Vol.2014-MUS-103 No.34
2014/5/24

音響イベント列を利用した音響シーン分析のためのモデル学習とオンライン化の検討

井本 桂右[1,a)]　Vincent Gagnon Shaigetz[2,1]　植松 尚[1]　大室 仲[1]

概要：本研究では，音響イベント列として表現された長時間の音響信号から，ユーザ行動や音響信号が収録された周囲の場所，状況など (これらをまとめて音響シーンと称する) などを推定する技術を提案する．従来法による音響シーンの推定手法では，音響シーンをモデル化，推定する際，事前に音響信号を全て用意しておく必要があるという問題点や，大規模な量の音響信号を利用してモデル学習を行うため，モデル化に要する演算時間が非常に大きくなるといった問題点があった．そこで本稿では，逐次的に得られる音響信号から音響シーンの推定を可能とするオンライン型の音響シーン分析手法を提案する．提案手法では，音響シーンの分析性能を劣化させることなくオンライン処理を可能とするため，崩壊型変分ベイズ法に基づいたオンライン型の音響シーンモデル学習手法を提案する．また，実環境で収録した音響信号を用いて提案手法の評価を行い，従来のオフライン型手法と同等の性能を，数十から数百分の一の演算時間で実現できることを確認した．

1. はじめに

近年のメディア情報量の爆発的な増加に伴い，これらを自動で分類，タグ付けするメディア処理技術に注目が集まっている．音響分野においても，音声や楽音に加え，従来では分析の主対象とされてこなかった環境音をも含めた様々な音を分析する Acoustic event detection (AED) の研究が注目されており [1–5]，音響信号を含むメディアコンテンツの自動分類，検索，推薦等への応用が期待されている．また，ユーザ行動や場所，時間，状況等のユーザプロフィールや周囲環境 (以降ではこれらをまとめて音響シーンと称する) の推定に AED を利用する研究も増えつつある [6,7].

音響イベントを利用して音響シーンを推定する手法として，Samuel [8], Imoto [9], Lee [10] は，複数の音響イベント (例えば，足音，ドアの開閉音，音声，楽器の音など音のオブジェクトパターン) の組み合わせにより音響シーンが特徴付けられる事に着目している．これは例えば，「料理」という音響シーンは，「水が流れる音」「包丁の音」「フライパンを熱する音」等の音響イベントの出現頻度や共起関係によって特徴付ける事が可能となることを示し，これらの手法では，音響シーンと音響イベントの関係をベイズモデルの枠組みによりモデル化する．つまり，図 1 に示すように，音響信号を構成する音響イベントの時系列中に，音響シーンを表す潜在的な状態 (これを本稿では音響トピックと称する) が存在し，それらの潜在的な状態が音響イベントの生成分布を決定すると考える．また，これらの手法では，事前に混合ガウスモデル (GMM) や隠れマルコフモデル (HMM) を利用して音響信号と音響イベントを対応づける音響イベントモデルを学習し，短時間 (数十 msec. から

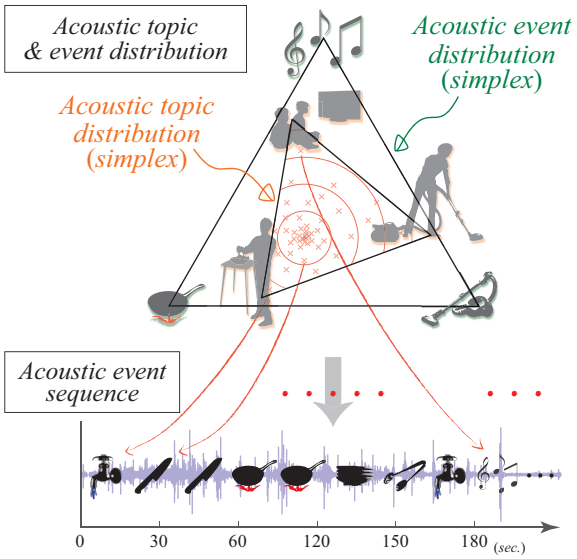

図 1 Relation between acoustic signal, acoustic topic, and acoustic event sequence.

