Neural Virtual Microphone Estimator: Application to Multi-Talker Reverberant Mixtures

Hanako Segawa*, Tsubasa Ochiai†, Marc Delcroix†, Tomohiro Nakatani†,
Rintaro Ikeshita†, Shoko Araki†, Takeshi Yamada*, Shoji Makino‡*
* University of Tsukuba, Japan † NTT corporation, Japan ‡ Waseda University, Japan

***Abstract*—The performance of array processing is limited when the number of available microphones is insufficient. Virtual microphone estimation (VME) aims at tackling this limitation by virtually increasing the number of microphones. Recently, we proposed the neural network-based VME approach (NN-VME), which uses a neural network to predict the signal at a virtual microphone position given observed microphone signals. In our previous study, we only evaluated NN-VME under a single-talker noisy acoustic condition for the supervised beamforming, but its applicability to more difficult acoustic conditions has not been fully explored. In this paper, we apply NN-VME to more challenging acoustic conditions and use different array processing approaches; 1) an underdetermined multi-talker scenario, 2) a far-field reverberant scenario, and 3) blind source separation (BSS). Experimental results demonstrate the applicability of NN-VME to 1) estimating the virtual microphones with high accuracy even when multiple speakers are simultaneously speaking, 2) working in the presence of a certain level of reverberation, although the VME accuracy decreases as the reverberation time becomes larger, and 3) enabling BSS approaches to work under the underdetermined condition where they could not be originally applied.**

I. Introduction

Array signal processing techniques using multiple microphones play an important role in the development of various applications such as noise reduction, source separation, and source localization. However, the performance of these techniques is highly dependent on the number of microphones, and the achievable performance is limited when the number of microphones is insufficient. Therefore, it is desirable to increase the number of microphones on the array to achieve high performance. On the other hand, the number of microphones that can actually be used in a device is limited due to the small size requirements of many devices, as well as structural or cost limitations.

Virtual microphone estimation (VME) [1], [2] is an approach that could compensate for the limitation in the number of available microphones. VME estimates the observed signals at locations where microphones do not actually exist (i.e., virtual microphone signals) given a few actually recorded observed signals (i.e., real microphone signals). Augmenting the array observation with the generated virtual microphone signals (i.e., virtually increasing the number of microphones) could improve the array processing performance when the number of available microphones is limited.

Early VME studies [1], [2] estimated the virtual microphone signals by linearly interpolating the phases of two real microphone signals under the assumptions of 1) plane wave propagation, 2) W-disjoint orthogonality of the sources [3], and 3) a small inter-microphone distance to avoid spatial aliasing. However, these acoustic assumptions do not always hold in real conditions (e.g., W-disjoint orthogonality does not hold under diffuse noise conditions). Recently, we proposed a neural network-based VME (NN-VME) [4], as an alternative approach that does not explicitly rely on the above assumptions. NN-VME adopts a recent time-domain neural network, that can simultaneously estimate the amplitude and phase of a virtual microphone signal as it operates directly on the time-domain signals. Assuming that during training recordings at the location of the virtual microphone are available as training targets, NN-VME learns to estimate the signals at the virtual microphone locations from the other real microphone signals in a data-driven manner based on the supervised learning framework.

In our prior work [4], we confirmed that NN-VME could effectively estimate virtual microphone signals and contribute the performance of beamforming [4] through experiments using real recordings of single-talker in diffuse noise conditions (i.e., CHiME-3 task [5]). However, these experiments were limited to single-speaker diffuse noise conditions and beamforming. NN-VME framework is not dependent on specific acoustic conditions or array signal processing techniques. Therefore, it could potentially be applied to other acoustic conditions such as the simultaneous speech of multiple speakers. In the single-speaker condition, it is sufficient to estimate the spatial information (e.g., speaker position and transfer function) for one speaker from the observed real microphone signal, but in the multi-speaker condition, it is necessary to estimate spatial information of mixed speakers, which would makes the estimation more difficult. In addition, it has been reported that, in the source separation task, the performance of the time-domain neural network could degrade in the presence of reverberation, and thus it is unclear whether NN-VME could achieve sufficient accuracy of virtual microphone estimation under reverberant conditions. In this regard, the applicability in NN-VME has not yet been fully explored.

In this paper, we verify the applicability of NN-VME framework to more challenging and diverse recording conditions. The contributions of this paper are as follows:

1) We apply NN-VME framework for underdetermined (i.e., smaller number of microphones than sources) multi-talker acoustic conditions that are more complex

than the single-talker condition. Through the experiment, we confirm that NN-VME can estimate virtual microphones with sufficient quality even under the multi-talker condition, where three speakers are simultaneously speaking, and can contribute to improving the performance of subsequent array processing.

