Online Sound Structure Analysis Based on Generative Model of Acoustic Feature Sequences

Keisuke Imoto*, Nobutaka Ono†‡, Masahiro Niitsuma*, Yoichi Yamashita*
* Ritsumeikan University, Japan, † National Institute of Informatics, Japan
‡ SOKENDAI (The graduate university for advanced studies)

***Abstract*—We propose a method for the online sound structure analysis based on a Bayesian generative model of acoustic feature sequences, with which the hierarchical generative process of the sound clip, acoustic topic, acoustic word, and acoustic feature is assumed. In this model, it is assumed that sound clips are organized based on the combination of latent acoustic topics, and each acoustic topic is represented by a Gaussian mixture model (GMM) over an acoustic feature space, where the components of the GMM correspond to acoustic words. Since the conventional batch algorithm for learning this model requires a huge amount of calculation, it is difficult to analyze the massive amount of sound data. Moreover, the batch algorithm does not allow us to analyze the sequentially obtained data. Our variational Bayes-based online algorithm for this generative model can analyze the structure of sounds sound clip by sound clip. The experimental results show that the proposed online algorithm can reduce the calculation cost by about 90% and estimate the posterior distributions as efficiently as the conventional batch algorithm.**

I. Introduction

The amount of media information, such as sound, video, and text data, has increased recently, and it has become more important to analyze them and explain their structure automatically. Acoustic scene analysis is such a technique for analyzing the sound structure and extracting valuable information (e.g., *What is someone doing and where and when? Who is this someone?*) from different types of sounds (e.g., environmental sounds, voice, music), in which much interest has been expressed recently [1], [2], [3], [4]. Particularly, some methods are focused on the fact that many sounds are characterized by a combination of multiple types of sounds. For example, the sound of ”*cooking*” is marked by a combination of sounds including “*cutting with a knife*,” “*heating a skillet*,” and “*running water*.” On the basis of this idea, Kim *et al.* [5], Lee *et al.* [6], and Imoto *et al.* [7], [8] proposed generative probabilistic models of acoustic word sequences (that consist of multiple acoustic words) for analyzing the sound structure, which are called acoustic topic models (ATMs). Note that an acoustic word is defined as a label of the sound type given time-frame-by-time-frame. Conventional ATMs preliminarily estimate acoustic word sequences from long-term sound clips time-frame-by-time-frame using Gaussian mixture models (GMMs) or hidden Markov models (HMMs). They then model a probabilistic generative process of an acoustic word sequence over sound clips and analyze the sound structure with the models. However, these conventional ATMs model the generative process of acoustic words and acoustic features separately; therefore, they do not capture the acoustic similarity between acoustic words or the variance of each acoustic word in the generative model. To address this problem, Imoto *et al.* [7] proposed a generative model of acoustic word sequences that can precisely capture the acoustic similarity between acoustic words and the variance of each acoustic word, as shown in Figs. 1 and 2. They called their model latent acoustic topic and event allocation (LATEA).

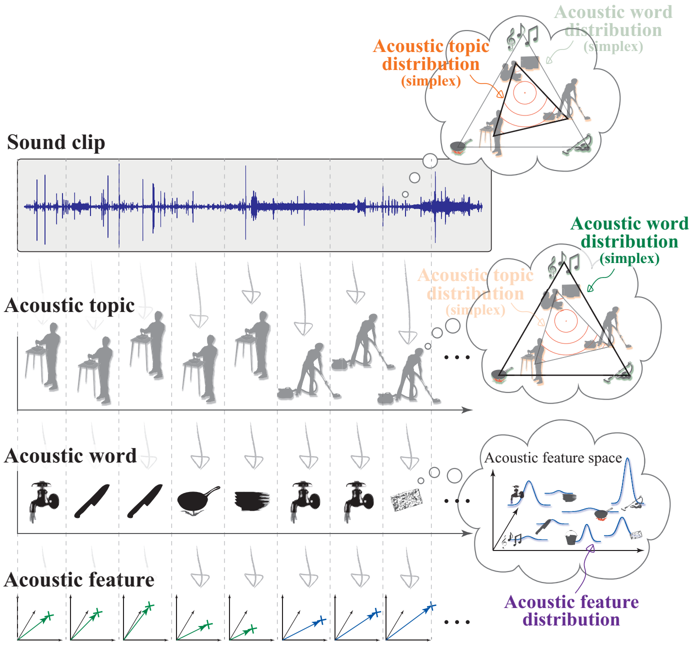
Acoustic word distribution (simplex)
Acoustic topic distribution (simplex)
Sound clip
Acoustic word distribution (simplex)
Acoustic topic distribution (simplex)
Acoustic topic
Acoustic word
Acoustic feature space
Acoustic feature
Acoustic feature distribution

Fig. 1. Relation between sound clip, acoustic topic, acoustic word, and acoustic feature sequence

These models require estimating the posterior distributions after the entire sound corpus has been observed; therefore, it is difficult to apply them to sequentially obtained acoustic signals. Moreover, they incur high calculation cost to estimate the optimal posterior distributions because it is necessary to run the estimation algorithm iteratively over an entire sound corpus. On the contrary, online learning algorithms in ATMs were proposed by Kim *et al.* [9] and Imoto *et al.* [10]. However, it can model only a generative process of the acoustic word sequence except for the acoustic similarity between acoustic words or the variance of each acoustic word in the model.