1 NTT メディアインテリジェンス研究所
NTT Media Intelligente Laboratories
2 Ecole Polytechnique of Montreal
a) keisuke.imoto@ieee.org

数 sec. 程度) の長さを持つフレーム毎に唯一つの音響イベントを割り当てる．その後，音響シーンによって音響イベントの時系列の生成確率が規定されると仮定したモデルを導入することで，音響シーンの分析・推定を行う．

しかしながら，これらの手法では，モデル化する音響シーンを含む音響信号を事前に取得しておく必要があることや，モデル化するデータが大規模になると演算時間が非常に大きくなるなどの問題点がある．そこで本稿では，音響イベント列の生成モデルを逐次的に得られた音響信号から構築可能であり，また，モデル学習を精度良く行うことが可能な，崩壊型変分ベイズ法に基づくオンライン型の音響シーンモデル化手法を提案する．

2. Acoustic Topic Model による音響イベント列の生成過程のモデル化

Samuel らが提案したモデルは Acoustic Topic Model (ATM) と呼ばれ，図 1 に示すように，潜在的な状態として表された音響シーンから音響イベント列が生成される確率的な生成モデルを仮定するもので，そのパラメータを推定することによって音響シーンを分析する．ATM は，自然言語処理において単語の組み合わせから文書の話題を推定する際にしばしば利用される，Latent Dirichlet Allocation [11,12] と等価なモデルとして知られており，「音響信号」と「文書」が，「音響シーン」と「文書の話題 (トピック)」が，「音響イベント列」と「単語列」が対応すると考えることができる．つまり，例えば「家の中で収録された音響信号 (新聞という文書)」からある確率に従って「料理をしているという音響シーン (料理というトピック)」が選択され，そこからさらに「包丁で食材を切るという音響イベント (包丁という単語)」が選択されるという過程が繰り返され，最終的に「音響イベント列 (単語列)」が生成されるという一連の生成モデルとして ATM(LDA) は説明することができ，生成モデルのパラメータを与えられた音響イベント列 (単語列) から推定することで音響シーン (文書の話題，トピック) が推定可能となる．

またこのとき，ATM における音響イベント列 $\boldsymbol{e}_s$ の生成過程は以下のように表現できる．なお，各変数の定義は表 1 に示す通りである．

1. Choose $\boldsymbol{\theta}_s$ $\sim$ $Dir(\alpha)$
   Choose $\boldsymbol{\phi}_t$ $\sim$ $Dir(\gamma)$
2. Choose $z_i \mid \boldsymbol{\theta}_s$ $\sim$ $Discrete(\boldsymbol{\theta}_s)$
3. Choose $e_i \mid z_i, \boldsymbol{\phi}_{z_i}$ $\sim$ $Discrete(\boldsymbol{\phi}_{z_i})$,

ATM では，音響シーンが持つ潜在的な構造をモデル化し，抽出可能とする為の潜在変数 $\boldsymbol{z}$ を導入し，これを文書の話題推定に準えて音響トピックと称する．上記の生成過程に従うと，ATM ではまず音響信号 s 毎に，音響トピックの生成分布 $\boldsymbol{\theta}_s$ が，パラメータ α をもつ Dirichlet 分布から選択され，さらに，音響イベント列 $\boldsymbol{e}_s$ の i 番目の音響イベント $\boldsymbol{e}_i$ に対する音響トピック z_i が，多項分布 $\boldsymbol{\theta}_s$ から選択される．さらに，z_i に対する多項分布 $\boldsymbol{\phi}_{z_i}$ に従い，音響イベント列 $\boldsymbol{e}_s$ の i 番目の音響イベント e_i が選択される．この生成過程が $\boldsymbol{e}_s$ に含まれる音響イベントの個数 N_s だけ繰り返されることにより，音響イベント列 $\boldsymbol{e}_s$ が生成される．またこのとき，ATM の結合分布は以下のように表現可能である．

