Accelerated Convolutive Transfer Function-Based Multichannel NMF Using Iterative Source Steering

Xuemai Xie*, Xianrui Wang*, Liyuan Zhang*, Yichen Yang* and Shoji Makino*
* Waseda University, Japan
E-mails: xuemaixie@ruri.waseda.jp, wangxianrui97@gmail.com
ly.zhang@akane.waseda.jp, yang_yichen@mail.nwpu.edu.cn
s.makino@waseda.jp

***Abstract*—Among numerous blind source separation (BSS) methods, convolutive transfer function-based multichannel nonnegative matrix factorization (CTF-MNMF) has demonstrated strong performance in highly reverberant environments by modeling multi-frame correlations of delayed source signals. However, its practical deployment is hindered by the high computational cost associated with the iterative projection (IP) update rule, which requires matrix inversion for each source. To address this issue, we propose an efficient variant of CTF-MNMF that integrates iterative source steering (ISS), a matrix inversion-free update rule for separation filters. Experimental results show that the proposed method achieves comparable or superior separation performance to the original CTF-MNMF, while significantly reducing the computational complexity.**

***Index Terms*—Blind source separation, nonnegative matrix factorization, convolutive transfer function, fast algorithm.**

I. Introduction

Blind source separation (BSS) is a technique that recovers the source signals from only the observed sensor signals, without prior knowledge of the mixing process or the source characteristics [1]–[3]. Based on how signals are mixed, BSS can be categorized into two types: instantaneous BSS and convolutional BSS [4]–[6]. Although the instantaneous mixing model is computationally efficient and conceptually simple, it fails to adequately capture real-world reverberation structures, resulting in terrible performance in environments with pronounced delays and reflections [1], [7], [8].

As popular instantaneous BSS techniques, auxiliary function-based independent vector analysis (AuxIVA) [9]–[12] and independent low-rank matrix analysis (ILRMA) [13] are widely used due to their stable separation performance. ILRMA replaces the source model used in AuxIVA with the nonnegative matrix factorization (NMF) [14] model to capture deeper harmonic structures in sources. Both AuxIVA and ILRMA adopt the rank-1 spatial model to enable efficient separation, where each source's spatial image is modeled as a scaled steering vector. However, these algorithms will remain effective only when the Short-Time Fourier Transform(STFT) window fully encompasses the dominant part of the acoustic impulse response(AIR). Once this condition is violated, e.g., in highly reverberant environments, their performance quickly degrades.

Recently, convolutive transfer function-based multichannel non-negative matrix factorization (CTF-MNMF) [15]–[17] has shown superior separation performance effectively, especially in highly reverberant conditions. By explicitly modeling multi-frame correlations of delayed source signals, the convolutive transfer function(CTF) model retains a finite set of delayed taps of the CTF filter in the STFT domain. Therefore, early reflections are integrated into an extended instantaneous mixing matrix, which enables efficient instantaneous BSS updates while modeling the actual mixing process more accurately. Besides, since the CTF model can efficiently model long AIR using short-time frames, CTF-MNMF also relaxes the restriction on STFT window lengths. However, CTF-MNMF suffers from significant computational complexity due to the introduction of additional parameters, especially since its iterative projection (IP) [10] based demixing filter update requires matrix inversion for each source. This computational burden increases substantially with longer CTF filters.

To address these computational challenges, we propose an efficient variant of CTF-MNMF [15] by integrating the iterative source steering (ISS) [18] algorithm, termed CTF-MNMF-ISS. The ISS update rule completely avoids matrix inversion, significantly reducing computational complexity. Experimental results demonstrate that the proposed CTF-MNMF-ISS achieves comparable or superior separation performance relative to the original CTF-MNMF-IP method, while substantially enhancing computational efficiency.