2) We perform a comprehensive evaluation in terms of reverberation time by creating evaluation utterances whose reverberation times vary among 0, 100, 200, and 300 ms. Through the experiment, we reveal that NN-VME can estimate virtual microphones with high accuracy when the reverberation time is short (e.g., 0 ms, 100 ms), but the estimation performance decreases when the reverberation time becomes longer (e.g., 300 ms). However, even with longer reverberation, the estimated virtual microphone signal contributes to improving the beamformer's performance in terms of both speech enhancement and automatic speech recognition (ASR) metrics.
3) We experiment with NN-VME using two well-studied blind source separation (BSS) approaches (i.e., independent vector analysis (IVA) [6] and independent low-rank matrix analysis (ILRMA) [7]) in addition to the supervised beamforming approach [8]. Through this experiment, we confirm that the proposed NN-VME can be used independently of the subsequent microphone array processing approach. Moreover, we show that NN-VME enabled us to use IVA and ILRMA even under underdetermined conditions where they cannot be applied in theory.

The remainder of this paper is summarized as follows: In Section II, we first review NN-VME framework. Then in Section III, we briefly explain the array processing approaches used for the evaluation. In Sections IV and V, we describe the experiments under (underdetermined) multi-talker and reverberant acoustic conditions. We conclude the paper in Section VI.

II. Neural network-based virtual microphone estimator

A. General procedure

Figure 1 shows a basic processing flow for estimating a virtual microphone signal and applying it to array processing. In the figure, for simplicity, the network receives two input channels corresponding to the observed real microphone (i.e., channels 1 and 3), and it generates one output channel corresponding to the estimated virtual microphone (i.e., channel 2). Let $\overline{\mathbf{r}}_c \in \mathbb{R}^{\mathcal{T}}$ be the time-domain waveform of the observed signal for the c-th real microphone and $\overline{\mathbf{v}}_{c'} \in \mathbb{R}^{\mathcal{T}}$ denote the estimated signal for the c'-th virtual microphone, where $\mathcal{T}$ denotes the length of the waveform (i.e., number of samples). In the following, symbols with an over-line, such as $\overline{\mathbf{r}}_c$ and $\overline{\mathbf{v}}_{c'}$ represent time-domain signals.

NN-VME estimates the virtual microphone signals $\overline{\mathbf{v}} = \{\overline{\mathbf{v}}_{c'=1}, \ldots, \overline{\mathbf{v}}_{c'=C'}\} \in \mathbb{R}^{\mathcal{T}\times C'}$ given the real microphone signals $\overline{\mathbf{r}} = \{\overline{\mathbf{r}}_{c=1}, \ldots, \overline{\mathbf{r}}_{c=C}\} \in \mathbb{R}^{\mathcal{T}\times C}$ as input, where C denotes the number of channels for real microphones and C' denotes the number of channels for estimated virtual microphones.

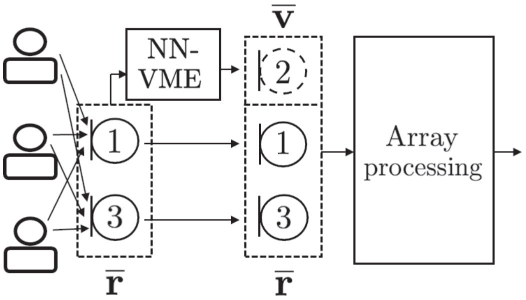

Fig. 1. Basic processing flow of array processing with NN-VME

The estimated virtual microphones are combined with the real microphones, and we obtain the augmented microphone array signal $\overline{\mathbf{x}} = [\overline{\mathbf{r}}, \overline{\mathbf{v}}] \in \mathbb{R}^{\mathcal{T}\times(C+C')}$. We can then use $\overline{\mathbf{x}}$ as the input microphone signals to subsequent microphone-array processing. By applying array processing to the augmented signal $\overline{\mathbf{x}}$, it is expected that the array processing performance could be improved compared to using only the real microphone observations $\overline{\mathbf{r}}$.

B. Network architecture

Given the real microphone signal $\overline{\mathbf{r}}$ as input, NN-VME uses a neural network to estimate the virtual microphone signal $\overline{\mathbf{v}}$ as:

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

where NN-VME$(\cdot)$ denotes the neural network model for estimating the virtual microphone.

The network architecture for NN-VME is based on the time-dilated convolutional network (TDCN) such as the fully-convolutional time-domain audio separation network (Conv-TasNet) [9], which can estimate time-domain signals with high accuracy. As shown on the left side of Fig. 2, the network is composed of a 1d-convolution encoder layer, an internal convolution block, and a 1d-deconvolution decoder layer. The encoder layer directly maps the time-domain signals to an intermediate representation, which is then further processed by the internal convolution block, and finally the decoder layer directly remaps this intermediate representation back to time-domain signals. Here, the output of the network is the estimate of the virtual microphones [4], instead of the separated signals in the original paper [9].

