SPATIALLY INFORMED INDEPENDENT VECTOR ANALYSIS FOR SOURCE EXTRACTION BASED ON THE CONVOLUTIVE TRANSFER FUNCTION MODEL

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

[1]CIAIC and Shaanxi Provincial Key Laboratory of Artificial Intelligence,
Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China
[2]Multimedia Communications and Signal Processing,
Friedrich-Alexander-Universität, Erlangen-Nürnberg, Germany

ABSTRACT

Spatial information can help improve source separation performance. Numerous spatially informed source extraction methods based on the independent vector analysis (IVA) have been developed, which can achieve reasonably good performance in non- or weakly reverberant environments. However, the performance of those methods degrades quickly as the reverberation increases. The underlying reason is that those methods are derived based on the multiplicative transfer function model with a rank-1 assumption, which does not hold true if reverberation is strong. To circumvent this issue, this paper proposes to use the convolutive transfer function (CTF) model to improve the source extraction performance and develop a spatially informed IVA algorithm. Simulations demonstrate the efficacy of the developed method even in highly reverberant environments.

***Index Terms*—** Independent vector analysis, spatially informed source extraction, convolutive transfer function.

1. INTRODUCTION

In acoustic applications, the signals of interest are often contaminated by interference, reverberation, and noise, which will not only impair the speech quality and intelligibility but also degrade the performance of automatic speech recognition (ASR) [1, 2]. Blind source separation (BSS) aims at recovering the source signals from observed mixtures with minimal prior information about the source activities and the mixing system. A popular class of BSS algorithms is rooted in frequency-domain independent component analysis (FD-ICA), which attempts to separate source signals by exploiting statistical independence between different sources in every frequency bin individually [3–5]. But the permutation problem is inherent to FD-ICA though a post-processing step can help mitigate this issue [6, 7]. As a multivariate extension of ICA, independent vector analysis (IVA) not only exploits the independence among sources, but also enforces the dependence among different frequency components of every extracted signal [8, 9]. Theoretically, IVA should not suffer from the inner permutation problem; but in practice, IVA may fail to align distant frequency components as they are not strongly dependent, leading to the so-called block permutation problem [10]. Besides, the output signals are arbitrarily ordered, which is known as the outer permutation problem.

Most source extraction or separation applications focus on separating different source signals coming from different directions. It is straightforward to think about how to use this spatial information in source separation since spatial information of the sound sources is known or can be well-estimated [11, 12]. Spatial information associated with microphone arrays has been widely used in beamforming-based source extraction methods [13–15]. Intuitively, incorporating spatial information into ICA or IVA-based source extraction should help improve performance, which has attracted considerable attention. For example, studies showed that spatially informed source extraction methods can extract the desired sources and solve both the block and the outer permutation problems with a faster convergence speed [16–19]. In [20, 21], a unified probabilistic framework for spatially informed source separation and extraction was proposed, which was also generalized to incorporate statistics of the background noise [22].

Most of the IVA-based source extraction methods are relying on a rank-1 assumption, which requires that the analysis window length exceeds the effective length of the acoustic impulse response (AIR) so that the time-domain signal can be represented by the multiplicative transfer function (MTF) model in the short-time Fourier transform (STFT) domain. However, this requirement cannot be guaranteed in practical reverberant environments, leading to dramatic performance degradation [23]. To overcome this limitation, full-rank covariance matrix analysis (FCA) was introduced [24] where each source image is modeled with a time-invariant full-rank spatial covariance matrix (SCM) and a time-varying variance. However, FCA models are challenging to identify as a large number of parameters are involved. To decrease the number of free parameters, non-negative matrix factorization (NMF) [25] was leveraged to model the power spectrogram, leading to many multichannel NMF (MNMF) methods [26–28]. However, the computational burden of those algorithms is heavy, and it is difficult to incorporate with prior spatial information.

An alternative model for the signals in reverberant environments is based on the convolutive transfer function (CTF) [29], which uses convolution rather than multiplication in the STFT domain to represent the time-domain convolution. Recently, a CTF-based MNMF (CTF-MNMF) method was proposed, which showed superiority in highly reverberant environments [30]. However, its convergence speed is low and it restricts the length of the CTF filters to be equal to the number of microphones. To circumvent the limitations of the method in [30], we develop a method in this paper, which can be viewed as the extension of the work in [20–22] with the CTF model. The proposed spatially informed source extraction method can achieve better source extraction performance in reverberant environments.

