Accelerate Algorithm of Simultaneous Source Separation and Dereverberation based on Time-decorrelation and Iterative Source Steering

Xun Tian, Boyuan Wang, Jiahui Hu, Shoji Makino

Waseda University
2-7 Hibikino, Wakamatsu-ku, Kitakyushu, Fukuoka 808-0135, Japan
E-mail: {xun_tian@fuji, t.ueda@akane, boyuan@ruri, hujiahui22@toki, s.makino@}.waseda.jp

Abstract

This paper introduces a hybrid method to accelerate simultaneous source separation and dereverberation, addressing limitations in existing methods. The original methods, Weighted Prediction Error (WPE) combined with Independent Vector Analysis (IVA) using Iterative Source Steering (ISS) and IVA with Time-decorrelation (IVA-T) using ISS, exhibit certain drawbacks: WPE-IVA-ISS suffers from slow convergence due to matrix inversion in the dereverberation part, while IVA-T-ISS achieves a completely matrix-inversion-free algorithm but delivers lower source separation performance than WPE-IVA-ISS. To overcome these issues, the proposed method integrates Efficient IVA-T-ISS (EIVA-T-ISS) and WPE-IVA-ISS for rapid convergence and enhanced separation performance. Experimental results demonstrate that this hybrid method achieves significantly faster convergence while maintaining source separation performance comparable to WPE-IVA-ISS, offering a balanced and efficient solution for blind source separation tasks.

1. Introduction

Independent Vector Analysis (IVA) [1, 2] and Weighted Prediction Error (WPE) [3] are famous methods for blind source separation and blind dereverberation, respectively. To combine these processes, two main approaches have been proposed: 1) alternately optimizing source separation and dereverberation in different frameworks, known as WPE-IVA [4], and 2) optimizing them in a joint framework, known as IVA with Time-decorrelation (IVA-T) [1] [6]. Both methods can separate individual source signals from a microphone observation without prior information.

Previous research has focused on reducing computational complexity in updating parameters for source separation and dereverberation. This is because both IVA-T and WPE-IVA include matrix inversions, which increase computational complexity as we increase the number of microphones. For that problem, Iterative Source Steering (ISS) [7] has been highly attracted. ISS enables updating parameters without matrix inversions and reduces the computational complexity while maintaining source separation performance in the field of source separation. By combining this ISS, WPE-IVA-ISS[2] [8] and IVA-T-ISS [9] have been proposed, which contribute to reducing their computational complexity.

Unlike the previous research, this research focuses on improving the convergence speed of parameters (distinguishing from computational complexity) and source separation performance after convergence. For example, WPE-IVA-ISS [8] still has a low convergence speed for their parameter updating since it still holds matrix inversion in the dereverberation part. On the other hand, IVA-T-ISS [9] has achieved a completely matrix-inversion-free algorithm, so its computational complexity is low. However, the parameter convergence speed is slower than before using ISS (e.g., IVA-T) [9], and the source separation performance after convergence is lower than WPE-IVA-ISS [8]. Although the slow convergence speed has been recovered by Efficient-IVA-T-ISS (EIVA-T-ISS) [10], which updates only the parameters for source separation during the iterative updates, the source separation performance after convergence is still lower than WPE-IVA-ISS [8]. In addition to the above discussions, it is worth comparing methods based on WPE-IVA ([4] and [8]) and those based on IVA-T ([6], [9], and [10]) in terms of convergence speed and source separation performance after convergence.

From the above background, we propose a hybrid method, which uses EIVA-T-ISS [10] and WPE-IVA-ISS [8] in the former and the latter half of updating parameters. Since EIVA-T-ISS and WPE-IVA-ISS have advantages in high convergence and high separation performance after convergence respectively, we hope this hybrid method can hold both advantages. Simultaneously, we compare the proposed method with five

[1] Strictly speaking, they use Independent Low-Rank Matrix Analysis (ILRMA) [5] instead of IVA and propose ILRMA-T. However, we use IVA-T in this paper to concentrate on our subjects rather than the difference between IVA and ILRMA.

[2] Strictly speaking, they included geometric constraints (GC) and called GC-WPE-IVA-ISS for their proposal. However, we exclude GC in this paper to concentrate on our subjects rather than the difference between with and without GC.

other methods, WPE-IVA [4], IVA-T [6], WPE-IVA-ISS [8], IVA-T-ISS [9], and EIVA-T-ISS [10], to analyze the convergence speed and source separation performance comprehensively, and to confirm the compatibility of the proposed method's advantages of fast convergence speed and source separation performance.

