



# Bayesian Extension of MUSIC for Sound Source Localization and Tracking

Takuma Otsuka<sup>1</sup>, Kazuhiro Nakadai<sup>2</sup>, Tetsuya Ogata<sup>1</sup>, Hiroshi G. Okuno<sup>1</sup>

<sup>1</sup>Graduate School of Informatics, Kyoto University, Kyoto, Japan

<sup>2</sup>Honda Research Institute Japan, Co., Ltd., Saitama, Japan

{ohtsuka, ogata, okuno}@kuis.kyoto-u.ac.jp, nakadai@jp.honda-ri.com

## Abstract

This paper presents a Bayesian extension of MUSIC-based sound source localization (SSL) and tracking method. SSL is important for distant speech enhancement and simultaneous speech separation for improving speech recognition, as well as for auditory scene analysis by mobile robots. One of the drawbacks of existing SSL methods is the necessity of careful parameter tunings, e.g., the sound source detection threshold depending on the reverberation time and the number of sources. Our contribution consists of (1) automatic parameter estimation in the variational Bayesian framework and (2) tracking of sound sources with reliability. Experimental results demonstrate our method robustly tracks multiple sound sources in a reverberant environment with RT20 = 840 (ms).

**Index Terms:** simultaneous sound source localization, MUSIC algorithm, variational Bayes, particle filter

## 1. Introduction

Auditory information holds an important place in the human perception. Needless to say oral speech as a communication channel, humans perceive audio signals emitted by surrounding objects to understand their situation. For example, the sound of footsteps may inform people that somebody is approaching or moving away without any glance. Achieving a computational auditory function will help people, especially hearing-impaired people, to have enhanced auditory awareness by showing trajectories of audio sources or presenting enhanced speech signals [1].

Sound source localization (SSL) is the most fundamental and important function for distant speech enhancement and simultaneous speech separation using a microphone array [2], presentation of sound sources to the operator of a tele-presence robot [3], and the detection and mapping of sound sources by a mobile robot [4]. Figure 1 illustrates a robot standing in an auditory dynamic environment. There may be moving and multiple sound sources surrounding the robot or the microphone array system. These system should robustly localize and track each sound source without a time-consuming parameter tuning.

For SSL with a microphone array, two methods have been widely exploited; beamforming [5] and multiple signal classification (MUSIC) [6, 7, 8]. Between these two methods, MUSIC is said to produce better SSL performances because the evaluation function for the direction of arrival detection called MUSIC spectrum has much sharper peaks at the directions of sound sources than the evaluation function of the beamforming method. Furthermore, MUSIC is capable of detecting multiple sound sources on condition that the number of sound sources is less than that of microphones.

In the frame work of MUSIC-based SSL, the threshold should be carefully set for the MUSIC spectrum to detect active

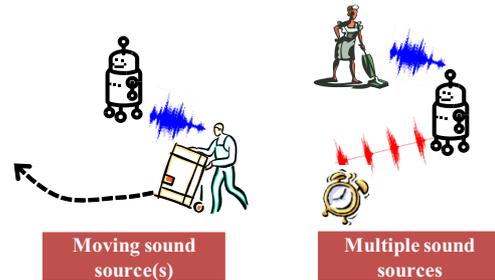


Figure 1: Sound source localization in a dynamic environment

sound sources. The problem is that this threshold is inevitably dependent on the number of sound sources and the reverberation time of the environment. The estimation of the number of sound sources have so far been tackled with Akaike information criterion [8] or a support vector machine [9]. However, an elaborate setting of the threshold is still necessary for the robust detection and tracking of sound sources. Typically, the threshold should be empirically tuned by looking into the MUSIC spectrum of the recording of the environment in question.

This paper presents a Bayesian extension of MUSIC-based multiple sound source localization and tracking. This method dispenses with most parts of the manual and empirical parameter tuning that is critical to existing frameworks. Our method consists of two stages: (1) The parameters for the localization and tracking are automatically estimated using a pre-recorded audio signal as the learning data based on the variational Bayesian hidden Markov model (VB-HMM) [10]. (2) Our method incrementally localizes and tracks multiple sound sources with previously estimated parameters based on a particle filter [11].

