INDEPENDENT VECTOR ANALYSIS ASSISTED ADAPTIVE BEAMFOMRING FOR SPEECH SOURCE SEPARATION WITH AN ACOUSTIC VECTOR SENSOR

Yichen Yang[1], *Xianrui Wang*[1], *Wen Zhang*[1], *Jingdong Chen*[1],*Chaoyu Shi*[2], *Mengyao Zhu*[2], *and Chunjian Li*[2]

[1]Center of Intelligent Acoustics and Immersive Communications,
School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an, China
[2] Audio Department, Huawei CBG, Shanghai, China

ABSTRACT

Acoustic vector sensor (AVS), as a compact sensor with the capability of forming a frequency-invariant spatial beampattern over the 3D space, has potential in source separation. A straightforward way to achieve source separation with AVS is through adaptive beamforming. Such a method requires the direction-of-arrival (DOA) information, which is challenging to estimate accurately in reverberant environments. To circumvent this issue, we present a framework jointly implementing adaptive beamforming and independent vector analysis (IVA). Different from the conventional beamforming, the presented method only require rough DOA estimation for initialization. It iteratively refines the estimates of source DOA and signal statistics. The proposed method has great advantages of improving source separation performance and enhancing DOA estimation accuracy. Simulations demonstrate the properties of the developed method.

Index Terms— Adaptive beamforming, independent vector analysis, acoustic vector sensor

1. INTRODUCTION

Speech enhancement is of great importance for audio processing as speech signals recorded by microphones are normally contaminated by reverberation and interference. Various methods have been developed over the last few decades [1], among which source separation is one important category of methods that have to be used for speech enhancement in environments where there exist multiple sources simultaneously [2].

Adaptive beamforming has been widely used for source separation when the microphone array geometry is known [1, 3, 4]. Generally speaking, adaptive beamforming needs to estimate certain parameters and signal statistics, e.g., DOA of each source, covariance matrices of noise and interferences. Many efforts have been devoted to the estimation of such parameters and statistics in adversarial environments including the latest ones that attempt to estimate those parameters using deep neural network (DNN) [5, 6, 7]. However, the estimation performance often suffers from significant degradation in practical environments regardless of whether traditional or deep methods are used.

In contrast to adaptive beamforming, blind source separation (BSS) achieves source signal separation from observations based on independent component analysis (ICA) [9, 10], which requires little or no *a priori* information [8]. The multivariate extension of ICA, i.e., IVA, has been successfully applied to BSS of speech signals [11, 12] as speech signals are broadband in nature and IVA can help circumvent the problem of permutation, which is inherent to the ICA based algorithms. To improve the robustness and accelerate the convergence of IVA, the auxiliary function-based IVA (AuxIVA) is proposed [13]. In order to further improve the separation performance, geometry constrained separation methods [14, 15, 16, 17] have been developed, which impose the beamforming constraint on the demixing filters of BSS. However, given that the DOA estimates easily become biased in real multi-source environments, the optimal demixing filter is difficult if not impossible to achieve. Some other methods attempt to improve the separation performance by combining BSS algorithms and beamforming. For example, in [18, 19], null beamforming is introduced in the frequency-domain ICA and in [20, 21] beamforming is cascaded directly with BSS. The work of [21] presents a framework that unifies differential beamforming and independent low-rank matrix analysis (ILRMA). This framework, however, faces two challenging issues: 1) the DOA of source signals must be accurately estimated, and 2) the white noise gain of the differential beamformers at low frequency bins should be sufficiently high.

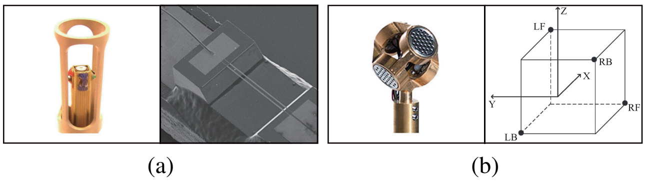

Fig. 1. (a) 3D intensity probe (left) and scanning electron microscope photo (right) of *Microflown* [27]. (b) *SENNHEISER AMBEO VR mic* (left) and corresponding microphone array setup (right).