C. Training

NN-VME is trained based on a supervised training framework to enable the neural network model to predict the virtual microphones. In NN-VME framework, it is assumed that we have fewer constraints on the number of microphones during the system development (i.e., collection of training data) than during actual deployment. Based on this assumption, NN-VME framework prepares a set of input and target signals $\overline{\mathbf{r}}, \overline{\mathbf{t}}$ for

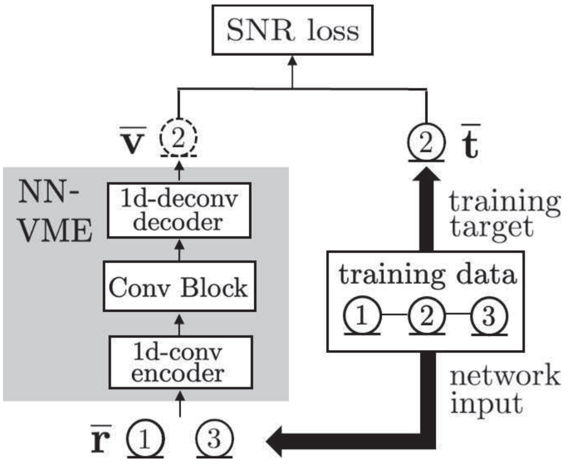

Fig. 2. Network architecture and supervised training procedure for NN-MVE

supervised training, where $\bar{\mathbf{t}} = \{\bar{\mathbf{t}}_{c'=1}, \ldots, \bar{\mathbf{t}}_{c'=C'}\} \in \mathbb{R}^{\mathcal{T}\times C'}$, and $\bar{\mathbf{t}}_{c'} \in \mathbb{R}^{\mathcal{T}}$ denote the time-domain target signal corresponding to the c'-th virtual microphone. Figure 2 illustrates the network architecture and the training procedure of NN-VME. For training, we assign a subset of the three-channel recorded microphone (i.e., channels 1 and 3) as network input $\bar{\mathbf{r}}$ and another subset (i.e., channel 2) as network target $\bar{\mathbf{t}}$.

We adopt the signal-to-noise ratio (SNR) between the target signal (real microphone signal at virtual microphone position) $\bar{\mathbf{t}}$ and estimated virtual microphone signal $\bar{\mathbf{v}} = \text{NN-VME}(\bar{\mathbf{r}})$ as the training objective:

$$\mathcal{L}_{\text{VM}} = \sum_{c'=1}^{C'} 10\log_{10}\frac{\|\bar{\mathbf{t}}_{c'}\|^2}{\|\bar{\mathbf{t}}_{c'} - \bar{\mathbf{v}}_{c'}\|^2}. \tag{2}$$

To enable the estimation of virtual microphones robust to acoustic conditions, the training dataset includes a variety of acoustic conditions, such as different source and array positions, noise conditions, and reverberation conditions.

III. Array processing approaches used for evaluation

In this paper, to investigate the applicability of NN-VME to multi-talker (underdetermined) and reverberant acoustic conditions, we evaluate two types of popular array processing approaches; 1) neural network-supported mask-based beamforming approach [8] and 2) blind source separation approach [6], [7], [10]. This section briefly overviews the two approaches.

A. Problem formulation

In this paper, we assume a situation where I source signals are recorded by M microphones. Let $s_{i,t,f} \in \mathbb{C}$ denote the short-time Fourier transform (STFT) coefficients of the i-th source at time-frequency bin (t, f) and $\mathbf{a}_{i,f} = [a_{1,i,f}, \ldots, a_{M,i,f}]^{\mathsf{T}} \in \mathbb{C}^M$ denote the time-invariant transfer function of the i-th source at frequency f. In the STFT domain, the observed signal[1] at each time-frequency bin $\mathbf{x}_{t,f} = [x_{1,t,f}, \ldots, x_{M,t,f}]^{\mathsf{T}} \in \mathbb{C}^M$ is modeled as:

$$\mathbf{x}_{t,f} = \sum_{i=1}^{I}\mathbf{a}_{i,f}s_{i,t,f} + \mathbf{n}_{t,f}, \tag{3}$$

$$= \mathbf{A}_f\mathbf{s}_{t,f} + \mathbf{n}_{t,f}, \tag{4}$$

where $\mathbf{s}_{t,f} = [s_{1,t,f}, \ldots, s_{I,t,f}]^{\mathsf{T}} \in \mathbb{C}^I$ are the source signals, $\mathbf{n}_{t,f} = [n_{1,t,f}, \ldots, n_{M,t,f}]^{\mathsf{T}} \in \mathbb{C}^M$ are the additive noise signals, and $^{\mathsf{T}}$ represents a transpose operation. $\mathbf{A}_f = [\mathbf{a}_{1,f}, \ldots, \mathbf{a}_{I,f}] \in \mathbb{C}^{M\times I}$ denotes the mixing matrix containing the transfer functions for each speaker.

