INDEPENDENT LOW-RANK MATRIX ANALYSIS BASED ON MULTIVARIATE COMPLEX EXPONENTIAL POWER DISTRIBUTION

Rintaro Ikeshita and Yohei Kawaguchi

Hitachi, Ltd. Research & Development Group, Tokyo, Japan

ABSTRACT

Independent low-rank matrix analysis (ILRMA), a unified method of independent vector analysis (IVA) and nonnegative matrix factorization (NMF), is a state-of-the-art blind source separation method for convolutive mixtures. Although ILRMA provides high separation performance for music signals whose spectra can be well modeled by NMF, speech spectra do not have low-rank properties, and modeling them by NMF is not appropriate. In this paper, to stably improve the separation performance of ILRMA for speech mixtures, a source spectrum model in ILRMA is generalized to explicitly model the strong higher-order correlations between neighboring frequency bins of speech signals. In addition, multivariate complex exponential power distributions, which are recognized to have high performance with IVA, are introduced as source distributions assumed in ILRMA. Experimental results show the effectiveness of the proposed method over the original ILRMA when separating speech mixtures.

***Index Terms*—** Blind source separation, independent component analysis, independent vector analysis, nonnegative matrix factorization, multivariate exponential power distribution

1. INTRODUCTION

Blind source separation (BSS) is a technique that estimates original source signals from a given mixture without any knowledge of mixing systems or microphone positions. The problem of BSS for convolutive mixtures is generally addressed in the time-frequency (TF) domain using the short-term Fourier transform [1]. For the determined situation where the number of sources is less than or equal to the number of microphones, frequency-domain independent component analysis (FD-ICA [1]) and independent vector analysis (IVA [2, 3, 4]) are common, and these methods have been applied to preprocessing for speech recognition tasks for multiple speakers.

FD-ICA and IVA are the methods that perform BSS by relying only on independence between source signals as a clue for separation. In recent years, in addition to independence between sources, efforts to exploit the time-frequency structure of source signals, or source spectrum, have been made to improve the separation performance of FD-ICA and IVA. Among them, low-rank approximation of source spectra by using nonnegative matrix factorization (NMF [5, 6, 7, 8, 9, 10]) has received attention and has been incorporated into the spectrum models assumed in FD-ICA and IVA. This approach is called independent low-rank matrix analysis (ILRMA [11, 12, 13]), and it was reported to outperform the performance of FD-ICA and IVA when separating music signals, which have remarkable co-occurrence of temporal frequency components [11, 12]. On the other hand, since the spectra of speech signals do not have low-rank properties, NMF modeling in ILRMA [11, 12, 13] is not necessarily appropriate for separating speech mixtures.

When ILRMA was first proposed by Kitamura *et al*. [11, 12], a (time-varying) complex Gaussian distribution was used as a source distribution. This Gaussian ILRMA can be viewed as a Multichannel NMF (MNMF [14, 15]), which is a multichannel extension of Itakura-Saito NMF [6], rewritten as an optimization problem of the demixing system. Since the degree of freedom of the model is large in MNMF, the sensitivity of the optimization of MNMF to the initial model parameters and the rather large processing time required for optimization were reported as cumbersome problems [12, 14, 15]. On the other hand, since ILRMA has fewer parameters than MNMF and can utilize the fast and stable optimization algorithm based on the auxiliary-function-based IVA (AuxIVA [16, 17, 18]), ILRMA can achieve a more efficient and robust separation compared with MNMF.

Recently, in order to improve the stability and the separation performance of NMF and MNMF, distributions with heavier tails than the complex Gaussian distribution have been introduced as source distributions [7, 8, 9, 10, 19, 20]. As for ILRMA, inspired by the studies to extend the source distribution to the complex Student's t-distribution in NMF [8] and MNMF [19], ILRMA based on the complex Student's t-distribution, called t-ILRMA, was proposed [13]. t-ILRMA can realize the Gaussian ILRMA [11, 12] and ILRMA based on the complex Cauchy distribution, which has a heavy tail as a result of adjusting the degree of freedom parameter in Student's t-distribution. However, since the separation performance of t-ILRMA with heavy tails depends on the initial model parameters and the degree of freedom parameter, it is difficult to stably improve the separation performance of Gaussian ILRMA [13]. Furthermore, t-ILRMA still uses a low-rank approximation of source spectra by using NMF, and so the model in t-ILRMA is not appropriate for speech signals.

