Acoustic Scene Analysis Based on Hierarchical Generative Model of Acoustic Event Sequence

Keisuke IMOTO(a), Nonmember and Suehiro SHIMAUCHI(b), Member

SUMMARY We propose a novel method for estimating acoustic scenes such as user activities, e.g., “cooking,” “vacuuming,” “watching TV,” or situations, e.g., “being on the bus,” “being in a park,” “meeting,” utilizing the information of acoustic events. There are some methods for estimating acoustic scenes that associate a combination of acoustic events with an acoustic scene. However, the existing methods cannot adequately express acoustic scenes, e.g., “cooking,” that have more than one subordinate category, e.g., “frying ingredients” or “plating food,” because they directly associate acoustic events with acoustic scenes. In this paper, we propose an acoustic scene estimation method based on a hierarchical probabilistic generative model of an acoustic event sequence taking into account the relation among acoustic scenes, their subordinate categories, and acoustic event sequences. In the proposed model, each acoustic scene is represented as a probability distribution over their unsupervised subordinate categories, called “acoustic sub-topics,” and each acoustic sub-topic is represented as a probability distribution over acoustic events. Acoustic scene estimation experiments with real-life sounds showed that the proposed method could correctly extract subordinate categories of acoustic scenes.

key words: acoustic signal processing, acoustic event detection, acoustic scene analysis, bayesian inference

1. Introduction

Much interest has been expressed recently in applications such as monitoring elderly people [1], security surveillance [2], automatic classification of user activities and contexts [3], [4], and multimedia retrieval [5], which utilizes the information that can be obtained by analyzing various acoustic signals. There are some significant techniques for these applications; acoustic event detection (AED) analyzes various types of sounds, e.g., “footsteps,” “running water,” “music,” “voice,” to detect specific types of sounds [6], [7], and acoustic scene analysis (ASA) analyzes acoustic scenes such as user activities, e.g., “cooking,” “vacuuming,” “watching TV,” or situations, e.g., “being on the bus,” “being in a park,” “meeting,” on the basis of the information of acoustic events [8]. In this paper, we focus on automatic estimation of acoustic scenes, which enables automatic life-logging, the monitoring of elderly people, or multimedia retrieval.

One of the most fundamental approaches for analyzing acoustic scenes is focusing on a single acoustic event that obviously characterizes a certain acoustic scene and detecting this noticeable acoustic event. However, many acoustic scenes are characterized not by a single acoustic event but by a combination of multiple acoustic events. For example, an acoustic scene, “cooking,” is marked by a combination of acoustic events including “running water,” “cutting with a knife,” and “heating a skillet.” On the basis of this consideration, Ahaikh et al. [9] proposed a supervised user activity estimation method using a combination of multiple acoustic events. Heittola et al. [10] also proposed a content recognition method using a combination of multiple acoustic events.

An alternative strategy for acoustic scene analysis is focusing on a model of generating acoustic events from acoustic scenes. With a simple supervised generative model of acoustic events, it can be assumed that each acoustic scene has a probability distribution of acoustic events directly, and each acoustic event is generated from its distribution. We call this model the “acoustic scene model (ASM).” Kim et al. [11], Lee et al. [12], and Imoto et al. [13], [14] proposed another unsupervised generative model for analyzing acoustic scenes on the basis of the “acoustic topic model (ATM).” The ATM is an unsupervised generative model that generates acoustic event sequences from acoustic scenes through the latent structure of the acoustic scene lying in a combination of acoustic events, which is called the “acoustic topic.” Specifically, the ATM characterizes acoustic scenes as a distribution of acoustic topics without utilizing the information of acoustic scenes explicitly, and this enables unsupervised analysis of acoustic scenes.

We point out here that some acoustic scenes have more than one subordinate category of an acoustic scene as shown in Fig.1, and therefore, directly characterizing acoustic scenes with combinations of multiple acoustic events such as with the ASM or ATM is not always as expressive in modeling and estimating these acoustic scenes. For example, we consider that the acoustic scene “cooking” is marked by its subordinate categories including “preparation of ingredients,” “frying ingredients,” and “plating food,” and each of the subordinate categories is characterized by a combination of acoustic events such as “opening a fridge,” “running water,” and “cutting with a knife.”

In this paper, we propose an acoustic scene estimation method for an acoustic scene that has more than one subordinate category of acoustic scenes. In the proposed method, we model the generative process of acoustic event sequences as a hierarchical structure assuming that an acoustic scene is represented as a probabilistic distribution of their sub-
ordinate categories, and each subordinate category of the acoustic scene is described as the probabilistic distribution of acoustic events. We call this model the “acoustic scene and sub-topic model (ASTM).”

The rest of this paper is as follows. In Sect. 2, we introduce the conventional probabilistic generative models of acoustic event sequences for acoustic scene analysis. In Sect. 3, we describe our proposed probabilistic generative model of an acoustic event sequence with acoustic scenes and their subordinate categories. In Sect. 4, we present an acoustic scene estimation method that includes acoustic event estimation and the parameter estimation method of our proposed model. In Sect. 5, we discuss the results of the experiments, and Sect. 6 concludes this paper.

