Performance Evaluation of Independent Low-rank Matrix Analysis for Short Signals

Taishi Nakashima, Robin Scheibler, Yukoh Wakabayashi, Nobutaka Ono

PERFORMANCE EVALUATION OF INDEPENDENT LOW-RANK MATRIX ANALYSIS FOR SHORT SIGNALS

Taishi Nakashima **Robin Scheibler** **Yukoh Wakabayashi** **Nobutaka Ono**
Department of Computer Science, Graduate School of Systems Design,
Tokyo Metropolitan University, Japan
nakashima-taishi@ed.tmu.ac.jp, robin@tmu.ac.jp, wakayuko@tmu.ac.jp, onono@tmu.ac.jp

ABSTRACT

In this paper, we evaluate the performance of independent low-rank matrix analysis (ILRMA) for short signals. ILRMA is a state-of-the-art blind source separation (BSS) technique based on the assumption that sources are statistically independent and their spectrograms can be approximately expressed as low-rank matrices. Because ILRMA estimates many parameters such as demixing matrices, spectral bases, and source activations, it requires observations of sufficient length for stable estimation. Thus, the performance of ILRMA could degrade when the available signals are short. Toward overcoming this problem, we apply ILRMA to a short mixture and investigate the dependence of the performance on the signal length.

1. INTRODUCTION

Blind source separation (BSS) is a technique of estimating the source signals from a mixture of sources using only the observed signals without any other information. BSS is useful for many applications such as acoustic event detection and hearing-aid devices. As a state-of-the-art BSS method, independent low-rank matrix analysis (ILRMA) has recently been proposed [1]. In ILRMA, a low-rank matrix model of the spectrogram obtained from each source is assumed. ILRMA can be interpreted as a method that unifies auxiliary-function-based independent vector analysis (AuxIVA) [2] and multichannel non-negative matrix factorization (MNMF) [3–10]. ILRMA can achieve better performance than AuxIVA and is more stable than MNMF. In ILRMA, many parameters such as demixing matrices, spectral bases, and source activations must be estimated. Thus, it is considered that ILRMA requires sufficiently long observations for stable estimation.

However, in some cases, BSS must be applied to signals of insufficient length. For example, in online BSS, a well-used approach is to divide the observed signal into several short-length blocks to satisfy the assumption of a time-invariant system, and then update the demixing matrices in every block [11–16]. In this case, the length of each block can be 1 or 2 s. In this paper, as a preliminary investigation to realize online real-time ILRMA, we evaluate the performance of ILRMA for short signals and examine its dependence on the parameters setting, such as the STFT length and the number of bases, to achieve better separation.

The rest of this paper is organized as follows. In Section 2, we formulate the multichannel BSS problem and describe conventional ILRMA. In Section 3, we show the experimental results and discuss them. Finally, in Section 4, we present our conclusions.

2. BACKGROUND

2.1 Formulation

Let K and M be the numbers of sources and microphones, respectively. In this paper, henceforth, we consider the determined case, $K = M$. We respectively define the STFT representations of the source, observed, and estimated signals as

$$\boldsymbol{s}_{f\tau} = \left[s_{1,f\tau} \ \cdots \ s_{k,f\tau} \ \cdots \ s_{K,f\tau}\right]^\top \in \mathbb{C}^{K\times 1}, \quad (1)$$

$$\boldsymbol{x}_{f\tau} = \left[x_{1,f\tau} \ \cdots \ x_{k,f\tau} \ \cdots \ x_{K,f\tau}\right]^\top \in \mathbb{C}^{K\times 1}, \quad (2)$$

$$\boldsymbol{y}_{f\tau} = \left[y_{1,f\tau} \ \cdots \ y_{k,f\tau} \ \cdots \ y_{K,f\tau}\right]^\top \in \mathbb{C}^{K\times 1}, \quad (3)$$

where $f \in \{1, \ldots, F\}$, $\tau \in \{1, \ldots, T\}$, and $k \in \{1, \ldots, K\}$ are the indices of frequency bins, time frames, and channels, respectively, and $^\top$ denotes the vector/matrix transpose. When the STFT window is sufficiently longer than the impulse response, we can represent the observed signal $\boldsymbol{x}_{f\tau}$ as

