CARN-Conformer: Conformer in Attention Spectral Mapping Based Convolutional Recurrent Networks for Speech Enhancement

Bo Fang[1(✉)], Hongqing Liu[1], Yi Zhou[1], Yizhuo Jiang[2], and Lu Gan[2]

[1] School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing, China
s200131155@stu.cqupt.edu.cn

[2] College of Engineering, Design and Physical Science, Brunel University, London UB8 3PH, UK

Abstract. In recent years, the attention transformer model has been widely used in the field of speech enhancement. With the introduction of a convolutionally enhanced transformer (Conformer), it models both the local and the global information of the speech sequence to achieve a better performance. In this paper, we propose a speech enhancement structure using conformer with time-frequency (TF) domain in DCCRN. To that aim, the second layer LSTM in DCCRN is replaced with TF-Conformer. By doing this, information between and within frames can be better utilized. An attention convolution path between the convolutional encoder and decoder is also developed to better convey nonlinear information. The results show that the model's PESQ surpasses DCCRN and DCCRN+ on the testset of Interspeech 2020 Deep Noise Suppression (DNS) Challenge, with the best model size of 2.3 M. At the same time, the excellent results have been obtained on the blind test set of ICASSP 2021 DNS Challenge, and the overall MOS score exceeds the winner team by 0.06.

Keywords: Speech enhancement · Attention · Time-frequency domain

1 Introduction

Noise is widespread in the natural environments and interferes with our language communications. Speech enhancement aims at suppressing noise components from noisy speech, and to improve the quality and intelligibility of audio. An excellent speech enhancement system can help people or machines better understand the meaning of speech such as speech recognition, hearing aids, and sentiment analysis [6]. Traditional speech enhancement algorithms usually estimate the noise spectrum theoretically based on statistical signals, and use filters to suppress the noise. The rise of deep learning has turned the speech enhancement into a

F. Gao et al. (Eds.): ChinaCom 2022, LNICST 500, pp. 273–282, 2023.
https://doi.org/10.1007/978-3-031-34790-0_21

data-driven supervised learning task. In recent years, DNN-based speech enhancement [18] methods have achieved great results in noise suppression. However, in practical applications, it is still difficult to suppress the noise in the case of low signal-to-noise ratios (SNR). The ICASSP 2021 deep noise suppression (DNS) challenge [12] is dedicated to speech enhancement task in harsh environments, and the corresponding dataset and evaluation metric (DNSMOS) [10] are provided for participants. For single-channel speech enhancement, the algorithms can be mainly divided into time domain [7,8] and time-frequency domain [21] structures. The time-domain algorithms mainly use convolutional encoder and transposed convolutional decoder to simulate the Fourier and inverse transforms, and directly estimate the clean target speech sequence through an end-to-end data-driven concept. Although it is not necessary to estimate the phase in the time domain, it is difficult to model the long speech sequences. The time-frequency domain algorithms extract the complex spectrum or power spectrum of the target speech from the noisy speech. After the short-time Fourier transform (STFT), the spectrums are processed to remove the noise in frequency domain.

In terms of the training goal, there are two common methods in deep learning: direct mapping and mask estimation. Due to the limited dynamic ranges of mask as a training target, the convergence speed is very fast, which is usually desired. The common masks mainly include ideal ratio mask (IRM) [11], ideal binary mask (IBM) [17], and target magnitude spectrum (TMS) [18]. However, many of these methods ignore the phase information of speech, and ambiguous phase information can make the modeling of recovered speech difficult [19]. The recent studies have shown that phase ratio mask (PSM) [2] and complex ratio mask (CRM) [20] achieved a good performance in phase estimation. The CRM respectively acts on the real and imaginary parts of the spectrogram, which estimates the phase information more accurately and improves the speech quality. With this concept and using complex operations, DCCRN [4] combines the advantages of two networks of DCUNET [1] and CRN [13], and develops a new complex speech enhancement network [14], where LSTM layers between convolutional encoder and decoder model the temporal context to reconstruct speech by simultaneously enhancing the real and imaginary parts of the speech in frequency domain.