2. Problem Formulation

Assume that there are N sources and $M(=N)$ microphones. The source signals and observed signals can be expressed in a vector form:

$$\boldsymbol{s}_{f,t} = [s_{1,f,t}, \ldots, s_{N,f,t}]^{\mathsf{T}} \in \mathbb{C}^N, \tag{1}$$

$$\boldsymbol{x}_{f,t} = [x_{1,f,t}, \ldots, x_{M,f,t}]^{\mathsf{T}} \in \mathbb{C}^M, \tag{2}$$

where $(\cdot)^{\mathsf{T}}$ denotes the transpose. $s_{n,f,t}$ and $x_{m,f,t}$ denote the short-time Fourier transformer coefficients of the nth source and the mth microphone. $f = 1, \ldots, F$ and $t = 1, \ldots, T$ are the numbers of frequency bins and frames, respectively. $\boldsymbol{x}_{f,t}$ can be expressed as

$$\boldsymbol{x}_{f,t} = \sum_{\tau=0}^{L} \boldsymbol{A}_{f,\tau} \boldsymbol{s}_{f,t-\tau}, \tag{3}$$

where $\boldsymbol{A}_{f,\tau}$ stands for $M \times N$ mixing matrix at time lag τ and L denotes the order of time-lagged mixing matrix.

Our goal is to obtain source estimates $\hat{\boldsymbol{s}}_{f,t}$ accurately with rapid convergence.

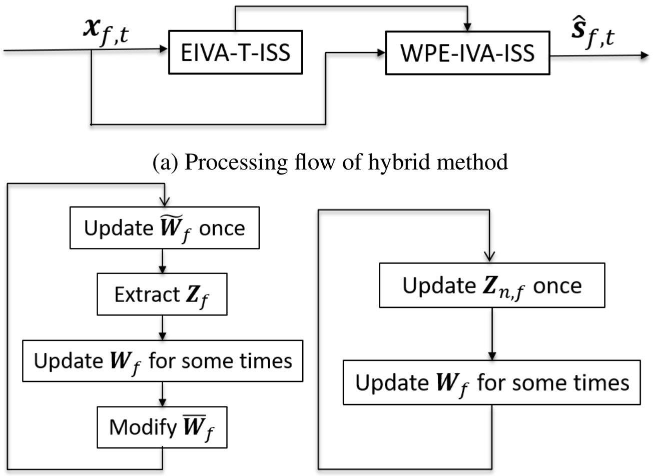
3. Proposed method

In this section, we propose a hybrid method, which uses EIVA-T-ISS [10] and WPE-IVA-ISS [8] in the former and the latter processes in obtaining $\hat{\boldsymbol{s}}_{f,t}$. We show an overview of our proposed method in Fig. 1. As shown in Fig. 1a, we use EIVA-T-ISS [10] to obtain parameters for $\hat{\boldsymbol{s}}_{f,t}$. Then, we use the parameters for initializing those used in WPE-IVA-ISS [8]. Finally, we obtain $\hat{\boldsymbol{s}}_{f,t}$ by further updating the parameters by WPE-IVA-ISS. We hope this hybrid approach can hold both advantages of fast convergence and high separation performance.

Hereafter, we explain the processes of our proposed method in detail.

In EIVA-T-ISS [10], $\hat{\boldsymbol{s}}_{f,t}$ is obtained by the following estimation model:

$$\hat{\boldsymbol{s}}_{f,t} = \tilde{\boldsymbol{W}}_f \tilde{\boldsymbol{x}}_{f,t}, \tag{4}$$


$\boldsymbol{x}_{f,t}$
EIVA-T-ISS
WPE-IVA-ISS
$\hat{\boldsymbol{s}}_{f,t}$
Update $\tilde{\boldsymbol{W}}_f$ once
Extract $\boldsymbol{Z}_f$
Update $\boldsymbol{W}_f$ for some times
Modify $\overline{\boldsymbol{W}}_f$
Update $\boldsymbol{Z}_{n,f}$ once
Update $\boldsymbol{W}_f$ for some times

(a) Processing flow of hybrid method

(b) Updating flow of EIVA-T-ISS [10]

(c) Updating flow of WPE-IVA-ISS [8]

Figure 1: The updating and processing flows of the hybrid method.

