INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Real-time Multi-channel Speech Enhancement Based on Neural Network Masking with Attention Model

Cheng Xue[1], *Weilong Huang*[1], *Weiguang Chen*[1], *Jinwei Feng*[2]

[1]Speech Lab, Alibaba Group, China
[2]Speech Lab, Alibaba Group, USA
{xuecheng.xc, yuankai.hwl, chenweiguang.cwg, jinwei.feng}@alibaba-inc.com

Abstract

In this paper, we propose a real-time multi-channel speech enhancement method for noise reduction and dereverberation in far-field environments. The proposed method consists of two components: differential beamforming and mask estimation network. The differential beamforming is employed to suppress the interference signals from non-target directions such that a relatively clean speech can be obtained. The mask estimation network with an attention model is developed to capture the signal correlation among different channels in the feature extraction stage and enhance the feature representation that needs to be reconstructed into the target speech in the estimation mask stage. In the inference phase, the spectrum after differential beamforming is filtered by the estimated mask to obtain the final output. The spectrum after differential beamforming can provide a higher signal-to-noise ratio (SNR) than the original spectrum, so the estimated mask can more easily filter out the noise. We conducted experiments on the ConferencingSpeech2021 challenge (INTERSPEECH 2021) dataset to evaluate the proposed method. With only 2.9M parameters, the proposed method achieved competitive performance.

Index Terms: real-time, multi-channel speech enhancement, beamforming, deep neural network

1. Introduction

With the advancement of video conferencing technology, we can communicate face-to-face with people from all over the world at any time. However, in a video conference, the speech quality will be dramatically degraded by the noise and reverberation in the speaker's environment. In addition, video conference also faces the problems of far-field and complex conference room environment. Therefore, it is necessary to design a real-time and effective speech enhancement algorithm to improve the speech quality and intelligibility in the video conference.

Recently, the performance of speech enhancement methods based on deep neural networks has been significantly improved compared to methods based on conventional signal processing. The speech enhancement method based on deep neural network can be regarded as a supervised learning problem. Given the clean speech and background noise, the speech enhancement method takes the simulated noisy speech as input and the clean speech as the target. Although many speech enhancement methods based on deep neural networks have been proposed in the time domain [1, 2, 3], more work is currently focused on building speech enhancement methods in the time-frequency domain through short-time Fourier transform (STFT). Since time-frequency representations are easier to understand and analyze, there are currently many time-frequency domain speech enhancement methods that have achieved state-of-the-art performance. Previous studies [4, 5] focused on enhancing the amplitude spectrum by the neural network while reusing the noise phase spectrum. Since phase and amplitude information are both important to improve the quality and intelligibility of speech, recent methods [6, 7] consider both estimated amplitude and phase. More recently, speech enhancement methods based on U-Net [8] and convolutional recurrent network (CRN) [9] architecture combined with deep complex network (DCN) [10] were proposed to simulate complex-valued operations and achieved the best performance. Besides, Fullsubnet [11] proposed a method of fusing full-band and sub-band models to further improve speech quality and intelligibility.

Although the above methods have achieved the best performance, they all only work in the single-channel signal input method. However, in a noisy and reverberant far-field environment, the performance of monaural speech enhancement methods is limited. Recently, multi-channel speech enhancement methods based on deep neural networks have been proposed, one of which is called neural network beamformer [12, 13]. It first uses mono speech enhancement technology to estimate the time-frequency mask, thereby calculating the second-order statistics of speech and noise. Then beamformer applies the statistics to linearly combine multi-channel noisy speech to produce clean speech. In essence, the neural network beamformer still performs linear spatial filtering at each time-frequency bin, so its performance is limited by the nature of beamforming. Another type of method [14, 15] uses spatial features and spectral information to estimate the time-frequency mask. Although these methods have achieved the best performance, it is difficult to meet the requirements of real-time systems.

