ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10448366

AN MVDR-EMBEDDED U-NET BEAMFORMER FOR EFFECTIVE AND ROBUST MULTICHANNEL SPEECH ENHANCEMENT

Ching-Hua Lee[⋆]*, Kashyap Patel*[†]*, Chouchang Yang*[⋆]*, Yilin Shen*[⋆]*, Hongxia Jin*[⋆]

[⋆]Samsung Research America [†]Department of ECE, University of Texas at Dallas

ABSTRACT

In multichannel speech enhancement (SE) systems, deep neural networks (DNNs) are often utilized to directly estimate the clean speech for effective beamforming. This approach, however, may not generalize adequately to new acoustic or noise conditions. Alternatively, DNNs can indirectly perform SE by predicting the time-frequency masks of speech and noise patterns to assist classic statistical beamformers. Despite being robust, its effectiveness is constrained by the later statistical component relying on certain modeling assumptions, e.g., covariance-based modeling in the minimum-variance-distortionless-response (MVDR) beamformer. In this paper, we propose a novel integration of the two types of methodology, by introducing an *intra-MVDR* module embedded in the U-Net beamformer, that encompasses the merits of both, i.e., effectiveness and robustness. Experiments show that intra-MVDR leads to improvements that are not achievable by simply enlarging the baseline SE network.

Index Terms— Multichannel speech enhancement, neural beamforming, MVDR, time-frequency mask, spatial filtering

1. INTRODUCTION

Based on the estimation target and input modalities of deep neural networks (DNNs), beamforming-based multichannel speech enhancement (SE) systems have appeared in several types [1, 2, 3, 4, 5, 6]. For SE systems using only microphone arrays, the mainstream beamforming-based approach utilizes the DNN to directly estimate the clean speech for effectiveness, e.g., [7, 8, 9, 10, 11, 12, 13, 14], as depicted in Fig. 1 (a). Such methods are fully data-driven, capable of achieving quite effective performance when test conditions are similar to those seen in training, but could suffer from a lack of generalization to unseen room acoustic and noise conditions.

Another popular type of approach is the time-frequency (T-F) mask-based statistical beamforming, where the DNN performs SE indirectly by estimating some T-F masks representing the speech and noise patterns in the T-F domain [15, 16, 17, 18, 19, 20]. The estimated T-F masks are subsequently leveraged to estimate the signal and noise statistics for utilization in classical beamformers, e.g., the minimum-variance-distortionless response (MVDR) filter [21], as illustrated in Fig. 1 (b). Methods of this type, though potentially being more robust because here the DNN only has to perform the relatively simple task of T-F mask prediction, usually have their efficacy bounded by the subsequent statistical beamforming algorithm relying on several modeling assumptions.

In this paper, we study integration of the two method types to encompass their respective merits, i.e., *effectiveness* and *robustness*, as in Fig. 1 (c). To this end, we introduce an *intra-MVDR* module embedded in a U-Net direct beamformer network to incorporate T-F mask-based statistical beamforming. It is found that by placing the mask-based MVDR module in the middle of the encoder and decoder layers of U-Net, spatial features of multichannel signals can be effectively exploited for improved SE outcomes with only minor increase in model size. Moreover, it is found that integrating intra-MVDR with U-Net in a multi-level manner leads to further improvements by exploiting multi-scale spatial features for processing.

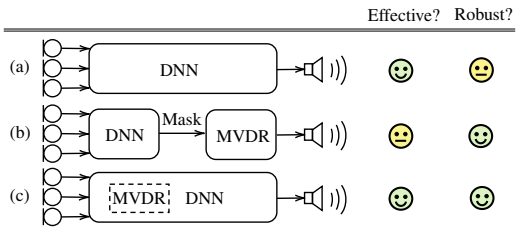

Fig. 1. Illustration of different multichannel SE systems: (a) DNN direct beamformer; (b) DNN followed by statistical beamformer (e.g., MVDR); (c) MVDR-embedded DNN beamformer (proposed).

