EMBEDDING AND BEAMFORMING: ALL-NEURAL CAUSAL BEAMFORMER FOR MULTICHANNEL SPEECH ENHANCEMENT

Andong Li[⋆†]*, Wenzhe Liu*[⋆†]*, Chengshi Zheng*[⋆†]*, Xiaodong Li*[⋆†]

[⋆] Key Laboratory of Noise and Vibration Research, Institute of Acoustics, Chinese Academy of Sciences, Beijing, China
[†] University of Chinese Academy of Sciences, Beijing, China

ABSTRACT

Standing upon the intersection of traditional beamformers and deep neural networks, we propose a causal neural beamformer paradigm called *Embedding and Beamforming*, and two core modules are devised accordingly, namely EM and BM. For EM, instead of estimating spatial covariance matrix explicitly, the 3-D embedding tensor is learned with the network, where the spatial-spectral discriminative information can be implicitly represented. For BM, a network is directly leveraged to derive the beamforming weights so as to implement filter-and-sum operation. To further improve the speech quality, a post-processing module is introduced to further suppress the residual noise. Based on the DNS-Challenge dataset, we conduct the experiments for multichannel speech enhancement and the results show that the proposed system outperforms previous advanced baselines by a large margin in terms of multiple evaluation metrics.

***Index Terms*—** Multichannel speech enhancement, neural beamformer, embedding, causal, post-processing

1. INTRODUCTION

Speech enhancement (SE) attempts to extract the target speech from the mixture signals. Due to the utilization of spatial information to distinguish between target and interference, a plethora of beamforming-based multichannel speech enhancement algorithms have been widely proposed in a diverse set of applications, ranging from audio-video conferencing to human-machine interaction [1, 2].

With the renaissance of deep neural networks (DNNs), neural beamformers have propitiated wide interest due to their promising performance in speech restoration and automatic speech recognition (ASR) systems [3–5]. A typical strategy is to combine DNNs with traditional beamforming techniques. Specifically, a single-channel SE network is first adopted to parallelly estimate time-frequency (T-F) masks *w.r.t.* speech and noise for each channel. The spatial covariance matrices are then calculated to derive the optimal weights for beamformers based on statistical optimization criteria [3, 6, 7], like minimum variance distortionless response (MVDR) beamformers and multichannel Wiener filter (MWF) beamformers. It is obvious that the second stage is purely based on statistical theory and the two stages are optimized with different optimization criteria and implementation methods, where the mask estimation error in the first stage may heavily hamper the subsequent beamforming results. More recently, regression-based approaches began to thrive. Instead of obtaining the beamformers' weights, the spatial information is represented either manually or implicitly, which serves as the auxiliary aspect of spectral feature to assist the speech recovery in either time domain [8, 9] or T-F domain [10, 11]. The overall network topology is akin to the single-channel case. Nonetheless, it is still far from affirmative how to combine the spatial and spectral features efficiently [9]. Moreover, the potential of spatial filtering is not fully exploited, which can limit the overall enhancement performance in real acoustic scenarios.

Recently, another research line follows the guidance of estimating the beamforming weights with DNNs. As an early trial, Xiao *et al.* [12] proposed to use GCC features to estimate the weights with DNNs. In [13], Luo *et al.* proposed the FasNet to implement the filter-and-sum operation in the time domain. Compared with advanced T-F domain based works, it lacks adequate robustness and superiority [11]. More recently, an all-deep-learning beamforming paradigm called ADL-MVDR was proposed, where the mask estimation, spatial covariance calculation, and framewise beamforming are integrated into a whole network and trained in an end-to-end manner [14]. While achieving impressive performance in speech quality and ASR accuracy, as only the supervision *w.r.t.* target speech is provided, it remains agnostic whether the internal signal-theory based operations follow the expected physical definitions.

