Neural Network-Assisted Joint DOA Estimation and Beamforming with First-Order Reflection Modeling

Yichen Yang*†, Chao Pan*, Qiang Gao*, Jacob Benesty‡, Shoji Makino†, Jingdong Chen*
*Center of Intelligent Acoustics and Immersive Communications, Northwestern Polytechnical University, Xi'an, China
†Graduate School of Information, Production and Systems, Waseda University, Kitakyushu, Japan
‡INRS-EMT, University of Quebec, Montreal, QC H5A 1K6, Canada

***Abstract*—Source signal extraction and direction-of-arrival (DOA) estimation are core tasks in microphone array signal processing. However, their performance degrades significantly in reverberant environments: DOA estimation struggles to distinguish the direct path from strong reflections, while source extraction fails to effectively exploit the spatial cues in early reflections. This paper proposes a novel formulation of the source array-manifold vector (AMV) as a linear combination of free-field AMVs corresponding to the direct path and the four first-order reflections from the surrounding walls. A neural network is developed to jointly estimate the DOA and the combination weights of these component AMVs. The resulting composite AMV is then used in a minimum variance distortionless response (MVDR) beamformer for enhanced source extraction. Additionally, we introduce a location-based training strategy that enables end-to-end learning to automatically identify the direct-path direction. Simulation results show that the proposed method significantly improves both speech quality and DOA estimation accuracy in reverberant environments.**

***Index Terms*—Beamforming, real-and-virtual source localization, DOA estimation, deep neural network.**

I. Introduction

Direction-of-arrival (DOA) estimation and source extraction are essential techniques in a wide range of applications [1]–[6], including video broadcasting, acoustic scene analysis, surveillance, human-robot interaction, and speech recognition. Accurate DOA estimation enables dynamic steering, while effective source extraction enhances signal clarity and intelligibility by isolating the target source from noisy mixtures. This paper addresses the joint integration of DOA estimation and source extraction, with a focus on improving performance under reverberant acoustic environments.

Over recent years, significant efforts have been made to integrate parameter learning with microphone array-based source extraction, giving rise to the field of neural beamforming [7]–[14]. For example, In [11], a data-driven minimum-variance-distortionless-response (MVDR) beamformer was proposed, in which both the source array-manifold vector (AMV) and the noise covariance matrix are learned directly from array observations. In [10], a method was introduced for learning a time-varying interference AMV based on a given source AMV and a noise coherence matrix. A two-stage neural beamforming approach was developed in [15], which combines multiple beamforming techniques [16], [17] with learned source directions using a delay-and-sum beamformer as the backbone. The work in [9] approximates the source AMV using only the direct-path free-field model, ignoring early reflections though this simplification may lead to performance degradation in reverberant environments [18].

Despite substantial progress, most neural beamforming methods are designed primarily for source extraction and often overlook simultaneous DOA estimation. This leads to the following limitations.

- **Lack of explicit DOA representation**: DOA information is not explicitly encoded in the learned parameters, preventing direct extraction of DOA estimates through array processing.
- **Oversimplified AMV models**: The learned AMVs, when tied to fixed or encoded DOAs, are often too simplistic to fully leverage the spatial diversity of array observations, particularly in the presence of reverberation.
- **Inability to resolve direct-path DOA**: These methods typically do not distinguish between different DOAs, making it difficult to isolate the direct path: a capability crucial for applications such as dynamic camera or beamformer steering.

In this paper, we propose a method that models the target speech AMV as a linear combination of AMVs corresponding to the direct path and first-order reflections, with associated transfer functions that capture key propagation characteristics. A multi-task neural network is employed to jointly estimate the DOAs of individual sources and their respective transfer functions from the observed microphone signals. The resulting AMV is then used to construct an MVDR beamformer for source extraction. Simulation results demonstrate the strong potential of the proposed approach in reverberant environments.

