# Multiple Moving Speaker Tracking by Microphone Array on Mobile Robot

*Masamitsu Murase*†, *Shunichi Yamamoto*†, *Jean-Marc Valin*‡, *Kazuhiro Nakadai**, *Kentaro Yamada**, *Kazunori Komatani*†, *Tetsuya Ogata*†, *Hiroshi G. Okuno*†

†Graduate School of Informatics, Kyoto University, Sakyo, Kyoto, Japan.
‡Université de Sherbrooke, Sherbrooke, Quebec, Canada.
* Honda Research Institute Japan Co., Ltd., Wako, Saitama, Japan
{murase,shunichi,komatani,ogata,okuno}@kuis.kyoto-u.ac.jp,
{nakadai,yamaken}@jp.honda-ri.com, Jean-Marc.Valin@USherbrooke.ca

## Abstract

Real-world applications often require tracking multiple moving speakers for improving human-robot interactions and/or sound source separation. This paper presents multiple moving speaker tracking using an 8ch microphone array system installed on a mobile robot. This problem is difficult because the system does not assume that sound sources and/or the microphone array are fixed. Our solutions consist of two key ideas – time delay of arrival estimation, and multiple Kalman filters. The former localizes multiple sound sources based on beamforming in real time. Non-linear movements are tracked by using a set of Kalman filters with different history lengths in order to reduce errors in tracking multiple moving speakers under noisy and echoic environments. For quantitative evaluation of the tracking, motion references of sound sources and a mobile robot, called SIG2, were measured accurately by ultrasonic 3D tag sensors. As a result, we showed that the system tracked three simultaneous sound sources even when SIG2 moved in a room with large reverberation due to glass walls.

## 1. Introduction

Tracking multiple moving speakers is a critical function for realizing robust human-robot interactions in real-world environments. People may not stay at the same place but move during talking. Or, people do not move but a robot moves. Tracking moving speakers is also needed as a clue for separating speech signals from noises and/or interfering speech signals [1]. Separated speech signals may be used to recognize what each speaker says.

We developed a system that gives a humanoid robot the ability to localize, separate and recognize simultaneous speakers [2]. An 8-channel microphone array (Figure 1) is used along with a real-time dedicated implementation of Geometric Source Separation (GSS) and a multi-channel post-filter that gives us a further reduction of interferences from other sources. An automatic speech recognizer based on the Missing Feature Theory recognizes separated sounds in real-time by generating missing feature masks automatically from the post-filtering step.

This system works well for stationary speakers, but it may often fail with moving speakers due to wrong tracking of them. The previous study on localization of simultaneous moving sound source using frequency-domain steered beamformer approach showed that it localized moving sound sources well at each time frame [3]. However, it did not generate a sound stream originating from the same sound source or speaker. Therefore, it could not track sound sources correctly in case of crossing moving speakers or approaching-then-leaving ones.

Nakadai *et al.* [4] implemented the real-time tracking system of multiple speakers by integrating audio and vision signals. The poor results of auditory localization by a pair of microphones were usually improved by the visual localization by stereo cameras. Since auditory localization gave only azimuths (directions in the horizontal plane), such results were not be used as a clue for separating speech signals.

Asoh, Asano, *et al.* [5] developed the system of tracking human speech signals using a particle filter. It integrated audio and vision signals by a particle filter and showed a good performance of tracking. However, it did not run in real-time on conventional hardware. Furthermore, they did not report how their method worked for crossing or approaching-then-leaving speakers.

This paper presents the tracking mechanism by using Kalman filter and continuity of harmonic structures of speech signals. The robot on which the system was installed successfully tracks multiple moving speakers and disambiguates whether two speakers are crossing or approaching and leaving. The accuracy of localization was evaluated by measuring the precise position of the robot and moving speakers in a room where a sensory network of ultrasonic 3D tag sensors was installed.

## 2. Overview of multiple sound sources tracking system

We used the following procedures in multiple sound source tracking system.

1. Steered beamformer with microphone array [3] localizes sound sources at each time frame.
2. The positions of speakers are estimated by using multiple Kalman filters and localization results.
3. The localization results, which are estimated as the same speakers, are given the same labels.

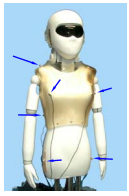


Figure 1: SIG2 robot & 8-channel microphones

The details of the algorithms in each step are described in the following sections.

