2013 IEEE/RSJ International Conference on
Intelligent Robots and Systems (IROS)
November 3-7, 2013. Tokyo, Japan

Nested iGMM Recognition and Multiple Hypothesis Tracking of Moving Sound Sources for Mobile Robot Audition

Yoko Sasaki[1], Naotaka Hatao[1], Kazuyoshi Yoshii[1], Satoshi Kagami[1]

Abstract— **The paper proposes two modules for a mobile robot audition system: 1) recognizing surrounding acoustic event, 2) tracking moving sound sources. We propose nested infinite Gaussian mixture model (iGMM) for recognizing frame based feature vectors. The main advantage is that the number of classes is allowed to increase without bound, if necessary, to represent unknown audio input. The multiple hypothesis tracking module provides time-series of separated audio stream using localized directions and recognition results at each frame. Not only for continuous sounds, the proposed tracker automatically detects appearing and disappearing point of stream from multiple hypothesis. These two modules are connected to microphone array based sound localization and separation, and the combined robot audition system achieved tracking of multiple moving sounds including intermittent sound source.**

I. INTRODUCTION

Environmental sounds are vital to improve autonomous tasks of robots. Daily activities involve many auditory cues: ringing phones, home electronics, barking dogs, passing trucks, and voices. All these sounds notify an individual of environmental changes. It is important for autonomous robots to obtain timely information of the surrounding environment.

Minimizing previous knowledge is an important factor to understand varied audio signals for practical application, and dealing with moving sources is required for a mobile robot. We propose two functions of sound-tracking system for a mobile robot: 1) recognizing surrounding known and unknown acoustic signals, 2) multiple hypothesis tracking of moving sound sources. These are combined with microphone-array-based sound localization and separation.

A common approach for recognizing sound is to generate a fixed model, such as an expanding automatic speech recognition (ASR) model [1] or a Gaussian mixture model (GMM) [2][3]. Recognizing daily sounds using a support vector machine (SVM) has recently been investigated for robot application [4]. A major barrier to recognizing varied audio signals in a real environment is the existence of unknown parameters. It is not always clear how many audio events there are (number of classes) or how many mixtures are appropriate to describe a given audio event (model complexity). Using predefined parameters means that the model is far from realistic.

The recognition module of the proposed system uses a frame-based (i.e., instant of time) input signal and works for frame-based sound localization and separation outputs when the robot or sound sources are moving. The tracked time-series of the separated audio stream can then be recognized using the above conventional recognition models. For example, a conventional ASR system is useful for a tracked voice stream.

A particle-filter-based sound tracking system was proposed and enabled multiple continuous sources tracking [5]. Usually, audio signals are not only continuous events. A signal sometimes stops and new signals appear. Therefore, the number of sound sources changes by frame and detecting the stream appearing or disappearing are important for audio signal tracking.

In Section 2, we explain our moving-sound-tracking system for a mobile robot and give a brief explanation of the microphone-array-based sound localization and separation module. We propose the audio-event recognition method in Section 3 and the multiple sound tracking in Section 4.

II. MICROPHONE ARRAY BASED MOBILE AUDITION

The section gives a basic overview of our moving-sound-tracking system for a mobile robot.

A. System Overview

Fig. 1 shows the system flow and passed variables between modules.

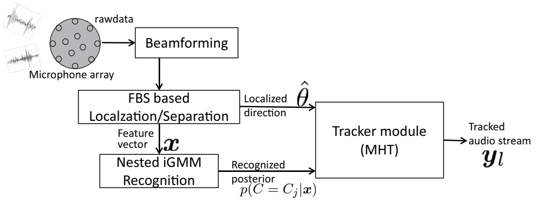


Fig. 1. Overview of the proposed moving-sound-tracking system

The moving sound tracking system we propose in the paper is composed of three parts:

- Microphone array based localization and separation
- Frame based separated sound recognition
- Multiple hypothesis tracking (MHT)

A microphone array is used to localize the sound source direction of arrival and separate the localized sound sources from mixed observation signals. We use beamforming based localization and separation in this paper, but any frame-based method is usable for remaining recognition and tracking modules. Separated signals on each frame are then recognized on our proposed nested infinite Gaussian mixture model (iGMM) module explained in the next section. The

[1] 1: Digital Human Research Center, National Institute of Advanced Industrial Science and Technology, 2-3-26 Aomi, Koto-ku, Tokyo 135-0064, Japan. {y-sasaki, n.hatao, k.yoshii s.kagami}@aist.go.jp

978-1-4673-6358-7/13/$31.00 ©2013 IEEE

module outputs the posterior probability distribution of a sound event class. The MHT module generates audio streams from sound directions and recognition results. Section 4 explains this module in detail.

B. Microphone Array

To extract target audio signals, we use a microphone array embedded in a mobile robot. Using an array of microphones enables accurate and reliable extraction of multiple sound sources. We previously developed a 32-channel omnidirectional microphone array for a mobile robot [6]. It can localize which direction sounds come from and separate localized sound sources.

