A NESTED INFINITE GAUSSIAN MIXTURE MODEL FOR IDENTIFYING KNOWN AND UNKNOWN AUDIO EVENTS

Yoko Sasaki, Kazuyoshi Yoshii, Satoshi Kagami

National Institute of Advanced Industrial Science and Technology (AIST)

ABSTRACT

This paper presents a novel statistical method that can classify given audio events into known classes or recognize them as an unknown class. We propose a nested infinite Gaussian mixture model (iGMM) to represent varied audio events in real environment. One of the main problems of conventional classification methods is that we need to specify a fixed number of classes in advance. Therefore, all audio events are forced to be classified into known classes. To solve the problem, the proposed method formulates a infinite Gaussian mixture model (iGMM) in which the number of classes are allowed to increase without bound. Another problem is that the complexity of each audio event is different. Then, the nested iGMM using nonparametric Bayesian approach is applied to adjust the needed dimension of each audio model. Experimental results show the effectiveness for these two problems to represent the given audio events.

1. INTRODUCTION

Environment recognition is a fundamental research topic for developing autonomous robots that are supposed to work in the real world. Although a lot of effort has been devoted to improving *visual* functions of those robots, *auditory* functions have gained relatively less attention of researchers. Auditory information plays an important role especially in an initial stage of environment recognition.

This paper aims to develop an auditory function for identifying what kinds of audio events happen around the mobile robot. Many promising methods have been proposed for special purposes such as automatic speech recognition (ASR) [1], speaker diarization [2], and music transcription [3]. A key feature of our study, on the other hand, is to classify a wide range of environmental sounds in the real world, not limited to speech and music. As a similar purpose, some studies on specific audio event recognition are proposed [4–6]. These can identify human voice, animal sounds, or noise from audio streams. This means that all events are forced to be classified into predefined classes, even if those events have significantly different characteristics.

To overcome this limitation, we propose a nested infinite Gaussian mixture model (iGMM) for identifying known and unknown audio events. Conventionally, a fixed number of GMMs corresponding to known classes are trained by using acoustic features extracted from labeled audio data. The trained model are then used for classifying given audio events into those classes [7]. On the other hand, the proposed method can train an infinite mixture of iGMMs. The number of iGMMs (*i.e.*, the number of classes) is automatically increased according to the complexity of *semi-labeled* audio data if necessary. The method also adjust the number of Gaussians required for representing acoustic features of each class. It has a solid mathematical foundation [8] and provides a principled way to distinguish whether given audio events are considered to be known or unknown without using an ad-hoc thresholding procedure.

2. AUDIO EVENT IDENTIFICATION

This section explains a Bayesian approach to audio event identification in case that the number of event classes, K, is determined in advance. Suppose we have a series of acoustic feature vectors that are partially given class labels. We aim to infer the classes of unlabeled feature vectors in a semi-supervised manner.

2.1. Problem Specification

Audio event is defined as a label that people would use to describe a recognizable event in a region of the sound [9]. We assume that input audio signals (mixed sounds) are spatially separated into individual audio sources by using a microphone array on a mobile robot before audio event identification.

First, we extract following 33 dimensional acoustic feature from the input audio signal: Mel-frequency cepstral coefficients (12 dims.), their delta components (12 dims.), logarithmic energy (1 dim.), its delta component (1 dim.), the zero-crossing rate of the signal, and the flux, centroid, variance, entropy, skewness, and kurtosis of the spectrum (7 dims.) for each frame. Let N be the total number of frames. 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 ($c_n \in \{1, \cdots, K\}$). Here, $\boldsymbol{X}$ is assumed *semi-labeled* data, *i.e.*, a part of $\boldsymbol{C}$ is given as observed data in advance and the other part is unobserved.

Our goal is to infer the unobserved part of $\boldsymbol{C}$ by using $\boldsymbol{X}$ and the observed part of $\boldsymbol{C}$. This is a standard framework of semi-supervised learning. We focus on classifying each frame (feature vector) into one of K classes. Such frame-based identification would be useful for detecting the durations of audio events to be noticed and tracking moving audio events in an audio stream.

