EURASIP Journal on
Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Nonparametric Bayesian sparse factor analysis for frequency domain blind source separation without permutation ambiguity

Kohei Nagira*, Takuma Otsuka and Hiroshi G Okuno

Abstract

Blind source separation (BSS) and sound activity detection (SAD) from a sound source mixture with minimum prior information are two major requirements for computational auditory scene analysis that recognizes auditory events in many environments. In daily environments, BSS suffers from many problems such as reverberation, a permutation problem in frequency-domain processing, and uncertainty about the number of sources in the observed mixture. While many conventional BSS methods resort to a cascaded combination of subprocesses, e.g., frequency-wise separation and permutation resolution, to overcome these problems, their outcomes may be affected by the worst subprocess. Our aim is to develop a unified framework to cope with these problems. Our method, called permutation-free infinite sparse factor analysis (PF-ISFA), is based on a nonparametric Bayesian framework that enables inference without a pre-determined number of sources. It solves BSS, SAD and the permutation problem at the same time. Our method has two key ideas: unified source activities for all the frequency bins and the activation probabilities of all the frequency bins of all the sources. Experiments were carried out to evaluate the separation performance and the SAD performance under four reverberant conditions. For separation performance in the BSS_EVAL criteria, our method outperformed conventional complex ISFA under all conditions. For SAD performance, our method outperformed the conventional method by 5.9–0.5% in *F*-measure under the condition $RT_{20} = 30$–600 [ms], respectively.

1 Introduction

Computational auditory scene analysis (CASA) aims to find auditory events and extract valuable information from captured sound signals [1,2]. An overview of CASA system is depicted in Figure 1. First, the CASA system captures sound signals by using a microphone array. Then, it detects sound activities of each source and separates the mixture into individual sources. Finally, it visualizes the auditory events or recognizes these separated sound sources. This article focuses on the source activity detection (SAD) and sound source separation. SAD is useful for CASA systems because this function helps these systems discover audio sources especially when a huge amount of archived audio signals is analyzed. Another example of the benefit of the SAD is compatibility with automatic speech recognition. For accurate automatic speech recognition, it is necessary to extract the voiced part, which is referred to as voice activity detection [3,4]. Sound source separation is essential for CASA systems because we often observe a mixture of multiple sound sources in our daily environment. Our goal is to develop a simultaneous sound activity detection and sound source separation system for CASA.

The combination of sound source separation and source activity detection should overcome the following difficulties for real-world applications:

1. unknown mixing processes,
2. source number uncertainty,
3. reverberation, and
4. performance degradation caused by mutually dependent functions.

The first one indicates that the CASA system should work without information specific to a certain environment or a situation such as the environment's impulse responses or the sound source locations. The second one expresses that the CASA system should achieve robust estimation under the condition that the number of sources is unknown. The

*Correspondence: knagira@kuis.kyoto-u.ac.jp
Graduate School of Informatics, Kyoto University, Sakyo, Kyoto, Japan

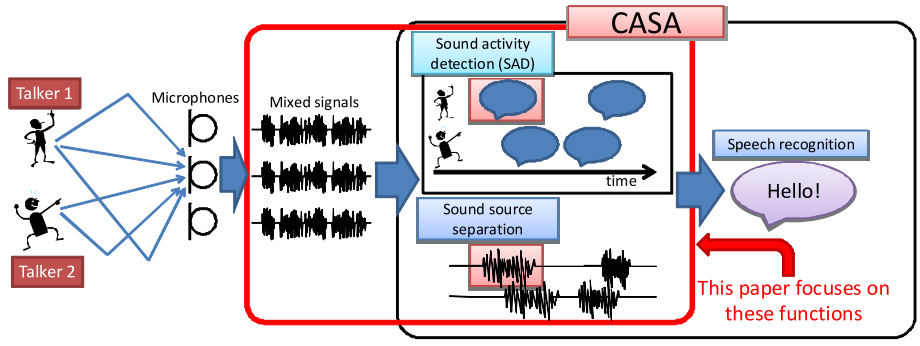

Figure 1 Overview of CASA system and location of our method.

third one means that the mixture of audio signals captured in a room may contain reverberations that affect the microphone array processing. The last one means that cascaded processing to cope with the above-mentioned difficulties may be severely affected by the worst subprocess of the CASA system. When source separation processing is performed in the frequency domain, the output signals are affected by the permutation problem, which is ambiguity in the output order for different frequency bins. Conventional methods take the cascade approach. The mixed signals are separated for each frequency bin first, and then the permutation problem is solved. As mentioned above, the overall performance is limited to the performance of the worst subprocess.

Our solution for overcoming these difficulties is as follows. The mixing process is modeled stochastically and inferred on the basis of this model. To handle source number uncertainty, we introduce a nonparametric Bayesian approach. The reverberation is absorbed by using frequency-domain processing. Unified analysis of the source separation and permutation resolution is used to optimize these mutually dependent functions.

This article presents a permutation-free infinite sparse factor analysis (PF-ISFA): a joint estimation method that simultaneously achieves frequency-domain source separation and SAD using a minimum amount of prior information. PF-ISFA achieves robust estimation without using prior information about the number of sources. PF-ISFA extends the frequency-domain ISFA [5], which is a nonparametric Bayesian frequency-domain source separation method. We build a generative process that explains the observed sound source mixture and derive a Bayesian inference to retrieve respective sound sources and sound activities. The key idea of PF-ISFA is that all the frequency bins of signals are processed at the same time to avoid the permutation problem. In particular, a unified source activity for all frequency bins is introduced into its generative model.

The rest of this article is organized as follows. Section 2 summarizes the main problem treated, and introduces study related to our method. Section 3 explains conventional ISFA in the time and frequency domains and then introduces our new method PF-ISFA. Section 4 gives detailed posterior inferences of PF-ISFA. Section 5 presents experimental results, and Section 6 concludes this article.

2 Problem statement and related study

This section starts by summarizing the problem that is solved in this article and the assumptions needed to solve it. After that, the study related to this problem, especially concerning source separation, the permutation problem, and sound detection methods, are introduced.

2.1 Problem statement

The problem statement is briefly summarized below.

Input:

- Sound mixtures of K sources captured by D microphones.

Output:

- Estimated K source signals,
- Detected source activities of source signals.

Assumptions:

1. The number of sources K is not more than the number of microphones D.
2. The locations of the sources do not change.

The sound activity represents whether or not sound is active in each time frame. This sound activity estimation enables sound detection. The system estimates the source activities of K source signals and separates the D mixed signals captured by the microphones into K sources without prior information, such as locations, microphone locations, and impulse responses between sound sources and microphones. The first assumption means that this system deals with a determined or overdetermined problem. The second assumption means that

the mixing process from the sources to the microphones is unchanged.

2.2 Requirements

This system should fulfill some requirements in order to work in daily environments. These requirements are summarized as follows.

1. Blind source separation,
2. Frequency domain processing,
3. Permutation resolution,
4. Robust estimation without source knowledge, and
5. Unified approach.

These requirements are described in detail below.

2.2.1 Blind source separation

One of the system's major requirements is to work with the minimum amount of prior information. This is because getting prior information, such as the direction of arrival of sound or the reverberation level of the room, in advance is a troublesome task for the system. In addition, even if the prior information can be obtained, the separation performance is severely affected by the quality of the information. The system should not be dependent on such information. The source separation method that uses the minimum prior information is called blind source separation (BSS).

2.2.2 Frequency domain processing

There are two reasons why frequency domain processing is inevitable for CASA. One is to deal with reverberation and the other is to model source signals using the sparseness of sound energy.

The mixing process of speech signals in our daily surroundings is modeled as a convoluted mixture [6]. The signals captured by the microphones consist of a mixture of ones from various sources and they are contaminated by reflections, reverberations, and arrival time lags at the microphones. To model these time-delayed signals, a convoluted mixture is often used.

Attempts to solve a BSS problem involving convoluted mixtures of signals mainly use, frequency domain processing. This is because the convoluted mixture in the time domain can be explained in a simplistic form in the frequency domain. Specifically, the short time Fourier transform (STFT) can convert a convoluted mixture in the time domain into instantaneous mixtures for all frequency bins. In other words, STFT can absorb the reverberation of the source signals within the window length. Thus, frequency domain processing is effective when BSS is applied to audio signals in practical situations.

2.2.3 Permutation resolution

As mentioned above, the convoluted mixture in the time domain is converted into instantaneous mixtures for individual frequency bins. Many frequency-domain BSS methods independently separate the mixed signals for all the frequency bins; thus, an ambiguity arises in the output order. The system must arrange the separated signals in the correct order for the frequency bins. This is called the "permutation problem". The permutation problem should be solved in order to achieve frequency-domain BSS.

