Integrated DNN-Based Parameter Estimation for Multichannel Speech Enhancement

Sein Cheong, Minseung Kim, and Jong Won Shin, *Senior Member, IEEE*

***Abstract*—One of the popular configurations for the statistical model-based multichannel speech enhancement (SE) is to apply a spatial filter such as the minimum-variance distortionless response beamformer followed by a single channel post-filter, and some of the deep neural network (DNN)-based approaches mimic it. While a number of DNN-based SE focused on direct estimation of clean speech features or the masks to estimate clean speech, some of the efforts were devoted to estimate the statistical parameters. DNN-based parameter estimation with two DNNs for a beamforming stage and a post-filtering stage has demonstrated impressive performance, but the parameter estimation for a beamformer and that for a post-filter operate separately, which may not be optimal in that the post-filter cannot utilize spatial information from multi-microphone signals. In this letter, we propose integrated DNN-based parameter estimation for multichannel SE based on both the beamformer output and multi-microphone signals. The speech presence probability and the power spectral densities for speech and noise estimated in the beamforming stage are utilized in the post-filtering stage for better parameter estimation. We also adopt the dual-path conformer structure with an encoder and decoders to enhance the performance. Experimental results show that the proposed method marked the best wideband perceptual evaluation of speech quality (PESQ) scores on the CHiME-4 dataset among all methods with comparable computational complexity.**

***Index Terms*—Multichannel speech enhancement, MVDR beamformer, post-filter, DNN-based parameter estimation.**

I. Introduction

The goal of multichannel speech enhancement (SE) is to improve the perceptual quality of noisy and reverberant multichannel speech [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30]. One widely-used structure for multichannel SE is to process the input with a spatial filter such as the minimum variance distortionless response (MVDR) beamformer exploiting spatial diversity of sound sources, and then apply a single channel post-filter to the output of the spatial filter to further reduce the residual noise [2], [3], [4], [5]. Recently, various studies have been proposed to utilize deep neural networks (DNNs) for multichannel SE, which have shown significant performance improvement. Early approaches simply adopt DNN-based single channel SE to obtain spatial covariance matrices (SCMs) for speech or noise [6], [7], and several efforts have been made to estimate beamformers directly [8], [9], [10], [11], [12], [13]. While a majority of recent DNN-based multichannel SE methods focus on direct estimation of clean features or masks applied to noisy features [14], [15], [16], [17], [18], [19], [20], [21], [22], there have been attempts to estimate parameters required in the traditional framework such as SCMs, relative transfer functions (RTFs), power spectral densities (PSDs) for speech and noise, and signal-to-noise ratios (SNRs) [6], [11], [12], [13], [23], [24], [25], [26], [27], [28], [29].

As for the post-filter in the DNN-based multichannel SE systems, a Wiener filter was adopted in [23] and DNN-based post-filters which directly estimate masks or clean features from the beamformer output and/or microphone signals were employed in [9], [10], [25], [26], [27], [28]. Recently, a DNN-based parameter estimation approach was proposed in [29], where one DNN estimates parameters for an MVDR beamformer and the other produces parameters required for a post-filter. While it achieved a good performance with a moderate number of parameters demonstrating the potential of DNN-based parameter estimation for the traditional statistical model-based multichannel SE framework, the parameter estimation in the post-filtering stage only takes the beamformer output signal. As the multi-microphone signals available only for the beamforming stage would have information on speech and noise that the beamformed signal does not include, it may have limited the performance of this method.

In this letter, we propose an integrated DNN-based parameter estimation incorporating the minimum mean square error (MMSE) estimators of the PSDs for speech and residual noise based on both beamformer output and the multi-microphone signals. Specifically, the *a posteriori* speech presence probability (SPP), speech PSD, and noise PSD estimated in the beamforming stage based on the microphone signals are integrated in the parameter estimation for post-filtering to exploit spatial and spectro-temporal information. We also adopt the dual-path conformer structure [31] as the network structures for the two DNNs to further enhance the performance. Experimental results on the CHiME-4 dataset demonstrate that the proposed method showed the highest perceptual evaluation of speech quality (PESQ) scores [32].

Received 3 May 2025; revised 6 August 2025; accepted 8 August 2025. Date of publication 14 August 2025; date of current version 27 August 2025. This work was supported by the IITP (Institute of Information & Communications Technology Planning & Evaluation)-ITRC (Information Technology Research Center) funded by Korea Government (Ministry of Science and ICT) under Grant IITP-2025-RS-2021-II211835. The associate editor coordinating the review of this article and approving it for publication was Prof. Konrad Kowalczyk. *(Corresponding author: Jong Won Shin.)*

The authors are with the Department of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 61005, South Korea (e-mail: seiinjung@gm.gist.ac.kr; a99756867@gmail.com; jwshin@gist.ac.kr).

