Proceedings of the 2006 IEEE/RSJ
International Conference on Intelligent Robots and Systems
October 9 - 15, 2006, Beijing, China

Real-Time Tracking of Multiple Sound Sources by Integration of In-Room and Robot-Embedded Microphone Arrays

Kazuhiro Nakadai*, Hirofumi Nakajima†, Masamitsu Murase‡,
Hiroshi G. Okuno‡, Yuji Hasegawa* and Hiroshi Tsujino*

* HONDA Research Institute Japan Co., Ltd., 8-1 Honcho, Wako-shi, Saitama 351-0114, JAPAN
† Nittobo Acoustic Engineering Co., Ltd., 1-13-12 Midori, Sumida-ku, Tokyo 130-0021, JAPAN
‡ Graduate School of Informatics, Kyoto University, Yoshidahonmachi, Sakyo-ku, Kyoto 606-8501, JAPAN
{nakadai, yuji.hasegawa, tsujino}@jp.honda-ri.com, nakajima@noe.co.jp, {murase, okuno}@kuis.kyoto-u.ac.jp

***Abstract*— Real-time and robust sound source tracking is an important function for a robot operating in a daily environment, because the robot should recognize where a sound event such as speech, music and other environmental sounds originate from. This paper addresses real-time sound source tracking by real-time integration of an in-room microphone array (IRMA) and a robot-embedded microphone array (REMA). The IRMA system consists of 64 ch microphones attached to the walls. It localizes multiple sound sources based on weighted delay-and-sum beamforming on a 2D plane. The REMA system localizes multiple sound sources in azimuth using eight microphones attached to a robot's head on a rotational table. The localization results are integrated to track multiple sound sources by using a particle filter in real-time. The experimental results show that particle filter based integration improved accuracy and robustness in multiple sound source tracking even when the robot's head was in rotation.**

I. Introduction

Integration of various information improves robustness and accuracy of perception. In human perception, plenty of evidence on such integration has been reported. For example, temporal integration [1], McGurk effect in speech recognition [2] and audio-visual localization [3] are commonly known as the evidence on audio-visual integration, Also, in sound source localization, two different cues such as interaural phase difference and interaural intensity difference are integrated to robustly localize sounds with wide ranges of frequencies [4]. Some researchers reported auditory-visually integrated systems inspired by the evidence to deal with real-world problems [5], [6]. This means that integration is essential for a robot which is expected to work in the real world to improve the robustness of perception. Actually, we have been reported a robot audition system that localizes, separates and recognizes a mixture of speech uttered by three persons simultaneously by using audio-visual integration, and showed the effectiveness of the system in application to human-robot communication [9].

The system, however, relied on only two microphones at the position of the robot's ears. The system can use not only microphones embedded in a robot, but also those deployed in the surrounding environments to improve the performance. Thus, we propose *spatial integration*, which means the integration of multiple microphone arrays for better sound processing.

A. Two Types of Microphone Arrays and Their Integration

We considered two types of microphone arrays – a robot-embedded microphone array, and an in-room microphone array (hereafter, referred to as *REMA* and *IRMA*, respectively). REMA is promising to improve robot audition directly. Actually, some work [7], [8] has been reported that an 8 ch REMA system has better performance for sound source localization and separation than a binaural approach such as [9]. However, it has two defects. One is that the performance, while the robot is in motion, is worse because it is difficult to synchronize signal capturing with motion precisely and to adapt to acoustic environmental changes after a robot's motion. The other is that it does not give any solution to extract accurate information from a distant talker due to the small size of the microphone array. On the other hand, IRMA can solve these problems, because a microphone array is always stationary, and the microphones are distributed throughout the room. Since this type of microphone array can compensate for the above two defects, it is effective to improve robot audition indirectly. Large size microphone arrays for sound source localization and separation reported in [10], [11], [12] can be used for this purpose. For REMA, we reported a binaural auditory system [9] and *Geometric Source Separation (GSS)* based microphone array [13] so far. Although both systems work in real-time, they were not robust enough against acoustic environmental changes. In this paper, we adopt an adaptive beamformer, *Multiple Signal Classification (MUSIC)* [14] for our 8 ch REMA system. It has better performance for sound source localization and separation than that of the above non-adaptive methods, because it can adapt to some environmental changes. In addition, it can work in real-time by making use of pre-measured impulse responses.

As an algorithm for IRMA, we proposed *weighted delay-and-sum beamforming (WDS-BF)* [15]. This algorithm can estimate directivity patterns and locations of sound sources.

1-4244-0259-X/06/$20.00 ©2006 IEEE

Directivity pattern estimation can be applied to detect actual voice and sound orientation. These functions are useful for human-robot communication, because it enables a robot to distinguish between voices from TV and those uttered by a user and it provides information to move the robot face-to-face with others. Actually, we constructed a 64 ch IRMA with this algorithm and showed the effectiveness in terms of these functions. However, it did not work in real-time, because such a large number of microphones made its computational cost very expensive. So, we introduced a sub-array method to attain real-time processing.

To integrate localization results from REMA and IRMA, we propose a particle filter refined for multiple sound source tracking. Particle filtering is well-known to track moving objects in robot vision and to solve a *Simultaneous Localization And Mapping (SLAM)* problem [16]. It was not applied to multiple sound source tracking and integration of microphone arrays. Our proposed particle filter can cope with difficulties in microphone array integration and works in real-time.

We will show the effectiveness of a spatial integration system, which consists of an 8 ch REMA system, a 64 ch IRMA system and a particle filter based integrator, through multiple sound source localization and tracking.

