Experimental Analysis of Optimal Window Length for Independent Low-Rank Matrix Analysis

Daichi Kitamura*, Nobutaka Ono†*, and Hiroshi Saruwatari‡
*SOKENDAI (The Graduate University for Advanced Studies), Kanagawa 240–0193, Japan
†National Institute of Informatics, Tokyo 101–8430, Japan
‡The University of Tokyo, Tokyo 113–8656, Japan

***Abstract*—In this paper, we address the blind source separation (BSS) problem and analyze the optimal window length in the short-time Fourier transform (STFT) for independent low-rank matrix analysis (ILRMA). ILRMA is a state-of-the-art BSS technique that utilizes the statistical independence between low-rank matrix spectrogram models, which are estimated by nonnegative matrix factorization. In conventional frequency-domain BSS, the modeling error of a mixing system increases when the window length is too short, and the accuracy of statistical estimation decreases when the window length is too long. Therefore, the optimal window length is determined by both the reverberation time and the number of time frames. However, unlike classical BSS methods such as ICA and IVA, ILRMA enables the full modeling of spectrograms, which may improve the robustness to a decrease in the number of frames in a longer-window case. To confirm this hypothesis, the optimal window length for ILRMA is experimentally investigated, and the difference between the performances of ILRMA and conventional BSS is discussed.**

I. Introduction

Source separation is a technique for estimating specific source signals from observed mixture signals. Many approaches have been developed for single-channel and multichannel observations. Blind source separation (BSS) in determined and overdetermined cases (number of channels ≥ number of sources) has been well studied so far [1]–[10]. BSS does not require any prior information about the recording environment or the locations of sources or sensors. In particular, independent component analysis (ICA) [1] and its extensions, frequency-domain ICA (FDICA) [2]–[7] and independent vector analysis (IVA) [8]–[10], are the most popular methods for solving the BSS problem of audio signals. These methods exploit the statistical independence between specific sources and estimate a demixing matrix for the separation. For both ICA and IVA, fast and stable update rules, which are derived by an auxiliary function technique, have been proposed [11], [12].

As another means of solving audio source separation, nonnegative matrix factorization (NMF) [13], [14] is widely used for both blind and informed source separation [15]–[20]. NMF is a parts-based low-rank decomposition and can extract some meaningful spectral patterns (bases) with their time-varying gains (activations) from an observed spectrogram. In [21] and [22], a multichannel extension of NMF (multichannel NMF: MNMF) was proposed, which clusters the decomposed bases and activations into each source using estimated spatial parameters.

Recently, a new BSS method that unifies IVA and NMF was proposed by the authors [23]–[25], which is called *independent low-rank matrix analysis (ILRMA)* in this paper[1]. Similarly to MNMF, ILRMA exploits the NMF decomposition of the estimated source spectrograms as a low-rank spectral model and optimizes the frequency-wise demixing matrix based on the independence between the spectral models. This NMF-based spectral model in ILRMA (low-rank matrix) can be interpreted as a natural extension of those in FDICA (scalar) and IVA (vector).

The separation result of all ICA-based frequency-domain BSS methods strongly depends on the length of the analysis window in the short-time Fourier transform (STFT). This is because the modeling error of a mixing system increases when the window length is too short, and the accuracy of statistical estimation decreases when the window length is too long (fewer time frames) [4], [26]. However, unlike classical BSS methods such as ICA and IVA, ILRMA enables the full modeling of spectrograms, which may improve the robustness to a decrease in the number of frames in a longer-window case. In this paper, to confirm this hypothesis, we experimentally compare the optimal window lengths for FDICA, IVA, and ILRMA, and discuss the difference in their performances.