2. SIGNAL MODEL

We consider a reverberant acoustic scenario with N source signals picked up by M microphones ($M \geq N$). The observed signal at the mth microphone can be expressed as

$$\begin{aligned} x_m(t) &= \sum_{n=1}^{N} h_{mn} * s_n(t) + v_m(t) \\ &= \sum_{n=1}^{N} c_{m,n}(t) + v_m(t),\ m = 1, 2, \ldots, M, \end{aligned} \tag{1}$$

where t is the discrete-time index, h_{mn} is the AIR from the nth source to the mth microphone, $s_n(t)$ denotes the nth source signal, $*$ represents the linear convolution, $v_m(t)$ is the additive background noise at the mth microphone, and $c_{m,n}(t) = h_{mn} * s_n(t)$ stands

for the contribution of the nth source to the mth microphone. It is assumed that source signals are mutually independent and the background noise is uncorrelated to the source signals.

In the STFT domain, the signal model in (1) can be expressed as

$$x_{m,i,j} = \sum_{n=1}^{N} c_{m,n,i,j} + v_{m,i,j}, \tag{2}$$

where $i \in \{1, 2, \ldots, I\}$ is the frequency index, $j \in \{1, 2, \ldots, J\}$ denotes the time-frame index, I, J are the numbers of frequency bins and time frames, $c_{m,n,i,j}$ and $v_{m,i,j}$ are the STFTs of $c_{m,n}(t)$ and $v_m(t)$, respectively. With the CTF approximation [29], we have

$$c_{m,n,i,j} = \sum_{l=0}^{L_n-1} h_{m,n,i,l}\, s_{n,i,j-l}, \tag{3}$$

where $h_{m,n,i,l}$ are the band-to-band filter coefficients and L_n is the length of the CTF filter. In a vector form, the CTF mixing system can be represented as

$$\begin{aligned}\mathbf{x}_{i,j} &= \sum_{n=1}^{N} \mathbf{c}_{n,i,l} + \mathbf{v}_{i,j} \\ &= \sum_{n=1}^{N} \sum_{l=0}^{L_n-1} \mathbf{h}_{n,i,l}\, s_{n,i,j-l} + \mathbf{v}_{i,j} \\ &= \tilde{\mathbf{H}}_i \tilde{\mathbf{s}}_{i,j} + \mathbf{v}_{i,j},\end{aligned} \tag{4}$$

where

$$\begin{aligned}\mathbf{x}_{i,j} &= [x_{1,i,j}, \quad x_{2,i,j}, \quad \ldots, \quad x_{M,i,j}]^T \in \mathbb{C}^{M\times 1}, \\ \mathbf{c}_{n,i,l} &= [c_{1,n,i,j}, \quad c_{2,n,i,j}, \quad \ldots, \quad c_{M,n,i,j}]^T \in \mathbb{C}^{M\times 1}, \\ \mathbf{v}_{i,j} &= [v_{1,i,j}, \quad v_{2,i,j}, \quad \ldots, \quad v_{M,i,j}]^T \in \mathbb{C}^{M\times 1}, \\ \mathbf{h}_{n,i,l} &= [h_{1,n,i,l}, \quad h_{2,n,i,l}, \quad \ldots, \quad h_{M,n,i,l}]^T \in \mathbb{C}^{M\times 1}, \\ \tilde{\mathbf{H}}_i &= [\mathbf{H}_{1,i}, \quad \mathbf{H}_{2,i}, \quad \ldots, \quad \mathbf{H}_{N,i}] \in \mathbb{C}^{M\times L}, \\ \mathbf{H}_{n,i} &= [\mathbf{h}_{n,i,0}, \quad \mathbf{h}_{n,i,1}, \quad \ldots, \quad \mathbf{h}_{n,i,L_n-1}] \in \mathbb{C}^{M\times L_n}, \\ \tilde{\mathbf{s}}_{n,i,j} &= [\check{\mathbf{s}}_{1,i,j}, \quad \check{\mathbf{s}}_{2,i,j}, \quad \ldots, \quad \check{\mathbf{s}}_{N,i,j}]^T \in \mathbb{C}^{L\times 1}, \\ \check{\mathbf{s}}_{n,i,j} &= [s_{n,i,j}, \quad s_{n,i,j-1}, \quad \ldots, \quad s_{n,i,j-L_n+1}] \in \mathbb{C}^{1\times L_n},\end{aligned}$$

and $L = \sum_{n=1}^{N} L_n \leq M$. If the STFT window is sufficiently long and can cover the effective part of the AIRs, L_n can be set to 1 and (4) degenerates to the conventional MTF-based mixing model.