表 1 Definition of variables

Symbol	Definition
S	音響信号の総数
S'	mini batch として同時に処理を行う音響信号の数
T	音響トピックのクラス数
M	音響イベントのクラス数
N_s	音響信号 s に含まれる音響特徴量の数
s, s'	音響信号のインデックス
t	音響トピックのクラスインデックス
m	音響イベントのクラスインデックス
i	音響イベント列に含まれる音響イベントのインデックス
$\boldsymbol{z}$	音響シーンを表す潜在変数 (音響トピック)
$\boldsymbol{e}$	音響イベントを表す変数
$\boldsymbol{\theta}_s$	音響イベント列 s における音響トピック出現分布
$\boldsymbol{\phi}_t$	音響トピック t におけるイベント出現分布のパラメータ (Gauss 分布の平均と分散)
α, β	Dirichlet 分布の超パラメータ
$\mathcal{D}(\cdot)$	Dirichlet 分布
$\Gamma(\cdot)$	Gamma 分布
n_{st}	音響信号 s において，音響トピック t に割り当てられた音響イベントの数
n_{tm}	音響トピック t において，音響イベント m に割り当てられた音響イベントの数

$$
\begin{aligned}
p(\boldsymbol{e}) &= \prod_{s=1}^{S}\prod_{i=1}^{N_s} p(\boldsymbol{e}_i|\boldsymbol{\theta}_s, \boldsymbol{\phi}_t; \alpha, \beta) \qquad (1)\\
&= \prod_{s=1}^{S}\prod_{i=1}^{N_s}\sum_{t=1}^{T} p(z_i|\boldsymbol{\theta}_s)\mathcal{D}(\boldsymbol{\theta}_s;\alpha)p(e_i|\boldsymbol{\phi}_t, z_i)\mathcal{D}(\boldsymbol{\phi}_t;\beta)\\
&= \prod_{i=1}^{N_s}\prod_{s=1}^{S}\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\prod_{t=1}^{T}\theta_{st}^{\alpha-1+n_{st}}.\prod_{t=1}^{T}\frac{\Gamma(M\beta)}{\Gamma(\beta)^M}\prod_{m=1}^{M}\phi_{tm}^{\beta-1+n_{tm}}
\end{aligned}
$$

3. CVB0 による Acoustic Topic Model の変分事後分布の推定

ATM を用いて音響信号から音響シーンを分析するためには，$\boldsymbol{\theta}_s, \boldsymbol{\phi}_t, z_i$ 等のパラメータを推定する必要がある．本章では，collapsed variational Bayes with $0th$-order ap-

proximation (CVB0) [13] を利用した batch 型 ATM の事後分布の推定法をまず示し，次章でさらに，CVB0 を用いた online 型 ATM の事後分布推定手法を導出する．CVB0 によるパラメータの推定法は，collapsed Gibbs sampling (CGS) 法 [14] よりも高速にパラメータ推定が可能であり，また，variational Bayes 法 [15] よりも精度良くパラメータの推定が可能であることが知られている．本稿では，CVB0 と online 型ベイズ推定法を組み合わせることにより，逐次的に得られる音響イベント列を利用して，さらに，従来法よりも高速に，精度を劣化させる事なくパラメータの事後分布が推定可能な手法を提案する．

提案モデルに対する真の事後分布 $p(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi}|\boldsymbol{e})$ を直接推定することは容易でないため，CVB0 による推定法では，変分事後分布 $q(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi})$ を定義し，この変分事後分布を繰り返し最適化することによって真の事後分布に近づける．まず，Jensen の不等式より，全ての未知量に対する周辺対数尤度の下限値 $\mathcal{F}[q]$ を考える．

$$\begin{aligned}\mathcal{L}(\boldsymbol{e}) &\equiv \log p(\boldsymbol{e};\alpha,\beta)\\ &= \iint \sum_{\boldsymbol{z}} \log q(\boldsymbol{z},\boldsymbol{\phi},\boldsymbol{\theta}) \frac{p(\boldsymbol{e},\boldsymbol{z},\boldsymbol{\phi},\boldsymbol{\theta};\alpha,\beta)}{q(\boldsymbol{z},\boldsymbol{\phi},\boldsymbol{\theta})} d\boldsymbol{\phi} d\boldsymbol{\theta}\\ &\geq \iint \sum_{\boldsymbol{z}} q(\boldsymbol{z},\boldsymbol{\phi},\boldsymbol{\theta}) \log \frac{p(\boldsymbol{e},\boldsymbol{z},\boldsymbol{\phi},\boldsymbol{\theta};\alpha,\beta)}{q(\boldsymbol{z},\boldsymbol{\phi},\boldsymbol{\theta})} d\boldsymbol{\phi} d\boldsymbol{\theta}\\ &\equiv \mathcal{F}[q] \end{aligned} \tag{2}$$

