An Application of a Particle Filter to Bayesian Multiple Sound Source Tracking with Audio and Video Information Fusion

Hideki Asoh, Futoshi Asano, Takashi Yoshimura
Yoichi Motomura, Naoyuki Ichimura, Isao Hara, Jun Ogata
Information Technology RI
National Institute of Industrial Science and Technology
AIST Central 2, 1-1-1 Umezono
Tsukuba, Ibaraki 305-8568, Japan
h.asoh@aist.go.jp

Kiyoshi Yamamoto
Institute of
Information Science & Electronics
University of Tsukuba
1-1-1 Tennodai
Tsukuba, Ibaraki 305-8573, Japan

Abstract – *A particle filter is applied to the problem of detecting and tracking multiple sound sources by Bayesian inference using combined audio and video information. The problem is formulated within a general framework of Bayesian hidden variable sequence estimation by fusing observed information. The particle filter is then introduced as an approximation of Bayesian inference. Experiments using real-world data demonstrate that the proposed method works well in ordinary environments such as a meeting room. The computational cost of estimation is reduced significantly compared to exact Bayesian inference, while maintaining the quality of estimation.*

Keywords: Particle Filter, Bayesian Multiple Target Tracking, Sound Source Tracking.

1 Introduction

The importance of statistical inference using probabilistic models in multi-modal information fusion has been pointed out by many researchers, and the Bayesian approach in particular has been promoted and applied to various problems in recent years. In the Bayesian approach, all variables in models are treated as random variables and both inference and learning can be treated in a unified way as a computation of conditional probabilities of the target (hidden) variables or parameters conditioned by observed information.

Although the Bayesian approach is very simple and powerful in principle, the central drawback in practice is that it often requires intractably large amount of computations, mainly for the execution of integrations in a very high-dimensional space of random variables. Many ideas for reducing the amount of computation by approximation have been proposed and investigated. Monte Carlo methods using randomized point mass representations of probability distributions are one of the most general and promising techniques, and are most effective when the target probability distributions are localized into several points in high-dimensional space.

For the estimation of a sequence of hidden variables, the estimation needs to executed iteratively at every time step. Particle filters (Monte Carlo filters) are a very efficient framework for executing such incremental estimations using Monte Carlo methods [1, 2, 3]. The simplicity of the idea and the operation of particle filters have attracted many researchers, and various kinds of particle filters have been proposed and applied to many problems, including radar/visual target tracking and robot localization [4, 5, 3].

In this paper, a particle filter is applied to the problem of multiple sound source tracking based on audio and video information fusion. The tracking of user speech events and enhancement of target speech signals are indispensable for constructing robust speech/multi-modal user interface systems that can work in real environments, which are typically full of background noise and interference. In order to focus on user speech among other speech signals from interference sound sources such as television, we have proposed to combine audio and video information using a Bayesian network [6, 7]. In this scheme, a microphone array is employed to localize sound sources, and a stereo camera is used to localize user locations. Both sets of information are then combined to compute the posterior conditional probability distribution of the time and location of human speech events using the Bayesian network.

In previous work, it was assumed that the speakers are stationary during each utterance. In this study, the scheme is extended to moving speakers. The problem is formulated within the framework of the Bayesian hidden variable sequence estimation, equivalent to extending the Bayesian network to a dynamic Bayesian network to account for the dynamics of status of sound sources. As this extension results in an increase in computational cost, a particle filter is applied to facilitate a real-time implementation. Although particle filters have been applied to many problems such as target tracking using video or radar information, the effectiveness of these filters for processing audio information or in multi-modal information fusion problems has yet to be explored in any detail.

This paper is organized as follows. In Section 2, a very general formulation of the Bayesian approach to the problem of hidden variable sequence estimation with information fusion is presented, and several Markov assumptions are introduced in order to reduce the size of computation and training data. In Section 3, a basic particle filter algorithm for the above problem is described, and in Section 4, the framework is applied to the speech event tracking problem. The detailed design of the probability model specified for the problem is also shown. The results of experiments using data obtained in a real office environment are shown in Section 5, and the efficiency of the particle filter is evaluated in comparison to rigorous Bayesian computation. Section 6 provides a discussion of the results and some conclusions.

2 Bayesian Approach to Hidden Variable Sequence Estimation with Information Fusion

Many problems of information fusion can be formulated as problems of estimating hidden variable sequences from sequences of multi-modal sensory observations. Let a hidden state sequence and an observed sequence from time $t = 1$ to T be denoted by $\mathbf{X}_{1|T} = \mathbf{X}(1), ..., \mathbf{X}(T)$ and $\mathbf{Y}_{1|T} = \mathbf{Y}(1), ..., \mathbf{Y}(T)$, respectively. Here, $\mathbf{X}(t) = (X_1(t), ..., X_{N_s}(T))'$ and $\mathbf{Y}(t) = (Y_1(t), ..., Y_{N_o}(T))'$, where $'$ is the transform of a vector. N_s is the number of hidden state variables, and N_o is the number of observed variables. The problem then becomes the estimation of the value of a hidden variable sequence $\mathbf{X}_{1|T}$ from an observed sequence $\mathbf{Y}_{1|T} = \mathbf{y}_{1|T}$.

