

NEURAL NETWORK-BASED VIRTUAL MICROPHONE ESTIMATION WITH VIRTUAL MICROPHONE AND BEAMFORMER-LEVEL MULTI-TASK LOSS

Hanako Segawa¹, Tsubasa Ochiai², Marc Delcroix², Tomohiro Nakatani²,
Rintaro Ikeshita², Shoko Araki², Takeshi Yamada¹, Shoji Makino^{1,3}

¹University of Tsukuba, Japan, ²NTT corporation, Japan, ³Waseda University, Japan

ABSTRACT

Array processing performance depends on the number of microphones available. Virtual microphone estimation (VME) has been proposed to increase the number of microphone signals artificially. Neural network-based VME (NN-VME) trains an NN with a VM-level loss to predict a signal at a microphone location that is available during training but not at inference. However, this training objective may not be optimal for a specific array processing back-end, such as beamforming. An alternative approach is to use a training objective considering the array-processing back-end, such as a loss on the beamformer output. This approach may generate signals optimal for beamforming but not physically grounded. To combine the advantages of both approaches, this paper proposes a multi-task loss for NN-VME that combines both VM-level and beamformer-level losses. We evaluate the proposed multi-task NN-VME on multi-talker underdetermined conditions and show that it achieves a 33.1 % relative WER improvement compared to using only real microphones and 10.8 % compared to using a prior NN-VME approach.

Index Terms— Virtual microphone estimation, array processing, multi-task learning

1. INTRODUCTION

Array signal processing [1, 2, 3] utilizing spatial information captured with multiple microphones has been actively studied for several decades. It plays a key role in developing various audio processing applications such as noise reduction, source separation, and source localization. However, the achievable performance of an array signal-processing back-end, such as beamforming (BF), relies on the number of available microphones. For example, BF with C microphones enhances (separates) a specific sound source by producing $C - 1$ nulls to reduce the interference sources. Accordingly, it cannot suppress all of the interference sources if the number of sound sources I exceeds the number of microphones C , i.e., underdetermined condition ($C < I$).

To mitigate such performance limitations due to the number of available microphones, the virtual microphone estimation (VME) approach has been studied [4, 5, 6]. VME virtually increases the number of microphones by generating virtual observations at positions where there are no real microphones (virtual microphone, VM) given a few real observations (real microphone, RM).

Earlier studies [4, 5] have estimated the VM signals by linearly interpolating the phases of two RMs while relying on physical model-based assumptions: 1) plane wave propagation, 2) W-disjoint orthogonality of the sources [7], and 3) short inter-microphone distances to avoid spatial aliasing. However, such assumptions may not always hold in realistic acoustic conditions, such as under reverberant and diffuse noise conditions.

A previous work [6] proposed a fully data-driven neural network-based VME framework (NN-VME) as an alternative VME approach that does not explicitly rely on the above assumptions. NN-VME exploits the success of recent time-domain NN to estimate the waveform of the VM directly. In other words, it can estimate both the amplitude and phase based on the supervised learning framework. NN-VME does not make physical model-based assumptions but instead assumes that we can access RM observations at VMs' locations during the training stage, which is missing during the inference stage due to structural constraints and cost restrictions. It trains the NN with a VM-level loss, which consists of minimizing the distance between the estimated VM and the RM observation at that location. Consequently, it can estimate a signal that is close to the RM signal at the VM's location and virtually increases the number of microphones by augmenting the RMs with the estimated VMs. This training scheme does not depend on a specific array signal processing back-end, such as BF. Consequently, the VM signals generated by NN-VME could be applied to arbitrary array processing back-ends [6, 8].

Training an NN-VME with a VM-level loss offers versatility, but it may not be optimal for a specific array processing back-end. If the array processing back-end is determined in advance, we could estimate VM signals better suited for that back-end by adopting a training loss on the output of the array processing back-end. In this case, however, the estimated VM signal may not be interpretable as a virtually recorded observation signal. For example, when using a BF-level training loss, the NN-VME may learn an extreme behavior, such as estimating only the target source signals, because the BF-level loss could still be improved even if the estimated VM signal is close to the target source signal. Arguably, a signal trained with such an extreme tendency cannot be called a VM.