For this study, we propose an online sound structure analysis

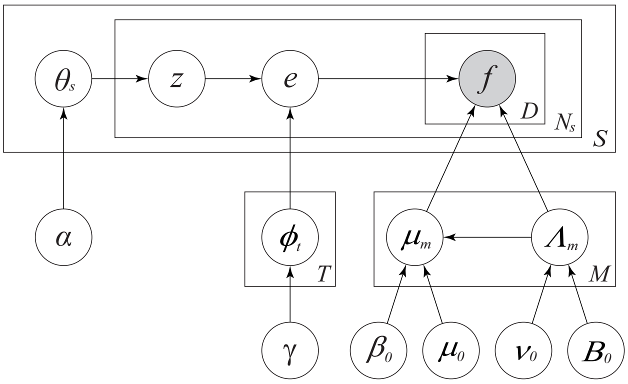

Fig. 2. Graphical model representation of LATEA

algorithm based on LATEA that can be applied to sequentially obtained acoustic signals and precisely capture the relations among sound clips, acoustic topics, acoustic words, and acoustic features.

The rest of this paper is structured as follows. In Section 2, we introduce LATEA and a batch parameter estimation algorithm for the model. In Section 3, the parameter estimation algorithm for LATEA based on an online variational Bayes (VB) is described. In Section 4 and 5, we discuss experimental results and conclude this paper.

II. Generative model of acoustic feature sequences

LATEA [7] is a generative probabilistic model of acoustic feature sequences for modeling the latent structures of acoustic topics and words simultaneously. An acoustic topic is defined as a latent structure time-frame-by-time-frame, which represents the sound structure in an unsupervised manner. As shown in Figs. 1 and 2, LATEA models a generative process of acoustic feature sequences hierarchically; it represents sound clips as categorical distributions over acoustic topics, acoustic topics as categorical distributions over acoustic words, and acoustic words as Gauss-Wishart distributions over acoustic features. This means that we regard each acoustic word as a Gaussian component of a GMM and represent acoustic topics as the mixture weights of the GMM in the acoustic feature space. Following the stochastic method, we can describe these relations with a fully Bayesian approach, as shown in Fig. 2. The definitions of the variables in this paper are listed in Table I.

TABLE I
Definitions of variables in this paper

Symbol	Definition
$\boldsymbol{z}$	Latent acoustic topic variables
$\boldsymbol{w}$	Latent acoustic word variables
$\boldsymbol{f}$	Set of all acoustic feature sequences
$\boldsymbol{\theta}_s$	Acoustic topic distributions over sound clip s
$\boldsymbol{\phi}_t$	Acoustic word distributions over topic t
$\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m$	Acoustic feature distribution over word m (Mean and variance of Gaussian distribution)
α, γ	Hyperparameter of Dirichlet distribution
$\beta_0, \boldsymbol{\mu}_0$	Hyperparameter of Gaussian distribution
$\nu_0, \boldsymbol{B}_0$	Hyperparameter of Wishart distribution
C_m	Regularization term of Wishart distribution
n_{st}	acoustic word counts assigned to t in s
n_{tm}	acoustic word counts assigned to m in t
S, s	Total number and index of sound clips
T, t	Total number of classes and index of acoustic topic
M, m	Total number of classes and index of acoustic word
D, d	Dimension and dimension index of acoustic feature
N_s, n	Total number and index of acoustic features in s
$\mathcal{D}(\cdot)$	Dirichlet distribution
$\mathcal{C}(\cdot)$	Categorical distribution
$\mathcal{N}(\cdot)$	Gaussian distribution
$\mathcal{W}(\cdot)$	Wishart distribution
$\Gamma(\cdot)$	Gamma function
$\psi(\cdot)$	Digamma function
τ_0	Time shift coefficient
κ	Forgetting factor

In the generative process of LATEA, $z_{s,i}$ is first sampled from categorical distribution $\boldsymbol{\theta}_s$ for every $\boldsymbol{f}_{s,i}$ in $\boldsymbol{f}_s$, where $\boldsymbol{\theta}_s$ has a Dirichlet prior of parameter α. This $z_{s,i}$ then samples $w_{s,i}$ from categorical distribution $\boldsymbol{\phi}_{z_{s,i}}$ over acoustic topics associated with $z_{s,i}$, where $\boldsymbol{\phi}_{z_{s,i}}$ has a Dirichlet prior of parameter β. This $w_{s,i}$ samples an $\boldsymbol{f}_{s,i}$ from a $\mathcal{N}(\boldsymbol{\mu}_{w_{s,i}}, \boldsymbol{\Lambda}_{w_{s,i}})$, where $\mathcal{N}(\boldsymbol{\mu}_{w_{s,i}}, \boldsymbol{\Lambda}_{w_{s,i}})$ has a Gaussian-Wishart prior of parameters $\beta_0, \boldsymbol{\mu}_0, \nu_0$ and $\boldsymbol{B}_0$. This generative process is repeated for N_s to generate $\boldsymbol{f}_s$. The joint distribution of LATEA is expressed by the following equation,