2.2.4 Robust estimation without source knowledge

Many CASA systems and many source separation methods use prior knowledge about source signals for robust estimation to improve the performance. For instance, HARK [7] localizes the sound sources before separation by using the number of sources. When independent component analysis (ICA), a well-known BSS method, is applied to the input signals, principal component analysis (PCA) is commonly used as preprocessing for ICA [8]. This is because the number of dimensions of ICA's input signals can be reduced. However, getting prior knowledge about sources is difficult for the system, so robust estimation without source knowledge is desirable. A nonparametric Bayesian framework is helpful for robust inference without knowing the number of sources.

2.2.5 Unified approach

A unified estimation method enables effective processing because it makes the most of the information available from the observed signals. Many source separation frameworks use a cascaded approach. For instance, HARK [7] localizes the sources first and then separates the observed signals into individual sources; the conventional frequency-domain ICA separates the observations and then resolves the permutation problem. One of the critical weak points of these cascaded approaches is that the separation performance is limited to the performance of the worst subprocess.

2.3 Related study

2.3.1 Source separation method of speech signals

Source separation is being actively studied for signal processing. Some methods use the source and microphone locations. Delay-and-sum beamforming and null beamforming are methods that emphasize or suppress the signal from a specific direction. These methods can be implemented with less computational complexity. HARK uses geometric higher-order decorrelation-based source separation (GHDSS) [9]. GHDSS separates mixed signals by using a higher-order decorrelation between the sound source signals and geometric constraints derived from the positional relationships among the microphones. The weak point of these methods is that they require the source and microphone locations. This prior information cannot easily be obtained in advance.

Many BSS methods have already been introduced. One well-known BSS method is ICA, which separates mixed signals on the basis of the statistical independence between of different source signals. Many algorithms are used for ICA, such as the minimization mutual information [10], Fast ICA [11], and JADE [12]. For BSS for speech signals, frequency-domain ICA is commonly used [13]. While ICA does achieve BSS, it does not detect the activities of individual sources; moreover, frequency-domain ICA is plagued by the permutation problem.

ISFA [14] is a BSS method based on the nonparametric Bayesian approach. It achieves SAD and BSS simultaneously, but it is modeled in the time domain, so it is vulnerable to the reverberation that often appears in our daily surroundings.

Frequency-domain ISFA (FD-ISFA), which we proposed in our previous study [5], can handle a convoluted mixture that contains room reverberation. One problem for FD-ISFA is the permutation problem. Conventional FD-ISFA independently separates the signals for all the frequency bins, so it cannot avoid permutation ambiguity.

2.3.2 Permutation problem

Some methods solve the permutation problem by post processing. One method is based on estimation of the direction of arrival and inter-frequency correlation of the signal envelopes [15]; another uses the power ratio of the signals as a dominance measure [16].

Other methods avoid this problem by using a unified criterion from among all frequency bins. Independent vector analysis (IVA) [17] and permutation-free ICA [18] are BSS methods that avoid the permutation problem. These methods are based on ICA and cannot simultaneously achieve sound source detection.

2.3.3 BSS framework achieving SAD

Some BSS frameworks obtain SAD information simultaneously. Switching ICA [19] is a BSS method which can achieve SAD. Switching ICA employs a hidden Markov model (HMM) on its model to represent whether the source is active or not. The SAD information is obtained from these estimated hidden variables of HMM. Nonstationary Bayesian ICA [20] achieves dynamic source separation by estimating the sources and the mixing matrices for each time frame on the basis of variational Bayesian inference. The SAD information is obtained from automatic relevance determination (ARD) parameters, which are the precision parameters of the probabilistic density of the mixing matrix. Since these methods are time-domain approaches, it is not appropriate for speech separation of convoluted mixtures.

The combination of a maximize signal-to-noise ratio beamformer, a voice activity detector and online clustering achieves BSS and SAD [21]. This method is a cascade approach. It achieves SAD and the time-difference of arrival estimation first and then separates signals using this them. As mentioned above, the weak point of cascaded approach is that the separation performance is limited to the performance of the worst subprocess.

3 ISFA

This section first summarizes conventional methods for ISFA: Section 3.1 shows the model of ISFA in the time domain [14], and Section 3.2 explains its expansion into the frequency domain (FD-ISFA) [5] and its problems. Then, Section 3.3 describes a model of FD-ISFA without permutation ambiguity (PF-ISFA).

3.1 ISFA in time domain

ISFA [14] achieves BSS of instantaneous mixtures of time-domain signals without knowing the number of sources. It is based on the following instantaneous mixture model, which expresses that $D \times T$ observed data $\mathbf{X}$ is composed of a linear combination of $K \times T$ source signals $\mathbf{S}$.

$$\mathbf{X} = \mathbf{A}(\mathbf{Z} \odot \mathbf{S}) + \mathbf{E}, \tag{1}$$

where $\mathbf{A}$ is a $D \times K$ mixing matrix, $\mathbf{E}$ is a $D \times T$ Gaussian noise term, and $\mathbf{Z}$ is a binary mask on $\mathbf{X}$. $\odot$ denotes element-wise multiplication. Let $x_{\mathrm{dt}}, a_{\mathrm{dk}}, z_{\mathrm{kt}}, s_{\mathrm{kt}}, \varepsilon_{\mathrm{dt}}$ be the elements of $\mathbf{X}$, $\mathbf{A}$, $\mathbf{Z}$, $\mathbf{S}$, and $\mathbf{E}$, respectively. The generative model of ISFA is shown in Figure 2. σ_A^2 and σ_ε^2 are the variance parameters of the elements of $\mathbf{A}$ and $\mathbf{E}$.

The priors of these parameters are as follows:

$$\boldsymbol{\varepsilon}_t \sim \mathcal{N}(0, \sigma_{\boldsymbol{\varepsilon}}^2 \mathbf{I}),\ \sigma_{\boldsymbol{\varepsilon}}^2 \sim \mathcal{IG}(p_{\boldsymbol{\varepsilon}}, q_{\boldsymbol{\varepsilon}}), \tag{2}$$

$$s_{\mathrm{kt}} \sim \mathcal{N}(0, 1), \tag{3}$$

$$\mathbf{a}_k \sim \mathcal{N}(0, \sigma_{\mathbf{A}}^2 \mathbf{I}),\ \sigma_{\mathbf{A}}^2 \sim \mathcal{IG}(p_A, q_A),\ \text{and} \tag{4}$$

$$\mathbf{Z} \sim \mathrm{IBP}(\alpha),\ \alpha \sim \mathcal{G}(p_\alpha, q_\alpha). \tag{5}$$

Here, $\mathbf{a}_k$ is the kth row of $\mathbf{A}$, and $p_{\boldsymbol{\varepsilon}}, q_{\boldsymbol{\varepsilon}}, p_A, q_A, p_\alpha$, and q_α are the hyperparameters. IBP(α) is the Indian buffet process (IBP) [22] with concentration parameter α. IBP [22] is a stochastic process that can deal with a potentially infinite number of signals. It is used in order to achieve separation without using prior knowledge about the number of sources.

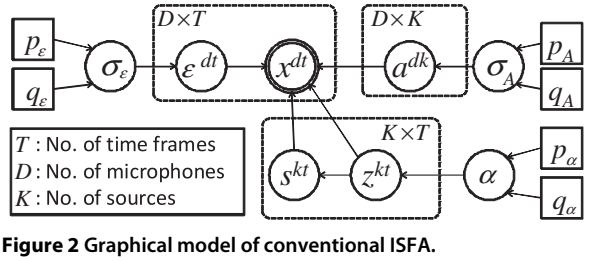

Figure 2 Graphical model of conventional ISFA.

In the time domain, each element of **X**, **A**, **S**, and **E** is a real-valued variable. Each of these variables has a normal distribution as a prior. $\mathcal{N}(\mu, \sigma^2)$ is a normal distribution with mean μ and variance σ^2. The probability density function of this normal distribution is

$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \quad (6)$$

The IBP concentration parameter has a gamma prior, and the variance parameters of A and E have inverse gamma priors. $\mathcal{G}(b, \theta)$ and $\mathcal{IG}(b, \theta)$ are gamma distribution and the inverse gamma distribution with shape parameter b and scale parameter θ, respectively. The probability density functions of these distributions are

$$\mathcal{G}(x; b, \theta) = \frac{x^{b-1}}{\Gamma(b)\,\theta^b} \exp\left(-\frac{x}{\theta}\right), \text{ and} \quad (7)$$

$$\mathcal{IG}(x; b, \theta) = \frac{x^{-(b+1)}}{\Gamma(b)\,\theta^b} \exp\left(-\frac{1}{\theta x}\right). \quad (8)$$

A Bayesian hierarchical model aims at explaining the uncertainty in the model from the observed data by treating latent variables as a probabilistic variable rather than a fixed value. In our model, we place a gamma prior on the concentration parameter of IBP so that the emergence of sources in **Z** can be controlled by the data we have.