ここで，$\mathcal{L}(\boldsymbol{e})$ と $\mathcal{F}[q]$ は以下のような関係式が成り立つ．

$$\begin{aligned}&\mathcal{L}(\boldsymbol{e}) - \mathcal{F}[q]\\ &= \log p(\boldsymbol{e};\alpha,\beta) - \iint \sum_{\boldsymbol{z}} q(\boldsymbol{z},\boldsymbol{\phi},\boldsymbol{\theta}) \log \frac{p(\boldsymbol{e},\boldsymbol{z},\boldsymbol{\phi},\boldsymbol{\theta};\alpha,\beta)}{q(\boldsymbol{z},\boldsymbol{\phi},\boldsymbol{\theta})} d\boldsymbol{\phi} d\boldsymbol{\theta}\\ &= \iint \sum_{\boldsymbol{z}} q(\boldsymbol{z},\boldsymbol{\phi},\boldsymbol{\theta}) \log \frac{q(\boldsymbol{z},\boldsymbol{\phi},\boldsymbol{\theta})}{p(\boldsymbol{z},\boldsymbol{\phi},\boldsymbol{\theta}|\boldsymbol{e})} d\boldsymbol{\phi} d\boldsymbol{\theta}\\ &= KL\big(q(\boldsymbol{z},\boldsymbol{\phi},\boldsymbol{\theta}) || p(\boldsymbol{z},\boldsymbol{\phi},\boldsymbol{\theta}|\boldsymbol{e})\big) \end{aligned} \tag{3}$$

(3) 式から，変分事後分布と真の事後分布の KL ダイバージェンスが最小となるような変分事後分布は，下限値 $\mathcal{F}[q]$ を最大化することで得られることが分かる．つまり，真の事後分布を変分事後分布で近似するためには，$\mathcal{F}[q]$ を最大化するように変分事後分布 $q(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi})$ を決めれば良い．

ここで CVB0 では，変分事後分布を以下のように分解し，$q(\boldsymbol{\theta},\boldsymbol{\phi}|\boldsymbol{z})$ が真の事後分布 $p(\boldsymbol{\phi},\boldsymbol{\theta}|\boldsymbol{e},\boldsymbol{z};\alpha,\beta)$ と一致すると仮定する．

$$\begin{aligned} q(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\phi}) &= q(\boldsymbol{\theta},\boldsymbol{\phi}|\boldsymbol{z}) q(\boldsymbol{z})\\ &= p(\boldsymbol{\phi},\boldsymbol{\theta}|\boldsymbol{e},\boldsymbol{z};\alpha,\beta) q(\boldsymbol{z}) \end{aligned} \tag{4}$$

online CVB0 algorithm for ATM

set α, β, κ, τ_0

initialize $N_{tm}^{(0)}, N_t^{(0)}, \rho^{(0)}$

iterate $k \leftarrow 1$ **to** S

1-1: **initialize** $N_{st}^{(0)}, \hat{\gamma}_{se_it}$

1-2: **iterate until convergence**

$\hat{\gamma}_{se_it}^{(k)} = \big(\alpha + N_{st}^{(k)/si}\big)\big(\beta + N_{tm}^{(k)/si}\big)\big(M\beta + N_t^{(k)/si}\big)^{-1}$

for all s, t **and** m

$\hat{\gamma}_{smt}^{(k)} = \hat{\gamma}_{smt}^{(k)} \Big/ \sum_t \hat{\gamma}_{smt}^{(k)}$

$N_{st}^{(k)} = N_{st}^{(k-1)} - \sum_m n_{sm} \hat{\gamma}_{smt}^{(k-1)} + \sum_m n_{sm} \hat{\gamma}_{smt}^{(k)}$

1-3: $N_{tm}^{(k)} = (1-\rho^{(k)}) N_{tm}^{(k-1)} + \rho^{(k)} \dfrac{S}{S'} \sum_{s'}^{S'} n_{s'm} \hat{\gamma}_{s'mt}^{(k)}$

