DiffCBF: A Diffusion Model with Convolutional Beamformer for Joint Speech Separation, Denoising, and Dereverberation

Rino Kimura*, Tetsuya Ueda*, Tomohiro Nakatani†, Naoyuki Kamo†,
Marc Delcroix†, Shoko Araki†, Shoji Makino*
*Waseda University, 2-7 Hibikino, Wakamatsu-ku, Kita-Kyushu 808-0135, Japan †NTT Corporation, Japan
Email: kimura.r@suou.waseda.jp, t.ueda@akane.waseda.jp, tnak@ieee.org, s.makino@ieee.org

***Abstract*—This paper proposes a new multi-channel Speech Enhancement (SE) method that simultaneously performs denoising, dereverberation, and source separation. The method combines a diffusion model-based approach, the multi-stream Score-based Generative Model for Speech Enhancement (ms-SGMSE), with a signal processing-based approach, the Convolutional Beamformer (CBF). We refer to this integrated method as the Diffusion model with CBF (DiffCBF). By leveraging the strengths of both methods, it improves estimation accuracy through iterative refinement. Additionally, it can process captured signals regardless of the number of speech sources as long as the source count is provided, making it highly versatile. Experimental results demonstrate that the proposed method significantly enhances the quality of clean speech from noisy and reverberant speech mixtures, greatly outperforming a conventional diffusion model-based source separation method.**

***Index Terms*—Diffusion model, source separation, denoising, dereverberation, multi-input multi-output, microphone array**

I. Introduction

This paper proposes a new multi-channel Speech Enhancement (SE) approach based on the diffusion model [1]–[3]. Speech signals captured by distant microphones are often degraded by noise, reverberation, and overlapping speech. The goal of multi-channel SE in this paper is to restore clean speech signals from such degraded inputs by combining denoising, dereverberation, and speech separation.

For multi-channel SE, deterministic Neural Network (NN)-based and Signal Processing (SP)-based approaches have been developed [4]. NN-based SE [5]–[7] learns a mapping from distorted speech to clean speech using training data, achieving high-quality SE when training and test conditions align. SP-based SE (e.g., [8], [9]), on the other hand, relies on assumptions about room acoustics and signal characteristics, typically not requiring prior training to achieve SE. It demonstrates high adaptability to various test conditions. Recognizing the complementary strengths of these approaches, researchers have actively explored techniques that integrate them, often leading to state-of-the-art SE performance [10]–[13].

Recently, the diffusion model [1], [2], an emerging probabilistic NN-based technology, has shown success in applications such as denoising [14], dereverberation [3], [15], and source separation [16], partially surpassing the deterministic NN approach. It models the conditional density of clean speech given distorted speech and samples clean speech estimates from this density. Compared to deterministic NN-based SE, it enhances perceptual speech quality [15] and improves robustness against training-test mismatches [3]. It can also significantly improve signal-level distortion metrics like the Signal-to-Distortion Ratio (SDR) [17] using ensemble inference [18], [19]. A representative method based on the diffusion model is the Score-based Generative Model for Speech Enhancement (SGMSE) [3]. It has been extended to a single-channel source separation method, DiffSep [20]. In addition, the multi-stream SGMSE (ms-SGMSE) enables multi-channel processing for simultaneous denoising and dereverberation [21], and can be integrated with other SE methods to enhance estimation accuracy [18], [22], [23].

Despite its potential, diffusion model-based SE has yet to be tested in challenging recording conditions that require simultaneous source separation, denoising, and dereverberation. For instance, as our experiments will show, applying DiffSep [20] in such conditions does not yield satisfactory results. This limitation prevents the approach from reaching its full potential.

Based on the above research background, this paper proposes a new multi-channel SE method based on ms-SGMSE. Unlike DiffSep, which independently performs source separation, the proposed method combines an SP-based SE technique, the Convolutional Beamformer (CBF) [9], with ms-SGMSE. While CBF performs simultaneous denoising, dereverberation, and source separation (without prior training), ms-SGMSE refines each CBF output using the diffusion model. We refer to this integrated method as the Diffusion model with CBF (DiffCBF). DiffCBF utilizes iterative estimation to combine the strengths of both methods. Additionally, DiffCBF can effectively process a captured signal regardless of the number of speech sources as long as the source count is known. Experimental results show that DiffCBF effectively addresses challenging noisy and reverberant speech mixtures, greatly outperforming the conventional DiffSep method.