Recently, AVS [22] has been investigated for source separation [23, 24, 25, 26]. Given the fact that it can simultaneously record the sound pressure and particle velocity at the same spatial point, AVS can form requency-invariant beampatterns and steer the main lobe towards any direction, so the speech distortion after enhancement is generally small. This make AVS a good choice for speech source separation in the 3D space. Figure 1 shows several types of AVS sensors.

In this paper, we present an approach with AVS that jointly optimizes adaptive beamforming and AuxIVA. It performs speech separation in an iterative manner.IVA is firstly used as a parameter estimator, whose outputs are subsequently used by adaptive beamforming to extract target signals. The beamforming outputs are then fed to

This work was supported in part by the National Key Research and Development Program of China under Grant No. 2018AAA0102200 and in part by the Key Program of National Science Foundation of China (NSFC) under Grant No. 61831019.

978-1-6654-6867-1/22/$31.00 ©2022 IEEE

IVA to update statistical estimation. Simulation results demonstrate that this presented method outperforms two widely studied baseline methods.

2. SIGNAL MODEL AND PROBLEM FORMULATION

Let us consider the multi-source scenario under reverberant conditions, when the window length in the short-time Fourier transform (STFT) is longer than that of the room impulse responses (RIRs), an instantaneous mixture model in the STFT domain can be written as

$$\mathbf{x}(t,f)=\sum_{k=1}^{K}\mathbf{a}_k(f)s_k(t,f)+\mathbf{v}(t,f), \tag{1}$$

where $s_k(t,f)$, $\mathbf{x}(t,f)$, and $\mathbf{v}(t,f)$ denote, respectively, the source signal, and the observation and additive noise signals, $f=1,\dots,F$ and $t=1,\dots,T$ are the indexes of the frequency bins and time frames, respectively, k and K denote, respectively, the index and the number of sources, $\mathbf{a}_k(f)\in\mathbb{C}^{M\times 1}$ denotes a set of the RIRs and M denotes the number of microphone channels. Now the problem of source separation is to extract the source signals, i.e., $s_k(t,f)$, given the microphone signals $\mathbf{x}(t,f)$.

3. SOURCE SEPARATION ALGORITHMS

3.1. Adaptive beamforming

The source signal extraction through beamforming can be expressed as

$$z(t,f)=\mathbf{h}^{\mathrm{H}}(f)\mathbf{x}(t,f), \tag{2}$$

where $\mathbf{h}(f)\in\mathbb{C}^{M\times 1}$ and $z(t,f)$ represent the spatial filter and the output signal for the source signal of interest, and $^{\mathrm{H}}$ denotes the conjugate-transpose operator. For adaptive beamforming, given that the covariance matrix of the noise and interference is difficult to estimate accurately, the minimum power distortionless response (MPDR) beamformer is often used to extract the desired signal from the look direction with the minimum output power, for which the cost function is formulated as

$$\min_{\mathbf{h}(f)}\mathbf{h}^{\mathrm{H}}(f)\mathbf{\Phi}_{\mathbf{xx}}(f)\mathbf{h}(f)\text{ subject to }\mathbf{h}^{\mathrm{H}}(f)\mathbf{d}(\Theta_k)=1, \tag{3}$$

where $\mathbf{\Phi}_{\mathbf{xx}}(f)$ is the covariance matrix of the microphone signals and $\mathbf{d}(\Theta_k)$ is the steering vector of AVS. As shown in [], the steering vector pointing at azimuth angle ϕ_k and elevation angle θ_k can be denoted as

$$\mathbf{d}(\Theta_k)=\left[1,\quad \cos\theta_k\cos\phi_k,\quad \cos\theta_k\sin\phi_k,\quad \sin\theta_k\right]^{\mathrm{T}}, \tag{4}$$

where $^{\mathrm{T}}$ stands for the transpose operator.

$$\mathbf{h}_{\Theta_k}(f)=\frac{\mathbf{\Phi}_{\mathbf{xx}}^{-1}(f)\mathbf{d}(\Theta_k)}{\mathbf{d}^{\mathrm{H}}(\Theta_k)\mathbf{\Phi}_{\mathbf{xx}}^{-1}(f)\mathbf{d}(\Theta_k)}, \tag{5}$$

When the DOAs of the sources can be accurately estimated, MPDR can achieve good separation performance [26]; however, in multiple-source scenarios, the estimates of DOAs and covariance matrices are generally biased, especially in reverberant and noisy environments [28][29], leading to significant degradation in beamforming performance. In this work, we propose to incorporate BSS as a pre-processing for parameter estimation.