$N_t^{(k)} = \sum_m N_{tm}^{(k)}$

1-4: **set** $k \leftarrow k+1$, $\rho^{(k)} = (k+\tau_0)^{-\kappa}$

図 **2** Estimation procedure of posterior distributions

(4) 式を (2) 式に代入することにより，$\mathcal{F}[q]$ を以下のように得る．

$$\begin{aligned}\mathcal{F}[q] &= \iint \sum_{\boldsymbol{z}} q(\boldsymbol{z},\boldsymbol{\phi},\boldsymbol{\theta}) \log \frac{p(\boldsymbol{e},\boldsymbol{z};\alpha,\beta) p(\boldsymbol{\phi},\boldsymbol{\theta}|\boldsymbol{e},\boldsymbol{z};\alpha,\beta)}{q(\boldsymbol{z},\boldsymbol{\phi},\boldsymbol{\theta})} d\boldsymbol{\phi} d\boldsymbol{\theta}\\ &= \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \iint q(\boldsymbol{\phi},\boldsymbol{\theta}|\boldsymbol{z}) \log \frac{p(\boldsymbol{e},\boldsymbol{z};\alpha,\beta)}{q(\boldsymbol{z})} d\boldsymbol{\phi} d\boldsymbol{\theta}\\ &= \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{p(\boldsymbol{e},\boldsymbol{z};\alpha,\beta)}{q(\boldsymbol{z})} \end{aligned} \tag{5}$$

(5) 式を $\hat{\gamma}_{se_it} = \hat{q}(z_{se_it} = t)$ について解く [16] と，以下の更新式が得られる．

$$\hat{\gamma}_{se_it} = \frac{\exp\{E_{q(\boldsymbol{z}^{/si})}[\log(\alpha + n_{s\cdot t}^{/si}) + \log(\beta + n_{\cdot e_it}^{/si}) + \log(M\beta + n_{\cdot\cdot t}^{/si})]\}}{\sum_t \exp\{E_{q(\boldsymbol{z}^{/si})}[\log(\alpha + n_{s\cdot t}^{/si}) + \log(\beta + n_{\cdot e_it}^{/si}) + \log(M\beta + n_{\cdot\cdot t}^{/si})]\}} \quad \cdots (6)$$

但し，$n_{s\cdot t}^{/si}$ は s 番目の音響イベント列の i 番目の音響イベントを除いた音響イベントのうち，音響トピック t となる音響イベントの数を表す．(6) 式に含まれる $E_{q(\boldsymbol{z}^{/si})}[\log(\alpha + n_{s\cdot t}^{/si})]$ 等の期待値を厳密に計算するには大きな演算コストを要するため，CVB0 では [13] に示すように，$E_{q(\boldsymbol{z}^{/si})}[\log(\alpha + n_{s\cdot t}^{/si})]$ の $n_{s\cdot t}^{/si}$ に対する 0 次近似のテイラー展開および，$E_{q(\boldsymbol{z}^{/si})}[n_{s\cdot t}^{/si}])$ に対するガウス近似を利用し，以下を得る．

$$\begin{aligned} E_{q(\boldsymbol{z}^{/si})}[\log(\alpha + n_{s\cdot t}^{/si})] &\approx \log(\alpha + E_{q(\boldsymbol{z}^{/si})}[n_{s\cdot t}^{/si}])\\ &= \log(\alpha + N_{st}^{/si}) \end{aligned} \tag{7}$$

表 2 Experimental conditions

Sampling rate / quantization	16 kHz / 16 bits
Frame size / shift	512 / 256
Frame overlap rate	0.5
Acoustic event size	8 - 512
Mini batch size S'	20
Hyperparameter α / β	3.33 / 0.1
Parameter τ_0 / κ	5.0 / 0.7

但し，$N_{st}^{/si} = \sum_{s',i' \neq s,i} \hat{\gamma}_{se_{i't}}$ である．また，他の期待値も同様の近似により計算することが可能である．(7) 式を用いると ATM の更新式 (6) は，最終的に以下のようになる．

$$\hat{\gamma}_{se_it} \propto \left(\alpha + N_{st}^{/si}\right)\left(\beta + N_{tm}^{/si}\right)\left(M\beta + N_{t}^{/si}\right)^{-1} \quad (8)$$