At this point, we would like to answer a question, *i.e., how to guarantee a neural system that can generate frame-level weights for beamforming?* We argue that it should meet two requirements. First, it should incorporate abundant spatial information to distinguish the sources from different directions. Besides, as the filters need to be updated frame-by-frame, it should learn the T-F cue to assist the separation between speech and interference especially when the spatial cue is absent and inaccurate. To this end, we propose a generalized framework with the causal setting called **E**mbedding **a**nd **B**eamforming **Net**work (**EaBNet**) for all-neural beamforming. Two core modules are devised, namely *Embedding Module* (EM) and *Beamforming Module* (BM). In the first module, the network aims to extract the feature from the spectral and spatial perspectives and obtain the 3-D embedding tensor that can latently distinguish between speech and noise components in both senses. In the second module, rather than obtain the filter weights following the statistically optimal beamformer theory, we adopt a network to accomplish the process. It has been illustrated that DNNs can better learn the optimal filter weights than following the traditional beamformer formulas [15]. Note that, in contrast to [15] that the second-order spatial statistics need to be explicitly calculated, our approach chooses to directly obtain the temporal-spatial embedding so that it can potentially learn higher-order spatial statistics with data-driven. We generate the multichannel dataset based on DNS-Challenge corpus, the experimental results show that our system outperforms previous state-of-the-art (SOTA) baselines by a large margin and also surpasses the MVDR beamformer with oracle masks.

The rest of the paper is organized as follows. In Section 2, we

Chengshi Zheng is the corresponding author.

formulate the physical model. In Section 3, the proposed system is introduced in detail. Section 4 gives the experimental setup, and the experimental results and analysis are provided in Section 5. Some conclusions are drawn in Section 6.

2. PHYSICAL MODEL

Let us assume $x^{(p)}(t)$, with $p = 0, \cdots, P-1$, denotes the time-domain noisy and reverberant speech signal at the pth microphone. The physical model in the short-time Fourier transform (STFT) domain can be given by

$$\mathbf{X}_{f,t} = \mathbf{S}_{f,t} + \mathbf{N}_{f,t} = \mathbf{c}_f S_{f,t} + \mathbf{r}_f N_{f,t}, \tag{1}$$

where $\{\mathbf{X}_{f,t}, \mathbf{S}_{f,t}, \mathbf{N}_{f,t}\} \in \mathbb{C}^{P\times 1}$ respectively denote the reverberant mixture, target speech and noise of P channels with frequency index of $f \in \{1, \cdots, F\}$ and time index of $t \in \{1, \cdots, T\}$. Without loss of generality, the first channel is selected as the reference channel by default. $\{\mathbf{c}_f, \mathbf{r}_f\} \in \mathbb{C}^{P\times 1}$ denote the relative transfer function (RTF) of the speech and that of noise, respectively. $\{S_{f,t}, N_{f,t}\} \in \mathbb{C}$ are the complex values of target speech and that of noise in the reference channel. Note that although we foucs on noise reduction in this study, it also works to directional speaker interference case, which is left as future work.

Different from previous beamformers operated at either utterance-[3] or chunk-level [16], we aim to estimate the framewise filter weights, which can support the real-time processing at run-time. Therefore, the beamforming output can be formulated as

$$\widetilde{S}_{f,t} = \sum_{p=0}^{P} \left(M_{f,t}^{(p)}\right)^* X_{f,t}^{(p)}, \tag{2}$$

where $M_{f,t}^{(p)} \in \mathbb{C}$ denotes the beamforming weight of the pth microphone. * stands for the conjugate operator. Note that only the noise suppression is considered and dereverberation is not addressed in this paper.

3. PROPOSED SYSTEM

3.1. Forward Stream

In this study, the target speaker is assumed to be static within each utterance, *i.e.*, the direction of arrivals (DOAs) of the target and noise remains unchanged. However, due to the highly dynamic property of speech and noise distribution, it is quite difficult to accurately separate them with spatial-only cues. To this end, we propose an embedding-and-beamforming paradigm to learn the beamforming weights from both spatial and spectral perspectives. The overall diagram of the proposed framework is shown in Fig. 1(a). It consists of three parts, namely the embedding module (EM), beamforming module (BM), and the post-processing module (PostNet). For EM, it adaptively aggregates the information across the T-F spectrum and different channels and obtain the 3-D tensor where both spatial and spectral discriminative information are represented. For BM, it is employed to replace the traditional beamforming step and directly infer the filter weights and apply them to each channel. The filtered spectra are then summed together to obtain the expected speech. After the beamforming process, it may still contain the residual noise. Therefore, the PostNet is adopted to further suppress these residual noise components and improve the speech quality. In a nutshell, the whole procedure can be formulated as

$$\widetilde{\mathbf{E}} = EMet(Cat(\mathbf{X}^{(0)}, \cdots, \mathbf{X}^{(P-1)})), \tag{3}$$

