Sound Source Localization Using Joint Bayesian Estimation With a Hierarchical Noise Model

Futoshi Asano, *Member, IEEE*, Hideki Asoh, *Member, IEEE*, and Kazuhiro Nakadai, *Member, IEEE*

***Abstract*—The performance of sound source localization is often reduced by the presence of colored noise in the environment, such as room reverberation. In this study, a method for estimating the noise spatial covariance using a hierarchical model is proposed and its performance is evaluated. By employing the hierarchical model in joint Bayesian estimation, robust estimation of the covariance is expected with a relatively small amount of data. Moreover, a method of jointly estimating the number of sources is introduced so that it can be used for cases in which the number of active sources dynamically changes, for example, speech signals. The results of the experiments performed using actual room reverberation show the effectiveness of the proposed method.**

***Index Terms*—Colored noise, hierarchical model, joint Bayesian estimation, reversible jump MCMC method, the Markov chain Monte Carlo method.**

I. Introduction

SOUND source localization is a technique useful for various applications such as distant speech recognition [1], sound interfaces for robots [2], machine diagnostics in factories, and sound source detection in disaster environments. Among the localization methods, the parametric modeling approach is widely used. In this method, an observation model is built and optimized under some criterion so that it matches the actual data observed. Then, the parameters of interest, such as the location of sound sources, are extracted. Among these approaches, the maximum likelihood (ML) method (e.g., [3]), in which the likelihood of the observation model is maximized as a criterion, is the most straight forward. The widespread delay-and-sum beamformer (steered beamformer response power in some literature [1]), which is basically a heuristic method, can also be viewed as a special form of the ML estimator, as briefly described in Appendix A. The well-known MUSIC method [4] is also a model-based approach in which the orthogonality of the basis for the signal and noise is used as a criterion. In this study, the observation model defined in the frequency domain is used in the same manner as in many previous general-purpose localization studies (e.g., [3]), because frequency domain modeling has the following advantages:

- The observation model in the frequency domain is simpler than that in the time domain as shown in Section II-A, because convolution in the time domain is replaced by a simple product in the frequency domain.
- A large number of methods developed for narrowband applications, such as radar/sonar and communication, can be directly introduced.

On the other hand, the time domain approach, which is not discussed in this paper, is also widely used in sound source localization. This approach is typically based on the estimation of the time delay of arrival (TDOA) [1], [5]–[7].

In the model-based approach, model precision is an important issue. In terms of the model precision, the model-based approach has the following problems when applied to acoustic source localization:

- The noise included in the observation is usually spatially colored (i.e., there is some correlation between the sensors), and this makes it difficult to build a precise model. A typical example of the colored noise is room reverberation, which is the main focus of this study.
- The number of sound sources, which corresponds to the dimension of the model, is usually unknown.

Regarding the problem of noise, the noise in the model is typically assumed to be spatially white in most model-based approaches even when the actual noise is spatially colored. With this assumption, the estimation algorithm becomes much simpler. However, as discussed in Section II-B, this mismatch of the model causes some degradation in performance, such as a reduction in spatial resolution. To overcome this problem, the prewhitening method using generalized eigenvalue decomposition (GEVD) was proposed by Roy *et al.* [8]. In this approach, the spatial covariance of the noise must be known in advance. However, for room reverberation, it is difficult to obtain the noise covariance from the observation because the reverberation cannot be observed in isolation. An alternative approach is to employ a more realistic noise model such as the isotropic noise model [3], which is considered to be an appropriate approximation of the reverberant sound field. However, as shown later in Fig. 6, the effect of this model on sound source localization is sometimes limited for actual room reverberation, which is not always isotropic.

Regarding the number of sources, Wax *et al.* proposed an ML-based method using the AIC/MDL criterion [10]. However, this method greatly depends on the assumption that the noise is spatially white and is not effective for acoustic source localization [11]. An approach using pattern classification of the eigenvalue distribution of the observation covariance matrix is somewhat effective for spatially colored noise [12]. However, the disadvantage of this method is that it requires training for each target environment.

Manuscript received September 21, 2012; revised January 31, 2013; accepted May 04, 2013. Date of publication May 14, 2013; date of current version July 12, 2013. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Sharon Gannot.

F. Asano and H. Asoh are with Intelligent Systems Research Institute, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba 305-8568, Japan (e-mail: f.asano@aist.go.jp; h.asoh@aist.go.jp).

K. Nakadai is with Honda Research Institute Japan Co., Ltd., Wako-shi 351-0188, Japan (e-mail: nakadai@jp.honda-ri.com).

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/TASL.2013.2263140

As described above, the performance of the model-based approach is limited by the two unknown parameters of the observation model, i.e., the noise covariance and the number of sources. Therefore, these parameters as well as the location parameter should be estimated from the observation data to enhance the estimation performance. This is the motivation for this study. To estimate multiple parameters, the Bayesian approach, which can provide a framework for several types of joint estimation [13], [14], is considered to be promising. The Bayesian approach has already been applied to localization/tracking applications. In this paper, the localization and tracking functions are distinguished for the sake of simplicity, such that the localization assumes that the sources are static (without movements), while the tracking assumes that they are dynamic. A method for implementing the Bayesian approach for source localization is the Markov chain Monte Carlo (MCMC) method (e.g., [14]). Andrieu *et al.* [15] proposed a method of source localization using the MCMC method, in which the location and number of sources are jointly estimated. However, in this method, the noise was assumed to be spatially white and the noise covariance was integrated out as a nuisance parameter. Our method extends the method in [15] to explicitly include spatially colored noise. On the other hand, in the Bayesian approach for source tracking, the sequential Monte Carlo (SMC) method, which includes the well-known particle filter, is widely used (e.g., [16]–[18] and the references therein). The particle filter approach has already been employed for sound source tracking [19], [20]. For source tracking, both the observation model (measurement equation [21]) and the motion model (process equation) are considered. In this sense, the tracking problem is more difficult than that in static source localization. The scope of this paper is limited to the improvement of the observation model precision in sound source localization. However, the discussion in this paper can be extended to the tracking problem because tracking also uses the observation model.

To estimate the noise covariance, the hierarchical model is introduced as a main feature. In some applications, the time period in which the location of target sources can be assumed to be stationary is short. In this case, the amount of data for sound source localization might be insufficient, resulting in large estimation variance. However, as long as the observations are obtained in the same environment (room), the spatial covariance of room reverberation is expected to have a common factor. This is because some environmental parameters such as resonant frequencies and the mode of a room are independent of the sensor/source location. By estimating this common factor in the noise covariance using a hierarchical model, the estimation of the noise covariance is expected to be robust even for limited amounts of data.

This paper is organized as follows. In Section II, the observation model in the frequency domain is introduced. Moreover, the problem associated with the assumption of spatially white noise and how it is resolved by the application of a precise noise model is illustrated by using GEVD-MUSIC as an example. In Section III, the basic joint estimation procedure using the MCMC method in [15] is extended so that the noise covariance can also be estimated [22]. In Section IV, a hierarchical model of the noise covariance is introduced [23]. In Section V, the estimation of the number of sound sources is discussed. The reversible jump MCMC method proposed by Green [24], which is also used in the MCMC source localization in [15], is introduced in our hierarchical model approach [25]. In Section VI, the proposed method is evaluated using the data with the reverberation measured in an actual room.

II. Problem Statement

A. Observation Model

In this study, the estimations are performed in the frequency domain. The observation vector consists of the short-time Fourier transform (STFT) of the sensor inputs, which is given as

$$\boldsymbol{z}_{j,k} = [Z_1(\omega, j, k), \ldots, Z_M(\omega, j, k)]^T \tag{1}$$

where $Z_m(\omega, j, k)$ denotes the STFT of the mth sensor input at the frequency ω and the time frame index k. The symbol j denotes the index for the time block that consists of L observations (frames) as $\boldsymbol{Z}_j = [\boldsymbol{z}_{j,1}, \ldots, \boldsymbol{z}_{j,L}]$. The source direction $\boldsymbol{\theta}_j = [\theta_{j,1}, \ldots, \theta_{j,N_j}]^T$ within the block is assumed to be invariant. The symbols M and N_j denote the number of sensors and sources, respectively. The number of sources, N_j, may vary at the different blocks. The observation vector is assumed to be modeled as

$$\boldsymbol{z}_{j,k} = \boldsymbol{A}(\boldsymbol{\theta}_j)\boldsymbol{s}_{j,k} + \boldsymbol{v}_{j,k} \tag{2}$$

where $\boldsymbol{A}(\boldsymbol{\theta}_j)$ denotes the array manifold matrix. The matrix $\boldsymbol{A}(\boldsymbol{\theta}_j)$ is sometimes denoted as $\boldsymbol{A}_j$ for the sake of simplicity. The nth column vector of $\boldsymbol{A}_j$ (the array manifold vector, AMV) has the form of $\boldsymbol{a}(\theta_{j,n}) = [\exp(-j\omega\tau_{1,n}), \ldots, \exp(-j\omega\tau_{M,n})]^T$ when microphones are located in free field and the sources are distantly located from the array (far-field condition) . The symbol $\tau_{m,n}$ denotes the propagation time between the nth source to the mth microphone. When this free field and far field assumption dose not hold, the AMV will include factors other than the time delay. For example, when microphones are mounted on a rigid surface as discussed in Section VI, the AMV should reflect the diffraction of the surface. The symbols $\boldsymbol{s}_{j,k}$ and $\boldsymbol{v}_{j,k}$ are the source vector and noise vector, respectively. Assuming that $\boldsymbol{s}_{j,k}$ and $\boldsymbol{v}_{j,k}$ are uncorrelated, the observation covariance matrix can be modeled as

$$\boldsymbol{R}_j = E[\boldsymbol{z}_{j,k}\boldsymbol{z}_{j,k}^H] = \boldsymbol{A}_j\boldsymbol{\Gamma}_j\boldsymbol{A}_j^H + \boldsymbol{K}_j \tag{3}$$

where $\boldsymbol{\Gamma}_j = E[\boldsymbol{s}_{j,k}\boldsymbol{s}_{j,k}^H]$ is the source covariance and $\boldsymbol{K}_j = E[\boldsymbol{v}_{j,k}\boldsymbol{v}_{j,k}^H]$ is the noise covariance. When $\boldsymbol{v}_{j,k}$ is spatially white, the noise covariance has the form of $\boldsymbol{K}_j = \sigma^2\boldsymbol{I}$, and the estimation algorithm becomes much simpler. However, as described in the next section, when $\boldsymbol{v}_{j,k}$ is spatially colored, the mismatch between the model and the observation sometimes degrades the performance.

