Acoustic Topic Model for Scene Analysis With Intermittently Missing Observations

Keisuke Imoto, *Member, IEEE*, and Nobutaka Ono, *Senior Member, IEEE*

***Abstract*—We propose a sophisticated method of acoustic scene analysis with intermittently missing observations, which analyzes acoustic scenes and restores missing observations simultaneously on the basis of the temporal correlation between acoustic words. One effective strategy for analyzing acoustic scenes is to characterize them as a combination of acoustic words. An acoustic topic model (ATM) is one of the techniques, which models the process generating multiple acoustic words. Here, an acoustic word corresponds to a sound category, while it has a homogenous time duration and is defined time frame by time frame. In the ATM, it is assumed that all acoustic words are observed, and therefore, it cannot be applied if any acoustic observations are missing. However, acoustic observations may sometimes be missing because of poor recording conditions, transmission loss, or privacy reasons. In the proposed method, focusing on the fact that acoustic words are temporally correlated, we consider the transition of acoustic words in two ways: First, by modeling the temporal transition of acoustic words directly using a Markov process and finally, by modeling the temporal transition of hidden states that generate acoustic words using a hidden Markov model. We then incorporate each transition model in a process generating acoustic words based on the ATM. The proposed method allows us to analyze acoustic scenes from acoustic words by restoring missing acoustic words. In our experiments, the proposed method exhibited a classification accuracy of acoustic scenes close to that for the case of no missing observations even when 50% of the observations were missing. Moreover, the model considering the hidden-state transition can classify acoustic scenes more accurately than the model considering the acoustic word transition directly.**

***Index Terms*—Acoustic scene analysis, missing data analysis, Markov model, acoustic topic model.**

I. Introduction

In RECENT years, the use of acoustic sensors, such as in smartphones, wearable devices, and surveillance cameras, has rapidly increased. Utilizing these acoustic sensors, more sounds are being recorded and analyzed to realize useful applications such as advanced media retrieval [1]–[4], automatic surveillance [5]–[8], monitoring of elderly people [9], [10], automatic life logging [11], [12], and medical systems based on sound [13].

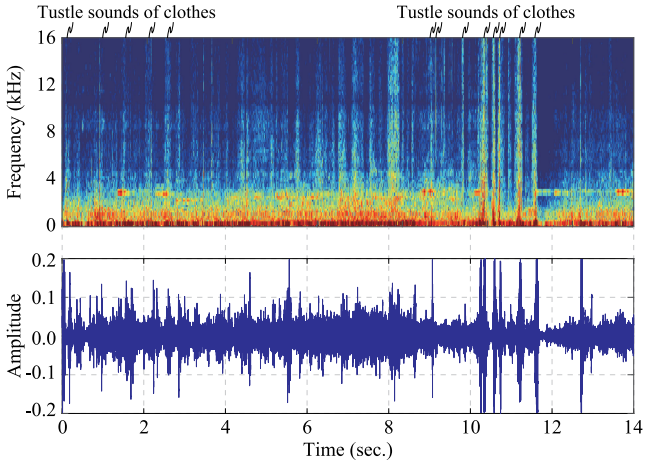
One important technique used in these applications is acoustic event detection (AED), which classifies or identifies types of sounds, such as footsteps, running water, music, or voices from short-term observations [14]–[16]. Another important technique is acoustic scene analysis (ASA), which classifies or identifies acoustic scenes from relatively long-term observations, where the acoustic scene is an environment or situation in which sounds are produced such as cooking, watching TV, an emergency, on the bus, or a meeting. In ASA, many approaches based on machine learning techniques are proposed [17]. For instance, Eronen *et al.* [18] have proposed a method using the mel-frequency cepstral coefficients (MFCCs) as an input feature and Gaussian mixture model (GMM) as a classifier. Han *et al.* [19] and Jallet *et al.* [20] have proposed methods using mel-spectrogram as input features, and the convolutional neural network (CNN) or recurrent convolutional neural network (RCNN) as classifiers. Another reliable approach to ASA is to characterize an acoustic scene as a combination of acoustic events in a long-term sound; for instance, an acoustic scene involving cooking can be characterized by a combination of acoustic events including cutting with a knife, heating a skillet, running water. To utilize this idea, a long-term sound is preliminarily represented as a temporal sequence of acoustic events using a Gaussian mixture model (GMM) or hidden Markov model (HMM), and then machine learning techniques are applied to analyze the acoustic scenes. Heittola *et al.* [21] and Guo and Li [22] proposed acoustic scene classification methods based on a histogram of acoustic events and a support vector machine (SVM) [23]. In these methods, the feature of each acoustic scene is represented by a histogram of acoustic events, and then a multiclass classifier of acoustic scenes is learned by using a multiclass SVM. Considering that the generative probabilities of the type of sounds vary according to the acoustic scene, Lee and Ellis [24] proposed ASA based on a generative model of acoustic words. Here, an acoustic word is related to an acoustic event, while the acoustic word is defined time frame by time frame without time overlapping. On the basis of this generative model, they analyzed acoustic scenes through the maximum a posteriori (MAP) estimation of model parameters. However, since it is not easy to collect sufficient sounds that contain all possible acoustic scenes, in these methods overfitting is one of the most serious concerns. To solve this problem, Kim *et al.* [25] and Imoto and coworkers [26], [27] proposed Bayesian

Manuscript received March 6, 2018; revised June 21, 2018, September 6, 2018, and October 24, 2018; accepted October 24, 2018. Date of publication November 7, 2018; date of current version November 29, 2018. 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 associate editor coordinating the review of this manuscript and approving it for publication was Prof. Simon Doclo. *(Corresponding author: Keisuke Imoto.)*

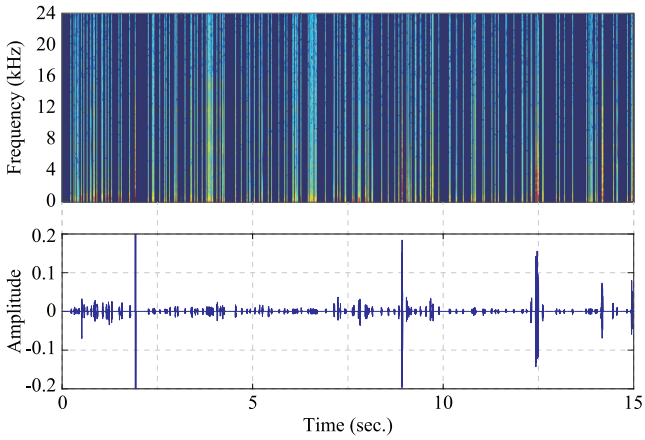
K. Imoto is with the School of Information Science and Engineering, Ritsumeikan University, Kusatsu 525-8577, Japan (e-mail: keisuke.imoto@ieee.org).

N. Ono is with the Graduate School of System Design, Tokyo Metropolitan University, Tokyo 192-0364, Japan (e-mail: onono@tmu.ac.jp).

Digital Object Identifier 10.1109/TASLP.2018.2879855



(a) Sound of a station platform recorded by a smartphone in a trouser pocket (lower) and its spectrogram (upper). The recorded sound includes the sounds of rustling clothes, which are unrelated to the acoustic scene.



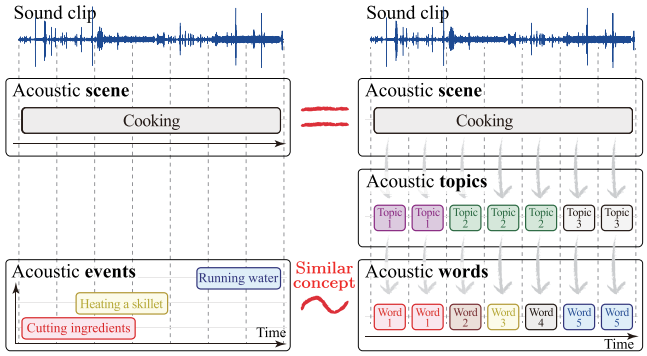
(b) Partially recorded cooking sound (lower) and its spectrogram (upper). To protect privacy, we sometimes need to analyze acoustic scenes with such an acoustic signal.

Fig. 1. Examples of observations with some parts missing.

generative probabilistic models of acoustic word sequences associated with acoustic scenes; these models are called acoustic topic models (ATMs). In ATMs, by introducing prior distributions of parameters of acoustic word sequences, the overfitting of input data can be avoided.

On the other hand, as shown in Fig. 1(a), sounds recorded with smart devices or surveillance cameras often have intermittent unreliable observations caused by wind noise, rustling sounds, clipping, or completely missing parts caused by packet loss in data transmission over the network. In addition, to protect privacy, continuous recording is sometimes not preferable, and therefore, we may need to analyze acoustic scenes from partially recorded sounds as shown in Fig. 1(b). The conventional ATMs cannot cope well with missing acoustic words, and simply applying the ATM technique while ignoring missing observations degrades the performance of acoustic scene analysis.

To address this problem, we propose novel methods for simultaneously analyzing acoustic scenes and estimating missing observations. In the proposed methods, we focus on the temporal correlation between acoustic words, and model it as processes generating acoustic word sequences with 1) an



(a) Acoustic scene and events in real world

(b) Generative process of supervised acoustic topic model (ATM)

Fig. 2. Relationship between acoustic scene, topic, word, and event.

acoustic word transition or 2) a transition of hidden-state generating acoustic words. Moreover, we model processes generating acoustic word sequences with supervised manners in acoustic scenes, which estimate model parameters using training data and acoustic scene labels unlike our previous work [28]. With these models, we can analyze acoustic scenes while restoring missing words, as a result of which, improved performances of acoustic scene analysis can be expected.

The rest of this paper is structured as follows. In the next section, we introduce some conventional methods for analyzing acoustic scenes on the basis of the ATM. In Section III, we introduce the concept of the proposed models and formulate them, then in Section IV, we describe the parameter estimation method employed in the proposed models. In Section V, we present and discuss experimental results and Section VI concludes this paper.

II. Conventional Methods for Acoustic Scene Analysis

A. Unsupervised Acoustic Topic Model

On the basis of the idea that the sound categories that occur depend on the acoustic scene, an ATM [25] models the relationship between an acoustic scene and acoustic words as a process that generates acoustic word sequences from acoustic scenes in an unsupervised manner. Here, the name ATM is derived from the term "topic model" in natural language processing [29], [30], which is an equivalent model to the ATM.

Let us consider a continuous acoustic signal of, for example, 1h length that is divided into sound clips of 10s length, which are the units of acoustic word sequences and are equivalent to the "documents" in the original topic model. As preprocessing of the ATM, each sound clip is represented by a temporal sequence of acoustic words defined time frame by time frame; the number of acoustic words in a sound clip $N_{\boldsymbol{w}_s}$ is equal to the number of time frames in the sound clip. Here, the relationship between the terms acoustic scene, topic, event, and word is shown in Fig. 2. Note that an acoustic event can be related to several different acoustic words, while the ATM handles these acoustic words as different ones. Then, the ATM assumes that the process generating acoustic word sequences can be represented by a hierarchical process that generates acoustic topics

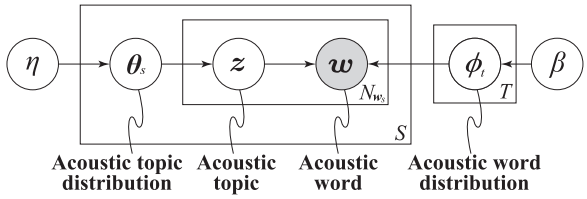


Fig. 3. Graphical model representation of ATM.

and words. Here, the acoustic topic is defined as the latent state time frame by time frame and is implicitly associated with the acoustic scene; thus, in the ATM, the acoustic scenes are indirectly characterized by the distribution of acoustic topics instead of the combination of acoustic words.

In particular, the process generating acoustic word sequences can be modeled as shown in Fig. 3 and Table II. Here, $\boldsymbol{\theta}_s$ represents the acoustic topic distribution that depends on an acoustic word sequence s, and the other symbols are defined in Table I. According to the generative process of the ATM, each acoustic topic distribution $\boldsymbol{\theta}_s$ varies from sound clip to sound clip, and in each sound clip, an acoustic topic $z_{s,i}$ is generated from the topic distribution $\boldsymbol{\theta}_s$ frame by frame. Then, each acoustic word $w_{s,i}$ is generated frame by frame from an acoustic word distribution $\boldsymbol{\phi}_t$ that depends on the acoustic topic t. These distributions $\boldsymbol{\theta}_s$ and $\boldsymbol{\phi}_t$ have Dirichlet priors; that is, hyperparameters η and β control the sparseness of the acoustic topic and word distribution, which prevents overfitting of the model to the given data. This simple ATM does not include any temporal constraints on the acoustic word sequence, and thus, it is assumed that the acoustic word sequence can be described as a "bag of acoustic words," which corresponds to the "bag of words" assumption in natural language processing [31].

To classify acoustic scenes using the ATM, Kim *et al.* [25] first estimated the distributions of acoustic topics for each acoustic scene, and then classified acoustic scenes on the basis of a the multiclass SVM that utilized acoustic scene labels and the parameters of acoustic topic distributions.

B. Supervised Acoustic Topic Model

To explicitly model the relationship between acoustic scenes and a word sequence and to classify acoustic scenes using the model itself, a supervised model for generating acoustic word sequences has been proposed [26]. We call this supervised model the supervised acoustic topic model (sATM). In the sATM, possible acoustic scene labels are preliminarily given to each acoustic word sequence explicitly, and an acoustic scene is sampled randomly from them in its generative process. Furthermore, this model assumes that each acoustic scene has a different acoustic topic distribution, and an acoustic topic is then sampled from its distribution. The other part of the generative process is the same as that in the unsupervised ATM. Thus, the generative process of the sATM is represented as shown in Table III.