In this paper, we focus on the popular BF approach as the array processing back-end and consider a multi-talker scenario. We extend the NN-VME framework to adopt the BF-level loss by additionally assuming the availability of reference single-talker sources. To make the NN-VME work for arbitrary positions of the target and interference sources, we propose combining the NN-VME with a mask-based frequency-domain BF [9, 10] and the permutation invariant training (PIT) [11] schemes. Moreover, we propose a novel multi-task training objective for the NN-VME. It combines both VM-level and BF-level losses to take advantage of both training objectives by assuming the availability of both reference RM observations at the locations of the VMs and reference single-talker sources as the training targets.

We evaluate the effectiveness of the proposed multi-task NN-VME on the underdetermined multi-talker and reverberant acoustic conditions using three criteria: 1) estimation accuracy of the VM signal, 2) estimation accuracy of the beamformed signal augmented

with the VME, and 3) speech recognition accuracy of the beamformed signal. The experiments confirm that the NN-VME trained with the proposed multi-task loss successfully takes advantage of both loss functions and achieves higher performance than the NN-VMEs trained with the single-task losses.

2. RELATED WORK

Prior to this work, Yamaoka et al. [5] proposed the adoption of a BF-level training objective for the physical model-based VME framework, where the VM's amplitude is estimated using neural networks. However, their experimental validation was relatively limited. For example, their experiments assume that the direction of the target source is known (i.e., the oracle steering vector of the target source is available) and that the positions of the target source and interference sources are fixed for all of the experiments. Moreover, their experiments showed that such a BF-level training objective results in a large performance improvement for a closed (training) dataset but only a small improvement for an open (evaluation) dataset due to the severe nonlinear noise (artifact) in the estimated VM, probably as a result of the over-fitting [5]. Therefore, the previous study [5] has not fully revealed the effectiveness of the BF-level training objective for the VME framework in a practical use case, i.e., the direction of the target source is unknown and variable.

This paper incorporates a BF-level training objective in a recent fully data-driven NN-VME framework [6, 8] by combining 1) the mask-based frequency-domain BF [9, 10] that does not use any prior information of the sources and 2) the PIT-based multi-talker reconstruction loss [11]. We experimentally confirmed that the proposed NN-VME with the BF-level loss successfully improves BF performance for an open evaluation dataset that contains variable source positions.

3. PROPOSED METHOD: MULTI-TASK TRAINING LOSS FOR NN-VME

In this paper, we assume a situation where C microphones record I source signals. Let $\mathbf{s}_i \in \mathbb{R}^T$ denote a time-domain source signal of the i -th source, where T is the length of the waveform (i.e., number of samples). The observed signal $\mathbf{y}_c \in \mathbb{R}^T$ is modeled as $\mathbf{y}_c = \sum_{i=1}^I \mathbf{x}_{c,i} + \mathbf{n}_c$, where $\mathbf{x}_{c,i}$ denotes a reverberant source signal (i.e., spatial image) of the i -th source recorded at the c -th channel, and $\mathbf{n}_c \in \mathbb{R}^T$ is the additive noise signal.

3.1. General procedure of NN-VME

The VME framework consists of two steps: 1) estimating the VM and 2) applying the array signal processing technique.

In the NN-VME [6], a time-domain signal estimation neural network [12] is used to estimate the amplitude and phase of the VM signal simultaneously. Let $\mathbf{r}_c \in \mathbb{R}^T$ denote the time-domain observed signal recorded by the c -th RM, and $\hat{\mathbf{v}}_{c'} \in \mathbb{R}^T$ denote the estimated signal corresponding to the c' -th VM. Given the RM observation $\mathbf{r} = \{\mathbf{r}_{c=1}, \dots, \mathbf{r}_{c=C_r}\} \in \mathbb{R}^{T \times C_r}$, the VM observation $\hat{\mathbf{v}} = \{\hat{\mathbf{v}}_{c'=1}, \dots, \hat{\mathbf{v}}_{c'=C_v}\} \in \mathbb{R}^{T \times C_v}$ is estimated as:

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

where $\text{NN-VME}(\cdot)$ denotes the neural network model, and C_r and C_v denote the number of RMs used as input and the number of VMs to be estimated, respectively.

