Online Acoustic Scene Analysis Based on Nonparametric Bayesian Model

Keisuke Imoto* and Nobutaka Ono†*
*SOKENDAI (The Graduate University for Advanced Studies), Kanagawa, Japan
†National Institute of Informatics, Tokyo, Japan

***Abstract*—In this paper, we propose a novel online method for analyzing acoustic scenes from sequentially obtained sounds. One prospective method for analyzing acoustic scenes is the use of a generative model of acoustic topics and event sequences in observed sounds, where the acoustic topic represents the latent structure of acoustic events associating an acoustic scene and acoustic events. This generative model is called an acoustic topic model (ATM). However, the conventional ATM employs a batch technique for estimating model parameters and cannot model sequentially obtained acoustic event sequences. Moreover, the number of classes of acoustic topics that lies in acoustic event sequences needs to be predetermined before observing acoustic events. However, the necessary number of acoustic topics for representing acoustic scenes varies in accordance with their contents, and this causes a mismatch between the actual number of classes of acoustic topics and the predetermined number of classes. In our method, the number of classes of acoustic topics can be automatically inferred from sequentially obtained acoustic event sequences on the basis of the online and nonparametric Bayesian technique. The experimental results of online acoustic scene estimation using real-life sounds indicated that the proposed method performed of acoustic scene classification better than the conventional ATM. In addition, the proposed method produced an efficient computation performance.**

I. Introduction

A lot of attention has been drawn recently to applications for monitoring elderly people [1], security surveillance [2], automatic classification of user activities and contexts [3], [4], and multimedia retrieval [5], which utilize the information obtained from various acoustic signals. There are some useful techniques to realize these applications. One is acoustic event detection (AED), which analyzes various types of acoustic events (e.g., *"footsteps," "running water," "music,"*) for detecting or classifying specific types of sounds [6], [7], and another is acoustic scene analysis (ASA), which analyzes a scene in which sounds are produced such as user activities (e.g., *"cooking," "vacuuming," "watching TV"*) or situations (e.g., *"on the train," "in a meeting"*) [8]. In this paper, we focus on automatic estimation of acoustic scenes, especially user activities.

One simple approach for analyzing acoustic scenes is focused on a combination of acoustic events. For example, an acoustic scene "*cooking*" is marked by a combination of acoustic events including "*running water*," "*cutting with a knife*," and "*heating a skillet*." On the basis of this idea, Heittola *et al.* [9] and Guo and Li [10] proposed acoustic scene classification methods based on the histogram of acoustic events and support vector machine (SVM) [11]. Kim *et al.* [12], Lee and Ellis [13], and Imoto *et al.* [14], [15] focused on the fact that the probabilities of acoustic events depend on the acoustic scenes and proposed generative models to represent acoustic event sequences contained within long-term sounds. Their models are called an acoustic topic model (ATM). In the ATM, the relationship between acoustic scenes and acoustic events is represented by a generative model using an acoustic topic, which is a latent variable to control a generative probability of acoustic events. The original ATM is a batch model that needs to prepare many acoustic event sequences preliminarily for analyzing them, and therefore, it suffers from the disadvantage that sequentially obtained acoustic event sequences are difficult to model.

Then, even though an online method of ATM can be introduced with reference to [16], [17], it still needs to determine the number of classes of acoustic topics that lies in acoustic event sequences before observing acoustic event sequences. Therefore, the simple online ATM causes a mismatch between the actual number of classes of acoustic topics in obtained acoustic event sequences and the predetermined number of classes. This leads to modeling inappropriate relations between acoustic scenes and events, and as a consequence, the degradation in the performance of the acoustic scene analysis.

In this paper, we propose a probabilistic generative model of acoustic event sequences and propose a parameter estimation method that can estimate the number of classes of acoustic topics from sequentially obtained acoustic event sequences. Our proposed model can adaptively estimate the number of acoustic topics in accordance with the content of each acoustic event sequence.

The present paper is divided as follows. Section 2 contains an overview describing the conventional and proposed generative models of acoustic event sequences. Section 3 summarizes parameter estimation for the proposed model. In section 4, the results from real environment experiments are discussed, and in section 5, we conclude this paper.

