ROBUST TRACKING OF MULTIPLE SOUND SOURCES BY SPATIAL INTEGRATION OF ROOM AND ROBOT MICROPHONE ARRAYS

Kazuhiro Nakadai [*], *Hirofumi Nakajima* [†], *Masamitsu Murase* [‡], *Satoshi Kaijiri* [‡], *Kentaro Yamada* [*],
Takahiro Nakamura [*], *Yuji Hasegawa* [*], *Hiroshi G. Okuno* [‡] *and Hiroshi Tsujino* [*]

∗ Honda Research Institute Japan Co.,Ltd., Saitame, Japan
† Nittobo Acoustic Engineering Co., Ltd., Tokyo, Japan
‡ Graduate School of Informatics, Kyoto University, Kyoto, Japan
nakadai@jp.honda-ri.com

ABSTRACT

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 originates from. This paper addresses sound source tracking by integrating a room and a robot microphone array. The room microphone array consists of 64 microphones attached to the walls. It provides 2D (x-y) sound source localization based on a weighted delay-and-sum beamforming method. The robot microphone array consists of eight microphones installed on a robot head, and localizes multiple sound sources in azimuth. The localization results are integrated to track sound sources by using a particle filter for multiple sound sources. The experimental results show that particle filter based integration reduces localization errors and provides accurate and robust 2D sound source tracking.

1. INTRODUCTION

Robust and real-time robot audition in the real world is essential to realize natural human-robot communication, because humans use a lot of information obtained from environmental sounds including speech signals for their communication. To realize such robot audition, we propose *spatial integration*. Spatial integration means the use of multiple microphone arrays, and integrates them for better sound processing. We considered two types of microphone arrays – a robot-embedded microphone array, and a microphone array installed in a room (hereafter, referred to as robot-MA and room-MA, respectively). The former is promising to improve robot audition directly. Actually, some work [1, 2] has been reported that an 8 ch robot-MA has better performance in sound source localization and separation. 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 microphone array. On the other hand, the latter can solve these problems, because a microphone array is always stationary, and the microphones are distributed over the room. Since this type of microphone array can compensate for the defects, it is effective to improve robot audition indirectly. We can select large size microphone arrays reported in [3] for this purpose, although they focus on only sound source localization and separation.

We used *MUSIC* [1] for a robot-MA, and proposed *weighted delay-and-sum beamforming (WDS-BF)*[8] for a room-MA. However, because of a large number of microphones for the room-MA, the computational cost of the WDS-BF became large. In this paper, we extended the WDS-BF to be faster by using a sub-array method. The sub-array method selects a microphone subset which highly contributes to localize sound sources. Therefore, it makes the beamforming process faster while keeping high performance. The localization results by robot-MA and room-MA are integrated to track sound sources. For integration, we propose a particle filter (PF) for multiple array integration. The PF is a popular method for object tracking and *Simultaneous Localization And Mapping (SLAM)*[4], because it can deal with non-linear motion of an object and the processing speed can be controlled by the number of particles. The PF is basically easy to apply to track a sound source, because the PF needs only probabilistic models on a transition and an observation of the internal states [5, 6, 2, 7]. We constructed an 8 ch robot-MA and a 64 ch room-MA, and show the effectiveness through sound source localization, and sound source tracking by the PF based integration of these microphone arrays.

2. SIGNAL PROCESSING FOR MICROPHONE ARRAYS

2.1. Algorithm for 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 signal 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, that we reported in [8], is generalized to be able 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 $X_n(\omega)$.

1-4244-0469-X/06/$20.00 ©2006 IEEE

In this paper, we introduce the sub-array method. It has two advantages. One is faster processing speed because a subset of microphones is used for localization. The other is improvement in localization accuracy because only channels with high contribution to localization is used. 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.