When using the VM with a microphone array processing back-end, the estimated VM signal is combined with the RM signal to

obtain an augmented microphone array signal $\bar{\mathbf{y}} = [\mathbf{r}, \hat{\mathbf{v}}] \in \mathbb{R}^{T \times C}$, where $C = C_r + C_v$. We expect the array processing performance will improve when using augmented microphone array signal $\bar{\mathbf{y}}$, whose number of microphones is virtually increased, compared to using only the real array observation \mathbf{r} .

This paper focuses on the source separation task and adopts a mask-based frequency-domain BF [9, 10] as the array processing back-end. Given the augmented array observation $\bar{\mathbf{y}}$, the BF computes the enhanced signal $\hat{\mathbf{x}}_i^{\text{BF}} \in \mathbb{R}^T$ of the i -th source as:

$$\hat{\mathbf{x}}_i^{\text{BF}} = \text{MaskBF}_i(\bar{\mathbf{y}}), \quad (2)$$

where $\text{MaskBF}_i(\cdot)$ denotes the functional representation of the mask-based BF for the i -th source. Specifically, in this paper, we adopt the formulation of the Minimum Variance Distortionless Response (MVDR) BF of [13]. We also follow prior studies [6, 8] in implementing mask-based BF with NN-VME.

3.2. Loss function of NN-VME

In this section, we first overview two levels of loss functions (i.e., VM-level and BF-level) and then introduce the proposed multi-task loss function. Note that the training objective determines the property of the estimated VM signals. For example, when using the VM-level loss [6], the VM signal mimics the RM observation that would be captured at the position of the VM. Here, the RM observations at the location of the VM are assumed to be known in the training stage but missing in the inference stage. When using the BF-level loss [5], the VM signal becomes a signal that improves the array processing performance when combined with the RM observations.

3.2.1. Virtual microphone (VM)-level training loss

The original NN-VME [6] adopts the VM-level training loss that brings the estimated VM signals $\hat{\mathbf{v}}$ close to the target signals \mathbf{v} , i.e., RM observations at the locations of the VMs.

The NN-VME framework assumes that we have fewer constraints on the number of microphones during system development (i.e., training stage) than during actual deployment (i.e., inference stage). It divides the observed signal $\mathbf{y} = \{\mathbf{y}_{c=1}, \dots, \mathbf{y}_{c=C}\} \in \mathbb{R}^{T \times C}$ into two subsets: one for network input (i.e., $\mathbf{r} \in \mathbb{R}^{T \times C_r}$) and the other for network target (i.e., $\mathbf{v} \in \mathbb{R}^{T \times C_v}$). It assumes that a set of input and target signals $\{\mathbf{r}, \mathbf{v}\}$ is available for training the model, while only \mathbf{r} is available for the inference stage, i.e., \mathbf{v} is missing due to structural constraints or cost restrictions. Given the set of input and target signals $\{\mathbf{r}, \mathbf{v}\}$, the VM-level loss is defined based on the classical signal-to-noise ratio (SNR) loss as:

$$\mathcal{L}_{\text{VM}} = \text{SNR}(\mathbf{v}, \hat{\mathbf{v}}), \quad (3)$$

where $\hat{\mathbf{v}}$ is the output of the NN-VME(\cdot) module, as in Eq. (1). Here, we adopt the widely used classical SNR [14] as a loss metric. Given the time-domain reference signal $\mathbf{z}_{\text{ref}} \in \mathbb{R}^T$ and estimated signal $\mathbf{z}_{\text{est}} \in \mathbb{R}^T$, the SNR loss is computed as $\text{SNR}(\mathbf{z}_{\text{ref}}, \mathbf{z}_{\text{est}}) = -10 \log_{10}(\|\mathbf{z}_{\text{ref}}\|^2 / \|\mathbf{z}_{\text{ref}} - \mathbf{z}_{\text{est}}\|^2)$.