Delay and Sum BeamForming (DSBF) is a basic array processing method for localization and separation. Aligning the phase of each microphone amplifies the desired direction's signal and attenuates ambient noise, i.e., the microphone array ”focuses” on the specific direction. Scanning the focus to all azimuth directions results in obtaining the sound pressure distribution called the ”spatial spectrum”.

C. FBS Based Multiple Sound Localization and Separation

After beamforming, Frequency Band Selection (FBS) [7] is used for multiple sound localization and separation. FBS is a kind of binary mask to filter out a detected sound's signal to localize other sources simultaneously, and the filtered signal is the separated source at each frame.

The sound localization process using DSBF and FBS is as follows. The first step involves finding the loudest sound source from the spatial spectrum as the maximum total power. The second step involves filtering out the first sound signal using FBS and finding the second strongest sound source from the spectrum. When the frequency component of the DSBF-enhanced signal of the first sound direction is higher than that of any other directions, the FBS module filters out the spectrum at each frequency. For more than three sound sources, the module finds the third strongest sound source, and so on, after filtering out the second strongest sound signal.

When two sound sources are close (usually about 10 deg), false positive detections appear between sources because of wide directivity of DSBF on the spatial spectrum. This problem can be solved in the tracking phase.

III. Nested Infinite GMM for Recognition

This section explains the recognition module for separated sound sources at each frame. We propose a nested iGMM for recognizing varied sound sources in environment. The model is based on Bayesian nonparametrics [8] to avoid the model selection problem. The appropriate number of mixtures varies by audio event. For example, a monotonous sound like the sound of a ventilation fan may be described with just a few Gaussian distributions, but a human voice requires more Gaussian distributions because it contains many phonemes. The model should thus be able to automatically adjust the number of distributions to the complexity of the audio event. In addition, it is important to produce a new class when a previously unknown audio event is detected.

A. Audio Feature Extraction and Model Generation

From the separated sound source at each frame, the feature vector is calculated. We use the following 33 dimensional feature vectors: 12 bands Mel-frequency cepstral coefficients (MFCC), delta MFCC, log energy (E), delta E, zero cross rate (zcr), flux, centroid, variance, entropy, skewness, and kurtosis. Let $\boldsymbol{x}_n$ be the n-th frame feature vector.

As for model generation, we use semi-labeled feature vectors. Let $\boldsymbol{X} = \{\boldsymbol{x}_n\}_{n=1}^{N}$ be a set of feature vectors (observed data) and $\boldsymbol{C} = \{c_n\}_{n=1}^{N}$ be a set of the corresponding class labels ($c_n \in \{1, \cdots, K\}$). N is the total number of training data and K is the number of classes. Because it is semi-supervised training, a part of $\boldsymbol{C}$ is given as observed data in advance and the others are unobserved. The model generation process is summarized in Fig. 2.

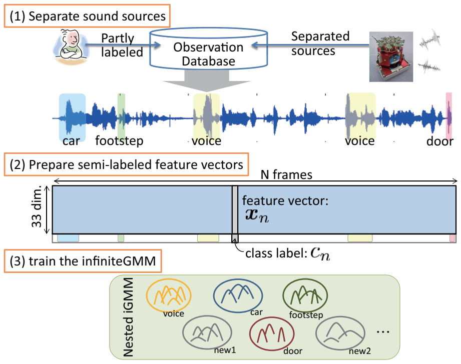


Fig. 2. Model generation process

B. Nested iGMM Formulation

We propose nested Gaussian Mixture Model $\mathcal{M}$, which is mixture of each class model $\mathcal{M}_k$ to generate model using all training data at once. The proposed modeling method has two key component:s 1) It trains K GMMs at the same time by estimating the unobserved class labels, 2) The model has infinite number of classes(K) and dimensions(M).

The proposed model is expressed as follows:

$$\mathcal{M}(\boldsymbol{x}) = \sum_{k=1}^{\infty} \pi_k \sum_{m=1}^{\infty} \tau_{km} \mathcal{N}(\boldsymbol{x} \mid \mu'_{km}, \Lambda_{km}^{-1}). \qquad (1)$$

It is the infinite extension ($K \to \infty, M \to \infty$) of following equations:

$$\mathcal{M}_k(\boldsymbol{x}) = \sum_{m=1}^{M} \tau_{km} \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_{km}, \boldsymbol{\Lambda}_{km}^{-1}), \qquad (2)$$

$$\mathcal{M}(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{M}_k(\boldsymbol{x}), \qquad (3)$$

where τ_{km}, $\boldsymbol{\mu}_{km}$, and $\boldsymbol{\Lambda}_{km}$ are the mixing ratio, mean vector, and precision matrix of the m-th Gaussian in the k-th GMM. Those parameters can be estimated from the observed class labels in $\boldsymbol{C}$ and the corresponding feature vectors in $\boldsymbol{X}$ by using the expectation-maximization (EM) algorithm [9]. π_k is a mixing ratio of the k-th GMM. The parameters $\boldsymbol{\pi}$, $\boldsymbol{\tau}$, $\boldsymbol{\mu}$,

$\boldsymbol{\Lambda}$ are estimated from *incomplete* data that includes missing values (unobserved class labels) by using the EM algorithm as in the case of supervised learning. Note that a feature vector $\boldsymbol{x}$ is classified into one of the K classes, c, such that $c = \mathrm{argmax}_k \pi_k \mathcal{M}_k(\boldsymbol{x})$.

