INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

DNN-based Parameter Estimation for MVDR Beamforming and Post-filtering

Minseung Kim, Sein Cheong, and Jong Won Shin

School of Electrical Engineering and Computer Science,
Gwangju Institute of Science and Technology, Gwangju 61005, Korea
kms0603@gm.gist.ac.kr, seiinjung@gm.gist.ac.kr, jwshin@gist.ac.kr

Abstract

Multi-channel speech enhancement systems usually consist of spatial filtering such as minimum-variance distortionless-response (MVDR) beamforming and post-processing, which require acoustic parameters including relative transfer function (RTF), noise spatial covariance matrix (SCM), and *a priori* and *a posteriori* signal-to-noise ratios (SNRs). In this paper, we propose a deep neural network (DNN)-based parameter estimation for MVDR beamforming and post-filtering. Specifically, we propose to use a DNN to estimate the interchannel phase differences of the clean speech and the speech presence probability (SPP), which are used to estimate the RTF and the noise SCM for MVDR beamforming. As for the post-processing, we adopt the iDeepMMSE framework in which another DNN is employed to estimate the *a priori* SNR, speech power spectral density, and SPP used to compute spectral gains. The proposed method outperformed several previous approaches especially in the PESQ scores for the CHiME-4 dataset.

Index Terms: multi-channel speech enhancement, interchannel phase difference, RTF estimation, Deep Xi, iDeepMMSE.

1. Introduction

In the presence of background noise and reverberation, speech enhancement is indispensable to ensure satisfactory perceptual quality and intelligibility of the speech signal [1–22] and performance of the subsequent speech applications such as speech recognition [1–3, 7, 8], speaker recognition [23], meeting summarization using smartphones [24], smart speakers [25], and hearing aids [26]. Nowadays many devices are equipped with multiple microphones, which enables multi-channel speech enhancement exploiting spatial information in addition to the spectro-temporal characteristics [1–6, 9–17]. One of the popular configurations is to apply spatial filtering such as the minimum-variance distortionless response (MVDR) beamforming and then process the output with a single-channel post-filtering [3, 4, 7, 22].

Recently, various deep learning approaches to the multi-channel speech enhancement have been proposed [1–6, 12–15]. These approaches can be classified into two categories. One is the neural beamforming in which a deep neural network (DNN) directly learns the relationship between multiple noisy inputs and output in an end-to-end way [2, 5, 6, 12–15]. The other approach is to combine deep learning with conventional beamforming methods, which often require the estimation of the acoustic parameters representing spatial and spectro-temporal characteristics such as the direction of arrival (DoA), relative transfer function (RTF), spatial covariance matrices (SCMs), and power spectral densities (PSDs) for speech and noise [1, 3, 4].

Many studies in the second category estimate SCMs using masks or signals estimated by DNNs [1, 3, 4]. In [3], the complex spectral mapping approach was proposed, but the DNNs which estimate the signals used to obtain the SCMs operated separately for each microphone without utilizing spatial information. While [4] uses the dominant eigenvector of the noisy SCM as a DNN input which considers the spatial information, it still had a limitation that the output is limited to the masks or SPPs. Also, since they computed the SCMs by averaging masked signals for all the frames, it was difficult to deal with the abrupt changes in signal statistics.

In this paper, we propose a DNN-based acoustic and statistical parameter estimation for the MVDR beamformer and post-filter. Specifically, we utilize a DNN to estimate the phase differences for clean speech and *a posteriori* speech presence probability (SPP), which are used to obtain the RTF and estimate the noise SCM through uni- or bi-directional multi-channel minima-controlled recursive averaging (MC-MCRA) for the MVDR beamformer. We also exploit another DNN to estimate the *a priori* signal-to-noise ratios (SNRs), speech PSD, and *a posteriori* SPP used for a single-channel post-filter adopting the iDeepMMSE framework [20]. The proposed method exhibited superior multi-channel speech enhancement performance compared to the previously proposed approaches in the experiments on the CHiME-4 dataset.