In this paper, we propose a complex convolutional network [4], which uses DCCRN as a backbone structure. The difference is the conformer [3] attention mechanism, we use Multihead-Attention in frequency domain to process the relevant information between frequencies in the frame. In addition to that, we use the skip connection attention mechanism of attention convolution to improve the information aggregation between the encoder and decoder, which is also different from DCCRN+ [9]. To train the network, the weighted loss of mean square error (MSE) and scale-invariant signal-to-noise ratio (SI-SNR) [8] is developed to better balance the speech distortion and noise suppression. Under the framework of the proposed model, we perform comparisons on the DNS test sets, and it is found that the proposed model outperforms DCCRN in all scenarios, with a significantly less computation. Under the P808 subjective evaluation system, the proposed model outperforms DCCRN and TSCN-PP [5].

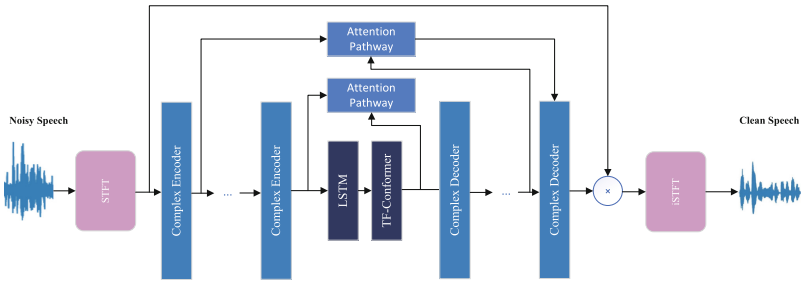


Fig. 1. CARN-Conformer Network

2 The CARN-Conformer Network

As shown in Fig. 1, our time-frequency domain network adopts a multi-layer convolution structure, where the complex convolution structure is utilized. The attention pathway is added between the encoder and decoder. LSTM and one layer TF-Conformer Module is used as the middle layer. The detailed information for each layer is provided below.

2.1 Time-Frequency Conformer

The conformer is widely used in the field of speech recognition(ASR), which adds a convolution module to the original transformer. We use Conformer to extract information and intra-frame frequency correlations between temporal contexts. The TF-Conformer Module is mainly composed of a stack of five blocks, namely two feed forward modules (FFM), the multi-head self attention (MHSA) module, the convolution module, and the LayerNormlization, depicted in Fig. 2. We use $X(t, f)$ to represent the first input of TF domain features, and suppose the input of the ith block is z and the output of the ith block is

$$\hat{z} = z + \frac{1}{2}\mathrm{FFM}(z), \tag{1}$$

$$z' = \mathrm{MHSA}(\hat{z}) + \hat{z}, \tag{2}$$

$$z'' = conv(z') + z', \tag{3}$$

$$output = layernorm(z'' + \frac{1}{2}\mathrm{FFM}(z)). \tag{4}$$

The FFM follows a pre-normalized residual unit and applies layer normalization on the input within the residual unit. It is similar to a macaron-structured network and consists of a linear layer that includes a swish activation function and dropout. We improve MHSA to explore contextual information in the frequency dimension. Therefore, the frequency dimension is used as the sequence dimension, the time dimension is combined into batch size, and the channel dimension is used as the feature dimension. The reshape operators are applied before and after MHSA layer between the features of Z $\in R^{B\times C\times T\times F}$ and Z $\in$

$R^{BT\times F\times C}$, where B, C, T, F are the dimension of batch size, channel, time and frequency, BT $= B \times T$ is the combined dimension of batch size and time. The improved MHSA can be formulated as follows:

$$Q_i = Z^Q W_i^Q, K_i = Z^K W_i^K, V_i = Z^V W_i^V, \tag{5}$$

$$head_i = Attention(Q_i, K_i, V_i) = softmax(\frac{Q_i K_i^T}{\sqrt{d}})V_i, \tag{6}$$

$$MultiHead = Concat(head_1, ..., head_h)W^O, \tag{7}$$

In the formula, i ∈ [1, h] is the head index. Z^Q, Z^K, $Z^V \in R^{F\times C}$ are the input features with length F and dimension C, Q_i, K_i, $V_i \in R^{F\times C/h}$ are queries, keys and values. W^Q, W^K, $W^V \in R^{C\times C/h}$ and $W^O \in R^{C\times C}$ are parameter matrices.

