Differentiable Max-Directivity Beamforming Normalization for Independent Vector Analysis

Shoichiro Takeda
Media Intelligence Lab.
NTT Corporation
Kanagawa, Japan
shoichiro.takeda.us@hco.ntt.co.jp

Kenta Niwa
Media Intelligence Lab.
NTT Corporation
Tokyo, Japan
kenta.niwa.bk@hco.ntt.co.jp

Shinya Shimizu
Media Intelligence Lab.
NTT Corporation
Kanagawa, Japan
shinya.shimizu.te@hco.ntt.co.jp

***Abstract*—Independent vector analysis (IVA) minimizes an objective function to estimate separation filters that separate mixture signals into individual source signals. Unfortunately, IVA often suffers from the well-known block permutation problem. To mitigate that problem, the use of geometry knowledge has been studied, but two crucial issues remain: the necessity of non-differential processes outside the minimization and of high-level geometrical clues such as the directions of arrival (DOAs) of the source signals. This paper thus presents a novel IVA method whose objective function has a differentiable max-directivity beamforming normalization (MDBN) term. This term uses geometry knowledge from only a low-level geometrical clue (the positions of the microphone array) via the traditional beamforming (BF) concept that each separation filter should have a maximum gain for each specific DOA across all frequency bins. Thus, our overall objective function can be minimized by gradient descent, and the MDBN term encourages the separation filters to focus on specific directions, which implicitly estimates the most reasonable DOAs of the source signals at each iteration. Therefore, our method uses geometry knowledge while avoiding the above two issues and estimates good separation filters mitigating the permutation problem. Several experiments show that our method outperforms the conventional BF and IVA methods.**

***Index Terms*—independent vector analysis, geometry knowledge, beamforming, maximum a posterior, normalization term, chain rule**

I. Introduction

Services that use speech signals, such as teleconference services, hands-free car communication services, and AI smart speakers, are widely spread. In real environment, however, surrounding noises and diffuse reverberation are often mixed with source signals. Therefore, such services require pre-processing to separate the resulting mixture signals into the target individual source signals [1], [2].

For this purpose, blind source separation (BSS) has been studied as a possible approach. One BSS approach, called frequency domain independent component analysis (FDICA) [3]–[6], assumes a statistically independent distribution model for source signals, such as a Laplacian distribution model, and minimizes its objective function to estimate separation filters that separate mixture signals into source signals. The objective function with respect to (w.r.t.) the separation filters is defined by, for example, Kullback-Leibler divergence or negative logarithm of maximum likelihood (ML) estimation, and its minimization enables the separation filters to be estimated. Unfortunately, FDICA often suffers from the permutation problem, in which permutation mismatch occurs in adjacent frequency bins of the separated source signals.

To mitigate this permutation problem, independent vector analysis (IVA) has been attracting attention [7]–[12]. This is similar to FDICA except that it considers a super-Gaussian distribution model in which higher-order frequency co-occurrence exists in each source signal across all frequency bins. Moreover, various distribution models for source signals have been studied in IVA, e.g., by combining it with non-negative matrix factorization (NMF) [11] and adding components that depend on pilot source signals [12]. While those methods may be effective in terms of mitigating the permutation problem between adjacent frequency bins having co-occurrence, another permutation problem remains, called the block permutation problem, in which permutation mismatch occurs in each grouped frequency bins of the separated source signals.

To overcome these permutation problems in FDICA and IVA, the use of geometry knowledge has been studied in the BSS community, e.g., by replacing/interpolating the separation filters with traditional beamforming (BF) filters [13], [14], applying geometrical binary masking in post-processing [15], and adding geometric normalization to the IVA objective function with high-level geometrical clues, such as the directions of arrival (DOAs) of the source signals [16], [17]. However, two crucial issues remain: the necessity of non-differential processes and of high-level geometrical clues. That is, to use geometry knowledge, some approaches require non-differential processes outside the minimization in FDICA and IVA [13]–[15], [18]. Other approaches require high-level geometrical clues, thus limiting their use in practical situations [16], [17], [19].

