An Evaluation of Supervised Virtual Microphone Estimators in Reverberant Sound Fields

Kimihiro Hattori, Wen-Chin Huang, Kazuya Takeda and Tomoki Toda
Nagoya University, Japan
E-mail: hattori.kimihiro@g.sp.m.is.nagoya-u.ac.jp

***Abstract*—This paper evaluates the generalization performance of supervised virtual microphone estimation under reverberant environments. While the previous studies have shown that virtual microphone estimators trained in reverberant time conditions can operate effectively, detailed investigations on their performance have not been conducted from various perspectives. In this paper, we clarify the generalization capability of the estimator by examining both estimation performance and its impact on array signal processing under simulated environments with different reverberation times. Furthermore, we explored strategies to improve the estimator, including different model architectures and loss functions. Experimental results show that as the reverberation time increases, the estimation performance decreases and the enhancement performance by MVDR based on steering vectors degrades significantly. Furthermore, the results suggest that enhancement performance can be improved by properly designing the model architecture and loss function.**

I. Introduction

Array signal processing uses the spatial information of multi-channel audio signals to perform various tasks, such as sound source localization and sound source separation. In many cases, it is designed based on linear processing and has advantages such as structural simplicity, ease of theoretical analysis, applicability to real-time processing, and less acoustic distortion compared to non-linear processing when performing spatial operations. On the other hand, it assumes recording with the number of channels greater than or equal to the number of sound sources included in the mixed sound, and there is a constraint that the expected performance cannot be obtained under conditions where the number of channels is insufficient (i.e., underdetermined conditions). Although there have been proposed various methods that can be applied to underdetermined conditions, such as time-frequency masking for sound source separation [1], they basically involve trade-off between noise reduction performance and low distortion.

As a method to alleviate such constraints, virtual microphone signal estimator is attracting attention [2]. Virtual microphone estimator aims to improve the performance of array signal processing by estimating acoustic signals at unobserved positions and treating them as if they were physically present. Several methods for virtual microphone estimator have been proposed so far. Early methods proposed a technique that estimates virtual microphone signals at arbitrary points within the array by assuming spatial sparsity of acoustic signals in the time-frequency domain and performing non-linear interpolation for amplitude and linear interpolation for phase [2][3]. Subsequently, some methods were proposed that interpolate the amplitude components in the time-frequency domain using a convolutional neural network (CNN) conditioned on positional information [4][5]. These methods increase flexibility by learning amplitude interpolation in a data-driven manner instead of explicit probabilistic model-based interpolation. However, these time-frequency domain-based approaches estimate amplitude and phase separately, which can lead to errors due to mismatches between the two. In particular, the accuracy of phase information is directly linked to the quality of the virtual microphone signal. Therefore, if phase interpolation is not sufficiently accurate, it can be a factor limiting overall estimation performance.

In recent years, instead of the approaches that individually estimate amplitude and phase in the time-frequency domain, direct time-domain estimation using neural networks has been gaining attention. In particular, a Neural Network-based Virtual Microphone Estimator (NN-VME) has been proposed, demonstrating the direct estimation of time-domain virtual microphone waveforms through supervised learning [6]. NN-VME is known to operate in a setting where it estimates one virtual microphone signal in a two-microphone environment, and it has been reported to work effectively even in reverberant environments or under conditions where multiple speakers talk simultaneously [7]. Although its generalization performance has been confirmed by training with various reverberant conditions, its performance is also known to degrade as reverberation time increases, and the processing limits under such conditions have not been fully clarified. Besides, the effects of various reverberant time conditions in the training phase on the performance of NN-VME and the detailed properties of the estimated virtual microphone signals have not been sufficiently investigated. Furthermore, NN-VME employs the architecture of Conv-TasNet [8], known as a monaural sound separation model, but the masking process in the feature domain appears suboptimal for virtual microphone estimator. Looking ahead to applications in more challenging signal processing tasks using four or more microphones, it is important to understand the characteristics of NN-VME in more detail and clarify its general-purpose performance under various conditions.

