人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B202-7

# 無限混合ガウスモデルを用いた未知クラスに対応可能な実環境音分類法

**Nested Infinite Gaussian Mixture Model for Environmental Audio Signal Recognition**

○ 佐々木洋子 ⋆, 吉井和佳 †, 加賀美聡 ⋆

**Yoko SASAKI, Kazuyoshi YOSHII and Satoshi KAGAMI**

産業技術総合研究所 ⋆ デジタルヒューマン工学研究センター/† 情報技術研究部門

**Digital Human Research Center/Information Technology Research Institute, AIST**

{y-sasaki, k.yoshii, s.kagami}@aist.go.jp

## Abstract

The paper proposes a nonparametric Bayesian audio signal modeling based on nested infinite Gaussian mixture model. It can describe a variety of sound sources for recognizing known and unknown sounds in surrounding environment. So far, various methods of audio signal recognition have been proposed for classifying given audio events into a fixed number of categories that are manually defined in advance. Therefore, audio events of unseen classes are forced to be classified into the known classes although those events have distinct acoustic features. To solve this problem, our model can increase the number of classes unboundedly, to represent the given audio signals. Experimental results showed the effectiveness of the proposed model.

## 1 はじめに

実環境中の様々な音の中で何の音がしたかを理解する技術は，ロボット聴覚分野において主要な課題の一つである．聴覚は視覚に比べ情報量は少ないがより広い範囲に伝播するため，呼びかけに応える，物音に反応して振り返る，など環境変化の初期知覚として有用である．一方で音を認識する技術として，人の声を対象とした音声認識や楽曲中の楽器音推定などが広く発展している．こうした特定種類の音に限らず様々な音が混在する実環境の多様な音響信号を扱うためには，これらの認識技術の前段処理としても，まず何の音なのか理解する技術が役立つ．一例として，一連の音ストリームから人の声，動物の鳴き声，物音などを書き起こす音響イベント検出も盛んに研究が行われている [1] [2] [3].

本研究では，マイクロホンで収録された音響信号が何の音なのか理解するための音源の識別手法について扱う．一般の音を対象とした認識では，既存の音声認識手法を利用したもの [4] や，各種周波数特徴量の GMM [5] [6] など，あらかじめ設計されたモデルの学習が主流である．一方で観測データから直接モデルを設計するアプローチも提案されており [7]，大規模データからベクトル量子化によるデータのランク付け，自動分類が行われている．いずれも正解付のデータに基づく教師あり学習となっている．

様々な条件が想定される実環境に対応するためには，音源の種類数や各音源を表すモデルの次元数など，事前知識は最少であることが望ましい．本稿では，実環境中の多様な音を表現し，未知の音を検出可能なモデルの生成を目指し，ノンパラメトリックベイズ学習 [8] に基づくネスト型無限混合ガウスモデルを提案する．

## 2 ネスト型混合モデルによる環境音分類法

本章では，実環境中の多様な音を認識することをめざし，複雑さの異なる様々な音を一度に学習可能なモデル生成法を提案する．

### 2.1 特徴量の抽出

本研究では音響信号の振幅スペクトルを基にした，フレーム単位の特徴量による音源の識別を対象とする．フレーム単位の特徴量を用いることで，時系列情報を扱うことはできない．一方で各フレームの識別結果は，後段のセグメンテーションや移動音源のトラッキングに利用可能である．

振幅スペクトルの局所的な特徴量として，12 次元 MFCC (Mel-Frequency Cepstrum Coefficient), ΔMFCC, 対数エネルギー E, ΔE, ゼロクロス，フラックス, セントロイド, 分散, エントロピー, 歪度 (skewness), 尖度 (kurtosis) の計 33 次元ベクトルを用いる．

