Multiple Sound Source Localization Based on Stochastic Modeling of Spatial Gradient Spectra

Natsuki Ueno[1,2] and Hirokazu Kameoka[1]
[1]*NTT Communication Science Laboratories, Atsugi, Japan*
[2]*Tokyo Metropolitan University, Hino, Japan*
natsuki.ueno@ieee.org, hirokazu.kameoka.uh@hco.ntt.co.jp

***Abstract*—We propose source localization methods for multiple sound sources. The proposed method requires only an observation of a sound pressure and its spatial gradient at one fixed point, which can be realized by a small microphone array. The key idea is to utilize the partial differential equation relating the observed signals and the source position, which was originally proposed for the direct method for the single source localization problem. We extend this framework using stochastic modeling and proposed a method for the mutliple source localization in the presence of noises. Two source localization methods are proposed: one is the expectation-minimization algorithm for a given number of sources, and the other is the variational Bayesian inference for an unknown number of sources. By numerical experiments, the localization accuracies of the two proposed methods are compared with the baseline method.**

***Index Terms*—source localization, microphone array, expectation-maximization algorithm, variational Bayesian inference algorithm.**

I. Introduction

Sound source localization technique using a microphone array has a wide variety of applications such as sonar, robot audition, and hearing aids [1], [2]. Especially, it is one of the major challenges to localize multiple sources using a small microphone array. Various source localization methods have been proposed in the literature, such as Multiple Signal Classification (MUSIC) algorithm [3], Generalized Cross-Correlation methods with Phase Transform (GCC-PHAT) [4], method based on the sound source constraint partial differential equation (SSC-PDE) [5], stochastic methods considering reverberant envionment [6]–[9], and methods using deep neural network (DNN) [10], [11].

In most source localization methods, a large array with a large number of microphones is generally preferable since the source positions are estimated on the basis of the time difference of arrivals of signals observed by the multiple microphones. On the other hand, the SSC-PDE-based method, which utilizes the partial differential equation relating the observed signal and the source position, enables source localization only from an instantanesous observation of a sound pressure and its spatial gradient at one observation position. Therefore, this method can be realized theoretically by using a small microphone array with a small number of microphones. The SSC-PDE is formulated for a direct method of single source localization, and by assuming time-frequency sparsity of the source signals, it can be applied to each time-frequency component even in the presence of multiple sources [5]. However, its application to multiple source localization problem, including a clustering scheme of the estimated positions obtained for each time-frequency component, is still to be established. We also refer to DNN-based methods [10], [11], whose approaches are different from those based only on physical properties of the wave propagation, such as the SSC-PDE-based method.

Focusing on the capability of the SSC-PDE-based method, we extend its framework to multiple source localization problem in the presence of noises based on stochastic modeling of the observation. As well as the conventional SSC-PDE-based method, the proposed method is applicable to a small microphone array as long as a sound pressure and its spatial gradient at one observation point can be obtained, which has an advantage in practical feasibility over other methods used mainly with a large microphone array. We first derive the expectation-maximization (EM) algorithm [12] for a given number of sources, and then extend it to the variational Bayesian inference (VBI) algorithm [13] for an unknown number of sources. By numerical experiments of multiple source localization under a reverberant environment, the performance comparison of the proposed EM and VBI algorithms were investigated with the baseline MUSIC algorithm.

This work was supported by JST CREST Grant Number JPMJCR19A3, Japan.

II. Partial-Differential-Equation-Based Stochastic Modeling for Source Localization

As preliminaries for the proposed methods, this section describes the observation model for the multiple source localization problem, which relates the observation of a sound pressure and its spatial gradient to the multiple source positions.

A. Sound Source Constraint Partial Differential Equation

We begin by breifly introducing basic theories of the SSC-PDE, which was proposed by Ando et al. [5]. Let $\boldsymbol{r}_0 \in \mathbb{R}^3$ be the position vector of the sound source. By assuming the free-field (spherical-wave) propagation from the point source, the sound pressure at the observation position $\boldsymbol{r} \in \mathbb{R}^3$ and time $t \in \mathbb{R}$, denoted by $f(\boldsymbol{r}, t)$, is represented by

$$f(\boldsymbol{r},t) = \frac{1}{\|\boldsymbol{r}-\boldsymbol{r}_0\|} g\left(t - \frac{\|\boldsymbol{r}-\boldsymbol{r}_0\|}{c}\right), \tag{1}$$

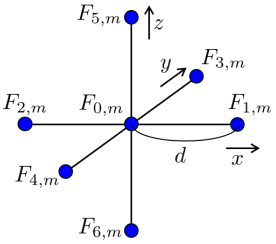


Fig. 1: Example of array configuration. For example, the gradient of F with respect to x direction is approximated by $F_{x,m} \approx (F_{1,m} - F_{2,m})/2d$.