The convolution module helps extract the information that the model extracts the temporal context, starting with point convolution and gated linear units, Then a 1D depthwise convolution where normalization helps the model better deep training.

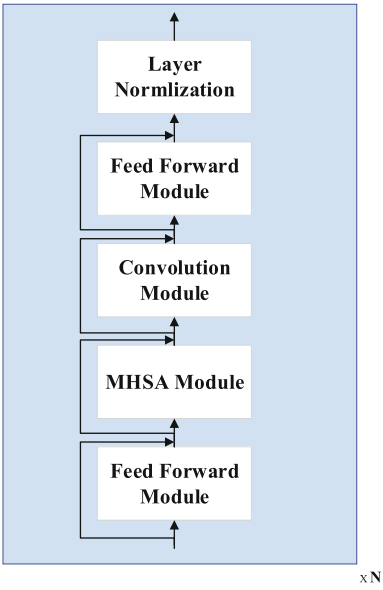


Fig. 2. TF-Conformer Module.

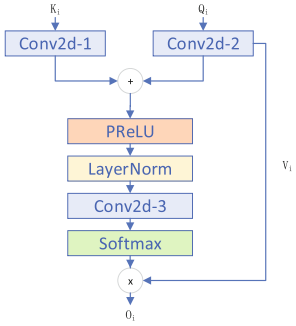


Fig. 3. Attention pathway.

2.2 Attention Pathway (AP)

In past DCCRN networks, skip connections were used to map encoder features directly to decoders. This paper proposes a new skip connection method based

on the attention mechanism, which uses the output of the encoder and decoder (conformer) of the *ith* layer as the input of the attention path, and obtains the input of the $(i-1)th$ layer. In the Fig. 3, K_i is the output of the *Conv2d* encoder and Q_i is the output of the decoder. There are two causal 2D convolutions with kernel (3, 2) and stride (1, 1) to process K_i and Q_i. The first layer Conv2d doubles the output channels to extract high-dimensional spatial features, referring as W_1 and W_2, and then sums the outputs. The output A_i of high-dimensional spatial features can be described as:

$$A_i = LayerNorm(PReLU(W_1 \cdot K_i + W_2 \cdot Q_i)) \tag{8}$$

After that, utilize the PReLU activation function and normalization layer, then recover the number of channels and generate the attention mask using causal *Conv2d* with the same kernel and stride, refering as W_3. Finally, the feature V_i of the original input to the decoder is multiplied by the attention mask element, and the resulting mask feature is concatenated with the output Q_i to get the final output to the next decoder. The output B_i of AP can be described as:

$$B_i = Softmax(PReLU(W_3 \cdot A_i) \cdot V_i \tag{9}$$

2.3 Learning Target and Loss Function

In our model, we use CRM [20] as our training target. During training, CARN-Conformer estimates CRM. Given the complex-valued STFT spectrogram of clean speech S and noisy speech Y, CRM is defined as

$$CRM = \frac{Y_r S_r + Y_i S_i}{Y_r^2 + Y_i^2} + j\frac{Y_r S_i + Y_i S_r}{Y_r^2 + Y_i^2}, \tag{10}$$

where Y_r and Y_i refer to the real and imaginary parts of the noisy spectrum, respectively, and S_r and S_i represent the real and imaginary parts of the clean complex spectrum, and M_r and M_i denote the real and imaginary parts of the CRM, respectively.

We multiply the input noisy speech $X = X_r + jX_i$ and the CRM mask to generate the enhanced spectrogram

$$\hat{S} = X_r M_r - X_i M_i + i(X_r M_i + X_i M_r). \tag{11}$$

Applying inverse Fourier transform (iSTFT) to $\hat{S}$ yields a time-domain waveform $\hat{s}$

$$\hat{s} = iSTFT(\hat{S}). \tag{12}$$

To train the network, the weighted loss functions is developed by considering both SI-SNR and MSE. The SI-SNR is defined

$$\begin{cases} s_{target} := \frac{<\hat{s}, s> \cdot s}{||s||_2^2} \\ e_{noise} := \hat{s} - s_{target} \\ L_{SI-SNR} := -10log_{10}(\frac{||s_{target}||_2^2}{||e_{noise}||_2^2}) \end{cases} \tag{13}$$

