RECURSIVE AND PARTIALLY SUPERVISED ALGORITHMS FOR SPEECH ENHANCEMENT ON THE BASIS OF INDEPENDENT VECTOR EXTRACTION

Tomáš Kounovský, Zbyněk Koldovský and Jaroslav Čmejla

Acoustic Signal Analysis and Processing Group,
Faculty of Mechatronics, Informatics and Interdisciplinary Studies,
Technical University of Liberec, Liberec, Czech Republic

ABSTRACT

This paper introduces a recursive variant of the recently proposed independent vector extraction algorithm suitable for on-line blind source separation. Two partially supervised variants are proposed and tested. Both variants exploit known direction of arrival (DOA) of the source of interest (SOI). The first variant uses a pre-separated output of a DOA-steered MPDR beamformer as a pilot component to ensure the extraction of the SOI. In the second variant, a geometrical constraint is imposed to ensure that the separating vector does not stray too far from the assumed direction of the SOI. Experiments using simulated and real-world recordings are demonstrated. Both supervised variants show improvements compared to the unsupervised algorithm in terms of convergence stability and separation accuracy.

Index Terms— Blind Audio Source Separation, Speech Enhancement, Independent Component Analysis, Independent Vector Analysis, Independent Vector Extraction

1. INTRODUCTION

Independent Component Analysis (ICA) is a popular method used for blind separation of audio signals [1]. Typically, it is applied to the signals obtained by an array of microphones that are transformed by a Short-term Fourier Transform (STFT) [2]. In the kth frequency bin, $k = 1, \ldots, K$, where K is the frequency resolution of the STFT, the signal mixture can be described through

$$\mathbf{X}^k = \mathbf{A}^k \mathbf{S}^k, \tag{1}$$

where $\mathbf{S}^k$ and $\mathbf{X}^k$ denote $d \times T$ matrices of the original and mixed signals within the kth frequency bin, respectively; the rows of the matrices correspond to signals while the columns correspond to STFT frames (time). $\mathbf{A}^k$ is the $d \times d$ non-singular mixing matrix. In ICA, the same number of microphones and of the original sources is assumed, nevertheless, ICA methods may be successfully deployed also in situations that do not obey the mixing model exactly; see, e.g., [3].

Since ICA is applied independently on each frequency band, the separated frequency components of the signals have to be clustered as their order is random [4]. Another popular solution to this permutation problem is based on Independent Vector Analysis (IVA) [5]. Here, the mixtures $\mathbf{X}^1, \ldots, \mathbf{X}^K$ are separated jointly by forcing frequency components related to the same source to be as dependent as possible; an on-line algorithm for audio source separation was proposed, e.g., in [6].

When extracting only one particular source of interest (SOI), which is the case for most speaker enhancement applications, ICA and IVA are less effective in the sense that they separate d independent signals. Therefore, we have recently proposed an approach called *Independent Vector Extraction* (IVE) where the mixing model (1) is re-parameterized only to extract the SOI [7]. Here, a simple gradient algorithm, referred to as OGIVE, has been introduced.

The goal of this work is to design a recursive version of OGIVE that is capable to extract the SOI in an on-line regime. Here, a crucial problem resides in the convergence stability: It is necessary to control that the algorithm extracts the SOI, not any other source. To this end, we apply preconditioning of the input signals and consider two modifications of OGIVE performing partial supervision of the algorithm. The first one, referred to as S-OGIVE, is piloted by the output of a Minimum Power Distortionless Beamformer (MPDR), which is directed towards the assumed SOI location [8]. In the second variant, GC-OGIVE, a geometrical penalty term is imposed as in [9]; see also [10].

In the following section, the IVE problem and the OGIVE algorithm are briefly described. Section 3 is devoted to the proposal of the recursive variant of OGIVE and to its supervised variants. Section 4 is devoted to an experimental validation of the approaches, and Section 5 concludes the paper.

This work was supported by the United States Department of the Navy, Office of Naval Research Global, through Project No. N62909-18-1-2040, and by the Student Grant Scheme 2018 project of the Technical University in Liberec.

