ACOUSTIC SCENE ANALYSIS FROM ACOUSTIC EVENT SEQUENCE WITH INTERMITTENT MISSING EVENT

Keisuke Imoto[†], *Nobutaka Ono*[‡†]

[†] The Graduate University for Advanced Studies, Kanagawa, Japan,
[‡] National Institute of Informatics, Tokyo, Japan.

ABSTRACT

We propose a novel method for analyzing acoustic scenes that can sophisticatedly estimate acoustic scenes from an acoustic event sequence with intermittent missing events. On the basis of the idea that acoustic events are temporally correlated, we model the transition of acoustic events using a hidden Markov model (HMM) and estimate missing acoustic events. Then, we incorporate the transition of acoustic events in a generative process of acoustic event sequence associated with the acoustic scenes based on acoustic topic model (ATM). Since the proposed method allows us to analyze acoustic scenes from acoustic event sequences while estimating missing acoustic events, we can estimate acoustic scenes successfully and restore missing acoustic events. Evaluation results indicate that the proposed method achieves an estimation accuracy for acoustic scenes comparable to that obtained when there is no missing data. Additionally, the proposed model can estimate acoustic events that are strongly correlated with acoustic scenes in an acoustic event sequence.

***Index Terms*—** Acoustic event detection (AED), Acoustic scene analysis, Missing data, Hidden Markov model (HMM)

1. INTRODUCTION

In recent years, the amount of media data such as sound and video has increased rapidly, and the realization of advanced media tagging, media summarizing, surveillance, and the monitoring of elderly people through the use of media data has attracted increasing attention. Acoustic event detection (AED) is an important technique for these applications, which extracts acoustic event information (e.g., environmental sounds, voice, music) or acoustic scene information (e.g., place, time, user activities) from acoustic signals, and considerable interest in AED has been expressed recently [1–6].

To analyze acoustic scenes from sounds, some methods focus on the fact that many acoustic scenes are characterized not by a single acoustic event but by a combination of multiple acoustic events [7, 8]; for instance, the acoustic scene "*cooking*" can be considered as a combination of acoustic events including "*cutting with a knife*," "*heating a skillet*," and "*running water*." Thus these methods model the abstract and complex phenomena of acoustic scenes by combining the characteristics of simple acoustic events. Focusing on the fact that when representing acoustic scenes using a combination of acoustic events, each acoustic scene has the sparsity in the acoustic event feature space, Lee *et al.* [9] proposed an efficient and effective acoustic scene modeling method that uses the low-rank approximation of an acoustic event feature space. However, this method is subject to overfitting of the input data in cases where the acoustic scenes are analyzed using a small dataset. To efficiently model the such complex phenomena of acoustic scenes with higher generalization performance, Kim *et al.* [10, 11] and Imoto *et al.* [12, 13] proposed generative probabilistic models of acoustic event sequences (consisting of multiple acoustic events) associated with acoustic scenes; these models are called acoustic topic models (ATMs). In ATMs, by introducing prior distributions of the parameters of acoustic events and scenes, overfitting input data can be avoided and ATMs can achieve generalization ability.

On the other hand, a large amount of multimedia data is now recorded by the general public, and it often has intermittent missing parts caused by wind noise, saturation of the sound pressure level, packet loss in data transmission, or the observation of a unknown acoustic events or those irrelevant to the acoustic scene. However, since a conventional ATM cannot take into account these missing acoustic events, it must exclude them, inevitable leading to performance degradation in the analysis of acoustic scenes.

To address this problem, we propose a novel method for estimating missing acoustic events and jointly analyzing acoustic scenes. In the proposed method, we focus on the temporal transition of acoustic events and estimate missing acoustic events from the temporally surrounding acoustic events based on a hidden Markov model (HMM) [14]. Then, we combine a conventional ATM and the HMM to analyze the transition of acoustic events and model the entire generative process of acoustic event sequences. Using this model, we can analyze acoustic scenes using not only observed acoustic events but also restored missing events, and we can also expect an improvement in the performance of acoustic scene analysis.

The rest of this paper is structured as follows. In Section 2, we introduce the ideas in the proposed model and formulate the model. In Section 3, we describe the parameter estimation method employed in the proposed model. In Section 4, we present and discuss experimental results and Section 5 concludes this paper.