When classifying acoustic scenes using the sATM, we first estimate the distributions of acoustic topics $\boldsymbol{\theta}_a$ and acoustic words $\boldsymbol{\phi}_t$ using a training dataset, and then we estimate the acoustic scene indicated by the test data by determining the acoustic scene with the highest posterior probability as follows.

$$\arg\max_a p(a|\boldsymbol{\theta}_a, \boldsymbol{\phi}_t, \boldsymbol{w}_s, \alpha, \beta) \tag{1}$$

TABLE I
DEFINITIONS OF SYMBOLS

Symbol	Definition
S	# acoustic word sequences (# sound clips)
A	# classes of acoustic scenes
T	# classes of acoustic topics
M	# classes of acoustic words
$N_{\boldsymbol{w}_s}$	# acoustic words in acoustic word sequence $\boldsymbol{w}_s$
s	Index of sound clip
a	Class index of acoustic scene
t	Class index of acoustic topic
m	Class index of acoustic word
i	Order index of acoustic word in each acoustic word sequence
a_s	Acoustic scene in sth sound clip
$z_{s,i}$	Acoustic topic at ith order in sth sound clip
$w_{s,i}$	Acoustic words at ith order in sth sound clip
$w_{s,i}^{obs}, w_{s,i}^{mis}$	Observed or missing acoustic word at ith order in sth sound clip
$\mathbf{a}_s$	Possible acoustic scenes in sth sound clip
$\boldsymbol{w}_s$	Acoustic word sequence in sth sound clip ($\boldsymbol{w}_s = \{w_{s,1}, \cdots, w_{s,N_{\boldsymbol{w}_s}}\}$)
$\boldsymbol{a}$	Collection of acoustic scenes ($\boldsymbol{a}=\{a_1, \cdots, a_S\}$)
$\boldsymbol{z}$	Collection of acoustic topics ($\boldsymbol{z} = \{z_{1,1}, \cdots, z_{s,i}, \cdots, z_{S,N_{\boldsymbol{w}_S}}\}$)
$\boldsymbol{w}$	Collection of acoustic words ($\boldsymbol{w} = \{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_S\}$)
$\theta_{a,t}$	Occurrence probability of acoustic topic t in acoustic scene a
$\phi_{t,m}$	Occurrence probability of acoustic word m in acoustic topic t
π_{m^-,m^+}	Transition probability of acoustic word from m^- to m^+
π_{t^-,t^+}	Transition probability of acoustic topic from t^- to t^+
$\boldsymbol{\theta}_s$	Acoustic topic distribution of acoustic word sequence (sound clip) s
$\boldsymbol{\theta}_a, \boldsymbol{\theta}_{a_s=a}$	Acoustic topic distribution of acoustic scene a ($\boldsymbol{\theta}_a = \{\theta_{a,1}, \theta_{a,2}, \cdots \theta_{a,T}\}$)
$\boldsymbol{\phi}_t, \boldsymbol{\phi}_{z_{s,i}=t}$	Acoustic word distribution of acoustic topic t ($\boldsymbol{\phi}_t = \{\phi_{t,1}, \phi_{t,2}, \cdots \phi_{t,M}\}$)
$\boldsymbol{\pi}_m, \boldsymbol{\pi}_{w_{s,i}=m}$	Transition probabilities of acoustic words from acoustic word m ($\boldsymbol{\pi}_m = \{\pi_{m,1}, \pi_{m,2}, \cdots \pi_{m,M}\}$)
$\boldsymbol{\pi}_t, \boldsymbol{\pi}_{z_{s,i}=t}$	Transition probabilities of acoustic topics from acoustic topic t ($\boldsymbol{\pi}_t = \{\pi_{t,1}, \pi_{t,2}, \cdots \pi_{t,T}\}$)
$\alpha, \beta, \gamma, \eta$	Hyperparameters for Dirichlet distribution
$n_t^{(a)}, n_m^{(t)}$	# acoustic words assigned to acoustic topic t in acoustic scene a, etc.
$n_{m^+}^{(m^-)}$	# instances in which acoustic words transit from m^- to m^+
$n_\cdot^{(a)}, n_\cdot^{(t)}$	# acoustic words in acoustic scene a, etc.
$n_\cdot^{(m^-)}$	# instances in which acoustic words transit from m^- to any acoustic event
$\backslash s, i$	Exclude ith acoustic word in $\boldsymbol{w}_s$
$\mathcal{D}(\cdot)$	Dirichlet distribution
$\Gamma(\cdot)$	Gamma function

TABLE II
GENERATIVE PROCESS OF ACOUSTIC WORD SEQUENCE IN ATM

for $t = 1$ to T **do**
  Choose $\boldsymbol{\phi}_t \sim \text{Dirichlet}(\beta)$
end for
for $s = 1$ to S **do**
  Choose $\boldsymbol{\theta}_s \sim \text{Dirichlet}(\eta)$
  for $i = 1$ to $N_{\boldsymbol{w}_s}$ **do**
    Choose $z_{s,i} \mid \boldsymbol{\theta}_s \sim \text{Categorical}(\boldsymbol{\theta}_s)$
    Choose $w_{s,i} \mid \boldsymbol{\phi}_{z_{s,i}=t}, z_{s,i} \sim \text{Categorical}(\boldsymbol{\phi}_{z_{s,i}=t})$
  end for
end for

TABLE III
GENERATIVE PROCESS OF ACOUSTIC WORD SEQUENCE IN SATM

A set of possible acoustic scenes $\mathbf{a}_s$ **is given,**
for $a = 1$ to A **do**
  Choose $\boldsymbol{\theta}_a \sim \text{Dirichlet}(\alpha)$
end for
for $t = 1$ to T **do**
  Choose $\boldsymbol{\phi}_t \sim \text{Dirichlet}(\beta)$
end for
for $s = 1$ to S **do**
  Choose $a_s \sim \text{Uniform}(\mathbf{a}_s)$
  for $i = 1$ to $N_{\boldsymbol{w}_s}$ **do**
    Choose $z_{s,i} \mid \boldsymbol{\theta}_{a_s=a}, a_s \sim \text{Categorical}(\boldsymbol{\theta}_{a_s=a})$
    Choose $w_{s,i} \mid \boldsymbol{\phi}_{z_{s,i}=t}, z_{s,i} \sim \text{Categorical}(\boldsymbol{\phi}_{z_{s,i}=t})$
  end for
end for

III. ACOUSTIC SCENE MODELING CONSIDERING TEMPORAL CONTINUITY OF ACOUSTIC WORDS

A. Motivation and Strategy

One limitation of the conventional methods for ASA is that they require the complete observations. However, because of wind noise, clipping, packet loss in data transmission, or privacy reasons, we often need to analyze acoustic scenes from sounds including missing parts. In this paper, we consider a situation in which some parts of observations are missing, i.e., we cannot observe some acoustic words completely, while we know which acoustic words are missing.

To overcome this limitation, we consider new ATMs with temporal dynamics, which focus on the fact that acoustic words in the short term are not independent of each other. Our motivation is to restore the missing words in a probabilistic manner by utilizing the temporal correlation of acoustic words, which is expected to contribute to improving the performance of acoustic scene analysis. That is, we model missing acoustic words as the latent states and estimate them in the parameter estimation. Specifically, we investigate the following two types of model.

- An acoustic scene model considering the temporal transition of acoustic words.
- An acoustic scene model considering the temporal transition of acoustic topics.

The former directly models the temporal dynamics of acoustic words, while the latter models it as a hidden structure such as an HMM.

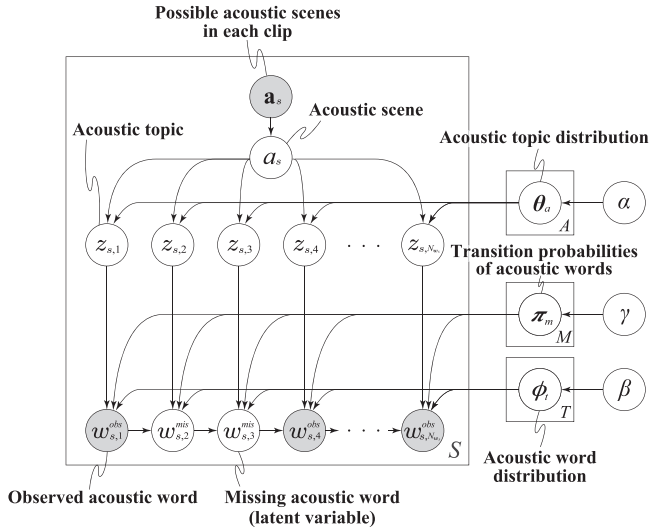


Fig. 4. Graphical representation of model generating acoustic topics and words considering acoustic word transition.

B. Acoustic Scene Modeling Considering Temporal Transition of Acoustic Words

We first discuss a model for generating acoustic word sequences that directly considers an acoustic word transition as shown in Fig. 4. This model is also based on a hierarchical process generating acoustic topics and words similarly to the sATM, while it differs from the conventional sATM in that the generation of each acoustic word is conditioned by not only the acoustic topic but also the previous acoustic word. In particular, we assume that an acoustic word transition can be modeled by using a simple Markov process, which is known as a promising method for modeling a dynamic correlation stochastically. Note that we assume that whether or not an acoustic word is observed is unrelated to the generative process of acoustic word sequences, and thus, it is considered when the model parameters are estimated.

The generative process of this model can be described as shown in Table IV, and we call this model the word-transition-based supervised acoustic topic model (word-transition sATM). Here, missing acoustic words are represented as latent variables, whereas observed acoustic words are represented as observed variables. This representation of missing acoustic words enables them to be estimated and restored using the proposed model. In the word-transition sATM, we assume that the generation of each acoustic word is represented by the product of an acoustic word occurrence probability $p(w_{s,i}|\boldsymbol{\phi}_{z_{s,i}=t})$ and transition probability $p(w_{s,i}|\boldsymbol{\pi}_{w_{s,i-1}=m})$, both of which are represented by the categorical distribution and have Dirichlet priors. Thus, the generative probability of all acoustic word sequences $\boldsymbol{w}$ can be represented as follows.

$$p(\boldsymbol{w}|\alpha,\beta,\gamma,\mathbf{a}_s) = \prod_{s=1}^{S}\prod_{i=1}^{N_{\boldsymbol{w}_s}}\sum_{\boldsymbol{a}}\sum_{\boldsymbol{z}}\sum_{\boldsymbol{m}} p(w_{s,i}|w_{s,i-1},z_{s,i},\alpha,\beta,\gamma,a_s) \cdot p(z_{s,i}|a_s,\alpha)p(a_s|\mathbf{a}_s)$$

TABLE IV
GENERATIVE PROCESS OF ACOUSTIC WORD SEQUENCE IN WORD-TRANSITION SATM

A set of possible acoustic scenes $\mathbf{a}_s$ is given,
for $a = 1$ to A **do**
  Choose $\boldsymbol{\theta}_a \sim \mathrm{Dirichlet}(\alpha)$
end for
for $t = 1$ to T **do**
  Choose $\boldsymbol{\phi}_t \sim \mathrm{Dirichlet}(\beta)$
end for
for $m = 1$ to M **do**
  Choose $\boldsymbol{\pi}_m \sim \mathrm{Dirichlet}(\gamma)$
end for
for $s = 1$ to S **do**
  Choose $a_s \sim \mathrm{Uniform}(\mathbf{a}_s)$
  for $i = 1$ to $N_{\boldsymbol{w}_s}$ **do**
    Choose $z_{s,i} \mid \boldsymbol{\theta}_{a_s=a}, a_s \sim \mathrm{Categorical}(\boldsymbol{\theta}_{a_s=a})$
    Choose $w_{s,i} \mid \boldsymbol{\phi}_{z_{s,i}=t}, \boldsymbol{\pi}_{w_{s,i-1}=m}, z_{s,i}, w_{s,i-1} \sim \mathrm{Categorical}(\boldsymbol{\phi}_{z_{s,i}=t}), \mathrm{Categorical}(\boldsymbol{\pi}_{w_{s,i-1}=m})$
  end for
end for

$$
\begin{aligned}
&= \prod_{s=1}^{S}\Bigg[p(a_s|\mathbf{a}_s) \sum_{\boldsymbol{a}} \int p(\boldsymbol{\theta}_a|a_s,\alpha) \prod_{i=1}^{N_{\boldsymbol{w}_s}} \\
&\quad \Bigg\{ \sum_{\boldsymbol{z}} p(z_{s,i}|\boldsymbol{\theta}_a) \int \mathcal{D}(\boldsymbol{\phi}_t|\beta) \\
&\quad \cdot \int \sum_{\boldsymbol{m}} \mathcal{D}(\boldsymbol{\pi}_m|\gamma) p(w_{s,i}|\boldsymbol{\phi}_t, \boldsymbol{\pi}_m, w_{s,i-1}, z_{s,i}) d\boldsymbol{\pi}_m \, d\boldsymbol{\phi}_t \Bigg\} d\boldsymbol{\theta}_a \Bigg] \\
&= \frac{1}{A} \prod_{s=1}^{S}\Bigg[\int \frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T} \prod_{i=1}^{N_{\boldsymbol{w}_s}} \Bigg\{ \prod_{t=1}^{T} \theta_{a,t}^{\alpha-1+n_t^{(a)}} \\
&\qquad \int \frac{\Gamma(M\beta)}{\Gamma(\beta)^M} \prod_{m=1}^{M} \phi_{t,m}^{\beta-1+n_m^{(t)}} \\
&\quad \cdot \int \frac{\Gamma(M\gamma)}{\Gamma(\gamma)^M} \prod_{m^+=1}^{M} \pi_{m^-,m^+}^{\gamma-1+n_{m^+}^{(m^-)}} d\boldsymbol{\pi}_m \, d\boldsymbol{\phi}_t \Bigg\} d\boldsymbol{\theta}_a \Bigg] \qquad (2)
\end{aligned}
$$

When classifying acoustic scenes, it is necessary to infer model parameters that maximize their posterior probabilities in advance for the training dataset. Then, acoustic scenes and missing acoustic words are estimated in a similar manner to in the sATM,

$$\arg\max_{a, w_{s,i}} p(a, w_{s,i}|\boldsymbol{\theta}_a, \boldsymbol{\phi}_t, \boldsymbol{\pi}_m, \boldsymbol{w}_s, \alpha, \beta, \gamma). \qquad (3)$$

In this paper, we derive a parameter estimation method based on collapsed Gibbs sampling, as described in the next section.