II. Probabilistic Generative Model of Acoustic Event Sequence

A. Conventional Acoustic Topic Model

In this paper, we define an acoustic event as a label for representing kinds of sounds at each frame, which can be obtained by an unsupervised clustering as a pre-processing. Then, a sound clip (for example with 30 seconds) is represented as an acoustic event sequence, which is a time series of acoustic

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

events. We define an acoustic scene as a label for each acoustic event sequence. In modeling a relationship between acoustic scenes and acoustic event sequences, the acoustic topic model (ATM) [12] has been proposed to model a generative process of acoustic event sequences in an unsupervised manner. This kind of generative model is originally proposed as the latent Dirichlet allocation (LDA) in the area of natural language processing [18], whereas the ATM uses it for an acoustic event sequence. The ATM assumed that a generative process of acoustic event sequences can be represented by a hierarchical generative process of acoustic topics and events, where the acoustic topic represents the latent structure of acoustic events that associates an acoustic scene and an acoustic event sequence. That is, in ATM acoustic scenes are characterized through acoustic topics indirectly.

In particular, ATM models the generative process of an acoustic event sequence $\boldsymbol{e}_s$ as follows,

$$\begin{aligned} \boldsymbol{\theta}_s &\sim \mathrm{Dir}(\beta) \\ \boldsymbol{\phi}_t &= \mathrm{Dir}(\sigma) \\ z_{si}\ (\in T) \mid \boldsymbol{\theta}_s &\sim \mathrm{Mult}(\boldsymbol{\theta}_s) \\ e_{si}\ (\in M) \mid \boldsymbol{\phi}_{z_{si}}, z_{si} &= \mathrm{Mult}(\boldsymbol{\phi}_t) \end{aligned} \tag{1}$$

where β and σ are hyperparameters of the Dirichlet distributions. The $\boldsymbol{\theta}_s$ and $\boldsymbol{\phi}_t$ are probability distributions over an acoustic topic in s and an acoustic event in t, respectively. The definitions of other variables used in this paper are listed in Table I. The ATM models acoustic scenes by using the predetermined number of acoustic topics (T) for every acoustic event sequence; however, the necessary number of acoustic topics for representing acoustic scenes varies in accordance with their contents. Moreover, the necessary number and combination of acoustic topics for describing each event sequence varies in accordance with the content of each acoustic event sequence. In the original ATM, an appropriate number of acoustic topics needs to be anticipated and set from a preliminarily obtained dataset, and therefore, sequentially obtained acoustic event sequences are difficult to model with the appropriate number of acoustic topics.

We thus can consider an online version of ATM referring to [17]. However, there remains a problem that it needs to set the number of acoustic topics before observing acoustic event sequences. This causes a mismatch between the actual number of acoustic topics in sequentially obtained acoustic event sequences and the predetermined number of acoustic topics. This may result in degrading the accuracy of the acoustic scene analysis and may cause an unnecessary calculation cost.

B. Acoustic Topic Model with Adaptively Estimating the Number of Acoustic Topics

We address the issue of the original ATM by proposing an online acoustic topic model that can infer the appropriate number of acoustic topics from sequentially obtained data. In this paper, to apply ATM to the sequentially obtained data, we preliminarily segment a continuously obtained acoustic event sequence into multiple acoustic event sequences. Then, it is assumed that we observe the data sequentially in this segmented acoustic event sequence unit. Thus, this segmented acoustic event sequence is called an "acoustic event sequence" and its index is referred by s in this paper.

