Independent Low-Rank Matrix Analysis for Fast Joint Blind Source Extraction and Dereverberation in Diffuse Noise Environments

Shuo Feng
Waseda University
Kitakyushu, Japan
hinanawipeach@asagi.waseda.jp

Liyuan Zhang
Waseda University
Kitakyushu, Japan
ly.zhang@akane.waseda.jp

Shoji Makino
Waseda University
Kitakyushu, Japan
s.makino@waseda.jp

***Abstract*—This paper addresses a blind source extraction (BSE) problem in reverberant and diffuse noise environments. Although traditional Diffuse-Noise-Aware Independent Low-Rank Matrix Analysis (NoisyILRMA) achieves efficient extraction in diffuse noise scenarios, the performance degrades in reverberant environments. To address this limitation, we propose NoisyILRMA-conv, which extends NoisyILRMA by integrating a Weighted Prediction Error dereverberation module under a unified optimization scheme. Our proposed method enables joint source extraction and dereverberation while maintaining the computational efficiency of NoisyILRMA.**

***Index Terms*—Blind source extraction, diffuse noise environments, dereverberation, joint optimization**

I. Introduction

Blind source extraction (BSE) is a technique that extracts target speech sources and removes redundant components from only the observed microphone mixtures. Conventional methods such as independent vector extraction (IVE) [1] perform well in common BSE problems but struggle in reverberant and diffuse noise scenarios.

Recently, Ref. [2] merged IVE with a weighted prediction error (WPE) [3] module to handle reverberation, referred to as IVE for convolutive mixtures (IVE-conv). IVE-conv enables efficient joint extraction and dereverberation optimization. However, it has difficulty finding the target source in diffuse noise environments, where noises come from all directions.

To tackle the problem of diffuse noise in BSE, diffuse-noise-aware independent low-rank matrix analysis for fast BSE (NoisyILRMA) [4] was proposed. It can effectively and efficiently extract a single target source under diffuse noise conditions. However, NoisyILRMA's performance deteriorates easily in reverberant environments because its mixing system only holds when the short-time Fourier transform (STFT) window size is sufficiently longer than the reverberation [4].

Inspired by these two methods, we propose *NoisyILRMA for convolutive mixtures (NoisyILRMA-conv)*. The relationship between our proposal and conventional methods is shown in Fig. 1. Experiments indicate that our proposal achieves enhanced extraction and dereverberation performance in diffuse noise environments.

II. Signal model and problem formulation

We consider a reverberant acoustic environment in which one target speech source and $N-1$ background noise sources are captured by $M=N$ microphones. The observed signal in the STFT domain is modeled as a convolutive mixture:

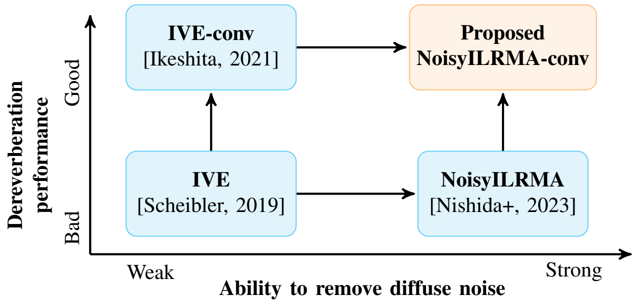

Fig. 1. Relationship between proposed and conventional methods.