2. Conventional Generative Model of Acoustic Event Sequence

In this section, we discuss two models for generating acoustic event sequences for associating acoustic scenes and combinations of acoustic events. Here, definitions of variables used in this paper are listed in Table 1.

2.1 Acoustic Scene Model

When modeling the relation between an acoustic scene and an acoustic event sequence with a single step generative process, it can be assumed that the acoustic scene has a probability distribution over the acoustic events, and each acoustic event is generated following its probability distribution. For instance, the ASM models a generative process of acoustic event sequences using acoustic scene information explicitly. In the ASM, the generative process of each acoustic event sequence \( e_s \) of acoustic signal index \( s \) (= 1, 2, ..., \( S \)) can be described as shown in Fig. 2. In this model, possible acoustic scenes in \( s \) are given in advance, and one of them is generated in each time frame. Each acoustic scene has the distribution \( \phi_{a_s} \) over acoustic events, and each acoustic event \( e_{si} \) is generated from this distribution. That is, acoustic scenes and events are associated through this distribution \( \phi_{a_s} \). Moreover, each acoustic event distribution has a Dirichlet prior \( \text{Dir}(\beta) \), which is intended to avoid overfitting to a dataset. In practice, the generative process of the ASM can be expressed as follows.

**Possible acoustic scenes \( a_s \) are given**

Iterate \( a \) over # of acoustic scene categories \( A \)

1. Choose \( \phi_{a_s} \sim \text{Dir}(\beta) \)

Iterate \( s \) over # of acoustic event sequences \( S \)

Iterate \( i \) over # of events in each event sequence \( N_{e_i} \)
Acoustic scene estimation methods that focus on the latent structure of an acoustic scene lying in acoustic event sequences have also been proposed, and these are called the “acoustic topic model (ATM)” [11]–[13]. In the ATM, acoustic scenes are represented as a distribution of latent variables called “acoustic topics,” and a generative process of acoustic event sequences $e_s$ is modeled not by using acoustic scenes directly but through distributions of acoustic topics as shown in Fig. 3. In the ATM, each acoustic event sequence $e_s$ has a distribution $\theta_s$ of acoustic topics, and each acoustic topic $z_{si}$ is generated from this distribution. Each topic also has the distribution $\phi_z$ over acoustic events, and each acoustic event $e_{si}$ is then generated from this distribution. That is, in this model, acoustic scenes and events are associated through the distributions $\theta_s$ and $\phi_z$. Moreover, each acoustic topic distribution and event distribution has a Dirichlet prior $\text{Dir}(\eta)$ and $\text{Dir}(\beta)$, which are also intended to avoid overfitting to a dataset. For instance, the generative process of the ATM can be represented as follows.

Iterate $t$ over # of acoustic topic categories $T$
1. Choose $\phi_t \sim \text{Dir}(\beta)$

Iterate $s$ over # of acoustic event sequences $S$
2. Choose $\theta_s \sim \text{Dir}(\eta)$

Iterate $i$ over # of events in each event sequence $N_{esi}$
3. Choose $z_{si} | \theta_t \sim \text{Mult}(\theta_t)$
4. Choose $e_{si} | \phi_{z_{si}}, z_{si} \sim \text{Mult}(\phi_{z_{si}})$.

The ATM models acoustic scenes through the distribution of acoustic topics characterized by the occurrence and co-occurrence frequency of acoustic events without the category labels of acoustic scenes. Therefore, the ATM can model acoustic scenes more flexibly, and there is a possibility that it can fit the model to the subordinate categories of acoustic scenes.

To classify acoustic scenes using the ATM, Kim et al. [20] estimated the distributions of acoustic topics for each acoustic scene and then classified acoustic scenes on the basis of a multiclass SVM that utilized acoustic scene labels and the parameters of acoustic topic distributions.

2.3 Discussion

As discussed above, the ASM is a supervised learning model of the relation between acoustic scenes and acoustic event sequences. In comparison, the ATM is an unsupervised one for modeling the latent structure of acoustic scenes lying in acoustic event sequences.

Some acoustic scenes have more than one subordinate category as shown in Fig. 1; however, it is not easy to arrange a data set that has all subordinate category labels of acoustic scenes. Therefore, it is impractical to model these subordinate categories of acoustic scenes using the ASM. Meanwhile, the ATM can model acoustic scenes more flexibly, and it may be able to fit the model to the subordinate categories of acoustic scenes. However, it cannot utilize the information of acoustic scenes explicitly, and therefore, we cannot control the model fitting to the subordinate categories of acoustic scenes. From these characteristics, the ASM and ATM are insufficient for modeling acoustic scenes that have more than one subordinate category, and they sometimes degrade the estimation performance of acoustic scenes.

3. Proposed Hierarchical Generative Model of Acoustic Event Sequence

To address the problems in conventional models, we propose a hierarchical generative model that can associate acoustic event sequences with acoustic scenes and subordinate categories of acoustic scenes as shown in Fig. 4, and we call this model the “acoustic scene and sub-topic model (ASTM).” In the ASTM, to model an acoustic scene, we apply the acoustic scene information given by acoustic scene labels, and we treat the subordinate categories of the acoustic scenes as latent variables to capture automatically underlying structures of acoustic events.