II. Signal Model And Problem Formulation

Assume that N sources are recorded by M microphones. For the overdetermined condition where $M > N$, the signals observed at the m-th microphone with time index t is expressed as

$$x_m(t) = \sum_{n=1}^{N} h_{m,n} * s_n(t), \tag{1}$$

where $h_{m,n}$ is the time-invariant AIR from the n-th source to the m-th microphone, $x_m(t)$ and $s_n(t)$ are the m-th microphone signal and the n-th source signal, respectively, and $*$ represents linear convolution. The STFT of the observed signals (1) is derived as a sum of linear convolutions using the

979-8-3315-7206-8/25/$31.00 ©2025 IEEE

CTF assumption [8], [15], [19], [20]

$$x_{m,i,j} = \sum_{n=1}^{N} \sum_{l=0}^{L_n-1} h_{m,n,i,l}\, s_{n,i,j-l}, \tag{2}$$

where $i = 1 \dots I$ and $j = 1 \dots J$ are the frequency index and time-frame indexes, respectively, with I and J being the total number of frequency bins and time frames, $x_{m,i,j}$ and $s_{n,i,j}$ are the STFTs of $x_m(j)$ and $s_n(j)$, respectively, $h_{m,n,i,l}$ is the band-to-band filter coefficient and L_n is the length of the CTF filter. For simplicity, we rewrite (2) in vector form as

$$\mathbf{x}_{i,j} = \tilde{\mathbf{H}}_{\mathbf{i}}\, \mathbf{s}_{i,j}, \tag{3}$$

where

$$\begin{aligned}
\mathbf{x}_{i,j} &= [x_{1,i,j},\, x_{2,i,j},\, \cdots,\, x_{M,i,j}]^\top \in \mathbb{C}^{M\times 1},\\
\mathbf{h}_{n,i,l} &= [h_{1,n,i,l},\, h_{2,n,i,l},\, \cdots,\, h_{M,n,i,l}]^\top \in \mathbb{C}^{M\times 1},\\
\mathbf{H}_{n,i} &= [\mathbf{h}_{n,i,0},\, \mathbf{h}_{n,i,1},\, \cdots,\, \mathbf{h}_{n,i,L-1}] \in \mathbb{C}^{M\times L_n},\\
\tilde{\mathbf{H}}_{\mathbf{i}} &= [\mathbf{H}_{1,i},\, \mathbf{H}_{2,i},\, \cdots,\, \mathbf{H}_{N,i}] \in \mathbb{C}^{M\times L},\\
\tilde{\mathbf{s}}_{n,i,j} &= [s_{n,i,j},\, s_{n,i,j-1},\, \cdots,\, s_{n,i,j-L_n+1}] \in \mathbb{C}^{1\times L_n},\\
\mathbf{s}_{i,j} &= [\tilde{\mathbf{s}}_{1,i,j},\, \tilde{\mathbf{s}}_{2,i,j},\, \cdots,\, \tilde{\mathbf{s}}_{N,i,j}]^\top \in \mathbb{C}^{L\times 1}.
\end{aligned}$$

Here, $\mathbf{H}_i$ is the mixing matrix for the i-th frequency bin, $\mathbf{s}_{i,j}$ stacks the delayed source signals, and $(\cdot)^\top$ denotes the transpose. Following [15], we set $L = \sum_{n=1}^{N} L_n = M$. This choice makes $\mathbf{H}_i \in \mathbb{C}^{M\times M}$ square and full-rank, a prerequisite for the IP update rule, which requires matrix inversion at every frequency bin. Consequently, the demixing matrix can be defined as $\mathbf{W}_i = \mathbf{H}_i^{-1}$ as

$$\mathbf{W}_i = \left[\tilde{\mathbf{W}}_{1,i},\, \cdots,\, \tilde{\mathbf{W}}_{N,i}\right]^H \in \mathbb{C}^{L\times M}$$

where

$$\tilde{\mathbf{W}}_{n,i} = \left[\mathbf{w}_{n,0,i},\, \cdots,\, \mathbf{w}_{n,L_n-1,i}\right]^H \in \mathbb{C}^{M\times L_n}$$

is the group of filters corresponding to source n and each $\mathbf{w}_{n,l,i}$ is an M-dimensional column vector. $(\cdot)^H$ stands for Hermitian transpose. Now, the demixing process is denoted as

$$y_{n,i,j,l} = \mathbf{w}_{n,l,i}^H \mathbf{x}_{i,j} \quad \text{for } n = 1, \dots, N, \tag{4}$$

where $y_{n,i,j,l}$ is the estimated source signal with l taps delay. Each source is modeled as a complex Gaussian random variable with zero mean and time-varying variance $\lambda_{n,i,j}$. The power spectral density (PSD) is represented using NMF as [21]

$$\lambda_{n,i,j-l} = \sum_{k=1}^{K_n} b_{n,i,k}\, v_{n,k,j-l},$$

where $b_{n,i,k}$ and $v_{n,k,j-l}$ are the NMF basis and activation components for the n-th source, respectively, with $k = 1, \dots, K_n$, where K_n is the number of latent spectral bases for source n.