2. BACKGROUND

Signal model: We consider an acoustic scenario with one desired speech source and several interfering noise signals in a reverberant environment. Our system is developed in the T-F domain using the short-time Fourier transform (STFT). Let f, t stand for the frequency and time frame indexes (with F frequency bins and T time frames in total). We consider an additive noise model where the i-th microphone noisy signal STFT $\mathbf{X}_i \in \mathbb{C}^{F\times T}$ of an N-element local microphone array whose (f,t)-th entry is $X_i(f,t)$ and can be written as [22]: $X_i(f,t) = S_i(f,t) + V_i(f,t), \forall f,t$, where $S_i(f,t)$ and $V_i(f,t)$ are the (f,t)-th entry of speech component $\mathbf{S}_i \in \mathbb{C}^{F\times T}$ and noise component $\mathbf{V}_i \in \mathbb{C}^{F\times T}$ received by microphone i, respectively. In this paper, we consider the goal of recovering the speech component $\mathbf{S} = \mathbf{S}_r \in \mathbb{C}^{F\times T}$ of a selected reference microphone $r \in \{1, \ldots, N\}$ given the microphone signals $\mathbf{X}_1, \ldots, \mathbf{X}_N$.

DNN direct beamformers (direct BF): Multichannel SE systems typically perform the "filter-and-sum" operation, or "beamforming" – linearly combining the multichannel signals to extract the target signal $\mathbf{S}$ from background noise [21]. In the T-F domain, it can be expressed as: $\hat{\mathbf{S}} = \sum_{i=1}^{N} \mathbf{W}_i \odot \mathbf{X}_i$, where $\mathbf{W}_i \in \mathbb{C}^{F\times T}$ is the corresponding set of filter weights for microphone i, $\hat{\mathbf{S}}$ is the enhanced signal, and $\odot$ denotes element-wise multiplication. In many DNN-based approaches [7, 8, 9, 10, 11, 12, 13, 14], the networks are utilized to imitate such beamforming processes (in the time or T-F domain) to directly output the clean speech trained by minimizing some signal reconstruction loss. Methods of this type are effective as the networks learn to directly model the noisy-clean mapping from data, but may not generalize adequately to unseen testing conditions.

T-F mask-based statistical beamformers: Another class of approaches utilizes the DNNs to estimate some set of T-F masks that represent the speech and noise patterns in the STFT domain, which are subsequently used to estimate the signal statistics (i.e., the power spectral density (PSD) matrices) for assisting the conventional beamforming algorithms that rely on accurate PSD estiamtion, e.g., the MVDR beamformer [21]. Methods of this type are usually referred to as T-F mask-based neural beamformers [15, 16, 17, 18, 19, 20]. Specifically, the DNN-predicted T-F masks are leveraged to weight the noisy signals for estimating the speech and noise PSD matrices $\mathbf{\Phi}_s(f), \mathbf{\Phi}_v(f) \in \mathbb{C}^{N\times N}$, e.g.,

$$\mathbf{\Phi}_s(f) = \frac{1}{\sum_{t=1}^{T} M_s(f,t)} \sum_{t=1}^{T} M_s(f,t)\mathbf{x}(f,t)\mathbf{x}^H(f,t), \quad (1)$$

$$\mathbf{\Phi}_v(f) = \frac{1}{\sum_{t=1}^{T} M_v(f,t)} \sum_{t=1}^{T} M_v(f,t)\mathbf{x}(f,t)\mathbf{x}^H(f,t), \quad (2)$$

where $\mathbf{M}_s, \mathbf{M}_v \in \mathbb{R}^{F\times T}$ are the estimated speech and noise masks, respectively, whose (f,t)-th entry $M_s(f,t)$ and $M_v(f,t)$ typically have values in $[0,1]$, and $\mathbf{x}(f,t) = [X_1(f,t),\ldots,X_N(f,t)]^T \in \mathbb{C}^{N\times 1}$ is the noisy signal snapshot at the (f,t)-th bin. In these methods, the DNNs are trained to predict such masks by minimizing some distance measure (e.g., mean absolute error) between the predicted mask and some pre-defined target mask [18, 19]. As the DNNs only have to estimate the intermediate masks, they may generalize better to unseen acoustic and noise conditions. However, the overall SE performance is often bounded by the later statistical component.

3. PROPOSED METHOD

Our system is depicted in Fig. 2, where the proposed intra-MVDR module(s) shown in Fig. 3 are incorporated into the backbone U-Net direct BF model. It operates in the T-F domain, taking the noisy signal STFTs $\mathbf{X}_i$ as input and predicting the filter weights $\mathbf{W}_i$ for estimating clean speech. Due to the complex nature of STFT, we further utilize complex-valued network operations following [23, 24], while noting that real-valued U-Nets are also adoptable as the backbone model with the same design concept. Next, we describe in detail the major components proposed to improve over the baseline U-Net.

