Infinite Sparse Factor Analysis for Blind Source Separation in Reverberant Environments

Kohei Nagira, Takuma Otsuka, and Hiroshi G. Okuno

Graduate School of Informatics, Kyoto University, Kyoto, Japan
{knagira,ohtsuka,okuno}@kuis.kyoto-u.ac.jp

Abstract. Sound source separation in a real-world indoor environment is an ill-formed problem because sound source mixing is affected by the number of sounds, sound source activities, and reverberation. In addition, blind source separation (BSS) suffers from a permutation ambiguity in a frequency domain processing. Conventional methods have two problems: (1) impractical assumptions that the number of sound sources is given, and (2) permutation resolution as a post processing. This paper presents a non-parametric Bayesian BBS called permutation-free infinite sparse factor analysis (PF-ISFA) that solves the two problems simultaneously. Experimental results show that PF-ISFA outperforms conventional complex ISFA in all measures of BSS_EVAL criteria. In particular, PF-ISFA improves Signal-to-Interference Ratio by 14.45 dB and 5.46 dB under $RT_{60} = 30$ ms and $RT_{60} = 460$ ms conditions, respectively.

Keywords: Blind source separation, Reverberant mixtures, Infinite sparse factor analysis, Non-parametric Bayes.

1 Introduction

Machine listening functions, e.g. a robot audition system [1] or a distant speech recognition [2], cannot dispense with a sound source separation technique because we often observe a mixture of sound sources. For instance, HARK [1],a robot audition software, provides functions of source localization, separation, and recognition of separated speech signals. Since HARK may be deployed to various kinds of acoustic environments, parameter tuning is critical to avoid performance degradation.

In order to maximize the availability of a source separation function, the following requirements should be fulfilled for the application to practical environments:

1. source separation under an unknown mixing process dependent on the locations of sources and microphones,
2. separation under the condition of unknown number of sources,
3. robustness against the reverberation.

Many source separation methods need prior information such as the number of sources or the mixing process. Since prior information is usually difficult to obtain in advance, source separation methods should work without prior information, or at least with minimal prior information. Such a separation method is called **blind source separation** (BSS).

G.L. Gimel' farb et al. (Eds.): SSPR & SPR 2012, LNCS 7626, pp. 638–647, 2012.

For sound source separation in a practical environment, the system should separate mixtures of reverberant speeches. This is because the mixed signals captured by microphones are affected by room reverberation.

Frequency domain processing is effective to separate reverberant mixed signals, and a lot of frequency domain BSS systems are proposed. One of the problems of frequency domain processing is a permutation problem [3]. Conventional frequency domain BSS systems separates signals for all frequency bins independently, and consequently, a permutation ambiguity arises in output orders for all frequency bins. The source separation system should resolve this permutation ambiguity to reconstruct separated signals.

Independent component analysis (ICA) [4] is a well known BSS method. Frequency domain ICA [5] fulfills the first and the third requirements. However, ICA assumes the number of sources because ICA cannot detect source activities. This means that ICA does not satisfy the second requirement. In addition, Frequency domain ICA suffers from the permutation problem. Independent vector analysis (IVA) [6] and permutation free ICA [7] are BSS methods avoiding permutation problem. These methods are based on ICA and also assume the number of sources. Thus they do not satisfy the second requirement. In our previous work, frequency domain infinite sparse factor analysis (FD-ISFA) is proposed [8]. This method achieves all the three requirements, but the separation quality of FD-ISFA may be deteriorated by the subsequent permutation resolution process.

This paper presents permutation free ISFA (PF-ISFA), a BSS method which meets all the requirements and offers permutation resolution. PF-ISFA is based on nonparametric Bayesian framework, which allows BSS under the uncertainty of source numbers. The key idea of our method is that all frequency bins of signals are processed at a time by introducing a unified source activity variable for the joint optimization of the separation and permutation resolution.

2 BSS in Frequency Domain

2.1 Problem Statement of BSS

The problem of BSS is stated as below:

Input: Sound mixtures of K sources captured by D microphones.
Output: Estimated K source signals
Assumption: K is not more than D.
The locations of microphones and those of sources are fixed.

The system extracts K source signals from the mixture signal captured with D microphones without prior information of mixing process such as the location of sources, the location of microphones, and impulse responses between microphone and sound sources.