The objective function is obtained by calculating the negative log-likelihood function as in [15]:

$$\begin{aligned}
\mathcal{L}(\mathbf{W}, \mathbf{\Lambda}) = \sum_{i,j,n,l} &\left(\log \lambda_{n,i,j-l} + \frac{|y_{n,i,j,l}|^2}{\lambda_{n,i,j-l}}\right)\\
&-2J \sum_i \log|\det \mathbf{W}_i| + cst.
\end{aligned} \tag{5}$$

The task is now transformed into minimizing the objective function (5) with respect to the demixing matrices $\{\mathbf{W}_i\}$ and PSD parameters $\{\lambda_{n,i,j}\}$, which yields the following optimization formulation:

$$\{\mathbf{W}^\star, \mathbf{\Lambda}^\star\} = \arg\min_{\mathbf{W},\mathbf{\Lambda}} \mathcal{L}(\mathbf{W}, \mathbf{\Lambda}). \tag{6}$$

III. Proposed Method

A. Optimization algorithm

1) Update of $\mathbf{W}_i$*:* In the following, we deduce update rules for optimizing (5) based on the auxiliary function technique [10]. It can be obtained that

$$\mathcal{L}^+ = -2\sum_{i=1}^{I} \log|\det \mathbf{W}_i| + \sum_{i=1}^{I}\sum_{n,l} \mathbf{w}_{n,l,i}^H\, \mathbf{Q}_{n,l,i}\, \mathbf{w}_{n,l,i}, \tag{7}$$

where the weighted covariance matrix $\mathbf{Q}_{n,l,i}$ is defined as

$$\mathbf{Q}_{n,l,i} = \frac{1}{J}\sum_{j=1}^{J} \frac{\mathbf{x}_{i,j}\,\mathbf{x}_{i,j}^H}{\lambda_{n,i,j-l}}. \tag{8}$$

The original IP-based Optimization goes as in [15]

$$\mathbf{w}_{n,l,i} \leftarrow \left(\mathbf{W}_i\, \mathbf{Q}_{n,l,i}\right)^{-1} \mathbf{e}_{(L_1+L_2\cdots+L_{n-1}+l+1)}, \tag{9}$$

$$\mathbf{w}_{n,l,i} \leftarrow \mathbf{w}_{n,l,i}\left(\mathbf{w}_{n,l,i}^H \mathbf{Q}_{n,l,i} \mathbf{w}_{n,l,i}\right)^{-1/2}, \tag{10}$$

where $\mathbf{e}_{(L_1+L_2\cdots+L_{n-1}+l+1)}$ is a unit column vector whose $(L_1+L_2\cdots+L_{n-1}+l+1)$th element equals to one. To simplify the ISS update formulation, we flatten the two-dimensional index (n, l) into a single index $r = (L_1 + L_2 \cdots + L_{n-1}) + l + 1$, where $n = 1, \dots, N$, $l = 0, \dots, L_n - 1$, and $r = 1, \dots, L$. Hence, $\mathbf{w}_{n,l,i}$ can be denoted as $\mathbf{w}_{r,i}$ for notational convenience. The proposed method employs the ISS updating rule [18], which updates the entire filter using a rank-1 matrix as

$$\mathbf{W}_i \leftarrow \mathbf{W}_i - \mathbf{z}_{r,i}\, \mathbf{w}_{r,i}^H, \tag{11}$$

where r is the index indicating the rank-1 updates applied sequentially to each source and delay tap, and $\mathbf{z}_{r,i} = [z_{1,r,i},\, \dots,\, z_{L,r,i}]^\top \in \mathbb{C}^{L\times 1}$ is a vector to estimate and $(\cdot)^*$ denotes complex conjugate.