3.2. Independent vector analysis

In IVA, the demixing system can be described as

$$\hat{\mathbf{s}}(t,f)=\mathbf{W}(f)\mathbf{x}(t,f), \tag{6}$$

where $\mathbf{W}(f)\in\mathbb{C}^{K\times M}$ is the demixing matrix and $\hat{\mathbf{s}}(t,f)\in\mathbb{C}^{K\times 1}$ is the separated signals. The demixing matrix $\mathcal{W}=\{\mathbf{W}(f)\}_{f=1}^{F}$ can be estimated by minimizing the following cost function

$$J(\mathcal{W})=\sum_{t=1}^{T}\sum_{k=1}^{K}\left[G\left(\hat{\mathbf{s}}_k(t)\right)\right]-2J\sum_{f=1}^{F}\log|\det\mathbf{W}(f)|, \tag{7}$$

where $G(\hat{\mathbf{s}}_k(t))=-\log p(\hat{\mathbf{s}}_k(t))$ is the contrast function, which can be expressed for a typical spherical multivariate distribution as

$$G\left[\hat{\mathbf{s}}_k(t)\right]=G_R(r_k(t)), \tag{8}$$

$$r_k(t)=\|\hat{\mathbf{s}}_k(t)\|_2=\sqrt{\sum_f|\hat{s}_k(t,f)|^2}, \tag{9}$$

where $\|\cdot\|_2$ represents the L_2 norm and $G_R(r_k(t))$ represents a real-valued function.

While it has demonstrated great potential, IVA faces two main limitations: 1) all parameters are estimated blindly though the geometry of the array is often given as the *a priori* information; 2) IVA extracts the source mainly by placing a null towards the interference direction [30], which has limited separation performance in reverberation conditions.

4. PROPOSED IVA-ASSISTED ADAPTIVE BEAMFORMING SYSTEM

In this work, we propose the following IVA-assisted adaptive beamforming method, as illustrated in Fig. 2. After IVA, the multichannel back-projection (BP) [31] is implemented to deal with the issue of scale ambiguity and recover the spatial information. By incorporating the spatial information, the adaptive beamformer is able to further improve the separation performance.

To obtain the demixing matrix of IVA, the auxiliary function-based optimization [13] is used and the auxiliary function is described by omitting the constant term as

$$Q(\mathcal{W},\mathcal{V})=\frac{1}{2}\sum_{f,k=1}^{F,K}\mathbf{w}_k^{\mathrm{H}}(f)\mathbf{V}_k(f)\mathbf{w}_k(f)-\sum_{f=1}^{F}\log|\det\mathbf{W}(f)|, \tag{10}$$

where $\mathcal{V}=\{\mathbf{V}_k(f)\}_{f=1,k=1}^{F,K}$ is the auxiliary variable. Using the auxiliary function, the demixing matrix can be optimized by updating the auxiliary variables, i.e.,

$$r_k(t)=\|\mathbf{y}_k(t)\|_2, \tag{11}$$

$$\mathbf{V}_k(f)=\mathbb{E}\left[\frac{G_R'(r_k(t))}{r_k(t)}\mathbf{x}_k(t)\mathbf{x}_k^{\mathrm{H}}(t)\right], \tag{12}$$

and updating the demixing matrix, i.e.,

$$\mathbf{w}_k(f)\leftarrow\left[\mathbf{W}(f)\mathbf{V}_k(f)\right]^{-1}\mathbf{e}_k, \tag{13}$$

$$\mathbf{w}_k(f)\leftarrow\mathbf{w}_k(f)/\sqrt{\mathbf{w}_k^{\mathrm{H}}(f)\mathbf{V}_k(f)\mathbf{w}_k(f)}, \tag{14}$$

