A Complex Neural Network Adaptive Beamforming for Multi-channel Speech Enhancement in Time Domain

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

[1] School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing, China
s190101065@stu.cqupt.edu.cn

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

Abstract. This paper presents a novel end-to-end multi-channel speech enhancement using complex time-domain operations. To that end, in time-domain, Hilbert transform is utilized to construct a complex time-domain analytic signal as the training inputs of the neural network. The proposed network system is composed of complex adaptive complex neural network beamforming and complex fully convolutional network (CNAB-CFCN). The real and imaginary parts (RI) of the clean speech analytic signal are used as training targets of the CNAB-CFCN network, and the weights of the CNAB-CFCN network are updated by calculating the scale invariant signal-to-distortion ratio (SI-SDR) loss function of the enhanced RI and clean RI. It is fundamentally different from the complex frequency domain single channel approach. The experimental results show that the proposed method demonstrates a significant improvement in end-to-end multi-channel speech enhancement scenarios.

Index Terms: End-to-end · Multi-channel · Speech enhancement · Complex operations

1 Introduction

The purpose of the speech enhancement algorithms is to suppress the background noise and to improve the quality and intelligibility of speech [1]. Recent studies show that deep learning based single-channel speech enhancement methods have achieved a great success, for example, the convolutional recurrent network (CRN) in [2] and dual-signal transformation LSTM network (DTLN) in [3] demonstrate promising results. The further studies indicate that these methods can also be applied to multi-channel speech enhancement. Due to the availability of multiple microphones, multi-channel signals contain spatial information, which can improve the system performance over single-channel speech enhancement, if utilized properly.

H. Gao et al. (Eds.): ChinaCom 2021, LNICST 433, pp. 129–139, 2022.
https://doi.org/10.1007/978-3-030-99200-2_11

In supervised learning, the techniques of estimating the time-frequency mask have become popular in both multi-channel and single-channel scenarios. In [2], intra-channel and inter-channel features are used as the input of the model to estimate phase sensitive mask (PSM) [4,5] and then applied to the reference channel of dual-channel speech. However, this method ignores the phase information and directly uses the phase from the noisy signal to reconstruct the enhanced speech, which may result in phase distortion of the enhanced speech. Several methods that utilize phase information have been proposed [5,6], but they still operate in the real number domain. This limits the upper limit of speech enhancement when the phase information estimation is not accurate enough.

To overcome the lack of properly utilizing phase information, deep complex U-net [7] that combines the advantages of deep complex network and U-net [8] is developed to process spectrogram with complex values to further improve the performance of speech enhancement. In [9], a complex number network is designed to simulate complex value operations, termed as deep complex convolution recurrent network (DCCRN), where both CNN and RNN structures handle the complex-valued operations. In DCCRN, the real and imaginary (RI) parts of the complex STFT spectrogram of the mixture are used as the input to the network, and both the amplitude and phase of the spectrogram can be reconstructed by the estimated RI. However, this is a single channel based approach and it is in frequency domain. In 2020, Wang [10] proposed a complex spectral mapping combined with minimum variance distortion-less response (MVDR) beamforming [11] for multi-channel speech enhancement approach. The enhanced signal spectrum is predicted in the neural network, and by calculating the covariance matrix of the signal and noise, the beamforming filter coefficients are produced. Without computing the covariance matrix, neural network adaptive beamforming (NAB) directly learns beamforming filters from noisy data, which avoids estimating the direction of arrival (DOA) [12], and the results demonstrate that NAB outperforms the traditional beamforming methods such as MVDR.

In this work, we propose an end-to-end multi-channel speech enhancement using complex operations that implicitly explore the phase information. To that aim, we first develop a complex neural network adaptive beamforming (CNAB) to predict complex time domain beamforming filter coefficients. It is worth noting that the coefficient will be updated according to the changes of the noisy dataset during the training process, which is different from the fixed filter in [13,14]. After that, the obtained complex beamforming filter coefficients by the CNAB are convolved with the input of each channel. The resulting signal is now single channel and to further process the signal, we develop a second time-domain complex network, called complex full convolutional network (CFCN), to predict the complex time-domain information of the enhanced speech. The proposed network is called CNAB-CFCN and results show that the proposed network demonstrates a superior performance over the current networks.