In this paper, to stably improve the separation performance of ILRMA, we explicitly model in ILRMA the property of speech spectra in which the statistical dependencies between neighboring frequency bins are stronger than the dependencies between distant bins. The attempts to model the above speech property were successful in improving the performance of IVA [21, 22, 23, 24], and we will generalize this approach for ILRMA (see Subsection 2.2). With this modeling, the proposed ILRMA is expected to alleviate the adverse effect of low-rank modeling of speech spectra in the original ILRMA, and to stably improve the separation performance. Furthermore, we generalize the source distribution assumed in Gaussian ILRMA to multivariate complex exponential power (MEP) distribution (see Subsection 2.3), which is a different extension from t-ILRMA. The MEP can represent a distribution with a heavy tail and a large kurtosis, and was reported to demonstrate high separation performance when used in IVA [18]. Therefore, the proposed ILRMA is expected to further improve the separation performance. The experimental results show that the proposed method is more effective than the Gaussian ILRMA and IVA when separating speech mixtures.

2. PROPOSED GENERATIVE MODEL

2.1. Formulation

Suppose that N sources are observed by N microphones. The source signal and the microphone observation in each time-frequency slot $(f,t) \in [N_F] \times [N_T]$ are denoted as

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

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

where $[N_F] := \{1, \ldots, N_F\}$ and $[N_T] := \{1, \ldots, N_T\}$ denote the set of frequency bins and time frames, respectively, and $\cdot^\top$ means the matrix transpose. This paper deals with the linear mixing system given by

$$\boldsymbol{x}_{f,t} = A_f \boldsymbol{s}_{f,t}, \quad \boldsymbol{s}_{f,t} = W_f^h \boldsymbol{x}_{f,t}, \tag{3}$$

where $A_f \in \mathbb{C}^{N \times N}$ and $W_f \in \mathbb{C}^{N \times N}$ denote the mixing and demixing matrices for frequency $f \in [N_F]$, respectively, and $\cdot^h$ means the matrix conjugate transpose. Note that the demixing matrix W_f for frequency $f \in [N_F]$ is composed of the separation filters $\boldsymbol{w}_{n,f} \in \mathbb{C}^N$ for each source $n \in [N] := \{1, \ldots, N\}$ as follows:

$$W_f = [\boldsymbol{w}_{1,f}, \ldots, \boldsymbol{w}_{N,f}] \in \mathbb{C}^{N \times N}. \tag{4}$$

In the following, for the sake of simplicity, we define

$$\boldsymbol{s}_{n,F,t} := [s_{n,f_1,t}, \ldots, s_{n,f_k,t}]^\top \in \mathbb{C}^{|F|} \tag{5}$$

for the set of frequency bins $F = \{f_1, \ldots, f_k\} \subseteq [N_F]$.

2.2. Introduction of frequency range division into ILRMA

Speech signals have a property in which the dependencies between neighboring frequency bins are stronger than those between distant bins. In the previous studies, by incorporating this property into the source models, efforts to improve the separation performance of IVA have been made [21, 22, 23, 24].

The source distribution assumed in [21, 22, 23] is given by

$$p(\{\boldsymbol{s}_{n,F,t}\}_{n,F,t}) \propto \prod_{n \in [N]} \prod_{t \in [N_T]} \exp\left\{ - \sum_{F \in \mathcal{F}} \left(\frac{\|\boldsymbol{s}_{n,F,t}\|^2}{\alpha_{n,F,t}} \right)^\beta \right\} \tag{6}$$

by using a hypergraph $([N_F], \mathcal{F})$[1] satisfying the following two conditions (C1) and (C2):

(C1) $\bigcup_{F \in \mathcal{F}} F = [N_F]$;

(C2) There is a path between i and j for arbitrary $i, j \in [N_F]$. [2]

Here, $\alpha_{n,F,t}$ and β are constants, and $\| \cdot \|$ denotes the L^2-norm. Owing to (C1) and (C2), each pair of frequency bins has the higher-order correlation in (6), and hence the approaches in [21, 22, 23] can also avoid the permutation problem in the same principle as the original IVA [3, 4]. However, since the normalization term in (6) cannot generally be obtained analytically because of (C2), even if the source spectrum, or its counterpart $\{\alpha_{n,F,t}\}_{n,F,t}$, is modeled by, e.g., NMF, the model parameters cannot be optimized in a statistical sense unlike the case in ILRMA.