II. Signal Model and Problem Formulation

Let us consider a reverberant environment, where the speech source signal is captured by a microphone array consisting of M elements, then the noisy observation in the short-time Fourier transform (STFT) domain can be expressed as

$$\begin{aligned}\mathbf{y}(\omega,t) &= \mathbf{x}(\omega,t)+\mathbf{v}(\omega,t)\\ &= \mathbf{a}(\omega,t)S(\omega,t)+\mathbf{r}(\omega,t)+\mathbf{v}(\omega,t),\end{aligned} \tag{1}$$

where $\mathbf{y}(\omega,t)\in\mathbb{C}^{M\times 1}$, $\mathbf{x}(\omega,t)\in\mathbb{C}^{M\times 1}$, and $\mathbf{v}(\omega,t)\in\mathbb{C}^{M\times 1}$ are, respectively, the noisy, clean speech, and additive noise received at the array, ω is the index of frequency bins, t is the time-frame index, $S(\omega,t)$ represents the source signal, $\mathbf{a}(\omega,t)$ is the associated AMV that captures both the direct path

This work was supported in part by the National Key Research and Development Program of China under Grant No. 2024YFF0505502 and in part by the NSFC Grants: 62171373, 61831019, and 62192713, and in part by the China Scholarship Council (CSC).

of the actual source and early reflections modeled as virtual sources, and $\mathbf{r}(\omega,t)$ is the late reverberation not accounted for in the AMV.

To extract the source signal of interest from the observation signal vector, $\mathbf{y}(\omega,t)$, a spatial filter, $\mathbf{h}(\omega,t)$, is applied as

$$Z(\omega,t) = \mathbf{h}^H(\omega,t)\mathbf{y}(\omega,t) \tag{2}$$

$$= Z_{\text{fd}}(\omega,t) + Z_{\text{rr}}(\omega,t) + Z_{\text{rn}}(\omega,t), \tag{3}$$

where $Z_{\text{fd}}(\omega,t) = \mathbf{h}^H(\omega,t)\mathbf{a}(\omega,t)S(\omega,t)$, $Z_{\text{rr}}(\omega,t) = \mathbf{h}^H(\omega,t)\mathbf{r}(\omega,t)$, and $Z_{\text{rn}}(\omega,t) = \mathbf{h}^H(\omega,t)\mathbf{v}(\omega,t)$ are the filtered desired signal, residual reverberation, and residual noise, respectively. Over the past decades, numerous methods have been developed to design the spatial filter in (2), resulting in a variety of beamforming approaches [18]–[25]. In this work, we adopt the widely used and highly effective MVDR beamformer for source signal extraction, which is expressed as

$$\mathbf{h}_{\text{MVDR}}(\omega,t) = \frac{\mathbf{\Phi}_{\mathbf{u}}^{-1}(\omega)\mathbf{a}(\omega,t)}{\mathbf{a}^H(\omega,t)\mathbf{\Phi}_{\mathbf{u}}^{-1}(\omega)\mathbf{a}(\omega,t)}, \tag{4}$$

where $\mathbf{\Phi}_{\mathbf{u}}(\omega) = \mathbf{\Phi}_{\mathbf{r}}(\omega) + \mathbf{\Phi}_{\mathbf{v}}(\omega)$ represents the spatial covariance matrix of the undesired signal components, including late reverberation $\mathbf{r}(\omega,t)$ and additive noise $\mathbf{v}(\omega,t)$. As evident from (4), the AMV, $\mathbf{a}(\omega,t)$, plays a central role in the design of the MVDR beamformer.

As shown in [26], the AMV governs the extraction of the target source signal from array observations. This AMV is generally influenced not only by the direct-path component but also by early reflections and source statistics. In this work, we parameterize the source AMV using the DOAs of the direct path and first-order reflections of the desired source. This formulation enables the integration of DOA estimation and beamforming, leading to improved performance in reverberant environments. Further details are provided in the following section.

III. Proposed Neural Estimator

To estimate the AMV, $\mathbf{a}(\omega,t)$, required for the optimal MVDR beamformer, we first model it as a weighted sum of AMVs corresponding to the direct path and early reflections, modulated by their respective transfer functions. A multi-task neural network is then employed to jointly estimate the DOAs and transfer functions of these components. The estimated AMV is subsequently used in (4) to compute the optimal beamformer. An overview of the proposed system is illustrated in Fig. 1.