II. Related Frequency-Domain BSS Algorithms

A. Formulation

Let N and M be the numbers of sources and channels, respectively. The complex-valued source, observed, and estimated signals are defined as $\boldsymbol{s}_{ij} = (s_{ij,1}, \cdots, s_{ij,N})^{\mathrm{T}}$, $\boldsymbol{x}_{ij} = (x_{ij,1}, \cdots, x_{ij,M})^{\mathrm{T}}$, and $\boldsymbol{y}_{ij} = (y_{ij,1}, \cdots, y_{ij,N})^{\mathrm{T}}$, where $i = 1, \cdots, I$; $j = 1, \cdots, J$; $n = 1, \cdots, N$; and $m = 1, \cdots, M$ are the integral indexes of the frequency bins, time frames, sources, and channels, respectively, and $^{\mathrm{T}}$ denotes a transpose. We also describe the spectrograms of the source, observed, and estimated signals as $\boldsymbol{S}_n \in \mathbb{C}^{I\times J}$, $\boldsymbol{X}_m \in \mathbb{C}^{I\times J}$, and $\boldsymbol{Y}_n \in \mathbb{C}^{I\times J}$, whose elements are $s_{ij,n}$, $x_{ij,m}$, and $y_{ij,n}$, respectively. In FDICA, IVA, and ILRMA, the following mixing system is assumed:

$$\boldsymbol{x}_{ij} = \boldsymbol{A}_i \boldsymbol{s}_{ij}, \tag{1}$$

where $\boldsymbol{A}_i = (\boldsymbol{a}_{i,1} \ \cdots \ \boldsymbol{a}_{i,N}) \in \mathbb{C}^{M\times N}$ is a frequency-wise mixing matrix and $\boldsymbol{a}_{i,n}$ is the steering vector for the nth source. This

[1]Note that ILRMA was called *rank-1 MNMF* in [23]–[25]. We have renamed the method to clarify that ILRMA is a natural extension of IVA.

ISBN 978-0-9928626-7-1 © EURASIP 2017

mixing system is called a linear time-invariant mixture or the rank-1 spatial model [27]. Thus, the estimated signal $\boldsymbol{y}_{ij}$ can be obtained by assuming $M=N$ and estimating the frequency-wise demixing matrix $\boldsymbol{W}_i=(\boldsymbol{w}_{i,1}\ \cdots\ \boldsymbol{w}_{i,N})^{\mathrm{H}}=\boldsymbol{A}_i^{-1}$ as

$$\boldsymbol{y}_{ij}=\boldsymbol{W}_i\boldsymbol{x}_{ij}, \tag{2}$$

where $\boldsymbol{w}_{i,n}$ is the demixing filter for the nth source and $^{\mathrm{H}}$ denotes a Hermitian transpose. The objective in FDICA, IVA, or ILRMA is to estimate both $\boldsymbol{W}_i$ and $\boldsymbol{y}_{ij}$ from only the observation $\boldsymbol{x}_{ij}$ assuming the statistical independence between $s_{ij,n}$ and $s_{ij,n'}$, where $n'\neq n$.

B. FDICA and IVA

In FDICA [2]–[7], a robust BSS method for reverberant observations, ICA is applied to the frequency-wise signal $(x_{i1,m},\cdots,x_{iJ,m})$ while assuming a non-Gaussian source distribution $p(s)\approx p(y)$. Since the permutation of the estimated signals at each frequency must be aligned, various permutation solvers have been proposed. IVA [8]–[10] is one of the most elegant solutions of the permutation problem. IVA formulates the frequency components as a vector $\bar{\boldsymbol{x}}_{j,m}=(x_{1j,m},\cdots,x_{Ij,m})^{\mathrm{T}}$ and applies multivariate ICA to the vector signal $(\bar{\boldsymbol{x}}_{1,m},\cdots,\bar{\boldsymbol{x}}_{J,m})$ to estimate the frequency-wise demixing matrix $\boldsymbol{W}_i$, where the source vector $\bar{\boldsymbol{s}}_{j,n}=(s_{1j,n},\cdots,s_{Ij,n})^{\mathrm{T}}$ is assumed to have a spherical I-dimensional non-Gaussian source distribution $p(\bar{\boldsymbol{s}})$ [10]. This spherical property ensures higher-order dependences among the frequency components in $\bar{\boldsymbol{s}}_{j,n}$, thus avoiding the permutation problem.