3.2.2. Beamformer (BF)-level training loss

In addition to the above VM-level loss function, we can consider the BF-level training loss that makes the estimated beamformed signals $\hat{\mathbf{x}}_i^{\text{BF}}$ close to the target signals $\mathbf{x}_{c_{\text{ref}},i}$, i.e., the single-talker reverberant source.

In the supervised source separation, we assume that a set of input and target signals $\{\mathbf{r}, \mathbf{x}\}$ is available for training the model, where

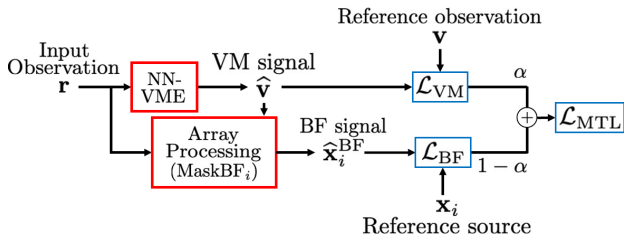


Fig. 1: Multi-task learning with VM-level and BF-level losses.

$\mathbf{x} = [\mathbf{x}_{c_{\text{ref}}, i=1}, \dots, \mathbf{x}_{c_{\text{ref}}, i=I}] \in \mathbb{R}^{T \times I}$ denotes the single-talker reverberant signals for each source at the reference channel c_{ref} . Given the set of input and target signals $\{\mathbf{r}, \mathbf{x}\}$, the BF-level loss is defined based on the classical SNR loss as:

$$\mathcal{L}_{\text{BF}} = \min_{p \in \text{perm}(I)} \sum_{i=1}^I \text{SNR}(\mathbf{x}_{c_{\text{ref}}, i}, \hat{\mathbf{x}}_{p_i}^{\text{BF}}), \quad (4)$$

where $\hat{\mathbf{x}}_{p_i}^{\text{BF}}$ is the output of the $\text{MaskBF}_i(\cdot)$ module constructed from the augmented array observation $\bar{\mathbf{y}}$ with the NN-VME, as in Eq. (2). $\text{perm}(I)$ produces all possible permutations, and $p: \{1, \dots, I\} \rightarrow \{1, \dots, I\}$ is a permutation that maps i to $p_i \in \{1, \dots, I\}$. Here, we adopt the PIT scheme [11] to handle the multiple sources. We can expect that adopting the reconstruction loss for all sources would impose the constraint that the estimated VM signals maintain the spectral and spatial information of all sources, unlike [5].

3.2.3. Proposed multi-task training loss combining VM-level and BF-level losses

Figure 1 shows a schematic diagram of the proposed multi-task learning scheme. To take advantage of the two different levels of training objectives, we introduce a multi-task loss function \mathcal{L}_{MTL} , which combines the VM-level \mathcal{L}_{VM} and BF-level \mathcal{L}_{BF} loss functions:

$$\mathcal{L}_{\text{MTL}} = \alpha \mathcal{L}_{\text{VM}} + (1 - \alpha) \mathcal{L}_{\text{BF}}, \quad (5)$$

where $0 \leq \alpha \leq 1$ represents the interpolation weight hyperparameter that controls the trade-off between VM-level and BF-level losses.

By combining the VM-level and BF-level losses, we expect the multi-task loss to suppress the over-fitting issues reported when using only BF-level loss [5] and to improve the source separation performance compared to using only the VM-level loss. Moreover, it could maintain the properties of the VM, i.e., obtaining an estimated signal close to the signal that would be captured by an RM at that position. Finally, we may also benefit from improved generalization thanks to the multi-task learning effect [15].