In the Bayesian approach, the relationship between $\mathbf{X}_{1|T}$ and $\mathbf{Y}_{1|T}$ is modeled by a joint probability distribution $P(\mathbf{X}_{1|T}, \mathbf{Y}_{1|T})$, and estimation is preformed by computing the posterior probability distribution $P(\mathbf{X}_{1|T}|\mathbf{Y}_{1|T} = \mathbf{y}_{1|T})$. In the following, we write $P(\mathbf{X}_{1|T}|\mathbf{Y}_{1|T} = \mathbf{y}_{1|T})$ as $P(\mathbf{X}_{1|T}|\mathbf{y}_{1|T})$.

There are several types of problem settings [8]. The first is *filtering*, which involves estimating the value of $\mathbf{X}(t)$ using observations up to time t, that is, $\mathbf{y}_{1|t}$. In the Bayesian sense, the problem is the computation of $P(\mathbf{X}(t)|\mathbf{y}_{1|t})$. The second setting is *prediction*, which is similar to filtering, but the target is to estimate the value of $\mathbf{X}(t + \lambda)$ for some $\lambda > 0$ using observations up to time t. In particular, the case of $\lambda = 1$ is often investigated as "one-step look ahead prediction". The third setting is *smoothing*, in which the information from observations for later than time t can be used. In typical settings, smoothing involves the estimation of $\mathbf{X}_{1|T}$ using all observations $\mathbf{y}_{1|T}$. In the Bayesian formulation, computing the joint conditional probability distribution $P(\mathbf{X}_{1|T}|\mathbf{y}_{1|T})$, or computing the distribution for each variable $P(\mathbf{X}(t)|\mathbf{y}_{1|T})$ for all $t = 1, ..., T$ is the problem.

The following treatment concentrates on the filtering problem. However, the results can be extended to other problem settings without difficulty.

2.1 Markov Assumptions

If no prior knowledge of the joint probability distribution $P(\mathbf{X}_{1|T}, \mathbf{Y}_{1|T})$ is held, it is very difficult to estimate the distribution from data alone because the degree of freedom of the distribution is very large. Hence, some prior knowledge or heuristics is usually introduced in order to restrict the form of the distribution. One of the most often-used assumptions is the Markov assumption. Although there are many variations of the Markov assumption, the typical form is described here and is used in the experiments.

Using the definition of conditional probability repeatedly, the joint distribution can be decomposed as follows.

$$P(\mathbf{X}_{1|T}, \mathbf{Y}_{1|T}) = \prod_{t=1}^{T} P(\mathbf{X}(t), \mathbf{Y}(t)|\mathbf{X}_{1|t-1}, \mathbf{Y}_{1|t-1}).$$

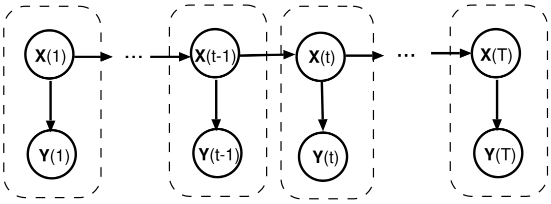

Fig. 1: Dynamic Bayesian network for hidden variable sequence estimation

Here, dummy variables $\mathbf{X}(0)$ and $\mathbf{Y}(0)$ are introduced for convenience to write $P(\mathbf{X}(1), \mathbf{Y}(1))$ as $P(\mathbf{X}(1), \mathbf{Y}(1)|\mathbf{X}(0), \mathbf{Y}(0))$. On the right-hand side of the decomposition, each $P(\mathbf{X}(t), \mathbf{Y}(t)|\mathbf{X}_{1|t-1}, \mathbf{Y}_{1|t-1})$ can be further decomposed to

$$P(\mathbf{Y}(t)|\mathbf{X}_{1|t}, \mathbf{Y}_{1|t-1})P(\mathbf{X}(t)|\mathbf{X}_{1|t-1}, \mathbf{Y}_{1|t-1}).$$

The Markov assumptions can then be introduced. For the first term $P(\mathbf{Y}(t)|\mathbf{X}_{1|t}, \mathbf{Y}_{1|t-1})$, it is assumed that

$$P(\mathbf{Y}(t)|\mathbf{X}(t)) = P(\mathbf{Y}(t)|\mathbf{X}_{1|t}, \mathbf{Y}_{1|t-1})$$

holds for all t. For the second term, it is assumed that

$$P(\mathbf{X}(t)|\mathbf{X}(t-1)) = P(\mathbf{X}(t)|\mathbf{X}_{1|t-1}, \mathbf{Y}_{1|t-1})$$

holds. Then $P(\mathbf{X}(t)|\mathbf{X}(t-1))$ and $P(\mathbf{Y}(t)|\mathbf{X}(t))$ are called the (hidden) state transition probability (dynamics model) and the observation probability (sensor model), respectively.

In summary, we obtain a simplified model of the joint probability distribution:

$$P(\mathbf{X}_{1|T}, \mathbf{Y}_{1|T}) = \prod_{t=1}^{T} P(\mathbf{Y}(t)|\mathbf{X}(t))P(\mathbf{X}(t)|\mathbf{X}(t-1)).$$

In order to specify the joint probability distribution, only the state transition probabilities and observation probabilities need to be given. This significantly reduces the amount of training data required to specify the joint distribution. This equation also means that the joint probability distribution can be written as a simple dynamic Bayesian network (DBN), as illustrated in Figure 1.