where

$$\tilde{\boldsymbol{W}}_f = [\boldsymbol{W}_f, \overline{\boldsymbol{W}}_f] \in \mathbb{C}^{N \times M(L+1)}, \tag{5}$$

$$\tilde{\boldsymbol{x}}_{f,t} = \begin{bmatrix} \boldsymbol{x}_{f,t} \\ \overline{\boldsymbol{x}}_{f,t} \end{bmatrix} \in \mathbb{C}^{M(L+1)}, \tag{6}$$

and $\overline{\boldsymbol{x}}_{f,t} = [\boldsymbol{x}_{f,t-1}^{\mathsf{T}}, \ldots, \boldsymbol{x}_{f,t-L}^{\mathsf{T}}]^{\mathsf{T}} \in \mathbb{C}^{ML}$ contains the past observed signals. In this estimation model, we update $\tilde{\boldsymbol{W}}_f$ once per several updates for $\boldsymbol{W}_f$, as shown in Fig. 1b. We can update $\boldsymbol{W}_f$ and $\tilde{\boldsymbol{W}}_f$ by ISS [7,9,10]. Note that after updating $\boldsymbol{W}_f$, we need to update $\overline{\boldsymbol{W}}_f$ as $\overline{\boldsymbol{W}}_f \leftarrow \boldsymbol{W}_f \boldsymbol{U}_f^{-1} \overline{\boldsymbol{W}}_f$, where $\boldsymbol{U}_f$ is the separation matrix after updating $\tilde{\boldsymbol{W}}_f$. This approach comes from our preliminary experiments that the update of separation matrix $\boldsymbol{W}_f$ should reflect $\overline{\boldsymbol{W}}_f$ when only the matrix $\boldsymbol{W}_f$ is updated [10]. In addition, the inversion of separation matrix $\boldsymbol{W}_f$ can be obtained similarly by ISS. Thus, it is unnecessary to conduct matrix inverse operation directly to obtain $\boldsymbol{U}_f^{-1}$.

After convergence of $\tilde{\boldsymbol{W}}_f$ using EIVA-T-ISS [10], we use $\boldsymbol{W}_f$ for initializing that used in WPE-IVA-ISS [8]. In WPE-IVA-ISS, $\hat{\boldsymbol{s}}_{f,t}$ is obtained by the following estimation model:

$$\hat{s}_{n,f,t} = \boldsymbol{w}_{n,f}^{\mathsf{H}} (\boldsymbol{x}_{f,t} - \boldsymbol{Z}_{n,f}^{\mathsf{H}} \overline{\boldsymbol{x}}_{f,t}), \tag{7}$$

where $\{\boldsymbol{w}_{n,f}\}_{n,f}$ is initialized by $\boldsymbol{W}_f = [\boldsymbol{w}_{1,f}, \ldots, \boldsymbol{w}_{M,f}]$ in (5) estimated by EIVA-T-ISS. $(\cdot)^{\mathsf{H}}$ denotes the conjugate transpose. $\boldsymbol{Z}_{n,f}$ is a dereverberation matrix for nth separated signal $\hat{s}_{n,f,t}$. After the initialization, we repeat updating $\{\boldsymbol{Z}_{n,f}\}_{n,f}$ once per several updates for $\{\boldsymbol{w}_{n,f}\}_{n,f}$, as shown in Fig. 1b. Here, $\{\boldsymbol{Z}_{n,f}\}_{n,f}$ is updated by WPE [3]:

$$\boldsymbol{Z}_{n,f} \leftarrow \left(\sum_t \frac{\overline{\boldsymbol{x}}_{f,t} \overline{\boldsymbol{x}}_{f,t}^{\mathsf{H}}}{r_{n,t}} \right)^{-1} \left(\sum_t \frac{\overline{\boldsymbol{x}}_{f,t} \boldsymbol{x}_{f,t}^{\mathsf{H}}}{r_{n,t}} \right), \tag{8}$$

where $r_{n,t}$ is variance of Time-Varying Gaussian source calculated as $\frac{1}{F} \sum_f |\hat{s}_{n,f,t}|^2$. The update rule of $\{\boldsymbol{w}_{n,f}\}_{n,f}$ is the same as that used for EIVA-T-ISS.