C. ILRMA

ILRMA extends the source model $p(\bar{\boldsymbol{s}})$ in IVA to the following time-varying distribution:

$$\prod_{i,j}p(y_{ij,n})=\prod_{i,j}\frac{1}{\pi r_{ij,n}}\exp\left(-\frac{|y_{ij,n}|^2}{r_{ij,n}}\right), \tag{3}$$

where the local distribution $p(y_{ij,n})$ is defined as a circularly symmetric (isotropic) complex Gaussian distribution, namely, the probability of $p(y_{ij,n})$ only depends on the power of the complex value $y_{ij,n}$. Also, $r_{ij,n}$ is a time-frequency-varying nonnegative variance and corresponds to the expectation of the power of $y_{ij,n}$, namely, $\mathrm{E}[|y_{ij,n}|^2]$. This is because $p(y_{ij,n})$ is isotropic in the complex plane. Since the variance $r_{ij,n}$ can fluctuate depending on the time frames, (3) becomes a non-Gaussian distribution. The negative log-likelihood function $\mathcal{L}$ based on (3) can be obtained as follows by assuming the independence between each source and each time frame:

$$\mathcal{L}=\text{const.}-2J\sum_i\log|\det\boldsymbol{W}_i|+\sum_{i,j,n}\left(\log r_{ij,n}+\frac{|y_{ij,n}|^2}{r_{ij,n}}\right). \tag{4}$$

ILRMA applies Itakura–Saito-divergence-based NMF (IS-NMF) to $\boldsymbol{Y}_n$. In ISNMF [28], the decomposition $y_{ij,n}=\sum_l c_{ij,nl}$ is assumed, where $l=1,\cdots,L$ is the integral index and L is set to a much smaller value than $\min(I,J)$. The components $c_{ij,nl}$ are assumed to be mutually independent and obey

$$p(c_{ij,nl})=\frac{1}{\pi t_{il,n}v_{lj,n}}\exp\left(-\frac{|c_{ij,nl}|^2}{t_{il,n}v_{lj,n}}\right), \tag{5}$$

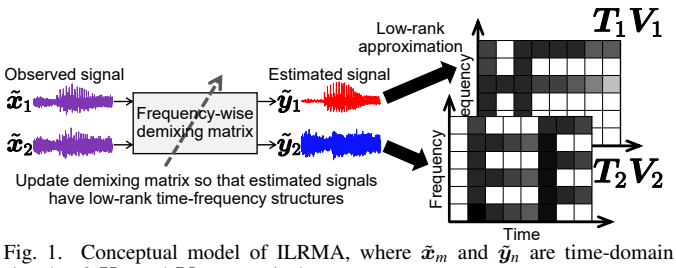

Fig. 1. Conceptual model of ILRMA, where $\tilde{\boldsymbol{x}}_m$ and $\tilde{\boldsymbol{y}}_n$ are time-domain signals of $\boldsymbol{X}_m$ and $\boldsymbol{Y}_n$, respectively.

where $t_{il,n}$ and $v_{lj,n}$ are the basis and activation, respectively, and $t_{il,n}v_{lj,n}=\mathrm{E}[|c_{ij,nl}|^2]$. Because of the reproductive property of (5), $y_{ij,n}(=\sum_l c_{ij,nl})$ obeys (3) with the variance $r_{ij,n}=\sum_l t_{il,n}v_{lj,n}$. This fact means that the additivity of the power spectrogram holds in an expectation sense [28], which provides a justification for decomposing the power spectrogram. Therefore, the power spectrogram of the estimated source is approximately decomposed with a fixed number of bases and activations as $|\boldsymbol{Y}_n|^{\cdot 2}\approx\boldsymbol{T}_n\boldsymbol{V}_n$, where the absolute value and the dotted exponent for a matrix denote an element-wise absolute and exponent, respectively, and $\boldsymbol{T}_n\in\mathbb{R}_{\geq 0}^{I\times L}$ and $\boldsymbol{V}_n\in\mathbb{R}_{\geq 0}^{L\times J}$ are the basis and activation matrices for the nth source, respectively. The estimation of $\boldsymbol{W}_i$, $\boldsymbol{T}_n$, and $\boldsymbol{V}_n$ can consistently be carried out by minimizing (4) in a fully blind manner. Note that ILRMA is theoretically equivalent to conventional MNMF only when the rank-1 spatial model is assumed, which yields a stable and computationally efficient algorithm for ILRMA. This issue and the convergence-guaranteed fast update rules for $\boldsymbol{W}_i$, $\boldsymbol{T}_n$, and $\boldsymbol{V}_n$ can be found in [25].