II. Signal Processing For Microphone Arrays

A. Algorithm for Robot-Embedded Microphone Array

The MUSIC implementation for our REMA system was developed by the National Institute of Advanced Industrial Science and Technology (AIST) [7]. It was specially developed for a humanoid robot operating in the real world. In their implementation, pre-measured impulse responses are used as transfer functions to overcome the diffraction of the robot and to realize faster adaptation. This approach is more accurate and faster in processing speed than model based ones such as [9]. The detail algorithm is described in [14].

B. Algorithm for In-Room Microphone Array

Generally, output spectrum $Y_{\boldsymbol{p}}(\omega)$ for a typical microphone array system is defined by

$$Y_{\boldsymbol{p}}(\omega) = \sum_{n=1}^{N} G_{n,\boldsymbol{p}}(\omega) X_n(\omega) \tag{1}$$

$$X_n(\omega) = H_{\boldsymbol{p},n}(\omega) X(\omega) \tag{2}$$

where $X(\omega)$ denotes the spectrum of a sound source S located at $\boldsymbol{p}$. $H_{\boldsymbol{p},n}(\omega)$ denotes a transfer function from S to the n-th microphone. $X_n(\omega)$ is the spectrum captured by the n-th microphone. $G_{n,\boldsymbol{p}}(\omega)$ denotes a filter function to estimate the sound spectrum at $\boldsymbol{p}$ from the spectrum of the input signal to the n-th microphone. The WDS-BF is generalized to be able to use various kinds of transfer functions such as measured impulse responses and simulated transfer functions which take reverberation and diffraction into account. Also, the norm of $G_{n,\boldsymbol{p}}(\omega)$ is minimized, so the WDS-BF is robust against the dynamic changes of $H_{\boldsymbol{p},n}$ and distorted $X_n(\omega)$. We introduce sub-array processing to use only channels with high contribution to localization for faster processing and improving the localization accuracy. The criteria for channel selection is decided by the distance between the sound source and each microphone, r_n. When r_n is less than r_{th}, n-th microphone is selected. Otherwise, n-th microphone is excluded in beamforming and every transfer function for n-th microphone is set to 0. The WDS-BF is applied to estimate directivity pattern estimation by replacing $\boldsymbol{p}$ with $\boldsymbol{p'} = (\boldsymbol{p}, \theta)$ in Eqs. (1) and (2). The detail algorithm of sound source localization with directivity pattern is described in [15].

III. Integration of Microphone Arrays

To integrate the results of two types of microphone arrays, we propose to use a particle filter[17]. The two main advantages of the particle filter are that it can deal with non-linear motion of an object and the processing speed can be controlled by the number of particles. It is basically easy to apply to track a sound source [18], [19], [20], because the particle filter needs only probabilistic models on a transition and an observation of the internal states. Valin[21] extended the particle filter to track multiple sound sources with an 8 ch microphone array in a mobile robot, by using the techniques on source-observation assignment in multiple object tracking[18], [22], [23]. In addition, the particle filter is extended for multi-modal integration[22], for example, Asoh *et al.*[24] suggests that this technique is useful for integration of audio and video data to track speakers. However, it is difficult to apply their methods to integrate multiple sound source localization results obtained from two types of microphone arrays, i.e., REMA and IRMA, because the following two issues related to the coordinates are not taken into account in their methods.

A. Issues in Integration of Microphone Arrays

For microphone array integration, we should consider two issues – the robot vs. the world coordinates, and the polar vs. the Cartesian coordinates.

REMA moves, while IRMA is stationary. This means that a sound is observed in the robot coordinates for REMA, so the coordinate conversion to the world coordinates is necessary to integrate REMA with IRMA. This requires that accurate time synchronization between sound processing and a robot's motion is crucial. We can consider two approaches to solve this synchronization problem. One is a software-based approach, i.e, the particle filter itself can solve this synchronization problem by using a probabilistic model concerning the time difference. The other is a hardware-based approach that uses the architecture with mechanically and electronically accurate synchronization between sound processing and robot motion. The former looks smart because one particle filter can be applied for time synchronization, but some synchronization errors occur inevitably because they use probabilistic representation. So, we chose the latter approach to solve this issue (see Section IV-A).

The second issue was caused by the fact that the coordinate types were different. To integrate them, we propose two types of likelihood functions for the polar and the Cartesian

coordinates. These functions output likelihood independent from the coordinates. So, the two likelihood values can be easily integrated. The detail is explained below.

B. Particle Filter

In the particle filter, the transition model $p(\boldsymbol{x}(t)|\boldsymbol{x}(t-1))$ and the observation model $p(\boldsymbol{y}(t)|\boldsymbol{x}(t))$ of internal state $\boldsymbol{x}(t)$ are defined as probabilistic representation. $\boldsymbol{y}(t)$ denotes an observation vector. Because the particle filter allows a non-linear transition model, it is more flexible than other linear filtering methods such as the Kalman filter. A particle plays a role of an agent to track a target source. The i-th particle includes the internal states $\boldsymbol{x}_i(t)$ and the importance weight $w_i(t)$, which is an index to show how the particle contributes to tracking and is usually defined as likelihood. The density of a set of the particles approximates posterior probability $p(\boldsymbol{x}(t)|\boldsymbol{y}(t))$. In other words, the posterior probability is sampled by the particles. That is why the particle filter is also called a sampling method. In our case, two types of observations, $\boldsymbol{Y}_{\mathrm{REMA}}(t)$ and $\boldsymbol{Y}_{\mathrm{IRMA}}(t)$, are obtained from the microphone arrays at time t.

