Speech Enhancement Based on Bayesian Low-Rank and Sparse Decomposition of Multichannel Magnitude Spectrograms

Yoshiaki Bando, *Student Member, IEEE*, Katsutoshi Itoyama, *Member, IEEE*, Masashi Konyo, *Member, IEEE*, Satoshi Tadokoro, *Fellow, IEEE*, Kazuhiro Nakadai, *Senior Member, IEEE*, Kazuyoshi Yoshii, *Member, IEEE*, Tatsuya Kawahara, *Fellow, IEEE*, and Hiroshi G. Okuno, *Fellow, IEEE*

***Abstract*—This paper presents a blind multichannel speech enhancement method that can deal with the time-varying layout of microphones and sound sources. Since nonnegative tensor factorization (NTF) separates a multichannel magnitude (or power) spectrogram into source spectrograms without phase information, it is robust against the time-varying mixing system. This method, however, requires prior information such as the spectral bases (templates) of each source spectrogram in advance. To solve this problem, we develop a Bayesian model called robust NTF (Bayesian RNTF) that decomposes a multichannel magnitude spectrogram into target speech and noise spectrograms based on their sparseness and low rankness. Bayesian RNTF is applied to the challenging task of speech enhancement for a microphone array distributed on a hose-shaped rescue robot. When the robot searches for victims under collapsed buildings, the layout of the microphones changes over time and some of them often fail to capture target speech. Our method robustly works under such situations, thanks to its characteristic of time-varying mixing system. Experiments using a 3-m hose-shaped rescue robot with eight microphones show that the proposed method outperforms conventional blind methods in enhancement performance by the signal-to-noise ratio of 1.03 dB.**

***Index Terms*—Multichannel speech enhancement, low-rank and sparse decomposition, Bayesian signal processing.**

Manuscript received May 1, 2017; revised August 11, 2017 and September 28, 2017; accepted October 24, 2017. Date of publication November 10, 2017; date of current version December 11, 2017. This study was supported in part by the ImPACT Tough Robotics Challenge and in part by JSPS KAKENHI under Grants 24220006 and 15J08765. This paper was presented in part at the Proceedings of the 24th European Signal Processing Conference, Hotel Hilton, Budapest, September 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Roland Badeau. *(Corresponding author: Yoshiaki Bando.)*

Y. Bando, K. Itoyama, K. Yoshii, and T. Kawahara are with the Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan (e-mail: yoshiaki@kuis.kyoto-u.ac.jp; itoyama@kuis.kyoto-u.ac.jp; yoshii@kuis.kyoto-u.ac.jp; kawahara@i.kyoto-u.ac.jp).

M. Konyo and S. Tadokoro are with the Graduate School of Information Science, Tohoku University, Sendai 980-8579, Japan (e-mail: konyo@rm.is.tohoku.ac.jp; tadokoro@rm.is.tohoku.ac.jp).

K. Nakadai is with Tokyo Institute of Technology, Saitama 351-0114, Japan, and also with the Honda Research Institute Japan Co., Ltd, Wako 351-0188, Japan (e-mail: nakadai@jp.honda-ri.com).

H. G. Okuno is with the Graduate Program for Enbodiment Informatics, Waseda University, Tokyo 169-0072, Japan (e-mail: okuno@nue.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2017.2772340

I. Introduction

SPEECH enhancement has been studied for various applications such as hearing aids, speech recognition, and speech telecommunication [2]–[8]. In these applications, it is often difficult to assume the typical usage situation, such as noise characteristics and the relative layout of sources and microphones. To enhance noisy speech signals with few assumptions on the usage situation, blind speech enhancement has been studied by focusing on some statistical structures of observed signals [6]–[11]. Single-channel speech enhancement, for example, focuses on the spectral pattern difference between speech and noise signals [2], [8]. Multichannel speech enhancement focuses on the inter-channel correlation difference between them, which depends on the relative layout of sources and microphones [6], [7], [11].

This study addresses to develop a blind multichannel speech enhancement method that is robust against the time-varying layout of sources and microphones. While most of the existing blind speech enhancement (or source separation) methods assume that the mixing system is time-invariant, this assumption does not always hold [11]–[14]. A possible way is to enhance speech in magnitude (or power) spectrogram domain, which is insensitive to relatively small changes of the layout. Non-negative tensor factorization (NTF), for example, can separate a multichannel magnitude spectrogram into source spectrograms [15]–[18]. NTF, however, requires prior information such as the spectral bases (templates) of each source spectrogram, and thus it is not completely blind.

This paper presents blind multichannel speech enhancement that works in magnitude spectrogram domain based on low-rank and sparse decomposition. Low-rank and sparse decomposition, such as robust non-negative matrix factorization (RNMF), can decompose a magnitude spectrogram into low-rank and sparse spectrograms without any prior training [10], [19]–[23]. The low-rank spectrogram corresponds to a noise spectrogram that can be represented by a small number of spectral bases (e.g. motor noises). The sparse spectrogram corresponds to a speech spectrogram that has harmonic structures. Our method is inspired by NTF and RNMF, and decomposes a multichannel

*Demo page: http://sap.ist.i.kyoto-u.ac.jp/members/yoshiaki/demo/vb-srntf/

2329-9290 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

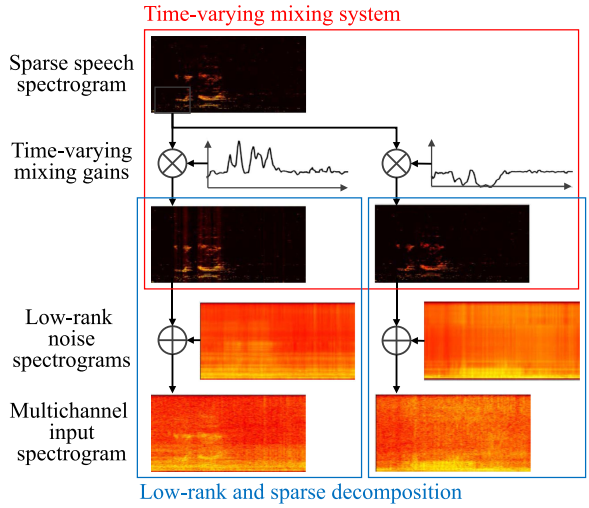


Fig. 1. Overview of the proposed Bayesian RNTF.

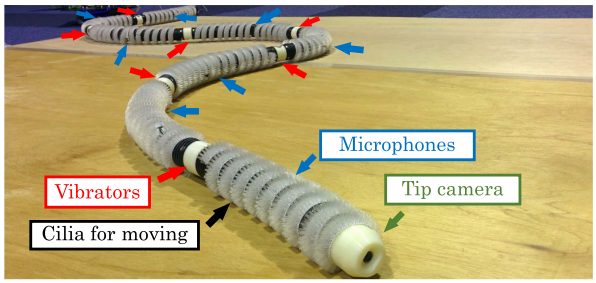


Fig. 2. Hose-shaped rescue robot with an eight-channel microphone array.

magnitude spectrogram into channel-wise low-rank noise spectrograms and sparse speech spectrogram common to all the channels (Fig. 1). It is formulated as a Bayesian generative model called Bayesian robust NTF (Bayesian RNTF). Since its mixing system is independently estimated at each time frame, it is robust against the time-varying layout of sources and microphones.

Bayesian RNTF is applied to speech enhancement with a microphone array on a hose-shaped rescue robot. This kind of robots, which are characterized by a thin, long and flexible body, have been developed for penetrating narrow gaps under collapsed buildings [24], [25]. The Active Scope Camera (ASC) robot, for example, moves forward by the vibrating cilia covering its body [24] (Fig. 2). While a robot operator searches for survived victims using a microphone array and a tip camera equipped on the robot, speech signals captured by microphones on the robot are contaminated by non-stationary ego-noise (e.g., motor and friction noise). The naïve "stop-and-listen" strategy for avoiding the ego-noise prevents a robot operator from finding survived victims quickly. It is important to check whether the human voice is included in the captured signal. Speech enhancement is helpful to prevent the operator from failing to detect the voice and to improve the performance of its automatic detection.

The speech enhancement for a hose-shaped rescue robot imposes the following three challenging problems:

1) *Environment-dependence of ego-noise:* The ego-noise changes over time depending on the robot's movements and surrounding materials.
2) *Deformable layout of microphones:* The relative positions of the microphones change over time because of the vibration and deformation of the robot body.
3) *Partial occlusion of microphones:* Some of the microphones often fail to capture target speech when they are shaded by rubble around the robot.

These problems make it impossible to use conventional supervised methods [3]–[5], [26]–[28], and degrade the conventional blind methods that assume a time-invariant mixing system [11]–[14]. On the positive side, since the ego-noise is generated from the vibration motors, the noise spectrogram has repetitive structures and is, thus, considered as low rank. The proposed Bayesian RNTF is based on the low-rank and sparse decomposition and time-varying mixing system, and thus it is robust against the first two problems. Moreover, it can deal with the occlusion problem because it estimates the speech level at each microphone.

In actual rescue activities searching for victims, real-time speech enhancement is crucial. Bayesian RNTF is extended to a state-space model called Bayesian streaming RNTF (Bayesian SRNTF) that represents the dynamics of the latent variables. The Bayesian inference of our method is conducted in a mini-batch manner with a variational Bayesian (VB) framework [29], [30]. We show that our Bayesian SRNTF works in real time on a mobile general-purpose graphics processing unit (GPGPU).

The rest of the paper is organized as follows: Section II reviews related work on multichannel blind source separation and low-rank and sparse decomposition. Section III explains the proposed Bayesian RNTF model, and Section IV derives its Bayesian inference algorithm. Section V reports the experimental evaluations of Bayesian RNTF with simulated data, and Section VI reports the experimental results obtained using recorded signals. Finally Section VII summarizes the key findings.

II. Related Work

This section overviews related work on multichannel blind source separation and low-rank and sparse decomposition.

A. Multichannel Blind Source Separation

Blind source separation based on the phase differences between the microphones can be used without prior knowledge about microphones and sources [7], [11]–[14], [31]. Multichannel non-negative matrix factorization (MNMF) [11], [13], [32], for example, decomposes given multichannel complex spectrograms into multiple low-rank source spectrograms and their transfer functions. Each of source spectrograms is represented as a product of spectral basis vectors and their temporal activation vectors. Kounades-Bastian *et al.* [32] extended MNMF for moving sources by assuming a Markov chain of time-varying transfer functions. Its performance may, however, be degraded by unexpected moving of sources.

One way to avoid estimating the time-varying transfer functions of sound sources is to perform source separation over multichannel magnitude (or power) spectrograms, which are insensitive to relatively small motions. NTF has been used

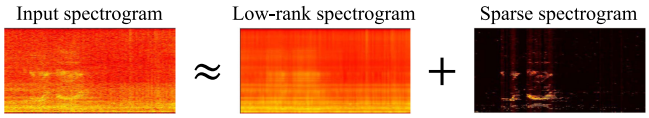


Fig. 3. Speech enhancement by low-rank and sparse decomposition. Speech signals that have a sparse structure are separated from low-rank noise signals.

for decomposing the multichannel spectrogram into source spectrograms and their magnitude transfer functions [15]–[18]. Murata *et al.* [16] proposed an NTF by marginalizing out the phase term of MNMF. This method, however, requires the basis vectors for each source in advance. Although another NTF that does not need information about the basis vectors was proposed [17], it requires the volume level ratio of each source in the channels in advance. To make NTF be completely blind, it is necessary to import other separation criteria that can remove the constraints.

B. Low-Rank and Sparse Decomposition

Low-rank and sparse decomposition is a popular approach to suppressing non-stationary periodic noise and enhancing target speech without prior training (Fig. 3) [10], [20]–[23]. Let $\mathbf{Y} \in \mathbb{R}^{F\times T}$, $\mathbf{L} \in \mathbb{R}^{F\times T}$, and $\mathbf{S} \in \mathbb{R}^{F\times T}$ be input, low-rank and sparse matrices (magnitude spectrograms with F frequency bins and T time frames), respectively. This decomposition was originally proposed in robust principal component analysis (RPCA) and is conducted by solving the following minimization problem with the augmented Lagrange multiplier framework [23]:

$$\underset{\mathbf{L}*,\mathbf{S}}{\operatorname{argmin}} \|\mathbf{Y}\|_* + \lambda\|\mathbf{S}\|_1 \text{ s.t. } \mathbf{Y} = \mathbf{L} + \mathbf{S}, \tag{1}$$

where $\|\cdot\|_*$ is the nuclear norm representing the low-rankness, $\|\cdot\|_1$ is the L1 norm representing the sparsity, and λ represents a scale parameter controlling the sparseness of $\mathbf{S}$. To reduce the processing time of RPCA, the following relaxed problem of (1) is proposed by replacing the equality constraint with a penalty term [33], [34]:

$$\underset{\mathbf{L},\mathbf{S}}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{Y} - \mathbf{L} - \mathbf{S}\|_F^2 + \lambda_1\|\mathbf{L}\|_* + \lambda_2\|\mathbf{S}\|_1, \tag{2}$$

where $\|x\|_F$ is the Frobenius norm and λ_1 and λ_2 are the scale parameters. When these scale parameters are small enough, the solutions to (2) approach the solutions to (1).