On the other hand, the source model that we proposed in [24] is assumed to satisfy the following (C3) instead of (C2) for the set of hyperedges $\mathcal{F}$:

(C3) $F_1 \cap F_2 = \emptyset$ for every $F_1, F_2 \in \mathcal{F}$.

The set of hyperedges $\mathcal{F}$ satisfying (C1) and (C3) is called the *frequency range division* [24].

In this paper, by using the frequency range division $\mathcal{F}$, the distribution for $\{\boldsymbol{s}_{n,F,t}\}_{n,F,t}$ is decomposed as follows:

$$p(\{\boldsymbol{s}_{n,F,t}\}_{n,F,t}) = \prod_{n \in [N]} \prod_{F \in \mathcal{F}} \prod_{t \in [N_T]} p(\boldsymbol{s}_{n,F,t}). \tag{7}$$

Thanks to the decomposition (7) based on (C3), it is possible to analytically obtain the normalization term in (7) at the same time as modeling higher-order correlations in each frequency range $F \in \mathcal{F}$. This in turn enables us to upgrade the scale parameters (such as $\{\alpha_{n,F,t}\}_{n,F,t}$ in (8) below which define the time-frequency structure of source signals) from the time-invariant constants to time-varying variables to be estimated and model them by NMF in the same way as ILRMA (see Subsection 2.3 for details).

By modeling source spectra with both the frequency range division and NMF, the proposed model can express the stronger dependencies within neighboring frequency bins as well as the co-occurrence relation of the frequency components between distant frequency bins. We will call this approach ILRMA as well (for a detailed comparison with the original ILRMA [12], see Section 4).

2.3. ILRMA based on multivariate complex exponential power distribution

It is known that speech signals are better explained by distributions with greater kurtosis than Gaussian distribution. Therefore, following the studies for IVA [18, 24], as the distribution assumed in ILRMA, we use the multivariate complex exponential power (MEP) distribution (see, e.g., [25] for real-valued MEP) defined as

$$p(\boldsymbol{s}_{n,F,t}) = \frac{\Gamma(1+|F|) \cdot \exp\left\{ - \left(\frac{\|\boldsymbol{s}_{n,F,t}\|^2}{\alpha_{n,F,t}} \right)^\beta \right\}}{(\pi \alpha_{n,F,t})^{|F|} \cdot \Gamma(1 + \frac{|F|}{\beta})}, \tag{8}$$

where $\Gamma(\cdot)$ is the gamma function, $|F|$ denotes the cardinality of set $F \in \mathcal{F}$, and $\alpha_{n,F,t} \in \mathbb{R}_{>0}$ and $\beta \in \mathbb{R}_{>0}$ are the scale and shape parameters in MEP, respectively. Note that MEP with $\beta = 1$ is nothing but multivariate complex Gaussian distribution, and the smaller the value of β, the greater the kurtosis of the MEP.

As mentioned in Subsection 2.2, the scale parameters $\{\alpha_{n,F,t}\}_{F,t}$ for each source $n \in [N]$, which encodes the prior information of the source spectrum, are modeled by NMF as follows:

$$\alpha_{n,F,t} = \left(\sum_{k=1}^{K_n} u_{n,F,k} \cdot v_{n,k,t} \right)^d, \tag{9}$$

where K_n, $\{u_{n,F,k}\}_F$, and $\{v_{n,k,t}\}_t$ denote the number of NMF bases, k-th nonnegative base, and k-th nonnegative activation in NMF for source $n \in [N]$, respectively. Also, $d \in \mathbb{R}_{>0}$ is the heuristic parameter that defines the domain subject to NMF as it was used in [13]. If $d = 1$ or $d = 2$, then the domain of NMF corresponds to the power spectrum $\|s_{n,F,t}\|^2$ or the amplitude spectrum $\|s_{n,F,t}\|$, respectively.