As shown in Fig. 4, in the ASTM, possible acoustic scenes in $s$ are given in advance, and one of them is assigned in each time frame. The assigned acoustic scene $x_{si}$ has the distribution $\theta_x$ of sub-topics, and each sub-topic $z_{si}$ is then generated from this distribution. This sub-topic also has the distribution $\phi_{z_{si}}$ over acoustic events, and each acoustic event $e_{si}$ is generated from this distribution. As with the
ASM or ATM, each sub-topic distribution and event distribution has a Dirichlet prior Dir(α) and Dir(β), which are intended to avoid overfitting. In practice, the following generative process is assumed with the ASTM.

**Possible acoustic scenes** $a_S$ **are given**

**Iterate $a$ over # of acoustic scene categories $A$**
1. Choose $θ_a$ ~ Dir($α$)

**Iterate $t$ over # of acoustic sub-topic categories $T$**
2. Choose $φ_t$ ~ Dir($β$)

**Iterate $s$ over # of acoustic event sequences $S$**

3. Choose $x_{si}$ | $a_s$ ~ Uni
4. Choose $z_{si}$ | $θ_{xa}$, $x_{si}$ ~ Mult($θ_{xa}$)
5. Choose $e_{si}$ | $φ_{zs}$, $z_{si}$ ~ Mult($φ_{zs}$).

Since acoustic scene labels are available for acoustic scene modeling explicitly, the ASTM has an advantage in that it can tailor the subordinate category layer modeling to fit each acoustic scene. This hierarchical modeling offers the combined benefits of the ASM and ATM to our proposed model.

The generative probability of $e_i$: $p(e_i | α, β, a_s)$ can be written as

$$p(e_i | α, β, a_s) = \prod_{s=1}^{N_s} p(e_{si} | α, β, a_s)$$

$$= \prod_{s=1}^{N_s} \int \prod_{t=1}^{A_s} \sum_{a_t} \sum_{r=1}^{T} \text{Uni}(x_{si}|a_s)p(z_{si}|x_{si}, θ_{xa})\text{Dir}(θ_{xa}|α)$$

$$\cdot p(e_{si}|z_{si}, φ_{zs})\text{Dir}(φ_{zs}|β) dφ dθ$$

$$= \prod_{s=1}^{N_s} \int \prod_{a_t} \frac{Γ(A_s)}{Γ(α)} \prod_{r=1}^{T} \theta_{xa}^{α-1} \sum_{z_{si}} p(z_{si}) \cdot \int \prod_{m=1}^{M} \frac{Γ(M)}{Γ(β)} \prod_{r=1}^{T} φ_{zs}^{β-1} dφ dθ,$$

where $A_s$ is the number of possible acoustic scenes in each acoustic event sequence, $φ_{zs}$ is the generative probability of $e_{si}$ in $z_{si}$, and $θ_{xa}$ is the generative probability of $z_{si}$ in $x_{si}$.

4. **Acoustic Scene Estimation Based on ASTM**

4.1 **Acoustic Scene Estimation Overview**

In this section, we introduce an acoustic scene estimation method with the ASTM. To estimate the acoustic scene that an acoustic event sequence $e_i^∗$ indicates with the ASTM, we need to infer an acoustic scene that maximizes the posterior probability $p(x_{si} = a[e_i^∗, γ])$. As shown in Fig. 5, the acoustic scene estimation system has a three-step process: recognizing acoustic event sequences from acoustic signals, estimating acoustic scene model parameters, and predicting acoustic scenes with estimated model parameters. We now discuss each step as follows.

4.2 **Acoustic Feature Extraction and Event Recognition**

To estimate $e_i$ or $e_i^∗$, we segment long-term input signals and extract acoustic feature vectors frame by frame. Then, we recognize acoustic events in every frame and output an $e_i$ or $e_i^∗$ (composed of multiple acoustic event recognition results).

As the acoustic features, we choose Mel-frequency cepstral coefficients (MFCCs). These are well known features for speech recognition and speech synthesis, and MFCC features have also been used in AED tasks [6], in which they performed well.

Many acoustic event recognition approaches are based on the Gaussian mixture model (GMM) or hidden Markov model (HMM) [6], [7], which defines acoustic events with manually labeled data. However, it is difficult to label every possible acoustic event since there are many different types of sounds. In view of this difficulty, our proposed acoustic scene model represents acoustic scenes and sub-topics as distributions over multiple acoustic sub-topics and events, respectively, so there is no need to label each acoustic event for acoustic scene modeling. Therefore, we apply the unsupervised technique of Gaussian clustering (GC) to model acoustic events. MFCC feature vectors in the learning set for acoustic event modeling are classified by

![Graphical model representation of proposed method (acoustic scene and sub-topic model: ASTM)](image)

![Acoustic scene estimation system overview](image)
using GC with the expectation-maximization (EM) algorithm [17], and each Gaussian component is modeled as a single acoustic event. To recognize acoustic events, we estimate acoustic events that maximize posterior probability for the model, and the acoustic event recognition step outputs the acoustic event sequence of multiple event recognition results.