$$\boldsymbol{x}_{f\tau} = \boldsymbol{A}_f \boldsymbol{s}_{f\tau}, \quad (4)$$

where $\boldsymbol{A}_f \in \mathbb{C}^{K\times K}$ is a mixing matrix. If $\boldsymbol{A}_f$ is invertible, we can define the demixing matrix as

$$\boldsymbol{W}_f = \left[\boldsymbol{w}_{1,f} \quad \cdots \quad \boldsymbol{w}_{K,f}\right]^{\mathsf{H}} = \boldsymbol{A}_f^{-1}, \quad (5)$$

where $\boldsymbol{w}_{k,f} \in \mathbb{C}^{K\times 1}$ $(k = 1, \ldots, K)$ are the demixing vectors and $^{\mathsf{H}}$ denotes the Hermitian transpose. Therefore, the estimated signal $\boldsymbol{y}_{f\tau}$ can be represented as

$$\boldsymbol{y}_{f\tau} = \boldsymbol{W}_f \boldsymbol{x}_{f\tau}. \quad (6)$$

2.2 ILRMA

As introduced earlier, ILRMA is a determined BSS technique unifying IVA and NMF. In ILRMA, the demixing matrices are updated under the assumption that the complex spectrogram of the kth source is approximately represented as the product of two non-negative matrices, $\boldsymbol{B}_k \in \mathbb{R}_+^{F\times L}$ and $\boldsymbol{H}_k \in \mathbb{R}_+^{L\times T}$, where $\boldsymbol{B}_k$ and $\boldsymbol{H}_k$ are

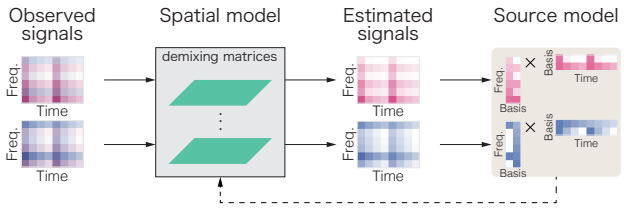


Figure 1: Overview of source separation in ILRMA (e.g., $K = 2$).

the basis and activation matrices, respectively, and $\mathbb{R}_+$ denotes the set of non-negative real numbers. To estimate the source model parameters $\boldsymbol{B}_k$ and $\boldsymbol{H}_k$, we consider the following complex Gaussian distribution as the generative model of the kth source:

$$p(\bar{\boldsymbol{y}}_{1,\tau}, \ldots, \bar{\boldsymbol{y}}_{k,\tau}) = \prod_{k,f} \frac{1}{\pi r_{k,f\tau}} \exp\left(-\frac{|y_{k,f\tau}|^2}{r_{k,f\tau}}\right), \quad (7)$$

$$r_{k,f\tau} = \sum_{l} b_{k,f\ell}\, h_{k,\ell\tau}, \quad (8)$$

where $r_{k,f\tau}$ is its variance, $\bar{\boldsymbol{y}}_{k,\tau}$ is the estimated vector that consists of all frequency components defined as $\bar{\boldsymbol{y}}_{k,\tau} = \begin{bmatrix} y_{k,1\tau} & \cdots & y_{k,F\tau} \end{bmatrix}^\top$; $b_{k,f\ell} \in \mathbb{R}_+$ and $h_{k,\ell\tau} \in \mathbb{R}_+$ are the (f,ℓ)th element of $\boldsymbol{B}_k$ and (ℓ,τ)th element of $\boldsymbol{H}_k$, respectively; and $\ell \in \{1, \ldots, L\}$ denotes the index of the bases. Figure 1 gives an overview of ILRMA.

Next, by using the demixing model Eq. (6) and the source model Eq. (7), and calculating the negative log-likelihood of the observed signals while omitting constants, we obtain the objective function of ILRMA as

$$\mathcal{J}(\mathsf{W}, \mathsf{B}, \mathsf{H}) = \sum_{k,f,\tau} \left[\frac{|\boldsymbol{w}_{k,f}^{\mathsf{H}} \boldsymbol{x}_{f\tau}|^2}{r_{k,f\tau}} + \log r_{k,f\tau}\right] - 2T \sum_{f} \log|\det \boldsymbol{W}_f|, \quad (9)$$

where W, B, and H are the tensors composed of all $\boldsymbol{W}_f$, $\boldsymbol{B}_k$, and $\boldsymbol{H}_k$, respectively.