In this study, we focus on the source separation problem in the underdetermined condition where the number of real microphone signals C is smaller than the number of sources I, i.e., $C < I$. By augmenting the array signals using VME, the number of microphones M is virtually increased from $M = C$ $(< I)$ to $M = C + C'$ $(\geq I)$, and we can thus use array processing methods assuming determined $(M = I)$ or overdetermined $(M > I)$ condition.

B. Mask-based beamforming

Beamformers enhance signals arriving from a specific direction. Given an observed signal $\mathbf{x}_{t,f}$, the enhanced signal $y^{\text{BF}}_{t,f} \in \mathbb{C}$ is calculated as follows:

$$y^{\text{BF}}_{t,f} = \mathbf{w}^{\mathsf{H}}_f\mathbf{x}_{t,f}, \tag{5}$$

where $\mathbf{w}_f \in \mathbb{C}^M$ denotes the beamforming filter coefficients at frequency f, and $^{\mathsf{H}}$ denotes a conjugate transpose. In this paper, we adopt the MVDR beamformer formulation of [11], and the time-invariant filter $\mathbf{w}_f$ is computed as:

$$\mathbf{w}_f = \frac{(\mathbf{\Phi}^{\text{N}}_f)^{-1}\mathbf{\Phi}^{\text{S}}_f}{\text{Tr}((\mathbf{\Phi}^{\text{N}}_f)^{-1}\mathbf{\Phi}^{\text{S}}_f)}\mathbf{u}, \tag{6}$$

where $\mathbf{\Phi}^{\text{S}}_f \in \mathbb{C}^{M\times M}$ and $\mathbf{\Phi}^{\text{N}}_f \in \mathbb{C}^{M\times M}$ denote the spatial covariance matrix for the target and noise signals, respectively. $\mathbf{u}$ denotes a one-hot vector representing the reference microphone, and $\text{Tr}(\cdot)$ denotes a matrix trace operation. The spatial covariance matrices $\mathbf{\Phi}^{S_i}_f$,$\mathbf{\Phi}^{N_i}_f$ for each speaker $i = 1, \ldots, I$ can be approximately estimated using the time-frequency mask $m_{i,t,f}$ as [8]:

$$\mathbf{\Phi}^{S_i}_f = \frac{1}{\sum_{t'=1}^{T} m_{i,t',f}}\sum_{t=1}^{T} m_{i,t,f}\mathbf{x}_{t,f}\mathbf{x}^{\mathsf{H}}_{t,f}, \tag{7}$$

$$\mathbf{\Phi}^{N_i}_f = \frac{1}{\sum_{t'=1}^{T}(1 - m_{i,t',f})}\sum_{t=1}^{T}(1 - m_{i,t,f})\mathbf{x}_{t,f}\mathbf{x}^{\mathsf{H}}_{t,f}, \tag{8}$$

where $m_{i,t,f} \in [0, 1]$ denotes the time-frequency mask, and $m_{i,t,f} = 1$ indicates that the time-frequency bin is dominated by the i-th source. We can use a generative model-based approach such as the complex Gaussian mixture model (cGMM) [12] or neural networks [8] to estimate the time-frequency

[1] The observed signal can be augmented with the virtual microphone signal generated by the VME approaches as described in Section II-A.

mask. In the experiments of Section V-B, we adopt the latter option, which we refer to as *neural network-supported beamformer (NN-BF)*.

In theory, a beamformer can produce at most $C-1$ nulls and thus cannot cancel more than $C-1$ interference signals. Therefore, it is in principle impossible to eliminate all of the interference speech in the underdetermined condition ($C<I$), and the achievable enhancement performance could be limited due to the number of available microphones. By adding virtual microphone signals to the real microphones, we expect to virtually increase the number of nulls the beamformer can generate, which may lead to better enhancement performance.