2.2. Robot Microphone Array

MUSIC for robot-MA is developed for a humanoid robot operating in a daily environment by National Institute of Advanced Industrial Science and Technology (AIST) [1]. In their implementation, pre-measured impulse responses are used as transfer functions to overcome the diffraction of the robot's head and body. This approach is more accurate than model based ones such as [9].

3. INTEGRATION BY PARTICLE FILTER

In the PF, 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. Since the PF allows non-linear transition, 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 our case, two types of observations, $\boldsymbol{Y_{rob}}$ and $\boldsymbol{Y_{room}}$, are obtained from the microphone arrays.

$$\boldsymbol{Y}_{rob}(t) = \{\boldsymbol{y}_{a_1}(t), \cdots, \boldsymbol{y}_{a_l}(t), \cdots, \boldsymbol{y}_{a_L}(t)\}, \quad (3)$$

$$\boldsymbol{Y}_{room}(t) = \{\boldsymbol{y}_{b_1}(t), \cdots, \boldsymbol{y}_{b_m}(t), \cdots, \boldsymbol{y}_{b_M}(t)\} \quad (4)$$

where L and M are the number of observations by room-MA and robot-MA at time t. $\boldsymbol{y}_{a_l}$ includes the localized azimuth in the world polar coordinates, the estimated power. In the world coordinates, the origin is P_0, and 0° is specified as a vector (1,0) and the direction of positive rotation is counterclockwise. $\boldsymbol{y}_{b_m}$ includes the sound location in the world Cartesian coordinates, its orientation and the estimated power.

Our PF consists of the following five steps – "Initialization", "Source Check", "Importance Sampling", "Selection" and "Output".

Initialization makes the initial states of a particle. As the internal states of i-th particle, we defined $\boldsymbol{x}_i(t)$ consisting of the position of a sound source, the velocity and the orientation of the sound source. At the initial state, the particles were distributed uniformly/randomly. To deal with multiple sound sources, we used the initialization of the importance weight,

$$\sum_{i \in P_k} w_i = 1, \quad \sum_{k=1}^{S} N_k = N, \quad (5)$$

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.

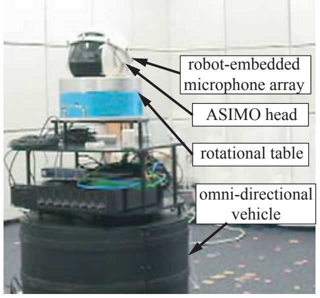


Fig. 1. Wheel-based Robot

Source Check is newly added to support multiple sound sources. The internal state of the particle group P_k is defined by

$$\hat{\boldsymbol{x}}_k(t) = \sum_{i \in P_k} \boldsymbol{x}_i(t) \cdot w_i(t). \quad (6)$$

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. (5) is maintained.

Importance Sampling, first, estimates $\boldsymbol{x}_i(t)$ from $\boldsymbol{x}_i(t-1)$ by using $p(\boldsymbol{x}(t)|\boldsymbol{x}(t-1))$. Secondly, $w_i(t)$ is updated. Finally, $w_i(t)$ is normalized to keep the conditions shown in Eq. (5). 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}, the system uses the transition model based on random walk. Otherwise, the system uses the transition model based on Newton's equation of motion. The likelihood for the observation model is defined as

$$l_{room}(t) = exp\left(-||\boldsymbol{x}_i(t) - \boldsymbol{y}_{b_m}(t)||/2R_{room}\right) \quad (7)$$

$$l_{rob}(t) = exp\left((-\angle(\boldsymbol{x}_i(t)) - \angle(\boldsymbol{y}_{a_l}(t)))^2/2R_{rob}\right) \quad (8)$$

where R_{room} and R_{rob} are variances of room and robot localization. $\angle$ is a function to calculate sound direction in the world polar coordinates. They are integrated into $l_I(t)$.

$$l_I(t) = \alpha_l l_{rob}(t) + (1 - \alpha_l) l_{room}(t) \quad (9)$$

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

$$w_i(t) = l_I(t) w_i(t-1). \quad (10)$$

Selection propagates and removes particles per particle group based on *Sampling Importance Resampling (SIR)* algorithm [10] according to the importance weight w_i.