2.2. Maximum Likelihood Approach

We here explain a maximum-likelihood approach to audio event identification. A common way to represent a distribution of feature vectors of each class is to use a Gaussian mixture model (GMM). Let M be the complexity of the GMM, *i.e.*, the number of Gaussians. We formulate K GMMs $\{\mathcal{M}_k\}_{k=1}^K$ that correspond to individual classes. In the case of supervised learning, K GMMs are first trained by using completely-labeled feature vectors. Those models are then used for classifying a feature vector $\boldsymbol{x}$ into one of K class, c, such that $c = \text{argmax}_k \mathcal{M}_k(\boldsymbol{x})$, where $\mathcal{M}_k(\boldsymbol{x})$ is the likelihood of $\boldsymbol{x}$ with respect to the class k. More specifically, $\mathcal{M}_k$ is given by

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

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 [10].

Instead of performing the training and prediction steps independently as described above, we propose to train K GMMs at the same time of 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 nested GMM is given by

$$\mathcal{M}(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{M}_k(\boldsymbol{x}), \tag{2}$$

where π_k is a mixing ratio of the k-th GMM. The parameters $\boldsymbol{\pi}$, $\boldsymbol{\tau}$, $\boldsymbol{\mu}$, $\boldsymbol{\Lambda}$ can be 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 K class, c, such that $c = \mathrm{argmax}_k \pi_k \mathcal{M}_k(\boldsymbol{x})$.

2.3. Bayesian Approach

We explain 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 training data, one of M elements $\{z_{nc_nm}\}_{m=1}^{M}$ must be one.

The joint distribution of the nested GMM is defined as follows:

$$\begin{aligned} & p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \\ & = p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\boldsymbol{Z}|\boldsymbol{\pi}, \boldsymbol{\tau}) p(\boldsymbol{\pi}) p(\boldsymbol{\tau}) p(\boldsymbol{\mu}, \boldsymbol{\Lambda}), \end{aligned} \tag{3}$$

where the first two terms are likelihood functions of $\boldsymbol{X}$ and $\boldsymbol{Z}$ and the other three terms are prior distributions of the parameters. The likelihood terms are formulated as follows:

$$p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{nkm} \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_{km}, \boldsymbol{\Lambda}_{km}^{-1})^{z_{nkm}} \tag{4}$$

$$p(\boldsymbol{Z}|\boldsymbol{\pi}, \boldsymbol{\tau}) = \prod_{nkm} (\pi_k \tau_{km})^{z_{nkm}}. \tag{5}$$

We now introduce conjugate prior distributions as follows:

$$p(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\alpha\boldsymbol{\nu}) \propto \prod_k \pi_k^{\alpha\nu_k - 1} \tag{6}$$

$$p(\boldsymbol{\tau}) = \prod_k \mathrm{Dir}(\boldsymbol{\tau}_k|\beta\boldsymbol{\upsilon}) \propto \prod_{km} \tau_{km}^{\beta\upsilon_m - 1} \tag{7}$$

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{km} \mathcal{N}(\boldsymbol{\mu}_{km}|\boldsymbol{m}_0, (b_0\boldsymbol{\Lambda}_{km})^{-1}) \mathcal{W}(\boldsymbol{\Lambda}_{km}|\boldsymbol{W}_0, c_0), \tag{8}$$

where Dir indicates the Dirichlet distribution and $\mathcal{NW}$ indicates the Gaussian-Wishart distribution. As for the hyper-parameters, α and β are called concentration parameters, $\boldsymbol{\nu}$ and $\boldsymbol{\upsilon}$ sum to unity, $\boldsymbol{m}_0$ and b_0 are the mean vector and the precision-matrix scale of a Gaussian distribution, $\boldsymbol{W}_0$ and c_0 are the scale matrix and degree of freedom of the Wishart distribution. In this paper, the hyper-parameters are set to define noninformative prior distributions.