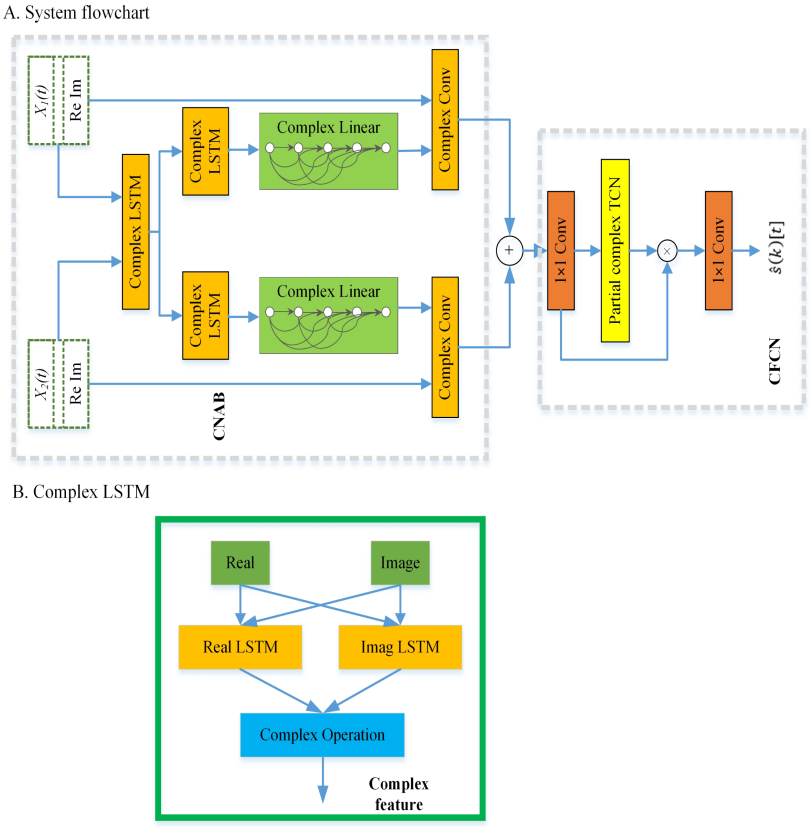

Fig. 1. (A) System flowchart. The input of each channel contains RI components. The entire system is composed of CNAB and CFCN. (B) Complex LSTM. The operation process of RI feature in complex LSTM.

2 Complex Neural Network Adaptive Beamforming

The proposed end-to-end time-domain multi-channel speech enhancement using complex value operation network framework is depicted in Fig. 1. It is of interest to point out that the flowchart provides a dual-channel description, but the extension to multi-channel is straightforward. It consists of complex neural network adaptive beamforming and a complex fully convolutional network (CNAB-CFCN). The input is complex time domain waveform, and the CNAB-CFCN model is updated by calculating scale invariant signal-to-distortion ratio (SI-SDR) loss [15].

2.1 The Formulation of Input Signal

Let $x_c(k)[t]$, $s(k)[t]$, $n_c(k)[t]$ represent noisy, clean speech, and noise, respectively, where $c \in \{0, 1\}$ is used to distinguish signals from different microphones. Note that $c = 0$ indicates the channel of the reference microphone. The relationship between them in the room is

$$x_0(k)[t] = s(k)[t] * h_0(k)[n] + n_0(k)[t], \tag{1}$$

$$x_1(k)[t] = s(k)[t] * h_1(k)[n] + n_1(k)[t], \tag{2}$$

where $t \in \{0, 1, \ldots, N-1\}$ is sample index in each frame $k \in \{0, 1, \ldots, M-1\}$, $h_0(k)[n] \in \mathbb{R}^{K\times 1}$ and $h_1(k)[n] \in \mathbb{R}^{K\times 1}$ are the room impulse responses (RIRs) corresponding to the microphones, $\mathbb{R}$ represents the set of real numbers. We choose the speech received by the reference microphone as the target source for training the network.

From (1) and (2), we only have real time domain waveform available. The important step now is to generate the complex time domain signals to prepare the input of the network. To that end, in this work, Hilbert transform is explored to construct the analytic function $x_{ac}(k)[t] \in \mathbb{C}^{N\times 1}$, where $\mathbb{C}$ represents the complex number set, given by

$$\begin{aligned} x_a[t] &= x[t] + \mathcal{H}(x[t]) \\ &= x[t] + j\hat{x}[t], \end{aligned} \tag{3}$$

where $\mathcal{H}$ indicates Hilbert operator, and we omit the channel c and the frame number k for convenience. From (3), it can be found that the imaginary part $\hat{x}[t] \in \mathbb{R}^{N\times 1}$ is obtained by Hilbert transform of $x[t] \in \mathbb{R}^{N\times 1}$. We now have both the real and imaginary time domain waveforms for the proposed network.

