allocation coefficients. When the input information at time t is more similar to the target information, the attention layer assigns more weight to the time t, that is, the output of the sequence is more dependent on the time t. In the experiment of this paper, Self-Attention mechanism is used, which can find the internal connection of the sequence in the training process, so as to ensure the similarity between the output sequence and the input sequence. So, we use Scaled Dot-Product Attention [18], the implementation equation is

$$Attention(Q, K, V) = softmax(\frac{Q \times K^T}{\sqrt{d_k}}) \times V \tag{4}$$

Here, K and V are the values of mask audio data after Self-Attention, Q is the data that corresponds to the label by masked Self - Attention after the value, d_k is the number of channels in the input sequence, used as a normalization.

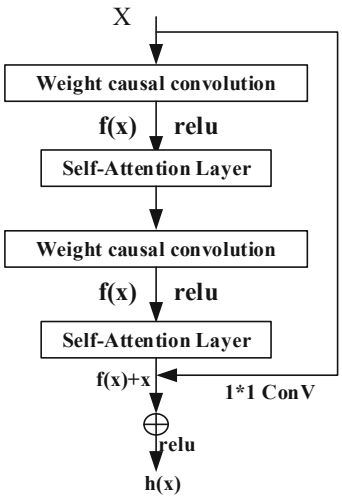


Fig. 3. Attention residual learning block.

3 Experiment Design

3.1 Data Pre-processing

In this part, all the audio data in the data set are circulated and filled in for 4s (the original data set lasts for 1s and the sampling frequency is 16 kHZ). Then, the librosa library is used to perform short-time Fourier transform to extract mel spectrogram. The parameters in the process of mel spectrogram extraction are as follows: the window width $w = 25$ ms, the window shift 10 ms, and n_{mels} = 128 mel frequency bands.

3.2 Experimental Setting

In our experiment, we mainly used three network learning models: CRNN, CTCNs and CTCNs with utterance level features. We will describe these three network structures in detail below. It should be noted that due to the limitation of server storage space, the batchsize of all our experiments is 64.

CRNN. In this model, we first used CNNs to extract features from the mel spectrograms, considering the effect of the preceding sequence on the prediction of the following sequence, we use BLSTM to remember information through forward propagation and backward propagation, so as to make the predicted results more robust. At the same time, after BLSTM layer, add the attention layer to allocate the feature weight, so that the network can focus on the features that play a key role in the classification effect. Finally, the spatial features extracted from the convolutional layer and the sequence features after the attention layer are fused as the final classification features, and the classification is carried out through the full connection layer containing softmax function. More specifically, our network model is described in Table 1.

Table 1. Our network structure

Nework layers	Parameter
Conv1	16, 7 * 7 kernels, 1 stride
Pooling	2 * 2 pooling, 2 stride
Dropout	0.25
Conv2	16, 5 * 5 kernels, 1 stride
Pooling	2 * 2 pooling, 2 stride
Dropout	0.25
Conv3	32, 5 * 5 kernels, 1 stride
Pooling	2 * 2 pooling, 2 stride
Dropout	0.25
MaxPooling	BLSTM/
	TCN blocks: 3 * 3 kernels,d: [1, 2, 4...]
	Self-attention layer
Features concatenation	
Full-connected Layer	4096 units
Classification Layers	Softmax

CTCNs. In this network structure, we use TCN and attention layer to build the network structure. In the TCN module, we mainly used the structure in Fig. 3. Multi-layer stacking was performed in residual block in Fig. 3 to build the main part of the network. Of course, with the stack of blocks, the number of layers in

the network would be deepened, and the attention layer would be added after the last layer. This approach is to achieve similar functions to BLSTM, enabling the network to extract time series features. Considering the impact of spatial features on the classification results, we added three convolutional layers at the beginning of the network. The features extracted by the convolutional layer were on the one hand input into the TCN network module, and on the other hand retained and fused with the sequence features extracted by the TCN module, thus forming the features of final progressive classification. The final features are sorted through the full connection layer of 4096 units by softmax. The detailed network structure is shown in Table 1.

CTCNs with Utterance Level Features. In the experiment, we mainly used the manually designed features of low level descriptors (LLDs) and high level statistics functions (HSFs), obtained utterance level features by making statistics on the voice features at the frame level, such as maximum value and mean value, and so on. Here, opensmile toolkit is used to extract utterance level features, and the feature Set used is ComParE.