4.3 ASTM Parameter Estimation

The ASTM includes model parameters $x$, $z$, $\theta_e$, $\phi$, to be estimated for analyzing acoustic scenes. In this paper, we introduce a parameter estimation method that uses the collapsed variational Bayes with 0th-order approximation (CVB0) [18]. The Bayesian inference based on CVB0 shows a lower calculation cost than collapsed Gibbs sampling (CGS) and naive variational Bayes (VB) [19] and a slightly worse performance than that of CGS [20], [21]. Since we give weight to the balance between a calculation cost and parameter estimation performance, we thus apply CVB0 to estimate parameters of the ASTM.

To infer the model parameters of the ASTM, we need to estimate the posterior distribution of all unknown parameters for joint distribution. However, it is intractable to estimate a true posterior distribution $p(x, z, \theta, \phi | e)$. Therefore, we introduce a variational distribution $q(x, z, \theta, \phi)$ and approximate it to a posterior distribution iteratively. In CVB0, we first calculate a lower bound on a marginal log likelihood $\mathcal{F}[q]$ by using Jensen’s inequality.

$$L(e) \equiv \log p(e | \alpha, \beta, a_s)$$

$$= \log \int \sum_x \sum_z q(x, z, \theta, \phi) \frac{p(e, x, z, \theta, \phi | x, z, \theta, \phi)}{q(x, z, \theta, \phi)} d\theta d\phi$$

$$\geq \int \sum_x \sum_z q(x, z, \theta, \phi) \log \frac{p(e, x, z, \theta, \phi | x, z, \theta, \phi)}{q(x, z, \theta, \phi)} d\theta d\phi$$

$$= \mathcal{F}[q].$$

Here, $L(e)$ and $\mathcal{F}[q]$ can be represented as the following relational expression using the KL divergence.

$$L(e) - \mathcal{F}[q] = \log p(e | \alpha, \beta, a_s) - \int \sum_x \sum_z q(x, z, \theta, \phi)$$

$$\cdot \log \frac{p(e, x, z, \theta, \phi | x, z, \theta, \phi)}{q(x, z, \theta, \phi)} d\theta d\phi$$

$$= \int \sum_x \sum_z q(x, z, \theta, \phi) \log \frac{p(e, x, z, \theta, \phi | x, z, \theta, \phi)}{p(x, z, \theta, \phi | x, z, \theta, \phi, \alpha, \beta, a_s)} d\theta d\phi$$

$$= KL(q(x, z, \theta, \phi) || p(x, z, \theta, \phi | x, z, \theta, \phi, \alpha, \beta, a_s)).$$

This means that we can obtain the appropriate variational parameters by maximizing the lower bound $\mathcal{F}[q]$, that is, minimizing the KL divergence. For this $\mathcal{F}[q]$, we apply the following mean field approximation.

$$q(x, z, \theta, \phi) = q(\theta, \phi | x, z)q(x, z)$$

$$= p(\theta, \phi | e, x, z, \alpha, \beta, a_s)q(x, z).$$

Here, $q(\theta, \phi | x, z)$ is defined with no restriction, and therefore, $\mathcal{F}[q]$ is maximized when $q(\theta, \phi | x, z) = p(\theta, \phi | e, x, z, \alpha, \beta, a_s)$. By substituting Eq. (4) into Eq. (2), we obtain $\mathcal{F}[q]$ as follows.

$$\mathcal{F}[q] = \int \sum_x \sum_z q(x, z, \theta, \phi)$$

$$\cdot \log \frac{p(e, x, z | \alpha, \beta, a_s)p(\theta, \phi | e, x, z, \alpha, \beta, a_s)}{q(x, z, \theta, \phi)} d\theta d\phi$$

$$= \sum_x \sum_z q(x, z) \int \log \frac{p(e, x, z | \alpha, \beta, a_s)}{q(x, z, \theta, \phi)} d\theta d\phi$$

$$= \sum_x \sum_z q(x, z) \log p(e, x, z | \alpha, \beta, a_s)$$

$$- \sum_x \sum_z q(x, z) \log q(x, z).$$

Considering that $p(x = a | a_s)$ is assumed to be a uniform distribution and that the joint distribution of the Dirichlet distribution and the multinomial distribution can be represented by using the gamma function, the first term of the right side in Eq. (5) can be calculated as follows.

$$\sum_x \sum_z q(x, z) \log p(e, x, z | \alpha, \beta, a_s)$$

$$= \sum_a \sum_x \sum_z q(x, z | a) \log \frac{\prod_i (\Gamma(T_a) \Gamma(\alpha + n_{s,i}^a))}{\Gamma(T_a + \alpha + n_{s,i}^a)} \prod_i \Gamma(M \beta + n_{s,i}^a) \prod_a \Gamma(M \beta + n_{s,i}^a) \prod_i \Gamma(M \beta + n_{s,i}^a).$$

where $q(x_{si}, z_{si}) = \gamma_{s,e,si:a}^\alpha$ and $n_{s,si:a}$ are the expected probability and the number of acoustic events in acoustic event sequence $s$ and event $e_{si}$ assigned to acoustic scene $a$ and sub-topic $t$. The symbol “$\cdot$” in the subscript indicates the summation in the corresponding variable. Detailed derivations of Eqs. (6), (8), and (10) are shown in Appendix 3. Then, the second term of the right side in Eq. (5) can be also calculated as follows.