3.1. The intra-MVDR module within direct BF network

As depicted in Fig. 2, the proposed system features intra-MVDR modules embedded in the backbone U-Net direct BF to take advantage of conventional statistical beamforming. Each intra-MVDR module consists of a T-F mask estimation network followed by a mask-based MVDR processing, as Fig. 3 illustrates. Here, the mask estimation network takes the encoder feature maps and predicts the speech and noise masks $\mathbf{M}_s$ and $\mathbf{M}_v$ for estimating the PSD matrices as (1) and (2) for performing MVDR filtering on noisy STFTs.

In Fig. 2, placed between the encoder and decoder units as an intermediate feature enhancer at each level, the intra-MVDR module combines the noisy signal STFTs and encoder feature maps for extracting useful spatial features from multichannel data, subsequently passed to the decoder unit. Notably, the original MVDR was derived from a criterion to minimize the output signal variance constrained on zero speech distortion [21]. Here, MVDR is integrated as a network module and all the learnable parameters are jointly optimized for the final clean signal reconstruction loss as in typical direct BF.

3.2. Exploiting MVDR-filtered signals at all microphones

In typical mask-based neural beamformers where the MVDR is usually the final processing stage, *only the reference channel signal gets enhanced through the MVDR filtering.* While in our method, utilized as an intermediate module we can have more flexibility for exploiting the spatial filtering properties from MVDR by *performing MVDR filtering with respect to* ***all microphones***, i.e., taking each microphone i as the reference channel and generating the corresponding MVDR-filtered signal $\mathbf{Z}_i \in \mathbb{C}^{F\times T}, \forall i = 1,\ldots,N$, where,

$$Z_i(f,t) = \mathbf{h}_i^H(f)\mathbf{x}(f,t), \quad (3)$$

$\forall f,t$, with $\mathbf{x}(f,t)$ being the input noisy signal snapshot and $\mathbf{h}_i(f) \in \mathbb{C}^{N\times 1}$ the vector of MVDR beamformer weights given by [21] :

$$\mathbf{h}_i(f) = \frac{\mathbf{\Phi}_v^{-1}(f)\mathbf{\Phi}_s(f)}{\text{Trace}(\mathbf{\Phi}_v^{-1}(f)\mathbf{\Phi}_s(f))}\mathbf{e}_i, \quad (4)$$

where the PSD matrices $\mathbf{\Phi}_s(f)$, $\mathbf{\Phi}_v(f)$ are computed based on using (1) and (2) and $\mathbf{e}_i = [0,\ldots,1,\ldots,0]^T$ with the 1 located at the i-th position for microphone i. In other words, the MVDR filtering (3) is performed N times to obtain the filtered signals with respect to all N microphones, where each time, the filter coefficients $\mathbf{h}_i(f)$ are computed for each specific channel i using (4). Finally, the generated $\mathbf{Z}_i \in \mathbb{C}^{F\times T}, \forall i = 1,\ldots,N$ are fed into the decoder units as additional spatial features for improved model learning.

3.3. Multi-scale beamforming with intra-MVDR

Bridging the encoder and decoder units of U-Net, the intra-MVDR module naturally fits into a multi-scale design by equipping each level of the U-Net with one intra-MVDR module. As illustrated in Fig. 2, the input STFTs $\mathbf{X}_i$ are downsampled to $\mathbf{X}_i^{'}$, $\mathbf{X}_i^{''}$, and $\mathbf{X}_i^{'''}$ via max-pooling, and subsequently fed into the encoder and the

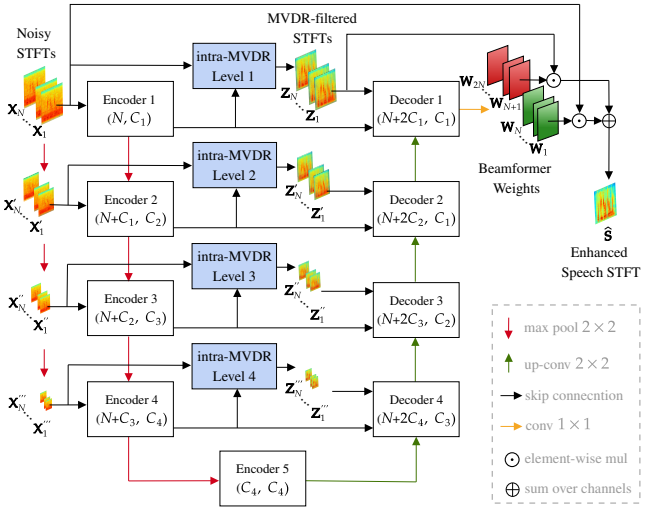

Fig. 2. The proposed MVDR-embedded U-Net beamformer for SE.