Equation (2) can be interpreted as a likelihood function ($\frac{1}{2}\|\mathbf{Y} - \mathbf{L} - \mathbf{S}\|_F^2$) with priors for the latent variables ($\lambda_1\|\mathbf{L}\|_*$ and $\lambda_2\|\mathbf{S}\|_1$). Bayesian RPCA has been studied for dealing with uncertainty of latent low-rank and sparse components [29], [35]. Babacan *et al.* [29] derived a VB algorithm for Bayesian RPCA (VB-RPCA) to reduce the computational cost. Bayesian RPCA represents the low-rank matrix $\mathbf{L}$ as the product of K basis vectors $\mathbf{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K] \in \mathbb{R}^{F\times K}$ and their coefficient vectors $\mathbf{H} \in \mathbb{R}^{K\times T}$ as follows:

$$\mathbf{L} = \mathbf{W}\mathbf{H}. \tag{3}$$

Note that the rank of the low-rank matrix $\mathbf{L}$ is constrained to be K or less. Using this low-rank model, the likelihood function is defined with a Gaussian distribution (denoted by $\mathcal{N}$) as follows:

$$p(\mathbf{Y}|\mathbf{W},\mathbf{H},\mathbf{S}) = \prod_{f,t}\mathcal{N}\left(y_{ft}\left|\sum_k w_{fk}h_{kt} + s_{ft}, \sigma\right.\right) \propto \exp\left(-\frac{1}{\sigma}\|\mathbf{Y} - \mathbf{W}\mathbf{H} - \mathbf{S}\|_F^2\right), \tag{4}$$

where σ is a variance parameter and is simultaneously estimated with other parameters. The low-rankness and sparseness of $\mathbf{L}$ and $\mathbf{S}$ are controlled by this structural constraint and their prior distributions. Ding *et al.* [35] proposed a Bayesian RPCA whose prior distribution of sparse components has Markovian constraint. This model was used for separating background and foreground images from video streams and reduced salt-and-pepper noise of estimated foreground images. Application of RPCA to audio or image data, however, is not physically justified because RPCA allows observation, low-rank, and sparse spectrograms or images to take negative values.

By constraining the low-rank and sparse matrices to be non-negative, RNMF was proposed for analyzing audio spectrograms or video images [8]–[10], [19], [22]. Since the Frobenius norm (Euclidean distance) in (2) and (4) often causes over-emphasis of high-energy components in a magnitude spectrogram, Li *et al.* [8] proposed an RNMF with the Kullback-Leibler (KL) divergence, which has been widely used in audio source separation:

$$\underset{\mathbf{W},\mathbf{H},\mathbf{S}}{\operatorname{argmin}} \operatorname{KL}(\mathbf{Y}|\mathbf{W}\mathbf{H} + \mathbf{S}) + \lambda\|\mathbf{S}\|_1, \tag{5}$$

where $\operatorname{KL}(\cdot|\cdot)$ represents the KL divergence. Min *et al.* [9] proposed an RNMF with the Itakura-Saito divergence, which is derived from a statistical generative model of acoustic signals. Like Bayesian RPCA, Bayesian formulation of RNMF is expected to allow for further extensions such as multichannel signal processing.

III. Bayesian Model of Robust Non-Negative Tensor Factorization

This section describes the proposed RNTF model that represents a multichannel magnitude spectrogram by channel-wise low-rank components and sparse components common to all the channels as shown in Fig. 1. Since our method does not use the phase information, the phase differences across channels do not affect the result. To derive the proposed multichannel model, we first formulate a Bayesian reformulation of RNMF (Bayesian RNMF) that is inspired by Bayesian NMF [36] and Bayesian RPCA [29]. We then formulate its multichannel extension (Bayesian RNTF) as a statistical generative model, and finally derive the mini-batch extension called Bayesian SRNTF by reformulating the batch Bayesian RNTF to a state-space model.

A. Bayesian RNMF for Single-Channel Enhancement

We first formulate an offline single-channel enhancement model called Bayesian RNMF. The problem of Bayesian RNMF is defined as follows:

Input: Single-channel magnitude spectrogram $\mathbf{Y} \in \mathbb{R}_+^{F \times T}$
Output: Denoised magnitude spectrogram $\mathbf{S} \in \mathbb{R}_+^{F \times T}$
Assumption:
The following values are given in advance:
A) Possible maximum rank of noise spectrogram $K \in \mathbb{N}$
B) Hyperparameters $\alpha^w \in \mathbb{R}_+$, $\beta^w \in \mathbb{R}_+$, $\alpha^h \in \mathbb{R}_+$, $\beta^h \in \mathbb{R}_+$, and $\alpha^s \in \mathbb{R}_+$

where F and T indicate number of frequency bins and time frame bins, respectively. The magnitude spectrogram is defined as the absolute values of the short-time Fourier transform (STFT) of a time-domain signal. Interpretations of the hyperparameters are explained below.

1) Overview: As in existing low-rank and sparse decomposition methods ((2), (4), and (5)), Bayesian RNMF approximates an input spectrogram $\mathbf{Y} \in \mathbb{R}_+^{F \times T}$ as the sum of a low-rank spectrogram $\mathbf{L} \in \mathbb{R}_+^{F \times T}$ (noise) and a sparse spectrogram $\mathbf{S} \in \mathbb{R}_+^{F \times T}$ (target speech) as follows:

$$\mathbf{Y} \approx \mathbf{L} + \mathbf{S}. \tag{6}$$

The low-rank spectrogram is represented by the product of K spectral basis vectors $\mathbf{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K] \in \mathbb{R}_+^{F \times K}$ and their temporal activation vectors $\mathbf{H} = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_T] \in \mathbb{R}_+^{K \times T}$:

$$\mathbf{Y} \approx \mathbf{W}\mathbf{H} + \mathbf{S}. \tag{7}$$

The low-rankness and sparseness of each term can be controlled in a Bayesian manner as explained below.

2) Likelihood Function: Bayesian RNMF tries to minimize the approximation error for the input spectrogram by using the KL divergence. Since the maximization of a Poisson likelihood corresponds to the minimization of a KL divergence, the likelihood function is defined as follows:

$$p(\mathbf{Y}|\mathbf{W}, \mathbf{H}, \mathbf{S}) = \prod_{ft} \mathcal{P}\left(y_{ft} \left| \sum_k w_{fk} h_{kt} + s_{ft} \right.\right), \tag{8}$$

where $\mathcal{P}(x|\lambda) \propto \frac{1}{\Gamma(x+1)} \lambda^x e^{-\lambda}$ denotes a Poisson distribution with a rate parameter $\lambda \in \mathbb{R}_+$. Although the discrete Poisson distribution can be used by quantizing the observation y_{ft}, it has been empirically shown that NMF with the continuous Poisson likelihood performs as well as those of the discrete distribution [37].

3) Prior Distributions on Low-Rank Components: Our low-rank modeling is inspired by Bayesian NMF [36] that has been studied for low-rank decomposition of audio spectrograms. Since the gamma distribution is a conjugate prior for the Poisson distribution, gamma priors are put on the basis and activation matrices of the low-rank components as follows:

$$p(\mathbf{W}|\alpha^w, \beta^w) = \prod_{f,k} \mathcal{G}(w_{fk}|\alpha^w, \beta^w), \tag{9}$$

$$p(\mathbf{H}|\alpha^h, \beta^h) = \prod_{k,t} \mathcal{G}(h_{kt}|\alpha^h, \beta^h), \tag{10}$$

where $\mathcal{G}(x|\alpha, \beta)$ denotes a gamma distribution with a shape parameter α and a rate parameter β; $\alpha^w \in \mathbb{R}_+$, $\beta^w \in \mathbb{R}_+$, $\alpha^h \in \mathbb{R}_+$, and $\beta^h \in \mathbb{R}_+$ are the hyperparameters which should be appropriately set in advance. Setting the shape parameters α^w and α^h to 1.0 or less forces the basis and activation matrices to be sparse [36], which means that the low-rank component $\mathbf{L}$ is forced to be low-rank. These prior distributions enhance the low-rankness of this component compared to the original RNMF.

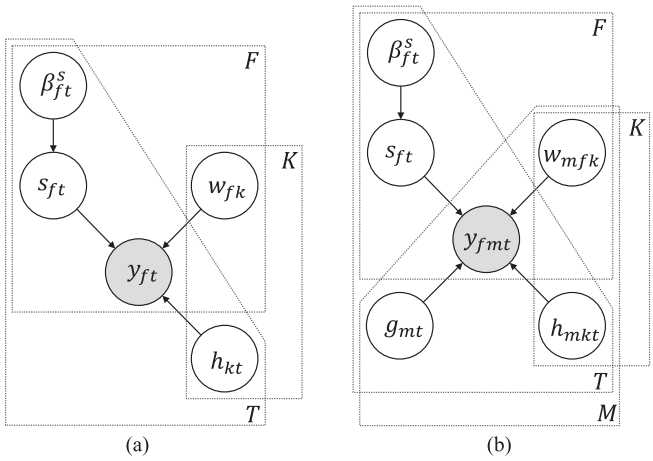


Fig. 4. Graphical models for Bayesian RNMF and RNTF. (a) Bayesian RNMF. (b) Bayesian RNTF.

4) Prior Distributions on Sparse Components: In Bayesian RPCA, Gaussian priors with the Jeffreys hyperpriors are put on sparse components [29]. To force the sparse components to take non-negative values, gamma priors are put on the sparse components as follows:

$$p(\mathbf{S}|\alpha^s, \boldsymbol{\beta}^s) = \prod_{f,t} \mathcal{G}(s_{ft}|\alpha^s, \beta^s_{ft}), \tag{11}$$

where $\alpha^s \in \mathbb{R}_+$ and $\beta^s_{ft} \in \mathbb{R}_+$ represent the shape and rate hyperparameters of the gamma distributions, respectively. To estimate the rate hyperparameters, the Jeffreys hyperpriors are put on them as follows:

$$p(\beta^s_{ft}) \propto (\beta^s_{ft})^{-1}. \tag{12}$$

The rate hyperparameters are independently defined at individual time-frequency bins. The significance of each time-frequency bin is automatically estimated by optimizing the rate hyperparameter as in Bayesian RPCA [29]. The shape hyperparameter α^s, on the other hand, controls the sparseness of the sparse component $\mathbf{S}$ and should be set appropriately in advance. The complete graphical model that represents the probabilistic dependency of the latent variables is shown in Fig. 4(a).

B. Bayesian RNTF for Multichannel Enhancement

We then formulate a multichannel extension of Bayesian RNMF called Bayesian RNTF. The problem in this section is defined as follows:
where m represents the microphone index. Interpretations of the hyperparameters are explained below. Our method is designed for enhancing speech sounds coming from one direction at each

Input: M-channel magnitude spectrograms $\mathbf{Y}_m \in \mathbb{R}_+^{F\times T}$
Output: Denoised magnitude spectrogram $\mathbf{S} \in \mathbb{R}_+^{F\times T}$
Assumption:
The following values are given in advance:
A) Possible maximum rank of noise spectrogram $K \in \mathbb{N}$
B) Hyperparameters $\alpha^w \in \mathbb{R}_+$, $\beta^w \in \mathbb{R}_+$, $\alpha^h \in \mathbb{R}_+$, $\beta^h \in \mathbb{R}_+$, $\alpha^g \in \mathbb{R}_+$, and $\alpha^s \in \mathbb{R}_+$

time frame. This is considered to be reasonable because multiple speakers located at different directions may not talk simultaneously in disaster situations. Even when a few people speak simultaneously from the same direction, the overlapping speech sounds could be enhanced because those sounds still have sparse harmonic structures and the fine time-frequency fluctuations of speech spectrograms violate the low-rank assumption.

1) Overview: Bayesian RNTF approximates an input spectrogram at each channel $\mathbf{Y}_m \in \mathbb{R}_+^{F\times T}$ as the sum of channel-wise low-rank spectrogram and channel-wise sparse spectrogram $\mathbf{S}'_m \in \mathbb{R}_+^{F\times T}$:

$$\mathbf{Y}_m \approx \mathbf{W}_m\mathbf{H}_m + \mathbf{S}'_m, \tag{13}$$

where $\mathbf{W}_m \in \mathbb{R}_+^{F\times K}$ and $\mathbf{H}_m \in \mathbb{R}_+^{K\times T}$ are channel-wise basis and activation matrices for the low-rank spectrogram, respectively.

The relationship between the target speech signal $\mathbf{S} \in \mathbb{R}_+^{F\times T}$ and its observation at each microphone $\mathbf{S}'_m$ is assumed to be a time-variant and frequency-invariant linear system:

$$s'_{mft} \approx g_{mt}s_{ft}, \tag{14}$$

where $g_{mt} \in \mathbb{R}_+$ represents a gain of the target speech signal at microphone m and time t. According to (13) and (14), Bayesian RNTF decomposes the input spectrogram $\mathbf{Y}_m$ into the following four components:

$$y_{mft} \approx \sum_k w_{mfk}h_{mkt} + g_{mt}s_{ft}. \tag{15}$$

where $\boldsymbol{g}_m = [g_{m1}, \ldots, g_{mT}]$ is a gain vector. Although magnitude spectrograms are insensitive to relatively small motions [16], the gain g_{mt} depends on the motion of microphones and target speech. The gain g_{mt} is, therefore, independently estimated at each time frame to deal with the movement of microphones and sources.