C. Acoustic Scene Modeling Considering Temporal Transition of Acoustic Topics

In the second model, the temporal transition of acoustic topics is assumed as shown in Fig. 5. In this model, the temporal correlation between acoustic words is derived from the temporal transition of a hidden structure such as an HMM.

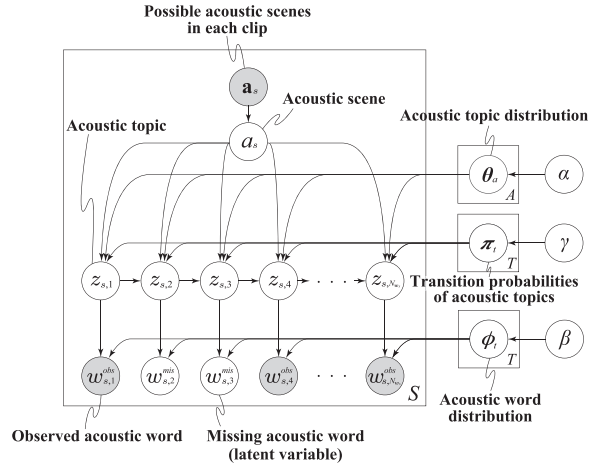


Fig. 5. Graphical representation of model generating acoustic topics and words considering acoustic topic transition.

TABLE V
GENERATIVE PROCESS OF ACOUSTIC WORD SEQUENCE IN TOPIC-TRANSITION SATM

A set of possible acoustic scenes $\mathbf{a}_s$ is given,
for $a = 1$ to A **do**
  Choose $\boldsymbol{\theta}_a \sim \mathrm{Dirichlet}(\alpha)$
end for
for $t = 1$ to T **do**
  Choose $\boldsymbol{\phi}_t \sim \mathrm{Dirichlet}(\beta)$
  Choose $\boldsymbol{\pi}_t \sim \mathrm{Dirichlet}(\gamma)$
end for
for $s = 1$ to S **do**
  Choose $a_s \sim \mathrm{Uniform}(\mathbf{a}_s)$
  for $i = 1$ to $N_{\boldsymbol{w}_s}$ **do**
    Choose $z_{s,i} \mid \boldsymbol{\theta}_{a_s=a}, \boldsymbol{\pi}_{z_{s,i-1}=t}, a_s, z_{s,i-1} \sim \mathrm{Categorical}(\boldsymbol{\theta}_{a_s=a}), \mathrm{Categorical}(\boldsymbol{\pi}_{z_{s,i-1}=t})$
    Choose $w_{s,i} \mid \boldsymbol{\phi}_{z_{s,i}=t}, z_{s,i} \sim \mathrm{Categorical}(\boldsymbol{\phi}_{z_{s,i}=m})$
  end for
end for

The generative process of this model is represented in Table V, and we call this model the topic-transition-based supervised acoustic topic model (topic-transition sATM). In the topic-transition sATM, we assume that the generationt of each acoustic topic is represented by the product of an acoustic topic occurrence probability $p(z_{s,i}|\boldsymbol{\theta}_{a_s=a})$ and transition probability $p(z_{s,i}|\boldsymbol{\pi}_{z_{s,i-1}=t})$. Thus, the probability of generating all acoustic word sequences $\boldsymbol{w}$ can be represented as follows.

$$
\begin{aligned}
&p(\boldsymbol{w}|\alpha,\beta,\gamma,\mathbf{a}_s) \\
&= \prod_{s=1}^{S} \prod_{i=1}^{N_{\boldsymbol{w}_s}} \sum_{\boldsymbol{a}} \sum_{\boldsymbol{z}} p(w_{s,i}|z_{s,i},\beta) p(z_{s,i}|a_s, z_{s,i-1}, \alpha, \gamma) p(a_s|\mathbf{a}_s) \\
&= \prod_{s=1}^{S}\Bigg[p(a_s|\mathbf{a}_s) \sum_{\boldsymbol{a}} \int p(\boldsymbol{\theta}_a|a_s,\alpha) \prod_{i=1}^{N_{\boldsymbol{w}_s}} \Bigg\{ \sum_{\boldsymbol{z}} \int \mathcal{D}(\boldsymbol{\pi}_z|\gamma) \\
&\quad \cdot p(z_{s,i}|\boldsymbol{\theta}_a, \boldsymbol{\pi}_t, z_{s,i-1})
\end{aligned}
$$

$$\times \int \mathcal{D}(\boldsymbol{\phi}_t|\beta)p(w_{s,i}|\boldsymbol{\phi}_t, z_{s,i})d\boldsymbol{\phi}_t d\boldsymbol{\pi}_t \Bigg\} d\boldsymbol{\theta}_a \Bigg]$$

$$= \frac{1}{A}\prod_{s=1}^{S}\Bigg[\int \frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\prod_{i=1}^{N_{\boldsymbol{w}_s}}\Bigg\{\prod_{t=1}^{T}\theta_{a,t}^{\alpha-1+n_t^{(a)}}$$

$$\times \int \frac{\Gamma(T\gamma)}{\Gamma(\gamma)^T}\prod_{t^+=1}^{T}\pi_{t^-,t^+}^{\gamma-1+n_{t^+}^{(t^-)}}$$

$$\cdot \int \frac{\Gamma(M\beta)}{\Gamma(\beta)^M}\prod_{m=1}^{M}\phi_{t,m}^{\beta-1+n_m^{(t)}} d\boldsymbol{\phi}_t d\boldsymbol{\pi}_t \Bigg\} d\boldsymbol{\theta}_a \Bigg] \tag{4}$$

In topic-transition sATM, the classification of acoustic scenes and the estimation of missing acoustic words are also achieved in a similar manner to in word-transition sATM.

IV. Estimation of Parameters in Proposed Models

To classify acoustic scenes and estimate missing acoustic words, we must infer the model parameters $\boldsymbol{\theta}_a$, $\boldsymbol{\phi}_t$, $\boldsymbol{\pi}_t/\boldsymbol{\pi}_m$, $\boldsymbol{a}$, $\boldsymbol{z}$, and $\boldsymbol{w}^{mis}$ that maximize the posterior probabilities of the models. However, it is difficult to infer them analytically in word-transition sATM and topic-transition sATM, and therefore, we introduce Bayesian inference based on collapsed Gibbs sampling (CGS) [30], [32], using which we can expect a more unbiased result for the predictive probability than by using other variational Bayes (VB) [33] or expectation propagation (EP) methods [34]. In CGS, latent variables corresponding to acoustic scenes, topics, and missing acoustic words are first initialized by giving them arbitrary values. Then, CGS iteratively samples latent variables in accordance with the conditional posterior probability of given acoustic word sequences, which do not involve updated acoustic scenes, topics, and missing words. This sampling is repeated until the iterative update converges. Then, the posterior distributions of the acoustic topic, word, and topic/word transitions are estimated from the inferred latent variables.

A. Parameter Estimation for Word-Transition sATM

When estimating the model parameters of the word-transition sATM using CGS, 1) $z_{s,i}, w_{s,i}^{mis}$ and 2) a_s are sampled separately since we assume that each acoustic topic $z_{s,i}$ and word $w_{s,i}$ is generated once every time frame s, i and that each acoustic scene a_s is generated once every sound clip s. We discuss the update for each case in detail below.

1) Posterior Probability of Acoustic Topic and Word: We first consider the joint posterior probability of an acoustic topic and word $p(w_{s,i}^{mis}, z_{s,i}|\boldsymbol{w}_{\backslash s,i}, \boldsymbol{z}_{\backslash s,i}, \boldsymbol{a}, \alpha, \beta, \gamma)$ for their sampling, which can be written as

$$p(w_{s,i}^{mis}, z_{s,i}|\boldsymbol{w}_{\backslash s,i}, \boldsymbol{z}_{\backslash s,i}, \boldsymbol{a}, \alpha, \beta, \gamma)$$

$$= \frac{p(\boldsymbol{w}|\boldsymbol{z}, \boldsymbol{a}, \alpha, \beta, \gamma)}{p(\boldsymbol{w}_{\backslash s,i}|\boldsymbol{z}_{\backslash s,i}, \boldsymbol{a}, \alpha, \beta, \gamma)} \cdot \frac{p(\boldsymbol{z}|\boldsymbol{a}, \alpha, \beta, \gamma)}{p(\boldsymbol{z}_{\backslash s,i}|\boldsymbol{a}, \alpha, \beta, \gamma)}$$

$$= \frac{p(\boldsymbol{w}|\boldsymbol{z}, \beta, \gamma)}{p(\boldsymbol{w}_{\backslash s,i}|\boldsymbol{z}_{\backslash s,i}, \beta, \gamma)} \cdot \frac{p(\boldsymbol{z}|\boldsymbol{a}, \alpha)}{p(\boldsymbol{z}_{\backslash s,i}|\boldsymbol{a}, \alpha)}. \tag{5}$$

Then, considering that $p(z_{s,i}|\boldsymbol{\theta}_a, a_s)$, $p(w_{s,i}|\boldsymbol{\phi}_{z_{s,i}}, z_{s,i})$, and $p(w_{s,i}|\boldsymbol{\pi}_{w_{s,i-1}}, w_{s,i-1})$ have the categorical distribution and that $p(\boldsymbol{\theta}|\alpha)$, $p(\boldsymbol{\phi}|\beta)$, and $p(\boldsymbol{\pi}|\gamma)$ have the Dirichlet distribution, $p(\boldsymbol{z}|\boldsymbol{a}, \alpha)$ and $p(\boldsymbol{w}|\boldsymbol{z}, \beta, \gamma)$ can be represented as

$$p(\boldsymbol{z}|\boldsymbol{a}, \alpha) = \int p(\boldsymbol{z}, \boldsymbol{\theta}|\boldsymbol{a}, \alpha)d\boldsymbol{\theta}$$

$$= \left(\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\right)^A \prod_{a=1}^{A}\frac{\prod_{t=1}^{T}\Gamma(n_t^{(a)} + \alpha)}{\Gamma(n_\cdot^{(a)} + T\alpha)} \tag{6}$$

$$p(\boldsymbol{w}|\boldsymbol{z}, \beta, \gamma) = \int\int p(\boldsymbol{w}, \boldsymbol{\phi}, \boldsymbol{\pi}|\boldsymbol{z}, \beta, \gamma)d\boldsymbol{\phi}d\boldsymbol{\pi}$$

$$= \prod_{s=1}^{S}\prod_{i=1}^{N_{\boldsymbol{w}_s}}\int\int p(w_{s,i}, \boldsymbol{\phi}|\boldsymbol{z}, \beta) \times p(w_{s,i}, \boldsymbol{\pi}|w_{s,i-1}, \gamma)d\boldsymbol{\phi}d\boldsymbol{\pi}, \tag{7}$$

where

$$\prod_{s=1}^{S}\prod_{i=1}^{N_{\boldsymbol{w}_s}}\int p(w_{s,i}, \boldsymbol{\phi}|\boldsymbol{z}, \beta)d\boldsymbol{\phi}$$

$$= \left(\frac{\Gamma(M\beta)}{\Gamma(\beta)^M}\right)^T \prod_{t=1}^{T}\frac{\prod_{m=1}^{M}\Gamma(n_m^{(t)} + \beta)}{\Gamma(n_\cdot^{(t)} + M\beta)} \tag{8}$$

$$\prod_{s=1}^{S}\prod_{i=1}^{N_{\boldsymbol{w}_s}}\int p(w_{s,i}, \boldsymbol{\pi}|w_{s,i-1}, \gamma)d\boldsymbol{\pi}$$

$$= \left(\frac{\Gamma(M\gamma)}{\Gamma(\gamma)^M}\right)^M \prod_{m^-=1}^{M}\frac{\prod_{m^+=1}^{M}\Gamma(n_{m^+}^{(m^-)} + \gamma)}{\Gamma(n_\cdot^{(m^-)} + M\gamma)}. \tag{9}$$

$n_t^{(a)}$ and $n_\cdot^{(a)}$ are the number of acoustic words assigned to acoustic topic t in acoustic scene a and the number of acoustic words in acoustic scene a. $n_m^{(t)}$, $n_\cdot^{(t)}$, $n_{m^+}^{(m^-)}$, and $n_\cdot^{(m^-)}$ are also defined in the same way as $n_t^{(a)}$ and $n_\cdot^{(a)}$. The detailed derivation of Eqs. (6)–(9) is given in Appendix A-1). Substituting these equations into each term of Eq. (5) and using $\Gamma(x+1)/\Gamma(x) = x$ for the gamma function, the update for an acoustic topic and acoustic word can be obtained as

$$p(w_{s,i}^{mis}, z_{s,i}|\boldsymbol{w}_{\backslash s,i}, \boldsymbol{z}_{\backslash s,i}, \boldsymbol{a}, \alpha, \beta, \gamma) \propto \left(n_{(\backslash s,i),t}^{(a)} + \alpha\right) \cdot \frac{n_{(\backslash s,i),m}^{(t)} + \beta}{n_{(\backslash s,i),\cdot}^{(t)} + M\beta}$$

$$\cdot \frac{\left(n_{(\backslash s,i),w_{s,i}}^{(w_{s,i-1})} + \gamma\right)\left\{n_{(\backslash s,i),w_{s,i+1}}^{(w_{s,i})} + \delta_{w_{s,i-1},w_{s,i}} \cdot \delta_{w_{s,i},w_{s,i+1}} + \gamma\right\}}{n_{(\backslash s,i),\cdot}^{(w_{s,i})} + \delta_{w_{s,i-1},w_{s,i}} + M\gamma}, \tag{10}$$

where $n_{(\backslash s,i),w_{s,i}}^{(w_{s,i-1})}$ represents the number of acoustic words that transit from $w_{s,i-1}$ to $w_{s,i}$ in all acoustic words excluding $w_{s,i}$. $\delta_{w_{s,i-1},w_{s,i}}$ is the Kronecker delta function, which is 1 if $w_{s,i-1} = w_{s,i}$ and 0 otherwise. The detailed derivation of Eq. (10) is given in Appendix A-2).