A. Parameterization of AMV

By explicitly incorporating the DOAs of the direct path and early reflections, the AMV can be parameterized as

$$\mathbf{a}(\omega,t) = \sum_{n=0}^{N} \mathbf{d}(\omega,\theta_n) A_{n,1}(\omega,t), \tag{5}$$

where

$$\mathbf{d}(\omega,\theta_n) = \begin{bmatrix} 1 & e^{-\jmath\omega\tau_2(\theta_n)} & \cdots & e^{-\jmath\omega\tau_M(\theta_n)} \end{bmatrix}^T \tag{6}$$

denotes the time-invariant free-field AMV corresponding to θ_n, $\tau_m(\theta_n)$ represents the time delay at the mth microphone for the nth path (also referred to as the nth virtual source) relative to the reference microphone ($m=1$), N is the total number of propagation paths, $n = 0, 1, 2, \ldots, N$ indexes both the direct path ($n=0$) and early reflection paths, and $A_{n,1}(\omega,t)$ is the corresponding time-variant transfer function at the reference channel.

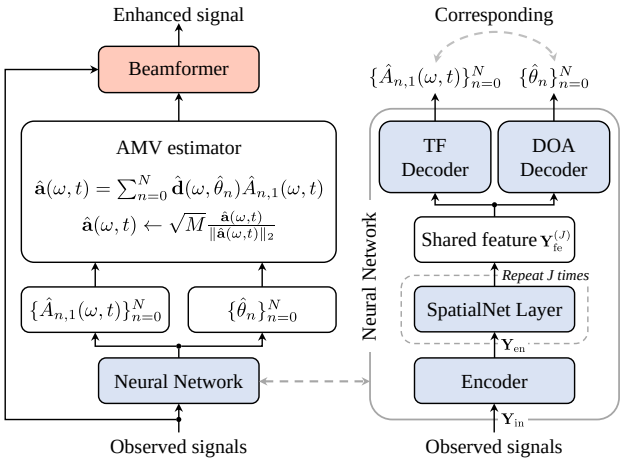

Fig. 1. Diagram of the proposed neural estimator.

B. Model Architecture

To jointly capture the DOAs of both the direct and early reflection paths along with their corresponding transfer functions, a multi-task neural network is employed. It consists of two downstream tasks: one estimates the set of $\{\theta_n\}_{n=0}^{N}$ and the other estimates the corresponding transfer functions $\{A_{n,1}(\omega,t)\}_{n=0}^{N}$. Both tasks leverage shared features learned directly from the microphone array observations.

Specifically, the real-valued input feature of the network, denoted as $\mathbf{Y}_{\text{in}} \in \mathbb{R}^{\Omega\times\mathcal{T}\times 2M}$ is first derived from the observed signals, which consists of $\mathcal{T}$ neighboring frames, i.e.,

$$\mathbf{Y}_{\text{in}} = \{\mathbf{Y}(\omega)\}_{\omega=1}^{\Omega}, \tag{7}$$

$$\mathbf{Y}(\omega) = [\overline{\mathbf{y}}(\omega,t) \;\; \overline{\mathbf{y}}(\omega,t-1) \;\; \cdots \;\; \overline{\mathbf{y}}(\omega,t-\mathcal{T}+1)]^T, \tag{8}$$

$$\overline{\mathbf{y}}(\omega,t) = \left[\mathbf{y}_{\text{re}}^T(\omega,t) \quad \mathbf{y}_{\text{im}}^T(\omega,t)\right]^T, \tag{9}$$

where $\mathbf{y}_{\text{re}}(\omega,t)$ and $\mathbf{y}_{\text{im}}(\omega,t)$ are the real and imaginary parts of $\mathbf{y}(\omega,t)$, respectively, and Ω is the number of frequency bins.

1) Encoder: The encoder consist of a set of narrow band one dimensional convolution layer (Conv1d) along the time frame axis. Given $\mathbf{Y}_{\text{in}}$ as input, the encoder produces an output expressed as

$$\mathbf{Y}_{\text{en}} = \text{Conv1d}\left[\mathbf{Y}_{\text{in}}\right], \tag{10}$$

where $\mathbf{Y}_{\text{en}} \in \mathbb{R}^{\Omega\times\mathcal{T}\times F_{\mathbf{e}}}$ is the output of the encoder and $F_{\mathbf{e}}$ is the feature dimension of the Conv1d.

