GEOMETRICALLY CONSTRAINED INDEPENDENT VECTOR ANALYSIS FOR DIRECTIONAL SPEECH ENHANCEMENT

Li Li[1]*, Kazuhito Koishida*[2]

[1] University of Tsukuba, Japan [2] Microsoft Corporation, USA

ABSTRACT

This paper addresses the multichannel directional speech enhancement problem with geometrically constrained independent vector analysis (GCIVA), where we aim to combine the high separation performance from blind source separation and the capability of directional focus from beamforming. The proposed method exploits geometric constraints composed from the spatial information of sources to guide the target speech to the desired output channel. A convergence-guaranteed parameter estimation algorithm is derived from the framework of auxiliary function-based IVA (AuxIVA) to take advantage of fast convergence, low computational cost, and no step-size tuning. We propose a dual-microphone speech enhancement system based on the proposed method and investigate its effectiveness with objective metrics. The experimental evaluations revealed that the proposed system outperformed the conventional beamforming and the standard AuxIVA in a large margin in terms of source-to-distortion and source-to-interference ratios.

Index Terms— Speech enhancement, independent vector analysis, geometric constraints, multichannel, auxiliary function approach

1. INTRODUCTION

Speech enhancement is a crucial technology for extracting the target speech from recorded noisy signals since the presence of diffuse noise and directional interference can significantly degrade the performances of many speech processing applications. Various speech enhancement algorithms [1] have been developed with the goal of overcoming this problem.

Blind source separation (BSS) is one promising approach, which copes with multichannel scenarios. BSS algorithms, including a variety of independent component analysis (ICA) methods [2, 3, 4, 5], estimate source signals using only the observed signals based on the assumption that source signals are statistically independent with each other. Independent vector analysis (IVA) [3, 4] is one such method, which models the whole frequency components as a multivariate variable following a spherical multivariate distribution so that permutation ambiguity can be avoided by exploiting higher-order dependencies of signals. Owing to the high separation performance, IVA has attracted much attention and been widely studied, which promotes the practicability of the approach in various scenarios. A fast and stable algorithm based on the auxiliary function approach, referred to as AuxIVA [6], has been recently developed and experimentally demonstrated to perform well in both offline and online cases with low computational costs [7, 8, 9, 10]. However, when considering a practical application of speech enhancement, an additional process is necessary for selecting the target speech after the separation, which is typically performed by utilizing the spatial information, i.e., the direction of arrival (DOA) of the target. Moreover, it is reported that block permutation problem [11] occurs between the low- and high-frequency bands in IVA, which results in the degradation of the performance.

To improve the performance of BSS algorithms and avoid the permutation problem, exploiting spatial information to guide the demixing matrices is one promising method. [12] derives IVA in a maximum a posteriori (MAP) fashion so that a spatially informed prior of demixing matrices can be incorporated into the optimization. Another well-known framework is the geometrically constrained BSS [13, 14, 15, 16]. In this framework, beamforming-based geometric constraints derived from prior spatial information of source signals and the sensor geometry are combined with the optimization problem of BSS, which makes it possible to manually control the spatial and frequency responses of the demixing filter estimated by BSS. In [17], a penalty term restricting the Euclidean angle between the separation filter and the far-field steering vector calculated from the desired source DOA is combined with IVA to force the desired signal always being outputted at the corresponding channel, which has been shown to improve the performance of IVA in directional speech enhancement. However, there are two drawbacks to prevent this method from a wide adoption to real applications. Firstly, a relatively large number of microphones are needed to meet the constraints of forming a sharp beam and suppressing interferences at the same time. Secondly, the step-size parameter of the gradient-based algorithm must be carefully tuned to make the system work under different real use cases.

In this paper, we propose an approach of geometrically constrained IVA (GCIVA), which combines linear constraints that restrict far-field responses of demixing filters [13] with IVA. To preserve the advantages of fast convergence and no step-size tuning from AuxIVA, we derive a convergence-guaranteed algorithm based on the auxiliary function approach with adopting the idea of vectorwise coordinate descent (VCD) [18] to obtain the closed-form solution. The proposed method is called "GCAV (Geometrically Constrained Auxiliary-function with VCD)-IVA". We introduce a dual-microphone system based on the proposed method. In the system, the interference channel is constrained in such a way that a null is formed toward the target direction which is assumed known. Regarding the constraint on the target channel, it is found that a null constraint toward the estimated interference DOA is the best option, where the standard AuxIVA is used for this DOA estimation purpose. Experimental results show that the proposed GCAV-IVA system can offer higher performance than the conventional beamforming and the standard AuxIVA in the dual-microphone setting.