II. Problem definition

Suppose an array of M microphones captures a mixture of N reverberant speech signals with diffuse noise. In this paper, multi-channel SE refers to the process of estimating the direct component of each speech signal, referred to as a clean speech signal, from such a captured signal. We denote the

captured signal as $\boldsymbol{y} \in \mathbb{C}^{F\times T\times M}$ and each clean speech signal as $\boldsymbol{x}_n \in \mathbb{C}^{F\times T\times M}$ for $1 \le n \le N$ in the complex spectrum domain, where F and T denote the numbers of frequencies and time frames, respectively.

III. Conventional methods

This section describes two conventional methods: CBF [9] and ms-SGMSE [21], [23]. They will be incorporated into our proposed method in the next section.

A. Convolutional Beamformer (CBF)

CBF is a multi-channel SE method that can work with no prior training on the signals or recording conditions [9]. Instead of relying on prior training, it assumes that the captured signal is a convolutional mixture of N ($\le M$) speech signals and stationary noise signals. To recover the clean speech estimates, it uses Multi-Channel Linear Prediction (MCLP) to dereverberate the captured mixture [24] and Beamformer (BF) to separate and denoise the dereverberated mixture into clean speech estimates [25], [26].

Let $\boldsymbol{y}_{t,f} \in \mathbb{C}^M$ represent the captured signal at a time frame t and a frequency f. CBF then performs multi-channel SE at each frequency:

$$\hat{\boldsymbol{x}}_{t,f}^{CBF} = \boldsymbol{W}_f^{\mathsf{H}} \left(\boldsymbol{y}_{t,f} - \sum_{\tau=D}^{L} \boldsymbol{G}_{\tau,f}^{\mathsf{H}} \boldsymbol{y}_{t-\tau,f} \right), \tag{1}$$

where $\hat{\boldsymbol{x}}_{t,f}^{CBF} = [\hat{x}_{1,t,f}^{CBF}, \ldots, \hat{x}_{N,t,f}^{CBF}, \hat{v}_{1,t,f}^{CBF}, \ldots, \hat{v}_{M-N,t,f}^{CBF}]^T \in \mathbb{C}^M$ is a vector containing estimated N speech signals $\{\hat{x}_{n,t,f}^{CBF}\}_{n=1}^N$ and noise signals $\{\hat{v}_{n,t,f}^{CBF}\}_{n=N+1}^M$, and $(\cdot)^{\mathsf{H}}$ and $(\cdot)^{\top}$ denote the conjugate and non-conjugate transpose operators. The terms in parentheses in (1) correspond to MCLP with prediction matrices $\boldsymbol{G}_{\tau,f} \in \mathbb{C}^{M\times M}$ for $D \le \tau \le L$, where L and D are the prediction order and delay. $\boldsymbol{W}_f \in \mathbb{C}^{M\times M}$ is a BF matrix applied to the output of MCLP.

To estimate the prediction matrices and the BF matrices, CBF assumes that each speech signal at each Time-Frequency (TF) point follows a complex Gaussian distribution with a TF-dependent variance, $\lambda_{n,t,f}$, and that each noise signal is stationary Gaussian. By further assuming mutual independence between the signals and noise over TF points, we derive the following log-likelihood function for the estimation.