 978-1-5386-8151-0/18/$31.00

2. INDEPENDENT VECTOR EXTRACTION

Without any loss of generality, let the SOI correspond to the first original source in (1). For the problem of blind extraction, it is suitable to describe the mixing model by

$$\mathbf{X}^k = \mathbf{a}^k \mathbf{s}^k + \mathbf{Y}^k, \qquad k = 1, \ldots, K, \tag{2}$$

where, in view of the mixing model (1), $\mathbf{a}^k$ is the first column of $\mathbf{A}^k$, which will be referred to as the *mixing vector*, $\mathbf{s}^k$ denotes the first row of $\mathbf{S}^k$, a row vector containing the samples of the SOI, and $\mathbf{Y}^k$ denotes the other signals within the mixture.

IVE is based on an algebraic model of $\mathbf{X}^k$ having only two parameter $d \times 1$ vectors $\mathbf{a}^k$ and $\mathbf{w}^k$, where the latter one is referred to as *separating vector* as it extracts the kth output $\widehat{\mathbf{s}}^k = (\mathbf{w}^k)^H \mathbf{X}^k$. Next, there is a statistical model of the signals, which is based on the assumption that $\mathbf{s}^k$ and $\mathbf{Y}^k$ are independent. While $\mathbf{Y}^k$ is assumed to be Gaussian, $\mathbf{s}^1, \ldots, \mathbf{s}^K$ are assumed to have a joint non-Gaussian probability density function (pdf), which will be denoted by $p(\cdot)$. This is similar to the model used in IVA [11], however, IVA aims at decomposing $\mathbf{Y}^k$ to the other independent vector components; for further details on IVE see [7].

OGIVE is a gradient-update algorithm that performs maximization of a contrast function $\mathcal{J}(\cdot)$ based on the maximum quasi-likelihood principle [7]. In every iteration, OGIVE performs a small step in the direction of the gradient subject to the separating vector $\mathbf{w}^k$, i.e., $\mathbf{w}^k \leftarrow \mathbf{w}^k + \mu \frac{\partial \mathcal{J}}{\partial \mathbf{w}^{k\,H}}$, where μ is a step-length parameter,

$$\frac{\partial \mathcal{J}}{\partial \mathbf{w}^{k\,H}} = \mathbf{a}^k - \frac{1}{N} \mathbf{X}^k \phi^k\big(\widehat{\mathbf{s}}^1, \ldots, \widehat{\mathbf{s}}^K\big)^T, \tag{3}$$

where the mixing vector $\mathbf{a}^k$ is linked with $\mathbf{w}^k$ through

$$\mathbf{a}^k = \frac{\widehat{\mathbf{C}}^k_{\mathbf{x}} \mathbf{w}^k}{\mathbf{w}^{k\,H} \widehat{\mathbf{C}}^k_{\mathbf{x}} \mathbf{w}^k}, \tag{4}$$

where $\widehat{\mathbf{C}}^k_{\mathbf{x}} = \mathbf{X}^k (\mathbf{X}^k)^H / N$ is the sample covariance matrix of $\mathbf{X}^k$. This constraint is necessary to avoid local extrema of the contrast function for which $\mathbf{w}^k$ and $\mathbf{a}^k$ do not correspond to the same source [12]. Next, the nonlinearity ϕ^k in (3) is defined as

$$\phi^k(\xi^1, \ldots, \xi^K) = -\frac{\partial \log f(\xi^1, \ldots, \xi^K)}{\partial \xi^k}, \tag{5}$$

which is the kth score function related to a model density function $f(\cdot)$ that is a substitute of the density $p(\cdot)$, which is not known in the blind scenario. The fact that ϕ^k is a function of the outputs for all $k = 1, \ldots, K$ means that the frequency components of the extracted source are extracted jointly.