2.4. Variational Bayesian Inference

The goal of Bayesian inference is to calculate a posterior distribution over the latent variables and parameters $p(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{X})$ 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. If we assume that the latent variables are independent from the parameters, the posterior distribution could be factorized as follows:

$$q(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\boldsymbol{Z}) q(\boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda}), \tag{9}$$

where $q(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ is called a variational posterior distribution. Note that $q(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ is essentially different from the *true* posterior distribution $p(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{X})$ because of the independence assumption. Nonetheless, we aim to make $q(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ as close to $p(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{X})$ as possible such that the Kullback-Leibler (KL) divergence is minimized. Each factor of $q(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ can be iteratively optimized as follows:

$$q(\boldsymbol{Z}) \propto \exp(\mathbb{E}_{q(\boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda})}[\log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda})]) \tag{10}$$

$$q(\boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \propto \exp(\mathbb{E}_{q(\boldsymbol{Z})}[\log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda})]) \tag{11}$$

As a result, it turns out that $q(\boldsymbol{Z})$ can be further factorized as follows:

$$q(\boldsymbol{Z}) = \prod_n q(\boldsymbol{z}_n) \tag{12}$$

This means that an KM-dimensional posterior discrete distribution $p(\boldsymbol{z}_n|\boldsymbol{X})$ can be approximated by $q(\boldsymbol{z}_n) = \prod_{km} z_{nkm}^{\gamma_{nkm}}$, where $\{\gamma_{nkm}\}_{k=1}^{K}{}_{m=1}^{M}$ is a set of KM variational parameters. An unlabeled vector $\boldsymbol{x}_n$ is thus classified into one of K classes, c_n, such that $c_n = \mathrm{argmax}_k \sum_m \gamma_{nkm}$.

3. NONPARAMETRIC BAYESIAN APPROACH

This section explains our approach to audio event identification. There are two main problems of the conventional approach described in Section 2. The first problem is that all time frames are forced to be classified into predefined K classes even if they have distinct acoustic features. A naive solution would be to use a thresholding process for adding a new class if the acoustic features are significantly different from those of known classes. However, it is difficult to determine a threshold in a principled manner. The second problem is that the number of Gaussians is fixed as M regardless of the acoustic characteristics of each class. For example, monotonous sounds like ventilation-fan noise can be described with just a few Gaussians, but human voices require more Gaussians because they consist of many phonemes. These problems are caused by lack of flexibility of the nested GMM consisting of KM Gaussians.

To solve these problems, we propose a nonparametric Bayesian model called an *infinite* nested GMM that consists of infinitely many GMMs ($K \rightarrow \infty$) each of which consists of infinitely many Gaussians ($M \rightarrow \infty$). The term "nonparametric" means that the number of parameters of the model is neither fixed nor limited and the model considers infinite complexity. If we have an infinite amount of observed data ($N \rightarrow \infty$), 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 observed data. This enables us to avoid determining K and M in advance.

3.1. Model Formulation

The infinite nested GMM is defined by taking the infinite limit of Eq. (2) as both K and M diverge to infinity as follows:

$$\mathcal{M}(\boldsymbol{x}) = \sum_{k=1}^{\infty} \pi_k \sum_{m=1}^{\infty} \tau_{km} \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_{km}, \boldsymbol{\Lambda}_{km}^{-1}). \tag{13}$$

A technical problem here is how to design prior distributions over infinite-dimensional vectors $\boldsymbol{\pi}$ and $\boldsymbol{\tau}_k$.

First, let M goes to infinity, *i.e.*, consider an infinite-dimensional Dirichlet distribution for each class k in Eq. (7). 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, we can write G as follows:

$$G = \sum_{m=1}^{\infty} \tau_{km} \delta_{\boldsymbol{\mu}_{km}, \boldsymbol{\Lambda}_{km}}, \tag{14}$$

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 [11]. The set of mixing weights $\boldsymbol{\tau}_k$ can be explicitly represented as follows:

$$\tau_{km} = \upsilon_{km} \prod_{m'=1}^{m-1} (1 - \upsilon_{km'}), \quad \upsilon_{km} \sim \mathrm{Beta}(1, \beta). \tag{15}$$

If we let K approach infinity, the same idea can be used as follows:

$$\pi_k = \lambda_k \prod_{k'=1}^{k-1} (1 - \lambda_{k'}), \quad \lambda_k \sim \mathrm{Beta}(1, \alpha). \tag{16}$$

