GEOMETRICALLY CONSTRAINED SOURCE EXTRACTION AND DEREVERBERATION BASED ON JOINT OPTIMIZATION

Yichen Yang[1], *Xianrui Wang*[1,2], *Andreas Brendel*[2], *Wen Zhang*[1],
Walter Kellermann[2], *and Jingdong Chen*[1]

[1]Center of Intelligent Acoustics and Immersive Communications,
Northwestern Polytechnical University, Xi'an, China
[2]Multimedia Communications and Signal Processing,
Friedrich-Alexander-Universität, Erlangen-Nürnberg, Germany

ABSTRACT

Source extraction, which aims at extracting the target source signals from the observed reverberant mixtures, plays an important role in voice communication and human-machine interfaces. Among the numerous source extraction methods that have been developed, the geometrically constrained (GC) one, which incorporates the direction-of-arrival (DOA) information of the target signals, has demonstrated great potential. However, this method generally suffers from significant performance degradation in strong reverberant environments since it is challenging to obtain in such environments accurate DOA estimates that are needed by the algorithm. To address this problem, we present in this work an iterative algorithm, which integrates the source-wise weighted prediction error (WPE)-based dereverberation principle with the geometrically constrained source extraction method. We show that this algorithm is able to improve the DOA estimation accuracy as well as the source extraction performance.

Index Terms— Blind source separation, semi-blind source extraction, geometrical constraint, dereverberation.

1. INTRODUCTION

Blind source separation (BSS), which is a particular problem of speech enhancement [1–4], is an essential component of audio processing systems as speech signals recorded under realistic conditions by microphones often consist of signals from multiple sources. BSS aims at separating source signals from observed mixtures with minimal prior information [5]. Independent component analysis (ICA) [6, 7] is one of the most widely-used BSS approaches, which estimates the demixing system under the assumption of statistically independent source signals. As speech signals are broadband in nature, a multivariate extension of ICA, i.e., independent vector analysis (IVA) [8–10], has been proposed to resolve the well-known inner permutation problem by considering higher-order relationship among frequency components of the source signals.

However, those BSS methods in their original form work only for the determined cases where the number of sources is restricted to be equal to the number of sensors. To extract sources in overdetermined scenarios, i.e., more sensors than sources, independent vector extraction (IVE) [11–13] has been proposed by introducing a statistical model for the background noise. But IVE still suffers from a number of major issues including: 1) all the target signals are separated simultaneously with a random output order, which is well-known as the outer permutation problem; 2) it is difficult to determine the desired sources when some undesired sources are similarly strong as the target ones; 3) the extraction performance degrades significantly in highly reverberant environments, which are not uncommon in practical applications.

To deal with the aforementioned issues, the geometrically constrained (GC) methods [14, 15] were proposed, which can improve the separation performance and at the same time mitigate the outer permutation problem. In [16, 17], an iterative projection (IP)-like optimization method, which uses the vector-wise coordinate descent (VCD) algorithm, was introduced to the GC-based extraction framework to accelerate its convergence. Recently, a spatially informed Bayesian framework [18, 19] based on the maximum a posteriori (MAP) principle was proposed to incorporate both spatial prior and background noise prior. Moreover, the principle of adaptive beamforming was combined with AuxIVA [20] for further performance improvement.

Although so much effort has been devoted to it, the existing methods still suffer from great performance degradation in highly reverberation environments. In order to overcome the degradation of extraction performance caused by reverberation, several methods were proposed recently. In [21], a blind dereverberation method, WPE, is used as a preprocessing step for BSS, which leads to a cascaded optimization method. To further mitigate the effect of reverberation, several methods, which jointly optimize the dereverberation and BSS modules, have been proposed [22–25]. Meanwhile, a convolutional beamformer (CBF) [26, 27] has been developed to extract the target source in highly reverberant environments. The so-called source-wise factorization CBF was then developed [27] to reduce computational complexity, in which the global filter is factorized into a source-wise WPE filter and an extraction filter. The extraction filter is updated with the IP method while the WPE filter is updated though minimizing the mean square error (MSE) [28–30]. But these methods still suffer from the problem of slow convergence and outer permutation.

