Direction-guided Spatial Attention for Multichannel Speech Enhancement

Shuai Nie* Yaran Chen† Shan Liang† Jiaming Xu* and Runyu Shi*
* Xiaomi Corporation, China
E-mail: {nieshuai, xujiaming, shirunyu}@xiaomi.com
† Xi'an Jiaotong-Liverpool University, China
E-mail: {yaran.chen, shan.liang}@xjtlu.edu.cn

***Abstract*—The direction of target speech is a crucial clue for multichannel speech enhancement. But its real-time estimation is usually very challenging, especially when the source is moving and in far-field reverb environments. Instead of directly estimating the direction, we propose a direction-guided spatial attention to automatically focus on the target source and incorporate it into masking neural beamformers for speech enhancement. Neural beamformers are trained to learn to preserve signals from the look direction areas and suppress signals from other direction areas as much as possible. Spatial attention is guided by an extra classification objective of the target direction to focus on the target direction area and select an appropriate neural beamformer to listen to the direction area of target speech. Systematic experiments demonstrate that the proposed approach improves the performances of both speech enhancement and far-field speech recognition against prior methods.**[1]

I. Introduction

Multichannel speech enhancement has been demonstrated to significantly improve the perceptual quality and intelligibility of speech [1]–[3] and be very beneficial for far-field speech recognition[3]. However the target speech enhancement are still very challenging when the direction is unknown or the target source is moving. Although there are many methods that do not need to know the direction of target speech in advance, such as the minimum variance distortionless response (MVDR) and the parameterized multichannel non-causal Wiener filter (PMWF)[1], their performances rely heavily on complex covariance matrix estimation and its inverse operation for each frequency[4], which is very difficult in far-field environments and time-consuming.

The direction of target speech is a crucial clue that can be exploited to significantly improve the performance of multichannel speech enhancement. On the one hand, directional beamforming technology is capable of enhancing signals from the target direction and suppressing signals from other directions[5]. And it has been demonstrated to effectively avoid speech distortion and benefits for far-field speech recognition[6]. On the other hand, when the taget direction is known, many directional information can be exploited for speech enhancement, such as directional feature[7], directional power ratio (DPR) and directional signal-to-noise ratio (DSNR)[8], which have been widely leveraged for deep learning-based mask prediction and significantly improve the performance of speech enhancement. Therefore, the direction of arrival (DOA) estimation is usually regarded as an indispensable component in many speech enhancement systems. These systems usually use a clip of signals to estimate the DOA of target speech before speech enhancement, such as the audio clip of wake-up word. But the real-time estimation of DOA is very difficult, especially when the source is moving and in a far-field noisy environment, and the estimation errors may significantly degrade the performance of speech enhancement.

Attention mechanism is a promising scheme to automatically focus on the target direction where the direction is unknown or the target source is moving. In previous works[9], spatial attention was lack of guidance, resulting in attention instability. In this paper, we propose a direction-guided spatial attention mechanism to exploit direction information of target speech. Spatial attention is guided by a classification objective between attention weights and labeled target direction areas and also considers its temporal correlation. An encoder-decoder masking neural beamformers is further proposed for speech enhancement. The features extracted from each predefined direction area are firstly fed into the shared encoder for high-level feature representations. The high-level feature representations of multiple direction areas are aggregated into a vector by the target direction-guided spatial attention. The aggregated vector is then fed into a decoder for mask estimation. Following the decoder, several neural beamformers directing to the predefined directional areas are constructed by complex block affine layers to perform beamforming operations. According to attention weights, the neural beamformer directing to the target direction area is selected to perform beamforming on the multichannel observed signals. Finally, the estimated mask and the beamformed signal are fed into a elementwise multiplication operator to obtain the final enhanced signal. Through data-driven supervised training, spatial attention, encoder-decoder mask prediction network and neural beamformer are jointly optimized by the spectrum and waveform approximation objective in an end-to-end manner.