3.2 ISFA in frequency domain

Since the convoluted mixture is converted into complex spectra by using STFT, the elements of **X**, **S**, **A**, and **E** become complex-valued variables. FD-ISFA is a model for complex values that arises in frequency-domain processing. It can deal with an instantaneous mixture of complex spectra.

The generative model is the same as for time-domain ISFA. However, the priors of these complex-valued elements are different from those of time-domain ISFA.

$$\boldsymbol{\varepsilon}_t \sim \mathcal{N}_C(0, \sigma_{\boldsymbol{\varepsilon}}^2 \mathbf{I}), \quad (9)$$

$$s_{\mathrm{kt}} \sim \mathcal{N}_C(0, 1), \text{ and} \quad (10)$$

$$\mathbf{a}_k \sim \mathcal{N}_C(0, \sigma_{\mathbf{A}}^2 \mathbf{I}) \quad (11)$$

Here, instead of the normal distribution, a univariate complex normal distribution $\mathcal{N}_C$ is used for complex-valued parameters. The probability density functions of this distribution is

$$\mathcal{N}_C(x; \mu, \sigma^2) = \frac{1}{\pi\sigma^2} \exp\left(-\frac{|x-\mu|^2}{\sigma^2}\right). \quad (12)$$

Conjugacy is one of the helpful properties of Bayesian inference. If we choose a conjugate prior, a closed-form expression can be given for the posterior. The variances $\sigma_{\boldsymbol{\varepsilon}}^2$ and $\sigma_{\mathbf{A}}^2$ have a conjugate inverse gamma prior, and the Gaussian conjugate prior can be used for the mixing matrix **A**. For simplicity, the univariate complex normal distribution is introduced as a conjugate prior of source signal **S**. It is noted that a super-Gaussian prior, such as student-t or Laplace distribution, should be used for speech signals. The complex extension of these distributions is non-trivial. We don't deal with the complex super-Gaussian prior in this article and this is one of our future study.

The processing flow of FD-ISFA is as follows. After STFT, the complex spectra are whitened in each frequency bin, and FD-ISFA is applied for each frequency bin of these complex spectra independently. FD-ISFA is plagued by two well-known ambiguities of frequency domain BSS: the scaling ambiguity and permutation ambiguity. The scaling ambiguity is that the amplitude of the output signals may not equal that of the original sources. Some post-processing methods are needed to resolve these two ambiguities. The projection back method [23] is an effective solution for the scaling ambiguity. The permutation ambiguity is solved by using the methods mentioned above [15,16]. After these problems have been solved, estimated complex spectra are assembled into source signals by using inverse STFT.

3.3 New method: PF-ISFA

Our new method, permutation-free ISFA (PF-ISFA), achieves both BSS and SAD without being affected by the permutation problem. Its key idea for avoiding the permutation problem is unified activity for each time frame. Conventional ISFA is applied independently to each frequency bin. That is to say, conventional ISFA does not consider any relations across frequency bins. This is the main reason for the permutation problem. By contrast, in the PF-ISFA model, all frequency bins are unified by the activity matrix. Since this unified activity controls the output order of source signals, PF-ISFA is not affected by the permutation problem.

The flow of PF-ISFA is depicted in Figure 3, and the generative process of PF-ISFA is described in Figure 4. Let F be the number of frequency bins. PF-ISFA is also based on instantaneous mixture for each frequency bin. PF-ISFA deals with the F-tuple frequency bins at the same time. The elements of **Z**, **X**, **S**, **E**, and **A** are defined as x_{fdt}, a_{fdk}, z_{fkt}, s_{fkt}, $\varepsilon_{\mathrm{fdt}}$, respectively.

The following model is introduced to unify the activities of all frequency bins.

$$z_{\mathrm{fkt}} = b_{\mathrm{kt}}\phi, \qquad \phi \sim \mathrm{Bernoulli}(\psi_{\mathrm{kf}}), \quad (13)$$

where Bernoulli(x) is the Bernoulli distribution with parameter x. b_{kt} is the unified source activity of source k at time t, and $\boldsymbol{\Psi}$ is the probability of the source k becoming active (activation probability) in the fth frequency bin. **B** represents the $K \times T$ matrix of b_{kt} and $\boldsymbol{\Psi}$ means the $K \times F$ matrix of ψ_{kf}. Let β be the hyperparameter.

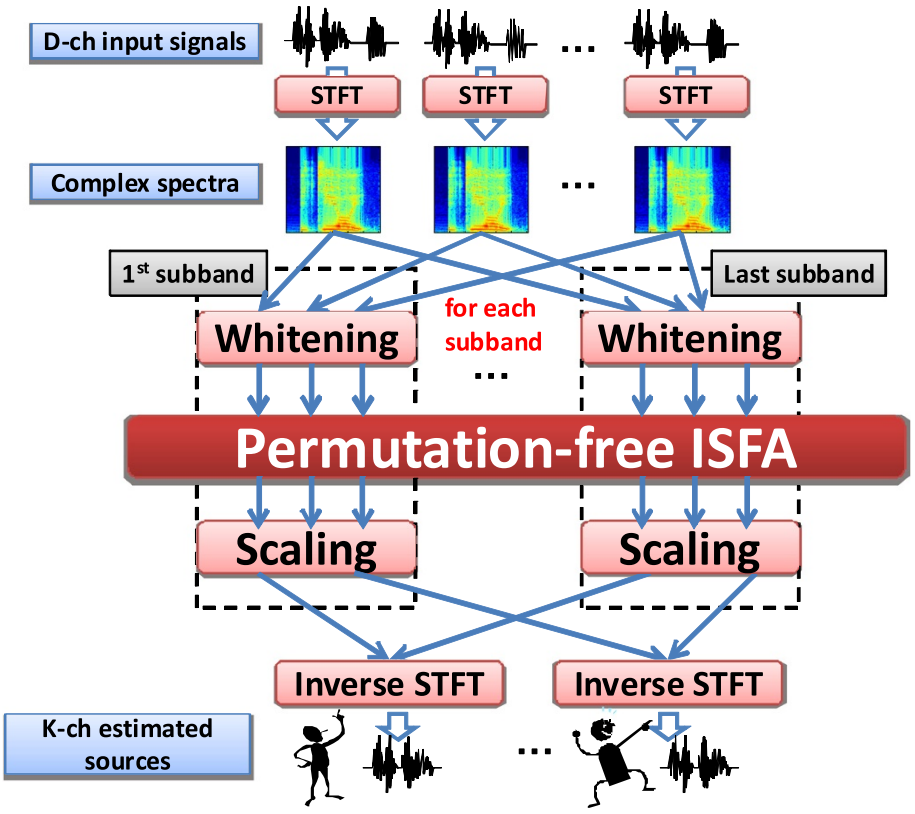

Figure 3 Schematic overview for our method.

The prior distributions of the newly introduced variables are assumed to be as follows:

$$\mathbf{B} \sim \mathrm{IBP}(\alpha),\ \alpha \sim \mathcal{G}(p_\alpha, q_\alpha),\ \text{and} \tag{14}$$

$$\mathbf{\Psi} \sim \mathrm{Beta}\left(\frac{\beta}{K}, \frac{\beta(K-1)}{K}\right). \tag{15}$$

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

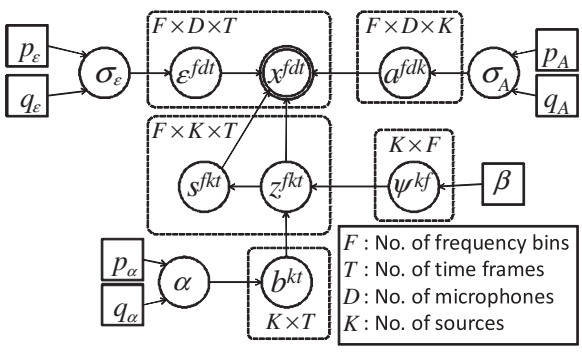

Figure 4 Graphical model of PF-ISFA.

One of the main differences between this PF-ISFA model and conventional ISFA model is the unified activity matrix for each time frame $\mathbf{B}$ and the activation probability matrix for each frequency bin $\mathbf{\Psi}$. A graphical model of conventional ISFA is shown in Figure 2. Whereas each frequency bin is independently estimated in the conventional ISFA model, all frequency bins are bundled together by the unified activity matrix in the PF-ISFA model.