[1] A *hypergraph* is a pair $(V, \mathcal{E})$ where V is a finite set and $\mathcal{E}$ is a set of non-empty subsets of V, namely, $\mathcal{E} \subseteq 2^V \setminus \{\emptyset\}$. The elements of V are called *vertices* and the elements of $\mathcal{E}$ are called *hyperedges*.

[2] In the hypergraph $(V, \mathcal{E})$, the *path* between $i \in V$ and $j \in V$ is said to exist if the following condition holds: There exists a sequence of hyperedges $e_1, \ldots, e_n$ and a sequence of vertices $v_0 (= i), v_1, \ldots, v_n (= j)$ such that $v_{k-1}, v_k \in e_k$ holds for $k = 1, \ldots, n$.

In the generative model defined by (3), (7), (8), and (9), the set of parameters for the demixing system is given by

$$\Theta = \{W_f,\, u_{n,F,k},\, v_{n,k,t}\}_{n,f,F,t,k}, \tag{10}$$

and it will be optimized by the maximum likelihood criterion:

$$\min_{\Theta} J(\Theta) := -\frac{1}{N_T}\sum_{n,F,t}\log p(\boldsymbol{s}_{n,F,t}) - 2\sum_{f}\log|\det W_f|. \tag{11}$$

Substituting (8)–(9) into the cost function $J(\Theta)$, we obtain

$$J(\Theta) = \frac{1}{N_T}\sum_{n,F,t}\left[|F|\cdot d\cdot\log\sum_{k}u_{n,F,k}\cdot v_{n,k,t} + \frac{r^{\beta}_{n,F,t}}{\left(\sum_k u_{n,F,k}\cdot v_{n,k,t}\right)^{\beta d}}\right] - 2\sum_{f}\log|\det W_f| + C, \tag{12}$$

where C is independent of the parameters Θ, and we define

$$r_{n,F,t} := \|\boldsymbol{s}_{n,F,t}\|^2 = \sum_{f\in F}|\boldsymbol{w}^h_{n,f}\boldsymbol{x}_{f,t}|^2. \tag{13}$$

In the following, we call this approach $(\mathcal{F},\beta,d)$-ILRMA since it is characterized by the frequency range division $\mathcal{F}$, the scale parameter β in MEP, and the NMF domain parameter d in (9).

3. OPTIMIZATION OF THE MODEL

In this section, an algorithm for solving the optimization problem (11) is derived. The update rules for the NMF parameters and the separation filters are derived in Subsections 3.1 and 3.2, respectively. To obtain an efficient algorithm for the separation filters, the shape parameter in MEP is considered only when $0<\beta\le 1$.

After the convergence of the optimization, the separated signals are obtained by (3), and the amplitude ambiguities can be restored by applying the projection back technique [26, 27] as follows:

$$s_{n,f,t}A_f e_n = (\boldsymbol{w}^h_{n,f}\boldsymbol{x}_{f,t})(W^h_f)^{-1}e_n \in \mathbb{C}^N, \tag{14}$$

where e_n is a unit vector with the n-the element equal to one and the others equal to zero.

3.1. Optimization of NMF parameters

The update rules for the NMF parameters $\{u_{n,F,k}, v_{n,k,t}\}_{n,F,k,t}$ can be derived by the majorization-minimization (MM) algorithm in the same manner as the conventional NMF. For the majorization function J^{+}_{NMF} of the cost J with respect to the NMF parameters,

$$\begin{aligned} J(\Theta) &\le J^{+}_{\mathrm{NMF}}(\Theta, \{\lambda_{n,F,t,k}, \mu_{n,F,t}\}_{n,F,t,k}) \\ &:= \frac{1}{N_T}\sum_{n,F,t,k}\lambda^{1+\beta d}_{n,F,t,k}\frac{r^{\beta}_{n,F,t}}{(u_{n,F,k}\cdot v_{n,k,t})^{\beta d}} \\ &\quad + \frac{d}{N_T}\sum_{n,F,t,k}|F|\cdot\frac{u_{n,F,k}\cdot v_{n,k,t}}{\mu_{n,F,t}} + C \end{aligned} \tag{15}$$