II. Proposed architecture

Fig 1 illustrates a block diagram of the proposed spatial attention-based masking neural beamformer, which is composed of feature extractor, spatial attention-based encoder-decoder mask prediction network and neural beamformer. The multi-channel time-domain signals are firstly windowed and then fed

[1]Corresponding Author: Yaran Chen

979-8-3315-7206-8/25/$31.00

into short-time Fourier transform (STFT) layer to obtain their time-frequency (T-F) representations. For a linear microphone array, we uniformly divide the whole area into 5 direction areas with center angles of $\Omega = \{0^o, 45^o, 90^o, 135^o, 180^o\}$. For each direction area $\theta \in \Omega$, the feature extractor extracts the spectral and spatial features using directional beamformers desinged in advance. The spectral and spatial features are concatenated into the shared TDNN encoder to further refine and encode the high-level feature representation of each direction area, respectively. Spatial attention aggregates the high-level feature representations of multiple direction areas into a feature vector that is then fed into LSTM-based decoder for mask prediction. Neural beamformers directing to different direction areas follows the decoder. The neural beamformer corresponding to the maximum attention weight is selected to listen to the directional areas of target speech. Neural beamformers are implemented by complex block affine layers to enhance the signal from target directional areas by performing beamforming operation on the multichannel observed signals. Finally, the estimated mask and the beamformed signal are fed into a elementwise multiplication operator to obtain the final enhanced signal. Both mask prediction network and neural beamformers are learnable and jointly optimized by the spectrum and waveform approximation objective in an end-to-end manner. In addition, we also explore a cross-entropy classification objective with the target direction to guide spatial attention to better focus on the target source.

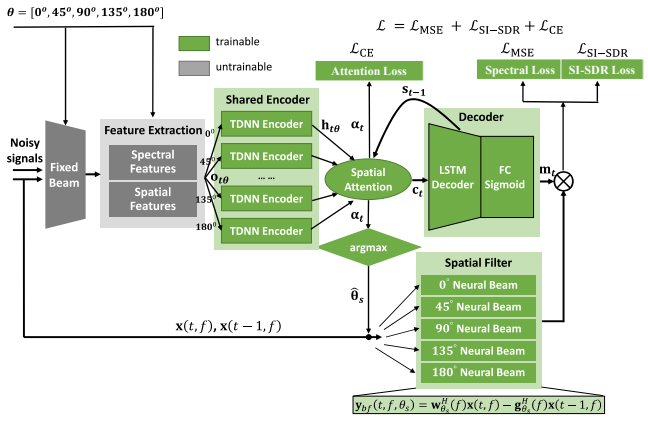

Fig. 1. Spatial attention-based masking neural beamformer.

A. Feature extraction

When multiple microphones are available, both spectral and spatial features are available to provide complementary discriminations for speech enhancement. Fixed beamformers are well known to have good spatial enhancement and directional discrimination capacity, which can be used to exploit spectral and spatial features. We use the logarithm power spectrum (LPS) of beamformed signal as the spectral feature and explore several spatial features based on fixed beamforming, including DPR[8], DSNR[8] and directional coherence features (DCF)[10]. DPR and DSNR use the output power of multi-look fixed beamformers to design spatial features with directional discriminations. DCF is inspired by coherent-to-diffuse ratio (CDR)[11] and uses a pair of complementary beamformers with the spatial separation capacity to exploit spatial features. One beamformer focuses on suppressing the diffuse noise, while the other is expected to suppress the directive interference with a null direction, but both of them preserve the target signals coming from the given target direction. In addition, cosIPD and sinIPD[12] are also taken as the spatial features.

B. Spatial attention-based encoder-decoder network

Figure 1 illustrates an overview of the spatial attention-based encoder-decoder mask prediction network, which consists of an encoder network, a spatial attention and a decoder network. The encoder network maps the spatial and spectral features extracted from each predefined direction area into a higher-level feature representation. The spatial attention automatically aggregates the feature representations of multiple predefined direction areas into a vector. The decoder network predicts the T-F mask of desired speech using the aggregated vector.