Fig. 1 shows the conceptual model of ILRMA. When original sources have a low-rank spectrogram $|\boldsymbol{S}_n|^{\cdot 2}$, the spectrogram of their mixture $|\boldsymbol{X}_m|^{\cdot 2}$ should be more complicated, namely, the rank of $|\boldsymbol{X}_m|^{\cdot 2}$ will be greater than that of $|\boldsymbol{S}_n|^{\cdot 2}$. On the basis of this assumption, in ILRMA, the low-rank constraint for each estimated spectrogram $|\boldsymbol{Y}_n|^{\cdot 2}$ is introduced by employing NMF. The demixing matrix $\boldsymbol{W}_i$ is estimated so that the spectrogram of estimated signal $|\boldsymbol{Y}_n|^{\cdot 2}$ becomes a low-rank matrix modeled by $\boldsymbol{T}_n\boldsymbol{V}_n$, whose rank is at most L.

III. Experimental Analysis of Optimal Window Length

A. Motivation

In the practical use of frequency-domain BSS, the length of the analysis window in STFT directly affects the separation performance. For instance, a decrease in performance for shorter- or longer-window cases in FDICA was reported in [4]. When the window length is too short, the separation fails because the mixing assumption (1) does not hold owing to the reverberation. In contrast, when the window length is too long, the statistical estimation in ICA fails because the number of time frames J decreases. IVA and ILRMA also suffers from this problem because they obviously cannot estimate the demixing matrix $\boldsymbol{W}_i$ when $J=1$. However, the full modeling of the $I\times J$ spectrogram in ILRMA may improve the robustness to a decrease in the number of frames in a longer-window case (fewer time frames). In this section, we experimentally

ISBN 978-0-9928626-7-1 © EURASIP 2017

TABLE I
MUSIC AND SPEECH SOURCES OBTAINED FROM SISEC2011

Signal	Data name	Source (1/2)	Length [s]
Music	bearlin-roads	acoustic_guit_main/vocals	14.6
Music	another_dreamer-the_ones_we_love	guitar/vocals	25.6
Music	fort_minor-remember_the_name	violins_synth/vocals	24.6
Music	ultimate_nz_tour	guitar/synth	18.6
Speech	dev1_female4	src_1/src_2	10.0
Speech	dev1_female4	src_3/src_4	10.0
Speech	dev1_male4	src_1/src_2	10.0
Speech	dev1_male4	src_3/src_4	10.0

compare the optimal window lengths for FDICA, IVA, and ILRMA and discuss the difference in their performances.

B. Dataset and Experimental Conditions

In this experiment, we used four music and four speech observations, as shown in Table I, where each observation includes two sources. These dry sources were obtained from professionally produced music and underdetermined separation tasks in SiSEC2011 [29]. To simulate the reverberant mixture, the observed signals were produced by convoluting the impulse response E2A (T_{60} = 300 ms) or JR2 (T_{60} = 470 ms), which was obtained from the RWCP [30], with each source. Fig. 2 shows the recording conditions of the impulse responses. Note that all the separation tasks are determined, namely, $N = M = 2$.

We compared three BSS methods, namely, FDICA, IVA, and ILRMA. For FDICA, two blind and ideal permutation solvers were employed and compared: FDICA+DOA and FDICA+IPS. FDICA+DOA solves the permutation problem by clustering the components using the relative locations of microphones and the estimated direction of arrival (DOA) [3], and FDICA+IPS utilizes the reference (oracle) source spectrograms $\boldsymbol{S}_n$ to align the permutations, which is an ideal permutation solver (IPS). All the optimizations in FDICA, IVA, and ILRMA were based on an auxiliary function technique [11], [12], [25]. The other experimental conditions are described in Table II. As an evaluation score of the separation performance, we used the improvement of the signal-to-distortion ratio (SDR) [31].