This structure is a popular one. As is well known, when variables $\mathbf{X}(t)$ and $\mathbf{Y}(t)$ are continuous and the state transition and observation probabilities can be modeled by a linear Gaussian model, the model becomes the Kalman filter. When hidden variables $\mathbf{X}(t)$ take discrete states as values, the model becomes the hidden Markov model (HMM), which is often used in speech recognition.

2.2 Recursions

From the structure introduced into the joint probability distribution, several recursions can be derived for efficient computation of the various probability distributions. For example, $P(\mathbf{Y}_{1|t}, \mathbf{X}(t))$ can be computed recursively as

$$P(\mathbf{Y}_{1|t}, \mathbf{X}(t)) = P(\mathbf{Y}(t)|\mathbf{X}(t)) \times \sum_{\mathbf{X}(t-1)} P(\mathbf{X}(t)|\mathbf{X}(t-1))P(\mathbf{Y}_{1|t-1}, \mathbf{X}(t-1)).$$

From this recursion, another recursion for the posterior distribution $p(\mathbf{X}(t)|\mathbf{Y}_{1|t})$ can be easily derived:

$$P(\mathbf{X}(t)|\mathbf{Y}_{1|t}) = \frac{1}{Z} P(\mathbf{Y}(t)|\mathbf{X}(t)) \times \sum_{\mathbf{X}(t-1)} P(\mathbf{X}(t)|\mathbf{X}(t-1))\ P(\mathbf{X}(t-1)|\mathbf{Y}_{1|t-1}).$$

Here, Z is a normalization constant that is independent of $\mathbf{X}(t)$.

In this recursion, there is a summation taken over $\mathbf{X}(t-1)$. For the Kalman filter model, this summation can be reduced to the operation of the mean vector and the covariance matrix of Gaussian distribution. However, in other cases, such as discrete hidden variables or non-linear non-Gaussian (multi-modal) state transition/observation distributions, the computational cost increases exponentially according to the dimension, and it becomes intractable to compute the posterior distribution exactly when the dimension of $\mathbf{X}(t-1)$ is large. An approximation method to obtain the posterior is therefore necessary, for which particle filters are commonly used.

3 Particle Filters

Monte Carlo methods were originally developed for statistical physics, where they were used for stochastic simulation of complex systems with large degree of freedom through the generation of random numbers. As such methods provide a very flexible framework, their application very quickly spread to many different areas. In recent years, Monte Carlo methods have been very often used in Bayesian statistical inference to represent probability distributions and to facilitate inference in high-dimensional state spaces.

N independent and identically distributed random samples (particles) $\{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(N)}\}$ can be generated according to a probability distribution $P(\mathbf{X})$, and the distribution P can be approximated from these samples as follows.

$$P(\mathbf{X}) \approx \frac{1}{N} \sum_{i=1}^{N} \delta(\mathbf{X} - \mathbf{x}^{(i)}),$$

where $\delta()$ denotes Dirac's delta function. Using this approximation, for example, the mean of an arbitrary function $f(\mathbf{X})$ of $\mathbf{X}$ can be approximated as

$$\sum_{\mathbf{X}} f(\mathbf{X}) P(\mathbf{X}) \approx \sum_{i=1}^{N} \frac{1}{N} f(\mathbf{x}^{(i)}).$$

This approximation is unbiased, and the rate of convergence to the true value is independent of the dimension of the variable $\mathbf{X}$. The set of particles represents the distribution function in this sense. This is the simplest version of Monte Carlo sampling, and is called the perfect sampling method.

When it is difficult to sample efficiently from the target distribution $P(\mathbf{X})$ itself, the importance sampling method can be used. Considering an arbitrary proposal distribution $Q(\mathbf{X})$, an importance weight function can be defined as

$$w(\mathbf{x}) = \frac{P(\mathbf{x})}{Q(\mathbf{x})}.$$

Then, if N i.i.d samples $\{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(N)}\}$ can be obtained according to the proposal distribution $Q(\mathbf{X})$, an approximate representation can be obtained as

$$P(\mathbf{X}) \approx \sum_{i=1}^{N} \tilde{w}^{(i)} \delta(\mathbf{X} - \mathbf{x}^{(i)}).$$

Here,

$$\tilde{w}^{(i)} = \frac{w(\mathbf{x}^{(i)})}{\sum_{j=1}^{N} w(\mathbf{x}^{(j)})}$$

is the normalized weight of the ith particle. The approximation of the mean value of $f(\mathbf{X})$ then becomes

$$\sum_{\mathbf{X}} f(\mathbf{X}) P(\mathbf{X}) \approx \sum_{i=1}^{N} \tilde{w}^{(i)} f(\mathbf{x}^{(i)}).$$

The set of weighted particles represents the distribution. The perfect sampling method is a special case in which all normalized weights are equal to $1/N$. The design of the proposal distribution is an important research issue for deriving efficient approximations.

In the present Bayesian inference problem, the problem is to obtain a set of weighted particles to approximate the posterior $p(\mathbf{X}(t)|\mathbf{y}_{1|t})$. It is obvious that perfect sampling cannot be used in this case because only the transition probabilities and observation probabilities are known, and the approximation was introduced to represent the posterior itself. However, by repeated sampling using recursion, the particles can be generated incrementally.

