CHALLENGES IN DEPLOYING A MICROPHONE ARRAY TO LOCALIZE AND SEPARATE SOUND SOURCES IN REAL AUDITORY SCENES

Yoshiaki Bando[1], *Takuma Otsuka*[2], *Katsutoshi Itoyama*[1], *Kazuyoshi Yoshii*[1], *Yoko Sasaki*[3], *Satoshi Kagami*[3], *and Hiroshi G. Okuno*[4]

[1]Kyoto University, [2]NTT Communication Science Laboratories
[3]Advanced Institute of Science and Technology, [4]Waseda University

ABSTRACT

Analyzing the auditory scene of real environments is challenging partly because an *unknown number and type* of sound sources are observed at the same time and partly because these sounds are observed on a *significantly different sound pressure level* at the microphone. These are difficult problems even with state-of-the-art sound source localization and separation methods. In this paper, we exploit two such methods using a microphone array: (1) Bayesian nonparametric microphone array processing (BNP-MAP), which is capable of separating and localizing sound sources when the number of sound sources is unspecified, and (2) robot audition software "HARK" is capable of separating and localizing in real time. Through experimentation, we found that BNP-MAP is more robust against differences in the sound pressure levels of the source signals and in the spatial closeness of source positions. Experiments analyzing real scenes of human conversations recorded in a big exhibition hall and bird calling recorded at a natural park demonstrate the efficacy and applicability of BNP-MAP.

Index Terms— Auditory scene analysis, Bayesian nonparametrics, simultaneous sound source localization and separation, sounds of different volume, unknown time-varying number of sources

1. INTRODUCTION

Computational auditory scene analysis (CASA) [1, 2] in real-world environments is crucial for analyzing and understanding scenes by sounds, performing surveillance, maintaining security, and monitoring environmental changes such as human and animal behaviors [3–7]. For these tasks, sound source localization (SSL) and sound source separation (SSS) are critical functions of CASA systems. The four main challenges with real-world environments are (1) *unknown number and type* of sound sources, (2) *interference by reverberation, reflection, and noise*, and (3) *significant difference in sound pressure levels* of sound sources, (4) *real-time processing* for certain applications.

To address these challenges, many microphone array-based methods have been developed [8–12]. For example, a robot audition software called "HARK" [13] is capable of localizing and separating sound sources in real time to overcome challenge (4). This efficient method is designed as a cascade approach: HARK first carries out SSL using a multiple signal classification (MUSIC) method [14, 15] and then SSS is executed on the basis of the localized results using an algorithm to estimate the separation matrix [16, 17]. While effective with regard to computation time, HARK sometimes requires manual parameter tuning to optimize the performance depending on the acoustic environment, which is problematic in terms of challenges (1) and (2).

In order to overcome challenge (1), Bayesian nonparametric microphone array processing (BNP-MAP) has been developed [18]. BNP-MAP enjoys a robust SSS performance in various indoor environments even if the number of sound sources is unknown which satisfies challenges (1) and (2). The drawback of BNP-MAP is its lengthy computation time, which fails to satisfy the fourth challenge.

In this work, we extensively investigate the third challenge: the impact of the difference of the sound pressure level of constituent source signals on an SSS task through a comparison of HARK and BNP-MAP. In addition to an ordinary speech separation benchmark using a mixture of speech signals played over loudspeakers, our experimental materials pose challenges (1–3) by using recordings collected from an actual exhibition site and recordings of bird songs in a natural park. The former were captured by a 32-channel microphone array embedded on a robot called "Peacock" and the latter by a 7-channel microphone array called "Microcone" manufactured by Dev-Audio. We found through experimentation that BNP-MAP outperforms HARK in terms of separation quality.

2. BACKGROUND AND RELATED WORK

Machine listening systems or robot audition usually hear a mixture of sounds. Robot audition open software "HARK" [13] provides various of signal processing algorithms to solve three fundamental problems of CASA: *sound source localization*, *sound source separation*, and *recognition of separated sounds*.

HARK provides an adaptive beamforming algorithm called multiple signal classification (MUSIC) that robustly localizes multiple sound sources in real environments [14, 15]. It requires steering vectors, which are transfer functions between a sound source and each microphone, to exploit the advantages of the sub-space method. HARK provides the MUSIC localization algorithm via these vectors. It also provides pre-measured steering vectors for the Dev-Audio Microcone (7-channel).