where g is the source signal, c is the speed of sound, and $\|\cdot\|$ denotes the Euclidean norm. Therefore, the spatial gradient of f, denoted by ∇f, is given by

$$\nabla f(\boldsymbol{r}, t) = \frac{1}{R^2} g\left(t - \frac{R}{c}\right)\boldsymbol{n} + \frac{1}{cR}\dot{g}\left(t - \frac{R}{c}\right)\boldsymbol{n}, \tag{2}$$

where $R \coloneqq \|\boldsymbol{r} - \boldsymbol{r}_0\|$, $\boldsymbol{n} \coloneqq (\boldsymbol{r} - \boldsymbol{r}_0)/\|\boldsymbol{r} - \boldsymbol{r}_0\|$, and $\dot{g}$ denotes the temporal derivative of g. On the other hand, the temporal derivative of f, denoted by $\dot{f}$, is given by

$$\dot{f}(\boldsymbol{r}, t) = \frac{1}{\|\boldsymbol{r} - \boldsymbol{r}_0\|}\dot{g}\left(t - \frac{\|\boldsymbol{r} - \boldsymbol{r}_0\|}{c}\right). \tag{3}$$

From (2) and (3), we obtain the partial differential equation

$$\nabla f(\boldsymbol{r}, t) = \left(\frac{1}{R} f(\boldsymbol{r}, t) + \frac{1}{c}\dot{f}(\boldsymbol{r}, t)\right)\boldsymbol{n}, \tag{4}$$

which includes only the observed signal and its spatial and temporal gradients. This equation is referred to as the *sound source constraint partial differential equation*.

B. Stochastic Modeling for Single Sound Source

The SSC-PDE can be rewritten in the frequency domain as

$$-\nabla F(\boldsymbol{r}, \omega) + \left(\frac{1}{R} + \mathrm{j}\frac{\omega}{c}\right) F(\boldsymbol{r}, \omega)\boldsymbol{n} = 0, \tag{5}$$

where F is the temporal Fourier transform of f, $\omega \in \mathbb{R}$ denotes the angular frequency, and j denotes the imaginary unit. Here, suppose the temporal spectra of the sound pressure and its spatial gradient are observed by a small microphone array, for example, as shown in Fig. 1. Various other types of array, such as commercially available ambisonic microphone arrays [14] and acoustic vector sensors [15], [16], can also be used for this purpose. Let $\{\omega_m\}_m$ denote the set of discrete frequencies, and $F_{0,m}$ and $[F_{x,m}, F_{y,m}, F_{z,m}]^{\mathsf{T}}$ denote respectively the observation of $F(\boldsymbol{r}, \omega_m)$ and $\nabla F(\boldsymbol{r}, \omega_m)$ at a given fixed observation position $\boldsymbol{r}$. Then, we have where $\epsilon_{x,m}, \epsilon_{y,m}, \epsilon_{z,m} \in \mathbb{C}$ are the observation errors caused by the sensor noises and approximation of the spatial gradient by the spatial subtraction.

Here, we assume that $\epsilon_{x,m}, \epsilon_{y,m}, \epsilon_{z,m}$ follow independently the (circularly-symmetric) complex Gaussian distribution $\mathcal{N}_{\mathbb{C}}(0, \sigma_m^2)$ and that $F_{0,m}$ follows the complex Gaussian distribution $\mathcal{N}_{\mathbb{C}}(0, \sigma_{0,m}^2)$. Then, from (5), we have

$$\boldsymbol{C}_m(\theta)\boldsymbol{y}_m \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_m) \tag{6}$$

with $\theta \coloneqq \{R, \boldsymbol{n}\}$, i.e., θ denotes the source position, and

$$\boldsymbol{C}_m(\theta) \coloneqq \begin{bmatrix} -1 & & & \left(\frac{1}{R} + \mathrm{j}\frac{\omega_m}{c}\right) n_x \\ & -1 & & \left(\frac{1}{R} + \mathrm{j}\frac{\omega_m}{c}\right) n_y \\ & & -1 & \left(\frac{1}{R} + \mathrm{j}\frac{\omega_m}{c}\right) n_z \\ & & & 1 \end{bmatrix}, \tag{7}$$

$$\boldsymbol{y}_m \coloneqq [F_{0,m}, F_{x,m}, F_{y,m}, F_{z,m}]^{\mathsf{T}}, \tag{8}$$

$$\boldsymbol{\Sigma}_m \coloneqq \mathrm{diag}\left(\sigma_m^2, \sigma_m^2, \sigma_m^2, \sigma_{0,m}^2\right). \tag{9}$$

Since $\boldsymbol{C}(\theta)$ is invertible from $\det(\boldsymbol{C}(\theta)) = -1$, we obtain the probabilistic distribution of the observed signal $\boldsymbol{y}_m$ under the given θ as

$$p(\boldsymbol{y}_m|\theta) = \mathcal{N}_{\mathbb{C}}(\boldsymbol{y}_m; \boldsymbol{0}, \boldsymbol{C}_m(\theta)^{-1}\boldsymbol{\Sigma}_m\boldsymbol{C}_m(\theta)^{-\mathsf{H}}), \tag{10}$$

where $(\cdot)^{-\mathsf{H}}$ denotes the inverse of the Hermitian conjugate of a matrix.

C. Stochastic Modeling for Multiple Sound Sources

Various kinds of real-world acoustic signals such as human speech and music have sparsity in the time-frequency representation. Therefore, many practical situations satisfy well an time-frequency disjointness of the sources, which means each time-frequency component of the observed signal is assumed to be dominated by at most one sound source.