2) Feature extraction: To extract features from $\mathbf{Y}_{\text{en}}$, this work employs J SpatialNet layers [27] to construct the feature extraction block. Each layer consists of a cross-band and a narrow-band module designed to capture time-frequency domain characteristics. For simplicity, the output of the jth SpatialNet layer, denoted as $\mathbf{Y}_{\text{fe}}^{(j)} \in \mathbb{R}^{\Omega\times\mathcal{T}\times F_{\mathbf{e}}}$ for $j =$

$1, 2, \ldots, J$, is expressed as

$$\mathbf{Y}_{\text{fe}}^{(j)} = \text{SpatialNet}_j \left[\mathbf{Y}_{\text{fe}}^{(j-1)} \right], \tag{11}$$

where $\mathbf{Y}_{\text{fe}}^{(0)} \triangleq \mathbf{Y}_{\text{en}}$ and $\text{SpatialNet}_j[\cdot]$ indicates the jth SpatialNet layer [27]. After J SpatialNet layers, the output, $\mathbf{Y}_{\text{fe}}^{(J)}$, is treated as a shared feature for all downstream tasks.

3) Decoder for DOA estimation: Prior to being fed into the decoder, layer normalization is applied to the shared feature, followed by mean pooling along both the frequency and frame dimensions to produce a feature vector of shape $F_{\mathbf{e}}$. The DOA estimation decoder consists of three fully connected linear layers, each interleaved with batch normalization and ReLU activation functions. The dimensionality of each linear layer is set to $2F_{\mathbf{e}}$. To address angular discontinuities and improve training stability, a sine-cosine encoding scheme is employed to represent the directions. The embedding corresponding to a direction θ is given by

$$\mathbf{c}(\theta) = \begin{bmatrix} \sin\theta & \cos\theta \end{bmatrix}^T, \tag{12}$$

where the mapping $\mathbf{c}(\theta)$ is a numerically continuous and uniquely defined function for each $\theta \in [0°, 360°]$. Accordingly, an additional linear layer is used to project the feature vector to a dimension of $2(N+1)$, followed by a hyperbolic tangent activation to constrain the output within the range $[-1, 1]$. This process can be summarized as

$$\left\{ \mathbf{c}(\hat{\theta}_n) \right\}_{n=0}^{N} = \text{Decoder}_{\text{DOA}} \left[\mathbf{Y}_{\text{fe}}^{(J)} \right] \tag{13}$$

and the estimated DOAs are calculated as

$$\hat{\theta}_n = \arctan \left[\frac{\mathbf{c}^T(\hat{\theta}_n)\mathbf{e}_1}{\mathbf{c}^T(\hat{\theta}_n)\mathbf{e}_2} \right], \tag{14}$$

where $\mathbf{e}_1 = [1 \;\; 0]^T$ and $\mathbf{e}_2 = [0 \;\; 1]^T$.

4) Decoder for transfer function estimation: The decoder for transfer function estimation consists of two components: an additional SpatialNet layer, followed by a fully connected linear layer that projects the feature dimension to $\Omega \times \mathcal{T} \times 2(N+1)$, representing the real and imaginary parts. This can be expressed as

$$\left\{ \hat{\mathbf{A}}_{n,\text{re}}, \hat{\mathbf{A}}_{n,\text{im}} \right\}_{n=0}^{N} = \text{Decoder}_{\text{TF}} \left[\mathbf{Y}_{\text{fe}}^{(J)} \right] \tag{15}$$

and the transfer function for nth source $\hat{\mathbf{A}}_n \in \mathbb{C}^{\Omega \times \mathcal{T}}$ is obtained as

$$\begin{aligned} \hat{\mathbf{A}}_n &= \hat{\mathbf{A}}_{n,\text{re}} + \jmath \hat{\mathbf{A}}_{n,\text{im}} \\ &\triangleq \begin{bmatrix} \hat{A}_{n,1}(1,t) & \cdots & \hat{A}_{n,1}(1, t-\mathcal{T}+1) \\ \vdots & \ddots & \vdots \\ \hat{A}_{n,1}(\Omega,t) & \cdots & \hat{A}_{n,1}(\Omega, t-\mathcal{T}+1) \end{bmatrix}. \end{aligned} \tag{16}$$