2. ACOUSTIC TOPIC MODEL CONSIDERING TEMPORAL TRANSITION OF ACOUSTIC EVENTS

Considering that each sound recording consists of some acoustic scenes and that the generative probabilities of acoustic events vary according to the acoustic scenes, we can assume a generative model of an acoustic event sequence associated with sound clips and acoustic scenes. In a conventional ATM [10], it is assumed that sound recordings can be modeled as a discrete probability distribution of acoustic scenes and that acoustic scenes can also be modeled as a distribution of acoustic events. The whole generative process of the ATM can be written as

1. **Iterate # acoustic topics**
   Choose $\boldsymbol{\phi}_t \sim Dirichlet(\beta)$

Iterate # acoustic event sequences

2. Choose $\boldsymbol{\theta}_s \sim Dirichlet(\alpha)$

   Iterate # events in each acoustic event sequence

3. Choose $z_i \mid \boldsymbol{\theta}_s \sim Discrete(\boldsymbol{\theta}_s)$

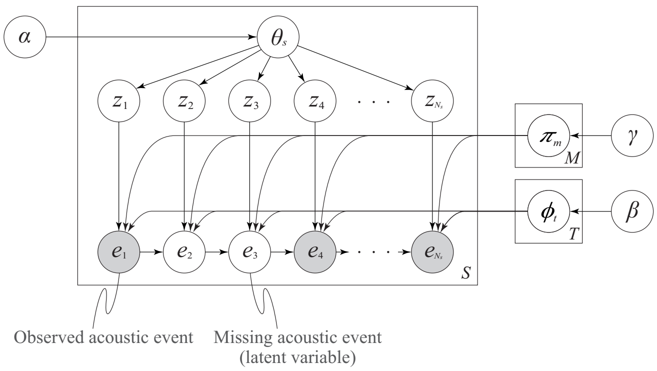

Fig. 1. Graphical model representation of acoustic event HMM-ATM

4. Choose $e_i \mid \boldsymbol{\phi}_{z_i}, z_i \;\sim\; Discrete(\boldsymbol{\phi}_{z_i})$.

The definition of each symbol is shown in Table 1. In this model, the latent variable $\boldsymbol{z}_i$ used to capture the latent structure in the acoustic scene (called the acoustic topic) is introduced, and then the relations between acoustic scenes and combinations of acoustic events are modeled via acoustic topics.

In a conventional ATM, since the missing acoustic events are not considered, missing acoustic events must be excluded before analyzing acoustic scenes. Moreover, since a conventional ATM assumes sparsity in terms of the type and number of acoustic events in each scene and represents acoustic topics using a small numbers of dimensions, missing acoustic events lead to a significant loss of information for analyzing acoustic scenes, and consequently, significant degradation of the performance of acoustic scene analysis.

To solve this problem, we focus on the fact that successive acoustic events in a short range are strongly associated, and therefore, each acoustic event can be estimated using surrounding events. Considering that the transition of acoustic events in a short range can be modeled using an HMM based on a simple Markov process, we can model a generative process of acoustic event sequences that is associated not only with sound clips/acoustic scenes but also with transition probabilities of acoustic events, as shown in Fig. 1, and the following generative process:

1. **Iterate # acoustic topics**
   Choose $\boldsymbol{\phi}_t \quad \sim Dirichlet(\beta)$
2. **Iterate # types of acoustic events**
   Choose $\boldsymbol{\pi}_m \quad \sim Dirichlet(\gamma)$

Iterate # acoustic event sequences

3. Choose $\boldsymbol{\theta}_s \quad \sim Dirichlet(\alpha)$

Iterate # events in each acoustic event sequence

4. Choose $z_i \mid \boldsymbol{\theta}_s \quad \sim Discrete(\boldsymbol{\theta}_s)$
5. Choose $e_i \mid \boldsymbol{\phi}_{z_i}, z_i, \pi_{e_{i-1}} \;\sim Discrete(\boldsymbol{\phi}_{z_i}), Discrete(\boldsymbol{\pi}_{e_{i-1}})$

where the proposed model represents missing acoustic events as latent variables, similar to acoustic topics. We call this generative model the "acoustic event HMM-ATM". In the acoustic event HMM-ATM, the generation of each acoustic event is based on the product of an acoustic event distribution $\boldsymbol{\phi}_{z_i=t}$ associated with an acoustic topic t and an acoustic event transition probability $\boldsymbol{\pi}_{e_i=m}$, both of which have Dirichlet priors. The remainder of the generative process is similar to that of a conventional ATM.