Output estimates the posterior probability $p(\boldsymbol{x}(t)|\boldsymbol{y}_m(t))$ from the density of the updated particles. The internal states of a set of particles for sound source k is estimated as Eq. (6). These steps are repeated until the tracking process finishes.

4. SYSTEM IMPLEMENTATION

Our spatial integration system consists of three systems – a robot with the robot-MA, the room-MA and an ultra sonic 3D tag system for quantitative evaluation. For the robot-MA system, we developed a wheel-based robot shown in Fig. 1. The robot consists of the head of Honda ASIMO with an 8 ch robot-MA, a rotational

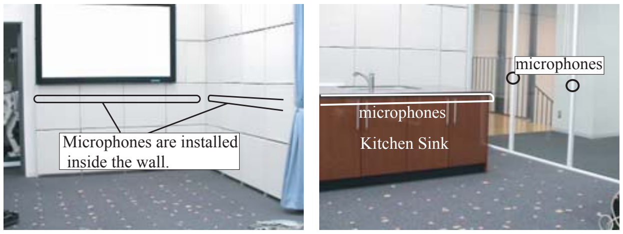


Fig. 2. Room Microphone Array

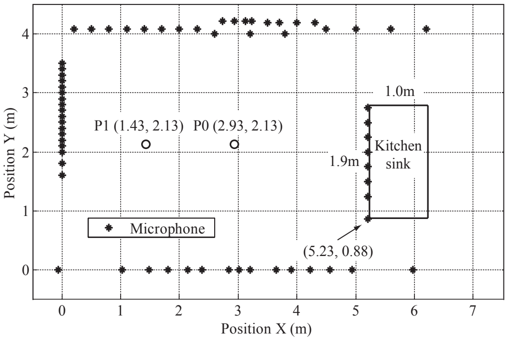


Fig. 3. Layout of Microphones

table with an encoder on an omni-directional vehicle. Each microphone in robot-MA is embedded in a rubber head-band for the head of ASIMO at the same interval. The angle resolution of the rotational table is 0.015 °. For the room-MA system, we constructed a 64 ch microphone array which captures 64 ch signals synchronously at a sampling rate of 16 kHz. Fig. 2 shows a 4.0 m × 7.0 m room with a kitchen sink which room-MA is installed. The three walls are covered with sound absorbing material, and the other wall is made of glass with high sound reflection. Fig. 3 illustrates the layout of microphones in the room. The asterisks are microphone positions in plan view. The height of microphones is 1.2 m. This layout maximizes the number of digitized orientation angles which are able to be estimated. The room is also equipped with an Ultrasonic Three Dimensional Tag System (U3D-TS) [11], which provides in-door GPS function. In our implementation, when the distance between two tags which are located around the center of the room is less than 1 m and from 1 m to 3 m, the errors are around 1 cm and 8 cm, respectively. These errors become 6 cm and 13 cm near the walls.

4.1. Design of beamformer for room microphone array

To design beamformers, the sound position is digitized at the interval of $25\,cm$. The digitizing area is 1.0 m – 5.0 m for X axis, and 0.5 m – 3.5 m for Y axis. The hight (Z axis) is fixed to 1.2 m. So, the total number of digitized points is 221. At each point, the sound orientation is digitized at the interval of $45°$. We, then, designed four types of WDS-BF. We designed a beamformer from measured transfer functions (hereafter, "M-BF"). The transfer functions are obtained by measurements of impulse responses at every digitized point. We then designed the sub-array version of the M-BF ("MS-BF") of which r_{th} is set to 3.5 m. MS-BF is expected to reduce 30% of the computational cost. The other two beamformers are 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 based on a kind of adaptation technique described in [12].