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

$$\begin{aligned} P(\mathbf{X}|\mathbf{A},\mathbf{S},\mathbf{Z}) &= \prod_{f=1}^{F}\prod_{t=1}^{T} P(\mathbf{x}_{\mathrm{ft}}|\mathbf{A}_f, \mathbf{s}_{\mathrm{ft}}, \mathbf{z}_{\mathrm{ft}}) \\ &= \prod_{f=1}^{F}\prod_{t=1}^{T} \mathcal{N}_C(\mathbf{x}_{\mathrm{ft}}; \mathbf{A}_f(\mathbf{z}_{\mathrm{ft}} \odot \mathbf{s}_{\mathrm{ft}}), \sigma_{\boldsymbol{\varepsilon}}^2 \mathbf{I}) \\ &= \prod_{f=1}^{F} \frac{1}{(\pi\sigma_{\boldsymbol{\varepsilon}}^2)^{TD}} \exp\left(-\frac{\mathrm{tr}(\mathbf{E}_f^{\mathrm{H}}\mathbf{E}_f)}{\sigma_{\boldsymbol{\varepsilon}}^2}\right), \end{aligned} \tag{16}$$

where

$$\mathbf{E}_f = \mathbf{X}_f - \mathbf{A}_f(\mathbf{Z}_f \odot \mathbf{S}_f). \tag{17}$$

Here, all data points are assumed to be independent and identically distributed. The smaller the sum of the noise terms is, the higher the likelihood of PF-ISFA is.

4 Inference of PF-ISFA

The model parameters of PF-ISFA are estimated by using an iterative algorithm based on the nonparametric Bayesian model. Sound source separation and SAD are achieved by estimating s_{kft} and b_{kt}, respectively. The parameter update algorithm is given as follows.

1. Initialize parameters using their priors.
2. At each time t, carry out the following:
   - 2-1 For each source k, sample b_{kt} from Equation (26).
   - 2-2 If $b_{\mathrm{kt}} = 1$, sample z_{kft} from Equation (20) and for each frequency bin f; otherwise $z_{\mathrm{kft}} = 0$.
   - 2-3 If $z_{\mathrm{kft}} = 1$, sample s_{kft} from Equation (18); otherwise $s_{\mathrm{kft}} = 0$.
   - 2-4 Determine the number of new classes κ_t, and initialize the parameters.
3. For each source k and frequency bin f, sample the activation probability ψ_{kf} from Equation (28).
4. For each source k and frequency bin f, sample mixing matrix $\mathbf{a}_{\mathrm{kf}}$ from Equation (29).
5. If there is a source that is always inactive, remove it.
6. Update $\sigma_{\boldsymbol{\varepsilon}}^2$, $\sigma_{\mathbf{A}}^2$, and α from Equations (30), (31), and (32), respectively.
7. Go to 2.

This method is based on the Metropolis-Hastings algorithm [24]. The posterior distributions of the latent variables are derived from Bayes' theorem by multiplying the priors by the likelihood function.

4.1 Sound sources

When z_{fkt} is active, s_{fkt} is sampled by using the following posterior.

$$\begin{aligned} P(s_{\mathrm{fkt}}|\mathbf{A}_f, \mathbf{s}_{-\mathrm{fkt}}, \mathbf{x}_{\mathrm{ft}}\mathbf{z}_{\mathrm{ft}}) &\propto P(\mathbf{x}_{\mathrm{ft}}|\mathbf{A}_f, \mathbf{s}_{\mathrm{ft}}, \mathbf{z}_{\mathrm{ft}}, \sigma_{\boldsymbol{\varepsilon}}^2) P(s_{\mathrm{fkt}}) \\ &= \mathcal{N}_C\left(s_{\mathrm{fkt}}; \mu_{s,f}, \sigma_{s,f}^2\right), \end{aligned} \tag{18}$$

where

$$\sigma_{s,f}^2 = \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \mathbf{a}_{\mathrm{fk}}^{\mathrm{H}}\mathbf{a}_{\mathrm{fk}}}, \quad \mu_{s,f} = \frac{\mathbf{a}_{\mathrm{fk}}^{\mathrm{H}}\varepsilon_{-\mathrm{fkt}}}{\sigma_\varepsilon^2 + \mathbf{a}_{\mathrm{fk}}^{\mathrm{H}}\mathbf{a}_{\mathrm{fk}}}.$$

Here, $\mathbf{s}_{-\mathrm{fkt}}$ means $\mathbf{s}_{\mathrm{ft}}$ except for s_{fkt}, and $\boldsymbol{\varepsilon}_{-\mathrm{fkt}}$ means $\boldsymbol{\varepsilon}|_{z_{\mathrm{fkt}}=0}$.

4.2 Source activity of each time-frequency frame

If $b_{\mathrm{kt}} = 1, z_{\mathrm{fkt}}$ is sampled from its posterior distribution. The posterior of z_{fkt} is calculated as follows.

$$P(z_{\mathrm{fkt}}|b_{\mathrm{kt}}, \psi_{\mathrm{kf}}, z_{-\mathrm{fkt}}, \mathbf{x}_{\mathrm{ft}}, \mathbf{s}_{\mathrm{ft}}, \mathbf{A}_f) \propto P_p P_l \tag{19}$$

where

$$P_l = P(x_{\mathrm{ft}}|\mathbf{A}_f, \mathbf{s}_{\mathrm{ft}}, \mathbf{z}_{\mathrm{ft}}, \sigma_{\boldsymbol{\varepsilon}}^2)$$

is the probability of likelihood, and

$$P_p = P(z_{\mathrm{fkt}}|b_{\mathrm{kt}}, \psi_{\mathrm{kf}})$$

is the probability of prior.

Then, the following posterior distribution is derived.

$$P(z_{\mathrm{fkt}}|b_{\mathrm{kt}}, \psi_{\mathrm{kf}}, z_{-\mathrm{fkt}}, \mathbf{x}_{\mathrm{ft}}, \mathbf{s}_{\mathrm{ft}}, \mathbf{A}_f) = \mathrm{Bernoulli}\left(\frac{p_1}{p_0 + p_1}\right), \tag{20}$$

where

$$\log(p_1) = \log(\psi_{\mathrm{kf}}) + \frac{2\mathrm{Re}(s_{\mathrm{fkt}}^* \mathbf{a}_{\mathrm{kf}}^H \boldsymbol{\varepsilon}_{-\mathrm{fkt}}) + |s_{\mathrm{fkt}}|^2 \mathbf{a}_{\mathrm{fk}}^H \mathbf{a}_{\mathrm{fk}}}{\sigma_{\boldsymbol{\varepsilon}}^2} \tag{21}$$

$$\log(p_0) = \log(1 - \psi_{\mathrm{kf}}). \tag{22}$$

4.3 Unified activity for each time frame

To calculate the ratio of the probability that b_{kt} becomes active to the probability that b_{kt} becomes inactive,we use Equation (23). This ratio r is divided into two parts: the ratio of prior r_p and the ratio of the likelihood of fth frequency bin $r_{l,f}$.

$$\begin{aligned} r &= \frac{P(b_{\mathrm{kt}} = 1|\mathbf{A}, \mathbf{S}_{-kt}, \mathbf{X}_t, \mathbf{S}_{-\mathrm{kt}})}{P(b_{\mathrm{kt}} = 0|\mathbf{A}, \mathbf{S}_{-\mathrm{kt}}, \mathbf{X}_t, \mathbf{Z}_{-\mathrm{kt}})} \\ &= r_p \prod_{f=1}^{F} r_{l,f}, \end{aligned} \tag{23}$$

where

$$\begin{aligned} r_p &= \frac{P(b_{\mathrm{kt}} = 1|\mathbf{b}_{\mathrm{kt}})}{P(b_{\mathrm{kt}} = 0|\mathbf{b}_{\mathrm{kt}})}, \text{and} \\ r_{l,f} &= \frac{P(\mathbf{x}_{\mathrm{ft}}|\mathbf{A}_f, \mathbf{s}_{-\mathrm{fkt}}, \mathbf{x}_{\mathrm{ft}}, \mathbf{z}_{-\mathrm{fkt}}, \mathbf{b}_{-\mathrm{kt}}, b_{\mathrm{kt}} = 1, \psi_{\mathrm{kf}}, \sigma_{\boldsymbol{\varepsilon}}^2)}{P(\mathbf{x}_{\mathrm{ft}}|\mathbf{A}_f, \mathbf{s}_{-\mathrm{fkt}}, \mathbf{x}_{\mathrm{ft}}, \mathbf{z}_{-\mathrm{fkt}}, \mathbf{b}_{-\mathrm{kt}}, b_{\mathrm{kt}} = 0, \psi_{\mathrm{kf}}, \sigma_{\boldsymbol{\varepsilon}}^2)}. \end{aligned}$$

Here, $\mathbf{X}_t$ is $\mathbf{x}_{1t}, \ldots, \mathbf{x}_{\mathrm{Ft}}$ and $\mathbf{S}_{-\mathrm{kt}}$ and $\mathbf{Z}_{-\mathrm{kt}}$ are $\mathbf{S}$ and $\mathbf{Z}$ except for $s_{1\mathrm{kt}}, \ldots, s_{\mathrm{Fkt}}$ and $z_{1\mathrm{kt}}, \ldots, z_{\mathrm{Fkt}}$, respectively.