If the acoustic words $w_{s,i}$ are not missing, we only have to sample the acoustic topic $z_{s,i}$ in each update. In this case, the update for the acoustic topic $z_{s,i}$ is given as Eq. (11) because

the last term in Eq. (10) becomes a constant when the acoustic word is observed.

$$p(z_{s,i}|\boldsymbol{w}_{\backslash s,i},\boldsymbol{z}_{\backslash s,i},\boldsymbol{a},\alpha,\beta,\gamma) \propto \left(n^{(a)}_{(\backslash s,i),t}+\alpha\right)\cdot\frac{n^{(t)}_{(\backslash s,i),m}+\beta}{n^{(t)}_{(\backslash s,i),\cdot}+M\beta} \tag{11}$$

2) Posterior Probability of Acoustic Scene: We next derive the posterior probability for the sampling of an acoustic scene $p(a_s|\boldsymbol{w},\boldsymbol{z},\boldsymbol{a}_{\backslash s},\alpha,\beta,\gamma)$. Given that $p(\boldsymbol{a}|\alpha)$ has a uniform distribution and $p(\boldsymbol{w})$ does not directly depend on $\boldsymbol{a}$, the posterior probability is described by

$$\begin{aligned} &p(a_s|\boldsymbol{w},\boldsymbol{z},\boldsymbol{a}_{\backslash s},\alpha,\beta,\gamma)\\ &\quad=\frac{p(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{a},\beta,\gamma)}{p(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{a}_{\backslash s},\beta,\gamma)}\cdot\frac{p(\boldsymbol{z}|\boldsymbol{a},\alpha)}{p(\boldsymbol{z}|\boldsymbol{a}_{\backslash s},\alpha)}\cdot\frac{p(\boldsymbol{a}|\alpha)}{p(\boldsymbol{a}_{\backslash s}|\alpha)}\\ &\quad\propto\frac{p(\boldsymbol{z}|\boldsymbol{a},\alpha)}{p(\boldsymbol{z}|\boldsymbol{a}_{\backslash s},\alpha)}. \end{aligned} \tag{12}$$

Similarly to in the above discussion, we obtain the posterior probability of an acoustic scene a_s as

$$p(a_s|\boldsymbol{w},\boldsymbol{z},\boldsymbol{a}_{\backslash s},\alpha,\beta,\gamma)\propto\frac{n^{(a)}_{(\backslash s),t}+\alpha}{n^{(a)}_{(\backslash s),\cdot}+T\alpha}. \tag{13}$$

Since the proposed method assumes that each acoustic word sequence contains a single acoustic scene, the acoustic scene is sampled with respect to each acoustic word sequence.

3) Updates for Distributions: Given the posterior probabilities for sampling, the posterior distributions of an acoustic topic, word, and word transition can be estimated through the assignments of sufficiently updated latent variables using Eqs. (10), (11), and (13). In practice, the parameters of the generative distributions can be approximated as the following means of distributions of multiple samples:

$$\overline{\theta}_{a,t}=\frac{1}{N_G}\sum_{j=1}^{N_G}\left\{\frac{\sum_S\sum_{N_{\boldsymbol{w}_s}}\delta_{\hat{a}_{j,s},a}\cdot\delta_{\hat{z}_{j,s,i},t}+\alpha}{\sum_S\sum_{N_{\boldsymbol{w}_s}}\sum_t\delta_{\hat{a}_{j,s},a}\cdot\delta_{\hat{z}_{j,s,i},t}+T\alpha}\right\} \tag{14}$$

$$\overline{\phi}_{t,m}=\frac{1}{N_G}\sum_{j=1}^{N_G}\left\{\frac{\sum_s\sum_{N_{\boldsymbol{w}_s}}\delta_{\hat{z}_{j,s,i},t}\cdot\delta_{\hat{w}_{j,s,i},m}+\beta}{\sum_s\sum_{N_{\boldsymbol{w}_s}}\sum_m\delta_{\hat{z}_{j,s,i},t}\cdot\delta_{\hat{w}_{j,s,i},m}+M\beta}\right\} \tag{15}$$

$$\overline{\pi}_{m^-,m^+}=\frac{1}{N_G}\sum_{j=1}^{N_G}\left\{\frac{\sum_s\sum_{N_{\boldsymbol{w}_s}}\delta_{\hat{w}_{j,s,i},m^-}\cdot\delta_{\hat{w}_{j,s,i+1},m^+}+\gamma}{\sum_s\sum_{N_{\boldsymbol{w}_s}}\sum_{m^+}\delta_{\hat{w}_{j,s,i},m^-}\cdot\delta_{\hat{w}_{j,s,i+1},m^+}+M\gamma}\right\}, \tag{16}$$

where N_G and δ are the number of samplings and the Kronecker delta function, respectively. $\hat{a}_{j,s}$ is the acoustic scene of the sth sound clip in the jth sampling. $\hat{z}_{j,s,i}$ and $\hat{w}_{j,s,i}$ are the ith acoustic topic and acoustic word in the sth sound clip and the jth sampling, respectively.

B. Parameter Estimation for Topic-Transition sATM

In the CGS for topic-transition sATM, posterior probabilities for sampling $w^{mis}_{s,i}$, $z_{s,i}$, and a_s can be derived in similar manner to in word-transition sATM. We discuss them separately by considering the posterior probabilities of 1) $w^{mis}_{s,i}$, $z_{s,i}$ and 2) a_s.

1) Posterior Probability of Acoustic Topic and Word: In topic-transition sATM, the joint posterior probability of the acoustic topic and word $p(w^{mis}_{s,i},z_{s,i}|\boldsymbol{w}_{\backslash s,i},\boldsymbol{z}_{\backslash s,i},\boldsymbol{a},\alpha,\beta,\gamma)$ can be written as

$$\begin{aligned} &p(w^{mis}_{s,i},z_{s,i}|\boldsymbol{w}_{\backslash s,i},\boldsymbol{z}_{\backslash s,i},\boldsymbol{a},\alpha,\beta,\gamma)\\ &\quad=\frac{p(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{a},\alpha,\beta,\gamma)}{p(\boldsymbol{w}_{\backslash s,i}|\boldsymbol{z}_{\backslash s,i},\boldsymbol{a},\alpha,\beta,\gamma)}\cdot\frac{p(\boldsymbol{z}|\boldsymbol{a},\alpha,\beta,\gamma)}{p(\boldsymbol{z}_{\backslash s,i}|\boldsymbol{a},\alpha,\beta,\gamma)}\\ &\quad=\frac{p(\boldsymbol{w}|\boldsymbol{z},\beta)}{p(\boldsymbol{w}_{\backslash s,i}|\boldsymbol{z}_{\backslash s,i},\beta)}\cdot\frac{p(\boldsymbol{z}|\boldsymbol{a},\alpha,\gamma)}{p(\boldsymbol{z}_{\backslash s,i}|\boldsymbol{a},\alpha,\gamma)}. \end{aligned} \tag{17}$$

Considering that $p(z_{s,i}|\boldsymbol{\theta}_a,a_s)$, $p(w_{s,i}|\boldsymbol{\phi}_{z_{s,i}},z_{s,i})$, and $p(z_{s,i}|\boldsymbol{\pi}_{z_{s,i-1}},z_{s,i-1})$ have the categorical distribution and that $p(\boldsymbol{\theta}|\alpha)$, $p(\boldsymbol{\phi}|\beta)$, and $p(\boldsymbol{\pi}|\gamma)$ have the Dirichlet distribution, $p(\boldsymbol{z}|\boldsymbol{a},\alpha,\gamma)$ and $p(\boldsymbol{w}|\boldsymbol{z},\beta)$ can be represented as

$$\begin{aligned} p(\boldsymbol{z}|\boldsymbol{a},\alpha,\gamma)&=\int\int p(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\pi}|\boldsymbol{a},\alpha,\gamma)d\boldsymbol{\theta}d\boldsymbol{\pi}\\ &=\prod_{s=1}^{S}\prod_{i=1}^{N_{\boldsymbol{w}_s}}\int\int p(z_{s,i},\boldsymbol{\theta}|\boldsymbol{a},\alpha)\\ &\quad\times p(z_{s,i},\boldsymbol{\pi}|z_{s,i-1},\gamma)d\boldsymbol{\theta}d\boldsymbol{\pi} \end{aligned} \tag{18}$$

$$\begin{aligned} p(\boldsymbol{w}|\boldsymbol{z},\beta)&=\int p(\boldsymbol{w},\boldsymbol{\phi}|\boldsymbol{z},\beta)d\boldsymbol{\phi}\\ &=\left(\frac{\Gamma(M\beta)}{\Gamma(\beta)^M}\right)^T\prod_{t=1}^{T}\frac{\prod_{m=1}^{M}\Gamma(n^{(t)}_m+\beta)}{\Gamma(n^{(t)}_{\cdot}+M\beta)}, \end{aligned} \tag{19}$$

where

$$\begin{aligned} &\prod_{s=1}^{S}\prod_{i=1}^{N_{\boldsymbol{w}_s}}\int p(z_{s,i},\boldsymbol{\theta}|\boldsymbol{a},\alpha)d\boldsymbol{\theta}\\ &\quad=\left(\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\right)^A\prod_{a=1}^{A}\frac{\prod_{t=1}^{T}\Gamma(n^{(a)}_{(t)}+\alpha)}{\Gamma(n^{(a)}_{\cdot}+T\alpha)} \end{aligned} \tag{20}$$

$$\begin{aligned} &\prod_{s=1}^{S}\prod_{i=1}^{N_{\boldsymbol{w}_s}}\int p(z_{s,i},\boldsymbol{\pi}|z_{s,i-1},\gamma)d\boldsymbol{\pi}\\ &\quad=\left(\frac{\Gamma(T\gamma)}{\Gamma(\gamma)^T}\right)^T\prod_{t^-=1}^{T}\frac{\prod_{t^+=1}^{T}\Gamma(n^{(t^-)}_{t^+}+\gamma)}{\Gamma(n^{(t^-)}_{\cdot}+T\gamma)}. \end{aligned} \tag{21}$$

Substituting these equations into each term of Eq. (17) and using $\Gamma(x+1)/\Gamma(x)=x$ for the gamma function, the update for an acoustic topic and acoustic word can be obtained as

$$\begin{aligned} p(w^{mis}_{s,i},z_{s,i}|\boldsymbol{w}_{\backslash s,i},\boldsymbol{z}_{\backslash s,i},\boldsymbol{a},\alpha,\beta,\gamma)&\propto\left(n^{(a)}_{(\backslash s,i),t}+\alpha\right)\\ &\quad\cdot\frac{n^{(t)}_{(\backslash s,i),m}+\beta}{n^{(t)}_{(\backslash s,i),\cdot}+M\beta}\\ &\quad\cdot\frac{\left(n^{(z_{s,i-1})}_{(\backslash s,i),z_{s,i}}+\gamma\right)\left\{n^{(z_{s,i})}_{(\backslash s,i),z_{s,i+1}}+\delta_{z_{s,i-1},z_{s,i}}\cdot\delta_{z_{s,i},z_{s,i+1}}+\gamma\right\}}{n^{(z_{s,i})}_{(\backslash s,i),\cdot}+\delta_{z_{s,i-1},z_{s,i}}+T\gamma}. \end{aligned} \tag{22}$$

TABLE VI
POSTERIOR PROBABILITIES FOR GIBBS SAMPLING. TO INCREASE READABILITY, SUBSCRIPT

Variable	Word-transition sATM	Topic-transition sATM
Acoustic topic & word (when acoustic word is missing) $p(w_i^{mis}, z_i \mid \boldsymbol{w}_{\setminus i}, \boldsymbol{z}_{\setminus i}, \boldsymbol{a}, \alpha, \beta, \gamma)$	$\propto (n^{(a)}_{(\setminus i),t} + \alpha) \cdot \frac{n^{(t)}_{(\setminus i),m} + \beta}{n^{(t)}_{(\setminus i),\cdot} + M\beta} \cdot (n^{(w_{i-1})}_{(\setminus i),w_i} + \gamma) \cdot \frac{n^{(w_i)}_{(\setminus i),w_{i+1}} + \delta_{w_{i-1},w_i} \cdot \delta_{w_i,w_{i+1}} + \gamma}{n^{(w_i)}_{(\setminus i),\cdot} + \delta_{w_{i-1},w_i} + M\gamma}$ (10)	$\propto (n^{(a)}_{(\setminus i),t} + \alpha) \cdot \frac{n^{(t)}_{(\setminus i),m} + \beta}{n^{(t)}_{(\setminus i),\cdot} + M\beta} \cdot (n^{(z_{i-1})}_{(\setminus i),z_i} + \gamma) \cdot \frac{n^{(z_i)}_{(\setminus i),z_{i+1}} + \delta_{z_{i-1},z_i} \cdot \delta_{z_i,z_{i+1}} + \gamma}{n^{(z_i)}_{(\setminus i),\cdot} + \delta_{z_{i-1},z_i} + T\gamma}$ (22)
Acoustic topic (when acoustic word is observed) $p(z_i \mid \boldsymbol{w}, \boldsymbol{z}_{\setminus i}, \boldsymbol{a}, \alpha, \beta, \gamma)$	$\propto (n^{(a)}_{(\setminus i),t} + \alpha) \cdot \frac{n^{(t)}_{(\setminus i),m} + \beta}{n^{(t)}_{(\setminus i),\cdot} + M\beta}$ (11)	
Acoustic scene $p(a_s \mid \boldsymbol{w}, \boldsymbol{z}, \boldsymbol{a}_{\setminus s}, \alpha, \beta, \gamma)$	$\propto \frac{n^{(a)}_{(\setminus s),t} + \alpha}{n^{(a)}_{(\setminus s),\cdot} + T\alpha}$ (13)	

where $n^{(z_{s,i-1})}_{(\setminus i),z_{s,i}}$ represents the number of acoustic topics that transit from $z_{s,i-1}$ to $z_{s,i}$ in all acoustic topics excluding $z_{s,i}$. The detailed derivation of Eq. (22) is given in Appendix B.