If it is assumed that a set of weighted particles $\mathbf{x}_{t-1}^{(i)}$ and $\tilde{w}_{t-1}^{(i)} (i = 1 \cdots N)$ are obtained, approximating $P(\mathbf{X}(t-1)|\mathbf{y}_{1|t-1})$, it is possible to sample $\mathbf{x}_t^{(i)}$ from $P(\mathbf{X}(t)|\mathbf{x}_{t-1}^{(i)})$ and set the weight of the sample as follows.

$$\tilde{w}_t^{(i)} = \frac{1}{Z} \tilde{w}_{t-1}^{(i)} P(\mathbf{y}(t)|\mathbf{x}_t^{(i)}),$$

where Z is a normalization constant satisfying $\sum_i \tilde{w}_t^{(i)} = 1$. This method is called sequential importance sampling.

This framework can be generalized even further. Instead of using the sequence of posterior distributions, other sequence of proposal distributions $Q(\mathbf{X}(t)|\mathbf{Y}_{1|t})$ can also be employed. Assume that $Q(\mathbf{X}(t)|\mathbf{Y}_{1|t})$ can be decomposed as follows.

$$Q(\mathbf{X}(t)|\mathbf{Y}_{1|t}) = Q(\mathbf{X}(t)|\mathbf{X}(t-1), \mathbf{Y}(t)) Q(\mathbf{X}(t-1)|\mathbf{Y}_{1|t-1}).$$

Then, if $Q(\mathbf{X}(t)|\mathbf{X}(t-1), \mathbf{Y}(t))$ is known, sequential importance sampling can be performed using Q. That is, $\mathbf{x}_t^{(i)}$ is sampled from $Q(\mathbf{X}(t)|\mathbf{x}_{t-1}^{(i)}, \mathbf{y}(t))$ and the weight is set according to

$$\tilde{w}_t^{(i)} = \frac{1}{Z} \tilde{w}_{t-1}^{(i)} \frac{P(\mathbf{y}(t)|\mathbf{x}_t^{(i)}) P(\mathbf{x}_t^{(i)}|\mathbf{x}_{t-1}^{(i)})}{Q(\mathbf{x}_t^{(i)}|\mathbf{x}_{t-1}^{(i)}, \mathbf{y}(t))}.$$

A serious practical problem encountered with sequential importance sampling is that the variance of weights very quickly converges to 0 as time t increases. This means that the weights for very few particles are close to 1, while the weights of all others are 0. In order to remedy this problem, an additional resampling step is introduced. Here, the samples obtained in the above sequential sampling step are denoted $\tilde{\mathbf{x}}_t^{(i)}$. In the resampling step, resampling is performed using new N particles $\mathbf{x}_t^{(i)}$ taken from the set of samples $\tilde{\mathbf{x}}_t^{(i)}(i = 1, ..., N)$ using the normalized weight $\tilde{w}_t^{(i)}$ as the sampling probability. By this resampling step, the weights for the new particles are reset to $1/N$.

In summary, an algorithm for a simple particle filter can be described as follows [9].

Initialization

Sample $\mathbf{x}_1^{(i)}$ from $P(\mathbf{X}(1))$ for $i = 1, ..., N$ and set $t = 2$.

Iteration

Repeat the following steps.

Importance sampling

Sample $\tilde{\mathbf{x}}_t^{(i)}$ from $P(\mathbf{X}(t)|\mathbf{x}_{t-1}^{(i)})$ and set the normalized importance weights

$$\tilde{w}_t^{(i)} = \frac{1}{Z} P(\mathbf{y}(t)|\mathbf{x}_t^{(i)})$$

Resampling

Resample new N particles $\mathbf{x}_t^{(i)}$ from the set of samples $\tilde{\mathbf{x}}_t^{(j)}(j = 1, ..., N)$ according to the normalized weight $\tilde{w}_t^{(j)}$.

This algorithm can be easily modified for smoothing or prediction problems.

4 Application to Multiple Sound Source Tracking

Now the Bayesian hidden variable sequence estimation using particle filter is applied to the problem of multiple sound source tracking. As is described before, tracking users' speech events and enhancing target speech signals are indispensable for robust speech/multi-modal user interface systems, which can work in real environments full of background noise and interference. Importance of the technologies is increasing recently as various personal information tools such as PDA, personal robot, and so on, become popular. In addition, from the technical point of view, the problem of tracking sound sources has its own difficulties, in comparison with visual tracking and tracking with antenna array. Because sound waves have lower frequency than light or radio wave, the effect of reverberation is much more serious, and information fusion is a very promising way to reduce the difficulty. These are the reasons why we choose this problem as an application.

As the framework of the Bayesian estimation is a very general one, the random variables $\mathbf{X}(t)$ and $\mathbf{Y}(t)$ should be refined in the model and the relationship between the variables should be designed to fit to the problem. In the present problem setting, the raw observations are audio signals from a microphone array and video signals from a stereo camera. Figure 2 shows the array and the camera used to obtain data for the following experiments.