2.2 Frequency Domain Processing and Permutation Problem

In real environment with reverberation, the mixing process of speech signal is convolutive. The observed signals consist of a mixture of sources and they are contaminated

by their reverberations. To model these time-delayed signals, the convoluted mixture is often employed.

$$\overline{\mathbf{x}}(t) = \sum_{j=0}^{J} \overline{\mathbf{A}}(j)\overline{\mathbf{s}}(t-j) \tag{1}$$

where $\overline{\mathbf{x}}(t)$, $\overline{\mathbf{s}}(t)$, and $\overline{\mathbf{A}}(j)$ are observed signals, source signals, and transfer function coefficients in the time domain, respectively. BSS problem aims to retrieve that constituent sound sources $\overline{\mathbf{s}}(t)$ only given the observation $\overline{\mathbf{x}}(t)$ where the mixing process including the reverberation $\overline{\mathbf{A}}(j)$ is unknown.

When solving a BSS problem involving convoluted mixtures of signals, short time Fourier transform (STFT) is often applied in order to convert a convoluted mixture in a time domain into an instantaneous mixture in a frequency domain. In the case that signals are separated for each frequency bin independently in frequency domain, the permutation ambiguity of output order of separated signals has to be solved. This is called "permutation problem" [3]. The permutation problem is one of the well-known problems of frequency-domain BSS.

Some methods are proposed to solve this problem. One method is based on the direction of arrival estimation and the inter-frequency correlation of signal envelopes [3], and another uses power ratio of signals as dominance measure [9]. However, existing permutation resolution as a post-processing of a frequency-wise separation process, the resulting sound sources may severely be affected by the preceding separation quality. For example, if the frequency-wise separation is deteriorated by the reverberation, the permutation resolution by the signal envelopes fails, which results in the failure of BSS as well.

3 Permutation-Free ISFA

3.1 Outline of Our System

The flow of PF-ISFA is depicted in Fig. 1. After STFT, the complex spectra are whitened in each frequency bin, and PF-ISFA is applied to these whitened signals. The output order of PF-ISFA is already aligned, but the amplitude of the output signals may not equals to that of original sources. This is called scaling ambiguity, and this is another well-known problem of frequency domain BSS. The projection back method [10] is an effective solution for this problem. After projection back processing, the separated signals are reconstructed by inverse STFT.

3.2 Generative Model and Likelihood of PF-ISFA

Let K, D, F, and T be the number of sources, the number of microphones, the number of frequency bins, and the length of the source signals, respectively. The ISFA model is based on the instantaneous mixture model:

$$\mathbf{X}_f = \mathbf{A}_f(\mathbf{Z}_f \odot \mathbf{S}_f) + \mathbf{E}_f \ (f = 1, \cdots, F), \tag{2}$$

where $\mathbf{Z}_f = [\mathbf{z}_{f1}, \cdots \mathbf{z}_{fT}]$, $\mathbf{X}_f = [\mathbf{x}_{f1}, \cdots \mathbf{x}_{fT}]$, $\mathbf{S}_f = [\mathbf{s}_{f1}, \cdots \mathbf{s}_{fT}]$, $\mathbf{E}_f = [\varepsilon_{f1}, \cdots \varepsilon_{fT}]$, $\mathbf{x}_{ft} = [x_{1ft}, x_{2ft}, \cdots, x_{Dft}]^{\mathrm{T}}$ is a mixed signal vector at time t, $\mathbf{s}_{ft} = [s_{1ft}, s_{2ft}, \cdots, s_{Kft}]^{\mathrm{T}}$