B. Effect of Spatially Colored Noise

In this subsection, using the MUSIC estimator as an example, it is briefly shown how information in the noise covariance $\boldsymbol{K}_j$ affects source localization. Fig. 1 shows an example of the eigenvalue distribution of $\boldsymbol{R}_j$ and the corresponding MUSIC

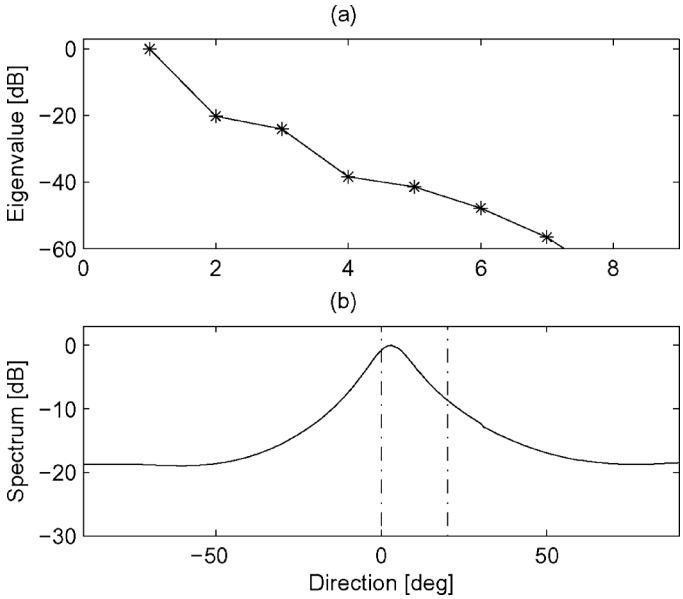

Fig. 1. Eigenvalues and the MUSIC spectrum using SEVD. (a) Eigenvalue. (b) MUSIC Spectrum.

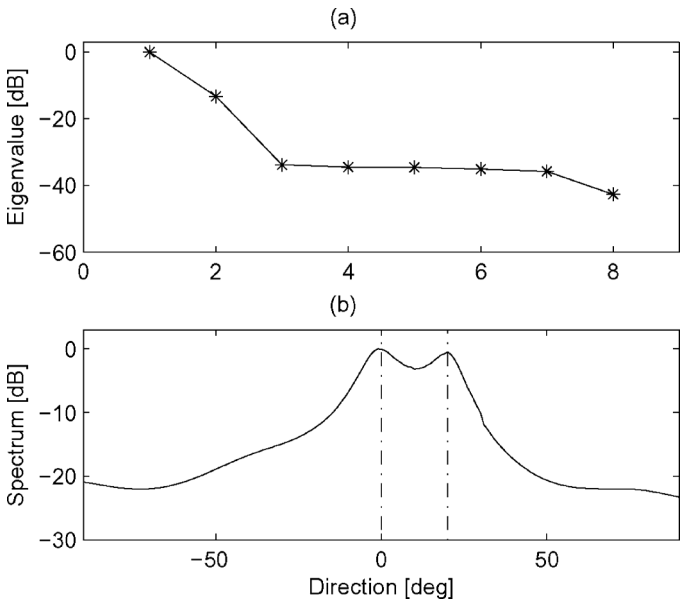

Fig. 2. Eigenvalues and the MUSIC spectrum using GEVD with a known noise covariance. (a) Eigenvalue. (b) MUSIC Spectrum.

spectrum. The location of the two sources is (0°, 20°), as depicted by the dash-dot line in the figure. The observation condition is the same as that in the experiment described in Section VI. The eigenvalues and eigenvectors are obtained from the standard eigenvalue decomposition (SEVD) problem $\boldsymbol{R}_j\boldsymbol{e} = \lambda\boldsymbol{e}$, where λ and $\boldsymbol{e}$ denote the eigenvalue and eigenvector, respectively. The SEVD-based MUSIC estimator assumes that the noise $\boldsymbol{v}_{j,k}$ is spatially white, whereas the actual noise in the observation consists of room reverberation and is spatially colored. It can be seen that the two peaks that should appear at (0°, 20°) were merged into a single peak.

Fig. 2 shows the case when GEVD is employed. In GEVD, the eigenvalues and eigenvectors are obtained from the generalized eigenvalue problem $\boldsymbol{R}_j\boldsymbol{e} = \lambda\boldsymbol{K}_j\boldsymbol{e}$, and the noise whitening process is included in the joint diagonalization of $\boldsymbol{R}_j$ and $\boldsymbol{K}_j$ [26]. From Fig. 2(a), two dominant eigenvalues corresponding to the number of sources $N_j = 2$ can be seen, whereas the other eigenvalues are almost flat. This is attributed to noise whitening. In Fig. 2(b), two peaks appear at (0°, 20°). From these, it can be seen that the spatial resolution of the estimator is improved by the information included in the noise covariance. To obtain $\boldsymbol{K}_j$ in this example, the measured room impulse responses were divided into direct sound and reverberation, and the responses corresponding to the reverberation were then convolved with the source signal to obtain the noise observation (reverberation) $\boldsymbol{v}_k$ separately. However, in a real application, $\boldsymbol{K}_j$ is not available. Therefore, in the present study, $\boldsymbol{K}_j$ is estimated using the joint Bayesian estimation and the hierarchical model.

III. Basic MCMC Estimation

A. Overview

In Section III, a basic method for estimating the multiple parameters $\{\boldsymbol{\theta}_j, \boldsymbol{S}_j, \boldsymbol{K}_j\}$ within a block is proposed [22]. The symbol $\boldsymbol{S}_j$ denotes the source signals in the jth block as $\boldsymbol{S}_j = [\boldsymbol{s}_{j,1}, \ldots, \boldsymbol{s}_{j,L}]$. When multiple parameters are present, it is often difficult to estimate them simultaneously because the joint posterior distribution $p(\boldsymbol{\theta}_j, \boldsymbol{S}_j, \boldsymbol{K}_j|\boldsymbol{Z}_j)$ might not be tractable. However, in some cases, it is easy to obtain samples of the parameters from the full conditional distribution (e.g., the full conditional distribution of $\boldsymbol{\theta}_j$ is $p(\boldsymbol{\theta}_j|\boldsymbol{Z}_j, \boldsymbol{S}_j, \boldsymbol{K}_j)$). In such cases, it is known that the joint posterior distribution $p(\boldsymbol{\theta}_j, \boldsymbol{S}_j, \boldsymbol{K}_j|\boldsymbol{Z}_j)$ can be approximated by the Gibbs sampler [14]. On the other hand, in cases where the sample cannot be easily obtained from its full conditional distribution, the Metropolis algorithm [14] can be used. In the Metropolis algorithm, the sample is drawn from an arbitrary distribution called the proposal distribution. However, the proposed sample is accepted/rejected on the basis of an acceptance ratio. Gibbs sampling and the Metropolis algorithm are related in that Gibbs sampling is an ideal case of the Metropolis algorithm when the full conditional distribution is used as the proposal distribution. In this study, the samples of $\boldsymbol{S}_j$ and $\boldsymbol{K}_j$ can be drawn from their full conditional distributions. On the other hand, the samples of $\boldsymbol{\theta}_j$ are obtained using the Metropolis algorithm because of the nonlinearity between $\boldsymbol{Z}_j$ and $\boldsymbol{\theta}_j$ [15]. Sections III-C – III-F describe the sampling model used for each parameter, whereas Section III-G describes the entire sampling procedure.

Assumptions

In preparation for developing the proposed algorithm, the assumptions required for it are described.

A-1) The source vectors $\{\boldsymbol{s}_{j,1}, \ldots, \boldsymbol{s}_{j,L}\}$ are i.i.d. (independent and identically distributed) and have a complex Gaussian prior distribution.

A-2) The noise vectors $\{\boldsymbol{v}_{j,1}, \ldots, \boldsymbol{v}_{j,L}\}$ are i.i.d. and has a complex Gaussian prior distribution.

A-3) The source direction $\boldsymbol{\theta}_j$ has a uniform prior distribution.

A-4) The noise covariance $\boldsymbol{K}_j$ has an inverse Wishart prior distribution.

A-5) The random variables $\boldsymbol{\theta}_j$, $\boldsymbol{S}_j$ and $\boldsymbol{K}_j$ are mutually independent.

Regarding the choice of the prior distribution, it will be ideal to select a distribution that is close to the actual distribution.

For example, for the application of speaker tracking, the speech source is known to have a sharper distribution than the Gaussian distribution. However, its true distribution is usually unknown. Therefore, in this paper, the prior distribution is selected from the mathematical viewpoint in which conjugacy is an important issue. The prior distribution is called conjugate if the prior and posterior (conditional) distributions belong to the same class (e.g., Gaussian). The conjugate prior distribution makes the algorithm tractable.

B. Likelihood

Assuming that $\boldsymbol{v}_{j,k}$ has the complex Gaussian distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{K}_j)$ (Assumption A-2 in Section III-B), the likelihood for the observation $\boldsymbol{z}_{j,k}$ is written as follows:

$$p(\boldsymbol{z}_{j,k}|\boldsymbol{\theta}_j, \boldsymbol{s}_{j,k}, \boldsymbol{K}_j) \propto |\boldsymbol{K}_j|^{-1} \times \exp\left[-\left(\boldsymbol{z}_{j,k} - \boldsymbol{A}_j\boldsymbol{s}_{j,k}\right)^H \boldsymbol{K}_j^{-1}\left(\boldsymbol{z}_{j,k} - \boldsymbol{A}_j\boldsymbol{s}_{j,k}\right)\right] \quad (4)$$

Using Assumptions A-1 and A-2 (Section III-B), the likelihood of the block observation $\boldsymbol{Z}_j$ can be written as follows (e.g., [14]):

$$\begin{aligned} p(\boldsymbol{Z}_j|\boldsymbol{\theta}_j, \boldsymbol{S}_j, \boldsymbol{K}_j) &= \prod_{k=1}^{L} p(\boldsymbol{z}_{j,k}|\boldsymbol{\theta}_j, \boldsymbol{s}_{j,k}, \boldsymbol{K}_j) \\ &\propto |\boldsymbol{K}_j|^{-L} \exp\left[-\mathrm{tr}\left(\boldsymbol{C}_j\boldsymbol{K}_j^{-1}\right)\right] \end{aligned} \quad (5)$$

where $\mathrm{tr}(\cdot)$ indicates the trace of a matrix and

$$\boldsymbol{C}_j = \sum_{k=1}^{L} \left[\boldsymbol{z}_{j,k} - \boldsymbol{A}_j\boldsymbol{s}_{j,k}\right]\left[\boldsymbol{z}_{j,k} - \boldsymbol{A}_j\boldsymbol{s}_{j,k}\right]^H \quad (6)$$

C. Conditional Distribution of $\boldsymbol{s}_{j,k}$

Assuming that the signal $\boldsymbol{s}_{j,k}$ has the complex Gaussian prior distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Phi}_0)$ (Assumption A-1 in Section III-B), its full conditional distribution becomes the following Gaussian distribution (see Appendix A for the derivation) :

$$p(\boldsymbol{s}_{j,k}|\boldsymbol{Z}_j, \boldsymbol{\theta}_j, \boldsymbol{K}_j) = \mathcal{N}(\boldsymbol{\mu}_{j,k}, \boldsymbol{\Phi}_j) \quad (7)$$

where

$$\boldsymbol{\mu}_{j,k} := \boldsymbol{\Phi}_j\boldsymbol{A}_j^H\boldsymbol{K}_j^{-1}\boldsymbol{z}_{j,k} \quad (8)$$

$$\boldsymbol{\Phi}_j := \left(\boldsymbol{A}_j^H\boldsymbol{K}_j^{-1}\boldsymbol{A}_j + \boldsymbol{\Phi}_0^{-1}\right)^{-1} \quad (9)$$

The reason for the choice of the prior $p(\boldsymbol{s}_{j,k})$ is the consideration of conjugacy, as described in Section III-B. With respect to the parameter $\boldsymbol{\Phi}_0$, $\boldsymbol{\Phi}_0 = c\boldsymbol{I}$ where c is a scaling constant, when assuming that the source signals are mutually uncorrelated. Because $\boldsymbol{\Phi}_0$ only appears in (9), the constant c is determined in terms of regularization so that the existence of $\boldsymbol{\Phi}_j$ is guaranteed. In this paper, $c = 10^{-4\sim-6}\times$ the largest eigenvalue of $(\boldsymbol{A}_j^H\boldsymbol{K}_j^{-1}\boldsymbol{A}_j)^{-1}$.