Consider to the separation of $\boldsymbol{M}$ sound sources with $\boldsymbol{N}$ microphones, where $\boldsymbol{N} \geq \boldsymbol{M}$. The spectrum vector of $\boldsymbol{M}$ sources at time t and frequency f and the mixing matrix are denoted as $\mathbf{s}_{tf}$ and $\mathbf{A}_f$, respectively. The observed signals captured by the $\boldsymbol{M}$ microphones at time t and frequency f are denoted as $\mathbf{x}_{tf}$, which is then calculated as $\mathbf{x}_{tf} = \mathbf{A}_f \mathbf{s}_{tf}$. *Sound source separation* aims to find the separation matrix, $\mathbf{W}_f$, that satisfies the equation $\mathbf{y}_{tf} = \mathbf{W}_f \mathbf{x}_{tf}$ under a condition requires the output signal $\mathbf{y}_{tf}$ to be the same as $\mathbf{s}_{tf}$ for any t, possibly with a permutation in the order of the elements.

Thanks to JSPS Kakenhi No.24220006 for funding.

Blind source separation (BSS) solves this problem by obtaining an optimal separation matrix $\mathbf{W}_f^{opt}$ without using any prior information such as $\mathbf{A}_f$. $\mathbf{W}_f^{opt}$ is estimated by minimizing a cost function $J(\mathbf{y}_{1:T,f})$ that denotes the mixture degree of the output $\mathbf{y}_{t,f}$ for $1 \leq t \leq T$. To obtain $\mathbf{W}_f^{opt}$, we use a gradient method to minimize $J(\mathbf{y}_{1:T,f})$ by using $\mathbf{W}_f^{j+1} = \mathbf{W}_f^j - \mu J'(\mathbf{W}_f^j)$, where $J'(\mathbf{W}_f^j)$ defines the derivative of the objective function with regard to $\mathbf{W}_f$ and μ is the stepsize parameter. HARK provides adaptive stepsize control (GHDSS-AS) [16] to attain low-computational cost and improve the sound source separation performance.

3. BAYESIAN NONPARAMETRIC SOUND SOURCE SEPARATION AND LOCALIZATION

The BNP-MAP method can cope with sound source separation and localization even if the number of sound sources is uncertain [18]. In order to consistently cope with an arbitrary number of sound sources regardless of the number of microphones, BNP-MAP uses a time-frequency (TF) masking approach [19–21]. The key question here is how many TF masks should be used when the number of sound sources is unknown. The Bayesian nonparametric model circumvents this problem by allowing in theory for an infinite number of TF masks.

3.1. Model

In this section, we outline the observation model and the major latent parameters used for the separation and localization. Let $\mathbf{x}_{tf}$ be the observed M-channel mixture signal, an M-dimensional complex-valued vector in the TF domain with t and f being the time and frequency index, respectively. Each element of $\mathbf{x}_{tf}$ corresponds to the signal observed by each microphone. In this method, in addition to the multichannel observation, the steering vectors of the microphone array are used for the localization. The localization is carried out in a discrete manner: for example, in our implementation, we prepare steering vectors with a 5° resolution on the azimuth plane, which results in 72 distinct directions.

This TF masking-based model assumes that at most one sound source signal is dominant at time t and frequency f (TF point). Soft TF masks corresponding to respective sound sources are generated to extract the sound sources by calculating the probability of which sound source each TF point belongs to. At the same time, each TF mask is assigned to a certain direction for the localization. This model involves two types of latent variables, z_{tf} and w_k, for the separation and localization, respectively. By $z_{tf} = k$, we mean that sound source k is dominant at TF point $\mathbf{x}_{tf}$, whereas $w_k = d$ means that sound source k arrives from direction d, where k and d denote source index and discrete direction index, respectively.