$$\boldsymbol{x}_{f,t}=\sum_{\tau=0}^{T-1}\left[\boldsymbol{a}_{f,\tau}^{(s)}s_{f,t-\tau}+\mathbf{A}_{f,\tau}^{(z)}\boldsymbol{z}_{f,t-\tau}\right], \tag{1}$$

where $\boldsymbol{a}_{f,\tau}^{(s)}\in\mathbb{C}^{M\times 1}$ and $\mathbf{A}_{f,\tau}^{(z)}\in\mathbb{C}^{M\times(M-1)}$ are the acoustic transfer functions of the target speech and background noises, respectively, superscripts $^{(s)}$ and $^{(z)}$ denote components related to the target source and noises, respectively. The indices $f=1,\ldots,F$ and $t=1,\ldots,T$ denote frequency bins and time frames, respectively, $s_{f,t}\in\mathbb{C}$ denotes the target source, and $\boldsymbol{z}_{f,t}\in\mathbb{C}^{(M-1)\times 1}$ denotes the noise sources. Hereafter, $(\cdot)^{\mathsf{T}}$ and $(\cdot)^{\mathsf{H}}$ denote transpose and Hermitan transpose, respectively.

A. IVE-conv

For the demixing model, IVE-conv adopts a joint BSE and dereverberation (BSE-DR) framework. In the dereverberation step using WPE, the augmented observation vector is defined as $\hat{\boldsymbol{x}}_{f,t}=\left[\boldsymbol{x}_{f,t}^{\mathsf{T}}\quad\boldsymbol{x}_{f,t-D_1}^{\mathsf{T}}\quad\cdots\quad\boldsymbol{x}_{f,t-D_2}^{\mathsf{T}}\right]^{\mathsf{T}}\in\mathbb{C}^{M+L}$, $0<D_1\leq D_2$, D_1 is the prediction delay, and $L=M(D_2-D_1+1)$. The estimated signal is denoted as $\boldsymbol{y}_{f,t}=\left[s_{f,t}\quad\boldsymbol{z}_{f,t}^{\mathsf{H}}\right]^{\mathsf{H}}$ and the BSE-DR process is $\boldsymbol{y}_{f,t}=\hat{\mathbf{W}}_f^{\mathsf{H}}\hat{\boldsymbol{x}}_{f,t}$, where $\hat{\mathbf{W}}_f=[\hat{\boldsymbol{w}}_f^{(s)}\ \hat{\mathbf{W}}_f^{(z)}]=\begin{bmatrix}\mathbf{W}_f\\-\mathbf{Z}_f\mathbf{W}_f\end{bmatrix}\in\mathbb{C}^{(M+L)\times M}$, $\mathbf{Z}_f$ is the WPE prediction matrix, and $\mathbf{W}_f$ is the extraction matrix.

B. NoisyILRMA

In NoisyILRMA, the target source is modeled as $y_{f,t,1}\sim\mathcal{N}_{\mathbb{C}}\left(0,r_{f,t}^{(s)}+r_{f,t}^{(z)}\lambda_f^{(z)}\right)$, and the noise sources are modeled as $y_{f,t,n} \sim \mathcal{N}_{\mathbb{C}}\left(0, r_{f,t}^{(z)}\right)$ $(n \in 2, \dots, M)$, where $r_{f,t}$ is the time-variant variance, and $\lambda_f^{(z)} > 0$ denotes the weight of the noise components in $y_{f,t,1}$. The BSE process is $\boldsymbol{y}_{f,t} = \mathbf{W}_f^{\mathsf{H}} \boldsymbol{x}_{f,t}$. The cost function of NoisyILRMA is as follows:

$$\begin{aligned} L(\Theta) = \sum_{f,t} \Bigg[& -2 \log |\det \mathbf{W}_f| \\ & + \log(r_{f,t}^{(s)} + r_{f,t}^{(z)} \lambda_f^{(z)}) + (M-1) \log r_{f,t}^{(z)} \\ & + \frac{|y_{f,t,1}|^2}{r_{f,t}^{(s)} + r_{f,t}^{(z)} \lambda_f^{(z)}} + \sum_{n=2}^{M} \frac{|y_{f,t,n}|^2}{r_{f,t}^{(z)}} \Bigg] + \text{const.} \end{aligned} \quad (2)$$

Here, $\Theta = \{\mathbf{W}_f, r_{f,t}^{(s)}, r_{f,t}^{(z)}, \lambda_f^{(z)}\}$.