D. Conditional Distribution of $\boldsymbol{K}_j$

It is assumed that the covariance $\boldsymbol{K}_j$ has a complex inverse Wishart distribution as its prior distribution (Assumption A-4 in Section III-B):

$$\begin{aligned} p(\boldsymbol{K}_j) &= \text{inv-Wishart}\left(\nu_0, (\nu_0\boldsymbol{K}_0)^{-1}\right) \\ &\propto |\boldsymbol{K}_j|^{-(\nu_0+M)} \exp\left[-\mathrm{tr}(\nu_0\boldsymbol{K}_0\boldsymbol{K}_j^{-1})\right] \end{aligned} \quad (10)$$

where ν_0 is the virtual sample size. The full conditional distribution of $\boldsymbol{K}_j$ is then the following inverse Wishart distribution (e.g., [14]):

$$\begin{aligned} p(\boldsymbol{K}_j|\boldsymbol{Z}_j, \boldsymbol{S}_j, \boldsymbol{\theta}_j) &\propto p(\boldsymbol{K}_j)p(\boldsymbol{Z}_j|\boldsymbol{\theta}_j, \boldsymbol{S}_j, \boldsymbol{K}_j) \\ &\propto |\boldsymbol{K}_j|^{-(\nu_0+M)} \exp\left[-\mathrm{tr}(\nu_0\boldsymbol{K}_0\boldsymbol{K}_j^{-1})\right] \\ &\quad \times |\boldsymbol{K}_j|^{-L} \exp\left[-\mathrm{tr}(\boldsymbol{C}_j\boldsymbol{K}_j^{-1})\right] \\ &\propto \text{inv-Wishart}\left[\nu_0 + L, (\nu_0\boldsymbol{K}_0 + \boldsymbol{C}_j)^{-1}\right] \end{aligned} \quad (11)$$

The reason for the choice of the prior $p(\boldsymbol{K}_j)$ is the consideration of conjugacy (see Section III-B). The parameters $\{\nu_0, \boldsymbol{K}_0\}$ are estimated using the hierarchical model as described later in Section IV.

E. Conditional Distribution of $\boldsymbol{\theta}_j$

As described in Section III-A, the samples for $\boldsymbol{\theta}_j$ are obtained using the Metropolis algorithm (e.g., [14]). The proposal distribution used in the Metropolis algorithm is the following Gaussian distribution:

$$J(\boldsymbol{\theta}_j^*|\boldsymbol{\theta}_j^{(p)}) = \mathcal{N}(\boldsymbol{\theta}_j^{(p)}, \sigma_\theta^2\boldsymbol{I}) \quad (12)$$

where p is the index for the iteration in the Gibbs sampling described in Section III-G. The symbol σ_θ^2 is an appropriate constant which controls the search width. The proposed sample $\boldsymbol{\theta}^*$ is accepted/rejected as follows:

$$\boldsymbol{\theta}_j^{(p+1)} = \begin{cases} \boldsymbol{\theta}_j^* & \text{with probability } \min(r, 1) \\ \boldsymbol{\theta}_j^{(p)} & \text{with probability } 1 - \min(r, 1) \end{cases} \quad (13)$$

where r is the acceptance ratio defined as

$$\begin{aligned} r &= \frac{p(\boldsymbol{\theta}_j^*|\boldsymbol{Z}_j, \boldsymbol{S}_j^{(p+1)}, \boldsymbol{K}_j^{(p+1)})}{p(\boldsymbol{\theta}_j^{(p)}|\boldsymbol{Z}_j, \boldsymbol{S}_j^{(p+1)}, \boldsymbol{K}_j^{(p+1)})} \\ &= \frac{p(\boldsymbol{Z}_j|\boldsymbol{\theta}_j^*, \boldsymbol{S}_j^{(p+1)}, \boldsymbol{K}_j^{(p+1)})}{p(\boldsymbol{Z}_j|\boldsymbol{\theta}_j^{(p)}, \boldsymbol{S}_j^{(p+1)}, \boldsymbol{K}_j^{(p+1)})} \frac{p(\boldsymbol{\theta}_j^*)}{p(\boldsymbol{\theta}_j^{(p)})} \end{aligned} \quad (14)$$

In (14), Assumption A-5 in Section III-B is used. The ratio $p(\boldsymbol{\theta}_j^*)/p(\boldsymbol{\theta}_j^{(p)})$ is assumed to be unity in this paper. Equation (13) can be accomplished by sampling $u \sim \mathcal{U}(0, 1)$ and accepting $\boldsymbol{\theta}_j^*$ when $r > u$ [14]. Regarding the prior distribution $p(\boldsymbol{\theta}_j)$, a uniform distribution is assumed (Assumption A-3 in Section III-B) because no assumption can be made for the source location.

F. Joint Parameter Estimation Using the MCMC Method

The iterative algorithm for the joint estimation is as follows:

1) Set $\boldsymbol{K}_j^{(1)}$ and $\boldsymbol{\theta}_j^{(1)}$
2) Sample $\boldsymbol{s}_{j,k}^{(p+1)} \sim p(\boldsymbol{s}_{j,k}|\boldsymbol{Z}_j, \boldsymbol{\theta}_j^{(p)}, \boldsymbol{K}_j^{(p)}) \quad \forall k$

3) Sample $\boldsymbol{K}_j^{(p+1)} \sim p(\boldsymbol{K}_j|\boldsymbol{Z}_j, \boldsymbol{S}_j^{(p+1)}, \boldsymbol{\theta}_j^{(p)})$
4) Sample $\boldsymbol{\theta}_j^{(p+1)}$ as follows:$\boldsymbol{\theta}_j^* \sim J(\boldsymbol{\theta}_j^*|\boldsymbol{\theta}_j^{(p)})$, and accept/reject $\boldsymbol{\theta}_j^*$ according to (13).
5) Go back to Step 2 with $p \leftarrow p+1$.

IV. Hierarchical Model

A. Overview

From the results in the previous section, we now have the sample estimates of the noise covariance $\{\boldsymbol{K}_1, \ldots, \boldsymbol{K}_J\}$ for J observation blocks. However, in some blocks where the signal-to-noise ratio (SNR) is low, the precision of the estimation might be low. Moreover, when the number of observations in each block is small, the variance of the estimation might be large. As described in Section I, the covariance for the room reverberation may have a common factor between the blocks as long as the observations are taken in the same environment (room). Based on this consideration, a method for increasing the precision and the stability of the estimation of the noise covariance is proposed in this section by introducing the hierarchical model. In the hierarchical model, it is assumed that $\{\boldsymbol{K}_1, \ldots, \boldsymbol{K}_J\}$ are the samples from the same population, as described later in Section IV-B, and the distribution of this population is estimated. By considering this population, the probability of generating samples far from the mean of this population is reduced.

B. Sampling Model

We assume the following sampling model:

$$\boldsymbol{K}_1, \ldots, \boldsymbol{K}_J \sim \text{i.i.d. inv-Wishart}(\nu_0, (\nu_0 \boldsymbol{K}_0)^{-1}) \tag{15}$$

where the parameter set $\{\nu_0, \boldsymbol{K}_0\}$ is common between the J block observations. Fig. 3 provides a schematic depiction of this model.

C. Conditional Distribution of $\boldsymbol{K}_0$

According to the sampling model (15), the full conditional distribution of $\boldsymbol{K}_0$ can be decomposed as follows:

$$p(\boldsymbol{K}_0|\boldsymbol{K}_1, \ldots, \boldsymbol{K}_J, \nu_0) \propto p(\boldsymbol{K}_0) \prod_{j=1}^{J} p(\boldsymbol{K}_j|\boldsymbol{K}_0, \nu_0) \tag{16}$$

Assuming that $\boldsymbol{K}_0$ has the complex Wishart distribution $p(\boldsymbol{K}_0) = \text{Wishart}(\eta, \boldsymbol{\Psi})$ as the prior distribution, the full conditional distribution becomes

$$\begin{aligned} &p(\boldsymbol{K}_0|\boldsymbol{K}_1, \ldots, \boldsymbol{K}_J, \nu_0) \\ &\propto \text{Wishart}(\boldsymbol{K}_0; \eta, \boldsymbol{\Psi}) \prod_{j=1}^{J} \text{inv-Wishart}(\boldsymbol{K}_j; \nu_0, (\nu_0 \boldsymbol{K}_0)^{-1}) \\ &\propto |\boldsymbol{K}_0|^{\eta + J\nu_0 - M} \exp\left\{-\text{tr}\left(\boldsymbol{K}_0 \boldsymbol{\Lambda}^{-1}\right)\right\} \\ &\propto \text{Wishart}\left(\boldsymbol{K}_0; \eta + J\nu_0, \boldsymbol{\Lambda}\right) \end{aligned} \tag{17}$$

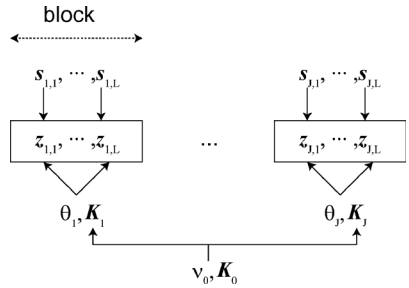

Fig. 3. Schematic of the hierarchical model of the noise covariance matrix.

where

$$\boldsymbol{\Lambda} := \left(\boldsymbol{\Psi}^{-1} + \nu_0 \sum_{j=1}^{J} \boldsymbol{K}_j^{-1}\right)^{-1} \tag{18}$$

Equation (18) is analogous to the harmonic mean in the case of a scalar parameter [14]. The reason for the choice of the prior distribution $p(\boldsymbol{K}_0)$ is the consideration of conjugacy (see Section III-B). Regarding the parameter $\boldsymbol{\Psi}$, $\boldsymbol{\Psi} = c\boldsymbol{I}$ where c is a scaling constant. The parameter η functions as a weighting factor for $p(\boldsymbol{K}_0)$ in (16), and is chosen to have a small value compared to $J\nu_0$ ($\eta = 1$ in this paper.)