## 2. MUSIC-based SSL

This section specifies the problem and explains the MUSIC algorithm in general. We assume the SSL problem on the azimuth plane using a circular microphone array, as illustrated by Figure 2. In our configuration, the localization resolution is set 5 (deg). The problem statement is given as follows:

**Input:**  $M$ -channel audio signal,  $D$  transfer functions<sup>1</sup> for each direction and frequency bin,

**Output:**  $N$  directions where sound sources exist,

**Assumption:** the maximum number of sources is less than the number of microphones ( $N \leq N_{max} < M$ ).

<sup>1</sup> $D = 72$  in our implementation since the localization resolution is 5 (deg) in Fig. 2

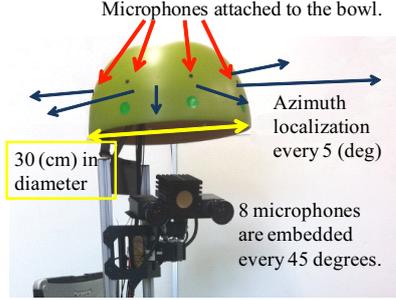


Figure 2: SSL on the azimuth plane using an 8-channel MEMS microphone array on a mobile robot called “Kappa”. The directions of sound sources are localized as blue arrows show.

Here, we briefly outline the procedures of an ordinary MUSIC algorithm. Detailed explanation is provided in [6, 8]. The MUSIC algorithm is applied in the frequency-time domain<sup>2</sup>. Let  $\mathbf{x}_{\tau,\omega}$  denote the amplitude of input  $M$ -channel audio signal at time  $\tau$  and frequency  $\omega$ . For each frequency  $\omega$  and time  $t$  at  $\Delta T$  interval, (1) the correlation matrix  $\mathbf{R}_{t,\omega}$  of the input signal is calculated. (2) Then, the eigenvalue decomposition of  $\mathbf{R}_{t,\omega}$  is obtained. (3) Finally, the MUSIC spectrum is calculated using the eigenvectors and the transfer functions.

(1) The correlation matrix is calculated by averaging over observed samples for  $\Delta T$  (sec) as follows:

$$\mathbf{R}_{t,\omega} = \frac{1}{\Delta T} \sum_{\tau=t-\Delta T}^t \mathbf{x}_{\tau,\omega} \mathbf{x}_{\tau,\omega}^H, \quad (1)$$

where  $H$  is the conjugate transpose operator. The vector  $\mathbf{x}_{\tau,\omega}$  has  $M$  elements corresponding to each channel.

(2) The eigenvalue decomposition of  $\mathbf{R}_{t,\omega}$  is given by

$$\mathbf{R}_{t,\omega} = \mathbf{E}_{t,\omega} \mathbf{Q}_{t,\omega} \mathbf{E}_{t,\omega}^H, \quad (2)$$

where  $\mathbf{E}_{t,\omega}$  is the eigenvector matrix and  $\mathbf{Q}_{t,\omega}$  is the eigenvalue matrix. The column vectors of  $\mathbf{E}_{t,\omega}$  are the eigenvectors of  $\mathbf{R}_{t,\omega}$ , that is,  $\mathbf{E}_{t,\omega} = [\mathbf{e}_{t,\omega}^1 \dots \mathbf{e}_{t,\omega}^M]$ .  $\mathbf{Q}_{t,\omega}$  is a diagonal matrix with eigenvalues of  $\mathbf{R}_{t,\omega}$ , that is,  $\mathbf{Q}_{t,\omega} = \text{diag}(q_{t,\omega}^1 \dots q_{t,\omega}^M)$ . The eigenvectors are arranged in descending order.

When we observe  $N$  sound sources, eigenvalues from  $q_{t,\omega}^1$  to  $q_{t,\omega}^N$  have larger values corresponding to the power of each sound source; whereas the rest  $q_{t,\omega}^{N+1}$  to  $q_{t,\omega}^M$  have smaller values corresponding to the power of microphone measurement noises. The important feature of the eigenvalues is that the noise eigenvectors  $\mathbf{e}_{t,\omega}^{N+1}, \dots, \mathbf{e}_{t,\omega}^M$  are orthogonal to the transfer function vectors that correspond to the directions of sound sources [6].