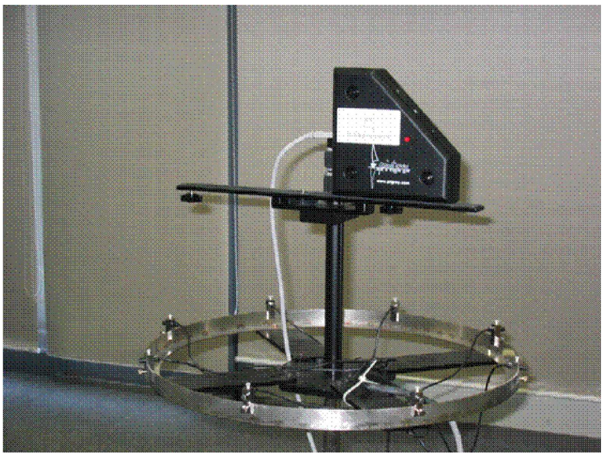

Fig. 2: Microphone array and stereo camera

The hidden variables to be estimated are the location (direction) and status (speaking or silent) of human speakers. Additional hidden variables that influence the observations are the location and status of interference sound sources such as televisions and loudspeakers. The maximum number of speakers and noise sources are assumed to be known. The speaker locations are written as $Lh_1(t), ..., Lh_{N_h}(t)$, the speaking status (speaking or not) is written as $Sh_1(t), ..., Sh_{N_h}(t)$, the location of interference sources is written as $Ln_1(t), ..., Ln_{N_n}(t)$, and the status of the interference sources (making a sound or not) is written as $Sn_1(t), ..., Sn_{N_n}(t)$. Here, N_h and N_n are the maximum number of humans and the maximum number of sound sources, respectively. $\mathbf{X}(t)$ denotes the set of all hidden variables.

The relationship between the observed data and the hidden variables is simplified by first extracting essential information from the raw audio and video data. From the audio signals, the locations of sound sources are extracted by a sound source localization method. From the video signals, the locations of speakers are extracted. The extracted information are referred to as features, following the convention of pattern recognition. The features are denoted by binary variables $A_1(t), ..., A_{N_a}(t)$ and $V_1(t), ..., V_{N_v}(t)$, which take values of 0 or 1. $A_i(t) = 1$ means that there is at least one sound source in the ith direction, and $V_j(t) = 1$ means that there is at least one human in the jth direction. These parameters function as virtual sensors. N_a and N_v are the number of discretized directions for sound source localization and human detection, respectively. $\mathbf{Y}(t)$ denotes the set of all features.

The relationship between the features and the hidden variables is modeled by $P(\mathbf{Y}(t)|\mathbf{X}(t))$, and the dynamics of hidden variables is modeled by $P(\mathbf{X}(t)|\mathbf{X}(t-1))$. When the location of speakers and interference sources are discretized into N_l areas, $P(\mathbf{Y}(t)|\mathbf{X}(t))$ is represented as a conditional probability table of size $(N_l \times 2)^{(N_h+N_n)} \times 2^{(N_a+N_v)}$. As such a table is huge, the naive Bayes assumption is introduced. That is, conditional independence

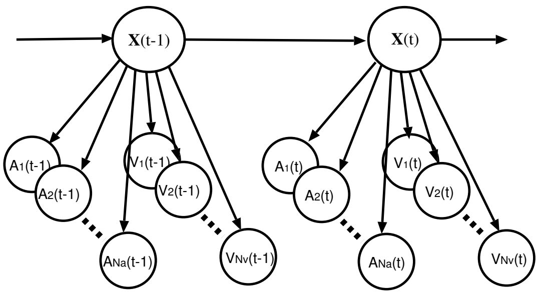

Fig. 3: Dynamic Bayesian network for audio-video information fusion and speech event tracking

of all $A_i(t)$ and $V_j(t)$ is assumed given the value of hidden variables:

$$P(A_1, \cdots, A_{N_a}, V_1, \cdots, V_{N_v}|\mathbf{X}) = \prod_{i=1}^{N_a} P(A_i|\mathbf{X}) \prod_{i=1}^{N_v} P(V_i|\mathbf{X}).$$

However, as the table is still very large, each element is further approximated by

$$P(A_i|\mathbf{X}) \approx \prod_{j=1}^{N_h} P(A_i|Lh_j, Sh_j) \times \prod_{j=1}^{N_n} P(A_i|Ln_j, Sn_j),$$

and

$$P(V_i|\mathbf{X}) \approx \prod_{j=1}^{N_h} P(V_i|Lh_j),$$

respectively. By these approximations, the size of the table is reduced to $N_l(4(N_h+N_n)N_a+2N_hN_v)$. The transition probability $P(\mathbf{X}(t)|\mathbf{X}(t-1))$ is also approximated as

$$P(\mathbf{X}(t)|\mathbf{X}(t-1)) \approx \prod_{i=1}^{N_h} P(Lh_j(t)|Lh_j(t-1)P(Sh_j(t)|Sh_j(t-1)) \times \prod_{i=1}^{N_n} P(Ln_j(t)|Ln_j(t-1)P(Sn_j(t)|Sn_j(t-1)).$$

This means that each object (speaker and noise source) behaves independently. The dynamic Bayesian network in Figure 3 describes the structure of the resultant probability model.