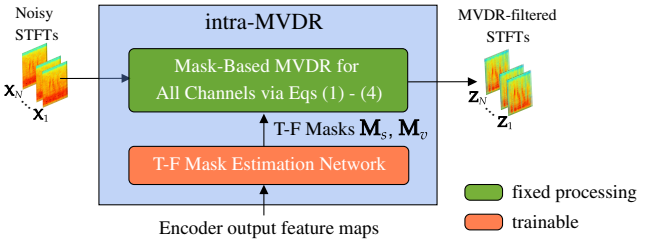

Fig. 3. The proposed intra-MVDR module (Level 1) in details.

intra-MVDR module at the respective level. To be more exact, we sequentially perform 2-D max-pooling operations to the STFT spectrograms in alignment with the downsampling operations in the first half of U-Net. In this way, after the pooling operation, the noisy spectrograms will become the same size as the downsampled feature maps that are passed to the encoder unit of the corresponding level k. Then the N spectrograms can augment the C_k feature maps via concatenation to become $N+C_k$ input feature maps for the encoder. Finally, intra-MVDR is applied on the corresponding noisy spectrograms by taking the encoder output feature maps as input to the T-F mask estimation network. The overall SE model can thus exploit coarse- and fine-grained spatial features from various resolutions.

3.4. Combine MVDR-filtered signals for target reconstruction

Since the MVDR-filtered signals $\mathbf{Z}_i$ are enhanced versions of the noisy inputs, including them in the final filtering stage could provide the model with additional flexibility to estimate the clean signal through refinement. We hence combine $\mathbf{Z}_i$ for reconstructing the target speech at the network output. As depicted in Fig. 2, the U-Net outputs $2N$ sets of filter weights $\mathbf{W}_i$, $i = 1, \ldots, 2N$. The first N sets are the coefficients to be multiplied with the noisy STFTs $\mathbf{X}_i$, same as the typical filter-and-sum scheme; the rest of the filter weights $\mathbf{W}_{N+i}$ are used to weight the MVDR outputs $\mathbf{Z}_i$, $\forall i = 1, \ldots, N$. The enhanced speech is given by the weighted sum:

$$\hat{\mathbf{S}} = \sum_{i=1}^{N} \mathbf{W}_i \odot \mathbf{X}_i + \mathbf{W}_{N+i} \odot \mathbf{Z}_i. \quad (5)$$

Empirically, we have observed (5) leads to better reconstructed speech than only considering filter-and-sum of the noisy STFTs.

Network architecture details: Our modifications to the original U-Net [25] for the model in Fig. 2 are below: Convolutional layers of U-Net are all replaced by *complex* convolutional layers. For the activation function, complex leaky ReLU, i.e., an activation function that applies leaky ReLU on both real and imaginary values, is utilized. The number of feature maps in each layer is also modified. Each encoder or decoder layer consists of two stacks of "(complex) 3×3 convolution→batch normalization→leaky ReLU" with (C_{in}, C_{out}) specifying the number of input and output channels: the convolution layer of the first stack takes C_{in} feature maps and outputs C_{out} feature maps; the convolution layer of the second stack takes C_{out} feature maps and outputs C_{out} feature maps. Regarding the T-F mask estimation network in Fig. 3, it consists of "(complex) 3×3 convolution→batch normalization→leaky ReLU→1×1 convolution→Sigmoid" layers. The 3×3 convolution layer takes C_{out} features and outputs C_{out} features; the 1×1 convolution layer takes C_{out} features and outputs the two T-F masks, where C_{out} is the number of the encoder output feature maps at a given level.

4. EXPERIMENTS

We experimentally show that the proposed components in Section 3 lead to improved SE over baseline methods of Fig. 1 (a) and (b).