C. Comparison Using Ideal Initialization

To compare the net separation ability for each setting of the window length, in this subsection, the initial values of the spatial and spectral parameters in each BSS method are set to their ideal values. For the spatial parameter, the initial demixing matrix $\boldsymbol{W}_i^{(\mathrm{initial})}$ is set to its optimal value

$$\boldsymbol{W}_i^{(\mathrm{initial})} = \left(\sum_j \boldsymbol{s}_{ij}\boldsymbol{s}_{ij}^{\mathrm{H}}\right)\left(\sum_j \boldsymbol{x}_{ij}\boldsymbol{s}_{ij}^{\mathrm{H}}\right)^{-1}, \tag{6}$$

which gives the best separation performance under the linear mixing assumption (1). In addition, only for ILRMA, source-wise initial basis and activation matrices, $\boldsymbol{T}_n^{(\mathrm{initial})}$ and $\boldsymbol{V}_n^{(\mathrm{initial})}$, are pretrained by ISNMF using the oracle power spectrogram given by

$$\left(\boldsymbol{T}_n^{(\mathrm{initial})}, \boldsymbol{V}_n^{(\mathrm{initial})}\right) = \underset{\boldsymbol{T}_n, \boldsymbol{V}_n}{\arg\min}\, \mathcal{D}_{\mathrm{IS}}\left(|\boldsymbol{S}_n|^2 \| \boldsymbol{T}_n\boldsymbol{V}_n\right), \tag{7}$$

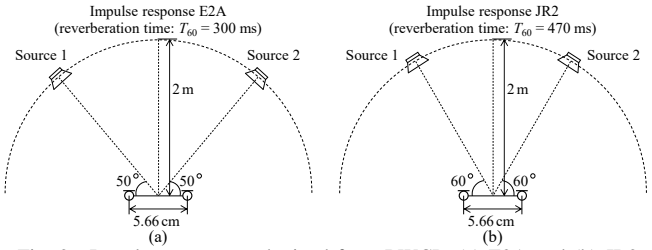

Fig. 2. Impulse responses obtained from RWCP: (a) E2A and (b) JR2.

TABLE II
EXPERIMENTAL CONDITIONS

Window function	Hamming window
Window length	32/64/128/256/512/768/1024/ 1280/1536/1792/2048 ms
Window shift length	Quarter of window length
Number of bases for each source in ILRMA	5/10/30/50 for music signals and 2/3/4/10 for speech signals
Number of iterations	100

where $\mathcal{D}_{\mathrm{IS}}(\cdot\|\cdot)$ is the element-wise Itakura–Saito divergence. Therefore, in this experiment, FDICA+DOA and IVA are based on the spatial oracle initialization, and FDICA+IPS and ILRMA are based on the spatial and spectral oracle initialization. Since the separation performance is obviously maximized for the initial parameter given by (6), this experiment illustrates how the performance decreases at the converged solution for the model used in each method.

The results are shown in Figs. 3 and 4, where the scores are averaged over the observed signals with the same impulse response. As already mentioned in Sect. III-A, the separation with a shorter window is highly limited in all the methods because the assumption of a linear mixture model (1) collapses (the reverberation time exceeds the window length). For the longer-window case, the performance of FDICA and IVA deteriorates when the length exceeds $2T_{60}$ even if the oracle source spectrogram is employed in FDICA+IPS. This instability in the statistical estimation is caused by the insufficient number of time frames J [4]. On the other hand, for the music signals (Fig. 3), ILRMA maintains its separation accuracy even for windows longer than 1 s. This is a benefit of employing the full modeling of time-frequency dependences, and the robustness to fewer time frames is improved by the low-rank spectrogram modeling. From this result, we can confirm that a longer window length exceeding $2T_{60}$ is preferable for music source separation using ILRMA, whereas FDICA achieves the highest performance when the length is set to less than $2T_{60}$. However, this behavior does not appear in the results for speech signals (Fig. 4). This is because the low-rank assumption in ILRMA does not apply to the speech signals, and the spectral model cannot capture the precise source spectrogram during the optimization.