Instead of performing the training and prediction steps independently as described above, we propose to train K GMMs at the same time by estimating the unobserved class labels in a semi-supervised manner. To do this, we formulate a nested GMM that is a weighted mixture of K GMMs. This enables us to take into account how likely each class is to occur.

The model consists of infinitely many GMMs ($K \to \infty$) each of which consists of infinitely many Gaussians ($M \to \infty$). If we have an infinite amount of observed data ($N \to \infty$), an infinite number of Gaussians would be required because the data shows infinite variety. In reality, however, we have only a finite amount of observed data. Therefore, the necessary parameters are a finite part of the infinitely many parameters. In other words, the *effective* complexity of the model is automatically adjusted according to the observed data. This enables us to avoid determining K and M in advance. A technical problem here is how to design prior distributions over infinite-dimensional vectors $\boldsymbol{\pi}$ and $\boldsymbol{\tau}_k$.

First, let M go to infinity. This prior can generate an infinite-dimensional vector of mixing weights $\boldsymbol{\tau}_k$. Most entries of $\boldsymbol{\tau}_k$ take extremely small values because all entries must sum to unity. On the other hand, infinitely many Gaussians are stochastically drawn from the Gaussian-Wishart distribution.

This stochastic process is called the Dirichlet process (DP) [8]. Let $\mathrm{DP}(\beta, G_0)$ be a DP with a concentration parameter β and a base measure G_0. In this study G_0 is a *continuous* distribution (Gaussian-Wishart distribution) over Gaussians ($\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$). A *discrete* distribution G over Gaussians can be drawn as $G \sim \mathrm{DP}(\alpha, G_0)$, where G_0 is an expectation of G and β controls the inverse variance around G_0. More specifically, G is expressed as follows:

$$G = \sum_{m=1}^{\infty} \tau_{km} \delta_{\boldsymbol{\mu}_{km}, \boldsymbol{\Lambda}_{km}}, \tag{4}$$

where δ is the Dirac delta function. Therefore, the parameters of G form an infinite GMM of class k.

One of popular ways to implement the DP is known as the stick-breaking construction [10]. The set of mixing weights $\boldsymbol{\tau}_k$ can be explicitly represented as follows:

$$\tau_{km} = \upsilon_{km} \prod_{m'=1}^{m-1} (1 - \upsilon_{km'}), \quad \upsilon_{km} \sim \mathrm{Beta}(1, \beta). \tag{5}$$

The same idea can be used for K approaching infinity:

$$\pi_k = \lambda_k \prod_{k'=1}^{k-1} (1 - \lambda_{k'}), \quad \lambda_k \sim \mathrm{Beta}(1, \alpha). \tag{6}$$

We then discuss how to determine the concentration parameters α and β. These unknown parameters control the numbers of classes and Gaussians required for representing the observed data. Therefore, non-informative gamma priors are used with a shape parameter d_0 and a rate parameter e_0 on α and β as follows:

$$p(\alpha) = \mathrm{Gamma}(\alpha | d_0, e_0), \quad p(\beta) = \mathrm{Gamma}(\beta | d_0, e_0). \tag{7}$$

The graphical model of the proposed iGMM is summarized in Fig. 3. Some variables are explained in the next subsection.

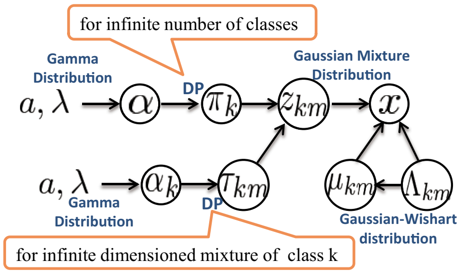


Fig. 3. Graphical representation of infinite Gaussian model

C. Variational Bayesian Inference

This section explains Bayesian treatment of the nested GMM for audio event identification. The Bayesian approach is more robust to over fitting than the maximum-likelihood approach. Since the nested GMM is a kind of mixture models, each observed vector $\boldsymbol{x}_n$ is assumed to be drawn from one of KM Gaussians. Let $\boldsymbol{Z} = \{\boldsymbol{z}_n\}_{n=1}^{N}$ be latent variables that indicate class labels, where $\boldsymbol{z}_n$ is a KM-dimensional vector such that $z_{nkm} = 1$ when $\boldsymbol{x}_n$ is generated from the m-th Gaussian of the k-th GMM and it is otherwise zero ($z_{nk'm'} = 0$ if $k' \neq k, m' \neq m$). If c_n is given as observed data, one of M elements $\{z_{nc_nm}\}_{m=1}^{M}$ must be one.