In this part, we added utterance level features to integrate the deep features extracted from deep learning for classification. The extraction of deep features is based on the experiment in Sect. 3.2, and the features extracted from its full connection layer are used.

Refer to the specific network structure Fig. 4.

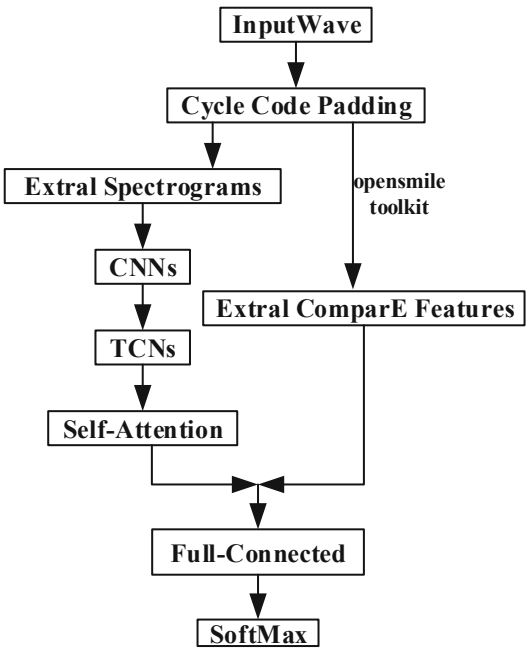


Fig. 4. TCNs with utterance level features network structure.

Experimental Result. In this paper, we will use unweighted average recall (UAR) to evaluate the experimental results of various network structures. As this is a Sub-Challenge task, all our results are obtained on the development set. For Sect. 3.2, we conducted experiments with [3,10] different attention residual blocks, and the experimental results are shown in Table 2. It can be seen from the table that when residual blocks is 4, the experimental UAR is 66.4%, which is the best result. The number of channels and experiment time for each block are also shown in Table 2. The confusion matrix corresponding to the experiment is shown in the Fig. 5.

Table 2. The result of CTCNs network structure on the development set

Nework blocks	Channels	WAR (%)	UAR (%)	Time (s)
3	[64, 128, 256]	64.6	65.1	3692.98
4	[64, 128, 256, 512]	66.3	**66.4**	4874.648
5	[64, 128, 256, 512, 1024]	64.7	64.8	4218.848
6	[64, 128, 256, 512, 1024, 2048]	66.3	65.6	8350.363
7	[64, 128, 256, 512, 1024, 2048, 4096]	65.4	65.0	12205.034
8	[64, 64, 128, 128, 256, 256, 512, 1024]	66.6	66.0	10779.831
9	[64, 64, 64, 128, 128, 128, 256, 256, 512]	64.8	64.4	5451.703
10	[64, 64, 64, 128, 128, 128, 256, 256, 512, 1024]	65.2	65.1	15232.913

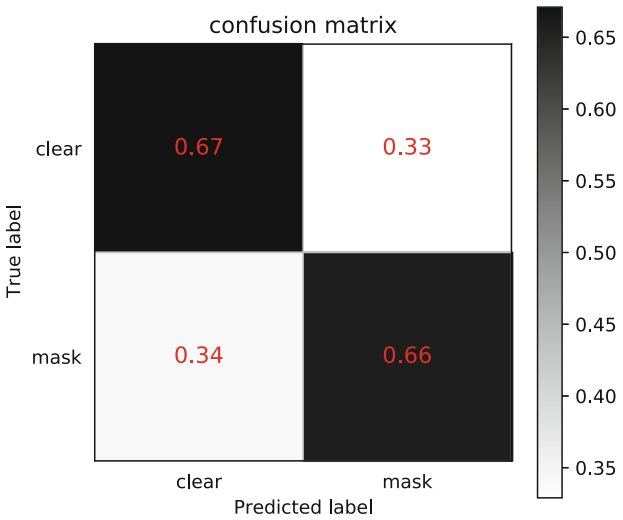


Fig. 5. Confusion matrix graph on the development set.