2) Likelihood Function and Prior Distributions: The likelihood function and prior distributions except for those on the gain parameters g_{mt} are formulated in the same manner as in Bayesian RNMF ((8)–(12)). A gamma prior is put on g_{mt} assuming that its mean is 1:

$$p(g_{mt}|\alpha^g) = \mathcal{G}(g_{mt}|\alpha^g, \alpha^g), \tag{16}$$

where $\alpha^g \in \mathbb{R}_+$ is a hyperparameter controlling the variance of the gain parameters. The complete graphical model is shown in Fig. 4(b).

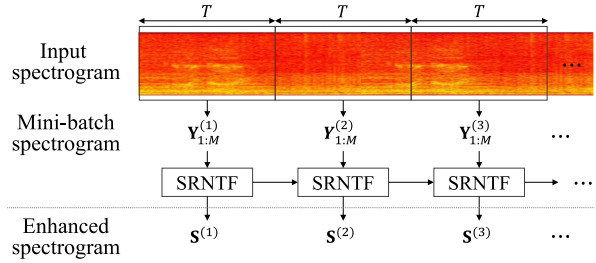


Fig. 5. Mini-batch processing flow of Bayesian SRNTF.

C. Bayesian Streaming RNTF for Real-Time Enhancement

This section describes the Bayesian SRNTF. It is formulated as a state-space model representing the latent variable as time-varying latent variables.

1) Overview: Bayesian SRNTF sequentially enhances target speech for T frames of mini-batch audio inputs (Fig. 5). The problem in this section is defined as follows:

Input:
1. M-channel magnitude spectrograms $\mathbf{Y}_m^{(n)} \in \mathbb{R}_+^{F\times T}$
2. Posterior distribution at the previous $(n-1)$ mini-batch.
Assumption:
The following values are given in advance:
A) Possible maximum rank of noise spectrogram $K \in \mathbb{N}$
B) Hyperparameters $\alpha^w \in \mathbb{R}_+$, $\beta^w \in \mathbb{R}_+$, $\alpha^h \in \mathbb{R}_+$, $\beta^h \in \mathbb{R}_+$, $\alpha^g \in \mathbb{R}_+$, $\alpha^s \in \mathbb{R}_+$, and $\gamma \in \mathbb{R}_+$

where n indicates the mini-batch index ($n = 1, 2, 3, \ldots$). As explained below, the posterior distribution at the prevous mini-batch is used for the prior information of the current latent variables. Interpretations of the hyperparameters are explained below.

Bayesian SRNTF decomposes the mini-batch audio spectrogram $y_{mft}^{(n)}$ into low-rank and sparse components in the same manner as in Bayesian RNTF:

$$y_{mft}^{(n)} \approx \sum_k w_{mfk}^{(n)}h_{mkt}^{(n)} + g_{mt}^{(n)}s_{ft}^{(n)}. \tag{17}$$

where $\mathbf{W}_m^{(n)} \in \mathbb{R}_+^{F\times K}$, $\mathbf{H}_m^{(n)} \in \mathbb{R}_+^{K\times T}$, $\boldsymbol{g}_m^{(n)} \in \mathbb{R}_+^{1\times T}$, and $\mathbf{S}^{(n)} \in \mathbb{R}_+^{F\times T}$ are the latent variables for the basis and activation matrices, gain, and sparse matrix notated in the same manner as in Bayesian RNTF, respectively. Let $\boldsymbol{\Theta}^{(n)}$ be a set of all the latent variables at the n-th mini-batch $\{\mathbf{W}_{1:M}^{(n)}, \mathbf{H}_{1:M}^{(n)}, \boldsymbol{g}_{1:M}^{(n)}, \mathbf{S}^{(n)}, \boldsymbol{\beta}^{s(n)}\}$. We formulate an observation model $p(\mathbf{Y}_{1:M}^{(n)}|\boldsymbol{\Theta}^{(n)})$ and a state update model $p(\boldsymbol{\Theta}^{(n)}|\boldsymbol{\Theta}^{(n-1)})$ for a state-space model (Fig. 6) that represents the relationship between the observation and latent variables and the dynamics of the latent variables.

2) Observation Model: The observation model of Bayesian SRNTF $p(\mathbf{Y}_{1:M}^{(n)}|\boldsymbol{\Theta}^{(n)})$ is formulated with a Poisson distribution

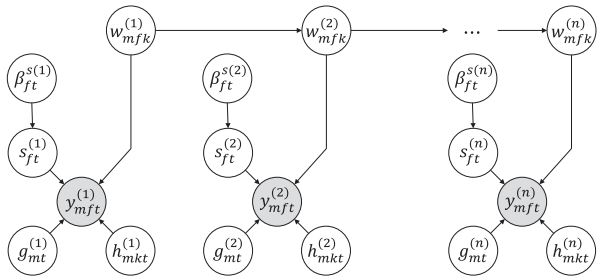


Fig. 6. Graphical model for Bayesian SRNTF.

in the same manner as in Bayesian RNTF:

$$p\left(\mathbf{Y}_{1:m}^{(n)}\middle|\boldsymbol{\Theta}^{(n)}\right) = \prod_{m,f,t}\mathcal{P}\left(y_{mft}^{(n)}\middle|\sum_{k}w_{mfk}^{(n)}h_{mkt}^{(n)}+g_{mt}^{(n)}s_{ft}^{(n)}\right). \quad (18)$$

3) State Update Model: Since the latent variables for the sparse component ($\boldsymbol{g}_{1:M}^{(n)}$, $\mathbf{S}^{(n)}$, and $\boldsymbol{\beta}^{s(n)}$) and the activation matrix for the low-rank component ($\mathbf{H}_{1:M}^{(n)}$) are time-independent, only the basis matrix $\mathbf{W}_{m}^{(n)}$ depends on the previous state $\mathbf{W}_{m}^{(n-1)}$ in our state update model:

$$p\left(\boldsymbol{\Theta}^{(n)}\middle|\boldsymbol{\Theta}^{(n-1)}\right) = p\left(\mathbf{W}_{1:M}^{(n)}\middle|\mathbf{W}_{1:M}^{(n-1)}\right)p\left(\mathbf{H}_{1:M}^{(n)}\right) \times p\left(\boldsymbol{g}_{1:M}^{(n)}\right)p\left(\mathbf{S}^{(n)}\right)p\left(\boldsymbol{\beta}^{s(n)}\right). \quad (19)$$

The priors for $\mathbf{H}_{m}^{(n)}$, $\boldsymbol{g}_{m}^{(n)}$, $\mathbf{S}^{(n)}$, and $\boldsymbol{\beta}^{s(n)}$ are formulated in the same way as in the batch Bayesian RNTF (Section III-B-2).

In this study, the state update model for $\mathbf{W}_{1:M}^{(n)}$ is independently formulated on each of its elements $w_{mfk}^{(n)}$:

$$p\left(\mathbf{W}_{m}^{(n)}\middle|\mathbf{W}_{m}^{(n-1)}\right) = \prod_{m,f,k}p\left(w_{mfk}^{(n)}\middle|w_{mfk}^{(n-1)}\right). \quad (20)$$

The state update model $p(w_{mfk}^{(n)}|w_{mfk}^{(n-1)})$ represents how $w_{mfk}^{(n)}$ varies from the previous state $w_{mfk}^{(n-1)}$. It has the following properties. The mean of $w_{mfk}^{(n)}$ should not be changed from that of $w_{mfk}^{(n-1)}$ because we assume no bias on the update. The variance of $w_{mfk}^{(n)}$, on the other hand, should be increased from that of $w_{mfk}^{(n-1)}$ because its uncertainty increases over time. As proposed in [38], such an update model can be formulated with a multiplicative process noise $v_{mfk}^{(n)}\in\mathbb{R}_{+}$ as follows:

$$w_{mfk}^{(n)} = v_{mfk}^{(n)}w_{mfk}^{(n-1)}. \quad (21)$$

A beta prior distribution is put on $v_{mfk}^{(n)}$ as follows:

$$p\left(v_{mfk}^{(n)}|\alpha_{mfk}^{(n-1)},\gamma\right) = \mathcal{B}\left(v_{mfk}^{(n)}\gamma\middle|\gamma\alpha_{mfk}^{(n-1)},(1-\gamma)\alpha_{mfk}^{(n-1)}\right), \quad (22)$$

where $\mathcal{B}(\alpha,\beta)$ represents a beta distribution with two shape parameters α and β, and $\gamma\in\mathbb{R}_{+}$ is a rate parameter controlling the variance of $w_{mfk}^{(n)}$. From (21) and (22), the update model $p(w_{mfk}^{(n)}|w_{mfk}^{(n-1)})$ can be derived as follows:

$$p\left(w_{mfk}^{(n)}\middle|w_{mfk}^{(n-1)}\right) = \mathcal{B}\left(\gamma\frac{w_{mfk}^{(n)}}{w_{mfk}^{(n-1)}}\middle|\gamma\hat{\alpha}_{mfk}^{(n-1)},(1-\gamma)\hat{\beta}_{mfk}^{(n-1)}\right). \quad (23)$$

As shown later (Section IV), the posterior $p(w_{mfk}^{(n-1)}|\mathbf{Y}^{(1:n-1)})$ is a gamma distribution $\mathcal{G}(w_{mfk}^{(n-1)}|\hat{\alpha}_{mfk}^{(n-1)},\hat{\beta}_{mfk}^{(n-1)})$ with a shape parameter $\hat{\alpha}_{mfk}^{(n-1)}\in\mathbb{R}_{+}$ and a rate parameter $\hat{\beta}_{mfk}^{(n-1)}\in\mathbb{R}_{+}$. As proven in [38], the predictive distribution $p(w_{mfk}^{(n)}|\mathbf{Y}^{(1:n-1)})$ is calculated from $p(w_{mfk}^{(n-1)}|\mathbf{Y}^{(1:n-1)})$ as follows:

$$\begin{aligned}p(w_{mfk}^{(n)}|\mathbf{Y}^{(1:n-1)}) &= \int p\left(w_{mfk}^{(n)}|w_{mfk}^{(n-1)}\right) \times p\left(w_{mfk}^{(n-1)}|\mathbf{Y}^{(1:n-1)}\right)dw_{mfk}^{(n-1)} \\ &= \mathcal{G}\left(\gamma\hat{\alpha}_{mfk}^{(n-1)},\gamma\hat{\beta}_{mfk}^{(n-1)}\right). \quad (24)\end{aligned}$$

Note that the mean of this distribution is the same as that of the one in the previous state and its variance is γ^{-1} times larger than that of the one in the previous state.

IV. Speech Enhancement Based on Bayesian Robust Non-Negative Tensor Factorization

This section derives Bayesian inferences of the proposed Bayesian RNMF, RNTF, and SRNTF, and describes the processing flow of speech enhancement based on Bayesian RNTF. The inferences of these models are derived with the VB framework. In summary, the enhancement methods we propose are VB-RNMF, VB-RNTF, and VB-SRNTF collectively.

A. Variational Inference

Our goal is to calculate the full posterior distributions of the proposed models. Since the true posterior is analytically intractable, we approximate it by using a VB algorithm [29], [36]. This section describes the main update rules. The detailed derivations of them are summarized in Appendix.

1) VB-RNMF: The target full posterior distribution of Bayesian RNMF is $p(\mathbf{W},\mathbf{H},\mathbf{S},\boldsymbol{\beta}|\mathbf{Y})$. Let $\boldsymbol{\Theta}$ be a set of all the parameters and $q(x)$ be a variational posterior distribution of x. Then, the true posterior distribution is approximated as follows:

$$p(\boldsymbol{\Theta}|\mathbf{Y})\approx q(\mathbf{W})q(\mathbf{H})q(\mathbf{S})q(\boldsymbol{\beta}^{s}). \quad (25)$$

The VB algorithm estimates the parameters of each variational distribution by minimizing the KL divergence between the true and approximated distributions.

Each variational posterior distribution is alternately and iteratively updated by fixing the other distributions as follows:

$$q(w_{fk}) = \mathcal{G}\left(\alpha^w + \sum_t y_{ft}\phi_{ftk}, \beta^w + \sum_t \langle h_{kt}\rangle\right), \quad (26)$$

$$q(h_{kt}) = \mathcal{G}\left(\alpha^h + \sum_f y_{ft}\phi_{ftk}, \beta^h + \sum_f \langle w_{fk}\rangle\right), \quad (27)$$

$$q(s_{ft}) = \mathcal{G}\left(\alpha^s + y_{ft}\psi_{ft}, \left\langle \beta^s_{ft}\right\rangle + 1\right), \quad (28)$$

$$q(\beta^s_{ft}) = \mathcal{G}(\alpha^s, \langle s_{ft}\rangle), \quad (29)$$

$$\phi_{ftk} = \frac{\mathbb{G}[w_{fk}]\mathbb{G}[h_{kt}]}{\sum_k \mathbb{G}[w_{fk}]\mathbb{G}[w_{kt}] + \mathbb{G}[s_{ft}]}, \quad (30)$$

$$\psi_{ft} = \frac{\mathbb{G}[s_{ft}]}{\sum_k \mathbb{G}[w_{fk}]\mathbb{G}[h_{kt}] + \mathbb{G}[s_{ft}]}, \quad (31)$$

where ϕ_{mftk} and ψ_{mft} are auxiliary variables and $\mathbb{G}[x] = \exp(\langle \log x\rangle)$ represents the geometric expectation.