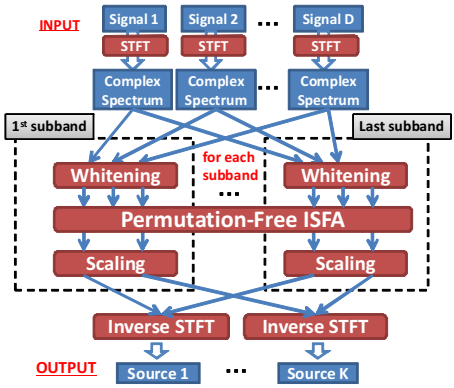

Fig. 1. Schematic overview for our method

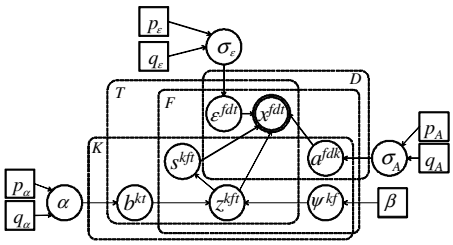

Fig. 2. Graphical model of PF-ISFA

is the source signal vector, and $\varepsilon_{ft} = [\varepsilon_{1ft}, \varepsilon_{2ft}, \cdots, \varepsilon_{Dft}]^{\mathrm{T}}$ is the Gaussian noise vector. Here, $\mathbf{A}_f$ is the $D \times K$ mixing matrix, $\mathbf{z}_{ft} = [z_{1ft}, z_{2ft}, \cdots, z_{Kft}]^{\mathrm{T}}$ is the activity of each source at time t in f-th frequency bin, and source activity z_{kft} is a binary variable: $z_{kft} = 1$ if source k is active at time t in f-th frequency bin, otherwise $z_{kft} = 0$. Operator $\odot$ indicates the element-wise product. PF-ISFA deals with F-tuple frequency bins at the same time. $\mathbf{Z}$, $\mathbf{X}$, $\mathbf{S}$. $\mathbf{E}$, and $\mathbf{A}$ are defined as $\mathbf{Z} = [\mathbf{Z}_1, \cdots \mathbf{Z}_F]$, $\mathbf{X} = [\mathbf{X}_1, \cdots \mathbf{X}_F]$, $\mathbf{S} = [\mathbf{S}_1, \cdots \mathbf{S}_F]$, $\mathbf{E} = [\mathbf{E}_1, \cdots \mathbf{E}_F]$, and $\mathbf{A} = [\mathbf{A}_1, \cdots \mathbf{A}_F]$, respectively.

To unify activities of all frequency bins, the following model is introduced.

$$z_{kft} = b_{kt}\phi, \qquad \phi \sim \mathrm{Bernoulli}(\psi_{kf}), \tag{3}$$

where $\mathrm{Bernoulli}(x)$ is the Bernoulli distribution with parameter x. b_{kt} is the unified source activity of source k at time t, and ψ_{kf} is a activation probability of source k in the f-th frequency bin. $\mathbf{B}$ represents $K \times T$ matrix of b_{kt} and Ψ means $K \times F$ matrix of ψ_{kf}.

PF-ISFA estimates the source signals $\mathbf{S}$, their time-frequency activities $\mathbf{Z}$, mixing matrix $\mathbf{A}$, unified activities $\mathbf{B}$, activation probability Ψ, and other parameters by using only the observed signal $\mathbf{X}$.

The prior distributions of the variables are assumed as follows:

$$\varepsilon_{ft} \sim \mathcal{N}_C(0, \sigma_\varepsilon^2 \mathbf{I}),\ \sigma_\varepsilon^2 \sim \mathcal{IG}(p_\varepsilon, q_\varepsilon), \tag{4}$$

$$s_{kft} \sim \mathcal{N}_C(0, 1), \tag{5}$$

$$\mathbf{a}_{kf} \sim \mathcal{N}_C(0, \sigma_{\mathbf{A}}^2 \mathbf{I}),\ \sigma_{\mathbf{A}}^2 \sim \mathcal{IG}(p_A, q_A), \tag{6}$$

$$\mathbf{B} \sim \mathrm{IBP}(\alpha),\ \alpha \sim \mathcal{G}(p_\alpha, q_\alpha),\ \text{and} \tag{7}$$

$$\Psi \sim \mathrm{Beta}\,(\beta/K, \beta(K-1)/K). \tag{8}$$

Table 1. Algorithm for estimating model parameters of Permutation-Free ISFA

Input: Observed signals $\mathbf{X}$, Output: Source signals $\mathbf{S}$.

1. Initialize parameters using their priors.
2. At each time t, carry out the following:
   2-1 In each source k, sample b_{kt} from Eq. (14).
   2-2 If $b_{kt} = 1$, sample z_{kft} from Eq. (11) and for each frequency bin f; otherwise $z_{kft} = 0$.
   2-3 If $z_{kft} = 1$, sample s_{kft} from Eq. (10); otherwise $s_{kft} = 0$.
   2-4 Determine the number of new classes κ_t, and initialize the parameters.
3. In each source k and frequency bin f, sample the probability of activation ψ_{kf} from Eq. (16).
4. In each source k and frequency bin f, sample mixing matrix $\mathbf{a}_{kf}$ from Eq. (17).
5. If there is a source that is always inactive, remove it.
6. Update σ_ε^2, $\sigma_{\mathbf{A}}^2$, and α.
7. Go to 2.