$$\widetilde{\mathbf{M}} = BFNet(\widetilde{\mathbf{E}}), \tag{4}$$

$$\widetilde{\mathbf{S}}^{(1)} = \sum_{p=0}^{(p)} \left(\widetilde{\mathbf{M}}^{(p)}\right)^{\mathsf{H}} \widetilde{\mathbf{X}}^{(p)}, \tag{5}$$

$$\widetilde{\mathbf{S}}^{(2)} = PostNet(Cat(\widetilde{\mathbf{S}}^{(1)}, \mathbf{X}^{(0)})), \tag{6}$$

where $\{EMet, BFNet, PostNet\}$ denote the network topology of three modules, respectively. Cat refers to the concatenation operation along the channel axis. $\widetilde{\mathbf{E}} \in \mathbb{C}^{F\times T\times C}$ and $\widetilde{\mathbf{M}} \in \mathbb{C}^{F\times T\times P}$ are respectively the estimated 3-D embedding and beamforming tensors and C is the embedding channel dimension. Superscripts (1) and (2) denote the output of the beamformer and PostNet.

3.2. Embedding Module

Motivated by our preliminary works [17, 18], the convolutional "Encoder-TCN-Decoder" topology is employed in the EM. The encoder gradually extracts features with multiple downsampling operations. The decoder has the mirror structure except that all the convolution layers are replaced by the deconvolution (De) versions and recover to the original resolution. Temporal convolution networks (TCNs) serve as the bottleneck, where multiple temporal convolutional modules (TCMs) are stacked for long-term sequence modeling and we adopt the squeezed version herein to decrease the parameter burden [17], *i.e.*, S-TCM, as shown in Fig. 1(c).

The real and imaginary (RI) components of P microphones are concatenated along the channel dimension as the network input, *i.e.*, $\mathbf{X} \in \mathbb{C}^{F\times T\times 2P}$. To better capture spatial-spectral correlation, the U^2-Encoder and U^2-Decoder are employed, which consists of multiple recalibration encoder/decoder layers (RELs/RDLs). The details are shown in Fig. 1(b)(d). Take ith REL/RDL as an example, it mainly consists of a 2D-(De)GLU [19], instance normalization (IN), PReLU [20], and a UNet-block with the residual connection [21]. The input $\mathbf{I}_i$ is first encoded by the (de)convolution operation. Afterward, the UNet-block receives the encoded feature map $\mathcal{K}_i(\mathbf{I}_i)$ as the input and then further recalibrates the information distribution with a light-weight sub-UNet. The process can be given by

$$\mathcal{K}_i(\mathbf{I}_i) = GLU(\mathbf{I}_i), \tag{7}$$

$$\mathbf{O}_i = UNet\text{-}block(\mathcal{K}_i(\mathbf{I}_i)) + \mathcal{K}_i(\mathbf{I}_i), \tag{8}$$

3.3. Beamforming Module

Having obtained the estimated 3D-embedding tensor $\widetilde{\mathbf{E}}$, the BM is leveraged to derive the framewise beamforming weights. Different from most of existing beamformers, we leverage the mapping capability of networks to directly estimate the beamforming weights, which can avoid explicitly computing the spatial covariance matrix and its inversion, and thus to improve the system stability. Two types of modules are investigated for completeness herein, namely convolutional-based beamformer (C-BF) and recurrent-based beamformer (R-BF), as shown in Fig. 1(a). For the first type, a pointwise 2D-Conv is adopted to transform the channel dimension from C to $2P$, which obtains the real and imaginary components of P-channel filters. For the second type, the layer norm (LN) [22] is first adopted to normalize the embedding tensor, followed by two LSTM layers to simulate the beamforming process updated frame by frame, as shown in Fig. 1(e). Two fully-connected (FC) layers with hidden ReLU nonlinearity activation function are used to estimate the coefficients. Note that in the previous literature [3], the frequency dimension serves as the feature input of the LSTM. However, in this paper, the LSTM is shared by different frequency subbands to simulate the operation of traditional beamformers that apply to each frequency subband independently. The above process can be expressed as

$$\widetilde{\mathbf{M}} = Conv\left(\widetilde{\mathbf{E}}\right), \text{for C-BF}, \tag{9}$$

$$\widetilde{\mathbf{M}} = FC\left(LSTM\left(LayerNorm\left(\widetilde{\mathbf{E}}\right)\right)\right), \text{for R-BF}, \tag{10}$$

Following filter-and-sum operation, the complex-valued filters are then applied to each channel and the filtered spectra are summed together to obtain the expected speech.