4. EXPERIMENT

4.1. Experimental conditions

In this experiment, all of the training and evaluation data consisted of simulated reverberant noisy three-speaker mixtures using speech from the Wall Street Journal (WSJ) corpus [16] and noise from the CHiME-3 corpus [17]. The room impulse responses were generated using the image method [18]. The reverberation time (T_{60}) was randomly selected from 0 ms to 300 ms for both training and evaluation data. For each mixture, we randomly sampled the position of the speakers and the microphone array. We simulated square rooms with width and depth randomly set to 2.5 ~ 10 m, and the height set to 2.5

~ 5 m. The signal-to-interference ratio (SIR) for interfering speakers was randomly set within the range of -3 dB ~ 3 dB with respect to the first speaker, and the signal-to-noise ratio (SNR) of the diffuse noise was set to 20 dB. We generated 30,000, 5,000, and 5,000 mixtures for the training, development, and evaluation sets, respectively. The microphone geometry is a rectangular microphone array with six channels corresponding to the CHiME-3's tablet device [17] (refer to the figure in [17] for details). In the following experiments, we used the three bottom channels, i.e., channels 4, 5, and 6, linearly arranged at 10 cm intervals, and assumed that channels 4 and 6 are the RMs and channel 5 is the VM.

4.2. Evaluation systems

The network architecture of the NN-VME was based on the time-domain convolutional network (TDCN) [12] as in our prior work [6, 8]. According to the notation of a previous study [12], the hyperparameters are set to $N = 256$, $L = 20$, $B = 256$, $H = 512$, $P = 3$, $X = 8$, and $R = 4$. For the optimization, we used the Adam algorithm [19] with an initial learning rate of 0.0001 and a batch size of 8. The duration of each training sample is 4 seconds. We stopped the training procedure after 100 epochs.

We also prepared a TDCN-based source separation model (i.e., convolutional time-domain audio separation network (Conv-TasNet) [12]) based on the PIT scheme [11] to estimate the time-frequency masks for each source required to construct the mask-based BF [9, 10]. Following a previous study [20], the time-frequency mask for each source is computed by applying STFT to the time-domain observed signal and the separated signal of each source and then taking the ratio of the magnitudes between them.

The network and optimization configurations of the separation model were basically the same as those of NN-VME. The difference is that NN-VME has a single output and estimates the speech mixture at the locations of the VM, while the separation model has multiple outputs and estimates the clean speech for each source.

In this paper, we separately trained the parameters of NN-VME and the source separation model. We first trained the source separation model and then optimized only the parameters of the NN-VME while constructing the BF with the estimated time-frequency masks.

4.3. Evaluation metrics

To evaluate the accuracy of the estimated VMs and their effectiveness for the array processing, we adopted two types of signal-to-distortion ratio (SDR) and word error rate (WER) by following a previous study [6]. Given an estimated signal $\mathbf{z}_{\text{est}} \in \mathbb{R}^T$ and a reference signal $\mathbf{z}_{\text{ref}} \in \mathbb{R}^T$, the SDR is defined as $\text{SDR}(\mathbf{z}_{\text{ref}}, \mathbf{z}_{\text{est}}) = 10 \log_{10}(\|\mathbf{z}_{\text{tgt}}\|^2 / \|\mathbf{z}_{\text{tgt}} - \mathbf{z}_{\text{est}}\|^2)$, where \mathbf{z}_{tgt} is computed by orthogonally projecting the estimated signal \mathbf{z}_{est} onto the reference signal \mathbf{z}_{ref} [21].

First, we evaluated the accuracy of the estimated VM signal using $\text{SDR}_{\text{VM}} = \text{SDR}(\mathbf{v}, \hat{\mathbf{v}})$, where $\mathbf{v} \in \mathbb{R}^T$ denotes the RM observation at the position of the VM as the reference signal, and $\hat{\mathbf{v}} \in \mathbb{R}^T$ denotes the estimated VM observation.