This work was performed while Li Li was an intern at Microsoft Corporation.

2. FORMULATION OF GEOMETRICALLY CONSTRAINED IVA

Let us consider a determined situation where I sources are observed by I microphones. Let $x_i(\omega, t)$ and $y_j(\omega, t)$ denote the short-time Fourier transform (STFT) coefficients of the signal observed at the i-th microphone and the j-th estimated sources, respectively. Here ω and t are the frequency and time indices, respectively. We denote the frequency-wise vector representation of the observations and the estimated sources by

$$\boldsymbol{x}(\omega, t) = [x_1(\omega, t), \ldots, x_I(\omega, t)]^{\mathsf{T}} \in \mathbb{C}^I, \tag{1}$$

$$\boldsymbol{y}(\omega, t) = [y_1(\omega, t), \ldots, y_J(\omega, t)]^{\mathsf{T}} \in \mathbb{C}^J, \tag{2}$$

where $J = I$ and $(\cdot)^{\mathsf{T}}$ denotes the transpose. When the STFT window length is sufficiently longer than the impulse responses between sources and microphones, the relationship between the observations and the estimated sources can be expressed with the time-invariant instantaneous mixture model as:

$$\boldsymbol{y}(\omega, t) = \boldsymbol{W}(\omega)\boldsymbol{x}(\omega, t), \tag{3}$$

where $\boldsymbol{W}(\omega) = [\boldsymbol{w_1}(\omega), \ldots, \boldsymbol{w_I}(\omega)]^{\mathsf{H}}$ is an $I \times I$ demixing matrix and $(\cdot)^{\mathsf{H}}$ denotes Hermitian transpose.

IVA assumes that sources follow a multivariate distribution and thus dependencies over frequency components can be exploited to avoid the permutation problem. The demixing matrices $\mathcal{W} = \{\boldsymbol{W}(\omega)\}_\omega$ are estimated by minimizing the following objective function

$$J_{\text{IVA}}(\mathcal{W}) = \sum_{j=1}^{J} \mathbb{E}[G(\boldsymbol{y}_j(t))] - \sum_{\omega=1}^{\Omega} \log|\det \boldsymbol{W}(\omega)|, \tag{4}$$

where Ω denotes the number of frequency bins. $\mathbb{E}[\cdot]$ denotes the expectation operator and $\boldsymbol{y}_j(t)$ is the source-wise vector representation defined as

$$\boldsymbol{y}_j(t) = [y_j(1, t), \ldots, y_j(\Omega, t)]^{\mathsf{T}} \in \mathbb{C}^\Omega. \tag{5}$$

Here, $G(\boldsymbol{y}_j(t))$ is the contrast function having a relationship of $G(\boldsymbol{y}_j(t)) = -\log p(\boldsymbol{y}_j(t))$, where $p(\boldsymbol{y}_j(t))$ represents a multivariate probability density function of the j-th source. One typical choice of the contrast function is using spherical multivariate distribution [3, 4, 6], which is expressed as

$$G(\boldsymbol{y}_j(t)) = G_R(r_j(t)), \tag{6}$$

$$r_j(t) = ||\boldsymbol{y}_j(t)||_2 = \sqrt{\sum_{\omega} |y_j(\omega, t)|^2}. \tag{7}$$

Here, $||\cdot||_2$ denotes L_2 norm of a vector.