Here, $\mathbf{a}_{fk}$ is the kth row of $\mathbf{A}_f$, and p_ε, q_ε, p_A, q_A, p_α, q_α, and β are the hyperparameters. $\mathcal{N}_C$, $\mathcal{G}$, $\mathcal{IG}$ are the univariate complex normal, gamma, inverse gamma distributions, respectively. The prior for the variance of each parameter is the inverse Gamma since the inverse Gamma distribution is conjugate to the normal distribution. IBP(α) is the Indian buffet process (IBP) [11] with parameter α. IBP is a stochastic process that provides the probability distribution over sparse binary matrices with infinite number of columns. Therefore, IBP can deal with a potentially infinite number of signals.

The likelihood function of PF-ISFA is written as follows.

$$P(\mathbf{X}|\mathbf{A},\mathbf{S},\mathbf{Z}) = \prod_{f=1}^{F}\prod_{t=1}^{T} P(\mathbf{x}_{ft}|\mathbf{A}_f,\mathbf{s}_{ft},\mathbf{z}_{ft}) = \prod_{f=1}^{F}\prod_{t=1}^{T} \mathcal{N}_C(\mathbf{x}_{ft};\mathbf{A}_f(\mathbf{z}_{ft}\odot\mathbf{s}_{ft}),\sigma_\varepsilon^2\mathbf{I})$$
$$= \prod_{f=1}^{F} \frac{1}{(\pi\sigma_\varepsilon^2)^{TD}} \exp\left(-\frac{\mathrm{tr}(\mathbf{E}_f^{\mathrm{H}}\mathbf{E}_f)}{\sigma_\varepsilon^2}\right). \tag{9}$$

where $\mathbf{E}_f = \mathbf{X}_f - \mathbf{A}_f(\mathbf{Z}_f \odot \mathbf{S}_f)$. Here, each data point is assumed to be independent and identically distributed.

3.3 Source Separation through the Inference of Latent Variables

The model parameters of PF-ISFA are estimated using an iterative algorithm. The algorithm is given in Table 1, and a graphical model of PF-ISFA is shown in Fig. 2. This method is based on the Metropolis-Hastings algorithm. Posterior distributions of latent variables are derived from Bayes' theorem by multiplying priors by the likelihood function.

Sound Sources. When z_{kft} is active, s_{kft} is sampled by the following posterior.

$$P(s_{kft}|\mathbf{A}_f,\mathbf{s}_{-kft},\mathbf{x}_{ft}\mathbf{z}_{ft}) \propto P(\mathbf{x}_{ft}|\mathbf{A}_f,\mathbf{s}_{ft},\mathbf{z}_{ft},\sigma_\varepsilon^2)P(s_{kft}) = \mathcal{N}_C\left(s_{kft};\mu_{s,f},\sigma_{s,f}^2\right), \tag{10}$$

where

$$\sigma_{s,f}^2 = \sigma_\varepsilon^2/\left(\sigma_\varepsilon^2 + \mathbf{a}_{kf}^{\mathrm{H}}\mathbf{a}_{kf}\right), \quad \mu_{s,f} = \mathbf{a}_{kf}^{\mathrm{H}}\varepsilon_{-kft}/\left(\sigma_\varepsilon^2 + \mathbf{a}_{kf}^{\mathrm{H}}\mathbf{a}_{kf}\right).$$

$\mathbf{s}_{-kft}$ means $\mathbf{s}_{ft}$ except for s_{kft}, and ε_{-kft} means $\varepsilon|_{z_{kft}=0}$.