Table 1. Definition of symbols in generative probability model

Symbol	Definition
S	# acoustic event sequences (sound recordings)
T	# classes of acoustic topics
M	# classes of acoustic events
$N_{\boldsymbol{e}_s}$	# acoustic events in acoustic event sequence $\boldsymbol{e}_s$
t	Class index of acoustic topic
m	Class index of acoustic event
i	Order index of acoustic event in each acoustic event sequence
$\mathcal{S}$	Acoustic event sequence set
$\boldsymbol{z}$	Acoustic topics (latent variables)
$\boldsymbol{e}_s$	sth acoustic event sequence
$\tilde{\boldsymbol{e}}_s$	Missing acoustic event class in $\boldsymbol{e}_s$
$\boldsymbol{\theta}_s$	Acoustic topic distribution of $\boldsymbol{e}_s$
θ_t^s	Occurrence probability of acoustic topic t in $\boldsymbol{e}_s$
$\boldsymbol{\phi}_t$	Acoustic event distribution of acoustic topic t
ϕ_m^t	Occurrence probability of acoustic event m in acoustic topic t
$\boldsymbol{\pi}_m$	transition probabilistic distribution of acoustic event for acoustic event m
$\pi_{m^+}^{m^-}, \pi_{e_i}^{e_{i-1}}$	Transition probability of acoustic event from m^- to m^+
α, β, γ	Hyperparameter for Dirichlet distribution
$n_t^s, n_m^t, n_{m^+}^{m^-}$	# acoustic event assigned to acoustic topic t in $\boldsymbol{e}_s$, et cetera
$n_\cdot^s, n_\cdot^t, n_\cdot^{m^-}$	# acoustic event in $\boldsymbol{e}_s$, et cetera
$\backslash s, i$	Exclude ith acoustic event in $\boldsymbol{e}_s$
$\mathcal{D}(\cdot)$	Dirichlet distribution
$\Gamma(\cdot)$	Gamma distribution

Additionally, the generative probability of all acoustic event sequences $\mathcal{S}$ in a dataset can be represented as follows:

$$
\begin{aligned}
p(\mathcal{S}) &= \prod_{s=1}^{S} \prod_{i=1}^{N_{\boldsymbol{e}_s}} p(\boldsymbol{e}_i | \boldsymbol{\theta}_s, \boldsymbol{\phi}_t, \boldsymbol{\pi}_m; \alpha, \beta, \gamma, \tilde{\boldsymbol{e}}_s) \\
&= \prod_{s=1}^{S} \prod_{i=1}^{N_{\boldsymbol{e}_s}} \sum_{t=1}^{T} \sum_{\tilde{\boldsymbol{e}}_s} \mathcal{D}(\boldsymbol{\theta}_s; \alpha) \mathcal{D}(\boldsymbol{\phi}_t; \beta) \mathcal{D}(\boldsymbol{\pi}_m; \gamma) \\
&\qquad \cdot p(z_i | \boldsymbol{\theta}_s) p(e_i | \boldsymbol{\phi}_t, \boldsymbol{\pi}_m, z_i, \tilde{\boldsymbol{e}}_s) \\
&= \prod_{i=1}^{N_{\boldsymbol{e}_s}} \prod_{s=1}^{S} \frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T} \prod_{t=1}^{T} (\theta_t^s)^{\alpha-1+n_t^s} \cdot \prod_{t=1}^{T} \frac{\Gamma(M\beta)}{\Gamma(\beta)^M} \prod_{m=1}^{M} (\phi_m^t)^{\beta-1+n_m^t} \\
&\qquad \cdot \prod_{m^-=1}^{M} \frac{\Gamma(M\gamma)}{\Gamma(\gamma)^M} \prod_{m^+=1}^{M} (\pi_{m^+}^{m^-})^{\gamma-1+n_{m^+}^{m^-}} \qquad (1)
\end{aligned}
$$

To estimate missing acoustic events and topics from acoustic event sequences, it is necessary to infer model parameters that maximize their posterior distributions for given acoustic event sequences.