To circumvent the slow convergence and outer permutation problem of the algorithm in [27], we present in this work a spatially informed joint source extraction and dereverberation framework, which can be regarded as an extension of our previous work [20]. Unlike other geometrically constrained source extraction methods, the proposed framework integrates a WPE module to improve the extraction performance, particularly, in highly reverberant environments. Moreover, the DOA prior is refined during iterations to enhance the geometric constraint. Simulations demonstrate that the proposed method is able to significantly improve the source extraction performance as compared to other geometrically constrained methods in highly reverberation conditions.

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 Grants 61831019, 62192713, and 62271401.

2. SIGNAL MODEL AND PROBLEM FORMULATION

Consider the scenario where there are N sound sources in a reverberant environment and we use an array consisting of M microphones ($N \leq M$) to capture the signals. The signals observed by the microphone array in the short-time Fourier transform (STFT) domain can then be expressed as

$$\mathbf{x}(t,f) = \sum_{\tau=0}^{L_a-1} \mathbf{A}(\tau,f)\mathbf{s}(t-\tau,f), \tag{1}$$

where $t = 1,\ldots,T$ and $f = 1,\ldots,F$ denote, respectively, the time-frame and frequency-bin indexes, T denotes the total number of the time frames, F denotes the total number of the frequency bins, $\mathbf{x}(t,f) \in \mathbb{C}^{M\times 1}$ and $\mathbf{s}(t,f) \in \mathbb{C}^{N\times 1}$ represent, respectively, the observed mixture signals and the source signals in STFT domain, $\mathbf{A}(\tau,f) \in \mathbb{C}^{M\times N}$ is the mixing matrix at time lag τ also in the STFT domain, and L_a is the number of time lagged mixing matrices. Based on the multi-input multi-output (MIMO) CBF described in [27], the demixing rule can be written in the STFT domain as

$$\mathbf{y}(t,f) = \mathbf{W}(0,f)\mathbf{x}(t,f) + \sum_{\tau=D}^{L-1} \mathbf{W}(\tau,f)\mathbf{x}(t-\tau,f), \tag{2}$$

where $\mathbf{W}(0,f) \in \mathbb{C}^{M\times M}$ and $\mathbf{W}(\tau,f) \in \mathbb{C}^{M\times M}$ are matrices composed of the convolutional beamformer coefficients, $\mathbf{y}(t,f) \in \mathbb{C}^{M\times 1}$ contains the separated signals, D is the prediction delay and L is the length of the CBF filters, with $0 < D \leq L-1$. As shown in [28–30], the mth component of $\mathbf{y}(t,f)$ in (2) can be rewritten through the source-wise factorization of the MIMO CBF as [27]

$$\mathbf{z}_m(t,f) = \mathbf{x}(t,f) - \mathbf{G}_m^{\mathrm{H}}(f)\overline{\mathbf{x}}(t-D,f), \tag{3}$$

$$y_m(t,f) = \mathbf{q}_m^{\mathrm{H}}(f)\mathbf{z}_m(t,f), \tag{4}$$

where

$$\overline{\mathbf{x}}(t-D,f) = \begin{bmatrix}\mathbf{x}^{\mathrm{T}}(t-D,f) & \ldots & \mathbf{x}^{\mathrm{T}}(t-L+1,f)\end{bmatrix}^{\mathrm{T}}, \tag{5}$$

is a $M(L-D)\times 1$ vector consisting of past samples of the mixture signals.

The left-hand sides of (3) and (4) correspond to the mth component of $\mathbf{y}(t,f)$ in (2) with $\mathbf{q}_m(f) = \mathbf{w}_m(0,f)$ and $\mathbf{G}_m(f)\mathbf{q}_m(f) = -[\mathbf{w}_m^{\mathrm{T}}(D,f),\ldots,\mathbf{w}_m^{\mathrm{T}}(L-1,f)]^{\mathrm{T}}$, where $\mathbf{w}_m(0,f)$ and $\mathbf{w}_m(\tau,f)$ are, respectively, the mth column of $\mathbf{W}^{\mathrm{H}}(0,f)$ and $\mathbf{W}^{\mathrm{H}}(\tau,f)$ for $\tau = D,\ldots,L-1$. Therefore, $\mathbf{G}_m(f)$ and $\mathbf{z}_m(t,f)$ are indeed the prediction filter of WPE and the dereverberated signal of the mth output, respectively. $\mathbf{Q}(f) = [\mathbf{q}_1(f),\ldots,\mathbf{q}_M(f)]^{\mathrm{H}}$ equals to $\mathbf{W}(0,f)$ representing a demixing matrix to separate all signals from dereverberated signals.