In this paper, we propose a real-time multi-channel speech enhancement method combining beamforming and neural network masking. The proposed method contains two major contributions. The first contribution is that we propose a complex attention model, which contains two modules: a complex channel attention module and a complex spectral attention module. By combining the U-Net architecture and the complex attention model, a complex-valued network that considers both amplitude and phase information is designed to estimate the time-frequency mask. Different from the previous complex attention model [16, 17], the complex attention model we developed has the following characteristics: *1)* The proposed complex channel attention module and complex spectral attention module are applied at different stages of the U-Net architecture. The complex channel attention module is used in the multi-channel feature extraction stage to encode the correlation between multi-channel features. The complex spectral attention module is used in the stage of estimating the time-frequency mask to select the time-frequency bin that needs to be reconstructed into the target speech. *2)* A complex-valued attention operation is introduced to achieve a balance between computation cost and performance. The second contribution is that we propose a strategy for combining beamforming and neural network masking. First,

http://dx.doi.org/10.21437/Interspeech.2021-2266

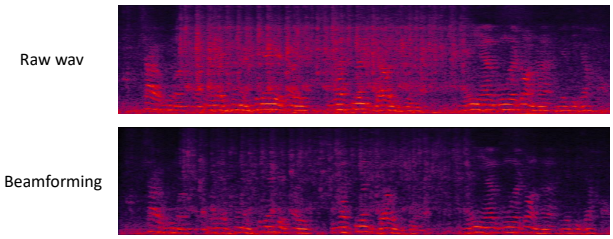

Figure 1: *Illustration of the advantage of introducing beamforming technology.*

design differential beamforming [18] to output a spectrum with a higher SNR. Then use the estimated time-frequency mask to recover the target speech on the spectrum output by the differential beamforming. The advantage of this strategy is that it allows us to suppress noise as much as possible, as shown in Figure 1. The proposed model is tested on Intel Core i5, 1.60GHz CPU with a single thread, the result of real-time factors is 0.421. With only 2.9M parameters, the model has higher computational efficiency. Experiments on the ConferencingSpeech2021 challenge dataset show that the proposed method has achieved competitive performance.

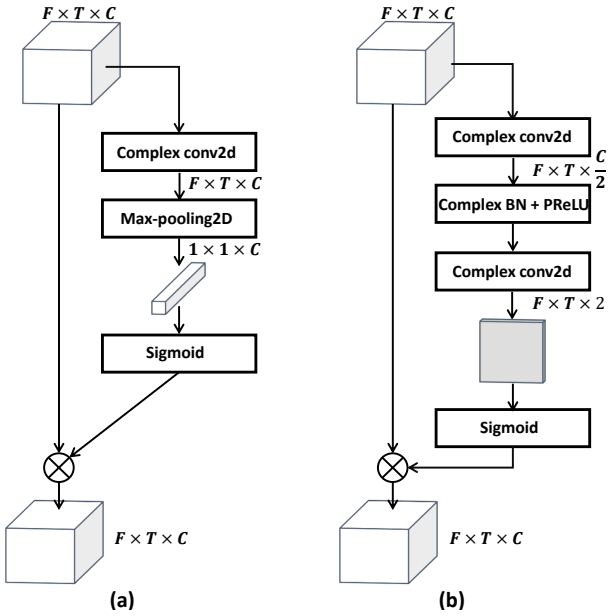

Figure 2: *Diagram of complex attention model, (a) indicates complex channel attention module and (b) indicates complex spectral attention module. '⊗' represents complex multiplication*

2. Method

2.1. Complex Attention Model

The mask attention model is based on previous work [16, 17], and it contains two modules: complex channel attention module and complex spectral attention module [1] The attention mechanism consists of a squeeze operation and an excitation operation. The squeeze operation is used for learning the effective feature representation and aggregating it. The excitation operation uses the aggregated feature representations to generate attention weights to determine which feature representations need to be enhanced.

Denote the input complex-valued feature map as $E = E_r + jE_i \in \mathbb{C}^{F\times T\times C}$, where F, T and C are the number of frequency bins, time frames and channel features, respectively. For the complex channel attention module, in the stage of extracting multi-channel input features, it is designed to capture the correlation between multi-channel features and enhance the multi-channel feature representation, as shown in Figure 2(a). It first uses a complex-valued convolution operation and a 2-D max-pooling operation to encode the correlation between channel features, which is regarded as a squeeze operation. Denote the complex-valued convolution filter as $W = W_r + jW_i \in \mathbb{C}^{F\times T\times C}$, where W_r and W_i represent the real and imaginary part of a complex convolution kernel, respectively. The complex output $P \in \mathbb{C}^{1\times 1\times C}$ of the squeeze operation can be obtained as:

$$\begin{aligned} P_r &= MaxPool(E_r * W_r - E_i * W_i) \\ P_i &= MaxPool(E_r * W_i + E_i * W_r) \end{aligned} \tag{1}$$

where '*' represents real-valued convolutional operation. After the squeeze operation, P is used to generate the attention weight, and the attention weight is multiplied by the input feature to enhance the effective channel feature representation in the input feature. Finally, the complex output $Q \in \mathbb{C}^{F\times T\times C}$ of the complex channel attention model can be obtained:

$$\begin{aligned} Q_r &= \sigma(P_r) * E_r - \sigma(P_i) * E_i \\ Q_i &= \sigma(P_i) * E_r - \sigma(P_r) * E_i \end{aligned} \tag{2}$$

where $\sigma(\cdot)$ represents sigmoid function.

The complex spectral attention module is designed to focus on the time-frequency bins that need to be reconstructed into the target speech at the spectrum level, as shown in Figure 2(b). Compared with the channel attention mechanism focus on channel feature representation, spectral attention is focused on the feature representation of the time-frequency bin. The complex output $G \in \mathbb{C}^{F\times T\times 2}$ of the squeeze operation in the complex spectral attention module can be formulated as:

$$\begin{aligned} G_r &= E_r * \bar{W}_r - E_i * \bar{W}_i \\ G_i &= E_r * \bar{W}_i + E_i * \bar{W}_r \end{aligned} \tag{3}$$

Where $\bar{W}$ represents two complex convolution filters, and the number of the filter are $C/2$, 2, respectively. Then the complex output $U \in \mathbb{C}^{F\times T\times C}$ of the complex spectral attention module can be obtained as:

$$\begin{aligned} U_r &= \sigma(G_r) * E_r - \sigma(G_i) * E_i \\ U_i &= \sigma(G_i) * E_r - \sigma(G_r) * E_i \end{aligned} \tag{4}$$

It is worth noting that the squeeze operation of the complex channel attention module is implemented using the 2-D max-pooling operation in the time and frequency dimensions, while the squeeze operation of the complex spectral attention module is implemented using the complex convolution operation in the channel dimension. The more specific operation of the complex attention model is shown in Figure 2.

[1] For ease of understanding, spatial attention is named spectral attention.

2.2. Mask Estimation Structure

The mask estimation network is based on Deep Complex Convolution Recurrent Network (DCCRN) [19] and the proposed complex attention model, as shown in Figure 3. DCCRN is a variant of the U-Net architecture, which consists of three components: encoder, skip connection and decoder. For multi-channel input, we combine the encoder and complex channel attention module to learn the correlation between channel features. For mask estimation, we insert the complex spectral attention module into the skip connection and decoder to enhance the feature representation that needs to be reconstructed into the target speech on the spectrum. Recent studies [17] have shown that using the tanh activation function to make unbounded complex-valued masks bounded can smooth the training process. Therefore, we add the tanh function to estimate the complex mask component. In order to make full use of the advantages of DCCRN, the mask estimation network only outputs a single-channel complex-valued mask.

2.3. The Full Speech Enhancement Method

The proposed multi-channel speech enhancement method is illustrated in Figure 4. It consists of two components: differential beamforming and mask estimation network. The noisy speech signal received by microphone c at discrete time n can be recorded as:

$$y^c(n) = g_s^c(n) * s(n) + g_d^c(n) * d(n) \tag{5}$$

where $\boldsymbol{y(n)} = [y^1(n), ..., y^C(n)] \in \mathbb{R}^{N\times C}$, $c \in \{1, ..., C\}$, $g_s^c(n)$ and $g_d^c(n)$ respectively represent the room impulse response of the target sound source and noise source to the c-th microphone, $s(n)$ and $d(n)$ represent the clean speech and noise, respectively. The goal of the multi-channel speech enhancement method is to estimate $s(n)$ from $\boldsymbol{y(n)}$. Let $Y \in \mathbb{C}^{F\times T\times C}$, $S \in \mathbb{C}^{F\times T}$ denote the STFT representation of $\boldsymbol{y(n)}$ and $s(n)$, respectively. We select the signal received by the first microphone as the reference signal, denoted as $Y^{ref} \in \mathbb{C}^{F\times T}$, and denote the estimated complex-valued mask as $\hat{M} = \hat{M}_r + j\hat{M}_i \in \mathbb{C}^{F\times T}$. In the training phase, the estimated clean speech spectrum $\hat{S} \in \mathbb{C}^{F\times T}$ can be obtained as follows:

$$\hat{S} = (Y_r^{ref} \cdot \hat{M}_r - Y_i^{ref} \cdot \hat{M}_i) + j(Y_r^{ref} \cdot \hat{M}_i + Y_i^{ref} \cdot \hat{M}_r) \tag{6}$$

The estimated clean speech $\hat{s}(n)$ in the time domain can be obtained by applying inverse STFT on $\hat{S}(t, f)$. The complex ratio masking (CRM) [7] method is designed to estimate the complex-valued mask $M = M_r + jM_i \in \mathbb{C}^{F\times T}$, so that the clean speech $S = M \otimes Y^{ref}$ can be obtained, where '$\otimes$' represents complex multiplication. The ground-truth CRM can be obtained as follows:

$$M = \frac{Y_r^{ref} \cdot S_r + Y_i^{ref} \cdot S_i}{(Y_r^{ref})^2 + (Y_i^{ref})^2} + j\frac{Y_r^{ref} \cdot S_i - Y_i^{ref} \cdot S_r}{(Y_r^{ref})^2 + (Y_i^{ref})^2} \tag{7}$$

In the training phase, we expect the estimated mask to retain the speech component while filtering out the noise component. In addition, we also expect to minimize the difference between the output speech and the target speech. Therefore, we use the MSE loss function to force the estimated mask to approximate the ground-truth mask. The mask loss function is defined as follows:

$$\mathcal{L}_{\text{Mask}}(M, \hat{M}) = \sum_{t,f}[(\hat{M}_r - M_r)^2 + (\hat{M}_i - M_i)^2] \tag{8}$$

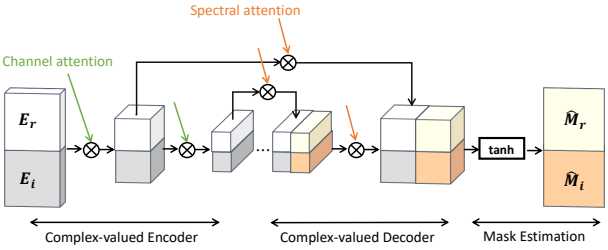

Figure 3: *Illustration of the mask estimation network.*

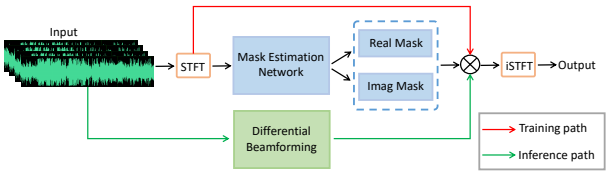

Figure 4: *Illustration of the multi-channel speech enhancement method combining differential beamforming and neural network masking. The red arrow and the green arrow respectively represent the data flow that only exists in the training phase and the data flow that only exists in the inference phase.*

For the estimated speech, the SI-SNR loss function [19] is used to approximate the target speech in the time domain. The SI-SNR loss function is defined as:

$$\begin{aligned} \mathcal{L}_{\text{SI-SNR}}(y, \hat{y}) &= 10\log10(\frac{||y_{target}||_2^2}{||v_{noise}||_2^2}) \\ y_{\text{target}} &= (<\hat{y}, y> \cdot y)/||y||_2^2 \\ v_{\text{noise}} &= \hat{y} - y_{target} \end{aligned} \tag{9}$$

where $<\cdot, \cdot>$ denotes dot product and $||\cdot||_2$ is L2 norm. Finally, the total loss of the model in the training phase is defined as:

$$\mathcal{L} = \lambda\mathcal{L}_{\text{Mask}} + (1-\lambda)\mathcal{L}_{\text{SI-SNR}} \tag{10}$$

where λ is a balance factor and $\lambda = 0.5$.