$$\mathcal{L}(\theta) = \log p\left(\boldsymbol{y}; \theta\right) + \log p\left(\{\lambda_{n,t,f}\}_{n,t,f}; \theta_\lambda\right), \tag{2}$$

$$= -\sum_{t,f}\left[\sum_{n=1}^{N}\left(\log \lambda_{n,t,f} + \frac{|\hat{x}_{n,t,f}^{CBF}|^2}{\lambda_{n,t,f}}\right) + \sum_{n=1}^{M-N} |\hat{v}_{n,t,f}^{CBF}|^2\right]$$

$$+ 2T\sum_f \log|\det \boldsymbol{W}_f| + \sum_{n,t,f} \log p(\lambda_{n,t,f}; \theta_\lambda) + \text{const.}, \tag{3}$$

where $\theta = \{\{\lambda_{n,t,f}\}_{n,t,f}, \{\boldsymbol{W}_f\}_f, \{\boldsymbol{G}_{\tau,f}\}_{\tau,f}\}$ is a set of parameters to be estimated by maximizing (3), and $\log p(\{\lambda_{n,t,f}\}_{n,t,f}; \theta_\lambda)$ is an optional term representing the prior distribution of $\lambda_{n,t,f}$ based on its prior knowledge, θ_λ.

An advantage of CBF, which will be leveraged in our proposed method, lies in its flexible mechanism for improving estimation based on the availability of the prior knowledge, θ_λ [11]. Without θ_λ, we can estimate θ in an unsupervised learning manner, disregarding the optional term. On the other hand, if reliable estimates of clean speech power spectra, $\sigma_{n,t,f}^2 \simeq E\{|x_{n,t,f}|^2\}$, are available, such as those obtained from a NN, we can enhance estimation accuracy by incorporating the prior term. The prior distribution is defined by using the inverse Gamma distribution:

$$p(\lambda_{n,t,f}; \theta_\lambda) = IG(\lambda_{n,t,f}; \alpha, \beta), \tag{4}$$

with the shape parameter $\alpha > 0$ and setting the scale parameter $\beta = \sigma_{n,t,f}^2$. In both scenarios, we can derive computationally efficient estimation algorithms for CBF (see [9] for details).

Note that CBF can provide a multi-channel clean speech estimate for each n, denoted as $\hat{\boldsymbol{x}}_n^{CBF} \in \mathbb{C}^{F\times T\times M}$, through a post-processing step called projection back [27]. This version of CBF output is used in our proposed method.

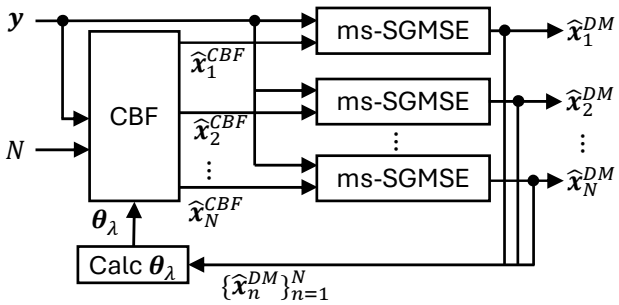

Fig. 1. Processing flow of DiffCBF. The symbols, $\hat{\boldsymbol{x}}_n^{CBF}$ and $\hat{\boldsymbol{x}}_n^{DM}$, denote the outputs of CBF and ms-SGMSE (i.e., Diffusion Model), respectively. All ms-SGMSE blocks in the figure utilize the same score model.

B. ms-SGMSE

SGMSE is a single-channel SE method based on the diffusion model [3]. It models the conditional density of a clean speech signal given the captured signal, i.e., $p(\boldsymbol{x}|\boldsymbol{y})$, using the diffusion process. The method performs SE by sampling a clean speech estimate from this density. This is accomplished by solving a reverse Stochastic Differential Equation (SDE), with the conditional score estimated by a NN.

ms-SGMSE is a multi-stream extension of SGMSE [21], [23]. It incorporates additional signal streams, denoted by χ, into the density to be modeled, i.e., $p(\boldsymbol{x}|\boldsymbol{y}, \chi)$, as additional conditions. It enables more precise distribution modeling, leading to improved SE accuracy. It has been demonstrated that ms-SGMSE can realize multi-channel denoising and dereverberation by adding microphone signals as additional streams [21]. Furthermore, it can leverage enhanced signals obtained from various other SE methods as additional streams, proving highly effective in improving SE accuracy [23].

IV. Proposed method: DiffCBF

This subsection describes our proposed method, DiffCBF, which realizes accurate multi-channel SE by integrating CBF and ms-SGMSE. Refer to the advantages summarized in Section IV-C for the motivation behind the approach.