2.2.1 *Update of the spatial model*

To minimize the objective function Eq. (9) with respect to the demixing matrix $\boldsymbol{W}_f$, we obtain the following function $\mathcal{Q}$ by extracting terms that are dependent on W from Eq. (9):

$$\mathcal{Q}(\mathsf{W}, \mathsf{U}) = \sum_{k,f} \boldsymbol{w}_{k,f}^{\mathsf{H}} \boldsymbol{U}_{k,f} \boldsymbol{w}_{k,f} - \sum_{f} \log|\det \boldsymbol{W}_f|, \quad (10)$$

$$\boldsymbol{U}_{k,f} = \frac{1}{T} \sum_{\tau} \frac{1}{2r_{k,f\tau}} \boldsymbol{x}_{f\tau} \boldsymbol{x}_{f\tau}^{\mathsf{H}}, \quad (11)$$

where $\boldsymbol{U}_{k,f} \in \mathbb{C}^{K\times K}$ and U are the covariance matrix and the tensor composed of all $\boldsymbol{U}_{k,f}$, respectively. Henceforth, we omit the frequency bin index f for simplicity.

By calculating $\partial\mathcal{Q}/\partial\boldsymbol{w}_k = 0$ $(k = 1, \ldots, K)$ and rearranging it, we can obtain the following system of quadratic equations:

$$\boldsymbol{w}_\ell^{\mathsf{H}} \boldsymbol{U}_k \boldsymbol{w}_k = \delta_{\ell k} \quad (k, \ell = 1, \ldots, K), \quad (12)$$

where $\delta_{\ell k}$ is the Kronecker delta. Equation (12) is referred to as hybrid exact-approximate joint diagonalization (HEAD) [17], and no closed-form solution for $K \geq 3$ has yet been found [18]. Instead of solving HEAD directly, we minimize Eq. (10) with respect to only one demixing vector $\boldsymbol{w}_m$ while keeping the other $\boldsymbol{w}_k$ $(k \neq m)$ fixed. In this case, the problem can be solved in a closed-form as [2]

$$\boldsymbol{w}_m \leftarrow (\boldsymbol{W}\boldsymbol{U}_m)^{-1} \boldsymbol{e}_m, \quad (13)$$

$$\boldsymbol{w}_m \leftarrow \boldsymbol{w}_m (\boldsymbol{w}_m^{\mathsf{H}} \boldsymbol{U}_m \boldsymbol{w}_m)^{-\frac{1}{2}}, \quad (14)$$

where $\boldsymbol{e}_m \in \mathbb{R}^K$ denotes the canonical basis vector with the mth element unity. This method is called iterative projection [1].

2.2.2 *Update of the source model*

By applying the auxiliary function method [19] to Eq. (9), we can obtain the following multiplicative update rules for the source model parameters $\boldsymbol{B}_k$ and $\boldsymbol{H}_k$ [1, 7]:

$$b_{k,f\ell} \leftarrow b_{k,f\ell} \left[\frac{\sum_\tau |y_{k,f\tau}|^2 h_{k,\ell\tau} (\sum_{\ell'} b_{k,f\ell'} h_{k,\ell'\tau})^{-2}}{\sum_\tau h_{k,\ell\tau} (\sum_{\ell'} b_{k,f\ell'} h_{k,\ell'\tau})^{-1}}\right]^{\frac{1}{2}}, \quad (15)$$

$$h_{k,\ell\tau} \leftarrow h_{k,\ell\tau} \left[\frac{\sum_f |y_{k,f\tau}|^2 b_{k,f\ell} (\sum_{\ell'} b_{k,f\ell'} h_{k,\ell'\tau})^{-2}}{\sum_f b_{k,f\ell} (\sum_{\ell'} b_{k,f\ell'} h_{k,\ell'\tau})^{-1}}\right]^{\frac{1}{2}}. \quad (16)$$

3. EXPERIMENTS

We conducted BSS experiments and evaluated the separation performance to investigate its dependence on the parameters of ILRMA, such as the signal length, the STFT length, and the number of bases.

3.1 Experimental setup

We conducted experiments on speech signals. We used a mixture speech of three speakers obtained from the Japanese Newspaper Article Sentences (JNAS) dataset [20]. The sampling rate was 16 kHz, the signal length was 30 s, and the observed mixture speech was divided into blocks of various lengths. For example, when the block length was 1 s, we had 30 blocks of the signal, each with a length of 1 s. Then, we applied ILRMA to every block.