In the inference phase, we propose to use differential beamforming technology to do beamforming in multiple directions, and select the beamforming output with the highest SNR as Y^{ref}. Compared with selecting the spectrum of the first channel as a reference in the training phase, using the spectrum output by the differential beamforming technology can further improve the speech quality and intelligibility.Since differential beamforming technology is essentially a linear filter, its operation is very efficient and meets real-time requirements.

3. Experiments

3.1. Datasets

We use the dataset provided by the ConferencingSpeech2021 challenge for model training and performance evaluation. The provided data is divided into three parts: training set, development test set, and evaluation test set. The data of the training set is simulated by using the specified clean speech, noise set, and Room Impulse Responses (RIR), and the training data can be generated through the official script [2]. The clean training speech is collected in 4 open-source speech datasets: AISHELL-1

[2]https://github.com/ConferencingSpeech/ConferencingSpeech2021

[20], AISHELL-3 [21], VCTK [22], Librispeech(train-clean-360) [23], and only the speech utterances with SNR larger than 15dB are selected for training. The total duration of clean training speech is around 550 hours. The noise set consists of two parts: one part is collected from the open-source speech datasets MUSAN [24] and Audioset [25], and the other part is from real recordings. The duration of the noise set is around 120 hours. The RIR is simulated by the Image method, and there are about 20,000 RIRs in total. The development test set and evaluation test set are used to evaluate the performance of the proposed method. For more details on these two test sets, please refer to [26]. All audio files are resampled to 16k Hz.

3.2. Training setup

For the proposed method, the window length and hop size are 20ms and 10ms, and the FFT length is 512. Following [19], the STFT and inverse STFT operations are implemented by using 1-dimensional convolution and deconvolution layers. For the DCCRN built in the mask estimation network, we use 4 complex convolutional layers in the encoder and 4 complex deconvolutional layers in the decoder. The number of output channels for the 4 complex convolution operations is set to 32, 64, 128, and 128. The number of output channels for the 4 complex deconvolution operations is set to 128, 128, 64, and 32 in turn. Both the convolution filter size and the deconvolution filter size are set to 5×2, and the stride size is set to 2×1. For the parameter settings of the complex attention model, please refer to the settings in Figure 2. We use Pytorch to train the neural network and use the Adam optimizer. The initial learning rate is set to 0.001, and it decays by 0.5 when the validation score does not increase.

Table 1: *The result of objective evaluation on development test set. Para. represents the size of model parameters.*

Method	Para.(M)	PESQ
Noisy	-	1.97
Fullsubnet	5.6	2.09
The proposed	2.9	2.15

Table 2: *The result of subjective evaluation on evaluation test set.* Δ *represents the performance gain of the proposed method.*

Method	MOS	S-MOS	N-MOS
Noisy	2.56	2.93	3.03
The challenge baseline	3.43	3.55	3.55
The proposed	3.56	3.59	3.63
Δ	1.00	0.66	0.60

3.3. Results

To evaluate the performance of the proposed method on the development test set, we use PESQ[3] [27] to evaluate the proposed method. We choose the current best single-channel speech enhancement method Fullsubnet [11] as the baseline. To make a fair comparison, we re-trained the open-source code provided by Fullsubnet on the ConferencingSpeech2021 challenge dataset with the same hyperparameters and tested it in online mode. Table 1 shows the results of the objective measurement evaluation on the development test set. It can be seen from the results that the performance of the proposed method is better than that of Fullsubnet, which shows that the proposed method can effectively use the multi-channel feature to enhance the speech quality. Besides, it should be noted that the parameter amount of the proposed method is about half of Fullsubnet, which also reflects the high efficiency of the proposed method.

To evaluate the proposed method's performance on the evaluation test set, the organizers of the challenge use Absolute Category Ratings (ACR) to estimate the Mean Opinion Score (MOS) through the Tencent Media Subjective Evaluation platform to evaluate the proposed method. The performance of the proposed method on the evaluation test set is shown in Table 2. MOS, S-MOS, and N-MOS respectively represent a global subjective evaluation metric, a subjective evaluation metric that only focuses on speech signals, and a subjective evaluation metric that only focuses on noise signals. The above experimental results all show that the proposed method has achieved competitive performance.