Table 1. Error of Localization

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

Table 2. The Error of Localization and Tracking

	Avg.	S.D.
Robot-MA	7.46 deg	7.83 deg
Room-MA	0.234 m	0.200 m
PF(Room-MA)	0.180 m	0.133 m
PF(Integrated)	0.170 m	0.123 m

5. EVALUATION

We performed two types of evaluations for the spatial integration system, that is, the performance of sound source localization, and sound source tracking.

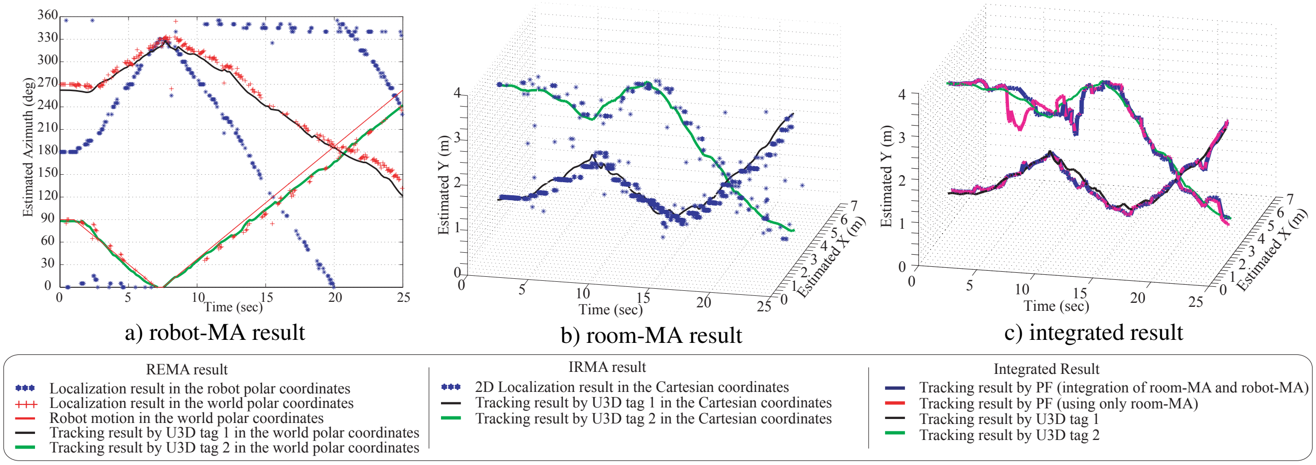
In the first evaluation, a single sound source is localized by the room-MA and robot-MA. As a sound source, we used the recorded voices played by a loudspeaker GENELEC 1029A located at at P1 shown in Fig. 3. The sound source is localized with M-BF, MS-BF, Sim-BF and RSim-BF. Tab. 1 shows the average error and the standard deviation in localization.

In the second evaluation, the performance of tracking of speaking two persons (Mr. A and Mr. B) along the circle with center P_0 and radius 1.5 m. They are asked to speak Japanese sentences continuously and to face with the robot. Mr. A starts at (2.93, 0.63), i.e., 90° in the world polar coordinates, and walks clockwise to 0°. Just before arriving at 0°, he turns back and walks counterclockwise to 270°. Mr. B starts at (2.93, 3.63) and walks in a mirrored way. They approach and recede at 0°, and cross at 180°. The heading of the robot located at P_0 is always kept to face with Mr. A. MS-BF is used as a beamformer for the room-MA. The parameters for the PF is decided experimentally. The number of particles is 2,000. R_{room} and R_{rob} are 0.35 and 0.1. α_l is 0.5. To obtain accurate location of the moving sound sources as reference data, we attached a U3D tag to each sound source. Fig. 4 shows the results of localization and tracking. Tab. 2 shows localization and tracking errors for the robot-MA and the room-MA, and the PF.