Now, let us consider a geometric constraint [13] that restricts the far-field response of the j-th demixing filter estimated by IVA at the direction θ, which is described as

$$\mathcal{J}_c(\mathcal{W}) = \sum_{j=1}^{J} \lambda_j \sum_{\omega=1}^{\Omega} |\boldsymbol{w}_j^{\mathsf{H}}(\omega)\boldsymbol{d}_j(\omega, \theta) - c_j|^2. \tag{8}$$

Here, $\boldsymbol{d}_j(\omega, \theta)$ is the steering vector pointing to the direction θ, c_j is the nonnegative-valued constraint, and $\lambda_j \geq 0$ is a parameter weighing the importance of the constraint. This concept is used in the linearly constrained minimum variance (LCMV) beamformer [19]. Note that (8) with $c_j = 1$ forces the spatial filter to form a conventional delay-and-sum beamformer steering at the direction θ to preserve the target source while a small value of c_j essentially creates a spatial null towards the target direction θ aiming at suppressing the target source and preserving all other sources. The null constraint on the target direction can also serve as a blocking matrix (BM) [20], so that the corresponding channel can produce good estimate of interference and noise. Such estimate would have potential benefit of better handling under/overdetermined cases compared to traditional BSS methods. The objective function of the proposed GCIVA is summarized as

$$J(\mathcal{W}) = J_{\text{IVA}}(\mathcal{W}) + J_c(\mathcal{W}). \tag{9}$$

3. INFERENCE ALGORITHM WITH AUXILIARY FUNCTION APPROACH

In this section, we derive an iterative algorithm for parameter estimation of (9) with the auxiliary function approach [21], which has already been employed in IVA and yielded the fast convergence and stable performance. In the approach, an auxiliary function $J^+(\mathcal{W}, \mathcal{V})$ is designed in such a way that $J(\mathcal{W}) = \min_{\mathcal{V}} J^+(\mathcal{W}, \mathcal{V})$ is satisfied. Then, instead of directly optimizing the original objective function (9), which is difficult to be analytically solved, the auxiliary function $J^+(\mathcal{W}, \mathcal{V})$ is minimized in terms of $\mathcal{W}$ and $\mathcal{V}$ alternatingly. Since the geometric constraints are linear, we can simply obtain the auxiliary function that upper-bounds (9) by combining the original AuxIVA's auxiliary function [6] with these linear constraints:

$$J^+(\mathcal{W}, \mathcal{V}) \overset{c}{=} \sum_{j=1}^{J} \sum_{\omega=1}^{\Omega} \Big\{ \frac{1}{2} \sum_{j} \boldsymbol{w}_j^{\mathsf{H}}(\omega)\boldsymbol{V}_j(\omega)\boldsymbol{w}_j(\omega) - \log|\det \boldsymbol{W}(\omega)| \Big\} + J_c(\mathcal{W}), \tag{10}$$

where $\boldsymbol{V}_j(\omega)$ is the weighted covariances expressed as

$$\boldsymbol{V}_j(\omega) = \mathbb{E}\Big[\frac{G_R'(r_j(t))}{r_j(t)}\boldsymbol{x}(\omega)\boldsymbol{x}^{\mathsf{H}}(\omega)\Big] \tag{11}$$

and $\overset{c}{=}$ denotes equality up to constant terms. Here, $(\cdot)'$ denotes the derivative operator.

The update rule for $\mathcal{V}$ is obtained straightforwardly by applying (7) into (11). Here we focus on deriving the update rule for $\mathcal{W}$. The indices of ω and θ are omitted hereafter for the notation simplicity. Due to the linear constraint terms, the equation $\partial J^+(\mathcal{W}, \mathcal{V})/\partial \boldsymbol{w}_j^* = 0$ cannot be solved as Hybrid Exact-Approximate Joint Diagonalization (HEAD) problem anymore, where $(\cdot)^*$ denotes the complex conjugate. To obtain the optimal $\boldsymbol{w}_j$ of (10) with fixed $\mathcal{V}$, inspired by the vectorwise coordinate descent (VCD) method [18], we embrace the idea of arranging the term $\log|\det \boldsymbol{W}|$ by using the property of cofactor expansion

$$\boldsymbol{B} = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_J] \overset{\text{def}}{=} (\det \boldsymbol{W})\boldsymbol{W}^{-1}, \tag{12}$$

where $\boldsymbol{b}_j$ is the j-th column of the adjugate matrix of $\boldsymbol{W}$ de-

fined as

$$\boldsymbol{B}_{pq} = (-1)^{p+q}\tilde{\boldsymbol{W}}_{qp}. \tag{13}$$

Here, the index pq denotes the (p, q) entry of $\boldsymbol{B}$ and $\tilde{\boldsymbol{W}}_{qp}$ is the (q, p) minor determinant of $\boldsymbol{W}$. We can then obtain $\det \boldsymbol{W} = \boldsymbol{w}_j^{\mathsf{H}}\boldsymbol{b}_j$. The partial derivative of (10) w.r.t. $\boldsymbol{w}_j^*$ is calculated as