In addition, we measured the source separation performance of the beamformed signals using $\text{SDR}_{\text{BF}} = \text{SDR}(\mathbf{x}, \hat{\mathbf{x}}^{\text{BF}})$, where $\mathbf{x} \in \mathbb{R}^T$ denotes the reference single-talker reverberant signal (i.e., spatial image), and $\hat{\mathbf{x}}^{\text{BF}} \in \mathbb{R}^T$ denotes the beamformed signal. In the evaluation, we used the fourth channel as the reference microphone. The SDR_{BF} was computed for each of the three speakers in the mixture and then averaged to obtain the total score, where the permutations between the estimates and references were determined based on the SIR score.

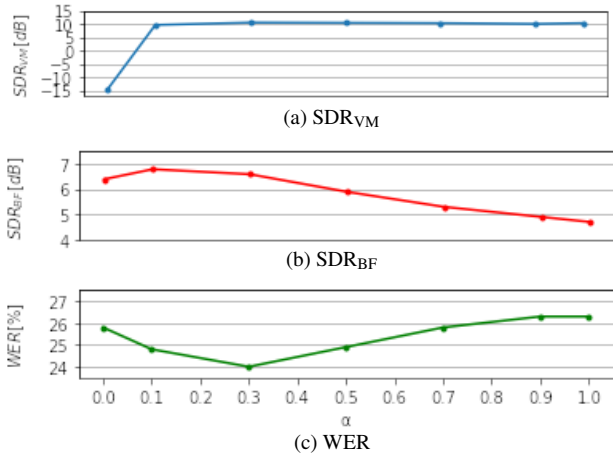


Fig. 2: Impact of multi-task weight α on SDR_{VM} , SDR_{BF} [dB] (higher is better) and WER [%] (lower is better).

Finally, to evaluate the automatic speech recognition (ASR) performance of the beamformed signals, we built a deep neural network-hidden Markov model (DNN-HMM) hybrid ASR system [22, 23] based on the Kaldi’s CHiME-4 recipe [24]. The acoustic model was trained with a lattice-free maximum mutual information framework [25]. As training data, we used 1) noisy single-talker signals, 2) beamformed signals using three RMs, and 3) beamformed signals using two RMs and one VM estimated by the NN-VME ($\alpha = 1.0$). We used a trigram language model for decoding.

4.4. Evaluation of multi-task interpolation weight

First, we explore the impact of the multi-task interpolation weight α in Eq. (5) on the evaluation measures. Figure 2 shows SDR_{VM} , SDR_{BF} , and WER scores for the evaluated BF systems, where the x-axis indicates the multi-task interpolation weight α and the y-axis indicates each of the evaluation measures. We varied the weight hyperparameter in the ranges of $\alpha = \{0.0, 0.1, 0.3, 0.5, 0.7, 0.9, 1.0\}$. Here, $\alpha = 1.0$ and $\alpha = 0.0$ correspond to the conventional single-task loss function; $\alpha = 1.0$ is equivalent to using only the VM-level loss \mathcal{L}_{VM} as in Eq. (3), while $\alpha = 0.0$ is equivalent to using only the BF-level loss \mathcal{L}_{BF} as in Eq. (4).

From the figure, we confirm that using the BF-level loss ($\alpha = 0.0$) outperforms using only the VM-level loss ($\alpha = 1.0$) in terms of SDR_{BF} and WER. However, the VM signal is not related to the RM as shown by the low value of SDR_{VM} for $\alpha = 0.0$. By appropriately tuning α , the proposed NN-VME with multi-task loss (e.g., $\alpha = 0.3$) outperforms the single-task models, particularly in terms of WER.