C. NN-Based MVDR Beamforming

With the learning-based DOAs and transfer functions for the direct and N early reflection paths, the AMV can be rewritten from (5) as

$$\hat{\mathbf{a}}(\omega, t) = \sum_{n=0}^{N} \mathbf{d}(\omega, \hat{\theta}_n) \hat{A}_{n,1}(\omega, t), \tag{17}$$

where $\hat{\theta}_n$ and $\hat{A}_{n,1}(\omega, t)$ are the DOA and transfer function estimates obtained through (14) and (16), respectively, and $\mathbf{d}(\omega, \hat{\theta}_n)$ is calculated according to (6). Hence, the neural network (NN) assisted MVDR beamformer is obtained by substituting the DOA, transfer function and AMV estimates into (4). The enhanced signal is finally obtained by applying the MVDR beamformer to the array observation, $\mathbf{y}(\omega, t)$, as shown in (2).

D. Training Strategy

The entire multi-task neural network is trained using two loss functions: a mean-squared error (MSE) loss for DOA estimation, and a time-domain scale-invariant signal-to-distortion ratio (SI-SDR) loss [28] for speech enhancement, which implicitly guides the estimation of the transfer function.

To resolve the permutation ambiguity associated with first-order early reflections, location-based training (LBT) [29] is employed. LBT enforces a predefined order in the output alignment, corresponding to reflections from the left wall, right wall, front wall, and back wall. With this ordering, the DOA estimation loss using the MSE criterion can be expressed as

$$\begin{aligned} \mathcal{L}_{\text{DOA}} = \frac{1}{2} \Bigg\{ & \frac{1}{2} \left\| \mathbf{c}(\theta_0) - \mathbf{c}(\hat{\theta}_0) \right\|_2^2 \\ & + \frac{1}{2N} \sum_{n=1}^{N} \left\| \mathbf{c}(\theta_n) - \mathbf{c}(\hat{\theta}_n) \right\|_2^2 \Bigg\}, \end{aligned} \tag{18}$$

where $\|\cdot\|_2$ represents the ℓ_2 norm. As seen in (18), the direct-path component and all early reflections are treated with equal importance during training, despite the fact that the direct-path signal $\mathbf{d}(\omega, \theta_0) A_{0,1}(\omega, t) S(\omega, t)$ typically carries more energy in the observation mixture. In addition, the loss function for speech enhancement, based on the SI-SDR criterion, is given by

$$\mathcal{L}_{\text{SI-SDR}} = -10 \log_{10} \left[\frac{\mathbb{E}\left(|\alpha \mathbf{s}|^2 \right)}{\mathbb{E}\left(|\alpha \mathbf{s} - \mathbf{z}|^2 \right)} \right], \tag{19}$$

where $\mathbf{s}$ and $\mathbf{z}$ are the time-domain source and enhanced signals constructed from its STFT domain estimation $S(\omega, t)$ and $Z(\omega, t)$, respectively, and $\alpha = \mathbf{z}^T \mathbf{s} / \|\mathbf{s}\|_2^2$ is a scaling factor. Finally, the entire loss function is defined as

$$\mathcal{L} = \mathcal{L}_{\text{SI-SDR}} + \lambda \mathcal{L}_{\text{DOA}}, \tag{20}$$

where λ is a weighting factor. Using this joint loss function, the DOAs and their corresponding transfer functions, including those for both the direct path and early reflections of the target speech, can be directly estimated from the observed signals.