If the acoustic word $w^{obs}_{s,i}$ is not missing, only the acoustic topic $z_{s,i}$ is sampled in each update. Since the update for sampling the acoustic topic depends on $z_{s,i} = t$, it can also be described by Eq. (22).

2) Posterior Probability of Acoustic Scene: Considering that $p(\boldsymbol{a}|\alpha)$ has a uniform distribution and that $p(\boldsymbol{w})$ does not directly depend on $\boldsymbol{a}$, the posterior probability for an acoustic scene $p(a_s|\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{a}_{\setminus s}, \alpha, \beta, \gamma)$ can be also represented by Eq. (13). Finally, the procedures of Gibbs sampling for updating the parameters are summarized in Tables VI, VII, and VIII.

V. EXPERIMENTAL EVALUATION

A. Data Preparation

We conducted experiments to evaluate how effectively the proposed method can classify acoustic scenes and can estimate missing words. As an illustration of acoustic scene analysis indoors, nine acoustic scenes; chatting, cooking, eating dinner, operating a PC, reading a newspaper, vacuuming, walking, washing dishes, and watching TV, that occur frequently in a living room, were selected. Here, each acoustic scene also typically included the sounds listed in Table IX, and there are around 100 types of acoustic events in the recorded sounds. The recording conditions of the microphone arrangement and the sound source positions are shown in Fig. 6. To record these sounds, we used Sony ECM-55B microphones, a Grace Design m802 microphone preamplifier, and an MOTU 24I/O A/D converter. Each sound was recorded with a start cue of one of these nine acoustic scenes, and then the sound was labeled with its acoustic scene label. The recording has 49.36 h of sounds, which were segmented into 11,105 sound clips with an equal duration. Among the 11,105 recorded sound clips, 9,802 sound clips were used for learning the model parameters and 1,303 sound clips were used for evaluation.

TABLE VII
GIBBS SAMPLING PROCEDURE FOR WORD-TRANSITION SATM

[Step 1] Initialization
set hyperparameters α, β, γ
initialize latent variables $\boldsymbol{a}$, $\boldsymbol{z}$, $\boldsymbol{w}^{mis}$
initialize $n^{(a_s)}_{z_{s,i}}$, $n^{(a_s)}_{\cdot}$, $n^{(z_{s,i})}_{w_{s,i}}$, $n^{(z_{s,i})}_{\cdot}$

[Step 2] Gibbs sampling for word-transition sATM
repeat
  for $s = 1$ to S **do**
    for $i = 1$ to $N_{\boldsymbol{w}_s}$ **do**
      $n^{(a_s)}_{z_{s,i}} = n^{(a_s)}_{z_{s,i}} - 1$, $n^{(z_{s,i})}_{w_{s,i}} = n^{(z_{s,i})}_{w_{s,i}} - 1$, $n^{(z_{s,i})}_{\cdot} = n^{(z_{s,i})}_{\cdot} - 1$
      if $w_{s,i}$ is missing **then**
        Sample $w^{mis}_{s,i}$ and $z_{s,i}$ using Eq. (10)
      else if $w_{s,i}$ is observed **then**
        Sample $z_{s,i}$ using Eq. (11)
      end if
      $n^{(a_s)}_{z_{s,i}} = n^{(a_s)}_{z_{s,i}} + 1$, $n^{(z_{s,i})}_{w_{s,i}} = n^{(z_{s,i})}_{w_{s,i}} + 1$, $n^{(z_{s,i})}_{\cdot} = n^{(z_{s,i})}_{\cdot} + 1$
    end for
    Sample a_s using Eq. (13)
  end for
until convergence condition is satisfied

[Step 3] Calculation of posterior distributions
for $j = 1$ to N_G **do**
  for $s = 1$ to S **do**
    for $i = 1$ to $N_{\boldsymbol{w}_{j,s}}$ **do**
      if $\hat{w}_{j,s,i}$ is missing **then**
        Sample $\hat{w}^{mis}_{j,s,i}$ and $\hat{z}_{j,s,i}$ using Eq. (10)
      else if $\hat{w}_{j,s,i}$ is observed **then**
        Sample $\hat{z}_{j,s,i}$ using Eq. (11)
      end if
    end for
    Sample $\hat{a}_{j,s}$ using Eq. (13)
  end for
end for
Calculate $\overline{\theta}_{a,t}$, $\overline{\phi}_{t,m}$, and $\overline{\pi}_{m^-,m^+}$ using Eqs. (14)–(16)

TABLE VIII
GIBBS SAMPLING PROCEDURE FOR TOPIC-TRANSITION SATM

[Step 1] Initialization
set hyperparameters α, β, γ
initialize latent variables $\boldsymbol{a}$, $\boldsymbol{z}$, $\boldsymbol{w}^{mis}$
initialize $n^{(a_s)}_{z_{s,i}}$, $n^{(a_s)}_{\cdot}$, $n^{(z_{s,i})}_{w_{s,i}}$, $n^{(z_{s,i})}_{\cdot}$
[Step 2] Gibbs sampling for topic-transition sATM
repeat
  for $s = 1$ to S **do**
    for $i = 1$ to $N_{\boldsymbol{w}_s}$ **do**
      $n^{(a_s)}_{z_{s,i}} = n^{(a_s)}_{z_{s,i}} - 1$, $n^{(z_{s,i})}_{w_{s,i}} = n^{(z_{s,i})}_{w_{s,i}} - 1$, $n^{(z_{s,i})}_{\cdot} = n^{(z_{s,i})}_{\cdot} - 1$
      if $w_{s,i}$ is missing **then**
        Sample $w^{mis}_{s,i}$ and $z_{s,i}$ using Eq. (22)
      else if $w_{s,i}$ is observed **then**
        Sample $z_{s,i}$ using Eq. (22)
      end if
      $n^{(a_s)}_{z_{s,i}} = n^{(a_s)}_{z_{s,i}} + 1$, $n^{(z_{s,i})}_{w_{s,i}} = n^{(z_{s,i})}_{w_{s,i}} + 1$, $n^{(z_{s,i})}_{\cdot} = n^{(z_{s,i})}_{\cdot} + 1$
    end for
    Sample a_s using Eq. (13)
  end for
until convergence condition is satisfied
[Step 3] Calculation of posterior distributions
for $j = 1$ to N_G **do**
  for $s = 1$ to S **do**
    for $i = 1$ to $N_{\boldsymbol{w}_s}$ **do**
      if $\hat{w}_{j,s,i}$ is missing **then**
        Sample $\hat{w}^{mis}_{j,s,i}$ and $\hat{z}_{j,s,i}$ using Eq. (22)
      else if $\hat{w}_{j,s,i}$ is observed **then**
        Sample $\hat{z}_{j,s,i}$ using Eq. (22)
      end if
    end for
    Sample $a_{j,s}$ using Eq. (13)
  end for
end for
Calculate $\overline{\theta}_{a,t}$, $\overline{\phi}_{t,m}$, and $\overline{\pi}_{m^-,m^+}$ using Eqs. (14)–(16)

TABLE IX
NUMBER OF SOUND CLIPS AND TYPICAL SOUNDS IN EACH ACOUSTIC SCENE

Acoustic scene	# sound clips (Train/Test)	Typical sounds
Vacuuming	502 / 78	whine of cleaner, footsteps
Cooking	2987 / 346	cutting, sizzling, running water, clattering dishes
Dishwashing	1027 / 155	running water, clattering dishes
Eating	2113 / 259	clattering dishes, voices, coughing
Newspaper	366 / 63	flipping newspaper, footsteps
PC	941 / 164	clicking mouse, clacking keyboard, fan noise, sound effects
Chatting	567 / 74	voices, coughing
TV	969 / 122	voices, music, sound effects, cheering
Walking	330 / 42	footsteps, coughing

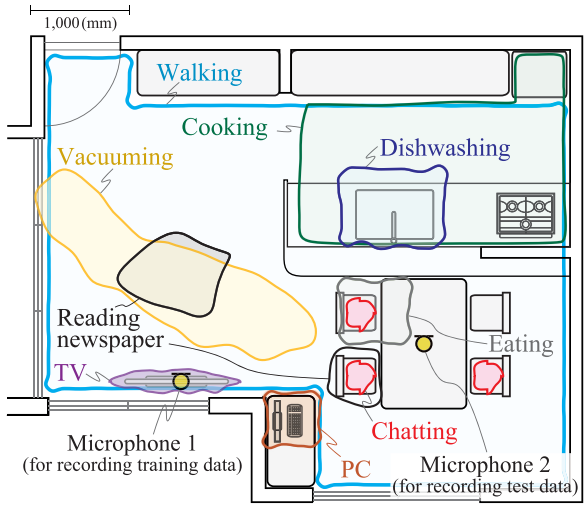


Fig. 6. Microphone arrangement and sound source positions in acoustic scene classification experiment using real-life environmental sounds.

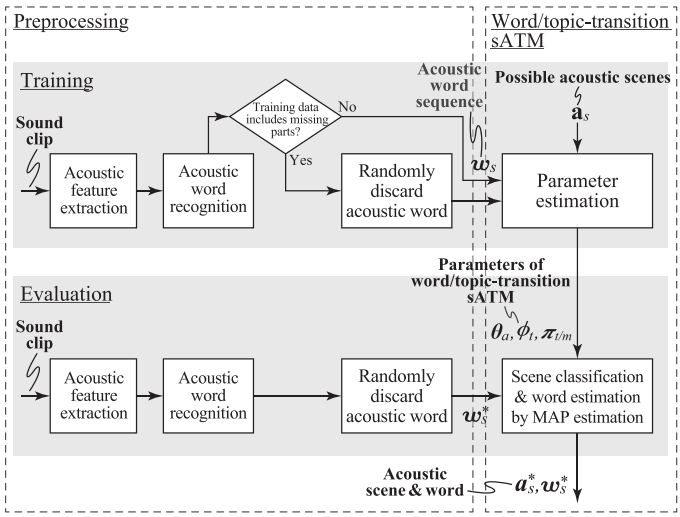


Fig. 7. Process used for acoustic scene classification and acoustic word estimation.

B. Evaluation Procedure and Experimental Conditions

The performance of the proposed methods of acoustic scene classification and missing word estimation was evaluated by the process shown in Fig. 7. As preprocessing, acoustic feature vectors were extracted from input sound clips frame by frame. In this experiment, the 12-dimensional Mel-frequency cepstral coefficients (MFCCs) were calculated as acoustic feature vectors, which are widely employed in ASA and AED. Acoustic words were then modeled and recognized using a GMM in an unsupervised manner. Thus, all the acoustic features in the training dataset were clustered by the GMM and each Gaussian component was defined to represent a different acoustic word. In Fig. 8, an example of an acoustic word histogram, in which each acoustic word was learned by the GMM is shown. The figure shows that the acoustic word histogram has a sparse structure and some acoustic words well characterize particular acoustic scenes; for

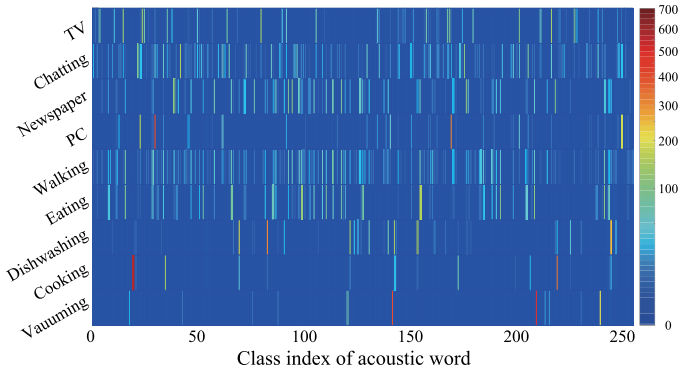


Fig. 8. Example of acoustic word histogram in each acoustic scene.

TABLE X
EXPERIMENTAL CONDITIONS

Sampling rate / quantization	16 kHz / 16 bits
Frame size	512
Acoustical feature	MFCCs (12 dim.)
# acoustic words in sound clip	1,000
# classes of acoustic word (# mixture of GMM)	256
# classes of acoustic topic	20
Hyperparameter α	3.33
Hyperparameter β	0.1
Hyperparameter γ	0.5

instance, the acoustic word of class index $= 20$ is active almost only in cooking. After recognizing acoustic words, to simulate observations with missing parts of various proportions, we discarded various proportions of acoustic words randomly and formed acoustic word sequences $\boldsymbol{w}_s$ and $\boldsymbol{w}_s^*$. Note that these acoustic word sequences include information on which acoustic words are missing.

For acoustic scene classification and word estimation, the parameters of the word-transition sATM and topic-transition sATM were estimated using acoustic word sequences $\boldsymbol{w}_s$ and acoustic scene label $\mathbf{a}_s$. Thus, acoustic scenes were finally classified and missing acoustic words were estimated through the MAP estimation of their corresponding parameters. The parameter estimation and classification test were conducted ten times with random initial parameters and randomly missing data.

The other experimental conditions are listed in Table X, among which the numbers of classes of acoustic words and topics, the hyperparameters α, β, and γ, and the parameters of the comparison methods were chosen on the basis of a preliminary experiment, which was conducted using a cross-validation setup with the training dataset.