$$\sum_x \sum_z q(x, z) \log q(x, z)$$

$$= \sum_x \sum_z \prod_i \prod_a q(x_{si}, z_{si}) \log \prod_i \prod_a q(x_{si}, z_{si})$$

$$= \sum_x \sum_z \prod_i \prod_a \gamma_{s,e,si:a}^{\alpha} \sum_i \sum_a \sum_{a_s} \log \gamma_{s,e,si:a}^{\alpha}.$$
\[ \frac{\partial}{\partial \gamma_{se,at}} \sum_x \sum_z q(x, z) \log p(x, z) \]
\[ = \frac{\partial}{\partial \gamma_{se,at}} \sum_x \sum_z \prod_{i} \prod_{j} \gamma^{s_i \rightarrow j}_{e_{ij},e_{ij}} + \log \gamma_{se,at} \]
\[ \propto \log \gamma_{se,at} + \text{const.} \]  

(9)

Substituting Eqs. (8) and (9) into \( \frac{\partial F}{\partial y_{se,at}}(x_i, z_i) = 0 \) and maximizing \( F(x_i, z_i) \) for \( \gamma_{se,at} \) [22], we obtain the following update.

(10)

Because the computation cost of each expectation in \( \gamma_{se,at} \) is expensive, for the ASM, we apply the Taylor series approximation to \( E_{q(s_i, z_i)}[\log(\alpha + n^{s_i}_{at})] \) at \( E_{q(s_i, z_i)}[n^{s_i}_{at}] \) as follows.

(11)

Furthermore, we utilize the zero-th order term of the Taylor series and obtain the following approximation.

(12)

where we consider that \( E_{q(s_i, z_i)}[n^{s_i}_{at}] \approx \sum_{x,z} q(x,z) \gamma_{e_{ij},e_{ij}} \) when \( n^{s_i}_{at} \gg 0 \) [18]. By applying the same approximation to \( E_{q(s_i, z_i)}[\log(\beta + n^{s_i}_{at})] \) and \( E_{q(s_i, z_i)}[\log(\gamma_{se,at})] \), we finally obtain the following update for ASM.

(13)

To estimate appropriate parameters, this update is repeated until a convergence condition is satisfied. After this iteration, we can estimate \( \theta_{at} \) and \( \phi_{tm} \) in the following equations

(14)

(15)

4.4 Acoustic Scene Estimation

To estimate acoustic scenes for acoustic signal \( s^* \), for which possible acoustic scenes are unlabeled, we need to calculate \( \arg \max_a p(a|e^*_a) \) as follows.

(16)

where \( a \in a_S \) is all possible acoustic scenes for all data sets. This is achieved by the MAP estimation of \( \gamma_{se,at}^* \) using the trained ASTM parameters. In practice, we maximize \( q(x^*, z^*) \) in the following equation in a similar way described above.

(17)

Referring to the derivation of the ASM parameters in Sect. 4.3, we can finally obtain the update for the \( \gamma_{se,at}^* \).

(18)

In expectation of the \( \gamma_{se,at}^* \) of the maximum likelihood, this update is repeated until a convergence condition is satisfied.

5. Experiment

5.1 Experimental Conditions

We evaluated the acoustic scene estimation performance for the proposed and conventional models with two kinds of sound data sets, both of which were recorded in several real-life situations. As one data set, we recorded 11,105 sounds in a living room (22.8 m², reverberation time: 0.31 sec.), which included 9 categories of user activities: “chatting,” “cooking,” “eating dinner,” “operating PC,” “reading a newspaper,” “vacuuming,” “walking,” “washing dishes,” and “watching TV.” We separated them into 9,802 sounds for parameter training and 1,303 sounds for evaluation. As
Table 2 Experimental conditions

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling rate/quantization</td>
<td>16 kHz/16 bits</td>
</tr>
<tr>
<td>Frame size/shift</td>
<td>512/256</td>
</tr>
<tr>
<td>Acoustic event class size</td>
<td>32–512</td>
</tr>
<tr>
<td>Acoustic topic class size</td>
<td>20, 50, 100</td>
</tr>
<tr>
<td>Acoustic event sequence size</td>
<td>16s, 1,000 frames</td>
</tr>
<tr>
<td>Hyperparameter $\alpha / \beta$</td>
<td>3.3333/0.1</td>
</tr>
<tr>
<td>Iteration number of CVB0</td>
<td>1,000</td>
</tr>
</tbody>
</table>

The other data set, we recorded 25,277 sounds which included 8 categories of situations: “bicycle,” “bus,” “car,” “meeting,” “office,” “park,” “street,” and “train.” We separated them into 21,461 sounds for parameter training and 3,816 sounds for evaluation. In the acoustic event recognition stage, 12-dimensional MFCC features were calculated from every segmented acoustic signal with 50% overlap, and acoustic event models were learned by using GC with 32–512 acoustic event classes. Each $e_s$ was calculated from a 16-s signal that included 1,000 acoustic events. The experimental conditions for the acoustic scene estimation are listed in Table 2.