### 2.1. Sound source localization at each time frame

We use the steered beamformer as a sound source localization method. The basic idea of this method is to direct a beamformer

in all possible directions and look for maximal output. The beamformer searches a spherical space around the microphone array which is divided into 5,120 triangle grids with 2,562 vertices. The beamformer energy is computed for each vertex by incremental refinements from a large triangle to smaller ones. The direction of a sound source is estimated as that of the region with the maximal energy. Thus, this method localized sound source accurately for stationary or moving sound sources [3].

Although this method provided directional information at each time frame, a temporal grouping of the same sound source was not attained. Therefore, it is difficult to track multiple moving sound sources, in particular, to determine whether crossing or approaching-then-leaving speakers. We will attain this kind of disambiguation by using temporal information as described in the next section.

### 2.2. Kalman filter for accurate tracking

To exploit temporal information in tracking multiple sound sources, we use a set of Kalman filters [6] with different history lengths. Kalman filter assumes that the noise is convoluted to the linear transition system. The system is described as follows:

$$x_{k+1} = Fx_k + Gw_k, \quad (1)$$

$$y_k = Hx_k + v_k, \quad (2)$$

where $y_k$ is a vector of observed states at time $k$, $x_k$ is a vector of internal states of the system, $F$ is a matrix for updating internal states, $H$ is a matrix for projecting internal states on observed states, $w_k$ is a process noise, and $v_k$ is an observation noise. In this experiment, variance ratio $\sigma_w/\sigma_v$ is 0.01.

In this paper, a feature vector at time $k$, $p_k$, is defined by

$$p_k = (\theta_k, \phi_k), \quad (3)$$

where $\theta_k$ is an azimuth of a speaker's position, and $\phi_k$ is an elevation of a speaker's position. Then the internal states $x_k$ is defined as follows:

$$x_k = (p_k, p_{k-1}, p_{k-2}, \cdots, p_{k-l}). \quad (4)$$

The internal states $x_k$ means a history in the past $l$ frames.

Suppose that $p_{k+1}$ can be approximated by $p_k$ and $p_{k-l}$, $p_{k+1}$ is defined as follows:

$$p_{k+1} = p_k + \frac{p_k - p_{k-l}}{l}. \quad (5)$$

Then, $F$, $G$, and $H$ are defined as follows:

$$F = \begin{pmatrix} (1+1/l)I & 0 & \dots & 0 & (-1/l)I \\ I & 0 & \dots & 0 & I \\ \vdots & & \ddots & & \vdots \\ 0 & & \dots & & 0 \end{pmatrix}, \quad (6)$$

$$G = \begin{pmatrix} I & 0 & \cdots & 0 \end{pmatrix}^T, \quad (7)$$

$$H = \begin{pmatrix} I & 0 & \cdots & 0 \end{pmatrix}, \quad (8)$$

where $I$ is a unit matrix of $2 \times 2$.

### 2.3. Multiple Kalman filters with different history length

Kalman filter assumes that the state transition is linear. Because this assumption may not hold in the real-world, the performance deteriorates severely. For example, the motion of a speaker moves is non-linear, speech is interrupted, or the number of speakers changes. We use a set of Kalman filters with different history lengths to deal with these nonlinear factors.

If the velocity of a speaker is constant, a filter with a longer history length is appropriate. This is because the motion of a speaker is linear for a long period of time. On the other hand, if the velocity of a speaker drastically changes, a filter with a shorter history length should be used. This is because the motion of the speaker is nonlinear.

We use a set of multiple Kalman filters with different history lengths to cope with a wide variety of motions by selecting an appropriate history length. Multiple Kalman filters with different history length estimate next states in parallel, and provides a set of estimates. The current estimate is obtained by the filter which estimated the state with the minimal error in the previous frame. Therefore, the method can cope with the two cases that the velocity of a speaker is constant and drastically changed.

Let $p(t)$ and $\hat{K}_l(t)$ be the observed value and the estimated value obtained by filter $l$ at time $t$, respectively. The estimation algorithm with $N$ filters is as follows:

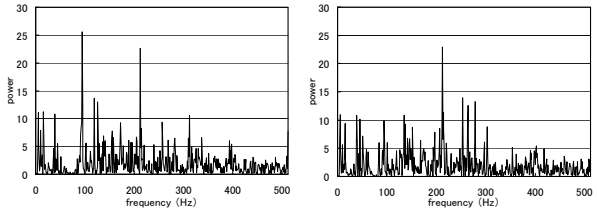
1. When the number of steps is less than or equal that of histories in Kalman filter, the observed value which is the nearest to the previous observed one is selected. The selected value is used in order to update Kalman filters.
2. In the later step, the estimation at time $t$ is as follows:
   (a) When the error between the observed value $p(t-1)$ and the estimated value $\hat{K}_l(t-1)$ is minimal, the estimated value $\hat{K}_l(t-1)$ obtained by Kalman filter $K_l$ is selected.
   $$l = \underset{i=0,\cdots,N-1}{\operatorname{argmin}} \left\| \hat{K}_i(t-1) - p(t-1) \right\|. \quad (9)$$
   (b) The observed values of speakers' position are obtained by a sound source localization. The observed values whose difference with the estimated value $\hat{K}_l(t)$ is less than or equal to a threshold $\delta$ are selected. The value which is acoustically the nearest to the estimated value $\hat{K}_l(t)$ are estimated as the speaker's observed value.
   (c) If the observed value whose difference with the estimated value $\hat{K}_l(t)$ does not exist, $l$th Kalman filter is excluded, and return to step 2(a).
3. By using the obtained observed value $p(t)$, all Kalman filters are updated.
4. The true value is estimated by using the observed value $p(t)$, and return to step 2.

In this experiment, we use three Kalman filters, whose history lengths are 3, 5 and 10 frames (120, 200 and 400 [ms]).

## 3. Further improvement of tracking by using acoustic features of speech signals

The continuity of localization for each speaker is forced by using acoustic features of separated speech signals. We focus on a power spectrum as acoustic features in each frame in order to reduce the ambiguities in tracking moving speakers. In the step 2(b) of the algorithm described in Section 2.3, the observed value whose power spectrum is similar to the past one of a speaker is selected as the observed value of the speaker.

The powere spectrum of a seperated sound is calculated by using a delay-and-sum beamformer, which uses a localization as a clue. If the separated speech of each moving speaker




(a) Speaker1 (b) Speaker2

Figure 2: The power spectrum of two separated sounds

is available, the fundamental frequency of the speech may be used for the accurate selection of the observed value. This is a chicken-and-egg problem, because the sound source separation needs an accurate sound source localization. Therefore, we do not use another sound source separation system.

Suppose that two sound souces are estimated in the direction of $d_1$ and $d_2$ at time $t$. The spectrum of a sound seperated by a delay-and-sum beamformer in the direction of $d_s$, $D_{d_s}$, is calculated as follows:

$$D_{d_s} = \sum_{i=0}^{M-1} x_i(t - \tau_{d_s,i}). \quad (10)$$

where $M$ is the number of microphones, and $\tau_{d_s,i}$ is a time delay of arrival for the microphone $i$ from the direction $d_s$. FFT is then applied to this enhanced spectrum as follows:

$$\begin{aligned} \mathcal{F}[D_{d_s}] &= \mathcal{F}\left[\sum_{i=0}^{M-1} x_i(t - \tau_{d_s,i})\right] \\ &= \sum_{i=0}^{M-1} \exp\left(\frac{-2\pi i k \tau_{d_s,i}}{L} X_i(k)\right), \end{aligned} \quad (11)$$

where $\mathcal{F}[g]$ is the resulting function of $g$ applied by FFT with $L$ points of window, and $X_i(k)$ is the value of FFT for $x_i(t)$.

The cosine similarity between the spectra is used to define their distance. Let $S$ be the number of localization candidates. When $d_0, d_1, \cdots, d_{S-1}$, are obtained by sound source localization, the most plausible direction $d_s$ of a speaker is defined as follows:

$$s = \underset{i=0,\cdots,S-1}{\operatorname{argmin}} \mathcal{F}[D_{d_i}] \cdot \mathcal{F}[D]_{t-1}, \quad (12)$$

where $\mathcal{F}[D]_{t-1}$ is the power spectrum of the speaker in the previous frame.

Let's consider the situation that two speakers are approaching each other up to 15 degrees. The power spectra of two speakers obtained by this method are shown in Figure 2(a) and (b), respectively. Acoustic signals sampled by 48 kHz were analyzed by FFT with 1,024 points of window. When two speakers have different pitches (fundamental frequencies), the obtained power spectra are apparantly different.

# 4. Evaluation

Our system is evaluated with the humanoid robot, SIG2, on which an 8-channel microphone array is installed. We also use a sensor-network room where a precice position of an object attached by an ultrasonic 3D tag sensor can be measured. The ultrasonic 3D tag sensor system is originally developped by AIST [8]. In our instllation, the error of an ultrasonic 3D tag sensor is about 5 cm at the center of the room and it is about 10 cm near the walls. The sampling rate of the ultrasonic 3D tag system is 20 Hz. This room has large reverberation due to glass walls. We