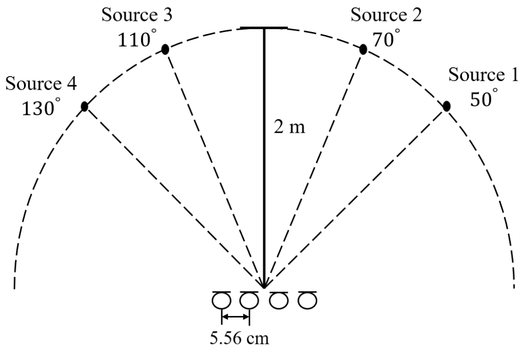


Figure 2: Layout of sound sources and microphones

4. Experiment

4.1 Experiment conditions

The observed signals were generated by convolving speech signals from the TIMIT database [11] with room impulse responses (RIRs) from the RWCP dataset [12]. For each mixed signal, 2 to 4 speech segments, randomly selected from different speakers, were concatenated to form 10 s long clean speech signals. The room reverberation time (T_{60}) was set to 300 ms and 600 ms. White Gaussian noise was introduced to adjust the signal-to-noise ratio (SNR) to 30 dB. The layout of sources and microphones is illustrated in Fig. 2.

The directions of arrival were set at 50° and 70° for the 2-source case, 50°, 70°, and 110° for the 3-source case, and 50°, 70°, 110°, and 130° for the 4-source case. The distance between the uniform linear microphone array center and the sources was 2 m.

To evaluate the separation performance, 25 Monte Carlo simulations were conducted. All observed signals were sampled at 16 kHz, and the short-time Fourier transform was performed using a Hann window of 64 ms (1024 samples) with a 16 ms (256 samples) window shift. For the dereverberation filter, the time delay (D) and filter length (L) were set to 2 and 10, respectively. Results for all methods were obtained over 60 iterations.

All algorithms were implemented on a workstation powered by AMD EPYC 9654. The improvement in signal-to-distortion ratio (ΔSDR) and signal-to-interference ratio (ΔSIR) [13] were used as metrics to assess separation performance.

The separation performance of the proposed algorithm was compared with the five other methods, WPE-IVA [4], IVA-T [6], WPE-IVA-ISS [8], IVA-T-ISS [9], and EIVA-T-ISS [10].

4.2 Results

Figure 3 shows the ΔSDR and ΔSIR of different methods with 4 channels. Figure 4 presents the average ΔSDR under different reverberant conditions. In the reverberation condition with $T_{60} = 300$ ms, ΔSDR of the proposed method has no obvious advantages over the conventional methods. However, as the reverberation becomes stronger (with T_{60} = 600 ms), the separation performance of EIVA-T-ISS decreases significantly. The proposed hybrid method shows the equivalent or higher ΔSDR than the conventional methods.

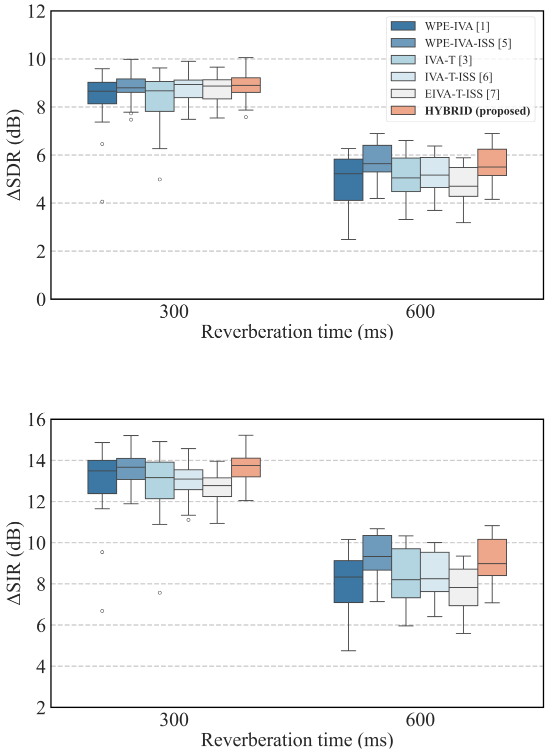


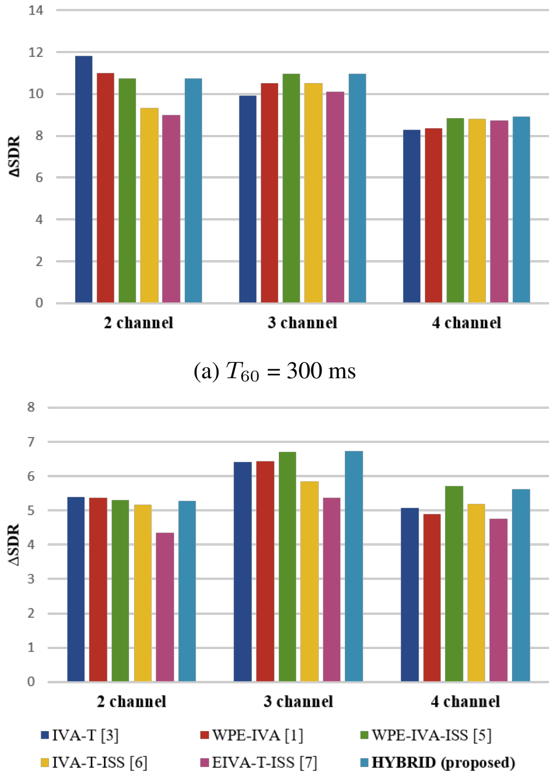
Figure 3: The separation performance of different methods.