(3) The MUSIC spectrum is calculated as:

$$P(t,d,\omega) = \frac{\|\mathbf{a}_d^H(\omega)\mathbf{a}_d(\omega)\|}{\sum_{m=N_{max}+1}^M \|\mathbf{a}_d^H(\omega)\mathbf{e}_{t,\omega}^m\|^2}, \quad (3)$$

where  $\mathbf{a}_d(\omega)$  is the  $M$ -dimensional transfer function for the  $d$ th direction and frequency  $\omega$ . These transfer functions are measured in advance as a calibration of the microphone array. When we assume the maximum number of sound sources  $N_{max}$ , the eigenvectors  $\mathbf{e}_{t,\omega}^{N_{max}+1}$  through  $\mathbf{e}_{t,\omega}^M$  are orthogonal to the transfer functions  $\mathbf{a}_d(\omega)$  with the direction  $d$  where sound source exists. Thus, the denominator in Eq. (3) becomes close to zero at  $d$ ; in other words, a salient peak of the MUSIC spectrum  $P(t,d,\omega)$  is observed at  $d$ . However, in practice, the peaks in the MUSIC spectrum are smoothed partly because the reverberation in the environment virtually adds sound sources from all directions.

<sup>2</sup>The short-time Fourier transform is carried out with the sampling rate 16,000 (Hz), the window length 512 (pt), the hop size 160 (pt).

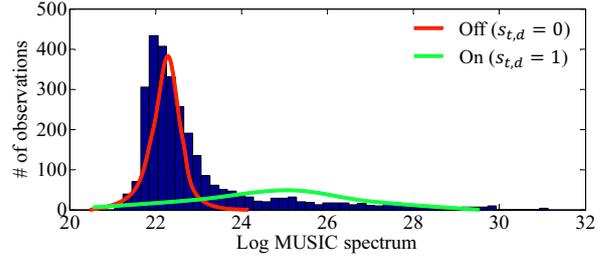


Figure 3: Blue: histogram of logarithmic MUSIC spectrum; Red: a Gaussian for non-active direction; Green: a Gaussian for active direction.

To account for a range of frequency bins, we integrate the MUSIC spectrum for each  $\omega$  as follows:

$$P'(t,d) = \sum_{\omega=\omega_{min}}^{\omega_{max}} \sqrt{q_{t,\omega}^1} P(t,d,\omega), \quad (4)$$

where  $q_{t,\omega}^1$  is the largest eigenvalue at frequency  $\omega$ . To target at speech signals, we set  $\omega_{min} = 500$ ,  $\omega_{max} = 2800$  (Hz).

Basically, we can carry out a localization by detecting the direction  $d$  with  $P'(t,d) > P_{thres}$ , where  $P_{thres}$  is the threshold to determine whether a sound source is active. Since  $P_{thres}$  is dependent on  $N_{max}$  and the reverberation, this threshold is set empirically.

### 3. Bayesian SSL And Tracking

This section presents our Bayesian extension of MUSIC-based SSL and tracking method. Our method consists of two steps: (1) off-line posterior estimation with the variational Bayesian hidden Markov model (VB-HMM), (2) on-line tracking of multiple sound sources using a particle filter. The point in the HMM is that the state vector is a  $D$ -dimensional binary vector whose element indicates whether the sound source at direction  $d$  is active or not. The counterpart of  $P_{thres}$  is automatically obtained through the training of the VB-HMM.

For the observation model, we employ a Gaussian mixture model. We approximate the MUSIC spectrum by a Gaussian distribution partly because the spectrum is a sum over the frequency bins specified in Eq. (4) and partly because analytic computation is possible for this distribution. Figure 3 shows the MUSIC spectrum on the logarithmic scale. As illustrated in Fig. 3, a cluster of non-active MUSIC spectrum is found in a lower area; on the other hand, the spectrum values for active sources scatter over a higher area. Through the training of HMM, we obtain the posterior distribution of the parameters for Gaussian distributions in Figure 3.