D. Conditional Distribution of ν_0

Assuming that ν_0 has the prior distribution $p(\nu_0) \propto \exp(-\alpha\nu_0)$ [14], its full conditional distribution becomes

$$\begin{aligned} &p(\nu_0|\boldsymbol{K}_0, \boldsymbol{K}_1, \ldots, \boldsymbol{K}_J) \\ &\quad \propto p(\nu_0) \prod_{j=1}^{J} p(\boldsymbol{K}_j|\nu_0, \boldsymbol{K}_0) \\ &\quad \propto \exp(-\alpha\nu_0) \prod_{j=1}^{J} \frac{|\nu_0 \boldsymbol{K}_0|^{\nu_0}}{\Gamma_M(\nu_0)} |\boldsymbol{K}_j|^{-(\nu_0 + M)} \\ &\qquad \times \exp\left\{-\text{tr}(\nu_0 \boldsymbol{K}_0 \boldsymbol{K}_j^{-1})\right\} \end{aligned} \tag{19}$$

where $\Gamma_M(\nu_0) = \pi^{M(M-1)/2} \prod_{m=1}^{M} \Gamma(\nu_0 - m + 1)$. The symbol $\Gamma(\cdot)$ denotes the Gamma function. The term $|\nu_0 \boldsymbol{K}_0|^{\nu_0}/\Gamma_M(\nu_0)$ in (19) is the normalizing constant in the inverse Wishart distribution. This normalizing constant was omitted in the previous equations such as (10) when it does not affect the conditional distribution to be obtained. The reason for the choice of the prior distribution $p(\nu_0)$ is to impose an exponentially decaying weight on $p(\nu_0|\boldsymbol{K}_0, \boldsymbol{K}_1, \ldots, \boldsymbol{K}_J)$ so that ν_0 does not grow infinitely.

E. Iterative Algorithm

The procedure used to obtain samples of $\boldsymbol{K}_0$ and ν_0 is as follows:

1) Set $\boldsymbol{K}_0^{(1)}$ and $\nu_0^{(1)}$.
2) Sample $\{\boldsymbol{K}_1^{(p+1)}, \ldots, \boldsymbol{K}_J^{(p+1)}\}$ using the procedure described in Section III-G.

TABLE I
COMPUTATIONAL COMPLEXITY OF THE PROPOSED METHOD AND THE CONVENTIONAL ML METHOD FOR OBTAINING A SINGLE SAMPLE

Method	Complexity	
	Matrix operation	Sampling
Proposed	$O(M^3) \times 9$	$O(M^2) \times 2$
ML	$O(M^3) \times 4$	-

3) Sample $\boldsymbol{K}_0$ as

$$\boldsymbol{K}_0^{(p+1)} \sim p(\boldsymbol{K}_0 | \boldsymbol{K}_1^{(p+1)}, \ldots, \boldsymbol{K}_J^{(p+1)}, \nu_0^{(p)})$$

4) Sample ν_0 as

$$\nu_0^{(p+1)} \sim p(\nu_0 | \boldsymbol{K}_0^{(p+1)}, \boldsymbol{K}_1^{(p+1)}, \ldots, \boldsymbol{K}_J^{(p+1)})$$

5) Go back to Step 2 with $p \leftarrow p+1$.

F. Computational Complexity

Table I shows a rough estimate of the computational complexity for obtaining a single sample in the proposed method (Sections III–IV). In the ML method (estimation of $\boldsymbol{\theta}_j$ only), the complexity for obtaining a value of the likelihood at a single grid point is indicated for comparison. The operations with low complexities ($\leq O(M)$) are omitted. From this, it can be seen that for a single sample the computational complexity of the proposed method is approximately 2–3 times higher than that of the conventional ML method. The total complexity depends on the sampling strategy. For the ML method with full grid search, the total number of samples (grid points) grows exponentially with the number of sources N_j. On the other hand, for the proposed method, this exponential growth with N_j is avoided by employing the Monte Carlo sampling.

V. ESTIMATION OF THE NUMBER OF SOURCES

A. Overview

Thus far in the preceding discussion, the number of sources N_j has been assumed to be known. However, in many practical applications, N_j is not known in advance. Moreover, when using the frequency domain approach for a wideband time-varying signal such as speech, the number of *active* sources in each frequency bin is also time varying. Therefore, the joint estimation of N_j at every block and every frequency bin is inevitable.

In this section, the reversible jump MCMC method [24] is introduced in the proposed method to jointly estimate the number of sources. The number of sources, N_j, corresponds to the dimension of the parameter space. In the reversible jump MCMC method, the samples of the parameters are obtained by the MCMC method with jumps between the parameter subspaces having different dimensions. This jump is implemented using the {Birth,Death,Update} moves [15]. In the birth move, a new source with an arbitrary location is proposed, and the current parameter subspace is switched to the higher-dimensional subspace. In the death move, one of the current sources is removed and the parameter subspace is switched to the lower-dimensional subspace. In the update move, the number of sources, and hence the parameter subspace, is unchanged.

B. Reversible Jump MCMC Algorithm

To introduce the reversible jump MCMC method, Step 4 of the joint estimation procedure described in Section III is replaced by the following {Birth,Death,Update} moves:

- **Birth**:
  1) Increase the number of sources: $N_j^* = N_j^{(p)} + 1$
  2) Propose a new source with the location θ_{j,N_j^*} randomly selected from the possible locations and add this to the parameter vector: $\boldsymbol{\theta}_j^* = [\boldsymbol{\theta}_j^{(p)},\ \theta_{j,N_j^*}]$.
  3) Evaluate the acceptance ratio r described in Section V-C.
  4) Accept the proposal (i.e., $N_j^{(p+1)} = N_j^*$ and $\boldsymbol{\theta}_j^{(p+1)} = \boldsymbol{\theta}_j^*$) with probability $\min(r, 1)$ in the same way as (13).
- **Death**:
  1) Decrease the number of sources: $N_j^* = N_j^{(p)} - 1$.
  2) Eliminate one of the sources randomly from $\boldsymbol{\theta}_j^{(p)}$ to yield $\boldsymbol{\theta}_j^*$.
  3) Evaluate the acceptance ratio r.
  4) Accept the proposal with probability $\min(r, 1)$ in the same way as (13).
- **Update**:
  1) Conduct Step 4 in Section III with $N_j^{(p+1)} = N_j^{(p)}$.

One of these three moves is randomly selected during the iteration.

C. Acceptance Ratio

The acceptance ratio in the reversible jump MCMC method [15], [24] is defined as

$$r = \text{posterior ratio} \times \text{proposal ratio} \tag{20}$$

In this paper, the proposal ratio is assumed to be unity for the sake of simplicity. Thus, the acceptance ratio is given by the same expression as (14). However, $\boldsymbol{S}_j^{(p+1)}$ cannot be used in (14) because the dimension of $\boldsymbol{S}_j^{(p+1)}$ may be changed by the move. Therefore, $\boldsymbol{S}_j^{(p+1)}$ must be eliminated from (14) through integration.

From the integration of $\boldsymbol{S}_j$ in $p(\boldsymbol{\theta}_j, \boldsymbol{S}_j | \boldsymbol{Z}_j, \boldsymbol{K}_j)$ and the omission of unnecessary terms (see Appendix B for the derivation and definition of the symbols), $p(\boldsymbol{\theta}_j | \boldsymbol{Z}_j, \boldsymbol{K}_j)$ becomes

$$p(\boldsymbol{\theta}_j | \boldsymbol{Z}_j, \boldsymbol{K}_j) \propto |\boldsymbol{\Phi}_j|^L \exp\left[-\mathrm{tr}\left(\bar{\boldsymbol{R}}_j \boldsymbol{P}_j\right)\right] \tag{21}$$

From this, the logarithm of the acceptance ratio becomes

$$\begin{aligned} \log r &= \log p(\boldsymbol{\theta}_j^* | \boldsymbol{Z}_j, \boldsymbol{K}_j^{(p+1)}) - \log p(\boldsymbol{\theta}_j^{(p)} | \boldsymbol{Z}_j, \boldsymbol{K}_j^{(p+1)}) \\ &\propto L\left(\log |\boldsymbol{\Phi}_j^*| - \log |\boldsymbol{\Phi}_j^{(p)}|\right) - \left(\mathrm{tr}\left[\bar{\boldsymbol{R}}_j \boldsymbol{P}_j^*\right] - \mathrm{tr}\left[\bar{\boldsymbol{R}}_j \boldsymbol{P}_j^{(p)}\right]\right) \end{aligned} \tag{22}$$

As shown in Section VI, the dependency of $|\boldsymbol{\Phi}_j|$ on $\boldsymbol{\theta}_j$ is low[1]. Thus, when the number of sources is unchanged by the move (update), $\log|\boldsymbol{\Phi}_j^*| \simeq \log|\boldsymbol{\Phi}_j^{(p)}|$. In this case, the acceptance ratio $\log r$ is mainly determined by the second term, $-\left(\mathrm{tr}\left[\bar{\boldsymbol{R}}_j\boldsymbol{P}_j^*\right] - \mathrm{tr}\left[\bar{\boldsymbol{R}}_j\boldsymbol{P}_j^{(p)}\right]\right)$.

When $N_j^* > N_j^{(p)}$ (birth move), the second term tends to increase. This can be explained by the fact that the likelihood generally increases as the degree of freedom in the model increases. Therefore, when the model order is determined on the basis of the likelihood, a "penalty" term is usually introduced as in AIC/MDL (e.g., [3]). In (22), the first term $L\left(\log|\boldsymbol{\Phi}_j^*| - \log|\boldsymbol{\Phi}_j^{(p)}|\right)$ functions as a penalty. The determinant of $\boldsymbol{\Phi}_j$ can be decomposed using its eigenvalues as follows:

$$\log|\boldsymbol{\Phi}_j| = \sum_{i=1}^{N_j} \log\lambda_i \tag{23}$$

where $\{\lambda_i\}$ denotes the eigenvalues of $\boldsymbol{\Phi}_j$. By the appropriate scaling of data $\boldsymbol{Z}_j$, the eigenvalues can also be scaled as $\lambda_i < 1, \forall i$. Thus, when $N_j^* > N_j^{(p)}$, the first term in (22) decreases. Similarly, in the case of the death move, the first term increases.

In the first term of (22), the constant L can be considered as the factor that controls the magnitude of the penalty, and an appropriate value should be selected. However, L is the number of frames in a block and is usually determined by the application. Therefore, in the proposed method, instead of using the actual value of L, it is replaced by the arbitrary constant κ in (22). The value of κ can be chosen experimentally to improve the performance.