2. MVDR beamforming and post-filtering

Suppose that an array of M microphones captures speech signal from a speaker in the presence of additive noises and reverberation. The observed microphone signals can be expressed in the short-time Fourier transform (STFT) domain as

$$\begin{aligned}\mathbf{y}(l,k) &= \mathbf{s}(l,k) + \mathbf{v}(l,k)\\ &= \mathbf{g}(l,k)S_1(l,k) + \mathbf{v}(l,k), \end{aligned} \tag{1}$$

for $l = 1, 2, ..., L$, and for $k = 1, 2, ..., K$, where L is the number of frames, K is the number of frequency bins, and $\mathbf{y}(l,k) = [Y_1(l,k), Y_2(l,k), ..., Y_M(l,k)]^T$, $\mathbf{s}(l,k) = [S_1(l,k), S_2(l,k), ..., S_M(l,k)]^T$, and $\mathbf{v}(l,k) = [V_1(l,k), V_2(l,k), ..., V_M(l,k)]^T$, in which $Y_m(l,k)$, $S_m(l,k)$, and $V_m(l,k)$ are the STFT coefficients of the noisy speech, clean speech, and noises including reverberations at the m-th microphone, respectively, and $\mathbf{g}(l,k) = [1, g_2(l,k), ..., g_M(l,k)]^T$ is the RTF vector. When we assume $S_m(l,k)$ and $V_m(l,k)$ are uncorrelated, the SCMs for the $\mathbf{y}(l,k)$, $\mathbf{s}(l,k)$, and $\mathbf{v}(l,k)$, $\Phi_{\mathbf{y}}(l,k) = E[\mathbf{y}(l,k)\mathbf{y}^H(l,k)]$, $\Phi_{\mathbf{s}}(l,k) = E[\mathbf{s}(l,k)\mathbf{s}^H(l,k)]$, and $\Phi_{\mathbf{v}}(l,k) = E[\mathbf{v}(l,k)\mathbf{v}^H(l,k)]$, are related as $\Phi_{\mathbf{y}}(l,k) = \Phi_{\mathbf{s}}(l,k) + \Phi_{\mathbf{v}}(l,k)$.

The objective of the multi-channel speech enhancement is to obtain clean speech at the reference microphone $S_1(l,k)$

10.21437/Interspeech.2023-420

from the observed microphone signals $\mathbf{y}(l,k)$. One of the popular approaches to the multi-channel speech enhancement is to apply spatial filtering first and then employ single-channel post-filtering. One of the widely-used spatial filtering methods is the MVDR beamformer. The MVDR beamformer is a linear filter applied to $\mathbf{y}(l,k)$ to produce the output $Z(l,k)$, i.e.,

$$Z(l,k) = \mathbf{w}_{mvdr}^H(l,k)\mathbf{y}(l,k), \tag{2}$$

where the beamformer weights $\mathbf{w}_{mvdr}$ are designed to minimize the noise power at the output with the constraint on distortionless target speech. The MVDR beamformer can be described as a function of $\Phi_{\mathbf{v}}(l,k)$ and $\mathbf{g}(l,k)$ as

$$\mathbf{w}_{mvdr}(l,k) = \frac{\Phi_{\mathbf{v}}^{-1}(l,k)\mathbf{g}(l,k)}{\mathbf{g}^H(l,k)\Phi_{\mathbf{v}}^{-1}(l,k)\mathbf{g}(l,k)}. \tag{3}$$

As the spatial filtering cannot completely suppress noises, the post-filtering often follows the spatial filtering to further enhance the speech signal. Let us denote the residual noise at the beamformer output as $O(l,k)$, i.e.,

$$Z(l,k) = S_1(l,k) + O(l,k). \tag{4}$$

Adopting the minimum mean square error log-spectral amplitude (MMSE-LSA) [21] clean speech estimator, the clean speech spectrum can be estimated by applying the gain function $G_{mmse-lsa}(l,k)$ to the beamformer output $Z(l,k)$ as

$$\hat{S}_1(l,k) = G_{mmse-lsa}(l,k)Z(l,k), \tag{5}$$

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

$$G_{mmse-lsa}(l,k) = \frac{\xi(l,k)}{\xi(l,k)+1}\exp\left\{\frac{1}{2}\int_{v(l,k)}^{\infty}\frac{e^{-t}}{t}dt\right\}, \tag{6}$$

in which $v(l,k) = [\xi(l,k)/(\xi(l,k)+1)]\gamma(l,k)$, $\xi(l,k)$ is the *a priori* SNR and $\gamma(l,k)$ is the *a posteriori* SNR defined as

$$\xi(l,k) = \frac{\phi_s(l,k)}{\phi_o(l,k)}, \gamma(l,k) = \frac{|Z(l,k)|^2}{\phi_o(l,k)}, \tag{7}$$

where $\phi_s(l,k) = E[|S_1(l,k)|^2]$ and $\phi_o = E[|O(l,k)|^2]$ are the PSDs of the $S_1(l,k)$ and $O(l,k)$, respectively.