3. PARAMETER INFERENCE FOR ACOUSTIC EVENT HMM-ATM

Bayesian inference allows us to estimate the posterior distribution of model parameters in the acoustic event HMM-ATM, and in this paper we introduces an estimation method for the model parameters based on collapsed Gibbs sampling (CGS) [15,16]. This method iteratively samples and updates the latent variables for missing acoustic events and topics in accordance with a conditional posterior distribution for all acoustic events in given acoustic event sequences, which is not involving the updated acoustic events and topics. The sampling is repeated until the iterative update converges and then the posterior distributions of the acoustic topic, event, and event transition $\boldsymbol{\theta}, \boldsymbol{\phi}$ and $\boldsymbol{\pi}$, respectively, are estimated from the inferred latent variables. In the acoustic event HMM-ATM, since the posterior distribution depends on whether or not an acoustic event is missing, we must apply a different update to each case. We discuss the update for each case in detail below.

3.1. Update of latent variables for case of no missing event

When no acoustic events are missing, we only have to sample each acoustic topic in each update. In this case, the update for the ith acoustic topic in $\boldsymbol{e}_s$ is given as the posterior distribution $p(z_{s,i}|\boldsymbol{z}_{\setminus s,i}, \boldsymbol{e})$ of the ith acoustic topic given the assignments of all acoustic events and acoustic topics except for the ith acoustic event and topic. Since this update is the same as that in a conventional ATM [10], we simply give the formula for $p(z_{s,i}|\boldsymbol{z}_{\setminus s,i}, \boldsymbol{e})$ without its derivation.

$$p(z_{s,i}|\boldsymbol{z}_{\setminus s,i}, \boldsymbol{e}) \propto \frac{n^{t}_{(\setminus s,i),m} + \beta}{n^{t}_{(\setminus s,i),\cdot} + M\beta} \cdot (n^{s}_{(\setminus s,i),t} + \alpha) \qquad (2)$$

3.2. Update of latent variables for case of missing event

When an acoustic event is missing, we must sample the missing acoustic event and acoustic topic in each update. In this case, the update for the ith acoustic event and topic in $\boldsymbol{e}_s$ is given as the posterior distribution $p(z_{s,i}, e_{s,i}|\boldsymbol{z}_{\setminus s,i}, \boldsymbol{e}_{\setminus s,i})$ of the ith acoustic event and topic given the assignments of all acoustic events and acoustic topics except for the ith acoustic event and topic. Representing $p(z_{s,i}, e_{s,i}|\boldsymbol{z}_{\setminus s,i}, \boldsymbol{e}_{\setminus s,i})$ as the components separately related to $\boldsymbol{\theta}$, $\boldsymbol{\phi}$, and $\boldsymbol{\pi}$, the posterior distribution can be written as

$$\begin{aligned} & p(e_{s,i}, z_{s,i}|\boldsymbol{e}_{\setminus s,i}, \boldsymbol{z}_{\setminus s,i}) \\ &= p(z_{s,i}|\boldsymbol{e}_{\setminus s,i}, \boldsymbol{z}_{\setminus s,i}, e_{s,i}) p(e_{s,i}|\boldsymbol{e}_{\setminus s,i}, \boldsymbol{z}_{\setminus s,i}) \\ &= \frac{p(\boldsymbol{e}|\boldsymbol{z})p(\boldsymbol{z})}{p(\boldsymbol{e}|\boldsymbol{z}_{\setminus s,i})p(\boldsymbol{z}_{\setminus s,i})} \cdot p(e_{s,i}|\boldsymbol{e}_{\setminus s,i}, \boldsymbol{z}_{\setminus s,i}) \\ &= \frac{p(\boldsymbol{e}|\boldsymbol{z})p(\boldsymbol{z})}{p(e_{s,i}|\boldsymbol{z}_{\setminus s,i})p(\boldsymbol{e}_{\setminus s,i}|\boldsymbol{z}_{\setminus s,i})p(\boldsymbol{z}_{\setminus s,i})} \cdot p(e_{s,i}|\boldsymbol{e}_{\setminus s,i}, \boldsymbol{z}_{\setminus s,i}). \quad (3) \end{aligned}$$