A. Processing flow

Figure 1 illustrates the processing flow of DiffCBF, which iteratively alternates between CBF and ms-SGMSE to improve SE progressively.

In the first iteration, given the captured signal $\boldsymbol{y}$ and the number of speech signals N, the method begins by applying CBF to $\boldsymbol{y}$ in an unsupervised manner, without using the prior term in (3). This yields N clean speech estimates, denoted as $\{\hat{\boldsymbol{x}}_n^{CBF}\}_{n=1}^N$. Next, for each n, ms-SGMSE is applied to $\boldsymbol{y}$, where the output $\hat{\boldsymbol{x}}_n^{CBF}$ from CBF serves as an additional stream in the diffusion process. This results in improved clean speech estimates, $\{\hat{\boldsymbol{x}}_n^{DM}\}_{n=1}^N$. In subsequent iterations, CBF additionally receives, as prior knowledge θ_λ, the power spectra of the clean speech estimates obtained by ms-SGMSE in the previous iteration. As a result, each new iteration allows CBF to utilize the inverse Gamma prior with the given θ_λ in (4) and thus can enhance estimation accuracy.

In the flow described above, ms-SGMSE performs SE by solving the following reverse SDE [2], [3], [23], [28]:

$$\begin{aligned}\mathrm{d}\boldsymbol{z}_k = &\left[-\gamma\mathbf{f}(\boldsymbol{z}_k,\boldsymbol{y}) + g(k)^2\boldsymbol{r}(k,\boldsymbol{z}_k,\boldsymbol{y},\hat{\boldsymbol{x}}_n^{CBF})\right]\mathrm{d}k \\ &+ g(k)\mathrm{d}\bar{\boldsymbol{w}},\end{aligned} \tag{5}$$

where $\boldsymbol{z}_k \in \mathbb{C}^{F\times T\times M}$ is the state of the diffusion process at a state index k, $\mathbf{f}(\boldsymbol{z}_k,\boldsymbol{y}) = \boldsymbol{y}-\boldsymbol{z}_k$ and $g(k) = ca^k$ for constants $c, a > 0$ are the drift and noise scheduling functions, $\gamma > 0$ is a stiffness parameter, and $\bar{\boldsymbol{w}}$ is a standard Wiener process in reverse time. $\boldsymbol{r}$ is a NN, called the score model, which approximates a score of the diffusion process:

$$\boldsymbol{r}(k,\boldsymbol{z}_k,\boldsymbol{y},\hat{\boldsymbol{x}}_n^{CBF}) \simeq \nabla_{\boldsymbol{z}_k} p_k(\boldsymbol{z}_k|\boldsymbol{y},\hat{\boldsymbol{x}}_n^{CBF}). \tag{6}$$

ms-SGMSE initializes $\boldsymbol{z}_k$ at $k = K$ as $\boldsymbol{z}_K = \boldsymbol{y} + \delta_K\boldsymbol{u}$, where δ_k is the standard deviation of the perturbation kernel of the diffusion process and $\boldsymbol{u}$ is a sampled complex white Gaussian noise with an identity covariance matrix. Then it realizes SE by iteratively solving the reverse SDE from $k = K$ to $k = 0$. The clean speech estimate is finally obtained as $\hat{\boldsymbol{x}}_n^{DM} = \boldsymbol{z}_0$.

B. Training of score model

DiffCBF requires prior training only for the score model. Given a set of training data, each consisting of $\boldsymbol{x}_n$, $\boldsymbol{y}$, and $\hat{\boldsymbol{x}}_n^{CBF}$, we adopt the following loss to train the score model based on the diffusion model framework [2], [23]:

$$\mathcal{C}(\omega) = E_{k,(\boldsymbol{x}_n,\boldsymbol{y},\hat{\boldsymbol{x}}_n^{CBF}),\boldsymbol{u}}\left\|\boldsymbol{r}(k,\boldsymbol{z}_k,\boldsymbol{y},\hat{\boldsymbol{x}}_n^{CBF};\omega) + \frac{\boldsymbol{u}}{\delta_k}\right\|_2^2, \tag{7}$$

where ω is a set of parameters of the score model, and $E_{k,(\boldsymbol{x}_n,\boldsymbol{y},\hat{\boldsymbol{x}}_n^{CBF}),\boldsymbol{u}}$ denotes the expectation over k, $(\boldsymbol{x}_n,\boldsymbol{y},\hat{\boldsymbol{x}}_n^{CBF}) \sim p(\boldsymbol{x}_n,\boldsymbol{y},\hat{\boldsymbol{x}}_n^{CBF})$, and $\boldsymbol{u}$.