4. Acoustic Topic Model における変分事後分布推定のオンライン化

逐次的に得られた音響イベント列に対して変分事後分布を逐次的に更新するために，オンライン型 ATM(online ATM) では，(5) 式を最大化する変分事後分布を直接推定する代わりに，以下のように $\mathcal{F}[q]$ を音響イベント列毎の和の形で表現し，音響イベント列が得られる度に変分事後分布を逐次更新していく．

$$\begin{aligned}\mathcal{F}[q] &= \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{p(\boldsymbol{e}, \boldsymbol{z}; \alpha, \beta)}{q(\boldsymbol{z})} \\ &= \sum_{s} \sum_{\boldsymbol{z}_s} q(\boldsymbol{z}_s) \log \frac{p(\boldsymbol{e}_s, \boldsymbol{z}_s; \alpha, \beta)}{q(\boldsymbol{z}_s)} \\ &= \sum_{s} l_s(\boldsymbol{e}_s, \boldsymbol{z}_s) \end{aligned} \quad (9)$$

つまり，音響イベント列が逐次的に得られる度に，(8) 式のうち各音響イベント列が寄与する分だけ繰り返し更新し，更新が終わればその音響イベント列の情報は破棄する．具体的には，各時刻に得られた音響信号に対して，N_{tm}, N_t を固定したまま，$\hat{\gamma}_{se_it}$ および N_{st} を繰り返し最適化し，その後，最適化された $\hat{\gamma}_{se_it}, N_{st}$ を用いて N_{te}, N_t を更新する．N_{tm}, N_t は複数の音響イベント列に係る変数であるため，その更新においては，時間変数 k 及び時間シフト係数 τ_0, 減衰係数 κ を用いて，更新重み係数を $\rho^{(k)} = (k+\tau_0)^{-\kappa}$ と設定し，

$$N_{tm}^{(k)} = (1-\rho^{(k)}) N_{tm}^{(k-1)} + \rho^{(k)}\, S\, n_{sm}\, \hat{\gamma}_{smt}^{(k)} \quad (10)$$

$$N_{t}^{(k)} = \sum_{m} N_{tm}^{(k)} \quad (11)$$

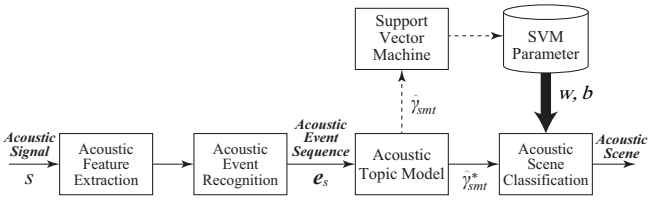

図 3 Acoustic scene estimation system

とする．

さらに本稿では，逐次的に得られた S' 個の音響イベント列を用いて，同時に $\sum_{S'} l_{S'}$ を最大化する mini batch method [17] を導入する．mini batch method を用いることにより，音響イベント列に含まれるノイズの影響を低減する事ができ，より頑健な音響シーンの分析が期待できる．mini batch method を用いた場合の最終的な更新アルゴリズムは図 2 のようになる．

5. 評価実験

5.1 実験条件

音響シーンの一例として，屋内における 9 種類のユーザの行動 (「会話」「料理」「食事」「PC の操作」「読書」「掃除」「移動」「皿洗い」「TV 鑑賞」) によって発生する 11,105 の実環境収録音を用い，音響シーンの推定実験を行った．収録した音響信号のうち，9,802 の信号をモデル学習用に，1,303 の信号を評価用に用いた．なお，本実験では Intel®Core i7-3820QM CPU および DDR3 SDRAM(16GB, 1600MHz) が搭載された PC により評価を行った．また，各種実験条件を表 2 に示す．