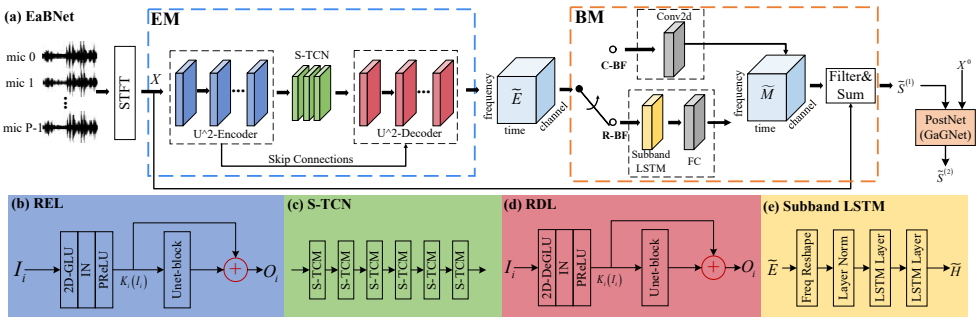

Fig. 1. The diagram of the proposed framwork EaBNet. It mainly consists of two modules, namely EM and BM. Besides, the post-processing module is also introduced to further suppress the residual noise component. Different modules are remarked with different colors.

3.4. Post-processing Module

After the beamforming stage, due to the performance limitation, some residual noise components may exist, which hinders the speech quality. To this end, a post-processing module is proposed to further suppress the remaining noise. Theoretically, any single-channel SE system can be selected. In this paper, we choose our newly proposed GaGNet as the PostNet due to its promising performance in noise reduction at low computational cost. Due to the space limit, we may refer the readers to [23] for more details.

4. EXPEREIMENTAL SETUP

4.1. Dataset Preparation

The DNS-Challenge dataset is chosen to convolve with multichannel RIRs to generate microphone-array signals for evaluation.[1] More specifically, for clean speech, the *neutral clean speech* set is selected, which consists of around 562 hours by 11,350 speakers. We randomly split it into two non-overlap parts, namely for model training and evaluation. The average duration of the utterance is chunked into around 6 seconds. For the noise set, similar to [17], around 20,000 types of noises are randomly selected for training, whose duration is around 55 hours. Multichannel RIRs are generated with image method [24] based on a uniform linear array with 9 microphones and the distance of two adjacent microphones being 4cm. The room size ranges from 3m-3m-2.5m to 10m-10m-3m (length-width-height). The reverberation time (RT_{60}) ranges from 0.05s to 0.7s. Note that the DOA difference between target speech and interference noise is at least 5° during the data generation and the distance from the source to the array is randomly selected from $\{0.5m, 1m, 2m, 3m\}$. The relative signal-to-noise ratio (SNR) ranges $[-6\text{dB}, 6\text{dB}]$ with 2dB interval. Totally, we create around 80,000, and 4000 mix-clean pairs for training and validation.

For model evaluation, four challenging noises are selected, namely babble, factory1, white noises from NOISEX92 [25] and cafe noise from CHiME3 noise set [26]. Four SNRs are set, namely $\{-5\text{dB}, -2\text{dB}, 0\text{dB}, 2\text{dB}\}$, with 600 mix-clean pairs in each case.

4.2. Baselines

Six approaches are selected as the baselines, namely CTSNet [17], GaGNet [23], FasNet-TAC [27], MC-ConvTasNet [8], MIMO-UNet [28], and oracle MB-MVDR. CTSNet and GaGNet are two SE systems that achieved state-of-the-art (SOTA) performance in the monaural scenario. FasNet-TAC and MC-ConvTasNet are two time-domain based methods that exploit multiple raw-waveforms to extract the target speech. MIMO-UNet ranked first in the Far-field Multi-Channel Speech Enhancement Challenge for Video Conferencing. For a fair comparison, all the non-causal settings are replaced by the causal versions. Besides, the scale-invariant SNR (SI-SNR) loss in FasNet-TAC and MC-ConvTasNet is replaced by classical SNR loss to mitigate the free change of the magnitude level [29]. For MB-MVDR, we utilize the oracle ideal ratio mask (IRM) to calculate the spatial covariance matrix and then derive MVDR weights for beamforming.[2]