The design of the likelihood model of the multichannel observation is based on the covariance model [22], where the observation vector follows a Gaussian distribution with zero mean and a time-varying covariance matrix. The covariance matrix factorizes into two parts: the time-varying scale corresponding to the power of the dominant source signal and the matrix corresponding to the propagation of the sound source from a certain direction. The likelihood is given as:

$$\mathbf{x}_{tf}|z_{tf}, w_k \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \lambda_{tf}\mathbf{H}_{f\ w_{z_{tf}}}), \tag{1}$$

where $\mathcal{N}_{\mathbb{C}}(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ is the complex Gaussian distribution with mean $\boldsymbol{\mu}$ and precision matrix $\boldsymbol{\Lambda}$[1]. Note that the subscript $w_{z_{tf}}$ is the direction index: $w_{z_{tf}} = d$ if $z_{tf} = k$ and $w_k = d$. Scalar λ_{tf} represents the time-varying scale of the source signal and $\mathbf{H}_{fd}$ represents the propagation matrix corresponding to direction d. To simplify the inference, the scale parameter is fixed as $\lambda_{tf} = \frac{1}{\mathbf{x}_{tf}^H \mathbf{x}_{tf}}$, where $\cdot^H$ denotes Hermitian transpose. We use the complex Wishart distribution as a prior of $\mathbf{H}_{fd}$ as

$$\mathbf{H}_{fd}|\nu_{fd}, \mathbf{G}_{fd} \sim \mathcal{W}_{\mathbb{C}}(\nu_{fd}, \mathbf{G}_{fd}), \tag{2}$$

where hyperparameters ν_{fd} and $\mathbf{G}_{fd}$ are the degree of freedom and the scale matrix, respectively. The degree of freedom is set as $\nu_{fd} = M$. Scale matrix $\mathbf{G}_{fd}$ is constructed from the M-dimensional steering vector $\mathbf{q}_{fd}$ as $\mathbf{G}_{fd}^{-1} = \mathbf{q}_{fd}\mathbf{q}_{fd}^H + \varepsilon \mathbf{I}_M$ with $\varepsilon = 0.01$. This means that the steering vector corresponding to direction d is used as the prior information to form the propagation matrix of the direction.

Next, we present the prior for the discrete latent parameters z_{tf} and w_k. The hierarchical Dirichlet process (HDP) [23] is used as the prior of z_{tf} so as to deal with an infinite number of TF masks. That is, HDP allows z_{tf} to take $1, \ldots, \infty$. The localization variable w_k follows a finite categorical distribution with the range $w_k = 1, \ldots, D$, where D is the number of directions given as the steering vectors. The formal expression is given as

$$\boldsymbol{\beta}|\gamma \sim \mathrm{GEM}(\gamma), \qquad \boldsymbol{\pi}_t|\alpha, \boldsymbol{\beta} \sim \mathrm{DP}(\alpha, \boldsymbol{\beta}), \quad z_{tf}|\boldsymbol{\pi}_t \sim \boldsymbol{\pi}_t, \tag{3}$$

$$\boldsymbol{\varphi}|\kappa \sim \mathcal{D}\left(\frac{\kappa}{D}\mathbf{1}_D\right), \qquad w_k|\boldsymbol{\varphi} \sim \boldsymbol{\varphi}, \tag{4}$$

where $\mathrm{GEM}(\gamma)$ is the Griffiths-Engen-McCloskey distribution with concentration γ and $\mathrm{DP}(\alpha, \boldsymbol{\beta})$ denotes the Dirichlet process (DP) with a concentration α and a base measure $\boldsymbol{\beta}$. $\mathbf{1}_D$ is a D-dimensional vector with all elements being 1 and $\mathcal{D}(\boldsymbol{\alpha})$ is a Dirichlet distribution with parameter $\boldsymbol{\alpha}$. Localization variable w_k is chosen in accordance with the D-dimensional probability vector $\boldsymbol{\varphi}$ whose elements add up to one. Separation variable z_{tf} is the infinite-dimensional extension: for each time frame t, an infinite-dimensional probability vector $\boldsymbol{\pi}_t$ is generated from the DP with base measure $\boldsymbol{\beta}$, where this base measure is again an infinite-dimensional probability vector. This hierarchical structure is used to create a temporal synchronization of the emergence of a certain source across all frequency bins.

3.2. Parameter inference and source extraction

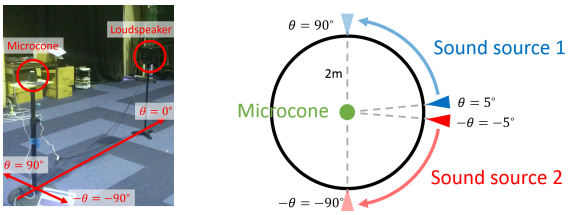
Given the observation $\mathbf{x}_{tf}$, we compute the posterior probability of the latent parameters z_{tf} and w_k. We use a Markov chain Monte Carlo method to generate samples from the posterior distribution of these latent variables. Specifically, we use a Gibbs sampling method with the propagation matrix $\mathbf{H}_{fd}$ being marginalized out.