OGIVE proceeds as follows. The separating vectors are set to initial values. In each iteration, the mixing vectors $\mathbf{a}^k$, $k = 1, \ldots, K$, are computed according to (4), and the current outputs are evaluated as $\widehat{\mathbf{s}}^k = (\mathbf{w}^k)^H \mathbf{X}^k$. Then, the nonlinearities $\phi^k(\widehat{\mathbf{s}}^1, \ldots, \widehat{\mathbf{s}}^K)$ are computed and normalized so that $\widehat{\mathbf{s}}^k \phi^k(\widehat{\mathbf{s}}^1, \ldots, \widehat{\mathbf{s}}^K)^T / N = 1$, $k = 1, \ldots, K$. The normalization is necessary to ensure that the desired solution is a stationary point of the algorithm. Finally, the separating vectors are updated as $\mathbf{w}^k \leftarrow \mathbf{w}^k + \mu (\mathbf{D}^k)^H \mathbf{D}^k \frac{\partial \mathcal{J}}{\partial \mathbf{w}^{k\,H}}$, $k = 1, \ldots, K$, where $\mathbf{D}^k$ denotes a preconditioning matrix for the kth frequency[1]. The algorithm stops when the norm of the gradient is smaller than a selected tolerance for all k. More details are provided in [7].

3. PROPOSED RECURSIVE ALGORITHMS

In this section, we propose recursive variants of OGIVE capable of processing signals continuously in time. Particular descriptions of supervised modifications will be given in separate subsections below.

The recursive OGIVE performs a single update of the separating vectors with every new (the tth) STFT frame using a batch of Q previous frames in the following steps. For every $k = 1, \ldots, K$,

1. the sample covariance matrix of input data is updated according to

$$\widehat{\mathbf{C}}^k_{\mathbf{x}} \leftarrow (1 - \lambda) \widehat{\mathbf{C}}^k_{\mathbf{x}} + \lambda \mathbf{X}^k_t (\mathbf{X}^k_t)^H / Q, \tag{6}$$

   where $\lambda \in [0, 1]$ is a learning parameter, and $\mathbf{X}^k_t$ denotes the $d \times Q$ matrix of new STFT frames for the kth frequency;

2. the mixing vectors are updated using (4);

3. current outputs are computed as $\widehat{\mathbf{s}}^k_t = (\mathbf{w}^k)^H \mathbf{X}^k_t$;

4. the normalization constants are updated as

$$\nu^k \leftarrow (1 - \lambda) \nu^k + \lambda \widehat{\mathbf{s}}^k_t \phi^k(\widehat{\mathbf{s}}^1_t, \ldots, \widehat{\mathbf{s}}^K_t)^T / Q; \tag{7}$$

5. the separating vectors are updated as

$$\mathbf{w}^k \leftarrow \mathbf{w}^k + \mu^k (\mathbf{D}^k)^H \mathbf{D}^k \Delta^k,$$

   where $\Delta_k = \mathbf{a}^k - \mathbf{X}^k_t \phi^k(\widehat{\mathbf{s}}^1_t, \ldots, \widehat{\mathbf{s}}^K_t)^T / \nu^k / Q$, and μ^k and $\mathbf{D}^k$ denote, respectively, the step-length parameter and the preconditioning matrix selected for the kth frequency.

The algorithm is initialized with $\widehat{\mathbf{C}}^k_{\mathbf{x}} = \mathbf{I}_d$, where $\mathbf{I}_d$ is the $d \times d$ identity matrix, and with $\nu^k = 1$.

[1] The preconditioned version of OGIVE has been introduced in a paper submitted to this conference [13], which is available for reviewers at https://github.com/tomaskounovsky/iwaenc2018

3.1. Piloted Algorithm: S-OGIVE

The idea of pilot-assisted separation is that some information about the SOI is injected into the nonlinearity in (5), forcing the extraction algorithm to keep the convergence to the SOI [8]. Our choice is

$$\phi^k(\widehat{s}^1, \ldots, \widehat{s}^K, P) = \frac{(\widehat{s}^k)^*}{\sqrt{(1-\beta)\sum_{k=1}^K |\widehat{s}^k|^2 + \beta\gamma|P|^2}}. \tag{8}$$

where β is a parameter controlling the influence of the pilot component P, and γ is a scale-correction parameter.

In this work, the pilot component P is computed from the pre-extracted source by the MPDR beamformer, namely,

$$P = \sqrt{\sum_{k=1}^{K} |(\mathbf{w}^k_{\text{MPDR}})^H X^k|^2}, \tag{9}$$

where

$$\mathbf{w}^k_{\text{MPDR}} = \frac{\widehat{\mathbf{C}}^k_{\mathbf{x}}\mathbf{g}^k}{\mathbf{g}^{k\,H}\widehat{\mathbf{C}}^k_{\mathbf{x}}\mathbf{g}^k}, \tag{10}$$

where $\mathbf{g^k}$ denotes the steering direction vector [14].