3. PROPOSED METHOD

3.1. Probabilistic model

Instead of attempting to separate all the active sources simultaneously from the observed mixtures, we consider in this work to extract only one source of interest (SOI) or some SOIs in strong reverberation conditions. Without loss of generality, let us consider to extract K of the N sources ($K \leq N$). The separated signal vector $\mathbf{y}(t,f)$ and the demixing matrix $\mathbf{Q}(f)$ can then be expressed as

$$\mathbf{y}(t,f) = \begin{bmatrix}\mathbf{y}^{\mathrm{SOI}}(t,f)\\ \mathbf{y}^{\mathrm{RS}}(t,f)\\ \mathbf{n}(t,f)\end{bmatrix}, \quad \mathbf{Q}(f) = \begin{bmatrix}\mathbf{Q}^{\mathrm{SOI}}(f)\\ \mathbf{Q}^{\mathrm{RS}}(f)\\ \mathbf{B}(f)\end{bmatrix}, \tag{6}$$

where $\mathbf{y}^{\mathrm{SOI}}(t,f) \in \mathbb{C}^{K\times 1}$, $\mathbf{y}^{\mathrm{RS}}(t,f) \in \mathbb{C}^{(N-K)\times 1}$, and $\mathbf{n}(t,f) \in \mathbb{C}^{(M-N)\times 1}$ are, respectively, the separated SOIs, the other *remaining sources* (RS), and the *background* (BG) noise, and $\mathbf{Q}^{\mathrm{SOI}}(f) \in \mathbb{C}^{K\times M}$, $\mathbf{Q}^{\mathrm{RS}}(f) \in \mathbb{C}^{(N-K)\times M}$, and $\mathbf{B}(f) \in \mathbb{C}^{(M-N)\times M}$ are the filters to extract the corresponding signals.

Based on the MAP criterion to derive the demixing filters [19], the following cost function is formulated

$$\begin{aligned}\mathcal{J} = &-2T\sum_f \log|\det\mathbf{Q}(f)| - \sum_{t,m}\log p\left(\underline{\mathbf{y}}_m(t)\right)\\ &-\sum_t \log p\left(\underline{\mathbf{n}}(t)\right) - \log p(\mathcal{Q}),\end{aligned} \tag{7}$$

where $\underline{\mathbf{y}}_m(t) = [y_m(t,1),\ldots,y_m(t,F)]^{\mathrm{T}}$ for $1 \leq m \leq N$ is the extracted signal vector of the mth source, and, analogously, $\underline{\mathbf{n}}(t) = [\underline{\mathbf{n}}_{N+1}^{\mathrm{T}}(t),\ldots,\underline{\mathbf{n}}_M^{\mathrm{T}}(t)]^{\mathrm{T}}$ denotes the remaining BG signals, $\mathcal{Q} = \{\mathbf{Q}(f)\}_{f=1}^{F}$ is the set of all extraction filters, and $p(\mathcal{Q})$ denotes the prior of the extraction filters.

The source m, $1 \leq m \leq N$, is assumed to follow a time-varying complex circular Gaussian distribution and the BG signal m',$N+1 \leq m' \leq M$, is assumed to follow a multivariate standard complex Gaussian distribution. Under the I.I.D. assumption, the cost function in (7) can be rewritten as

$$\begin{aligned}\mathcal{J}(\Theta) = &-2T\sum_f \log|\det\mathbf{Q}(f)| + \sum_{t,f,m}\left(\log r_m(t) + \frac{|y_m(t,f)|^2}{r_m(t)}\right)\\ &+\sum_{t,f,m'}|n_{m'}(t,f)|^2 - \sum_f \log p(\mathbf{Q}(f)).\end{aligned} \tag{8}$$

where $\Theta = \{\Theta_{\mathbf{G}},\Theta_{\mathbf{Q}},\Theta_{\mathbf{r}}\}$ contains all dereverberation matrix $\mathbf{G}_m(f)$, separation matrix $\mathbf{Q}(f)$ and the broadband time-varying variance r_m.