To overcome these issues, we propose a novel IVA method whose objective function has a differentiable max-directivity beamforming normalization (MDBN) term. We first formulate a maximum a posterior (MAP) estimation by extending the standard IVA objective function based on ML estimation. As the MAP estimation allows us to normalize the objective function via prior knowledge, we thus incorporate the differentiable MDBN term. This term uses geometry knowledge from only a low-level geometrical clue (the positions of the microphone array) via the traditional beamforming (BF) concept that each

separation filter should have a maximum gain for each specific DOA across all frequency bins. Thus, our overall objective function, which is based on negative logarithm of MAP estimation, can be minimized by gradient descent, and the MDBN term encourages the separation filters to focus on specific directions, which implicitly estimates the most reasonable DOAs of the source signals at each iteration. Therefore, our method uses geometry knowledge while avoiding the above two issues and estimates good separation filters mitigating the permutation problems. Through several experiments, we confirmed the effectiveness of our proposed method as compared with the conventional BF and IVA methods.

In section 2, we introduce notations of our problem and the standard IVA objective function based on ML estimation. In section 3, we formulate the MAP estimation with our proposed MDBN using geometry knowledge. In section 4, to demonstrate effectiveness of our method, we report the experimental results. We conclude in section 5 with a brief summary. Note that $(\cdot)^{\mathsf{T}}$, $(\cdot)^*$, and $(\cdot)^{\mathsf{H}}$ denote the transposition, complex conjugate, and Hermitian transpose, respectively.

II. Conventional IVA based on maximum likelihood estimation

Given N (≥ 2) source signals, M (≥ 2) mixture signals captured by M microphones, and the positions of a microphone array, we consider short-time Fourier transformed (STFT) mixture signals $x_{ftm} \in \mathbb{C}$ at a frequency bin $f = 1, \ldots, F$, a time frame $t = 1, \ldots, T$, and the m-th microphone. Let $\mathbf{x}_{ft} = [x_{ft1}, \ldots, x_{ftM}]^{\mathsf{T}} \in \mathbb{C}^M$ be a vector of STFT mixture signals. In frequency domain BSS, which includes FDICA and IVA, the mixture signals $\mathbf{x}_{ft}$ are separated into individual source signals for each frequency bin f by assuming that the source signals are statistically independent as

$$\mathbf{y}_{ft} = \mathbf{W}_f \mathbf{x}_{ft}. \tag{1}$$

Here, $\mathbf{y}_{ft} = [y_{ft1}, \ldots, y_{ftN}]^{\mathsf{T}} \in \mathbb{C}^N$ is a vector of N separated signals in the STFT domain, and $\mathbf{W}_f$ is an $N \times M$ separation matrix as

$$\mathbf{W}_f = [\mathbf{w}_{1f}, \ldots, \mathbf{w}_{Nf}]^{\mathsf{H}} \tag{2}$$

with a separation filter $\mathbf{w}_{nf}^{\mathsf{H}} \in \mathbb{C}^{1\times M}$ that outputs the n-th separated signal from the M mixture signals.

On the assumption that a frequency bins vector of the n-th separated signals $\mathbf{y}_{tn} = [y_{1tn}, \ldots, y_{Ftn}]^{\mathsf{T}} \in \mathbb{C}^F$ has co-occurrence across all the frequency bins and follows a multivariate super-Gaussian distribution model $p(\mathbf{y}_{tn})$ (we used the independent Laplacian distribution model [7] in this paper), IVA estimates the separation matrix by maximum likelihood (ML) estimation as

$$\underset{\{\mathbf{W}_f\}_{f=1}^F}{\operatorname{argmax}} \prod_{t=1}^{T} p\left([\mathbf{x}_{t1}, \ldots, \mathbf{x}_{tM}]^{\mathsf{T}} \middle| \{\mathbf{W}_f\}_{f=1}^F\right), \tag{3}$$

where $\mathbf{x}_{tm} = [x_{1tm}, \ldots, x_{Ftm}]^{\mathsf{T}} \in \mathbb{C}^F$. To minimize negative logarithm of this objective function (Eq.(3)), the natural gradient [20] and the auxiliary function [9] methods are often used. Because the co-occurrence in $\mathbf{y}_{tn}$ often exists only between adjacent frequency bins, however, IVA causes the block permutation problem in which permutation mismatch occurs in each grouped frequency bins.

To overcome these permutation problems in FDICA and IVA, the use of geometry knowledge has been studied [13]–[17], [19]. However, some approaches require non-differentiable processes to use geometry knowledge outside the minimization in FDICA and IVA [13]–[15]. Other approaches require high-level geometrical clues, such as the DOAs of the source signals, thus limiting their use in practical situations [16], [17], [19]. Therefore, two crucial issues remain: the necessity of non-differential processes and of high-level geometrical clues.