TABLE I
DEFINITION OF VARIABLES

Symbol	Definition
S	Total number of acoustic event sequences
K	Maximum number of acoustic topic categories in all acoustic event sequences
T	Maximum number of acoustic topic categories in each acoustic event sequence
M	Number of acoustic event categories
$N_{\boldsymbol{e}_s}$	Number of acoustic events in $\boldsymbol{e}_s$
s	Index of acoustic event sequences
k	Class index of acoustic topics in corpus level
t	Class index of acoustic topics in event sequence level
m	Class index of acoustic events
i	Order index of acoustic event in each event sequence
$\mathcal{E}$	Class of acoustic event sequences
G_0	Prior of acoustic topic distribution in corpus level
G_s	Acoustic topic distribution in sequence level
θ_k	Corpus level acoustic topic atom
η_t	Event sequence level acoustic topic atom
β_k	Weight of atom ϕ_k (corpus level topic distribution)
π_{st}	Weight of atom η_{si} (sequence level topic distribution)
z_{si}	Acoustic topic for acoustic event e_{si}
e_{si}	ith acoustic event in acoustic event sequence s
ξ_{st}	Multinomial variational parameter for c_{st}
ζ_{si}	Multinomial variational parameter for z_{si}
λ_t	Dirichlet variational parameter for ϕ_k
c_{st}	Relation indicator between corpus and acoustic event sequence level atoms
γ, α_0	Hyperparameter of beta distribution
u_k, w_k	Parameters of beta distribution relevant to corpus level topic distribution
a_{st}, b_{st}	Parameters of beta distribution relevant to acoustic event sequence level topic distribution
$\mathrm{Dir}(\cdot)$	Dirichlet distribution
$\mathrm{Mult}(\cdot)$	Multinomial distribution
$\mathrm{Beta}(\cdot)$	Beta distribution
H	Symmetric Dirichlet distribution over acoustic event

In the proposed model, we introduce the hierarchical Dirichlet process [19], [20] to ATM, which can estimate the appropriate number of acoustic topics for a dataset. We call this model nonparametric ATM (nATM). We first discuss a generative process of acoustic event sequences on the basis of nATM in this section, and then introduce an online parameter estimation method for nATM in the next section.

The nATM explicitly includes a process of determining the number of acoustic topics in a generative process of acoustic event sequences, and therefore, the nATM can adaptively estimate the number of acoustic topics to the dataset. In particular, we model the generative process of the acoustic event sequences with two stick-breaking construction processes [21], [22] for determining the number of acoustic topics in two levels: all-acoustic-event-sequences level (corpus level) and each-acoustic-event-sequence level (sequence level). This multi-level generative process enables to determine the appropriate number of acoustic topics in both of the corpus level and the sequence level.

In nATM, the prior of acoustic topics that corresponds to the corpus level acoustic topics is first generated.

$$\begin{aligned} \beta'_k &\sim \text{Beta}(1,\gamma) \\ \beta_k &= \beta'_k \textstyle\prod_{c=1}^{k-1}(1-\beta'_c) \\ \theta_k &\sim H \\ G_0 &= \textstyle\sum_{k=1}^{\infty}\beta_k\delta_{\theta_k} \end{aligned} \tag{2}$$

Here, θ_k and β_k indicate the common topic "seeds" called corpus level atoms in all acoustic event sequences and the topic distribution for each θ_k. Then, topic distributions in each acoustic event sequence are generated from the corpus level acoustic topics as follows,

$$\begin{aligned} \eta_{st} &\sim G_0 \\ \pi'_{st} &\sim \text{Beta}(1,\alpha_0) \\ \pi_{st} &= \pi'_k \textstyle\prod_{c=1}^{t-1}(1-\pi'_{sc}) \\ G_s &= \textstyle\sum_{k=1}^{\infty}\pi_{st}\delta_{\eta_{st}}, \end{aligned} \tag{3}$$

where η_{st} indicates the topic "seeds" called acoustic event sequence level atoms in each acoustic event sequence and π_{st} denotes the topic distribution for each η_{st}. Two stick-breaking construction processes enable modeling of acoustic event sequences with numbers and types of acoustic topics tailored to each acoustic event sequence. After that, topic distributions in each acoustic event sequence are generated from the corpus level acoustic topics as follows.

$$\begin{aligned} c_{st} &\sim \text{Beta}(\boldsymbol{\beta}) \\ z_{si} &\sim \text{Beta}(\boldsymbol{\pi}_s) \\ \phi_{si} &= \eta_{sz_{si}} = \theta_{c_{sz_{si}}} \\ e_{si} &\sim \text{Beta}(\phi_{si}), \end{aligned} \tag{4}$$

where we introduce the relation indicator c_{st} for mapping corpus level atoms to acoustic event sequence level atoms. Finally, acoustic topics and events are generated from the topic and event distributions.