We use a particle filter for efficient incremental localization and tracking. The reasons why we use a particle filter are: (1) The number of active sound sources in a state vector is easily capped at  $N_{max}$ . (2) Only local peaks of  $P'(t,d)$  can be activated as a sound source using a proposal distribution. Further explanation is given in Section 3.2

#### 3.1. Off-line Parameter Learning

We use a logarithmic MUSIC spectrum as an observation vector defined as:

$$x_{t,d} = 10 \log_{10} P'(t,d). \quad (5)$$

Let  $s_{t,d}$  be a binary variable. When  $s_{t,d} = 1$ , the sound source at direction  $d$  and time  $t$  is active.

Figure 4 shows the graphical model for the VB-HMM. The difference from the ordinary HMM is that the parameters for the state transition  $\theta_k$  and the observation  $\mu$  and  $\lambda$  are probability

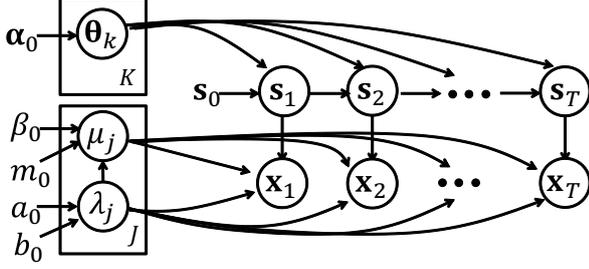


Figure 4: Graphical model for VB-HMM

variables instead of deterministic values. By taking account of many possibility of the parameters as probability variables, the training and the subsequent tracking produce better results than maximum likelihood-based HMM.

### 3.1.1. Observation Model

The observation model is defined as:

$$p(\mathbf{x}_t | \mathbf{s}_t, \boldsymbol{\mu}, \boldsymbol{\lambda}) = \prod_{d=1}^D \prod_{j=0}^1 \mathcal{N}(x_{t,d} | \mu_j, \lambda_j^{-1})^{\delta_j(s_{t,d})}, \quad (6)$$

where  $\delta_j(s_{t,d}) = 1$  iff  $s_{t,d} = j$ , and  $\mathcal{N}(\cdot | \mu, \lambda^{-1})$  denotes the Gaussian distribution with the mean  $\mu$  and precision  $\lambda$ . We use the Gaussian-gamma distribution for  $\boldsymbol{\mu}$  and  $\boldsymbol{\lambda}$  which is the conjugate prior distribution of the Gaussian distribution:

$$p(\boldsymbol{\mu}, \boldsymbol{\lambda} | \beta_0, m_0, a_0, b_0) = \prod_{j=0}^1 \mathcal{N}(\mu_j | m_0, (\beta_0 \lambda_j)^{-1}) \mathcal{G}(\lambda_j | a_0, b_0) \quad (7)$$

where  $\mathcal{G}(\cdot | a, b)$  denotes the gamma distribution with the shape  $a$  and rate  $b$ .

### 3.1.2. State Transition Model

To account for moving sound sources, the state transition model can be divided into 4 cases as summarized in Table 1. These 4 cases are the combination of previous states, that is, they depend on whether the previous state  $s_{t-1,d}$  is active and whether the previous adjacent states are both inactive ( $s_{t-1,d-1} s_{t-1,d+1} = 0$ ) or not. The state transition probability is defined as:

$$p(\mathbf{s}_t | \mathbf{s}_{t-1}, \boldsymbol{\theta}) = \prod_{d=1}^D \prod_{k=1}^4 \prod_{j=0}^1 \left( \theta_k^{s_{t,d}} (1 - \theta_k)^{1 - s_{t,d}} \right)^{f_k(s_{t-1,d})} \quad (8)$$

where  $f_k(s_{t-1,d})$  is a classifier that returns 1 if when  $k$  matches the condition of the previous state values from  $s_{t-1,d-1}$  to  $s_{t-1,d+1}$  as specified in Table 1, and returns 0 otherwise. As the initial state,  $s_{0,d}$  is set 0 for all  $d$ .