5 Experiments

5.1 General Conditions

The performance of the proposed method was evaluated using data obtained in a medium-sized meeting room with a reverberation time of about 0.5 s. As shown in Figure 2, the microphone array was composed of 8 omni-directional microphones arranged in a circular configuration with a diameter of 0.5 m. The camera was a Digiclops (Pointgray Research Inc.). Other parameters for audio and video signal processing are summarized in Table 1.

Table 1: Parameters of audio and video processing

Audio	
Number of microphones	8
Shape of array	Circular
Diameter of array	50 cm
Sampling frequency	16 kHz
FFT length	512
Window overlap	128
Frequencies of interest	500 ~ 3000 Hz
Interval of ML estimation	0.1 s
Number of assumed sources, M	2
Covering angle	$-90° \sim +90°$
Number of directions, N_a	19
Video	
Camera device	Digiclops
Interval of Human Detection	0.1 s
Visual angle	about $-30° \sim +30°$
Number of directions, N_v	10

5.2 Audio and Video Processing

For sound source localization, the expectation-maximization (EM) based maximum likelihood (ML) method [10] was employed. In this method, the spatial spectrum of sound sources is iteratively estimated by the EM algorithm. The advantage of this method is that a smaller amount of data (e.g., snapshots in 0.1 s) are required for the estimation compared to standard methods such as MUSIC [11]. This is a desirable feature for tracking moving targets. Candidates for sound source locations were determined by searching for peaks in the spatial spectrum. In the following experiment, we assumed that the number of sound sources is 2.

Humans in a video image were detected using a simple background subtraction method based on the range image data obtained by the stereo camera, as in previous work. However, any other method that provides the position of the humans in an observed image could also be used. In the following experiment, we detected two humans per a frame.

5.3 Setting Probability Values

The observation probability $P(\mathbf{Y}(t))|\mathbf{X}(t))$ was set using training data obtained in the same meeting room. The number of true locations of speakers was set at $N_l = 9$, at positions of {under $-$ 30°, $-30°$, $-20°$, $-10°$, 0°, 10°, 20°, 30°, and over30°}. Both of the maximum number of humans N_h and the maximum number of noise sources N_n were set at $N_h = N_n = 1$. In the training data, a single speaker stood still and uttered some words. The location of the speaker was then varied between $-30°$ and $+30°$ in 5° increments. The probabilities for missing data cases were filled manually. The state transition probability $P(\mathbf{X}(t)|\mathbf{X}(t-1))$ was also set manually according to the

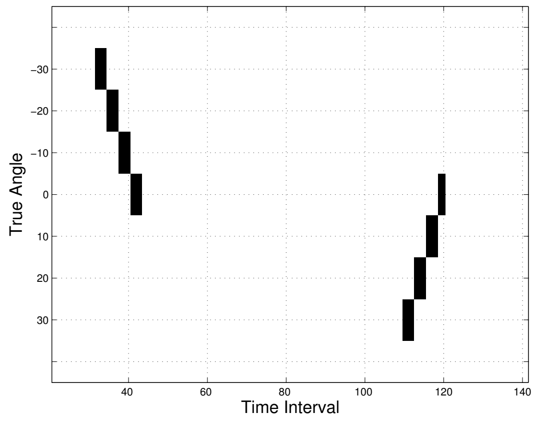

Fig. 4: True speech events for a walking speaker

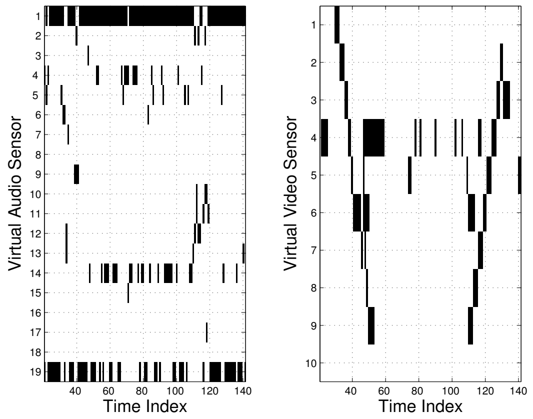

Fig. 5: Audio and video features

random walk and random utterance assumptions. That is, the value of $P(\mathbf{X}(t)|\mathbf{X}(t-1))$ depends only on the distance between $\mathbf{X}(t-1)$ and $\mathbf{X}(t)$, and the value monotonically decreases as the distance becomes large.

5.4 Results

The method was evaluated using test data for a speaker walking around the microphone at a distance of 1.5 m from the camera. The speaker uttered a short sentence twice during walking. The speed of walking was about $30°/\text{s}$. As an interference source, a loudspeaker was located at $-90°$ and played music continuously. The signal-to-noise ratio (SNR) was approximately 0 dB.

True speech events for the walking speaker are shown in Figure 4, where each time unit is 0.1 s. From 203 frames (20.3 s) of data, only the important frames (from 20 to 140) are shown. The black area indicates when the speaker is in that direction and speaking.

Figure5 show the feature vectors extracted from the audio and video signals. In the video features, the walking speaker can be detected as two inclined lines. In the audio features, two sound sources are detected and localized. The audio feature is rather noisy, and the important information is buried in the noise. When computing the ML estimation, the number of sound sources should be set: here, two sound sources are assumed. Hence, for all frames two sound sources are detected even if only one source is generating sound.