where, considering that $e_{s,i}$ is independent of $\boldsymbol{z}_{\setminus s,i}$, we can represent $p(e_{s,i}\,|\,\boldsymbol{z}_{\setminus s,i})$ as $p(e_{s,i})$ and $p(e_{s,i}\,|\,\boldsymbol{e}_{\setminus s,i}, \boldsymbol{z}_{\setminus s,i})$ as $p(e_{s,i}\,|\,\boldsymbol{e}_{\setminus s,i})$. Moreover, since $p(e_{s,i})$ is independent of $p(z_{s,i}|\boldsymbol{z}_{\setminus s,i}, \boldsymbol{e})$, we can consider that $p(e_{s,i})$ is constant. As above, we can express the posterior $p(z_{s,i}, e_{s,i}|\boldsymbol{z}_{\setminus s,i}, \boldsymbol{e}_{\setminus s,i})$ as the product of the contributions of the distributions $p(\boldsymbol{e}|\boldsymbol{z}), p(\boldsymbol{z})$, and $p(\boldsymbol{e})$.

$$\begin{aligned} p(e_{s,i}, z_{s,i}|\boldsymbol{e}_{\setminus s,i}, \boldsymbol{z}_{\setminus s,i}) &\propto \frac{p(\boldsymbol{e}|\boldsymbol{z})p(\boldsymbol{z})}{p(\boldsymbol{e}_{\setminus s,i}|\boldsymbol{z}_{\setminus s,i})p(\boldsymbol{z}_{\setminus s,i})} \cdot p(e_{s,i}|\boldsymbol{e}_{\setminus s,i}) \\ &= \frac{p(\boldsymbol{e}|\boldsymbol{z})p(\boldsymbol{z})}{p(\boldsymbol{e}_{\setminus s,i}|\boldsymbol{z}_{\setminus s,i})p(\boldsymbol{z}_{\setminus s,i})} \cdot \frac{p(\boldsymbol{e}_{\setminus s,i}|e_i)p(e_{s,i})}{p(\boldsymbol{e}_{\setminus s,i})} \\ &= \frac{p(\boldsymbol{e}|\boldsymbol{z})}{p(\boldsymbol{e}_{\setminus s,i}|\boldsymbol{z}_{\setminus s,i})} \cdot \frac{p(\boldsymbol{z})}{p(\boldsymbol{z}_{\setminus s,i})} \cdot \frac{p(\boldsymbol{e})}{p(\boldsymbol{e}_{\setminus s,i})} \qquad (4) \end{aligned}$$

$\frac{p(\boldsymbol{e}|\boldsymbol{z})}{p(\boldsymbol{e}_{\setminus s,i}|\boldsymbol{z}_{\setminus s,i})} \cdot \frac{p(\boldsymbol{z})}{p(\boldsymbol{z}_{\setminus s,i})}$ in Eq. (4) corresponds to the posterior distribution of the acoustic topic in the case outlined Section 3.1, and therefore, each part of this product can be written as follows:

$$\frac{p(\boldsymbol{e}|\boldsymbol{z})}{p(\boldsymbol{e}_{\setminus s,i}|\boldsymbol{z}_{\setminus s,i})} = \frac{n^{t}_{(\setminus s,i),m} + \beta}{n^{t}_{(\setminus s,i),\cdot} + M\beta}, \qquad (5)$$

$$\frac{p(\boldsymbol{z})}{p(\boldsymbol{z}_{\setminus s,i})} = \frac{n^{s}_{(\setminus s,i),t} + \alpha}{n^{s}_{(\setminus s,i),\cdot} + T\alpha}. \qquad (6)$$