In this paper, we investigate a virtual microphone estimator under reverberant conditions and various estimation models. In simulated sound fields, we conduct 1) signal estimation experiments with NN-VME specialized for specific sound fields, and 2) investigations on the estimation performance when the model structure and loss function are changed from the conventional

979-8-3315-7206-8/25/$31.00 ©2025 IEEE

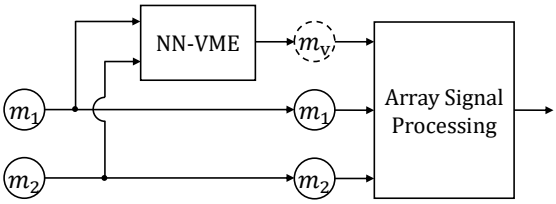

Fig. 1. Overview of array signal processing using NN-VME. m_1, m_2: input microphones, m_{v}: estimated virtual microphone

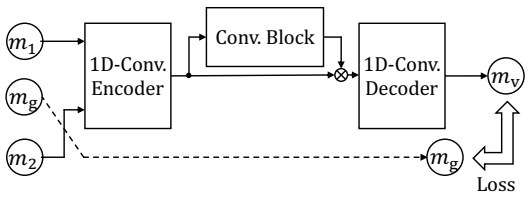

Fig. 2. Network architecture of NN-VME. m_1, m_2: input microphones, m_{v}: estimated virtual microphone m_{g}: teacher microphone

NN-VME. Furthermore, we use speech enhancement as a downstream task and evaluate its separation performance to assess the effectiveness of the generated virtual microphone signals in a practical application. From the experimental results, it was confirmed that even when NN-VME is optimized for a specific sound field, the estimation performance decreases as reverberation time is longer. It was also suggested that by appropriately setting the model structure and optimization loss, the consistency regarding the virtual microphone position can be improved. On the other hand, the degradation in estimation performance in the high-frequency band remains a challenge, and it was found that further verification is necessary for the applicability of array signal processing including virtual microphone signals.

II. Background knowledge

A. NN-VME

Fig. 1 shows an overview of NN-VME. NN-VME uses a neural network to estimate virtual microphone signals directly in the time domain. The observed signals of an array consisting of C microphones are denoted as $\mathbf{r} = \{\mathbf{r}_i \mid i = 1, ..., C\}$. Here, $\mathbf{r}_i$ indicates the observed signal at the i-th microphone. NN-VME takes this set of observed signals as input and outputs C' virtual microphone signals $\mathbf{v} = \{\mathbf{v}_{i'} \mid i' = 1, ..., C'\}$. That is, the estimation process by NN-VME is represented by (1).

$$\mathbf{v} = \text{NN-VME}(\mathbf{r}). \tag{1}$$

In the subsequent array signal processing, the observed signals $\mathbf{r}$ from the array and the estimated virtual microphone signals $\mathbf{v}$ are combined, and the extended array $\mathbf{r}' = [\mathbf{r}, \mathbf{v}]$ is used as input.

NN-VME adopts the architecture of Conv-TasNet [8], which shows high performance in monaural sound separation, as shown in Fig. 2. Specifically, it has a structure where features extracted by 1D-Convolution are subjected to masking processing based on a convolution block in the intermediate layer, and then restored to a time-domain signal by 1D-transposed convolution.

The model is optimized by supervised learning, minimizing the loss based on the Signal-to-Noise Ratio (SNR) between each virtual microphone signal output by NN-VME and the corresponding teacher signal. The loss function is defined as the sum of errors for all virtual microphone channels. Specifically, based on the SNR with the teacher signal $\mathbf{v}_{\text{true}}$, it is expressed by (2):

$$\mathcal{L} = \sum_{i'=1}^{C'} 10 \log_{10} \left(\frac{\|\mathbf{v}_{i'}^{\text{true}}\|^2}{\|\mathbf{v}_{i'}^{\text{true}} - \mathbf{v}_{i'}\|^2} \right). \tag{2}$$

B. MVDR Beamformer

In this paper, we use the Minimum Variance Distortionless Response (MVDR) beamformer [9], a speech enhancement method, as the array signal processing task to evaluate the virtual microphone signal estimator. The MVDR beamformer is a time-invariant linear filtering method that adaptively suppresses interference components while passing the target sound source components without distortion.