VI. Experiment

A. Condition

Sensor observations were generated by convolving the measured room impulse responses with the source signal. In Sections VI-B and VI-C, Gaussian noise was employed as the source, whereas in Section VI-D, a speech signal was used. The room used for the measurement of the impulse response was a medium-sized meeting room (8 m × 9 m × 2.5 m) with a reverberation time of approximately 0.5 s. The sound sources were located on a circle of radius 1.5 m. The angular distance between the sources was 20° and the location of the source-set consisting of multiple sources with a 20° interval was randomly selected on the circle. The number of sources in the source-set is shown in Table II. Twenty observation blocks with different source-set locations were used for the estimation of the hierarchical model, i.e., $J = 20$. A microphone array with 8 elements mounted on the head of a robot was placed at the center of the circle. The microphone array configuration is shown in Fig. 4. Array manifold vectors $\{\boldsymbol{a}(\theta)\}$, which are candidates of the column vectors of $\boldsymbol{A}(\boldsymbol{\theta}_j)$, were prepared for the angle range $[-90°, +90°]$ at every 1°. The components of $\boldsymbol{a}(\theta)$ were generated by the Fourier transform of the direct part of the measured impulse response. The head-related transfer function (HRTF) of the robot was thereby considered. The impulse responses for the array manifold were measured independently of those used for generating the observations.

[1]An exception is the case when the locations of two or more of the sources are identical and the column of $\boldsymbol{\Phi}_j$ is linearly dependent.

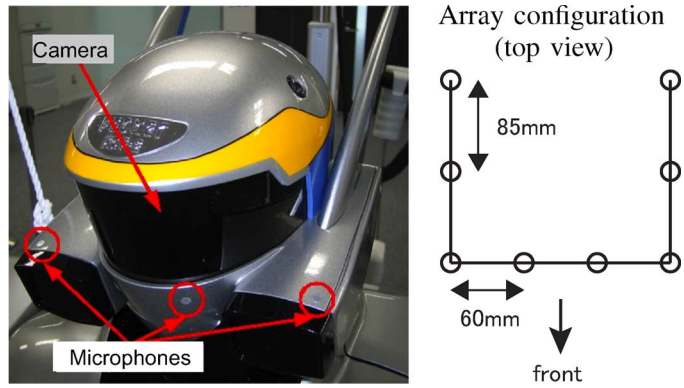

Fig. 4. Microphone array mounted on the head of robot HRP-2.

TABLE II
Parameters of the Experiment and the Signal Analysis

Parameter	Value
Number of sources	2 (VI-B, VI-D) / 1-3 (VI-C))
Source signal	Gaussian noise (VI-B, VI-C) / Speech (VI-D)
Sampling frequency	16 kHz
Frame length (STFT length)	32 ms (512 points)
Frame shift	8 ms (128 points)
Block length	0.2 s †
Number of iterations	1000
Frequency	1500 Hz (VI-B, VI-C) / 1250-2156Hz (VI-D)

†In the baseline experiment without the hierarchical model in Section VI-C (Case-A), a block length of 2.0 s was employed.

The parameters used for the signal analysis are summarized in Table II. The frequency range was selected such that the effectiveness of the proposed method could be demonstrated on the basis of the preliminary experiment described in [22]. In the lower frequency range, both the proposed and the conventional methods showed low spatial resolutions owing to a small phase difference between the microphones, which is a physical limitation of the array used. In the higher frequency range, the influence of the reverberation is considered to be smaller because at higher frequencies, the sound absorption by the walls of the room used in the experiment is larger. For the initial value $\boldsymbol{\theta}^{(1)}$, the ML estimate fluctuated by adding a Gaussian noise was employed. For the initial value $\boldsymbol{K}_j^{(1)}$, the identity matrix $\boldsymbol{I}$ was employed. The number of iterations (samples) was selected to be 1000 on the basis of the results of the preliminary test.

B. Hierarchical Model

Fig. 5 shows the variation of $\boldsymbol{\theta}_j^{(p)}$ over the course of the iterations. It is observed that the sample values converged on the true values (the dotted line) with a small number of iterations. The final estimate $\hat{\boldsymbol{\theta}}_j$ was obtained as the mean of the samples.

Fig. 6 compares the mean absolute error (MAE) of the proposed method with those of the conventional location estimators described in Appendix C. Thirty trials were conducted, i.e., $N_{trial} = 30$. MAE was calculated as $\mathrm{MAE} = 1/(J \times N_{trial}) \sum_{t,j} |\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j|$, where t indicates the trial index. The proposed method has the smallest MAE compared with the other

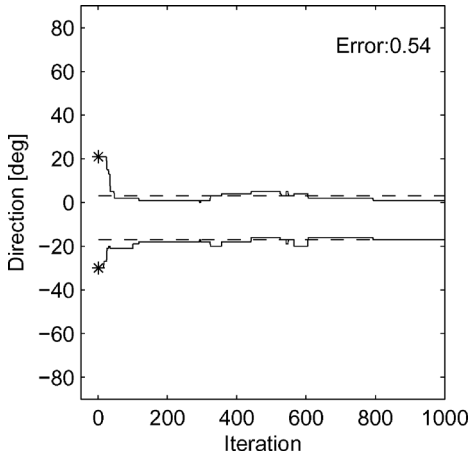

Fig. 5. Variation of $\boldsymbol{\theta}_j^{(p)}$ over the course of the iterations. The dashed lines show the true $\boldsymbol{\theta}_j$.

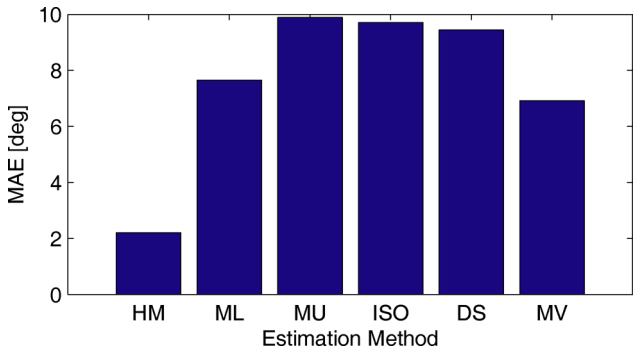

Fig. 6. MAE for the different parameter estimation methods. "HM" denotes the proposed method using hierarchical model. See Appendix C for the abbreviations of the other methods.

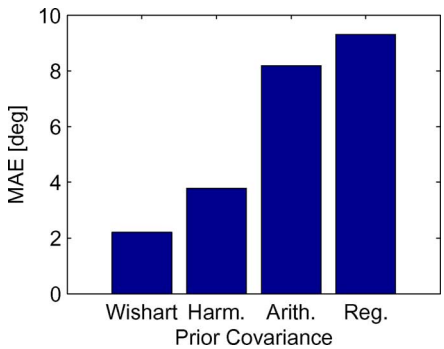

Fig. 7. MAE for the different methods used to obtain $\boldsymbol{K}_0$.

methods. It is worth noting that the MV adaptive beamformer, in which the noise model is adapted to the actual noise, shows the second smallest MAE. Therefore, we deduced that the estimation/adaptation of the noise model to the actual noise (reverberation) is essential to improve the spatial resolution under the current experimental condition. The difference in MAE between the proposed method and the MV beamformer is considered to be the difference in the precision of the model due to the amount of data available. The proposed method estimates the model using the data in J observation blocks with the hierarchical model while the MV beamformer estimates it from a single block.

Fig. 7 indicates MAE for the different methods used to obtain $\boldsymbol{K}_0$. "Wishart" corresponds to the proposed method described

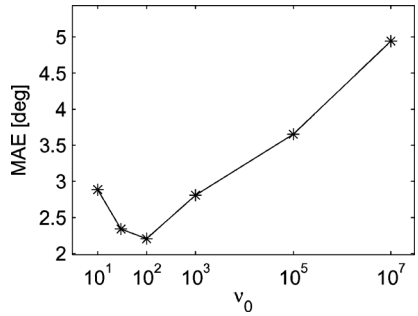

Fig. 8. MAE for different ν_0 values.

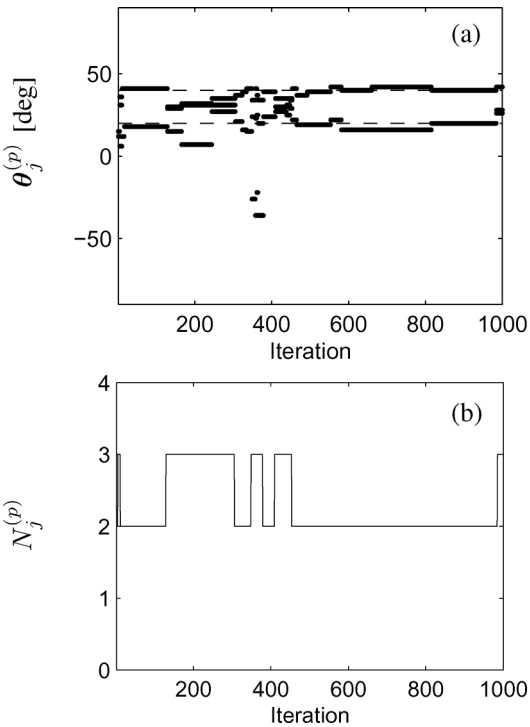

Fig. 9. Variation of $\boldsymbol{\theta}_j^{(p)}$ and $N_j^{(p)}$ over the course of the iterations. The dashed lines in (a) show the true $\boldsymbol{\theta}_j$.