$$\boldsymbol{Y}_{\mathrm{REMA}}(t) = \{\boldsymbol{y}_{a_1}(t), \cdots, \boldsymbol{y}_{a_l}(t), \cdots, \boldsymbol{y}_{a_{L_t}}(t)\}, \quad (3)$$

$$\boldsymbol{Y}_{\mathrm{IRMA}}(t) = \{\boldsymbol{y}_{b_1}(t), \cdots, \boldsymbol{y}_{b_m}(t), \cdots, \boldsymbol{y}_{b_{M_t}}(t)\} \quad (4)$$

where L_t and M_t are the number of observations by REMA and IRMA at time t. $\boldsymbol{y}_{a_l}$ and $\boldsymbol{y}_{b_m}$ are denoted as

$$\boldsymbol{y}_{a_l} = \{a_{\theta_l}, a_{p_l}\}, \quad (5)$$

$$\boldsymbol{y}_{b_m} = \{b_{x_m}, b_{y_m}, b_{o_m}, b_{p_m}\}, \quad (6)$$

where a_{θ_l} is the azimuth in the world coordinates. b_{x_m} and b_{y_m} are the position in the world coordinates, and b_{o_m} is the orientation of a sound source. a_{p_l} and b_{p_m} are the estimated power in dB.

Our particle filter consists of the following five steps: *Initialization*, *Source check*, *Importance sampling*, *Selection*, and *Output*. The following sections describe each step in detail.

Step 1 – Initialization: This step makes the initial states of a particle. We defined $(x_i(t),\ y_i(t), v_i(t), o_i(t))$ as the internal states of i-th particle. $(x_i(t), y_i(t))$ denotes the position of a sound source. $v_i(t)$ and $o_i(t)$ are the velocity and the orientation of the sound source. At the initial state, the particles were distributed uniformly at random. To deal with multiple sound sources, we initialized the importance weight defined by

$$\sum_{i \in P_k} w_i = 1, \quad \sum_{k=1}^{S} N_k = N, \quad (7)$$

where N_k is the number of particles for k-th particle group P_k, and S is the number of sound sources. N is the fixed value which shows the total number of particles.

Step 2 – Source Check: This step is newly added to support multiple sound sources. The internal state of the particle group P_k is defined by

$$\hat{\boldsymbol{x}}_k(t) = \sum_{i \in P_k} \boldsymbol{x}_i(t) \cdot w_i(t) \quad (8)$$

When $\boldsymbol{y}_m(t)$ satisfies $||\hat{\boldsymbol{x}}_k(t) - \boldsymbol{y}_m(t)|| < D_{th}$, $\boldsymbol{y}_m(t)$ is associated with P_k. When no particle group is found for $\boldsymbol{y}_m(t)$, a new particle group is generated. When no observation is found for the particle group P_k for more than time T_{th}, P_k is terminated. In both cases, the particles are re-distributed so that Eq. (7) is maintained.

Step 3 – Importance Sampling: In this step, first, $\boldsymbol{x}_i(t)$ is estimated from $\boldsymbol{x}_i(t-1)$ by using the transition model $p(\boldsymbol{x}(t)|\boldsymbol{x}(t-1))$. Secondly, $w_i(t)$ is updated by Eq. (20). Finally, $w_i(t)$ is normalized subject to Eq. (7).

For the transition model, we switched two models based on random walk and Newton's equation of motion according to the velocity of the sound source. When the velocity is less than v_{th}, which is empirically set to $2\,m/s$, the system uses the transition model based on random walk defined by

$$x_i(t) = x_i(t-1) + r_x, \quad (9)$$

$$y_i(t) = y_i(t-1) + r_y, \quad (10)$$

$$v_i(t) = v_i(t-1) + r_v, \quad (11)$$

$$o_i(t) = o_i(t-1) + r_o, \quad (12)$$

where r_* denotes white noise.

When the velocity is larger than v_{th}, the system uses the transition model based on Newton's equation of motion defined by

$$\begin{aligned} x_i(t) &= x_i(t-1) \\ &+ v_i(t-1) \cdot \cos(o_i(t-1)) + r_x, \end{aligned} \quad (13)$$

$$\begin{aligned} y_i(t) &= y_i(t-1) \\ &+ v_i(t-1) \cdot \sin(o_i(t-1)) + r_y, \end{aligned} \quad (14)$$

$$\begin{aligned} v_i(t) &= \alpha \cdot v_i(t-1) \\ &+ (1-\alpha) \cdot \sqrt{\Delta x_i(t)^2 + \Delta y_i(t)^2} + r_v, \end{aligned} \quad (15)$$

$$\begin{aligned} o_i(t) &= \alpha o_i(t-1) \\ &+ (1-\alpha) \cdot \tan^{-1}\left(\frac{\Delta y_i(t)}{\Delta x_i(t)}\right) + r_\theta, \end{aligned} \quad (16)$$

where α is an weight defined experimentally. $\Delta x_i(t)$ and $\Delta y_i(t)$ are defined by

$$\Delta x_i(t) = x_i(t) - x_i(t-1),$$

$$\Delta y_i(t) = y_i(t) - y_i(t-1).$$

The likelihood is defined as follows:

$$l_{\mathrm{REMA}}(t) = exp\left(-\frac{\left(\angle\left(\boldsymbol{x}_i(t) - \boldsymbol{P}_{\mathrm{REMA}}(t)\right) - \theta_l\right)^2}{2R_{\mathrm{REMA}}}\right) \quad (17)$$

$$l_{\mathrm{IRMA}}(t) = exp\left(-\frac{||\boldsymbol{x}_i(t) - \boldsymbol{y}_{b_m}(t)||^2}{2R_{\mathrm{IRMA}}}\right) \quad (18)$$

where $\angle(\boldsymbol{x})$ denotes the angular coordinate for $\boldsymbol{x}$. R_{REMA} and R_{IRMA} are variances of localization results by REMA