Let m, l, and $k \in \mathbb{N}$ denote indices of the frequency, time, and sound source, respectively, where $k = 0$ denotes the noise signal and $k \neq 0$ denotes the signal derived from the point source. We assume that only the $z_{m,l}$th source is active at the frequency m and time l. Then, under the given $z_{m,l} = k$, the probability density function of the observation of the sound pressure and its spatial gradient obtained by the short-time Fourier transform, denoted by $\boldsymbol{y}_{m,l} \coloneqq [F_{0,m,l}, F_{x,m,l}, F_{y,m,l}, F_{z,m,l}]^{\mathsf{T}} \in \mathbb{C}^4$, is given in a similar manner with (10) by

$$p(\boldsymbol{y}_{m,l}|z_{m,l} = k, \vartheta) = \mathcal{N}_{\mathbb{C}}(\boldsymbol{y}_{m,l}; \boldsymbol{0}, \boldsymbol{\Lambda}_{m,l}^{(k)}) \tag{11}$$

with

$$\boldsymbol{\Lambda}_{m,l}^{(k)} \coloneqq \begin{cases} \boldsymbol{C}_m(\theta^{(k)})^{-1}\boldsymbol{\Sigma}_{m,l}^{(k)}\boldsymbol{C}_m(\theta^{(k)})^{-\mathsf{H}} & (k \neq 0) \\ \nu_{m,l}^2\boldsymbol{W}_m & (k = 0) \end{cases}, \tag{12}$$

$$\boldsymbol{\Sigma}_{m,l}^{(k)} \coloneqq \mathrm{diag}\left(\sigma_m^2, \sigma_m^2, \sigma_m^2, \sigma_{0,m,l}^{(k)2}\right). \tag{13}$$

Here, $\theta^{(k)} \coloneqq \{R^{(k)}, \boldsymbol{n}^{(k)}\}$ denotes the position of the kth source, and ϑ denotes the set of model parameters consisting of $\{\theta^{(k)}\}_k$, $\{\sigma_m^2\}_m$, $\{\sigma_{0,m,l}^{(k)2}\}_{m,l,k}$, $\{\nu_{m,l}^2\}_{m,l}$. The covarianvce matrix for the noise component is modeled by $\nu_{m,l}^2\boldsymbol{W}_m$, where $\boldsymbol{W}_m$ represents the normalized time-invariant covariance and $\nu_{m,l}^2$ represents the power of the noise depending on both time and frequency. For example, we can define $\boldsymbol{W}_m$ for the diffuse noise field in accordance with [17].

III. Proposed Methods

On the basis of stochastic modeling described in Sec. II-C, we propose the EM and VBI algorithms for multiple sound source localization.

A. EM Algorithm for Given Number of Sources

Let $K \in \mathbb{N}$ denotes the number of sources, which is assumed to be given in this algorithm, and $z_{m,l}$ is assumed to follow the categorical distribution:

$$z_{m,l} \sim \mathrm{Categorical}(\pi_{m,l}^{(0)}, \pi_{m,l}^{(1)}, \dots, \pi_{m,l}^{(K)}). \tag{14}$$

Here, $\pi_{m,l}^{(0)}, \pi_{m,l}^{(1)}, \dots, \pi_{m,l}^{(K)} \in [0, \infty)$ are nonnegative parameters satisfying $\sum_{k=0}^{K} \pi_{m,l}^{(k)} = 1$. Then, the likelihood function with respect to ϑ and $\pi \coloneqq \{\pi_{m,l}^{(k)}\}_{m,l,k}$ is given by

$$p(Y|\vartheta, \pi) = \prod\nolimits_{m,l} \sum\nolimits_{k=0}^{K} \pi_{m,l}^{(k)} \mathcal{N}_{\mathbb{C}}(\boldsymbol{y}_{m,l}; \boldsymbol{0}, \boldsymbol{\Lambda}_{m,l}^{(k)}) \tag{15}$$

with $Y \coloneqq \{\boldsymbol{y}_{m,l}\}_{m,l}$. Therefore, the source localization problem can be reduced to the maximization problem of the log likelihood function:

$$\underset{\vartheta, \pi}{\text{maximize}} \ \log p(Y|\vartheta, \pi). \tag{16}$$

This maximization problem is difficult to solve in a closed form; however, the local solution can be obtained by the EM algorithm [12], where ϑ and π are updated alternately. The update rules are summarized in Appendix A.

B. VBI Algorithm for Unknown Number of Sources

The number of sources are often unknown in practical situations. For such cases, we extend the observation model in Sec. III-A to unknown number of sources using Dirichlet process mixture model in a similar manner with [18].