The goal of Bayesian inference is to calculate a posterior distribution over the latent variables and parameters $p(\boldsymbol{Z}, \boldsymbol{\lambda}, \boldsymbol{\upsilon}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \alpha, \beta | \boldsymbol{X})$ from the observed data $\boldsymbol{X}$. Since the posterior distribution cannot be calculated analytically, we instead approximate it by using an iterative method called the variational Bayes (VB). The computational cost of the VB algorithm is similar to that of the EM algorithm, which is usually used for the maximum-likelihood estimation of the GMM. By using VB, variational posterior distribution is expressed as follows:

$$q(\boldsymbol{Z}, \boldsymbol{\lambda}, \boldsymbol{\upsilon}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \alpha, \beta) = q(\boldsymbol{Z}) q(\boldsymbol{\lambda}, \boldsymbol{\upsilon}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) q(\alpha, \beta). \tag{8}$$

The updating formulas of the VB algorithm are as follows:

$$q(\boldsymbol{Z}) \propto \exp(\mathbb{E}_{q(\boldsymbol{\lambda}, \boldsymbol{\upsilon}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \alpha, \beta)}[\log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\lambda}, \boldsymbol{\upsilon}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \alpha, \beta)]),$$
$$q(\boldsymbol{\lambda}, \boldsymbol{\upsilon}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \propto \exp(\mathbb{E}_{q(\boldsymbol{z}, \alpha, \beta)}[\log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\lambda}, \boldsymbol{\upsilon}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \alpha, \beta)]),$$
$$q(\alpha, \beta) \propto \exp(\mathbb{E}_{q(\boldsymbol{z}, \boldsymbol{\lambda}, \boldsymbol{\upsilon}, \boldsymbol{\mu}, \boldsymbol{\Lambda})}[\log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\lambda}, \boldsymbol{\upsilon}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \alpha, \beta)]).$$

In practice, we set K and M to sufficiently large numbers and gradually remove unnecessary classes and Gaussians whose weights are sufficiently small ($\pi_k \approx 0$ and $\tau_{km} \approx 0$) at each iteration.

IV. Multiple Hypothesis Tracking for moving sounds

This section describes MHT [11] module from the sound localization and recognition results at each frame. MHT provides tracks hypotheses of multiple dynamic objects. An audio signal is intermittent information and is important to detect appearing and disappearing points of sound sources. The MHT module generates multiple hypotheses of a tracked audio stream in a time-space model. The advantage is that it automatically detects the start and end points of a stream from noisy observation signals.

The audition system provides multiple sound directions and feature vectors at each frame. We define $\boldsymbol{o}(n,j) = \{\hat{\theta}(n,j), \boldsymbol{x}(n,j)\}$ for the j-th observation at frame n. $\hat{\theta}$ is the localized direction and $\boldsymbol{x}$ is the feature vector of separated sound source. We explain direction only MHT in Section IV.B, then expand the model for audio feature in Section IV.C.

A. DOA Transition Model

To explain MHT, let us consider simple directional audio tracking from observation at each frame. When the estimated direction and angular velocity of the l-th audio stream $\boldsymbol{y}_l(n) = (\theta_l(n), \dot{\theta}_l(n))^T$ corresponds to the j-th observed direction $\hat{\theta}(n,j)$ at frame n, state and observation equations are modeled as follows:

$$\boldsymbol{y}_l(n) = F\boldsymbol{y}_l(n-1) + \boldsymbol{\zeta}(n), \tag{9}$$

$$\hat{\theta}(n,j) = H\boldsymbol{y}_l(n) + \omega(n), \tag{10}$$

$$F = \begin{pmatrix} 1 & \Delta t \\ 1 & 0 \end{pmatrix} \qquad H = \begin{pmatrix} 1 & 0 \end{pmatrix}, \tag{11}$$

where $\boldsymbol{\zeta}(n)$ is the system noise, which has a mean of 0 and covariance matrix Q, and $\omega(n)$ is the observation noise, which has mean of 0 and covariance matrix R. Δt is the time interval between frame n and $n-1$. By using the Kalman filter, the last stream position $\boldsymbol{y}_l(n)$ at frame n is estimated as follows:

$$\boldsymbol{y}_l(n|n-1) = F\boldsymbol{y}_l(n-1|n-1), \tag{12}$$

$$P_l(n|n-1) = FP_l(n-1|n-1)F^T + Q(n), \tag{13}$$

$$\nu_l(n) = \hat{\theta}(n,j) - H\boldsymbol{y}_l(n|n-1), \tag{14}$$

$$S_l(n) = HP_l(n|n-1)H^T + R, \tag{15}$$

$$W_l(n) = P_l(n|n-1)H^T S_l(n)^{-1}, \tag{16}$$

$$\boldsymbol{y}_l(n|n) = \boldsymbol{y}_l(n|n-1) - W_l(n)\nu_l(n), \tag{17}$$

$$P_l(n|n) = P_l(n|n-1) - W_l(n)S_l(n)W_l^T(n), \tag{18}$$

where $|\cdot|^T$ is the transpose index, ν and S are innovation vector and innovation covariance, W is Kalman gain, P is the covariance matrix of X, and F and H are the update and observation (measurement) matrices, respectively.

In the next subsection, we discus the MHT at frame n and abbreviate $*(n)$ to $*$ in all equations.