One issue in the training process is how to prepare $\hat{\boldsymbol{x}}_n^{CBF}$. In our experiments, we first obtained $\hat{\boldsymbol{x}}_n^{CBF}$ using CBF without the prior term and trained the score model. Next, we generated $\hat{\boldsymbol{x}}_n^{CBF}$ using CBF with the prior term, based on the output from ms-SGMSE in the first iteration. Finally, we retrained the score model using the outputs from both versions of CBF: one with and one without the prior term.

TABLE I
SE TASKS USED FOR EVALUATION

Task	#Sources (N)	Training-Test	#Utterances Train	Valid	Eval
Source separation	2	Matched	3569	2500	166
Single-speech enhancement	1	Mismatched	-	-	333

C. Advantages of DiffCBF

A key feature of DiffCBF is its ability to process any number 'N' ($\leq M$) of speech signals, provided that the number N is known and a pre-trained score model is available. Specifically, CBF can yield a specified number of clean speech estimates both with and without the prior term. Meanwhile, ms-SGMSE extracts each speech estimate individually from the captured signal, using the single pre-trained score model conditioned by each CBF output.

Furthermore, DiffCBF achieves highly accurate multi-channel SE by leveraging the ms-SGMSE framework, conditioned on the enhanced speech obtained from CBF, and incorporating the iterative estimation scheme in the method.

V. Experiments

We experimentally evaluated DiffCBF for SE tasks involving simultaneous denoising, dereverberation, and source separation. As a baseline, we used the single-channel source separation method, DiffSep. We also assessed DiffCBF's performance with increased iterations and when the number of speech sources differed from that in the training data.

A. SE tasks used for evaluation

We created two SE tasks for the experiments, as shown in Table I. In the source separation task, we trained and tested SE methods using a dataset containing signals with two speech sources (i.e., $N = 2$). In the single-speech enhancement task, we used a different dataset containing signals with a single speech source (i.e., $N = 1$) to test the SE methods trained for the source separation task. As a result, the single-speech enhancement task involved a mismatch in the number of speech sources between training and testing. Table I also shows the number of utterances included in the training and validation sets used for model training and the evaluation set used for testing, for each task.

We simulated each captured signal in the datasets by mixing N ($= 1$ or 2) speech signal(s) randomly taken from the Wall Street Journal (WSJ0) dataset [29] and 10 noise signals from the CHiME3 dataset [30] after convolving each signal with a Room Impulse Response (RIR). We also simulated each clean speech target using the same RIR truncated at 2 ms after the direct signal. We generated the RIRs using the image method, setting the room size to $5 \times 5 \times 2$ m. The speakers, the array, and the noise sources were randomly positioned for each utterance; the speaker-array distance was constrained between 0.5 and 1.5 m. We used three microphones ($M = 3$) to simulate each utterance. These microphones were randomly selected from a set of eight microphones, equally spaced on a linear array with a 2 cm distance between each, following

the Multiple Array-Geometry (MAG) training [21]. The reverberant speech mixture to the noise ratio and the reverberation time (T60) varied from 10 to 14 dB and 0.2 to 1.0 s. The sampling frequency was set at 8 kHz.

B. Methods to be compared and analysis conditions

We compared our proposed method, DiffCBF, with DiffSep. We also examined the output of CBF in DiffCBF.