The ratio of prior r_p is calculated by using:

$$r_p = \frac{P(b_{\mathrm{kt}} = 1|\mathbf{b}_{-\mathrm{kt}})}{P(b_{\mathrm{kt}} = 0|\mathbf{b}_{-\mathrm{kt}})} = \frac{m_{k,-t}}{T - m_{k,-t}}, \tag{24}$$

where $m_{k,-t} = \sum_{t' \neq t} b_{kt'}$. This is derived from the priors of source activity based on IBP [22].

The ratio of likelihood $r_{l,f}$ is calculated by using Equation (25).

$$\begin{aligned} r_{l,f} &= \frac{P(\mathbf{x}_{\mathrm{ft}}|\mathbf{A}_f, \mathbf{s}_{-\mathrm{fkt}}, \mathbf{x}_{\mathrm{ft}}, \mathbf{z}_{-\mathrm{fkt}}, \mathbf{b}_{-\mathrm{kt}}, b_{\mathrm{kt}} = 1, \psi_{\mathrm{kf}}, \sigma_{\boldsymbol{\varepsilon}}^2)}{P(\mathbf{x}_{\mathrm{ft}}|\mathbf{A}_f, \mathbf{s}_{-\mathrm{fkt}}, \mathbf{x}_{\mathrm{ft}}, \mathbf{z}_{-\mathrm{fkt}}, \mathbf{b}_{-\mathrm{kt}}, b_{\mathrm{kt}} = 0, \psi_{\mathrm{kf}}, \sigma_{\boldsymbol{\varepsilon}}^2)} \\ &= \psi_{\mathrm{kf}} \sigma_{s,f}^2 \exp\left(\frac{|\mu_{s,f}|^2}{\sigma_{s,f}^2}\right) + (1 - \psi_{\mathrm{kf}}). \end{aligned} \tag{25}$$

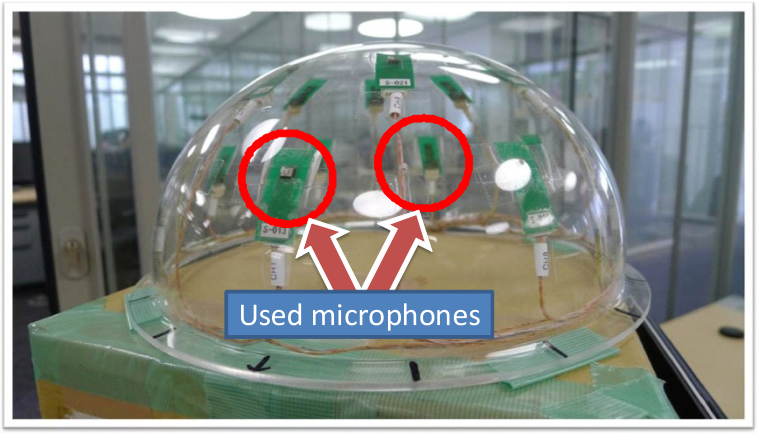

Figure 5 Microphone array for measuring impulse response.

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

$$P(b_{\mathrm{kt}} = 1|\mathbf{A}, \mathbf{S}_{-\mathrm{kt}}, \mathbf{X}_t, \mathbf{Z}_{-\mathrm{kt}}, \mathbf{b}_{-\mathrm{kt}}) = \frac{r}{1+r} \tag{26}$$

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

4.4 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. This κ_t is sampled with the Metropolis-Hastings algorithm.

First, the prior distribution of κ_t is $P(\kappa_t|\alpha) = \mathrm{Poisson}\left(\frac{\alpha}{T}\right)$. After sampling κ_t, we initialize the new sources and their activities. Next, we decide whether this update is acceptable or not. Let ξ and ξ^* be the current state (i.e., the condition before transition) and the next state candidate (the condition after transition), respectively. The acceptance probability of the transition is $\min(1, r_{\xi\to\xi^*})$. According to Meeds [25] and Knowles [14], $r_{\xi\to\xi^*}$ becomes the ratio of the likelihood of the current state to that of the next state. This ratio can be calculated as follows.

$$r_{\xi\to\xi^*} = \prod_{f=1}^{F} (\det\Lambda_{\xi,f})^{-1} \exp\left(\mu_{\xi,f}^{\mathrm{H}} \Lambda_{\xi,f} \mu_{\xi,f}\right), \tag{27}$$

where

$$\Lambda_{\xi,f} = \mathbf{I} + \frac{\mathbf{A}_f^{*\mathrm{H}} \mathbf{A}_f^*}{\sigma_{\boldsymbol{\varepsilon}}^2}, \Lambda_{\xi,f}\mu_{\xi,f} = \frac{1}{\sigma_{\boldsymbol{\varepsilon}}^2} \mathbf{A}_f^{*\mathrm{H}} \boldsymbol{\varepsilon}_{\mathrm{ft}}.$$

Here, $\mathbf{A}_f^*$ is the $D \times \kappa_t$ matrix of the additional part of $\mathbf{A}_f$. When new κ_t sources appear, the mixing matrix should be expanded from $D \times K$ to $D \times (K+\kappa_t)$. $\mathbf{A}_f^*$ means the mixing matrix for these new sources.

4.5 Activation probability for each frequency bin

ψ_{kf} is sampled by the following posterior.

$$\begin{aligned} P(\psi_{\mathrm{kf}}|\mathbf{z}_{\mathrm{kf}}, \Psi_{-\mathrm{kf}}, \mathbf{B}_{-\mathrm{kt}}) &\propto P(\psi_{\mathrm{kf}}|\beta) \prod_{t=1}^{T} P(z_{\mathrm{kft}}|\psi_{\mathrm{kf}}, b_{\mathrm{kt}}) \\ &= \mathrm{Beta}\left(\frac{\beta}{K} + n_{\mathrm{kf}}, \frac{\beta(K-1)}{K} + m_k - n_{\mathrm{kf}}\right), \end{aligned} \tag{28}$$

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

4.6 Mixing matrix

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

$$\begin{aligned} P(\mathbf{a}_{\mathrm{fk}}|\mathbf{A}_{f,-k}, \mathbf{S}_f, \mathbf{X}_f, \mathbf{Z}_f) &\propto P(\mathbf{X}_f|\mathbf{A}_f, \mathbf{S}_f, \mathbf{Z}_f, \sigma_{\boldsymbol{\varepsilon}}^2) P(\mathbf{a}_{\mathrm{fk}}|\sigma_{\mathbf{A}}^2) \\ &= \mathcal{N}_C(\mathbf{a}_{\mathrm{fk}}; \mu_{\mathbf{A}}, \Lambda_{\mathbf{A}}^{-1}), \end{aligned} \tag{29}$$

where

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

4.7 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.

$$\begin{aligned} P(\sigma_{\boldsymbol{\varepsilon}}^2|\mathbf{E}) &\propto P(\mathbf{E}|\sigma_{\boldsymbol{\varepsilon}}^2) P(\sigma_{\boldsymbol{\varepsilon}}^2|p_{\boldsymbol{\varepsilon}}, q_{\boldsymbol{\varepsilon}}) \\ &= \mathcal{IG}\left(\sigma_{\boldsymbol{\varepsilon}}^2; p_{\boldsymbol{\varepsilon}} + \mathrm{FTD}, \frac{q_{\boldsymbol{\varepsilon}}}{(1 + q_{\boldsymbol{\varepsilon}} \sum_{f=1}^{F} \mathrm{tr}(\mathbf{E}_f^{\mathrm{H}} \mathbf{E}_f))}\right). \end{aligned} \tag{30}$$

Table 1 Experimental conditions

No. of sources K	2
Sampling rate	16 [kHz]
STFT window length	64 [ms]
STFT shift length	32 [ms]
Iterations	300 [times]
Hyperparameters	$(p_{\boldsymbol{\varepsilon}}, q_{\boldsymbol{\varepsilon}}) = (10000,\ 1.0)$
	$(p_A, q_A) = (1.1,\ 0.1)$
	$(p_\alpha, q_\alpha) = (3.2,\ 0.21)$
	$\beta = 0.5$

$$P(\sigma_{\mathbf{A}}^2|\mathbf{A}) \propto P(\mathbf{A}|\sigma_{\mathbf{A}}^2)P(\sigma_{\mathbf{A}}^2|p_{\mathbf{A}}, q_{bfA})$$
$$= \mathcal{IG}\left(\sigma_{\mathbf{A}}^2; p_{\mathbf{A}} + \mathrm{FDK}, \frac{q_{\mathbf{A}}}{1 + q_{\mathbf{A}} \sum_{f=1}^{F} \mathrm{tr}(\mathbf{A}_f^{\mathrm{H}} \mathbf{A}_f)}\right) \tag{31}$$

4.8 Concentration parameter of IBP

The posterior distribution of concentration parameter α is

$$p(\alpha|\mathbf{B}) \propto P(\mathbf{B}|\alpha)P(\alpha|p_\alpha, q_\alpha)$$
$$= \mathcal{G}\left(\alpha; K_+ + p_\alpha, \frac{q_\alpha}{1 + q_\alpha H_T}\right), \tag{32}$$

where K_+ is the active number of sources, and $H_n = \sum_{j=1}^{n} \frac{1}{j}$ is the nth harmonic number.