Dataset and evaluation: We use the public CHiME-3 dataset [26] which is a 6-microphone recording of talkers speaking in noisy environments, sampled at 16 kHz. It consists of 7,138 and 1,320 simulated utterances for training and testing, respectively. We follow many existing works to take the 5-th channel as the reference microphone for the SE target. For evaluation, we use: **PESQ:** Perceptual Evaluation of Speech Quality [27] (value: -0.5 to 4.5). **STOI:** Short-Time Objective Intelligibility [28] (value: 0 to 1). **SNR:** Signal-to-Noise Ratio. In all metrics, a higher score indicates better SE.

Table 1. Comparison of different multichannel SE schemes. For the direct BF (Fig. 1 (a)) and mask-based MVDR (Fig. 1 (b)) approaches we show results for a base (1.27M) and a larger (1.62M) U-Net models. For our method (Fig. 1 (c)) we present results for incorporating the intra-MVDR modules at different levels into the base (1.27M) U-Net model, corresponding to the system in Fig. 2.

Methods		# Params	PESQ	STOI	SNR
Direct BF	(base)	1.27M	2.39	0.962	17.76
	(larger)	1.62M	2.44	0.965	18.31
Mask-based MVDR	(base)	1.27M	2.00	0.966	16.67
	(larger)	1.62M	2.01	0.966	16.81
Oracle MVDR		-	2.01	0.970	18.42
Direct BF w/ intra-MVDR	Level 1	1.30M	2.55	0.970	18.93
	Levels 1,2	1.38M	2.57	0.973	20.43
	Levels 1,2,3	1.47M	2.60	**0.974**	**20.80**
	Levels 1,2,3,4	1.56M	**2.64**	**0.974**	20.63

Model settings: We set $C_1, C_2, C_3, C_4 = 32, 64, 64, 64$ for the number of channels in Fig. 2, resulting in a base U-Net model of 1.27M parameters. For STFT processing, we use the Hann window with a window size of 1024 and a hop size of 256. During training, 4-second long segments are randomly cropped from the training samples, while during testing the whole utterances are used. For training, the Adam optimizer with a learning rate of 0.001 and decreased to 0.0001 at the 50-th epoch, with a total of 80 epochs, is adopted. A batch size of 4 is used. For direct beamforming schemes, the network is trained to minimize the signal reconstruction loss: $\mathcal{L}(\hat{\mathbf{S}}, \mathbf{S}) = 0.3\|\hat{\mathbf{S}}^{0.3} - \mathbf{S}^{0.3}\|_F^2 + 0.7\||\hat{\mathbf{S}}|^{0.3} - |\mathbf{S}|^{0.3}\|_F^2$, i.e., the combined power-law compressed mean squared error loss in [29].

We compare the SE performance of the following cases based on using the same backbone (1.27M) U-Net model:

i) **Direct BF (Fig. 1 (a)):** the U-Net is trained to directly estimate the clean speech by predicting the beamformer filter weights for filtering the input noisy signals

ii) **Mask-based MVDR (Fig. 1 (b)):** the U-Net is trained to estimate the speech and noise ideal ratio masks [18]. During inference, the estimated masks are used in MVDR filtering

iii) **Direct BF w/ intra-MVDR (Fig. 1 (c)):** the proposed intra-MVDR module(s) embedded in the U-Net direct BF

For the two baseline approaches of direct BF and mask-based MVDR, we also provide results with using a larger U-Net model (1.62M) where $C_1, C_2, C_3, C_4 = 36, 72, 72, 72$ for comparison.

Results: Table 1 presents the results. We first compare the two baseline approaches and see that the direct BF performs generally better than the mask-based MVDR, as the network is trained to directly reconstruct clean speech for effectiveness. On the other hand, while the mask-based MVDR may be robust, its performance is upper bounded by the oracle MVDR performance (also shown in the table) which uses the ground truth signal PSD matrices for computing the optimal filter. Finally, the proposed direct BF utilizing intra-MVDR only at the first level (1.30M) already attains higher scores than the two baselines. We also see that equipping the intra-MVDR blocks at subsequent resolutions (Levels) of U-Net further improves the performance, indicating that multi-scale spatial features are helpful. Note that the improvement is not purely due to the increased model size of the added intra-MVDR modules, as we can see despite using a larger model size (1.62M), the two baseline approaches do not match the proposed method's performance. The results reveal that by combining the merits of statistical filters (robustness)