C. Blind source separation

Suppose that the number of sources is same as that of microphones ($M=I$). BSS approaches such as IVA and ILRMA estimate demixing matrix $\mathbf{A}_f^{-1} \approx \mathbf{W}_f \in \mathbb{C}^{M\times M}$, which separates a mixture of speech sources into each of its original components based only on the observed microphone signals, by assuming statistical independence between sources. Given an observed signal $\mathbf{x}_{t,f}$, separated signals $\mathbf{y}_{f,t}^{\mathrm{BSS}} = [y_{1,t,f},\dots,y_{M,t,f}]^{\mathsf{T}} \in \mathbb{C}^M$ for all speakers is calculated as:

$$\mathbf{y}_{t,f}^{\mathrm{BSS}} = \mathbf{W}_f \mathbf{x}_{t.f}. \tag{9}$$

The demixing matrix $\mathbf{W}_f$ is estimated by maximizing the log-likelihood as:

$$\mathcal{L}_{\mathrm{BSS}} = \sum_{i=1}^{M}\sum_{t=1}^{T} \log p(\tilde{\mathbf{y}}_{i,t}) + 2T\sum_{f=1}^{F}\log|\det \mathbf{W}_f|, \tag{10}$$

where $\tilde{\mathbf{y}}_{i,t} = [y_{i,t,f},\dots,y_{i,t,F}]^{\mathsf{T}} \in \mathbb{C}^F$ denotes the vector of the i-th source at time frame t, which includes the STFT coefficients for all frequencies, and F denotes the number of frequency bins. $p(\tilde{\mathbf{y}}_{i,t})$ denotes the the probability density of i-th source.

In this paper, we experiment with IVA [6], [10] and ILRMA [7]. These two approaches assume different source spectrum models $p^{\mathrm{IVA}}(\tilde{\mathbf{y}}_{i,t})$, $p^{\mathrm{ILRMA}}(\tilde{\mathbf{y}}_{i,t})$ as:

$$p^{\mathrm{IVA}}(\tilde{\mathbf{y}}_{i,t}) = \frac{1}{Z}\exp\left(-\sqrt{\sum_f |y_{i,t,f}|^2}\right), \tag{11}$$

$$p^{\mathrm{ILRMA}}(\tilde{\mathbf{y}}_{i,t}) = \prod_f \frac{1}{\pi r_{i,t,f}}\exp\left(-\frac{|y_{i,t,f}|^2}{r_{i,t,f}}\right), \tag{12}$$

where $p^{\mathrm{IVA}}(\cdot)$ is a multivariate Laplace distribution and $p^{\mathrm{ILRMA}}(\cdot)$ is a complex Gaussian distribution. Here, Z denotes a normalization constant. $r_{i,t,f}$ denotes a variance, and it is modeled by non-negative matrix factorization (NMF) [13].

IVA and ILRMA are BSS approaches applicable only to the determined or overdetermined case where the number of sources I is equal to or less than the number of microphones C ($I \leq C$). By increasing the number of microphones from $M=C$ ($<I$) to $M=I$ using VME, we expect to apply these methods even for the underdetermined conditions ($C<I$).

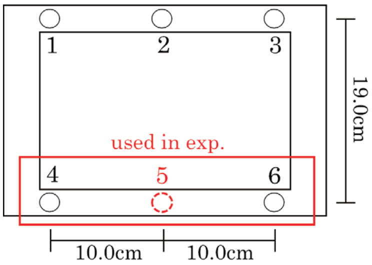

Fig. 3. Microphone array geometry for CHiME-3 corpus. Channel 5 is used as the virtual microphone (*VM*).

IV. Experimental conditions

A. Evaluated dataset

We created a dataset of simulated reverberant noisy multi-talker speech mixtures using the speech sources of the Wall Street Journal (WSJ) corpus [14] and the noise sources of the CHiME-3 corpus [5]. To create the reverberant speech sources, we randomly generated simulated room impulse responses based on the image method [15]. The reverberation time (T_{60}) was randomly selected from 0 to 300 ms for each utterance when creating the training data, and we created the four evaluation datasets with different reverberation times, i.e., fixed at 0, 100, 200, and 300 ms. The SIR for interfering speakers was randomly set to between -3 dB and 3 dB to the first speaker, and the SNR for diffuse noise was set to 20 dB. We generated 30,000, 5,000, and 5,000 mixtures for training, development, and evaluation sets, respectively. The microphone geometry (i.e., CHiME-3's tablet device [5]) is shown in Fig. 3. We used the three bottom channels of the tablet device (i.e., channels 4, 5, and 6) in the following experiments.

B. Evaluated systems

We adopted the time-domain convolutional network (TDCN)-based architecture [9] for our NN-VME. By following the notations of a previous work [9], the length of the filters was set to $L=20$, and the number of filters was set to $N=256$. The internal convolution block consisted of a stack of convolutional blocks with $H=512$ channels, kernel size $P=3$, and $B=256$ bottleneck channels. The stack was repeated $R=4$ times, and we used $X=8$ convolutional blocks per stack. We adopted the Adam algorithm [16] and gradient clipping [17] with an initial learning rate of 0.0001, and we stopped the training procedure after 200 epochs. We implemented the TDCN-based NN-VME based on the software tools provided in the GitHub repository [18].