The Gibbs sampler has been extensively described in [18], so here we briefly explain how the sound sources are extracted from the observed mixture $\mathbf{x}_{tf}$. Let $\{z_{tf}^{(i)}, w_k^{(i)}\}_{i=1}^I$ be a set of samples generated from the Gibbs sampler, where I and i are the number of samples and the index of the Markov chain, respectively. Note that the instantiated source index is upper-bounded by K such that $1 \leq z_{tf}^{(i)} \leq K$ because we have a finite amount of data (finite time frames and frequency bins). Using this samples, the source signal coming from direction d is estimated as

$$\hat{\mathbf{s}}_{tf}^d = \frac{1}{I}\sum_{i=1}^{I} \delta(w_{z_{tf}}, d)\mathbf{x}_{tf}, \tag{5}$$

where $\delta(i, j) = 1$ if $i = j$ and 0 otherwise.

[1] A precision matrix is the inverse of a covariance matrix. We opted for the precision-based notation so that we could use the Wishart distribution for the prior of the precision matrix.



a) Room for benchmarks b) Configuration of two sound sources

Fig. 1. Experiment room and configuration of two sound sources.

Table 1. SDR of sounds separated by BNP-MAP and GHDSS-AS. Rows and columns indicate direction of sound sources θ and SNR to another source p, respectively. Separation performance is significantly degraded when the direction or SNR is low (bold area).

a) BNP-MAP

SNR p [dB]	Direction θ [deg]																	
	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90
10	3.7	6.5	6.2	6.4	6.2	6.0	6.1	6.0	6.8	6.7	7.3	7.9	6.8	6.5	8.0	7.2	7.3	7.3
8	3.4	6.4	6.1	6.4	6.0	5.7	6.1	6.0	6.5	6.6	7.2	7.6	6.5	6.4	7.7	7.0	7.2	6.6
6	3.5	6.0	5.4	5.8	5.9	5.4	5.7	5.5	6.2	6.5	6.9	7.3	6.2	5.9	7.2	7.0	6.8	7.2
4	3.2	5.6	5.1	5.5	5.9	5.4	4.9	5.4	5.4	6.1	6.4	6.9	6.0	6.1	6.7	7.1	6.2	6.7
2	3.7	5.0	5.0	5.7	5.8	5.1	5.2	5.2	5.4	5.3	6.0	6.5	5.7	5.6	6.3	6.7	5.7	6.0
0	3.8	4.5	4.9	5.3	5.4	4.9	4.6	4.8	4.8	4.8	5.7	5.4	5.6	5.3	5.7	6.1	6.0	6.0
-2	3.2	3.9	5.1	4.9	5.2	4.6	3.9	4.8	4.7	4.2	5.5	4.2	5.1	4.8	5.2	5.9	5.1	5.6
-4	2.3	3.1	4.5	5.1	4.7	4.0	3.5	3.9	3.9	3.6	5.2	3.9	4.1	3.8	4.7	5.2	4.7	5.0
-6	1.6	2.7	4.2	4.4	4.7	4.0	2.5	3.9	3.0	2.8	4.5	3.4	3.7	3.6	3.3	4.6	3.6	3.8
-8	0.2	1.6	3.6	3.7	4.0	3.1	1.3	3.6	1.8	2.4	3.8	2.8	2.3	2.8	2.2	4.0	2.5	3.6
-10	**-1.8**	0.9	3.0	2.9	3.4	2.2	1.3	2.6	1.8	1.5	2.4	1.5	1.6	1.8	1.7	3.7	2.0	3.2