To derive the optimal update direction, we substituting the rank-1 update (11) into the auxiliary objective function (7), we get the new optimization objective:

$$\begin{aligned}
\mathcal{L}_{\text{ISS}}(\mathbf{z}_{r,i}) = -2\sum_{i=1}^{I} \log\left|\det\left(\mathbf{W}_i - \mathbf{z}_{r,i}\,\mathbf{w}_{r,i}^H\right)\right| + \sum_i \sum_{p=1}^{L}\\
\left(\mathbf{w}_{p,i} - z_{p,r,i}^*\,\mathbf{w}_{r,i}\right)^H \mathbf{Q}_{p,i}\left(\mathbf{w}_{p,i} - z_{p,r,i}^*\,\mathbf{w}_{r,i}\right),
\end{aligned} \tag{12}$$

where p is a dummy index ranging from 1 to L that enumerates all rows, by the matrix determinant lemma $\det\left(\mathbf{A} - \mathbf{u}\mathbf{z}^H\right) = \det(\mathbf{A})\left(1 - \mathbf{z}^H\mathbf{A}^{-1}\mathbf{u}\right)$, we have

$$\det\left(\mathbf{W}_i - \mathbf{z}_{r,i}\mathbf{w}_{r,i}^H\right) = \det\left(\mathbf{W}_i\right)\left(1 - z_{r,r,i}\right). \tag{13}$$

Taking the derivative of $\mathcal{L}$ with respect to $z_{p,r,i}^*$ and setting it to zero, we consider two cases.
First, when $r \neq p$, we can obtain

$$\frac{\partial \mathcal{L}_{\text{ISS}}}{\partial z_{p,r,i}^*} = -\mathbf{w}_{r,i}^H \mathbf{Q}_{p,i} \mathbf{w}_{p,i} + z_{p,r,i}\, \mathbf{w}_{r,i}^H \mathbf{Q}_{p,i} \mathbf{w}_{r,i} \tag{14}$$

Next, when $r = p$, the partial derivative of $\mathcal{L}$ is

$$\frac{\partial \mathcal{L}_{\text{ISS}}}{\partial z_{p,r,i}^*} = \frac{2}{1 - z_{r,r,i}} - 2(1 - z_{r,r,i})\, \mathbf{w}_{r,i}^H \mathbf{Q}_{r,i} \mathbf{w}_{r,i} \tag{15}$$

Setting these expressions to zero yields the closed-form solution, then the update of $z_{r,p,j}$ can be obtained that

$$z_{p,r,i} = \begin{cases} \dfrac{\mathbf{w}_{r,i}^H \mathbf{Q}_{p,i} \mathbf{w}_{p,i}}{\mathbf{w}_{r,i}^H \mathbf{Q}_{p,i} \mathbf{w}_{r,i}}, & p \neq r, \\ 1 - \left(\mathbf{w}_{r,i}^H \mathbf{Q}_{r,i} \mathbf{w}_{r,i}\right)^{-1/2}, & p = r. \end{cases} \tag{16}$$

Then, by using the original ISS update rule, the demixing matrix W is solved.

2) Update of $\mathbf{\Lambda}$*:* With $\mathbf{W}$ fixed, minimising $\mathcal{L}$ over $\{b_{n,i,k}, v_{n,k,j}\}$ is equivalent to minimising a sum of Itakura–Saito divergences between the $|y_{n,i,j,l}|^2$ and $\lambda_{n,i,j-l}$. Using a Majorize-Minimization (MM) framework yields the following Multiplicative Update (MU) rules [15]. For each n, i, k, the update of the basis and the activation can be obtained by

$$b_{n,i,k} \leftarrow b_{n,i,k} \sqrt{\frac{\sum_{j=1}^{J} \sum_{l=0}^{L-1} |y_{n,i,j,l}|^2\, v_{n,k,j-l} \lambda_{n,i,j-l}^{-2}}{\sum_{j=1}^{J} \sum_{l=0}^{L-1} v_{n,k,j-l} \lambda_{n,i,j-l}^{-1}}}, \tag{17}$$

$$v_{n,k,j} \leftarrow v_{n,k,j} \sqrt{\frac{\sum_{i=1}^{I} |y_{n,i,j,0}|^2\, b_{n,i,k} \lambda_{n,i,j}^{-2}}{\sum_{i=1}^{I} b_{n,i,k} \lambda_{n,i,j}^{-1}}}. \tag{18}$$

To resolve scale ambiguity among $b_{n,i,k}$, $v_{n,k,j}$ and the demixing matrix $\mathbf{W}$, we compute the average power:

$$\mu_{n,l} = \sqrt{\frac{1}{IJ} \sum_{i,j} |y_{n,i,j,l}|^2}, \tag{19}$$

then apply rescaling:

$$y_{n,i,j,l} \leftarrow y_{n,i,j,l}\, \mu_{n,l}^{-1}, \tag{20}$$

$$b_{n,i,k} \leftarrow b_{n,i,k}\, \mu_{n,0}^{-2}. \tag{21}$$

The same scaling is also applied to the corresponding rows of $\mathbf{W}$ to ensure consistent signal energy.