where $\mathbb{E}[\cdot]$ denotes the expectation operator. The output of IVA is

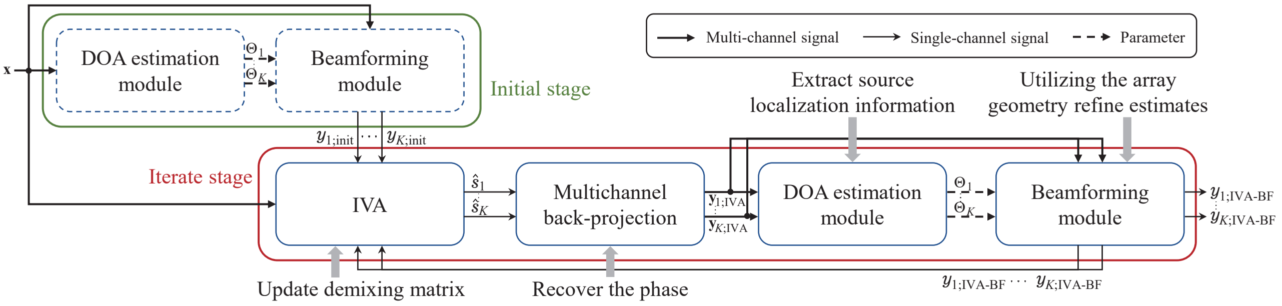

Fig. 2. Flowchart of the proposed method with joint optimization.

Input:	$\mathbf{x}(t,f)$
Initialize:	$\mathcal{W}$, $\boldsymbol{\Theta}$, $\mathbf{y}_{\text{init}}(t,f)$
Repeat	
Step 1:	Calculate the demixing matrix $\mathcal{W}$ of IVA using (11)–(14).
Step 2:	Calculate the multi-channel output of the IVA $\mathbf{Y}_{\text{IVA}}(t,f)$ using (6) and (15).
Step 3:	Calculate the optimal filter $\mathbf{H}_{\hat{\Theta}}$ and the output $\mathbf{y}_{\text{IVA-BF}}(t,f)$ of the MPDR using (16)–(18).
Until convergence	

Table 1. The proposed algorithm.

then

$$\mathbf{y}_{k;\text{IVA}}(t,f) = \mathbf{W}(f)^{-1}[\mathbf{e}_k \circ \hat{\mathbf{s}}(t,f)], \tag{15}$$

where $\mathbf{e}_k$ is the unit vector with the kth element being 1 and others being 0, and $\circ$ represents the Hadamard product.

The MPDR beamformer can be obtained as

$$\hat{\mathbf{\Phi}}_{\mathbf{yy}}(f) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{y}_{k;\text{IVA}}(t,f)\mathbf{y}_{k;\text{IVA}}^{\text{H}}(t,f), \tag{16}$$

$$\mathbf{h}_{\hat{\Theta}_k}(f) = \frac{\hat{\mathbf{\Phi}}_{\mathbf{yy}}^{-1}(f)\mathbf{d}(\hat{\Theta}_k)}{\mathbf{d}^{\text{H}}(\hat{\Theta}_k)\hat{\mathbf{\Phi}}_{\mathbf{yy}}^{-1}(f)\mathbf{d}(\hat{\Theta}_k)}, \tag{17}$$

where $\hat{\mathbf{\Phi}}_{\mathbf{yy}}(f)$ is the averaged covariance matrix of the IVA output in (15), and the DOA of kth source $\hat{\Theta}_k$ is estimated in the beamforming stage.

The final output of the proposed system is

$$y_{k;\text{IVA-BF}}(t,f) = \mathbf{h}_{\hat{\Theta}_k}^{\text{H}}(f)\mathbf{y}_{k;\text{IVA}}(t,f). \tag{18}$$

The proposed algorithm is summarized in Table 1.

5. EXPERIMENTAL EVALUATION

5.1. Experimental setup

In this section, we study the performance of the proposed method and compare it with two baseline methods in noisy and reverberant environments. A room of size (8×6×3) m is simulated, where an AVS is located at (4, 2, 1.5) m. The sources are set randomly in the 3D space in the room with the minimum angular septation of 30° between sources and the distance between the microphone and each source is randomly assigned to be larger than 1 m. Fifty two-source mixtures are simulated by convolving the speech randomly picked from the CMU ARCTIC dataset [32] with the RIRs generated by composing three bidirectional microphones along each coordinate axis in Cartesian coordinate system and one omnidirectional microphones at the origin. using the image-source method [33]. The reverberation time T_{60} varies from 200 to 400 ms and the white Gaussian noise is added with SNR of 30 dB. All signals are sampled at 16 kHz. The STFT is implemented with a Hanning window of length 256 ms, and the window shift is 64 ms.