Let's define $\mathbf{o}_{t\theta}$ as the feature vector concatenated by the spectral and spatial features for the direction area θ at the time step t. The outputs of encoder and spatial attention are defined as $\mathbf{h}_{t\theta}$ and $\mathbf{c}_t$, respectively. The mask predicted by decoder network is defined as $\mathbf{m}_t$. They are calculated as follows,

$$\mathbf{h}_{t\theta} = \text{Encoder}(\mathbf{o}_{t\theta}), \tag{1}$$

$$\mathbf{c}_t = \text{Attention}(\{\mathbf{h}_t\}_{\theta\in\Omega}, \mathbf{s}_{t-1}), \tag{2}$$

$$\mathbf{m}_t = \text{Decoder}(\mathbf{c}_t), \tag{3}$$

where $\mathbf{s}_{t-1}$ denotes the decoding state of the decoder network at time step $t-1$.

The spatial attention calculates the attention weight of each direction area in Ω, which is used as a weighted summation coefficient to aggregate the encoded vector $\mathbf{h}_{t\theta}$. Due to the temporal correlation of speech, the current attention should depend on the previous decoding state. Therefore, the spatial attention is formulated as follows,

$$\mathbf{e}_{t\theta} = \mathbf{w}^{\mathrm{T}} \tanh(\mathbf{U}\mathbf{s}_{t-1} + \mathbf{V}\mathbf{h}_{t\theta} + \mathbf{b}), \tag{4}$$

$$\alpha_{t\theta} = \exp(\beta \mathbf{e}_{t\theta}) \Big/ \sum_{\theta\in\mathbf{\Omega}} \exp(\beta \mathbf{e}_{t\theta}), \tag{5}$$

$$\mathbf{c}_t = \sum_{\theta\in\Omega} \alpha_{t\theta}\mathbf{h}_{t\theta}, \tag{6}$$

$$\hat{\theta}_s = \underset{\theta}{\text{argmax}}(\,\alpha_{t\theta}), \tag{7}$$

where $\mathbf{U}$ and $\mathbf{V}$ are the parameter matrices, and $\mathbf{w}$ and $\mathbf{b}$ are the parameter vectors. $\alpha_{t\theta}$ is the attention weight for the direction area θ at time step t, which suggests the direction of target speech. We take the direction corresponding to the maximum attention weight as the estimated target direction $\hat{\theta}_s$, as shown in Eq. (7), which determines a neural beamformer to enhance the desired speech.

C. Neural beamformer

In frequency domain, beamforming operation is formulated as

$$\mathbf{y}_{bf}(f,\theta_s) = \mathbf{w}_{\theta_s}^H(f)\mathbf{x}(f), \tag{8}$$

where H is the conjugate transpose operator, and $\mathbf{w}_{\theta_s}(f)$ is the complex weight vector of the beamformer for θ_s direction. The complex vector multiplication can be converted to the real-valued matrix multiplication:

$$\begin{aligned}\mathrm{Re}(\mathbf{y}(f,\theta_s)) &= \mathrm{Re}(\mathbf{w}_{\theta_s}^H(f))\mathrm{Re}(\mathbf{x}(f)) - \mathrm{Im}(\mathbf{w}_{\theta_s}^H(f))\mathrm{Im}(\mathbf{x}(f)) \\ \mathrm{Im}(\mathbf{y}(f,\theta_s)) &= \mathrm{Re}(\mathbf{w}_{\theta_s}^H(f))\mathrm{Im}(\mathbf{x}(f)) + \mathrm{Im}(\mathbf{w}_{\theta_s}^H(f))\mathrm{Re}(\mathbf{x}(f)),\end{aligned} \tag{9}$$

where $\mathrm{Re}(\cdot)$ and $\mathrm{Im}(\cdot)$ denote the real-part and imaginary-part operators, respectively. Obviously, we can readily incorporate the beamforming operation into neural network as a learnable layer, and train it's filtering coefficients by supervised learning to preserve signals from θ_s and suppress signals from other directions as much as possible. Considering that the existing beamformers does not utilize historical information, we are inspired by Weighted Prediction Error minimization (WPE)[13] and introduce an additional filter $\mathbf{g}_{\theta_s}(f)$ that models historical signals to further reduce the residual reverberation and noise.

$$\mathbf{y}_{bf}(t,f,\theta_s) = \mathbf{w}_{\theta_s}^H(f)\mathbf{x}(t,f) - \mathbf{g}_{\theta_s}^H(f)\mathbf{x}(t-1,f), \tag{10}$$

Although the filtering coefficients of beamformer are closely related to the direction, it is not necessary to construct the beamformer for each direction. For linear microphone arrays, we build 5 neural beamformers directing to 5 predefined direction areas with center angles of $\Omega = \{0^o, 45^o, 90^o, 135^o, 180^o\}$, which can be initialized with superdirective beamformers designed in advance.