We used the `pyroomacoustics` Python package [21] to simulate a rectangular room and create convolutive mixtures. Figure 2 shows the room used for the simulation and the locations of sources and the microphone array. The reverberation time was approximately 200 ms. There were three sound sources and three microphones. The microphone array was uniformly linear with a spacing of 2.83 cm.

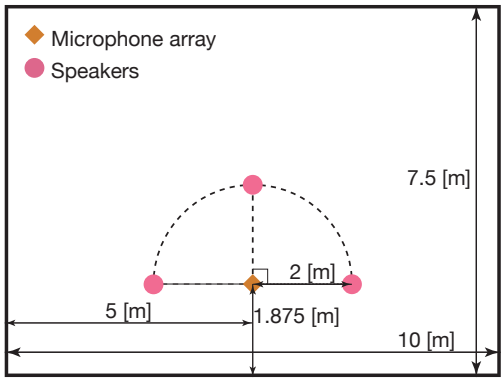


Figure 2: Setup of experiment.

Table 1: Parameters of the simulation.

Number of bases L	$1, 2, \ldots, 10$
STFT length	$64, 128, \ldots, 1024$
Frame shift of STFT	Half overlap
Initial demixing matrices $\boldsymbol{W}_f$	Identity matrix
Number of iterations	100
STFT window function	Hamming

Moreover, for numerical stability, we set the initial values of the source models $\boldsymbol{B}_k$ and $\boldsymbol{H}_k$ to $0.9\boldsymbol{Z}+0.1\boldsymbol{I}$, where $\boldsymbol{Z}$ and $\boldsymbol{I}$ are the matrix of values uniformly distributed over $[0, 1)$ and the matrix of ones, respectively. Table 1 shows the other parameters.

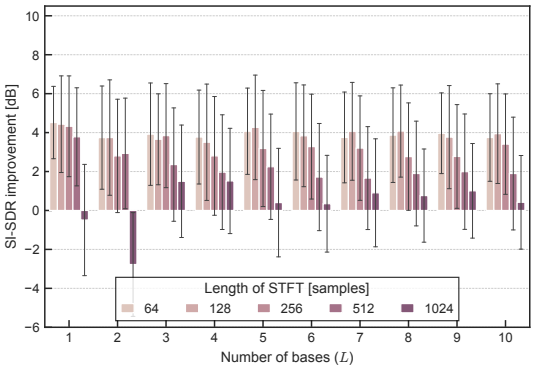
3.2 Separation performance

We applied ILRMA to each block and calculated the average scale-invariant signal-to-distortion ratio (SI-SDR) improvements [22] for all blocks and sources. We set the block length to 1 s, 2 s, and 5 s and conducted experiments. Figure 3 shows the results.

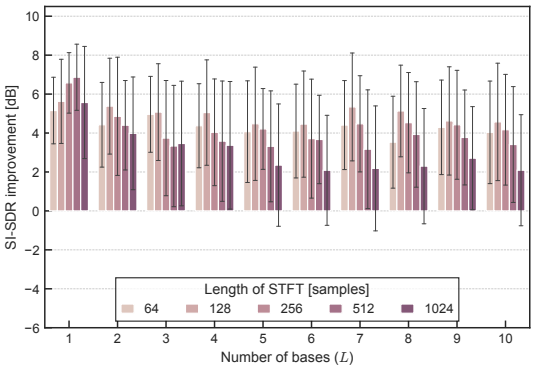
On the whole, the separation performance improved with increasing block length. Even when the block length was 1 s, the separation performance did not significantly degrade in terms of SI-SDR. Regarding the STFT length, the SI-SDR decreased with increasing STFT length when the block length was 1 s or 2 s, whereas SI-SDR increased when it was 5 s. This was because when the block length was short and the STFT was long, the number of frames of the observed signal decreased to a number insufficient for the estimation. In addition, the separation performance did not significantly improve with increasing number of bases. These results are consistent with those discussed in [1].