### 2.2 既知の種類の音源への識別

まず音の種類として $K$ 個のクラスが事前に定義されているとして，与えられた特徴量 $\boldsymbol{x}$ がどのクラスに属するかを予測する教師ありクラス分類問題について考える．一般的には学習データ中に含まれる各クラス $k$ $(1 \leq k \leq K)$ の特徴量の分布を，混合ガウスモデル (GMM) $\mathcal{M}_k$ を用いてあらかじめ独立に学習し，与えられた特徴量 $\boldsymbol{x}$ に対する尤度 $\mathcal{M}_k(\boldsymbol{x})$ を計算することで，最も尤度の高いクラスに分類することが行われる．ここで各クラスに対応するモデル $\mathcal{M}_k$ の混合数（ガウス分布の個数）は $M$ であるとしておく．

本研究では各クラスに対応するモデル $\mathcal{M}_k$ をさらに混合したモデル $\mathcal{M}$ を考え，学習データを一挙に与えて**一度に学習**することを提案する．すなわち本モデルは，各クラスの特徴量の分布を $M$ 混合 GMM として表現し，それら $K$ 個をさらに混合したものとして学習する．これは $K$ 混合 GMM の各要素分布が $M$ 混合 GMM となっているものであると言ってもよい．こうすることで各クラス $k$ の出現率（混合比）を加味した分類ができると期待される．

具体的にはあるクラス $k$ $(1 \leq k \leq K)$ の特徴量 $\boldsymbol{x}$ の分布を混合数 $M$ の有限混合ガウス分布

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

で表現する．ここでパラメータ $\boldsymbol{\mu}_{km}$ および $\boldsymbol{\Lambda}_{km}$ は，多次元ガウス分布の平均ベクトル，精度行列であり，$\tau_{km}$ は足して 1 になるように正規化された各ガウス分布の相対強度（混合比）を表す．さらに $K$ 個のクラスにわたる特徴量の分布を表現するため，$\mathcal{M}_k(\boldsymbol{x})$ をさらに混合することでネスト型 GMM

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

を得る．すなわち本モデルは，$KM$ 個のガウス分布からなる混合分布として得られる．特徴量 $\boldsymbol{x}$ は，$KM$ 個中のいずれかのガウス分布から生成されることになる．学習データを用いてパラメータ $\boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda}$ を求めることができれば，新たに与えられた特徴量 $\boldsymbol{x}$ がどのクラスから生成されたものであるかの事後分布が計算できるようになる．

### 2.3 ネスト型混合ガウスモデル

観測データ $\boldsymbol{X}$ の生成過程を表現するネスト型混合ガウスモデルを，ベイズモデルとして定式化する．ベイズモデルでは各パラメータに対して事前分布を導入することで，通常の最尤推定に比べて過学習しにくく，汎化能力の高いモデルを学習が可能である．

いま，学習データに含まれる特徴量は全体で $N$ サンプル（$N$ フレーム）あるとして，観測変数全体を $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$ で表す．同様に $\boldsymbol{X}$ に対する潜在変数を $\boldsymbol{Z} = \{\boldsymbol{z}_1, \cdots, \boldsymbol{z}_N\}$ とする．ここで $\boldsymbol{z}_n$ $(1 \leq n \leq N)$ は $KM$ 次元のベクトルであり，クラス $k$ に対応するモデル $\mathcal{M}_k$ を構成する $m$ 番目のガウス分布から $\boldsymbol{x}_n$ が生成された場合に，$z_{km} = 1$ となり，それ以外の要素はゼロ（$z_{k'm'} = 0$ if $k' \neq k, m' \neq m$）となる．

まず完全な同時分布は次式で与えられる．

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

ここで右辺の前二項はパラメータが与えられたときの観測変数 $X$ および潜在変数 $Z$ の尤度であり，後ろの三項はパラメータの事前分布である．尤度項はそれぞれ，

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

$$p(\boldsymbol{Z} \mid \boldsymbol{\pi}, \boldsymbol{\tau}) = \prod_{nkm} (\boldsymbol{\pi}_k \boldsymbol{\tau}_{km})^{z_{nkm}} \tag{5}$$