5.1. Observations

From the first evaluation, the best beamformer is MS-BF. It has the small localization errors of 15 cm - 20 cm. It is almost the same as M-BF. So, we can say that the sub-array method reduced the computational cost while keeping localization accuracy. MUSIC has the error of about 4.5 deg. This is equivalent to 12 cm at a point 1.5 m away from the robot. It is almost the same accuracy of the room-MA. That is why we use 0.5 for integration weight parameter α_l in the second evaluation.

Fig. 4a) and b) show that, compared with tracking by U3D-TS, both microphone arrays can localize at least two simultaneous speech signals properly even the sources are in motion. In the case of the robot-MA, accurate time synchronization is achieved because the coordinate-converted localization results fit those obtained from U3D-TS. However, some outliers can be seen, and



a) robot-MA result

b) room-MA result

c) integrated result

Fig. 4. Tracking Results

data association between each localization result and the corresponding sound source is not done yet. In tracking of multiple sound sources, this association is essential because miss association causes a fatal tracking error. Fig. 4c) shows that the PF solves this problem. In addition, Tab. 2 shows that the PF improves localization accuracy and robustness because the average errors are reduced 2 cm – 9 cm and the standard deviations are reduced about 10 cm on average. The effect of the microphone array integration looks small, but the integration contributes to improvement in the robustness of the tracking. For example, the tracking result using only room-MA localization results (red lines) have large errors from 5 sec to 10 sec in Fig. 4c), while the integrated ones (blue lines) do not include the large errors.

6. CONCLUSION

We proposed the integration of the robot-MA and the room-MA to enhance robot audition. For the room-MA, we propose subarray based weighted delay-and-sum beamforming. For the integration of microphone arrays, we proposed the PF for multiple sound sources, which can integrate multiple localization results utilizing a probabilistic integration method. The evaluations using a 64 ch room-MA and an 8 ch robot-MA show that the microphone arrays localize multiple sound sources accurately, and sound source tracking with the PF solves the data association problem in case of multiple sound sources, and improves the accuracy and the robustness of sound source localization. In this paper, 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. To relax or remove these restrictions is remaind as future work.

Acknowledgement

We thank Dr. Futoshi Asano and Dr. Hideki Asoh, AIST, and Shunichi Yamamoto, Kyoto University for their valuable discussion.

7. REFERENCES

[1] I. Hara *et al.*, “Robust speech interface based on audio and video information fusion for humanoid hrp-2,” in *Proc. of the IEEE/RAS Intl. Conf. on Intelligent Robots and Systems (IROS-2004)*. 2004, pp. 2404–2410, IEEE.

[2] J.-M. Valin, *Auditory System for Robot*, Ph.D. thesis, Universitè de Sherbrooke, 2005.

[3] H.F. Silverman *et al.*, “Performance of real-time source-location estimators for a large-aperture microphone array,” *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 4, pp. 593 – 605, 2005.

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

[5] E. A. Lehmann *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.

[6] J. Vermaak and A. Blake, “Nonlinear filtering for speaker tracking in noisy and reverberant environments,” in *Proc. of ICASSP 2001*, IEEE, Ed., 2001, vol. 5, pp. 3021–3024.

[7] H. Asoh *et al.*, “An application of a particle filter to bayesian multiple sound source tracking with audio and video information fusion,” in *Proc. of Intl. Conf. on Information Fusion*, 2004, pp. 805–812.

[8] K. Nakadai *et al.*, “Sound source tracking with directivity pattern estimation
using a 64 ch microphone array,” in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS 2005)*, 2005, pp. 196–202.

[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] 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.

[11] Y. Nishida *et al.*, “3D ultrasonic tagging system for observing human activity,” in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS 2003)*, IEEE, Ed., 2003, pp. 785–791.

[12] H. Nakajima *et al.*, “Minimum sidelobe beamforming based on mini-max criterion,” *Journal of Acoustic Science and Technology*, pp. 486–488, 2004.