IV. Experimental Evaluation

A. Experimental Setup

A simulated reverberant room environment is considered, where a target speech signal is captured by a microphone array in the presence of additive noise. The room dimensions, including length, width, and height, are uniformly sampled from the ranges $[8 \text{ m}, 12 \text{ m}]$, $[6 \text{ m}, 8 \text{ m}]$, and $[3 \text{ m}, 4 \text{ m}]$, respectively. The microphone array consists of seven channels: six microphones arranged in a circle with a diameter of 8.5 cm, and one microphone placed in the center. The array

TABLE I
THE AVERAGE SDR (DB) AND PESQ FOR THE ENHANCED SIGNAL OF THE PROPOSED AND BASELINE METHODS.

Methods	SDR (dB)						PESQ					
RT_{60}	300 ms			600 ms			300 ms			600 ms		
SNR	25 dB	15 dB	5 dB	25 dB	15 dB	5 dB	25 dB	15 dB	5 dB	25 dB	15 dB	5 dB
Unprocessed	15.64	12.49	4.66	10.26	9.02	3.66	1.92	1.38	1.08	1.49	1.27	1.08
DBNet [9]	23.25	20.13	12.48	19.19	17.14	11.25	2.98	2.13	1.32	2.38	1.87	1.27
Proposed	**23.46**	**21.74**	**16.71**	**20.42**	**19.19**	**15.40**	**3.89**	**3.27**	**2.09**	**3.49**	**3.00**	**1.97**

TABLE II
THE AVERAGE RMSE OF THE DOA ESTIMATION ERROR.

Methods	RMSE (deg)					
RT_{60}	300 ms			600 ms		
SNR	25 dB	15 dB	5 dB	25 dB	15 dB	5 dB
[38]	37.23	39.45	40.82	35.68	38.40	40.39
[39]	2.70	2.76	3.15	2.72	2.78	3.28
Proposed	**2.37**	**2.41**	**2.55**	**2.37**	**2.41**	**2.58**

is randomly positioned within a 1 m $\times$ 1 m square region in the center of the room, at a fixed height of 1.5 m. A point source representing the desired speech is placed at a random DOA and at a random distance from the array center within the range $[1.5\text{ m}, 2\text{ m}]$. The clean speech signals are randomly selected from the Wall Street Journal (WSJ0)-2mix dataset [30], which contains $20,000$, $5,000$, and $3,000$ speech clips for training, validation, and test, respectively. The value of $\mathcal{T}$ is set to 2, following the setup in [10]. All time-domain signals are transformed into the STFT domain using a 32-ms Hann window with 50% overlap, where the fast Fourier transform (FFT) length is set equal to the frame length. Room impulse responses (RIRs) between the source and microphones are generated using the image-source method [31], [32], and in the training stage, the room surface reflection coefficients are calculated based on reverberation times (RT_{60}) [33], randomly drawn from $[200\text{ ms}, 900\text{ ms}]$. Additive noise includes diffuse noise [34] and white Gaussian noise, mixed with a random energy ratio between $[15\text{ dB}, 25\text{ dB}]$. The overall signal-to-noise ratio (SNR) is randomly selected from the range $[5\text{ dB}, 35\text{ dB}]$. In this work, N is set to 4, corresponding to the first-order reflections from the left, right, front, and back walls.

The Adam optimizer [35] is employed to train the network from scratch with an initial learning rate of 10^{-3}, which decays by a factor of 0.99 at each epoch. The weighting coefficient λ in (20) is set to 10. The number of SpatialNet layers J used in the feature extraction module is set to 6, and the feature dimension F_{e} is fixed at 96. Training is conducted for up to 50 epochs, with early stopping applied if the validation loss does not improve for 10 consecutive epochs.

B. Results and Discussion

Figure 2 presents an example of the log-magnitude spectra for the observation, clean, and enhanced signals, and the frame-wise beampattern obtained using the proposed learning-based AMV approach. The spectra demonstrate that undesired components, such as background noise and late reverberation, are effectively suppressed by the MVDR beamformer when guided by the learned AMVs. Additionally, the frame-wise beampattern reveals that, beyond the direct-path signal, the proposed system also accounts for four first-order reflections from the surrounding walls, taking advantage of these strong early reflections to enhance signal quality.