$\boldsymbol{\theta} = [\theta_1, \dots, \theta_4]$  conforms to the beta distribution which is the conjugate prior distribution of Eq. (8).

$$p(\boldsymbol{\theta} | \boldsymbol{\alpha}_0) = \prod_{k=1}^4 \mathcal{B}(\theta_k | \alpha_{0,1}, \alpha_{0,2}), \quad (9)$$

where  $\mathcal{B}(\cdot | c, d)$  is the beta distribution with parameters  $c$  and  $d$ .

Table 1: State transition probabilities with adjacent values

previous state	adjacent states	transition probability
$s_{t-1,d}$	$1 - s_{t-1,d-1} s_{t-1,d+1}$	$p(s_{t,d} = 1   s_{t-1,d-1:d+1})$
0 (off)	0	$\theta_1$
0 (off)	1	$\theta_2$
1 (on)	0	$\theta_3$
1 (on)	1	$\theta_4$

### 3.1.3. Inference of Posterior Distribution

Here, the off-line training of the VB-HMM parameters means the estimation of posterior distribution  $p(\mathbf{s}_{1:T}, \boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\lambda} | \mathbf{x}_{1:T})$ . We approximate this posterior by a factorized distribution:

$$p(\mathbf{s}_{1:T}, \boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\lambda} | \mathbf{x}_{1:T}) \approx q(\mathbf{s}_{1:T}, \boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\lambda}), \quad (10)$$

$$q(\mathbf{s}_{1:T}, \boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = q(\mathbf{s}_{1:T}) q(\boldsymbol{\theta}) q(\boldsymbol{\mu}, \boldsymbol{\lambda}), \quad (11)$$

where  $\cdot_{1:T}$  denotes a set of values with the time from 1 to  $T$ . [10] explains the general inference algorithm in detail. We simply show the update equations due to space limitations.  $q(\boldsymbol{\theta})$  is updated to the beta distribution with parameters  $\hat{\alpha}_{k,0}$  and  $\hat{\alpha}_{k,1}$  for each  $k$ , while  $q(\boldsymbol{\mu}, \boldsymbol{\lambda}) = \prod_j q(\mu_j, \lambda_j)$  is updated to the Gaussian-gamma distribution with parameters  $\hat{\beta}_j, \hat{m}_j, \hat{a}_j, \hat{b}_j$ .

$$\hat{\alpha}_{k,j} = \alpha_{0,j} + \sum_{t,d} \langle s_{t,d,j} f_k(\mathbf{s}_{t-1,d}) \rangle, \quad (12)$$

$$\hat{\beta}_j = \beta_0 + w_j, \hat{m}_j = (\beta_0 m_0 + w_j \bar{x}_j) / (\beta_0 + w_j), \quad (13)$$

$$\hat{a}_j = a_0 + \frac{w_j}{2}, \hat{b}_j = b_0 + \frac{w_j S_j^2}{2} + \frac{\beta_0 w_j (\bar{x}_j - m_0)^2}{2(\beta_0 + w_j)}, \quad (14)$$

where  $s_{t,d,j}$  is equal to  $s_{t,d}$ , if  $j = 1$  and  $1 - s_{t,d}$ , if  $j = 0$ . The quantities in Eqs. (13,14) are  $w_j = \sum_{t,d} \langle s_{t,d,j} \rangle$ ,  $\bar{x}_j = \frac{\sum_{t,d} \langle s_{t,d,j} x_{t,d} \rangle}{w_j}$ , and  $S_j^2 = \frac{\sum_{t,d} \langle s_{t,d,j} (x_{t,d} - \bar{x}_j)^2 \rangle}{w_j}$ .  $\langle \cdot \rangle$  is the expectation over Eq. (11).  $\langle s_{t,d,j} \rangle$  and  $\langle s_{t,d,j} f_k(\mathbf{s}_{t-1,d}) \rangle$  are calculated as:

$$\langle s_{t,d,j} \rangle \propto \alpha(s_{t,d,j}) \beta(s_{t,d,j}), \quad (15)$$

$$\langle s_{t,d,j} f_k(\mathbf{s}_{t-1,d}) \rangle \propto \tilde{\alpha}(s_{t-1,d,k}) \tilde{p}(s_{t,d} | \mathbf{s}_{t-1}) \tilde{p}(x_{t,d} | s_{t,d}) \beta(s_{t,d,j}), \quad (16)$$

where  $\alpha(s_{t,d,j})$  and  $\beta(s_{t,d,j})$  are forward backward recursions.