Source Activity of Each Time-Frequency Frame. If $b_{kt}=1$, z_{kft} is sampled from its posterior distribution. The posterior of z_{kft} is calculated as follows.

$$\begin{aligned} P(z_{kft}|b_{kt},\psi_{kf},z_{-kft},\mathbf{x}_{ft},\mathbf{s}_{ft},\mathbf{A}_f) &\propto P(z_{kft}|b_{kt},\psi_{kf})P(x_{ft}|\mathbf{A}_f,\mathbf{s}_{ft},\mathbf{z}_{ft},\sigma_\varepsilon^2) \\ &= \text{Bernoulli}\,(p_1/(p_0+p_1)), \end{aligned} \tag{11}$$

where

$$\begin{aligned} \log(p_1) &= \log(\psi_{kf}) + (2\,\text{Re}(s_{kft}^* \mathbf{a}_{kf}^H \varepsilon_{-kft}) + |s_{kft}|^2 \mathbf{a}_{kf}^H \mathbf{a}_{kf})/\sigma_\varepsilon^2 \\ \log(p_0) &= \log(1-\psi_{kf}) \end{aligned}$$

Unified Activity for Each Time Frame. The ratio of the probability that b_{kt} becomes active to the probability that b_{kt} becomes inactive is calculated by Eq. (12). This ratio r consists of the ratio of prior r_p and the ratio of likelihood of each frequency bin $r_{l,f}$.

$$r = \frac{P(b_{kt}=1|\mathbf{A},\mathbf{S}_{-kt},\mathbf{X}_t,\mathbf{S}_{-kt})}{P(b_{kt}=0|\mathbf{A},\mathbf{S}_{-kt},\mathbf{X}_t,\mathbf{Z}_{-kt})} = r_p \prod_{f=1}^{F} r_{l,f}. \tag{12}$$

where

$$r_p = \frac{P(b_{kt}=1|\mathbf{b}_{kt})}{P(b_{kt}=0|\mathbf{b}_{kt})} = \frac{m_{k,-t}}{T-m_{k,-t}}, \text{and}$$

$$r_{l,f} = \frac{P(\mathbf{x}_{ft}|\mathbf{A}_f,\mathbf{s}_{-kft},\mathbf{z}_{-kft},\mathbf{b}_{-kt},b_{kt}=1,\psi_{kf})}{P(\mathbf{x}_{ft}|\mathbf{A}_f,\mathbf{s}_{-kft},\mathbf{z}_{-kft},\mathbf{b}_{-kt},b_{kt}=0,\psi_{kf})} = \psi_{kf}\sigma_{s,f}^2 \exp\left(\frac{|\mu_{s,f}|^2}{\sigma_{s,f}^2}\right) + (1-\psi_{kf}). \tag{13}$$

where $m_{k,-t}=\sum_{t'\neq t} b_{kt'}$. Here, $\mathbf{X}_t$ is $\mathbf{x}_{1t},\cdots,\mathbf{x}_{Ft}$, and $\mathbf{S}_{-kt}$ and $\mathbf{Z}_{-kt}$, are $\mathbf{S}$ and $\mathbf{Z}$ except for $s_{k1t},\cdots,s_{kFt}$ and $z_{k1t},\cdots,z_{kFt}$, respectively. The ratio of prior r_p is derived from the priors of source activity based on IBP [11].

The posterior probability of $z_{kt}=1$ is calculated using ratio r.

$$P(b_{kt}=1|\mathbf{A},\mathbf{S}_{-kt},\mathbf{X}_t,\mathbf{Z}_{-kt},\mathbf{b}_{-kt}) = r/(1+r) \tag{14}$$

To decide whether or not b_{kt} is active, we sample u from Uniform(0,1) and compare it to $r/(1+r)$. If $u \leq r/(1+r)$, b_{kt} becomes active; otherwise it is not.

Number of New Sources. Some source signals that were not active at the beginning are active at time t for the first time. Let κ_t be the number of these sources.

First, the prior distribution of κ_t is $P(\kappa_t|\alpha) = \text{Poisson}\left(\frac{\alpha}{T}\right)$. After sampling κ_t, we initialize new sources and their activities. Next, we decide whether this update is acceptable or not. The acceptance probability of the transition is $\min(1, r_{\xi\to\xi^*})$. According to Meeds [12] and Knowles [13], $r_{\xi\to\xi^*}$ becomes the ratio of the likelihood of the current state to that of the next state. Let $\mathbf{A}_f^*$ be the $D\times\kappa_t$ matrix of the additional part of $\mathbf{A}_f$. The ratio can be calculated as follows.