is obtained, where $\{\lambda_{n,F,t,k}, \mu_{n,F,t}\}_{n,F,t,k}$ are auxiliary variables, and C is a constant independent of the NMF parameters. The inequality in (15) holds if and only if

$$\lambda_{n,F,t,k} = \frac{u_{n,F,k}\cdot v_{n,k,t}}{\sum_k u_{n,F,k}\cdot v_{n,k,t}} \tag{16}$$

$$\mu_{n,F,t} = \sum_k u_{n,F,k}\cdot v_{n,k,t}. \tag{17}$$

By (15), (16), and (17), we have the following update rules:

$$u_{n,F,k} \leftarrow u_{n,F,k}\left[\frac{\beta\sum_t r^{\beta}_{n,F,t}\cdot v_{n,k,t}\cdot(\mu_{n,F,t})^{-1-\beta d}}{|F|\cdot\sum_t v_{n,k,t}\cdot(\mu_{n,F,t})^{-1}}\right]^{\frac{1}{1+\beta d}} \tag{18}$$

$$v_{n,k,t} \leftarrow v_{n,k,t}\left[\frac{\beta\sum_F r^{\beta}_{n,F,t}\cdot u_{n,F,k}\cdot(\mu_{n,F,t})^{-1-\beta d}}{\sum_F |F|\cdot u_{n,F,k}\cdot(\mu_{n,F,t})^{-1}}\right]^{\frac{1}{1+\beta d}}, \tag{19}$$

where $\{r_{n,F,t}\}_{n,F,t}$ is defined by (13).

3.2. Optimization of separation filters

We will derive the fast and stable update rules for the separation filters $\{W_f\}_f$ based on the MM algorithm in a similar manner to [11, 12, 13, 16, 17, 18]. When the shape parameter β in MEP satisfies $0<\beta\le 1$, the function $r^{\beta}_{n,F,t}$ in (12) is concave with respect to $r_{n,F,t}\in\mathbb{R}_{>0}$. Therefore, by using the tangent line inequality,[3] we can get the majorization function J^{+}_{IVA} of J with respect to the separation filters as follows:

$$\begin{aligned} J(\Theta) &\le J^{+}_{\mathrm{IVA}}(\Theta, \{\tilde{\boldsymbol{w}}_{n,f}\}_{n,f}) \\ &:= \sum_{n,f}\boldsymbol{w}^h_{n,f}R_{n,f}\boldsymbol{w}_{n,f} - 2\sum_f\log|\det W_f| + C, \end{aligned} \tag{20}$$

where $\{\tilde{\boldsymbol{w}}_{n,f}\}_{n,f}$ are the auxiliary variables, C is a constant independent of the separation filters, and we also define

$$R_{n,f} = \frac{1}{N_T}\sum_t\left[\phi_{n,F,t}\boldsymbol{x}_{f,t}\boldsymbol{x}^h_{f,t}\right],\ f\in F\in\mathcal{F} \tag{21}$$

$$\phi_{n,F,t} = \frac{\beta}{(\alpha_{n,F,t})^{\beta}\cdot\left(\sum_{f\in F}|\tilde{\boldsymbol{w}}^h_{n,f}\boldsymbol{x}_{f,t}|^2\right)^{1-\beta}}. \tag{22}$$

Here, $\alpha_{n,F,t}$ in (22) is given by (9). In (20), the inequality holds with the equality when $\boldsymbol{w}_{n,f}=\tilde{\boldsymbol{w}}_{n,f}$ for all $n\in[N]$ and $f\in[N_F]$. The minimization of (20) can iteratively be performed by a block coordinate descent method for each separation filter $\boldsymbol{w}_{n,f}$ as follows:

$$\boldsymbol{w}_{n,f} \leftarrow \left(W^h_f R_{n,f}\right)^{-1}e_n \tag{23}$$

$$\boldsymbol{w}_{n,f} \leftarrow \boldsymbol{w}_{n,f}/\sqrt{\boldsymbol{w}^h_{n,f}R_{n,f}\boldsymbol{w}_{n,f}}. \tag{24}$$