We consider infinitely many sources and assume $z_{m,l}$ follows the categorical distribution for countably infinite indices:

$$z_{m,l} \sim \mathrm{Categorical}(\{\pi_k\}_{k=0}^{\infty}). \tag{17}$$

Here, $\pi_k \in [0, \infty)$ is a nonnegative parameter satisfying $\sum_{k=0}^{\infty} \pi_k = 1$, which represents the probability that the kth source is active. The probability π_0 for the noise component is modeled as a random variable following the Beta distribution with hyper parameters $\alpha_{\mathrm{n}}, \beta_{\mathrm{n}} \in [0, \infty)$:

$$\pi_0 = v_0 \sim \mathrm{Beta}(\alpha_{\mathrm{n}}, \beta_{\mathrm{n}}). \tag{18}$$

Moreover, the probabilities $\{\pi_k\}_{k=1}^{\infty}$ for the point sources are modeled as random variables following the stick-breaking process [19] with a hyper parameter $\beta_{\mathrm{s}} \in [0, \infty)$:

$$\pi_k = v_k \prod\nolimits_{j=0}^{k-1} (1 - v_k), \tag{19}$$

$$v_k \sim \mathrm{Beta}(1, \beta_{\mathrm{s}}). \tag{20}$$

Then, the probability density function of the observed signals Y under the given parameters ϑ and $Z \coloneqq \{z_{m,l}\}_{m,l}$ is given by

$$p(Y|\vartheta, Z) = \prod\nolimits_{m,l} \mathcal{N}_{\mathbb{C}}(\boldsymbol{y}_{m,l}; \boldsymbol{0}, \boldsymbol{\Lambda}_{m,l}^{(z_{m,l})}). \tag{21}$$

Moreover, from the stick-breaking process, we have

$$p(Z|V) = \prod\nolimits_{m,l} \pi_{z_{m,l}} \quad \left(\pi_k = v_k \prod\nolimits_{j=0}^{k-1} (1 - v_k)\right), \tag{22}$$

$$p(V) = \mathrm{Beta}(v_0; \alpha_{\mathrm{b}}, \beta_{\mathrm{b}}) \prod\nolimits_{k=1}^{\infty} \mathrm{Beta}(v_k; 1, \beta_s) \tag{23}$$

with $V \coloneqq \{v_k\}_{k=0}^{\infty}$. Therefore, the joint distribution $p(\Theta, Y)$ and the posterior distribution $p(\Theta|Y)$ are obtained using the prior distribution $p(\vartheta)$ as

$$p(\Theta, Y) = p(Y|\vartheta, Z) p(Z|V) p(V) p(\vartheta), \tag{24}$$

$$p(\Theta|Y) = p(\Theta, Y)/p(Y), \tag{25}$$

where $\Theta \coloneqq \{\vartheta, Z, V\}$ denotes all the model parameters. The posterior distribution $p(\Theta|Y)$ is difficult to obtain in a closed form; however, its approximate distribution can be obtained by the VBI algorithm [13]. In the VBI algorithm, the distribution $q(\Theta)$ can be obtained iteratively so that the Kullback–Leibler (KL) divergence between $q(\Theta)$ and $p(\Theta|Y)$, defined as

$$\mathrm{KL}(q(\Theta)|p(\Theta|Y)) = \int q(\Theta) \log \frac{q(\Theta)}{p(\Theta, Y)} \, \mathrm{d}\Theta, \tag{26}$$

is locally minimized under the constraint that $q(\Theta)$ can be decomposed as

$$q(\Theta) = q(N) q(\rho) q(\lambda) q(\zeta) q(\gamma) q(Z) q(V). \tag{27}$$

Here, the model parameters N, ρ, λ, ζ, and γ are defined as

$$N = \{\boldsymbol{n}^{(k)}\}_k, \tag{28}$$

$$\rho = \{\rho^{(k)}\}_k \quad (\rho^{(k)} = 1/R^{(k)}), \tag{29}$$

$$\lambda = \{\lambda_m\}_m \quad (\lambda_m = 1/\sigma_m^2), \tag{30}$$

$$\zeta = \{\zeta_{m,l}^{(k)}\}_{m,l,k} \quad (\zeta_{m,l}^{(k)} = 1/\sigma_{0,m,l}^{(k)2}), \tag{31}$$

$$\gamma = \{\gamma_{m,l}\}_{m,l} \quad (\gamma_{m,l} = 1/\nu_{m,l}^2). \tag{32}$$

The distributions $q(N)$, $q(\rho)$, $q(\lambda)$, $q(\zeta)$, $q(\gamma)$, $q(Z)$, and $q(V)$ are updated alternately in the following forms:

$$q(N) = \prod\nolimits_{k=1}^{K'} \mathrm{vMF}(\boldsymbol{n}^{(k)}; \bar{\boldsymbol{\xi}}^{(k)}, \bar{\kappa}^{(k)}), \tag{33}$$

$$q(\rho) = \prod\nolimits_{k=1}^{K'} \mathcal{N}_{\mathbb{R}}(\rho^{(k)}; \bar{\mu}^{(k)}, \bar{\eta}^{(k)}), \tag{34}$$

$$q(\lambda) = \prod\nolimits_{m} \mathrm{Gamma}(\lambda_m; \bar{\alpha}_{\lambda,m}, \bar{\beta}_{\lambda,m}), \tag{35}$$