2.2 Complex Adaptive Spatial Filtering

The dual-channel noisy speech signal is subjected to Hilbert transform to obtain the sequence of real and imaginary parts as the input of CNAB model architecture. The purpose of CNAB is to estimate the beamforming filters, which also includes real and imaginary parts. The complex convolution of the beamforming filter coefficients and the input RI is

$$\begin{aligned} y_a[t] =& (conv_r(Re(x_a[t])) - conv_i(Im(x_a[t]))) \\ &+ j(conv_r(Im(x_a[t]) + conv_i(Re(x_a[t])), \end{aligned} \tag{4}$$

where $conv$ denotes the convolution operation, and the subscripts r and i are the real and imaginary parts of the CNAB, respectively, $y_a[t]$ is the output of the complex convolution in one channel, and $Re(\cdot)$ and $Im(\cdot)$ respectively takes the real and imaginary part of a complex signal.

Summing the results of different channels yields the final output

$$N_{out} = y_{a0}[t] + y_{a1}[t], \tag{5}$$

where $y_{a0}[t]$ and $y_{a1}[t]$ respectively represent the beamforming results of two channels, and N_{out} is the final output of CNAB.

2.3 CNAB-CFCN Architecture

In our CNAB-CFCN framework, we all use the rules of complex number operations. In Fig. 1, for the CNAB part, complex LSTM is utilized to estimate the filter coefficients, where the rule of complex LSTM is provided in Fig. 1(B). The first layer of the CNAB is a complex number LSTM, termed as complex shared-LSTM, which takes complex time domain waveforms generated by Hilbert transform as input. The next layer has two separated complex LSTMs, which process two corresponding futures of each channel, called complex splitted-LSTM. Finally, the beamforming filters are produced by complex linear activations, and the enhanced speech features are estimated through complex convolutions. The specific operations of complex LSTM is

$$L_r = LSTM_r(Re(x_a[t])) - LSTM_i(Im(x_a[t])), \tag{6}$$

$$L_i = LSTM_r(Re(x_a[t])) + LSTM_i(Re(x_a[t])), \tag{7}$$

$$L_{out} = L_r + jL_i, \tag{8}$$

where $LSTM_r$ and $LSTM_i$ are two ordinary LSTM networks, representing the real part and imaginary part of complex LSTM, respectively, $L_r \in \mathbb{R}^{N\times 1}$ and $L_i \in \mathbb{R}^{N\times 1}$ are feature mappings of real part and imaginary part. The feature mapping output $L_{out} \in \mathbb{C}^{N\times 1}$ by a complex LSTM is still a complex feature.

The output of CNAB now is a single channel time domain waveform and to further improve the system performance, we develop another complex network, called complex fully convolutional network (CFCN). In Fig. 1, for the CFCN part, 1×1 *conv* first separately processes the real and imaginary parts of N_{out}, and then the output features are stacked together. After that, in partial complex TCN, we repeat X 1-D convolution blocks with the dilated convolution factor $d = \{1, 2, 4, \ldots, 2^{X-1}\}$ for R times, and the size of kernel is P. Note that only the last 1-D convolution blocks is a complex network. The operation rules are as follows.

$$\begin{aligned} C_{out} =& (conv_r(M_r) - conv_i(M_i)) \\ &+ j(conv_r(M_i) + conv_i(M_r)), \end{aligned} \tag{9}$$

where $conv_r$ represents the feature mapping function corresponding to the real convolution layer, and $conv_i$ represents the feature mapping function corresponding to the imaginary convolution layer, M is the feature map output by the upper layer of the network, and $C_{out} \in \mathbb{C}^{T\times 1}$ is the output of the CFCN, which is also the final output of the whole network.