3.3. Summary of the proposed algorithm

The following is the overall procedure of the proposed algorithm:

1. Set the frequency range division $\mathcal{F}$, $\beta\in(0,1]$, and $d>0$.
2. Initialize the model parameters Θ.
3. Iterate the following steps until convergence.
   (a) Calculate $\{r_{n,F,t}\}_{n,F,t}$ by (13).
   (b) Update $\{u_{n,F,k}, v_{n,k,t}\}_{n,F,k,t}$ by (16)–(19).
   (c) Calculate $\{\alpha_{n,F,t}\}_{n,F,t}$ by (9).
   (d) Update $\{W_f\}_f$ by (21)–(24) with $\tilde{\boldsymbol{w}}_{n,f}=\boldsymbol{w}_{n,f}$ in (22).
4. Calculate the separated signals by (3) and (14).

[3] $f(r_{n,F,t}) \le f'(\tilde{r}_{n,F,t})\cdot(r_{n,F,t}-\tilde{r}_{n,F,t}) + f(\tilde{r}_{n,F,t})$

Table 1: Experimental conditions

Sampling rate	16 kHz
Frame length / Frame shift	4096 points (256 ms) / 1024 points (64 ms)
Window function	Hanning
Signal length	10 s

4. RELATION TO PRIOR WORK

If the frequency range division is given by $\mathcal{F}_{\mathrm{IVA}} = \{[N_F]\}$, then the proposed $(\mathcal{F}_{\mathrm{IVA}}, \beta, d)$-ILRMA is only the auxiliary-function-based IVA (AuxIVA [17, 18]) based on MEP with the time-varying scale parameters $\{\alpha_{n,[N_F],t}\}_{n,t}$. Specifically, $(\mathcal{F}_{\mathrm{IVA}}, 1, 1)$-ILRMA is nothing but the Gaussian AuxIVA [18]. In the same way, if $\mathcal{F}_{\mathrm{ICA}} = \{\{f\}\,;\, f \in [N_F]\}$, then $(\mathcal{F}_{\mathrm{ICA}}, \beta, d)$-ILRMA turns out to be ILRMA based on a 1-dimensional exponential power distribution. In particular, the original Gaussian ILRMA [11, 12] is realized by $(\mathcal{F}_{\mathrm{ICA}}, 1, 1)$-ILRMA. In this respect, the proposed method can be regarded as a simultaneous extension of Gaussian ILRMA and AuxIVA based on MEP.

5. EXPERIMENT

5.1. Conditions

To evaluate the performance of the proposed method, we carried out an experiment using the dataset provided by SiSEC2008 [28]. We used the liverec speech data in the dev1 dataset, with a reverberation time of 130 ms/250 ms and a microphone spacing of 5 cm/1 m. Since the dev1 task of SiSEC is an underdetermined BSS and the provided data are stereo recordings, we used only the first and the second clean spatial images to obtain the determined stereo mixture signals for each sample, and 16 mixtures in total were obtained.

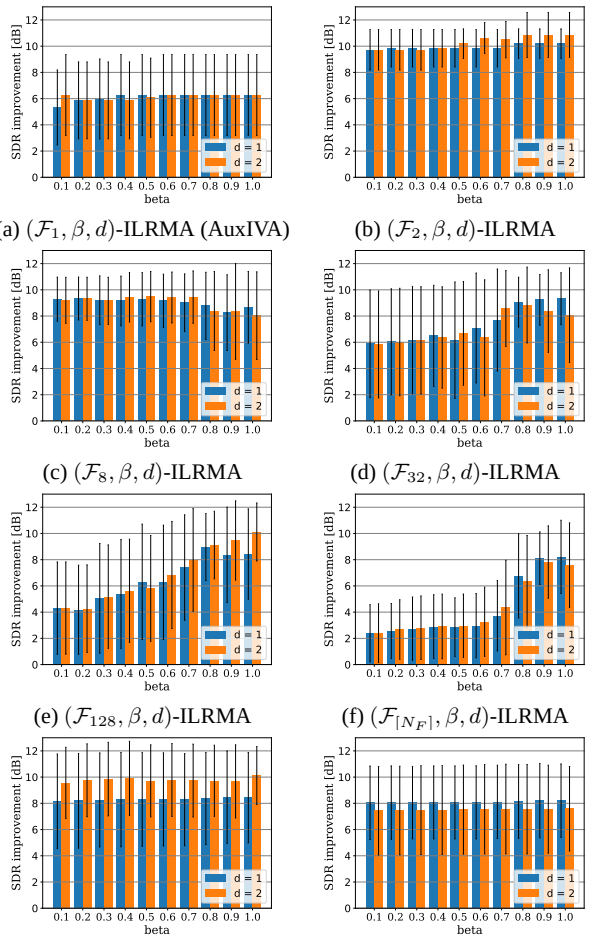
In the experiment we compared the proposed $(\mathcal{F}, \beta, d)$-ILRMA with $d \in \{1, 2\}$ and 10 varieties of β as shown in Figure 1. Also, we investigated the 6 types of the frequency range division $\mathcal{F}_k = \{F_i \subseteq [N_F] \mid i = 1, \ldots, k\}$ $(k \in \{1, 2, 8, 32, 128, [N_F]\})$, where

