User Activity Estimation Method Based on Probabilistic Generative Model of Acoustic Event Sequence with User Activity and Its Subordinate Categories

Keisuke Imoto, Suehiro Shimauchi, Hisashi Uematsu, and Hiitoshi Ohmuro

NTT Media Intelligence Laboratories, NTT Corporation, Tokyo, Japan

Abstract

We propose a method for estimating user activities by analyzing long-term (more than several seconds) acoustic signals represented as acoustic event temporal sequences. The proposed method is based on a probabilistic generative model of an acoustic event temporal sequence that is associated with user activities (e.g., "cooking") and subordinate categories of user activities (e.g., "fry ingredients" or "plate food") in which each user activity is represented as a probability distribution over unsupervised subordinate categories of user activities called activity-topics, and each activity-topic is represented as a probability distribution over acoustic events. This probabilistic generative model can express user activities that have more than one subordinate category of the user activities, which a model that takes into account only user activities cannot express adequately. User activity estimation with this model is achieved using a two-step process: frame-by-frame acoustic event estimation to output an acoustic event temporal sequence and user activity estimation with the proposed probabilistic generative model. Activity estimation experiments with real-life sounds indicated that the proposed method improved user activity estimation accuracy and stability of "unseen" acoustic event temporal sequences. In addition, the experiment showed that the proposed method could extract correct subordinate categories of user activities.

Index Terms: user activity estimation, probabilistic generative model, Acoustic Event Detection (AED)

1. Introduction

Much interest has been expressed recently in analyzing different types of acoustic events (e.g., footsteps, running water, vacuuming) in order to extract valuable information and this is referred to as Acoustic Event Detection (AED). It is a challenging research topic because there are so many types of acoustic events that have different properties. There are also many relevant applications for this research such as in monitoring elderly people [1], security surveillance [2], automatic classification of user activities and contexts [3], and multimedia retrieval [4]. For this study, we focus on automatic classification of user activities for life-logging or monitoring elderly people based on AED.

One approach in this area is focused on a single acoustic event that obviously characterizes a certain context or scene and detect this noticeable acoustic event. However, many contexts or scenes, particularly user activities, are characterized not by a single acoustic event but by a combination of multiple acoustic events. For example, the activity "cooking" is marked by a combination of sounds including "running water," "cutting with a knife," and "heating a skillet;" that is, we can consider that user activities determine the probability distributions over acoustic events. Based on this consideration, Ahaikh et al. [5] proposed a user activity estimation method using a combination of multiple acoustic events. Lee et al. [4] and Heittola et al. [6] also proposed content recognition method using a combination of multiple acoustic events.

On the other hand, we can assume that some user activities have subordinate categories of user activities as shown in Fig. 1, therefore, characterizing directly user activities by combinations of multiple acoustic events is not as expressive in these user activities. For example, we consider that the activity "cooking" is marked by its subordinate categories including "preparation of ingredients," "fry ingredients" and "plate food," and the its subordinate category "preparation of ingredients" is characterized by sounds including "opening a fridge," "running water" and "cutting with a knife."

For this study, we propose a probabilistic generative model that can associate acoustic event temporal sequences with user activities and unsupervised subordinate categories of user activity called activity-topics. We also propose a user activity estimation method based on this probabilistic generative model. In user activity estimation, we preliminarily estimate acoustic events frame by frame and represent them as an acoustic event temporal sequence. We then calculate an occurrence histogram of acoustic events in the acoustic event temporal sequence and use it for user activity estimation with the probabilistic generative model.

The rest of this paper is as follows. In Section 2, we describe the basic idea and our proposed probabilistic generative model of an acoustic event temporal sequence with user activities and activity-topics. In Section 3, we present an implemented user activity estimation system that includes acoustic event estimation and the parameter estimation method of our proposed model. In Section 4, we discuss the results of the experiments. In Section 5, we review related studies, and in Section 6, we conclude this paper and describe future works.
2. Probabilistic Generative Model of Acoustic Event Temporal Sequence

2.1. Basic Idea

On the basic idea, we assume that a user activity can be represented by a probability distribution over acoustic events, that is, we model a following generative process of an acoustic event temporal sequence \( e_s \) in an acoustic signal \( s \) that emanates with user activity. We call this the Acoustic Activity Model (AAM), \( \Pi \). To estimate user activities, which are the category of user activity, \( \Pi \) is the probability distribution of acoustic events in \( a \), \( x_i \) is the user activity that the \( i \)th acoustic event indicates, and \( e_i \) is the \( i \)th acoustic event in \( e_s \).