$$\alpha(s_{t,d,j}) \propto \sum_{k=1}^4 \tilde{\alpha}(s_{t-1,d,k}) \tilde{p}(s_{t,d} | \mathbf{s}_{t-1}) \tilde{p}(x_{t,d} | s_{t,d}), \quad (17)$$

$$\beta(s_{t,d,j}) = \sum_{j'=0}^1 \beta(s_{t+1,d,j'}) \tilde{p}(s_{t+1,d,j'} | s_{t,d,j}) \tilde{p}(x_{t,d} | s_{t,d}). \quad (18)$$

The smoothed transition probability is  $\tilde{p}(s_{t,d} = j | \mathbf{s}_{t,d-1:d+1}) \propto \exp\{\psi(\hat{\alpha}_{k,j}) - \psi(\hat{\alpha}_{k,0} + \hat{\alpha}_{k,1})\}$ .<sup>3</sup> The smoothed observation probability is  $\tilde{p}(x_{t,d} | s_{t,d}) \propto \prod_j \exp\left\{(\psi(\hat{a}_j) - \log \hat{b}_j - 1/\hat{\beta}_j)/2 - a_j(x_{t,d} - \hat{m}_j)^2/2\hat{b}_j\right\}^{s_{t,d,j}}$ .

Eqs. (15, 16) are normalized such that the summation over  $j$  or  $j, k$  becomes 1.  $\tilde{\alpha}(s_{t-1,d,k})$  is the probability regarding the condition  $k$ . Parameters are iteratively updated by Eqs. (12–16) until convergence. We start this iteration by setting  $\langle s_{t,d,j} \rangle$  and  $\langle s_{t,d,j} f_k(\mathbf{s}_{t-1,d}) \rangle$  as 1 or 0 with a threshold  $m_0$  on the observation  $x_{t,d}$ .

## 3.2. On-line Localization and Tracking using Particle Filter

This section explains an incremental tracking using a particle filter [11] with the parameters obtained Eqs. (12–14). Here, the posterior distribution of the sound source activation vector given a sequence of MUSIC spectrum is approximated by  $P$  particles:

$$p(\mathbf{s}_t | \mathbf{x}_{1:t}) \approx w_p \mathbf{s}_t^p, \quad (19)$$

where  $w_p$  and  $\mathbf{s}_t^p$  are the weight and state vector of particle  $p$ , respectively. These  $w_p$  and  $\mathbf{s}_t^p$  are obtained with two steps.

- (1) Draw  $\mathbf{s}_t^p$  from the proposal distribution:

$$\mathbf{s}_t^p \sim q(\mathbf{s}_t | \mathbf{x}_t, m, a, b), \quad (20)$$

$$q(\mathbf{s}_t^p | \mathbf{x}_t, \hat{m}, \hat{a}, \hat{b}) \propto \prod_d C(x_{t,d}) \exp(-\Delta_{d,j}^2/2)^{s_{t,d,j}^p}, \quad (21)$$

where  $C(x_{t,d}) = 1$  if  $x_{t,d}$  is a local peak with respect to the direction  $d$  and  $C(x_{t,d}) = 0$ , otherwise. The proposal weight is given by the Mahalanobis distance  $\Delta_{d,j}^2 = (x_{t,d} - \hat{m}_j)^2 \hat{a}_j / \hat{b}_j$ .

- (2) Calculate the weight  $w_p$  for each particle  $p$  as:

$$w_p \propto \frac{\tilde{p}(\mathbf{x}_t | \mathbf{s}_t^p) \tilde{p}(\mathbf{s}_t^p | \mathbf{s}_{t-1}^p)}{q(\mathbf{s}_t^p | \mathbf{x}_t, \hat{m}, \hat{a}, \hat{b})}, \quad (22)$$

<sup>3</sup> $\psi(\cdot)$  denotes the digamma function.