III. Optimization objectives

Although beamformer has the ability to suppress noise and reverberation, there is usually a lot of residual noise in the beamformed signal. Masking technology can significantly eliminate additive noise and is usually used as the post-processing step of speech enhancement. We apply the predicted mask $\mathbf{m}(t,f)$ to the beamformed signal to obtain the final enhanced speech signal.

$$\hat{\mathbf{y}}(t,f) = \mathbf{m}(t,f) \otimes \mathbf{y}_{bf}(t,f), \tag{11}$$

where $\otimes$ denotes an elementwise multiplication. The enhanced speech waveform $\hat{\mathbf{s}}$ is obtained by the inverse STFT. The MSE loss on the power-law compressed STFT spectrogram and scale-invariant signal-to-distortion ratio (SI-SDR)[14] between the enhanced and target speech signals are used as optimization objectives to jointly train the whole network. Notice that we take the beamformed signal on noise-free multichannel signals through the fixed beamformer as the target signal. The STFT coefficients and the waveform of target speech are denotes as $\mathbf{y}_{bf}^*$ and $\mathbf{s}^*$, respectively.

$$\mathcal{L}_{\text{MSE}} = \frac{1}{T}\sum_{t=0}^{T}\sum_{f=0}^{F}\left(\left|\mathbf{y}_{bf}^*(t,f)\right|^{0.3} - \left|\hat{\mathbf{y}}(t,f)\right|^{0.3}\right)^2, \tag{12}$$

TABLE I
THE CONFIGURES OF SIMULATION FOR DUAL-MICROPHONE.

Room size [L,W,H] (m)	$[3.0, 2.5, 2.5] \sim [9.0, 6.5, 4.0]$
T60(s)	$0.2 \sim 0.6$
Speech distance(m)	$0.5 \sim 5.5$
Interference distance(m)	$1.0 \sim 7.0$
Num of interferences	$1 \sim 3$
Angle Difference(^{o})	$30 \sim 180$
SIR(db)	$-5.0 \sim 15.0$
SNR(db)	$0.0 \sim 20$

$$\mathcal{L}_{\text{SI-SDR}} = -\text{SI-SDR} = -10\log_{10}\left(\frac{\left\|\frac{\hat{\mathbf{s}}^T\mathbf{s}^*}{\|\mathbf{s}^*\|^2}\mathbf{s}^*\right\|^2}{\left\|\frac{\hat{\mathbf{s}}^T\mathbf{s}^*}{\|\mathbf{s}^*\|^2}\mathbf{s}^* - \hat{\mathbf{s}}\right\|^2}\right), \tag{13}$$

In addition, we also consider using the direction of target speech to supervise spatial attention to focus on the target source further in training phase. According to the pre-defined direction areas Ω, the direction areas of target speech can be tagged with a label $y \in \{0,1,2,3,4\}$. A cross-entropy-based classification loss between attention weights and the direction labels is introduced into the optimization objective.

$$\mathcal{L}_{CE} = -\sum_{t=0}^{T}\ln(\boldsymbol{\alpha}_t[y]), \tag{14}$$

where $\boldsymbol{\alpha}_t$ is a vector consisting of $\alpha_{t\theta}$. Enhance, the final optimization objective function is written by

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{SI-SDR}} + \mathcal{L}_{\text{CE}}, \tag{15}$$

IV. Experiments

A. Dataset and simulation

We apply the proposed method to dual-channel speech enhancement and systematically evaluate its performances on a large-scale simulated dual-channel dataset. The inner distance of the dual-microphone is 4.0 cm. AISHELL-2[15] and AudioSet[16] are used as clean speech and interference sets, respectively. AISHELL-2 is a 1000-hour Mandarin Chinese Speech Corpus, and AudioSet consists of 632 audio event classes and 2,084,320 human-labeled 10-second sound clips drawn from YouTube videos. In addition to background noise, AudioSet also contains music, which can evaluate the ability to suppress music interferences. To simulate the realistic application scenario, we consider multiple interference sources and diffuse noise at the same time. 20,000 dual-channel nonstationary diffuse noise utterances are randomly simulated with NOISEX-92[17] through ANF-Generator[18]. Clean speech, interference and diffuse noises are randomly chosen to generate 520,000 dual-channel far-field noisy utterances by Pyroomacoustics[2], including 500,000 utterances for training, 10,000 utterances for testing and 10,000 utterances for development. For each far-field noisy utterances, its simulation parameters are randomly chosen from Table 1.