C. Comparison Methods

For comparison with the proposed methods, we conducted experiments using conventional sATMs with the following types of preprocessing. (i) Acoustic words were discarded randomly and a word histogram was created using only non-missing acoustic words. Then, the word histogram was fed to the sATM (referred to as sATM in Tables XI, XII, and Figs. 9–12). (ii) Missing

TABLE XI
AVERAGE CLASSIFICATION ACCURACY OF ACOUSTIC SCENES IN TERMS OF F-MEASURE IN THE CASE WHERE THE TRAINING DATASET HAS NO MISSING PARTS

Method	Missing proportion (%)		
	0%	**40%**	**80%**
Word-transition sATM	74.18%	76.39%	63.70%
Topic-transition sATM	**83.36%**	**82.08%**	**74.11%**
sATM	74.18%	67.65%	46.94%
sATM (Random)	74.18%	58.11%	32.73%
sATM (HMM)	74.18%	70.76%	49.07%
Histogram MLP	78.12%	70.33%	32.15%
Histogram MLP (HMM)	78.12%	69.47%	49.61%
Histogram CNN	72.16%	61.77%	36.25%
Histogram CNN (HMM)	72.16%	62.40%	36.89%
SVM	60.21%	45.05%	14.45%
SVM (HMM)	60.21%	53.39%	18.10%
GMM	66.42%	60.11%	35.24%
GMM (HMM)	66.42%	48.79%	24.88%
Mel-bank MLP	71.30%	64.43%	52.17%
Mel-bank CNN	76.13%	59.38%	44.20%

TABLE XII
NUMBER OF TRAINABLE PARAMETERS IN EACH MODEL

sATM, sATM (Random), sATM (HMM)	5.3k
Word-transition sATM	70.8k
Topic-transition sATM	5.7k
histogram MLP, MLP (HMM)	12.7k
histogram CNN, CNN (HMM)	229.5k
GMM, GMM (HMM)	2.1k
SVM, SVM (HMM)	677.0k

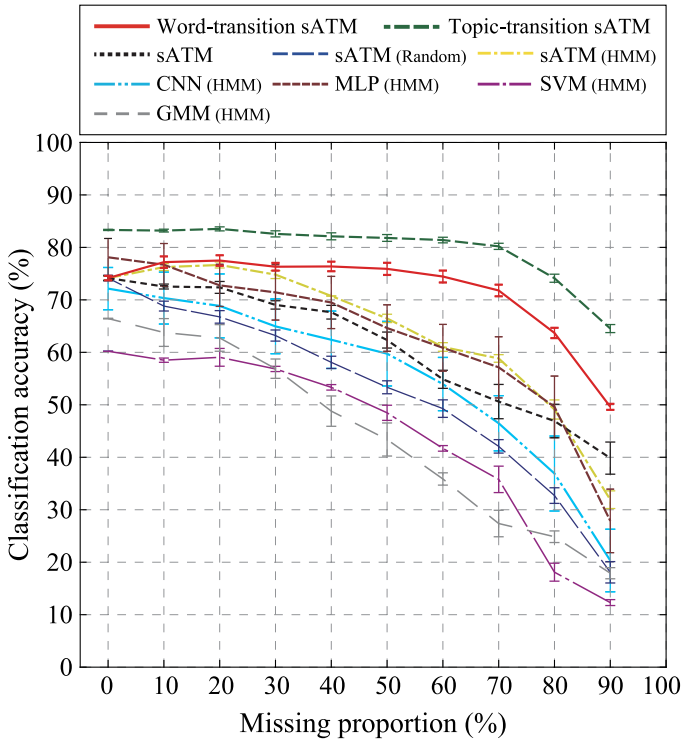


Fig. 9. Detailed classification accuracy of selected methods in the case where the training dataset has no missing parts.

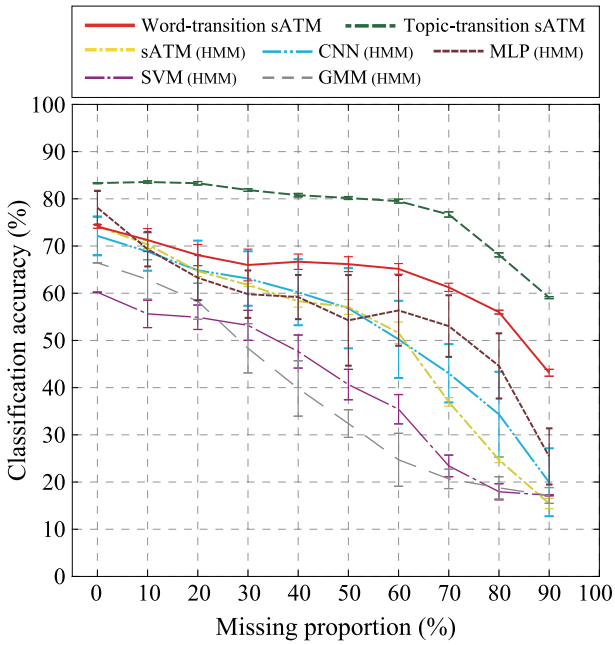


Fig. 10. Average classification accuracy of acoustic scenes in terms of F-measure in the case where the training dataset has missing parts.

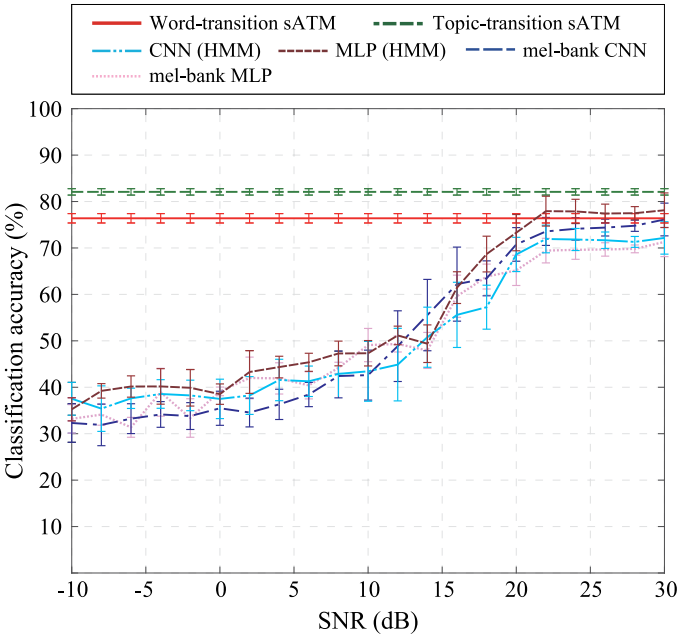


Fig. 11. Average classification accuracy of acoustic scenes where Gaussian noise is superimposed on 40% of time frames in the test dataset.

acoustic words were restored by random complementing and a word histogram was created by using the non-missing and restored acoustic words. Then, this word histogram was fed to the sATM (referred to as sATM (random)). (iii) Missing acoustic words were reconstructed using an HMM, which was modeled using all the training dataset. After that, a word histogram created using the non-missing and restored acoustic words was fed to the sATM (referred to as sATM (HMM)).

As other comparison methods, we also conducted experiments using an SVM, a GMM, a multilayer perceptron (MLP),

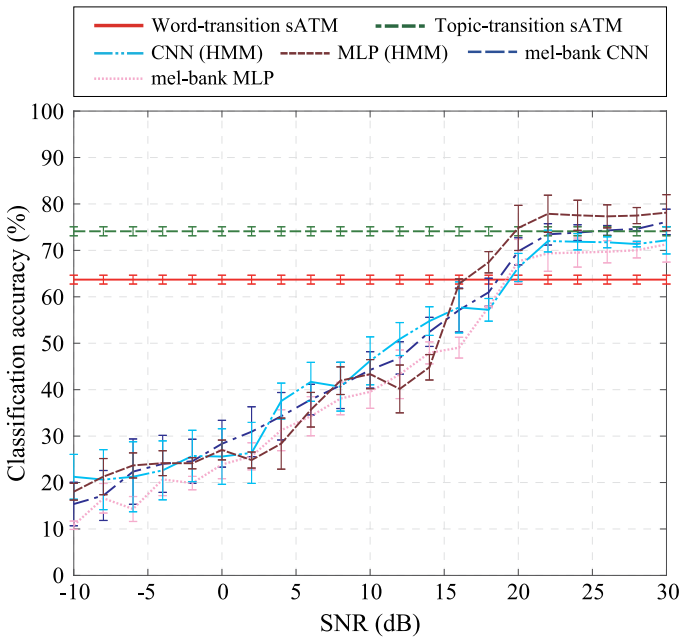


Fig. 12. Average classification accuracy of acoustic scenes where Gaussian noise is superimposed on 80% of time frames in the test dataset.

and a convolutional neural network (CNN). For these methods, (i) a word histogram created using only non-missing acoustic words and (ii) an acoustic word histogram reconstructed using HMM were utilized as input features. These methods are referred to as SVM, GMM, histogram MLP, histogram CNN, SVM (HMM), GMM (HMM), histogram MLP (HMM), and histogram CNN (HMM) in Tables XI, XII, and Figs. 9–12, respectively. For the SVM and SVM (HMM), we used the radial basis function (RBF) kernel. The histogram features were scaled to the range $[-1.0, 1.0]$. Then, a grid search of penalty and kernel parameters with the cross-validation setup using the training dataset was applied. For the GMM and GMM (HMM), each acoustic scene was modeled using four Gaussian mixture components with a diagonal covariance matrix. The histogram MLP and MLP (HMM) had one input layer, five hidden layers, and one output layer, where each hidden layer had 50 units and ReLUs as the activation functions. Softmax was applied for the activation function of the output layer. For parameter estimation of the MLP (HMM), we applied Adam as an optimizer and a dropout technique (dropout ratio $= 0.5$). The histogram CNN and CNN (HMM) had three convolution layers followed by five fully connected layers. In each convolution layer, batch normalization and maxpooling were applied. In each convolutional layer, the number of channels, kernel size, stride, and pooling size were 32, 1×4, 1, and 1×2, respectively. Each dense layer had 50 units. ReLUs were used as activation functions of the convolution and fully connected layers and softmax was used as the activation function of the output layer. Adam was applied as an optimizer.

Moreover, we also evaluated the MLP and CNN using a 128-dimensional mel-filter bank energy as the input feature. The mel-filter bank energy was extracted using a Hamming window of length 32 ms and 128 mel-bands in the frequency range of

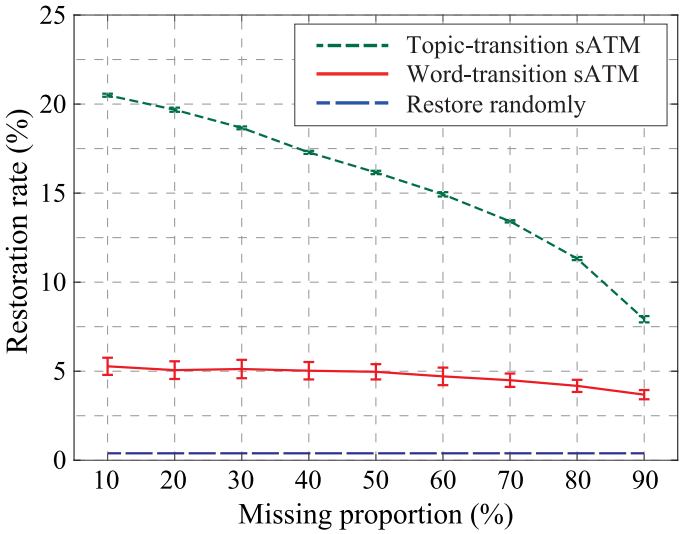


Fig. 13. Restoration rate of missing acoustic words in data restored by topic-transition sATM, and data restored by word-transition sATM, data restored by HMM, data restored randomly in the case where training dataset has no missing parts.

0–8,000 Hz. For the CNN, a 1,000 frame $\times$ 128-dimensional feature map was used as the input feature, and for the MLP, a concatenated vector of the feature map $(1,000 \cdot 128 \times 1)$ was used as the input feature. In these experiments, values of zero were assigned to acoustic features of unobserved time frames. These methods are referred to as mel-bank MLP and mel-bank CNN in Tables XI, XII, and Figs. 9–12, respectively. The mel-bank MLP had one input layer, five hidden layers, and one output layer, where the five hidden layers had 4,096, 2,048, 1,024, 256, and 64 units. The other settings of the mel-bank MLP are the same as those of MLP (HMM). The mel-bank CNN had three convolution layers followed by three fully connected layers. In each convolutional layer, the number of channels, kernel size, stride, and pooling size were 64, 3×3, 1, and 2×4, respectively. Each dense layer had 32 units. The other settings of the mel-bank CNN are the same as those of CNN (HMM). For all comparative methods, the model parameters were tuned using the same cross-validation setup as for the proposed methods.

D. Experimental Results

1) Scene Classification Performance for Training Data Without Missing Parts: The average classification accuracies (F-measure) of the nine acoustic scenes with various proportions of missing words are shown in Table XI and Fig. 9. In this experiment, we assume that the training dataset has no missing parts.

The results show that while the classification performances of the conventional methods decrease with increasing proportion of missing acoustic words, the two proposed methods achieve accurate classification even when the proportion of missing words is more than 50%. In particular, the topic-transition sATM achieves an average classification accuracy of 74.11% even when 80% of acoustic words are missing. This indicates that while in the conventional sATMs, the structure of acoustic words that characterizes an acoustic scene may collapse when more than 50% of the acoustic words are missing, the proposed methods can reconstruct them, enabling acoustic scene classification with reasonable accuracy. Additionally, the topic-transition sATM outperforms the word-transition sATM by about 7–8% on average. This result suggests that the topic-transition sATM can model the generation of acoustic words more realistically because it can model the variance of observed acoustic words similarly to an HMM. This is also indicated by comparing the results for no missing data (0% missing). Thus, the results indicate that modeling the transition of the latent structure of acoustic words is more reasonable than directly modeling the transition of observed acoustic words.

Table XII shows the number of trainable parameters in each model, where the trainable parameters are the parameters learned in the training stage and used in the evaluation stage such as $\boldsymbol{\theta}_a, \boldsymbol{\phi}_t, \boldsymbol{\pi}_t$, and $\boldsymbol{\pi}_m$. The results in this table indicate that the proposed methods achieve reasonable performance in acoustic scene classification with equivalent or fewer parameters than most of the conventional classification methods.

2) Scene Classification Performance for Training Data With Missing Parts: There are some situations in which it is difficult to prepare a training dataset without missing parts because of the recording environment or for reasons of privacy. Thus, we conducted a scene classification experiment using a training dataset with missing parts. The average classification accuracies (F-measures) are shown in Fig. 10. In this experiment, we compare the proposed methods with conventional sATM, a CNN, MLP, SVM, and GMM using a word histogram reconstructed using an HMM.

The results show that the two proposed methods achieve better performance than the conventional methods when acoustic words are missing. This indicates that even when the training dataset has missing parts, the proposed methods reconstruct the sound structure more accurately than the conventional methods. In particular, the topic-transition sATM achieves an average classification accuracy of 68.1% even when 80% of the acoustic words are missing.