2.4 Loss Function

We train CNAB-CFCN to estimate the real and imaginary parts of clean speech from noisy speech, and the weighted complex SI-SDR as a loss function is utilized to train our model, given by

$$\begin{aligned} \text{SI-SDR} = (1-\lambda) \times 10\log_{10} \frac{\|\beta_r \times Re(s_a[t])\|^2}{\|\beta_r \times Re(s_a[t]) - \hat{R}_t\|^2} \\ + \lambda \times 10\log_{10} \frac{\|\beta_i \times Im(s_a[t])\|^2}{\|\beta_i \times Im(s_a[t]) - \hat{I}_t\|^2}, \end{aligned} \tag{10}$$

$$\beta_r = \frac{\hat{R}_t^T Re(s_a(t))}{\|s_a[t]\|^2} = \arg\min_{\beta_r} \|\beta_r \times Re(s_a(t)) - \hat{R}_t\|^2, \tag{11}$$

$$\beta_i = \frac{\hat{I}_t^T Im(s_a(t))}{\|s_a[t]\|^2} = \arg\min_{\beta_i} \|\beta_i \times Im(s_a(t)) - \hat{I}_t\|^2, \tag{12}$$

where $\hat{R}_t \in \mathbb{R}^{N\times 1}$ and $\hat{I}_t \in \mathbb{R}^{N\times 1}$ indicate the real and imaginary parts of each frame estimated by the CNAB-CFCN model, $s_a[t]$ is the analytical signal of the clean speech from the reference channel, λ is a weighting constant in the range of $[0, 1]$. When $\lambda = 0$, the network only uses the real part information to update the network parameters, whereas when $\lambda = 1$ means that the network uses the imaginary part information to update the parameters. The experiments indicate that $\lambda = 0.5$ is a good empirical hyperparameter in this work.

Table 1. CNAB-CFCN model parameter configuration, where B is the batch size and L is the length of the feature mapping.

Layer name	Input size	Hyperparameters
Complex input	(B, 2, 16000) × 2	–
Complex sh-LSTM	(B, 2, 100, 160)	(160, 512)
Complex sp-LSTM × 2	(B, 2, 512)	(512, 256)
Complex linear	(B, 2, 256)	(256, 25)
N_{out}	(B, 2, 16000)	–
1 × 1 Conv	(B, 16000) × 2	(1, 256), 40, 20
Layernorm	(B, 2, 256, L)	BatchNorm2d
Complex 1 × 1 Conv	(B, 2, 256, L)	(1, 1), (5, 2), (2, 1)
1-D Conv × 23	(B, 256, L)	...
Complex 1-D Conv	(B, 128, L)	...
1 × 1 Conv	(B, 256, L)	(256, 512), 1, 1

3 Experimental Setup

To generate dual-channel speech, image method [16] was used. The azimuth of the target source is on the same horizontal line as the two microphones and close to the reference microphone. We define this angle as 0°. The noise direction angle is uniformly distributed between 0 and 90° at a 15° interval. The space between the two microphones is 3 cm and the clean speech and noise are 1 m

away from the microphone. The simulated reverberation room size is 10 × 7 × 3 m. The dual-channel clean and dual-channel noise generated by the above configuration are randomly mixed from −5 to 10 dB, and the signal-to-noise ratio (SNR) interval is 1 dB in the training set and the verification set. Note that noise and clean are also randomly mixed at different directions. The SNR of speech-noise mixtures include {−5, 0, 5, 10, 20} dB in the test set.

3.1 Production of Dual-Channel Dataset

In this section, we use the dataset provided by deep noise suppression (DNS) challenge [17] to train and evaluate our speech enhancement model, where all audio clips are 16 kHz. We use the script provided by the DNS to generate 75 h audio clips, and the size of each clip is 6 s long. In total, we generate 40400 clips for training set and 4600 clips for validation set. In addition, we generate 3500 clips for testing, and the speech and noise did not appear in the training set.

Table 2. Number of complex 1-D convolution blocks.