b) GHDSS-AS

SNR p [dB]	Direction θ [deg]																	
	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90
10	**-1.3**	**-1.1**	0.8	3.0	3.2	4.0	**-0.6**	2.4	3.5	5.4	7.0	5.8	6.7	5.8	6.7	6.5	6.2	6.8
8	**-2.1**	**-2.2**	**-0.0**	2.4	2.9	3.6	**-0.4**	1.8	3.6	5.1	6.5	5.1	6.6	5.7	6.5	6.0	5.8	6.5
6	**-3.1**	**-2.6**	**-0.9**	1.7	2.4	2.9	**-0.6**	1.4	3.3	4.6	6.4	4.4	6.3	5.4	6.1	5.5	5.3	5.9
4	**-4.1**	**-3.4**	**-1.4**	0.8	1.6	2.3	**-0.7**	1.2	2.8	4.0	5.7	3.5	5.5	4.2	5.2	4.7	4.8	5.1
2	**-5.5**	**-4.8**	**-2.9**	**-0.0**	0.6	1.4	**-1.2**	0.6	2.2	3.0	4.7	2.4	4.8	3.1	4.1	3.8	3.7	4.1
0	**-6.9**	**-6.3**	**-3.5**	**-1.2**	**-0.6**	0.3	**-1.7**	**-0.6**	1.3	2.0	3.6	1.3	3.7	1.6	3.1	2.1	2.3	2.9
-2	**-8.5**	**-7.6**	**-4.6**	**-2.4**	**-1.9**	**-1.0**	**-2.3**	**-2.0**	0.2	0.5	2.5	**-0.1**	2.6	**-0.1**	1.9	0.2	0.9	1.8
-4	**-10.1**	**-9.8**	**-6.1**	**-3.6**	**-3.0**	**-2.4**	**-3.5**	**-2.8**	**-1.4**	**-0.9**	0.8	**-1.4**	1.2	**-1.3**	0.2	**-1.6**	**-0.6**	0.3
-6	**-11.6**	**-10.7**	**-7.3**	**-5.1**	**-4.7**	**-4.1**	**-4.8**	**-4.1**	**-3.2**	**-2.1**	**-0.4**	**-2.5**	**-0.3**	**-3.2**	**-1.0**	**-3.1**	**-2.5**	**-1.2**
-8	**-13.3**	**-12.2**	**-9.0**	**-6.5**	**-6.2**	**-5.3**	**-5.4**	**-5.1**	**-4.5**	**-3.4**	**-2.2**	**-3.9**	**-1.9**	**-4.8**	**-2.7**	**-4.5**	**-4.3**	**-3.2**
-10	**-14.8**	**-13.8**	**-10.6**	**-8.3**	**-7.6**	**-7.0**	**-6.7**	**-6.6**	**-6.0**	**-4.8**	**-4.2**	**-5.2**	**-3.6**	**-5.6**	**-4.5**	**-6.6**	**-5.8**	**-4.9**

4. EXPERIMENTS

In this section, we present the experimental results that consist of the comparison between BNP-MAP and HARK, and the separation results of practical auditory scenes with BNP-MAP.

4.1. Evaluation with simulated sound

To deal with the third challenge, we evaluated the robustness of BNP-MAP and HARK from the perspectives of signal-to-noise ratio (SNR) and the spatial sparseness. The quality of sounds separated by BNP-MAP and HARK was measured in our experiment room and the results evaluated in terms of the signal-to-distortion ratio (SDR). The SDR measures the overall retrieval quality of sound sources from their mixture [24].

Evaluation settings The evaluation was conducted in our experiment room with the set-up shown in Fig. 1-(a). The reverberation time (RT_{60}) of the experiment room was 800 ms. To capture impulse responses, we used a 7-channel microphone array (Microcone, Dev-Audio Inc.) that captures multichannel sound signals at 16 kHz sampling. The input mixed sound was generated by convolving the target sounds with the impulse responses. The impulse responses were measured using a time-stretched pulse [25] with a length of 16,384 samples.

As shown in Fig. 1-(b), we assumed one microphone array and two sound sources located 2 m away from the microphone array. The sound sources were placed at $+\theta$ deg and $-\theta$ deg, respectively and mixed with the volume difference (SNR) of p dB. We changed the direction θ from 5 deg to 90 deg with a 5 -deg resolution. We also changed the SNR p from -10 dB to 10 dB with a 2 -dB resolution.

The two target sound sources were chosen from six human voices (three male, three female) in JNAS [26] phonetically balanced Japanese utterances. That is, 30 convolutive mixtures were generated from JNAS for each direction and SNR. Under all conditions, the clustering parameter K of BNP-MAP was set to 12. The localization and separation of HARK were conducted with MUSIC [15] and GHDSS-AS [16], respectively. The number of sound sources M, a parameter for MUSIC in HARK, was set to 3. These parameters were selected experimentally.