B. Multiple Hypothesis Tracking

The hypothesis in MHT is a set of association events at each frame. It is a tree structure in which the parent is the last frame's hypothesis. The h-th hypothesis Ω_h^n is expressed as follows:

$$\Omega_h^n = \{\psi_i, \Omega_{p(h)}^{n-1}\}, \tag{19}$$

where ψ_i is the i-th association event, and $\Omega_{p(h)}^{n-1}$ is the parent hypothesis of Ω_h^n. The likelihood of the stream is calculated on each hypothesis.

The stream has the following three states when we think about patterns of an association event.

- Matched: correctly connected to new observation
- Wrong Connection: present stream when it is not correctly connected
- Terminated: audio signal that disappeared

The observation (localized sound directions) has the following three states :

- Matched: correctly connected to estimated stream
- False Detection: present observation when it does not exist
- New Stream: newly appeared audio signal

From observations $\hat{\Theta}^n = \{\hat{\theta}(0), \hat{\theta}(1), \cdots, \hat{\theta}(n)\}$, the likelihood of each hypothesis $p(\Omega_h^n|\hat{\theta}^n)$ is calculated as follows:

$$\begin{aligned} p(\Omega_h^n|\hat{\Theta}^n) &= p(\psi_i, \Omega_{p(h)}^{n-1}|\hat{\theta}, \hat{\Theta}^{n-1}) \\ &= \eta p(\hat{\theta}|\psi_i, \Omega_{p(h)}^{n-1}) \cdot p(\psi_i, \Omega_{p(h)}^{n-1}|\hat{\Theta}^{n-1}) \\ &= \eta p(\hat{\theta}|\psi_i, \Omega_{p(h)}^{n-1}) \cdot p(\psi_i|\Omega_{p(h)}^{n-1}, \hat{\Theta}^{n-1}) \cdot p(\Omega_{p(h)}^{n-1}|\hat{\Theta}^{n-1}), \end{aligned} \tag{20}$$

where η is a normalization parameter. We assume that $\hat{\theta}$ and $\hat{\Theta}^{n-1}$ are independent. The first term in Eq (20) is expressed as follows.

$$p(\hat{\theta}|\psi_i, \Omega_{p(h)}^{n-1}) = V^{-(N_F+N_N)} \cdot \prod_{j=1}^{J} (\mathcal{L}[\hat{\theta}(j)])^{\tau_j}. \tag{21}$$

where N_F and N_N are the number of False Detection and New Stream, and V is the size of the search space. When False Detection and New Stream observations are equally distributed in the search space, V^{-1} is the probabilistic density function of the False Detection and New Stream. The term $V = 2\pi$ is for azimuth only tracking. $\tau_j = 1$ when the j-th observation is matched; otherwise, $\tau_j = 0$. J is the number of observations at a frame. From Eq (14) and Eq (15), the likelihood function for the Kalman filter is described as $\mathcal{L}[\hat{\theta}(j)] = \mathcal{N}(\nu_{lj}|0, S_{lj})$. $*_{lj}$ means combination of stream l and observation j.

The second term in Eq (20) is expressed as follows:

$$\begin{aligned} &p(\psi_i|\Omega_{p(h)}^{n-1}, \hat{\Theta}^{n-1}) = \\ &(I_M)^{N_M}(I_W)^{N_W}(I_T)^{N_T}(\lambda_F V)^{N_F}(\lambda_N V)^{N_N}, \end{aligned} \tag{22}$$

where I_M, I_W and I_T are the probability of Matched, Wrong Connection and Terminated, and their sum is 1. N_M, N_T and N_N are the number of Matched, Terminated and New Stream in the hypothesis. λ_F and λ_N are the incidence of False Detection and New Stream.

The third term in Eq (20) is the likelihood of the parent hypothesis. It is derived from Eq (21) and Eq (22) assigned to Eq (20).

If a hypothesis that selects the observation $\hat{\theta}(j)$ is assigned to a new stream, the Kalman filter is initialized as $\boldsymbol{y}_l(0|0) = (\hat{\theta}(j), 0), P_l(0|0) = P_0$. The parameter P_0 is initially defined as a 2x2 covariance matrix.

The image of MHT and some variables in this section are summarized in Fig. 4. The bottom stream $\boldsymbol{y}_l$ has four options for next frame, which is called "association event" ψ (i.e., connecting to one of two observations at n-th frame, terminated at the $n-1$ frame or n-th observation is missing). From two streams at $(n-1)$-th frame and two observations at n-th frame, h hypotheses Ω are generated as combinations of association events.

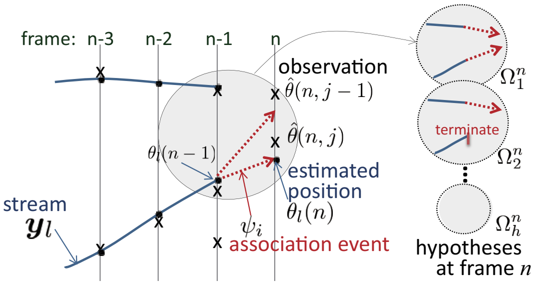


Fig. 4. Image and variables of MHT

C. Model Extension for Audition System

In this section, we introduce two extensions for the proposed MHT module. They can be applied to the tracking model simultaneously.