For NN-BF, we also created a neural source separation model based on the permutation invariant training [19]. We also used a Conv-TasNet model for the separation network, and the time-frequency mask for each source was calculated as the ratio of the magnitude between the TasNet output and the input mixture [20]. The configurations, such as the network and optimization hyperparameters, were basically the same as those for NN-VME. The difference between the two networks is that NN-VME has a single output, which corresponds to the speech mixture at the virtual microphone, whereas the separation network has multiple outputs, one for each source in the mixture.

For IVA and ILRMA, we used the implementations of pyroomacoustics [21], [22]. We ran 50 iterations for IVA and 100 iterations for ILRMA to estimate the demixing matrix. The number of NMF bases for ILRMA was set to 2.

For each evaluation set (i.e., each reverberation time), the window length of STFT was set to the one with the highest SDR on the development set among 32, 64, 128, 256, and 512 ms. The shift length was set to 1/4 of the window length.

C. Evaluation criteria

As the evaluation criteria, we used the signal-to-distortion ratio (SDR) of BSSEval [23] and word error rate (WER). Given an estimated signal $\bar{\mathrm{s}}_{\mathrm{est}} \in \mathbb{R}^{\mathcal{T}}$ and a reference signal $\bar{\mathrm{s}}_{\mathrm{ref}} \in \mathbb{R}^{\mathcal{T}}$ in the time domain, the SDR of BSSEval is defined as:

$$\mathrm{SDR}(\bar{\mathrm{s}}_{\mathrm{tgt}}, \bar{\mathrm{s}}_{\mathrm{est}}) = 10 \log_{10} \frac{\|\bar{\mathrm{s}}_{\mathrm{tgt}}\|^2}{\|\bar{\mathrm{s}}_{\mathrm{tgt}} - \bar{\mathrm{s}}_{\mathrm{est}}\|^2}. \tag{13}$$

where $\bar{\mathrm{s}}_{\mathrm{tgt}}$ is computed by orthogonally projecting the estimated signal $\bar{\mathrm{s}}_{\mathrm{est}}$ onto the reference signal $\bar{\mathrm{s}}_{\mathrm{ref}}$ [23].

To evaluate the accuracy of virtual microphone estimation, we used $\mathrm{SDR}_{\mathrm{VM}} = \mathrm{SDR}(\bar{\mathrm{t}}, \bar{\mathrm{v}})$, where $\bar{\mathrm{v}}$ denotes the virtual microphone signal estimated by NN-VME and $\bar{\mathrm{t}}$ denotes the real microphone signal at the position of the virtual microphone used as the reference signal.

In addition, to evaluate the effectiveness of the virtual microphone signal when applied to array processing (AP), we used $\mathrm{SDR}_{\mathrm{AP}} = \mathrm{SDR}(\bar{\mathrm{s}}, \bar{\mathrm{y}})$, where $\bar{\mathrm{y}} \in \mathbb{R}^{\mathcal{T}}$ denotes the time-domain signal obtained by applying iSTFT to the array processed signals (i.e., Eq. (5) and (9)) and $\bar{\mathrm{s}} \in \mathbb{R}^{\mathcal{T}}$ denotes the corresponding single-talker reverberant speech (i.e., spatial image) at channel 4 in Figure 3 as a reference. The total score was computed by averaging $\mathrm{SDR}_{\mathrm{AP}}$ for the three speakers in the mixture. The permutations of each estimated and reference signal were determined by BSS_EVAL based on the signal-to-interference ratio (SIR) scores.

To evaluate the speech recognition performance (i.e., WER), we created a deep neural network-hidden Markov model (DNN-HMM) hybrid acoustic model [24], [25] based on Kaldi's WSJ recipe [26], [27]. The system was trained with lattice-free maximum mutual information criterion [28] using the noisy single-talker speech data, beamformed signals constructed with three real microphones, and beamformed signals constructed with two real and one virtual microphone. We used a trigram language model for decoding.

V. Experimental results

A. Virtual microphone-level evaluation

First, we evaluate the VME accuracy in terms of $\mathrm{SDR}_{\mathrm{VM}}$ as described in Section IV-C. Table I shows the $\mathrm{SDR}_{\mathrm{VM}}$ scores for different reverberation conditions. In the table, "RM" denotes the real microphone signal, and "VM" denotes the virtual microphone signal estimated by NN-VME. The "eval ch" column indicates the channel index of the virtual or real microphone used as the estimated signal, while the "ref ch" column indicates the channel index of the real microphone used as the reference signal when calculating $\mathrm{SDR}_{\mathrm{VM}}$. In the "VM" column, a notation such as "5 (4,6)" indicates the virtual microphone signal at channel 5 estimated from the real microphone signals at channels 4 and 6.