III. Proposed Method

A. Maximum a posterior estimation for IVA

To mitigate the permutation problems by using geometry knowledge while avoiding the above two crucial issues, we first formulate a MAP estimation by extending the standard IVA objective function based on ML estimation (Eq.(3)) as

$$\begin{aligned} &\underset{\{\mathbf{W}_f\}_{f=1}^F}{\operatorname{argmax}} \prod_{t=1}^{T} p\left(\{\mathbf{W}_f\}_{f=1}^F \middle| [\mathbf{x}_{t1}, \ldots, \mathbf{x}_{tM}]^{\mathsf{T}}\right) \\ &\propto \underset{\{\mathbf{W}_f\}_{f=1}^F}{\operatorname{argmax}} \prod_{t=1}^{T} p\left([\mathbf{x}_{t1}, \ldots, \mathbf{x}_{tM}]^{\mathsf{T}} \middle| \{\mathbf{W}_f\}_{f=1}^F\right) p\left(\{\mathbf{W}_f\}_{f=1}^F\right), \end{aligned} \tag{4}$$

where $p\left(\{\mathbf{W}_f\}_{f=1}^F\right)$ is a prior distribution w.r.t. the separation matrices $\{\mathbf{W}_f\}_{f=1}^F$. Then, this objective function (4) can be reformulated into a negative logarithm minimization problem with the linear constraint of Eq.(1) as

$$\begin{aligned} \underset{\{\mathbf{W}_f\}_{f=1}^F}{\operatorname{argmin}} \sum_{t=1}^{T}\sum_{n=1}^{N} -\log(p(\mathbf{y}_{tn})) - 2T\sum_{f=1}^{F}\log|\det(\mathbf{W}_f)| \\ - \log\left(p\left(\{\mathbf{W}_f\}_{f=1}^F\right)\right). \end{aligned} \tag{5}$$

In this objective function, the first and second terms are associated with the standard IVA objective function based on ML estimation (Eq.(3)), which is effective to separate the mixture signals under the statistical independence assumption for the separated signals. These terms have been reported to encourage the separation filters to form spatial nulls for the DOAs of surrounding noises [13], [14], [21]. Additionally, this reformulated objective function (Eq.(5)) allows us to naturally normalize the original objective function (Eq.(3)) via the prior knowledge appearing in the third term $\log\left(p\left(\{\mathbf{W}_f\}_{f=1}^F\right)\right)$. For this normalization term, we present a new differentiable normalization called max-directivity beamforming normalization (MDBN) in the next subsection.

B. Max-directivity beamforming normalization (MDBN)

The differentiable MDBN term uses geometry knowledge from only a low-level geometrical clue (the positions of the microphone array) via the traditional BF concept that each separation filter should have a maximum gain for each specific

DOA across all frequency bins. Thus, by using MDBN for normalization in Eq.(5), the MDBN term encourages the separation filters to focus on specific directions. In turn, this implicitly estimates the most reasonable DOAs of source signals at each iteration during the minimization of Eq.(5). In this paper, we define the MDBN term by a composition of 5 simple functions as one implementation example:

$$\log\left(p\left(\{\mathbf{W}_f\}_{f=1}^F\right)\right) = \gamma g_1 \circ g_2 \circ g_3\left(\{g_{4f} \circ g_{5f}\left(\mathbf{W}_f\right)\}_{f=1}^F\right), \quad (6)$$

where γ (> 0) is a weight factor to be adjusted before starting the optimization process,

$$g_1\left(\overline{\boldsymbol{\psi}}\right) \coloneqq \left\|\overline{\boldsymbol{\psi}}\right\|_2^2,$$

$$\overline{\boldsymbol{\psi}} = g_2\left(\overline{\boldsymbol{\Psi}}\right) \coloneqq \begin{bmatrix} \max\left(\overline{\psi}_{11}, \ldots, \overline{\psi}_{1\Theta}\right) \\ \vdots \\ \max\left(\overline{\psi}_{N1}, \ldots, \overline{\psi}_{N\Theta}\right) \end{bmatrix}$$

where $\overline{\psi}_{n\theta}$ is the (n,θ)-th element of $\overline{\boldsymbol{\Psi}}$,