4.5. Overall evaluation

Table 1 summarizes the signal-level and ASR-level performance measures for all the evaluated BFs. Here, VM-BF denotes the BF constructed with two RMs (channels 4 and 6) and one VM (channel 5). System (1) indicates the performance of the observed signal without processing. Systems (2) and (3) denote the BF constructed with two (channels 4 and 6) and three (channels 4, 5, and 6) RMs, which would correspond to the lower-bound and upper-bound performances of the VM-BF. Note that we are dealing with challenging conditions, i.e., the separation of noisy and reverberant three-speaker mixtures using a limited amount of microphones, which is reflected by the relatively low performance of the baseline system (3). Systems (4) and (5) correspond to NN-VME trained with the BF-level

Table 1: SDR_{VM} , SDR_{BF} [dB] (higher is better) and WER [%] (lower is better) for evaluated beamforming systems.

	Method	α	SDR_{VM}	SDR_{BF}	WER
(1)	Mixture	-	-	-3.1	98.4
(2)	RM-BF (2ch)	-	-	3.0	35.9
(3)	RM-BF (3ch)	-	-	7.1	18.8
(4)	VM-BF (\mathcal{L}_{BF})	0.0	-14.3	6.4	25.8
(5)	VM-BF (\mathcal{L}_{VM})	1.0	10.5	4.7	26.9
(6)	VM-BF (\mathcal{L}_{MTL})	0.1	9.8	6.8	24.8
(7)	VM-BF (\mathcal{L}_{MTL})	0.3	10.7	6.6	24.0

and VM-level losses, respectively. Systems (6) and (7) correspond to NN-VME trained with the proposed multi-task loss with two different weights.

The performance of the baseline system (2) is relatively low due to the underdetermined condition ($C < I$). Using the VM, we can virtually increase the number of microphones and make the systems determined, which explains the significant boost in the performance of systems (4) to (7).

Note that unlike the previous study [5], our proposed NN-VME using PIT-based BF-level loss improves performance even for unseen conditions. Moreover, the VM-BF trained with BF-level loss (i.e., system (5)) achieved better SDR_{BF} and WER scores compared to the VM-BF trained with VM-level loss (i.e., system (4)), but the score of SDR_{VM} becomes very low. This can be expected because using only the BF-level loss \mathcal{L}_{BF} ($\alpha = 0.0$) does not directly specify the property of the estimated VM signals, and thus there is no guarantee that the output of the NN-VME module imitates the observed RM signals that are actually recorded at the specific microphone position (i.e., channel 5 in this experiment).

On the other hand, systems (6) and (7) trained with the proposed multi-task loss retain high SDR_{VM} scores. In addition, probably due to the effect of the multi-task learning scheme [15], they also achieved better SDR_{BF} and WER scores than systems (4) and (5) trained with the single-task losses. Here, the WER of system (7) with multi-task loss is significantly better than that of system (4) with single-task BF-level loss. This is probably because the VM signal generated only with the BF-level loss becomes quite artificial and may result in containing more nonlinear distortions in the beamformed signals, which is reported to have a negative effect on ASR performance [26, 27].

These results show the effectiveness of optimizing the NN-VME while considering the specific array processing back-end. Furthermore, they also show the effectiveness of using the multi-task training objective considering both the VM-level and BF-level losses, which enables the generation of interpretable VM signals and leads to better array processing performance.

5. CONCLUSIONS

This paper proposed a novel multi-task learning scheme for the NN-VME framework that combines the VM-level and BF-level losses. We evaluated the effectiveness of the proposed method on the underdetermined source separation task with mask-based BF. The experimental results show that the NN-VME with the proposed multi-task training loss achieved better source separation and speech recognition performances than the NN-VME with a single-task training loss, i.e., using only the VM-level loss or the BF-level loss. Future work will include the evaluation of the effectiveness of the proposed method for other array processing techniques, e.g., other types of beamforming, blind source separation, and source localization.