4.3. Implementation Details

4.3.1. Model Details

In the EM, the kernel size of 2D-(De)GLU is set as $(2, 3)$ with stride being $(1, 2)$ in the time and frequency axes. For each UNet-block, the kernel and stride size are $(1, 3)$ and $(1, 2)$, respectively. Let us define the number of (de)encoding layers within the UNet-block as Q, then $Q = \{4, 3, 2, 1, 0\}$ for U^2-Encoder and $Q = \{1, 2, 3, 4, 0\}$ for U^2-Decoder, respectively, where 0 means that no UNet-block is used. The number of channels in both encoder and decoder remains 64 by default. For bottleneck sequence modeling, 3 S-TCNs are stacked, each of which consists of 6 S-TCMs with kernel size and dilation rate being 5 and $\{1, 2, 4, 8, 16, 32\}$. When R-BF is switched for beamforming, two uni-directional LSTMs are utilized with 64 hidden nodes. For C-BF, the kernel size is set to $(1, 1)$. We provide the demo, which can be available at https://andong-li-speech.github.io/EaBNet-Demo/.

4.3.2. Training Details

All the utterances are sampled at 16 kHz. The 20 ms Hanning window is utilized with 50% overlap between adjacent frames. 320-point FFT is utilized, leading to 161-D features, *i.e.*, F=161. Most recently, the efficacy of power-compression is investigated for single-channel speech enhancement [23] and dereverberation tasks [30]. Here, we adopt it for both the input and target of each channel, *i.e.*, $|\mathbf{X}^{(p)}|^{0.5}e^{j\theta_{\mathbf{X}^{(p)}}}$, $|\mathbf{S}^{(p)}|^{0.5}e^{j\theta_{\mathbf{S}^{(p)}}}, p \in \{1 \cdots, P\}$. The rationale is that we only compress the magnitude and leave the phase unchanged. Therefore, spatial information can still be well preserved. MMSE with magnitude constraint is adopted as the loss function for training [17, 23]. All the models are trained with Adam optimizer [31] and the learning rate is initialized at 5e-4 and will be halved if the loss does not decrease for consecutive two epochs. The batch size is 8 and the number of epochs is 60.

5. RESULTS AND ANALYSIS

5.1. Ablation Study

We conduct the ablation study on EaBNet in terms of whether to use UNet-block, whether to output multiple filter weights (MO), the type of BF, and whether the magnitude compression is adopted, as shown in Table 1. Perceptual evaluation of speech quality (PESQ) [32], extended short-time objective intelligibility (ESTOI) [33], and signal-distortion ratio (SDR) [34] are adopted as evaluation metrics. Sev-

[1] github.com/microsoft/DNS-Challenge/tree/master/datasets

[2] https://pypi.org/project/beamformers

Table 1. Ablation study on the proposed EaBNet. The values are specified with PESQ/ESTOI(%)/SDR(dB) format. **BOLD** indicates the best score in each case. "Avg." denotes the average value among different SNRs in the test set.

System	ID	UNet-block	MO	BF Type	Compression	Para. (M)	-5dB	-2dB	0dB	2dB	Avg.
EaBNet	1	✗	✓	R-BF	✓	**2.19**	3.16/78.95/13.83	3.34/82.36/15.45	3.49/86.18/16.90	3.59/87.63/17.55	3.40/83.78/15.93
	2	✓	✗	✗	✓	2.77	3.10/77.22/12.26	3.28/80.80/13.74	3.44/84.65/15.09	3.54/86.28/15.94	3.34/82.24/14.26
	3	✓	✓	C-BF	✓	2.77	3.20/79.67/13.40	3.38/82.97/15.05	3.54/86.68/16.50	3.63/87.89/17.12	3.44/84.30/15.52
	4	✓	✓	R-BF	✗	2.84	2.93/76.66/**14.73**	3.12/80.88/**16.53**	3.29/84.83/**18.13**	3.39/86.21/**18.61**	3.18/82.15/**17.00**
	5	✓	✓	R-BF	✓	2.84	**3.30/81.75**/14.68	**3.47/84.66**/16.16	**3.61/88.04**/17.64	**3.70/89.19**/18.38	**3.52/85.91**/16.72

Table 2. Results comparison with advanced baselines.