$$\begin{aligned}
p(\boldsymbol{f}) &= \prod_{s=1}^{S}\prod_{i=1}^{N_s} p(\boldsymbol{f}_{s,i}|\boldsymbol{\theta}_s, \boldsymbol{\phi}_t, \boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m, \alpha, \gamma, \beta_0, \boldsymbol{\mu}_0, \nu_0, \boldsymbol{B}_0) \\
&= \prod_{s=1}^{S}\prod_{i=1}^{N_s}\sum_{t=1}^{T}\sum_{m=1}^{M} \mathcal{C}(z_{s,i}|\boldsymbol{\theta}_s)\mathcal{D}(\boldsymbol{\theta}_s|\alpha)\mathcal{C}(w_{s,i}|z_{s,i}, \boldsymbol{\phi}_{z_{s,i}}) \\
&\quad \cdot\mathcal{D}(\boldsymbol{\phi}_{z_{s,i}}|\gamma)\mathcal{N}(\boldsymbol{f}_{s,i}|w_{s,i}, \boldsymbol{\mu}_{w_{s,i}}, \boldsymbol{\Lambda}_{w_{s,i}})\mathcal{N}(\boldsymbol{\mu}_{w_{s,i}}|\beta_0, \mu_0, \boldsymbol{\Lambda}_{w_{s,i}}) \\
&\quad \cdot\mathcal{W}(\boldsymbol{\Lambda}_{w_{s,i}}|\nu_0, \boldsymbol{B}_0) \\
&= \prod_{s=1}^{S}\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\prod_{t=1}^{T}\theta_{st}^{\alpha-1+n_{st}} \cdot \prod_{t=1}^{T}\frac{\Gamma(M\gamma)}{\Gamma(\gamma)^M}\prod_{m=1}^{M}\phi_{tm}^{\gamma-1+n_{tm}} \\
&\quad \cdot \prod_{m=1}^{M}\frac{(\beta_m|\boldsymbol{\Lambda}_m|)^{1/2}}{(2\pi)^{D/2}}\exp\Big\{-\frac{\beta_m}{2}(\boldsymbol{\mu}_m-\boldsymbol{\mu}_0)^\intercal\boldsymbol{\Lambda}_m(\boldsymbol{\mu}_m-\boldsymbol{\mu}_0)\Big\} \\
&\quad \cdot C_m|\boldsymbol{\Lambda}_m|^{(\nu_m-D-1)/2}\exp\Big\{-\frac{1}{2}\mathrm{Tr}\big(\boldsymbol{B}_m\boldsymbol{\Lambda}_m\big)\Big\}, \qquad (1)
\end{aligned}$$

where we hypothesize that there is no temporal relation between acoustic features (acoustic words) because it can be considered that they are temporally exchangeable; therefore, we treat acoustic features as a "bag of acoustic features", which corresponds to the "bag of words" representation in natural language processing [11].

A. Batch VB algorithm for LATEA

In LATEA, the true posterior distribution of all unknown variables $p(\boldsymbol{z}, \boldsymbol{w}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{f})$ is intractable. Therefore, we estimate the posterior distribution with the VB method [12], which is faster in estimating these distributions than the collapsed Gibbs sampling (CGS) [13], [14] or expectation propagation (EP) methods [15]. With the VB method, the variational distribution $q(\boldsymbol{z}, \boldsymbol{w}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ is defined to estimate the true posterior distribution by iteratively approximating the variational distribution to the true distribution.

According to the VB method for LATEA, the appropriate variational parameters are obtained by maximizing the lower bound $\mathcal{F}(\boldsymbol{f})$ of the logarithm likelihood of all parameters through the update of $q(\boldsymbol{z}, \boldsymbol{w}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$. Specifically, to obtain an appropriate lower bound on the log likelihood of the distribution, Jensen's inequality is used as follows,

$$\begin{aligned}\mathcal{L}(\boldsymbol{f}) &\equiv \log p(\boldsymbol{f}|\alpha, \gamma, \mu_0, \beta_0, \nu_0, \boldsymbol{B}_0)\\ &= \iiiint \sum_{\boldsymbol{z}} \sum_{\boldsymbol{w}} \log q(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\phi}, \boldsymbol{\theta})\\ &\quad \cdot \frac{p(\boldsymbol{f}, \boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\phi}, \boldsymbol{\theta}|\alpha, \gamma, \mu_0, \beta_0, \nu_0, \boldsymbol{B}_0)}{q(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\phi}, \boldsymbol{\theta})} d\boldsymbol{\mu} d\boldsymbol{\Lambda} d\boldsymbol{\phi} d\boldsymbol{\theta}\\ &\geq \iiiint \sum_{\boldsymbol{z}} \sum_{\boldsymbol{w}} q(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\phi}, \boldsymbol{\theta})\\ &\quad \cdot \log \frac{p(\boldsymbol{f}, \boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\phi}, \boldsymbol{\theta}|\alpha, \gamma, \mu_0, \beta_0, \nu_0, \boldsymbol{B}_0)}{q(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\phi}, \boldsymbol{\theta})} d\boldsymbol{\mu} d\boldsymbol{\Lambda} d\boldsymbol{\phi} d\boldsymbol{\theta}\\ &\equiv \mathcal{F}[q]. \end{aligned} \tag{2}$$