3.2. Geometrically Constrained Algorithm: GC-OGIVE

The BSS techniques using geometrical constraint were proposed in [10]; an on-line IVA algorithm of this class was proposed in [9]. The idea is to constrain the Euclidean angle between $\mathbf{w}^k$ and $\mathbf{g^k}$. This is done by adding a penalty term to the contrast function. Then, the update rule for $\mathbf{w}^k$ is

$$\mathbf{w}^k \leftarrow \mathbf{w}^k + \mu^k(\mathbf{D}^k)^H\mathbf{D}^k\Delta^k + \lambda_{\text{GC}}\nabla\mathbf{w}^k, \tag{11}$$

where the penalty term $\nabla\mathbf{w}^k$ is computed as

$$\nabla\mathbf{w}^k = \frac{(\cos\Theta^k - 1)\left(\mathbf{g}^k - \frac{\mathbf{w}^k}{||\mathbf{w}^k||^2}\Re\left\{\mathbf{w}^{k\,H}\mathbf{g}^k\right\}\right)}{||\mathbf{w}^k|| \cdot ||\mathbf{g}^k||^2}, \tag{12}$$

where $\Theta^k = \arccos\big(\Re\{\mathbf{w}^{k\,H}\mathbf{g}^k\}/(||\mathbf{w}^k|| \cdot ||\mathbf{g}^k||)\big)$; $\Re\{\cdot\}$ denotes the real part of the argument.

4. EXPERIMENTS

We conducted two experiments: a simulated one in a semi-static environment and one with a real-world recording. In the first experiment, the image-source model was used to simulate the setup shown in Fig. 1, in a 5x5x2.5m room. The SOI is a 60 s long male speech played from the center the target area, which is 1 m distant from the microphone array. The interfering signal is a female speech played from positions 1, 2 and 3, respectively, each for 20 seconds. Four omnidirectional microphones were used in a linear array with 6.66 cm spacing. The reverberation time of the room was set to $T_{60} = 150$ ms, and the signals were mixed with a Signal-to-Interference Ratio (SIR) $= 0$ dB; the sampling rate was 16 kHz; the length of FFT in the STFT was 512 samples with 128 overlap; in the recursive algorithm we set Q corresponding to 2 seconds.

The extraction accuracy is evaluated in terms of SIR and SIR improvement. To compare, the performance of the MVDR beamformer (using oracle mixing vector and interference cov. matrix) is computed, which provides a performance bound. We also show the extraction performance of the non-oracle MPDR beamformer used to supervise S-OGIVE.

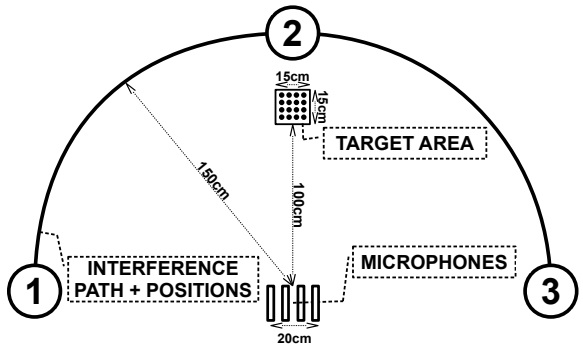

Fig. 1. Illustration of the experimental setup.

The setup of the experiment was designed to consider three situations. For position 1, the SOI and the interferer have different angular positions, and all methods, being initialized by the true direction vector, should be capable of extracting the SOI. For position 2, the angular positions of the sources are the same, which makes the blind extraction much more difficult, especially when no additional information is given. In position 3, we check the ability of the methods to solve the global permutation problem by re-initializing them towards the interferer, and we test whether they focus back on the SOI.