3. PROPOSED ALGORITHM

3.1. Source extraction

In order to extract target signals, we need to estimate a group of spatial filters [22, 31] to be applied to the sensor signals $\mathbf{x}_{i,j}$

$$\mathbf{W}_i = \begin{bmatrix} \mathbf{W}_i^{\mathrm{S}} \\ \mathbf{U}_i \end{bmatrix} \in \mathbb{C}^{M\times M}, \tag{5}$$

where

$$\mathbf{W}_i^{\mathrm{S}} = [\mathbf{W}_{1,i}, \quad \mathbf{W}_{2,i}, \quad \ldots, \quad \mathbf{W}_{N,i}]^H \in \mathbb{C}^{L\times M}$$

is the source extraction matrix and

$$\mathbf{W}_{n,i} = [\mathbf{w}_{n,i,0}, \quad \mathbf{w}_{n,i,1}, \quad \ldots, \quad \mathbf{w}_{n,i,L_n-1}] \in \mathbb{C}^{M\times L_n}$$

is the group of filters corresponding to source n.

The noise separation matrix $\mathbf{U}_i \in \mathbb{C}^{M\times K}$ is defined as

$$\mathbf{U}_i = [\mathbf{u}_{1,i}, \quad \mathbf{u}_{2,i}, \quad \cdots, \quad \mathbf{u}_{K,i}]^H,$$

where $\mathbf{u}_{k,i}$ is the kth background noise filter. We model the background noise as

$$\mathbf{v}_{i,j} = \mathbf{\Psi}_i \mathbf{z}_{i,j}, \tag{6}$$

where $\mathbf{\Psi}_i \in \mathbb{C}^{M\times K}$ $(K = M - L)$ is the noise transformation matrix. The columns of $\mathbf{\Psi}_i$ are assumed to be linearly independent of the columns of the signal mixing matrix $\tilde{\mathbf{H}}_i$ and

$$\mathbf{z}_{i,j} = [z_{1,i,j}, \quad , z_{2,i,j}, \quad \ldots, \quad z_{K,i,j}]^T \tag{7}$$

consists of decomposed noise components [31]. Under this assumption, the noise can be separated from source signals by applying $\mathbf{U}_i$ to the observations $\mathbf{x}_{i,j}$ in the form of (4). Since we do not aim at extracting the noise components, a specific structure of $\mathbf{U}_i$ can be chosen to derive a computationally efficient algorithmic solution. A straightforward choice is given in [22, 31]

$$\mathbf{U}_i = [\mathbf{J}_i, \quad -\mathbf{I}_K] \in \mathbb{C}^{K\times M}, \tag{8}$$

where $\mathbf{J}_i \in \mathbb{C}^{K\times L}$ contains the adjustable parameters and $\mathbf{I}_K$ is an identity matrix of size $K \times K$.

Applying these spatial filters, we obtain

$$y_{n,i,j,l} = \mathbf{w}_{n,i,l}^H \mathbf{x}_{i,j}, \tag{9}$$

$$\hat{z}_{k,i,j} = \mathbf{u}_{k,i}^H \mathbf{x}_{i,j}, \tag{10}$$

where $y_{n,i,j,l}$ is the estimated source signal with l taps delay and $\hat{z}_{k,i,j}$ is the filtered background noise which may stay mixed. Estimating source signals using (9) in reverberant environments will introduce spatial distortion [32], so we estimate the spatial images as in CTF-MNMF [30]. We define $\hat{\bar{\mathbf{H}}}_i = \mathbf{W}_i^{-1}$ and partition it with the same structure as the mixing matrix $\bar{\mathbf{H}} = \left[\tilde{\mathbf{H}}_i, \mathbf{\Psi}_i\right]$, i.e.,

$$\hat{\bar{\mathbf{H}}}_i = \left[\hat{\mathbf{H}}_{1,i}, \quad \hat{\mathbf{H}}_{2,i}, \quad \ldots, \quad \hat{\mathbf{H}}_{N,i}, \quad \hat{\mathbf{\Psi}}_i\right]. \tag{11}$$

Then the multichannel Wiener filter (MWF) is used to estimate the spatial images [27] [30]

$$\hat{\mathbf{c}}_{n,i,j} = \left(\hat{\mathbf{H}}_{n,i} \mathbf{\Lambda}_{n,i,j} \hat{\mathbf{H}}_{n,i}^H\right) \left(\mathbf{W}_i^H \mathbf{\Lambda}_{i,j}^{-1} \mathbf{W}_i\right) \mathbf{x}_{i,j}, \tag{12}$$

where

$$\begin{aligned}\mathbf{\Lambda}_{i,j} &= \begin{bmatrix} \mathbf{\Lambda}_{1,i,j} & & & & \\ & \mathbf{\Lambda}_{2,i,j} & & & \\ & & \cdots & & \\ & & & \mathbf{\Lambda}_{N,i,j} & \\ & & & & \mathbf{\Lambda}_{z,i,j} \end{bmatrix}, \\ &= \mathrm{diag}\left(\mathbf{\Lambda}_{1,i,j}, \quad \mathbf{\Lambda}_{2,i,j}, \quad \cdots, \quad \mathbf{\Lambda}_{N,i,j} \quad \mathbf{\Lambda}_{z,i,j}\right), \\ \mathbf{\Lambda}_{n,i,j} &= \mathrm{diag}\left(|y_{n,i,j,0}|^2, \quad |y_{n,i,j,1}|^2, \quad \cdots, \quad |y_{n,i,j,L_N-1}|^2\right), \\ \mathbf{\Lambda}_{z,i,j} &= \mathrm{diag}\left(|\hat{z}_{1,i,j}|^2, \quad |\hat{z}_{2,i,j}|^2, \quad \cdots, \quad |\hat{z}_{K,i,j}|^2\right).\end{aligned}$$