$$q(\zeta) = \prod\nolimits_{m,l} \prod\nolimits_{k=1}^{K'} \mathrm{Gamma}(\zeta_{m,l}^{(k)}; \bar{\alpha}_{\zeta,m,l}^{(k)}, \bar{\beta}_{\zeta,m,l}^{(k)}), \tag{36}$$

$$q(\gamma) = \prod\nolimits_{m,l} \mathrm{Gamma}(\gamma_{m,l}; \bar{\alpha}_{\gamma,m,l}, \bar{\beta}_{\gamma,m,l}), \tag{37}$$

$$q(Z) = \prod\nolimits_{m,l} \mathrm{Categorical}(z_{m,l}; \{\bar{\pi}_{m,l}^{(k)}\}_{k=0}^{\infty}), \tag{38}$$

$$q(V) = \prod\nolimits_{k=0}^{K'} \mathrm{Beta}(v_k; \bar{\alpha}_v^{(k)}, \bar{\beta}_v^{(k)}), \tag{39}$$

where $\mathrm{vMF}(\cdot)$, $\mathcal{N}_{\mathbb{R}}(\cdot)$, and $\mathrm{Gamma}(\cdot)$ denotes the von Mises–Fisher distribution [20], real Gaussian distribution, and gamma distribution, respctively. Note that the classes of the distributions from (33) to (37) are kept the same in the iterations by defining the initial distributions also in the above classes. Here, $K' \in \mathbb{N}$ represents the truncation of the stick-breaking process, which corresponds to the assumption $q(z_{m,l} = k) = 0$ $(k \geq K' + 1)$. Note that this truncation does not fix the complexity of the model but simply restricts the function space for $q(\Theta)$ to a certain extent. Therefore, we can aproximate $q(\Theta)$ well by setting large K'. The update rules for parameters in (33) to (37) are summarized in Appendix B.

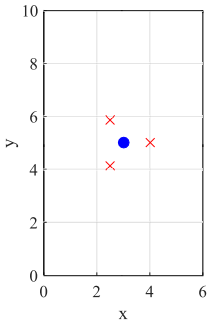


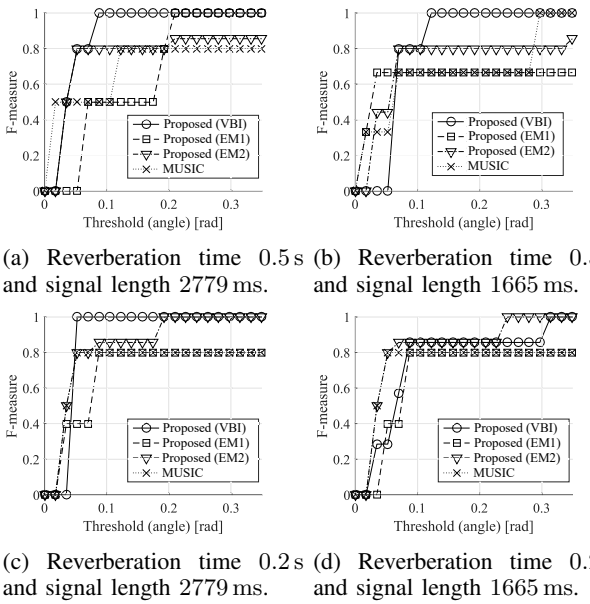
Fig. 2: Experimental settings. The blue circle denotes the center position of the microphone array, and the red crosses denote the true source positions.

IV. Numerical Experiments

We conducted numerical simulations of multiple source localization in a reverberant environment. A microphone array with seven microphones, whose configuration is given by Fig. 1 with $d = 0.03\,\mathrm{m}$, was located in the center of a rectangular room, whose size was $6\,\mathrm{m} \times 10\,\mathrm{m} \times 4\,\mathrm{m}$. The reflection coefficients were set as 0.7308 and 0.4566 so that the reverberation times from Sabine's formula were $0.5\,\mathrm{s}$ and $0.2\,\mathrm{s}$, respectively. Three sound sources were located around the microphone array, whose positions were $(1, 0, 0)\,\mathrm{m}$, $(-0.5, 0.87, 0)\,\mathrm{m}$, and $(-0.5, -0.87, 0)\,\mathrm{m}$ from the center of the room as shown in Fig. 2. Human speeches from SRV-DB [21] were used as the source signals. The sampling frequency of the microphones were $32\,\mathrm{kHz}$, and the frame length of the short-time Fourier transform were $64\,\mathrm{ms}$ with the half overlap.

We compared the four algorithms: the proposed VBI algorithm (denoted by VBI), the proposed EM algorithm with the correct number of sources $K = 3$ (denoted by EM1), the proposed EM algorithm with an incorrect number of sources $K = 6$ (denoted by EM2), and the conventional MUSIC algorithm as a baseline method. For a simple comparison with the MUSIC algorithm, only the source directions were evaluated and they were estimated on the xy-plane. The sound source was detected if $\sum_{m,l} \pi_{m,l}^{(k)} \boldsymbol{y}_{m,l}$ in the EM algorithm, $\sum_{m,l} \bar{\pi}_{m,l}^{(k)} \boldsymbol{y}_{m,l}$ in the VBI algorithm, or the angular spectrum in the MUSIC algorithm was larger than the threshold value. In the proposed VBI algorithm, the estimated source direction was given by the posterior mean, i.e., $\bar{\boldsymbol{\xi}}^{(k)}$. The accuracy of the source localization was evaluated by the F-measure, where the true positive was defined as the number of the estimated sources whose direction was within an angle $\pm\tau$ from some true source direction, the false positive was defined as the number of the estimated sources whose direction was not within an angle $\pm\tau$ from any true source direction, and the true negative was defined as the number of the true sources whose direction was not within an angle $\pm\tau$ from any estimated source direction.