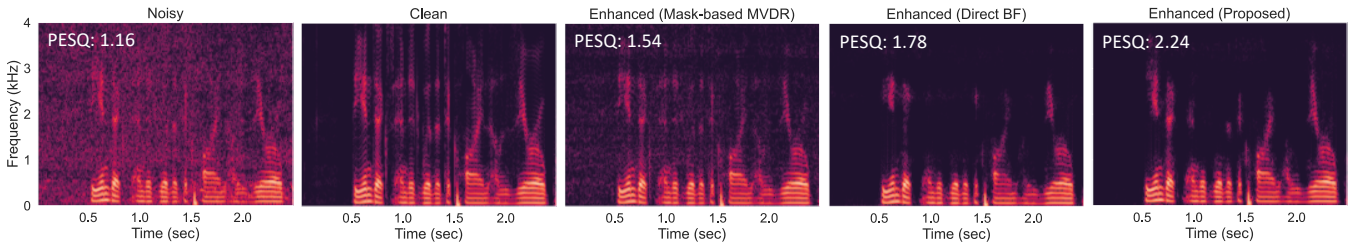

Fig. 4. Visualization of SE outputs. The proposed method has less residual noise (i.e., better noise suppression) as compared to Mask-based MVDR while preserving more speech components (i.e., less speech distortion) as compared to Direct BF, achieving the best speech quality.

and DNN direct BF (effectiveness), we can attain SE improvements that are not achievable by simply making the baseline network larger. Exemplary SE outputs processed by our full system (1.56M) and the two larger baselines (1.62M) for a noisy utterance taken from the CHiME-3 test set are visualized in Fig. 4 to facilitate comparison.

To better quantify the proposed method performance, we compare our full system (i.e., the 1.56M model in Table 1) results with existing multichannel SE systems also evaluated on the CHiME-3 test set, including **Neural BF** [16]: An MVDR beamformer with mask estimation through bidirectional-LSTM. **MVDR**$_{GC}$ [30]: An MVDR beamformer using a neural network-based method to identify and correct phase errors in the steering vector. **rSDFCN** [31]: A time-domain, fully convolutional network (FCN) with sinc and dilated convolutional layers for multichannel SE. **CA Dense U-Net** [32]: A time-frequency domain multichannel SE model that combines the merits of DenseNet, U-Net, and channel attention (CA) mechanism. **IC Conv-TasNet** [33]: A multichannel SE system based on a fully convolutional time-domain audio separation network by exploiting inter-channel relationships. In Table 2 we report the PESQ, STOI, and SNR taken from the corresponding papers, and the missing entries in the table indicate that the metric is not reported in the reference paper. It can be seen that the proposed method outperforms the other approaches in STOI and SNR, and is on par with the IC Conv-TasNet in PESQ despite using a smaller model, demonstrating its capability of achieving efficient SE.

Table 2. Comparison with existing multichannel SE methods.

Methods	# Params	PESQ	STOI	SNR
Noisy	-	1.27	0.870	6.50
Neural BF [16]	-	2.29	-	15.12
MVDR$_{GC}$ [30]	0.5M	-	0.952	-
rSDFCN [31]	2.1M	2.15	0.937	-
CA Dense U-Net [32]	>20M	2.44	-	18.64
IC Conv-TasNet [33]	1.67M	**2.67**	0.973	19.67
Proposed	1.56M	2.64	**0.974**	**20.63**

Besides the above quality/intelligibility scores, we evaluate the SE model as a front-end denoiser for automatic speech recognition (ASR) under noisy environments in Table 3. To this end, we pre-process the noisy CHiME-3 multichannel data through the well-trained SE model and feed the denoised audio separately to three pre-trained ASR engines taken from the NVIDIA NeMo toolkit[1]: Model 1: *Conformer-CTC* [34], Model 2: *Citrinet* [35], and Model 3: *Quartznet* [36]. We report the word error rate (WER) and character error rate (CER) for each ASR engine outcome. To demonstrate the advantages of the proposed method, we compare our full system with one existing direct BF approach, i.e., the Filter-and-Sum Network (FaSNet) [37]. We trained the FaSNet by ourselves[2] using the same signal reconstruction loss as for ours for fairness. The results show that our method achieves lower WER and CER than FaSNet across all three ASR engines with a smaller model size, indicating its efficacy for improved machine listening capabilities in noise.

Table 3. Comparison of SE models as a front-end denoiser for ASR.