For DiffCBF, we implemented ms-SGMSE by modifying the publicly available code for SGMSE.[1] The score model was trained as described in Section IV-B, with the number of iterations set to two. We used the Adam optimizer with a learning rate of 1.0×10^{-4} and applied exponential weight averaging. The hyperparameters were set to $\gamma = 1.5$, $c = 1.07{\times}10^{-1}$, and $a = 10.0$. We employed the short-time Fourier transform with transformed amplitudes [3]. To solve the reverse SDE, we used predictor-corrector sampling [2], with the number of diffusion steps set to 30. For CBF in DiffCBF, the hyperparameters were set to $D = 1$, $L = 10$, and $\alpha = 1$ with the analysis window length and shift set at 128 and 64 ms.

Although DiffSep is a single-channel source separation method, we adopted it as the baseline because, to our knowledge, no multi-channel SE methods based on diffusion models have been proposed that can cope with our SE task. To address our SE task, we configured DiffSep to separate three signals (two clean speech estimates and one signal containing noise and reverberation) following the configurations outlined for denoising and source separation in [20]. We used only the two speech estimates for evaluation. The publicly available code for DiffSep was used for the implementation.[2]

For both DiffCBF and DiffSep, we adopted ensemble inference [18], [19], which has been shown to improve the accuracy of diffusion model-based SE. We sampled the enhanced signals eight times with different random seeds for each method and computed their mean to obtain the final enhanced signal. Note that the order of sources estimated by DiffSep may differ over different samples. Thus, we aligned the order using the clean reference before performing the ensemble. This alignment is unnecessary for DiffCBF as each application of ms-SGMSE in DiffCBF yields only a single speech estimate.

The trainable parameter sizes of ms-SGMSE (used in DiffCBF) and DiffSep were 65.8 MB and 65.7 MB, respectively. With ensemble inference, their real-time factors (RTFs) were 4.50 and 4.07, respectively.

C. Evaluation metrics

For evaluation, we adopted improvements in the Scale-Invariant SDR from observation (SI-SDR-Imp) [31] as a signal distortion metric, the Perceptual Evaluation of Speech Quality (PESQ) [32] for overall speech quality, and Extended Short-Time Objective Intelligibility (ESTOI) [33] for speech intelligibility. We also used the overall (OVRL) metric of the Deep Noise Suppression Mean Opinion Score (DNSMOS) [34] as a non-intrusive speech quality metric.

[1]https://github.com/sp-uhh/sgmse

[2]https://github.com/fakufaku/diffusion-separation

TABLE II

EVALUATION RESULTS FOR SOURCE SEPARATION TASK, WHERE 'ITER' INDICATES ITERATION COUNTS FOR DIFFCBF. SI-SDR, PESQ, ESTOI, AND OVRL OF CAPTURED SIGNALS WERE -8.1 DB, 1.39, 0.32, AND 2.0.

	SI-SDR-Imp		PESQ		ESTOI		OVRL	
Iter	1	2	1	2	1	2	1	2
DiffSep	6.5 dB		2.06		0.42		3.0	
CBF (in DiffCBF)	5.9 dB	5.7 dB	1.71	1.74	0.53	0.54	2.4	2.4
DiffCBF (proposed)	14.6 dB	**15.1** dB	2.97	**3.04**	0.81	**0.83**	**4.0**	**4.0**

TABLE III

EVALUATION RESULTS FOR SINGLE-SPEECH ENHANCEMENT TASK, WITH A TRAINING-TEST MISMATCH IN THE NUMBER OF SPEECH SIGNALS INCLUDED IN THE CAPTURED SIGNAL. SI-SDR, PESQ, ESTOI, AND OVRL OF CAPTURED SIGNALS WERE -3.7 DB, 1.65, 0.46, AND 2.2.

	SI-SDR-Imp		PESQ		ESTOI		OVRL	
Iter	1	2	1	2	1	2	1	2
DiffSep	6.6 dB		2.80		0.60		3.7	
CBF (in DiffCBF)	3.7 dB	3.9 dB	2.04	2.10	0.65	0.68	2.7	2.7
DiffCBF (proposed)	12.6 dB	**12.7** dB	3.47	**3.51**	0.88	**0.89**	**4.1**	**4.1**

While DiffCBF generated three-channel speech estimates, DiffSep produced only a single-channel estimate. Thus, we utilized only the common channel shared between them for evaluation.