Figure 3 shows the F-measure against the angle τ. Here, for each τ, different threshold values for the source detection



(a) Reverberation time $0.5\,\mathrm{s}$ and signal length $2779\,\mathrm{ms}$. (b) Reverberation time $0.5\,\mathrm{s}$ and signal length $1665\,\mathrm{ms}$.

(c) Reverberation time $0.2\,\mathrm{s}$ and signal length $2779\,\mathrm{ms}$. (d) Reverberation time $0.2\,\mathrm{s}$ and signal length $1665\,\mathrm{ms}$.

Fig. 3: Estimation accuracy with respect to source directions.

were investigated, and the highest F-measures were plotted. In most cases, the proposed methods achieved close or higher accuracies than the baseline MUSIC algorithm. Among the proposed methods, the VBI algorithm achieved the highest F-measures in most cases, and the number of sources and their directions were estimated accurately even though no assumption on the number of sources was required.

V. Conclusion

We proposed sound source localization methods for multiple source sources. The proposed methods and the baseline MUSIC algorithm were evaluated and compared by the numerical experiments, and their results indicated that the proposed VBI algorithm was able to estimate both the number of sources and their directions. Expetimental comparison with other stochastic methods, further evaluation of source localization including the source distances, and an online extension of the proposed algorithm are considered as future works.

Appendix

A. Update Rules in Proposed EM Algorithm

The update rules in the proposed EM algorithms are as below.

$$\boldsymbol{n}^{(k)} \leftarrow \frac{\sum_{m,l} \pi_{m,l}^{(k)} \lambda_m \mathrm{Re}\left[\left(\rho^{(k)} + \mathrm{j}\frac{\omega_m}{c}\right) F_{0,m,l} \tilde{\boldsymbol{f}}_{m,l}^{*}\right]}{\left\|\sum_{m,l} \pi_{m,l}^{(k)} \lambda_m \mathrm{Re}\left[\left(\rho^{(k)} + \mathrm{j}\frac{\omega_m}{c}\right) F_{0,m,l} \tilde{\boldsymbol{f}}_{m,l}^{*}\right]\right\|}, \quad (40)$$

$$\rho^{(k)} (= 1/R^{(k)}) \leftarrow \frac{\sum_{m,l} \pi_{m,l}^{(k)} \lambda_m \mathrm{Re}\left[F_{0,m,l} \tilde{\boldsymbol{f}}_{m,l}^{\mathsf{H}}\right] \boldsymbol{n}^{(k)}}{\sum_{m,l} \pi_{m,l}^{(k)} \lambda_m |F_{0,m,l}|^2}, \quad (41)$$

$$\lambda_m (= 1/\sigma_m^2) \leftarrow 3 \sum_l \sum_{k=1}^{K} \pi_{m,l}^{(k)} / \sum_l \sum_{k=1}^{K} \pi_{m,l}^{(k)} A, \quad (42)$$

$$\zeta_{m,l}^{(k)} (= 1/\sigma_{0,m,l}^{(k)2}) \leftarrow 1/|F_{0,m,l}|^2, \quad (43)$$

$$\gamma_{m,l}(=1/\nu_{m,l}^2) \leftarrow 1/\boldsymbol{y}_{m,l}^{\mathsf{H}}\boldsymbol{W}_m^{-1}\boldsymbol{y}_{m,l}, \tag{44}$$

$$\pi_{m,l}^{(k)} \leftarrow \exp(\phi_{m,l}^{(k)})/\textstyle\sum_{j=0}^{\infty}\exp(\phi_{m,l}^{(j)}). \tag{45}$$

Here, $\tilde{\boldsymbol{f}}_{m,l} \coloneqq [F_{x,m,l}, F_{y,m,l}, F_{z,m,l}]^{\mathsf{T}}$, and $\phi_{m,l}^{(k)}$ and A are defined as

$$\phi_{m,l}^{(0)} = 4\log\gamma_{m,l} - \log\det(\boldsymbol{W}_m) - \gamma_{m,l}\boldsymbol{y}_{m,l}^{\mathsf{H}}\boldsymbol{W}_m^{-1}\boldsymbol{y}_{m,l}, \tag{46}$$

$$\phi_{m,l}^{(k)} = 3\log\lambda_m + \log\zeta_{m,l}^{(k)} - \lambda_m A - \zeta_{m,l}^{(k)}|F_{0,m,l}|^2 \quad (k\neq 0), \tag{47}$$

$$A = (\rho^{(k)2} + \omega_m^2/c^2)|F_{0,m,l}|^2 - 2\mathrm{Re}[(\rho^{(k)} + \mathrm{j}\omega_m/c)F_{0,m,l}\tilde{\boldsymbol{f}}_{m,l}^{\mathsf{H}}]\boldsymbol{n}^{(k)} + \|\tilde{\boldsymbol{f}}_{m,l}\|^2. \tag{48}$$