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. We therefore put noninformative gamma priors with a shape parameter d_0 and a rate parameter e_0 on α and β as follows:

$$p(\alpha) = \mathrm{Gamma}(\alpha|d_0, e_0), \quad p(\beta) = \mathrm{Gamma}(\beta|d_0, e_0). \tag{17}$$

3.2. Variational Bayesian Inference

We aim to calculate a posterior distribution over all unknown variables $p(\boldsymbol{Z}, \boldsymbol{\lambda}, \boldsymbol{\upsilon}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \alpha, \beta|\boldsymbol{X})$. Since this is analytically intractable as in Section 2.4, the VB algorithm is used by introducing the following variational posterior distribution:

$$q(\boldsymbol{Z}, \boldsymbol{\lambda}, \boldsymbol{\upsilon}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \alpha, \beta) = q(\boldsymbol{Z}) q(\boldsymbol{\lambda}, \boldsymbol{\upsilon}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) q(\alpha, \beta). \tag{18}$$

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.

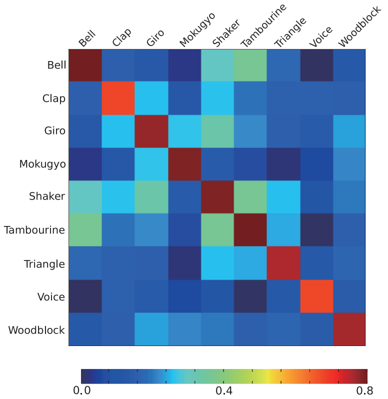


Fig. 1. Cosine similarity between nine kinds of sounds.

4. EVALUATION

This section reports our experiments that were conducted for evaluating the accuracy of audio event identification.

4.1. Experimental Conditions

The audio signals used for evaluation consisted of nine classes (seven percussions, hand claps, and human voice). The cosine similarity of acoustic features was shown in Fig. 1. The hand bell in the top row is similar to the shaker and tambourine. The human voice is the least similar to the other classes. The audio signals were recorded for 13 minutes each using a robot-embedded microphone array [12] with 16 bit, 16 kHz sampling. The first 10 minutes of each signal was used for training the nested iGMM, and class labels were estimated for the remaining 3 minutes.

To evaluate the performance, we conducted the experiments under the following three different conditions:

C1 The time frames contained in the 10 minutes of each audio signal were (partially) given class labels. We masked class labels at 0%, 30%, 50%, or 70% of those frames. The nested iGMM was trained in a semi-supervised manner.

C2 This is the same as **C1** except that one of the nine audio signals was not given class labels at all.

C3 This is the same as **C2** except that one of the nine audio signals itself was not used for training the nested iGMM.

In **C1**, we measured the frame-level accuracy of identification in the case that all nine classes were known in the training phase. For comparison, we trained conventional finite GMMs corresponding to the nine classes by using the Hidden Markov Model toolkit (HTK) [13]. The number of mixture is set to 12. In **C2** and **C3**, the estimated labels were judged as correct if the time frames of an unlabeled audio signal were identified as belonging to a new class. For all conditions, the proposed model has infinite number of GMMs (i.e., $K, M \to \infty$), and only activated GMMs are used for actual computation.

4.2. Experimental Results

Fig. 2 shows the experimental results of **C1** and **C2**. The proposed method performed better than the conventional method using HTK. One of main reasons is that the proposed method can automatically select an effective number of Gaussians in a principled manner. The

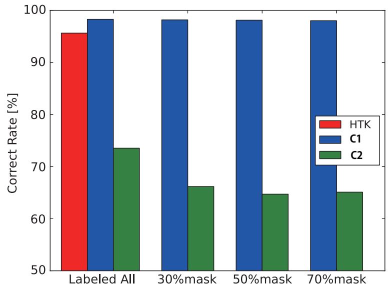
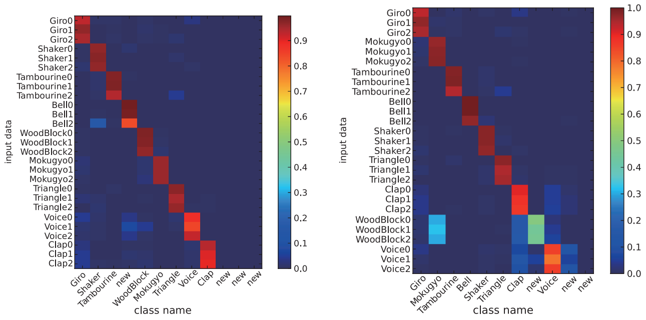