5 Experimental results

In this section, we evaluate the separation performance and the accuracy of the source activity. Section 5.1 presents the results of separation performance and SAD performance compared with FD-ISFA [5]. Section 5.2 shows the separation results compared with PF-ICA [18] using two or four microphones ($D = 2, 4$) and various source locations.

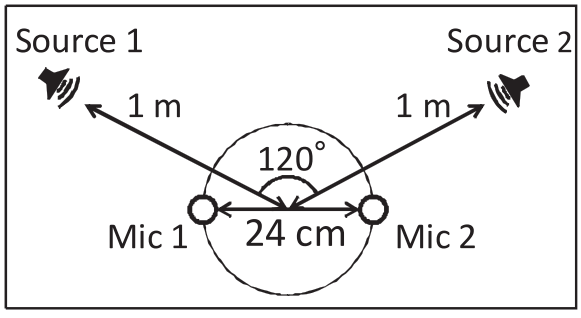

Figure 6 Locations of microphones and sources of Section 5.1.

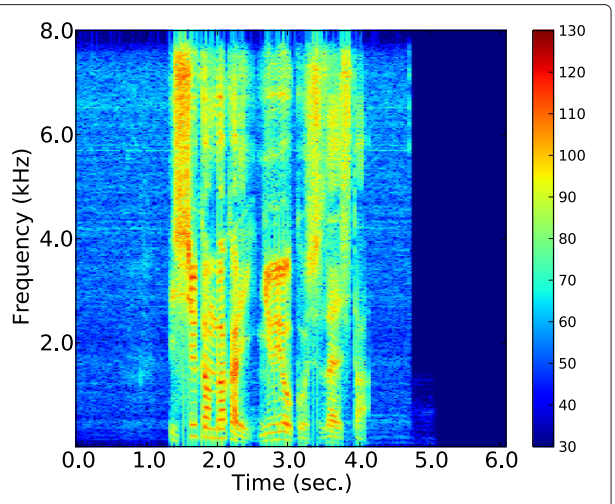

Figure 7 Spectrogram of source signal.

5.1 Compared with FD-ISFA

The experiments used simulated mixtures in four rooms with reverberation times of 20, 150, 400, and 600 [ms]. The simulated mixtures were generated by convoluting the impulse responses measured in the rooms. These impulse responses were recorded by using the microphone array depicted in Figure 5. We use two microphones in these experiments ($D = 2$). The microphone and source locations are shown in Figure 6, and experimental conditions are listed in Table 1. For each condition, 200 mixtures using JNAS phoneme-balanced sentences were tested.

The values of these hyperparameters are empirically-selected. The small σ_ε^2 means the smaller the noise term becomes. Therefore, $p_{\boldsymbol{\varepsilon}}$ and $q_{\boldsymbol{\varepsilon}}$ is set to 10000 and 1.0 in order to get smaller variance. In contrast, σ_A^2 should have a certain amount because σ_A^2 affects the amplitudes

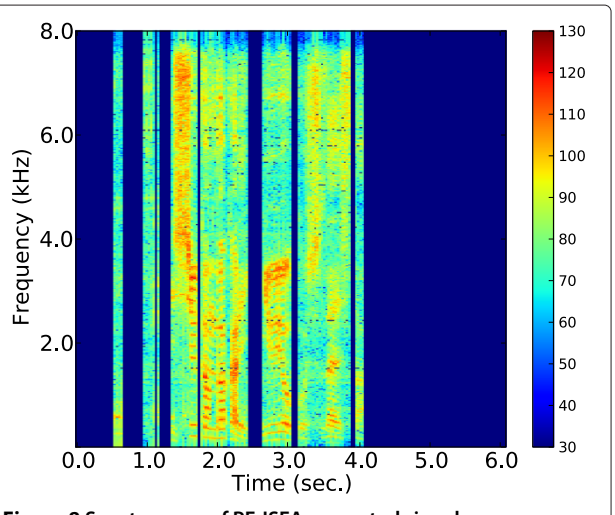

Figure 8 Spectrogram of PF-ISFA separated signal.

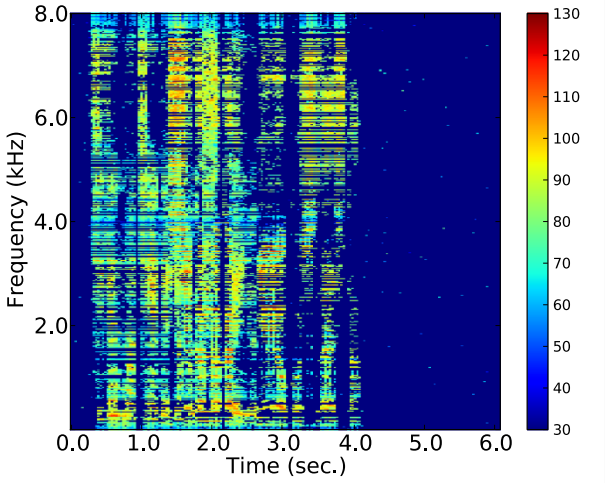

Figure 9 Spectrogram of FD-ISFA separated signal.

of output signals. If σ_A^2 is too large, the power of estimated signals become small, and then these signals are considered to be inactive.

5.1.1 Separation performance

First, an example of the experimental results obtained from the separation experiment using mixed signals ($D = 2$) in a room with reverberation time of 20 [ms] is shown. Spectrograms of a source signal, a signal separated using PF-ISFA, a signal separated using conventional FD-ISFA, and a permutation-aligned signal separated using FD-ISFA are shown in Figures 7, 8, 9, and 10, respectively.

When FD-ISFA is used, the results, shown in Figure 9, contained many horizontal lines; however, there are fewer of these lines in Figure 10. These lines are the spectrogram

Table 2 Separation result of Section 5.1 [dB]

	$RT_{20} = 20ms$			
	PF-ISFA		FD-ISFA	
	Perm	Non-perm	Perm	Solver
SDR	10.07	8.21	12.95	7.94
ISR	17.44	14.71	19.28	13.22
SIR	19.86	16.81	20.30	13.99
SAR	11.70	11.13	15.40	14.48
	$RT_{20} = 150\ ms$			
	PF-ISFA		FD-ISFA	
	Perm	Non-perm	Perm	Solver
SDR	6.85	4.71	6.68	4.21
ISR	11.62	9.30	11.33	8.35
SIR	11.16	8.15	10.84	7.24
SAR	11.38	10.15	11.32	10.62
	$RT_{20} = 400ms$			
	PF-ISFA		FD-ISFA	
	Perm	Non-perm	Perm	Solver
SDR	5.22	3.14	6.53	3.00
ISR	9.97	7.70	10.95	7.26
SIR	9.81	6.93	10.73	6.53
SAR	9.49	8.68	11.77	10.77
	$RT_{20} = 600\ ms$			
	PF-ISFA		FD-ISFA	
	Perm	Non-perm	Perm	Solver
SDR	3.57	1.56	4.26	1.16
ISR	8.07	5.89	8.67	5.37
SIR	7.01	4.07	7.72	3.37
SAR	8.21	7.73	9.19	8.41

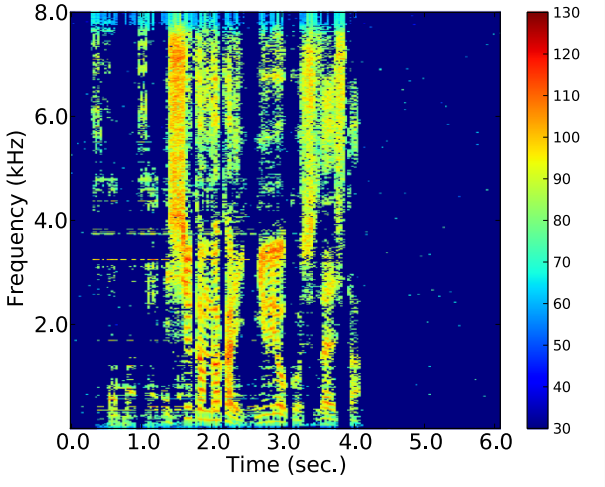

Figure 10 Spectrogram of permutation-aligned FD-ISFA separated signal.