Table 3. Results of different network structures on the development set

ID	Nework structure	UAR(%)
1	CRNN	65.5
2	CTCNs	**66.4**
3	CTCNs + ComparE	65.9
4	ComparE + SVM [14]	62.6
5	ComparE BOAW + SVM [14]	64.2
6	DeepSpectrum + SVM [14]	63.4
7	S2SAE + SVM [14]	64.4

As can be seen from the table, the lowest experimental result of our proposed method is 65.5%, while the experimental result of S2SAE model in the original paper is the best, with its UAR being 64.4%, which is lower than our lowest result by 1.1%, which fully proves the performance of our network structure.

4 Discussion

It can be seen from Table 2 that the network of 4-layer blocks has the best result on the development set. As the network deepens to 10 layer blocks, the UAR of the network is not as good as that of 4-layer blocks. This may be from the side that the deepening of the network makes the training gradient unstable. In Table 2, we can see that when blocks is 7 or 8, channels are the most and the experiment takes more than 10,000 s.

The experimental results of different network structures are shown in Table 3. The model 1, 2, 3 network structures are the three methods tried in this paper, and the model 4,5,6,7 are the experimental results of the original paper's network structures. The difference between model 2 and model 1 is that model 2 uses TCN to extract sequence features, while model 1 uses BLSTM, and it is finally found that the experimental results of model 2 are better than those of model 1, which maybe indicates that TCN has a better fitting on this data set. When we fusioned utterance level features (in this article, ComParE the features) into model 2, the experimental result is 65.9% in model 3, but this reduced the results by 0.5%. We consider the reasons for this result may be to join utterance level features, making increased certain features of the similarity between different categories, it increases the classification error, thus resulting in a loss of the experimental results. It may be possible to try other utterance level features for fusion, hoping to improve the classification result. Model 3 is about 3% higher than model 4, and it turns out that the TCN network extracts features that are useful for classification.

5 Conclusion

Mask Sub-Challenge detection is a challenging task. In this paper, we first adopted the cycle code padding method to process the raw audio, and then conducted experiments on the MASC data set through three different network structures, namely CRNN, CTCNs and CTCNs with utterance level features. CTCNs achieves the best performance on the development set.

The experimental result of model 4 is the lowest, which used only ComParE features, while model 2 adds spectral features on this basis, the results increased by 3.3%, which may indicate the advantage of mel spectrograms in this data set. All the deep feature extraction in this paper is based on the spectrograms extracted by the short-time Fourier Transform (STFT). However, window size in the process of STFT do not have adaptivity and cannot be optimized for specific problems, so better results may be obtained by using wavelet transform to extract spectrograms.

The experimental results of other models are better than model 4, which may reflect the good performance of deep learning. This suggests that we should not be confined to machine learning, and future research can be developed towards deep learning, perhaps with better results.

Acknowledgements. This work was partially supported by the National Natural Science Foundation of China (Grant No. 61702370), P. R. China, the Key Program of the Natural Science Foundation of Tianjin (Grant No. 18JCZDJC36300), P. R. China, the Open Projects Program of the National Laboratory of Pattern Recognition, P. R. China, the Zhejiang Lab's International Talent Fund for Young Professionals (Project HANAMI), P. R. China, the JSPS Postdoctoral Fellowship for Research in Japan (ID No. P19081) from the Japan Society for the Promotion of Science (JSPS), Japan, and the Grants-in-Aid for Scientific Research (No. 19F19081) from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

1. Bartz, C., Herold, T., Yang, H., Meinel, C.: Language identification using deep convolutional recurrent neural networks. In: Proceedings of the 24th International Conference of Neural Information Processing, pp. 880–889. Springer, Guangzhou, China (2017)
2. Cai, W., Cai, D., Huang, S., Li, M.: Utterance-level end-to-end language identification using attention-based cnn-blstm. In: Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5. IEEE, Brighton, UK (2019)
3. Chan, W., Jaitly, N., Le, Q., Vinyals, O.: Listen, attend and spell: a neural network for large vocabulary conversational speech recognition. In: Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4960–4964. IEEE, Shanghai, China (2016)
4. Chen, M., He, X., Yang, J., Zhang, H.: 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. IEEE Sig. Process. Lett. **25**(10), 1440–1444 (2018)