1) Error detection between two close streams: As shown in Section II-C, our mobile audition system tends to detect false positive observations between sources, when two sound sources are close. The tendency is constant regardless of changes in the environment. Therefore, these error observations can be removed in MHT module including the intermediate noise model.

By using the estimated state of a stream $\boldsymbol{y}_l(n|n-1) = (\theta_l(n|n-1), \dot{\theta}_l(n|n-1))$, the estimated position difference between a pair of streams $\rho = \{l_1, l_2\}$ is expressed as follows:

$$\Delta\theta_\rho = |\theta_{l1}(n|n-1) - \theta_{l2}(n|n-1)|. \tag{23}$$

We assume noise appears when $\Delta\theta_p$ is smaller than threshold. and the noise observations are distributed as a Gaussian distribution : $\hat{\theta}_\rho \sim \mathcal{N}(\hat{\theta}_\rho|\bar{\theta}_\rho, \Delta\theta_\rho/6)$ where $\bar{\theta}_\rho$ is the mean of $\theta(n|n-1)$. Let $I_{F'}$ be the probability of the appearance of intermediate noise, Eq (21) and Eq (22) are updated as follows:

$$p(\hat{\theta}|\psi_i, \Omega_{p(h)}^{n-1}) = V^{-(N_F+N_N)} \cdot \prod_{j=1}^{J} (\mathcal{L}[\hat{\theta}(j)])^{\tau_j} \cdot \prod_{\rho=1}^{J'} \mathcal{N}(\hat{\theta}_\rho|\theta_\rho, \Delta\theta_\rho/6) \tag{24}$$

$$p(\psi_i|\Omega_{p(h)}^{n-1}, \Theta^{n-1}) = (I_M)^{N_M}(I_W)^{N_W}(I_T)^{N_T}(\lambda_F V)^{N_F}(\lambda_N V)^{N_N} \cdot (I_{F'})^{J'}. \tag{25}$$

where J' is the amount of intermediate noises and it is the noise appearance probability.

2) Using Recognition Output: The proposed audition system provides categories of audio events of detected sound sources. We apply this information for tracking. This extension not only improves the performance of the tracker, but also makes sound source estimation possible in the MHT framework.

When a new stream is assigned with K categories of audio K different child hypotheses are generated. k-th hypothesis in them suppose that the new stream is corresponding to the k-th category.

C_j means the most probable category of the j-th sound. Using the assumption that direction of sound $\hat{\theta}_j$ and C_j are independent, Eq (21) is updated as follows:

$$p(\boldsymbol{x}|\psi_i, \Omega_{p(h)}^{n-1}) = V^{-(N_F+N_N)} \cdot \prod_{j=1}^{J} (\mathcal{L}[\hat{\theta}(j)] \cdot P(C_j|C_{lj}))^{\tau_j}, \tag{26}$$

where C_{t_j} is the category of the t_j-th stream. Each $P(C_j|C_{t_j})$ is calculated using the evaluation results of the sound recognition method. Eq (22) is not changed in the extension described in this section.

3) Hypothesis Pruning: The pruning function is necessary for online application because MHT provides a huge number of hypotheses. We use following rules to remove unwanted hypotheses.

- K-best pruning: L streams are registered at most and others are removed.
- Threshold pruning: pruned when the likelihood is smaller than the threshold.
- N-scan back: At frame n, if a pair of hypotheses has the same association events between frames $n-N$ and N and both differences of the initialized frames and positions are within the thresholds, these hypotheses are integrated with weighted averaging. The hypotheses likelihoods are used as weights.

V. EXPERIMENT

The section evaluates the performance of the proposed two modules; nested iGMM recognition and MHT tracking. The experiment is conducted on 32 channel microphone array [6]. It can localize multiple sound sources within 6 deg error even when two sources are within a 20 deg interval. On the other hand, small sounds are missing when two sources are very close.

A. Nested iGMM Recognition Performance

This section describes semi-supervised training. Nine audio recordings (seven percussion instruments, hand clapping, and human voice) were used as sources of audio events. Each was recorded for 13 minutes using a robot embedded with a microphone array [6] with 16 bit and 16 kHz sampling. The first 10 minutes of each source were used for training the nested iGMM, and the remaining 3 minutes of each source were used to evaluate recognition.

We evaluated the recognition performance of our proposed system using the following two conditions for the training data:

expA all nine sounds are included with (partially) given class labels

expB same as **expA**, but one class was completely unlabeled

In **expA**, the model was trained using all nine audio events, each with the correct label, and the rate of correct recognitions for the test data was calculated. In **expB**, the model was trained using all 9 audio events, but one class was unlabeled, i.e., the model learned a previously unknown audio input. For different training experiments, we masked (30, 50, or 70%) of the correct labels.

Fig. 5 shows the results for **expA** and **expB**. The input test data were determined to be correct if they were ascribed to the class for which they had the maximum likelihood from a K-dimensional probability distribution. For **expB**, it was correct when the test data of the unlabeled audio event were identified as belonging to a new class.