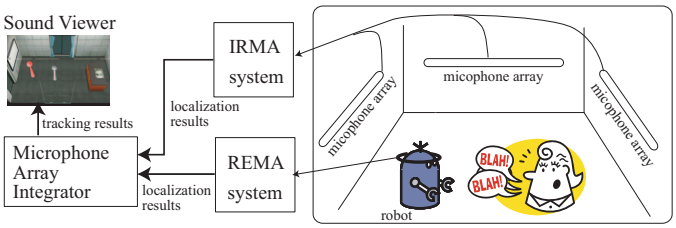


Fig. 1. Spatial Integration System

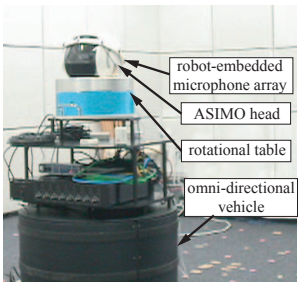


Fig. 2. Wheel-based Robot

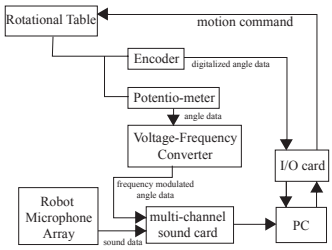


Fig. 3. The Architecture of REMA

and IRMA, $\boldsymbol{P}_{\mathrm{REMA}}$ is the position of the robot. They are integrated into $l_I(t)$.

$$l_I(t) = \alpha_l \cdot l_{\mathrm{REMA}}(t) + (1 - \alpha_l) \cdot l_{\mathrm{IRMA}}(t) \tag{19}$$

where α_l is an integration weight value. Finally w_i is updated by

$$w_i(t) = l_I(t) \cdot w_i(t-1). \tag{20}$$

Step 4 – Selection: According to the importance weight w_i, the selection step propagates and removes particles. When $i \in P_k$, the number of particles for i is updated by

$$N_{k_i} = round(N_k \cdot w_i). \tag{21}$$

In this case, some particles can remain. The number of such particles is calculated by

$$R_k = N_k - \sum_{i \in P_k} N_{k_i}. \tag{22}$$

These particles are also distributed according to the residue weight R_{w_i}.

$$R_{w_i} = w_i - N_{k_i} / \sum_{i \in P_k} N_{k_i} \tag{23}$$

For this, *Sampling Importance Resampling (SIR)* algorithm [17] is commonly used.

Step 5 – Output: From the density of the updated particles, the posterior probability is estimated as $p(\boldsymbol{x}(t)|\boldsymbol{y}_m(t))$. The internal states of a set of particles for sound source k is estimated as Eq. (8). Steps 2 – 5 are repeated until the tracking process finishes.

IV. System Implementation

Fig 1 shows the architecture of our spatial integration system. It consists of four components – a robot with a REMA system, an IRMA system, a microphone array integrator and a sound viewer. They are described in the following sections.

A. Robot with REMA System

For the REMA system, we developed a wheel-based robot shown in Fig. 2. The robot consists of the head of Honda ASIMO, a rotational table, an omni-directional vehicle and an 8 ch REMA system. The rotational table is controlled by a remote PC via an I/O card shown in Fig. 3. The angle of rotation is measured by an encoder accurately. Its resolution of the angle measurement is 0.0015°. The omni-directional vehicle also can be controlled by a remote PC via a wireless-LAN. The maximum load of the vehicle is 80 kg. The side of the vehicle is covered by sensors to detect collision. The REMA system consists of 8 microphones, and each microphone is embedded in a rubber head-band at the same interval. The head-band is installed on the head of ASIMO.

The captured sound signals by REMA and the signals of angle information from the encoder are sent to one PC. To localize sound signals in the world coordinates even when the robot is in rotation, these two types of signals should be synchronized precisely. The encoder outputs accurate information, but it takes small processing time to send the digitized data. This makes time difference between the captured sound signals and the corresponding encoder output signal. To make precise synchronization, we measured the time difference. As shown in Fig. 3, we installed a potentio-meter in the rotational table to detect rotation quickly in addition to the encoder. The potentio-meter was connected to the sound card to synchronously capture sounds from REMA via a voltage-frequency converter which prevents a DC component like potentio-meter output from being filtered out by the sound card. Because the potentio-meter produced analog output, the output included larger errors (0.95°), but the time difference was regarded as 0, since the data was captured with sounds from REMA simultaneously. Therefore, we measured the time difference by comparing angle data captured by the sound card with that sent from the encoder. As a result, we found that the angle data from the encoder was delayed 32.9 ms on average in comparison with captured sound signals. This delay was taken into account in the coordinate conversion from the robot coordinates to the world coordinates.