$$\overline{\boldsymbol{\Psi}} = g_3\left(\{\boldsymbol{\Psi}_f\}_{f=1}^F\right) \coloneqq \frac{1}{F}\sum_{f=1}^F \boldsymbol{\Psi}_f, \quad (7)$$

$$\boldsymbol{\Psi}_f = g_{4f}\left(\hat{\mathbf{W}}_f\right) \coloneqq \begin{pmatrix} \left|\hat{\mathbf{w}}_{1f}^{\mathsf{H}}\mathbf{a}_{1f}\right| & \cdots & \left|\hat{\mathbf{w}}_{1f}^{\mathsf{H}}\mathbf{a}_{\Theta f}\right| \\ \vdots & \ddots & \vdots \\ \left|\hat{\mathbf{w}}_{Nf}^{\mathsf{H}}\mathbf{a}_{1f}\right| & \cdots & \left|\hat{\mathbf{w}}_{Nf}^{\mathsf{H}}\mathbf{a}_{\Theta f}\right| \end{pmatrix}$$

where $\hat{\mathbf{w}}_{nf}^{\mathsf{H}}$ is the n-th row vector of $\hat{\mathbf{W}}_f$,

$$\hat{\mathbf{W}}_f = g_{5f}(\mathbf{W}_f) \coloneqq \mathbf{B}_f\mathbf{W}_f,$$

where $\mathbf{B}_f = \mathrm{diag}[b_{1f}, \ldots, b_{nf}]$ is a scaling matrix defined by the projection back method [22], which is necessary to avoid the scaling ambiguity problem inherent to BSS, $\mathbf{a}_{\theta f} = [a_{1\theta f}, \ldots, a_{M\theta f}]^{\mathsf{T}} \in \mathbb{C}^M$ is an array manifold vector given the positions of the microphone array assuming that a plane wave arrives from a θ-th discrete DOA indexed by $\theta = 1, \ldots, \Theta$, and the n-th row vector of $\boldsymbol{\Psi}_f$ is a beam pattern of the n-th separation filter $\mathbf{w}_{nf}^{\mathsf{H}}$ for each θ-th discrete DOA.

During the minimization, the MDBN term calculates the current averaged beam patterns $\overline{\boldsymbol{\Psi}}$ by using g_3 and $g_{4f} \circ g_{5f}$, and the maximum gains $\overline{\boldsymbol{\psi}}$ for the most reasonable DOAs by using g_2. Finally, g_1 calculates the total power, representing how much the estimated separation filters focus on the most reasonable DOAs of the source signals at each iteration.

By rewriting the third term in Eq.(5) with our proposed MDBN, we thus formulate a novel objective function $\mathcal{L}$ for IVA with MDBN as

$$\mathcal{L} = \sum_{t=1}^T\sum_{n=1}^N - \log(p(\mathbf{y}_{tn})) - 2T\sum_{f=1}^F \log|\det(\mathbf{W}_f)| - \gamma g_1 \circ g_2 \circ g_3\left(\{g_{4f} \circ g_{5f}\left(\mathbf{W}_f\right)\}_{f=1}^F\right). \quad (8)$$

This objective function can appropriately restrict the feasible region of the separation filters by using geometry knowledge via our proposed MDBN term. It thus estimates good separation filters mitigating the permutation problems.

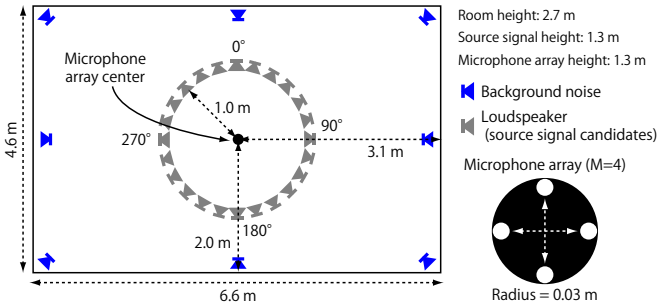

Fig. 1. Experimental conditions.

To simply validate the effectiveness of the proposed MDBN term, we do not forcibly apply advanced optimization algorithms in this paper (rather, we save it for future work). Instead, as all three terms in Eq.(8) are differentiable, we carefully optimize the equation by using the natural gradient method [20] with a small step size, as explained in the next subsection. Although the convergence could be slow, we limit our goal here to the performance regarding the eventually obtained separation filters.