No.	0	3	6	9	12
PESQ	3.257	3.291	2.501	2.497	2.506

3.2 Experimental Setting

In this study, we divide 16 kHz speech signal with a duration of 6 s into 1 s segments, and each segment contains 16000 sampling points. The Hilbert transform is performed on the input speech segments to create complex time-domain signals. Table 1 summarizes the parameter configurations of the model used, where the input feature is dual-channel, and sh-LSTM and sp-LSTM are short for shared-LSTM and split-LSTM, respectively. The format of hyperparameter $LSTM$ is *input* and *output channels*, and the format of $Conv$ is *input* and *output channels*, *kernelsize*, and *stride*. The structure of 1-D convolution blocks refers to [18]. A complex 1-D convolution block is composed of two 1-D convolution blocks, representing real 1-D Conv and imaginary 1-D conv, respectively. The input channel size of the complex 1-D convolution block structure is half of 1-D convolution blocks. To verify the effect of the number of complex 1-D convolution blocks, in Table 2, the PESQs of the proposed approach versus number of complex Conv are provided. It is found that increasing the number of complex 1-D Conv does not always improve the performance of the model, and at the same time, it will increase the amount of model parameters by using more complex Convs. Therefore, in current work, three complex 1-D Conv are utilized and remaining Convs are still real.

Table 3. PESQ and STOI (%) on the simulated DNS dataset.

Model	Mics	Para (M)	PESQ					STOI (%)				
			−5 dB	0 dB	5 dB	10 dB	20 dB	−5 dB	0 dB	5 dB	10 dB	20 dB
Noisy	–	–	1.37	1.67	1.98	2.32	2.99	66.91	75.57	83.26	89.32	96.43
C-TasNet	1	5.1	2.39	2.75	3.02	3.27	3.65	81.38	88.95	93.17	95.48	98.32
MFMVDR	1	5.3	2.45	2.83	3.08	3.35	3.70	84.73	90.52	93.34	96.25	98.43
CRN-i	2	0.08	1.56	1.89	2.28	2.63	3.24	69.51	78.68	86.27	91.77	97.36
CRN-ii	2	17.6	1.61	1.96	2.33	2.66	3.30	71.06	79.85	87.15	92.30	97.93
Prop.	2	9.2	**2.75**	**3.07**	**3.32**	**3.52**	**3.81**	**89.22**	**93.29**	**95.74**	**97.27**	**98.81**

3.3 Training Baseline and Training Results

For comparisons, we reproduce CRN [2], Conv-TasNet (C-TasNet) [18], and MFMVDR [19] based on our dataset for training and testing. Conv-TasNet is a single-channel speech enhancement, and one channel of our noisy datasets is used for training and testing. MFMVDR is trained according to the open source project provided in [19]. Since MFMVDR is also a single-channel speech enhancement, the datasets are the same as C-TasNet. The parameters of CRN-i are selected according to [2], while CRN-ii is the parameter we tuned based on the dataset with reference to [20]. The experimental results are provided in Table 3, where the best results are highlighted with bold numbers. Compared with the single-channel speech enhancement MFMVDR, our proposed method has a significant improvement, which shows the benefits of multiple channels. Compared with the dual-channel CRN model, the proposed method also demonstrates a superior performance, regardless the sizes of the CRN model.

Table 4. The influence of interference sources in different directions on PESQ.

Method	Interference direction					
	15°	30°	45°	60°	75°	90°
Noisy	2.07	2.01	1.98	1.97	1.96	1.95
MVDR	2.03	2.02	2.03	2.12	2.21	2.27
Prop.	**2.97**	**3.00**	**3.15**	**3.39**	**4.11**	**3.37**

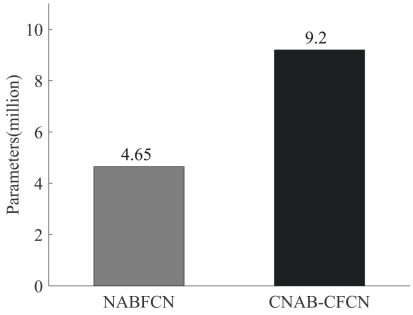

Fig. 2. The number of trainable parameters (unit: million).

Table 5. The influence of complex time domain network on speech quality in terms of PESQ and STOI.

Model	Metrics	SNR				
		−5 dB	0 dB	5 dB	10 dB	20 dB
NABFCN	PESQ	2.65	2.97	3.21	3.42	3.74
	STOI (%)	87.44	92.14	94.98	96.77	98.62
CNAB-CFCN	PESQ	**2.75**	**3.07**	**3.32**	**3.52**	**3.81**
	STOI (%)	**89.22**	**93.29**	**95.74**	**97.27**	**98.81**

In Table 4, we also analyze the effect of interference sources on the performance of CNAB-CFCN model at different azimuths. It indicates that the denoising performance is different at different directions, which agrees with the concept of the traditional beamforming. However, compared with traditional beamforming MVDR, the proposed neural beamforming indeed produces a better performance.