When performing speech enhancement on a multi-channel mixture $\mathbf{X} = \{\mathbf{x}_{f,t} \mid t \in [1, T],\ f \in [1, F]\}$ transformed into the time-frequency domain by Short-Time Fourier Transform (STFT), using a spatial filter $\mathbf{w}_f \in \mathbb{C}^{1\times M}$ at frequency f, the enhanced speech $y_{f,t} \in \mathbb{C}$ at a certain time-frequency bin is given by (3).

$$y_{f,t} = \mathbf{w}_f^{\mathrm{H}} \mathbf{x}_{f,t}. \tag{3}$$

where $(\cdot)^{\mathrm{H}}$ denotes the Hermitian transpose. There are largely two types of formulations for calculating the spatial filter $\mathbf{w}_f$ in (3) for the MVDR beamformer.

The first is a formulation based on steering vectors. The MVDR beamformer filter $\mathbf{w}_f^{\text{steering}}$ is calculated by (4).

$$\mathbf{w}_f^{\text{steering}} = \frac{(\mathbf{\Phi}_f^N)^{-1}\mathbf{a}_f}{\mathbf{a}_f^{\mathrm{H}}(\mathbf{\Phi}_f^N)^{-1}\mathbf{a}_f}. \tag{4}$$

where $\mathbf{\Phi}_f^N \in \mathbb{C}^{M\times M}$ is the spatial covariance matrix of the interference components, $\mathbf{a}_f \in \mathbb{C}^{1\times M}$ represents the steering vector or relative transfer function indicating the direction of the target sound source. This formulation requires prior information about the arrival direction of the target sound source, and the accuracy of the steering vector estimation directly affects performance.

The second is a formulation based on masking [10]. This method calculates the filter $\mathbf{w}_f^{\text{masking}}$ using the spatial covariance matrix of the target component $\mathbf{\Phi}_f^S \in \mathbb{C}^{M\times M}$ with (5).

$$\mathbf{w}_f^{\text{masking}} = \frac{(\mathbf{\Phi}_f^N)^{-1}\mathbf{\Phi}_f^S}{\mathrm{Tr}\left((\mathbf{\Phi}_f^N)^{-1}\mathbf{\Phi}_f^S\right)}\mathbf{u}. \tag{5}$$

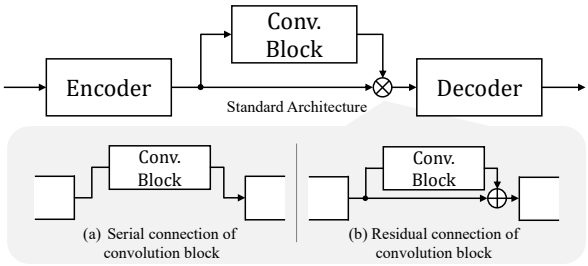

Fig. 3. Network architectures to be considered for virtual microphone estimator.

where $\mathbf{u}$ represents a one-hot vector that selects a specific microphone channel. $\mathbf{\Phi}_f^S$ and $\mathbf{\Phi}_f^N$ are estimated by masking the observed signal to extract the target sound component and the interference component. This formulation does not use steering vectors, thus having the advantage of not depending on prior information. On the other hand, since performance largely depends on the accuracy of mask estimation, performance is expected to degrade, especially in complex situations where the number of speakers increases and many speech components overlap.

III. Generalization Performance Evaluation of NN-VME

A. Performance Evaluation of NN-VME in Reverberation Conditions

In acoustic signal processing, the impact of reverberation time on the perceptibility of sound sources and the performance of various processing algorithms is significant and remains an important issue. It is essential to clarify the robustness of NN-VME against various reverberation time conditions in terms of both virtual microphone signal estimation and the downstream task. Specifically, we trained and examine NN-VME models with different reverberation time settings.