B. Update Rules in Proposed VBI Algorithm

The update rules in the proposed VBI algorithms are as below. Variables without overlines (e.g., $\boldsymbol{\xi}^{(k)}$ for $\bar{\boldsymbol{\xi}}^{(k)}$) denote parameters for the initial distribution.

$$\bar{\boldsymbol{\xi}}^{(k)} \leftarrow \boldsymbol{a}/\|\boldsymbol{a}\|,\ \bar{\kappa}^{(k)} \leftarrow \|\boldsymbol{a}\|, \tag{49}$$

$$\bar{\mu}^{(k)} \leftarrow B/C,\ \bar{\eta}^{(k)} \leftarrow C, \tag{50}$$

$$\bar{\alpha}_{\lambda,m} \leftarrow 3\textstyle\sum_l\sum_{k=1}^{K'}\bar{\pi}_{m,l}^{(k)} + \alpha_{\lambda,m}, \tag{51}$$

$$\bar{\beta}_{\lambda,m} \leftarrow \textstyle\sum_l\sum_{k=1}^{K'}\bar{\pi}_{m,l}^{(k)}D + \beta_{\lambda,m}, \tag{52}$$

$$\bar{\alpha}_{\zeta,m,l}^{(k)} \leftarrow \bar{\pi}_{m,l}^{(k)} + \alpha_{\zeta,m,l}^{(k)},\ \bar{\beta}_{\zeta,m,l}^{(k)} \leftarrow \bar{\pi}_{m,l}^{(k)}|Y_{0,m,l}|^2 + \beta_{\zeta,m,l}^{(k)}, \tag{53}$$

$$\bar{\alpha}_{\gamma,m,l} \leftarrow 4\bar{\pi}_{m,l}^{(0)} + \alpha_{\gamma,m,l}, \tag{54}$$

$$\bar{\beta}_{\gamma,m,l} \leftarrow 4\bar{\pi}_{m,l}^{(0)}\boldsymbol{y}_{m,l}^{\mathsf{H}}\boldsymbol{W}_m^{-1}\boldsymbol{y}_{m,l} + \beta_{\gamma,m,l}, \tag{55}$$

$$\bar{\pi}_{m,l}^{(k)} \leftarrow \exp(\bar{\phi}_{m,l}^{(k)})/\textstyle\sum_{j=0}^{\infty}\exp(\bar{\phi}_{m,l}^{(j)}), \tag{56}$$

$$\bar{\alpha}_v^{(k)} \leftarrow \begin{cases}\sum_{m,l}\bar{\pi}_{m,l}^{(0)} + \alpha_{\mathrm{n}} & (k=0)\\ \sum_{m,l}\bar{\pi}_{m,l}^{(k)} + 1 & (k\neq 0)\end{cases}, \tag{57}$$

$$\bar{\beta}_v^{(k)} \leftarrow \begin{cases}\beta_{\mathrm{n}} & (k=0)\\ \sum_{m,l}(1 - \sum_{j=0}^{k}\bar{\pi}_{m,l}^{(j)}) + \beta_{\mathrm{s}} & (k\neq 0)\end{cases}. \tag{58}$$

Here, $\bar{\phi}_{m,l}^{(k)}$, $\boldsymbol{a}$, B, C, and D are defined as

$$\bar{\phi}_{m,l}^{(0)} = 4(\Psi(\bar{\alpha}_{\gamma,m,l}) - \log\bar{\beta}_{\gamma,m,l}) + \Psi(\bar{\alpha}_v^{(k)}) - \Psi(\bar{\alpha}_v^{(k)} + \bar{\beta}_v^{(k)}) - \log\det(\boldsymbol{W}_m) - (\bar{\alpha}_{\gamma,m,l}/\bar{\beta}_{\gamma,m,l})\boldsymbol{y}_{m,l}^{\mathsf{H}}\boldsymbol{W}_m^{-1}\boldsymbol{y}_{m,l}, \tag{59}$$

$$\bar{\phi}_{m,l}^{(k)} = 3(\Psi(\bar{\alpha}_{\lambda,m}) - \log\bar{\beta}_{\lambda,m}) + \Psi(\bar{\alpha}_{\zeta,m,l}^{(k)}) - \log\bar{\beta}_{\zeta,m,l}^{(k)} - (\bar{\alpha}_{\lambda,m}/\bar{\beta}_{\lambda,m})D - (\bar{\alpha}_{\zeta,m,l}^{(k)}/\bar{\beta}_{\zeta,m,l}^{(k)})|F_{0,m,l}|^2 + \Psi(\bar{\alpha}_v^{(k)}) + \Psi(\bar{\beta}_v^{(k)}) - 2\Psi(\bar{\alpha}_v^{(k)} + \bar{\beta}_v^{(k)}), \tag{60}$$