B. IRMA System

We constructed a 64 ch IRMA system. The IRMA captures an acoustic signal from 64 microphones synchronously at a sampling rate of 16 kHz using four RASP II. Fig. 4 depicts a 4.0 m × 7.0 m room for IRMA. It is acoustically asymmetrical because three walls are covered with sound absorbing material, another wall is made of glass with high sound reflection, and there is a kitchen sink. The asterisks represent microphone positions in the room. The height of the microphones on the wall is 1.2 m. This microphone layout was decided because it covers as much of the room as possible. The sound position was digitized at an interval of 25 cm. The digitizing area was 1.0 m – 5.0 m for X axis, and 0.5 m – 3.5 m for Y axis. The height (Z axis) was fixed to 1.2 m. So, the number of $\boldsymbol{p}_m$ is 221. To design a beamformer for IRMA, we calculated beamforming coefficients from pre-measured transfer

TABLE I

THE EFFECT OF A SUB-ARRAY ON COMPUTATIONAL COST (SIMULATION)

r_{th}	computational cost (%)	# of ch to use	
		Max	Min
7	100	64	64
6	99.9	64	63
5	97.4	64	41
4	82.4	64	33
3.5	68.8	62	22
3	53.0	56	19
2.5	37.5	39	12
2	23.2	29	0

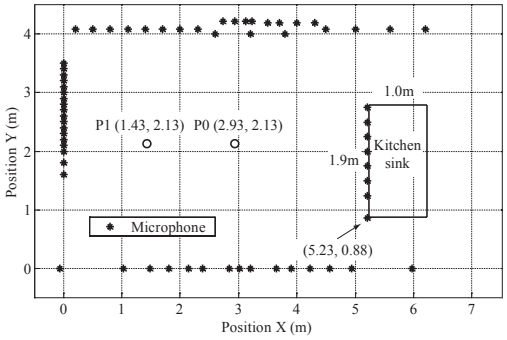


Fig. 4. Layout of Microphones

functions, because they can deal with any kind of acoustic environment and microphone layout. To obtain the transfer functions, we measured impulse responses at every $\boldsymbol{p}_m$ under the condition that a loudspeaker at $\boldsymbol{p}_m$ faced a robot placed at P_0. We call this beamformer "M-BF" in the latter sections. We then designed the sub-array version of the M-BF (hereafter, referred to as "MS-BF"). The distance threshold r_{th} described in Section II-B is set to 3.5 m. In this case, a 30% reduction of the computational cost would be expected as shown in Tab.I.

C. Microphone Array Integrator and Sound Viewer

Microphone array integrator integrates localization results from REMA and IRMA, and tracks sound sources by particle filtering described in Sec. III-B. The particle filter is implemented on PC Linux by using C language. The tracking results are sent to a sound viewer implemented with with Java3D and OpenFlight. It has the functions of real-time 3D visualization, online and off-line processing, and flexible change of a viewpoint so that we can understand sound scene in the room at a glance.

V. EVALUATION

Two types of evaluations for the spatial integration system using microphone arrays were performed as follows:

1) the basic performance of sound source localization, and
2) the performance of sound source tracking.

In the first evaluation, a single sound source was localized by the IRMA and REMA. As a sound source, we used the recorded voices played by a loudspeaker GENELEC 1029A located at P1 shown in Fig. 4. The average error and the standard deviation of localization were measured. The four types of beamformers – "M-BF", "MS-BF", "Sim-BF" and "RSim-BF" – were used for the IRMA. "M-BF" and "MS-BF" were already described in Section IV-B. The other two beamformers were based on simulation. "Sim-BF" is a beamformer which is designed by simply assuming a free space, while "RSim-BF" is a beamformer which takes room reverberations into account. RSim-BF is designed to reduce the power of sounds reflected by the walls. This is done by the adaptation of the RSim-BF coefficients to minimize the total system gain from the imaginary non-target points which are 0.2 m outside the room boundaries. The detailed algorithm is described in [25]. The orientation of the loudspeaker was 0° which was specified as a vector (1,0) and the direction of positive rotation was counterclockwise.

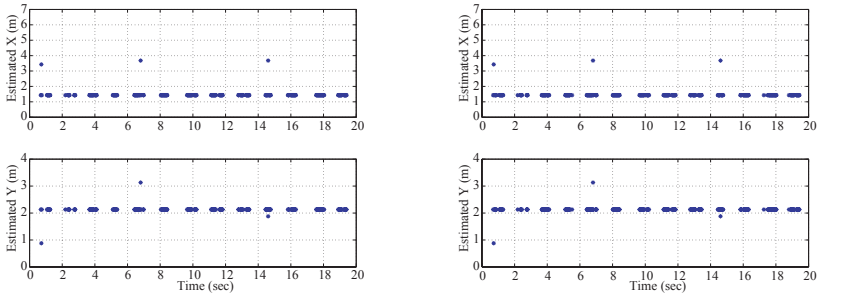
In the second evaluation, the performance to track moving sound sources was measured. Moving sound sources were recorded in five situations as follows:

- **Ex.2A**: The loudspeaker was moved from (2.93 m, 0.63 m) to (2.93 m, 3.63 m) along the arc of the circle with center P_0 and radius 1.5 m in the counterclockwise direction. The heading of the robot located at P_0 was fixed to 180°.
- **Ex.2B**: The loudspeaker was fixed at P_1. The robot was located at P_0. The heading was changed from 90° to 270°.
- **Ex.2C**: The loudspeaker was moved in the same way as **Ex.2A**. The heading of the robot located at P_0 was changed from 90° to 270° to face the loudspeaker.
- **Ex.2D**: Two persons (Mr. A and Mr. B) walked while speaking along the circle with center P_0 and radius 1.5 m. They were asked to say Japanese sentences continuously and to face the robot. Mr. A started at (2.93 m, 0.63 m), i.e., 90° in the robot coordinates, and walked clockwise to 0°. Just before arriving at 0°, he turned back and walked counterclockwise to 270°. Mr. B started at (2.93 m, 3.63 m) and walked in a mirrored way, that is, he first moved counterclockwise to 0°, and turned to 90° in the clockwise direction. They approached and receded at 0°, and crossed at 180°. The heading of the robot located at P_0 was fixed to 180°.
- **Ex.2E**: The motion of two persons was the same as **Ex.2D**. The heading of the robot located at P_0 was always kept facing Mr. A.