It is then assumed that the variational distribution can be expressed as follows using mean field approximation,

$$q(\boldsymbol{z}, \boldsymbol{w}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\boldsymbol{z}, \boldsymbol{w}) q(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}). \tag{3}$$

By substituting Eq. (3) into (2), the appropriate variational distributions can be expressed by

$$q(\boldsymbol{z}, \boldsymbol{w}) \equiv \exp\left\{ \left\langle \log p(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{w}|\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \right\rangle_{q(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})} \right\} \tag{4}$$

$$q(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \equiv \exp\left\{ \left\langle \log p(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{w}|\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \right\rangle_{q(\boldsymbol{z}, \boldsymbol{w})} \right\}, \tag{5}$$

where $\left\langle \log p(\cdot) \right\rangle_{q(\cdot)}$ is an expectation of $\log p(\cdot)$ to $q(\cdot)$. We obtain the variational posterior distributions by calculating Eqs. (4) and (5) iteratively until evidence of the lower bound converges. Finally, the variational inference procedure of LATEA is derived, as shown in Algorithm 1.

Algorithm 1: Batch VB algorithm for LATEA

```
[Step1] Initialization
set  $\alpha_0, \gamma_0, \beta_0, \boldsymbol{\mu}_0, \nu_0, \boldsymbol{B}_0, h = 0$
initialize  $N_{st}^{(h)}, \alpha_{st}^{(h)}, N_{tm}^{(0)}, \gamma_{tm}^{(0)}, \boldsymbol{\mu}_m^{(0)}, \nu_m^{(0)}, \boldsymbol{B}_m^{(0)}, g_{\boldsymbol{\mu}_m}^{(0)}, \boldsymbol{\Sigma}_{\boldsymbol{\mu}_m}^{(0)}$
[Step2] Parameter estimation based on batch VB algorithm
repeat
  set $h \leftarrow h + 1$
  for  $s, n, t, m = 1$ to $S, N_s, T, M$ do
    $u_{snm} = \frac{1}{2}\Big\{\sum_d \psi\Big(\frac{\nu_0 + \sum_t N_{tm}^{(h)} + 1 - d}{2}\Big) + D \log 2 - \log|\boldsymbol{B}_m^{(h)}|\Big\}$
        $-\frac{1}{2}\mathrm{Tr}\Big\{\nu_m \boldsymbol{B}_m^{-1}\Big(\frac{g_{\boldsymbol{\mu}_m}}{g_{\boldsymbol{\mu}_m} - 2}\boldsymbol{\Sigma}_{\boldsymbol{\mu}_m} + (\boldsymbol{f}_n - \boldsymbol{\mu}_m)(\boldsymbol{f}_n - \boldsymbol{\mu}_m)^{\mathsf{T}}\Big)\Big\}$
        $+\sum_t R_{snt}\Big(\psi(\sum_t \gamma_{tm}^{(h)}) - \psi(\sum_m \sum_t \gamma_{tm}^{(h)})\Big)$
    $U_{snm} = \exp\{u_{snm}\} \Big/ \sum_m \exp\{u_{snm}\}$
    $r_{snt} = \sum_m U_{snm}\Big(\psi(\gamma_{tm}^{(h)}) - \psi(\sum_m \gamma_{tm}^{(h)})\Big) + \psi(\alpha_{st}^{(h)}) - \psi(\sum_t \alpha_{st}^{(h)})$
    $R_{snt} = \exp\{r_{snt}\} \Big/ \sum_t \exp\{r_{snt}\}$
    $N_{st}^{(h)} = \sum_n R_{snt}, \ \ \alpha_{st}^{(h)} = \alpha_0 + N_{st}^{(h)}$
  end for
until convergence condition is satisfied
for  $t, m = 1$ to $T, M$ do
  $N_{tm}^{(h)} = \sum_s \sum_n U_{snm} R_{snt}, \ \ \gamma_{tm}^{(h)} = \gamma_0 + N_{tm}^{(h)}$
end for
for  $s, n, t, m = 1$ to $S, N_s, T, M$ do
  $\overline{\boldsymbol{f}}_{sn}^{(h)} = \frac{\sum_s \sum_n U_{snm} \boldsymbol{f}_{sn}}{\sum_t N_{tm}^{(h)}}, \ \ \nu_m^{(h)} = \nu_0 + \sum_t N_{tm}^{(h)}$
  $\boldsymbol{\mu}_m^{(h)} = \frac{\beta_0 \boldsymbol{\mu}_0 + \sum_t N_{tm}^{(h)} \overline{\boldsymbol{f}}_{sn}}{\beta_0 + \sum_t N_{tm}^{(h)}}, \ \ g_{\boldsymbol{\mu}_m}^{(h)} = \nu_m^{(h)} + 1 - D$
  $\boldsymbol{B}_m^{(h)} = \boldsymbol{B}_0 + \sum_s \sum_n \sum_t U_{s'nm} R_{s'nt}(\boldsymbol{f}_{sn} - \overline{\boldsymbol{f}}_{sn}^{(h)})(\boldsymbol{f}_{sn} - \overline{\boldsymbol{f}}_{sn}^{(h)})^{\mathsf{T}}$
      $+ \frac{\beta_0 \sum_t N_{tm}^{(h)}}{\beta_0 + \sum_t N_{tm}^{(h)}}(\boldsymbol{\mu}_0 - \overline{\boldsymbol{f}}_{sn}^{(h)})(\boldsymbol{\mu}_0 - \overline{\boldsymbol{f}}_{sn}^{(h)})^{\mathsf{T}}$
  $\boldsymbol{\Sigma}_{\boldsymbol{\mu}_m}^{(h)} = \frac{\boldsymbol{B}_m^{(h)}}{(\beta_0 + \sum_t N_{tm}^{(h)}) g_{\boldsymbol{\mu}_m}^{(h)}}$
end for
```