音響シーンのモデル化精度を評価するため，図 3 に示すような音響シーン分析の評価システムを構成した．評価システムではまず，全ての音響信号に対して，短時間のフレーム毎に音響特徴量 (12 次の MFCC 特徴量) を算出し，GMM クラスタリングを用いて音響イベントのモデル化と認識を行い，各音響信号を音響イベント列として表す．本稿で構築した評価システムでは，GMM クラスタリングにより推定されたガウス分布の各要素を 1 つの音響イベントとして定義した．次に，GMM により認識された音響イベントの時系列を，モデル学習用，評価用の順に online ATM アルゴリズムに入力し，推定された音響トピックの変分事後分布をそれぞれ $\hat{\gamma}_{smt}$ 及び $\hat{\gamma}^*_{smt}$ とした．同一の音響シーンから生成される音響トピックの分布は類似しているため，モデル学習用と評価用の変分事後分布 $\hat{\gamma}_{smt}, \hat{\gamma}^*_{smt}$ を比較することで音響シーンの推定が可能となる．本稿では，各音響イベント列に対応する行動ラベルと $\hat{\gamma}_{smt}$ を学習データとして，RBF カーネルを用いた多クラス SVM により音響シーン識別器を構成し，$\hat{\gamma}^*_{smt}$ を音響シーン識別器に入力する事で行動の分類精度を評価した．

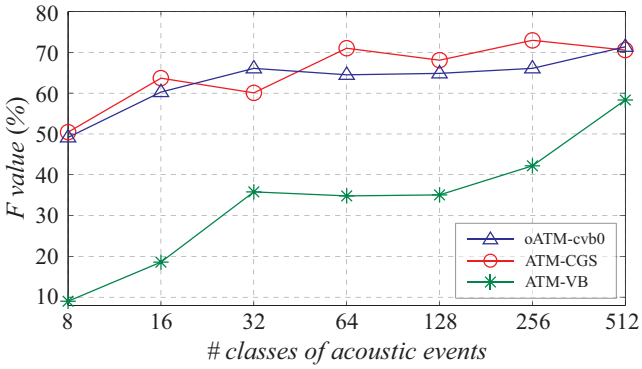

図 4 Acoustic scene estimation accuracy

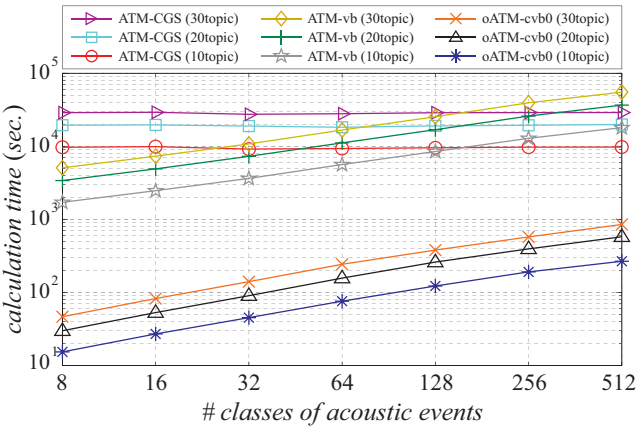

図 5 Calculation time for Acoustic topic modeling

5.2 実験結果

図 4 に音響シーンの平均推定精度を示す．本実験では，従来手法として batch 型の collapsed Gibbs sampling [12] 及び variational Bayes [11] により ATM の学習を行った結果を併せて示す．実験結果より，オフライン学習の中でも特に良い性能を示すとされている，collapsed Gibbs sampling(ATM-CGS) を用いたモデル学習と同等の推定精度が提案手法により実現可能であり，また，ナイーブな variational Bayes 法 (ATM-VB) と比較すると 20〜30 ポイント音響シーンの推定精度が向上していることが分かる．このとき，提案手法では，最も良い条件 (音響トピック数 20, 音響イベントのクラス数 256) で 74.9%の分類精度が得られた．

図 5 にモデル学習の計算時間を示す．オンライン型のモデル学習手法では，逐次的に得られる音響信号に対して音響シーンのモデル化が可能となることが利点として挙げられるが，提案手法ではそれに加え，従来型のオフライン学習法よりも数十から数百分の一の演算時間によりモデル学習が可能となった．例えば，音響トピック数 20, 音響イベントのクラス数 256 における演算時間を比較すると，collapsed Gibbs sampling 法や variational Bayes 法を用いた際は $3.0 \times 10^4 (sec.)$ 程度の演算時間を要する一方，提案手法では，$4.0 \times 10^2 (sec.)$ 程度の演算時間でモデル学習が可能となっている．また，従来手法では学習データの量が増加するにつれて，必要となるメモリ容量や記憶装置の容量が増加する，繰り返し最適の収束条件を満たすのに要する時間が長くなると言ったデメリットがあるが，提案手法では必要となるメモリ容量や記憶装置の容量が一定である，単位時間あたりの演算量が学習データの量に依存しないという利点もある．