[2]https://github.com/LCAV/pyroomacoustics

B. Evaluation metrics

We take the shorttime objective intelligibility (STOI)[19] and perceptual evaluation of speech quality (PESQ)[20] as evaluation metrics. We also evaluate the ASR performance of the enhanced speech. The used ASR system uses the TDNN chain model as acoustic model that is trained on multiple condition. The chain model is built and trained according to AISHELL-2[15] recipe. 4071-hour clean speech consisting of several open-source Mandarin Chinese Speech Corpuses[15], [21]–[23] and about 6000-hour simulated far-field noisy speech by Pyroomacoustics are used to train the acoustic model. Character error rate(CER) are used as the evaluation metrics. Higher values mean the better performances for SI-SDR and STOI which are the weighted means of all testing clips weighted by their lengths.

C. Baseline

We compare the proposed approaches with two baseline systems. The baseline models use the same encoder-decoder structure and features as the proposed methods. For each direction, the extracted BFLPS, DCF, DPR, DSNR, cosIPD and sinIPD features are concatenated into the TDNN encoder. The proposed methods use a spatial attention module to aggregate the encoded outputs of the multiple directions into a feature vector, and the first baseline system uses a concatenation module to concatenate the encoded outputs of the multiple directions into a feature vector. The feature vectors are then fed into LSTM-based decoder for mask prediction. The second baseline system use the same spatial attention as the proposed methods, but in contrast to the proposed methods, it uses the multi-tap MVDR[4] as the spatial filter. The L-tap in the multi-tap MVDR is empirically set to 48. For fair comparison, we use sigmoid real-valued mask to calculate the covariance matrix of speech and noise rather than the complex mask used in the original paper[4]. This experiment is used to verify the performance of the proposed neural beamformers.

D. Network and training configurations

All models use the same network architecture containing a shared encoder and a decoder. We take two TDNN layers, each with 512 nodes, as the shared encoder. The kernel size of each tdnn layer is set to 3, and the stride, dilation and padding are set to 1. The decoder is consists of two 512-node LSTM layers, one 257-node fully-connected layer followed by a sigmoid activation function. The inner product dimension of the spatial attention is set to 128. All networks are trained from a random initialization by the Adam optimizer. The learning rate is adjusted according to warmup policy and the warmup step is empirically set to 8000. The maximum epoch is set to 24 and the batchsize is set to 16. The input features is normalized to have zero mean and unit variance over the training set. 512-point STFT is applied to time-domain signals windowed by a 512-point Hamming window for time-domain decomposition. The window size is 32 ms and the hop size is 16 ms.

E. Results and discussions

Table 2 reports the performances on the PESQ and STOI for different systems and configurations. The 'fixed BF' means that the fixed beamformer designed in advance is used as the spatial filter and its weight coefficient are frozen during network training. And the 'neural BF' means the filter is constructed by Eq.(10), and its weight coefficients are randomly initialized and trained by data-driven supervised learning. The 'mtMVDR' means that multi-tap MVDR[4] is used as spatial filter for the multichannel speech enhancement. The 'masking' means that we apply the estimated mask to the beamformed signal to obtain the final enhanced speech signal.