To intuitively compare the speech involved in the experiment, in Figure 5, the spectrum of the original speech, the spectrum output by the differential beamforming, the spectrum output by the Fullsubnet, and the spectrum output by the proposed method are sequentially visualized. It can be seen from Figure 5 that the spectrum of the proposed method is clearer than other spectrums, especially in the place marked by the green box in the figure, the noise is obviously suppressed.

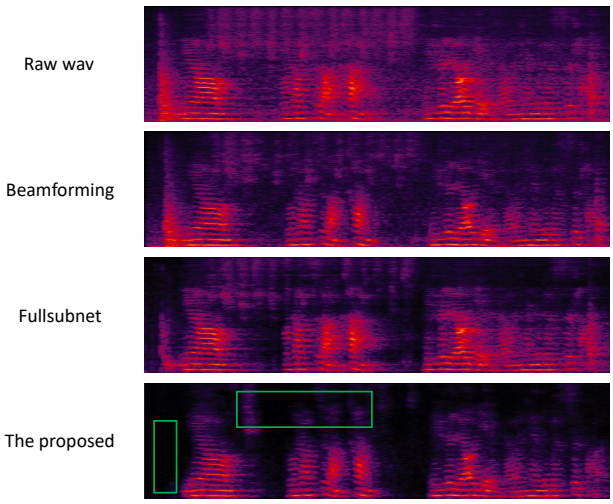

Figure 5: *Visualization of the spectrum involved in experiments.*

4. Conclusions

In this paper, we propose a real-time multi-channel speech enhancement method combining differential beamforming and mask estimation network. The estimated time-frequency mask is used to reconstruct the target speech on the high SNR time-frequency spectrum output by the differential beamforming. For mask estimation, a complex attention model was developed to enhance the feature representation of multi-channel signals and to enhance the ability of the mask to recover the target speech. The results of our experiments show that the proposed method is not only lightweight but also efficient.

[3]https://ecs.utdallas.edu/loizou/speech/software.htm

5. References

[1] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5069–5073.

[2] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1570–1584, 2018.

[3] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[4] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2014.

[5] N. Takahashi, N. Goswami, and Y. Mitsufuji, "Mmdenselstm: An efficient combination of convolutional and recurrent neural networks for audio source separation," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2018, pp. 106–110.

[6] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 708–712.

[7] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2015.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[9] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement." in *Interspeech*, 2018, pp. 3229–3233.

[10] C. Trabelsi, O. Bilaniuk, Y. Zhang, D. Serdyuk, S. Subramanian, J. F. Santos, S. Mehri, N. Rostamzadeh, Y. Bengio, and C. J. Pal, "Deep complex networks," in *International Conference on Learning Representations*, 2018.

[11] X. Hao, X. Su, R. Horaud, and X. Li, "Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," *arXiv preprint arXiv:2010.15508*, 2020.

[12] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux, "Improved mvdr beamforming using single-channel mask prediction networks." in *Interspeech*, 2016, pp. 1981–1985.

[13] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 196–200.

[14] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 1–5.

[15] S. Chakrabarty, D. Wang, and E. A. Habets, "Time-frequency masking based online speech enhancement with multi-channel data using convolutional neural networks," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2018, pp. 476–480.

[16] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[17] S. Zhao, T. H. Nguyen, and B. Ma, "Monaural speech enhancement with complex convolutional block attention module and joint time frequency losses," *arXiv preprint arXiv:2102.01993*, 2021.

[18] J. Benesty and C. Jingdong, *Study and design of differential microphone arrays*. Springer Science & Business Media, 2012, vol. 6.

[19] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," *Proc. Interspeech 2020*, pp. 2472–2476, 2020.

[20] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*. IEEE, 2017, pp. 1–5.

[21] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "Aishell-3: A multi-speaker mandarin tts corpus and the baselines," *arXiv preprint arXiv:2010.11567*, 2020.

[22] J. Yamagishi, C. Veaux, K. MacDonald *et al.*, "CSTR VCTK Corpus: English multi-speaker corpus for cstr voice cloning toolkit," *The Centre for Speech Technology Research (CSTR)*, 2016.

[23] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[24] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[25] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[26] W. Rao, L. Xie, Y. Wang, T. Yu, S. Watanabe, Z.-H. Tan, H. Bu, and S. Shang, "Conferencingspeech 2021 challenge evaluation plan," 2021.

[27] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2007.