C. Update

This subsection gives an update rule for the separation filters in minimizing our proposed objective function (Eq.(8)). As mentioned before, all terms in Eq.(8) are differentiable. Since the gradients of the first and second term are well known from many previous works [6]–[8], [16], we omit their derivation here. As for the third MDBN term, since it is a composition of 5 simple functions (Eq.(7)), its gradient can be obtained by back-propagation [22] based on the chain rule. Therefore, the gradient of the third MDBN term w.r.t. each complex conjugate element of the separation matrix $(w_{nmf}^*)_{n,m=1}^{N,M} = \mathbf{W}_f^*$ is calculated as

$$\begin{aligned} &\frac{\partial}{\partial w_{nmf}^*} g_1 \circ g_2 \circ g_3\left(\{g_{4f} \circ g_{5f}\left(\mathbf{W}_f\right)\}_{f=1}^F\right) \\ &= \frac{\partial g_1\left(\overline{\boldsymbol{\psi}}\right)}{\partial \overline{\boldsymbol{\psi}}} \frac{\partial \overline{\boldsymbol{\psi}}}{\partial \overline{\psi}_{n\hat{\theta}_n}} \frac{\partial \overline{\psi}_{n\hat{\theta}_n}}{\partial \psi_{n\hat{\theta}_n f}} \frac{\partial \psi_{n\hat{\theta}_n f}}{\partial \hat{w}_{nmf}^*} \frac{\partial \hat{w}_{nmf}^*}{\partial w_{nmf}^*} \\ &= \frac{1}{F}\overline{\psi}_{n\hat{\theta}_n} \frac{1}{\sqrt{\psi_{n\hat{\theta}_n f}}} a_{m\hat{\theta}_n f}\mathbf{a}_{\hat{\theta}_n f}^{\mathsf{H}}\hat{\mathbf{w}}_{nf}b_{nf}^* \\ &\approx \frac{1}{F}(\overline{\psi}_{n\hat{\theta}_n})' \frac{1}{\sqrt{\psi_{n\hat{\theta}_n f}}} a_{m\hat{\theta}_n f}\mathbf{a}_{\hat{\theta}_n f}^{\mathsf{H}}\hat{\mathbf{w}}_{nf}b_{nf}^*, \end{aligned} \quad (9)$$

where $\overline{\psi}_{n\hat{\theta}_n}$ is the n-th element of $\overline{\boldsymbol{\psi}}$ and the $(n,\hat{\theta}_n)$-th element of $\overline{\boldsymbol{\Psi}}$ with $\hat{\theta}_n = \mathrm{argmax}_{\theta=1,\ldots,\Theta}\,\overline{\psi}_{n\theta}$. Here, $\hat{\theta}_n$ indicates the most reasonable DOA index on which the separation filter $\mathbf{w}_{nf}^{\mathsf{H}}$ focuses at each iteration, and $\psi_{n\hat{\theta}_n f}$ is the $(n,\hat{\theta}_n)$-th element of $\boldsymbol{\Psi}_f$. We also present the approximation of this gradient with user-selected frequency bins $f = f_1, \ldots, f_2$ and $(\overline{\psi}_{n\hat{\theta}_n})' = 1/(f_2 - f_1) \cdot \sum_{f=f_1}^{f_2} \psi_{n\hat{\theta}_n f}$, which enables us to consider reliable frequency bins for the source signals' properties.

After the above calculation for each element w_{nmf}^* of $\mathbf{W}_f^*$, we can provide the update rule of the current separation matrix