Evaluation results Table 1 lists the mean SDR of separated sound sources for each direction and SNR. The SDRs of GHDSS-AS were significantly degraded when the direction θ was under 15 deg or the SNR p was under -6 dB. In contrast, BNP-MAP maintained the SDRs when the direction θ was over 5 deg and the SNR p was over -10 dB. BNP-MAP was more robust against spatial sparseness and SNR than GHDSS-AS in terms of SDR.

a) ET2013 Exhibition hall b) Peacock (AIST)

Fig. 2. Exhibition hall and Peacock mobile robot.

Fig. 3. Map generated by SLAM with LiDAR. Peacock was placed in the orange circle during recording.

4.2. Analysis of an actual recording

We recorded audio signals at Embedded Technology exhibition[2] (ET2013) held in convention center PACIFICO Yokohama, Japan in 2013. The recorded sound was analyzed by BNP-MAP to demonstrates its performance from the viewpoint of the challenges (1–3). In particular, a crowd of people talking with each other makes the number of sound sources extremely uncertain.

Analysis settings The recording was conducted with a mobile robot called Peacock in a big exhibition hall (Fig. 2). Peacock features a 32-channel microphone array on its top and a light detection and ranging (LiDAR) sensor under the array. A map of the hall generated by SLAM with the LiDAR is given in Fig. 3. Peacock was placed in the orange circle during the recording. We captured 32-ch sound signals at 16-kHz sampling and then analyzed 10 minutes

[2] http://www.jasa.or.jp/et/ET2013/english/

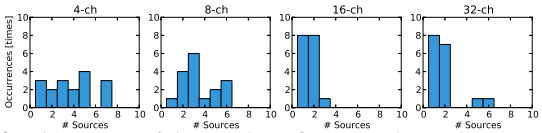


Fig. 4. Histograms of the number of separated sound sources when input mixture signal was 4-, 8-, 16-, and32-ch.

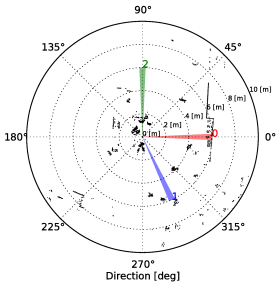


Fig. 5. Point clouds (black marks) obtained from LiDAR and the directions of separated sounds.

of the sound data by BNP-MAP. To reduce the computational cost of BNP-MAP, the 10-minute recording was divided into 30-second segments, each of which were individually analyzed.

Analysis results In this analysis, only 8 of the 32 microphones on the robot were used. Figure 4 provides histograms of the number of separated sound sources when the input signal was 4-, 8-, 16-, and 32-ch. As shown, BNP-MAP separated fewer sound sources as more microphones were used for the analysis. This is because, when the number of input channels of the input signal increases, the classification problem of sound sources is solved in high-dimensional space. Note that we reduced the number of microphones so that the residual microphones could form a circle.

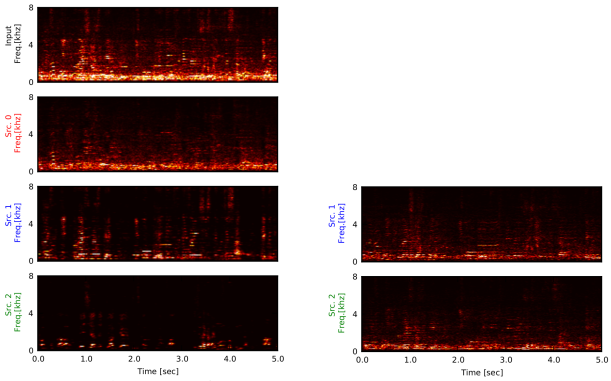
BNP-MAP separated the captured sound signals into various signals comprising talking voices, broadcasts, and background noises. Figure 5 shows the point clouds obtained from the LiDAR and the directions of the separated sounds in one the 30-second segments. Clusters and lines formed from the black points denote humans and walls, respectively. The input signal was separated into three sound sources: Src. 0, background noise, Src. 1., a female voice broadcast, and Src. 2, nearby talking male voices. The clusters at 90°, 1.7 m might belong to speakers from Src. 2. The BNP-MAP did not separate moving sounds such as conversation from walking people because it assumes that the sound sources are stable.