III. Proposed method: *NoisyILRMA-conv*

We propose to extend NoisyILRMA to reverberant environments by utilizing the BSE-DR framework in IVE-conv. $\mathbf{W}_f$ is decomposed as $[\boldsymbol{w}_f^{(s)} \; \mathbf{W}_f^{(z)}]$, where $\boldsymbol{w}_f^{(s)} \in \mathbb{C}^{M \times 1}$, $\mathbf{W}_f^{(z)} \in \mathbb{C}^{M \times (M-1)}$. The cost function with respect to $\hat{\mathbf{W}}_f$ is as follows:

$$\begin{aligned} L(\{\hat{\mathbf{W}}_f\}) = \sum_f \Big[& -2 \log |\det \mathbf{W}_f| + (\boldsymbol{w}_f^{(s)})^{\mathsf{H}} \mathbf{G}_f^{(s)} \boldsymbol{w}_f^{(s)} \\ & + \mathrm{Tr}\left((\mathbf{W}_f^{(z)})^{\mathsf{H}} \mathbf{G}_f^{(z)} \mathbf{W}_f^{(z)} \right) \Big] + \text{const.}, \end{aligned} \quad (3)$$

where $\mathbf{G}_f^{(l)} = \mathbf{R}_f^{(l)} - (\mathbf{P}_f^{(l)})^{\mathsf{H}} \bar{\mathbf{R}}_f^{(l)} \mathbf{P}_f^{(l)} \in \mathcal{S}_{++}^{M}$ $(l \in \{s, z\})$. Here, $\mathbf{R}_f^{(l)}$, $\bar{\mathbf{R}}_f^{(l)}$ and $\mathbf{P}_f^{(l)}$ are the submatrices of the spatial covariance matrices:

$$\mathbf{V}_f^{(s)} = \frac{1}{T} \sum_t \frac{\hat{\boldsymbol{x}}_{f,t} \hat{\boldsymbol{x}}_{f,t}^{\mathsf{H}}}{r_{f,t}^{(s)} + r_{f,t}^{(z)} \lambda_f^{(z)}}, \mathbf{V}_f^{(z)} = \frac{1}{T} \sum_t \frac{\hat{\boldsymbol{x}}_{f,t} \hat{\boldsymbol{x}}_{f,t}^{\mathsf{H}}}{r_{f,t}^{(z)}}, \quad (4)$$

which are decomposed as:

$$\mathbf{V}_f^{(l)} = \begin{bmatrix} \mathbf{R}_f^{(l)} & (\mathbf{P}_f^{(l)})^{\mathsf{H}} \\ \mathbf{P}_f^{(l)} & \bar{\mathbf{R}}_f^{(l)} \end{bmatrix} \in \mathcal{S}_{++}^{M+L}. \quad (5)$$

Here, $\mathbf{R}_f^{(l)} \in \mathcal{S}_{++}^{M}$, $\mathbf{P}_f^{(l)} \in \mathbb{C}^{L \times M}$ and $\bar{\mathbf{R}}_f^{(l)} \in \mathcal{S}_{++}^{L}$. $\mathbf{G}_f^{(s)}$, $\mathbf{G}_f^{(z)}$ satisfy $(\hat{\boldsymbol{w}}_f^{(s)})^{\mathsf{H}} \mathbf{V}_f^{(s)} \hat{\boldsymbol{w}}_f^{(s)} = (\boldsymbol{w}_f^{(s)})^{\mathsf{H}} \mathbf{G}_f^{(s)} \boldsymbol{w}_f^{(s)}$, $(\hat{\mathbf{W}}_f^{(z)})^{\mathsf{H}} \mathbf{V}_f^{(z)} \hat{\mathbf{W}}_f^{(z)} = (\mathbf{W}_f^{(z)})^{\mathsf{H}} \mathbf{G}_f^{(z)} \mathbf{W}_f^{(z)}$ [2]. Then we use the same update rules in NoisyILRMA and finally calculate the estimated signal $\boldsymbol{y}_{f,t} = \boldsymbol{b}_f^{(s)} \frac{r_{f,t}^{(s)}}{r_{f,t}^{(s)} + r_{f,t}^{(z)} \lambda_f^{(z)}} (\hat{\boldsymbol{w}}_f^{(s)})^{\mathsf{H}} \hat{\boldsymbol{x}}_{f,t}$, where $\boldsymbol{b}_f^{(s)} = (\mathbf{W}_f^{\mathsf{H}})^{-1} e_1$, $e_1 = [1, 0, 0, \cdots, 0]^{\mathsf{T}}$.