In practical situations, the precise *a priori* information for the demixing matrix is generally not accessible. One way to circumvent this issue is though using some information, e.g., the DOA information of the target sources, that can be estimated. This approach is adopted in this work and we investigate how to incorporate the estimates of the SOI DOAs into the priors of the SOI filters, i.e., $p(\mathbf{Q}^{\mathrm{SOI}}(f))$. Specifically, a prior based on the Euclidean distance between the extraction filter and the free-field steering vector of the SOIs [16, 18, 19] is applied to the K sources, which can be expressed using a negative log-likelihood form as follows

$$-\sum_f \log p(\mathbf{Q}^{\mathrm{SOI}}(f)) = \sum_{f,m}\lambda_m\|\mathbf{q}_m(f) - \mathbf{d}(f,\theta_m)\|_2^2 + \text{const.}, \tag{9}$$

where λ_m is a weighting coefficient, θ_m is the DOA of the mth SOI, and $\mathbf{d}(f,\theta_m)$ is the steering vector associated with the mth SOI. For the details of derivation, please refer to [18, 19]. For a uniform linear array (ULA), the steering vector can be written as

$$\mathbf{d}(f,\theta_m) = \begin{bmatrix}1 \; e^{-j\psi_m(f)} & \ldots & e^{-j(M-1)\psi_m(f)}\end{bmatrix}^{\mathrm{T}}, \tag{10}$$

where

$$\psi_m(f) = \frac{2\pi(f-1)f_{\mathrm{s}}d}{cN_{\mathrm{F}}}\cos\theta_m, \tag{11}$$

j is the imaginary unit, f_{s} is the sampling frequency, N_{F} is the length of the fast Fourier transform (FFT) in STFT, d is the microphone spacing, and c is the speed of sound in air.

The flowchart of the proposed algorithm is shown in Fig. 1, which can be viewed as a generalization of the geometric constrained BSS/BSE, the joint optimization of WPE and BSS/BSE, and the traditional IVA since the cost function used in the latter three algorithms are particular cases of the cost function given in (8).

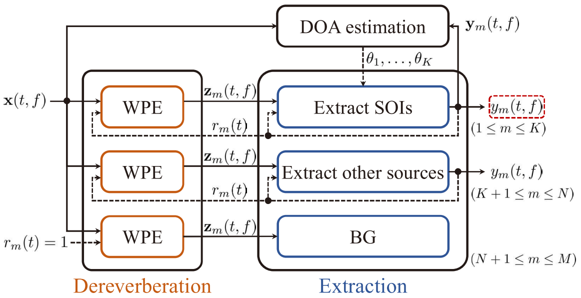

Fig. 1. Flowchart of the proposed method.

Specifically, there are three particular cases for the cost function given in (8).

- If $D = L$, in this case, the cost function in (8) degenerates to the one used in the typical geometric constrained BSS/BSE as in [19].
- If the uninformative prior of SOIs is used, in this case, (8) degenerates to the cost function for joint optimization of WPE and BSS/BSE as in [28–30].
- If both the previous conditions are satisfied and $K = N = M$, the cost function in (8) degenerates to the one used in the ordinary IVA [10].

3.2. Optimization algorithm

To optimize the cost function in (8), the coordinate ascent method [21] is used to iteratively update every parameter set in $\{\Theta_{\mathbf{G}}, \Theta_{\mathbf{Q}}, \Theta_{\mathbf{r}}\}$ while fixing the others until convergence.