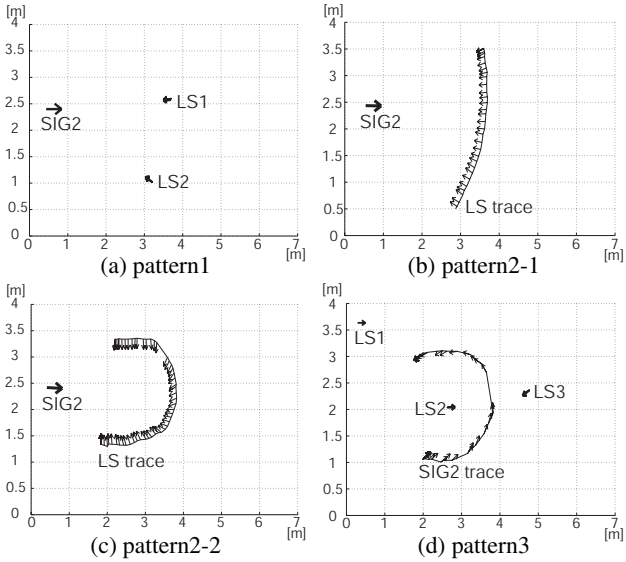



Figure 3: Allocation of SIG2 and loudspeakers ("LS" denotes "loudspeaker".)

used combinations of three different sentences selected from the ATR DataBase of phonetically-balanced Japanese sentences.

Since a precise position of only one object can be measured in this room, the following three patterns are evaluated.

**pattern1** SIG2 and a single loudspeaker are stationary. Their distance is 3m. The loudspeaker is placed:

1. in front of SIG2. (LS1 in Figure 3(a))
2. in the direction of 30 degrees to the right of the center of SIG2. (LS2 in Figure 3(a))

Hereafter, "pattern *x*-*y*" denotes experiment *y* in pattern *x*.

**pattern2** SIG2 is stationary and a single loudspeaker moves as follows:

1. it moves along circumference whose radius is about 3m. (Figure 3(b))
2. it moves around SIG2 non-linearly. (Figure 3(c))

**pattern3** SIG2 moves around stationary loudspeaker(s) non-linearly. As shown in Figure 3(d), we placed:

1. one loudspeaker. (LS1)
2. two loudspeakers. (LS1 and LS2)
3. three loudspeakers. (LS1, LS2 and LS3)

### 4.1. Analysis

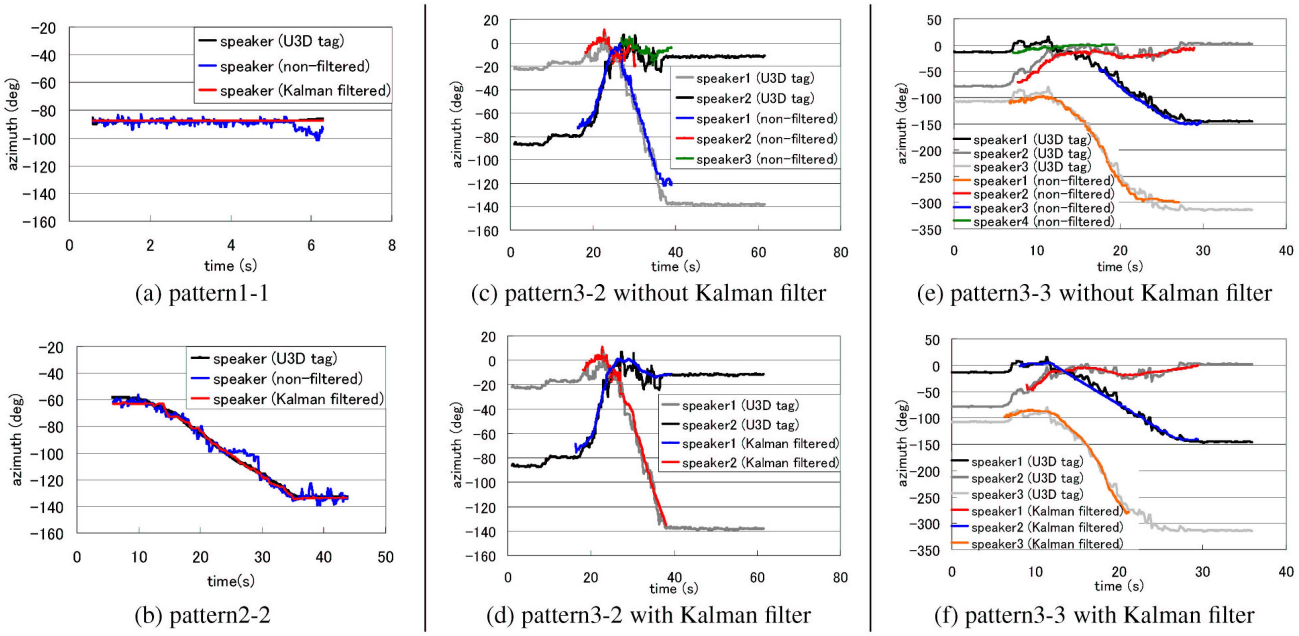
We localized loudspeakers with using 8ch sound recorded with a sampling rate of 48kHz. Then we compared accuracy of the method of giving labels based on a localization value predicted by multiple Kalman filters (*our method*) with accuracy of the method of giving the same speaker labels to near localization outputs (*baseline*). We obtained real positions of SIG2 and loudspeaker(s) by ultrasonic 3D tag sensors and we compared mean square errors of result of each method in order to evaluate accuracy of these methods.