5.2 Generalization Performance

We evaluated the generalization performance of the ASTM, ASM, and ATM using perplexity. Perplexity evaluates how well a model predicts a data set or acoustic scenes, and a lower perplexity indicates a better generalization performance. In particular, the perplexity for the ASTM was calculated as follows.

$$\text{Perplexity}(e_s|a_s) = \exp \left[ -\frac{1}{N_e} \sum_x \sum_z \sum_{\theta} p(x|\theta)p(z|\theta)p(e|z)p(\phi|\beta)p(e_s|\phi,z) \right]$$

For perplexity with the ASM and ATM, we calculated it in a manner similar to [23]. For the perplexities with the ASTM and ATM, we chose $T = 20, 50, 100$ as the number of acoustic topics, and for all models, we used the same test set including the above nine categories of user activities and eight categories of situations. As shown in Fig. 6, the ASTM decreased the perplexity more compared with the ASM and ATM in all number of classes of acoustic events; thus, the ASTM achieves better generalization performance in modeling acoustic scenes and sub-topics with the consequent $e_s$ since it can flexibly model acoustic scenes that have subordinate categories included with the acoustic scene information and the latent structure of acoustic event sequences.

5.3 Acoustic Scene Estimation Accuracy

The averaged acoustic scene estimation accuracies (F-measure) of the nine categories of user activities and eight categories of situations are shown in Figs. 7 and 8. Here, the F-measure is calculated by using the harmonic mean of the precision and recall as follows.

$$\text{F-measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

In Fig. 7, the estimation accuracy for the ASTM increased about three points compared with the ASM. For the best
results in the nine categories of user activities ($T = 100$, $M = 512$), the ASTM correctly classified 84.5% of the test data set, while the ASM ($M = 512$) and ATM ($T = 100$, $M = 512$) correctly classified 77.9% and 73.0%. For the eight categories of acoustic scenes, the estimation accuracy for the ASTM increased about two points with the ASM, and for the best results, the ASTM ($T = 100$, $M = 512$) correctly classified 85.8% of the test data set, while the ASM ($M = 512$) and ATM ($T = 50$, $M = 256$) correctly classified 83.7% and 80.2%. These results suggest that some acoustic scenes had more than one subordinate category of acoustic scenes; therefore, the ASTM offers advantages in modeling and estimating those kinds of acoustic scenes compared with the ASM.

5.4 Acoustic Sub-Topic Estimation

For detailed investigation, we compared sub-topic estimation results in some acoustic scenes with actual subordinate categories of their acoustic scenes. Figures 9 and 10 show the acoustic sub-topic estimation results of “cooking” with a test sound with the ASTM and ATM, respectively, and Figs. 11 and 12 indicate the estimation results for “bus” when we set $T = 50$, $M = 512$. Here, there is no acoustic topic variables for the ASM, and therefore, we cannot analyze acoustic topics with the ASM. The subordinate categories of acoustic scenes described in the upper part of the figures indicate actual manually labeled categories, and the color-coded acoustic signals denote the correspondence relationships between acoustic event sequences and acoustic sub-topics estimated by using the ASTM and ATM. In the case of the estimation result of “cooking,” for both models, the estimated acoustic sub-topics agreed well with the actual subordinate categories of acoustic scenes, and these results show that the ASTM and ATM can generally extract subordinate categories of acoustic scenes. For instance, in Fig. 9,
“preparation of ingredients” and “frying ingredients” agreed well with sub-topics 2, 3, and 8. However, in the case of the estimation result for “bus,” the ASTM extracted subordinate categories of acoustic scenes through less acoustic sub-topics more precisely than did the ATM. In Fig. 11, “running bus” and “stopping at bus stop” agreed well with sub-topics 4, 5, and 2, whereas in Fig. 12, they are sometimes confused. This result suggests that the proposed method can extract not only acoustic scenes but also their subordinate categories more precisely.

6. Conclusion and Future Work

In this paper, we proposed an acoustic scene estimation method based on a hierarchical generative model of acoustic event sequences that is associated with acoustic scenes and their sub-topics, in which each acoustic scene is represented by a probability distribution over acoustic sub-topics, and each acoustic sub-topic is represented as the probability distribution over acoustic events. We also proposed a parameter estimation method for the proposed model based on CVB0 and an acoustic scene estimation method that has a three-step process: recognizing frame-by-frame acoustic events to generate an acoustic event sequence, estimating acoustic scene model parameters, and predicting acoustic scenes in unlabeled acoustic signals with estimated model parameters.