3.2.1. Update of $\Theta_{\mathbf{G}}$

For the WPE parameters $\Theta_{\mathbf{G}}$, if we fix the other parameters and ignore the constant terms, the cost function in (8) can be rewritten as

$$\mathcal{J}(\Theta_{\mathbf{G}}) = \sum_{f,m} \left\| \left(\mathbf{G}_m(f) - \mathbf{R}_m^{-1}(f)\mathbf{P}_m(f)\right)\mathbf{q}_m(f) \right\|^2_{\mathbf{R}_m(f)}, \quad (12)$$

where

$$\mathbf{R}_m(f) = \sum_t \overline{\mathbf{x}}(t-D,f)\overline{\mathbf{x}}^{\mathrm{H}}(t-D,f)/r_m(t), \quad (13)$$

$$\mathbf{P}_m(f) = \sum_t \overline{\mathbf{x}}(t-D,f)\overline{\mathbf{x}}^{\mathrm{H}}(t,f)/r_m(t), \quad (14)$$

and $\|\mathbf{x}\|^2_{\mathbf{R}} \triangleq \mathbf{x}^{\mathrm{H}}\mathbf{R}\mathbf{x}$. Note that the update

$$\mathbf{G}_m(f) \leftarrow \mathbf{R}_m^{-1}(f)\mathbf{P}_m(f) \quad (15)$$

minimizes (12).

3.2.2. Update of $\Theta_{\mathbf{Q}}$

By fixing the WPE parameters, the cost function (8) is then equivalent to the cost function in [19]. Hence, by means of the MM algorithm, the original cost function is replaced by a simpler surrogate cost function, that is

$$\mathcal{J}(\Theta_{\mathbf{Q}}) = -2T\sum_f \log|\det \mathbf{Q}(f)| + \sum_{t,f,m} \mathbf{q}_m^{\mathrm{H}}(f)\mathbf{U}_m(f)\mathbf{q}_m(f) + \sum_{f,m} \lambda_m \left\|\mathbf{q}_m(f) - \mathbf{d}(f,\theta_m)\right\|_2^2, \quad (16)$$

where

$$\mathbf{U}_m(f) = \frac{1}{T}\sum_t \mathbf{z}_m(t,f)\mathbf{z}_m^{\mathrm{H}}(t,f)/r_m(t), \quad (17)$$

and $\mathbf{z}_m(t,f)$ is the mth dereverberated signal in (3). Note that the cost function above is equivalent to that of the spatially informed source extraction algorithm as in [16, 19] except from the observed signal $\mathbf{z}_m(t,f)$. Hence, the IP-based update rules can be applied for updating $\mathbf{Q}(f)$ [16]

$$\mathbf{p}_m(f) = \left(\mathbf{Q}(f)\hat{\mathbf{U}}_m(f)\right)^{-1}\mathbf{e}_m, \quad (18)$$

$$\widetilde{\mathbf{p}}_m(f) = \lambda_m \hat{\mathbf{U}}_m^{-1}(f)\mathbf{d}(f,\theta_m), \quad (19)$$

$$h_m(f) = \mathbf{p}_m^{\mathrm{H}}(f)\hat{\mathbf{U}}_m(f)\mathbf{p}_m(f), \quad (20)$$

$$\widetilde{h}_m(f) = \mathbf{p}_m^{\mathrm{H}}(f)\hat{\mathbf{U}}_m(f)\widetilde{\mathbf{p}}_m(f), \quad (21)$$

$$\mathbf{q}_m(f) \leftarrow \begin{cases} \frac{\mathbf{p}_m(f)}{\sqrt{h_m(f)}} + \widetilde{\mathbf{p}}_m(f), & \text{if } \widetilde{h}_m(f) = 0, \\ \frac{\widetilde{h}_m(f)}{2h_m(f)}\left(-1 + \sqrt{1 + \frac{4h_m(f)}{|\widetilde{h}_m(f)|^2}}\right)\mathbf{p}_m(f) + \widetilde{\mathbf{p}}_m(f), & \text{else}, \end{cases} \quad (22)$$

where $\hat{\mathbf{U}}_m(f) = \mathbf{U}_m(f) + \lambda_m\mathbf{I}_M$, $\mathbf{I}_M \in \mathbb{R}^{M\times M}$ is an identity matrix, and $\mathbf{e}_m$ is the unit vector with the mth element being 1 and the others being 0. For the remaining sources, the extraction filters are updated by regular IP updates rules as [10]

$$\mathbf{q}_m(f) \leftarrow \left(\mathbf{Q}(f)\mathbf{U}_m(f)\right)^{-1}\mathbf{e}_m, \quad (23)$$

$$\mathbf{q}_m(f) \leftarrow \mathbf{q}_m(f)/\sqrt{\mathbf{q}_m^{\mathrm{H}}(f)\mathbf{U}_m(f)\mathbf{q}_m(f)}, \quad (24)$$