B. Source image estimation

To avoid the spatial distortion caused by using (4) when estimating the source signal in a reverberant environment, after estimating the demixing matrix and the source PSDs, we reconstruct the spatial images $\hat{\mathbf{c}}_{n,i,j} \in \mathbb{C}^M$ using a multichannel Wiener filter (MWF) [22]. The goal is to minimize the mean squared error (MSE) between the estimated and true spatial images:

$$\mathbf{M}_{n,i,j}^{\text{opt}} = \arg \min_{\mathbf{M}_{n,i,j}} \mathbb{E}\left[\|\mathbf{c}_{n,i,j} - \mathbf{M}_{n,i,j} \mathbf{x}_{i,j}\|_2^2\right], \tag{22}$$

where $\mathbf{c}_{n,i,j} \in \mathbb{C}^M$ is the source image. The optimal estimator is given by:

$$\mathbf{M}_{n,i,j}^{\text{opt}} = \mathbb{E}[\mathbf{c}_{n,i,j} \mathbf{x}_{i,j}^H] \cdot \mathbb{E}^{-1}[\mathbf{x}_{i,j} \mathbf{x}_{i,j}^H]. \tag{23}$$

Under the proposed CTF-based spatial model, (23) becomes

$$\begin{aligned} \hat{\mathbf{c}}_{n,i,j} &= \mathbf{M}_{n,i,j}^{\text{opt}} \mathbf{x}_{i,j} \\ &= \mathbf{H}_{n,i} \mathbf{\Lambda}_{n,i,j} \mathbf{H}_{n,i}^H \left(\mathbf{H}_i \mathbf{\Lambda}_{i,j} \mathbf{H}_i^H\right)^{-1} \mathbf{x}_{i,j}, \end{aligned} \tag{24}$$

where $\mathbf{\Lambda}_{n,i,j} \in \mathbb{R}^{L \times L}$ is the diagonal PSD matrix of source n. Therefore, the MWF serves as the final stage to recover the spatial images.

IV. Computational Complexity Analysis

This section compares the computational costs of CTF-MNMF-IP and CTF-MNMF-ISS with respect to the demixing matrix $\mathbf{W}i$. As both methods employ identical MU rules for the NMF parameters, the computational complexity of this part is the same. Thus, the difference in the overall computational burden stems solely from the optimization strategy applied to $\mathbf{W}i$.

CTF-MNMF-IP updates the demixing matrix $\mathbf{W}_i$ by solving a full linear system. Inverting an $L \times M$ matrix in (9) requires $\mathcal{O}(M^3)$ floating-point operations. Repeating this for all signals across I frequency bins and L sources yields

$$\mathcal{C}_{\text{IP}} \propto \mathcal{O}\big(ILM^3\big). \tag{25}$$

CTF-MNMF-ISS rewrites the update as a rank-one steering step so the costly inversion disappears. The dominant operation becomes the matrix–vector product plus rank-one correction, whose computational complexity scales quadratically as $\mathcal{O}(M^2)$. Accounting again for all signals and I frequencies gives

$$\mathcal{C}_{\text{ISS}} \propto \mathcal{O}\big(ILM^2\big). \tag{26}$$

which is an order of magnitude lower than that of the IP method.

V. Experiment

In this section, we will compare the performance of CTF-ISS with the traditional method in [15]. For simplicity, we omit MNMF in this section.

A. Experimental setup

The observation signals are generated by convolving speech signals from the TIMIT database [23]. Each mixed signal consists of two speech segments, randomly selected from different speakers, concatenated to form an 8-second clean speech signal. To validate the performance under realistic reverberant environments, we utilize impulse responses from the RWCP dataset, specifically E2A with a reverberation time of $\text{RT}_{60} = 300$ ms, JR2 with $\text{RT}_{60} = 470$ ms, and E2B with $\text{RT}_{60} = 1300$ ms.