In this model, \( x_i \) is first sampled randomly from \( a_{e_s} \) for every event in \( e_s \). This activity \( x_i \) samples \( e_i \) from its multinomial distribution \( \phi_{s,x} \) over acoustic events associated with \( x_i \), and, \( \phi_{e} \) has a Dirichlet prior of parameter \( \beta \). This generative process is repeated for the \( N_{e_s} \) to generate \( e_s \).

In addition, to relax the temporal constraints of an acoustic event temporal sequence and simplify the representation of the model, we assume that the acoustic event temporal sequence can be described as a "bag of acoustic events" (BoE), - which corresponds to the "bag of words" representation in natural language processing [7]. BoE representation assumes that each acoustic event in the temporal sequence is exchangeable, that is, we hypothesize that there is no direct temporal relation between acoustic events in a user activity.

2.2. Acoustic Activity-Topic Model

As shown in Fig. 1, some user activities have more than one subordinate category, therefore, AAM is not as expressive in these user activities. Therefore, we propose a probabilistic generative model that can associate acoustic event temporal sequences with user activities and unsupervised subordinate categories of the user activities called activity-topics, as shown in Fig. 2, where \( \beta, a_{e_s}, x, e_i \) are the same as above, \( \alpha \) is a hyperparameter of the Dirichlet distribution, \( A \) and \( T \) are the size of the activity set and the size of the activity-topic set, respectively, \( \theta_{s,a} \) is the probability distribution of the activity-topics in the activity \( a \), \( \phi_{e} \) is the probability distribution of the acoustic events in the activity-topic \( t \), \( z_i \) is an activity-topic that the \( i \)th acoustic event indicates, \( N_{e_s} \) is the number of acoustic events in \( e_s \), and \( S \) is the size of acoustic event temporal sequence set. We call this model Acoustic Activity Topic Model (AATM). In AATM, we treat subordinate categories of the user activities as latent variables to capture underlying structures of user activities characterized by a combination of acoustic events.

The following generative process for each \( e_s \) is assumed with AATM.

The set of user activities \( a_{e_s} \) included in \( e_s \) are given,

1. Choose \( \phi_a \sim Dirichlet(\beta) \)
2. Choose \( x_i \sim Uniform(a_{e_s}) \)
3. Choose \( e_i | \phi_{x_i}, x_i \sim Discrete(\phi_{x_i}) \)
4. Choose \( e_i | \phi_{e}, z_i \sim Discrete(\phi_{e}) \).

\[ \mathcal{L}_{e_s} = \prod_{i=1}^{N_{e_s}} P(e_i | \theta_{s,a}, e_s, \alpha, \beta) \]
\[ = \prod_{i=1}^{N_{e_s}} \sum_{a=1}^{A} \sum_{t=1}^{T} P(z_i | x_i, \theta_{s,a}) \cdot Dir(\theta_{s,a} | \alpha) \cdot P(e_i | z_i, \phi_e) \cdot Dir(\phi_e | \beta) \cdot \text{Uni}(x_i | a_{e_s}) \]
\[ = \prod_{i=1}^{N_{e_s}} \frac{1}{A_{e_s}} \sum_{a \in a_{e_s}} \sum_{t=1}^{T} \phi_{e,z_i} \cdot \theta_{s,a} \cdot \phi_{e,z_i} \sim \text{Discrete}(\phi_{e}) \]

where \( A_{e_s} \) is the number of user activities in \( e_s \), Dir(\cdot | \beta) is the Dirichlet distribution with \( \beta \), Uni(\cdot | a_{e_s}) is the uniform distribution with \( a_{e_s}, \phi_{e,s} \) is the generative probability of \( e_i \) in \( z_i \), and \( \theta_{s,a} \) is the generative probability of \( z_i \) in \( x_i \).

3. User Activity Estimation System Based on AATM

3.1. User Activity Estimation System Overview

For this study, we also propose a user estimation method with AATM. To estimate user activities, which \( e_s \) indicates, with AATM, we calculate the activity that maximizes the posterior probability \( P(x_{a,s} = a | e_s) \) described as:

\[ e_s = \arg \max_a P(x_{a,s} = a | \theta_{s,a}, \phi_e, e_s) \]
\[ = \arg \max_a \prod_{i=1}^{N_{e_s}} P(x_i | z_i, \theta_{s,a}) \cdot Dir(\theta_{s,a} | \alpha) \cdot P(z_i | e_i, \phi_e) \cdot Dir(\phi_e | \beta) \]
\[ = \arg \max_a \frac{\prod_{i=1}^{N_{e_s}} P(x_i | z_i, \theta_{s,a}) \cdot Dir(\theta_{s,a} | \alpha) \cdot P(z_i | e_i, \phi_e) \cdot Dir(\phi_e | \beta)}{\sum_{\phi_e} P(e_i | z_i, \phi_e) \cdot Dir(\phi_e | \beta)} \]

where \( x_{a,s} \) is the user activity that \( e_s \) indicates. Note that if we find \( \theta_{s,a}, \phi_e \) and \( e_s \), we can estimate user activities with AATM.