Figure 6 shows the spectrograms of the sounds separated by BNP-MAP and GHDSS-AS from the same segment as Fig. 5. GHDSS-AS separated the captured signal into two sounds (the same sound sources as Src. 1 and 2 of BNP-MAP) in this period. While the signals separated by BNP-MAP contained very little noise and had clear harmonic structures, those of GHDSS-AS contained more noise and had obscure harmonic structures. The BNP-MAP separation is clearly superior to that of GHDSS-AS in this case.

4.3. Analysis of a bird chorus

We also analyzed the recorded signals of bird choruses with BNP-MAP posing the first challenge. The bird choruses were captured in natural park Higashi-mikawa Furusato Park, Japan in 2013 with the 7-channel Microcone microphone array. The 7-ch mixture sound signals were captured at 16 kHz sampling and then we analyzed one minute of the sound data by BNP-MAP.

Figure 7 shows an example of the separated sounds. BNP-MAP



a) BNP-MAP b) GHDSS-AS

Fig. 6. Example of separation results of recording captured at ET2013. Src. 0, 1, 2 are background noise, female voice broadcasts, and nearby talking male voices, respectively.

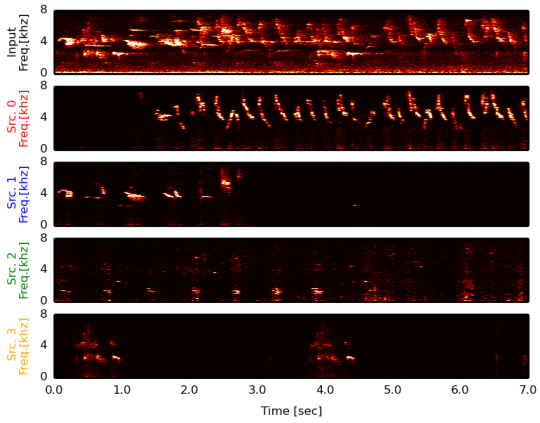


Fig. 7. Example of sounds of bird choruses separated by BNP-MAP. Src. 0, 1, 2, and 3 were choruses of a *Zosterops japonicus*, a *Ficedula narcissina*, a raven, and a *Hypsipetes amaurotis*, respectively.

separated the captured signal into 12 sound sources from which we selected sounds that contained bird songs (depicted in Fig. 7). Src. 0, 1, 2, and 3 refer to the choruses of a *Zosterops japonicus*, a *Ficedula narcissina*, a raven, and a *Hypsipetes amaurotis*, respectively.

5. CONCLUSION

In this work, we investigated the use of HARK and BNP-MAP on SSL and SSS in two real environments: an exhibition hall and a natural park. While BNP-MAP demonstrated proficient separation quality in the face of source number uncertainty, the analysis of real acoustic scenes poses further challenges. These include the separation of multiple sound sources with a spatial overlap and extraction of moving talkers around the microphone array.

Extraction of moving sound sources using Peacock mobile robot should be improved by integration with audio/visual analysis. We have already developed an audio/visual integrated frog chorus analyzer using BNP-MAP and the FireFly sound-imaging system [27] and demonstrated its robustness and accuracy with in-field experiments [28]. Similar to this system, we intend to integrate the results of BNP-MAP and LiDAR for analyzing human behaviors.

Acknowledgment We are grateful to Prof. Reiji Suzuki of Nagoya University for giving a recording of the bird-chorus.

6. REFERENCES

[1] D. F. Rosenthal and H. G. Okuno, *Computational Auditory Scene Analysis*, Lawrence Erlbaum, 1998.

[2] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano, "Active audition for humanoid," in *Proc. of 17th National Conference on Artificial Intelligence*, 2010, pp. 739–761.

[3] S. Kagami, S. Thompson, Y. Sasaki, H. Mizoguchi, and T. Enomoto, "2d sound source mapping from mobile robot using beamforming and particle filtering," in *Proc. of IEEE International Conf. on Acoustics, Speech and Signal Processing*, 2009, pp. 3689–3692.

[4] H. Asoh, I Hara, F. Asano, and K. Yamamoto, "Tracking human speech events using a particle filter," in *Proc. of IEEE International Conf. on Acoustics, Speech, and Signal Processing*, 2005, vol. 2, pp. 1153–1156.