5. Chernykh, V., Sterling, G., Prihodko, P.: Emotion recognition from speech with recurrent neural networks, pp.1–18 (2017). ArXiv:abs/1701.08071
6. Cummins, N., Amiriparian, S., Hagerer, G., Batliner, A., Steidl, S., Schuller, B.W.: An image-based deep spectrum feature representation for the recognition of emotional speech. In: Proceedings of the 25th ACM International Conference on Multimedia, pp. 478–484. Association for Computing Machinery, Seattle, USA (2017)
7. Eyben, F.: The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Trans. Affect. Comput. **7**(2), 190–202 (2016)
8. Greff, K., Srivastava, R.K., Koutník, J., Steunebrink, B.R., Schmidhuber, J.: LSTM: a search space odyssey. IEEE Trans. Neural Netw. Learn. Syst. **28**(10), 2222–2232 (2017)
9. Han, J., Qian, K., Song, M., Yang, Z., Ren, Z., Liu, S., Liu, J., Zheng, H., Ji, W., Koike, T., et al.: An early study on intelligent analysis of speech under Covid-19: Severity, sleep quality, fatigue, and anxiety. In: Proceedings of Interspeech, pp. 4946–4950. Shanghai, China (2020)
10. Li, P., Song, Y., McLoughlin, I.V., Guo, W., Dai, L.R.: An attention pooling based representation learning method for speech emotion recognition. In: Proceedings of Interspeech. ISCA, Hyderabad, India, pp. 3087–3091 (2018)
11. Matëjka, P., Glembek, O., Novotny, O., Plchot, O., Grézl, F., Burget, L., Cernocky, J.: Analysis of dnn approaches to speaker identification. In: Proceedings of the 41st IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5100–5104. IEEE, Shanghai, China (2016)
12. Qian, K., Li, X., Li, H., Li, S., Li, W., Ning, Z., Yu, S., Hou, L., Tang, G., Lu, J., Li, F., Duan, S., Du, C., Cheng, Y., Wang, Y., Gan, L., Yamamoto, Y., Schuller, B.W.: Computer audition for healthcare: opportunities and challenges. Front. Digit. Health **2**, 5 (2020)
13. Schmitt, M., Schuller, B.: openXBOW - introducing the Passau open-source crossmodal bag-of-words toolkit. J. Mach. Learn. Res. **18**(96), 1–5 (2017)
14. Schuller, B.W., et al.: The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly emotion, Breathing & Masks. In: Proceedings of Interspeech, pp. 2042–2046. Shanghai, China (2020)
15. Schuller, B.W., Schuller, D.M., Qian, K., Liu, J., Zheng, H., Li, X.: Covid-19 and computer audition: an overview on what speech & sound analysis could contribute in the SARS-CoV-2 corona crisis, pp. 1–7. arXiv preprint arXiv:2003.11117 (2020)
16. Shelhamer, E., Long, J., Darrell, T.: Fully convolutional networks for semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 640–651 (2017)
17. Snyder, D., Ghahremani, P., Povey, D., Garcia-Romero, D., Carmiel, Y., Khudanpur, S.: Deep neural network-based speaker embeddings for end-to-end speaker verification. In: Proceedings of the 2016 IEEE Spoken Language Technology Workshop (SLT), pp. 165–170. IEEE, San Juan, Puerto Rico (2016)
18. Vaswani, A., et al.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS), pp. 5998–6008. Curran Associates Inc., Long Beach, CA, USA (2017)
19. Villalba, J., Brümmer, N., Dehak, N.: Tied variational autoencoder backends for i-vector speaker recognition. In: Proceedings of Interspeech, pp. 1004–1008. ISCA, Stockholm, Sweden (2017)
20. Xie, J., Xu, X., Shu, L.: WT feature based emotion recognition from multi-channel physiological signals with decision fusion. In: Proceedings of the 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia), pp. 1–6. IEEE, Beijing, China (2018)

21. Xie, W., Nagrani, A., Chung, J.S., Zisserman, A.: Utterance-level aggregation for speaker recognition in the wild. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 5791–5795. IEEE, Brighton, UK (2019)
22. Yu, F., Koltun, V., Funkhouser, T.A.: Dilated residual networks. In: Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 636–644. IEEE, Honolulu, Hawaii (2017)