Note that estimates for $\Phi_{\mathbf{v}}$ and $\mathbf{g}$ are required to implement the MVDR beamforming in (2), and estimates for ξ and γ are needed for the post-filtering in (5). In this paper, we propose to incorporate DNNs to estimate these parameters for the multi-channel speech enhancement.

3. DNN-based parameter estimation for beamforming and post-filtering

The overall block diagram for the proposed multi-channel speech enhancement system is presented in Fig. 1. DNN$_1$ estimates the *a posteriori* SPP and the interchannel phase differences (IPDs) for the clean speech from the IPDs for the input signal and the magnitude spectrum of the reference microphone signal, which in turn are used to obtain $\widehat{\Phi}_{\mathbf{v}}$ and $\widehat{\mathbf{g}}$ for the MVDR beamformer. DNN$_2$ is employed to obtain the *a priori* SNR, speech PSD and *a posteriori* SPP needed for the post-filtering in the iDeepMMSE framework [20]. Detailed descriptions of the blocks are given in the following subsections.

3.1. DNN-based parameter estimation for MVDR beamforming

To implement $\mathbf{w}_{mvdr}(l,k)$ in (3), $\Phi_{\mathbf{v}}(l,k)$ and $\mathbf{g}(l,k)$ need to be estimated. Among various approaches to the noise SCM estimation, we adopt the MC-MCRA approach [17] and optionally extend it to a bi-directional version. Let two hypotheses H_0 and H_1 denote speech absence and presence, respectively. The estimate of $\Phi_{\mathbf{v}}(l,k)$, $\widehat{\Phi}_{\mathbf{v}}^f(l,k)$, may be updated under each hypothesis as

$$\begin{aligned} H_0(l,k): \widehat{\Phi}_{\mathbf{v}}^f(l,k) =& \alpha_v\widehat{\Phi}_{\mathbf{v}}^f(l-1,k) \\ &+ (1-\alpha_v)\mathbf{y}(l,k)\mathbf{y}^H(l,k), \end{aligned} \tag{8}$$

$$H_1(l,k): \widehat{\Phi}_{\mathbf{v}}^f(l,k) = \widehat{\Phi}_{\mathbf{v}}^f(l-1,k), \tag{9}$$

where α_v is a constant parameter. Combining the update equations under two hypotheses using the *a posteriori* SPP $p_s(l,k)$, the MC-MCRA noise SCM estimator can be derived as

$$\begin{aligned} \widehat{\Phi}_{\mathbf{v}}^f(l,k) =& \tilde{\alpha}_v(l,k)\widehat{\Phi}_{\mathbf{v}}^f(l-1,k) \\ &+ (1-\tilde{\alpha}_v(l,k))\mathbf{y}(l,k)\mathbf{y}^H(l,k) \end{aligned} \tag{10}$$

for $l = 1,2,...,L$, where $\tilde{\alpha}_v(l,k) = \alpha_v + p_s(l,k)(1-\alpha_v)$ is an SPP-dependent smoothing parameter. As for the offline applications such as meeting summarization [24], the MC-MCRA approach can also be applied in the backward direction, i.e.,

$$\begin{aligned} \widehat{\Phi}_{\mathbf{v}}^b(l,k) =& \tilde{\alpha}_v(l,k)\widehat{\Phi}_{\mathbf{v}}^b(l+1,k) \\ &+ (1-\tilde{\alpha}_v(l,k))\mathbf{y}(l,k)\mathbf{y}^H(l,k) \end{aligned} \tag{11}$$

for $l = L, L-1, ..., 1$ and $\widehat{\Phi}_{\mathbf{v}}^b(l,k)$ is the noise SCM estimate in the backward direction. The final estimate for the noise SCM in the bi-directional MC-MCRA (BMC-MCRA) is given by

$$\widehat{\Phi}_{\mathbf{v}}(l,k) = 0.5\cdot\widehat{\Phi}_{\mathbf{v}}^f(l,k) + 0.5\cdot\widehat{\Phi}_{\mathbf{v}}^b(l,k). \tag{12}$$

It is noted that the key parameter for both uni-directional and bi-directional MC-MCRA is the *a posteriori* SPP $p_s(l,k)$. In this paper, we employ a DNN to estimate $p_s(l,k)$. The training target for the $p_s(l,k)$ is constructed by thresholding the SNR:

$$p_s(l,k) = \begin{cases} 1 & \text{if } \frac{|S_1(l,k)|^2}{|V_1(l,k)|^2} > \eta \\ 0 & \text{otherwise,} \end{cases} \tag{13}$$

where η is the threshold.