で与えられる．また事前分布は共役事前分布を考える．

$$p(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi} \mid \alpha \boldsymbol{\nu}) \propto \prod_{\mathrm{k}} \boldsymbol{\pi}_{\mathrm{k}}^{\alpha \nu_{\mathrm{k}} - 1} \tag{6}$$

$$p(\boldsymbol{\tau}) = \prod_{k} \mathrm{Dir}(\boldsymbol{\tau}_{\mathrm{k}} \mid \beta \boldsymbol{\upsilon}) \propto \prod_{\mathrm{k,m}} \boldsymbol{\tau}_{\mathrm{km}}^{\beta \upsilon_{\mathrm{m}} - 1} \tag{7}$$

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{k,m} \mathcal{N}(\boldsymbol{\mu}_{km} \mid m_0, (b_0 \boldsymbol{\Lambda}_{km})^{-1}) \mathcal{W}(\boldsymbol{\Lambda}_{km} \mid \boldsymbol{W}_0, c_0) \tag{8}$$

ここで $p(\boldsymbol{\pi})$ は $K$ 次元ディリクレ分布，$p(\boldsymbol{\tau})$ は $M$ 次元のディリクレ分布の積である．また $p(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ はガウス・ウィシャート分布の積である．超パラメータに関しては，$\alpha$ および $\beta$ は集中度と呼ばれる正の実数であり，$\nu$ および $\upsilon$ はそれぞれ $K$ 次元ベクトルおよび $M$ 次元ベクトルであり，いずれも足して 1 になるよう正規化されている．$m_0$ および $b_0$ はガウス分布の平均および精度のスケール，$W_0$ および $c_0$ はウィシャート分布のスケール行列および自由度である．本研究では，事前分布ができるだけ無情報になるように超パラメータの値を設定した．

### 2.4 ベイズモデルの学習

ここでの我々の目的は，観測データ $\boldsymbol{X}$ が与えられたもとでの潜在変数およびパラメータの事後分布 $p(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda} | \boldsymbol{X})$ を求めることである．しかし真の事後分布は解析的には求めることはできないため，本研究では変分ベイズ法 (VB) を用いて事後分布を近似的に求めることにする．VB の計算量は，GMM の最尤推定に通常用いられる EM アルゴリズムと同程度であり，非常に効率的である．まず事後分布における潜在変数とパラメータとの独立性を仮定し，因子分解された形の変分事後分布

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

を仮定する．次に $p(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{X})$ の $q(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ に対するカルバック・ライブラー (KL) ダイバージェンスが最小化するように，$q(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{X})$ を反復最適化を行えばよい．このとき各ステップでの最適な変分事後分布は，期待値を $\mathbb{E}$ として

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

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

で与えられる．紙面の都合上，詳細な更新式は省略する．

### 2.5 モデルの生成と識別

モデル生成の流れを Figure 1 に図示する．まず (1) データの一部分に正解ラベルのついた音響信号を用意する．これに対し (2) 各フレームで求めた特徴量ベクトルを学習データ $X$ として，(3) 一部にのみ正解ラベルが付与された一連の学習データに対し，ラベルが与えられていない部分のクラスを推定しながらモデルの学習を行う．

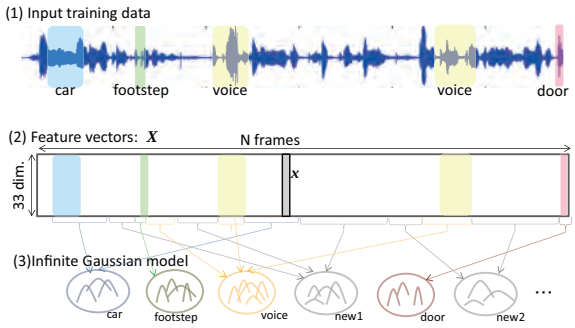


Figure 1: Semi-supervised training of a nested GMM

生成したモデルによる識別では，入力の特徴量ベクトル $\boldsymbol{x}$ がどのガウス分布から生成されたか，$KM$ 次元の離散分布を事後分布として計算する．本研究では，各クラスの次元数 $m$ について足し合わせ，$K$ 次元の離散分布として出力する．