The sound source localization by REMA and IRMA, and sound source tracking by the particle filter were performed from recorded sound. To obtain an accurate location of the moving sound sources as reference data, we used an *Ultrasonic 3D Tag System (U3D-TS)* which localized an ultrasonic 3D tag (U3D tag) with only several cm errors [15]. In every case, MS-BF was used as a beamformer for IRMA.

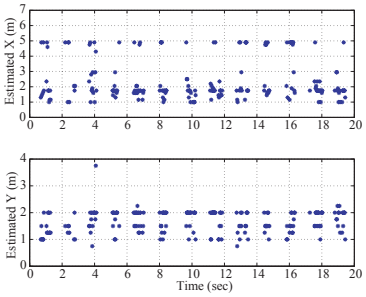
A. Results

Fig. 5a)-d) shows the localization performance of a single sound source by IRMA. The horizontal axis is time, the vertical axes are the estimated X and Y in meters. Fig. 5e)

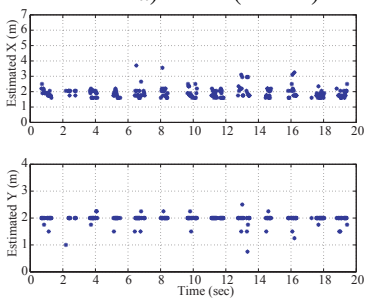


a) M-BF(IRMA)

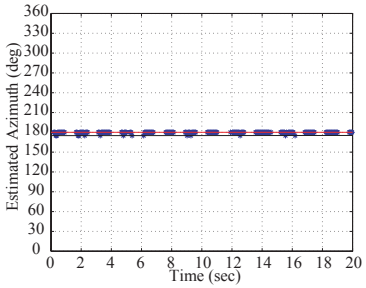
b) MS-BF(IRMA)



c) Sim-BF(IRMA)



d) RSim-BF(IRMA)



e) MUSIC(REMA)

Beamformer	Error	
	Avg.	S.D.
M-BF (IRMA)	0.016 m	0.039 m
MS-BF (IRMA)	0.019 m	0.041 m
Sim-BF (IRMA)	0.95 m	1.19 m
RSim-BF (IRMA)	0.50 m	0.52 m
MUSIC (REMA)	4.56°	1.41°

f) Error of Localization

Fig. 5. Sound Source Localization Results

TABLE II
LOCALIZATION ERROR WITH REMA AND IRMA

	REMA		IRMA	
	Avg.(deg)	S.D.(deg)	Avg.(m)	S.D.(m)
Ex.2A	4.01	16.18	0.217	0.157
Ex.2B	3.25	7.61	0.082	0.249
Ex.2C	5.96	3.16	0.190	0.303
Ex.2D	6.14	10.66	0.194	0.173
Ex.2E	7.46	7.83	0.234	0.200

TABLE III
TRACKING ERROR WITH PARTICLE FILTER

	IRMA Only		Integration of IRMA and REMA	
	Avg.(m)	S.D.(m)	Avg.(m)	S.D.(m)
Ex.2A	0.12	0.062	0.10	0.040
Ex.2B	0.06	0.012	0.06	0.012
Ex.2C	0.11	0.075	0.10	0.071
Ex.2D	0.16	0.084	0.16	0.083
Ex.2E	0.18	0.133	0.17	0.123

shows the result using REMA. The horizontal axis is time, the vertical axis is the estimated azimuth in degrees in the world polar coordinates. Fig. 5f) shows the average error and the standard deviation in localization.

Fig. 6 shows the results of localization and tracking. The first to the fifth row in Fig. 6 corresponds to the results of **Ex.2A** – **Ex.2E**. The left column shows the localization results by using REMA. The horizontal axis is time in seconds, and the vertical axis is estimated azimuth in degrees. The blue asterisks show the localization results in the robot coordinates. The red line shows the robot motion obtained from the encoder in the world polar coordinates. The red plus marks are the localization results after the coordinate conversion to the world polar coordinates. The black and green lines are the sound directions obtained from the U3D tags. The middle column shows the localization results by using IRMA. The blue asterisks show the localization results in the world Cartesian coordinates. The black and green lines are the sound tracks obtained from the U3D tags. The right column shows the tracking results using the particle filter. The red lines show the tracking results with the particle filter when only room localization data was used. The blue lines are those when room and robot localization data are integrated by the particle filter. The black and green lines are the sound tracks from the U3D-TS. Tab. II shows the average and the standard deviation of localization error in REMA and IRMA, and Tab. III shows the tracking errors with the particle filter.