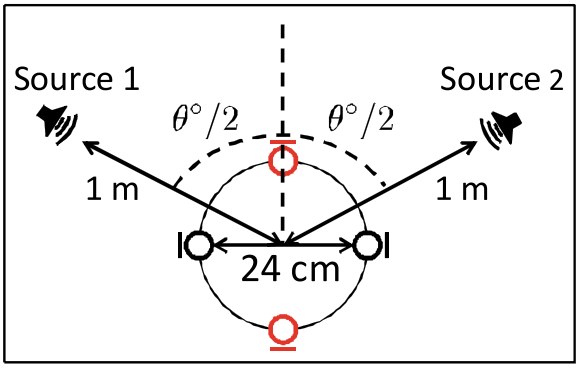

Figure 11 Locations of microphones and sources of Section 5.2.

Table 3 Average SAD performance: precision, recall, and *F*-measure

	Precision (%)		Recall (%)		*F*-measure (%)	
RT_{20}	FD-ISFA	PF-ISFA	FD-ISFA	PF-ISFA	FD-ISFA	PF-ISFA
20 [ms]	64.9	**77.6**	88.8	**84.6**	74.1	**80.0**
150 [ms]	67.6	**69.2**	93.5	**92.9**	77.5	**78.4**
400 [ms]	68.7	**75.1**	89.5	**84.3**	76.9	**78.6**
600 [ms]	71.7	**75.2**	90.7	**87.1**	79.3	**79.8**

of the other separated signal. This means that the output orders of the FD-ISFA results are not aligned for all frequency bins. However, there are no horizontal lines in the spectrogram of PF-ISFA (Figure 8). This shows that the output order is aligned; in other words, the permutation problem has been solved by using PF-ISFA.

The spectrogram shown in Figure 8 has vivid time structure. This indicates that the constraint on the unified activity is too strong and the activation probability for each frequency bin becomes almost one. In order to improve this phenomenon, we might introduce a hyperparameter which can control the activation probability appropriate to observed signals.

We also evaluated 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) [26]. SDR is an overall measure of the separation performance; ISR is a measure of the correctness of the inter-channel information; SIR is a measure of the suppression of the interference signals; and SAR is a measure of the naturalness of the separated signals.

The results are summarized in Table 2. Larger value means better separation. "Non-Perm" was calculated from the output signals themselves; in other words, their permutations were not aligned. "Solver" means that the permutations were aligned using inter-frequency correlation of signal envelope. "Perm" means that the output signals permutations are aligned using the correlation between the outputs and the original sources; in other words, the permutations were aligned by using the original source signals for reference.

Our method (PF-ISFA) outperformed FD-ISFA with permutation solver for all criteria except for SAR under all conditions. In particular, it improved the SIR by 2.82 dB under the condition $RT_{20} = 30$ [ms], 0.91 dB under $RT_{20} = 150$ [ms], 0.41 dB under $RT_{20} = 400$ [ms], and 0.70 dB under $RT_{20} = 600$ [ms].

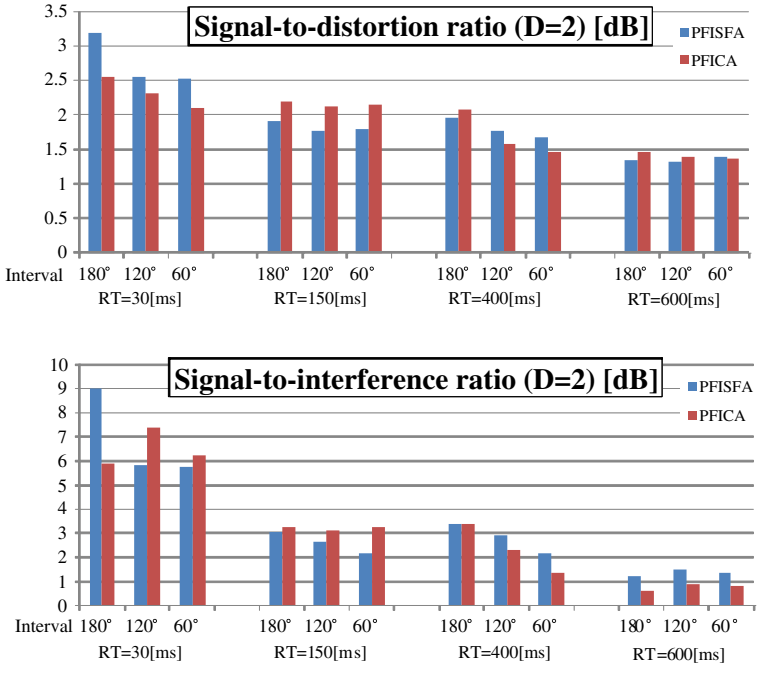

Figure 12 Separation results of Section 5.2 for each interval when $D = 2$ (upper: SDR, lower: SIR).

One of the reasons of the poor performance of FD-ISFA is due to the cascade approach. The results show that FD-ISFA achieves better performance if the permutation problem is perfectly solved. Therefore, this poor performance comes from the permutation solver. This indicates that the overall performance of cascade approach is severely affected by the performance of worst subprocess.

These results show that the performance in rooms with reverberation times of 150, 400, and 600 [ms] is worse than for $RT_{20} = 30$ [ms] reverberation. This is because the reverberation time of these rooms are longer than the STFT window length (64 [ms]). If the reverberation time is longer than the STFT window length, the reverberation affects multiple time frames, and this degrades the performance.

The result of PF-ISFA (Perm) and that of PF-ISFA (Non-Perm) is different. If the source activity results are poor, the activities of two separated signals become similar. In this case, the permutation ambiguity is likely to arise because the unified activity matrix becomes meaningless. In other words, PF-ISFA marks better result when each source signal has different activity.

5.1.2 SAD performance

Next, we evaluated our method in SAD accuracy. The SAD result of PF-ISFA was estimated as unified source activities, that is the parameter b_{ft} in Section 3.3. Since FD-ISFA estimated the sound activity for each frequency bin independently, we calculated the number of active bins for each time frame and determined the source activity of each time frame by using threshold processing.

The precision rate, recall rate, and F-measure of the source activity accuracy are listed in Table 3. PF-ISFA results are indicated by bold type. PF-ISFA outperformed FD-ISFA in precision rate and F-measure in all reverberant conditions. In particular, it improved the F-measure by 5.9 points, 0.9 points, 1.7 points, and 0.5 points under the conditions $RT_{20} = 30$ [ms], 150 [ms], 400 [ms], and 600 [ms], respectively.

Our method achieved a better precision rate and lower recall rate than FD-ISFA, and the results show that PF-ISFA achieved robust SAD performance under reverberant condition. This is because PF-ISFA estimates the source activities using a unified parameter for all frequency bins. PF-ISFA is less likely to determine that the time frame is active, even if some frequency bins have a certain power level.

5.2 Compared with PF-ICA

In second experiment, we used two or four microphones ($D = 2, 4$) to observe the two sound source mixture with interval $\theta = 60, 120,$ and180[deg]. For each interval, 20 mixtures were tested using JNAS phoneme-balanced sentences. The microphone and source locations is shown in Figure 11. We use red microphones when $D = 2$. In order to calculate SDR, ISR, SIR, and SAR, two signals which maximize SDR score are chosen from estimated signals when using four microphones.

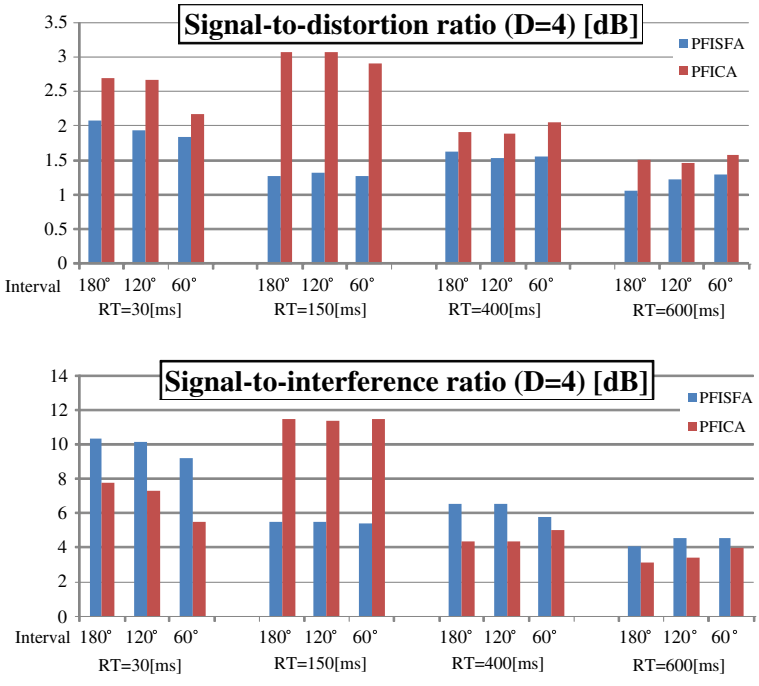

Figure 13 Separation results of Section 5.2 for each interval when $D = 4$ (upper: SDR, lower: SIR).