III. Online Inference of Variational Posterior Distribution

The batch VB algorithm for LATEA requires calculating all variational posterior distributions of whole sound clips in each iteration. Although the batch VB algorithm generally requires less calculation compared to the batch collapsed Gibbs sampling [13], [14], even this algorithm becomes computationally very slow as the sound corpus grows. Moreover, the batch VB algorithm cannot model sequentially obtained sound clips. We, therefore, propose an online algorithm that does not need to calculate all distributions in each iteration nor to store all of the sound clips.

In our online algorithm, better approximations of $\boldsymbol{\phi}, \boldsymbol{\mu}$, and $\boldsymbol{\Lambda}$ are obtained by optimizing $\boldsymbol{\theta}_s$, z, and w in each sound clip. Our goal for our online algorithm is to estimate the variational posterior distributions that maximize a summation of the contribution of each sound clip in $F[q]$ as follows,

$$F[q] \equiv \sum_{S} F[q(s)]. \tag{6}$$

The online variational inference procedure of LATEA is obtained, as shown in Algorithm 2. We iteratively repeat the optimization of $U_{s'nm}$, $R_{s'nt}$, and $\alpha_{s't}$ locally, while holding the other parameters fixed then update other posterior distributions with the optimized $U_{s'nm}$, $R_{s'nt}$, and $\alpha_{s't}$. The updating weight is controlled by the repeated count k, time-shift parameter τ_0, and forgetting factor κ. We also introduce a mini-batch technique [16] for denoising [17] the learning dataset, which updates variational parameters with multiple sound clips at the same time in this inference procedure. With this mini-batch technique, we update $N_{tm}^{(h)}$ and $\overline{\boldsymbol{f}}_{s'n}^{(h)}$ with the