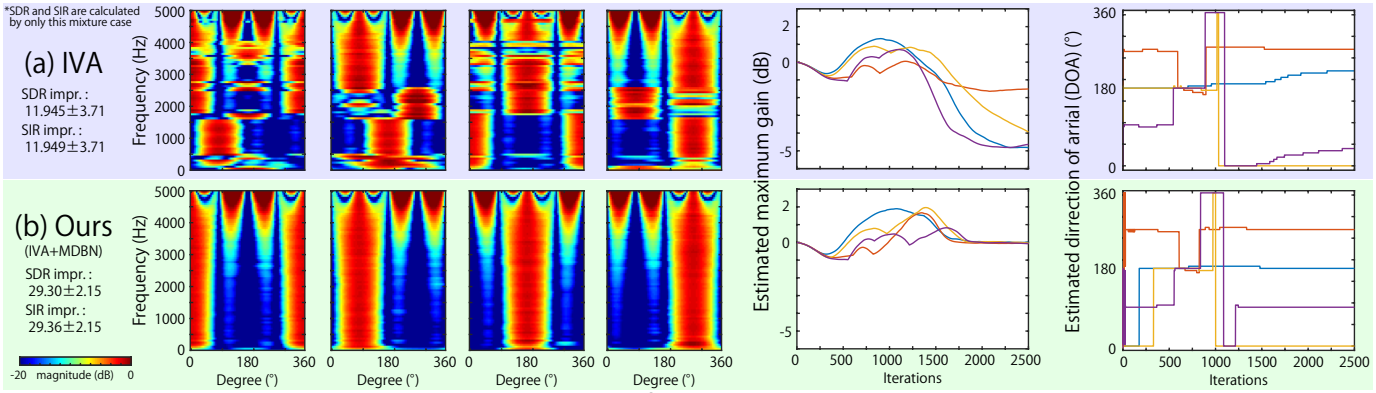

Fig. 2. Beam pattern for each estimated separation filter, $20\log_{10}\left(\psi_{n\theta f}\right)_{\theta,f=1}^{\Theta,F}$, in R1. The source signals arrive from the four loudspeakers at $0°, 90°, 180°$, and $270°$ degree. Whereas IVA caused the block permutation problems (top left panels), our proposed method (IVA+MDBN) could mostly mitigate those problems (bottom left panels). Moreover, in our method, the estimated maximum gains $\overline{\psi}_{n\hat{\theta}_n}$ stayed higher (bottom middle plot), and the $\hat{\theta}_n$-th estimated reasonable DOAs converged to the ideal DOAs of the source signals (bottom right plot).

TABLE I
SEPARATION PERFORMANCE IN TERMS OF AVERAGE ± STANDARD DEVIATION OF SDR AND SIR IMPROVEMENT FOR ALL MIXTURE CASES.

Algorithm	R1 ($RT_{60} = 0$ ms, *simulation)		R2 ($RT_{60} = 110$ ms)		R3 ($RT_{60} = 230$ ms)	
	SDR (dB)	SIR (dB)	SDR (dB)	SIR (dB)	SDR (dB)	SIR (dB)
None (mixtures)	-4.92 ± 2.31	-4.92 ± 2.31	-4.46 ± 1.63	-4.20 ± 1.67	-4.63 ± 1.81	-4.21 ± 1.88
	SDR improvement (dB)	SIR impr. (dB)	SDR impr. (dB)	SIR impr. (dB)	SDR impr. (dB)	SIR impr. (dB)
MVDR [23]	7.72 ± 5.28	7.73 ± 5.28	4.34 ± 2.59	4.37 ± 2.81	4.65 ± 2.96	4.76 ± 3.38
IVA [7]	17.95 ± **9.46**	17.98 ± **9.46**	7.55 ± **7.11**	10.13 ± **8.23**	6.97 ± **6.14**	9.65 ± **7.23**
Ours (IVA+MDBN)	**20.18** ± 8.91	**20.21** ± 8.91	**8.03** ± 6.75	**10.74** ± 7.93	**7.60** ± 5.91	**10.44** ± 6.92

$\mathbf{W}_f^{(k)}$ for minimizing the objective function $\mathcal{L}$ in Eq.(8) by following the natural gradient method [20] as

$$\mathbf{W}_f^{(k+1)} = \mathbf{W}_f^{(k)} - \alpha \frac{\partial \mathcal{L}}{\partial \mathbf{W}_f^*} \left(\mathbf{W}_f^{(k)}\right)^{\mathsf{H}} \mathbf{W}_f^{(k)}, \qquad (10)$$

where the step size α (> 0) is selected to be sufficiently small. The initial separation matrix was an identity matrix, but by repeating K iterations, we eventually obtain separation filters.