2) VB-RNTF: The target full posterior distribution of Bayesian RNTF, $p(\mathbf{W}_{1:m}, \mathbf{H}_{1:m}, \boldsymbol{g}_{1:m}, \mathbf{S}, \boldsymbol{\beta}|\mathbf{Y}_{1:m})$, is approximated in the same manner as that of Bayesian RNMF. The true posterior distribution is approximated as:

$$p(\boldsymbol{\Theta}|\mathbf{Y}_{1:M}) \approx \left\{\prod_m q(\mathbf{W}_m)q(\mathbf{H}_m)q(\boldsymbol{g}_m)\right\} q(\mathbf{S})q(\boldsymbol{\beta}^s). \quad (32)$$

The variational posterior distributions are calculated in the same way as in Bayesian RNMF, and each of them is alternately and iteratively updated as follows:

$$q(w_{mfk}) = \mathcal{G}\left(\alpha^w + \sum_t y_{mft}\phi_{mftk}, \beta^w + \sum_t \langle h_{mkt}\rangle\right), \quad (33)$$

$$q(h_{mkt}) = \mathcal{G}\left(\alpha^h + \sum_f y_{mft}\phi_{mftk}, \beta^h + \sum_f \langle w_{mfk}\rangle\right), \quad (34)$$

$$q(g_{mt}) = \mathcal{G}\left(\alpha^g + \sum_f y_{mft}\psi_{mft}, \alpha^g + \sum_f \langle s_{ft}\rangle\right), \quad (35)$$

$$q(s_{ft}) = \mathcal{G}\left(\alpha^s + \sum_m y_{mft}\psi_{mft}, \left\langle \beta^s_{ft}\right\rangle + \sum_m \langle g_{mt}\rangle\right), \quad (36)$$

$$q(\beta^s_{ft}) = \mathcal{G}(\alpha^s, \langle s_{ft}\rangle), \quad (37)$$

$$\phi_{mftk} = \frac{\mathbb{G}[w_{mfk}]\mathbb{G}[h_{mkt}]}{\sum_k \mathbb{G}[w_{mfk}]\mathbb{G}[h_{mkt}] + \mathbb{G}[g_{mt}]\mathbb{G}[s_{ft}]}, \quad (38)$$

$$\psi_{mft} = \frac{\mathbb{G}[g_{mt}]\mathbb{G}[s_{ft}]}{\sum_k \mathbb{G}[h_{mfk}]\mathbb{G}[h_{mkt}] + \mathbb{G}[g_{mt}]\mathbb{G}[s_{ft}]}. \quad (39)$$

where ϕ_{mftk} and ψ_{mft} are auxiliary variables.

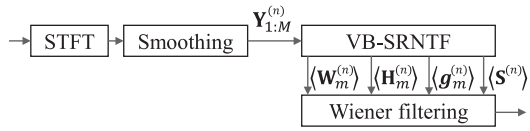


Fig. 7. Processing flow of the proposed speech enhancement.

3) VB-SRNTF: VB-SRNTF estimates the current posterior distribution recurrently in prediction and correction steps. The prediction step calculates $p(\boldsymbol{\Theta}^{(n)}|\mathbf{Y}^{(1:n-1)})$ from the previous posterior distribution $p(\boldsymbol{\Theta}^{(n-1)}|\mathbf{Y}^{(1:n-1)})$. The correction step, on the other hand, estimates the current posterior distribution $p(\boldsymbol{\Theta}^{(n)}|\mathbf{Y}^{(1:n)})$ from the observation $\mathbf{Y}^{(n)}$ and the predictive distribution $p(\boldsymbol{\Theta}^{(n)}|\mathbf{Y}^{(1:n-1)})$. In the prediction step, the following predictive distribution is calculated:

$$p\left(\boldsymbol{\Theta}^{(n)}\middle|\mathbf{Y}^{(1:n-1)}\right) = \prod_{m,f,k} \mathcal{G}\left(w^{(n)}_{mfk}\middle|\gamma\hat{\alpha}^{(n-1)}_{mfk}, \gamma\hat{\beta}^{(n-1)}_{mfk}\right) \times p\left(\mathbf{H}^{(n)}\right) p\left(\boldsymbol{g}^{(n)}_m\right) p\left(\mathbf{S}^{(n)}\right) p\left(\boldsymbol{\beta}^{s(n)}\right), \quad (40)$$

where $\hat{\alpha}^{(n-1)}_{mfk}$ and $\hat{\beta}^{(n-1)}_{mfk}$ are the shape and rate parameters of the gamma distribution $p(w^{(n-1)}_{mfk}|\mathbf{Y}^{(1:n-1)})$, respectively. In the correction step, the current posterior distribution is estimated in the same manner as in (33)–(39) by replacing the prior distribution of Bayesian RNTF with the predictive distribution. For the initial correction step ($n = 1$), we use the prior distribution of Bayesian RNTF (Section III-B-2) as the predictive distribution.

B. Speech Enhancement Based on VB-SRNTF

Fig. 7 shows the overall processing flow for the speech enhancement using VB-SRNTF. Our framework first takes the STFT of each microphone recording and obtains a multichannel magnitude spectrogram. Since each noisy input magnitude spectrogram includes fine fluctuations, the input spectrogram is smoothed for stable low-rank and sparse decomposition. Let $\mathbf{Y}'^{(n)}_m \in \mathbb{R}^{F\times T}_+$ be the raw magnitude spectrogram obtained with the STFT, this smoothing pre-processing is conducted as follows:

$$y^{(n)}_{mft} = \frac{1}{9}\sum_{f'=f-1}^{f+1}\sum_{t'=t-1}^{t+1} y'^{(n)}_{mf't'}. \quad (41)$$

After conducting VB-SRNTF, our framework reconstructs the target speech signal at each microphone with Wiener filtering because VB-SRNTF cannot estimate the absolute scale of the target signal. Finally, the time-domain output signal is obtained by taking the inverse STFT of the selected spectrogram.

V. Experimental Evaluation With Simulated Data

To analyze the performance of the proposed enhancement methods, and compare them with existing methods, we first evaluated them using simulated audio signals.

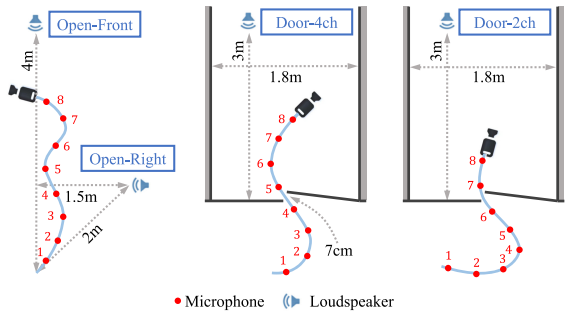


Fig. 8. Four conditions of robot and loudspeaker in experimental evaluation.

TABLE I
CONFIGURATIONS AND RESULTS OF BAYESIAN OPTIMIZATION

Parameters		α^w	α^h	α^g	α^s	γ	K
Search range	min	0.01	0.01	0.01	0.01	0.01	1
	max	1.0	1.0	10.0	2.0	1.0	10
VB-RNMF		0.71	0.19	–	0.53	–	7
VB-RNTF		0.30	0.35	6.3	2.0	–	6
VB-SRNTF (T=200)		0.88	0.38	7.8	1.6	0.76	5

A. Common Experimental Conditions

As shown in Fig. 2, the body of the hose shaped robot used in this evaluation was made from a corrugated tube of 38 mm in diameter and 3 m long. The entire surface of the robot was covered by cilia and seven vibrators used for moving forward by vibrating the cilia. This robot had an 8-ch synchronized microphone array whose microphones were distributed on its body at 40-cm intervals. The audio signals of these microphones were captured at 16 kHz and with 24-bit sampling.

The input signals were generated by mixing target speech and ego-noise signals at signal-to-noise ratios (SNRs) varying from -20 dB to $+5$ dB. As shown in Fig. 8, there were four conditions differing in the relative positions of the robot and the loudspeaker (target speech).

1) *Open-Front:* The robot was in an experimental room with no obstacles. The loudspeaker was in front of the robot. The reverberation time (RT_{60}) of the room was 750 ms.
2) *Open-Right:* Same as Open-Front except that the loudspeaker was to the right of the robot.
3) *Door-4ch:* The robot was caught by a door, the loudspeaker was in front of the robot, and four of the microphones were behind the door. The reverberation time was 990 ms.
4) *Door-2ch:* Same as Door-4ch except that six microphones were behind the door.

The ego-noise was recorded for 60 seconds under each condition while sliding the robot left and right by using vibrators and a hand. The loudspeaker was used for recording the impulse response. Multichannel speech signals were generated by convoluting clean speech signals and the impulse response, and then they were mixed with 20 seconds of the ego-noise recordings. The clean speech data consisted of 24 recordings of three male and three female speech, which were included in the JNAS phonetically balanced Japanese utterances database [39]. In this setting, we assume that the location of the target speech does not change as the target speech signal was generated with a single impulse response at each condition.

The enhancement performance was evaluated by using the source-to-distortion ratio (SDR) [40], [41]. The SDR measures the power ratio of the target speech and distortion component included in the output signals. Since VB-RNTF and VB-SRNTF outputs are obtained by applying Wiener filtering to one of the microphones, we simply evaluated the enhanced signals at the tip (8th) microphone.

The parameters for VB-RNMF, VB-RNTF, and VB-SRNTF were as follows. The shifting interval and window lengths of the STFT were set to 160 and 1024 samples, respectively. The hyperparameters α^w, α^h, α^g, α^s, γ and the number of bases K, which control the low-rankness and sparseness, were decided by using a Bayesian optimization method [42]. This method regards a target method as a black-box function that takes hyperparameters as input and outputs the value of average SDR. Assuming the function to follow a Gaussian process, the method searches for optimal hyperparameters that maximize the output of the function. We used noisy signals with SNRs of -10 dB and -5 dB and with the layout conditions of Open-Front and Door-4ch. Note that these signals were included in the data set of noisy signals used for the evaluation. We used 12 signals in the 24 noisy test signals at each SNR and layout condition. The search range and optimization results were summarized in Table I. The rate hyperparameters β^w and β^h were set to α^w and $\alpha^h K$, respectively. In this paper we iterated VB-RNMF and VB-RNTF 200 times and iterated VB-SRNTF 100 times. The latent variables were initialized randomly.

B. Evaluation of Batch VB-RNTF and VB-RNMF

VB-RNTF and VB-RNMF were compared with existing phase-based blind source separation methods [7], [13] and low-rank and sparse decomposition methods [8], [20], [29], [43]. The phase-based blind source separation methods we compared were MNMF [13] and independent vector analysis (IVA) [7]. The number of sources was set to eight for MNMF and IVA because seven vibrators generated noise and one target speech existed. This value corresponds to the maximum value tractable in these methods because they cannot perform under-determined source separation. Since these methods cannot distinguish the target speech source and other noise sources, the performance was determined by taking a maximum SDR value from all eight separation results. The low-rank and sparse decomposition methods we compared were conventional RPCA [20], [23], RNMF [8] and VB-RPCA [29]. The results of them were obtained by using the tip (8th) microphone signals. We also evaluated extended RPCA results that were obtained by taking median values of all the microphone results (Med-RPCA) [43], [44]. As a baseline, we evaluated an adaptive spectral subtraction (SS) method [45] that is applied to the tip microphone signals.