B. Examination of NN-VME Model Structure and Loss Functions

The original NN-VME adopted the Conv-TasNet as its model architecture, which is a monaural source separation model with a 1D convolution-based encoder-decoder structure that directly operates on the time domain signal with feature space mask estimation. While such a design exhibited high performance in source separation, it appears suboptimal for the task of virtual microphone signal estimation. Therefore, it is required to estimate a new signal component at an unobserved position from multiple observed microphone signals instead of masking the observed signal features to estimate.

In this section, as shown in Fig. 3, we compare the effectiveness of abolishing the masking process of Conv-TasNet by employing a structure where the intermediate network is connected serially with the feature decoder, or by introducing a residual connection.

Furthermore, we also reconsider the optimization loss function for NN-VME. Conventionally, an SNR-based loss has been used, but there is a possibility that estimation performance can be improved by introducing a loss function that considers the statistical properties of the estimated signal. In this section, we introduce the multi-resolution STFT loss [11], which is known to be effective in speech generation tasks, and verify the impact of optimization by combining it with the conventional SNR loss on the estimation results.

C. Applicability Evaluation of Virtual Microphone Signals by Beamforming

In evaluating speech enhancement by beamforming, it is important to verify the effectiveness of virtual microphone signals from the following two perspectives:

1) Whether or not the virtual microphone signal possesses accurate spatial information, i.e., the acoustic transfer function from the sound sources to the reference microphone position,
2) Whether or not the virtual microphone signal still possess useful information for array signal processing, i.e., dependent on the sound source signals but not linearly dependent on the input microphone signals even if its spatial information is not accurate.

The first point is an important indicator for judging whether the virtual microphone signal can be integrated with existing processing algorithms. The second point is directly related to the possibility of expanding the array configuration using virtual microphones. For example, even if the virtual microphone signal contains spatially independent information, if the estimation accuracy is insufficient, it possibly cause errors in filter design in highly position-dependent methods such as MVDR based on steering vectors, leading to a decrease in speech enhancement performance. In such cases, it is expected that while applicable to tasks not requiring spatial consistency, such as blind source separation, it would be unsuitable for processing that demands precise spatial estimation.

In previous studies [6][7][12][13], the masking-based MVDR has been used for evaluation, and this method has the advantage of being able to operate without depending on the positional accuracy of the virtual microphone signal. However, with this method, it is difficult to verify to what extent the virtual microphone signal is consistent with the actual physical microphone signal. Therefore, in this research, to evaluate in more detail the spatial information of the virtual microphone signal and the impact of estimation errors on subsequent processing, in addition to the masking-based MVDR beamforming, we verify the performance of steering vector-based MVDR and a linear separation filter optimized by the least-squares method using teacher signals. This filter shows the upper bound performance. By comparing the performance of these three linear filters, the masking-based MVDR, the steering vector-based MVDR, and the least-squares method separation filter, we will clarify which information is well estimated by the virtual microphone estimator under various reverberant conditions.

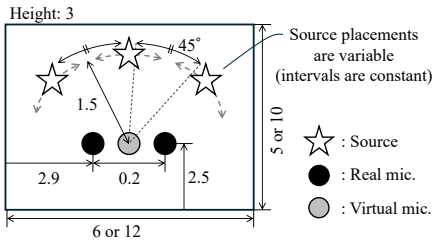

Fig. 4. Configuration of environment for acoustic simulation. (distance in meters [m])

IV. Experimental evaluation

A. Experimental Conditions

We conduct two experiments, 1) a reverberation change experiment to investigate the adaptability of NN-VME to reverberant environments and 2) a model change experiment to investigate the impact of model structure and loss functions on performance. As shown in Fig. 4, the sound source positions were set randomly for each sample, and the microphone array was fixed in a 2-microphone configuration with a 20 cm interval. Then, 3-mixture speech was generated by image source method simulation using Pyroomacoustics [14] and used for the experiments. In the reverberation change experiment, the room size was (6, 5, 3) m, and the reverberation time was varied in the range of 0.12 to 0.30 seconds. On the other hand, in the model change experiment comparisons were made under fixed conditions of room size (12, 10, 3) m and reverberation time 0.30 seconds. The sound sources used were from the clean subset of the LibriSpeech corpus [15]. As for the amount of data, 20,000, 1,000, and 3,000 samples were used for training, validation, and evaluation, respectively.