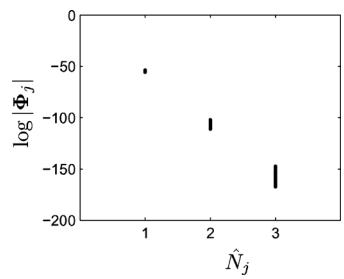

Fig. 10. Value of $\log|\boldsymbol{\Phi}_j|$ for different $\hat{N}_j$.

in Section IV-C. For "Harm," $\boldsymbol{\Lambda}$ in (18), which is the conditional mean of the Wishart distribution, was employed as $\boldsymbol{K}_0$. For "Arith," the arithmetic mean of $\{\boldsymbol{K}_1, \ldots, \boldsymbol{K}_J\}$ was employed. It should be noted that MAE was small for "Wishart" and "Harm." From these results, it can be deduced that the harmonic-mean-like operation in (18) is essential for the hierarchical modeling of $\boldsymbol{K}_0$. For "Reg," $c\boldsymbol{I}$ was employed as $\boldsymbol{K}_0$ where c is the scaling constant. In this case, only joint estimation without hierarchical modeling was conducted. The role of $\boldsymbol{K}_0 = c\boldsymbol{I}$ is the regularization of $\boldsymbol{C}_j$. By comparing "Wishart" and "Reg," it can be seen that the estimation performance was

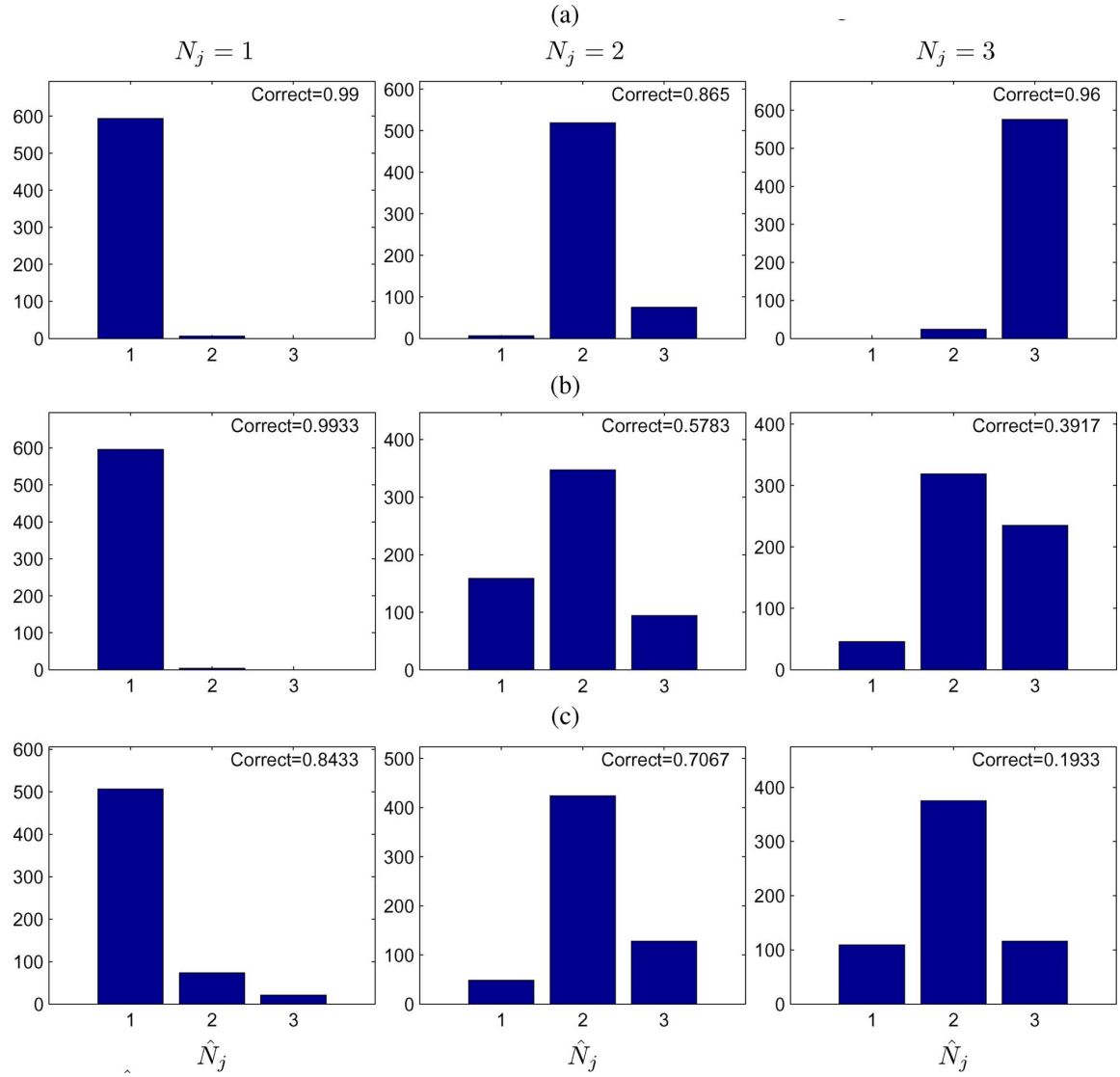

Fig. 11. Histograms of $\hat{N}_j$ for various true N_j when employing the reversible jump MCMC method. The value of "Correct" indicates the probability of a correct estimation of $\hat{N}_j$. (a) Case-A:Without hierarchical model, block length $= 2.0$ s. (b) Case-B: Without hierarchical model, block length $= 0.2$ s. (c) Case-C: With hierarchical model, block length $= 0.2$ s.

considerably improved by employing the proposed hierarchical model in the case where the quantity of data in a block was small.

Fig. 8 shows MAE when the value of ν_0 is fixed at a certain value during the iteration. It is observed that MAE has a minimum at around $\nu_0 = 10^2$. Regarding ν_0, it is difficult to obtain samples from (19). Therefore, this optimum value was employed in the other experiments.

C. Estimation of the Number of Sources

In this section, the estimation of the number of sources, $\hat{N}_j$, using the reversible jump MCMC method is examined.

First, the case without the hierarchical model was evaluated as a baseline (denoted as Case-A). A block length of 2.0 s, which is ten times longer than that of the other experiment, was employed. The number of blocks was $J = 1$. The true number of sources was $N_j = \{1, 2, 3\}$, and this number was selected from the uniform distribution. The estimate $N_j^{(p)}$ was also selected from the set {1, 2, 3}.

Fig. 9 shows an example of the variation of $N_j^{(p)}$ along with those of $\boldsymbol{\theta}_j^{(p)}$ over the course of the iterations. For the final estimate $\hat{N}_j$, $N_j^{(p)}$ with the highest frequency was employed. Then, $\{\boldsymbol{\theta}_j^{(p)}\}$ with $N_j^{(p)} = \hat{N}_j$ was averaged to obtain the final estimate $\hat{\boldsymbol{\theta}}_j$.

Fig. 10 shows the value of $\log|\boldsymbol{\Phi}_j|$ for different $\hat{N}_j$. From the figure, it can be seen that the value decreases as $\hat{N}_j$ increases while the variation in the same $\hat{N}_j$ is relatively small. From this observation, it can be understood that the term $|\boldsymbol{\Phi}_j|$ functions as the penalty in (22). The optimum value of κ was determined so that the sum of MAE for $\hat{\boldsymbol{\theta}}_j$ for all true $N_j = \{1, 2, 3\}$ is minimized.

Fig. 11(a) shows the histogram of $\hat{N}_j$ for 600 trials. From this, it can be seen that the correct $\hat{N}_j$ is estimated with high probability, as shown in the upper right corner of each panel.

Table III(a) shows MAE for different true N_j. MAE is defined as $(1/N_{trial})\sum_t |\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j|$. The values of C4 and C8 indicate the probabilities of MAE $\leq 4°$ and MAE $\leq 8°$, respectively. From

TABLE III
MAE AND THE PROBABILITY OF MAE BEING $\leq 4^\circ$ (C4) AND $\leq 8^\circ$ (C8) FOR DIFFERENT TRUE N_j WITH THE REVERSIBLE JUMP MCMC METHOD

(a) Case-A: Without hierarchical model, block length = 2.0 s			
N_j	MAE	C4	C8
1	13.01	0.76	0.77
2	**2.89**	0.80	0.95
3	7.54	0.39	0.70
(b) Case-B: Without hierarchical model, block length = 0.2 s			
N_j	MAE	C4	C8
1	34.34	0.13	0.19
2	**9.70**	0.48	0.75
3	11.71	0.41	0.72
(c) Case-C: With hierarchical model, block length = 0.2 s			
N_j	MAE	C4	C8
1	1.29	0.99	0.99
2	**4.44**	0.62	0.92
3	9.69	0.42	0.80

this table, it can be seen that MAE for $N_j = 2$ is comparable to that in Section VI-B, whereas MAE for $N_j = 1$ and $N_j = 3$ is slightly higher.

Next, the block length was reduced to 0.2 s (1/10th of that in Case-A), as employed in Section VI-B. In Case-B, the estimation was performed without the hierarchical model, whereas in Case-C, the hierarchical model was employed. In Case-C, the number of blocks was $J = 20$ as employed in Section VI-B. In a single trial (20 blocks), the true N_j was invariant while the location of the source-set was randomly selected (angular distance between the sources is 20°). Thirty trials were conducted so that the number of final estimates was the same as that in Case-A and Case-B.

Fig. 11(b) and (c) show the histograms of $\hat{N}_j$. For both Case-B and Case-C, the performance of the estimation of the number of sources was reduced to some extent by reducing the amount of data in each block.

Table III(b) and (c) show the MAE and C4/C8 values for $\boldsymbol{\theta}_j$. By comparing the values in (b) and (c) with that in (a), it can be seen that the case with the hierarchical model (Case-C) achieved a performance closer to the baseline case with a sufficient amount of data (Case-A) (e.g., when $N_j = 2$, MAE is 2.89, 9.68, and 4.44 in Case-A, B, and C, respectively, as indicated by the bold text in Table III.)

D. Time-Varying Source Signal

In this subsection, we examine the performance of the proposed method for a more realistic source signal. We chose speech as an example of a real source signal because speech signals are time varying and are sparse in the frequency domain. In this case, the number of *active* sources dynamically changes. This is a more realistic and challenging condition for the estimation of the number of sources.

As speech sources, single sentences included in the Japanese continuous speech corpus (JNAS [27]) were used. Because speech is time varying and sparse in the frequency domain as described above, the ratio of the direct signal to reverberation (hereafter denoted as DRR) dynamically changes. For example, the sections in which the direct sound of a consonant overlaps with the reverberation of a vowel in the previous sections have lower DRR. DRR also varies at different frequencies. Fig. 12 shows the distribution of DRR for all combinations of frequency bins and blocks. The number of frequency bins tested was 30 (1250–2156 Hz). DRR was calculated by splitting the measured room impulse into the direct sound and reverberation (reflection) components and convolving them with the speech source separately. When the direct sound is weak and DRR is low, the source is effectively off. In this experiment, the sources are assumed to be active when $DRR > DRR_{the}$, where DRR_{the} denotes the threshold, which was set at 0 dB. Fig. 13 shows the histogram of the number of active sources denoted as $\tilde{N}_j$. In the evaluation, the cases with $\tilde{N}_j = 0$ are omitted.

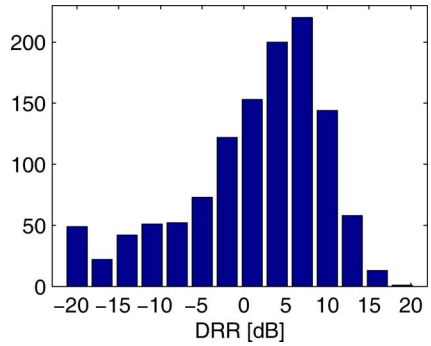

Fig. 12. DRR distribution of the observation for speech sources.

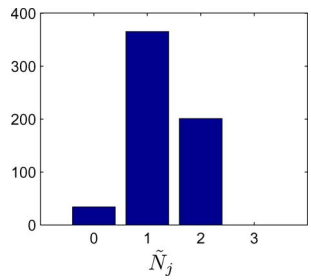

Fig. 13. Histgram of the number of active speech sources, $\tilde{N}_j$.