$$\frac{\partial J^+(\mathcal{W}, \mathcal{V})}{\partial \boldsymbol{w}_j^*} = \boldsymbol{D}_j\boldsymbol{w}_j - \frac{\boldsymbol{b}_j}{\boldsymbol{w}_j^{\mathsf{H}}\boldsymbol{b}_j} - \lambda_j c_j \boldsymbol{d}_j, \tag{14}$$

where $\boldsymbol{D}_j = \boldsymbol{V}_j + \lambda_j \boldsymbol{d}_j \boldsymbol{d}_j^{\mathsf{H}}$. Note that (14) has the same form with the equation (13) in [18], whose closed-form solution can be derived in the same procedure introduced in [18]. We omit the derivation here due to the space limitation. The update rules of $\boldsymbol{w}_j$ are summarized as follow:

$$\boldsymbol{u}_j = \boldsymbol{D}_j^{-1}\boldsymbol{W}^{-1}\boldsymbol{e}_j, \tag{15}$$

$$\hat{\boldsymbol{u}}_j = \lambda_j c_j \boldsymbol{D}_j^{-1}\boldsymbol{d}_j, \tag{16}$$

$$h_j = \boldsymbol{u}_j^{\mathsf{H}}\boldsymbol{D}_j\boldsymbol{u}_j, \tag{17}$$

$$\hat{h}_j = \boldsymbol{u}_j^{\mathsf{H}}\boldsymbol{D}_j\hat{\boldsymbol{u}}_j, \tag{18}$$

$$\boldsymbol{w}_j = \begin{cases} \frac{1}{\sqrt{h_j}}\boldsymbol{u}_j + \hat{\boldsymbol{u}}_j \ \ (\text{if } \hat{h}_j = 0), \\ \frac{\hat{h}_j}{2h_j}\Big[-1 + \sqrt{1 + \frac{4h_j}{|\hat{h}_j|^2}}\Big]\boldsymbol{u}_j + \hat{\boldsymbol{u}}_j \ \ (\text{o.w.}). \end{cases} \tag{19}$$

Here, $\boldsymbol{e}_j$ is the j-th column of the $I \times I$ identity matrix. These update rules are equivalent to those employed in AuxIVA when $\lambda_j = 0$. It is noteworthy that the algorithm takes benefits of the auxiliary function approach, namely, no step-size tuning and fast convergence. Moreover, the algorithm having similar updating procedures with AuxIVA allows us to adopt autoregressive estimation [9] to develop online systems, which is indispensable in real-time and low-latency applications. In the following sections, we adopt the proposed GCAV-IVA method to a dual-microphone system and evaluate the effectiveness via simulation.

4. SYSTEM FOR A DUAL-MICROPHONE CASE

To develop a dual-mirophone system, we take the following conditions into consideration:

- The correct DOA of the target speaker θ_{t} is known;
- Null constraints are employed, i.e. $c_j = 0$ or close to zero. It is a practical choice since only two microphones are available.

Fig. 1 shows an overview of the proposed system. Under the conditions above, we always apply a null constraint to the interference channel, where the null is formed toward the target speaker direction. For the target channel, we evaluate three options in the next section.

1. No constraint.
2. Null constraint at the interference direction from the oracle in 2-speaker case or at a dummy interference direction in 1-speaker case. This option is only for reference purpose.
3. Null constraint at the interference direction estimated by a separate AuxIVA system.

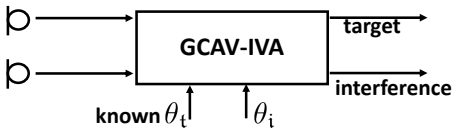

Fig. 1. Basic system structure.