As shown in Fig. 9, in the Open-Front and Open-Right conditions, the proposed VB-RNTF performed the best of all the evaluated methods. Fig. 9(a) shows the performances for all the

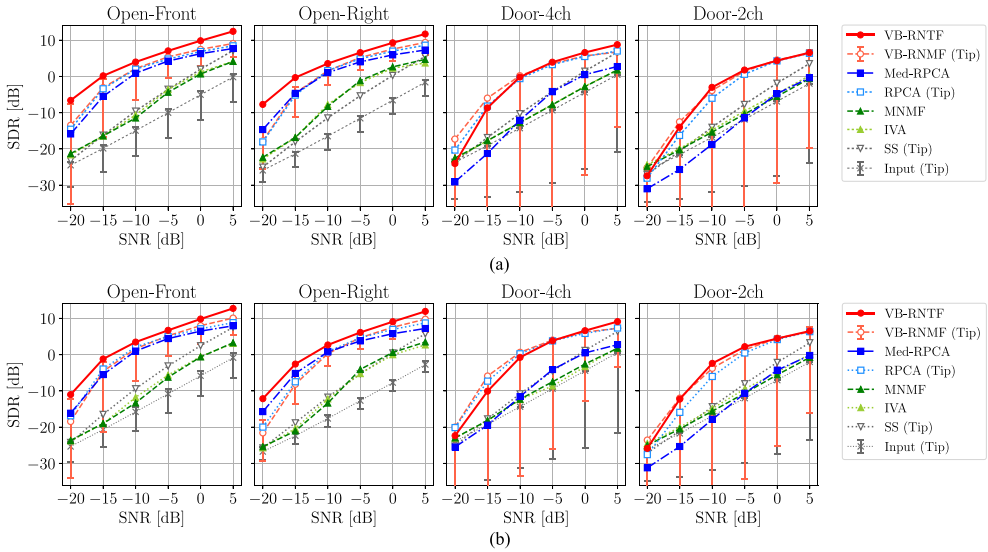


Fig. 9. Speech enhancement performances of VB-RNTF, VB-RNMF, and existing methods. Each lines indicates average SDR at the specified condition. Error bars for VB-RNMF and the input signal span the maximum and minimum SDRs in all the microphones. (a) Average SDRs for all the noisy test signals. (b) Average SDRs for the different set of noisy signals that were not used for the parameter optimization.

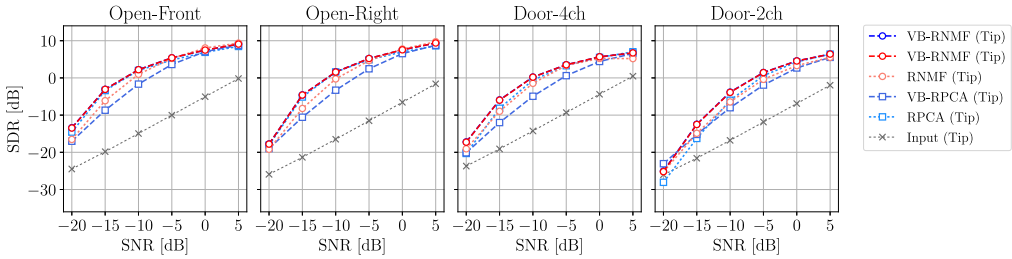


Fig. 10. Speech enhancement performances of VB-RNMF and existing low-rank and sparse decomposition methods. The SDR of the input signal (gray line) is that of the recordings of the tip microphone.

noisy signals. Fig. 9(b) shows those for a different set of noisy signals whose ego-noise and speech signals were not used for the parameter optimization. Compared with Fig. 9(a), Fig. 9(b) shows that our VB-RNTF and VB-RNMF can generalize to new data that were not used for the parameter optimization. The low-rank and sparse decomposition methods (VB-RNTF, VB-RNMF, RPCA, and Med-RPCA) significantly outperformed conventional phase-based methods (MNMF and IVA). In the Door-4ch and Door-2ch conditions where some of the microphones were shaded, Med-RPCA significantly degraded from the Open-Front and Open-Right conditions. VB-RNTF, on the other hand, outperformed other multichannel methods. Although VB-RNTF was also degraded in both of the Door conditions, its performance was comparable to those of single-channel VB-RNMF and RPCA in these condition except when the SNR was less than -10 dB.

The performances of single-channel VB-RNMF and the existing low-rank and sparse decomposition methods are compared in Fig. 10; where we see that VB-RNMF was comparable to the existing methods. This shows that VB-RNMF provides extensibility of RNMF in a Bayesian manner without performance degradation.

Fig. 11 illustrates excerpts of enhancement results by VB-RNTF and VB-RNMF in the Door-4ch condition. While the

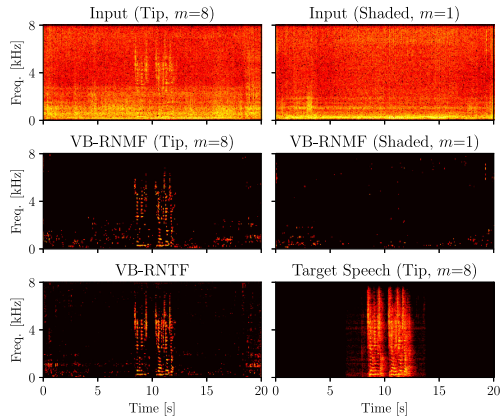


Fig. 11. Excerpts of enhancement results obtained by VB-RNTF and VB-RNMF when the robot layout was the Door-4ch condition and the SNR was -5 dB. VB-RNMF results on both the tip (8th) microphone and shaded (1st) microphone signals are shown.

VB-RNMF result applied to the tip (8th) microphone successfully enhanced the target speech, the result on the shaded (1st) microphone failed due to the low-SNR input. On the other hand, VB-RNTF using all the microphones robustly enhanced speech in this condition. Fig. 12 shows time-varying gains of the sparse spectrogram at each microphone g_{mt} estimated

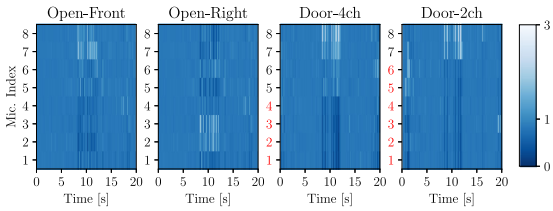


Fig. 12. Examples of time-varying gain of sparse spectrogram g_{mt} obtained by VB-RNTF at each microphone when SNR was -5 dB. Female speech was emitted between 8 s and 12 s. Microphones shaded in Door-4ch and Door-2ch conditions are highlighted in red.

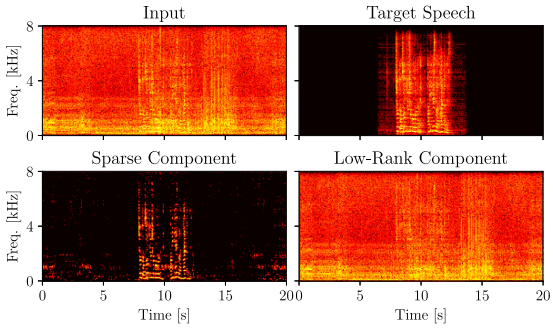


Fig. 13. Excerpts of estimated sparse and low-rank components ($m = 8$) obtained by VB-RNTF when the robot layout was the Door-4ch condition and the SNR was 0 dB.

by VB-RNTF. In the Door-4ch and Door-2ch conditions, the gains of the microphones that were separated from the sound source (highlighted in red) got significantly smaller. This shows that the estimated gain can be used as a reliability of each microphone.

VB-RNTF outperformed the existing methods under high reverberation ($\mathrm{RT}_{60} \geq 750$ ms). Fig. 13 shows that our method estimated the reverberant speech as the sparse component and separated a part of late reverberations into the low-rank component. This result shows that VB-RNTF deals with the reverberations by estimating the most prominent reverberant speech as a speech signal and separating other residuals into the low-rank components.

C. Evaluation of Mini-Batch VB-SRNTF

The performance of VB-SRNTF was evaluated with various mini-batch sizes. We tested the mini-batch sizes T of 300, 200, 100, 50, and 10 frames. VB-SRNTF was compared with batch VB-RNTF and the following two existing mini-batch inferences: Ind-VB-RNTF and SVI-RNTF. Ind-VB-RNTF simply and independently conducts VB-RNTF at each mini-batch observation. SVI-RNTF is based on the conventional mini-batch VB inference [46] of VB-RNTF. It corresponds to VB-SRNTF whose γ is set to 1.0. We also compared VB-SRNTF with a variant of VB-SRNTF (VB-SRNTF-Raw) that takes a raw magnitude spectrogram without smoothing as input.

As shown in Fig. 14, the enhancement performance became higher as the mini-batch size was increased except when the SNR was -15 dB or less. Fig. 14(b) shows that VB-SRNTF is also robust against the new data that were not used for the parameter optimization. When the mini-batch size was 200 frames or more, the SDR performances tended to be saturated. On the other hand, when the mini-batch size was 10 frames, the SDR performances were significantly degraded. Since a large mini-batch size leads to a large latency, there is a trade-off between performance and latency. These results show that a 2.0 -second mini-batch ($T = 200$) was needed for adequate performance.

Fig. 15 compares the proposed VB-SRNTF results and other mini-batch inference results. Compared with Ind-VB-RNTF, which did not consider the relationship between adjacent mini-batches, VB-SRNTF improved SDRs in the Door-4ch and Door-2ch conditions. Compared with SVI-RNTF, which did not consider the process noise of the basis vectors, the proposed VB-SRNTF improved SDRs when the SNR was less than -5 dB in the Open-Front and Open-Right conditions. Compared with VB-SRNTF-Raw, which did not smooth the input spectrogram, the proposed VB-SRNTF improved SDRs in all the conditions.

Fig. 16 shows the performances of VB-SRNTF with different values of K. The SDR degradation in the Open-Front and Open-Right conditions was less than 1.0 dB when K was 5 or more and 8 or less. On the other hand, K has to be set to 5 or 6 in the Door-2ch condition and only 5 in Door-4ch condition. The performance of our method was sensitive to the number of bases K when some of the microphones were shaded. We confirmed that the SDR degradation was less than 1 dB even when the hyperparameters α^w, α^h, and α^g were changed by 20% from the values of Table I. The α^s and γ had robustness against 10% and 5% changes, respectively.

D. Investigation of Gain Parameter Modeling

The gain parameter $g_{mt}^{(n)}$ of VB-SRNTF (14) ignores a frequency dependency and a temporal continuity. Since our formulation has only one gain parameter shared by frequency bins, the frequency characteristics are not considered. Our model also does not take into account the temporal continuity of the gains as shown in Fig. 12.

We investigated the gain parameter modeling by evaluating variants of VB-SRNTF with frequency-dependent gains and temporal-continuous gains. The following two variants with the frequency-dependent gains were evaluated:

1) VB-SRNTF-$g_{mft}^{(n)}$: The gain $g_{mft}^{(n)}$ is both frequency and time dependent.
2) VB-SRNTF-$g_{mf}^{(n)}$: The gain $g_{mf}^{(n)}$ is frequency dependent but time independent.

On the other hand, the following two variants with the temporal-continuous gains were evaluated:

3) VB-SRNTF-EV: The prior distribution of the current state is given with the expected value of the previous posterior distribution $\langle \boldsymbol{g}_m^{(n-1)} \rangle$ as follows:

$$p\left(g_{mt}^{(n)} \middle| \alpha^g, \left\langle \boldsymbol{g}_m^{(n-1)} \right\rangle\right) = \mathcal{G}\left(g_{mt}^{(n)} \middle| \alpha^g, \frac{\alpha^g T}{\sum_t \left\langle g_{mt}^{(n-1)} \right\rangle}\right), \tag{42}$$

where $\alpha^g \in \mathbb{R}_+$ is a hyperparameter that controls the strength of the dependencies.

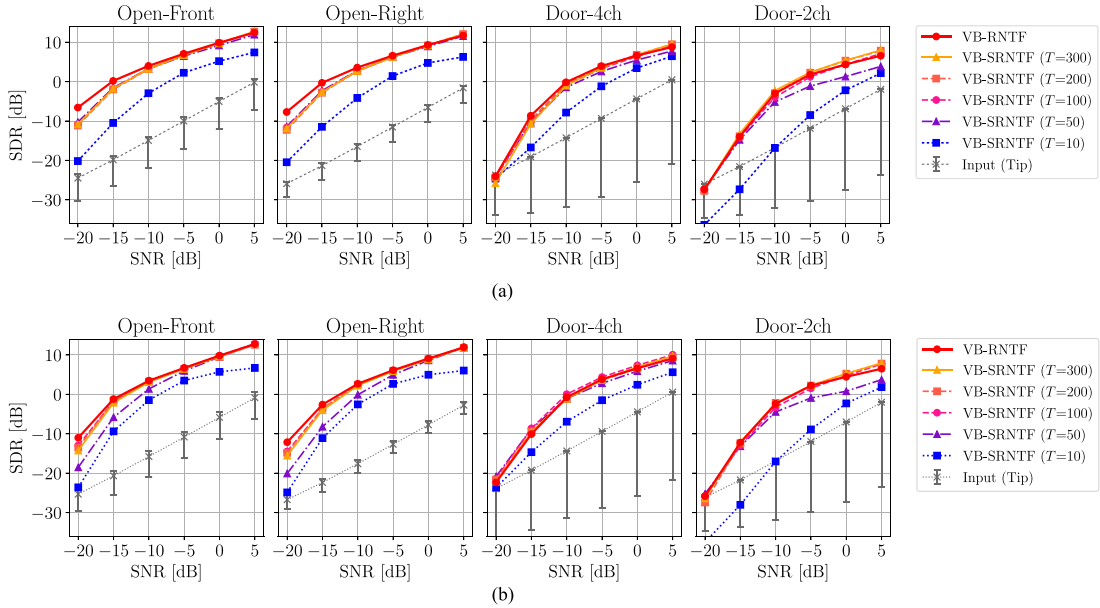


Fig. 14. SDR performances of VB-SRNTFs with different mini-batch sizes. (a) Average SDRs for all the noisy test signals (b) Average SDRs for the different set of noisy signals that were not used for the parameter optimization.