Since the NMF parameters are pretrained using (7), an increase in the number of bases directly improves the accuracy of the spectral model $\boldsymbol{T}_n\boldsymbol{V}_n$ and the separation performance of ILRMA. This means that improving the precision of the spectral model will provide a better estimation of $\boldsymbol{W}_i$, as predicted in [10].

ISBN 978-0-9928626-7-1 © EURASIP 2017

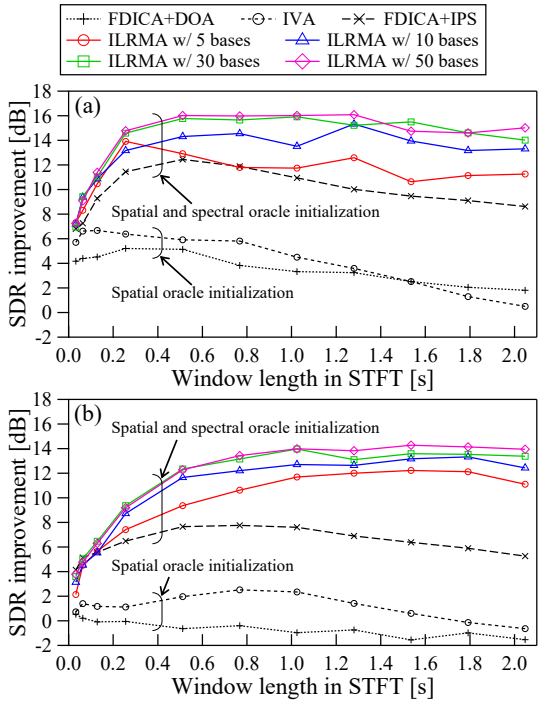

Fig. 3. Average results for music signals using ideal initialization: (a) E2A ($T_{60}=300$ ms) and (b) JR2 ($T_{60}=470$ ms).

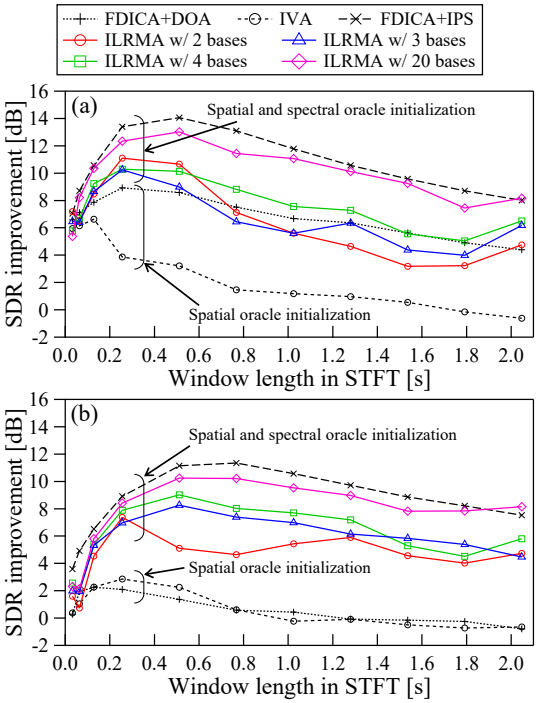

Fig. 4. Average results for speech signals using ideal initialization: (a) E2A ($T_{60}=300$ ms) and (b) JR2 ($T_{60}=470$ ms).

D. Comparison Using Random Initialization

In this subsection, the separation performance in a practical situation is compared for various window lengths. The initial demixing matrix $\boldsymbol{W}_i^{(\mathrm{initial})}$ was set to the identity matrix in all the methods, and the initial NMF matrices $\boldsymbol{T}_n^{(\mathrm{initial})}$ and $\boldsymbol{V}_n^{(\mathrm{initial})}$ were set to nonnegative uniform random values. Therefore, FDICA+DOA only utilizes the knowledge of the microphone spacing, FDICA+IPS still exploits $|\boldsymbol{S}_n|^{\cdot 2}$ for IPS, and the other methods are fully blind.

The results are shown in Figs. 5 and 6. In this experiment, ILRMA cannot maintain its accuracy for longer windows, and the optimal length in ILRMA is almost the same as those in FDICA and IVA. This means that the blind estimation of a precise spectral model is a difficult problem, and the robustness of ILRMA against fewer time frames is deteriorated.