提案手法は一部のラベル付きデータを利用した**半教師あり学習**である．次章では，本章で説明した音源のクラス数 $K$ および各クラスの混合数 $M$ を無限化し，事前にモデルの複雑さを設定する必要のない手法を提案する．観測データのうち正解ラベルのない部分のクラスを無限混合の状態で推定しながら学習することで，観測データに合わせた適切なモデルを生成可能となる．

## 3 未知クラスを含む音源の分類

特徴量の分布を表現するうえで適切な GMM の混合数は，音源のクラスごとに異なる．たとえば，換気扇など単調な音は少数のガウス分布で表現できるであろう．一方，人の話し声は様々な音素から構成されており，多数のガウス分布が必要であると考えられる．そのため各クラスを表現するためのモデルの混合数は，特徴量分布の複雑さに合わせて自動的に調節可能であることが望ましい．また既知の音源の特徴量分布からは非常に発生しにくいような特徴量をもつ音が発生したら，それを既知のクラスのいずれかに当てはめてしまうのではなく，未知の音として新たにクラスを生成する仕組みも重要である．

これらの問題に対処するため，前章で説明したネスト型有限混合ガウスモデルをノンパラメトリックベイズ理論を用いて無限モデルへ拡張することを提案する．「ノンパラメトリック」とは確率モデルの複雑さを表すパラメータ空間の次元が固定されておらず，無限の複雑さを考えることを意味する．もし観測データが無限であれば，その生成過程を表現するため無限個のパラメータが必要となる．ただし実際には観測データは有限であり，無限個のパラメータのうち一部を使うだけで十分である．

無限混合ガウスモデルでは，本来無限個存在するガウス分布のうち，与えられた観測データを表現するのに必要なガウス分布の個数が推定できる．無限個の異なる混合数のモデルが確率的に重なり合っており，混合数を一意に決定せずに学習や予測ができるため，モデル選択の問題が生じない．以降で提案するネスト型無限混合モデルは，ノンパラメトリックベイズ学習に基づいており，音源のクラス数 $K$ や各クラスにおける混合数 $M$ を事前に指定する必要がない．そのため各クラスの特徴量の複雑さに合わせた GMM を学習できるだけはなく，これまで学習していない未知の音が発生した際に，既知のクラスではない新たなクラスであると識別可能である．

### 3.1 混合数，クラス数の無限化

式 (2) を拡張し，音源識別のためのモデルを以下のようにネスト型無限混合ガウスモデルで表現する．

$$\mathcal{M}(\boldsymbol{x}) = \sum_{k=1}^{\infty} \boldsymbol{\pi}_k \sum_{m=1}^{\infty} \boldsymbol{\tau}_{km} \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_{km}, \boldsymbol{\Lambda}_{km}^{-1}) \quad (12)$$

まずクラス数 $K$ を無限大にする場合を考える．つまり，式 (6) で各クラス $k$ ごとに無限次元のディリクレ分布を考える．このような分布からサンプルされる混合比 $\boldsymbol{\pi}_k$ は各基底を選ぶ確率を要素として並べた無限次元のベクトルとなる．$\boldsymbol{\pi}_k$ は無限次元の離散分布であるため，無限個の要素の和が 1 となるよう正規化されている．実際には，ごく一部の要素のみが意味のある値をとり，残りの無限個の要素はほぼゼロに等しい．このような確率過程をディリクレ過程 (Dirichlet Process, DP) と呼ぶ．

集中度を $\alpha$ およびガウス・ウィシャート分布を基底測度 $G_0$ としたディリクレ過程 $\mathrm{DP}(\alpha, \mathrm{G}_0)$ を考える．可算無限個のガウス分布 $G$ は $G \sim \mathrm{DP}(\alpha, \mathrm{G}_0)$ にしたがって生成される．ここで，$G_0$ は $G$ の期待値となっている．$G_0$ からサンプルされた $G$（具体的にはパラメータ $\mu$ および $\Lambda$）の分布は，$\alpha$ が大きいほど $G_0$ に近くなる．したがって，$\alpha$ は逆分散のように振る舞う．ディリクレ過程の一つ