6. まとめ

本稿では，CVB0 に基づくオンライン型の音響シーンモデル化手法及び音響シーン分類手法を提案した．提案モデルにより，逐次的に得られた音響信号から音響トピックおよび音響イベントの生成分布を高い精度で推定する事が可能となるとともに，モデルの学習に要する演算時間を大幅に低減することが可能となる．

ユーザ行動に伴って発生する実環境音を用いた音響シーンの分析実験を行った結果，音響シーンのモデル学習が従来手法と比較して数十分の一から数百分の一の演算時間で実現できるとともに，音響シーンのモデル化精度を従来のオフライン学習法と同等の精度に保つ事が可能である事が明らかになった．

参考文献

[1] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, ”CLEAR Evaluation of Acoustic Event Detection and Classification Systems,” Springer Berlin Hidelberg, pp. 311-322, 2007.
[2] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, ”Acoustic Event Detection in Real Life Recordings,” in Proc. of the 18th European Signal Processing Conference (EUSIPCO 2010), pp. 1267-1271, 2010.
[3] Y. Peng, C. Lin, M. Sun, K. Tsai, ”Healthcare Audio Event Classification Using Hidden Markov Models and Hierarchical Hidden Markov Models,” in Proc. of the IEEE International Conference on Multimedia and Expo 2009 (ICME 2009), pp. 1218-1221, 2009.
[4] A. Harma, M. F. McKinney, and J. Skowronek, ”Automatic Surveillance of the Acoustic Activity in our Living Environment,” in Proc. of the IEEE International Conference on Multimedia and Expo 2005 (ICME 2005).
[5] Y. Ohishi, D. Mochihashi, T. Matsui, M. Nakano, H. Kameoka, T. Izumitani, and K. Kashino, ”Bayesian Semi-supervised Audio Event Transcription based on Markov Indian buffet Process,” in Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP2013).
[6] A. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, ”Audio-Based Context Recognition,” IEEE Trans. Audio, Speech, Language Process., pp.321-329, 2006.
[7] T. Heittola, A. Mesaros, A. Eronen, and A. Klapuri, ”Audio Content Recognition Using Audio Event Histograms”, in Proc. of the 18th European Signal Process-

ing Conference (EUSIPCO 2010), pp. 1272-1276, 2010.
[8] S. Kim, S. Narayanan, and S. Sundaram, "Acoustic topic models for audio information retrieval," in Proc. of the 2009 *IEEE* Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2009), pp. 37-40, 2009.
[9] K. Imoto, S. Shimauchi, H. Uematsu, and H. Ohmuro, "User Activity Estimation Method Based on Probabilistic Generative Model of Acoustic Event Sequence with User Activity and Its Subordinate Categories," in Proc. of INTERSPEECH 2013.
[10] K. Lee and D. P. W. Ellis, "Audio-Based Semantic Concept Classification for Consumer Video," IEEE Trans. Audio, Speech, Language Process., pp.1406-1416, 2010.
[11] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," J. Machine Learn. Res., 993-1022, 2003.
[12] T. L. Griffiths and M. Steyvers, "Finding Scientific Topics," PNAS, 5228-5235, 2004.
[13] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh, "On smoothing and inference for topic models," in Proc. of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence. AUAI Press, pp. 27-34. 2009.
[14] R. M. Neal, "Probabilistic inference using Markov chain Monte Carlo methods," Dept. of Comput. Sci., Univ. of Toronto, Tech. Rep. CRG-TR-93-1, 1993.
[15] H. Attias, "A Variational Bayesian Framework for Graphical Models," in Adv. Neural Inf. Proc. Syst., pp. 209-215, 2000.
[16] Y. W. Teh, D. Newman, and M. Welling, "A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation," in Adv. Neural Inf. Proc. Syst., pp. 1378-1385, 2006.
[17] M. D. Hoffman, D. M. Blei and F. Bach, "Online Learning for Latent Dirichlet Allocation," in Adv. Neural Inf. Proc. Syst., pp. 856-864, 2010.