Then, $\frac{p(\boldsymbol{e})}{p(\boldsymbol{e}_{\setminus s,i})}$ can be similarly obtained as

$$\begin{aligned} \frac{p(\boldsymbol{e})}{p(\boldsymbol{e}_{\setminus s,i})} &= \frac{n^{e_{s,i-1}}_{(\setminus s,i),e_{s,i}} + \gamma}{n^{e_{s,i-1}}_{(\setminus s,i),\cdot} + M\gamma} \\ &\cdot \frac{n^{e_{s,i}}_{(\setminus s,i),e_{s,i+1}} + \delta(e_{s,i-1} = e_{s,i}) \cdot \delta(e_{s,i} = e_{s,i+1}) + \gamma}{n^{e_{s,i}}_{(\setminus s,i),\cdot} + \delta(e_{s,i-1} = e_{s,i}) + M\gamma}, \quad (7) \end{aligned}$$

where $n^{e_{s,i-1}}_{(\setminus s,i),e_{s,i}}$ is the number of acoustic events in the transition from acoustic event $e_{s,i-1}$ to acoustic event $e_{s,i}$ in all acoustic events except for transitions $e_{s,i-1} \to e_{s,i}$ and $e_{s,i} \to e_{s,i+1}$. $\delta(e_{s,i-1} = e_{s,i})$ is the Kronecker delta function, which is 1 if $e_{s,i-1} = e_{s,i}$, and 0 otherwise. Finally, substituting Eqs. (5)–(7) into Eq. (4), the update for the missing case can be obtained as

$$\begin{aligned} p(e_{s,i}, z_{s,i}|\boldsymbol{e}_{\setminus s,i}, \boldsymbol{z}_{\setminus s,i}) &\propto (n^{s}_{(\setminus s,i),t} + \alpha) \cdot \frac{n^{t}_{(\setminus s,i),m} + \beta}{n^{t}_{(\setminus s,i),\cdot} + M\beta} \\ &\cdot \frac{(n^{e_{s,i-1}}_{(\setminus s,i),e_{s,i}} + \gamma) \cdot \{n^{e_{s,i}}_{(\setminus s,i),e_{s,i+1}} + \delta(e_{s,i-1} = e_{s,i}) \cdot \delta(e_{s,i} = e_{s,i+1}) + \gamma\}}{n^{e_{s,i}}_{(\setminus s,i),\cdot} + \delta(e_{s,i-1} = e_{s,i}) + M\gamma}. \end{aligned} \qquad (8)$$

3.3. Posterior distribution of parameters

Given the updates for the acoustic event HMM-ATM, the posterior distributions of parameters of the generative distributions can be estimated through the assignment of sufficiently updated latent variables sampled using Eqs. (2) and (8). In practice, the parameters of the generative distributions can be approximated as the following means of the distributions of multiple samples:

$$\overline{\theta}^{s}_{t} = \frac{1}{N_G} \sum_{j=1}^{N_G} \left\{ \frac{\sum_{N_{\boldsymbol{e}_s}} \hat{z}_{s,i,t,j} + \alpha}{\sum_{N_{\boldsymbol{e}_s}} \sum_{t} \hat{z}_{s,i,t,j} + T\alpha} \right\}, \qquad (9)$$

$$\overline{\phi}^{t}_{m} = \frac{1}{N_G} \sum_{j=1}^{N_G} \left\{ \frac{\sum_{s} \sum_{N_{\boldsymbol{e}_s}} \hat{z}_{s,i,t,j}\, \hat{e}_{s,i,m,j} + \beta}{\sum_{s} \sum_{N_{\boldsymbol{e}_s}} \sum_{m} \hat{z}_{s,i,t,j}\, \hat{e}_{s,i,m,j} + M\beta} \right\}, \qquad (10)$$

$$\overline{\pi}^{m^-}_{m^+} = \frac{1}{N_G} \sum_{j=1}^{N_G} \left\{ \frac{\sum_{s} \sum_{N_{\boldsymbol{e}_s}} \hat{e}_{s,i,m^-,j} \cdot \hat{e}_{s,i+1,m^+,j} + \gamma}{\sum_{s} \sum_{N_{\boldsymbol{e}_s}} \sum_{m^+} \hat{e}_{s,i,m^-,j} \cdot \hat{e}_{s,i+1,m^+,j} + M\gamma} \right\}, (11)$$

where N_G is the number of samplings and $\hat{\boldsymbol{z}}_{s,i,t,j}$ and $\hat{\mathbf{e}}_{s,i,t,j}$ are the sampled acoustic topic and event in the jth sampling, respectively, which are given a value of 1 if the acoustic topic or event index is t or m, and value of 0 otherwise.