Acoustic scene estimation experiments with real-life sounds showed that the proposed method improved perplexity; that is, the proposed method achieved better generalization performance in modeling acoustic scenes that have more than one subordinate category of acoustic scenes. Experiments also showed that the proposed method improved the acoustic scene estimation accuracy an average of more than 2–3 points compared with the existing acoustic scene estimation method, and the proposed method also enables the extraction of subordinate categories of acoustic scenes more correctly.

As future work, we need to quantitatively evaluate how precisely the proposed method extracts the subordinate categories of acoustic scenes through acoustic sub-topics. In addition, we will extend the proposed method to an online method that can apply to sequentially obtained sounds and can analyze acoustic scenes in real time.

References


Appendix: Detailed Discussion of ASTM Parameter Estimation

In this appendix, we discuss the derivation of the ASTM parameters in detail.

A.1 Derivation of Eq. (6)

The first equation of Eq. (6) can be rewritten as follows.

\[
\sum_x \sum_z q(x, z) \log p(e, z, x|\alpha, \beta, a_s) = \sum_x \sum_z q(x, z) \log \int p(x|a_s) p(\theta|\alpha) p(z|\theta, x) p(\phi|\beta) p(e|\phi, z) d\theta d\phi
\]

\[
= \sum_x \sum_z \prod_t \int \frac{\Gamma(T(a))}{\Gamma(T(\alpha))} \prod_t \frac{\Gamma(M(\beta))}{\Gamma(M(\beta))} \prod_m \frac{\Gamma(z_t)}{\Gamma(x_t)} \prod_s \frac{\Gamma(x_s)}{\Gamma(x_s)} \prod_s \frac{\Gamma(x_s)}{\Gamma(x_s)} \prod_t \frac{\Gamma(x_t)}{\Gamma(x_t)} \prod_t \frac{\Gamma(x_t)}{\Gamma(x_t)} d\theta d\phi.
\]

Now, we consider an integral of the Dirichlet distribution as follows.

\[
\int \text{Dir}(\mu|\zeta) d\mu = \int \prod_t \frac{\Gamma(T(a))}{\Gamma(T(\alpha))} \prod_t \frac{\Gamma(M(\beta))}{\Gamma(M(\beta))} \prod_m \frac{\Gamma(z_t)}{\Gamma(x_t)} \prod_s \frac{\Gamma(x_s)}{\Gamma(x_s)} \prod_t \frac{\Gamma(x_t)}{\Gamma(x_t)} \prod_s \frac{\Gamma(x_s)}{\Gamma(x_s)} d\mu = 1
\]

\[
\int \prod_t \frac{\Gamma(T(a))}{\Gamma(T(\alpha))} \prod_t \frac{\Gamma(M(\beta))}{\Gamma(M(\beta))} \prod_m \frac{\Gamma(z_t)}{\Gamma(x_t)} \prod_s \frac{\Gamma(x_s)}{\Gamma(x_s)} \prod_t \frac{\Gamma(x_t)}{\Gamma(x_t)} \prod_s \frac{\Gamma(x_s)}{\Gamma(x_s)} d\mu = \prod_t \frac{\Gamma(T(a))}{\Gamma(T(\alpha))} \prod_t \frac{\Gamma(M(\beta))}{\Gamma(M(\beta))} \prod_m \frac{\Gamma(z_t)}{\Gamma(x_t)} \prod_s \frac{\Gamma(x_s)}{\Gamma(x_s)} \prod_t \frac{\Gamma(x_t)}{\Gamma(x_t)} \prod_s \frac{\Gamma(x_s)}{\Gamma(x_s)} \prod_t \frac{\Gamma(x_t)}{\Gamma(x_t)} \prod_s \frac{\Gamma(x_s)}{\Gamma(x_s)}
\]

Substituting Eq. (A.2) into Eq. (A.1), Eq. (6) can be calculated as follows.

\[
\sum_x \sum_z q(x, z) \log p(e, z, x|\alpha, \beta, a_s)
\]

\[
\cdot \frac{\Gamma(T(a))}{\Gamma(T(\alpha))} \prod_t \frac{\Gamma(M(\beta))}{\Gamma(M(\beta))} \prod_m \frac{\Gamma(z_t)}{\Gamma(x_t)} \prod_s \frac{\Gamma(x_s)}{\Gamma(x_s)} \prod_t \frac{\Gamma(x_t)}{\Gamma(x_t)} \prod_s \frac{\Gamma(x_s)}{\Gamma(x_s)} \prod_t \frac{\Gamma(x_t)}{\Gamma(x_t)} \prod_s \frac{\Gamma(x_s)}{\Gamma(x_s)}
\]

A.2 Derivation of Eq. (8)

Then, we discuss the derivation of Eq. (8) in detail. We first consider the following joint probability.