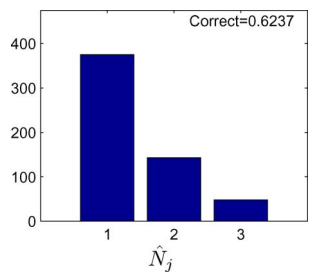

Fig. 14. Histogram of the estimated number of speech sources, $\hat{N}_j$.

TABLE IV
MAE AND THE PROBABILITY OF MAE BEING $\leq 4^\circ$ (C4) AND $\leq 8^\circ$ (C8) FOR SPEECH SOURCES

Method	MAE	C4	C8
Proposed	3.23	0.94	0.82
ML	17.2	0.68	0.54
MUSIC	37.18	0.57	0.48
ML†	4.56	0.90	0.78
MUSIC†	18.36	0.80	0.70

Fig. 14 shows the number of sources, $\hat{N}_j$, estimated by the proposed method. It can be seen that the histogram shown in Fig. 13 was approximately recovered. The correct rate was 62%. Table IV shows the MAE and C4/C8 values for the estimated direction $\hat{\boldsymbol{\theta}}_j$ together with those obtained using the

ML and MUSIC methods. For the ML and MUSIC methods, the number of sources was assumed to be two. For the case of ML†/MUSIC†, the number of active sources $\tilde{N}$, which is unknown in real applications, was given to the ML/MUSIC estimator. Thus, ML†/MUSIC† shows an achievable optimal performance for the ML/MUSIC estimator. It can be seen that the performance of the proposed method is slightly better than that of ML†/MUSIC†. The reason for the small difference between the proposed and ML†/MUSIC† methods relative to Fig. 6 is considered to be the trade-off that exists between the improvement in the directional resolution realized by the proposed method and an imperfect estimation of $\hat{N}_j$ using the reversible jump MCMC method.

VII. Discussion and Conclusion

In this study, a method for estimating the noise covariance matrix using a hierarchical model was proposed and applied to sound source localization in a reverberant environment. The experimental results showed that the spatial resolution was improved by the proposed method compared with the conventional methods, which assume a spatially white noise. The advantage of employing the hierarchical model is that high resolution can be achieved with a relatively small amount of data. Moreover, the reversible jump MCMC method was introduced into this method so that the number of sources would be jointly estimated. The experimental results indicated that the proposed method is effective in the case where the number of active sources dynamically changes.

In this paper, some basic aspects of the MCMC-based sound source localization technique were described and evaluated. However, for practical applications, further developments and detailed evaluation are required in the future. For application to speaker tracking, for example, an extension of the proposed method to dynamic environments is required. For this purpose, the unification of the MCMC and SMC approaches may be a possibility. In addition, a means of combining the results in each frequency bin should be addressed. The simplest way of combining them is to average the samples $\{\boldsymbol{\theta}_j^{(p)}\}$ with the same number of sources $N_j^{(p)}$ over the entire frequency range. However, the method of combination is application specific, and each individual application should therefore be discussed. Regarding the evaluation, real room reverberation (room impulse responses), a real microphone array mounted on the robot, and real speech source signals were used in the experiment discussed in Section VI-D. However, the movement of speakers, which is inevitable in speaker tracking applications, was not considered. The results of a brief test in which the proposed method is applied to the data recorded with moving speakers are available in [25][2]. However, a more detailed evaluation is required.

[2]It should be noted that when recorded observations are used instead of simulated data in the evaluation, the precision of the evaluation is limited to some extent. This is because the number of active sources $\tilde{N}_j$ in each frequency bin cannot be known. In [25], it was assumed that the estimated number of sources $\hat{N}_j$ was always true. This assumption affects the evaluation of the MAE of $\hat{\boldsymbol{\theta}}_j$

On the other hand, in applications such as machine diagnostics and sound source detection in disaster environments, the independent estimation of the number of sources and their locations in each frequency bin as in the proposed method is expected to be useful. The reason for this is that in these applications, the target sources (e.g., leaking sound of gas from a pipeline) may have certain resonant frequencies, and the identification of the frequency and location of sources will therefore lead to the source classifications. However, further evaluation in a realistic situation is also required.

Appendix A
Derivation of (8) and (9)

Using Bayes' theorem and the Assumptions A-1 and A-5,

$$\begin{aligned} &p(\boldsymbol{s}_{j,k}|\boldsymbol{Z}_j,\boldsymbol{\theta}_j,\boldsymbol{K}_j) \\ &\quad\propto p(\boldsymbol{z}_j|\boldsymbol{\theta}_j,\boldsymbol{s}_{j,k},\boldsymbol{K}_j)p(\boldsymbol{s}_{j,k}) \propto \exp\left(-\boldsymbol{s}_{j,k}^H\boldsymbol{\Phi}_0^{-1}\boldsymbol{s}_{j,k}\right) \\ &\qquad\cdot\exp\left[-\left(\boldsymbol{z}_{j,k}-\boldsymbol{A}_j\boldsymbol{s}_{j,k}\right)^H\boldsymbol{K}_j^{-1}\left(\boldsymbol{z}_{j,k}-\boldsymbol{A}_j\boldsymbol{s}_{j,k}\right)\right] \\ &\quad=\exp\left[-\boldsymbol{s}_{j,k}^H\left(\boldsymbol{A}_j^H\boldsymbol{K}_j^{-1}\boldsymbol{A}_j+\boldsymbol{\Phi}_0^{-1}\right)\boldsymbol{s}_{j,k}\right. \\ &\qquad\left.+\boldsymbol{s}_{j,k}^H\boldsymbol{A}_j^H\boldsymbol{K}_j^{-1}\boldsymbol{z}_{j,k}+\boldsymbol{z}_{j,k}^H\boldsymbol{K}_j^{-1}\boldsymbol{A}_j\boldsymbol{s}_{j,k}-\boldsymbol{z}_{j,k}^H\boldsymbol{K}_j^{-1}\boldsymbol{z}_{j,k}\right] \end{aligned} \tag{24}$$

Because the expressions inside of $\exp()$ have a quadratic form of $\boldsymbol{s}_{j,k}$, $p(\boldsymbol{s}_{j,k}|\boldsymbol{Z}_j,\boldsymbol{\theta}_j,\boldsymbol{K}_j)$ is a Gaussian distribution. Assuming this Gaussian distribution to be $\mathcal{N}(\boldsymbol{s}_{j,k};\boldsymbol{\mu}_{j,k},\boldsymbol{\Phi}_j)$, $p(\boldsymbol{s}_{j,k}|\boldsymbol{Z}_j,\boldsymbol{\theta}_j,\boldsymbol{K}_j)$ can also be expressed as follows:

$$\begin{aligned} &\mathcal{N}(\boldsymbol{s}_{j,k};\boldsymbol{\mu}_{j,k},\boldsymbol{\Phi}_j) \\ &\quad\propto\exp\left[-(\boldsymbol{s}_{j,k}-\boldsymbol{\mu}_{j,k})^H\boldsymbol{\Phi}_j^{-1}(\boldsymbol{s}_{j,k}-\boldsymbol{\mu}_{j,k})+C\right] \\ &\quad=\exp\left[-\boldsymbol{s}_{j,k}^H\boldsymbol{\Phi}_j^{-1}\boldsymbol{s}_{j,k}+\boldsymbol{s}_{j,k}^H\boldsymbol{\Phi_j}^{-1}\boldsymbol{\mu}_{j,k}\right. \\ &\qquad\left.+\boldsymbol{\mu}_{j,k}^H\boldsymbol{\Phi}_j^{-1}\boldsymbol{s}_{j,k}-\boldsymbol{\mu}_{j,k}^H\boldsymbol{\Phi}_j^{-1}\boldsymbol{\mu}_{j,k}+C\right] \end{aligned} \tag{25}$$

where C denotes a term that does not include $\boldsymbol{s}_{j,k}$ (a part of the normalizing constant). By comparing (24) with (25), the following terms can be identified as:

$$\boldsymbol{\Phi}_j=\left(\boldsymbol{A}_j^H\boldsymbol{K}_j^{-1}\boldsymbol{A}_j+\boldsymbol{\Phi}_0^{-1}\right)^{-1} \tag{26}$$

$$\boldsymbol{\mu}_{j,k}=\boldsymbol{\Phi}_j\boldsymbol{A}_j^H\boldsymbol{K}_j^{-1}\boldsymbol{z}_{j,k} \tag{27}$$

$$C=-\boldsymbol{z}_j^H\boldsymbol{P}_j\boldsymbol{z}_j \tag{28}$$

$$\boldsymbol{P}_j=\boldsymbol{K}_j^{-1}-\boldsymbol{K}_j^{-1}\boldsymbol{A}_j\boldsymbol{\Phi}_j\boldsymbol{A}_j^H\boldsymbol{K}_j^{-1} \tag{29}$$

It is worth noting that when the prior distribution of $\boldsymbol{s}_{j,k}$ is a uniform distribution ($p(\boldsymbol{s}_{j,k})=\text{const}$) instead of a Gaussian distribution, $\boldsymbol{\Phi}_0$ in (26) vanishes and (27) becomes the well-known ML beamformer (e.g., [3], extended to multiple sources):

$$\boldsymbol{\mu}_{j,k}=\left(\boldsymbol{A}_j^H\boldsymbol{K}_j^{-1}\boldsymbol{A}_j\right)^{-1}\boldsymbol{A}_j^H\boldsymbol{K}_j^{-1}\boldsymbol{z}_{j,k} \tag{30}$$

Furthermore, when the noise $\boldsymbol{n}_{j,k}$ is spatially white ($\boldsymbol{K}_j=c\boldsymbol{I}$) and the number of source is $N_j=1$, (30) is reduced to the delay-and-sum beamformer $\boldsymbol{\mu}_{j,k}=(\boldsymbol{a}_j^H\boldsymbol{a}_j)^{-1}\boldsymbol{a}_j^H\boldsymbol{z}_{j,k}$.