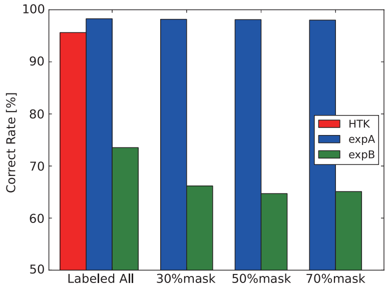


Fig. 5. Correct recognition rates of expA and expB

As in the conventional method, we mixed a fixed number of GMMs, based on the EM algorithm. The Hidden Markov Model Toolkit (HTK) [12] is used for building the GMM. As a result, the proposed method performed better than the HTK, since our proposed method can automatically select an effective number of mixtures for each class. The results of **expA** show a higher rate of correct classifications regardless of the masking level. The rate of correct classifications was 98% for the 70 %-masked model, compared to **expB**, which had a rate of 73.5% correct rate for the 100%-labeled model and 65% correct for the 70%-masked model.

Fig. 6 shows a result of **expB**. It is the posterior distributions of 3 × 9 classes test data for unlabeled Bell model. Bell was identified to new class and others were identified correctly.

B. Evaluation on Tracking

We evaluate 3 patterns of motion (static, cross, balance) as shown in Fig. 7. For each motion pattern, we tested 21 patterns of sound sources: randomly selected 13 pairs of continuous sources and 8 pairs of intermittent sources. In addition to these $3 \times 21 = 63$ patterns recordings of two sound sources, we evaluate 14 patterns of single source data including continuous and intermittent sounds.

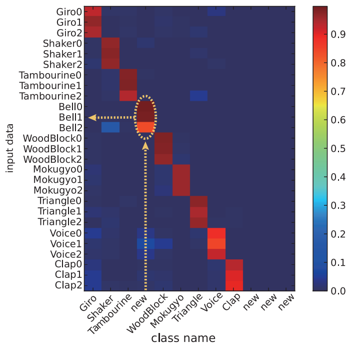
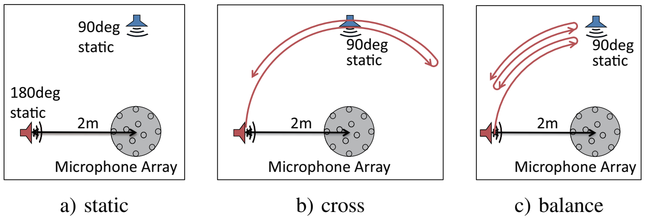


Fig. 6. Posterior for unlabeled bell model

TABLE I

SPECIFICATION OF TEST DATA

	#recordings	#streams	#close	#cross
single	14	42	–	–
static	21	98	0	0
cross	21	98	42	42
balance	21	98	42	42



a) static b) cross c) balance

Fig. 7. Motion patterns of the experiment

The specification of data is summarized in Table I. It indicates the data size of each pattern; from the left side, number of recordings, number of streams, number of closing and crossing point. Each recording is 30 [sec] and the length of a stream in recordings is 4 to 27 [sec].

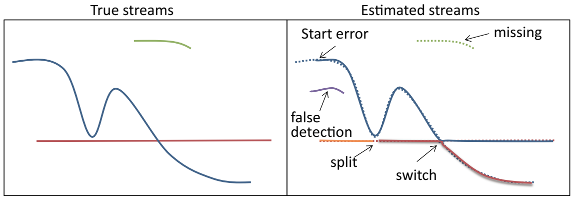
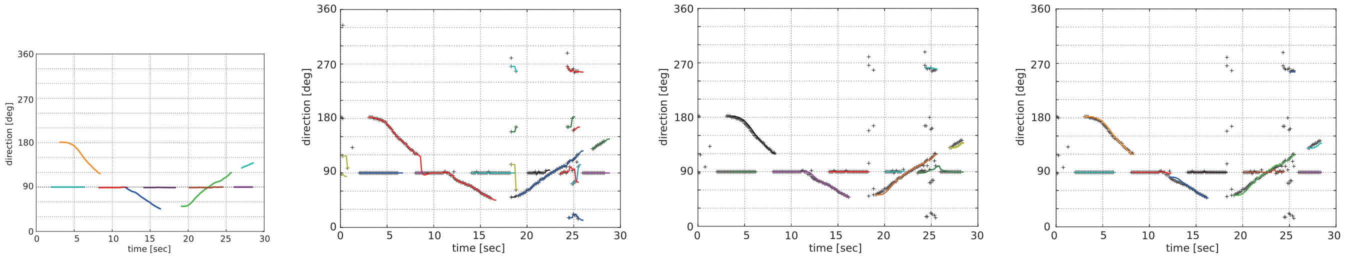


Fig. 8. The illustrative examples of errors

For evaluation of the tracking performance, we count following four errors:

- start/end error: start or end frame is wrong
- split/merge: divided single stream or merged two streams
- f.d./missing: false detection or missing streams
- switch: reversely connected crossing or close streams

Redundant errors are eliminated. For example, a *split* point has an *end* error of the first stream and a *start* error of the



a) ground truth b) nearest neighbor c) direction only MHT d) proposed MHT

Fig. 9. Tracking results for crossing motion of intermittent sources

second stream, but this point is counted as “1 split error”. Fig. 8 shows the illustrative examples of errors above.