We present the activity estimation system as shown in Fig. 3, which involves the two-step process: estimating an acous-
Acoustic Signal $S$ \hspace{1cm} Acoustic Event Recognition $e_s$ \hspace{1cm} BoE Calculation $\phi_t$ \hspace{1cm} Activity Estimation $\alpha$ \hspace{1cm} AATM Parameter Estimation $\theta_a$, $\theta_e$, $\alpha$, and $\beta$

Figure 3: Activity estimation system overview.

3.2. Acoustic Event Estimation

To calculate $e_s$, we segment long-term input signals and extract acoustic feature vectors frame by frame, and estimate acoustic events in every frame, and output an $e_s$ (composed of multiple event estimation results).

We choose Mel-Frequency Cepstral Coefficients (MFCCs) as the acoustic features. These are well known features for speech recognition and music information processing, and MFCC features have also been used in AED tasks [8], in which they performed well.

Many event estimation approaches are based on the Gaussian mixture model (GMM) or Hidden Markov model (HMM) [8, 9], 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. Fortunately, our user activity estimation method represents user activities and activity-topics as a distribution over multiple acoustic events.

Since there are many different types of sounds, it is difficult to label every possible acoustic event. Fortunately, our method performed well.

9,802 sounds for AA TM parameter training and 1,303 sounds for evaluation, both of which have nine user-activity categories: “chatting,” “cooking,” “eating dinner,” “operating PC,” “reading a newspaper,” “vacuuming,” “walking,” “washing dishes,” and “watching TV.”

In acoustic event estimation, 16-dimensional MFCC features were calculated from every segmented acoustic signal with 50% overlap, and acoustic event models were learned from the 9,802 sounds by using GC with 16-512 acoustic event sizes.

3.3. AATM Parameter Estimation

AATM includes the unknown parameters $\theta_a$ and $\phi_t$ for estimating user activities. For estimating the appropriate parameters, we calculate the $\theta_a$ and $\phi_t$ that maximizes $\sum_{t=1}^{T_a} L_{e_s}$ in the $e_s$ of the training set. Various algorithms have been proposed for parameter estimation such as the EM algorithm, Variational bayes (VB), and expectation propagation. We used Gibbs sampling (GS) which can avoid getting trapped by local optima and can estimate more appropriate parameters.

In GS, we initialize the $x$ and $z$ assignments, randomly (or by using prior information). Then, the sampling process with the following equation is performed for every acoustic event in the entire acoustic event sequence set. Due to the space limit, we show the equation without derivation:

$$ P(x_t = a, z_t = l|e_s = v, x_{t-1}, z_{t-1}, e_{t-1}, A, \alpha, \beta) \propto \frac{C_{VT}^{VT} + \beta}{\sum_{v'} C_{VT}^{VT} + E\beta}. $$

Table 1: Experimental conditions of activity estimation task.

<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 cluster size</td>
<td>16 - 512</td>
</tr>
<tr>
<td>Acoustic event sequence size</td>
<td>10-s</td>
</tr>
<tr>
<td>Hyperparameter $\alpha / \beta$</td>
<td>3.3333 / 0.1</td>
</tr>
<tr>
<td>Repeated number of Gibbs sampling</td>
<td>1,000</td>
</tr>
</tbody>
</table>

where $v$ is the category of acoustic event, $E$ is the size of the acoustic event set, $C_{VT}^{VT}$ is the number of times $v$ assigned to $a$ in $e_s$. $C_{VT}^{VT}$ are the number of times $v$ assigned to $t$ in $e_s$. $A$ is the set of all activities, and $x_{t-1}, z_{t-1}, e_{t-1}$ represent all activity, activity-topic and acoustic event assignments not including the $t$-th acoustic event.

This process is repeated a given number of times or until a convergence condition is satisfied. After this iteration, we can estimate $\theta_a$ and $\phi_t$ in the following equations with the resultant $C_{VT}^{VT}$ and $C_{VT}^{VT}$, respectively:

$$ \theta_a = \frac{C_{VT}^{VT} + \alpha}{\sum_{v'} C_{VT}^{VT} + T\alpha} $$
$$ \phi_t = \frac{C_{VT}^{VT} + \beta}{\sum_{v'} C_{VT}^{VT} + E\beta}. $$

4. Experiments

We evaluated activity estimation for life-logging or monitoring elderly people specifically in doors. We recorded 11,105 real-life sounds in a living room (22.8 m$^2$, reverberation time: 0.31 sec.) and separated them into 9,802 sounds for AATM parameter training and 1,303 sounds for evaluation, both of which have nine user-activity categories:

The other experimental conditions are listed in Table 1.