The motivation of third option is that, as demonstrated in Section 5, we find that the constraining both channels can lead to a higher enhancement performance. In this option, the interference DOA is obtained from a separate AuxIVA system. Since a BSS system can be interpreted as a set of adaptive null-beamformers [22], the directional nulls, which can be identified from the directivity patterns, usually point out the directions where the sources come from [14, 23, 24]. In the system, the DOA of the j-ch output sources is given as

$$\hat{\theta}_j = \underset{\theta}{\operatorname{argmin}} \sum_{\omega=1}^{\Omega/2} |\boldsymbol{w}_j^{\mathsf{H}}(\omega)\boldsymbol{d}(\omega, \theta)|. \tag{20}$$

The interference DOA $\hat{\theta}_{\mathrm{i}}$ can then be obtained by selecting the one far away from the target DOA θ_{t}:

$$\hat{\theta}_{\mathrm{i}} = \underset{\hat{\theta}_j}{\operatorname{argmax}} \left[|\hat{\theta}_j - \theta_{\mathrm{t}}|\right], \ j = 1, 2 \tag{21}$$

5. EXPERIMENTAL EVALUATIONS

5.1. Data and settings

To evaluate the effectiveness of the proposed GCAV-IVA method and the dual-microphone system, we conducted speech enhancement experiments in two situations: 2-speaker case where both target and interference speaker exist and 1-speaker case where only the target speaker exists.

We used speech samples of 4 speakers (2 females and 2 males) excerpted from Voice Conversion Challenge 2018 (VCC2018) database [25], which included 81 sentences for each speaker. The audio files were about 3-7 seconds long. The mixture signals were created by simulating two-channel recordings of two sources where the room impulse responses (RIRs) were synthesized using the image method [26]. Fig. 2 shows the positions of the sources and microphones. The interval of microphones was set at 5 cm. 2 DOA settings were investigated in the 2-speaker case, and 3 settings were investigated in the 1-speaker case. We tested two different reverberant conditions where the reverberation time (RT_{60}) was about 200 ms and 470 ms, which were controlled by setting the reflection coefficient of the walls at 0.4 and 0.8. To simulate the more realistic acoustic environment, 4 types of diffuse noise excerpted from DEMAND database [27], including park, office, cafeteria, and metro, were added to reverberant speech signals. We generated 1920 and 960 test samples for the 2-speaker and 1-speaker cases with various target-to-interference energy ratios and speech-to-noise energy ratios. The signal-to-noise ratios (SNRs) of the test samples in the 2-speaker case and 1-speaker case were between [-2, 6] dB and [0, 6] dB, respectively.

All the speech signals were sampled at 16 kHz. The STFT was computed using a Hanning window whose length was set at 32 ms, and the window shift was 16 ms. We compared the minimum power distortionless response (MPDR) beamformer [28] calculated with the far-field steering vectors,

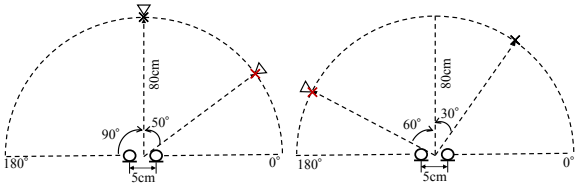

Fig. 2. Configurations of sources and microphones, where "×" and "△" denote source positions used for 2-speaker and 1-speaker case, respectively. Red "×" denotes the target.

Table 1. Summary of tested GCAV-IVA systems.

System #	θ_{i}	c_0	c_1	λ_0	λ_1
(1)	No constraint	—			
(2)	Known	0	0	2	10
(3)		0.5	0.2		
(4)	Estimated by AuxIVA	0	0		
(5)		0.5	0.2		

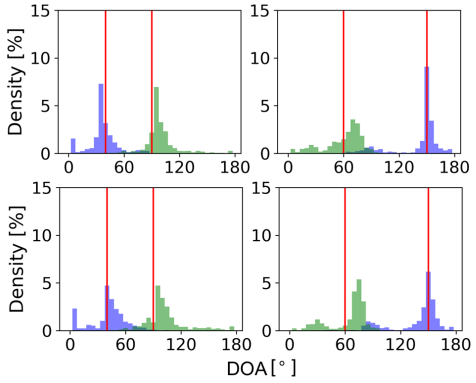

Fig. 3. DOA estimation results achieved by performing AuxIVA update for 3 times under reverberant conditions where $RT_{60} = 200$ ms (upper) and $RT_{60} = 470$ ms (bottom). Red lines show true DOAs. Blue and green graphs are estimated DOA histograms for two directions.

the AuxIVA using $G_R(r_j(t)) = r_j(t)$, and the GCAV-IVA method with various constraints. The specific settings of the tested systems are summarized in Table 1. Source-to-distortion ratios (SDR), source-to-interferences ratios (SIR) and sources-to-artifacts ratios (SAR) [29] were computed to evaluate the enhancement performance. For MPDR and GCAV-IVA, we evaluated the output from the target channel, whereas for AuxIVA, we evaluated outputs from all the channels and took the best score as the result.