Digital Object Identifier 10.1109/LSP.2025.3599455

II. Review of DNN-Based Parameter Estimation

A. MVDR Beamforming and Post-Filtering

Suppose that the desired speech is captured by M microphones in a noisy and reverberant environment. The observed noisy signal in the short-time Fourier transform (STFT) domain, $\mathbf{y}(l,k) = [Y_1(l,k), Y_2(l,k), \ldots, Y_M(l,k)]^T$, $1 \leq l \leq L, 1 \leq k \leq K$, in which $Y_m(l,k)$ is the k-th frequency component for the noisy speech at the m-th microphone in the l-th frame, can be written as a function of the clean speech $\mathbf{s}(l,k)$, noise including reverberation $\mathbf{v}(l,k)$, and RTF vector $\mathbf{g}(l,k)$:

$$\mathbf{y}(l,k) = \mathbf{s}(l,k) + \mathbf{v}(l,k) = \mathbf{g}(l,k)S_1(l,k) + \mathbf{v}(l,k). \tag{1}$$

To estimate clean speech at the reference microphone $S_1(l,k)$ from the observed signals $\mathbf{y}(l,k)$, DNN-based parameter estimation is proposed in [29] for the statistical model-based multichannel SE framework consisting of a spatial filter followed by a post-filter. One of the popular choices for the spatial filter is the MVDR beamformer minimizing the variance of the noise at the beamformer output under the constraint of the distortionless response for the desired direction :

$$\mathbf{w}_{mvdr}(l,k) = \frac{\Phi_{\mathbf{v}}^{-1}(l,k)\mathbf{g}(l,k)}{\mathbf{g}^H(l,k)\Phi_{\mathbf{v}}^{-1}(l,k)\mathbf{g}(l,k)} \tag{2}$$

in which $\Phi_{\mathbf{v}}(l,k) = \mathbb{E}[\mathbf{v}(l,k)\mathbf{v}^H(l,k)]$ is the SCM of $\mathbf{v}(l,k)$. The beamformer output $Z(l,k) = \mathbf{w}_{mvdr}^H(l,k)\mathbf{y}(l,k)$ does not perfectly match $S_1(l,k)$, and thus a post-filter to suppress $O(l,k) = Z(l,k) - S_1(l,k)$ is needed. One of the widely-used post-filter is the MMSE log-spectral amplitude (LSA) [33] clean speech estimator given by

$$\hat{S}_1(l,k) = G_{mmse-lsa}(l,k)Z(l,k), \tag{3}$$

where $G_{mmse-lsa}(l,k)$ is given by

$$G_{mmse-lsa}(l,k) = \frac{\xi(l,k)}{\xi(l,k)+1}\exp\left\{\frac{1}{2}\int_{v(l,k)}^{\infty}\frac{e^{-t}}{t}dt\right\}, \tag{4}$$

in which $v(l,k) = [\xi(l,k)/(\xi(l,k)+1)]\gamma(l,k)$, and $\xi(l,k)$ and $\gamma(l,k)$ are the *a priori* and the *a posteriori* SNRs defined as $\xi(l,k) = \phi_s(l,k)/\phi_o(l,k)$ and $\gamma(l,k) = |Z(l,k)|^2/\phi_o(l,k)$ where $\phi_s(l,k) = \mathbb{E}[|S_1(l,k)|^2]$ and $\phi_o(l,k) = \mathbb{E}[|O(l,k)|^2]$.

B. Parameter Estimation for Beamforming and Post-Filtering

In [29], a DNN-based parameter estimation approach was proposed, which provides higher tunability than the end-to-end approaches. Let $\psi_m(l,k)$ and $\theta_m(l,k)$, $2 \leq m \leq M$ be the interchannel phase differences (IPDs) between the m-th and the first microphone signals for the clean and noisy speech and $\Psi_m(l,k) = [\frac{\sin\psi_m(l,k)+1}{2}, \frac{\cos\psi_m(l,k)+1}{2}]$ and $\Theta_m(l,k) = [\frac{\sin\theta_m(l,k)+1}{2}, \frac{\cos\theta_m(l,k)+1}{2}]$ are sinusoidal functions of them [34] mapped into [0,1] range [29]. The first DNN (DNN$_1$) estimates the *a posteriori* SPP $p_s = P(H_1|\mathbf{y})$ and $\mathbf{\Psi} = \{\Psi_m(l,k), \forall m,l,k\} \in \mathbb{R}^{2(M-1)\times L\times K}$ from the magnitude spectrogram of the first microphone signal, $|\mathbf{Y}_1| = \{|Y_1(l,k)|, \forall l,k\} \in \mathbb{R}^{L\times K}$, and $\mathbf{\Theta} = \{\Theta_m(l,k), \forall m,l,k\} \in \mathbb{R}^{2(M-1)\times L\times K}$ for the MVDR beamformer:

$$[\widehat{p}_{s_{BF}}, \widehat{\mathbf{\Psi}}] = \text{DNN}_1(|\mathbf{Y}_1|, \mathbf{\Theta}), \tag{5}$$

in which $\widehat{p}_{s_{BF}}$ is the estimate of p_s in the beamforming stage. Each component of $\mathbf{g}(l,k)$ in (2) is approximated using $\widehat{\psi}_m(l,k)$ computed from $\widehat{\Psi}_m(l,k)$ as in [29] with the far-field assumption, i.e,

$$\widehat{g}_m(l,k) = \exp j \cdot \widehat{\psi}_m(l,k). \tag{6}$$

Another parameter, $\Phi_{\mathbf{v}}(l,k)$, is computed using the bidirectional multichannel minima-controlled recursive averaging (BMC-MCRA) approach [29], where $\hat{p}_{s_{BF}}$ determines the SPP-dependent smoothing parameter.

The second DNN (DNN$_2$) is dedicated to estimate the parameters to evaluate the gain function in (4). In [29], the iDeepMMSE proposed in [35] was adopted, which computes $\xi(l,k)$ and $\gamma(l,k)$ using the parameters obtained from DNN$_2$. DNN$_2$ estimates p_s, speech PSD ϕ_s, and *a priori* SNR ξ from the magnitude of the beamformer output $|\mathbf{Z}| = \{|Z(l,k)|, \forall l,k\} \in \mathbb{R}^{L\times K}$, i.e,

$$\left[\widehat{p}_{s_{PF}}, \hat{\bar{\phi}}_{s_{PF}}, \hat{\bar{\xi}}_{PF}\right] = \text{DNN}_2(|\mathbf{Z}|), \tag{7}$$

in which $\bar{\phi}_s$ and $\bar{\xi}$ are the training targets for ϕ_s and ξ mapped into [0,1] range by modeling the distribution of each variable with a Gaussian distribution and applying the cumulative distribution function as a mapping function [29], [35], and $\hat{\cdot}_{PF}$ denotes an estimate for the post-filter. $\hat{\phi}_{s_{PF}}$ and $\hat{\xi}_{PF}$ are obtained by applying the inverse mappings to $\hat{\bar{\phi}}_{s_{PF}}$ and $\hat{\bar{\xi}}_{PF}$. Then, the MMSE estimators for the power spectra of speech and residual noise can be obtained as

$$\begin{aligned}\widehat{|S_1|^2} &= \mathbb{E}(|S_1|^2|Z) \\ &= p(H_0|Z)\mathbb{E}(|S_1|^2|Z,H_0) + p(H_1|Z)\mathbb{E}(|S_1|^2|Z,H_1),\end{aligned} \tag{8}$$

$$\begin{aligned}\widehat{|O|^2} &= \mathbb{E}(|O|^2|Z) \\ &= p(H_0|Z)\mathbb{E}(|O|^2|Z,H_0) + p(H_1|Z)\mathbb{E}(|O|^2|Z,H_1),\end{aligned} \tag{9}$$

where H_0 and H_1 indicate the hypotheses for speech absence and presence, $p(H_0|Z) = 1 - p(H_1|Z)$ and

$$\mathbb{E}(|S_1|^2|Z,H_0) = 0, \quad \mathbb{E}(|O|^2|Z,H_0) = |Z|^2, \tag{10}$$

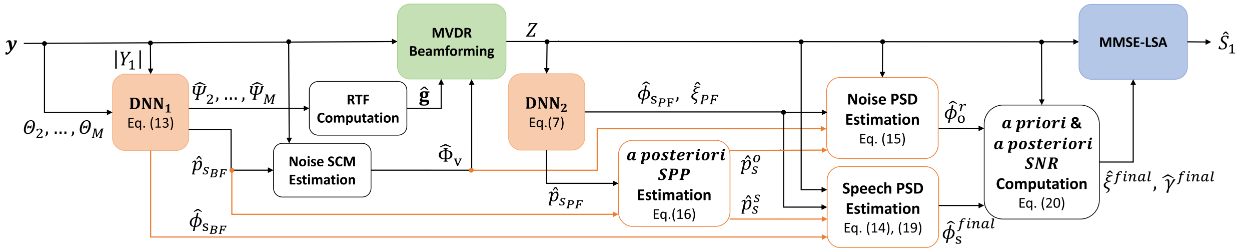
$$\mathbb{E}(|S_1|^2|Z,H_1) = \left(\frac{\xi}{1+\xi}\right)^2|Z|^2 + \frac{1}{1+\xi}\phi_s, \tag{11}$$