On the other hand, we estimate the RTF $\mathbf{g}(l,k)$ under the far-field assumption [16], i.e., $|g_m| = 1$ for $m \in \{1,...,M\}$. Then, the RTF is given as a simple function of the IPD for the clean speech, $\psi_m(l,k)$:

$$g_m(l,k) = \exp\{j\cdot\psi_m(l,k)\}, \tag{14}$$

where $\psi_m(l,k) = \angle S_m(l,k) - \angle S_1(l,k)$. Although the far-field assumption is not always met in practical scenarios, this simplified model may make the RTF estimation more robust in the presence of severe noises. In this paper, we propose to estimate the IPDs for the clean speech, $\psi_m(l,k)$, from the IPDs for the noisy signals, $\theta_m(l,k) = \angle Y_m(l,k) - \angle Y_1(l,k)$, using a DNN as in [27]. In [27], the sine and cosine functions of the IPDs are used as inputs and outputs of the DNN. In this paper, we slightly modify them so that the inputs and outputs

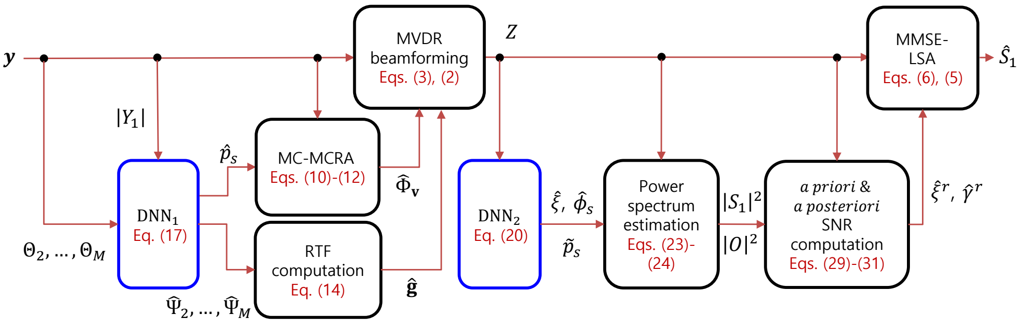

Figure 1: *Block diagram of the proposed multi-channel speech enhancement system.*

have values between 0 and 1:

$$\Psi_m(l,k) =[\frac{\sin\psi_m(l,k)+1}{2}, \frac{\cos\psi_m(l,k)+1}{2}], \tag{15}$$

$$\Theta_m(l,k) =[\frac{\sin\theta_m(l,k)+1}{2}, \frac{\cos\theta_m(l,k)+1}{2}], \quad m \in \{2,...,M\}, \tag{16}$$

which are the training targets and inputs, respectively.

Combining the tasks to estimate the *a posteriori* SPP p_s in (13) and sinusoidal functions of clean IPDs ψ_m in (15), we propose to exploit one DNN to estimate both of the parameters at once, given the magnitude spectrum of the noisy speech at the reference microphone and the sinusoidal functions of the noisy IPDs Θ_m in (16), which is expressed as

$$[\widehat{p}_s, \widehat{\Psi}_2, ..., \widehat{\Psi}_M] = \text{DNN}_1(|Y_1|, \Theta_2, ..., \Theta_M), \tag{17}$$

where $\widehat{p}_s = \{\widehat{p}_s(l,k)|1 \le k \le K\}$, $\widehat{\Psi}_m = \{\widehat{\Psi}_m(l,k)|1 \le k \le K\}$, $|Y_1| = \{|Y_1(l,k)||1 \le k \le K, 1 \le l \le L\}$, $\Theta_m = \{\Theta_m(l,k)|1 \le k \le K, 1 \le l \le L\}$, and $\text{DNN}_1(\cdot)$ can be any DNN architecture that can effectively deal with correlated signals. $\widehat{p}_s$ is used to compute $\widehat{\Phi}_{\mathbf{v}}$ in (10) or (12) and $\widehat{\Psi}_m$ is converted to $\widehat{\psi}_m$ using the four-quadrant inverse tangent function, which in turn is used to compute g_m as in (14).