Table 4 Separation result of Section 5.2 [dB]

	$RT_{20} = 20\ ms$			
	D=2		D=4	
	PF-ISFA	PF-ICA	PF-ISFA	PF-ICA
SDR	2.76	2.32	1.95	2.51
ISR	3.53	3.54	2.36	3.27
SIR	6.87	6.48	9.90	6.87
SAR	7.39	9.96	5.61	11.28
	$RT_{20} = 150\ ms$			
	D=2		D=4	
	PF-ISFA	PF-ICA	PF-ISFA	PF-ICA
SDR	1.82	2.15	1.28	3.01
ISR	2.71	3.02	1.64	3.34
SIR	2.64	3.19	5.50	11.45
SAR	6.37	7.67	3.58	9.55
	$RT_{20} = 400\ ms$			
	D=2		D=4	
	PF-ISFA	PF-ICA	PF-ISFA	PF-ICA
SDR	1.79	1.71	1.57	1.95
ISR	2.80	2.93	1.96	2.69
SIR	2.84	2.35	6.31	4.60
SAR	6.56	9.37	4.46	9.62
	$RT_{20} = 600\ ms$			
	D=2		D=4	
	PF-ISFA	PF-ICA	PF-ISFA	PF-ICA
SDR	1.36	1.40	1.19	1.51
ISR	2.38	2.52	1.59	2.22
SIR	1.36	0.79	4.40	3.50
SAR	5.81	9.41	3.69	8.55

The average SDR and SIR of separated signals are shown in Figures 12 and 13 for each interval when $D = 2$ and 4, respectively. Table 4 summarizes average SDR, ISR, SIR, and SAR of all intervals.

Table 4 indicates that PF-ISFA marks better average SIR except for the condition $RT_{20} = 150$ [ms]. This means that PF-ISFA can suppress the interference signal better than PFICA. PF-ISFA and PF-ICA marks similar results by the average SDR when $D = 2$, and The SDR score of PF-ISFA is lower than that of PF-ICA when $D = 4$. This is because these SDR scores are affected by the SAR scores. The output signals of PF-ICA are created by multiplying separation matrix by observed signals. Then, the artificial noise is not likely to emerge. In contrast, PF-ISFA estimates the source signals by sampling, and PF-ISFA output is based on the best one sample of all samples created during estimation.

6 Conclusion and future study

This article presented a joint estimation method of BSS and SAD in the frequency domain that also solves the permutation problem. It was designed by using a nonparametric Bayesian approach. Unified source activity was introduced to automatically align the permutations of the output order for all frequency bins.

Our method improves the average SIR by 2.82–0.41 dB compared with the baseline method based on FD-ISFA when separating convoluted mixtures of $RT_{20} = 30$ [ms]–600 [ms] room environments. It also outperforms FD-ISFA under reverberant conditions ($RT_{20} = 150, 400, 600$ ms). For SAD performance, our method outperforms the conventional method by 5.9–0.5% in F-measure under the condition $RT_{20} = 20$–600 [ms], respectively.

In the future, we will evaluate the separation performance of a mixture of signals from three or more talkers. We will attempt to develop a method that can separate mixtures with longer reverberations (i.e., longer than the STFT window length) robustly. Last but not least, the method should be sped up to achieve real-time processing so that it can be applied to robot applications.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This study was partially supported by KAKENHI and Honda Research Institute Japan Inc., Ltd.

Received: 16 June 2012 Accepted: 27 October 2012
Published: 22 January 2013

References

1. D Rosenthal, HG Okuno, *Computational auditory scene analysis* (CRC press, USA, 1998)
2. D Wang, G Brown, *Computational auditory scene analysis: principles, algorithms, and applications* (Wiley-IEEE press, USA, 2006)
3. J Sohn, N Kim, W Sung, A statistical model-based voice activity detection. IEEE Signal Process. Lett. **6**, 1–3 (1999)
4. J Ramırez, J Segura, C Benıtez, A De La Torre, A Rubio, Efficient voice activity detection algorithms using long-term speech information. Speech Commun. **42**(3), 271–287 (2004)
5. K Nagira, T Takahashi, T Ogata, HG Okuno, in *Proc. of International Conference on Latent Variable Analysis and Signal Separation*. Complex extension of infinite sparse factor analysis for blind speech separation, Tel-Aviv, 2012), pp. 388–396
6. MS Pedersen, J Larsen, U Kjems, LC Parra, in *Springer Handbook of Speech Processing*, ed. by J Benesty, MM Sondhi, and Y Huang. Convolutive blind source separation methods, Part I (Springer Press, 2008), pp. 1065–1094
7. K Nakadai, T Takahashi, H Okuno, H Nakajima, Y Hasegawa, H Tsujino, Design and implementation of robot audition system "HARK" open source software for listening to three simultaneous speakers. Adv. Robot. **24**(5), 739–761 (2010)
8. F Asano, S Ikeda, M Ogawa, H Asoh, N Kitawaki, Combined approach of array processing and independent component analysis for blind separation of acoustic signals. IEEE Trans. Speech Audio Process. **11**(3), 204–215 (2003)
9. H Nakajima, K Nakadai, Y Hasegawa, H Tsujino, in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*. Adaptive step-size

parameter control for real-world blind source separation, Las Vegas, 2008), pp. 149–152
10. P Comon, Independent component analysis, a new concept? Signal Process. **36**(3), 287–314 (1994)
11. A Hyvärinen, Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans. Neural Netw. **10**(3), 626–634 (1999)
12. J Cardoso, A Souloumiac, Blind beamforming for non-Gaussian signals. IEE Proceedings F Radar and Signal Processing. **140**(6), 362–370 (1993)
13. H Sawada, R Mukai, S Araki, S Makino, in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*. Polar coordinate based nonlinear function for frequency-domain blind source separation, Orlando, 2002), pp. 1001–1004
14. D Knowles, Z Ghahramani, in *Proc. of Independent Component Analysis and Signal Separation*. Infinite sparse factor analysis and infinite independent components analysis, London, 2007), pp. 381–388
15. H Sawada, R Mukai, S Araki, S Makino, A robust and precise method for solving the permutation problem of frequency-domain blind source separation. IEEE Trans. Speech Audio Process. **12**(5), 530–538 (2004)
16. H Sawada, S Araki, S Makino, in *Proc. of IEEE International Symposium on Circuits and Systems*. Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS, New Orleans, 2007), pp. 3247–3250
17. I Lee, T Kim, T Lee, Fast fixed-point independent vector analysis algorithms for convolutive blind source separation. Signal Process. **87**(8), 1859–1871 (2007)
18. A Hiroe, in *Proc. of International Conference on Independent Component Analysis and Blind Signal Separation*. Solution of permutation problem in frequency domain ICA, using multivariate probability density functions, Charleston, 2006), pp. 601–608
19. J Hirayama, S Maeda, S Ishii, Markov and semi-Markov switching of source appearances for nonstationary independent component analysis. IEEE Trans. Neural Netw. **18**(5), 1326–1342 (2007)
20. H Hsieh, J Chien, in *Proc. of IEEE International Conference on Acoustics Speech and Signal Processing*. Online Bayesian learning for dynamic source separation, Dallas, 2010), pp. 1950–1953
21. S Araki, H Sawada, S Makino, in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1. Blind speech separation in a meeting situation with maximum SNR beamformers, Honolulu, 2007), pp. 41–44
22. T Griffiths, Z Ghahramani, Infinite latent feature models and the Indian buffet process. Adv. Neural Inf. Process. Syst. **18**, 475–482 (2006)
23. N Murata, S Ikeda, A Ziehe, An approach to blind source separation based on temporal structure of speech signals. Neurocomputing. **41**, 1–24 (2001)
24. W Hastings, Monte Carlo sampling methods using Markov chains and their applications. Biometrika. **57**(1), 97–109 (1970)
25. E Meeds, Z Ghahramani, R Neal, S Roweis, Modeling dyadic data with binary latent factors. Adv. Neural Inf. Process. Syst. **19**, 977–984 (2007)
26. E Vincent, H Sawada, P Bofill, S Makino, J Rosca, in *Proc. of Independent Component Analysis and Signal Separation*. First stereo audio source separation evaluation campaign: data, algorithms and results, London, 2007), pp. 552–559

doi:10.1186/1687-4722-2013-4
Cite this article as: Nagira *et al.*: **Nonparametric Bayesian sparse factor analysis for frequency domain blind source separation without permutation ambiguity.** *EURASIP Journal on Audio, Speech, and Music Processing* 2013 **2013**:4.