The geometric configuration used in our experiments is illustrated in Fig. 1. Two sound sources are positioned 2 meters away from the microphone array center, forming an

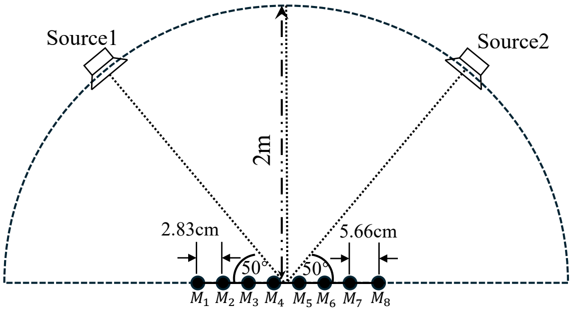

Fig. 1. Illustration of the simulation setup.

TABLE I
MICROPHONE CONFIGURATIONS AND FILTER LENGTHS

M	Filter length	Microphones configurations
4	$L_n = 2$	M_3, M_4, M_5, M_6
6	$L_n = 3$	M_2, M_3, M_4, M_5, M_6, M_7
8	$L_n = 4$	M_1, M_2, M_3, M_4, M_5, M_6, M_7, M_8

angular separation of 100° (at azimuth angles $-50°$ and $+50°$). The microphone array consists of eight omnidirectional microphones (M_1 to M_8) arranged in a linear formation. The experimental settings are identical to those of the conventional method. The specific microphone configurations for different array sizes utilized in this study is summarized in Table I. All recordings are sampled at 16 kHz. The time-frequency representation is obtained using the STFT with a 1024-point Hann window and a hop size of 25%. All of the separation matrices $\mathbf{W}_i$ are initialized as identity matrices, and the number of iterations is fixed at 100. The number of bases K_n is set to 3. The number of microphones used varies among 4, 6, and 8. Each microphone is regarded as an independent observation channel, and two sources are active in all scenarios. The CTF filter length L_n is set to 2, 3, and 4 taps for the 4-, 6-, and 8-channel microphone arrays, respectively, as shown in Table I. All experiments are conducted on a laptop with an AMD Ryzen 7 5800H CPU.

B. Experimental results

Figure 2 compares CTF-IP (red) and the proposed CTF-ISS (blue) in terms of SDR-improvement for three reverberation times (RT_{60} = 300, 470, and 1300 ms). Under all conditions, the median SDR decreases with increasing reverberation time, demonstrating the inherent difficulty of source separation in highly reverberant environments. Increasing the number of microphones consistently improves SDR performance by around 2–4 dB, with the most notable improvement observed at 1300 ms, where an 8-channel configuration outperforms the 4-channel case by nearly 4 dB.

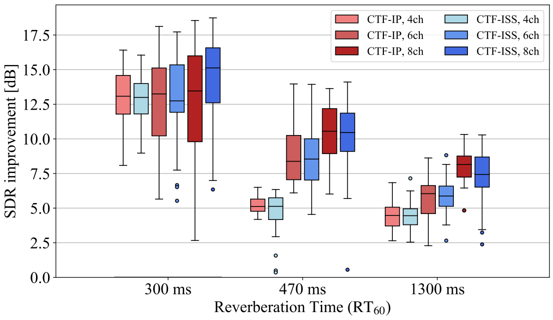

Fig. 2. Average SDR improvement under different reverberation and microphone setups.

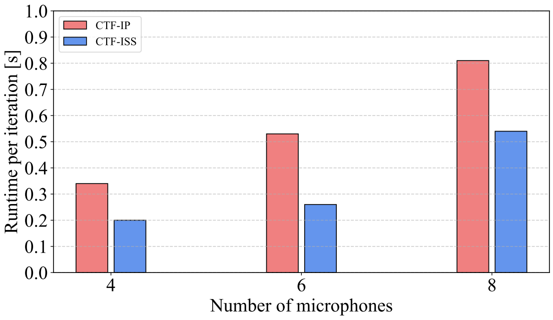

Fig. 3. Average runtime under different microphone setups.

Regarding the optimization strategy, CTF-ISS maintains comparable median performance to CTF-IP, with differences never exceeding 1 dB. However, CTF-ISS exhibits narrower inter-quartile ranges and fewer outliers, particularly noticeable at longer reverberation times of 470 ms and 1300 ms. This improvement indicates that ISS updates provide robustness and lessen sensitivity to initial conditions, yielding more consistent separation performance. Fig 3 further analyzes the computational efficiency of the two methods by presenting their average runtime across different microphone configurations. The proposed CTF-ISS consistently demonstrates lower computational costs compared to CTF-IP. Specifically, CTF-ISS achieves runtime reductions of approximately 41% , 51%, and 33%, respectively, for the 4-channel, 6-channel, and 8-channel setups. These significant reductions underscore the computational advantage of the ISS-based approach, which is particularly beneficial for real-time processing applications or systems with limited computational resources. Besides the CPU saving, ISS avoids repeated matrix inversions, leading to smaller memory usage and better numerical stability. Consequently, the proposed CTF–ISS variant achieves essentially the same separation accuracy as CTF–IP while running significantly faster and exhibiting lower run-to-run variance. The experimental results confirm that the proposed CTF-MNMF-ISS method not only preserves high-quality source separation comparable to conventional method but also offers enhanced stability and significantly reduced computational complexity, making it particularly advantageous as the number of microphones increases or under challenging reverberant conditions.