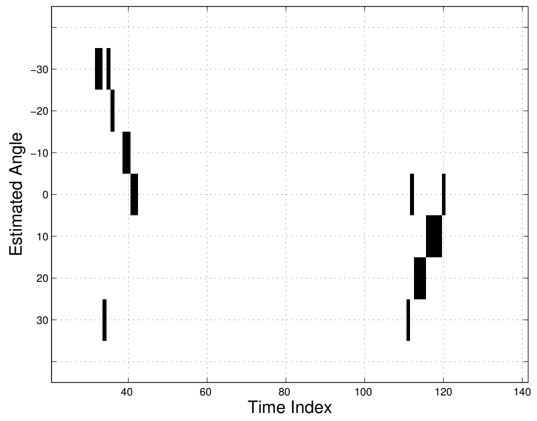

Fig. 6: Result of speech event detection by exact Bayesian inference

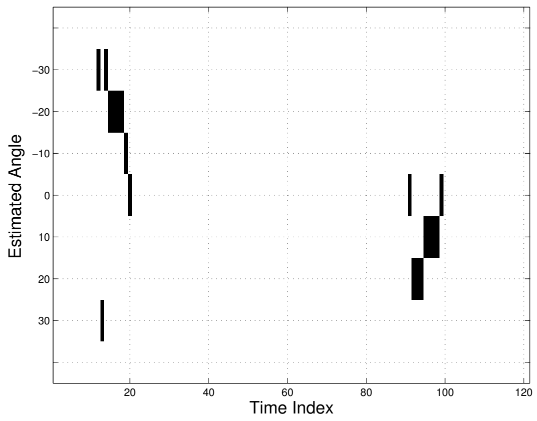

Fig. 7: Result of speech event detection by the particle filter scheme

Figures 6 and 7 show the results of speech event detection by exact Bayesian inference and the particle filter scheme, respectively. The number of particles for particle filtering was 100. From the posterior probability $P(\mathbf{X}(t)|\mathbf{Y}_{1:t})$, the most probable location and status of the speaker is decided and plotted only when the speaker is talking.

By comparing the results with Figure 4, it can be seen that both methods successfully detect certain parts of the speech events by combining the audio and video information.

As a more quantitative evaluation, the error rates for speech event detection were calculated, as summarize in Table 2. Type I error represents a misjudgment of a speech frame as a non-speech frame, whereas type II error is a misjudgment of a non-speech frame as a speech frame. The type I error rate is the number of type I errors divided by the number of speech frames, and the type II error rate is the number of type II errors divided by the number of non-speech frames. The total error rate is the sum of type I and type II errors divided by the total number of frames. From the 203 frames, only the frames when the speaker was within the camera view angle were used for the computation. For comparison, the rates for frame-wise inde-

Table 2: Error rates for moving speech event detection

Error Rate	Type I	Type II	Total
Inference	0.32	0.0	0.18
Bayesian Tracking	0.14	0.0	0.08
Particle Filter	0.18	0.0	0.10

Table 3: Computation time for moving speech event detection

	Computing time
Bayesian Tracking	64852 s
Particle Filter	610.97 s

pendent inference using the Bayesian network employed in previous work are also shown. Here, "Inference" refers to frame-wise independent inference, and "Bayesian Tracking" refers to exact Bayesian inference using a dynamic Bayesian network. The results demonstrate the effectiveness of introducing tracking for moving speakers, and show that the use of the particle filter does not degrade performance significantly. The detection rates obtained in the present study are in fact as good as those obtained for stationary speakers in previous works [12].

The computation times for each of these methods are listed in Table 3. Computations were performed using MATLAB on a personal computer (1 GHz Pentium processor). The particle filter with 100 particles is about 100 times faster than exact Bayesian inference for two sound sources, and this acceleration ratio is expected to increase exponentially with the number of sound sources.

6 Discussion and Conclusions

Particle filters were applied in this study to the problem of detecting and tracking multiple sound sources for moving human speakers by combining audio and video information. The proposed method was shown to work well for real-world data, with a computational cost much lower than that by exact Bayesian inference.

This method is currently being evaluated more extensively, and is being incorporated into a robust multi-modal user interface system for noisy environments [15]. In particular, in the experiments we assumed that $N_h = 1$ and $N_n = 1$. What happens in other conditions is one of the most interesting issues. When there are many moving targets and they intersect each other, data association problem should be solved. Ito et al. proposed a method to solve the association problem for audio and video information fusion[16].

Particle filters has not intensively applied to the problem of tracking sound sources yet. Vermaak et al. applied a particle filter to tracking single moving sound source robustly in reverberant environments[13]. Ward et al. extended the result and demonstrated better performance[14]. Compared to these works, the originality of this work is in the point that we use visual information to make the tracking robust. Also we can discriminate noise sources from human speakers at the same time.

Many problems of information fusion can be formulated as problems of estimating hidden variable sequences by integrating various kinds of observed information. For such problems, the Bayesian approach provides a powerful solution. However, in order to realize Bayesian solutions with a manageable computational load, techniques such as particle filtering will become increasingly important.

Much research is being conduction on improving the efficiency of particle filters, that is, to obtain better approximations with smaller numbers of particles, such as the optimal design of the proposal distribution and optimal selection of feature variables (Rao-Blackwellisation) [17]. Application of such techniques to the present problem will surely be an interesting issue.