3.2. iDeepMMSE framework for post-filtering

To evaluate the gain function in (6), the *a priori* and *a posteriori* SNRs should be estimated. In this paper, we adopt the iDeepMMSE framework [20] which employs a DNN to estimate the *a priori* SNR, speech PSD and *a posteriori* SPP from the noisy magnitude spectrogram. While the original iDeepMMSE framework was applied to the input signal for single-channel speech enhancement, we employ this approach for the enhancement of the beamformer output Z. In the signal model in (4), we want to estimate $S_1(l,k)$ in the presence of additive residual noise $O(l,k)$. Let ϕ_s and ϕ_o denote the PSDs for $S_1(l,k)$ and $O(l,k)$, respectively. The training target for the *a priori* SNR is set to be the instantaneous SNR given by [19]

$$\xi^{inst}(l,k) = \frac{|S_1(l,k)|^2}{|O(l,k)|^2}. \tag{18}$$

Instead of directly estimating $\xi^{inst}(l,k)$ which has a large dynamic range, a function of it, $\bar{\xi}(l,k)$, that has a value between 0 and 1 given by

$$\bar{\xi}(l,k) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{\xi^{inst}_{dB}(l,k) - \mu(k)}{\sigma(k)\sqrt{2}}\right)\right], \tag{19}$$

is estimated in [19], where $\text{erf}(\cdot)$ is the error function, $\xi^{inst}_{dB}(l,k) = 10\log_{10}(\xi^{inst}(l,k))$ and $\mu(k)$ and $\sigma(k)$ are the mean and standard deviation of $\xi^{inst}_{dB}(l,k)$ in the k-th frequency bin for the training data filtered with the MVDR beamformer with ground truth $\Phi_{\mathbf{v}}$ and $\mathbf{g}$. The training target for the *a posteriori* SPP is constructed in a similar way to (13) replacing V_1 with O, while that for the speech PSD is set through a dynamic range compression similar to (19) as described in [20].

In the inference phase of iDeepMMSE, mapped versions of the *a priori* SNR $\bar{\xi}$ and speech PSD $\bar{\phi}_s$ along with the *a posteriori* SPP $\tilde{p}_s$, are obtained from Z using another DNN:

$$[\hat{\bar{\xi}}, \hat{\bar{\phi}}_s, \tilde{p}_s] = \text{DNN}_2(|Z|), \tag{20}$$

where $\text{DNN}_2(\cdot)$ can also be any DNN architecture. It is noted that $\tilde{p}_s$ is not directly related to $\widehat{p}_s$ obtained in (17), although they may be combined in the future. The estimates for the *a priori* SNR and speech PSD are obtained using the inverse functions, respectively, as

$$\widehat{\xi}(l,k) = 10^{\{\sigma(k)\sqrt{2}\text{erf}^{-1}(2\hat{\bar{\xi}}(l,k)-1)+\mu(k)\}/10}, \tag{21}$$

$$\widehat{\phi}_s(l,k) = 10^{\{\sigma_s(k)\sqrt{2}\text{erf}^{-1}(2\hat{\bar{\phi}}_s(l,k)-1)+\mu_s(k)\}/10}, \tag{22}$$

where $\mu_s(k)$ and $\sigma_s(k)$ are estimated from the histogram of $|S_1|^2$ in the speech active time-frequency bins [20]. The estimated parameters are then utilized to evaluate the MMSE speech and noise power spectrum estimators. These estimators under the speech presence uncertainty are given as

$$\begin{aligned}\widehat{|S_1|^2} =&E(|S_1|^2|Z,\phi_s,\phi_o)\\ =&p(H_0|Z)\cdot E(|S_1|^2|Z,\phi_s,\phi_o,H_0)\\ &+p(H_1|Z)\cdot E(|S_1|^2|Z,\phi_s,\phi_o,H_1),\end{aligned} \tag{23}$$