Systems	Domain	Para. (M)	MACs (G/s)	RTF	Channel	-5dB	-2dB	0dB	2dB	Avg.
Noisy	-	-	-	-	-	1.45/29.62/-4.89	1.62/37.71/-1.93	1.74/41.85/0.05	1.87/49.42/2.05	1.67/39.65/-1.18
CTSNet	T-F	4.35	5.57	0.37	1	1.87/40.35/2.32	2.14/50.87/5.63	2.34/58.27/7.64	2.50/63.64/9.03	2.21/53.28/6.15
GaGNet	T-F	5.94	**1.63**	0.19	1	1.91/42.02/3.01	2.22/52.50/5.99	2.42/59.99/7.90	2.58/65.09/9.19	2.28/54.90/6.52
FasNet-TAC	T	3.82	7.56	0.67	9	2.40/63.39/11.43	2.62/68.99/13.44	2.77/73.89/14.88	2.88/76.50/15.54	2.67/70.69/13.82
MC-ConvTasnet	T	6.56	5.28	0.43	9	2.21/59.57/9.90	2.44/65.20/11.61	2.72/72.16/13.46	2.82/74.52/14.02	2.55/67.86/12.25
MIMO-UNet	T-F	**1.97**	4.09	**0.16**	9	2.39/60.93/8.96	2.61/66.99/11.17	2.75/71.29/12.49	2.85/73.71/13.20	2.65/68.23/11.45
MB-MVDR(oracle)	T-F	-	-	-	9	2.88/78.59/12.06	3.04/82.45/13.90	3.18/85.82/15.17	3.29/87.43/15.90	3.10/83.57/14.26
EaBNet*	T-F	2.91	8.46	0.80	9	3.24/80.16/13.60	3.41/83.49/15.15	3.56/86.91/16.51	3.65/88.11/17.15	3.46/84.67/15.60
EaBNet	T-F	2.84	7.38	0.59	9	3.30/81.75/14.68	3.47/84.66/16.16	3.61/88.04/17.64	3.70/89.19/18.38	3.52/85.91/16.72
EaBNet+PostNet	T-F	8.78	9.04	0.83	9	**3.44/83.33/15.13**	**3.58/85.86/16.58**	**3.71/89.04/18.05**	**3.79/90.03/18.72**	**3.63/87.06/17.12**

eral observations can be made. 1) Going from ID-1 to ID-5, consistent improvements are made for all cases, which show that the introduction of the UNet-block can well preserve spectral and spatial information and lead to better beamforming results. 2) When we directly estimate the complex-valued mask for the reference channel without explicit beamforming process, as shown from ID-5 to ID-2, consistent performance degradations in three metrics are observed, which emphasize the significance of the beamforming operation in multi-channel speech enhancement systems. 3) We compare the performance between different BF types, as shown in ID-3 and ID-5, one can find that R-BF yields relatively better performance over C-BF. This is because, in R-BF, the LSTM is leveraged to update the state frame by frame, which leads to better beamforming weights estimation when the spatial information is not accurate enough for separation. 4) Compared with ID-4, when the magnitude compression is employed, considerable improvements in PESQ and ESTOI are achieved while mild degradation in SDR is observed. This is because compression operation decreases the dynamic range of spectrum distribution and highlights the priority of low-energy regions [23, 30]. As the result, more residual noise can be suppressed and improve the speech quality. Meanwhile, with the nonlinear compression operation, the linear separability between different sources in the space may be destroyed. Therefore, it may cause more target distortion, which partly explains the degradation in the SDR.

5.2. Results Comparison with Advanced Baselines

The best configuration of EaBNet in Table 1, *i.e.,* ID-5, is chosen to compare with other baselines, whose results are presented in Table 2. To emphasize the effectiveness of the learned embedding in spectral-spatial information representation, we also set the reference dubbed EaBNet*, where we output the complex-valued masks *w.r.t.* speech and noise at the end of EB and then the spatial covariance matrices are calculated and concatenated as the input of R-BF [15]. For a fair comparison, the hidden nodes of the LSTM remain the same.