$$F_i = \{\lfloor \frac{N_F}{k} \cdot (i-1) \rfloor + 1, \ldots, \lfloor \frac{N_F}{k} \cdot i \rfloor\}. \tag{25}$$

Note that the $(\mathcal{F}_1, \beta, d)$-ILRMA corresponds to the AuxIVA [17, 18] based on the time-varying MEP, and the $(\mathcal{F}_{[N_F]}, 1, 1)$-ILRMA corresponds to the conventional Gaussian ILRMA [11, 12]. The number of NMF bases was set to 2 for each source, and the iteration number of the optimization (step-3 in Subsection 3.3) was set to 200 for all methods. In each method, the separation filters $\{W_f\}_f$ were initialized by the identity matrix. As for the NMF parameters, we tested the two cases: random initializations from the uniform distribution over $(0, 1)$, whose results are shown in Figure 1 (a)–(f), and initializations by the corresponding ILRMA of $\beta = 1$ shown in Figure 1 (g)–(h), meaning that $(\mathcal{F}, 1, d)$-ILRMA was performed for the first 100 iterations followed by $(\mathcal{F}, \beta, d)$-ILRMA for the second 100 iterations. The evaluation criterion is SDR [29] improvements averaged with 16 samples, which shows the overall separation quality. The other experimental conditions are described in Table 1.

5.2. Results

Figure 1 shows the average SDR improvements and their deviations. The proposed $(\mathcal{F}_2, \beta, d)$-ILRMA outperforms the conventional methods for all β and d, showing the effectiveness of the proposed approach using the frequency range division $\mathcal{F}$. Also, $(\mathcal{F}_8, \beta, d)$-ILRMA provides better results than the corresponding $(\mathcal{F}_8, 1, d)$-ILRMA for almost all shape parameters. This suggests



(a) $(\mathcal{F}_1, \beta, d)$-ILRMA (AuxIVA) (b) $(\mathcal{F}_2, \beta, d)$-ILRMA

(c) $(\mathcal{F}_8, \beta, d)$-ILRMA (d) $(\mathcal{F}_{32}, \beta, d)$-ILRMA

(e) $(\mathcal{F}_{128}, \beta, d)$-ILRMA (f) $(\mathcal{F}_{[N_F]}, \beta, d)$-ILRMA

(g) $(\mathcal{F}_{128}, \beta, d)$-ILRMA w/ initialization by $\beta = 1$ (h) $(\mathcal{F}_{[N_F]}, \beta, d)$-ILRMA w/ initialization by $\beta = 1$

Fig. 1: Average SDR improvements and their standard deviation for each $(\mathcal{F}_k, \beta, d)$-ILRMA. The horizontal line denotes the shape parameter β in MEP. In (g) and (h), the parameters Θ are initialized by using the corresponding $(\mathcal{F}_k, 1, d)$-ILRMA with $\beta = 1$.

the validity of incorporating MEP distributions in ILRMA when $|\mathcal{F}_k|$ is small. While the highest score is attained by $(\mathcal{F}_{128}, 1, 2)$-ILRMA, it is not robust to the shape parameter β. The same trend can be seen in the $(\mathcal{F}_k, \beta, d)$-ILRMA with large k as well when the parameters are randomly initialized. On the other hand, with the initializations by the corresponding $(\mathcal{F}_k, 1, d)$-ILRMA, as shown in Figure 1 (g) and (h), the proposed ILRMA with a MEP distribution turns out to be robust to β, but it does not show any SDR improvements from the corresponding $(\mathcal{F}_k, 1, d)$-ILRMA.