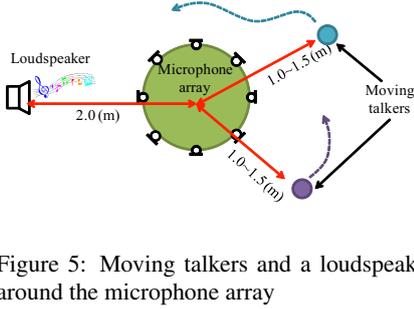


Figure 5: Moving talkers and a loudspeaker around the microphone array

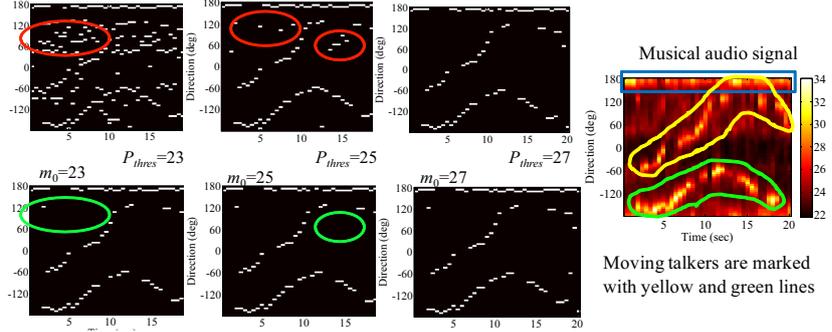


Figure 6: Plots of trajectories of sound sources. White plots are active audio sources. Top: static thresholding with  $P_{thres}$ . Bottom: Our method with  $m_0$ . Right: Observed logarithmic MUSIC spectrum. Musical audio signal is observed at close to 180 (deg).

$$\bar{p}(\mathbf{x}_t | \mathbf{s}_t^p) = \int p(\mathbf{x}_t | \mathbf{s}_t^p, \boldsymbol{\mu}, \boldsymbol{\lambda}) q(\boldsymbol{\mu}, \boldsymbol{\lambda}) d\boldsymbol{\mu} d\boldsymbol{\lambda}, \quad (23)$$

$$\bar{p}(\mathbf{s}_t^p | \mathbf{s}_{t-1}^p) = \int p(\mathbf{s}_t^p | \mathbf{s}_{t-1}^p, \boldsymbol{\theta}) q(\boldsymbol{\theta}). \quad (24)$$

The observation and state transition probabilities in Eqs. (23,24) are given by marginalizing out the parameters with the posterior distributions from those in HMM in Eqs. (6,8). These are analytically calculated as

$$\bar{p}(\mathbf{x}_t | \mathbf{s}_t^p) = \prod_d St(x_{t,d} | \hat{m}_j, \frac{\hat{\beta}_j \hat{a}_j}{(1 + \hat{\beta}_j) \hat{b}_j}, 2\hat{a}_j)^{s_{t,d}^p}, \quad (25)$$

$$\bar{p}(\mathbf{s}_t^p | \mathbf{s}_{t-1}^p) = \prod_d \prod_k \left( \hat{\alpha}_{k,s_{t,d}} / (\hat{\alpha}_{k,0} + \hat{\alpha}_{k,1}) \right)^{f_k(s_{t-1}^p, d)} \quad (26)$$

where  $St(\cdot | m, \lambda, \nu)$  denotes the Student's t-distribution with the mean  $m$ , precision  $\lambda$ , and the degree of freedom  $\nu$ . To keep the number of active sources under  $N_{max}$ , the observation probability is set 0 if that of active sources in  $\mathbf{s}_t^p$  exceeds  $N_{max}$ .

After calculating the weight of all particles, the weights  $w_p$  are normalized s.t.  $\sum_{p=1}^P w_p = 1$ . Then, the posterior is obtained as Eq. (19). In our implementation, the particles are resampled for each time  $t$  in proportion to the weight of each particle.