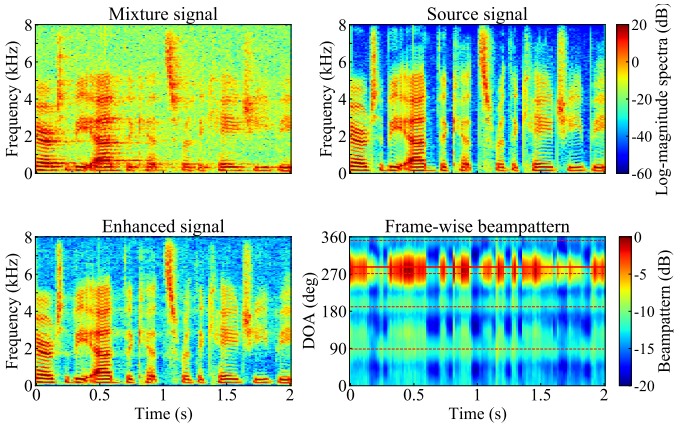

Fig. 2. The log-magnitude spectra of the mixture (top left), the source (top right), and the enhanced signal obtained by the proposed method (bottom left) are shown, along with the frame-wise beampattern of its optimal beamformer (bottom right). The red solid line indicates the estimated direct-path DOA of $287.16°$, with the ground-truth value being $288.67°$. And four red dotted lines represent the estimated first-order reflection DOAs from the surrounding walls at $190.84°$, $351.60°$, $271.23°$, and $88.41°$, with corresponding ground-truth values of $188.94°$, $348.56°$, $274.80°$, and $86.36°$, respectively.

To evaluate performance, DBNet [9] is adopted as a baseline for comparison. Unlike the proposed method, DBNet employs a single AMV directed only toward the estimated direct-path DOA. In contrast, the proposed method incorporates AMVs directed toward both the direct path and early reflections, using learning-based DOAs and corresponding transfer functions. For a fair comparison, both methods utilize the MVDR beamformer. Table I summarizes the average SDR (dB) [36] and perceptual evaluation of speech quality (PESQ) [37] scores of the enhanced signals under different reverberation times and input SNRs. The results show that the proposed method consistently outperforms DBNet, with particularly significant gains in PESQ, highlighting its ability to exploit informative early reflections.

Furthermore, Table II presents the average root MSE (RMSE) of the DOA estimation, compared with the independent subspace spatial matching (ISSM)-based MUSIC algo-

rithm [38] and a neural network-based MUSIC approach [39]. The proposed method achieves the lowest DOA estimation errors among all compared methods, demonstrating its robustness in distinguishing coherent sound sources in reverberant environments.

V. Conclusions

In this paper, we proposed a novel neural network-assisted adaptive beamforming method that jointly estimates the DOAs of both the direct path and the paths of four first-order reflections from the surrounding walls, along with their corresponding transfer functions, using a multi-task neural network. Through an end-to-end training strategy, these parameters are directly inferred from observed signals and subsequently applied to an MVDR beamformer to enhance the target speech. Experimental results demonstrate that the proposed method outperforms baseline approaches in both source signal enhancement and DOA estimation performance.

References

[1] J. Benesty, J. Chen, Y. Huang, and J. Dmochowski, "On microphone-array beamforming from a MIMO acoustic signal processing perspective," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1053–1065, Mar. 2007.

[2] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer Handbook of Speech Processing*. Berlin, Germany: Springer, 2008.

[3] J. Benesty, I. Cohen, and J. Chen, *Fundamentals of Signal Enhancement and Array Signal Processing*. Singapore: Wiley-IEEE Press., 2018.

[4] C. Pan, J. Chen, and J. Benesty, "Microphone array beamforming with high flexible interference attenuation and noise reduction," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 1865–1876, May 2022.

[5] Y. Lu, C. Pan, J. Chen, and J. Benesty, "A closed-form DOA estimator using spherical microphone arrays in the presence of interference," *IEEE Signal Process. Lett.*, vol. 31, pp. 1770–1774, 2024.

[6] C. Pan and J. Chen, "A framework of directional-gain beamforming and a white-noise-gain-controlled solution," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 2875–2887, 2022.

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

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

[9] A. Aroudi and S. Braun, "DBNet: DOA-driven beamforming network for end-to-end reverberant sound source separation," in *Proc. IEEE ICASSP*, 2021, pp. 211–215.