1) Evaluation of spatial attention: We evaluate the effectiveness of the proposed spatial attention in a masking-based MVDR framework. The spectral and spatial features are extracted from each pre-defined direction area and concatenated to fed into the shared encoder. The encoded high-level features of 5 pre-defined look direction areas are concatenated by concatenation operation or aggregated by spatial attention into a vector for mask prediction. The 3-th and 4-th rows of Table 2 report their performances of speech enhancement. Compared with concatenation operation, spatial attention significantly and consistently improves the enhancement performance, which suggests that spatial attention has the capacity to select and focus on the features of right directions for mask prediction. From $5 \sim 8$ rows in Table 2, we also observe that the cross-entropy loss between attention weights and the target direction can effectively supervise spatial attention to focus on the target source and further improve the enhancement performance for both fixed and neural beamforming. Figure 2 presents an example of the spatial attention weights. We observed that the spatial attention can track the target direction quickly and stably when the target speech appears, which is due to the target direction guidance and the consideration of the temporal correlation in spatial attention.

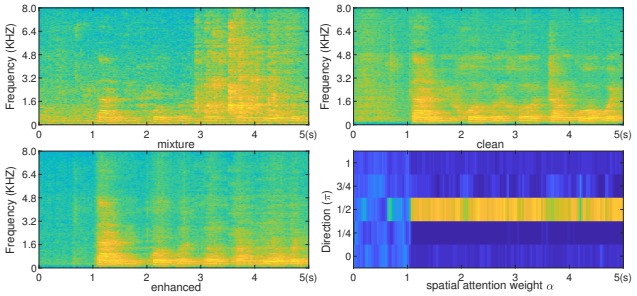

Fig. 2. An example of the spatial attention weights α. The speech is corrupted by interference signals at SIR -0.6 dB and diffuse noise at SNR 10.8 dB. The target speech is at 94^o.

2) Evaluation of neural beamformers: We compare the performances of fixed beamformer, neural beamformer and mtMVDR as spatial filters. The far-field noisy signals are firstly filtered by the spatial filters, and then the filtered signal is masked to obtain the final speech enhancement signal. The 5-th row VS 6-th row and the 7-th row VS

TABLE II
THE SPEECH ENHANCEMENT PERFORMANCES OF DIFFERENT MODELS.

Methods	Filtering	Connection	Loss	PESQ	STOI
Far-field-cln	fixed BF(reference)	–	–	–	–
Far-field-mix	–	–	–	2.02	0.67
5×(BFLPS+DCF+DPR+DSNR+cosIPD+sinIPD)	masking mtMVDR	concate	$\mathcal{L}_{\text{IRM\_MSE}}$	2.60	0.80
5×(BFLPS+DCF+DPR+DSNR+cosIPD+sinIPD)	masking mtMVDR	attention	$\mathcal{L}_{\text{IRM\_MSE}} + \mathcal{L}_{\text{CE}}$	2.68	0.81
5×(BFLPS+DCF+DPR+DSNR+cosIPD+sinIPD)	masking fixed BF	attention	$\mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{SI-SDR}}$	2.50	0.80
5×(BFLPS+DCF+DPR+DSNR+cosIPD+sinIPD)	masking neural BF	attention	$\mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{SI-SDR}}$	2.51	0.80
5×(BFLPS+DCF+DPR+DSNR+cosIPD+sinIPD)	masking fixed BF	attention	$\mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{SI-SDR}} + \mathcal{L}_{\text{CE}}$	2.80	0.84
5×(BFLPS+DCF+DPR+DSNR+cosIPD+sinIPD)	masking neural BF	attention	$\mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{SI-SDR}} + \mathcal{L}_{\text{CE}}$	**2.84**	**0.85**

TABLE III
THE ASR PERFORMANCE ACHIEVED BY DIFFERENT SPEECH ENHANCEMENT METHODS.

Methods	Filtering	Connection	Loss	CER(%)
Far-field-cln	fixed BF(reference)	–	–	13.61
Far-field-mix	–	–	–	43.13
5×(BFLPS+DCF+DPR+DSNR+cosIPD+sinIPD)	masking mtMVDR	concate	$\mathcal{L}_{\text{IRM\_MSE}}$	35.63
5×(BFLPS+DCF+DPR+DSNR+cosIPD+sinIPD)	masking mtMVDR	attention	$\mathcal{L}_{\text{IRM\_MSE}} + \mathcal{L}_{\text{CE}}$	34.10
5×(BFLPS+DCF+DPR+DSNR+cosIPD+sinIPD)	masking fixed BF	attention	$\mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{SI-SDR}}$	41.66
5×(BFLPS+DCF+DPR+DSNR+cosIPD+sinIPD)	masking neural BF	attention	$\mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{SI-SDR}}$	42.43
5×(BFLPS+DCF+DPR+DSNR+cosIPD+sinIPD)	masking fixed BF	attention	$\mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{SI-SDR}} + \mathcal{L}_{\text{CE}}$	33.22
5×(BFLPS+DCF+DPR+DSNR+cosIPD+sinIPD)	masking neural BF	attention	$\mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{SI-SDR}} + \mathcal{L}_{\text{CE}}$	**32.91**