3.2. Probabilistic Model

Now we construct a cost function for estimating the set of all spatial filters $\mathcal{W} = \{\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_I\}$. We model the source signals with a time-varying complex Gaussian distribution, which has been widely used in the literature [8, 9], i.e.,

$$\mathrm{p}(\mathbf{s}_{n,j}) = \frac{1}{\pi\sigma_{n,j}^2} \exp\left(-\frac{||\mathbf{s}_{n,j}||_2^2}{\sigma_{n,j}^2}\right), \tag{13}$$

where $||\cdot||_2$ stands for ℓ_2 norm, $\sigma_{n,j}^2$ is the broadband time-varying signal variance, and

$$\mathbf{s}_{n,j} = [s_{n,1,j}, \quad s_{n,2,j}, \quad \cdots, \quad s_{n,I,j}]^T \in \mathbb{C}^{I\times 1}.$$

The filtered background noise may stay mixed, thus it is assumed to be more stationary than the source signals and close to being normally distributed. Therefore, we assume

$$\hat{\mathbf{z}}_{i,j} = [\hat{z}_{1,i,j}, \quad \hat{z}_{2,i,j}, \quad \ldots, \quad \hat{z}_{K,i,j}]^T \tag{14}$$

to follow a time-invariant complex Gaussian distribution

$$p(\hat{\mathbf{z}}_{i,j}) = \frac{1}{\pi^K |\det \mathbf{R}_i|} \exp\left(-\hat{\mathbf{z}}_{i,j}^H \mathbf{R}_i^{-1} \hat{\mathbf{z}}_{i,j}\right), \tag{15}$$

where $\mathbf{R}_i \in \mathbb{C}^{K\times K}$ is the covariance matrix of the separated noise.

As shown in [29], the direct path and early reflections, which encode the source location and array geometry information, are mainly represented by $\mathbf{h}_{n,i,0}$ while the late reverberation, which is much more diffuse, is embedded in $\mathbf{h}_{n,i,l},\ l > 0$. Thus, we only impose constraints on $\mathbf{w}_{n,i,0}, n \in \mathcal{N}$, where $\mathcal{N}$ denotes the set of source signals whose direction is given or pre-estimated. We choose the prior distribution for $\mathbf{w}_{n,i,0}, n \in \mathcal{N}$ as in [20, 22]

$$p\left(\mathbf{w}_{n,i,0}\right) = \frac{\sqrt{\lambda_n^M}}{\sqrt{\pi^M}} \exp\left(-\sum_{i=1}^{I} \lambda_n ||\mathbf{w}_{n,i,0} - \mathbf{d}_{i,\theta_n}||_2^2\right), \quad (16)$$

where λ_n is a user-defined parameter and $\mathbf{d}_{i,\theta_n} \in \mathbb{C}^{M\times 1}$ is the free-field steering vector of the look direction θ_n, i.e.,

$$[\mathbf{d}_{i,\theta_n}]_m = \exp\left(j\frac{2\pi f_i}{c}\left\|\mathbf{r}_m - \mathbf{r}_1\right\|_2 \cos\theta_n\right), \quad (17)$$

f_i is the frequency, c is the speed of sound in air, and $\mathbf{r}_m$ is the location of the m th microphone.

By adopting the maximum a posteriori (MAP) framework [20–22], the spatially informed negated likelihood function is denoted as

$$\begin{aligned}\mathcal{J}\left(\mathcal{W}\right) = &-2J\sum_{i=1}^{I}\log|\det\mathbf{W}_i| \\ &+\sum_{n,i,j=1}^{N,I,J}\sum_{l=0}^{L_n-1}\left(\log r_{n,j-l} + \frac{||\mathbf{y}_{n,j,l}||_2^2}{r_{n,j-l}}\right) \\ &+\sum_{i,j=1}^{I,J}\left(\log|\det\mathbf{R}_i| + \hat{\mathbf{z}}_{i,j}^H\mathbf{R}_i^{-1}\hat{\mathbf{z}}_{i,j}\right) \\ &-\sum_{n\in\mathcal{N}}\sum_{i=1}^{I}\log p(\mathbf{w}_{n,i,0}) + C, \end{aligned} \quad (18)$$

where C is a constant and

$$\mathbf{y}_{n,j,l} = [y_{n,1,j,l}, \quad y_{n,2,j,l}, \quad \cdots, \quad y_{n,I,j,l}]^T \in \mathbb{C}^{I\times 1}.$$