where one can check that (23) and (24) are equal to (18)–(22) for $\lambda_m = 0$. Since we do not attempt to separate the BG components, $\mathbf{B}(f)$ can be decomposed based on the orthogonal constraints as proposed in [11] as

$$\mathbf{B}(f) = \left[\mathbf{J}(f), -\mathbf{I}_{M-N}\right], \quad (25)$$

where $\mathbf{J}(f) \in \mathbb{C}^{(M-N)\times N}$ and $\mathbf{I}_{M-N}$ is an identity matrix. By assuming orthogonality of the desired and background signal subspaces, $\mathbf{J}(f)$ is updated as

$$\mathbf{B}(f) \leftarrow \begin{pmatrix} \left(\mathbf{E}_N\mathbf{C}(f)\mathbf{Q}^{\mathrm{AS}}(f)\right)\left(\mathbf{E}_S\mathbf{C}(f)\mathbf{Q}^{\mathrm{AS}}(f)\right)^{-1} \\ -\mathbf{I}_{M-N} \end{pmatrix}, \quad (26)$$

where $\mathbf{E}_S = [\mathbf{I}_N, \mathbf{0}_{N\times(M-N)}]$, $\mathbf{E}_N = [\mathbf{0}_{(M-N)\times N}, \mathbf{I}_{M-N}]$. $\mathbf{Q}^{\mathrm{AS}}(f) = [(\mathbf{Q}^{\mathrm{SOI}}(f))^{\mathrm{T}}, (\mathbf{Q}^{\mathrm{RS}}(f))^{\mathrm{T}}]^{\mathrm{T}}$ contain the extraction filters for all *active sources* (AS), and $\mathbf{C}(f)$ is the covariance matrix of observed signals.

3.2.3. Update of Θ_r

After updating the extraction matrix $\mathbf{Q}(f)$, the time-varying variance $r_m(t)$ for all the active sources and BG signals can be updated by using the output of (4) as

$$r_m(t) \leftarrow \begin{cases} \frac{1}{F}\sum_f |y_m(t,f)|^2, & \text{if } 1 \le m \le N, \\ 1, & \text{else}. \end{cases} \quad (27)$$

Then, the updated $r_m(t)$ can be applied to iteratively update again the parameters $\Theta_{\mathbf{G}}$ and $\Theta_{\mathbf{Q}}$ according to (13), (14), and (17).

3.2.4. DOA Updates for the SOIs

Due to the fact that the target source DOAs are used to form the prior through (11), good estimation of DOAs is essential for the presented algorithm to extract the target SOIs. However, the estimates of DOAs in multiple-source scenarios are generally biased, especially in reverberant and noisy environments [31]. To ensure high accuracy for

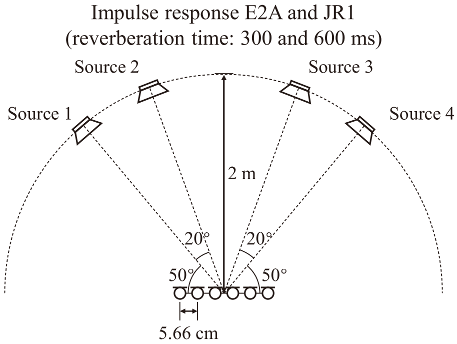
Impulse response E2A and JR1
(reverberation time: 300 and 600 ms)
Source 1
Source 2
Source 3
Source 4
2 m
20°
20°
50°
50°
5.66 cm

Fig. 2. Simulation setup with a ULA of 6 microphones and 4 sources.

DOA estimation, an additional DOA estimation module based on the Capon method [32] is adopted during the iterative update so that the DOA estimation error decreases iteratively. In every iteration, a multi-channel back-projection [33] is applied to calculate the covariance matrix associated with every SOI for corresponding DOA estimation, which has been discussed in our earlier work [20].