\[
p(e, x^{\alpha}, x^{\beta}, z^{\alpha}, z^{\beta} = t|\alpha, \beta, a_s)
\]

\[
\cdot \frac{\Gamma(T(a))}{\Gamma(T(\alpha))} \prod_t \frac{\Gamma(M(\beta))}{\Gamma(M(\beta))} \prod_m \frac{\Gamma(z_t)}{\Gamma(x_t)} \prod_s \frac{\Gamma(x_s)}{\Gamma(x_s)} \prod_t \frac{\Gamma(x_t)}{\Gamma(x_t)} \prod_s \frac{\Gamma(x_s)}{\Gamma(x_s)} \prod_t \frac{\Gamma(x_t)}{\Gamma(x_t)} \prod_s \frac{\Gamma(x_s)}{\Gamma(x_s)}
\]

Finally, we derive the update of Eq. (10). Substituting Eqs. (8) and (9) into \( \frac{\partial}{\partial y_{i, a_t}} \log p(e^{\alpha}, x^{\alpha}, z^{\alpha}|\alpha, \beta, a_s) = 0 \), we can obtain the following equation.

\[
\frac{\partial}{\partial y_{i, a_t}} \sum_x \sum_z q(x, z) \log p(e, z, x|\alpha, \beta, a_s)
\]

\[
- \frac{\partial}{\partial y_{i, a_t}} \sum_x \sum_z q(x, z) \log q(x, z) = 0
\]

\[
\sum_x \sum_z \frac{\partial}{\partial y_{i, a_t}} \sum_x \sum_z \frac{\Gamma(T(a))}{\Gamma(T(\alpha))} \prod_t \frac{\Gamma(M(\beta))}{\Gamma(M(\beta))} \prod_m \frac{\Gamma(z_t)}{\Gamma(x_t)} \prod_s \frac{\Gamma(x_s)}{\Gamma(x_s)} \prod_t \frac{\Gamma(x_t)}{\Gamma(x_t)} \prod_s \frac{\Gamma(x_s)}{\Gamma(x_s)} \prod_t \frac{\Gamma(x_t)}{\Gamma(x_t)} \prod_s \frac{\Gamma(x_s)}{\Gamma(x_s)}
\]

A.3 Derivation of the update for ASTM parameters

Finally, we derive the update of Eq. (10). Substituting Eqs. (8) and (9) into \( \frac{\partial}{\partial y_{i, a_t}} \log p(e^{\alpha}, x^{\alpha}, z^{\alpha}|\alpha, \beta, a_s) = 0 \), we can obtain the following equation.

\[
\frac{\partial}{\partial y_{i, a_t}} \sum_x \sum_z q(x, z) \log p(e, z, x|\alpha, \beta, a_s)
\]

\[
- \frac{\partial}{\partial y_{i, a_t}} \sum_x \sum_z q(x, z) \log q(x, z) = 0
\]

\[
\sum_x \sum_z \frac{\partial}{\partial y_{i, a_t}} \sum_x \sum_z \frac{\Gamma(T(a))}{\Gamma(T(\alpha))} \prod_t \frac{\Gamma(M(\beta))}{\Gamma(M(\beta))} \prod_m \frac{\Gamma(z_t)}{\Gamma(x_t)} \prod_s \frac{\Gamma(x_s)}{\Gamma(x_s)} \prod_t \frac{\Gamma(x_t)}{\Gamma(x_t)} \prod_s \frac{\Gamma(x_s)}{\Gamma(x_s)} \prod_t \frac{\Gamma(x_t)}{\Gamma(x_t)} \prod_s \frac{\Gamma(x_s)}{\Gamma(x_s)}
\]

\[
\cdot \log \frac{\Gamma(x | \alpha, \beta, a_s)}{\Gamma(x | \alpha, \beta, a_s) + \Gamma(x | \alpha, \beta, a_s)}
\]

\[
\frac{\Gamma(T(a))}{\Gamma(T(\alpha))} \prod_t \frac{\Gamma(M(\beta))}{\Gamma(M(\beta))} \prod_m \frac{\Gamma(z_t)}{\Gamma(x_t)} \prod_s \frac{\Gamma(x_s)}{\Gamma(x_s)} \prod_t \frac{\Gamma(x_t)}{\Gamma(x_t)} \prod_s \frac{\Gamma(x_s)}{\Gamma(x_s)} \prod_t \frac{\Gamma(x_t)}{\Gamma(x_t)} \prod_s \frac{\Gamma(x_s)}{\Gamma(x_s)}
\]

By introducing the form of expectation of \( q(x^{\alpha}, z^{\alpha}) \), we can derive Eq. (10) from Eq. (A.7).
Keisuke Imoto received B.E., M.E. degrees from the Kyoto University in 2008 and 2010. He joined the Nippon Telegraph and Telephone Corporation (NTT) in 2010. He is now a Ph.D. student at SOKENDAI (The Graduate University for Advanced Studies), and he has been engaged in research on acoustic event detection and microphone array signal processing. He is a member of IEEE, IPSJ, and ASJ.

Suehiro Shimauchi received B.E., M.E., and Ph.D. degrees from the Tokyo Institute of Technology in 1991, 1993, and 2007, respectively. Since joining the Nippon Telegraph and Telephone Corporation (NTT) in 1993, he has been engaged in research on acoustic signal processing. He is now a senior research engineer at NTT Media Intelligence Laboratories. He is a senior member of IEEE and a member of ASJ.