Algorithm 2: Online VB algorithm for LATEA

```
[Step1] Initialization
set  $\alpha_0, \gamma_0, \beta_0, \boldsymbol{\mu}_0, \nu_0, \boldsymbol{B}_0, \tau_0, \kappa, h = 0$
initialize  $N_{tm}^{(0)}, \gamma_{tm}^{(0)}, \boldsymbol{\mu}_m^{(0)}, \nu_m^{(0)}, \boldsymbol{B}_m^{(0)}, g_{\boldsymbol{\mu}_m}^{(0)}, \boldsymbol{\Sigma}_{\boldsymbol{\mu}_m}^{(0)}, \rho_h = (\tau_0)^{-\kappa}$
[Step2] Parameter estimation based on online VB algorithm
while sound clip is input do
  initialize $\alpha_{s't}^{(h)}, N_{s't}^{(h)}$
  repeat
    for  $s', n, t, m = 1$ to $S', N_s, T, M$ do
      $u_{s'nm} = \frac{1}{2}\Big\{\sum_d \psi\Big(\frac{\nu_0 + \sum_t N_{tm}^{(h)} + 1 - d}{2}\Big) + D \log 2 - \log |\boldsymbol{B}_m^{(h)}|\Big\}$
          $- \frac{1}{2}\mathrm{Tr}\Big\{\nu_m \boldsymbol{B}_m^{-1}\Big(\frac{g_{\boldsymbol{\mu}_m}}{g_{\boldsymbol{\mu}_m} - 2}\boldsymbol{\Sigma}_{\boldsymbol{\mu}_m} + (\boldsymbol{f}_n - \boldsymbol{\mu}_m)(\boldsymbol{f}_n - \boldsymbol{\mu}_m)^{\mathsf{T}}\Big)\Big\}$
          $+ \sum_t R_{s'nt}\Big(\psi(\sum_t \gamma_{tm}^{(h)}) - \psi(\sum_m \sum_t \gamma_{tm}^{(h)})\Big)$
      $U_{s'nm} = \exp\{u_{s'nm}\} \Big/ \sum_m \exp\{u_{s'nm}\}$
      $r_{s'nt} = \sum_m U_{s'nm}\Big(\psi(\gamma_{tm}^{(h)}) - \psi(\sum_m \gamma_{tm}^{(h)})\Big) + \psi(\alpha_{s't}^{(h)}) - \psi(\sum_t \alpha_{s't}^{(h)})$
      $R_{s'nt} = \exp\{r_{s'nt}\} \Big/ \sum_t \exp\{r_{s'nt}\}$
      $N_{s't}^{(h)} = \sum_n R_{s'nt}, \quad \alpha_{s't}^{(h)} = \alpha_0 + N_{s't}^{(h)}$
    end for
  until convergence condition is satisfied
  for  $t, m = 1$ to $T, M$ do
    $N_{tm}^{(h)} = \frac{S}{S'}\sum_{s'}\sum_{n} U_{snm} R_{s'nt}, \quad \tilde{\gamma}_{tm}^{(h)} = \gamma_0 + N_{tm}^{(h)}$
    $\gamma_{tm}^{(h)} = \gamma_{tm}^{(h-1)}(1 - \rho_k) + \tilde{\gamma}_{tm}^{(h)} \rho_k$
  end for
  for  $s', n, t, m = 1$ to $S', N_{s'}, T, M$ do
    $\overline{\boldsymbol{f}}_{s'n}^{(h)} = \frac{\frac{S}{S'}\sum_{s'}\sum_n U_{s'nm} \boldsymbol{f}_{s'n}}{\sum_t N_{tm}^{(h)}}, \quad \nu_m^{(h)} = \nu_0 + \sum_t N_{tm}^{(h)}$
    $\boldsymbol{\mu}_m^{(h)} = \frac{\beta_0 \boldsymbol{\mu}_0 + \sum_t N_{tm}^{(h)} \overline{\boldsymbol{f}}_{s'n}}{\beta_0 + \sum_t N_{tm}^{(h)}}, \quad g_{\boldsymbol{\mu}_m}^{(h)} = \nu_m^{(h)} + 1 - D$
    $\boldsymbol{B}_m^{(h)} = \boldsymbol{B}_0 + \frac{S}{S'}\sum_{s'}\sum_n \sum_t U_{s'nm} R_{s'nt} (\boldsymbol{f}_{s'n} - \overline{\boldsymbol{f}}_{s'n}^{(h)})(\boldsymbol{f}_{s'n} - \overline{\boldsymbol{f}}_{s'n}^{(h)})^{\mathsf{T}}$
        $+ \frac{\beta_0 \sum_t N_{tm}^{(h)}}{\beta_0 + \sum_t N_{tm}^{(h)}} (\boldsymbol{\mu}_0 - \overline{\boldsymbol{f}}_{s'n}^{(h)})(\boldsymbol{\mu}_0 - \overline{\boldsymbol{f}}_{s'n}^{(h)})^{\mathsf{T}}$
    $\boldsymbol{\Sigma}_{\boldsymbol{\mu}_m}^{(h)} = \frac{\boldsymbol{B}_m^{(h)}}{(\beta_0 + \sum_t N_{tm}^{(h)}) g_{\boldsymbol{\mu}_m}^{(h)}}$
  end for
  set set $h \leftarrow h + 1, \quad \rho_h = (h + \tau_0)^{-\kappa}$
end while
```

following equations:

$$N_{tm}^{(h)} = \frac{S}{S'}\sum_{s'}\sum_n U_{s'nm} R_{s'nt}, \tag{7}$$

$$\overline{\boldsymbol{f}}_{s'n}^{(h)} = \frac{\frac{S}{S'}\sum_{s'}\sum_n U_{s'nm} \boldsymbol{f}_{s'n}}{\sum_t N_{tm}^{(h)}}, \tag{8}$$

where S' is the number of sound clips in a mini-batch. When $S' = 1$, this algorithm corresponds to the simple online VB algorithm without the mini-batch technique, and when $S' = S$, it corresponds to the batch VB algorithm.

IV. Experiments

We evaluated the performance of our proposed online algorithm and its calculation efficiency by analyzing the sound structure using sounds recorded in a living room. We used 1,504 real-life sounds, which included nine categories of acoustic scenes: *"chatting," "cooking," "eating dinner," "operating a PC," "reading a newspaper," "vacuuming," "walking," "washing dishes," and "watching TV,"* though we analyzed sound structure without these acoustic scene labels. For acoustic features, 12-dimensional Mel-frequency cepstral coefficients (MFCCs) were calculated from every segmented sound clip with 50% overlap, and each sound clip was composed of 1,000 acoustic features. The other experimental conditions are listed in Table II. We also evaluated batch ATM, online ATM [18], and batch LATEA as comparative methods and used the same parameter set as the proposed online LATEA.

TABLE II
Experimental conditions

Sampling rate / quantization	16 kHz / 16 bits
Frame size / shift	512 / 256
Acoustic word size	8–256
Hyperparameter α / γ	1.0 / 1.0
Hyperparameter β_0 / $\boldsymbol{\mu}_0$	5.0 / **O**
Hyperparameter ν_0 / $\boldsymbol{B}_0$	13.0 / **I**
τ_0 / κ	2.0 / 0.75

TABLE III
Runtimes of online and batch LATEA algorithms (in min.)