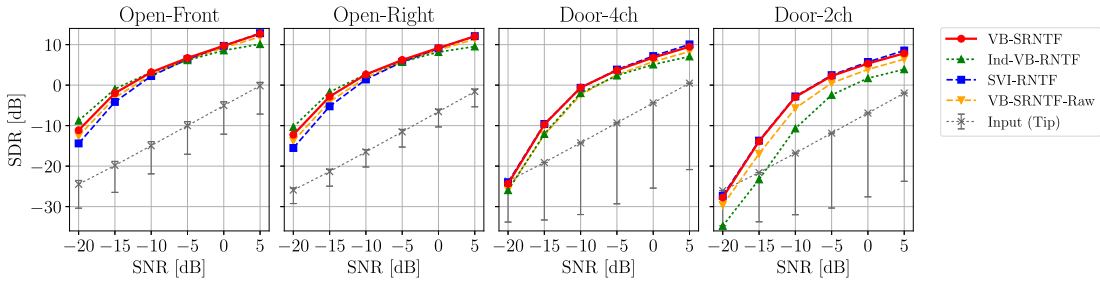


Fig. 15. Comparison of VB-SRNTF ($T = 200$) and existing mini-batch inferences of Bayesian RNTF.

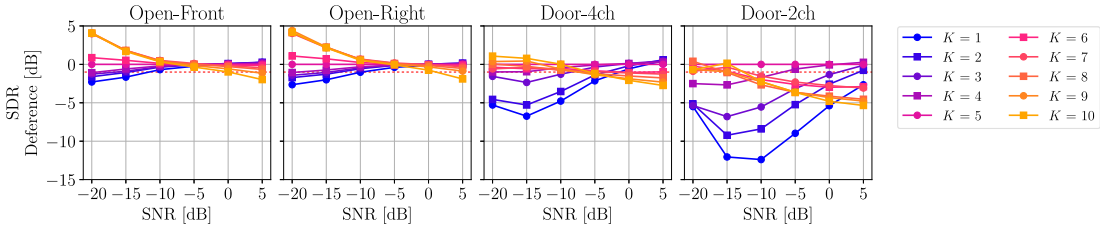


Fig. 16. VB-SRNTF performances with different values of K in SDR difference from that with the values in Table I.

4) VB-SRNTF-GMC: Markov dependencies between adjacent time frames are introduced with a gamma Markov chain prior [47] as follows:

$$p\left(g_{mt}^{(n)}\middle|\eta, z_{mt}^{(n)}\right) = \mathcal{G}\left(g_{mt}^{(n)}\middle|\eta, \eta z_{mt}^{(n)}\right), \quad (43)$$

$$p\left(z_{mt}^{(n)}\middle|\eta, g_{m(t-1)}^{(n)}\right) = \mathcal{G}\left(z_{mt}^{(n)}\middle|\eta, \eta g_{m(t-1)}^{(n)}\right), \quad (44)$$

$$p\left(z_{m1}^{(n)}\middle|\eta, g_{mT}^{(n-1)}\right) = \mathcal{G}\left(z_{m1}^{(n)}\middle|\eta, \eta g_{mT}^{(n-1)}\right), \quad (45)$$

where $z_{mt}^{(n)}$ is an auxiliary latent variable that makes Markov dependencies between $g_{mt}^{(n)}$ and $g_{m(t-1)}^{(n)}$ in a conjugate manner and $\eta \in \mathbb{R}_+$ is a hyperparameter that controls the strength of the dependencies.

The hyperparameters of these models were determined by using the Bayesian optimization method in the same way as in Section V-A.

Fig. 17 compares the SDR performances of the original VB-SRNTF and the variants of VB-SRNTF with the frequency-dependent gains and the temporal-continuous gains. The performance of the original VB-SRNTF was comparable to that of VB-SRNTF-$g_{mt}^{(n)}$ although small differences were found depending on experimental conditions. On the other hand, the SDR performance was degraded in the Door-2ch condition by VB-SRNTF-$g_{mft}^{(n)}$ whose gain has both frequency and time

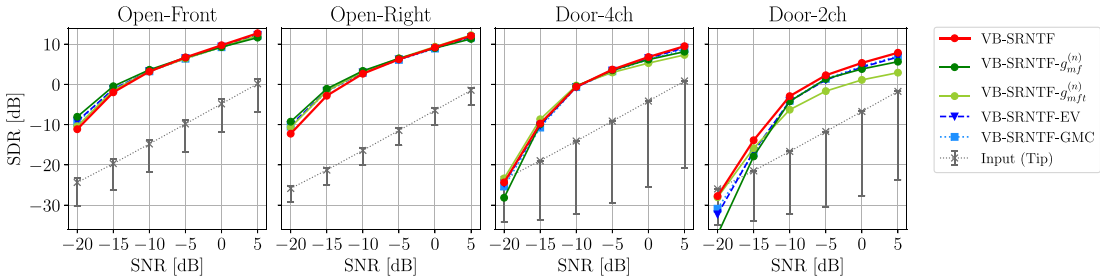


Fig. 17. Comparison of VB-SRNTF ($T = 200$) proposed in Section III-C and the variants of VB-SRNTF with the frequency-dependent gains and the temporal-continuous gains.

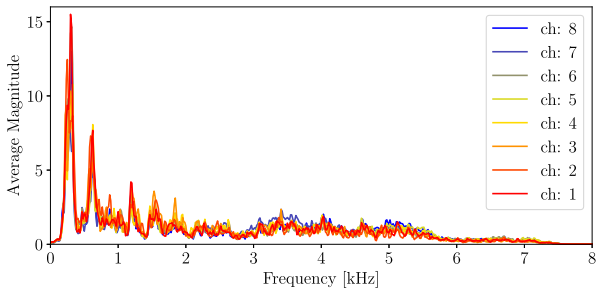


Fig. 18. Average magnitude spectrum of a speech spectrogram at each microphone. A female speech spectrum with the condition of Open-Front is shown. Note that scale differences over microphones are normalized.

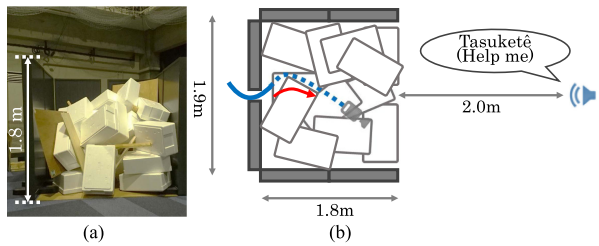


Fig. 19. Condition of rubble and target speech in experiments reported in Section VI. (a) Pile of rubble. (b) Condition of rubble and target speech.

dependencies. These results show that the performance did not deteriorate even if the frequency differences were ignored. We also see that the performance of VB-SRNTF was comparable to those of VB-SRNTF-EV and -GMC. Although the temporal continuity is one of the essential clues for blind source separation, the original VB-SRNTF is adequate in our task.

Although our gain model (14) ignores the differences of the magnitude spectral pattern across channels, the original VB-SRNTF was not degraded compared to VB-SRNTF-$g^{(n)}_{mft}$ and -$g^{(n)}_{mf}$. As shown in Fig. 18, this is because the differences were small enough to ignore them. The possible reason for the degradation of VB-SRNTF-$g^{(n)}_{mft}$ is VB-SRNTF-$g^{(n)}_{mft}$ has so many free parameters that it is difficult to estimate them properly. The performance of VB-SRNTF-$g^{(n)}_{mf}$, on the other hand, was comparable to VB-SRNTF because the speech source locations were stable over time in this evaluation. VB-SRNTF has an advantage over VB-SRNTF-$g^{(n)}_{mf}$ in the sense that it can deal with moving sound sources.

VI. Experiments With Recorded Data

This section reports experimental results obtained using data recorded in an environment with simulated rubble.

A. Experimental Conditions

We evaluated VB-RNTF and VB-SRNTF in the condition that the robot moved under simulated rubble. To simulate rubble disturbing sound propagation, styrene foam boxes and wooden plates were piled up (Fig. 19(a)). A loudspeaker for playing back target speech signals was put 2 m away from this rubble (Fig. 19(b)). The target signals were four male and female speech recordings screaming for rescue in Japanese (e.g., "Tasukete kudasai (Help me)" and "Kokoniimasu (I'm here)") and the loudspeaker was calibrated so that its sound pressure level for each utterance was 80 dB. The robot was inserted from behind the rubble and captured eight-channel audio signals (mixtures of ego-noise and each target speech) for 10 seconds during the insertion. In this experiment, the relative layout of the microphones and target speech source changed over time due to the insertion and vibration. The parameters of the proposed VB-RNMF, VB-RNTF, and VB-SRNTF were the same values as in Section V.

Since it was impossible to obtain clean speech signals captured by the robot microphones, we used the following SNR as an evaluation criterion in this experiment:

$$\mathrm{SNR}(\hat{\mathbf{S}}, \mathbf{S}, a) = 10 \log_{10} \frac{\sum_{f,t} a^2 s_{ft}^2}{\sum_{f,t} (\hat{s}_{ft} - a s_{ft})^2}, \qquad (46)$$

where $\mathbf{S} \in \mathbb{R}_+^{F \times T}$ and $\hat{\mathbf{S}} \in \mathbb{R}_+^{F \times T}$ represent the magnitude spectrograms of reference and estimated target speech signals, respectively, and a represents a gain parameter compensating for the level difference between $\mathbf{S}$ and $\hat{\mathbf{S}}$. This gain parameter was determined with minimum mean-square error estimation (MMSE) between $a\mathbf{S}$ and $\hat{\mathbf{S}}$.

B. Experimental Results

Fig. 20 shows that VB-RNTF and VB-SRNTF outperformed all of the other methods. VB-SRNTF improved the SNR by 1.0 dB more than VB-RNMF, which had the second-best performance. Fig. 21 shows the magnitude spectrogram of an observed signal (at the tip microphone) and the enhanced speech signals obtained by VB-RNTF and VB-SRNTF. These results showed that VB-RNTF and VB-SRNTF suppressed the time-varying ego-noise.

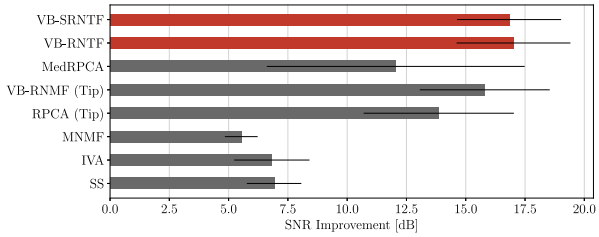


Fig. 20. Speech enhancement performances in terms of SNR improvement from the input signal (at the tip microphone). Error bars indicate the standard deviation of the results. The average SNR of the input signals was −19.7 dB.

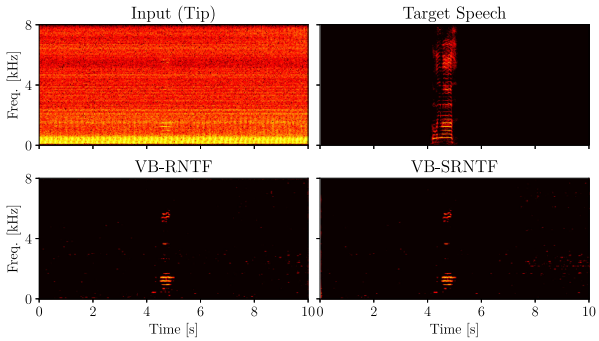


Fig. 21. Examples of enhancement results obtained in experiments reported in Section VI.

To realize a real-time mobile enhancement system, our enhancement system was implemented on an embedded GPGPU board (NVIDIA Jetson TX1) and a standard laptop computer (Dell XPS13). The proposed VB-SRNTF was implemented on the TX1 with GPGPU programming using C++ and CUDA 8.0. The elapsed time for VB-SRNTF with a 20.0 s input signal was 12.3 s when the batch size T was 200 frames. Since this value was small enough compared with the whole signal length, our method could work in real time.

VII. Conclusion

This paper presented a multichannel blind speech enhancement method based on low-rank and sparse decomposition. Our method is formulated as a Bayesian model called Bayesian RNTF. It separates a multichannel magnitude spectrogram into sparse and low-rank spectrograms (target speech and noise) without any prior training. Since Bayesian RNTF works without phase information, it can deal with the time-varying layout of microphones and sound sources. For real-time speech enhancement, Bayesian RNTF is extended to a state-space model called Bayesian SRNTF that represents the dynamics of the latent variables in a mini-batch manner. The Bayesian inferences of these models were derived with a VB framework, so the decomposition methods are abbreviated as VB-RNTF and VB-SRNTF.

VB-RNTF and VB-SRNTF are applied to speech enhancement for a microphone array distributed on a hose-shaped rescue robot. This speech enhancement needs to address three main problems: the environment-dependence of ego-noise, deformable layout of microphones, and partial occlusion of microphones. Since our method is based on the low-rank and sparse decomposition and time-varying mixing system, it is robust against the first two problems. In addition, it can deal with the shaded microphones because it estimates the speech level at each microphone. Experiments using a 3-m hose-shaped rescue robot with eight microphones showed that VB-SRNTF improves the SNR of a speech signal 1.03 dB more than conventional blind methods do. Using an embedded GPGPU board, we also confirmed that the proposed mini-batch VB-SRNTF was fast enough to work in real time.