We define the estimated source signal energy as

$$r_{n,j} = ||\mathbf{w}_{n,i,0}^H\mathbf{x}_{i,j}||_2^2. \quad (19)$$

3.3. Optimization

In the following, we deduce update rules for optimizing (18) based on the auxiliary function technique. Let us define p as the iteration index. As shown in (18), once $\hat{\mathbf{z}}_{i,j}$ and $\mathbf{R}_i$ are fixed, the noise term does not influence the estimation of the set of source extraction matrices $\mathcal{W}^s = \{\mathbf{W}_1^{\mathrm{S}}, \mathbf{W}_2^{\mathrm{S}}, \ldots, \mathbf{W}_I^{\mathrm{S}}\}$ anymore. Therefore, to derive update rules for $\mathcal{W}^{\mathrm{S}}$, we drop the noise term and the constant term in (18). Since the prior term is quadratic, we construct an auxiliary function as

$$\begin{aligned}\mathcal{J}_{\text{Aux}}^p\left(\mathcal{W}|\mathcal{W}^p\right) = &-2J\sum_{i=1}^{I}\log|\det\mathbf{W}_i| \\ &+\sum_{n,i,j=1}^{N,I,J}\sum_{l=0}^{L_n-1}\left(\mathbf{w}_{n,i,l}^H\mathbf{Q}_{n,i,l}^p\mathbf{w}_{n,i,l}^H\right) \\ &+\sum_{n\in\mathcal{N}}\sum_{i=1}^{I}\lambda_n||\mathbf{w}_{n,i,0} - \mathbf{d}_{i,\theta_n}||_2^2, \end{aligned} \quad (20)$$

where $\mathcal{W}^p$ is the estimate of $\mathcal{W}$ at the pth iteration and

$$\mathbf{Q}_{n,i,l}^p = \frac{1}{J}\sum_{j=1}^{J}\frac{\mathbf{x}_{i,j}\mathbf{x}_{i,j}^H}{r_{n,j-l}^p} \quad (21)$$

is the weighted covariance matrix of input signals at the pth iteration. The energy of the estimated source signal at the pth iteration, $r_{n,j-l}^p$, is calculated by (19) with $\mathbf{w}_{n,i,0}^p$. Now $\mathcal{W}^S$ can be updated by minimizing the auxiliary function.

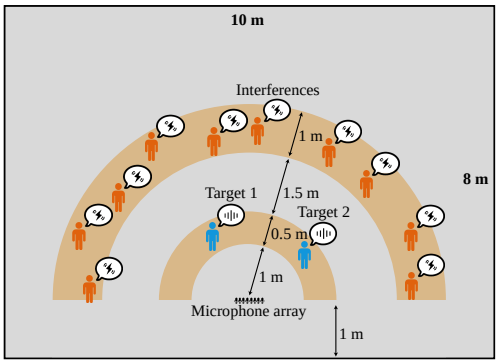

Fig. 1. Illustration of the simulation setup. Condition: $T_{60} = 0.7$ s, SNR $= 30$ dB, and SIR$\in [15\,\text{dB}, 30\,\text{dB}]$.

Firstly, for unconstrained sources, i.e., $n \notin \mathcal{N}$, the cost function (20) is identical to overdetermined IVA (OverIVA) [31]. Its optimization is known as the hybrid exact-approximate diagonalization (HEAD) problem, which can be efficiently solved with the IP algorithm [9]:

$$\tilde{\mathbf{w}}_{n,i,l}^p = \left(\mathbf{W}_i^p\mathbf{Q}_{n,i,l}^p\right)^{-1}\mathbf{e}_{\left(L_1+L_2+\ldots+L_{n-1}+l+1\right)}, \quad (22)$$

$$\mathbf{w}_{n,i,l}^{p+1} = \tilde{\mathbf{w}}_{n,i,l}^p / \left[\left(\tilde{\mathbf{w}}_{n,i,l}^p\right)^H\mathbf{Q}_{n,i,l}^p\tilde{\mathbf{w}}_{n,i,l}^p\right]^{-\frac{1}{2}}, \quad (23)$$

where $\mathbf{e}_{\left(L_1+L_2+\ldots+L_{n-1}+l+1\right)} \in \mathbb{R}^{M\times 1}$ is a unit column vector whose $(L_1 + L_2 + \ldots + L_{n-1} + l + 1)$th element equals to one.