III. Parameter Estimation Method for Online Nonparametric ATM

We next describe a model parameter estimation method of nATM based on the variational inference. A parameter estimation for batch nATM is first derived, and then a method for an online algorithm is provided.

A. Parameter Estimation for Batch nATM

To estimate model parameters of batch nATM, we need to find parameters that maximize $p(\boldsymbol{\beta}', \boldsymbol{\pi}', \boldsymbol{c}, \boldsymbol{z}, \boldsymbol{\phi}|\mathcal{E})$. In a variational inference of nATM, it is intractable to infer model parameters of nATM analytically. Therefore, we introduce a variational distribution $q(\boldsymbol{\beta}', \boldsymbol{\pi}', \boldsymbol{c}, \boldsymbol{z}, \boldsymbol{\phi})$ and approximate this to true posterior distributions of model parameters. For this variational distribution, we apply the following mean field approximation to $q(\boldsymbol{\beta}', \boldsymbol{\pi}', \boldsymbol{c}, \boldsymbol{z}, \boldsymbol{\phi})$.

$$\begin{aligned} q(\boldsymbol{\beta}', \boldsymbol{\pi}', \boldsymbol{c}, \boldsymbol{z}, \boldsymbol{\phi}) &= q(\boldsymbol{\beta}')\, q(\boldsymbol{\pi}')\, q(\boldsymbol{c})\, q(\boldsymbol{z})\, q(\boldsymbol{\phi}) \\ &= \prod_{k=1}^{K} q(\beta'_k|u_k, w_k) \cdot \prod_{s=1}^{S}\prod_{t=1}^{T} q(\pi'_{st}|a_{st}, b_{st}) \\ &\quad \cdot \prod_{s=1}^{S}\prod_{t=1}^{T} q(c_{st}|\xi_{st}) \cdot \prod_{s=1}^{S}\prod_{i=1}^{N_s} q(z_{si}|\zeta_{si}) \cdot \prod_{k=1}^{K} q(\phi_k|\lambda_k) \end{aligned} \tag{5}$$

Then, we consider the marginal log likelihood for obtained acoustic event sequences $\log p(\mathcal{E}|\gamma, \alpha_0, \upsilon)$. According to the variational inference [23], we can estimate appropriate posterior distributions by maximizing a lower boundary obtained by applying Jensen's inequality to $\log p(\mathcal{E}|\gamma, \alpha_0, \upsilon)$, and finally, we can obtain the following corpus level and acoustic event sequence level parameter updates.

Acoustic event sequence level parameter updates:

$$\begin{aligned} a_{st} &= 1 + \textstyle\sum_{i=1}^{N_s}\zeta_{sit} \\ b_{st} &= \alpha_0 + \textstyle\sum_{i=1}^{N_s}\sum_{d=t+1}^{T}\zeta_{sid} \\ \xi_{stk} &\propto \exp\big(\textstyle\sum_{i=1}^{N_s}\zeta_{sit}\mathbb{E}_q[\log p(e_{si}|\phi_k)] + \mathbb{E}_q[\log\beta_k]\big) \\ \zeta_{sik} &\propto \exp\big(\textstyle\sum_{k=1}^{K}\xi_{stk}\mathbb{E}_q[\log p(e_{si}|\phi_k)] + \mathbb{E}_q[\log\pi_{st}]\big) \end{aligned} \tag{6}$$

Corpus level parameter updates:

$$\begin{aligned} u_k &= 1 + \textstyle\sum_{s=1}^{S}\sum_{t=1}^{T}\xi_{stk} \\ w_k &= \gamma + \textstyle\sum_{s=1}^{S}\sum_{t=1}^{T}\sum_{c=k+1}^{K}\xi_{stc} \\ \lambda_{ke} &= \upsilon + \textstyle\sum_{s=1}^{S}\sum_{t=1}^{T}\xi_{stk}(\sum_{i=1}^{N_s}\zeta_{sit}\mathrm{I}[e_{si}=m]) \end{aligned} \tag{7}$$

where $\mathrm{I}[e_{si}=m]$ becomes 1 if $e_{si}=m$ and 0 otherwise. Corpus level and acoustic event sequence level updates are iteratively calculated for the batch nATM parameter estimation until a convergence condition is satisfied. A more detailed discussion of the parameter estimation of an equivalent generative model to nATM can be found in the work of Wang *et al.* [22].