$$\boldsymbol{a} = 2\textstyle\sum_{m,l}\bar{\pi}_{m,l}^{(k)}\dfrac{\bar{\alpha}_{\lambda,m}}{\bar{\beta}_{\lambda,m}}\mathrm{Re}\Big[\Big(\bar{\mu}^{(k)} + \mathrm{j}\dfrac{\omega_m}{c}\Big)F_{0,m,l}\tilde{\boldsymbol{f}}_{m,l}^{*}\Big] + \kappa^{(k)}\boldsymbol{\xi}^{(k)}, \tag{61}$$

$$B = 2\textstyle\sum_{m,l}\bar{\pi}_{m,l}^{(k)}\dfrac{\bar{\alpha}_{\lambda,m}}{\bar{\beta}_{\lambda,m}}\mathrm{Re}\left[F_{0,m,l}\tilde{\boldsymbol{f}}_{m,l}^{\mathsf{H}}\right]\bar{\boldsymbol{\xi}}^{(k)} + \eta^{(k)}\mu^{(k)}, \tag{62}$$

$$C = 2\textstyle\sum_{m,l}\bar{\pi}_{m,l}^{(k)}\dfrac{\bar{\alpha}_{\lambda,m}}{\bar{\beta}_{\lambda,m}}|F_{0,m,l}|^2 + \eta^{(k)}, \tag{63}$$

$$D = (\bar{\mu}^{(k)2} + \bar{\eta}^{(k)} + \omega_m^2/c^2)|F_{0,m,l}|^2 - 2\mathrm{Re}[(\bar{\mu}^{(k)} + \mathrm{j}\omega_m/c)F_{0,m,l}\tilde{\boldsymbol{f}}_{m,l}^{\mathsf{H}}]\bar{\boldsymbol{\xi}}^{(k)} + \|\tilde{\boldsymbol{f}}_{m,l}\|^2, \tag{64}$$

where $\Psi(\cdot)$ is the digamma function.

REFERENCES

[1] I. E. Robert, "Robust sound localization: An application of an auditory perception system for a humanoid robot," *MIT Department of Electrical Engineering and Computer Science*, 1995.

[2] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, Springer, New York, 2001.

[3] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.

[4] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, 1976.

[5] S. Ando, N. Ono, and T. Nara, "Direct algebraic method for sound source localization with finest resolution both in time and frequency," in *Proc. Int. Congress Sound Vibration*, Cairns, July 2007, pp. 3273–3280.

[6] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 382–394, 2010.

[7] O. Schwartz and S. Gannot, "Speaker tracking using recursive EM algorithms," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 122, no. 2, pp. 392–402, 2014.

[8] Z. Zohny, S. Naqvi, and J. Chambers, "Variational EM for clustering interaural phase cues in MESSL for blind source separation of speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, South Brisbane, April 2015, pp. 3966–3970.

[9] Y. Soussana and S. Gannot, "Variational inference for DOA estimation in reverberant conditions," in *Proc. IEEE European Signal Process. Conf.*, A Coruna, September 2019.

[10] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 1, pp. 34–48, 2019.

[11] S. Chakrabarty and E. A. P. Habets, "Multi-speaker DOA estimation using deep convolutional networks trained with noise signals," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 1, pp. 8–21, 2019.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from imcomplete data via the em algorithm," *J. Royal Statistical Soc., B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[13] H. Attias, "Inferring parameters and structure of latent variable models by variational Bayes," in *Proc. the Fifteenth Conf. Uncertainty in Artificial Intelligence*, Stockholm, July 1999, pp. 21–30.

[14] M. A. Poletti, "Three-dimensional surround sound systems based on spherical harmonics," *J. Audio Eng. Soc.*, vol. 53, no. 11, pp. 1004–1025, 2005.

[15] A. Nehorai and E. Paldi, "Acoustic vector-sensor array processing," *IEEE Trans. Signal Process.*, vol. 42, no. 9, pp. 2481–2491, 1994.

[16] D. Levin, E. A. P. Habets, and S. Gannot, "Maximum likelihood estimation of direction of arrival using an acoustic vector-sensor," *J. Acoust. Soc. Am.*, vol. 131, no. 2, pp. 1240–1248, 2012.

[17] R. K. Cook, R. V. Waterhouse, R. D. Berendt, S. Edelman, and M. C. T. Thompson Jr., "Measurement of correlation coefficients in reverberant sound fields," *J. Acoust. Soc. Am.*, vol. 27, no. 6, pp. 1072–1077, 1955.

[18] H. Kameoka, M. Sato, T. Ono, N. Ono, and S. Sagayama, "Bayesian nonparametric approach to blind separation of infinitely many sparse sources," *IEICE Trans. Fundamentals*, vol. E96-A, no. 10, pp. 1928–1937, 2013.

[19] J. Sethuraman, "A constructive definition of Dirichlet priors," *Statistica Sinica*, vol. 4, no. 2, pp. 639–650, 1994.

[20] K. V. Mardia and P. E. Jupp, *Directional Statistics*, John Wiley & Sons, Chichester, 2009.

[21] K. Takahashi, "SRV-DB," [Online], http://www.it.cei.uec.ac.jp/SRV-DB, (Accessed on March 7, 2022).