$$r_{\xi\to\xi^*} = \prod_{f=1}^{F} (\det \Lambda_{\xi,f})^{-1} \exp\left(\mu_{\xi,f}^{\mathrm{H}} \Lambda_{\xi,f} \mu_{\xi,f}\right), \tag{15}$$

where

$$\Lambda_{\xi,f} = \mathbf{I} + \mathbf{A}_f^{*\mathrm{H}} \mathbf{A}_f^*/\sigma_\varepsilon^2, \quad \Lambda_{\xi,f}\mu_{\xi,f} = \mathbf{A}_f^{*\mathrm{H}} \varepsilon_{ft}/\sigma_\varepsilon^2.$$

Probability of Activation for Each Frequency Bin. ψ_{kf} is sampled by the following posterior.

$$P(\psi_{kf}|\mathbf{z}_{kf},\Psi_{-kf},\mathbf{B}_{-kt}) \propto P(\psi_{kf}|\beta)\prod_{t=1}^{T}P(z_{kft}|\psi_{kf},b_{kt})$$
$$= \mathrm{Beta}\left(n_{kf}+\beta/K, m_k - n_{kf} + \beta(K-1)/K\right), \tag{16}$$

where $n_{kf} = \sum_{t=1}^{T} z_{kft}$ is the number of active time-frequency frames of source k in f-th frequency bin, and $m_k = \sum_{t=1}^{T} b_{kt}$ is the number of active time frames of source k.

Mixing Matrix. The mixing matrix is estimated in each column. The posterior distribution is

$$P(\mathbf{a}_{kf}|\mathbf{A}_{f,-k},\mathbf{S}_f,\mathbf{X}_f,\mathbf{Z}_f) \propto P(\mathbf{X}_f|\mathbf{A}_f,\mathbf{S}_f,\mathbf{Z}_f,\sigma_\varepsilon^2)P(\mathbf{a}_{kf}|\sigma_{\mathbf{A}}^2)$$
$$= \mathcal{N}_C(\mathbf{a}_{kf};\mu_{\mathbf{A}},\Lambda_{\mathbf{A}}^{-1}), \tag{17}$$

where

$$\Lambda_{\mathbf{A}} = \left(\frac{\mathbf{s}_{kf}^{\mathrm{H}}\mathbf{s}_{kf}}{\sigma_\varepsilon^2} + \frac{1}{\sigma_{\mathbf{A}}^2}\right)\mathbf{I}_D, \quad \mu_{\mathbf{A}} = \frac{\sigma_{\mathbf{A}}^2}{\mathbf{s}_{kf}^{\mathrm{H}}\mathbf{s}_{kf}\sigma_{\mathbf{A}}^2 + \sigma_\varepsilon^2}\mathbf{E}_f|_{\mathbf{a}_{kf}=0}\mathbf{s}_{kf}.$$

Variance of Noise and Mixing Matrix. The variance of noise corresponds to the noise level of the estimated signals, and the variance of the mixing matrix affects the scale of the estimated signals. Their posteriors are as follows.

$$P(\sigma_\varepsilon^2|\mathbf{E}) \propto P(\mathbf{E}|\sigma_\varepsilon^2)P(\sigma_\varepsilon^2|p_\varepsilon,q_\varepsilon) = \mathcal{IG}\left(\sigma_\varepsilon^2; p_\varepsilon + FTD, \frac{q_\varepsilon}{(1+q_\varepsilon\sum_{f=1}^{F}\mathrm{tr}(\mathbf{E}_f^{\mathrm{H}}\mathbf{E}_f))}\right). \tag{18}$$

$$P(\sigma_{\mathbf{A}}^2|\mathbf{A}) \propto P(\mathbf{A}|\sigma_{\mathbf{A}}^2)P(\sigma_{\mathbf{A}}^2|p_{\mathbf{A}},q_{\mathbf{A}}) = \mathcal{IG}\left(\sigma_{\mathbf{A}}^2; p_{\mathbf{A}} + FDK, \frac{q_{\mathbf{A}}}{1+q_{\mathbf{A}}\sum_{f=1}^{F}\mathrm{tr}(\mathbf{A}_f^{\mathrm{H}}\mathbf{A}_f)}\right). \tag{19}$$

Parameter of IBP. Since the IBP parameter α can be updated in the same way as FD-ISFA [8], the detailed explanation is omitted here.