4. CONCLUSION

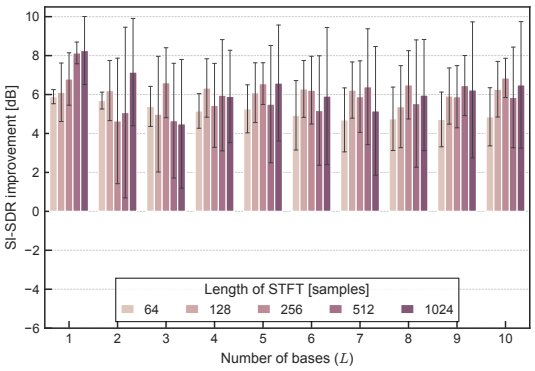
We evaluated the performance of ILRMA for short time signals to investigate its dependence on the parameters in ILRMA. More specifically, we performed a simulation experiment in a reverberant environment for a mixture of speech signals of three speakers. On the whole, the separation performance tended to increase with the STFT length when the block length was 2 s or 5 s but decrease when it was 1 s. Although the separation performance did not strongly depend on the number of bases in the NMF source models, the best average performance was obtained for a single basis, for which the optimum STFT lengths for block lengths of 1 s, 2 s, and 5 s were 64, 512, and 1024, respectively. Even when the block length was 1 s, SI-SDR was improved by approximately 4 dB. In our future work, we will further investigate the separation performance in a realistic environment. We will also apply the online update algorithm of demixing matrices [15, 16] for AuxIVA to ILRMA.



(a) Block length: 1 [s]



(b) Block length: 2 [s]



(c) Block length: 5 [s]

Figure 3: Average SI-SDR improvements with speech signals for $L = 1, 2, \ldots, 10$.

5. ACKNOWLEDGEMENT

This work was supported by JSPS KAKENHI Grant Number JP16H01735 and JST CREST Grant Number JP-

MJCR19A3, Japan.

6. REFERENCES

[1] 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 Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1622–1637, 2016.

[2] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2011, pp. 189–192.

[3] 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.

[4] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proceedings of Neural Information Processing Systems conference*, Dec. 2001, pp. 556–562.

[5] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, Mar. 2007.

[6] H. Kameoka, N. Ono, K. Kashino, and S. Sagayama, "Complex NMF: A new sparse representation for acoustic signals," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, Apr. 2009, pp. 3437–3440.

[7] C. Févotte, N. Bertin, and J. Durrieu, "Nonnegative matrix factorization with the Itakura–Saito divergence: With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, Mar. 2009.

[8] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, Mar. 2010.

[9] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, Sep. 2010.

[10] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 971–982, May 2013.

[11] A. Koutvas, E. Dermatas, and G. Kokkinakis, "Blind speech separation of moving speakers in real reverberant environments," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, Jun. 2000, pp. 1133–1136.

[12] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Robust real-time blind source separation for moving speakers in a room," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, Apr. 2003, pp. V–469–V–472.

[13] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Blind source separation for moving speech signals using blockwise ICA and residual crosstalk subtraction," *IEICE Transactions on Fundamentals*, vol. E87-A, no. 8, Aug. 2004.

[14] B. Sallberg, N. Grbic, and I. Claesson, "Complex-valued independent component analysis for online blind speech extraction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1624–1632, Nov. 2008.

[15] T. Taniguchi, N. Ono, A. Kawamura, and S. Sagayama, "An auxiliary-function approach to online independent vector analysis for real-time blind source separation," in *Proceedings of Hands-Free Speech Communication and Microphone Arrays*, May 2014, pp. 107–111.

[16] M. Sunohara, C. Haruta, and N. Ono, "Low-latency real-time blind source separation for hearing aids based on time-domain implementation of online independent vector analysis with truncation of non-causal components," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, Mar. 2017, pp. 216–220.

[17] A. Yeredor, "On hybrid exact-approximate joint diagonalization," in *Proceedings of IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, Dec. 2009, pp. 312–315.

[18] N. Ono, "Fast algorithm for independent component/vector/low-rank matrix analysis with three or more sources," in *Proceedings of 2018 Spring Meeting of Acoustical Society of Japan*, Mar. 2018 (in Japanese), pp. 437–438.

[19] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *The American Statistician*, vol. 58, no. 1, pp. 30–37, 2004.

[20] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *Journal of the Acoustical Society of Japan E*, vol. 20, no. 3, pp. 199–206, May 1999.

[21] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A Python package for audio room simulation and array processing algorithms," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, Apr. 2018, pp. 351–355.

[22] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR — half-baked or well done?" in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, May 2019, pp. 626–630.