Our methods based on the low-rank noise and sparse speech assumptions have the following limitations. Experimental results showed that the possible maximum rank of the low-rank component K should be given appropriately in advance. Our sparseness assumption, on the other hand, may cause our method to enhance not only speech signals but also other sparse noise signals. For example, if an input signal includes impact noise sounds caused by rubble-removal operations, our method extracts the noise as a speech signal. To relax the low-rank limitation, future work includes the estimation of K based on the non-parametric Bayesian framework [36]. We also plan to introduce speech-specific structures as prior information of the sparse component. Spectrograms of speech signals have dependencies between frequency bins (e.g. harmonic structures) and time frames (e.g. temporal continuity). A Bayesian RPCA model whose sparse component has time and frequency dependence [35] would be useful for this extension. In addition, for search-and-rescue activities, we will extend our method to localize a victim by using the estimated speech gains at microphones. VB-SRNTF also calculates a simple distribution of estimated speech gain at each microphone. It would be able to roughly estimate the location of a victim by using the gain differences across microphones.

Appendix
Derivations of VB Inference Algorithms

This appendix shows the derivations of VB inference algorithms for Bayesian RNMF, RNTF, and SRNTF. VB framework approximately estimates the analytically intractable posterior distribution of the target latent variables [29], [36].

We first show a brief description of the VB inference framework. Let $\mathbf{Y}$ be an observation variable, $\mathbf{Z}_i$ $(i = 1, \ldots, I)$ be parameters whose posterior distributions are estimated, and $\mathbf{\Theta} = \{\mathbf{Z}_1, \ldots, \mathbf{Z}_I\}$ be a set of all the parameters. Then, the true posterior distribution $p(\mathbf{\Theta}|\mathbf{Y})$ is approximated by the product of variational posterior distributions $q(\mathbf{Z}_i)$ as follows:

$$p(\mathbf{\Theta}|\mathbf{Y}) \approx \prod_i \mathbf{q}(\mathbf{Z}_i). \tag{47}$$

VB algorithm estimates the variational distributions $q(\mathbf{Z}_i)$ by maximizing the following lower bound $\mathcal{L}(q)$ of $\log p(\mathbf{\Theta})$:

$$\log p(\mathbf{Y}) = \log \int p(\mathbf{Y}, \mathbf{\Theta}) d\mathbf{\Theta} \tag{48}$$

$$\geq \int q(\mathbf{\Theta}) \log \frac{p(\mathbf{Y}, \mathbf{\Theta})}{q(\mathbf{\Theta})} d\mathbf{\Theta} \tag{49}$$

$$= \langle \log p(\mathbf{Y}, \mathbf{\Theta}) \rangle - \langle \log q(\mathbf{\Theta}) \rangle \stackrel{\text{def}}{=} \mathcal{L}(q), \tag{50}$$

where $\langle x \rangle$ is the expectation operation. This maximization corresponds to the minimization of the KL divergence between the true and approximated distributions. $\mathcal{L}(q)$ is maximized by alternately and iteratively updating each of $q(\mathbf{Z}_i)$ as follows:

$$q(\mathbf{Z}_i) = \exp \langle \log p(\mathbf{Y}, \mathbf{\Theta}) \rangle_{\mathbf{\Theta} \setminus \mathbf{Z}_i}, \tag{51}$$

where $\mathbf{\Theta} \setminus \mathbf{Z}_i$ represents a subset of $\mathbf{\Theta}$ obtained by removing $\mathbf{Z}_i$ from $\mathbf{\Theta}$.

A. VB-RNMF

The target full posterior distribution of Bayesian RNMF is $p(\mathbf{W}, \mathbf{H}, \mathbf{S}, \boldsymbol{\beta}[\mathbf{Y}])$. Let $\mathbf{\Theta}$ be a set of all the parameters, the true posterior distribution is approximated as follows:

$$p(\mathbf{\Theta}[\mathbf{Y}]) \approx q(\mathbf{W}) q(\mathbf{H}) q(\mathbf{S}) q(\boldsymbol{\beta}^s). \tag{52}$$

We take a lower bound of $\log p(\mathbf{Y})$ and estimate the variational posterior distributions by maximizing it.

Since the expectation of $\log p(\mathbf{Y}|\mathbf{W}, \mathbf{H}, \mathbf{S})$ includes the following intractable expectations:

$$\langle \log p(\mathbf{Y}|\mathbf{W}, \mathbf{H}, \mathbf{S}) \rangle = \sum_{f,t} y_{ft} \left\langle \log \left(\sum_k w_{fk} h_{kt} + s_{ft} \right) \right\rangle - \sum_{f,t,k} \langle w_{fk} h_{kt} \rangle - \sum_{f,t} \langle s_{ft} \rangle + \text{const.}, \tag{53}$$

this Poisson log-likelihood is lower-bounded by using Jensen's inequality [36]:

$$\begin{aligned} &\langle \log p(\mathbf{Y}|\mathbf{W}, \mathbf{H}, \mathbf{S}) \rangle \geq \\ &\sum_{f,t,k} y_{ft} \phi_{ftk} \left\langle \log \left(\frac{w_{fk} h_{kt}}{\phi_{ftk}} \right) \right\rangle + \sum_{f,t} y_{ft} \psi_{ft} \left\langle \log \left(\frac{s_{ft}}{\psi_{ft}} \right) \right\rangle \\ &\quad - \sum_{f,t,k} \langle w_{fk} h_{kt} \rangle - \sum_{f,t} \langle s_{ft} \rangle + \text{const.} \end{aligned} \tag{54}$$

$$\stackrel{\text{def}}{=} \langle p'(\mathbf{Y}|\mathbf{W}, \mathbf{H}, \mathbf{S}) \rangle \tag{55}$$

where $\phi_{ftk} \in \mathbb{R}_+$ and $\psi_{ft} \in \mathbb{R}_+$ ($\sum_k \phi_{ftk} + \psi_{ft} = 1$) are the auxiliary variables. Using $\langle p'(\mathbf{Y}|\mathbf{W}, \mathbf{H}, \mathbf{S}) \rangle$, $\log p(\mathbf{Y})$ is lower-bounded as follows:

$$\begin{aligned} \log p(\mathbf{Y}) \geq\ & \langle \log p'(\mathbf{Y}|\mathbf{W}, \mathbf{H}, \mathbf{S}) \rangle + \langle \log p(\mathbf{W}) \rangle + \langle \log p(\mathbf{H}) \rangle \\ & + \langle \log p(\mathbf{S}|\boldsymbol{\beta}^s) \rangle + \langle \log p(\boldsymbol{\beta}^s) \rangle \\ & - \langle \log q(\mathbf{W}) \rangle - \langle \log q(\mathbf{H}) \rangle \\ & - \langle \log q(\mathbf{S}) \rangle - \langle \log q(\boldsymbol{\beta}^s) \rangle \stackrel{\text{def}}{=} \mathcal{L}(q). \end{aligned} \tag{56}$$

The optimal ϕ_{ftk} and ψ_{ft} (30) and (31) are obtained by maximizing $\mathcal{L}(q)$ with the Lagrange multiplier method.

The update rules in (26)–(29) are obtained such that $\mathcal{L}(q)$ is maximized. The update rule for $q(\mathbf{W})$ is derived as follows:

$$\begin{aligned} \log q(\mathbf{W}) =& \langle \log p'(\mathbf{Y}|\mathbf{W}, \mathbf{H}, \mathbf{S}) \rangle + \langle \log p(\mathbf{W}) \rangle + \langle \log p(\mathbf{H}) \rangle \\ & + \langle \log p(\mathbf{S}|\boldsymbol{\beta}^s) \rangle + \langle \log p(\boldsymbol{\beta}^s) \rangle \end{aligned} \tag{57}$$

$$= \langle \log p'(\mathbf{Y}|\mathbf{W}, \mathbf{H}, \mathbf{S}) \rangle + \langle \log p(\mathbf{W}) \rangle + \text{const.} \tag{58}$$

$$= \sum_{f,t,k} \{ y_{ft} \phi_{ftk} \log w_{fk} + w_{fk} \langle h_{kt} \rangle \} + \sum_{f,k} \{ (\alpha^w - 1) \log w_{fk} - \beta^w w_{fk} \} + \text{const.} \tag{59}$$

$$= \sum_{f,k} \left(\alpha^w + \sum_t y_{ft} \phi_{ftk} - 1 \right) \log w_{fk} - \sum_{f,k} \left(\beta^w + \sum_t \langle h_{kt} \rangle \right) w_{fk} + \text{const.} \tag{60}$$

$$= \sum_{f,k} \log \mathcal{G} \left(w_{fk} \left| \alpha^w + \sum_t y_{ft} \phi_{ftk}, \beta^w + \sum_t \langle h_{kt} \rangle \right. \right), \tag{61}$$

where $\langle \cdot \rangle$ represents $\langle \cdot \rangle_{\mathbf{\Theta} \setminus \mathbf{W}}$. The update rules for $q(\mathbf{H})$, $q(\mathbf{S})$, and $q(\boldsymbol{\beta}^s)$ can be derived in the same way.

B. VB-RNTF

The variational posterior distributions are calculated in the same way as in VB-RNMF. The update rules for the variational posterior distributions and the auxiliary variables are summarized in (33)–(39).

C. VB-SRNTF

VB-SRNTF estimates the current posterior distribution recurrently in prediction and correction steps. The prediction step calculates $p(\mathbf{\Theta}^{(n)}|\mathbf{Y}^{(1:n-1)})$ from the previous posterior distribution $p(\mathbf{\Theta}^{(n-1)}|\mathbf{Y}^{(1:n-1)})$:

$$p\left(\mathbf{\Theta}^{(n)} \left| \mathbf{Y}^{(1:n-1)} \right.\right) = \int p\left(\mathbf{\Theta}^{(n)} \left| \mathbf{\Theta}^{(n-1)} \right.\right) p\left(\mathbf{\Theta}^{(n-1)} \left| \mathbf{Y}^{(1:n-1)} \right.\right) d\mathbf{\Theta}^{(n-1)}. \tag{62}$$

According to (19) and (20), the predictive distribution is calculated as (40). The correction step estimates the current posterior distribution $p(\mathbf{\Theta}^{(n)}|\mathbf{Y}^{(1:n)})$ from the observation $\mathbf{Y}^{(n)}$ and the predictive distribution $p(\mathbf{\Theta}^{(n)}|\mathbf{Y}^{(1:n-1)})$ as follows:

$$p\left(\mathbf{\Theta}^{(n)} \left| \mathbf{Y}^{(1:n)} \right.\right) \propto p\left(\mathbf{Y}^{(n)} \left| \mathbf{\Theta}^{(n)} \right.\right) p\left(\mathbf{\Theta}^{(n)} \left| \mathbf{Y}^{(1:n-1)} \right.\right). \tag{63}$$

In the same way as in VB-RNMF and VB-RNTF, this posterior distribution can be estimated with the VB algorithm.

ACKNOWLEDGMENTS

We thank Dr. I. Sato of The University of Tokyo for providing the code of Bayesian optimization.

REFERENCES

[1] Y. Bando *et al.*, "Variational Bayesian multi-channel robust NMF for human-voice enhancement with a deformable and partially-occluded microphone array," in *Proc. Eur. Signal Process. Conf.*, 2016, pp. 1018–1022.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[3] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2140–2151, Oct. 2013.

[4] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 196–200.

[5] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. INTERSPEECH*, 2013, pp. 436–440.

[6] S. Araki, M. Okada, T. Higuchi, A. Ogawa, and T. Nakatani, "Spatial correlation model based observation vector clustering and MVDR beamforming for meeting recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 385–389.

[7] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. IEEE Workshop Appl. Signal Process Audio Acoust.*, 2011, pp. 189–192.

[8] Y. Li, X. Zhang, M. Sun, and J. Pan, "Speech enhancement based on robust NMF solved by alternating direction method of multipliers," in *Proc. IEEE Int. Workshop Multimedia Signal Process.*, 2015, pp. 1–5.

[9] G. Min, X. Zhang, X. Zou, and M. Sun, "Mask estimate through Itakura-Saito nonnegative RPCA for speech enhancement," in *Proc. IEEE Int. Workshop Acoust. Signal Enhancement*, 2016, pp. 1–5.

[10] M. Sun, Y. Li, J. F. Gemmeke, and X. Zhang, "Speech enhancement under low SNR conditions via noise estimation using sparse and low-rank NMF with Kullback–Leibler divergence," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 7, pp. 1233–1242, Jul. 2015.

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

[12] T. Kim, "Real-time independent vector analysis for convolutive blind source separation," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 57, no. 7, pp. 1431–1438, Jul. 2010.

[13] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Efficient multichannel nonnegative matrix factorization exploiting rank-1 spatial model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 276–280.