Fig. 2 shows the achieved SIR improvement. Mean segmental SIR and SIR improvement for each position can be found in Table 1. The results show that all methods managed to extract the SOI for position 1, where the OGIVE variants yield comparable performance; S-OGIVE and OGIVE perform slightly better than GC-OGIVE, as the performance of the latter method is limited due to the constraint. In position 2, the performance of all methods is considerably lower. MPDR does not improve the SIR since the directional vector is the same for both the SOI and the interference. OGIVE-based methods improve the SIR by 0.2-1.56 dB, with the best performance being provided by the unsupervised OGIVE.

In position 3, OGIVE keeps tracking the interferer and focuses the SOI only during the last 5 s. This points to its lower stability in source tracking. S-OGIVE steers back to the SOI in about 5 seconds and outperforms OGIVE and GC-OGIVE from that point. GC-OGIVE focuses the SOI within 2 seconds, which points to the effectiveness of the constraint.

Table 1. Mean segmental SIR (SIR improvement) for simulated experiment (in dB)

Method	Position 1	2	3
Input	-1.89 (-)	1.79 (-)	0.275 (-)
MVDR	12.2 (14.8)	12.4 (10.4)	13.5 (13.1)
MPDR	5.81 (8.34)	1.75 (-0.23)	5.87 (5.45)
OGIVE	4.8 (7.33)	3.53 (1.56)	-4.05 (-4.47)
S-OGIVE	5.21 (7.74)	3.07 (1.09)	5.06 (4.64)
GC-OGIVE	4.48 (7.01)	2.25 (0.275)	4.06 (3.63)

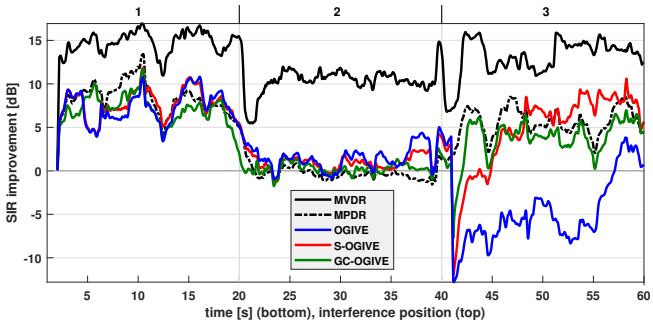

Fig. 2. SIR achieved in a simulated experiment. Detailed parameter settings: $\mu = 1000$, $\lambda = 0.5$, $\beta = 0.5$, $\lambda_{GC} = 10$.

The second experiment is realized in an office room having difficult reverberation time ($T_{60} \approx 700$ ms). The SOI is a male speech played back by a loudspeaker slowly moving within the target area in Fig 1. For the positions within the target area, a set of DOAs is available for the supervised algorithms. The interferer is represented by a female utterance played back by another loudspeaker initially located in position 1 where it stays static for about 20 s. It is then slowly moved to position 2 where it again stays still for 20 s. This way the interferer visits position 3 and goes back to positions 2 and 1, where the cycle repeats. Both sources were separately recorded using a linear array of four directional microphones; the recordings were mixed with initial SIR of 0 dB; the total length of the recordings is 8 min.

The algorithms were set to the same parameters as in the first experiment. The first 200 seconds of the results are shown in Fig. 3, the full graph can be seen at the previously mentioned github page; mean segmental SIRs are summarized in Table 2.

Whenever the interferer stays in position 2, the extraction performance of all methods drops down as the sources have very close angular positions. The performances in positions 1 and 3 are better. MPDR beamformer and GC-OGIVE show similar results across the entire recording. The unsupervised OGIVE and S-OGIVE perform similarly, outperforming MPDR and GC-OGIVE in almost all instances. We were unable to showcase the better convergence stability of the supervised methods on this recording because the global permutation problem did not occur. Overall, the results correspond to those from the first experiment, albeit at a lower scale due to the difficult conditions of the scenario.