6. CONCLUSION

To stably improve the separation performance for speech mixtures, a generative model of ILRMA is extended by using the frequency range division to explicitly model the strong dependencies between neighboring frequency bins of speech signals. Also, MEP distributions are introduced into ILRMA to model a source distribution with a large kurtosis. The proposed ILRMA based on MEP outperforms the conventional Gaussian ILRMA and IVA based on MEP.

7. REFERENCES

[1] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1, pp. 21–34, 1998.

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

[3] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 70–79, 2007.

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

[5] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788, 1999.

[6] 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.

[7] A. Liutkus, D. Fitzgerald, and R. Badeau, "Cauchy nonnegative matrix factorization," in *Proc. WASPAA*, 2015, pp. 1–5.

[8] K. Yoshii, K. Itoyama, and M. Goto, "Student's t nonnegative matrix factorization and positive semidefinite tensor factorization for single-channel audio source separation," in *Proc. ICASSP*, 2016, pp. 51–55.

[9] P. Magron, R. Badeau, and A. Liutkus, "Lévy NMF for robust nonnegative source separation," *arXiv:1608.01844*, 2016.

[10] P. Magron, R. Badeau, and A. Liutkus, "Separation of nonnegative alpha-stable sources," *IEEE Signal Processing Letters*, 2016.

[11] D. Kitamura *et al.*, "Efficient multichannel nonnegative matrix factorization exploiting rank-1 spatial model," in *Proc. ICASSP*, 2015, pp. 276–280.

[12] D. Kitamura *et al.*, "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. 1626–1641, 2016.

[13] S. Mogami *et al.*, "Independent low-rank matrix analysis based on complex Student's t-distribution for blind audio source separation," in *Proc. MLSP*, 2017.

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

[15] H. Sawada *et al.*, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 971–982, 2013.

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

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

[18] N. Ono, "Auxiliary-function-based independent vector analysis with power of vector-norm type weighting functions," in *Proc. APSIPA*, 2012, pp. 1–4.

[19] K. Kitamura *et al.*, "Student's-t multichannel nonnegative matrix factorization for blind source separation," in *Proc. IWAENC*, 2016.

[20] S. Leglaive *et al.*, "Alpha-stable multichannel audio source separation," in *Proc. ICASSP*, 2017.

[21] G.-J. Jang, I. Lee, and T.-W. Lee, "Independent vector analysis using non-spherical joint densities for the separation of speech signals," in *Proc. ICASSP*, 2007, vol. 2, pp. II–629.

[22] I. Lee and G.-J. Jang, "Independent vector analysis based on overlapped cliques of variable width for frequency-domain blind signal separation," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 1, pp. 113, 2012.

[23] Y. Liang, S. M. Naqvi, and J. Chambers, "Overcoming block permutation problem in frequency domain blind source separation when using AuxIVA algorithm," *Electronics letters*, vol. 48, no. 8, pp. 460–462, 2012.

[24] R. Ikeshita *et al.*, "Independent vector analysis with frequency range division and prior switching," in *Proc. EUSIPCO*, 2017.

[25] E. Gómez, M. A. Gomez-Viilegas, and J. M. Marin, "A multivariate generalization of the power exponential family of distributions," *Communications in Statistics-Theory and Methods*, vol. 27, no. 3, pp. 589–600, 1998.

[26] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1, pp. 1–24, 2001.

[27] K. Matsuoka, "Minimal distortion principle for blind source separation," in *Proc. SICE*, 2002, vol. 4, pp. 2138–2143.

[28] E. Vincent, S. Araki, and P. Bofill, "The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation," in *Proc. ICA*, 2009, pp. 734–741.

[29] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.