The virtual microphone was assumed to be at the center position of the two physical microphones. In the reverberation change experiment, training was performed with the same model structure and loss function as in previous studies [6][7]. In the model change experiment, in addition to the existing structure, structures where the intermediate network was connected serially (as shown in Fig. 3-(a)) or with a residual connection to the feature decoder (as shown in Fig. 3-(b)) were considered. For the loss function, in addition to SNR, multi-resolution STFT loss was introduced, and its combined effect was evaluated.

The parameters for Conv-TasNet were set as filter length $L = 16$ and number of filters $N = 512$. The internal convolution block had channel number $H = 512$, kernel size $P = 3$, and bottleneck channel number $B = 256$. Furthermore, the model was constructed with block repeat count $R = 3$ and number of blocks within each repeat $X = 8$. For the calculation of multi-resolution STFT loss, FFT sizes were set to [512, 1024, 2048], with corresponding hop sizes [50, 120, 240] and window widths [240, 600, 1200]. For the STFT loss, we considered introducing a loss whose weight increases from 0.1 to 2.0 linearly with increasing frequency. The Adam optimizer was used for training, with a learning rate of 0.00001 and 200 epochs.

TABLE I
Virtual microphone estimation performance SNR [dB] in reverberant environment. RT60 is the reverberation time of the training and evaluation data, and eval and ref indicate evaluation and reference microphones based on Fig. 2, respectively.

Evaluation mic.		RT60 [s]			
eval.	ref.	0.12	0.15	0.20	0.30
m_1	m_g	3.02	3.17	3.25	3.35
m_v	m_g	21.29	16.60	13.97	11.86
m_2	m_g	3.00	3.16	3.24	3.36

TABLE II
Speech enhancement performance SDR [dB] with virtual microphone array. Array configuration refers to the microphones used for processing based on Fig. 2.

Experimental Setting		RT60 [s]			
Method	Array configuration	0.12	0.15	0.20	0.30
MVDR (steering)	m_1-m_2	4.34	4.56	4.56	4.49
	m_1-m_v-m_2	7.96	3.88	0.88	-1.47
	m_1-m_g-m_2	45.93	32.51	24.24	18.18
MVDR (masking)	m_1-m_2	4.38	4.64	4.75	4.74
	m_1-m_v-m_2	11.43	9.95	8.59	7.40
	m_1-m_g-m_2	13.27	13.56	13.35	12.30
Least-squares	m_1-m_2	7.23	7.51	7.74	7.83
	m_1-m_v-m_2	15.05	12.62	10.87	9.59
	m_1-m_g-m_2	44.44	31.66	23.88	18.22

For array signal processing, STFT with a window width of 10240 samples and a hop size of 1/4 of that was used. To focus on the performance evaluation of the virtual microphone, all MVDR filters were calculated based on accurate information. Specifically, for steering vector-based MVDR, the room impulse response was used as the relative transfer function, and the spatial covariance matrix was obtained from the noise components. For the masking-based MVDR, the filter was calculated using accurate target sound and noise masks.

To evaluate the virtual microphone signal estimation accuracy, we calculated the SNR between the estimated virtual microphone signal and the ground truth signal. For the downstream speech enhancement performance, we calculated the signal-to-distortion ratio (SDR) [16][17] using the extended array.

B. Adaptability Evaluation of NN-VME for each Reverberation

Table I shows the estimation performance of NN-VME with respect to changes in reverberation time. For reference, the SNR calculated between the observed microphone and the teacher microphone is also shown. It can be observed that the virtual microphone estimation performance decreases as the reverberation time becomes longer. This result indicates that even when NN-VME is optimized for each acoustic condition, the estimation accuracy of the virtual microphone signal is affected by reverberation time.