$$\mathbb{E}(|O|^2|Z,H_1) = \left(\frac{\eta}{1+\eta}\right)^2|Z|^2 + \frac{1}{1+\eta}\phi_o \tag{12}$$

in which $\eta = 1/\xi$ is the noise-to-signal ratio. $\widehat{p}_{s_{PF}}$ is used as $p(H_1|Z)$, ϕ_s and ϕ_o are estimated by $\widehat{\phi}_{s_{PF}}$ and $\widehat{\phi}_{o_{PF}} = \frac{1}{1+\widehat{\xi}_{PF}}|Z|^2$, and ξ and η are given by $\widehat{\phi}_{s_{PF}}/\widehat{\phi}_{o_{PF}}$ and its reciprocal, respectively. The refined estimates for the speech and noise PSDs, $\widehat{\phi}_s^r$ and $\widehat{\phi}_o^r$, are obtained through temporal recursive smoothing of $\widehat{|S_1|^2}$ and $\widehat{|O|^2}$. The MMSE-LSA gain function in (4) is evaluated with $\widehat{\xi}^r(l,k) = \widehat{\phi}_s^r(l,k)/\widehat{\phi}_o^r(l,k)$ and $\widehat{\gamma}^r(l,k) = |Z(l,k)|^2/\widehat{\phi}_o^r(l,k)$.

III. Integrated DNN-Based Parameter Estimation for Multichannel Speech Enhancement

While the study in [29] showed that adopting DNN-based parameter estimation for MVDR beamforming and post-filtering in the statistical model-based SE framework can compete with the end-to-end DNN-based SE, the post-filtering stage in [29]



y
$|Y_1|$
$\Theta_2, \ldots, \Theta_M$
DNN1 Eq. (13)
$\widehat{\Psi}_2, \ldots, \widehat{\Psi}_M$
RTF Computation
$\widehat{\mathbf{g}}$
MVDR Beamforming
Noise SCM Estimation
$\widehat{\Phi}_\mathrm{v}$
$\widehat{p}_{s_{BF}}$
$\widehat{\phi}_{s_{BF}}$
Z
DNN2 Eq.(7)
$\hat{\bar{\phi}}_{s_{PF}}$, $\hat{\bar{\xi}}_{PF}$
$\widehat{p}_{s_{PF}}$
a posteriori SPP Estimation Eq.(16)
$\widehat{p}_s^o$
$\widehat{p}_s^s$
Noise PSD Estimation Eq. (15)
$\widehat{\phi}_\mathrm{o}^r$
Speech PSD Estimation Eq. (14), (19)
$\widehat{\phi}_\mathrm{s}^{final}$
a priori & a posteriori SNR Computation Eq. (20)
$\hat{\xi}^{final}$, $\widehat{\gamma}^{final}$
MMSE-LSA
$\hat{S}_1$

Fig. 1. Block diagram of the proposed multichannel speech enhancement system. Orange color indicates modified parts different from DeepPE [29].

was essentially a single channel SE operating independently of the beamforming stage. In this work, the MMSE estimators for the power spectra of speech and residual noise based on the beamformer output Z in (8) and (9) are replaced by the MMSE estimators for them based on not only Z but also the multichannel microphone signal $\mathbf{y}$ so that the spatial information can be exploited in the post-filtering stage. The estimate for the speech PSD is also further refined incorporating the estimate of it obtained in the beamforming stage. As the speech activity and speech PSD in the beamforming stage are the same as those in the post-filtering stage and the noise PSD estimated from multi-microphone signals bears a certain amount of information on the residual noise in the beamformed signal, the integrated parameter estimation combining the estimates from both of the stages will lead to better SE. We also propose a new architecture for DNN_1 and DNN_2 based on dual-path conformers modifying the generator of the CMGAN [31]. The block diagram of the proposed system is shown in Fig. 1, which is described in the following subsections in detail.

A. Architecture of the DNNs

The input and output for DNN_2 are the same as those in (7), while the output of DNN_1 is modified from the one in (5) used in [29]. One more parameter $\bar{\phi}_{s_{BF}}$, the speech PSD mapped into the range $[0, 1]$, is also estimated by DNN_1, i.e,

$$\left[\widehat{p}_{s_{BF}}, \hat{\bar{\phi}}_{s_{BF}}, \widehat{\boldsymbol{\Psi}}\right] = \text{DNN}_1(|\mathbf{Y}_1|, \boldsymbol{\Theta}). \tag{13}$$

The training target for $\hat{\bar{\phi}}_{s_{BF}}$ is essentially the same with that for $\hat{\bar{\phi}}_{s_{PF}}$, and the same inverse mapping is applied to $\hat{\bar{\phi}}_{s_{BF}}$ to obtain $\hat{\phi}_{s_{BF}}$, which is used in the post-filtering stage.