の実現方法として，ここでは 棒折り過程 (Stick-Breaking Construction, SBC) を用いる．SBC は変分ベイズ法を適用するうえで都合が良い DP の表現方法である．このとき，混合係数 $\boldsymbol{\pi}_k$ は次式で表現できる．

$$\boldsymbol{\pi}_k = \boldsymbol{\upsilon}_k \prod_{k'=1}^{k-1} (1 - \boldsymbol{\upsilon}_{k'}) \tag{13}$$

$$\boldsymbol{\upsilon}_k \sim \mathrm{Beta}(1, \alpha) \tag{14}$$

$K$ と同様に各クラスの混合数 $M$ を無限大にする場合を考える．下位の $\boldsymbol{\tau}_{km}$ について $\boldsymbol{\pi}_k$ と同様に変数変換し，次式で表現できる．

$$\boldsymbol{\tau}_{km} = \boldsymbol{\upsilon}_{km} \prod_{m'=1}^{m-1} (1 - \boldsymbol{\upsilon}_{km'}) \tag{15}$$

$$\boldsymbol{\upsilon}_{km} \sim \mathrm{Beta}(1, \alpha_{\mathrm{k}}) \tag{16}$$

ここで式 (14), 式 (16) の集中度 $\alpha, \alpha_k$ について考える．ディリクレ過程 $\mathrm{DP}(\alpha, \mathrm{G}_0)$ における集中度 $\alpha > 0$ は無限混合ガウス分布のハイパーパラメータであり，観測データを生成するのに実際に利用されたガウス分布の個数（混合数）に大きく影響する．適切な値は自明ではないので，$\alpha$ の事前分布 $p(\alpha)$ としてガンマ分布をおく．

$$p(\alpha) \propto \mathrm{Gam}(\alpha \mid \mathrm{a}, \lambda), \quad \mathrm{p}(\alpha_{\mathrm{k}}) \propto \mathrm{Gam}(\alpha_{\mathrm{k}} \mid \mathrm{a}, \lambda) \tag{17}$$

このとき，ハイパーパラメータ $a, \lambda$ に対しては無情報事前分布をおき，特に事前知識がないことを自然に表現する．

### 3.2 変分事後分布

2.4 節と同様に，観測データ $\boldsymbol{X}$ が与えられたときの事後分布 $q(\boldsymbol{Z}, \boldsymbol{\upsilon}, \boldsymbol{\mu}, \boldsymbol{\Lambda} \mid \boldsymbol{X})$ を求めることを考える．これを解析的に行うことは困難なので変分事後分布 $q(\boldsymbol{Z}, \boldsymbol{\upsilon}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ を導入し，真の事後分布に近づくよう最適化を行う．

事後分布において潜在変数とパラメータの独立性を仮定し，以下の因子分解を考える．

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

ここで $\boldsymbol{\alpha}$ は $(\alpha, \alpha_k)$ で表される上位 GMM，下位 GMM の集中度である．右辺の三項についてそれぞれ VB を用いて反復最適化を行えばよい．紙面の都合上詳細な更新式は省略するが，VB 各ステップでの変分事後分布は，

$$q(\boldsymbol{Z}) \propto \exp \mathbb{E}_{\boldsymbol{\upsilon}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\alpha}} [\log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\upsilon}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\alpha})] \tag{19}$$

$$q(\boldsymbol{\upsilon}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \propto \exp \mathbb{E}_{\boldsymbol{z}, \alpha} [\log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\upsilon}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\alpha})] \tag{20}$$

$$q(\boldsymbol{\alpha}) \propto \exp \mathbb{E}_{\boldsymbol{\upsilon}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\alpha}} [\log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\upsilon}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\alpha})] \tag{21}$$