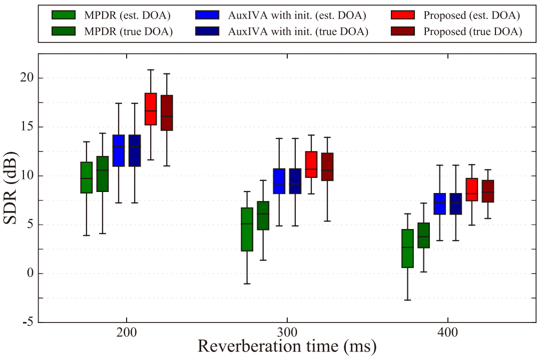

Fig. 3. The average SDR (dB) of the studied methodsin different reverberation conditions.

The Signal-to-Distortion Ratio (SDR) and signal-to-interferences ratio (SIR) [34] are used to evaluate the separation performance, and the perceptual evaluation of speech quality (PESQ) is used to evaluate the speech quality. The reference signals are obtained by convolving the speech signals with the RIRs truncated at 32 ms.

The separation performance of the proposed method is compared with two baseline methods: the MPDR beamformer with an input of the multiple source DOA estimation [26] and the AuxIVA algorithm initialized with the MPDR beamforming [20]. A hyperparameter-free algorithm based on the orthogonal constraints [35] is used in the over-determined situation ($M > K$) in our simulations. The multiple signal classification (MUSIC) algorithm is used for DOA estimation in the initial stage.

5.2. Results and discussion

The results of average SDR under different reverberation times are shown in Fig. 3. Compared with the two baseline methods, the proposed one achieved the best average SDR performance in all the

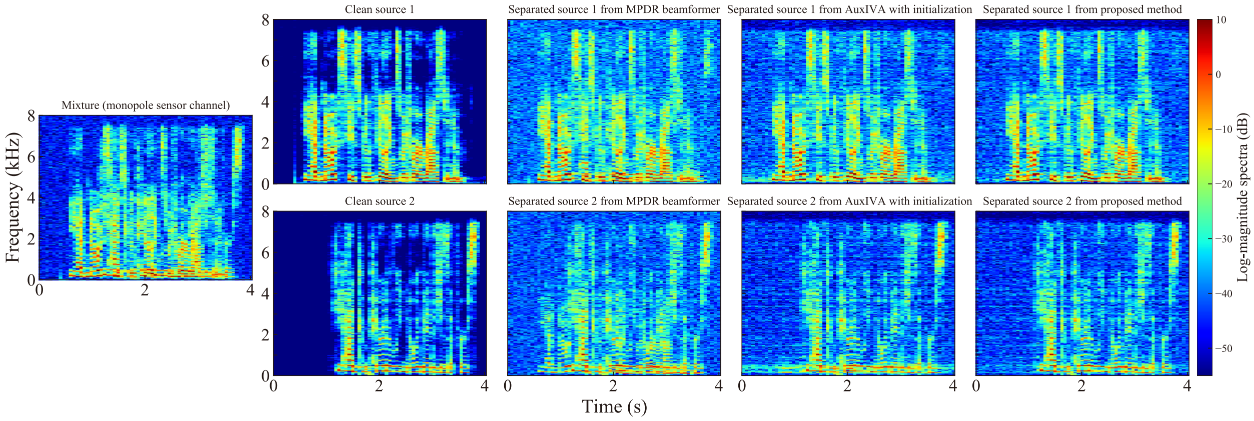

Fig. 4. A segment of log-magnitude spectra of a mixture and the separated signals.

T_{60}	200 ms	300 ms	400 ms
MPDR	1.57	1.33	1.23
AuxIVA with init.	2.05	1.70	1.47
Proposed	2.35	1.88	1.60

Table 2. The average PESQ of the studied methods.