4 Experimental Results

We test our method in a separation experiment using speech signals in order to evaluate the separation performance of our method. In this experiment, our method is compared with the baseline method, complex ISFA [8]. We use two kinds of mixed signals for this experiment: convoluted mixture with impulse responses measured in anechoic chamber, and convoluted mixture with impulse responses measured in meeting room (RT_{60} = 460 ms). Figure 4 shows the locations of the microphones and sources, and Table 2

Table 2. Experimental conditions

Number of sources K	2
Number of microphones D	2
Test set	ASJ-JNAS
Sampling rate	16000 Hz
Window length	64 ms
Shift length	32 ms
Iterations	300

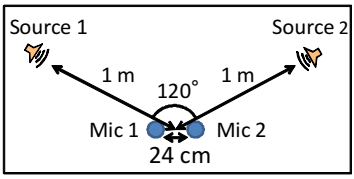

Fig. 3. Locations of microphones and sources in experiment

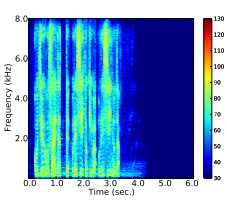

Fig. 4. Spectrogram of original source

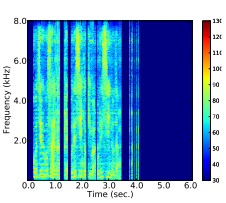

Fig. 5. Spectrogram of PF-ISFA result

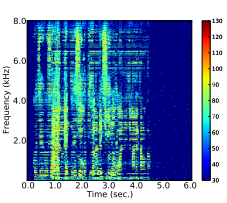

Fig. 6. Spectrogram of FD-ISFA result

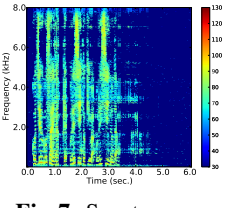

Fig. 7. Spectrogram of permutation-aligned FD-ISFA result

lists the conditions for this experiment. We used 200 utterances from JNAS phoneme balanced sentences on each condition.

First, an example of experimental results of separation experiment using mixed signals in the anechoic chamber is shown. Figures 4–7 show the spectrograms of a source signal, a signal separated with PF-ISFA, a signal separated with conventional FD-ISFA, and a permutation-aligned signal separated with FD-ISFA. In the results of FD-ISFA, many horizontal lines are seen in Figure 6, but in Figure 7, the number of these lines decrease. These lines are the spectrogram of the other separated signal. This means that the output orders of FD-ISFA result are not aligned for all frequency bins. In contrast, there is no horizontal line in the spectrogram of PF-ISFA (Figure 5). This shows that the output order is aligned, in other words the permutation problem is solved by PF-ISFA.

We also evaluate our method in terms of the Signal to Distortion Ratio (SDR), the Image to Spatial distortion Ratio (ISR), the Source to Interference Ratio (SIR), and the Source to Artifacts Ratio (SAR) [14]. Table 3 summarizes the results. “Non-Perm” is calculated by output signals themselves, in other words, their permutations are not aligned. “Perm” means that output signals are aligned their permutations using the correlation between outputs and original sources. In other word, permutation is aligned by using original source signals as reference. Our proposed method outperforms FD-ISFA by all criteria in Non-Perm case. Especially, proposed method improves SIR by 14.45 dB in anechoic chamber reverberations and 5.46 dB in meeting room reverberations.

One of the reasons of poor performance of FD-ISFA is caused by the permutation problem, because the difference between the performance of permutation-aligned results of FD-ISFA and that of FD-ISFA results without aligning permutations is large. In contrast, that of PF-ISFA results is smaller. This means that the permutations of outputs are automatically aligned when PF-ISFA is applied.

Table 3. Average separation performance [dB]

	Anechoic chamber				Meeting room ($RT_{60} = 460$ ms)			
	FD-ISFA		PF-ISFA		FD-ISFA		PF-ISFA	
	Non-Perm	Perm	Non-Perm	Perm	Non-Perm	Perm	Non-Perm	Perm
SDR	0.38	11.96	**10.26**	**12.59**	0.35	5.85	**3.56**	**5.31**
ISR	4.98	18.23	**15.96**	**18.75**	4.73	10.41	**8.08**	**9.88**
SIR	1.38	18.58	**15.83**	**19.20**	1.12	9.86	**6.58**	**9.22**
SAR	5.22	14.39	**13.91**	**15.16**	5.72	10.36	**9.30**	**10.36**

This results show that the performance in meeting room reverberation is worse than that in anechoic chamber reverberation. This is because the reverberation time of meeting room ($RT_{60} = 460$ ms) is longer than STFT window length (64 ms). If the reverberation time is longer than STFT window length, reverberation affects multiple time frames, and this degrades the performance.