Table 2. Experimental conditions

Sampling rate / quantization	16 kHz / 16 bits
Frame size	512
Acoustical feature	12-dimensional MFCCs
Acoustic event size	8 - 512
Acoustic topic size	20
Hyperparameter α / β	3.33 / 0.1
Hyperparameter γ	0.5

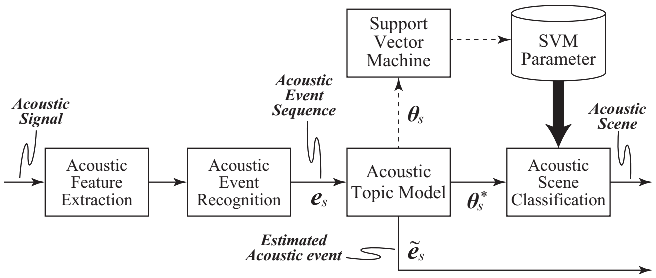

Fig. 2. Acoustic scene classification and acoustic event estimation system

4. EVALUATION

4.1. Experimental conditions

We evaluated the performance of the acoustic event HMM-ATM by using a real environmental sound dataset recorded in a house. The dataset contains 11,105 sounds that involve nine categories of acoustic scenes: *"chatting," "cooking," "eating dinner," "operating a PC," "reading a newspaper," "vacuuming," "walking," "washing dishes," and "watching TV."* These sounds are separated into 9,802 sounds for use as learning model parameters for each acoustic scene and 1,303 sounds for evaluation, then we arbitrarily create missing parts in only the test data.

To evaluate the performance of acoustic scene analysis and acoustic event estimation, we calculated the estimation accuracy of acoustic scenes and events using the acoustic scene classification and event estimation system shown in Fig. 2. The evaluation system first calculates acoustic feature vectors of input acoustic signals frame by frame and models acoustic events by Gaussian mixture model (GMM) clustering. For our evaluation, we define each Gaussian component modeled by the GMM as a single acoustic event. After recognizing acoustic events, we create missing parts in the acoustic event sequences, and then, we input them to the acoustic event HMM-ATM and estimate latent variables and parameters of generative distributions. In practical use, however, we must detect missing acoustic events by applying a clipping detection system or wind noise detection system [17] to acoustic signals. In this experiment, we artificially create acoustic event sequences in which every second (50%) acoustic event is missing, and this rate of missing events can be regarded as being higher than that expected in practical use. Since the parameters $\boldsymbol{\theta}_s$ for the acoustic topic distribution are similar if acoustic event sequences are generated from the same acoustic scene, we can estimate acoustic scenes by comparing $\boldsymbol{\theta}_s$ for learning data and $\boldsymbol{\theta}_s^*$ for evaluation data. In our evaluation, we compare the similarity of $\boldsymbol{\theta}_s$ and $\boldsymbol{\theta}_s^*$ by using the multiclass support vector machine (SVM) based on the radial basis function (RBF) kernel [18, 19]. The other experimental conditions are listed in Table 2.

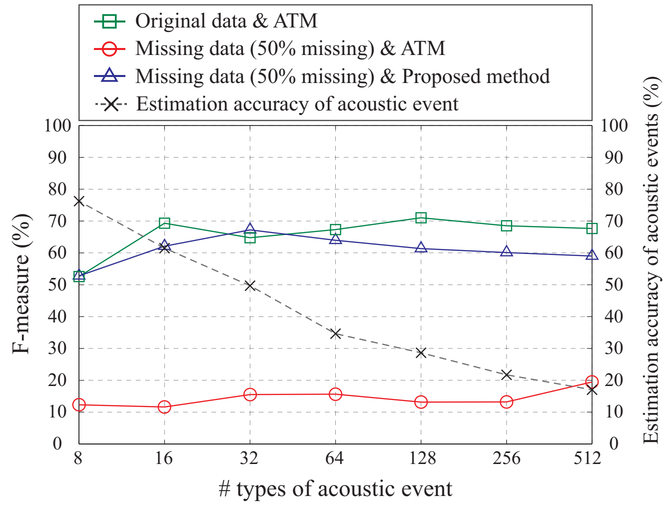

Fig. 3. Classification accuracy of acoustic scene and estimation accuracy of missing acoustic events