# acoustic word/topic	Batch LATEA	Online LATEA ($S' = 10$)	Online LATEA ($S' = 50$)
$M = 8,\ T = 10$	80.6	9.9	8.3
$M = 16,\ T = 10$	160.0	19.8	16.3
$M = 32,\ T = 20$	555.1	69.4	57.9
$M = 64,\ T = 20$	1,329.8	141.4	119.2
$M = 128,\ T = 20$	2,595.3	284.1	249.3
$M = 256,\ T = 50$	12,970.7	1,242.8	1127.0

Table III lists the runtimes of the batch and online LATEA algorithms with various parameters. The experimental results indicate that online LATEA can estimate posterior distributions at about a $1/10$ calculation cost. Moreover, it combines the calculation efficiency and performance of the acoustic topic estimation, as shown in Table III and Fig. 4. Figure 4 shows the acoustic topic estimation results of a sound associated with "cooking." Actual acoustic scenes are depicted in the upper part of the figures, and each color-coded acoustic signal denotes the acoustic topic estimated with batch ATM, online ATM, batch LATEA, and online LATEA. As shown in these figures, online LATEA can extract a sound structure represented by an acoustic scene as well as batch LATEA because the extracted acoustic topics agree equally well with actual acoustic scenes in both algorithms, while the conventional online ATM confused acoustic scenes in their acoustic topics. These results indicate that online LATEA better represents the acoustic similarity and variance between acoustic words and enables the better description of the relationship between acoustic scenes and topics.

We evaluated the perplexities for the LATEA algorithms that determine their generalization performance; they are calculated

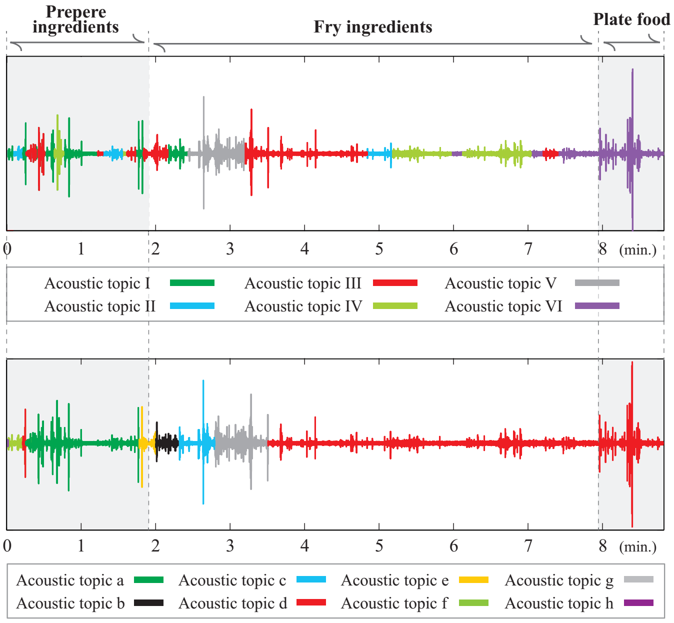

Fig. 3. Acoustic topic estimation results in acoustic scene ”cooking” with batch ATM (upper, $T = 20, M = 256$) and online ATM (lower, $T = 20, M = 256, S' = 50$)

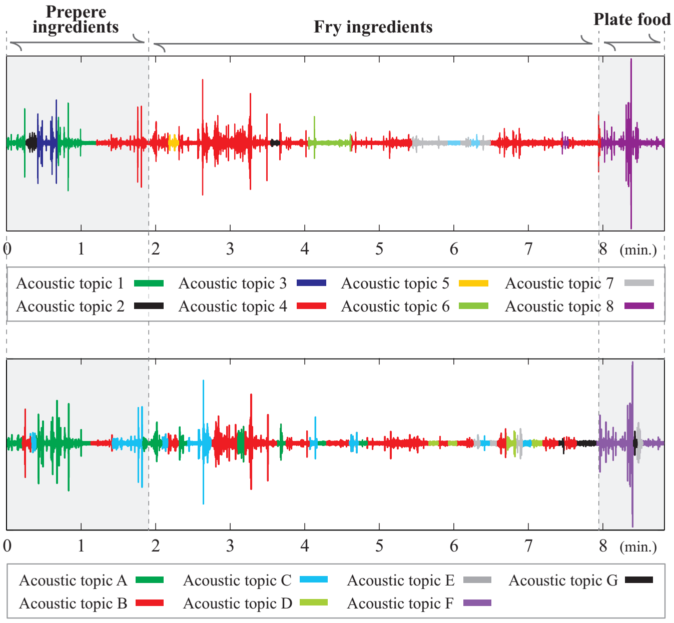

Fig. 4. Acoustic topic estimation results in acoustic scene ”cooking” with batch LATEA (upper, $T = 20, M = 256$) and online LATEA (lower, $T = 20, M = 256, S' = 50$)

as follows,

$$\text{Perplexity}(\boldsymbol{f}) = \exp\left[-\frac{\sum_{s=1}^{S} \log p(\boldsymbol{f}_s)}{\sum_{s=1}^{S} N_s}\right]. \tag{9}$$

For the perplexity with the LATEA algorithms, we chose $T = 20, 30, 50$, and $M = 8, 16, 32, 64, 128, 256$, and fit per-sound clip parameters $U_{s'nm}$, $R_{s'nt}$, and $\alpha_{s't}$ to test sound clips preliminarily. As shown in Fig. 5, these test sounds recorded in a living room can be modeled well with the proposed online LATEA when using a few hundred classes of acoustic words and a few dozens of acoustic topics because the results are close to the perplexity in the batch LATEA. Meanwhile, when using a few dozen classes of acoustic words, the online LATEA results in higher perplexities because the models confuse multiple classes of acoustic words, which is supposed to be separated into different classes of acoustic words. Moreover, when $M = 8$, it is estimated that the acoustic word distribution has degenerated to a couple of classes, and therefore, the perplexity falsely drop to a lower value.