Front-ends	# Params	WER / CER (%)		
		ASR Model 1	ASR Model 2	ASR Model 3
Unprocessed	-	7.40 / 4.25	9.18 / 5.64	16.75 / 8.28
FaSNet [37]	2.76M	5.21 / 2.63	5.65 / 3.41	10.20 / 4.87
Proposed	1.56M	**3.81 / 1.96**	**3.54 / 2.33**	**6.31 / 3.06**

With an aim to further validate the proposed method having the virtues of both mask-based statistical beamforming and DNN direct BF, we utilize Pyroomacoustics [38] to simulate training and test data for demonstrating effectiveness (on seen or similar acoustics and noise conditions) and robustness (generalization to unseen conditions) using speech corpus from the AVSpeech dataset [39]. We take 8308 speech utterances for training and 1099 for testing, each mixed with four types of noise profiles downloaded from YouTube. We create two test sets, with Test Set-1 having the same acoustics and noise circumstances as training, i.e., room sizes, reverberation times, noise types, i.e., {blender, vacuum, washer, baby cry}, and source locations. Meanwhile, Test Set-2 has different background noise types from training, i.e., {dog barking, kids playing, hair dryer, food sizzling}, and different room sizes, reverberation times, and source locations. The results in terms of PESQ of a 4-mic planar array are shown in Table 4. Compared to the two baselines, the proposed method achieves the highest result for Test Set-1, demonstrating effectiveness to seen conditions, and has less deteriorating performance on Test Set-2, suggesting robustness to unseen conditions.

Table 4. PESQ scores for comparing effectiveness on test data with seen room/noise conditions and robustness to unseen conditions.

Methods	# Params	Seen Cond.	Unseen Cond.
Noisy	-	1.21	1.22
Mask-based MVDR	1.62M	1.71	1.55
Direct BF	1.62M	2.02	1.66
Proposed	1.56M	**2.13**	**1.76**

5. CONCLUSION

In this paper, we presented a novel integration of DNN direct beamforming and mask-based statistical beamforming by introducing the intra-MVDR module embedded in a U-Net design. The new model encompasses the merits of the two method types, leading to the proposed MVDR-embedded U-Net beamformer better exploiting multi-scale spatial features. We showed that by incorporating intra-MVDR modules, improved SE effectiveness and robustness to seen and unseen room acoustics and noise conditions can be efficiently achieved.

[1]https://github.com/NVIDIA/NeMo

[2]We use the FaSNet-TAC model from: https://github.com/yluo42/TAC

6. REFERENCES

[1] K. Tesch and T. Gerkmann, "Insights into deep non-linear filters for improved multi-channel speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 563–575, 2022.

[2] K. Tan, Z.-Q. Wang, and D. Wang, "Neural spectrospatial filtering," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 605–621, 2022.

[3] Y. Xu, Z. Zhang, M. Yu, S.-X. Zhang, and D. Yu, "Generalized spatio-temporal RNN beamformer for target speech separation," in *Annual Conf. Int. Speech Comm. Assoc. (Interspeech)*, 2021, pp. 3076–3080.

[4] R. Gu, S.-X. Zhang, Y. Zou, and D. Yu, "Complex neural spatial filter: Enhancing multi-channel target speech separation in complex domain," *IEEE Signal Process. Lett.*, vol. 28, pp. 1370–1374, 2021.

[5] A. Aroudi and S. Braun, "DBnet: DOA-driven beamforming network for end-to-end farfield sound source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2021, pp. 211–215.

[6] Y. Xu *et al.*, "Neural spatio-temporal beamformer for target speech separation," in *Proc. Annual Conf. Int. Speech Comm. Assoc. (Interspeech)*, 2020, pp. 56–60.

[7] X. Xiao *et al.*, "Deep beamforming networks for multi-channel speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2016, pp. 5745–5749.

[8] T. N. Sainath *et al.*, "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 5, pp. 965–979, 2017.

[9] Z. Meng, S. Watanabe, J. R. Hershey, and H. Erdogan, "Deep long short-term memory adaptive beamforming networks for multichannel robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2017, pp. 271–275.

[10] Y. Koyama and B. Raj, "W-Net BF: DNN-based beamformer using joint training approach," *arXiv preprint arXiv:1910.14262*, 2019.

[11] A. Aroudi, S. Uhlich, and M. F. Font, "TRUNet: Transformer-recurrent-U network for multi-channel reverberant sound source separation," in *Proc. Annual Conf. Int. Speech Comm. Assoc. (Interspeech)*, 2021, pp. 911–915.