で与えられる．

## 4 環境音の分類実験

提案法による環境音の分類例として，自転車で走行しながら周囲の音を IC レコーダで録音し，環境音のモデル生成を行った．収録した音の説明および分類結果を Figure 2 に示す．上段が録音データの時間波形および付与した正解ラベル，下段が分類結果となっている．また時間波形の上部に各部分の主な音源の説明をつけた．

### 4.1 実験条件

実験用の音源として，自転車走行中にレコーダ (Roland R-09) で録音した，9 分 9 秒のデータを用いた．Figure 2 上部の説明の通り，坂道を下り，静かな通りを走行後，交通量の多い大通りを通り，再び静かな通りを走行した．データは 16bit, 44kHz で録音したものを 16kHz にダウンサンプルした．特徴量はフレーム長 100ms, シフト長 20ms で計算し，計 27490 フレームとなっている．

また Figure 2 上段の時間波形に色つきで示した 3 種類 4 か所のデータに正解ラベルを付与した．ラベル付きの部分は計 4314 フレームあり，残りの 23176 フレーム (白い部分) は未知の音として，モデル生成を行った．

### 4.2 分類結果

Figure 2 下段に，生成したモデルで求めた各クラスの確率分布を示す．ここではフレームごとに各クラスの混合数 $m$ について和をとった $K$ 次元離散分布となっている．

結果は正解ラベルを付与した 3 クラスに加え，新たに 3 クラス，計 6 クラスに分類された．新クラスのひとつめ (new1) には，大通りに出る手前で遠くから聞こえる車の走行音が主に分類された．二番目 (new2) に分類された 5 分 40 秒あたりと 7 分すぎの部分は，車の走行音がなく，アスファルト上を走る自転車のロードノイズが主な音源であった．三番目 (new3) には，すれ違う人の話し声 (女性) や，自転車が段差を越える際の金属音といった比較的高い音が含まれた．正解ラベルのない新しい音について，ほぼ適切に分類できているといえる．

全体では，冒頭と最後の静かな通りでは，主に自転車の走行音 (bicycle) と風切音 (wind) に分類され，中間の大通りでは主に車の走行音 (car) に分類された．一部に正解を付与した 3 クラスについて，正解ラベルのない部分も適切に推定できているといえる．ただし本稿で提案する混合モデルは，複数のクラスを同時にアクティベートできないため，混合音としてクラスを生成するか，分離された音源に適用するなどの工夫が必要である．

## 5 識別性能の評価実験

提案する無限混合モデルの識別性能を検証するため，分類・識別実験を行った．ここでは学習データの条件が異なる数種類のモデルを生成し，既知の音，未知の音に対する識別正解率を求め，結果を考察する．

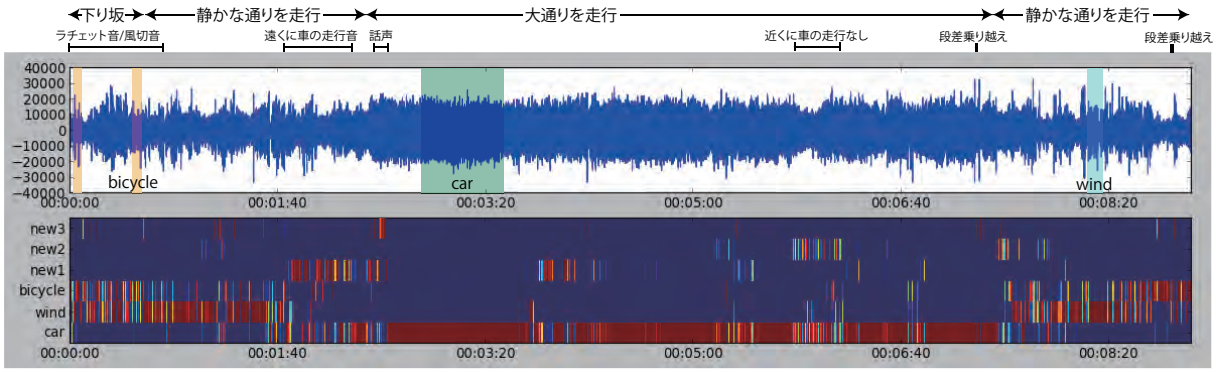


Figure 2: An example of environmental sound modeling

### 5.1 実験条件

評価のための音源として，7 種類のパーカッション楽器 (ハンドベル，ギロ，木魚，シェイカー，タンバリン，トライアングル，ウッドブロック) および，拍手，話し声の 9 種類を用いた．それぞれ 13 分ずつ録音し，10 分を学習データとしてモデル生成に用い，残り 3 分をテストデータとして音源識別の評価に用いた．