In this work, a target localization problem and target tracking problem were solved separately in order to simplify the information fusion process. However, the framework of Bayesian inference allows for more unified ways of solving these problems simultaneously [18]. In the works of tracking a moving sound source[19, 14] mentioned above, the localization problem and tracking problem are solved in an integrated manner. Fong et al. and Vermaak et al. proposed to use a particle filter/smoother to the problem of speech enhancement[20, 19]. Further unification with speech recognition [21] and user intention recognition in conjunction with the application of particle filters to such complex cases is also an interesting topic for further research.

References

[1] A. Doucet, N. Freitas, and N. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, 2001.

[2] G. Kitagawa. Monte carlo filter and smoother for non-gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, 5:1–25, 1996.

[3] M. Isard and A. Blake. Condensation – conditional density propagation for visual tracking. *International Journal of Computer Vision*, 28:5–28, 1998.

[4] D. Fox, S. Thrun, W. Burgard, and F. Dellaert. Particle filters for mobile robot localization. In A Doucet et al., editors, *Sequential Monte Carlo Methods in Practice*, pages 401–428. Springer-Verlag, 2001.

[5] N. Ichimura. Stochastic filtering for motion trajectory in image sequences using a monte carlo filter with estimation of hyper-parameters. In *Proceedings of International Conference on Pattern Recognition (ICPR 2002)*, pages 68–73, 2002.

[6] F. Asano, Y. Motomura, H. Asoh, T. Yoshimura, N. Ichimura, and S. Nakamura. Fusion of audio and video information for detecting speech events. In *Proceedings of the Sixth International Conference on Information Fusion (Fusion 2003)*, pages 386–393, 2003.

[7] F. Asano, Y. Motomura, H. Asoh, T. Yoshimura, N. Ichimura, K. Yamamoto, N. Kitawaki, and

S. Nakamura. Detection and separation of speech segment using audio and video information fusion. In *Proceedings of 8th European Conference on Speech Communication and Technology (Eurospeech 2003)*, pages 2257–2260, 2003.

[8] B. D. O. Anderson and J. B. Moore. *Optimal Filtering*. Prentice-Hall, Inc., 1979.

[9] A. Doucet, N. de Freitas, and N. Gordon. An introduction to sequential monte carlo methods. In A Doucet et al., editors, *Sequential Monte Carlo Methods in Practice*, pages 3–14. Springer-Verlag, 2001.

[10] M. I. Miller and D. R. Fuhrmann. Maximum-likelihood narrow-band direction finding and the em algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38:1560–1577, 1990.

[11] R. O. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas Propagation*, AP-34:276–280, 1986.

[12] T. Yoshimura, F. Asano, Y. Motomura, H. Asoh, N. Ichimura, K. Yamamoto, and S. Nakamura. Detection of speech events in real environments through fusion of audio and video information using bayesian networks. In *Proceedings of 2003 International Workshop on Acoustic Echo and Noise Control (IWAENC 2003)*, pages 319–322, 2003.

[13] J. Vermaark and A. Blake. Nonlinear filtering for speaker tracking in noisy and reverberant environments. In *Proceedings of 2001 IEEE International Conference on Acoustis, Speech, and Signal Processing*, 2001.

[14] D. B. Ward, E. A. Lehmann, and R. C. Williamson. Particle filtering algorithms for tracking an acoustic source in a reverberant environment. *IEEE Transactions on Speech and Audio Processing*, 11:826–836, 2003.

[15] K. Yamamoto, N. Kitawaki, F. Asano, I. Hara, J. Ogata, M. Goto, H. Furukawa, and T. Kamashima. Real-time implementation and evaluation of speech event detection and separation based on the fusion of audio and video information. In *Proceedings of Embedded Signal Processing Conference at the 2004 Global Signal Processing Expo.*, to be published.

[16] W. Ito, N. Ikoma, and H. Maeda. On association problem for sensor fusion in dynamic situation using particle filters. In *Science of Modeling — The 30th Anniversary of the Information Criterion (AIC) — Proceedings: ISM Report on Reserch and Education No.17*, pages 414–415. The Institute of Statistical Mathematics, Japan, 2003.

[17] A. Doucet, N. de Freitas, K. Murphy, and S. Russell. Rao-blackwellised particle filtering for dynamic bayesian networks. In *Proceedings of Sixteenth Conference on Uncertainty in Artificial Intelligence (UAI 2000)*, 2000.

[18] D. L. Stone, C. A. Barlow, and T. L. Corwin. *Bayesian Multiple Target Tracking*. Artech House Publishers, 1999.

[19] J. Vermaak, C. Andrieu, A. Doucet, and S. J. Godsill. Particle methods for bayesian modeling and enhancement of speech signals. *IEEE Transactions on Speech and Audio Processing*, 10:173–185, 2002.

[20] W. Fong, S. J. Godsill, A. Doucet, and M. West. Monte carlo smoothing with application to audio signal enhancement. *IEEE Transactions on Signal Processing*, 50:438–449, 2002.

[21] P. Heracleous, T. Yamada, S. Nakamura, and K. Shikano. Simultaneous recognition of multiple sound sources based on 3-d n-best search using microphone array. In *Proceedings of 6th European Conference on Speech Communication and Technology (Eurospeech 1999)*, pages 69–72, 1999.