Therefore, the final weighted loss is

$$\begin{aligned} Loss = L_{SI-SNR} + log(MSE(S_r, \hat{S}_r) + MSE(S_i, \hat{S}_i) \\ + MSE(|S|, |\hat{S}|)). \end{aligned} \tag{14}$$

where $<\cdot, \cdot>$ denotes the dot product between two vectors and $||\cdot||_2$ is Euclidean norm (L2 norm). The added MSE part measures the real, imaginary, and magnitude differences between the estimated spectrum and the true spectrum. We take the logarithm of the MSE loss to ensure it is of the same order of magnitude as the negative SNR.

3 Experiments

We first evaluate the performance of our model on both Interspeech 2020 and ICASSP 20201 DNS challenge datasets [12]. For the Interspeech 2020 dataset, it contains 180 h of noise sets, which includes 150 lessons and 65,000 noise clips, and more than 500 h of clean speech, which includes 2,150 speaker audio clips, and has 80,000 RIR clips. We generated 500 h each for the unreverb and reverberated datasets with a sampling rate of 16 kHz and 10 s segments, and the SNR was set in the range of −5 to 20 dB. The ratio of training set and test set is 9:1. The generation method for the ICASSP dataset is consistent with the method for the Interspeech dataset. Additionally, we also Voice Bank+DEMAND [15] dataset to test the our model, where a total of 824 samples from 8 speakers are used.

3.1 Training Setup and Parameters

For our model, the window length and frame shift used are 32 ms and 16 ms, respectively, corresponding to an STFT length of 512, and we stack the real and imaginary parts together as the input to the network. The model is trained using the Adam optimizer with a bach size of 6. The initial learning rate is 1e−3, the learning rate is halved if the validation loss does not drop within 3 epochs, and training is stopped if the validation loss does not drop within 10 epochs. We compare several models on dataset, described as follows.

DCCRN [4]**:** The number of channels of the encoder and decoder is {32, 64, 128, 128, 256,256}. The kernel size and stride of each layer are (5, 2) and (2, 1), respectively. The middle layer uses a two-layer LSTM with 256 nodes. There is a 1024 * 256 fully connected layer after the LSTM. Each encoder processes the previous frame and the current frame. An additional future frame is processed by a later layer of the decoder, and each layer only uses the previous and current frame.

CARN-Conformer: Three models consist of ours-1, ours-2 and ours-3. The number of channels of the encoding layer is [32, 32, 64, 64, 64, 64, 64, 64]. The kernel size and stride are (5, 2) and (2, 1) respectively, and all the convolutional encoders and decoders are causal. In the middle layer of ours-1, we set up a conformer block instead of LSTM, with 16 attention heads, 64 attention dimensions,

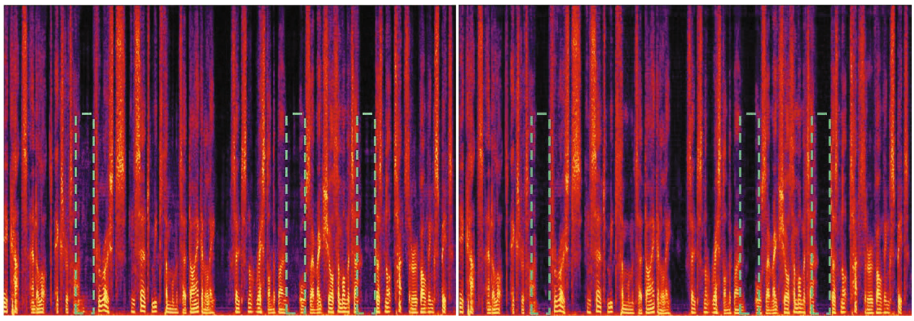
256 FFM dimensions, using relative position coding, dropout settings in FFM is 0.15. Based on ours-1, we set AP between the encoder and decoder to get ours-2. In ours-3, we replace the SI-SNR in ours-2 with our proposed constraint function.

Table 1. PESQ on DNS2020 synthetic test set.