5.2. DOA estimation results

First we investigated the potential of the standard AuxIVA as a DOA estimator. The AuxIVA had 3 update iterations and the DOA range was set at $[0°, 180°]$ with an interval of $5°$. Fig. 3 shows the estimation results in a histogram format, which were calculated from the 2-speaker dataset. It is revealed that more than 60% of the estimated directions is located in the range of $\pm 20°$ against the true DOA. In the next subsection, we will demonstrate the benefit of the DOA estimation in speech enhancement experiments.

Table 2. SDR, SIR, and SAR [dB] of 2-speaker case.

Method	$RT_{60} = 200$ ms			$RT_{60} = 470$ ms		
	SDR	SIR	SAR	SDR	SIR	SAR
unproc	1.46	1.61	23.02	0.78	1.47	12.11
MPDR	3.82	4.89	12.30	3.55	5.33	9.95
AuxIVA	7.12	8.98	**14.05**	4.96	7.42	**10.51**
GCAV-IVA(1)	8.42	11.19	13.33	6.47	10.33	9.86
GCAV-IVA(2)	8.71	11.50	13.53	6.51	10.34	9.89
GCAV-IVA(3)	8.75	11.62	13.49	6.55	**10.50**	9.84
GCAV-IVA(4)	8.72	11.52	13.52	6.53	10.36	9.93
GCAV-IVA(5)	**8.80**	**11.69**	13.51	**6.57**	**10.50**	9.88

Table 3. SDR, SIR, and SAR [dB] of 1-speaker case.

Method	$RT_{60} = 200$ ms			$RT_{60} = 470$ ms		
	SDR	SIR	SAR	SDR	SIR	SAR
unproc	3.03	3.37	21.61	2.14	3.06	12.48
MPDR	1.29	2.79	9.50	2.14	4.03	8.98
AuxIVA	6.04	8.00	**13.12**	4.07	6.65	**10.04**
GCAV-IVA(1)	7.00	10.20	11.73	5.47	10.20	8.76
GCAV-IVA(2)	7.37	10.33	12.23	5.60	10.30	8.90
GCAV-IVA(3)	7.32	10.40	12.20	5.55	10.36	8.75
GCAV-IVA(4)	7.39	10.27	12.37	5.71	10.41	9.03
GCAV-IVA(5)	**7.43**	**10.41**	12.31	**5.73**	**10.56**	8.93

5.3. Speech enhancement results

Table 2 and Table 3 summarize the speech enhancement results. The proposed GCAV-IVA method exceeded the conventional MPDR in terms of all criteria and achieved higher scores than AuxIVA in terms of SDRs and SIRs, which confirmed the advantage of the geometric constraints. Comparing the results achieved by system (1) with other systems, we found that constraining two channels led to higher enhancement performances, even in the situation where any interference speaker doesn't exist, i.e., 1-speaker case. The results also indicate that carefully tuned c_j was able to produce slightly higher SDR and SIR scores. Interestingly, the system exploiting interference DOA estimation outperformed the one using true DOAs. One possible reason is that, since the DOA estimate coming from the AuxIVA points out the direction including the most statistically independent components, suppressing that direction can result in a higher SIR.

6. CONCLUSIONS

In this paper, we proposed a geometrically constrained BSS method called GCAV-IVA, which combines IVA with a set of linear constraints restricting the far-field response of the demixing filter. We derived a convergence-guaranteed algorithm with the auxiliary function approach and showed the update rules of the parameter estimation by exploiting the idea introduced in the VCD method. A dual-microphone system, including GCAV-IVA for speech enhancement and AuxIVA for DOA estimation, was introduced and experimentally investigated. The experimental results confirmed that the proposed method outperformed the conventional MPDR beamformer and AuxIVA.

7. REFERENCES

[1] P. C. Loizou, "Speech enhancement: Theory and practice," *CRC press*, 2013.

[2] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1–3, pp. 21–34, 1998.

[3] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: An extension of ICA to multivariate components," in *Proc. ICA*, pp. 165–172, 2006.