4. SIMULATIONS

4.1. Simulation setup

The performance of the proposed algorithm is verified through simulations in this section. The observed signals are simulated by convolving the speech signals from the test set of the TIMIT database [34] with the room impulse responses (RIRs) from the RWCP dataset [35] (see Fig. 2 for the geometric setup of the recordings). Room reverberation time T_{60} from 300 ms and 600 ms are simulated and white Gaussian noise is added to control the signal-to-noise ratio (SNR) to 30 dB. A ULA consisting of $M = 6$ omnidirectional microphones with an inter-element spacing of 5.66 cm is used, and the number of simultaneously active sources is $N = 4$. The azimuth angles of the four sources are 50°, 70°, 110°, and 130°, which are all 2-meter away from the center of the microphone array.

Clean signals are randomly selected from the TIMIT database and then concatenated to form the source signals which are 10-s long. Twenty-five mixtures are then generated. To demonstrate the ability of extracting any number ($K \leq M$) of sources, two scenarios are evaluated: one with only a single SOI and the other with 4 SOIs. All signals are sampled at 16 kHz. The STFT is implemented with the von Hann window of 128-ms long and a window shift of 32 ms. The improvement of the signal-to-distortion ratio (SDR) [36] (denoted by ΔSDR) is used as the metric for evaluating the extraction performance where the reference signals are obtained by convolving the speech signals with the RIRs truncated to a length of 32 ms.

The extraction performance of the presented algorithm (denoted as WPE-GCIVE) is compared to the IVE [12], the joint optimization algorithm of WPE and IVE (WPE-IVE) [29, 30], and the classical geometric constrained independent vector extraction (GCIVE) [19]. Note that IVE and WPE-IVE do not work in the case when $K = 1$ as they can not distinguish the target signal and interferences. So, in this case, only GCIVE is compared. The weighting parameter of the geometric constraint is decreased over iterations according to [16] with the initial value set to 1, the delay D and the length L are set to 2 and 4, respectively. Taking the computational complexity into account, the parameters of WPE filters are updated every 10 iterations.

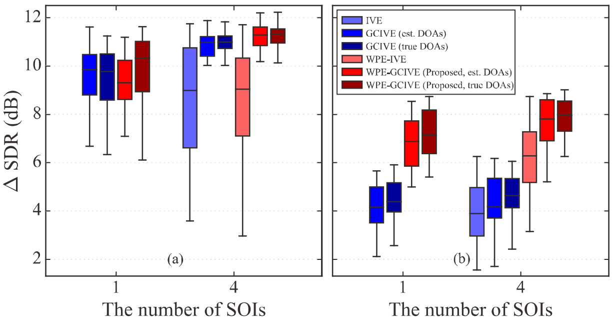

Fig. 3. Comparison of the average ΔSDR (dB) achieved by the studied methods with the different number of SOIs K in reverberant environments: (a) 300 ms, and (b) 600 ms.

T_{60}	Initialization	GCIVE	WPE-GCIVE
300 ms	9.73	4.88	**4.56**
600 ms		14.41	**7.91**

Table 1. The average angular localization error ($^\circ$) in different reverberation conditions with $K = 4$ SOIs.

4.2. Results and discussion

The extraction performance of the presented algorithm and the baselines under different reverberation times and different numbers of SOIs are presented in Fig. 3. For the GC-based methods, either the true DOAs are directly used as priors, or DOAs are estimated and refined during the parameter updating with a rough initialization, which is given by adding a random error uniformly distributed in $[-17.5^\circ, 17.5^\circ]$ to the ground truth. The average angular localization errors at initialization or after update through extraction approaches are shown in Table 1.

From these simulation results, one can see that including WPE yields small improvements for DOA estimation if reverberation is light (e.g., $T_{60} = 300$ ms) since the angular localization error is already acceptable for source extraction. However, as reverberation becomes stronger ($T_{60} = 600$ ms), the original GC-based methods cannot improve the performance since the bias of the angular localization error increases and the geometrical constraints fail to select the SOIs. By incorporating WPE into the GC-based method, the proposed algorithm only reduces the average DOA estimation error for the SOIs but also improves the extraction performance (see Fig. 3) as compared to the baseline methods. Moreover, because the proposed method can extract an arbitrary number of SOIs satisfying $K \leq M$, the outer permutation is solved naturally.