B. Observations

From the first evaluation, the best beamformers were M-BF and MS-BF. They had small localization errors of 15 cm – 20 cm. These beamformers were designed from measured transfer functions, so they were robust for the reverberation in the room. When considering processing speed, we can say that MS-BF is the best beamformer for our IRMA. As shown in Tab. I, the 30% computational cost was reduced while maintaining localization accuracy. Actually, the IRMA system attained around 16 fps of localization speed due to the introduction of the sub-array method. Because transfer functions were available at the discrete points which we measured in advance, a regression method was necessary to cope with sound localization at the points where transfer functions were unavailable. As such a method, RSim-BF could be substituted for M-BF and MS-BF. In the case of MUSIC, it had an error of about 4.5°. This is equivalent to 12 cm at a point 1.5 m away from the robot. It is almost the same accuracy as IRMA. Localization was more accurate for a closer sound source, and the resolution of localization was worse for the further sound sources. The distance between the robot and the sound sources was about 1.5 m in every experiment. So, we used 0.5 for integration weight parameter α_l in the last evaluation.

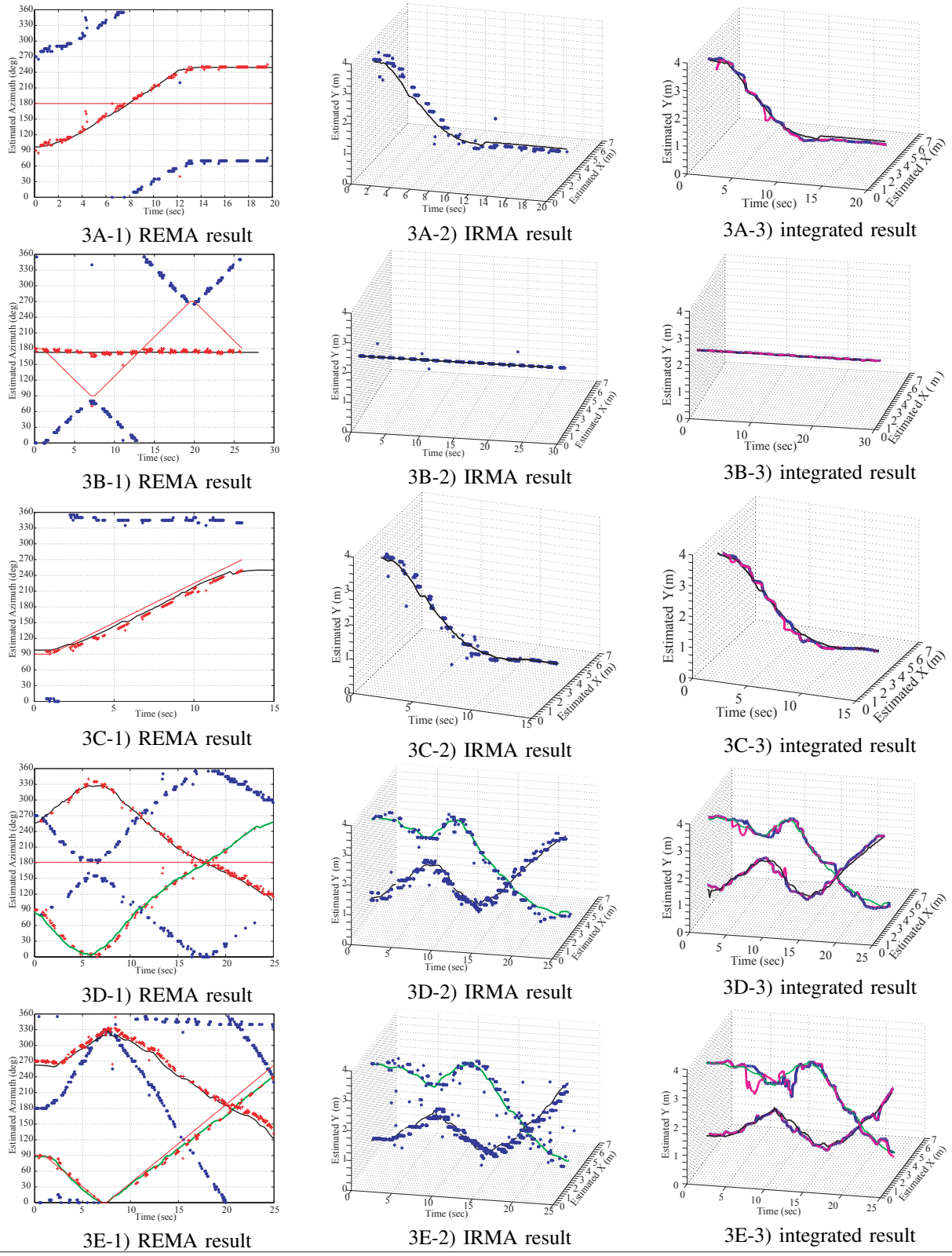


Fig. 6. Tracking Results

In the second evaluation, compared with tracking by U3D-TS, we can say that REMA and IRMA could localize at least two simultaneous speech signals properly even when the sources were in motion. In the case of REMA, accurate time synchronization was achieved because the coordinate-converted localization results fitted those obtained from U3D-TS. The localization error in Tab. II shows that it became large, when the number of sound source increases and/or the sound sources were moving, but the difference of the errors was within several cm or degrees. On the other hand, some outliers could be seen, and data association between each localization result and the corresponding sound source was not done yet. In tracking of multiple sound sources, this association was essential because miss association causes a fatal tracking error. This problem is also known as the permutation problem in sound source separation such as independent component analysis. The particle filter solved this problem from the right column in Fig. 6. In addition, Tab. III shows that the particle filter improved localization in accuracy and robustness because the average errors were reduced 2 cm – 9 cm and the standard deviations were reduced about 10 cm on average. From Tab. III, the effect of the microphone array integration looks small, but the integration contributes to an improvement in the robustness of tracking. For example, the tracks (red lines) using localization results by IRMA had large errors from 5 sec to 10 sec in Fig. 6 2D-3) and 2E-3), while the integrated tracks (blue lines) did not include the large errors.