Figs. 9 and 10 indicate that the classification performance of the proposed methods using training data with missing parts decreases by about 3–6% compared with that when using training data without missing parts. This indicates that the structure of acoustic words and topics collapses with increasing ratio of missing acoustic words in the training dataset.

3) Scene Classification Performance for Noisy Sound: In the proposed methods, it is assumed that we cannot observe some parts of observations completely, while even noisy observations may have information which is helpful for acoustic scene classification. Thus, we compared the acoustic scene classification performance of the proposed methods with the performance of conventional methods using a noisy dataset. In this experiment, we simulated a noisy environment by superimposing stationary white Gaussian noise on 40% or 80% of time frames in the test dataset with various SNRs. Here, the noise level in each frame was selected randomly in the range of ± 5 dB from the center SNR. For the proposed methods, we assume that we know which time frame is a noisy frame and we treat this time frame as the missing observation. As comparison methods, we evaluated a histogram CNN, histogram MLP, mel-bank CNN, and mel-bank

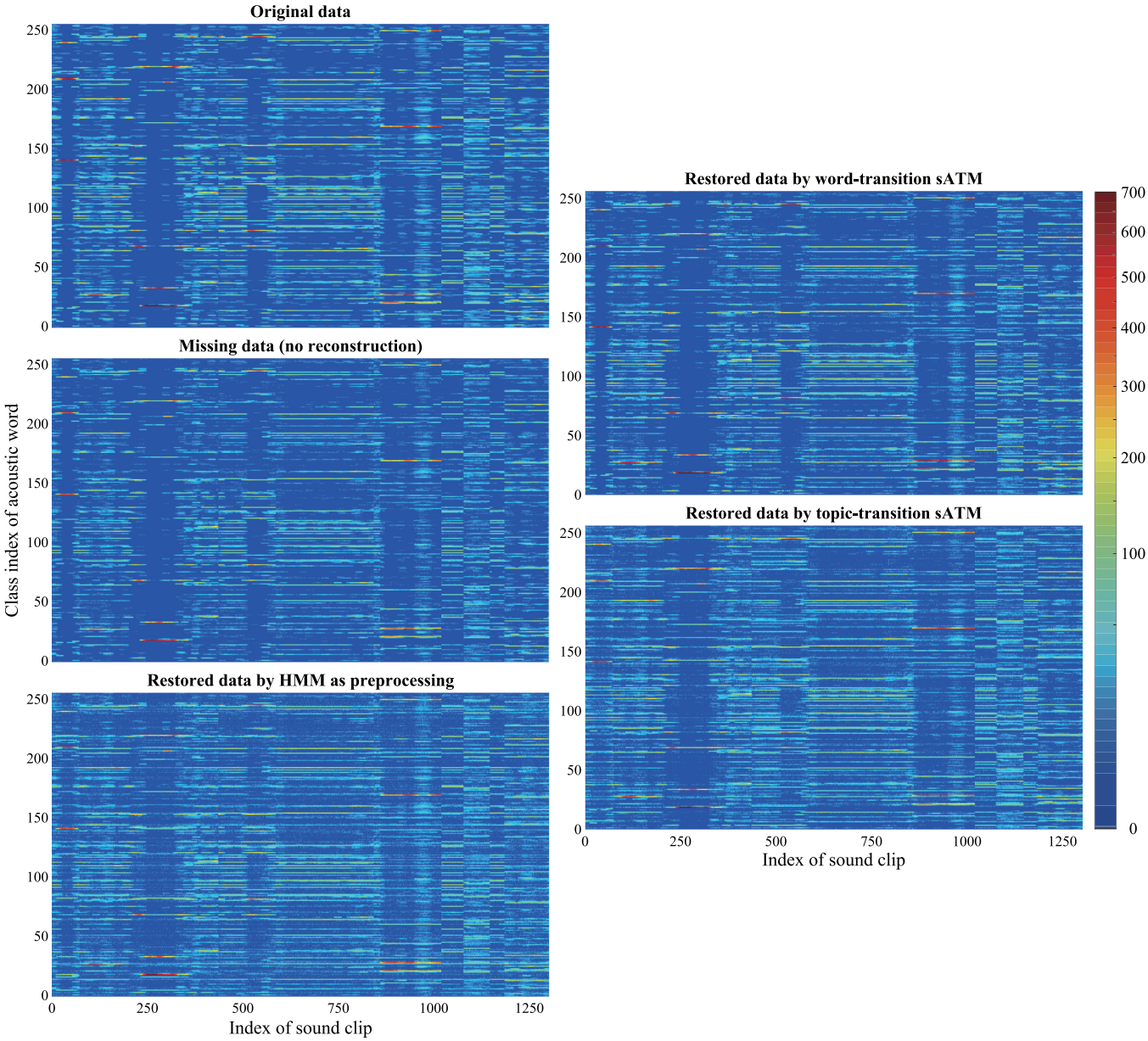


Fig. 14. Acoustic word histograms for the case of 40% missing data.

MLP. The other experimental conditions were the same as in the experiment of Sec. V-D-1).

Figs. 11 and 12 indicate that even when the SNR is high, the topic-transition sATM achieves a classification accuracy of acoustic scenes comparable to that of the conventional methods.

4) Estimation of Acoustic Words: The average restoration rate of missing acoustic words under similar conditions to those for Sec. V-D-1) are depicted in Fig. 13. Here, the restoration rate of missing acoustic words is defined as

$$(\text{Restoration rate}) = \frac{(\#\text{ correctly restored acoustic words})}{(\#\text{ missing acoustic words})}. \tag{23}$$

Both proposed methods achieve a higher restoration rate for acoustic words than random restoration. Moreover, Figs. 9 and 13 show that even when the proportion of missing acoustic words becomes high, the estimation accuracy of acoustic scenes using the topic-transition sATM achieves reasonable performance. This suggests that even when missing acoustic words are not estimated correctly, the proposed model can estimate acoustic words that are strongly correlated with an acoustic scene of the acoustic word sequence.

Further evaluations were conducted to investigate how the estimated acoustic words were distributed and the similarity between the original and estimated acoustic word histograms. Figs. 14 and 15 show acoustic word histograms of the original data, intermittently missing data, data restored by an HMM, data restored by the word-transition sATM, and data restored by the topic-transition sATM. When the proportion of missing words is 40%, the pattern of the acoustic words still remains. In this case, the three estimation methods can reconstruct the pattern of acoustic words accurately. On the other hand, when

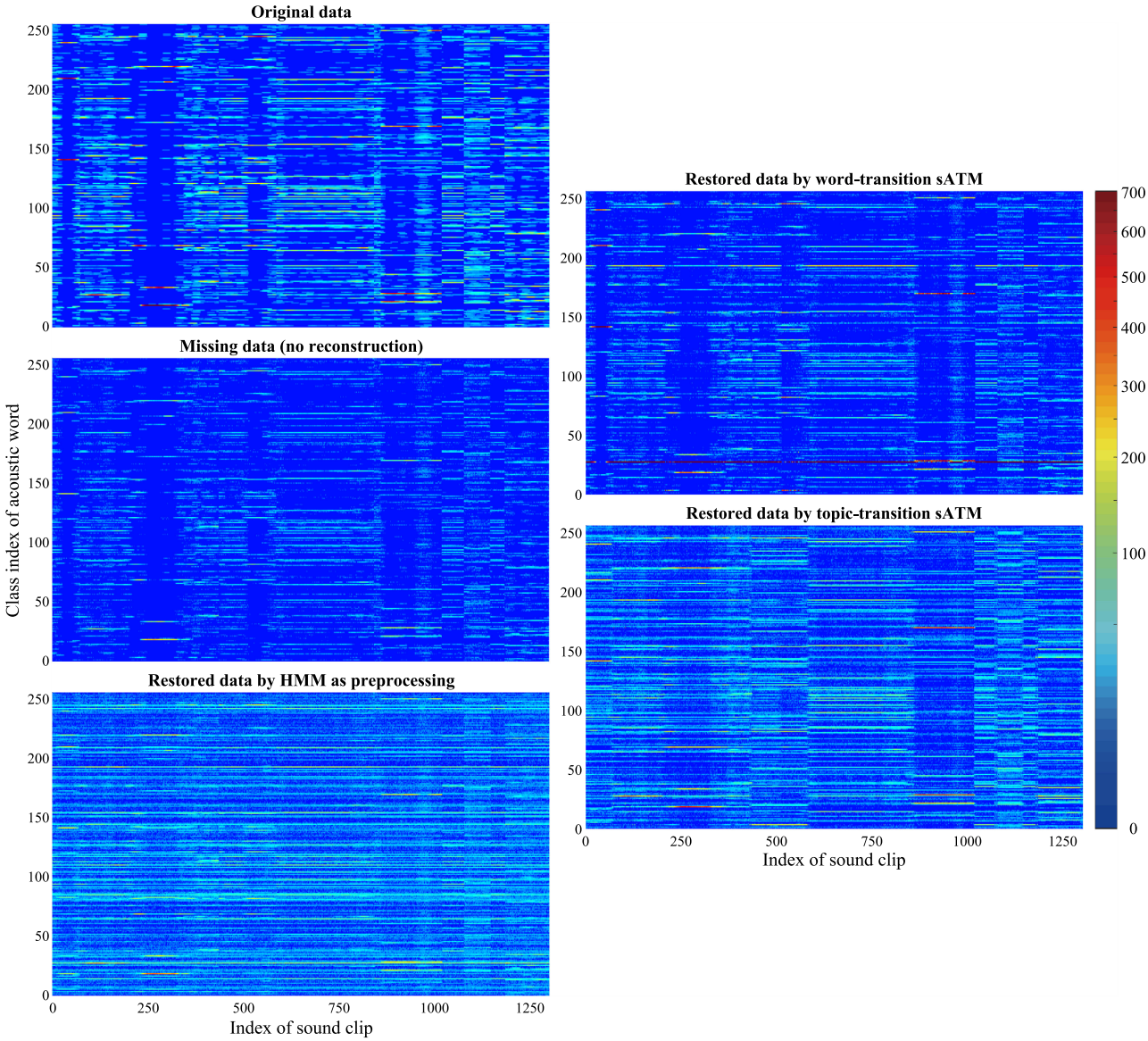


Fig. 15. Acoustic word histograms for the case of 80% missing data.

the proportion of missing words is 80%, many of the acoustic words cannot be identified. Even in this case, proposed methods can successfully reconstruct the histogram patterns of acoustic words, and the topic-transition sATM achieves a reasonably accurate reconstruction.

Table XIII shows the similarity between the original and estimated acoustic word histograms in terms of the cosine similarity and root mean squared error (RMSE). When the proportion of missing words is 40%, the acoustic word histograms are well reconstructed and have high similarity scores. Even when the proportion of missing words is 80%, the topic-transition sATM achieves high scores for both the cosine similarity and RMSE. Meanwhile, in the word-transition sATM, the RMSE between the original and reconstructed histograms is slightly higher than that for the HMM because the word-transition sATM mainly misreconstructed a single acoustic word (class index = 27 in Fig. 15). Nonetheless, the cosine distance score of the word-transition sATM is lower than that for the conventional methods and the missing acoustic words are well reconstructed.

TABLE XIII
SIMILARITY BETWEEN ORIGINAL AND RESTORED ACOUSTIC WORD HISTOGRAMS

Missing prop.	Similarity criterion	Method			
		Missing	**HMM**	**word-trans. sATM**	**topic-trans. sATM**
40%	Cosine similarity	0.959	0.960	0.965	0.962
	RMSE	7.489	5.281	4.201	5.194
80%	Cosine similarity	0.638	0.120	0.670	0.801
	RMSE	14.845	13.314	13.419	10.628

These results indicate that the proposed models can restore acoustic words that are strongly correlated with an acoustic scene in an acoustic word sequence.

VI. Conclusion

To estimate acoustic scenes and missing acoustic words, we have proposed novel models generating acoustic word sequences based on an acoustic topic model (ATM) that considers the temporal continuity of acoustic words or topics. In the proposed models, the temporal transition of acoustic words or topics is modeled in two ways: by modeling the temporal transition of acoustic words directly using a Markov process and by modeling the temporal transition of a hidden structure generating acoustic words such as an HMM. We then incorporate each transition model of acoustic words and topics in a process generating acoustic words based on the conventional sATM. Simultaneous modeling of the process generating acoustic words and the temporal transition enables acoustic scenes and missing acoustic words to be estimated by utilizing the mutual estimated information. We then introduced a parameter estimation method for the proposed models that is based on collapsed Gibbs sampling. Evaluation results of the proposed method indicate that it achieves a classification accuracy of acoustic scenes comparable to that obtained when there is no missing data. Additionally, the proposed models can estimate acoustic words that are strongly correlated with acoustic scenes in an acoustic word sequence.

Appendix

A. Calculation of Posterior Probability in Word-Transition sATM

In this section, detailed derivations of the posterior probabilities of the acoustic topic and word in the word-transition sATM are given.