APPENDIX B
DERIVATION OF (21)

Using the Bayes' theorem and Assumptions A-3 and A-5,

$$p(\boldsymbol{\theta}_j, \boldsymbol{S}_j|\boldsymbol{Z}_j, \boldsymbol{K}_j) \propto p(\boldsymbol{Z}_j|\boldsymbol{\theta}_j, \boldsymbol{S}_j, \boldsymbol{K}_j)p(\boldsymbol{S}_j) = \prod_{k=1}^{K} p(\boldsymbol{z}_{j,k}|\boldsymbol{\theta}_j, \boldsymbol{s}_{j,k}, \boldsymbol{K}_j)p(\boldsymbol{s}_{j,k}) \quad (31)$$

By using this and Appendix A, the integration of $\boldsymbol{S}_j$ in (31) yields

$$\begin{aligned} & p(\boldsymbol{\theta}_j|\boldsymbol{Z}_j, \boldsymbol{K}_j) \\ & \propto \int p(\boldsymbol{\theta}_j, \boldsymbol{S}_j|\boldsymbol{Z}_j, \boldsymbol{K}_j)d\boldsymbol{S}_j \\ & = \prod_{k=1}^{K} \int p(\boldsymbol{z}_{j,k}|\boldsymbol{\theta}_j, \boldsymbol{s}_{j,k}, \boldsymbol{K}_j)p(\boldsymbol{s}_{j,k})d\boldsymbol{s}_{j,k} \\ & = \prod_{k=1}^{K} \exp\left(-\boldsymbol{z}_j^H \boldsymbol{P}_j \boldsymbol{z}_j\right) \\ & \quad \cdot \int \exp\left[-(\boldsymbol{s}_{j,k} - \boldsymbol{\mu}_{j,k})^H \boldsymbol{\Phi}_j^{-1}(\boldsymbol{s}_{j,k} - \boldsymbol{\mu}_{j,k})\right] d\boldsymbol{s}_{j,k} \quad (32) \end{aligned}$$

A general integration of the exponential of the quadratic form yields

$$\int \exp\left[-(\boldsymbol{s}_{j,k} - \boldsymbol{\mu}_{j,k})^H \boldsymbol{\Phi}_j^{-1}(\boldsymbol{s}_{j,k} - \boldsymbol{\mu}_{j,k})\right] d\boldsymbol{s}_{j,k} = \pi^N |\boldsymbol{\Phi}_j| \quad (33)$$

This result can be easily verified from the general definition of the complex Gaussian distribution. From this, (32) becomes

$$p(\boldsymbol{\theta}_j|\boldsymbol{Z}_j, \boldsymbol{K}_j) \propto \pi^{NK} |\boldsymbol{\Phi}_j|^K \exp\left[-\mathrm{tr}(\bar{\boldsymbol{R}}_j \boldsymbol{P}_j)\right] \quad (34)$$

where

$$\bar{\boldsymbol{R}}_j := \sum_{k=1}^{K} \boldsymbol{z}_{j,k} \boldsymbol{z}_{j,k}^H \quad (35)$$

APPENDIX C
CONVENTIONAL LOCATION ESTIMATOR

In this appendix, the conventional location estimators used for comparison in Section VI are briefly described. The ML estimator with white noise assumption [28] is given by

$$\hat{\boldsymbol{\theta}}_j = \arg\max \mathrm{tr}\left[\boldsymbol{A}(\boldsymbol{\theta})\left(\boldsymbol{A}^H(\boldsymbol{\theta})\boldsymbol{A}(\boldsymbol{\theta})\right)^{-1} \boldsymbol{A}^H(\boldsymbol{\theta})\bar{\boldsymbol{R}}_j\right] \quad (36)$$

Equation (36) is an N_j-dimensional maximization problem. The other conventional estimators employ a spatial spectrum approach, in which a single dimensional spectrum is first estimated, and then N_j peaks are detected. The MUSIC (MU) spectrum is given by

$$P_{MU}(\theta) = \frac{\|\boldsymbol{a}(\theta)\|^2}{\|\boldsymbol{a}^H(\theta)\boldsymbol{E}_n\|^2} \quad (37)$$

where $\boldsymbol{E}_n$ is the noise-subspace eigenvector matrix. The spatial spectrum for a general adaptive beamformer [29] is given by

$$P_{AD}(\theta) = \boldsymbol{w}^H \bar{\boldsymbol{R}}_j \boldsymbol{w}, \boldsymbol{w} = \frac{\boldsymbol{Q}^{-1}\boldsymbol{a}(\theta)}{\boldsymbol{a}^H(\theta)\boldsymbol{Q}^{-1}\boldsymbol{a}(\theta)} \quad (38)$$

The matrix $\boldsymbol{Q}$ is the covariance matrix, and its choice is dependent on the algorithm. The isotropic noise model (ISO), which is a set of many random waves propagating in all possible direction with equal probability [3], is often introduced in (38) by employing $\boldsymbol{Q}$ corresponding to the isotropic noise [9]. In this paper, $\boldsymbol{Q} = \sum_{\theta\in\Theta} \boldsymbol{a}(\theta)\boldsymbol{a}^H(\theta) + c\boldsymbol{I}$ is used where $\Theta = [-90^\circ, +90^\circ]$ with an interval of 1°, and c is the regularization constant. On the other hand, when employing $\boldsymbol{Q} = \bar{\boldsymbol{R}}_j$, (38) becomes the minimum variance (MV) adaptive beamformer [3]. When employing $\boldsymbol{Q} = \boldsymbol{I}$, (38) becomes the delay-and-sum (DS) beamformer. It should be noted that in the MV adaptive beamformer, the noise model is adapted to the actual noise, while other conventional methods (MU,ISO,DS) use a fixed noise model without adaptation.

ACKNOWLEDGMENT

The authors would like to thank the reviewers for their thorough review of the manuscript and useful advice, which contributed to significant improvement of the paper.

REFERENCES

[1] M. Wölfel and J. McDonough, *Distant Speech Recognition*. New York, NY, USA: Wiley, 2009.
[2] K. Nakamura, K. Nakadai, F. Asano, Y. Hasegawa, and H. Tsujino, "Intelligent sound source localization for dynamic environments," in *Proc. IROS '09*, 2009, pp. 664–669.
[3] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.
[4] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. AP-34, no. 3, pp. 276–280, Mar. 1986.
[5] G. C. Carter, "Time delay estimation for passive sonar signal processing," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. ASSP-29, no. 3, pp. 463–470, Jun. 1981.
[6] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 3, pp. 288–292, May 1997.
[7] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *J. Acoust. Soc. Amer.*, vol. 107, no. 1, pp. 384–391, 2000.
[8] R. Roy and T. Kailath, "ESPRIT – estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 37, no. 7, pp. 984–995, Jul. 1989.
[9] H. Cox, R. M. Zeskind, and T. Kooij, "Practical supergain," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. ASSP-34, no. , pp. 393–398, Jun. 1986.
[10] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 387–392, Apr. 1985.
[11] F. Asano, *Array Signal Processing for Acoustics (in Japanese)*. Kobe, Japan: Corona, 2011.
[12] K. Yamamoto, F. Asano, W. Rooijen, T. Yamada, and N. Kitawaki, "Estimation of the number of sound sources using support vector machine and its application to sound source separation," in *Proc. ICASSP '03*, 2003, vol. V, pp. 485–488.

[13] C. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.
[14] P. D. Hoff, *A First Course in Bayesian Statistical Methods*. New York, NY, USA: Springer, 2009.
[15] C. Andrieu and A. Doucet, "Joint Bayesian model selection and estimation of noisysinusoids via reversible jump MCMC," *IEEE Trans. Signal Process.*, vol. 47, no. 10, pp. 2667–2676, Oct. 1999.
[16] L. Stone, C. Barlow, and T. Corwin, *Bayesian Multiple Target Tracking*. Norwood, MA, USA: Artech House, 1999.
[17] *Sequential Monte Carlo Methods in Practice*, A. Doucent, N. de Freitas, and N. Gordon, Eds. New York, NY, USA: Springer, 2001.
[18] B. Ristic, S. Arulampalam, and N. Gordon, *Beyond the Kalman Filter*. Norwood, MA, USA: Artech House, 2004.
[19] D. B. Ward, A. Lehmann, and R. C. Williamson, "Particle filter algorithm for tracking an acoustic source in a reverberant environment," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 826–836, Nov. 2003.
[20] H. Asoh, I. Hara, F. Asano, and K. Yamamoto, "Tracking human speech events using a particle filter," in *Proc. ICASSP '05*, 2005, vol. II, pp. 1153–1156.
[21] *Kalman Filtering and Neural Networks*, S. Haykin, Ed. New York, NY, USA: Wiley Interscience, 2001.
[22] F. Asano and H. Asoh, "Joint estimation of sound source location and noise covariance in spatially colored noise," in *Proc. Eusipco '11*, 2011.
[23] F. Asano, H. Asoh, and K. Nakadai, "Sound source localization in spatially colored noise using a hierarchical Bayesian model," in *Proc. ICASSP '12*, 2012, pp. 193–196.
[24] P. J. Green, "Reversible jump MCMC computation and Bayesian model determination," *Biometrika*, vol. 82, pp. 711–732, 1995.
[25] F. Asano, H. Asoh, and K. Nakadai, "Estimation of the number of sources and their locations in colored noise using reversible jump MCMC," in *Proc. EUSIPCO '12*, 2012.
[26] G. Strang, *Linear Algebra and Its Application*. Orlando, FL, USA: Harcourt Brace Jovanovich Inc., 1988.
[27] *ASJ Continuous Speech Corpus – Japanese Newspaper Article Sentences (JNAS)* [Online]. Available: http://www.mibel.cs.tsukuba.ac.jp/_090624/jnas/instruct.html
[28] M. Miller and D. Fuhrmann, "Maximum-likelihood narrow-band direction finding and the EM algorithm," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 38, no. 9, pp. 1560–1577, Sep. 1990.
[29] H. Cox, R. M. Zeskind, and M. M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. ASSP-35, no. 10, pp. 1365–1376, Oct. 1987.

Futoshi Asano received the B.E. degree in electrical engineering, and the M.E. and Ph.D. degrees in electrical and communication engineering from Tohoku University, Sendai, Japan, in 1986, 1988, and 1991, respectively. From 1991 to 1995, he was a Research Associate with the Research Institute of Electrical Communication at Tohoku University. From 1993 to 1994, he stayed at The Applied Research Laboratory of Pennsylvania State University as a Visiting Researcher. Currently, he is a Senior Researcher with Intelligent Systems Research Institute at National Institute of Advanced Industrial Science and Technology, Tsukuba, Japan. He is also a Visiting Researcher at Honda Research Institute Japan, Co., Ltd. His research interests include array signal processing and statistical signal processing.

Hideki Asoh received his B.Eng. in mathematical engineering and M.Eng. in information engineering from the University of Tokyo, in 1981 and 1983 respectively. In April 1983, he joined the Electrotechnical Laboratory as a researcher. From 1993 to 1994 he stayed at the German National Research Center for Information Technology as a visiting research scientist. He is currently a chief senior researcher in Intelligent Systems Research Institute at National Institute of Advanced Industrial Science and Technology (AIST). His research interests are in constructing intelligent systems which can learn through interactions with the real-world.

Kazuhiro Nakadai received the B.E. in electrical engineering in 1993, the M.E. in information engineering in 1995, and the Ph.D. in electrical engineering in 2003 from the University of Tokyo. He had been worked with Nippon Telegraph and Telephone and NTT Comware Corporation from 1995 to 1999. He was a researcher with Kitano Symbiotic Systems Project, ERATO, Japan Science and Technology Agency (JST) from 1999 to 2003. He is currently a principal researcher for Honda Research Institute Japan, Co., Ltd. From 2006 to 2010, he was concurrently Visiting Associate Professor at Tokyo Institute of Technology, and he is Visiting Professor at Tokyo Institute of Technology since 2011. He also has had another position of Visiting Professor at Waseda University since 2011. His research interests include AI, robotics, signal processing, computational auditory scene analysis, multi-modal integration and robot audition.