5. CONCLUSION

This paper presented a spatially guided framework, which jointly optimizes the WPE and geometric constrained source extraction cost functions through the source-wise CBF factorization. The source DOA estimation is also refined during the iterations to further improve the source extraction performance. The presented algorithm works well in highly reverberant environments and simulation results demonstrate the superiority of the proposed method over the compared baseline methods, which are the state-of-the-art source separation methods reported in the literature.

6. REFERENCES

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

[2] G. Huang, J. Chen, and J. Benesty, "Insights into frequency-invariant beamforming with concentric circular microphone arrays," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 12, pp. 2305–2318, Dec. 2018.

[3] G. Huang, J. Benesty, and J. Chen, "Fundamental approaches to robust differential beamforming with high directivity factors," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 3074–3088, 2022.

[4] 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, Jan. 2017.

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

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

[7] 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.

[8] 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.

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

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

[11] Z. Koldovskỳ and P. Tichavskỳ, "Gradient algorithms for complex non-gaussian independent component/vector extraction, question of convergence," *IEEE Trans. Signal Process.*, vol. 67, no. 4, pp. 1050–1064, Dec. 2018.

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

[13] R. Ikeshita, T. Nakatani, and S. Araki, "Overdetermined independent vector analysis," in *Proc. IEEE ICASSP*, 2020, pp. 591–595.

[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, Dec. 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] A. Brendel, T. Haubner, and W. Kellermann, "Spatially guided independent vector analysis," in *Proc. IEEE ICASSP*, 2020, pp. 596–600.

[19] A. Brendel, T. Haubner, and W. Kellermann, "A unified probabilistic view on spatially informed source separation and extraction based on independent vector analysis," *IEEE Trans. Signal Process.*, vol. 68, pp. 3545–3558, Jun. 2020.

[20] Y. Yang, X. Wang, W. Zhang, and J. Chen, "Independent vector analysis assisted adaptive beamforming for speech source separation on an acoustic vector sensor," in *Proc. IEEE IWAENC*, 2022.

[21] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. G. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 69–84, Jan. 2011.

[22] H. Kagami, H. Kameoka, and M. Yukawa, "Joint separation and dereverberation of reverberant mixtures with determined multichannel non-negative matrix factorization," in *Proc. IEEE ICASSP*, 2018, pp. 31–35.

[23] R. Ikeshita, N. Ito, T. Nakatani, and H. Sawada, "A unifying framework for blind source separation based on a joint diagonalizability constraint," in *Proc. EUSIPCO*, 2019, pp. 1–5.

[24] T. Nakashima, R. Scheibler, M. Togami, and N. Ono, "Joint dereverberation and separation with iterative source steering," in *Proc. IEEE ICASSP*, 2021, pp. 216–220.

[25] X. Wang, A. Brendel, G. Huang, Y. Yang, W. Kellermann, and J. Chen, "Spatially informed independent vector analysis for source extraction based on the convolutive transfer function model," in *Proc. IEEE ICASSP*, 2023.

[26] T. Nakatani and K. Kinoshita, "A unified convolutional beamformer for simultaneous denoising and dereverberation," *IEEE Signal Process. Lett.*, vol. 26, pp. 903–907, Apr. 2019.

[27] T. Nakatani, C. Boeddeker, K. Kinoshita, R. Ikeshita, M. Delcroix, and R. Haeb-Umbach, "Jointly optimal denoising, dereverberation, and source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2267–2282, Jul. 2020.

[28] T. Nakatani, R. Ikeshita, K. Kinoshita, H. Sawada, and S. Araki, "Computationally efficient and versatile framework for joint optimization of blind speech separation and dereverberation," in *Proc. Interspeech*, 2020, pp. 91–95.

[29] R. Ikeshita and T. Nakatani, "Independent vector extraction for fast joint blind source separation and dereverberation," *IEEE Signal Process. Lett.*, vol. 28, pp. 972–976, 2021.

[30] T. Nakatani, R. Ikeshita, K. Kinoshita, H. Sawada, and S. Araki, "Blind and neural network-guided convolutional beamformer for joint denoising, dereverberation, and source separation," in *Proc. IEEE ICASSP*, 2021, pp. 6129–6133.

[31] 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.

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

[33] 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.

[34] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, pp. 27403, Feb. 1993.

[35] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," in *Proc. Lang. Resour. Eval.*, 2000, pp. 965–968.

[36] 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, Jun. 2006.