1) Calculation of Eqs. (6)–(9): We first derive Eq. (6). $p(\boldsymbol{z},\boldsymbol{\theta}|\boldsymbol{a},\alpha)$ is calculated as

$$\begin{aligned} &p(\boldsymbol{z},\boldsymbol{\theta}|\boldsymbol{a},\alpha) \\ &= \prod_{a=1}^{A}\left[\left\{\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\prod_{t=1}^{T}(\theta_{a,t})^{\alpha-1}\right\}\left\{\prod_{s=1}^{S}\prod_{i=1}^{N_{\boldsymbol{w}s}}\prod_{t=1}^{T}(\theta_{a,t})^{\delta_{s,i,t}}\right\}\right] \\ &= \left(\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\right)^A\prod_{a=1}^{A}\left[\left\{\prod_{t=1}^{T}(\theta_{a,t})^{\alpha-1}\right\}\left\{\prod_{s=1}^{S}\prod_{i=1}^{N_{\boldsymbol{w}s}}\prod_{t=1}^{T}(\theta_{a,t})^{\delta_{s,i,t}}\right\}\right] \\ &= \left(\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\right)^A\prod_{a=1}^{A}\prod_{t=1}^{T}\left\{(\theta_{a,t})^{\alpha-1}\cdot(\theta_{a,t})^{\sum_{s=1}^{S}\sum_{i=1}^{N_{\boldsymbol{w}s}}\delta_{s,i,t}}\right\}, \end{aligned} \tag{24}$$

where $\delta_{s,i,t}$ is the Kronecker delta, which is 1 if $a_s = a$ and $z_{s,i} = t$. Considering $\sum_{s=1}^{S}\sum_{i=1}^{N_{\boldsymbol{w}s}}\delta_{s,i,t} = n_t^{(a)}$, Eq. (24) can be written as

$$p(\boldsymbol{z},\boldsymbol{\theta}|\boldsymbol{a},\alpha) = \left(\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\right)^A\prod_{a=1}^{A}\prod_{t=1}^{T}(\theta_{a,t})^{n_t^{(a)}+\alpha-1} \tag{25}$$

Here, we consider the following integral of the Dirichlet distribution:

$$\int\prod_j\theta_j^{\zeta_j-1}d\theta = \frac{\prod_j\zeta_j}{\Gamma(\sum_{k=1}^{K}\zeta_k)}. \tag{26}$$

Substituting Eqs. (25) and (26) into $\int p(\boldsymbol{z},\boldsymbol{\theta}|\boldsymbol{a},\alpha)d\boldsymbol{\theta}$, Eq. (6) is obtained. Similarly, Eqs. (8) and (9) are also obtained using the integral of the Dirichlet distribution as with the case of $\int p(\boldsymbol{z},\boldsymbol{\theta}|\boldsymbol{a},\alpha)d\boldsymbol{\theta}$.

2) Calculation of Eq. (10): To derivate Eq. (10) in the word-transition sATM, substituting Eqs. (6)–(9) into each term of Eq. (5) gives

$$\frac{p(\boldsymbol{z}|\boldsymbol{a},\alpha)}{p(\boldsymbol{z}_{\backslash s,i}|\boldsymbol{a},\alpha)} = \left(\frac{\Gamma(n_t^{(a)}+\alpha)}{\Gamma(n_{(\backslash s,i),t}^{(a)}+\alpha)}\right)\Bigg/\left(\frac{\Gamma(n_{\cdot}^{(a)}+T\alpha)}{\Gamma(n_{(\backslash s,i),\cdot}^{(a)}+T\alpha)}\right) \tag{27}$$

$$\frac{p(\boldsymbol{w}|\boldsymbol{z},\beta,\gamma)}{p(\boldsymbol{w}_{\backslash s,i}|\boldsymbol{z}_{\backslash s,i},\beta,\gamma)} = \frac{\frac{\Gamma(n_m^{(t)}+\beta)}{\Gamma(n_{(\backslash s,i),m}^{(t)}+\beta)}\frac{\Gamma(n_{m+}^{(m^-)}+\gamma)}{\Gamma(n_{(\backslash s,i),m+}^{(m^-)}+\gamma)}}{\frac{\Gamma(n_{\cdot}^{(t)}+M\beta)}{\Gamma(n_{(\backslash s,i),\cdot}^{(t)}+M\beta)}\frac{\Gamma(n_{\cdot}^{(m^-)}+M\gamma)}{\Gamma(n_{(\backslash s,i),\cdot}^{(m^-)}+M\gamma)}}. \tag{28}$$

Since $\Gamma(x+1)/\Gamma(x) = x$ for the gamma function, Eqs. (27) and (28) can be rewritten as

$$\frac{p(\boldsymbol{z}|\boldsymbol{a},\alpha)}{p(\boldsymbol{z}_{\backslash s,i}|\boldsymbol{a},\alpha)} = \frac{n_{(\backslash s,i),t}^{(a)}+\alpha}{n_{(\backslash s,i),\cdot}^{(a)}+T\alpha} \tag{29}$$

$$\begin{aligned} \frac{p(\boldsymbol{w}|\boldsymbol{z},\beta,\gamma)}{p(\boldsymbol{w}_{\backslash s,i}|\boldsymbol{z}_{\backslash s,i},\beta,\gamma)} &= \frac{n_{(\backslash s,i),m}^{(t)}+\beta}{n_{(\backslash s,i),\cdot}^{(t)}+M\beta}\cdot\frac{n_{(\backslash s,i),w_{s,i}}^{(w_{s,i-1})}+\gamma}{n_{(\backslash s,i),\cdot}^{(w_{s,i-1})}+M\gamma} \\ &\cdot\frac{n_{(\backslash s,i),w_{s,i+1}}^{(w_{s,i})}+\delta_{w_{s,i-1},w_{s,i}}\cdot\delta_{w_{s,i},w_{s,i+1}}+\gamma}{n_{(\backslash s,i),\cdot}^{(w_{s,i})}+\delta_{w_{s,i-1},w_{s,i}}+M\gamma}. \end{aligned} \tag{30}$$

Substituting these equations into Eq. (5) again, Eq. (10) is obtained.

B. Calculation of Posterior Probability in Topic-Transition sATM

Similarly to the word-transition sATM, substituting Eqs. (18)–(21) into each term of Eq. (17) and using $\Gamma(x+1)/\Gamma(x) = x$, we obtain

$$\begin{aligned} \frac{p(\boldsymbol{z}|\boldsymbol{a},\alpha,\gamma)}{p(\boldsymbol{z}_{\backslash s,i}|\boldsymbol{a},\alpha,\gamma)} &= \frac{n_{(\backslash s,i),t}^{(a)}+\alpha}{n_{(\backslash s,i),\cdot}^{(a)}+T\alpha}\cdot\frac{n_{(\backslash s,i),z_{s,i}}^{(z_{s,i-1})}+\gamma}{n_{(\backslash s,i),\cdot}^{(z_{s,i-1})}+T\gamma} \\ &\cdot\frac{n_{(\backslash s,i),z_{s,i+1}}^{(z_{s,i})}+\delta_{z_{s,i-1},z_{s,i}}\cdot\delta_{z_{s,i},z_{s,i+1}}+\gamma}{n_{(\backslash s,i),\cdot}^{(z_{s,i})}+\delta_{z_{s,i-1},z_{s,i}}+T\gamma} \end{aligned} \tag{31}$$

$$\frac{p(\boldsymbol{w}|\boldsymbol{z},\beta)}{p(\boldsymbol{w}_{\backslash s,i}|\boldsymbol{z}_{\backslash s,i},\beta)} = \frac{n_{(\backslash s,i),m}^{(t)}+\beta}{n_{(\backslash s,i),\cdot}^{(t)}+M\beta}. \tag{32}$$

Substituting these equations into Eq. (17) again, Eq. (22) is obtained.

REFERENCES

[1] Q. Jin, P. F. Schulam, S. Rawat, S. Burger, D. Ding, and F. Metze, "Event-based video retrieval using audio," in *Proc. INTERSPEECH*, 2012, pp. 2085–2088.

[2] T. Zhang and C. J. Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *IEEE Trans. Audio Speech Lang. Process.*, vol. 9, no. 4, pp. 441–457, May 2001.

[3] Y. Ohishi *et al.*, "Bayesian semi-supervised audio event transcription based on Markov Indian buffet process," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 3163–3167.

[4] J. Liang, L. Jiang, and A. Hauptmann, "Temporal localization of audio events for conflict monitoring in social media," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 1597–1601.

[5] A. Harma, M. F. McKinney, and J. Skowronek, "Automatic surveillance of the acoustic activity in our living environment," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2005, pp. 1–4.

[6] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "On acoustic surveillance of hazardous situations," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2009, pp. 165–168.

[7] R. Radhakrishnan, A. Divakaran, and P. Smaragdis, "Audio analysis for surveillance applications," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2005, pp. 158–161.

[8] T. Komatsu and R. Kondo, "Detection of anomaly acoustic scenes based on a temporal dissimilarity model," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 376–380.

[9] 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. Multimedia Expo*, 2009, pp. 1218–1221.

[10] P. Guyot, J. Pinquier, and R. André-Obrecht, "Water sound recognition based on physical models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 793–797.

[11] 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," in *Proc. 11th Int. Conf. Comput. Inf. Technol.*, 2008, pp. 294–299.

[12] K. Imoto and N. Ono, "Spatial cepstrum as a spatial feature using distributed microphone array for acoustic scene analysis," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 6, pp. 1335–1343, Jun. 2017.

[13] J. Schröder, J. Anemüller, and S. Goetze, "Classification of human cough signals using spectro-temporal Gabor filterbank features," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 6455–6459.

[14] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, "CLEAR evaluation of acoustic event detection and classification systems," in *Proc. Int. Eval. Workshop Classification Events, Activities Relationships, Springer, Berlin Heidelberg*, 2007, pp. 311–322.

[15] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real life recordings," in *Proc. 18th Eur. Signal Process. Conf.*, 2010, pp. 1267–1271.

[16] S. Adavanne and T. Virtanen, "A report on sound event detection with different binaural features," DCASE2017 Workshop, Tech. Rep., 2017.

[17] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 16–34, May 2015.

[18] A. Eronen *et al.*, "Audio-based context recognition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 1, pp. 321–329, Jan. 2006.

[19] Y. Han, J. Park, and K. Lee, "Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification," DCASE2017 Workshop, Tech. Rep., 2017.

[20] H. Jallet, E. Çakır, and T. Virtanen, "Acoustic scene classification using convolutional recurrent neural networks," DCASE2017 Workshop, Tech. Rep., 2017.

[21] T. Heittola, A. Mesaros, A. Eronen, and A. Klapuri, "Audio content recognition using audio event histograms," in *Proc. 18th Eur. Signal Process. Conf.*, 2010, pp. 1272–1276.

[22] G. Guo and S. Z. Li, "Content-based audio classification and retrieval by support vector machines," *IEEE Trans. Neural Netw.*, vol. 14, no. 1, pp. 209–215, Jan. 2003.

[23] K. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, "An introduction to kernel-based learning algorithms," *IEEE Trans. Neural Netw.*, vol. 12, no. 2, pp. 181–201, Mar. 2001.

[24] 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, Aug. 2010.

[25] S. Kim, S. Narayanan, and S. Sundaram, "Acoustic topic models for audio information retrieval," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2009, pp. 37–40.

[26] 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, Oct. 2016.

[27] K. Imoto, Y. Ohishi, H. Uematsu, and H. Ohmuro, "Acoustic scene analysis based on latent acoustic topic and event allocation," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, 2013, pp. 1–6.

[28] K. Imoto and N. Ono, "Acoustic scene analysis from acoustic event sequence with intermittent missing event," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 156–160.

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

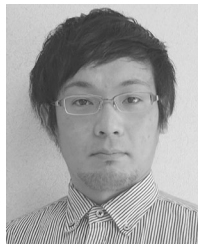
[30] T. L. Griffiths and M. Steyvers, "Finding scientific topics," in *Proc. Nat. Acad. Sci.*, 2004, vol. 1, pp. 5228–5235.

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

[32] R. M. Neal, "Probabilistic inference using Markov chain Monte Carlo methods," Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, Tech. Rep. CRG-TR-93-1, 1993.

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

[34] T. P. Minka and J. Lafferty, "Expectation propagation for the generative aspect model," in *Proc. 18th Conf. Uncertainty Artif. Intell.*, 2002, pp. 352–359.



Keisuke Imoto (M'12) received the B.E. and M.E. degrees from Kyoto University, Kyoto, Japan, in 2008 and 2010, respectively, and the Ph.D. degree from The Graduate University for Advanced Studies, Hayama, Japan, in 2017. He was the Nippon Telegraph and Telephone Corporation, in 2010. In 2017, He moved to Ritsumeikan University as an Assistant Professor. He has been engaged in research on acoustic event detection, acoustic scene analysis, and microphone array signal processing. He is a member of the IEEE Signal Processing Society and the Acoustical Society of Japan (ASJ). He was the recipient of the Awaya Award from ASJ, in 2013, and the TAF Telecom System Technology Awards, in 2018.

Nobutaka Ono (SM'13) received the B.E., M.S., and Ph.D degrees in mathematical engineering and information physics from the University of Tokyo, Tokyo, Japan, in 1996, 1998, 2001, respectively. He was the Graduate School of Information Science and Technology, the University of Tokyo, Japan, in April 2001, as a Research Associate and became a Lecturer, in April 2005. He moved to the National Institute of Informatics, Japan, as an Associate Professor, in April 2011, and became a Professor, in September 2017. He moved to Tokyo Metropolitan University in October 2017. He is the author or co-author of more than 180 articles in international journal papers and peer-reviewed conference proceedings. He was a Tutorial speaker, ISMIR 2010 and ICASSP 2018, a special session chair, in EUSIPCO 2013, 2015, 2017, and 2018, a chair of Signal Separation Evaluation Campaign evaluation committee, in 2013 and 2015. He was an Associate Editor for the IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING during 2012 to 2015. He has been a member of the IEEE AUDIO AND ACOUSTIC SIGNAL PROCESSING Technical Committee since 2014. His research interests include acoustic signal processing, specifically, microphone array processing, source localization and separation, machine learning and optimization algorithms for them. He is a senior member of the IEEE Signal Processing Society, and a member of the Acoustical Society of Japan (ASJ), the Institute of Electronics, Information and Communications Engineers, the Information Processing Society of Japan, and the Society of Instrument and Control Engineers (SICE) in Japan. He was the recipient of the Sato Paper Award and the Awaya Award from ASJ, in 2000 and 2007, respectively, the Igarashi Award at the Sensor Symposium on Sensors, Micromachines, and Applied Systems from IEEJ, in 2004, the best paper award from the IEEE ISIE, in 2008, Measurement Division Best Paper Award from SICE, in 2013, the best paper award from the IEEE IS3C, in 2014, the excellent paper award from IIHMSP, in 2014, the unsupervised learning ICA pioneer award from SPIE.DSS, in 2015, the Sato Paper Award from ASJ, and two TAF Telecom System Technology Awards, in 2018.