Table II shows the speech enhancement performance using an array extended with virtual microphone signals. Although performance degradation is observed for all enhancement methods as reverberation time increases, the results using steering

TABLE III
ESTIMATION PERFORMANCE SNR [DB] AND ENHANCEMENT PERFORMANCE SDR [DB] OF VIRTUAL MICROPHONE ARRAY FOR EACH MODEL.

Model	STFT Loss	Weighted STFT Loss	Estimation Performance	MVDR (steering)	MVDR (masking)	Least-squares
Previous	-	-	14.86	2.84	8.83	11.19
Serial	-	-	14.99	2.94	8.78	11.08
Residual	-	-	15.08	3.01	8.79	11.09
Previous	✓	-	14.86	3.35	8.83	11.19
Residual	✓	-	15.42	3.65	8.84	11.14
Residual	-	✓	**15.64**	**4.42**	**8.97**	**11.32**

vector-based MVDR are heavily affected by virtual microphone estimation errors, and almost no enhancement performance is obtained under conditions where the reverberation time exceeds 0.15 seconds. On the other hand, the masking-based MVDR, although inferior in performance to the steering vector type with an accurate microphone array (m_1-m_{v}-m_2), is robust against estimation errors of the virtual microphone signal and shows relatively high enhancement performance in the virtual microphone array. Furthermore, linear separation based on the least-squares method shows a significant drop in enhancement performance from an accurate microphone array when using a virtual microphone array, with the masking-based MVDR performing worse than this. These results suggest that the virtual microphone signal obtained by NN-VME not only loses spatial information but also, particularly, fails to maintain spatial consistency with the relative transfer function used for filter calculation as reverberation time increases.

C. Examination of NN-VME Model Structure and Loss Functions

Table III shows the results comparing the virtual microphone estimation performance of models with modified structures and added losses to Conv-TasNet. Previous, Serial and Residual refer to the standard, serial-connection and residual-connection architectures in Fig. 3, respectively. It can be seen that the combination of structural modification and STFT loss slightly improves the estimated performance from 14.86 dB to 15.64 dB.

Looking at the speech enhancement performance, a slight improvement of about 0.15 dB is observed for the masking-based MVDR and spatial filtering based on the least-squares method. On the other hand, the steering vector-based MVDR shows a clear performance improvement from 2.84 dB to 4.42 dB due to a combination of structural changes and additional losses. In a comparison the performance of the least-squares separation filter between standard model (11.19 dB) and residual-connection with weighted STFT Loss model (11.32 dB), we can see their performance difference is relatively small. This result implies there was almost no difference in the spatial information of the virtual microphone among the methods, suggesting improved consistency between the microphone and steering vectors estimated by the combination of structural changes and additional loss model.

Fig. 5 shows the results of evaluating the performance of the virtual microphone array for each frequency band. A tendency for the virtual microphone estimation performance

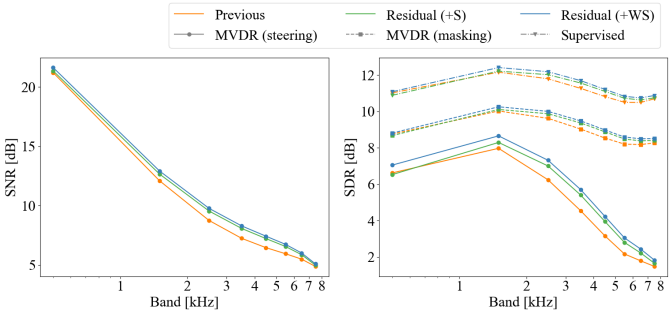

Fig. 5. Virtual microphone estimation performance per frequency band. (left: virtual microphone estimation performance, right: speech enhancement performance)

to monotonically decrease as the frequency increases is confirmed, suggesting that it is difficult for time-domain-based deep learning models to reproduce high-frequency components. This trend is also prominent in the speech enhancement performance of steering vector-based MVDR, resulting in a significant performance drop in the high-frequency band. This is thought to be because the estimation error of the virtual microphone signal in the high-frequency components impairs the phase consistency required for spatial filter design. On the other hand, for the masking-based MVDR and spatial filters based on the least-squares method, which do not depend on the consistency of positional information, almost no performance degradation due to frequency band is observed. Also, focusing on the impact of structural changes and the addition of STFT loss, the performance of steering vector-based MVDR improves across all bands, suggesting that the consistency regarding the positional information of the output signal can be improved over a wide band.