$$\begin{aligned}\widehat{|O|^2} =&E(|O|^2|Z,\phi_s,\phi_o)\\ =&p(H_0|Z)\cdot E(|O|^2|Z,\phi_s,\phi_o,H_0)\\ &+p(H_1|Z)\cdot E(|O|^2|Z,\phi_s,\phi_o,H_1),\end{aligned} \tag{24}$$

where $p(H_1|Z) = \tilde{p}_s$ and $p(H_0|Z) = 1 - p(H_1|Z)$, and

$$E(|S_1|^2|Z,\phi_s,\phi_o,H_0) =0, \tag{25}$$

$$E(|O|^2|Z,\phi_s,\phi_o,H_0) =|Z|^2, \tag{26}$$

$$\begin{aligned}E(|S_1|^2|Z,&\phi_s,\phi_o,H_1)\\ &=\left(\frac{\phi_s}{\phi_o+\phi_s}\right)^2|Z|^2+\frac{\phi_o}{\phi_s+\phi_o}\phi_s,\end{aligned} \tag{27}$$

$$E(|O|^2|Z,\phi_s,\phi_o,H_1) \\ = \left(\frac{\phi_o}{\phi_o+\phi_s}\right)^2|Z|^2+\frac{\phi_s}{\phi_s+\phi_o}\phi_o. \quad (28)$$

(27) and (28) are evaluated using $\widehat{\phi}_s$ and $\widehat{\phi}_o = \frac{1}{1+\widehat{\xi}}|Z|^2$. Then, the refined speech and noise PSDs can be obtained by applying temporal recursive smoothing to $\widehat{|S_1|^2}$ and $\widehat{|O|^2}$ as

$$\widehat{\phi}^r_s(l,k) = \alpha_s\widehat{\phi}^r_s(l-1,k)+(1-\alpha_s)\widehat{|S_1|^2}(l,k), \quad (29)$$

$$\widehat{\phi}^r_o(l,k) = \alpha_o\widehat{\phi}^r_o(l-1,k)+(1-\alpha_o)\widehat{|O|^2}(l,k), \quad (30)$$

where the superscript ^{r} denotes they are refined versions and α_s and α_o are smoothing parameters. Using $\widehat{\phi}^r_s(l)$ and $\widehat{\phi}^r_o(l)$, we can refine the estimates for the *a priori* and *a posteriori* SNRs as

$$\widehat{\xi}^r(l,k) = \frac{\widehat{\phi}^r_s(l,k)}{\widehat{\phi}^r_o(l,k)}, \widehat{\gamma}^r(l,k) = \frac{|Z(l,k)|^2}{\widehat{\phi}^r_o(l,k)}. \quad (31)$$

Finally, the gain function (6) is evaluated using $\widehat{\xi}^r$ and $\widehat{\gamma}^r$.

4. Experiments

4.1. Experimental settings

For the experiments, we used the simulated set with 6 microphones in the CHiME-4 database [28]. Four noise scenarios are considered in the dataset including the bus, cafe, pedestrian area, and street junction noises. The training set consists of 7,138 utterances from 83 speakers, while the development and evaluation sets are 1,640 and 1,320 utterances, respectively, spoken by 4 different speakers. The sampling rate was 16 kHz, and the frame size was 32ms with 50% overlap. A square-root Hann window was used for analysis and synthesis, and the 512-point STFT was applied. The fifth microphone located at the bottom center on the frontal surface of the tablet device was selected as the reference channel for the algorithms and the evaluations. The parameter values for α_v, α_s, α_o, and η were set to 0.9, 0.0, 0.0, and -12 dB, respectively.

As for the architectures for the DNNs in (17) and (20), we adopted the Conformer [29] for both of them, which can efficiently capture both the local and global sequential information. Each Conformer block includes a multi-head self-attention module and Convolution module sandwiched by two feed-forward modules, followed by layer normalization [29]. After the repeated stacks of Conformer blocks, a fully-connected layer with sigmoidal activation is placed to produce the outputs. The dimension of DNN output was 2827 ($= 257 \times 11$) for DNN$_1$ and 771 ($= 257 \times 3$) for DNN$_2$. The other configurations for the network structure were the same as what are described in [20]. The loss function to train both of the networks was the binary cross entropy. We used Adam optimizer [30] with CosineAnnealingWarmRestarts scheduler [31]. The maximum number of epochs was 500 and the mini-batch size was 5 and the best epoch was chosen based on the validation loss.