As for the model architectures for DNN_1 and DNN_2, we adopt the dual-path conformer structure employed for the generator of the CMGAN [31] as a backbone, which originally comprises an encoder that takes the magnitude and real and imaginary parts of the spectrogram, dual-path conformer blocks each of which applies the conformers along the time and frequency axes sequentially, and two decoders to produce real-valued masks and complex-valued clean feature estimates. The input variables to DNN_1 are concatenated along the channel direction and then fed into the encoder. Because the sinusoidal functions of the clean IPDs have quite different patterns from the *a posteriori* SPP and speech PSD which have high values for the time-frequency bins with high speech energies, we employ two separate decoders for $\widehat{\boldsymbol{\Psi}}$ and $[\widehat{p}_{s_{BF}}, \hat{\bar{\phi}}_{s_{BF}}]$, respectively, both with sigmoid activation functions. As for DNN_2, only one decoder is used to produce $\widehat{p}_{s_{PF}}$, $\hat{\bar{\phi}}_{s_{PF}}$, and $\hat{\bar{\xi}}_{PF}$ with a sigmoid activation function. We have used the weighted summation of the binary cross entropy (BCE) for each output variable as the loss function to train both of the networks as in [29], since all the output variables are scaled into [0,1].

B. Post-Filtering With Integrated Estimation of PSDs

In [29], the power spectra $|S_1|^2$ and $|O|^2$ are estimated with the MMSE criteria assuming that Z is the only observation as shown in (8) and (9). In this letter, we propose the MMSE estimators for them based on both Z and $\mathbf{y}$ given by

$$\begin{aligned}\widehat{|S_1|^2} = \mathbb{E}(|S_1|^2|Z, \mathbf{y}) = p(H_0|Z, \mathbf{y})\mathbb{E}(|S_1|^2|Z, \mathbf{y}, H_0)\\ +p(H_1|Z, \mathbf{y})\mathbb{E}(|S_1|^2|Z, \mathbf{y}, H_1),\end{aligned} \tag{14}$$

$$\begin{aligned}\widehat{|O|^2} = \mathbb{E}(|O|^2|Z, \mathbf{y}) = p(H_0|Z, \mathbf{y})\mathbb{E}(|O|^2|Z, \mathbf{y}, H_0)\\ +p(H_1|Z, \mathbf{y})\mathbb{E}(|O|^2|Z, \mathbf{y}, H_1).\end{aligned} \tag{15}$$

$\mathbb{E}(|S_1|^2|Z, \mathbf{y}, H_0), \mathbb{E}(|O|^2|Z, \mathbf{y}, H_0), \mathbb{E}(|S_1|^2|Z, \mathbf{y}, H_1)$ and $\mathbb{E}(|O|^2|Z, \mathbf{y}, H_1)$ have the same forms with the right-hand sides in (10)-(12) except the conditioning variable $\mathbf{y}$ affects the estimation of ξ and η. The estimates for the *a posteriori* SPP $p(H_1|Z, \mathbf{y})$ in (14) and (15), $\widehat{p}_s^s$ and $\widehat{p}_s^o$, can be obtained by combining the estimate of it from $\mathbf{y}$, $\widehat{p}_{s_{BF}}$, and another estimate of it from Z, $\widehat{p}_{s_{PF}}$, with two different weights $w_{p_s}^s$ and $w_{p_s}^o$ as

$$\widehat{p}_s^x = w_{p_s}^x \widehat{p}_{s_{BF}} + (1 - w_{p_s}^x)\widehat{p}_{s_{PF}},\ x \in \{s, o\}. \tag{16}$$