Then, for constrained sources, i.e., $n \in \mathcal{N}$, we first update the first column of corresponding extraction matrix, $\mathbf{W}_{n,i}$ (as the spatial constraint is only imposed on $\mathbf{w}_{n,i,0}$). Because of the quadratic term, it cannot be optimized with the IP method, so we utilize the vector-wise coordinate descent (VCD) algorithm [17, 22, 33]:

$$\mathbf{p}_{n,i}^p = \left(\mathbf{W}_i^p\tilde{\mathbf{Q}}_{n,i,0}^p\right)^{-1}\mathbf{e}_{(L_1+L_2+\ldots+L_{n-1}+1)}, \quad (24)$$

$$\tilde{\mathbf{p}}_{n,i}^p = \lambda_n\left(\tilde{\mathbf{Q}}_{n,i,0}^p\right)^{-1}\mathbf{d}_{i,\theta_n}, \quad (25)$$

$$\mu_{n,i}^p = \left(\mathbf{p}_{n,i}^p\right)^H\tilde{\mathbf{Q}}_{n,i,0}^p\mathbf{p}_{n,i}^p, \quad (26)$$

$$\tilde{\mu}_{n,i}^p = \left(\mathbf{p}_{n,i}^p\right)^H\tilde{\mathbf{Q}}_{n,i,0}^p\tilde{\mathbf{p}}_{n,i}^p, \quad (27)$$

$$\mathbf{w}_{n,i,0}^{p+1} = \begin{cases}\frac{\mathbf{p}_{n,i}^p}{\sqrt{\mu_{n,i}^p}} + \tilde{\mathbf{p}}_{n,i}^p, \text{ if } \tilde{\mu}_{n,i}^p = 0, \\ \frac{\tilde{\mu}_{n,i}^p}{2\mu_{n,i}^p}\left(-1+\sqrt{1+\frac{4\mu_{n,i}^p}{|\tilde{\mu}_{n,i}^p|^2}}\right)\mathbf{p}_{n,i}^p + \tilde{\mathbf{p}}_{n,i}^p, \text{ else,}\end{cases} \quad (28)$$

where $\tilde{\mathbf{Q}}_{n,i,0}^p = \mathbf{Q}_{n,i,0}^p + \lambda_n\mathbf{I}_M$. The remaining columns of $\mathbf{W}_{n,i}$ are updated with (22) and (23).

Finally, since we dropped the noise term in (20), we rely on an orthogonal constraint (OC) [31] to separate background noise from source signals. The filtered background noise and interference signals are assumed to be orthogonal to source signals, i.e,

$$\mathbf{W}_i^{\mathrm{S}}\mathbf{C}_i\mathbf{U}_i^H = \mathbf{0}_{L\times K}, \quad (29)$$

where

$$\mathbf{C}_i = \frac{1}{J}\sum_{j=1}^{J}\mathbf{x}_{i,j}\mathbf{x}_{i,j}^H$$

is the sample-averaged SCM of the observed signals. Substituting (8) into (29), we obtain

$$\left(\mathbf{W}_i^{\mathrm{S}}\mathbf{C}_i\mathbf{E}_S^H\right)\mathbf{J}_i^H = \mathbf{W}_i^{\mathrm{S}}\mathbf{C}_i\mathbf{E}_N^H, \quad (30)$$

where $\mathbf{E}_S = [\mathbf{I}_L, \mathbf{0}_{L\times K}]$ and $\mathbf{E}_N = [\mathbf{0}_{K\times L}, \mathbf{I}_K]$. Then $\mathbf{J}_i$ can be

updated by solving (30), resulting in

$$\mathbf{J}_i^{p+1} = \left\{ \mathbf{E}_N \mathbf{C}_i \left[\left(\mathbf{W}_i^{\mathrm{S}} \right)^{p+1} \right]^H \right\} \left\{ \mathbf{E}_S \mathbf{C}_i \left[\left(\mathbf{W}_i^{\mathrm{S}} \right)^{p+1} \right]^H \right\}^{-1}. \quad (31)$$

Note that the fundamental work of [22] is extended here to the case $L_n > 1$.