5 Conclusion and Future Work

This paper presented PF-ISFA based on a non-parametric Bayesian framework for reverberant environments. PF-ISFA achieves BSS without the assumptions about observations such as the number of sources, reverberation, and the mixing process. This method is processed in frequency domain to separate reverberant speeches without prior information, and it can avoid permutation problem. Experimental results show that PF-ISFA outperforms conventional FD-ISFA.

Future work includes the following. We focus on the source activity accuracy, and achieve voice activity detection using the source activity estimated by PF-ISFA for an effective speech recognition system. In addition, the time complexity of PF-ISFA should be reduced for an accelerated separation system. If we attain a real-time processing, PF-ISFA can be applied to many applications including robot audition.

Acknowledgement. This study was partially supported by the Grant-in-Aid for Scientific Research (S), and Honda Research Institute Japan Inc., Ltd.

References

1. Nakadai, K., et al.: Design and Implementation of Robot Audition System "HARK" Open Source Software for Listening to Three Simultaneous Speakers. Advanced Robotics 24(5-6), 739–761 (2010)
2. Wölfel, M., et al.: Distant Speech Recognition. Wiley (2009)
3. Sawada, H., et al.: A robust and precise method for solving the permutation problem of frequency-domain blind source separation. IEEE Trans. on Speech and Audio Processing 12(5), 530–538 (2004)
4. Hyvärinen, A., et al.: Independent component analysis. Wiley Interscience (2001)
5. Sawada, H., et al.: Polar coordinate based nonlinear function for frequency-domain blind source separation. In: Proc. of IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2002), pp. 1001–1004 (2002)

6. Lee, I., et al.: Fast fixed-point independent vector analysis algorithms for convolutive blind source separation. Signal Processing 87(8), 1859–1871 (2007)
7. Hiroe, A.: Solution of Permutation Problem in Frequency Domain ICA, Using Multivariate Probability Density Functions. In: Rosca, J.P., Erdogmus, D., Príncipe, J.C., Haykin, S. (eds.) ICA 2006. LNCS, vol. 3889, pp. 601–608. Springer, Heidelberg (2006)
8. Nagira, K., Takahashi, T., Ogata, T., Okuno, H.G.: Complex Extension of Infinite Sparse Factor Analysis for Blind Speech Separation. In: Theis, F., Cichocki, A., Yeredor, A., Zibulevsky, M. (eds.) LVA/ICA 2012. LNCS, vol. 7191, pp. 388–396. Springer, Heidelberg (2012)
9. Sawada, H., et al.: Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS. In: IEEE Intl. Symposium on Circuits and Systems, ISCAS 2007, pp. 3247–3250. IEEE (2007)
10. Murata, N., et al.: An approach to blind source separation based on temporal structure of speech signals. Neurocomputing 41(1-4), 1–24 (2001)
11. Griffiths, T., et al.: Infinite latent feature models and the Indian buffet process. Advances in Neural Information Processing Systems 18, 475–482 (2006)
12. Meeds, E., et al.: Modeling dyadic data with binary latent factors. Advances in Neural Information Processing Systems 19, 977–984 (2007)
13. Knowles, D., Ghahramani, Z.: Infinite Sparse Factor Analysis and Infinite Independent Components Analysis. In: Davies, M.E., James, C.J., Abdallah, S.A., Plumbley, M.D. (eds.) ICA 2007. LNCS, vol. 4666, pp. 381–388. Springer, Heidelberg (2007)
14. Vincent, E., Sawada, H., Bofill, P., Makino, S., Rosca, J.P.: First Stereo Audio Source Separation Evaluation Campaign: Data, Algorithms and Results. In: Davies, M.E., James, C.J., Abdallah, S.A., Plumbley, M.D. (eds.) ICA 2007. LNCS, vol. 4666, pp. 552–559. Springer, Heidelberg (2007)