IV. Experimental evaluations

In this section, we evaluated the performance of the proposed method using synthetic mixtures with reverberation and diffuse noise. Both the speech sources and the background noises were taken from the ATR Japanese Speech Database [5]. Five male and five female utterances were used as target sources, while cafe and station noises were used as diffuse background noises. The spatial arrangement of the sources and microphones followed the same configuration as in NoisyILRMA [4]. Reverberation times (RT_{60}) were set to 200 ms and 600 ms, and the direction of arrival (DOA) of the target source was set to 30°.

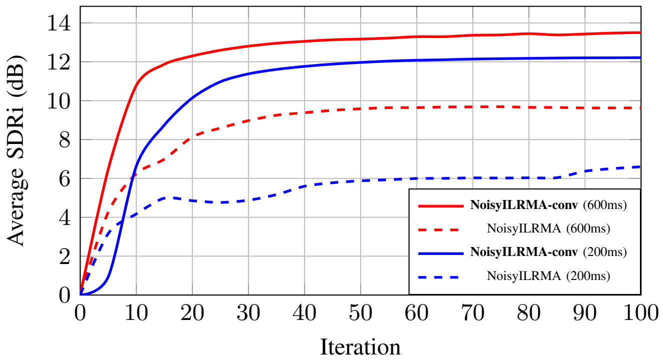

Fig. 2. Average SDRi over iterations for NoisyILRMA and NoisyILRMA-conv under different reverberation times (RT60 = 200 ms and 600 ms).

TABLE I
SDR (dB) under different RT_{60} conditions.

RT_{60} (ms)	Input SDR	Output SDR (conventional / proposed)
200	7.29	13.17 / **19.26**
600	0.58	10.16 / **13.74**

The extraction performance was evaluated in terms of the source-to-distortion ratio improvement (SDRi), defined as the difference between the SDR of the separated signal and that of the microphone observation. The proposed NoisyILRMA-conv was compared with the original NoisyILRMA under identical experimental settings. Figure 2 shows the average SDRi over 10 trials as a function of iteration. The proposed method achieved higher SDRi values in both reverberation conditions (RT_{60} = 200 ms and 600 ms) compared to the original NoisyILRMA [4]. Table 1 summarizes the numerical SDR values under these two RT_{60} conditions. These results demonstrate that the proposed integration of dereverberation improves extraction performance in reverberant and diffuse noise environments.

V. Conclusion

This paper proposed NoisyILRMA-conv, which extends NoisyILRMA by integrating a dereverberation module based on WPE, and the proposed method enables effective BSE in reverberant and diffuse noise environments. Experimental evaluations validated the effectiveness of our proposal.

References

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

[2] 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.

[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] K. Nishida, N. Takamune, R. Ikeshita, D. Kitamura, H. Saruwatari, and T. Nakatani, “Noisyilrma: diffuse-noise-aware independent low-rank matrix analysis for fast blind source extraction,” in *Proc. EUSIPCO*, 2023, pp. 925–929.

[5] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, “Atr japanese speech database as a tool of speech recognition and synthesis,” *Speech Commun.*, vol. 9, no. 4, pp. 357–363, 1990.