The performance of the proposed method is compared with recent papers reporting the performances for CHiME-4 dataset [1–6] in terms of the wideband (WB) and narrowband (NB) perceptual evaluation of speech quality (PESQ) scores [32], short-time objective intelligibility (STOI) [33], and scale-invariant signal-to-distortion ratio (SI-SDR) [34]. For the papers that reported performances for multiple configurations, the systems with the highest performance were compared.

Table 1: *The performance of multi-channel speech enhancement for the previous approaches and the proposed methods on the CHiME-4 dataset.*

	NB PESQ	WB PESQ	STOI	SI-SDR
Noisy	2.18	1.27	0.87	7.51
Li *et al.* [1]	2.68	-	0.95	14.10
Zhang *et al.* [2]	2.96	-	0.96	17.52
Wang *et al.* [3]	**3.68**	-	**0.986**	**22.0**
Pfeifenberger *et al.* [4]	-	1.86	-	-
Tolooshams *et al.* [5]	-	2.436	-	18.635
Lee *et al.* [6]	-	2.67	0.973	19.67
Proposed (MC-MCRA)	3.43	3.00	0.971	18.54
Proposed (BMC-MCRA)	3.48	**3.09**	0.974	19.31

4.2. Experimental results

Table 1 shows the performances for the previous approaches and the proposed methods with the MC-MCRA and BMC-MCRA noise SCM estimators on the CHiME-4 dataset. It is noted that some papers showed NB PESQ scores, while others reported WB PESQ scores. Among the compared methods, Wang *et al.* [3] showed the best performance with considerably more parameters of around 26 million compared with the proposed system which had 8.3 million parameters. Lee *et al.* [6] exhibited higher SI-SDR and similar STOI compared with the proposed method, but showed much lower WB PESQ scores. The proposed method outperformed other approaches in terms of the PESQ scores, STOI, and SI-SDR and the adoption of the bi-directional MC-MCRA further improved them. Especially, the average WB PESQ score for the proposed method with the BMC-MCRA was significantly higher than those for Lee *et al.* [6], Tolooshams *et al.* [5], and Pfeifenberger *et al.* [4] with margins of 0.42, 0.654, and 1.23, respectively.

5. Conclusion

In this paper, we propose a deep learning approach to the acoustic and statistical parameter estimation for multi-channel speech enhancement. We propose to adopt a DNN to estimate the *a posteriori* SPP and the IPDs for clean speech, which are used to estimate the noise SCM and RTF needed to construct the MVDR beamformer. In addition, we optionally apply the bi-directional MC-MCRA approach for the noise SCM estimation for offline applications. As for the post-filter, we follow the iDeepMMSE framework to estimate the *a priori* SNR, speech PSD, and *a posteriori* SPP with another DNN and compute spectral gains using them. Experimental results on the CHiME-4 dataset showed that the proposed method outperformed several previous approaches especially in the PESQ scores.

6. Acknowledgements

This work was supported in part by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2023-2021-0-01835) supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation) and by IITP grant funded by the Korea government (MSIT) (No.2019-0-01842, Artificial Intelligence Graduate School Program (GIST)).

7. References

[1] C. Li, J. Shi, W. Zhang, A. S. Subramanian, X. Chang, N. Kamo, M. Hira, T. Hayashi, C. Boeddeker, Z. Chen, and S. Watanabe, "ESPnet-SE: end-to-end speech enhancement and separation toolkit designed for ASR integration," in *IEEE Spoken Lang. Technology Workshop (SLT)*, 2021, pp. 785–792.

[2] W. Zhang, J. Shi, C. Li, S. Watanabe, and Y. Qian, "Closing the gap between time-domain multi-channel speech enhancement on real and simulation conditions," in *IEEE Workshop on Applications of Signal Proc. to Audio and Acoust. (WASPAA)*, 2021, pp. 146–150.

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

[4] L. Pfeifenberger, M. Zöhrer, and F. Pernkopf, "DNN-based speech mask estimation for eigenvector beamforming," in *IEEE International Conference on Acoust., Speech and Signal Proc. (ICASSP)*, 2017, pp. 66–70.

[5] B. Tolooshams, R. Giri, A. H. Song, U. Isik, and A. Krishnaswamy, "Channel-attention dense u-net for multichannel speech enhancement," in *IEEE International Conference on Acoust., Speech and Signal Proc. (ICASSP)*, 2020, pp. 836–840.