各音の収録にはロボットに搭載した 32ch マイクロホンアレイを用いて，16bit, 16kHz サンプリングで行った．特徴量計算に用いる振幅スペクトルは，フレーム長 256ms, シフト長 128ms, Hamming 窓の短時間フーリエ変換により求めた．

Figure 3 に，各音源の類似度を示す．尺度にはコサイン類似度を用いた．値が大きいほど類似していることを示している．たとえば，ひとつめのハンドベルはマラカス，シェイカー，タンバリンに比較的近い特徴量を持ち，最後から 2 番目の話し声は残り 9 種のどの音源にもあまり似ていないことがわかる．

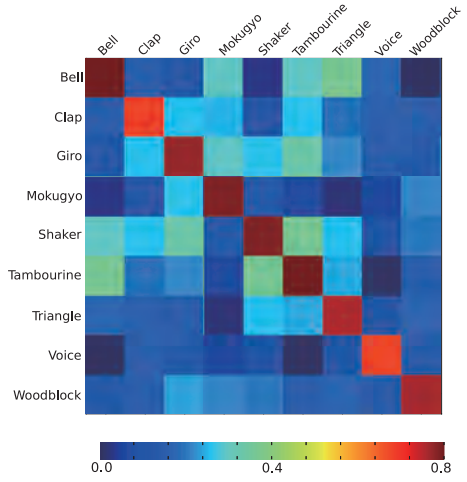


Figure 3: Cosine Similarity of experiment sounds

### 5.2 楽器音の識別

以下の 2 条件についてそれぞれ正解ラベルを含む割合を変化させて半教師あり学習を行った．

**expA** 正解ラベルを持つデータで学習した場合

**expB** 未知音として学習データに含まれる場合

**expA** では，9 種類のすべての音源について正解ラベル付きのデータでモデルを学習し．既知の音源に対する識別正解率を評価する．**expB** では，**expA** と同じ学習データセットに対し，特定の 1 クラスについて全く正解ラベルがない状態で学習し，このクラスのテストデータが新たなクラスとして学習されるかどうかを評価する．

また **expA**, **expB** それぞれについて学習データの音源ごとにそれぞれ一定の割合 (0, 30, 50, 70%) で正解ラベルをマスクし，正解ラベルを含む割合を変化させた各条件での識別率を比較した．

まず **expA** および **expB** で生成したモデルによる，テストデータの識別正解率を Figure 4 に示す．ここでは $K$ 次元の確率分布に対し，最大尤度のクラスに識別されたとして正解率を求めた．また **expB** については，新しいクラスと識別された場合に正解とした．

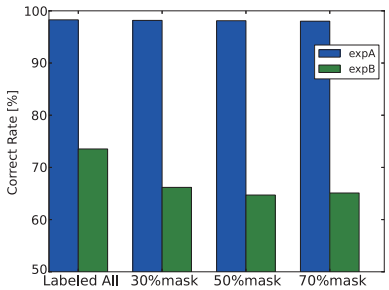


Figure 4: Correct recognition rate of **expA, expB**

**expA** では正解ラベルを持つデータの割合に関わらず高い正解率が得られており，各クラスの正解ラベルを 70% マスクした場合でも正解率 98.0% となった．**expB** については，既知のクラスが完全に正解ラベルを持つ場合に正解率 73.5%，既知クラスのラベルを一部マスクした場合は 65% 前後の正解率となった．