The number of bases L does not strongly affect the performance in the music separation task (Fig. 5). For the speech signals (Fig. 6), as reported in [25], a small number of bases is preferable, whereas spectrograms of speech signals do not have the low-rank property. For speech signals, the estimation of $\boldsymbol{T}_n\boldsymbol{V}_n$ using a large number of bases always fails to capture the precise source spectrograms $|\boldsymbol{S}_n|^{\cdot 2}$ because of the difficulty in optimization, and a rough and broad spectral model with a small number of bases can stably separate the speech sources.

Since FDICA+IPS achieves high separation accuracy even for speech signals, we have significant scope to improve speech BSS using the linear mixing model (1), which yields a computationally efficient solution. The blind capture of complicated (not low-rank) spectrograms requires another criterion, such as sparseness or time-varying speech structures, which can be considered as a further study.

IV. Conclusion

We presented an experimental analysis of optimal window lengths for FDICA, IVA, and ILRMA. Since ILRMA employs not only the independence between sources but also a time-frequency structure for the estimation of a demixing matrix, the robustness to long windows (fewer time frames) can be improved. However, in a practical situation, the optimal window length of ILRMA was similar to that in IVA or FDICA, which shows the difficulty of the blind estimation of a precise spectral model in ILRMA.

Acknowledgments

This work was partly supported by Grant-in-Aid for JSPS Fellows Number 26 · 10796, ImPACT Program of Council for Science, and SECOM Science and Technology Foundation.

References

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

[2] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, pp. 21–34, 1998.

[3] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura, "Evaluation of blind signal separation method using directivity pattern under reverberant conditions," in *Proc. ICASSP*, vol. 5, pp. 3140–3143, 2000.

[4] 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. SAP*, vol. 11, no. 2, pp. 109–116, 2003.

[5] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Convolutive blind source separation for more than two sources in the frequency domain," in *Proc. ICASSP*, pp. III-885–III-888, 2004.

[6] H. Buchner, R. Aichner, and W. Kellerman, "A generalization of blind source separation algorithms for convolutive mixtures based on second order statistics," *IEEE Trans. SAP*, vol. 13, no. 1, pp. 120–134, 2005.

ISBN 978-0-9928626-7-1 © EURASIP 2017

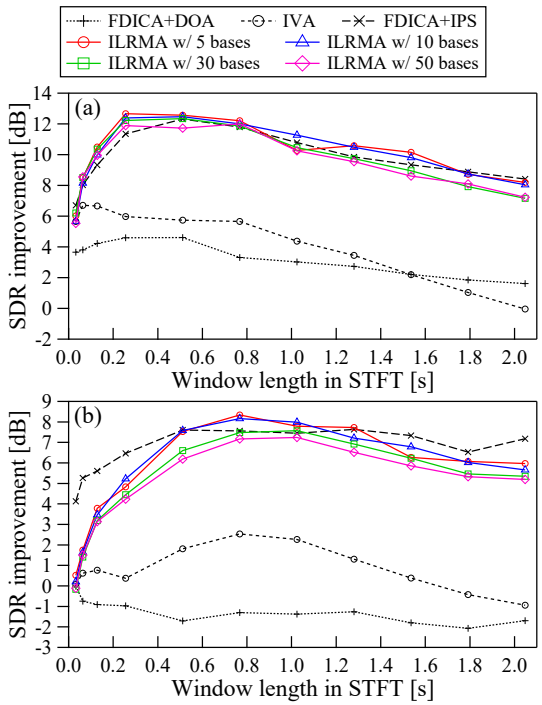

Fig. 5. Average results for music signals using random initialization: (a) E2A ($T_{60}=300$ ms) and (b) JR2 ($T_{60}=470$ ms).