Model	Para.(M)	no reverb	reverb	Ave.
Noisy	-	2.45	2.75	2.60
NSNet(Baseline)	1.3	3.07	2.81	2.94
DCUNET [1]	3.6	3.22	2.79	3.01
DCCRN [4]	3.7	3.26	3.20	3.23
DCCRN+ [9]	3.3	3.33	3.30	3.32
DCCRN +TF-conformer (ours-1)	2.1	3.28	3.26	3.27
+attention pathway(ours-2)	2.3	3.40	3.35	3.37
+new loss(ours-3)	2.3	**3.42**	**3.36**	**3.39**

3.2 Experimental Results and Discussions

We evaluate model performance in terms of perceptual evaluation of speech quality (PESQ) and DNSMOS [15], which is provided by the Challenge organizer. In Table 1, we perform ablation experiments on the models using the DNS-2020 synthetic test set, in terms of PESQ. It is seen that all proposed models outperform DCCRN [4] with the smaller parameters. After adding the conformer, it can be found that the parameters of the model are significantly reduced, and the noise reduction ability is improved to a certain extent. The model after adding the attention path only adds a small amount of parameters and the performance is significantly improved. It is seen that our best model outperforms DCCRN [4] by 0.16 and DCCRN+ [9] 0.07 on the test set.



(a) without AP (b) with AP

Fig. 4. The denoising result on a testing noisy clip for the cases with/without AP: (a) without AP, (b) with AP.

In Fig. 4, to visually see the performance of AP, we compare the spectrograms of a test segment without AP and with AP. We can find that the spectrogram with AP is clearer and the residual noise is better suppressed.

Table 2. DNSMOS on DNS2020 blind test set.

Model	Para.(M)	no reverb	reverb	realrec	Ave.
Noisy	-	3.13	2.64	2.83	2.85
NSNet(Baseline)	1.3	3.49	2.64	3.00	3.03
DCCRN[T1]	3.7	4.00	2.94	3.37	3.42
ours-1[T1]	2.1	4.00	3.16	3.28	3.49
ours-2[T1]	2.3	4.04	3.34	3.35	3.56
ours-3[T1]	2.3	**4.04**	**3.36**	**3.37**	**3.59**
DCCRN[T2]	3.7	3.90	2.96	3.34	3.38
ours-1[T2]	2.1	3.92	3.02	3.38	3.43
ours-2[T2]	2.3	3.97	3.14	3.40	3.50
ours-3[T2]	2.3	**3.98**	**3.16**	**3.42**	**3.52**

In Table 2, the blind test set is used to evaluate the performance in terms of DNSMOS [6], where [T1] and [T2] respectively represent the real-time track and non-real-time track. It is noted that the proposed models outperform the baselines. With attention path and new loss function, the performance of the proposed model is superior to that of DCCRN [4] in both tracks.

In Table 3, the experiments with the Voice Bank + DEMAND dataset [15] are provided. It is seen that our model has a significantly lower parameter quantity, and outperforms DCCRN [4] and DCCRN+ [9].

Table 3. PESQ on Voice Bank + DEMAND.

Model	Para.(M)	External Data	PESQ-WB
Noisy	-	-	1.97
RNNoise [16]	0.06	√	2.29
DCCRN [4]	3.7	√	2.68
DCCRN+ [4]	3.3	√	2.84
ours-3	2.3	√	**2.90**

Finally, we further compare the performances in more complex acoustic scenarios using 2021 DNS Challenge blind test set [12], and the results are provided in Table 4. The noise reduction performance of our best model outperforms the winning model of TSCN-PP [5] in all scenarios, which indicates our model generalizes better.

Table 4. DNSMOS on the DNS2021 blind test set.

Model	singing	Tinal	Non-English	English	Emotional	Overall
Noisy	2.96	3.00	2.96	2.80	2.67	2.86
NSnet2(Baseline)	3.10	3.25	3.28	3.30	2.88	3.21
TSCN-PP [5]	3.14	3.44	3.50	3.49	2.92	3.38
ours-3	**3.18**	**3.46**	**3.54**	**3.52**	**2.96**	**3.44**

4 Conclusion

In this paper, we propose a model CARN-Conformer that uses conformer and adopts attention pathway skip connections to perform speech enhancement. Conformer is used to combine local and global information for modeling and the new loss function constrains both the time and frequency domains. Especially, the attention pathway is utilized to map between the encoder and the decoder, and as a result, the ability of the model to suppress noise is improved. The proposed model achieves excellent MOS scores on the ICASSP 2021 DNS Challenge blind test set, proving the effectiveness of the model structure.