[14] J. Nikunen and T. Virtanen, "Direction of arrival based spatial covariance model for blind sound source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 3, pp. 727–739, Mar. 2014.

[15] D. FitzGerald, M. Cranitch, and E. Coyle, "Sound source separation using shifted non-negative tensor factorisation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2006, vol. V, pp. 653–656.

[16] N. Murata *et al.*, "Reverberation-robust underdetermined source separation with non-negative tensor double deconvolution," in *Proc. Eur. Signal Process. Conf.*, 2016, pp. 1648–1652.

[17] H. Chiba, N. Ono, S. Miyabe, Y. Takahashi, T. Yamada, and S. Makino, "Amplitude-based speech enhancement with nonnegative matrix factorization for asynchronous distributed recording," in *Proc. IEEE Int. Workshop Acoust. Signal Enhancement*, 2014, pp. 203–207.

[18] C. Févotte and A. Ozerov, "Notes on nonnegative tensor factorization of the spectrogram for audio source separation: Statistical insights and towards self-clustering of the spatial cues," in *Proc. Int. Symp. Comput. Music Multidisciplinary Res.*, 2010, pp. 102–115.

[19] C. Févotte and N. Dobigeon, "Nonlinear hyperspectral unmixing with robust nonnegative matrix factorization," *IEEE Trans. Signal Process.*, vol. 24, no. 12, pp. 4810–4819, Dec. 2015.

[20] C. Sun, Q. Zhang, J. Wang, and J. Xie, "Noise reduction based on robust principal component analysis," *J. Comput. Inf. Syst.*, vol. 10, no. 10, pp. 4403–4410, 2014.

[21] Z. Chen and D. P. Ellis, "Speech enhancement by sparse, low-rank, and dictionary spectrogram decomposition," in *Proc. IEEE Workshop Appl. Signal Process Audio Acoust.*, 2013, pp. 1–4.

[22] N. Dobigeon and C. Févotte, "Robust nonnegative matrix factorization for nonlinear unmixing of hyperspectral images," in *Proc. Workshop Hyperspectral Image Signal Process, Evolution Remote Sensing*, 2013, pp. 1–4.

[23] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, 2011, Art. no. 11.

[24] J. Fukuda, M. Konyo, E. Takeuchi, and S. Tadokoro, "Remote vertical exploration by active scope camera into collapsed buildings," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2014, pp. 1882–1888.

[25] S. Tadokoro *et al.*, "Application of Active Scope Camera to forensic investigation of construction accident," in *Proc. IEEE Adv. Robot. Its Social Impacts*, 2009, pp. 47–50.

[26] A. Deleforge and W. Kellermann, "Phase-optimized K-SVD for signal extraction from underdetermined multichannel sparse mixtures," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 355–359.

[27] B. Cauchi, S. Goetze, and S. Doclo, "Reduction of non-stationary noise for a robotic living assistant using sparse non-negative matrix factorization," in *Proc. Speech Multimodal Interaction Assistive Environ.*, 2012, pp. 28–33.

[28] G. Ince, K. Nakamura, F. Asano, H. Nakajima, and K. Nakadai, "Assessment of general applicability of ego noise estimation," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2011, pp. 3517–3522.

[29] S. D. Babacan, M. Luessi, R. Molina, and A. K. Katsaggelos, "Sparse Bayesian methods for low-rank matrix estimation," *IEEE IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 3964–3977, Aug. 2012.

[30] C. M. Bishop, "Pattern recognition," *Mach. Learn.*, vol. 128, 2006.

[31] N. Mae *et al.*, "Ego noise reduction for hose-shaped rescue robot combining independent low-rank matrix analysis and multichannel noise cancellation," in *Proc. Int. Conf. Latent Variable Anal. Signal Separation*, 2017, pp. 141–151.

[32] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. Horaud, "A variational EM algorithm for the separation of moving sound sources," in *Proc. IEEE Workshop Appl. Signal Process Audio Acoust.*, 2015, pp. 1–5.

[33] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma, "Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix," in *Proc. Int. Workshop Comput. Adv. Multi-Channel Sensor Array Process.*, 2009, pp. 1–18.

[34] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," in *Proc. Neural Inf. Process. Syst.*, 2009, pp. 2080–2088.

[35] X. Ding, L. He, and L. Carin, "Bayesian robust principal component analysis," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3419–3430, 2011.

[36] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Comput. Intell. Neurosci.*, vol. 2009, 2009, Art. no. 785152.

[37] M. D. Hoffman, "Poisson-uniform nonnegative matrix factorization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2012, pp. 5361–5364.

[38] D. Gamerman, T. R. Santos, and G. C. Franco, "A non-gaussian family of state-space models with exact marginal likelihood," *J. Time Series Anal.*, vol. 34, no. 6, pp. 625–645, 2013.

[39] K. Itou *et al.*, "The design of the newspaper-based Japanese large vocabulary continuous speech recognition corpus," in *Proc. Int. Conf. Spoken Lang. Process.*, 1998, p. 4.

[40] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

[41] C. Raffel *et al.*, "mir_eval: A transparent implementation of common MIR metrics," in *Proc. Int. Soc. Music Inf. Retrieval*, 2014, pp. 367–372.

[42] E. Contal, V. Perchet, and N. Vayatis, "Gaussian process optimization with mutual information," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 253–261.

[43] Y. Bando *et al.*, "Human-voice enhancement based on online RPCA for a hose-shaped rescue robot with a microphone array," in *Proc. IEEE Int. Symp. Safety, Security, Rescue Robot.*, 2015, pp. 1–6.

[44] Y. Bando *et al.*, "Low latency and high quality two-stage human-voice-enhancement system for a hose-shaped rescue robot," *J. Robot. Mechatron.*, vol. 29, no. 1, pp. 198–212, 2017.

[45] H. Nakajima, G. Ince, K. Nakadai, and Y. Hasegawa, "An easily-configurable robot audition system using histogram-based recursive level estimation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2010, pp. 958–963.

[46] T. Broderick, N. Boyd, A. Wibisono, A. C. Wilson, and M. I. Jordan, "Streaming variational Bayes," in *Proc. Neural Inf. Process. Syst.*, 2013, pp. 1727–1735.

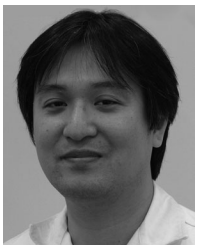
[47] A. T. Cemgil and O. Dikmen, "Conjugate gamma Markov random fields for modelling nonstationary sources," in *Proc. Int. Conf. Ind. Compon. Anal. Signal Separation*, 2007, pp. 697–705.

Yoshiaki Bando (S'17) received the M.S. degree in informatics from Kyoto University, Kyoto, Japan, in 2015. He is currently working toward the Ph.D. degree in the Department of Intelligence Science and Technology, Kyoto University. His research interests include microphone array signal processing, rescue robotics, and machine learning. He was the recipient of the Advanced Robotics Best Paper Award in 2016, the Most Innovative Paper Award at IEEE SSRR 2015, the Best Student Paper Award at IEEE SSRR 2014, and the IEEE RAS Japan Chapter Young Award at IEEE/RSJ IROS 2013. He is a member of the RSJ, IPSJ, and JSAI.

Katsutoshi Itoyama (M'07) received the B.E., M.S., and the Ph.D. degrees in informatics from Kyoto University, Kyoto, Japan, in 2006, 2008, and 2011, respectively. He is currently an Assistant Professor with the Graduate School of Informatics, Kyoto University. His research interests include musical sound source separation, music listening interfaces, and music information retrieval. He was the recipient of the 24th TAF Telecom Student Technology Award and the IPSJ Digital Courier Funai Young Researcher Encouragement Award. He is a member of the IPSJ and ASJ.

Masashi Konyo (M'05) received the B.S., M.S., and Ph.D. degrees in engineering from Kobe University, Kobe, Japan, in 1999, 2001, and 2004, respectively. He is currently an Associate Professor with the Graduate School of Information Sciences, Tohoku University, Sendai, Japan. His research interests include haptic interfaces, rescue robotics, and new actuators. He was the recipient of the Young Scientists Prize, the Commendation for Science and Technology by MEXT in 2015, the Best Paper Awards of Journal of Robotics and Mechatronics in 2010 and Advanced Robotics in 2016, the Best Poster Awards of the World Haptics Conference in 2007 and 2013, and the Best Hands on Demo Award at the Euro-Haptics 2008 and Haptics Symposium 2014. He is a member of the RSJ, JSME, SI, and VRSJ.

Satoshi Tadokoro (F'09) was an Associate Professor with Kobe University during 1993–2005, and has been a Professor with the Graduate School of Information Sciences, Tohoku University, Sendai, Japan. He is the President of International Rescue System Institute and IEEE RAS in 2016–2017. He was a Project Manager of MEXT DDT Project on rescue robotics during 2002–2007, having contribution of more than 100 professors nationwide, and NEDO Project that developed a rescue robot Quince, which is being used at the Fukushima-Daiichi Nuclear Power Plant Accident since June 2011. Since 2014, he is a Project Manager of Japan Government's ImPACT Project. He is an RSJ Fellow.



Kazuhiro Nakadai (SM'14) received the B.E. degree in electrical engineering, the M.E. degree in information engineering, and the Ph.D. degree in electrical engineering from the University of Tokyo, Tokyo, Japan, in 1993, 1995, and 2003, respectively. He was with Nippon Telegraph and Telephone for four years as a System Engineer from 1995 to 1999. After that, he worked on the Kitano Symbiotic Systems Project, ERATO, and JST as a Researcher from 1999 to 2003. He is currently a Principal Researcher with Honda Research Institute Japan, Co., Ltd, Wako, Japan. He has had a concurrent position at Tokyo Institute of Technology, as a Visiting Associate Professor from 2006 to 2010, a Visiting Professor from 2011 to 2017, and a specially appointed Professor since July 2017. He also has had a concurrent position as a Guest Professor at Waseda University since 2011. His research interests include AI, robotics, signal processing, computational auditory scene analysis, multimodal integration, and robot audition. He has been an Executive Board Member of the JSAI from 2015 to 2016, and of the RSJ from 2017 to 2018. He is also a member of the IPSJ, ASJ, HIS, ISCA, and ACM.

Kazuyoshi Yoshii (M'08) received the Ph.D. degree in informatics from Kyoto University, Kyoto, Japan, in 2008. He is currently a Senior Lecturer with Kyoto University and concurrently a Leader of Sound Scene Understanding Team at RIKEN Center for Advanced Intelligence Project (AIP). His research interests include audio signal processing and machine learning. He is a member of the Information Processing Society of Japan and Institute of Electronics, Information, and Communication Engineers.

Tatsuya Kawahara (F'17) received the B.E., M.E., and Ph.D. degrees, all in information science, from Kyoto University, Kyoto, Japan, in 1987, 1989, and 1995, respectively. From 1995 to 1996, he was a Visiting Researcher at Bell Laboratories, Murray Hill, NJ, USA. He is currently a Professor in the School of Informatics, Kyoto University. He has also been an Invited Researcher at ATR and NICT. He has authored or coauthored more than 300 technical papers on speech recognition, spoken language processing, and spoken dialogue systems. He has been conducting several speech-related projects in Japan including speech recognition software Julius and the automatic transcription system for the Japanese Parliament (Diet).

He was the recipient of the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology (MEXT) in 2012. From 2003 to 2006, he was a member of IEEE SPS Speech Technical Committee. He was a General Chair of IEEE Automatic Speech Recognition and Understanding workshop (ASRU 2007). He also served as a Tutorial Chair of INTERSPEECH 2010 and a Local Arrangement Chair of ICASSP 2012. He is an Editorial Board Member of Elsevier *Journal of Computer Speech and Language*, *APSIPA Transactions on Signal and Information Processing*, and IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING. He is the VP-Publications of APSIPA and a Board Member of ISCA.

Hiroshi G. Okuno (F'12) received the B.A. and Ph.D. degrees from the University of Tokyo, Tokyo, Japan, in 1972 and 1996, respectively. He was with NTT, JST, the Tokyo University of Science, and Kyoto University. He is currently a Professor in the Graduate Program for Embodiment Informatics, Waseda University, Tokyo, Japan, and a Professor Emeritus, Kyoto University, Kyoto, Japan. He was a visiting scholar at Stanford University, Stanford, CA, USA, from 1986 to 1988. He has done research in programming languages, parallel processing, and reasoning mechanisms in AI. He is currently engaged in computational auditory scene analysis, music scene analysis, and robot audition.

He coedited *Computational Auditory Scene Analysis* (CRC Press, 1998), *Advanced Lisp Technology* (Taylor & Francis, 2002), and New Trends in *Applied Artificial Intelligence (IEA/AIE)* (Springer, 2007). He was a recipient of various awards including 2013 JSAI Achievement Award, the 2nd Best Paper Award of Advanced Robotics, 2013 Science and Technology Award of Minister of Education, Culture, Sports, Science Technology, 2010 NTF Award for Entertainment Robots and Systems, the Best Paper Award of IEA/AIE-2001, 2005, 2010 and 2013, and the 1990 Best Paper Award of the JSAI. He is Fellow of the JSAI, IPSJ, and RSJ.