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

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

[8] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust mvdr beamforming using time-frequency masks for online/offline asr in noise," in *IEEE International Conference on Acoust., Speech and Signal Proc. (ICASSP)*, Mar. 2016, pp. 5210–5214.

[9] T. Dietzen, S. Doclo, M. Moonen, and T. van Waterschoot, "Square root-based multi-source early psd estimation and recursive retf update in reverberant environments by means of the orthogonal procrustes problem," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 755–769, Jan. 2020.

[10] ——, "Integrated sidelobe cancellation and linear prediction kalman filter for joint multi-microphone speech dereverberation, interfering speech cancellation, and noise reduction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 740–754, Jan. 2020.

[11] T. Dietzen, M. Moonen, and T. van Waterschoot, "Instantaneous psd estimation for speech enhancement based on generalized principal components," in *2020 28th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 191–195.

[12] D. Markovic, A. Defossez, and A. Richard, "Implicit neural spatial filtering for multichannel source separation in the waveform domain," *arXiv preprint arXiv:2206.15423*, 2022.

[13] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6394–6398.

[14] Y. Luo and N. Mesgarani, "Implicit filter-and-sum network for end-to-end multi-channel speech separation." in *Interspeech*, 2021, pp. 3071–3075.

[15] H. Kim, K. Kang, and J. W. Shin, "Factorized MVDR deep beamforming for multi-channel speech enhancement," *IEEE Signal Proc. Lett.*, vol. 29, Aug. 2022.

[16] S. Hwang, M. Kim, and J. W. Shin, "Dual microphone speech enhancement based on statistical modeling of interchannel phase difference," *IEEE/ACM Trans. on Audio, Speech, and Lang. Proc.*, vol. 30, Aug. 2022.

[17] M. Souden, J. Chen, J. Benesty, and S. Affes, "An integrated solution for online multichannel noise tracking and reduction," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 19, no. 7, pp. 2159–2169, Sep. 2011.

[18] M. Kim and J. W. Shin, "Improved speech enhancement considering speech PSD uncertainty," *IEEE/ACM Trans. on Audio, Speech, and Lang. Proc.*, vol. 30, Jun. 2022.

[19] A. Nicolson and K. K. Paliwal, "Deep learning for minimum mean-square error approaches to speech enhancement," *Speech Communication*, vol. 111, pp. 44–55, Aug. 2019.

[20] 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 *Interspeech*, Sep. 2022, pp. 181–185.

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

[22] M. Kim, S. Cheong, H. Song, and J. W. Shin, "Improved speech spatial covariance matrix estimation for online multi-microphone speech enhancement," *Sensors*, vol. 23, no. 1, p. 111, 2022.

[23] A. E. Rosenberg, "Automatic speaker verification: A review," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 475–487, Apr. 1976.

[24] G. Murray, S. Renals, and J. Carletta, "Extractive summarization of meeting recordings," in *EUROSPEECH*, Sep. 2005, pp. 593–596.

[25] R. Haeb-Umbach, S. Watanabe, T. Nakatani, M. Bacchiani, B. Hoffmeister, M. L. Seltzer, H. Zen, and M. Souden, "Speech processing for digital home assistants: Combining signal processing with deep-learning techniques," *IEEE Signal Proc. Magazine*, vol. 36, no. 6, pp. 111–124, Nov. 2019.

[26] J. M. Kates, *Digital Hearing Aids*. San Diego, CA, USA: Plural Publishing, 2008.

[27] J. Pak and J. W. Shin, "Sound localization based on phase difference enhancement using deep neural networks," *IEEE/ACM Trans. on Audio, Speech, and Lang. Proc.*, vol. 27, no. 8, pp. 1335–1345, 2019.

[28] 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," *Computer Speech & Lang.*, vol. 46, pp. 535–557, Nov. 2017.

[29] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

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

[31] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[32] *Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codec*, ITU Recommendation P.862.2, 2007.

[33] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *IEEE International Conference on Acoust., Speech and Signal Proc. (ICASSP)*, Mar. 2010, pp. 4214–4217.

[34] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" in *IEEE International Conference on Acoust., Speech and Signal Proc. (ICASSP)*, May 2019, pp. 626–630.