[12] A. Li, G. Yu, C. Zheng, and X. Li, "TaylorBeamformer: Learning all-neural beamformer for multi-channel speech enhancement from Taylor's approximation theory," in *Proc. Annual Conf. Int. Speech Comm. Assoc. (Interspeech)*, 2022, pp. 5413–5417.

[13] A. Li, W. Liu, C. Zheng, and X. Li, "Embedding and beamforming: All-neural causal beamformer for multichannel speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2022, pp. 6487–6491.

[14] Y. Yang, C. Quan, and X. Li, "McNet: Fuse multiple cues for multichannel speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2023.

[15] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2016, pp. 196–200.

[16] H. Erdogan, J. R. Hershey, S. Watanabe, M. Mandel, and J. Le Roux, "Improved MVDR beamforming using single-channel mask prediction networks," in *Proc. Annual Conf. Int. Speech Comm. Assoc. (Interspeech)*, 2016, pp. 1981–1985.

[17] T. Ochiai, S. Watanabe, T. Hori, J. R. Hershey, and X. Xiao, "Unified architecture for multichannel end-to-end speech recognition with neural beamforming," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 8, pp. 1274–1288, 2017.

[18] Z.-Q. Wang and D. Wang, "Mask weighted STFT ratios for relative transfer function estimation and its application to robust ASR," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2018, pp. 5619–5623.

[19] S. Chakrabarty and E. A. P. Habets, "Time-frequency masking based online multi-channel speech enhancement with convolutional recurrent neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 4, pp. 787–799, 2019.

[20] C.-H. Lee, C. Yang, Y. Shen, and H. Jin, "Improved mask-based neural beamforming for multichannel speech enhancement by snapshot matching masking," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2023.

[21] 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, 2009.

[22] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 18–30, 2015.

[23] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex U-Net," *arXiv preprint arXiv:1903.03107*, 2019.

[24] Y. Hu *et al.*, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Proc. Annual Conf. Int. Speech Comm. Assoc. (Interspeech)*, 2020, pp. 2472–2476.

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Medical Image Comput. Comput.-Assist. Interv. (MICCAI)*, 2015, pp. 234–241.

[26] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Analysis and outcomes," *Comput. Speech Lang.*, vol. 46, pp. 605–626, 2017.

[27] ITU-T Recommendation P.862.2, "Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs," *Int. Telecomm. Union*, 2005.

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

[29] S. Braun and I. Tashev, "A consolidated view of loss functions for supervised deep learning-based speech enhancement," in *Proc. Int. Conf. Telecomm. Signal Process. (TSP)*, 2021, pp. 72–76.

[30] S. Bu, Y. Zhao, and M.-Y. Hwang, "A novel method to correct steering vectors in MVDR beamformer for noise robust ASR," in *Proc. Annual Conf. Int. Speech Comm. Assoc. (Interspeech)*, 2019, pp. 4280–4284.

[31] C.-L. Liu, S.-W. Fu, Y.-J. Li, J.-W. Huang, H.-M. Wang, and Y. Tsao, "Multichannel speech enhancement by raw waveform-mapping using fully convolutional networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1888–1900, 2020.

[32] B. Tolooshams, R. Giri, A. H. Song, U. Isik, and A. Krishnaswamy, "Channel-attention dense U-Net for multichannel speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2020, pp. 836–840.

[33] D. Lee, S. Kim, and J.-W. Choi, "Inter-channel Conv-TasNet for multichannel speech enhancement," *arXiv preprint arXiv:2111.04312*, 2021.

[34] A. Gulati *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Annual Conf. Int. Speech Comm. Assoc. (Interspeech)*, 2020, pp. 5036–5040.

[35] S. Majumdar, J. Balam, O. Hrinchuk, V. Lavrukhin, V. Noroozi, and B. Ginsburg, "Citrinet: Closing the gap between non-autoregressive and autoregressive end-to-end models for automatic speech recognition," *arXiv preprint arXiv:2104.01721*, 2021.

[36] S. Kriman *et al.*, "Quartznet: Deep automatic speech recognition with 1D time-channel separable convolutions," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2020, pp. 6124–6128.

[37] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2020, pp. 6394–6398.

[38] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A Python package for audio room simulation and array processing algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2018, pp. 351–355.

[39] A. Ephrat *et al.*, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Trans. Graphics*, vol. 37, no. 4, pp. 109:1–109:11, 2018.