From the table, several observations can be obtained. First, compared with the single-channel case, when more channels are available, considerable improvements for three metrics can be achieved for all the multichannel based models. This indicates that the utilization of spatial information can facilitate the separation of different sources. Second, the proposed system outperforms the previous baselines by a large margin. For example, going from MIMO-UNet to EaBNet, average 0.87, 17.68%, and 5.67dB improvements are achieved in terms of PESQ, ESTOI, and SDR, respectively. It fully demonstrates the superiority of our system in speech recovery. Besides, we observe that the proposed system also surpasses the MB-MVDR with oracle IRM consistently, which reveals the feasibility of end-to-end framewise beamformers with DNNs over previous tandem-style schemes [3, 6, 7]. Third, it is interesting to find that compared with EaBNet*, when the embedding tensor is abstractly represented rather than follow the traditional signal-theory to calculate the spatial covariance matrices *w.r.t.* speech and noise, even better performance can be made, which inspires us to rethink the necessity of signal-theory in end-to-end neural beamformers. We can explain this phenomenon from several aspects. For one thing, the spatial covariance matrix is usually sparse and is often redundant, and thus it is unnecessary to estimate all its entries in theory to improve the robustness [35]. Meanwhile, it also tends to be less robust toward the real scenarios than the compact embedding. For another, compared with explicit second-order covariance matrix calculation, the implicit embedding is learned directly from the training data, it may thus potentially learn higher-order spatial statistics. Fourth, when the PostNet is adopted, further metric improvements can be achieved, which illustrates the necessity of post-processing in noise suppression and speech recovery.

We also provide the model size, the number of multiply-accumulate operations (MACs) per second, and real-time factor (RTF), as shown in Table 2. RTF is evaluated on an Intel Core(TM) i5-4300 CPU clocked at 1.90GHz. One can find that EaBNet has an overall light-weight model size (2.84M) than other baselines and the RTF is 0.59, which meets the real-time processing criterion. Despite more parameters and higher RTF are induced when the PostNet is added, we can decrease the overall burden by choosing more decent post-processing algorithms with less computational complexity.

6. CONCLUSIONS

In this paper, we propose a generalized causal framework called EaBNet, which enables the framewise neural beamforming for multichannel speech enhancement. Two modules are designed, namely EM and BM. In the EM, we directly generate the 3-D embedding tensor which contains both spatial-spectral discriminative information. In the BM, a network is directly utilized to output the filter weights. A post-processing module is also introduced to further suppress the residual noise and facilitate speech recovery. The experiments show that the proposed system yields state-of-the-art performance over previous baselines by a large margin.

References

[1] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio. Speech, Lang. Process.*, vol. 25, no. 4, pp. 692–730, 2017.

[2] B.D Van Veen and K.M Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, 1988.

[3] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. ICASSP*, pp. 196–200, 2016.

[4] Z. Wang, P. Wang, and D. Wang, "Complex spectral mapping for single-and multi-channel speech enhancement and robust ASR," *IEEE/ACM Trans. Audio. Speech, Lang. Process.*, vol. 28, pp. 1778–1787, 2020.

[5] T. Hori, Z. Chen, H. Erdogan, J. Hershey, J.-Le Roux, V. Mitra, and S. Watanabe, "Multi-microphone speech recognition integrating beamforming, robust feature extraction, and advanced DNN/RNN backend," *Comput. Speech & Lang.*, vol. 46, pp. 401–418, 2017.

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

[7] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, "BLSTM supported GEV beamformer front-end for the 3rd chime challenge," in *Proc. ASRU, 2015*, pp. 444–451, 2015.

[8] J. Zhang, C. Zorilă, R. Doddipatla, and J. Barker, "On end-to-end multi-channel time domain speech separation in reverberant environments," in *Proc. ICASSP*, pp. 6389–6393, 2020.

[9] R. Gu, L. Chen, S. Zhang, J. Zheng, Y. Xu, M. Yu, D. Su, Y. Zou, and D. Yu, "Neural spatial filter: Target speaker speech separation assisted with directional information," in *Proc. Interspeech*, pp. 4290–4294, 2019.

[10] Z. Wang and D. 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, 2018.

[11] Y. Fu, J. Wu, Y. Hu, M. Xing, and L. Xie, "DESNET: A multi-channel network for simultaneous speech dereverberation, enhancement and separation," in *Proc. SLT*, pp. 857–864, 2021.