T_{60}	200 ms	300 ms	400 ms
Initialization	12.94	24.33	29.40
Convergence	2.42	4.20	5.63

Table 3. The average angular localization error (°) of the proposed method.

studied conditions. The use of true DOAs dose bring the SDR improvement for the MPDR method given that a more accurate steering vector is used. However, it brings little help to AuxIVA because the DOAs are only used in the initial step. Interestingly, for the proposed method , the results using the DOA estimates during the iteration are slightly better that those using the true DOAs. One possible reason is that the most statistically independent components point to directions that are different from the true DOAs in reverberant environments [17].

To analyze speech quality, a segment of the log-magnitude spectra of the separated signals is shown in Fig. 4. Compared with two baseline methods, separated signals of the proposed method contain the most precise target signals and least interference, especially in the low-frequency region. The results of average PESQ for all the studied methods are shown in Table 2, which again confirms that the proposed method outperforms the two baseline method.

To further investigate the DOA estimation results using the proposed method, the average angular localization error during the iteration under different reverberation conditions are shown in Table 3. In comparison with the original multiple-source DOA estimation results obtained from the MUSIC algorithm, the average location error decreases significantly using the joint optimization method.

The average SDRs as a function of the iteration under the reverberation time of 200 ms are plotted in Fig. 5.2(a). To verify if MPDR can further suppress the interference components using the outputs of the IVA in the proposed method, Fig. 5.2(b) shows plots the average SIRs as a function of the iteration number. The results show that the output of MPDR contains less interference components as compared with the output of IVA. Meanwhile, IVA in the proposed method sightly improves the SIR as compared with the original AuxIVA because the auxiliary variable in the proposed method is refined based on the beamforming outputs during the joint optimization.

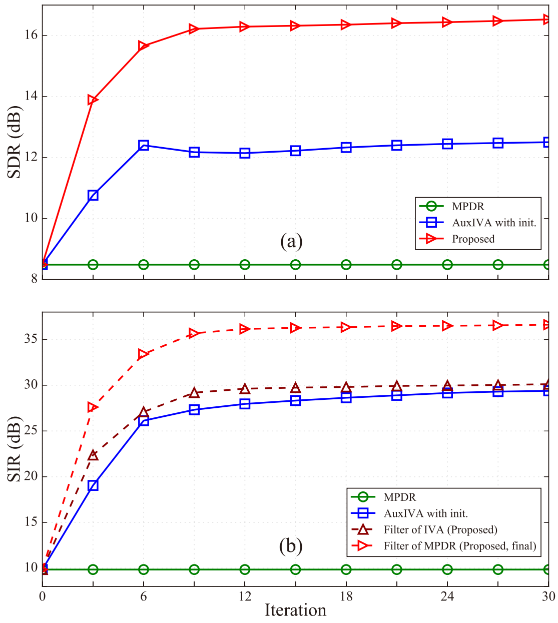

Fig. 5. The average: (a) SDR (dB) and (b) SIR (dB), both as a function of the iteration number.

6. CONCLUSION

This paper presented an IVA-assisted adaptive beamforming method with an AVS. With a rough estimates of the source DOAs, the developed method iteratively refines the estimates of the source DOAs and signal statistics. Simulation results demonstrated that significant improvement in terms of both separation and speech quality has been achieved in reverberant environments.

7. REFERENCES

[1] J. Benesty, I. Cohen, and J. Chen, *Fundamentals of Signal Enhancement and Array Signal Processing*. Singapore: Wiley-IEEE Press., 2018.

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

[3] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 2008.

[4] O. Schwartz, S. Gannot, and E. A. Habets, "Multispeaker LCMV beamformer and postfilter for source separation and noise reduction," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 5, pp. 940–951, May 2017.

[5] Y. Xu, C. Weng, L. Hui, J. Liu, M. Yu, D. Su, and D. Yu, "Joint training of complex ratio mask based beamformer and acoustic model for noise robust asr," in *Proc. IEEE ICASSP*, 2019, pp. 6745–6749.

[6] Y. Xu, M. Yu, S.-X. Zhang, L. Chen, C. Weng, J. Liu, and D. Yu, "Neural spatio-temporal beamformer for target speech separation," in *Proc. Interspeech*, 2020, pp. 56–60.