#### 4. Experimental Results

This section presents the experimental results. Our method is compared to the fixed threshold approach. The experimental setup is shown in Figure 5. For the off-line learning of VB-HMM, only one talker moves around the microphone array while talking. During the on-line tracking using the particle filter, two talkers move around while talking and a musical audio signal is played from the loudspeaker. Both signals are 20 seconds in length. The parameters are set as follows:  $N_{max} = 3$ ,  $\alpha_0 = [1, 1]$ ,  $\beta_0 = 1$ ,  $a_0 = 1$ ,  $b_0 = 500$ . The number of particles is set as  $P = 500$ . The reverberation time of the experiment chamber is  $RT_{20} = 840$  (msec).

Figure 6 shows the results with threshold  $P_{thres} = 23, 25, 27$  and  $m_0 = 23, 25, 27$ . Particle filtering results show the trajectories where the posterior probability exceeds 0.95. The fixed threshold approach with a low threshold produces enormous false detections, as red circles show in Figure 6. On the other hand, our method produces much more stable results as green circles show. We also confirmed that the resulting trajectories are almost the same as long as the threshold of the posterior is in 0.95–1.0. These results confirm that our method automatically converges to good parameters for the SSL and tracking. Furthermore, our method stably tracks multiple sound sources even if only one sound source is used in the training phase.

#### 5. Conclusion and Future Work

This paper presented a Bayesian extension of MUSIC-based SSL. Our method consists of (1) an automatic parameter learning in the VB-HMM framework, and (2) an incremental SSL and tracking using a particle filter. Future work includes the integration with the robot audition system HARK [2], and the application to auditory scene analysis with a mobile robot.

**Acknowledgment:** This research was partly supported by Kakuhni (S) and partly by JST-ANR Research Program on BIN-NAHR (Binaural Active Audition for Humanoid Robots).

#### 6. References

- [1] Y. Kubota et al., "Design and Implementation of 3D Auditory Scene Visualizer towards Auditory Awareness with Face Tracking," in *Proc. of IEEE Int'l Symposium on Multimedia (ISM-2008)*, 2008, pp. 468–476.
- [2] K. Nakadai et al., "Design and Implementation of Robot Audition System "HARK"," *Advanced Robotics*, vol. 24, no. 5–6, pp. 739–761, 2010.
- [3] T. Mizumoto et al., "Design and Implementation of Selectable Sound Separation on a Texai Telepresence System using HARK," in *Proc. of IEEE/RAS Int'l Conf. on Robotics and Automation (ICRA-2011)*, pp. 2130–2137, 2011.
- [4] Y. Sasaki et al., "Map-Generation and Identification of Multiple Sound Sources from Robot in Motion," in *Proc. of IEEE/RAS Int'l Conf on Intelligent Robots and Systems (IROS-2010)*, 2010, pp. 437–443.
- [5] S. Doclo and M. Moonen, *Microphone arrays*. Springer, 2001, ch. GSVD-based optimal filtering for multi-microphone speech enhancement, pp. 111–132.
- [6] R. O. Schmidt, "Multiple Emitter Location and Signal Parameter Estimation," *IEEE Trans. on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.
- [7] F. Asano et al., "Real-time Sound Source Localization and Separation System and Its Application to Automatic Speech Recognition," in *Proc. of Eurospeech2001*, 2001, pp. 1013–1016.
- [8] P. Danès and J. Bonnal, "Information-Theoretic Detection of Broadband Sources in a Coherent Beam-space MUSIC Scheme," in *Proc. of IROS-2010*, 2011, pp. 1976–1981.
- [9] K. Yamamoto et al., "Detection of Overlapping Speech in Meeting using Support Vector Machines and Support Vector Regression," *IEICE Trans. Fundamentals*, vol. E89-A, no. 8, pp. 2158–2165, 2006.
- [10] M. J. Beal, "Variational Algorithms for Approximate Bayesian Inference," Ph.D. dissertation, Gatsby Computational Neuroscience Unit, University College London, 2003.
- [11] M. Arulampalam et al., "A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking," *IEEE Transactions on Signal Proc.*, vol. 50, no. 2, pp. 174–189, 2002.