[4] A. Hiroe, "Solution of permutation problem in frequency domain ICA using multivariate probability density functions," in *Proc. ICA*, pp. 601–608, 2006.

[5] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation with independent low-rank matrix analysis," in *Audio Source Separation*, pp. 125–155, 2018.

[6] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. WASPAA*, pp. 189–192, 2011.

[7] N. Ono, "Fast stereo independent vector analysis and its implementation on mobile phone," in *Proc. IWAENC*, pp. 1–4, 2012.

[8] N. Ono, "Blind source separation on iphone in real environment," in *Proc. EUSIPCO*, pp. 1–5, 2013.

[9] T. Taniguchi, N. Ono, A. Kawamura, and S. Sagayama, "An auxiliary-function approach to online independent vector analysis for real-time blind source separation," in *Proc. HSCMA*, pp. 107–111, 2014.

[10] M. Sunohara, C. Haruta, and N. Ono, "Low-latency real-time blind source separation for hearing aids based on time-domain implementation of online independent vector analysis with truncation of noncausal components," in *Proc. ICASSP*, pp. 216–220, 2017.

[11] Y. Liang, SM Naqvi, and J Chambers, "Overcoming block permutation problem in frequency domain blind source separation when using AuxIVA algorithm," *Electronics letters*, vol. 48, no. 8, pp. 460–462, 2012.

[12] A. Brendel, T. Haubner, and W. Kellermann, "Spatially informed independent vector analysis," *eprint arXiv:1907.09972*, 2019.

[13] L. C Parra and C. V Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE Trans. SAP*, vol. 10, no. 6, pp. 352–362, 2002.

[14] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," *IEEE Trans. ASLP*, vol. 14, no. 2, pp. 666–678, 2006.

[15] M. Knaak, S. Araki, and S. Makino, "Geometrically constrained independent component analysis," *IEEE Trans. ASLP*, vol. 15, no. 2, pp. 715–726, 2007.

[16] Y. Zheng, K. Reindl, and W. Kellermann, "Analysis of dual-channel ICA-based blocking matrix for improved noise estimation," *EURASIP journal on Advances in Signal Processing*, vol. 2014, no. 1, pp. 26, 2014.

[17] A. H Khan, M. Taseska, and E. AP Habets, "A geometrically constrained independent vector analysis algorithm for online source extraction," in *Proc. LVA/ICA*, pp. 396–403, 2015.

[18] Y. Mitsui, N. Takamune, D. Kitamura, H. Saruwatari, Y. Takahashi, and K. Kondo, "Vectorwise coordinate descent algorithm for spatially regularized independent low-rank matrix analysis," in *Proc. ICASSP*, pp. 746–750, 2018.

[19] J. Bourgeois and W. Minker, Eds., "Linearly constrained minimum variance beamforming," pp. 27–38, 2009.

[20] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. ASLP*, vol. 25, no. 4, pp. 692–730, 2017.

[21] D. R Hunter and K. Lange, "A tutorial on MM algorithms," *The American Statistician*, vol. 58, no. 1, pp. 30–37, 2004.

[22] S. Araki, S. Makino, Y. Hinamoto, R. Mukai, T. Nishikawa, and H. Saruwatari, "Equivalence between frequency-domain blind source separation and frequency-domain adaptive beamforming for convolutive mixtures," *EURASIP Journal on Applied Signal Processing*, vol. 2003, pp. 1157–1166, 2003.

[23] A. Lombard, T. Rosenkranz, H. Buchner, and W. Kellermann, "Multidimensional localization of multiple sound sources using averaged directivity patterns of blind source separation systems," in *Proc. ICASSP*, pp. 233–236, 2009.

[24] Y. Zheng, A. Lombard, and W. Kellermann, "An improved combination of directional BSS and a source localizer for robust source separation in rapidly time-varying acoustic scenarios," in *Proc. HSCMA*, pp. 58–63, 2011.

[25] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," *eprint arXiv:1804.04262*, 2018.

[26] 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.

[27] J. Thiemann, N. Ito, and E. Vincent, "DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments," Supported by Inria under the Associate Team Program VERSAMUS, June 2013.

[28] H. L. Van Trees, "Optimum array processing," John Wiley & Sons, 2002

[29] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans.ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.