### 4.2. Results

The results of loudspeaker(s) tracking are shown in Figure 4.

Mean square errors for each pattern are shown in Table 1.

We compared the accuracies of two methods:




(a) pattern1-1 (c) pattern3-2 without Kalman filter (e) pattern3-3 without Kalman filter

(b) pattern2-2 (d) pattern3-2 with Kalman filter (f) pattern3-3 with Kalman filter

Figure 4: The results of tracking ("U3D tag" denotes "ultrasonic 3D tag".)

1. *baseline*: the method based on the proximity in localization, that is, the same speaker label is given to the nearest localization,
2. *our method*: the method with multiple Kalman filters.

Figure 4(b) shows that our method tracked a moving speaker continuously and well, while the baseline failed in tracking continuously and obtained two parts of a single speaker.

In pattern3, SIG2 observed that multiple speakers were crossing. Therefore, there was an ambiguity problem when tracking. Figure 4(c) and (d) show that our method correctly tracked multiple moving speakers, while baseline misunderstood that they were not crossing. Therefore, mean square errors of baseline cannot be computed for pattern3-2 and 3-3.

Table 1: Mean square errors

| pattern | *baseline* | *our method* |
|---|---|---|
| 1-1 | 29 ($deg^2$) | 0.34 ($deg^2$) |
| 1-2 | 35 ($deg^2$) | 1.0 ($deg^2$) |
| 2-1 | 35 ($deg^2$) | 2.1 ($deg^2$) |
| 2-2 | 22 ($deg^2$) | 3.6 ($deg^2$) |
| 3-1 | 53 ($deg^2$) | 25 ($deg^2$) |
| 3-2 | — | 26 ($deg^2$) |
| 3-3 | — | 28 ($deg^2$) |

"—" indicates failure of tracking.

Table 1 shows that our method reduces localization errors when a mobile robot moves or loudspeakers move.

## 5. Conclusions

We presented the design and implementation of the method that solved the following issues in tracking multiple moving speakers:

- pursuit of temporal continuity of speakers' labels,
- disambiguation of crossing speakers or approaching-then-leaving speakers.

We used multiple Kalman filters with different history lengths and selected the most plausible predicted value to solve these issues. The experiments showed that our method was effective in solving these issues. The performance for the second issue was further improved by using acoustic features in speech signals. As a result, multiple moving speakers could be tracked successfully by our method even when speakers and a mobile robot moved non-linearly.

These evaluations were based on the precise measuremens of localization of moving SIG2 robot or loudspeakers by using ultrasonic 3D tag sensors. The capability and performance of this room with sensor-networks will be reported by a separate paper. We believe that this kind of quantitative evaluation for moving speaker tracking has not been reported.

The future work includes integration of visual information for tracking, improvement of sound source separation of moving speakers, and automatic speech recognition of simultaneous moving speaker speech signals.

## 6. References

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

[2] S. Yamamoto, *et al.*, "Enhanced Robot Speech Recognition Based on Microphone Array Source Separation and Missing Feature Theory," *Proc. IEEE ICRA-2005*.

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

[4] K. Nakadai, *et al.*, "Real-time auditory and visual multiple-object tracking for robots," *Proc. IJCAI-2001*.

[5] H. Asoh, *al.*, "Tracking Human Speech Events using a Particle Filter," in *Proc. IEEE ICASSP-2005*, 1153–1156.

[6] E. Kalman, "A new approach to linear filtering and prediction problems," *Trans. ASME – J. of Basic Engineering*, vol. 82, 35–45, 1960.

[7] J.-M. Valin, *et al.*, "Enhanced robot audition based on microphone array source separation with post-filter," *Proc. IEEE/RSJ IROS-2004*.

[8] Y. Nishida, *et al.*, "3D Ultrasonic Tagging System for Observing Human Activity", *Proc. IROS 2003*.