B. Parameter Estimation for Online nATM

We then propose an online parameter estimation method for nATM. In the online parameter estimation, we need to update model parameters without storing all acoustic event sequences, and therefore, appropriate posterior distributions are estimated by sequentially maximizing the contribution of each acoustic event sequence to $\sum_{s=1}^{S}\log p(\boldsymbol{e}_s|\gamma,\alpha_0,\upsilon)$ instead of maximizing $\log p(\mathcal{E}|\gamma,\alpha_0,\upsilon)$ directly. To maximize $\sum_{s=1}^{S}\log p(\boldsymbol{e}_s|\gamma,\alpha_0,\upsilon)$, this paper introduces a stochastic optimization based on the natural gradient of the variational distribution [17], [24], [25], which allows a fast and a tractable online algorithm. Specifically, given a new acoustic event sequence, acoustic event level parameters are first updated with Eq. (6) iteratively while keeping the corpus level parameters fixed. Then, the corpus level parameters are updated with the following updates, which corresponds to the stochastic optimization of the natural gradient.

$$u_k^{(h)} = (1-\rho^{(h)})u_k^{(h-1)} + \rho^{(h)}(1 + \sum_{s=1}^{S}\sum_{t=1}^{T}\xi_{stk})$$

$$w_k^{(h)} = (1-\rho^{(h)})w_k^{(h-1)} + \rho^{(h)}(\gamma + \sum_{s=1}^{S}\sum_{t=1}^{T}\sum_{c=k+1}^{K}\xi_{stc})$$

TABLE II
EXPERIMENTAL CONDITIONS

Sampling rate / quantization	16 kHz / 16 bits
Frame size / shift	512 / 256
Acoustic event sequence size	16-s (including 1,000 events)
Hyperparameter γ / α_0	3.3333 / 0.1
Time shift parameter τ_0	2.0
Forgetting factor κ	0.6

$$\lambda_{ke}^{(h)} = (1-\rho^{(h)})\lambda_{ke}^{(h-1)} + \rho^{(h)}(\upsilon + \sum_{s=1}^{S}\sum_{t=1}^{T}\xi_{stk}(\sum_{i=1}^{N_s}\zeta_{sit}\mathrm{I}[e_{si}=m])) \quad (8)$$

where the updating weight is controlled by the repeated count h, time shift parameter τ_0, and forgetting factor κ.

IV. EXPERIMENTS

A. Experimental Conditions

We evaluated how efficiently and effectively online nATM can model acoustic scenes through the perplexity, computation cost, and classification accuracy of acoustic scenes compared with nATM and other conventional models. In this experiment, we used 11,105 real environment acoustic signals that include nine categories of user activities: *"chatting," "cooking," "eating dinner," "operating PC," "reading a newspaper," "vacuuming," "walking," "washing dishes," and "watching TV."* Of these acoustic signals, 9,802 were used for parameter estimation and 1,303 for evaluation. For each acoustic signal, 12-dimensional Mel-Frequency Cepstral Coefficients (MFCCs) were calculated from every segmented acoustic signal with 50% overlap, and then acoustic events were recognized by using a Gaussian Mixture Model (GMM) with 128 acoustic event classes frame by frame [14]. The other experimental conditions are listed in Table II.

B. Perplexity and Computation Time

In the first experiment, we evaluated the generalization performance using perplexity. Perplexity evaluates how well a model predicts a dataset, and a lower perplexity indicates a better generalization performance. In particular, the perplexity for the proposed method was calculated as follows.

$$\mathrm{Perplexity}(S) = \exp\left\{-\frac{\sum_{s=1}^{S}\log p(\boldsymbol{e}_s)}{\sum_{s=1}^{S}N_s}\right\} \quad (9)$$

We also compared how efficiently the proposed method models acoustic event sequences by using the computation time. The times for modeling 9,802 acoustic event sequences were measured by using a PC with an Intel Core i7-870 (2.93GHz) CPU. In this experiment, we compared four algorithms: online nATM (proposed), online ATM based on VB (Online ATM-VB), batch ATM based on VB (Batch ATM-VB), and batch ATM based on collapsed Gibbs sampling (Batch ATM-CGS) [26], [27].