D. Evaluation results

Table II presents the evaluation results for the source separation task.[3] While all methods showed improvements across all metrics from the captured signals, DiffCBF significantly outperformed the others. The iterative estimation also enhanced DiffCBF's performance. For instance, DiffCBF achieved a substantial SI-SDR-Imps, 14.6 and 15.1 dB, at iterations 1 and 2, respectively. In contrast, DiffSep achieved a more modest SI-SDR-Imp, 6.5 dB.

Table III presents the evaluation results for the single-speech enhancement task, where we used DiffSep and DiffCBF, both trained for the source separation task. Since DiffSep produced two speech estimates for each utterance, we selected the one with the higher SI-SDR for evaluation. In contrast, DiffCBF generated a single speech estimate as desired by setting $N = 1$. While all methods improved all metrics, DiffCBF significantly outperformed the others.

These results clearly highlight the effectiveness of DiffCBF for both tasks. Its superiority over DiffSep can be attributed to its multi-channel processing capability, facilitated by the ms-SGMSE framework, along with its integration with CBF and the iterative estimation scheme. Additionally, DiffCBF proved effective even with a training-test mismatch in the number of speech signals for the single-speech enhancement task.

VI. CONCLUDING REMARKS

This paper proposed a new multi-channel SE method, DiffCBF, simultaneously performing denoising, dereverbera-

[3]Sound examples from this paper's experiments can be found at https://www.kecl.ntt.co.jp/icl/signal/nakatani/demos/eusipco2025/demo.html.

tion, and source separation. It combines a diffusion model-based method, ms-SGMSE, with a signal processing-based method, CBF, to achieve highly accurate SE. Experimental results demonstrated that DiffCBF effectively recovered clean speech from noisy and reverberant mixtures, significantly outperforming the conventional diffusion model-based source separation method, DiffSep. Additionally, DiffCBF was shown to work effectively even when the number of speech sources in the captured signals differed from that in the training data.

VII. Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 23K28113.

References

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020.

[2] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole, "Score-based generative modeling through stochastic differential equations," in *Proc. International Conference on Learning Representations (ICLR)*, 2021.

[3] Julius Richter, Simon Welker, Jean-Marie Lemercier, Bunlong Lay, and Timo Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2351–2364, 2023.

[4] Reinhold Haeb-Umbach, Tomohiro Nakatani, Marc Delcroix, Christoph Boeddeker, and Tsubasa Ochiai, "Microphone array signal processing and deep learning for speech enhancement," *IEEE Signal Processing Magazine*, vol. 41, no. 6, pp. 12–23, 2025.

[5] Xingwei Sun, Risheng Xia, Junfeng Li, and Yonghong Yan, "A deep learning based binaural speech enhancement approach with spatial cues preservation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5766–5770.

[6] Cong Han, Yi Luo, and Nima Mesgarani, "Real-time binaural speech separation with preserved spatial cues," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6404–6408.

[7] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Francois Grondin, and Mirko Bronzi, "Exploring self-attention mechanisms for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2169–2180, 2023.

[8] Nobutaka Ono and Shigeki Miyabe, "Auxiliary-function-based independent component analysis for super-Gaussian sources," in *International Conference on Latent Variable Analysis and Signal Separation*, 2010, pp. 165–172.

[9] Tomohiro Nakatani, Rintaro Ikeshita, Keisuke Kinoshita, Hiroshi Sawada, and Shoko Araki, "Blind and neural network-guided convolutional beamformer for joint denoising, dereverberation, and source separation," in *Proc. ICASSP*, 2021, pp. 6129–6133.

[10] Christoph Boeddeker, Jens Heitkaemper, Joerg Schmalenstroeer, Lukas Drude, Jahn Heymann, and Reinhold Haeb-Umbach, "Front-end processing for the CHiME-5 dinner party scenario," in *CHiME5 Workshop, Hyderabad, India*, 2018.

[11] Naoki Makishima, Shinichi Mogami, Norihiro Takamune, Daichi Kitamura, Hayato Sumino, Shinnosuke Takamichi, Hiroshi Saruwatari, and Nobutaka Ono, "Independent deeply learned matrix analysis for determined audio source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 10, pp. 1601–1615, 2019.