Table 2. Mean segmental SIR/SIR improvement (dB) achieved in an office room

Method	SIR	SIR imp.
Input	0.42	-
MVDR	6.86	6.75
MPDR	0.67	0.56
OGIVE	2.69	2.58
S-OGIVE	2.66	2.54
GC-OGIVE	1.32	1.20

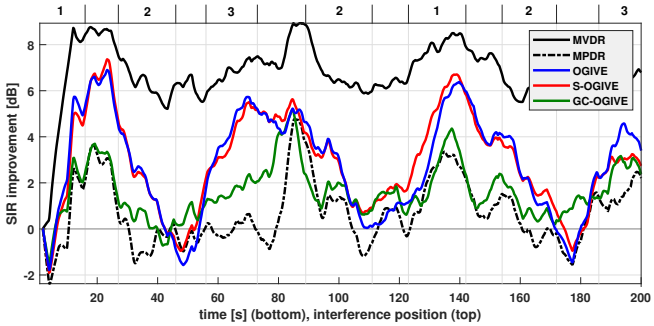

Fig. 3. SIR improvement achieved in an office room. Detailed settings: $\mu = 1000$, $\lambda = 0.5$, $\beta = 0.5$, $\lambda_{GC} = 10$.

5. CONCLUSIONS

In this study, we have proposed a new recursive variant of an algorithm for independent vector extraction capable of extracting the source of interest in an on-line regime. Two partially supervised variants of the algorithm were also presented. These variants exploit the knowledge of the direction of arrival of the source of interest to stabilize the extraction. A geometrically constrained version of the recursive OGIVE algorithm was shown to provide fast adaptation to the movements of the focused source, which was, however, payed for by a limited extraction performance. OGIVE piloted by a pre-separated output of an MPDR beamformer showed a slower steering ability but it provided superior extraction performance both in simulated and real conditions.

The current versions of OGIVE show significantly lower performance compared to an MVDR beamformer, which points to the fact that there is room for improvements. State-of-the-art BSS algorithms developed for audio source separation using advanced source modeling are capable to achieve better performance; see, e.g., [15, 16, 17]. (In the final version of this paper we plan to include the comparison with [9].) Nevertheless, this work gives a proof of concept that methods for independent vector extraction provide relevant alternatives for speech enhancement technologies.

6. REFERENCES

[1] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*, Independent Component Analysis and Applications Series. Elsevier Science, 2010.

[2] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, pp. 21–34, 1998.

[3] S. Araki, S. Makino, H. Sawada, and R. Mukai, "Blind separation of more speech than sensors using time-frequency masks and ICA," in *Proc. of 2004 NTT Workshop on Communication Scene Analysis*, Apr. 2004, vol. 1, p. AU4.

[4] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 530–538, Sept. 2004.

[5] Intae Lee, Taesu Kim, and Te-Won Lee, "Independent vector analysis for convolutive blind speech separation," in *Blind speech separation*, pp. 169–192. Springer, 2007.

[6] T. Kim, "Real-time independent vector analysis for convolutive blind source separation," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 57, no. 7, pp. 1431–1438, July 2010.

[7] Z. Koldovský and P. Tichavský, "Gradient algorithms for complex non-gaussian independent component/vector extraction," Mar. 2018, arXiv:1803.10108 [eess.SP].

[8] 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.

[9] Affan H. Khan, Maja Taseska, and Emanuël A. P. Habets, *A Geometrically Constrained Independent Vector Analysis Algorithm for Online Source Extraction*, pp. 396–403, Springer International Publishing, Cham, 2015.

[10] L. C. Parra and C. V. Alvino, "Geometric source separation: merging convolutive source separation with geometric beamforming," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 6, pp. 352–362, Sep 2002.

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

[12] Zbyněk Koldovský, Petr Tichavský, and Václav Kautský, "Orthogonally constrained independent component extraction: Blind MPDR beamforming," in *Proceedings of European Signal Processing Conference*, Sept. 2017, pp. 1195–1199.

[13] Jaroslav Čmejla and Zbyněk Koldovský, "Multichannel speech enhancement based on independent vector extraction," in *submitted to this conference*, 2018.

[14] Harry L. Van Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*, John Wiley & Sons, Inc., 2002.

[15] Y. Mitsui, D. Kitamura, S. Takamichi, N. Ono, and H. Saruwatari, "Blind source separation based on independent low-rank matrix analysis with sparse regularization for time-series activity," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 21–25.

[16] 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 *2014 4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, May 2014, pp. 107–111.

[17] M. Taseska and E. A. P. Habets, "Blind source separation of moving sources using sparsity-based source detection and tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 657–670, March 2018.