$\widehat{\xi}$ and $\widehat{\eta}$ are the estimated ratios of the PSDs for speech and residual noise, and we can combine the estimates for the PSDs of speech and residual noise from the beamforming and the post-filtering stages. Unlike the $\widehat{\xi}$ and $\widehat{\eta}$ in [29] which are the reciprocal of each other, we propose to estimate ξ and η separately for specific goals such as the preservation of speech and clean-up of the spectral valleys. ξ and η are estimated as

$$\widehat{\xi}_{Z,\mathbf{y}} = \frac{w_\xi \widehat{\phi}_{s_{BF}} + (1 - w_\xi)\widehat{\phi}_{s_{PF}}}{\widehat{\phi}_{o_{PF}}} \tag{17}$$

$$\widehat{\eta}_{Z,\mathbf{y}} = \frac{w_\eta \widehat{\phi}_{o_{BF}} + (1 - w_\eta)\widehat{\phi}_{o_{PF}}}{\widehat{\phi}_{s_{PF}}} \tag{18}$$

in which w_ξ and w_η are weights and $\widehat{\phi}_{o_{BF}}$, the noise PSD estimate from the beamformer stage, is computed from the (1,1)-th component of $\Phi_\mathbf{v}$ estimated by the BMC-MCRA approach, $[\widehat{\Phi}_\mathbf{v}]_{(1,1)}$, and the magnitude gain function $G_{BF} = \max(|Z|^2/|Y_1|^2, 1)$ as $\widehat{\phi}_{o_{BF}} = G_{BF}[\widehat{\Phi}_\mathbf{v}]_{(1,1)}$ [36].

Using the MMSE estimators for power spectra in (14) and (15), the refined estimates for speech and noise PSDs, $\widehat{\phi}_s^r$ and $\widehat{\phi}_o^r$, are obtained by temporal recursive smoothing of power spectra as in [29]. To further reduce the speech PSD underestimation in the speech present regions, we combine $\widehat{\phi}_s^r$ with $\widehat{\phi}_{s_{BF}}$ once more since $\widehat{\phi}_{s_{BF}}$ tends to underestimate ϕ_s less in our preliminary experiment. The final estimate of the speech PSD becomes

$$\widehat{\phi}_s^{final} = \tilde{w}_{\phi_s}\widehat{\phi}_{s_{BF}} + (1 - \tilde{w}_{\phi_s})\widehat{\phi}_s^r \tag{19}$$

where $\tilde{w}_{\phi_s} = w_{\phi_s}\hat{\bar{\phi}}_{s_{BF}}$ is a weighting factor depending on $\hat{\bar{\phi}}_{s_{BF}}$. It is noted that we have used $\hat{\bar{\phi}}_{s_{BF}}$ as a measure of speech presence because $\bar{\phi}_{s_{BF}}$ is in [0,1] and $\hat{\bar{\phi}}_{s_{BF}}$ shows higher peak-to-valley ratios than the estimates for the SPP. Finally, the MMSE-LSA gain function of the post-filter in (4) is computed using the $\widehat{\xi}^{final}(l,k)$ and $\widehat{\gamma}^{final}(l,k)$ given by

$$\widehat{\xi}^{final}(l,k) = \frac{\widehat{\phi}_s^{final}(l,k)}{\widehat{\phi}_o^r(l,k)}, \widehat{\gamma}^{final}(l,k) = \frac{|Z(l,k)|^2}{\widehat{\phi}_o^r(l,k)}. \tag{20}$$

IV. Experiments

A. Experimental Settings

To compare the performance of the proposed method with previously proposed ones, we have experimented on the simulated data in the CHiME-4 dataset [37] with 6 microphones. The simulated set in the CHiME-4, which is divided into training, validation, and test sets, includes 4 noisy environment including bus, cafe, pedestrian area, and street junction. Among the six microphones on the tablet device held by the speaker, the fifth microphone at the bottom center of the frontal surface was used as a reference microphone. The sampling rate was 16 kHz and the STFT was applied to a 32 ms window with a 50% overlap, in which the square-root Hann window was utilized for analysis and synthesis. The parameter values for $w_{p_s}^s$, $w_{p_s}^o$, w_ξ, w_η, and w_{ϕ_s} were determined to 0.5, 0, 0.01, 0.99, and 0.045, respectively, to maximize WB-PESQ scores for the validation set. All other parameters were set to be the same as those in [29].

As for DNN$_1$ and DNN$_2$, the batch size, number of channels at the encoder output, and number of two-stage conformer blocks were set to 5, 48, and 4, respectively. The weights on the BCEs for output variables were all 1 except the one for $\hat{\bar{\phi}}_{s_{BF}}$ in DNN$_1$, which was set to 8. Models were trained using Adam optimizer [38] with a learning rate of 5×10^{-4} for the first 400 epochs and 2.5×10^{-4} for the next 400 epochs, and the models with the best validation losses were chosen.