[7] Z. Zhang, Y. Xu, M. Yu, S.-X. Zhang, L. Chen, and D. Yu, "Adl-mvdr: All deep learning mvdr beamformer for target speech separation," in *Proc. IEEE ICASSP*, 2021, pp. 6089–6093.

[8] S. Makino, *Audio Source Separation*. Cham, Switzerland: Springer, 2018.

[9] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Netw.*, vol. 13, no. 4, pp. 411–430, Jun. 2000.

[10] B. Anthony and S. Terrence, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, no. 6, pp. 1129–1159, Nov. 1995.

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

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

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

[14] L. C. Parra and C. V. Alvino, "Geometric source separation: merging convolutive source separation with geometric beamforming," *IEEE Trans. Speech, Audio Process.*, vol. 10, no. 6, pp. 352–562, Sep. 2002.

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

[16] 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. IEEE ICASSP*, 2018, pp. 746–750.

[17] L. Li and K. Koishida, "Geometrically constrained independent vector analysis for directional speech enhancement," in *Proc. IEEE ICASSP*, 2020, pp. 846–850.

[18] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, and K. Shikano, "Blind source separation combining independent component analysis and beamforming," *EURASIP J. Appl. Signal Process.*, vol. 2003, no. 11, pp. 1135–1146, Nov. 2003.

[19] 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. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 666–678, Mar. 2006.

[20] A. Zamani, M. Klimke, G. Dartmann, and A. Schmeink, "Convolutive blind source separation with independent vector analysis and beamforming," in *Proc. ICECIE*, 2019, pp. 1–6.

[21] D. Li, G. Huang, Y. Lei, J. Chen, and J. Benesty, "Robust source separation with differential microphone arrays and independent low-rank matrix analysis," in *Proc. EUSIPCO*, 2021, pp. 291–295.

[22] A. Nehorai and E. Paldi, "Acoustic vector-sensor array processing," *IEEE Trans. Signal Process.*, vol. 42, no. 9, pp. 2481–2491, Sep. 1994.

[23] X. Zheng, C. Ritz, and J. Xi, "Collaborative blind source separation using location informed spatial microphones," *IEEE Signal Process. Lett.*, vol. 20, no. 1, pp. 83–86, Jan. 2013.

[24] X. Chen, W. Wang, Y. Wang, X. Zhong, and A. Alinaghi, "Reverberant speech separation with probabilistic time–frequency masking for B-format recording," *Speech Commun.*, vol. 68, pp. 41–54, Apr. 2015.

[25] M. Jia, J. Sun, C. Bao, and C. Ritz, "Separation of multiple speech sources by recovering sparse and non-sparse components from B-format microphone recordings," *Speech Commun.*, vol. 96, pp. 184–196, Feb. 2018.

[26] Y. Jia, M. Jia, L. Li, J. Gao, and S. Yang, "Multiple Speech Source Separation by Using MVDR for B-Format Recordings," in *Proc. ICCPR*, 2020, pp. 400–404.

[27] H.-E.de Bree, "An overview of microflown technologies," *Acta Acust. United Acust.*, vol. 89, no. 1, pp. 163–172, Jan. 2003.

[28] Y. Huang, J. Benesty, and G. W. Elko, "Passive acoustic source localization for video camera steering," in *Proc. IEEE ICASSP*, 2000, pp. 909–912.

[29] X. Wang, G. Huang, J. Benesty, J. Chen, and I. Cohen, "Time difference of arrival estimation based on a kronecker product decomposition," *IEEE Signal Process. Lett.*, vol. 28, pp. 51–55, 2020.

[30] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Trans. Speech, Audio Process.*, vol. 11, no. 2, pp. 109–116, Mar. 2003.

[31] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1, pp. 1–24, Oct. 2001.

[32] J. Kominek and A. W. Black, "Cmu arctic databases for speech synthesis," Carnegie Mellon Univ., Tech. Rep. CMU-LTI-03-177, 2003.

[33] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.

[34] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

[35] R. Scheibler and N. Ono, "Independent vector analysis with more microphones than sources," in *Proc. IEEE WASPAA*, 2019, pp. 185–189.