8-th row and in Table 2 show that neural beamformers randomly initialized outperform the fixed beamformers, which is suggested that the neural network can be fully competent for the traditional beamformers through the supervised learning. The fixed beamformers are usually desinged in advance under certain sound field assumption, which is hard to be met in real-world environments. And the data-driven neural beamformers learn the spatial filter coefficients from supervised data under the constraint of traditional beamformering operation and are jointly optimized with mask prediction network in an end-to-end manner. In addition, neural beamformers also utilize historical information compared to fixed beamformers. Therefore, neural beamformers show greater potentials for speech enhancement. We also observed that neural beamformers outperform the adaptive mtMVDR in masking-based multichannel speech enhancement framework. It is very valuable for applications on low-source devices because neural beamformers avoids the complex covariance matrix estimation and its inverse operation for each frequency.

Figure 3 shows the beampattern colormaps of neural beamformers learned from the supervised data. As expected, the filters $\mathbf{w}_{\theta_s}(f)$ preserve the signal coming from the desired directional areas and suppress the signal coming from other directional areas, which suggestes that data-driven neural beamformer has learned the expert knowledge of the hand-designed beamformer. We also observe that the filters $\mathbf{g}_{\theta_s}(f)$ further reduce the residual noise in the signals beamformered by $\mathbf{w}_{\theta_s}(f)$, which achieves the expectation of Eq (10). Intuitively, neural beamformers shows better noise reduction performance compared with the fixed beamformers designed by experts in advance.

3) ASR experiments: Speech enhancement is usually used as an essential front-end module for speech recognition systems to improve their ASR performance in far-field environments.

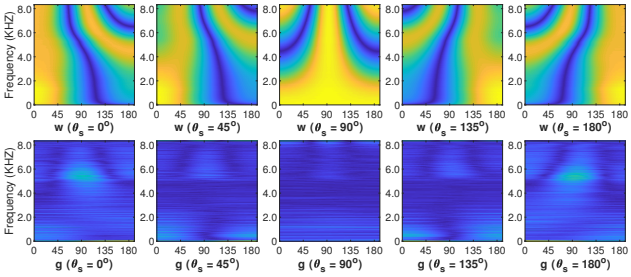

Fig. 3. The beampatterns of neural beamformers. The top figures are the beampatterns of the filters $\mathbf{w}_{\theta_s}(f)$ and the bottom figures are the beampatterns of the filters $\mathbf{g}_{\theta_s}(f)$. A brighter color means a larger filter response.

In order to comprehensively evaluate the performance of the proposed method, we conducted several ASR experiments. The ASR model is trained on multiple condition. Its training dataset contains clean speech and far-field noisy speech. The ASR model was not fine-tuned by the enhanced speech data. Table 3 reports the ASR performance of different speech enhancement systems. All results show that multichannel speech enhancement can significantly improve the ASR performance in far-field environments. Compared with the baseline systems, the proposed methods achieve lower CER, which indicates the bigger improvement in recognition accuracy. The proposed neural beamformer also outperforms the fixed beamformer and mtMVDR in terms of the ASR performance, which suggestes that neural beamformer have capability in reducing noise and avoiding speech distortion.

V. CONCLUSIONS

In this paper, we leverage a spatial attention mechanism to automatically focus on the target source and incorporate it into masking neural beamformers for speech enhancement. Neural beamformers considering historical signals are proposed and

trained to enhance speech signal from the target direction. A cross-entropy classification objective between attention weights and target directions is explored to supervise spatial attention to focus on the target source. Through data-driven supervised training, spatial attention, mask prediction network and neural beamformer are jointly optimized in an end-to-end manner. Systematic experiments demonstrate that data-driven neural beamformers have greater potentials for speech enhancement and spatial attention has the capacity to focus on the direction of target speech.