TABLE I
SDR [dB] (higher is better) for evaluating estimation accuracy of virtual microphone, in which observed mixture is used as reference

			T_{60}[ms]				
mictype	eval ch	ref ch	0	100	200	300	ave.
RM	4	5	2.9	3.7	2.6	2.2	2.9
RM	6	5	4.6	4.4	3.6	2.7	3.8
VM	5 (4,6)	5	17.8	15.9	12.5	9.9	14.0

TABLE II
SDR improvement [dB] (higher is better) for evaluating array processing performance, in which single-talker reverberant speech is used as reference. (baseline $\mathrm{SDR}_{\mathrm{AP}}$ of observed signals was -3.1 dB on average)

		used ch		T_{60} [ms]				
	Method	RM	VM	0	100	200	300	ave.
(1)	NN-BF	4,6	–	5.8	6.1	6.2	5.8	6.0
(2)		4,6	5 (4,6)	8.1	8.3	8.2	7.2	8.0
(3)		4,6,5	–	10.1	10.6	10.6	9.2	10.1
(4)	IVA	4,6	–	N/A	N/A	N/A	N/A	N/A
(5)		4,6	5 (4,6)	6,1	5.2	4.4	2.6	4.6
(6)		4,6,5	–	7.7	7.8	6.4	4.2	6.5
(7)	ILRMA	4,6	–	N/A	N/A	N/A	N/A	N/A
(8)		4,6	5 (4,6)	4.6	4.0	2.8	1.2	3.2
(9)		4,6,5	–	6.9	7.1	6.0	4.1	6.0

Table I shows that the virtual microphone signal (i.e., 5 (4,6)) estimated by NN-VME achieved a much higher $\mathrm{SDR}_{\mathrm{VM}}$ score compared to the adjacent real microphones (i.e., 4 or 6). This result demonstrates that, even for the simultaneous speech of multiple speakers, NN-VME framework has the potential to estimate the virtual microphone signals (i.e., spatial information of each speaker) from a few observed real microphone signals.

From the table, we also confirm that NN-VME achieves higher $\mathrm{SDR}_{\mathrm{VM}}$ when the reverberation time is short (i.e, $T_{60} = 0$ ms or $T_{60} = 100$ ms), while the VME accuracy tends to decrease as the reverberation time increases (although the $\mathrm{SDR}_{\mathrm{VM}}$ of the virtual microphone remains higher than those of the adjacent real microphones).

B. Array processing-level evaluation

We then evaluated the impact of using a VME on array processing. Table II shows the $\mathrm{SDR}_{\mathrm{AP}}$ improvement (described in Section IV-C) as a way to evaluate the source separation performances of the evaluated array processing approaches (i.e., NN-BF, IVA, and ILRMA). The "RM" and "VM" columns in "used ch" indicate the channel indices of real and

TABLE III
WER [%] (LOWER IS BETTER) FOR EVALUATED BEAMFORMERS

		used ch		T_{60} [ms]				
	Method	RM	VM	0	100	200	300	ave.
(1)	no-proc.	–	–	97.6	97.6	98.2	98.1	97.9
(2)	NN-BF	4,6	–	45.5	42.4	35.3	36.7	40.0
(3)		4,6	5 (4,6)	28.5	25.6	23.8	31.2	27.3
(4)		4,6,5	–	21.6	18.3	16.8	22.3	19.7

virtual microphones, respectively, used for running the array processing approaches. For example, the second row, "(2) NN-BF," shows results using NN-BF constructed with two real microphones (i.e., RM = 4,6) and one virtual microphone (i.e., VM = 5 (4,6)). Here, "(3) NN-BF," "(6) IVA," and "(9) ILRMA" show the results of each method when using three real microphones, which corresponds to the upper-bound performance that could be achieved by virtual microphone augmentation. "N/A" for "(4) IVA" and "(7) ILRMA" indicates that IVA and ILRMA are not applicable in underdetermined situations (i.e., $M = 2$ and $I = 3$).

Table II shows that "(2) NN-BF" with a virtual microphone achieved higher $\mathrm{SDR}_{\mathrm{AP}}$ improvement than "(1) NN-BF" with the same number of real microphones (i.e., using two real microphones). This result indicates that the estimation accuracy of the virtual microphones verified in Table I is sufficient to improve the array processing performance.

From the table, we also confirmed that "(5) IVA" and "(8) ILRMA" with virtual microphones improved $\mathrm{SDR}_{\mathrm{AP}}$ compared to the unprocessed observed mixture, especially in the less reverberant condition. This result demonstrates that NN-VME has the potential to make the BSS techniques such as IVA and ILRMA applicable for the underdetermined conditions, where they could not be originally applied.