B. Experimental Results

Table I summarizes the SE performances on the CHiME-4 dataset including wideband (WB) and narrowband (NB) PESQ scores [32], [39], short-time objective intelligibility (STOI) [40], and scale-invariant signal-to-distortion ratio (SI-SDR) [41], with the numbers of parameters. It is noted that the compared TF-GridNet is the 6 channel version with a different loss function in [20], not the original single channel separation model in [42]. A 6 channel version of the CMGAN [31] trained in an end-to-end way is also compared to the parameter estimation approach based on a CMGAN architecture. The average WB-PESQ score for the DNN-based parameter estimation (DeepPE) in [29] was higher than several recently proposed methods directly estimating clean features or masks despite the far-field assumption in (6), although it has more number of parameters. By adopting the proposed architecture using dual-path conformers, the WB-PESQ of DeepPE could be improved to 3.25, which was higher than that of the 6 channel CMGAN implying that the parameter estimation approach was competitive with end-to-end enhancement approaches. The average WB-PESQ score was improved by introducing each component of the integrated parameter estimation, and that for the proposed method (iDeepPE) with both the new architecture and the integrated estimation was 3.34, which was better than those for all compared methods except the 6 channel TF-GridNet. 6 channel TF-GridNet [20] achieved the same WB-PESQ score with significantly higher computational complexity, i.e., with 5.4 M parameters and 139 Giga floating point operations per second (GFLOPS) compared to iDeepPE requiring 1.98 M parameters and 39.8 GFLOPS. The performance improvement of iDeepPE verified that the integrated parameter estimation incorporating the estimated parameters in the beamforming stage which enabled exploitation of spatial and spectro-temporal information was beneficial. More experimental results and the code for the iDeepPE is available at https://github.com/CSeIn/iDeepPE.

TABLE I
SPEECH ENHANCEMENT PERFORMANCES ON THE CHIME-4 DATASET

Methods	#Param	WB PESQ	NB PESQ	STOI	SI SDR
Noisy	-	1.27	2.18	0.870	7.5
FT-JNF†[15]	1.2M	2.61	3.20	0.967	17.6
AFnet [13]	2.6M	2.72	-	0.972	-
McNET† [16]	1.9M	2.73	3.38	0.976	19.2
SpatialNet-s† [18]	1.6M	2.88	3.49	0.983	22.1
CMGAN (6ch)	2.01M	2.91	3.38	0.976	18.8
MCMamba [19]	-	2.98	3.49	0.982	-
MC-SEMamba [21]	2.3M	3.07	3.48	0.985	21.3
DeepPE [29]	8.43M	3.09	3.47	0.974	19.3
Wang et al. [27]	26M	-	3.68	0.986	22.0
USES [17]	3.05M	3.16	-	0.983	-
TF-GridNet (6ch) [20]	5.4M	3.34	-	0.987	22.9
DeepPE+DPConformer	1.98M	3.25	3.57	0.982	19.2
+ (14)	1.98M	3.28	3.59	0.982	19.2
+ (15)	1.98M	3.30	3.59	0.982	19.0
+ (19) (iDeepPE)	1.98M	3.34	3.61	0.982	19.0

† denotes that the scores and model sizes reported in [18].

V. Conclusion

In this letter, we propose an integrated DNN-based parameter estimation for multichannel SE, in which the *a posteriori* SPP and PSDs for speech and noise estimated during the beamforming stage are incorporated into the parameter estimation for the post-filter. The MMSE estimators for the power spectra of speech and residual noise based on both the beamformer output and the multi-microphone signals are employed to exploit spatial and spectro-temporal information. Furthermore, we adopt the dual-path conformer architecture for both of the DNNs to improve the performance with a reasonable number of parameters. Experimental results on the CHiME-4 dataset show that the proposed method outperformed compared methods with similar levels of computational complexity.

REFERENCES

[1] J. Benesty, *Microphone Array Signal Processing*. Berlin, Germany: Springer, 2008.

[2] S. Gannot and I. Cohen, “Adaptive beamforming and postfiltering,” in *Springer Handbook of Speech Processing*. Berlin, Germany: Springer, 2008, pp. 945–978.

[3] 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, Apr. 2017.

[4] P. Thüne and G. Enzner, “Maximum-likelihood approach with Bayesian refinement for multichannel-wiener postfiltering,” *IEEE Trans. Signal Process.*, vol. 65, no. 13, pp. 3399–3413, Jul. 2017.

[5] S. Cheong, M. Kim, and J. W. Shin, “Postfilter for dual channel speech enhancement using coherence and statistical model-based noise estimation,” *Sensors*, vol. 24, no. 12, 2024, Art. no. 3979.

[6] 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.*, 2016, pp. 196–200.

[7] 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.

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

[9] H. Kim, K. Kang, and J. W. Shin, “Factorized MVDR deep beamforming for multi-channel speech enhancement,” *IEEE Signal Process. Lett.*, vol. 29, pp. 1898–1902, 2022.

[10] 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.*, 2022, pp. 6487–6491.