REFERENCES

[1] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 2, pp. 260–276, 2010.

[2] D. L. Wang and J. T. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[3] R. Haeb-Umbach, S. Watanabe, T. Nakatani, *et al.*, "Speech processing for digital home assistants: Combining signal processing with deep-learning techniques," *IEEE Signal Processing Magazine*, vol. 36, no. 6, pp. 111–124, 2019.

[4] Y. Xu, M. Yu, S. X. Zhang, *et al.*, "Neural spatio-temporal beamformer for target speech separation," in *Interspeech 2020, Virtual Event, Shanghai, China, 25-29 October 2020*, 2020, pp. 56–60.

[5] C. A. Anderson, P. D. Teal, and M. A. Poletti, "Spatially robust far-field beamforming using the von mises(-fisher) distribution," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 12, pp. 2189–2197, 2015.

[6] C. Böddeker, H. Erdogan, T. Yoshioka, and R. Haeb-Umbach, "Exploring practical aspects of neural mask-based beamforming for far-field speech recognition," in *ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, 2018, pp. 6697–6701.

[7] Z. Q. Wang and D. L. Wang, "On spatial features for supervised speech separation and its application to beamforming and robust ASR," in *ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, 2018, pp. 5709–5713.

[8] R. Z. Gu, L. W. Chen, S. X. Zhang, *et al.*, "Neural spatial filter: Target speaker speech separation assisted with directional information," in *Interspeech 2019, Graz, Austria, September 15-19, 2019*, 2019, pp. 4290–4294.

[9] G. J. Li, S. Liang, S. Nie, W. J. Liu, and Z. L. Yang, "Deep neural network-based generalized sidelobe canceller for dual-channel far-field speech recognition," *Neural Networks*, vol. 141, pp. 225–237, 2021, ISSN: 0893-6080.

[10] S. Liang, G. J. Li, S. Nie, Z. L. Yang, W. J. Liu, and J. H. Tao, "Exploiting the directional coherence function for multichannel source extraction," *Speech Communication*, vol. 128, pp. 1–14, 2021, ISSN: 0167-6393.

[11] A. Schwarz and W. Kellermann, "Coherent-to-diffuse power ratio estimation for dereverberation," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 6, pp. 1006–1018, 2015.

[12] Z. Q. Wang and D. L. Wang, "Combining spectral and spatial features for deep learning based blind speaker separation," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 2, pp. 457–468, 2019.

[13] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Trans. Speech Audio Process.*, vol. 18, no. 7, pp. 1717–1731, 2010.

[14] M. Kolbæk, Z. H. Tan, S. H. Jensen, and J. Jensen, "On loss functions for supervised monaural time-domain speech enhancement," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 825–838, 2020.

[15] J. Y. Du, X. Y. Na, X. C. Liu, and H. Bu, "AISHELL-2: transforming mandarin ASR research into industrial scale," *CoRR*, vol. abs/1808.10583, 2018.

[16] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, *et al.*, "Audio set: An ontology and human-labeled dataset for audio events," in *ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, 2017, pp. 776–780.

[17] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication.*, vol. 12, no. 3, pp. 247–251, 1993.

[18] E. A. P. Habets, I. Cohen, and S. Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint," *The Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 2911–2917, 2008.

[19] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Speech Audio Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[20] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *ICASSP 2001, Utah, USA, May 7-11, 2001*, 2001, pp. 749–752.

[21] H. Bu, J. Y. Du, X. Y. Na, B. G. Wu, and H. Zheng, "AISHELL-1: an open-source mandarin speech corpus and a speech recognition baseline," in *O-COCOSDA 2017, Seoul, South Korea, November 1-3, 2017*, 2017, pp. 1–5.

[22] L. Beijing DataTang Technology Co., *Aidatatang 1505*, 2019. [Online]. Available: https://www.datatang.com/opensource.

[23] D. Wang, X. W. Zhang, and Z. Y. Zhang, *Thchs-30 : A free chinese speech corpus*, 2015. [Online]. Available: http://arxiv.org/abs/1512.01882.