4. SIMULATIONS

In this section, we study the performance of the proposed source extraction method and compare it with the source extraction method in [22, 31]. We consider a room of size $10\,\mathrm{m} \times 8\,\mathrm{m} \times 3\,\mathrm{m}$. For ease of exposition, we use the Cartesian coordinate system to denote the positions in the room and the bottom left corner of the room is chosen as the origin. An 8-element uniform linear microphone array with spacing of 5 cm is horizontally placed with its center at $(5\,\mathrm{m}, 1\,\mathrm{m}, 1\,\mathrm{m})$. Two target speakers are at the directions $(40^\circ + \epsilon_1)$ and $(120^\circ + \epsilon_2)$, respectively, where ϵ_1, ϵ_2 denote the DOA observation errors, which are assumed to be uniformly distributed in $[-5^\circ, 5^\circ]$. The horizontal distance between each source and the array is randomly generated, which is uniformly distributed in $[1\,\mathrm{m}, 1.5\,\mathrm{m}]$ as illustrated in Fig.1. There are ten interfering sources, whose positions are again randomly generated, with angles being uniformly distributed in $[0^\circ, 180^\circ]$ and the horizontal distance being uniformly distributed in $[3\,\mathrm{m}, 4\,\mathrm{m}]$. The heights of all the sources, including both targets and interferences, are also randomly generated and uniformly distributed in $[1.5\,\mathrm{m}, 1.8\,\mathrm{m}]$. The signal-to-interference ratio (SIR) is controlled to be in the range of $[15\,\mathrm{dB}, 30\,\mathrm{dB}]$.

Both target and interference speech are arbitrarily taken from the TIMIT database [34] with a sampling rate of 16 kHz. The AIRs are generated with the image model method [35], where the corresponding reverberation time, T_{60}, is approximately 0.7 s by using the Python toolkit gpuRIR [36]. The microphone signals are generated by convolving the AIRs with the clean signals taken from TIMIT. White Gaussian noise is then added at a signal-to-noise ratio (SNR) of 30 dB. One hundred Monte Carlo simulations are carried out and the average is used to measure the extraction performance. For the STFT analysis, the von Hann window of length 128 ms is used and the frame overlap is 75%. We consider two situations, the spatially informed case and the unconstrained (blind) case, i.e., $\lambda = 0$. For the unconstrained case, we consider two configurations for the proposed algorithm, CTF1 with $L_1 = L_2 = 2$ and CTF2 with $L_1 = L_2 = 3$. We choose [31] as the baseline and refer it to MTF for the unconstrained case. For the spatially informed case, the configuration of L for the proposed algorithm is set analogously to the unconstrained case and we refer them to CTF1-GC and CTF2-GC. We choose MTF-GC in [22] as the baseline for the spatially informed case. For the spatially informed methods, we set $\lambda_n^{\mathrm{init}} = 3$ and decrease the influence of the prior according to $\lambda_n^p = \left(1 - \left(\frac{p-1}{P}\right)\right)^\alpha \lambda_n^{\mathrm{init}}$. Where P is the number of total iterations. The parameter α is set to 3 for the first 5 iterations and 1 for the remaining iterations. The extraction performance is measured in terms of the improvement of the signal-to-distortion ratio (SDR), ΔSDR [37].

The convergence behavior of all the studied methods depicted in Fig. 2 shows that in both constrained and unconstrained conditions, the proposed methods based on the CTF model outperform the counterparts based on the MTF model in reverberant conditions. We define two runtime factors as

$$\gamma_n^{\mathrm{MTF}} = \frac{t_n}{t_{\mathrm{MTF}}}, \quad \gamma_n^{\mathrm{GC}} = \frac{t_n}{t_n^{\mathrm{UC}}}, \quad (32)$$

where t_n is the runtime per iteration of the compared method, t_{MTF} is the runtime per iteration of MTF, and t_n^{UC} is the runtime per iteration of the corresponding unconstrained method. The runtime

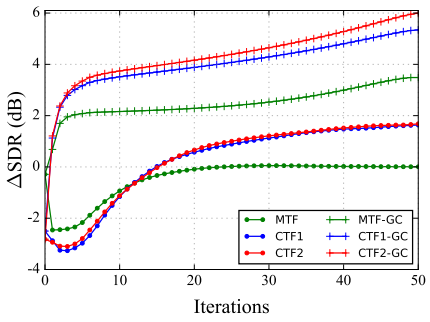

Fig. 2. Convergence behavior of the proposed method and compared methods in terms of average ΔSDR with two target sources.

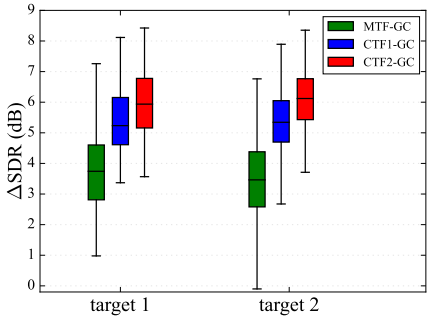

Fig. 3. ΔSDR of the proposed method and the baseline.

factors of all algorithms are shown in Table 1. The runtime factor γ_n^{GC} for MTF-GC, CTF1-GC and CTF2-GC are 2.02, 1.62 and 1.45, respectively. As can be seen in Fig. 2, spatially informed methods can achieve reasonable extraction performance within 5 iterations while the blind ones need more than 20 iterations to converge and are likely to fail to extract target signals. Therefore, the spatially informed methods not only yield significantly better extraction performance but also converge faster than the blind ones, showing the efficacy of (24)−(28).