The table shows that NN-BF achieved between 7 and 8 dB $\mathrm{SDR}_{\mathrm{AP}}$ improvement by using the virtual microphone regardless of the reverberation time, and that the performance of BSS with the virtual microphone (i.e., "(5) IVA" and "(8) ILRMA") significantly decreased, especially when reverberation time became longer (e.g., $T_{60} = 300$ ms)[2]. These results suggest that, as a future research direction, it would be worth mitigating the effect of reverberation, for example, by combining a dereverberation technique (e.g., [29]) with the NN-VME framework.

C. Speech recognition-level evaluation

Finally, we perform evaluations using NN-VME and NN-BF as a front-end for ASR. NN-BF is commonly used as a front-end for ASR systems [30]. Here, we evaluate the effectiveness of NN-VME for improving the ASR performance. Table III shows the WER scores of the evaluated NN-BF systems with and without a virtual microphone. In the table,"(1) no-proc." shows the scores for the unprocessed observed mixture. From the table, we confirm that, for the multi-talker simultaneous speaking scenario, "(3) NN-BF" with a virtual microphone successfully improves the ASR performance (i.e., WER) compared to "(2) NN-BF" with the same number of real microphones (i.e., using two real microphones).

VI. Conclusion

In this paper, we investigated the applicability of the NN-VME framework with difficult acoustic conditions (underdetermined multi-talker and far-field reverberant) and two array processing approaches (NN-BF and BSS). Our experimental results show that the NN-VME framework could estimate the virtual microphone signals in the acoustic condition where the multiple speakers are simultaneously speaking, and that it enabled blind source separation approaches (i.e., IVA and ILRMA) to work on the underdetermined condition where they could not be originally applied. Experimental results also reveal that the estimation accuracy of the virtual microphones tended to decrease as the reverberation time increased. However, even with reverberation time of more than 200 ms, the use of a virtual microphone consistently improved array-processing performance. Future work will include an investigation to mitigate the impact of reverberation, e.g., introducing dereverberation techniques as a pre-processing for the NN-VME framework.

Acknowledgment

This work was partially supported by JSPS KAKENHI Grant Number 19H04131.

[2] Note that this degradation can be attributed in part to the VME accuracy degradation under reverberant conditions as shown in Table I and the fact that IVA and ILRMA separation performances also degrade in higher reverberant conditions as shown by the performance degradation observed when using real microphones.

References

[1] 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.

[2] 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 *Proc. European Signal Processing Conference (EUSIPCO)*, 2017, pp. 2324–2328.

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

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

[5] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME'speech separation and recognition challenge: Dataset, task and baselines," in *Proc. Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2015, pp. 504–511.

[6] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 70–79, 2006.

[7] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1626–1641, 2016.

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

[9] 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.

[10] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011, pp. 189–192.

[11] M. Souden, J. Benesty, and S. Affes, "A study of the LCMV and MVDR noise reduction filters," *IEEE Transactions on Signal Processing*, vol. 58, no. 9, pp. 4925–4935, 2010.

[12] N. Ito, S. Araki, and T. Nakatani, "Complex angular central Gaussian mixture model for directional statistics in mask-based microphone array signal processing," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2016, pp. 1153–1157.

[13] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[14] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) complete," *Linguistic Data Consortium, Philadelphia*, 2007.

[15] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[16] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proc. International Conference on Learning Representations (ICLR)*, 2015.

[17] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proc. International Conference on Machine Learning (ICML)*, 2013, pp. 1310–1318.

[18] J. Wu, "funcwj/conv-tasnet," GitHub, https://github.com/funcwj/conv-tasnet (accessed Aug. 2022).

[19] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[20] T. Ochiai, M. Delcroix, R. Ikeshita, K. Kinoshita, T. Nakatani, and S. Araki, "Beam-Tasnet: Time-domain audio separation network meets frequency-domain beamformer," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6384–6388.

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

[22] R. Scheibler, I. Dokmanić, S. Barthe, E. Bezzam, and H. Pan, "LCAV/pyroomacoustics," GitHub, https://github.com/LCAV/pyroomacoustics (accessed Aug. 2022).

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

[24] H. A. Bourlard and N. Morgan, *Connectionist speech recognition: a hybrid approach*. Kluwer Academic Publishers, 1994.

[25] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[26] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *Proc. Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2011.

[27] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "kaldi-asr/kaldi," GitHub, https://github.com/kaldi-asr/kaldi/tree/master/egs/wsj/s5 (accessed Aug. 2022).

[28] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI." in *Proc. Interspeech*, 2016, pp. 2751–2755.

[29] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

[30] C. Boeddeker, H. Erdogan, T. Yoshioka, and R. Haeb-Umbach, "Exploring practical aspects of neural mask-based beamforming for far-field speech recognition," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6697–6701.