Fig. 9 shows a result of crossing motion of two intermittent sources. Cross marks are localized directions and colored lines are tracked streams. A static hand bell sound at 90 [deg], and a moving shaker sound are crossing at 12 and 23 [sec]. Compared to (a)correct streams, (b)nearest neighbor [13] has false positive streams at false positive observations in the localization module. (c)Direction only MHT tracked existing streams, but the second bell stream switched to crossing shaker stream and the fourth bell stream is split at the second crossing point. These errors are corrected on the proposed MHT using recognition result.

TABLE II

TRACKING ERROR EVALUATION

	start/end			split/merge			f.d./missing			switch		
	nn	doa	all	nn	doa	all	nn	doa	all	nn	doa	all
single	9	1	0	0	0	0	51	4	2		–	
static	10	1	0	1	0	0	31	2	1	0	0	0
cross	9	6	3	37	31	11	79	6	3	24	10	7
balance	10	5	5	18	5	4	76	6	2	25	28	8

The evaluation results are summarized in Table II. We compare three algorithms; 1) nearest neighbor(nn) 2) direction only MHT(doa) 3) MHT including recognition result(all). Compared to MHT, nearest neighbor has many false detections and merged error of two intermittent streams. Direction only MHT can track existing streams, but some split or switch errors occur especially at closing or crossing positions. These errors are reduced by using recognition results, and the MHT including recognition shows the best performance.

VI. CONCLUSIONS

The paper proposed two functions for mobile robot audition; 1) frame based sound recognition using nested iGMM, 2) multiple hypothesis tracking (MHT) of moving sound sources. These functions are connected to existing microphone array based sound localization and separation, and the combined system can track multiple sound sources including crossing motion or intermittent signals.

Nested iGMM can adaptively change the needed dimension of the GMM for each class and increase the number of classes to recognize new audio signals. The experimental results show that our model learned unknown classes, and its performance was better than that of the conventional fixed-dimensional model. From localized directions and recognition results, MHT module generates time-series of audio streams. The proposed MHT can detect start and end point of intermittent streams. By using recognition results, MHT can reduce tracking error especially at crossing point.

In this paper, we show multiple sound tracking using frame based observation, but its recognition performance is limited. For example, footstep has characteristic rhythm but its sound is varied depending on the combination of shoe sole and floor material. Future research is needed for understanding time-temporal audio signals.

REFERENCES

[1] Ramasubramanian V., Karthik R., Thiyagarajan S., and Cherla S. Continuous audio analytics by HMM and Viterbi decoding. In *Proceedings of ICASSP*, pp. 2396–2399, May 2011.

[2] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti. Scream and gunshot detection and localization for audio-surveillance systems. In *Proceedings of Advanced Video and Signal Based Surveillance*, pp. 21–26. IEEE, September 2007.

[3] A. Fleury, N. Noury, M. Vacher, H. Glasson, and J.-F. Serignat. Sound and speech detection and classification in a health smart home. In *EMBS*, pp. 4644–4647. IEEE, August 2008.

[4] Joseph M. Romano, Jordan P. Brindza, and Katherine J. Kuchenbecker. Ros open-source audio recognizer: Roar environmental sound detection tools for robot programming. *Autonomous Robots*, Vol. 34, , February 2013.

[5] Jean-Marc Valin, Francois Michaud, and Jean Rouat. Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering. *Robotics and Autonomous Systems Journal*, Vol. 55, No. 3, pp. 216–228, 2007.

[6] Yoko Sasaki, Tomoaki Fujihara, Satoshi Kagami, Hiroshi Mizoguchi, and Kyoichi Oro. 32-channel omni-directional microphone array design and implementation. *Journal of Robotics and Mechatronics*, Vol. 23, No. 3, pp. 378–385, 6 2011.

[7] Yuki Tamai, Yoko Sasaki, Satoshi Kagami, and Hiroshi Mizoguchi. Three ring microphone array for 3d sound localization and separation for mobile robot audition. In *Proceedings of 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2005)*, pp. 903–908, Edmonton, Canada, August 2005.

[8] T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, Vol. 1, pp. 209–230, 1973.

[9] David J.C. MacKay. *Information Theory, Inference, and Learning Algorithms*, chapter 33.7. Cambrigdge University Press, 2003.

[10] Jayaram Sethuraman. A constructive definition of dirichlet priors. *Statistica Sinica*, Vol. 4, pp. 639–650, 1994.

[11] D. B. Reid. An algorithm for tracking multiple targets. *IEEE Transactions on Automatic Control*, Vol. AC-24, No. 6, pp. 843–854, 1979.

[12] S.J. Young. The HTK hidden markov model toolkit: Design and philosophy. *Entropic Cambridge Research Laboratory, Ltd*, Vol. 2, pp. 2–44, 1994.

[13] Yaacov Bar-Shalom, Peter K. Willett, and Xin Tian. *Tracking and Data Fusion*, chapter 3.4. YBS Publishing, 2011.