[10] Y. Yang, N. Pan, W. Zhang, C. Pan, J. Benesty, and J. Chen, "Interference-controlled maximum noise reduction beamformer based on deep-learned interference manifold," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 4676–4690, Oct. 2024.

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

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

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

[14] C. Pan, "Fundamentals of data-driven approaches to acoustic signal detection, filtering, and transformation," *arXiv:2508.21470*, 2025.

[15] J. Chen, X. Wu, and T. Qu, "Early reflections based speech enhancement," in *Proc. ICICSP*, 2021, pp. 183–187.

[16] J. Flanagan, A. Surendran, and E. Jan, "Spatially selective sound capture for speech and audio processing," *Speech Commun.*, vol. 13, no. 1, pp. 207–222, 1993.

[17] T. Nishiura, S. Nakamura, and K. Shikano, "Speech enhancement by multiple beamforming with reflection signal equalization," in *Proc. IEEE ICASSP*, 2001, pp. 189–192.

[18] J. S. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms," *J. Acoust. Soc. Am.*, vol. 113, no. 6, pp. 3233–3244, May 2003.

[19] H. L. Van Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation*. New York, NY: Wiley, 2004.

[20] H. Cox, R. M. Zeskind, and T. Kooij, "Practical supergain," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 3, pp. 393–398, Jun. 1986.

[21] S. Yan and Y. Ma, "Robust supergain beamforming for circular array via second-order cone programming," *Appl. Acoust.*, vol. 66, no. 9, pp. 1018–1032, Sep. 2005.

[22] C. Pan, J. Chen, and J. Benesty, "Reduced-order robust superdirective beamforming with uniform linear microphone arrays," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1544–1555, Sep. 2016.

[23] J. Li, P. Stoica, and Z. Wang, "On robust Capon beamforming and diagonal loading," *IEEE Trans. Signal Process.*, vol. 51, no. 7, pp. 1702–1715, Jul. 2003.

[24] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 2008.

[25] C. Pan, J. Chen, and J. Benesty, "Performance study of the MVDR beamformer as a function of the source incidence angle," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 1, pp. 67–79, Jan. 2014.

[26] C. Pan, J. Chen, G. Shi, and J. Benesty, "On microphone array beamforming and insights into the underlying signal models in the short-time-Fourier-transform domain," *J. Acoust. Soc. Am.*, vol. 149, no. 1, pp. 660–672, 2021.

[27] C. Quan and X. Li, "SpatialNet: Extensively learning spatial information for multichannel joint speech separation, denoising and dereverberation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 1310–1323, Feb. 2024.

[28] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR Half-baked or Well Done?" in *Proc. IEEE ICASSP*, 2019, pp. 626–630.

[29] H. Taherian, K. Tan, and D. L. Wang, "Location-based training for multi-channel talker-independent speaker separation" in *Proc. IEEE ICASSP*, 2022, pp. 696–700.

[30] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE ICASSP*, 2016, pp. 31–35.

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

[32] C. Pan, L. Zhang, Y. Lu, J. Jin, L. Qiu, J. Chen, and J. Benesty, "An anchor-point based image-model for room impulse response simulation with directional source radiation and sensor directivity patterns," *arXiv:2308.10543*, 2023.

[33] W. C. Sabine, *Collected papers on acoustics*. Cambridge, MA: Harvard University Press, 1922.

[34] E. A. P. Habets and S. Gannot, "Generating sensor signals in isotropic noise fields," *J. Acoust. Soc. Am.*, vol. 122, no. 6, pp. 3464–3470, Dec 2007.

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

[36] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, June 2006.

[37] 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. IEEE ICASSP*, 2001, pp. 749–752.

[38] S. Argentieri and P. Danes, "Broadband variations of the MUSIC high-resolution method for sound source localization in robotics," in *Proc. IEEE/RSJ IROS*, 2007, pp. 2009–2014.

[39] H. Li, W. Zhang, and L. Zhang, "DoA estimation of room reflections using NN-based MUSIC algorithm" in *Proc. APSIPA ASC*, 2023, pp. 1960–1965.