Figure 1 shows perplexities for the calculation time in each model. For the conventional models, we calculated perplexities and computation times with $T = 10 - 100$. These results indicate that the proposed model clearly came out on top with its very good perplexity value (4.69) and it has advantages in terms of the generalization performance than the batch models. The reason that the proposed model achieved a lower perplexity than the online ATM-VB is that it can estimate parameters while adjusting the number of acoustic topics necessary for representing acoustic event sequences. Moreover, the results also indicate that the proposed model needs less time than the batch models because of same reason it achieved the lower perplexity value. Overall, the proposed model achieves a balance between good generalization performance and less calculation time.

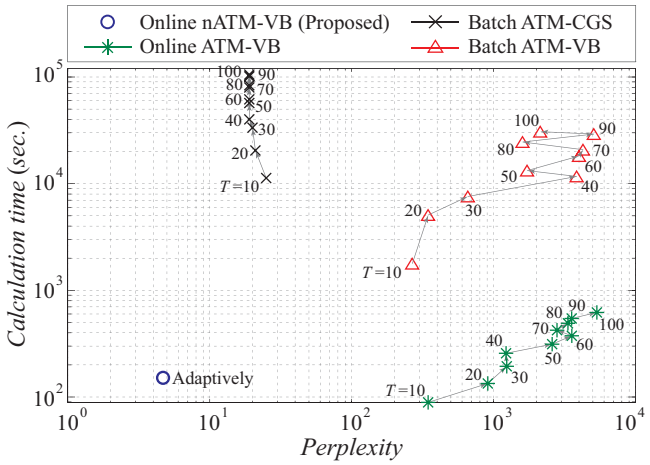

Fig. 1. Perplexity and calculation time for online nATM and conventional models

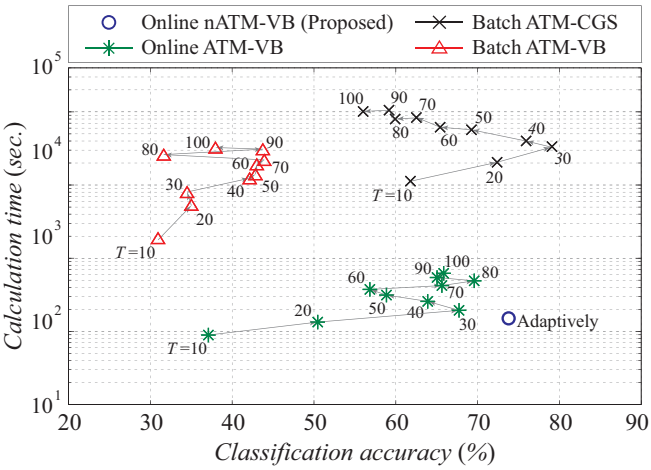

Fig. 2. Classification accuracy and calculation time for online nATM and conventional models

C. Classification Accuracy

The second experiment was conducted to evaluate how well the proposed online nATM models the relationship correctly between acoustic event sequences and their corresponding user activities through acoustic topics. For this experiment, we manually labeled acoustic event sequences with acoustic scenes as the ground truth. After the estimation of model parameters, the acoustic scene classification was conducted by using a multi-class support vector machine (SVM), in which acoustic topic distributions are used as the feature of acoustic

scenes in the same manner as [12]. We thus calculated the estimation accuracy of acoustic event sequences to a user activity as follows.

$$\text{Accuracy}(a) = \frac{\text{\# sequences correctly classified}}{\text{\# sequences classified to acoustic scene } a} \quad (10)$$

Figure 2 plots the average classification accuracy for the calculation time in each model. The proposed method achieved higher accuracy classification performance (73.80%) than the other VB based ATMs and comparable performance to the batch CGS based ATM, which a prominent method for estimating model parameters. The further investigation of the results shows that there is the appropriate number of acoustic topics for a dataset. If we select the inappropriate number of acoustic topics in the conventional method, the classification performance degrades or unnecessary calculation cost is required. On the other hand, the proposed method outperforms conventional models without predetermining the number of acoustic topics.