4.2. Experimental results

The average estimation accuracy in nine acoustic scenes and the estimation accuracy of missing acoustic events are shown in Fig. 3. For comparison with the proposed model, Fig. 3 also shows the results for a conventional ATM. This figure shows that while the result of the conventional ATM with 50% missing data decreases considerably (19.5%) compared with the result with all the original data (72.4%), the acoustic event HMM-ATM achieves an estimation accuracy of 67.2%, which is close to the result obtained using the conventional ATM with all the original data (72.4%). These results indicate that while the conventional ATM structure may collapse when 50% of acoustic event are missing, the acoustic event HMM-ATM can reconstruct the structure and estimate acoustic scenes with reasonable accuracy.

The accuracy of acoustic event estimation is 76.3% when we use a small number of types of acoustic events (8 types); however, the estimation accuracy obtained using 512 types of acoustic events is less than 20%. On the other hand, even when we use a large number of types of acoustic events, the estimation accuracy of acoustic scenes using the proposed model does not decrease. This suggests that even if missing acoustic events are not estimated correctly, the proposed model can estimate acoustic events that are strongly correlated with the acoustic scene in the acoustic event sequence.

5. CONCLUSION

To estimate acoustic scenes and missing acoustic events, we proposed a novel acoustic topic model (ATM) that considers the generative process of an acoustic event sequence including the temporal transition of acoustic events. In the proposed model, the temporal transition of acoustic events is modeled by an HMM and is incorporated into a conventional ATM. Moreover, we introduced a parameter estimation method for the proposed method based on collapsed Gibbs sampling. Evaluation results for the model performance indicate that the proposed method achieves an estimation accuracy of acoustic scenes comparable to that obtained when there is no missing data. Additionally, the proposed model can estimate acoustic events that are strongly correlated with acoustic scenes in an acoustic event sequence.

6. REFERENCES

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

[2] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, "CLEAR evaluation of acoustic event detection and classification systems," *Springer, Berlin and Heidelberg*, pp. 311–322, 2007.

[3] Y. Peng, C. Lin, M. Sun, and K. Tsai, "Healthcare audio event classification using hidden Markov models and hierarchical hidden markov models," *in Proc. IEEE Int. Conf. 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. IEEE International Conference on Multimedia and Expo 2005 (ICME 2005)*, 2005.

[5] 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.*, pp. 321–329, 2006.

[6] 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. IEEE Int. Conf. on Acoustics, Speech and Signal Processing. (ICASSP), 2013*, pp. 3163–3167, 2013.

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

[8] M. A. M. Ahaikh, M. K. I. Molla, and K. Hirose, "Automatic life-logging: A novel approach to sense real-world activities by environmental sound cues and common sense," *Proc. of the 11th International Conference on Computer and Information Technology (ICCIT 2008)*, pp. 294–299, 2008.

[9] K. Lee and D. P. W. Ellis, "Audio-based semantic concept classification for consumer video," *IEEE Trans. Audio Speech Lang. Process.*, pp. 1406–1416, 2010.

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

[11] S. Kim, P. G. Georgiou, S. Narayanan, and S. Sundaram, "Supervised acoustic topic model for unstructured audio information retrieval," *in Proc. 2010 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2010.

[12] 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.

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

[14] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *in Proc, of the IEEE*, vol. 77, pp. 257–286, 1989.

[15] 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.

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

[17] C. M. Nelke, N. Nawroth, M. Jeub, C. Beaugeant, and P. Vary, "Single microphone wind noise reduction using techniques of artificial bandwidth extension," *in Proc. 19th European Signal Processing Conference (EUSIPCO 2012)*, pp. 2328–2332, 2012.

[18] V. Franc and H. Vaclav, "Multi-class support vector machine," *in Proc. 16th Int. Conf. on Pattern Recognition 2002*, pp. 236–239, 2002.

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