[12] Rongzhi Gu, Shi-Xiong Zhang, Yuexian Zou, and Dong Yu, "Towards unified all-neural beamforming for time and frequency domain speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 849–862, 2022.

[13] Zhong-Qiu Wang, Samuele Cornell, Shukjae Choi, Younglo Lee, Byeong-Yeol Kim, and Shinji Watanabe, "TF-GridNet: Integrating full- and sub-band modeling for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[14] Yen-Ju Lu, Zhong-Qiu Wang, Shinji Watanabe, Alexander Richard, Cheng Yu, and Yu Tsao, "Conditional diffusion probabilistic model for speech enhancement," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7402–7406.

[15] Joan Serrà, Santiago Pascual, Jordi Pons, R. Oguz Araz, and Davide Scaini, "Universal speech enhancement with score-based diffusion," *arXiv:2206.03065v2*, 2022.

[16] Robin Scheibler, Yusuke Fujita, Yuma Shirahata, and Tatsuya Komatsu, "Universal score-based speech enhancement with high content preservation," in *Proc. Interspeech*, 2024, pp. 1165–1169.

[17] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, "Performance measurement in blind audio source separation," *IEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[18] Naoyuki Kamo, Marc Delcroix, and Tomohiro Nakatani, "Target speech extraction with conditional diffusion model," in *Proc. Interspeech*, 2023, pp. 176–180.

[19] Hao Shi, Naoyuki Kamo, Marc Delcroix, Tomohiro Nakatani, and Shoko Araki, "Ensemble inference for diffusion model-based speech enhancement," in *Proc. Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2024.

[20] Robin Scheibler, Youna Ji, Soo-Whan Chung, Jaeuk Byun, Soyeon Choe, and Min-Seok Choi, "Diffusion-based generative speech source separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[21] Rino Kimura, Tomohiro Nakatani, Naoyuki Kamo, Delcroix Marc, Shoko Araki, Tetsuya Ueda, and Shoji Makino, "Diffusion model-based MIMO speech denoising and dereverberation," in *Proc. Hands-free Speech Communication and Microphone Array (HSCMA)*, 2024, pp. 455–459.

[22] Jean-Marie Lemercier, Julius Richter, Simon Welker, and Timo Gerkmann, "StoRM: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2724–2737, 2023.

[23] Tomohiro Nakatani, Naoyuki Kamo, Marc Delcroix, and Shoko Araki, "Multi-stream diffusion model for probabilistic integration of model-based and data-driven speech enhancement," in *Proc. International Workshop on Acoustitc Echo and Noise Control (IWAENC)*, 2024, pp. 65–69.

[24] Takuya Yoshioka and Tomohiro Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 10, pp. 2707–2720, 2012.

[25] R. Scheibler and N. Ono, "Independent vector analysis with more microphones than sources," in *Proc. WASPAA*. IEEE, 2019, pp. 185–189.

[26] R. Ikeshita, T. Nakatani, and S. Araki, "Block coordinate descent algorithms for auxiliary-function-based independent vector extraction," *IEEE Trans. Signal Processing*, vol. 69, pp. 3252–3267, 2021.

[27] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1-4, pp. 1–24, Oct. 2001.

[28] Brian David and Outram Anderson, "Reverse-time diffusion equation models," *Stochastic Processes and their Applications*, vol. 12, no. 3, pp. 313–326, 1982.

[29] John S. Garofolo, David Graff, Doug Paul, and David Pallett, "CSR-I (WSJ0) complete," https://catalog.ldc.upenn.edu/LDC93S6A.

[30] Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe, "The third CHiME speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 504–511.

[31] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey, "SDR–half-baked or well done?," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.

[32] Antony W. Rix, John G. Beerends, Michael P. Hollier, and Andries P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2001, vol. 2, pp. 749–752.

[33] Asger Heidemann Andersen, Jan Mark de Haan, Zheng-Hua Tan, and Jesper Jensen, "Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions," *Speech Communication*, vol. 102, pp. 1–13, 2018.

[34] Chandan KA Reddy, Vishak Gopal, and Ross Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6493–6497.