V. Conclusion

A conventional acoustic topic model cannot model sequentially obtained acoustic event sequences and may require unnecessary calculation time. Moreover, the number of classes of acoustic scenes in acoustic event sequences needs to be predetermined before obtaining them, and this causes a mismatch between the actual number of classes of acoustic scenes and the predetermined number of classes. In this paper, we proposed an online acoustic topic model that can infer the number of classes of acoustic scenes from sequentially obtained acoustic event sequences. In the proposed model, a generative process of an acoustic event sequence is modeled with the hierarchical Dirichlet process to select the appropriate number of acoustic topics from a dataset automatically and estimate their parameters with a VB-based online parameter estimation method. Acoustic scene estimation experiments with real-life sounds indicated that the proposed method produced a more efficient computation and higher acoustic scene classification performance than the conventional VB-based ATMs.

Acknowledgments

This work was supported by a Grant-in-Aid for Scientific Research (A) (Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number 16H01735). The sound dataset used for this research was originally recorded by Nippon Telegraph and Telephone Corporation (NTT).

References

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

[2] A. Harma, M. F. McKinney, and J. Skowronek, "Automatic surveillance of the acoustic activity in our living environment," *Proc. the* IEEE *International Conference on Multimedia and Expo* (*ICME*), 2005.

[3] 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.

[4] K. Imoto and N. Ono, "Spatial-feature-based acoustic scene analysis using distributed microphone array," *Proc. the European Signal Processing Conference* (*EUSIPCO*), pp. 739–743, 2015.

[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," *Proc. the* IEEE *International Conference on Acoustics, Speech and Signal Process.* (*ICASSP*), pp. 3163–3167, 2013.

[6] 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.

[7] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real life recordings," *Proc. the 18th European Signal Processing Conference* (*EUSIPCO*), pp. 1267–1271, 2010.

[8] 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.

[9] T. Heittola, A. Mesaros, A. Eronen, and A. Klapuri, "Audio content recognition using audio event histograms," *Proc. the 18th European Signal Processing Conference* (*EUSIPCO*), pp. 1272–1276, 2010.

[10] G. Guo and S. Z. Li, "Content-based audio classification and retrieval by support vector machines," *IEEE Trans. Neural Netw.*, pp. 209–215, 2003.

[11] K. Muller, S. Mika, G. Ratsch, K. Tsukada, and B. Scholkopf, "An introduction to kernel-based learning algorithms," *IEEE Trans. Neural Netw.*, pp. 181–201, 2001.

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

[13] 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.

[14] 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," *Proc. INTERSPEECH*, 2013.

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

[16] A. Banerjee and S. Basu, "Topic models over text streams: A study of batch and online unsupervised learning," *Proc. SIAM International Conference on Data Mining*, pp. 437–442, 2007.

[17] M. D. Hoffman, D. M. Blei, and F. Bach, "Online learning for latent dirichlet allocation," *In Adv. Neural Inf. Proc. Syst. 23*, pp. 856–864, 2010.

[18] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Machine Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[19] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical dirichlet processes," *Journal of the American Statistical Association*, vol. 101, pp. 1566–1581, 2004.

[20] E. Zavitsanos, G. Paliouras, and G. A. Vouros, "Non-parametric estimation of topic hierarchies from texts with hierarchical dirichlet processes," *The Journal of Machine Learning Research*, vol. 12, pp. 2749–2775, 2011.

[21] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, "An hdp-hmm for systems with state persistence," *Proc. the 25th International Conference on Machine Learning* (*ICML*), pp. 312–319, 2008.

[22] C. Wang, P. J, and D. M. Blei, "Online variational inference for the hierarchical dirichlet process," *Proc. the International Conference on Artificial Intelligence and Statistics* (*AISTATS*), pp. 752–760, 2011.

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

[24] S. Amari, "Natural gradient works efficiently in learning," *Neural computation*, vol. 10(2), pp. 251–276, 1998.

[25] M. Sato, "Online model selection based on the variational bayes," *Neural computation*, vol. 13(7), pp. 1649–1681, 2001.

[26] 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.

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