Table 1. Runtime factors of algorithms

	MTF	CTF1	CTF2	MTF-GC	CTF1-GC	CTF2-GC
γ_n^{MTF}	1	1.82	2.62	2.02	2.95	3.81
γ_n^{GC}	1	1	1	2.02	1.62	1.45

The extraction performance of the spatially informed methods after 50 iterations is shown in Fig. 3. The number of cases corresponding to $\Delta\mathrm{SDR} \geq 5$ dB achieved by MTF-GC, CTF1-GC and CTF2-GC, are $18\%, 59\%, 79\%$ for target 1 and $16\%, 62\%, 90\%$ for target 2, respectively. Hence, the proposed spatially informed extraction method outperforms the state-of-the-art MTF. Even with $L_n = 2$, the extraction performance is significantly improved relative to the MTF baseline.

5. CONCLUSION

This paper studied the problem of source extraction in reverberant environments. Unlike the existing spatially informed extraction methods, which are based on the MTF model, we adopted the CTF model, which is able to better model the array signals in the STFT domain. We then developed an algorithm to estimate the demixing system with this CTF model. Simulations were carried out and the results show that the presented method has achieved reasonable extraction performance even in highly reverberant environments.

6. REFERENCES

[1] J. Benesty, M. M. Sondhi, Y. Huang, *et al.*, *Springer handbook of speech processing*, vol. 1. New York, NY, USA: Springer, 2008.

[2] S. Makino, *Audio Source Separation*. Springer, 2018

[3] P. Comon, "Independent component analysis, a new concept?," *Signal Process.*, vol. 36, no. 3, pp. 287–314, Apr. 1994.

[4] D. T. Pham, "Blind separation of instantaneous mixture of sources via an independent component analysis," *IEEE Trans. on Signal Process.*, vol. 44, no. 11, pp. 2768–2779, 1996.

[5] A. Hyvärinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Comput.*, vol. 9, no. 7, pp. 1483–1492, Oct. 1997.

[6] S. C. Douglas, H. Sawada, and S. Makino, "A spatio-temporal FastICA algorithm for separating convolutive mixtures," *Proc. IEEE ICASSP*, 2005, pp. 165–168.

[7] H. Sawada, S. Araki, and S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS," *Proc. IEEE Int. Sym. Circuits and Systems*, 2007, pp. 3247–3250.

[8] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind Source Separation Exploiting Higher-order Frequency Dependencies," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 70–79, 2007

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

[10] Y. Liang, S. 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, Apr. 2012.

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

[12] 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.

[13] 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, 2018.

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

[15] G. Huang, J. Benesty, and J. Chen, "On the design of frequency-invariant beampatterns with uniform circular microphone arrays," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 5, pp. 1140–1153, 2017.

[16] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From blind to guided audio source separation: How models and side information can improve the separation of sound," *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 107–115, May 2014.

[17] 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.

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

[19] D. Li, G. Huang, Y. Lei, J. Chen, and J. Benesty, "Robust source separation with differential microphone arrays and lndependent low-rank matrix analysis," in *Proc. EUSIPCO*, 2020.

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

[21] A. Brendel and W. Kellermann, "Informed source extraction based on independent vector analysis using eigenvalue decomposition," in *Proc. EUSIPCO*, 2021, pp. 875–879.

[22] 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, 2020.

[23] 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.

[24] N. Q. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1830–1840, Sept. 2010.

[25] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999.

[26] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 550–563, Mar. 2010.

[27] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 5, pp. 971–982, May 2013.

[28] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1626–1641, Sept. 2016.

[29] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 546–555, May 2009.

[30] T. Wang, F. Yang, and J. Yang, "Convolutive transfer function-based multichannel nonnegative matrix factorization for overdetermined blind source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 802–815, 2022.

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

[32] E. Vincent, S. Araki, F. J. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, B. V. Gowreesunker, D. Lutter, and N. Q. K. Duong, "The signal separation evaluation campaign (2007-2010): Achievements and remaining challenges," *Sig. Proc.*, vol. 92, no. 8, pp. 1928–1936, 2012.

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

[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"*, vol. 94. Nat. Inst. Standards Technol., Gaithersburg, MD, USA, Tech. Rep. NISTIR 4930, 1993.

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

[36] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "gpuRIR: A python library for room impulse response simulation with GPU acceleration," *Multimedia Tools Appl.*, vol. 80, no. 4, pp. 5653–5671, Oct. 2020.

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