References

1. Choi, H.S., Kim, J.H., Huh, J., Kim, A., Ha, J.W., Lee, K.: Phase-aware speech enhancement with deep complex u-net. In: International Conference on Learning Representations (2018)
2. Erdogan, H., Hershey, J.R., Watanabe, S., Le Roux, J.: Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 708–712. IEEE (2015)
3. Gulati, A., et al.: Conformer: convolution-augmented transformer for speech recognition. arXiv preprint arXiv:2005.08100 (2020)
4. Hu, Y., et al.: DCCRN: deep complex convolution recurrent network for phase-aware speech enhancement. arXiv preprint arXiv:2008.00264 (2020)
5. Li, A., Liu, W., Luo, X., Zheng, C., Li, X.: ICASSP 2021 deep noise suppression challenge: Decoupling magnitude and phase optimization with a two-stage deep network. In: 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), ICASSP 2021, pp. 6628–6632. IEEE (2021)
6. Li, J., et al.: Developing far-field speaker system via teacher-student learning. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5699–5703. IEEE (2018)
7. Luo, Y., Chen, Z., Yoshioka, T.: Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation. In: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), ICASSP 2020, pp. 46–50. IEEE (2020)
8. Luo, Y., Mesgarani, N.: Conv-TasNet: surpassing ideal time-frequency magnitude masking for speech separation. IEEE/ACM Trans. Audio Speech Lang. Process. **27**(8), 1256–1266 (2019)

9. Lv, S., Hu, Y., Zhang, S., Xie, L.: DCCRN+: channel-wise subband DCCRN with SNR estimation for speech enhancement. arXiv preprint arXiv:2106.08672 (2021)
10. Naderi, B., Cutler, R.: An open source implementation of ITU-T recommendation p. 808 with validation. arXiv preprint arXiv:2005.08138 (2020)
11. Narayanan, A., Wang, D.: Ideal ratio mask estimation using deep neural networks for robust speech recognition. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7092–7096. IEEE (2013)
12. Reddy, C.K., et al.: ICASSP 2021 deep noise suppression challenge. In: 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), ICASSP 2021, pp. 6623–6627. IEEE (2021)
13. Tan, K., Wang, D.: A convolutional recurrent neural network for real-time speech enhancement. In: Interspeech, vol. 2018, pp. 3229–3233 (2018)
14. Trabelsi, C., et al.: Deep complex networks (2017). arXiv preprint arXiv:1705.09792
15. Valentini-Botinhao, C., Wang, X., Takaki, S., Yamagishi, J.: Investigating RNN-based speech enhancement methods for noise-robust text-to-speech. In: SSW, pp. 146–152 (2016)
16. Valin, J.M.: A hybrid DSP/deep learning approach to real-time full-band speech enhancement. In: 2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP), pp. 1–5. IEEE (2018)
17. Wang, D.: On ideal binary mask as the computational goal of auditory scene analysis. In: Divenyi, P. (ed.) Speech Separation by Humans and Machines, pp. 181–197. Springer, Boston (2005). https://doi.org/10.1007/0-387-22794-6_12
18. Wang, D., Chen, J.: Supervised speech separation based on deep learning: an overview. IEEE/ACM Trans. Audio Speech Lang. Process. **26**(10), 1702–1726 (2018)
19. Wang, D., Lim, J.: The unimportance of phase in speech enhancement. IEEE Trans. Acoust. Speech Signal Process. **30**(4), 679–681 (1982)
20. Williamson, D.S., Wang, Y., Wang, D.: Complex ratio masking for monaural speech separation. IEEE/ACM Trans. Audio Speech Lang. Process. **24**(3), 483–492 (2015)
21. Zhao, Y., Wang, D., Merks, I., Zhang, T.: DNN-based enhancement of noisy and reverberant speech. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6525–6529. IEEE (2016)