IV. EXPERIMENTS

A. Setup

To evaluate the effectiveness of our proposed IVA with differentiable MDBN, we firstly conducted a numerical simulation. In this simulation (R1), we numerically generated impulse responses composed of a direct plane wave without background noise and room reverberation. Next, we conducted real experiments. We measured the impulse responses in two reverberation rooms with background pink noise as shown in Fig. 1. By changing the wall/ceiling material, the reverberation time was changed to $RT_{60} = 110$ ms (R2) and $RT_{60} = 230$ ms (R3) at 1 kHz. The background pink noise had power spectral density of the form $S(f) \propto \frac{1}{f}$, and its power was maintained at -40 dB w.r.t. the sum of the source signals.

In all situations (R1, R2, and R3), the impulse responses were measured from four loudspeakers ($N = 4$), which are randomly chosen from 24 loudspeakers placed at 15 degree intervals, to circular four microphones ($M = 4$) with the radius of 0.03 meters as shown in Fig.1. Note that the positions of the microphone array were given but the DOAs of the source signals were unknown. We made 20 mixture cases by convolving the simulated/measured impulse responses to 14-second dry speech source signals recorded with a 16 kHz sampling frequency. The signals consisted of three sentences randomly selected from 10 male/female speaker utterances. The frame length and frame shift of the STFT were set to 16 ms and 4 ms, respectively. We set the each θ-th discrete DOA to define the each array manifold vector $\{\mathbf{a}_{\theta f}\}_{\theta=1}^{\Theta}$ in Eq.(7) as $0°, 5°, \ldots, 350°, 355°$ that match the indices $\theta = 1, \ldots, \Theta$, the weight γ multiplied to MDBN term as $[0.01, 0.11]$ for each mixture case, the frequency range in the gradient approximation as $(f_1, f_2) = (500, 3000)$ Hz for speech enhancement, and the number of iterations as $K = 2000$ with a small step size $\alpha = 0.0125$. The separation performance was evaluated in terms of signal-to-distortion ratio (SDR) and signal-to-inference ratio (SIR) by using BSS-EVAL [21].

B. Results

Figure 2 shows the results for R1 at the four chosen loudspeakers $(0°, 90°, 180°, 270°)$. During and after minimization, we checked the beam pattern in terms of $20\log_{10}\left(\psi_{n\theta f}\right)_{\theta,f=1}^{\Theta,F}$ (left panels), the estimated maximum gain $\overline{\psi}_{n\hat{\theta}_n}$ (middle plots), and the $\hat{\theta}_n$-th estimated reasonable DOA at each iteration (right plots).

With IVA [7], the beam patterns seemed unstable due to the block permutation problem (top left panels). In contrast, with our proposed IVA with MDBN, the problem was mostly mitigated, and each beam pattern was clearly focused on the most reasonable DOA (bottom left panels). During the minimization, IVA and our method seemed to behave similarly until around 1000 – 1500 iterations (middle and right plots).

However, since IVA failed to form a beam pattern focused on a specific DOA, the estimated maximum gain decreased (top middle plot), and the estimated DOAs did not converge to the ideal DOAs of the source signals (top right plot). In contrast, as our method encouraged forming a beam pattern focused on a specific DOA with the imposition of MDBN, the estimated maximum gain stayed higher (bottom middle plot). Furthermore, the estimated DOAs converged, surprisingly, to the ideal DOAs of the source signals (bottom right plot), which was thanks to the good combination of the statistically independent with IVA terms (first and second terms in Eq.(8)) and geometry knowledge with the MDBN term (third term).

Table I shows separation performance. Our proposed IVA with MDBN could outperform the conventional MVDR [23] and IVA [7] methods, especially in R1. We consider the results to be caused by sameness of each microphone's sensitivity in R1 and appearance of reflected source signals in R2/R3. In those situations, as the MDBN term struggled to form maximum gains for the DOAs of source signals, our method might not have worked well, and the separation performance fell to around that of the conventional IVA. Note that SDR and SIR are similar in R1 because their definitions in BSS-EVAL [21] get close in such no noise and reverberation situation.

V. Conclusions

We have proposed IVA with differentiable MDBN using geometry knowledge to mitigate the block permutation problem of IVA. The differentiable MDBN term uses geometry knowledge from only a low-level geometrical clue (the positions of the microphone array) via the traditional BF concept that each separation filter should have a maximum gain for each specific DOA across all frequency bins. Thus, our overall objective function can be minimized by gradient descent, and the MDBN term encourages separation filters to focus on specific directions, which implicitly estimates the most reasonable DOAs of the source signals at each iteration. Several experiments demonstrated that IVA with MDBN can mitigate the permutation problem and outperforms the conventional BF and IVA methods. For a future work, we should make MDBN more robust to reverberation as in R2 and R3, and compare it more thoroughly to previous methods with geometry knowledge (in contrast, this paper focused on validating the effectiveness of MDBN for IVA).