V. Conclusion

We proposed an online learning algorithm for LATEA that can be applied to sequentially obtained acoustic signals and used to analyze the sound structures that are organized by the combination of the latent acoustic topics. In LATEA, a generative process of acoustic feature sequences is modeled hierarchically; it represents sound clips as categorical distributions over acoustic topics, and each acoustic topic is represented by a GMM over an acoustic feature space, which can capture the acoustic similarity between acoustic words or the variance of each acoustic word in the generative model. For the online learning of LATEA, we derived a VB-based online algorithm that sequentially estimates the appropriate posterior distributions of every several sound clips. The experimental results indicate that the proposed algorithm can estimate posterior distributions as efficiently as batch LATEA at a fraction of the calculation cost.

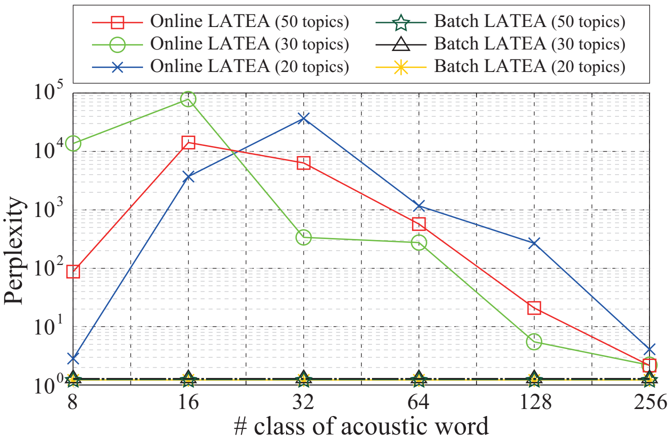

Fig. 5. Averaged perplexity for batch and online LATEA

References

[1] R. Radhakrishnan, A. Divakaran, and P. Smaragdis, “Audio analysis for surveillance applications,” *Proc. 2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (*WASPAA*), pp. 158–161, 2005.

[2] 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 Lang. Process.*, vol. 14, no. 1, pp. 321–329, 2006.

[3] Y. Peng, C. Lin, M. Sun, and K. Tsai, "Healthcare audio event classification using hidden Markov models and hierarchical hidden Markov models," *Proc. IEEE International Conference on Multimedia and Expo* (*ICME*), pp. 1218–1221, 2009.

[4] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Process. Mag.*, pp. 16–34, 2015.

[5] S. Kim, S. Narayanan, and S. Sundaram, "Acoustic topic models for audio information retrieval," *Proc. 2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (*WASPAA*), pp. 37–40, 2009.

[6] K. Lee and D. P. W. Ellis, "Audio-based semantic concept classification for consumer video," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 6, pp. 1406–1416, 2010.

[7] K. Imoto, Y. Ohishi, H. Uematsu, and H. Ohmuro, "Acoustic scene analysis based on latent acoustic topic and event allocation," *Proc. IEEE International Workshop on Machine Learning for Signal Processing* (*MLSP*), 2013.

[8] K. Imoto and S. Shimauchi, "Acoustic scene analysis based on hierarchical generative model of acoustic event sequence," *IEICE Trans. Inf. & Syst.*, vol. E99-D, no. 10, pp. 2539–2549, October 2016.

[9] S. Kim, P. Georgiou, and S. Narayanan, "On-line genre classification of tv programs using audio content," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 798–802, 2013.

[10] K. Imoto and N. Ono, "Online acoustic scene analysis based on nonparametric Bayesian model," *Proc. European Signal Processing Conference* (*EUSIPCO*), pp. 988–992, 2016.

[11] T. Joachims, "Learning to classify text using support vector machines: Methods, theory, and algorithms," *J. Comput. Linguist.*, vol. 29, pp. 655–664, 2003.

[12] H. Attias, "A variational bayesian framework for graphical models," *In Adv. Neural Inf. Proc. Syst. 12*, pp. 209–215, 2000.

[13] 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.

[14] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *PNAS*, vol. 1, pp. 5228–5235, 2004.

[15] T. P. Minka and J. Lafferty, "Expectation propagation for the generative aspect model," *in Proc. of the 17th conference on Uncertainty in artificial intelligence (UAI)*, 2002.

[16] M. D. Hoffman, D. M. Blei, and F. Bach, "Online learning for latent dirichlet allocation," *in Proc. of NIPS 2010*, pp. 856–864, 2010.

[17] L. Bottou and O. Bousquet, "The tradeoffs of large scale learning," *In Adv. Neural Inf. Proc. Syst. 20*, pp. 161–168, 2008.

[18] K. Imoto, Y. Ohishi, H. Uematsu, H. Ohmuro, and N. Ono, "Online acoustic scene analysis with sequentially obtained acoustic event sequence," *J. Acoust. Soc. Jpn*, vol. 72, no. 6, pp. 293–305, June 2016 (In Japanese).