We evaluated the perplexities with AATM and the basic model (AAM), by which we could evaluate their generalization performance, as shown in Fig. 4. We calculated the perplexity with AATM as follows.

$$ \text{Perplexity}(e_s|a_e) = \exp \left[ -\frac{\log \left\{ \prod_{i=1}^{N_{e_s}} \left( \frac{1}{\sum_{a_{e_s}|a_{e_s}} \sum_{C_{VT}^{VT} + T\alpha} + E\beta} \right) \right\} }{N_{e_s}} \right]. $$

For the perplexity with AATM, we chose $T = 10, 20, 50, 100$ as the number of is, and for all models, we used the same test set including the above nine user-activity categories. For the perplexity with AAM, we calculated the one with the similar manner in [10]. As shown in Fig. 4, AATM improved perplexity compared to AAM in all acoustic event sizes; thus, AATM indicated better generalization performance in modeling user activity, its subordinate categories and the consequent $e_s$ since AATM could flexibly model user activities that have subordinate categories.

The averaged estimation accuracy (F-value) of the nine user-activity categories with the 16-512 acoustic event sizes is
shown in Fig. 5. The user activity estimation accuracy with AATM, which has more topics than the number of user activities, increased about three points compared with AAM, while the user activity estimation with AAM, which has a similar number of topics as user activities, was only as effective as AAM. For example, AATM (acoustic event size = 512, activity-topic size = 50) correctly classified 86.1 % of nine user-activity categories, while AAM (acoustic event size = 512) correctly classified 83.1 %. This suggests that some user activities have more than one subordinate category; therefore, AATM offers advantages in user activity estimation.

Figure 6 shows an activity-topic estimation result of “cooking” with a test sound. The subordinate categories of user activity in upper part of the figure describe actual manually labeled categories, and color-coded acoustic signal denotes the correspondence relationships between $e_s$ and $t_s$ estimated with AATM. AATM can generally extracts subordinate categories of user activity in addition to estimating user activity.

5. Related Work

We discussed our method of estimating user activities based on a probabilistic generative model over multiple acoustic events and multiple activity-topics. Many methods of estimating user activities or acoustic scenes are focused on a single acoustic event that obviously characterizes these activities or scenes and detect this noticeable acoustic event [8,9]. There have also been many studies with these approaches for monitoring elderly people [1], security surveillance [2], and automatic classification of user activities and contexts [3].

However, many contexts or scenes, particularly user activities, are characterized not by a single acoustic event but by a combination of multiple acoustic events. Ahaikh et al. [5], Lee et al. [4] and Heittola et al. [6] proposed a user activity or content estimation method that characterized them by combinations of multiple acoustic events.

On the other hand, we proposed the user activity estimation method on the idea that some user activities have subordinate categories of user activities, therefore, we propose a probabilistic generative model that can associate acoustic event temporal sequences with user activities and unsupervised subordinate categories of user activity called activity-topics.

Probabilistic generative models also have been proposed in the field of natural language processing [11,12], where they model and estimate topics or authors in a document (corresponding to activity-topics or user activities) with words (analogous to acoustic events). Other researchers [13,14] have applied this technique to acoustic event estimation. However, they focused on probabilistic generative models of a single acoustic event to analyze overlapped acoustic events, where it is assumed that a single acoustic event is composed of a combination of frequency components.

6. Conclusion and Future Works

Some user activities (e.g. ”cooking”) have subordinate categories of user activities (e.g. ”preparation of ingredients,” ”fry ingredients”) and they are marked by their subordinate categories, and their subordinate categories (e.g. ”preparation of ingredients”) are characterized by combinations of multiple acoustic events (e.g. ”opening a fridge,” ”running water” and ”cutting with a knife.”)

We proposed a user activity estimation method based on a probabilistic generative model which is associated with user activity and activity-topics, which are represented by the probability distribution over activity-topics, and each activity-topic is represented as the probability distribution over acoustic events. User activity estimation with this model is achieved using a two-step process: frame-by-frame acoustic event estimation to generate an acoustic event temporal sequence and user activity estimation with the proposed model.

Activity estimation experiments with real-life sounds indicated that the proposed method improved the perplexity, that is, the proposed method can achieve better generalization performance in modeling user activity and consequent acoustic event temporal sequence. The experiments also suggested that the proposed method improved the activity estimation accuracy an average of three points compared to the basic method, and the proposed method enables the extraction of subordinate categories of user activity.

In the future, we plan to apply this user activity estimation method to sounds recorded in various living room or other rooms and apply to other user activities.
7. References