References

[1] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W. J. Fabian, M. Espi, T. Higuchi, S. Araki, and T. Nakatani, "The ntt chime-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec 2015, pp. 436–443.

[2] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *Trans. Audio, Speech and Lang. Proc.*, vol. 15, no. 7, pp. 2011–2022, Sept. 2007.

[3] H. Sawada, R. Mukai, S. Araki, and S. Makino, *Frequency-Domain Blind Source Separation*, pp. 299–327, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.

[4] S.S. Haykin, *Unsupervised Adaptive Filtering: Blind source separation*, Wiley series on adaptive and learning systems for signal processing, communications, and control. Wiley, 2000.

[5] S. Makino, T.W. Lee, and H. Sawada, *Blind Speech Separation*, Springer Netherlands, 2007.

[6] H. Sawada, N. Ono, H. Kameoka, D. Kitamura, and H. Saruwatari, "A review of blind source separation methods: two converging routes to ilrma originating from ica and nmf," *APSIPA Transactions on Signal and Information Processing*, vol. 8, pp. e12, 2019.

[7] T.Kim, H.T.Attias, S.Y.Lee, and T.W.Lee, "Blind source separation exploiting higher-order frequency dependencies," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, pp. 70 – 79, 02 2007.

[8] A. Hiroe, "Solution of permutation problem in frequency domain ica, using multivariate probability density functions," in *Independent Component Analysis and Blind Signal Separation*, J. Rosca, Deniz E., J. Príncipe, and S. Haykin, Eds., Berlin, Heidelberg, 2006, pp. 601–608, Springer Berlin Heidelberg.

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

[10] N. Ono, "Auxiliary-function-based independent vector analysis with power of vector-norm type weighting functions," in *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, Dec 2012, pp. 1–4.

[11] 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, Sep. 2016.

[12] F. Nesta and Z. Koldovský, "Supervised independent vector analysis through pilot dependent components," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 536–540.

[13] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ica and beamforming," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, pp. 666 – 678, 04 2006.

[14] K. Osako, Y. Mori, Y. Takahashi, H. Saruwatari, and K. Shikano, "Fast convergence blind source separation using frequency subband interpolation by null beamforming," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E91.A, no. 6, pp. 1357–1361, 2008.

[15] Y. Tachioka, T. Narita, and J. Ishii, "Semi-blind source separation using binary masking and independent vector analysis," *IEEJ Transactions on Electrical and Electronic Engineering*, vol. 10, 01 2015.

[16] A. Khan, M. Taseska, and E. Habets, "A geometrically constrained independent vector analysis algorithm for online source extraction," in *Latent Variable Analysis and Signal Separation*, 08 2015, vol. 9237, pp. 396–403.

[17] B. Andreas, H. Thomas, and K. Walter, "Spatially informed independent vector analysis," in *arXiv*, 2019.

[18] I. Himawan, I. McCowan, and M. Lincoln, "Microphone array beamforming approach to blind speech separation," in *Proceedings of the 4th International Conference on Machine Learning for Multimodal Interaction*. 2007, MLMI'07, p. 295–305, Springer-Verlag.

[19] W. Zhang and B. D. Rao, "Combining independent component analysis with geometric information and its application to speech processing," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2009, pp. 3065–3068.

[20] S. Amari, S. C. Douglas, A. Cichocki, and H. H. Yang, "Multichannel blind deconvolution and equalization using the natural gradient," in *First IEEE Signal Processing Workshop on Signal Processing Advances in Wireless Communications*, April 1997, pp. 101–104.

[21] S. Kurita, H.Saruwatari, S.Kajita, K.Takeda, and F.Itakura, "Evaluation of blind signal separation method using directivity pattern under reverberant conditions," in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*. IEEE, 2000, vol. 5, pp. 3140–3143.

[22] D.E.Rumelhart, G.E.Hinton, R.J.Williams, et al., "Learning representations by back-propagating errors," *Cognitive modeling*, vol. 5, no. 3, pp. 1, 1988.

[23] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug 1969.