[5] Y. Sasaki, M. Hatao, K. Yoshii, and S. Kagami, "Nested iGMM recognition and multiple hypothesis tracking of moving sound sources for mobile robot audition," in *Proc. of IEEE/RSJ International Conf. on Intelligent Robots and Systems*, 2013, pp. 3930–3936.

[6] G. Valenzise, L. Gerosa, M. Tagliasacchi, E. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *Proc. of IEEE Conf. on Advanced Video and Signal Based Surveillance*, 2007, pp. 21–26.

[7] V. M. Valin, J. Rouat, and F. Michaud, "Enhanced robot audition based on microphone array source separation with post-filter," in *Proc. of IEEE/RSJ International Conf. on Intelligent Robots and Systems*, 2004, vol. 3, pp. 2123–2128.

[8] Y. Li and B. Li, "A novel online algorithm for blind source separation with unknown number of sources," in *Proc. of International Conference on Consumer Electronics, Communications and Networks*, 2011, pp. 2885–2888.

[9] S. Araki, T. Nakatani, and H. Sawada, "Simultaneous clustering of mixing and spectral model parameters for blind sparse source separation," in *Proc. of IEEE International Conference on Acoustics Speech and Signal Processing*, 2010, pp. 5–8.

[10] Y. Zhang and J. Cao, "Dynamic blind source separation using subspace method," in *International Conf. on Web Information Systems and Mining*, 2010, vol. 1, pp. 433–436.

[11] Nobutaka Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2011, pp. 189–192.

[12] A. Masnadi-Shirazi and B. D. Rao, "An ica-sct-phd filter approach for tracking and separation of unknown time-varying number of sources," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 828–841, 2013.

[13] K. Nakadai, T. Takahashi, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "Design and implementation of robot audition system "HARK"," *Advanced Robotics*, vol. 24, no. 5–6, pp. 739–761, 2010.

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

[15] K. Nakamura, K. Nakadai, F. Asano, Y. Hasegawa, and H. Tsujino, "Intelligent sound source localization for dynamic environments," in *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2009, pp. 664–669.

[16] H. Nakajima, K. Nakadai, Y. Hasegawa, and H. Tsujino, "Blind source separation with parameter-free adaptive step-size method for robot audition," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1476–1485, 2010.

[17] H. Krim and M. Viberg, "Two decades of array signal processing research: The parametric approach," *IEEE Signal Processing Magazine*, vol. 13, no. 4, pp. 67–94, 1996.

[18] T. Otsuka, K. Ishiguro, H. Sawada, and H. G. Okuno, "Bayesian nonparametrics for microphone array processing," *IEEE/ACM Trans. on Audio, Speech and Language Processing*, vol. 22, no. 2, pp. 493–504, 2014.

[19] Jalil Taghia, Nasser Mohammadiha, and Arne Leijon, "A variational bayes approach to the underdetermined blind source separation with automatic determination of the number of sources," in *Proc. of IEEE International Conf. on Acoustics, Speech and Signal Processing*, 2012, pp. 253–256.

[20] Hiroshi Sawada, Shoko Araki, and Shoji Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 516–527, 2011.

[21] M. I. Mandel, D. PW Ellis, and T. Jebara, "An em algorithm for localizing multiple sound: Sources in reverberant environments," in *Proc. of Advances in neural information processing systems*, 2007, pp. 953–960.

[22] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Underdetermined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.

[23] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

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

[25] Y. Suzuki, F. Asano, H.Y. Kim, and T. Sone, "An optimum computergenerated pulse signal suitable for the measurement of very long impulse responses," *The Journal of the Acoustical Society of America*, vol. 97, no. 2, pp. 1119–1123, 1995.

[26] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "The design of the newspaper-based japanese large vocabulary continuous speech recognition corpus," in *Proc. of the 5th International Conference on Spoken Language Processing*, 1998, pp. 3261–3264.

[27] I. Aihara, T. Mizumoto, T. Otsuka, H. Awano, K. Nagira, H. G. Okuno, and K. Aihara, "Spatio-temporal dynamics in collective frog choruses examined by mathematical modeling and field observations," *Scientific Reports*, vol. 4, no. 3891, 2014.

[28] Y. Bando, T. Otsuka, K. Itoyama, K. Yoshii, and H. G. Okuno, "Recognition of in-field frog chorusing using bayesian nonparametric microphone array processing," 2015, *under review*.