6. REFERENCES

- [1] Jacob Benesty, Jingdong Chen, and Yiteng Huang, *Microphone array signal processing*, vol. 1, 2008.
- [2] Barry D Van Veen and Kevin M Buckley, “Beamforming: A versatile approach to spatial filtering,” *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988.
- [3] Michael Brandstein and Darren Ward, *Microphone arrays: Signal processing techniques and applications*, 2001.
- [4] Hiroki Katahira, Nobutaka Ono, Shigeki Miyabe, Takeshi Yamada, and Shoji 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.
- [5] Kouei Yamaoka, Li Li, Nobutaka Ono, Shoji Makino, and Takeshi Yamada, “CNN-based virtual microphone signal estimation for MPDR beamforming in underdetermined situations,” in *European Signal Processing Conference (EUSIPCO)*, 2019, pp. 1–5.
- [6] Tsubasa Ochiai, Marc Delcroix, Tomohiro Nakatani, Rintaro Ikeshita, Keisuke Kinoshita, and Shoko Araki, “Neural network-based virtual microphone estimator,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6114–6118.
- [7] Ozgur Yilmaz and Scott Rickard, “Blind separation of speech mixtures via time–frequency masking,” *IEEE Transactions on signal processing*, vol. 52, no. 7, pp. 1830–1847, 2004.
- [8] Hanako Segawa, Tsubasa Ochiai, Marc Delcroix, Tomohiro Nakatani, Rintaro Ikeshita, Shoko Araki, Takeshi Yamada, and Shoji Makino, “Neural virtual microphone estimator: Application to multi-talker reverberant mixtures,” in *Asia-Pacific Signal and Information Processing Association (APSIPA)*, 2022, pp. 293–299.
- [9] Jahn Heymann, Lukas Drude, and Reinhold Haeb-Umbach, “Neural network based spectral mask estimation for acoustic beamforming,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 196–200.
- [10] Hakan Erdogan, John R Hershey, Shinji Watanabe, Michael I Mandel, and Jonathan Le Roux, “Improved MVDR beamforming using single-channel mask prediction networks,” in *Interspeech*, 2016, pp. 1981–1985.
- [11] Morten Kolbæk, Dong Yu, Zheng-Hua Tan, and Jesper 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.
- [12] Yi Luo and Nima 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.
- [13] Mehrez Souden, Jacob Benesty, and Sofiene Affes, “On optimal frequency-domain multichannel linear filtering for noise reduction,” *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 18, no. 2, pp. 260–276, 2009.
- [14] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey, “SDR–half-baked or well done?,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.
- [15] Rich Caruana, *Multitask learning*, 1998.
- [16] Douglas B Paul and Janet Baker, “The design for the Wall Street Journal-based CSR corpus,” in *International Workshop on Speech and Natural Language*, 1992, pp. 357–362.
- [17] Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe, “The third ‘CHiME’ speech separation and recognition challenge: Dataset, task and baselines,” in *Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2015, pp. 504–511.
- [18] Jont B Allen and David 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.
- [19] Diederik P Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” in *International Conference on Learning Representations (ICLR)*, 2015.
- [20] Tsubasa Ochiai, Marc Delcroix, Rintaro Ikeshita, Keisuke Kinoshita, Tomohiro Nakatani, and Shoko Araki, “Beam-TasNet: Time-domain audio separation network meets frequency-domain beamformer,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6384–6388.
- [21] Emmanuel Vincent, Rémi Gribonval, and Cédric 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.
- [22] Herve A Bouchard and Nelson Morgan, *Connectionist speech recognition: A hybrid approach*, 1994.
- [23] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdelrahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara 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.
- [24] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, “The Kaldi speech recognition toolkit,” in *Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2011.
- [25] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur, “Purely sequence-trained neural networks for ASR based on lattice-free MMI,” in *Interspeech*, 2016, pp. 2751–2755.
- [26] Kazuma Iwamoto, Tsubasa Ochiai, Marc Delcroix, Rintaro Ikeshita, Hiroshi Sato, Shoko Araki, and Shigeru Katagiri, “How bad are artifacts?: Analyzing the impact of speech enhancement errors on ASR,” in *Interspeech*, 2022, pp. 5418–5422.
- [27] Cătălin Zorilă and Rama Doddipatla, “Speaker reinforcement using target source extraction for robust automatic speech recognition,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6297–6301.