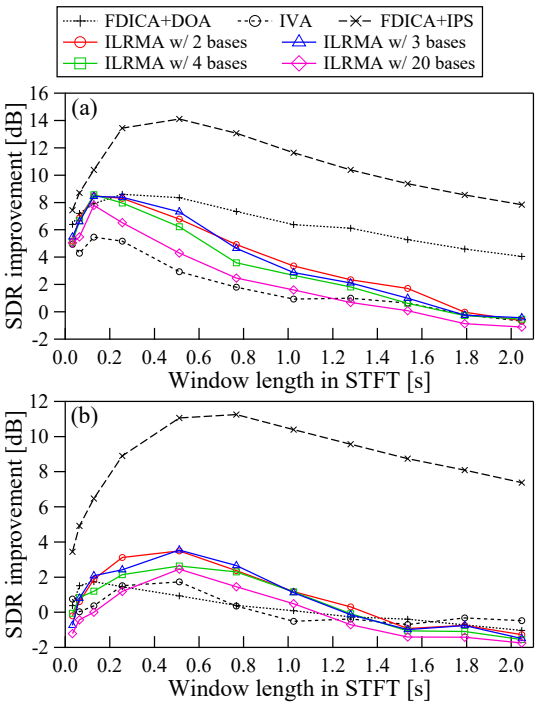

Fig. 6. Average results for speech signals using random initialization: (a) E2A ($T_{60}=300$ ms) and (b) JR2 ($T_{60}=470$ ms).

[7] 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. ASLP*, vol. 14, no. 2, pp. 666–678, 2006.

[8] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: an extension of ICA to multivariate components," in *Proc. ICA*, pp. 165–172, 2006.

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

[10] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. ASLP*, vol. 15, no. 1, pp. 70–79, 2007.

[11] N. Ono and S. Miyabe, "Auxiliary-function-based independent component analysis for super-Gaussian sources," in *Proc. LVA/ICA*, pp. 165–172, 2010.

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

[13] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[14] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. NIPS*, vol. 13, pp. 556–562, 2000.

[15] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. ASLP*, vol. 15, no. 3, pp. 1066–1074, 2007.

[16] A. Ozerov, C. Févotte, and M. Charbit, "Factorial scaled hidden Markov model for polyphonic audio representation and source separation," in *Proc. WASPAA*, pp. 121–124, 2009.

[17] H. Kameoka, M. Nakano, K. Ochiai, Y. Imoto, K. Kashino, and S. Sagayama, "Constrained and regularized variants of non-negative matrix factorization incorporating music-specific constraints," in *Proc. ICASSP*, pp. 5365–5368, 2012.

[18] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Proc. ICA*, pp. 414–421, 2007.

[19] D. Kitamura, H. Saruwatari, K. Yagi, K. Shikano, Y. Takahashi, and K. Kondo, "Music signal separation based on supervised nonnegative matrix factorization with orthogonality and maximum-divergence penalties," *IEICE Trans. Fundamentals*, vol. E97-A, no. 5, pp. 1113–1118, 2014.

[20] D. Kitamura, H. Saruwatari, H. Kameoka, Y. Takahashi, K. Kondo, and S. Nakamura, "Multichannel signal separation combining directional clustering and nonnegative matrix factorization with spectrogram restoration," *IEEE/ACM Trans. ASLP*, vol. 23, no. 4, pp. 654–669, 2015.

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

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

[23] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Efficient multichannel nonnegative matrix factorization exploiting rank-1 spatial model," in *Proc. ICASSP*, pp. 276–280, 2015.

[24] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Relaxation of rank-1 spatial constraint in overdetermined blind source separation," in *Proc. EUSIPCO*, pp. 1271–1275, 2015.

[25] 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. ASLP*, vol. 24, no. 9, pp. 1626–1641, 2016.

[26] T. Nishikawa, H. Saruwatari, and K. Shikano, "Blind source separation of acoustic signals based on multistage ICA combining frequency-domain ICA and time-domain ICA," *IEICE Trans. Fundamentals*, vol. E86-A, no. 4, pp. 846–858, 2003.

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

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

[29] S. Araki, F. Nesta, E. Vincent, Z. Koldovský, G. Nolte, A. Ziehe, and A. Benichoux, "The 2011 signal separation evaluation campaign (SiSEC2011):-audio source separation," in *Proc. LVA/ICA*, pp. 414–422, 2012.

[30] 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 *Proc. LREC*, pp. 965–968, 2000.

[31] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.

ISBN 978-0-9928626-7-1 © EURASIP 2017