VI. Conclusion

We proposed the particle filter based spatial integration of REMA and IRMA for general sound understanding to enhance a robot audition system. For IRMA, we extended reported weighted delay-and-sum beamforming to work in real-time by the introduction of a sub-array method. For real-time spatial integration, we newly proposed a particle filter for multiple sound sources, which can integrate multiple localization results utilizing a probabilistic integration method to track sound sources. We constructed a real-time spatial integration system including 64 ch IRMA and 8 ch REMA based on the particle filter. The evaluations of the system show that the two types of microphone arrays localized multiple sound sources accurately, and sound source tracking with the particle filter improved the accuracy and the robustness of sound source localization.

VII. Future Work

The particle filter used several parameters. We selected the best values for each parameter manually. These values should be optimized automatically. Also, we assumed that the number of sound sources is at most two. This restriction should be removed or relaxed. The microphone layout for IRMA should be considered more. For example, a combination of small microphone arrays can reduce the number of microphones while maintaining the total performance of IRMA. The integration of microphone arrays has a possibility to improve not only sound source localization/tracking but also other sound processing such as sound source separation and automatic speech recognition.

Acknowledgement

We would like to thank Satoshi Kaijiri and Shunichi Yamamoto, Kyoto University for their help. We would also like to thank Dr. Futoshi Asano and Dr. Hideki Asoh, AIST for their valuable discussion.

References

[1] Y. Sugita and Y. Suzuki, "Audiovisual perception: Implicit estimation of sound-arrival time," *Nature*, vol. **421**, p. 911, 2003.

[2] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. **264**, pp. 746–748, 1976.

[3] D. H. Mershon *et al.*, Mills, "Perceived loudness and visually-determined auditory distance," *Perception*, vol. **10**, pp. 531–543, 1981.

[4] L. Jeffress, "A place theory of sound localization," *Journal of Comparative Physiology and Psychology*, vol. **41**, pp. 35–39, 1948.

[5] G. Potamianos and C. Neti, "Stream confidence estimation for audio-visual speech recognition," in *ICSLP 2000*. ISCA, 2000, pp. 746–749.

[6] J. Hershey *et al.*, "Audio vision: Using audio-visual synchrony to locate sounds," *NIPS 2000* , vol. 12. MIT Press, pp. 813 – 819.

[7] I. Hara *et al.*, "Robust speech interface based on audio and video information fusion for humanoid HRP-2," *IROS-2004*. pp. 2404–2410.

[8] J.-M. Valin *et al.*, "Localization of simultaneous moving sound sources for mobile robot using a frequency-domain steered beamformer approach," *ICRA 2004*.

[9] K. Nakadai *et al.*, "Improvement of recognition of simultaneous speech signals using av integration and scattering theory for humanoid robots," *Speech Communication*, vol. 44, pp. 97–112, 2004.

[10] P. Aarabi and S. Zaky, "Robust sound localization using multi-source audiovisual information fusion," *Information Fusion*, vol. 2, no. 3, pp. 209–223, 2001.

[11] H. Silverman *et al.*, "The huge microphone array," LEMS, Brown University," Technical Report, 1996.

[12] E. Weinstein *et al.*, "Loud: A 1020-node modular microphone array and beamformer for intelligent computing spaces," MIT, MIT/LCS Technical Memo MIT-LCS-TM-642, 2004.

[13] S. Yamamoto *et al.*, "Improvement of robot audition by interfacing sound source separation and automatic speech recognition with missing feature theory," *ICRA-2004*, pp. 1517–1523.

[14] F. Asano *et al.*, "Real-time sound source localization and separation system and its application to automatic speech recognition," in *Eurospeech 2001*, ISCA, Ed., 2001, pp. 1013–1016.

[15] K. Nakadai *et al.*, "Sound source tracking with directivity pattern estimation using a 64 ch microphone array," *IROS 2005*, pp. 196–202.

[16] S. Thrun *et al.*, *Probabilistic Robotics*. The MIT Press, 2005.

[17] M. S. Arulampalam *et al.*, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Trans. on Signal Processing*, vol. 50, no. 2, pp. 174–188, 2002.

[18] D. B. Ward and R. C. Williamson, "Particle filtering beamforming for acoustic source localization in a reverberant environment," *ICASSP 2002*, vol. II, IEEE, pp. 1777–1780.

[19] D. B. Ward *et al.*, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 6, pp. 826 – 836, 2003.

[20] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," *ICASSP 2001*, vol. 5, IEEE, pp. 3021–3024.

[21] J.-M. Valin, "Auditory system for robot," Ph.D. dissertation, Universitè de Sherbrooke, 2005.

[22] J. MacCormick and A. Blake, "A probabilistic exclusion principle for tracking multiple objects," *Int'l Journal of Computer Vision*, vol. 39, no. 1, pp. 57–71, 2000.

[23] C. Hue *et al.*, "A particle filter to track multiple objects," in *IEEE Workshop on Multi-Object Tracking*, IEEE, Ed., 2001, pp. 61–68.

[24] H. Asoh *et al.*, "An application of a particle filter to bayesian multiple sound source tracking with audio and video information fusion," in *Int'l Conf. on Information Fusion*, 2004, pp. 805–812.

[25] H. Nakajima *et al.*, "Minimum sidelobe beamforming based on minimax criterion," *Journal of Acous. Sci. and Tech.*, pp. 486–488, 2004.