[12] X. Xiao, S. Watanabe, H. Erdogan, L. Lu, J. Hershey, M. Seltzer, G. Chen, Y. Zhang, M. Mandel, and D. Yu, "Deep beamforming networks for multi-channel speech recognition," in *Proc. ICASSP*, pp. 5745–5749, 2016.

[13] Y Luo, C. Han, N. Mesgarani, E. Ceolini, and S.-C. Liu, "Fasnet: Low-latency adaptive beamforming for multi-microphone audio processing," in *Proc. ASRU*, pp. 260–267, 2019.

[14] Z. Zhang, Y. Xu, M. Yu, S. Zhang, L. Chen, and D. Yu, "ADL-MVDR: All deep learning mvdr beamformer for target speech separation," in *Proc. ICASSP*, pp. 6089–6093, 2021.

[15] Y. Xu, Z. Zhang, M. Yu, S. Zhang, and D. Yu, "Generalized Spatial-Temporal RNN Beamformer for Target Speech Separation," in *Proc. Interspeech*, pp. 3076–3080, 2021.

[16] T. Higuchi, N. Ito, S. Araki, T. Yoshioka, M. Delcroix, and T. Nakatani, "Online MVDR beamformer based on complex gaussian mixture model with spatial prior for noise robust asr," *IEEE/ACM Trans. Audio. Speech, Lang. Process.*, vol. 25, no. 4, pp. 780–793, 2017.

[17] A. Li, W. Liu, C. Zheng, C. Fan, and X. Li, "Two heads are better than one: A two-stage complex spectral mapping approach for monaural speech enhancement," *IEEE/ACM Trans. Audio. Speech, Lang. Process.*, vol. 29, pp. 1829–1843, 2021.

[18] A. Li, W. Liu, X. Luo, G. Yu, C. Zheng, and X. Li, "A simultaneous denoising and dereverberation framework with target decoupling," in *Proc. Interspeech*, pp. 2801–2805, 2021.

[19] Y. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proc. ICML*, pp. 933–941, 2017.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. ICCV*, pp. 1026–1034, 2015.

[21] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. Zaiane, and M. Jagersand, "U2-Net: Going deeper with nested u-structure for salient object detection," *Pattern Recognit.*, vol. 106, pp. 107404, 2020.

[22] J. Ba, J. Kiros, and G. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[23] A. Li, C. Zheng, L. Zhang, and X. Li, "Glance and gaze: A collaborative learning framework for single-channel speech enhancement," *Appl. Acoust.*, vol. 187, p. 108499, 2022.

[24] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.

[25] A. Varga and H. 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 Commun.*, vol. 12, no. 3, pp. 247–251, 1993.

[26] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'chime'speech separation and recognition challenge: Dataset, task and baselines," in *Proc. ASRU*, pp. 504–511, 2015.

[27] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in *Proc. ICASSP*, pp. 6394–6398, 2020.

[28] X. Ren, X. Zhang, L. Chen, X. Xheng, C. Zhang, L. Guo, B. Yu, "A Causal U-net based Neural Beamforming Network for Real-Time Multi-Channel Speech Enhancement," in *Proc. Interspeech*, pp. 1832–1836, 2021.

[29] K. Kinoshita, T. Ochiai, M. Delcroix, and T. Nakatani, "Improving noise robust automatic speech recognition with single-channel time-domain enhancement network," in *Proc. ICASSP*, pp. 7009–7013, 2020.

[30] A. Li, C. Zheng, R. Peng, and X. Li, "On the importance of power compression and phase estimation in monaural speech dereverberation," *JASA Express Letters*, vol. 1, no. 1, pp. 014802, 2021.

[31] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[32] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, pp. 749–752, IEEE, 2001.

[33] J. Jensen and C. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio. Speech, Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, 2016.

[34] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. Rosca, "First stereo audio source separation evaluation campaign: data, algorithms and results," in *Proc. Int. Conf. Ind. Compon. Anal. Blind Source Separation*. Springer, 2007, pp. 552–559.

[35] C. Zheng, A. Deleforge, X. Li, and W. Kellermann, "Statistical Analysis of the Multichannel Wiener Filter Using a Bivariate Normal Distribution for Sample Covariance Matrices," *IEEE/ACM Trans. Audio. Speech, Lang. Process.*, vol. 26, no. 5, pp. 951–966, 2018.