[11] Y. Xu, Z. Zhang, M. Yu, S.-X. Zhang, and D. Yu, “Generalized spatio-temporal RNN beamformer for target speech separation,” in *Proc. INTERSPEECH*, pp. 3076–3080, 2021.

[12] C.-H. Lee, K. Patel, C. Yang, Y. Shen, and H. Jin, “An MVDR-embedded U-net beamformer for effective and robust multichannel speech enhancement,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2024, pp. 8541–8545.

[13] C.-H. Lee, C. Yang, Y. M. Saidutta, R.S. Srinivasa, Y. Shen, and H. Jin, “Better exploiting spatial separability in multichannel speech enhancement with an align-and-filter network,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2025, pp. 1–5.

[14] 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.*, 2020, pp. 836–840.

[15] 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, 2023.

[16] Y. Yang, C. Quan, and X. Li, “McNet: Fuse multiple cues for multichannel speech enhancement,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.

[17] W. Zhang, K. Saijo, Z.-Q. Wang, S. Watanabe, and Y. Qian, “Toward universal speech enhancement for diverse input conditions,” in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2023, pp. 1–6.

[18] 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, 2024.

[19] W. Ren et al., “Leveraging joint spectral and spatial learning with MAMBA for multichannel speech enhancement,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2025, pp. 1–5.

[20] Z.-Q. Wang, “ctPuLSE: Close-talk, and pseudo-label based far-field, speech enhancement,” 2024, *arXiv:2407.19485*.

[21] W.-Y. Ting, W. Ren, R. Chao, H.-Y. Lin, Y. Tsao, and F.-G. Zeng, “MC-SEMamba: A simple multi-channel extension of SEMamba,” 2024, *arXiv:2409.17898*.

[22] H. N. Chau, T. D. Bui, H. B. Nguyen, T. T. H. Duong, and Q. C. Nguyen, “A novel approach to multi-channel speech enhancement based on graph neural networks,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 1133–1144, 2024.

[23] L. Pfeifenberger, M. Zöhrer, and F. Pernkopf, “DNN-based speech mask estimation for eigenvector beamforming,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 66–70.

[24] Y. Liu, A. Ganguly, K. Kamath, and T. Kristjansson, “Neural network based time-frequency masking and steering vector estimation for two-channel MVDR beamforming,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 6717–6721.

[25] R. Cheng and C. Bao, “Speech enhancement based on beamforming and post-filtering by combining phase information,” in *Proc. Interspeech*, 2020, pp. 4496–4500.

[26] Z.-Q. Wang and D. Wang, “Multi-microphone complex spectral mapping for speech dereverberation,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 486–490.

[27] Z.-Q. 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.

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

[29] M. Kim, S. Cheong, and J. W. Shin, “DNN-based parameter estimation for MVDR beamforming and post-filtering,” in *Proc. INTERSPEECH*, Dublin, Ireland, pp. 20–24, 2023.

[30] S. Hwang, M. Kim, and J. W. Shin, “Dual microphone speech enhancement based on statistical modeling of interchannel phase difference,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 2865–2874, 2022.

[31] S. Abdulatif, R. Cao, and B. Yang, “CMGAN: Conformer-based metric-gan for monaural speech enhancement,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 2477–2493, 2024.

[32] *Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codec*, Rec. ITU-R P.862.2, International Telecommunication Union, Geneva, Switzerland, 2007.

[33] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. TASSP-33, no. 2, pp. 443–445, Apr. 1985.

[34] Z.-Q. 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, Feb. 2019.

[35] M. Kim, H. Song, S. Cheong, and J. W. Shin, “iDeepMMSE: An improved deep learning approach to MMSE speech and noise power spectrum estimation for speech enhancement,” in *Proc. Interspeech*, Sep. 2022, pp. 181–185.

[36] B. Lay and T. Gerkmann, “An analysis of the variance of diffusion-based speech enhancement,” in *Proc. INTERSPEECH*, Dublin, Ireland, 2023, pp. 2205–2209.

[37] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, “An analysis of environment, microphone and data simulation mismatches in robust speech recognition,” *Comput. Speech Lang.*, vol. 46, pp. 535–557, Nov. 2017.

[38] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. Int. Conf. Learn. Representations*, 2015.

[39] *Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs*, Rec. ITU-R P. 862, International Telecommunication Union, Geneva, Switzerland, 2001.

[40] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “A short-time objective intelligibility measure for time-frequency weighted noisy speech,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2010, pp. 4214–4217.

[41] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, “SDR–half-baked or well done?,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2019, pp. 626–630.

[42] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, “TF-GridNet: Integrating full-and sub-band modeling for speech separation,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 3221–3236, 2023.