VI. Conclusions

In this paper, we proposed an accelerated CTF-MNMF algorithm for overdetermined blind source separation, named

CTF-MNMF-ISS. By integrating the iterative source steering approach, we successfully circumvented the numerical instability associated with iterative projection updates, thus significantly enhancing the practicality and robustness of the original algorithm. CTF-MNMF-ISS substantially reduces computational complexity, achieving a runtime reduction of approximately 40%. It simultaneously maintains or even surpasses the separation performance of the original IP-based method. Furthermore, the ISS-based updates exhibited greater numerical stability and robustness against variations in initialization, particularly beneficial in highly reverberant environments.

References

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

[2] S. Makino, T.-W. Lee, and H. Sawada, *Blind Speech Separation*. Berlin, Germany: Springer, 2007.

[3] X. Wang, N. Pan, J. Benesty, and J. Chen, "On multiple-input/binaural-output antiphasic speaker signal extraction," in *Proc. IEEE ICASSP*, 2023, pp. 1–5.

[4] S. C. Douglas, H. Sawada, and S. Makino, "Natural gradient multichannel blind deconvolution and speech separation using causal FIR filters," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 92–104, 2004.

[5] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 120–134, 2004.

[6] Y. Yang, X. Wang, W. Zhang, and J. Chen, "Independent vector analysis assisted adaptive beamforming for speech source separation with an acoustic vector sensor," in *Proc. IEEE IWAENC*, 2022, pp. 1–5.

[7] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York, USA: Wiley, 2001.

[8] 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, pp. 1–5.

[9] T. Kim, I. Lee, and T.-W. Lee, "Independent vector analysis: Definition and algorithms," in *Proc. Asilomar Conf. Signals Syst. Comput.*, Pacific Grove, CA, USA, Nov. 2006, pp. 1393–1396.

[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] A. Hiroe, "Solution of permutation problem in frequency domain ICA using multivariate probability density functions," in *Proc. Int. Conf. Independent Component Anal. Blind Signal Separation*, 2006, pp. 601–608.

[12] Y. Yang, X. Wang, A. Brendel, W. Zhang, W. Kellermann, and J. Chen, "Geometrically constrained source extraction and dereverberation based on joint optimization," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2023, pp. 41–45.

[13] 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, 2016.

[14] 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, 1999.

[15] 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.

[16] X. Wang, Y. Yang, A. Brendel, T. Ueda, S. Makino, J. Benesty, W. Kellermann, and J. Chen, "On semi-blind source separation-based approaches to nonlinear echo cancellation based on bilinear alternating optimization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 2973–2987, May 2024.

[17] K. Lu, X. Wang, T. Ueda, S. Makino, and J. Chen, "A computationally efficient semi-blind source separation approach for nonlinear echo cancellation based on an element-wise iterative source steering," in *Proc. IEEE ICASSP*, 2024, pp. 756–760.

[18] R. Scheibler and N. Ono, "Fast and stable blind source separation with rank-1 updates," in *Proc. IEEE ICASSP*, 2020, pp. 236–240.

[19] Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1305–1319, May 2007.

[20] 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, 2009.

[21] K. Sekiguchi, Y. Bando, A. A. Nugraha, K. Yoshii, and T. Kawahara, "Fast multichannel nonnegative matrix factorization with directivity-aware jointly-diagonalizable spatial covariance matrices for blind source separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 2610–2625, 2020.

[22] 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, 2013.

[23] J. S. Garofolo, L. F. Lamel, W. M. Fisher, D. S. Pallett, N. L. Dahlgren, V. Zue, and J. G. Fiscus, "TIMIT acoustic-phonetic continuous speech corpus," *Linguistic Data Consort.*, 1993.