Table 1: Runtime [ms] per iteration (T_{60} = 600ms)

Method	2 ch	3 ch	4 ch
WPE-IVA [4]	117.73	208.44	300.11
WPE-IVA-ISS [8]	104.48	182.43	282.12
IVA-T [6]	297.73	659.04	940.82
IVA-T-ISS [9]	273.58	602.22	1245.10
EIVA-T-ISS [10]	37.03	82.58	156.92
HYBRID (proposed)	**72.13**	**131.63**	**217.41**

Figure 5 plots the convergence curves of EIVA-T-ISS, WPE-IVA-ISS, and the hybrid method with 4 channels. The results show the proposed method gets convergence with a much lower time cost than WPE-IVA-ISS.

Table 1 shows the runtime of each method averaged over iterations for separating a signal with $T_{60} = 600$ ms. We found that WPE-IVA-ISS took longer runtime than EIVA-T-ISS or the hybrid method, which demonstrates the superior computational efficiency of our proposed method.



(a) T_{60} = 300 ms

(b) T_{60} = 600 ms

Figure 4: Average ΔSDR [dB] under different reverberant conditions.

5. Conclusion

In this paper, we proposed a hybrid method that combined EIVA-T-ISS in the first stage and WPE-IVA-ISS in the second stage. To evaluate its effectiveness, we compared the proposed method against five existing methods, focusing on convergence speed and source separation performance. Experimental results demonstrated the superior computational efficiency. Furthermore, it achieved faster convergence than WPE-IVA-ISS while maintaining the comparable source separation performance.

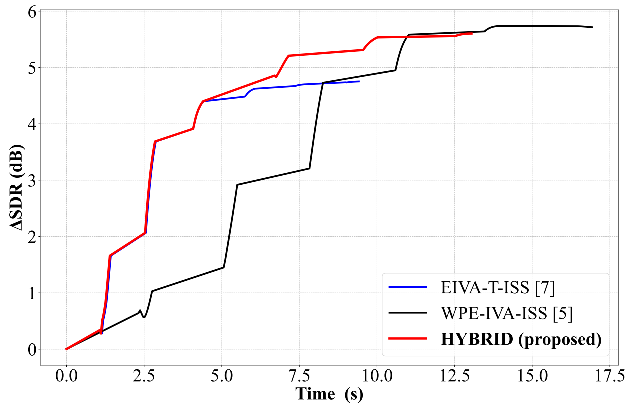


Figure 5: Convergence speed of EIVA-T-ISS, WPE-IVA-ISS, and the hybrid method.

References

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

[2] T. Kim, I. Lee, and T.-W. Lee, “Independent vector analysis: Definition and algorithms,” in *Proc. ACSSC*, 2006, pp. 1393–1396.

[3] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, “Speech dereverberation based on variance-normalized delayed linear prediction,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1717–1731, 2010.

[4] 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.

[5] 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. on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1626–1641, 2016.

[6] R. Ikeshita, N. Ito, T. Nakatani, and H. Sawada, “Independent low-rank matrix analysis with decorrelation learning,” in *Proc. ICASSP*, 2008, pp. 85–88.

[7] R. Scheibler and N. Ono, “Fast independent vector extraction by iterative sinr maximization,” in *Proc. ICASSP*, 2020, pp. 601–605.

[8] K. Mo, X. Wang, Y. Yang, T. Ueda, S. Makino, and J. Chen, “On joint dereverberation and source separation with geometrical constraints and iterative source steering,” in *Proc. APSIPA ASC*, 2023, pp. 1138–1142.

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

[10] B. Wang, K. Mo, T. Ueda, and S. Makino, “Accelerating algorithm of geometrically constrained source separation and dereverberation using iterative source steering,” in *Proc. HSCMA*, 2024, pp. 1–5 (non–archival track).

[11] 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, p. 27403, 1993.

[12] 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 *LREC*, 2000, pp. 965–968.

[13] 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, Jun. 2006.