Finally, we study the benefits brought by complex operations. The NABFCN model is the same structure as the proposed CNAB-CFCN, but with real operations. The model sizes of NABFCN and CNAB-CFCN are shown in Fig. 2. Under the same scenarios, in Table 5, it is seen that the proposed complex network outperforms its corresponding real one across all the input SNRs. Due to the complex operations, the CNAB-CFCN also increases the model size over the NABFCN, which needs further compression in real-time applications.

4 Conclusions

In this study, we propose a novel complex time domain means to perform an end-to-end multi-channel speech enhancement network, termed as CNAB-CFCN. The Hilbert transform is explored to generate complex time-domain waveforms. With the introductions of complex operation rules, the proposed model outperforms its corresponding real network, and other single- and dual-channel networks. In addition, the loss function SI-SDR considers both the real part and

imaginary part of the speech waveform to balance the speech quality. In the experiments, we only demonstrated the performance of the proposed network in the case of dual-channel, but the extension to more than two channels is straightforward, which is our future work.

References

1. Xia, Y., Braun, S., Reddy, C.K.A., Dubey, H., Cutler, R., Tashev, I.: Weighted speech distortion losses for neural-network-based real-time speech enhancement. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 871–875. IEEE (2020)
2. Tan, K., Zhang, X., Wang, D.L.: Real-time speech enhancement using an efficient convolutional recurrent network for dual-microphone mobile phones in close-talk scenarios. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5751–5755. IEEE (2019)
3. Westhausen, N.L., Meyer, B.T.: Dual-signal transformation LSTM network for real-time noise suppression. arXiv preprint arXiv:2005.07551 (2020)
4. Erdogan, H., Hershey, J.R., Watanabe, S., Roux, J.L.: 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)
5. Wang, Y., Wang, D.L.: A deep neural network for time-domain signal reconstruction. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4390–4394. IEEE (2015)
6. Liu, Y., Zhang, H., Zhang, X., Yang, L.: Supervised speech enhancement with real spectrum approximation. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5746–5750. IEEE (2019)
7. 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)
8. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
9. Hu, Y., et al.: DCCRN: deep complex convolution recurrent network for phase-aware speech enhancement. arXiv preprint arXiv:2008.00264 (2020)
10. Wang, Z.-Q., Wang, P., Wang, D.L.: Complex spectral mapping for single-and multi-channel speech enhancement and robust ASR. IEEE/ACM Trans. Audio Speech Lang. Process. **28**, 1778–1787 (2020)
11. Zhang, J., Chepuri, S.P., Hendriks, R.C., Heusdens, R.: Microphone subset selection for MVDR beamformer based noise reduction. IEEE/ACM Trans. Audio Speech Lang. Process. **26**(3), 550–563 (2017)
12. Benesty, J., Chen, J., Huang, Y.: Microphone Array Signal Processing, vol. 1. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-78612-2
13. Hoshen, Y., Weiss, R.J., Wilson, K.W.: Speech acoustic modeling from raw multichannel waveforms. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4624–4628. IEEE (2015)

14. Sainath, T.N., Weiss, R.J., Wilson, K.W., Narayanan, A., Bacchiani, M., et al.: Speaker location and microphone spacing invariant acoustic modeling from raw multichannel waveforms. In: 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 30–36. IEEE (2015)
15. Kolbæk, M., Tan, Z.-H., Jensen, S.H., Jensen, J.: On loss functions for supervised monaural time-domain speech enhancement. IEEE/ACM Trans. Audio Speech Lang. Process. **28**, 825–838 (2020)
16. Allen, J.B., Berkley, D.A.: Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. **65**(4), 943–950 (1979)
17. Reddy, C.K.A., et al.: The INTERSPEECH 2020 deep noise suppression challenge: datasets, subjective testing framework, and challenge results. arXiv preprint arXiv:2005.13981 (2020)
18. 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)
19. Tammen, M., Doclo, S.: Deep multi-frame MVDR filtering for single-microphone speech enhancement. arXiv preprint arXiv:2011.10345 (2020)
20. Tan, K., Wang, D.L.: A convolutional recurrent neural network for real-time speech enhancement. In: Interspeech, pp. 3229–3233 (2018)