Fig. 2. Identification performances in **C1** and **C2**.



a) Bell labels were masked

b) Woodblock labels were masked

Fig. 3. Posterior distributions in **C2**.

results of **C1** showed a higher accuracy of identification regardless of the masking level. In fact, the accuracy of known-class identification was kept around 98% even if we masked 70% of correct labels. As to **C2**, on the other hand, the accuracy of unknown-class identification was 73.5% in the 100%-labeled case and 65% in the 70%-masked case. We consider these results to be promising.

Fig. 3 shows average posterior distributions over K classes in the 100%-labeled case of **C2**. Fig. 3 a) shows the results in the case that the hand bell labels were not included in the training data. The three rows (Bell 0, 1, and 2) indicate that the bell was identified as a new class and the other rows were classified correctly. Fig. 3 b) shows the results in the case that the woodblock labels were not included in the training data. The three rows (Woodblock 0, 1, and 2) indicate that the woodblock was identified as a new class or sometimes wrongly classified into the mokugyo class because of the similar characteristics. The result indicates that it can generate more than one class when training data includes multiple unique unlabeled data. Note that such new-class identification was achieved without using any ad-hoc thresholding method.

Fig. 4 shows the results of **C3**. The input data is judged as correct when it is classified to never activated class in training process. The accuracies of unknown-class identification vary with class, and were better for characteristic classes, such as human voice. As in **C2**, the bell resulted in a better performance than did items from other classes, and the mokugyo and woodblock resulted in the worst performances.

5. CONCLUSION

This paper proposed a nonparametric Bayesian model for identifying known and unknown classes of audio events in a principled manner. Our model 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 showed that the proposed method can learn unknown classes, and its performance was better than that of the conventional fixed-dimensional model.

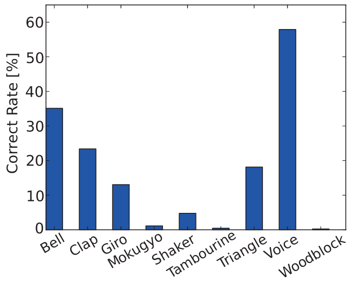


Fig. 4. Identification performances in **C3**.

Such frame-based identification is applicable for segmented or moving signals, but its recognition performance is limited. Future work is needed so that time-temporal information can be used.

6. REFERENCES

[1] Lawrence Rabiner and Biing-Hwang Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[2] G. Friedland, C. Yeo, and H. Hung, "Visual speaker localization aided by acoustic models," in *Proceedings of the ACM International Conference on Multimedia*, 2009, pp. 195–202.

[3] Simon Dixon, "Automatic extraction of tempo and beat from expressive performances," *Journal of New Music Research*, vol. 30, pp. 39–58, 2001.

[4] Andrey Temko and Climent Nadeu, "Acoustic event detection in meeting-room environments," *Pattern Recognition Letters*, vol. 30, no. 14, pp. 1281–1288, 2009.

[5] Taras Butko and Climent Nadeu, "Audio segmentation of broadcast news: A hierarchical system with feature selection for the albayzin-2010 evaluation," in *Proceedings of ICASSP*, 2011, pp. 357–360.

[6] Richard F. Lyon, Martin Rehn, Samy Bengio, Thomas C. Walters, and Gal Chechik, "Sound retrieval and ranking using sparse auditory representations," *Neural Computation*, vol. 22, no. 9, pp. 2390–2416, 2010.

[7] 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*. August 2008, pp. 4644–4647, IEEE.

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

[9] Annamaria Mesaros, Antti Eronen, and Tuomas Virtanen, "Acoustic event detection in real life recordings," in *Proceedings of European Signal Processing Conference*, August 2010.

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

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

[12] Yoko Sasaki, Tomoaki Hujihara, 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.

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