From these results, it is shown that the structure and loss function are effective in retaining positional information accuracy in virtual microphone signal estimation, but there is room for further improvement in the reconstruction of high-frequency components and consistency with existing signal processing methods.

V. CONCLUSIONS

In this research, we conducted investigations on NN-VME for the purpose of application to microphone array signal processing. First, to evaluate the adaptability of NN-VME to reverberant environments, experiments were conducted under

conditions with varying reverberation times. As a result, it was confirmed that even when NN-VME is trained and optimized for specific acoustic conditions, there is a consistent tendency for estimation performance and speech enhancement performance to decrease as the reverberation time is longer. In particular, it was found that steering vector-based MVDR beamforming is greatly affected by the estimation error of the virtual microphone signal, and practical performance cannot be obtained. On the other hand, it was found that the masking-based method can perform speech enhancement robustly against estimation errors. Next, we aimed to improve performance by reconsidering the model structure and loss function of NN-VME. Specifically, we modified the structure based on Conv-TasNet to be suitable for the construction of virtual microphone signals and added an STFT-based loss function. As a result, we were able to improve the speech enhancement performance of steering vector-based MVDR, suggesting an improvement in positional consistency. Evaluation for each frequency band also suggested that while the proposed improvements enhance overall performance, performance degradation in the high-frequency band remains a challenge. In the future, we will proceed with verifying its effectiveness through applications to advanced tasks with an increased number of virtual microphones and combinations with deep learning-based blind source separation.

Acknowledgment

Part of this work was supported by JST, CREST, JPMJCR19A3 and JST, AIP accelerated research, JPMJCR25U5.

References

[1] Ö. Yılmaz and S. T. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[2] H. Katahira, N. Ono, S. Miyabe, T. Yamada, and S. Makino, "Nonlinear speech enhancement by virtual increase of channels and maximum snr beamformer," *Eurasip Journal on Advances in Signal Processing*, vol. 2016, no. 1, pp. 1–8, 2016.

[3] K. Yamaoka, S. Makino, N. Ono, and T. Yamada, "Performance evaluation of nonlinear speech enhancement based on virtual increase of channels in reverberant environments," in *European Signal Processing Conference (EUSIPCO)*, 2017.

[4] K. Yamaoka, L. Li, N. Ono, S. Makino, and T. Yamada, "Cnn-based virtual microphone signal estimation for mpdr beamforming in underdetermined situations," in *European Signal Processing Conference (EUSIPCO)*, 2019, pp. 1–5.

[5] R. Takahashi, L. Li, S. Makino, and T. Yamada, "Vminnet: Interpolation of virtual microphones in optimal latent space explored by autoencoder," *Journal of Signal Processing*, vol. 25, no. 6, pp. 245–250, 2021.

[6] T. Ochiai, M. Delcroix, T. Nakatani, R. Ikeshita, K. Kinoshita, and S. Araki, "Neural network-based virtual microphone estimator," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6114–6118.

[7] H. Segawa, T. Ochiai, M. Delcroix, *et al.*, "Neural virtual microphone estimator: Application to multi-talker reverberant mixtures," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, pp. 293–299.

[8] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[9] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, 1972.

[10] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 260–276, 2010.

[11] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6199–6203.

[12] H. Segawa, T. Ochiai, M. Delcroix, *et al.*, "Neural network-based virtual microphone estimation with virtual microphone and beamformer-level multi-task loss," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 11 021–11 025.

[13] J. Wang and T. Toda, "Unsupervised training of neural network-based virtual microphone estimator," in *European Signal Processing Conference (EUSIPCO)*, 2024, pp. 256–260.

[14] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 351–355.

[15] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[16] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[17] R. Scheibler, "Sdr — medium rare with fast computations," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 701–705.