**expB** で既知クラスが全て正解ラベル付きだった場合について，テストデータ各 1 分の事後分布の平均値を Figure

5,6 に図示する．ハンドベルを正解ラベルなしの観測データとして学習した結果である Figure 5 では上 3 行のハンドベル (Bell) が高確率で新クラス (new) となっており，ほぼ確実に新クラスと推定されているといえる．また既知のクラスであるその他のテストデータに対しては，確率の高い部分が横軸のクラス名と一致しており，正しく識別されていることがわかる．

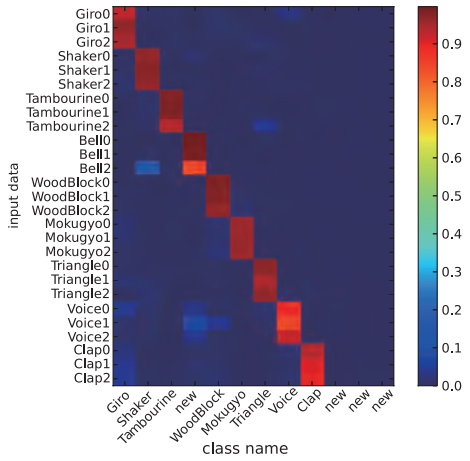


Figure 5: Posterior for unlabeled Bell model

一方，ウッドブロックを正解ラベルなしの観測データとして学習した結果である Figure 6 では，下 3 行のウッドブロック (Woodblock) が新クラス (new) に加え木魚クラス (Mokugyo) にも出現しており，既知クラスとして似たような音がある場合そちらにも分類されていることがわかる．その他の既知クラスについては Figure 5 と同様に正しく識別されている．

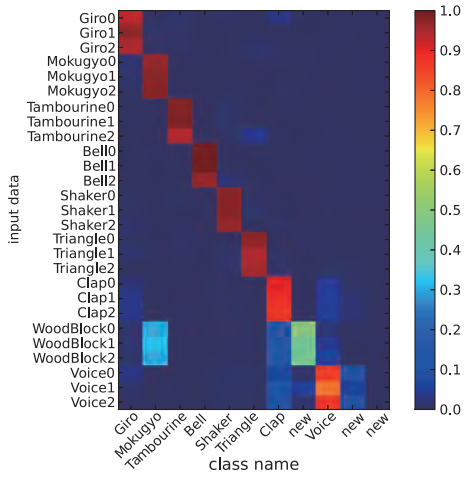


Figure 6: Posterior for unlabeled Woodblock model

## 6 結言

本稿では，知らない音を含む実環境中の様々な音源を認識することをめざし，観測データに合わせて自動生成可能な音響信号のモデル化方法を提案した．提案法は，従来のように対象の音源ごとに独立にモデルを学習させるのではなく，複雑さの異なる様々な音源を含む学習データを一度に学習させることがひとつめの特徴である．さらにノンパラメトリックベイズに基づき無限混合ガウスモデルをベイズ推定することで，本来未知であるはずの音の種類数や各音を表現するモデルの次元数を事前に決定することなく，観測データに合わせて柔軟なモデル生成が可能である．

実験では，既知の音に対する識別では正解ラベルが付与された割合によらず高い正解率が得られた．また正解ラベルを与えていない未知音に対しては，既知の似た音に分類されることもある一方で，既知の特徴量分布とは離れた音響イベントを新クラスと識別可能であることを確認した．

一方提案した無限混合モデルは，複数のクラスを同時にアクティベートできないため，実環境下での混合音の扱いには工夫が必要である．マイクロホンアレイによる音源定位・分離やロボットの移動を含めたロボット聴覚システムの一部として本手法を組み込むことで，時間・空間的に分離された音源の識別方法としての効果が期待される．また識別結果を利用した時系列方向の音のトラッキングや，他センサとの情報統合など，自律型ロボットシステム全体として提案法を活用することが今後の課題である．

## 参考文献


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

[2] 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*, pp. 357–360, 2011.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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. 2396–2416, August 2010.

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