BAYESIAN SEMI-SUPERVISED AUDIO EVENT TRANSCRIPTION BASED ON MARKOV INDIAN BUFFET PROCESS

Yasunori Ohishi†, *Daichi Mochihashi*‡, *Tomoko Matsui*‡,
Masahiro Nakano†, *Hirokazu Kameoka*†, *Tomonori Izumitani*†, *Kunio Kashino*†

†NTT Communication Science Laboratories, NTT Corporation
‡The Institute of Statistical Mathematics

ABSTRACT

We present a novel generative model for audio event transcription that recognizes "events" on audio signals including multiple kinds of overlapping sounds. In the proposed model, firstly, the overlapping audio events are modeled based on nonnegative matrix factorization into which Bayesian nonparametric approaches: the Markov Indian buffet process and the Chinese restaurant process, are incorporated. This approach allows us to automatically transcribe the events while avoiding the model selection problem by assuming a countably infinite number of possible audio events in the input signal. Then, Bayesian logistic regression annotates the audio frames with the multiple event labels in a semi-supervised learning setup. Experimental results show that our model can better annotate an audio signal in comparison with a baseline method. Additionally, we verify that our infinite generative model is also able to detect unknown audio events that are not included in the training data.

Index Terms— Audio event transcription, Generative model, Nonnegative matrix factorization, Bayesian nonparametric approach

1. INTRODUCTION

As the amount of available multimedia data increases, techniques for automatically extracting the significant information from audio or video data become crucial for application to multimedia search. Audio event detection/transcription is a technique where computational methods are used to separate and recognize mixtures of sounds in natural environments, and this approach has attracted increasing attention from the research community [1, 2, 3, 4, 5]. Here, an audio event is defined as a recognizable sound object in a given environment, e.g., human and animal sounds such as "speech", "coughing", and "dog barking" and natural and acoustic sounds such as "music", "traffic noise", and "office sounds".

We work on two audio event transcription problems. The first is modeling and detecting *overlapping* audio events. Most studies considered certain overlapping events as an acoustic category and constructed the acoustic models of the categories using Gaussian mixture models (GMM) and hidden Markov models (HMM), and features, e.g., mel-frequency cepstrum coefficients, which are calculated directly from the polyphonic mixture, and then detected the categories [6, 7, 8, 9, 10, 11]. These approaches are limited in terms of application to rich multisource environments. Recently, [12], [13], and [14] employed sound source separation methods as preprocessing techniques to separate an audio signal into some tracks and identified the events using acoustic features extracted from each track. However, the number of tracks must be manually adjusted depending on the sound environment. There has also been recent work on learning the number of tracks based on automatic relevance determination [15].

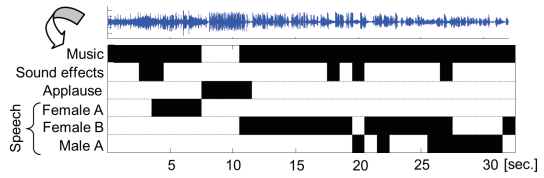

Fig. 1. Audio event transcription ("on": black, "off": white)

The second problem is lack of training samples. Compared with automatic speech recognition trained using hundreds or thousands of hours of manually transcribed speech, databases annotated in audio event transcription are still sparse. In addition, to detect a new audio event, we have to annotate the database with the label newly. Therefore semi-supervised and unsupervised learning is a promising approach for overcoming data sparseness and the handling of unknown audio events [16, 17, 18, 19].

To tackle these problems, we propose a novel generative model for audio event transcription that recognizes overlapping audio events, as shown in Fig. 1. Firstly, the overlapping audio events are modeled based on nonnegative matrix factorization (NMF) [20, 21, 22, 23] into which Bayesian nonparametric approaches: the Markov Indian buffet process (mIBP) [24] and the Chinese restaurant process (CRP) [25], are incorporated. NMF models an audio mixture signal as a sum of acoustic components which correspond, for example, to phonemes in speech, notes in music, and sound effects. The Bayesian approaches allow us to avoid the model selection problem by marginalizing out the unknown model parameters (including the number of acoustic components) and assuming that there are a countably infinite number of possible audio components in the input signal. Then, Bayesian logistic regression annotates the audio frames with the multiple event labels using a combination of these components in a semi-supervised learning setup. Finally, we derive an efficient inference algorithm for the model parameters based on the Gibbs sampler [26]. Experimental results show that our model can better annotate a real audio podcast in comparison with a baseline method. Furthermore, we verify that our infinite generative model has the ability to detect unknown audio events.

2. ACOUSTIC EVENT TRANSCRIPTION MODEL

Overlapping audio events are modeled based on an NMF approach with deformable bases [23] that represents time-varying spectra through state transitions. To characterize the timbre of each acoustic component, we apply NMF to mel-scaled filter bank outputs $\boldsymbol{Y} = (Y_{\omega,t})_{\Omega\times T} \in \mathbb{R}^{\geq 0,\Omega\times T}$ of the magnitude spectrogram of a mixture signal, where $\omega = 1,\ldots,\Omega$ is the center frequency bin

index of a mel-filter bank, $t = 1, \ldots, T$ is a frame index, and $d = 1, \ldots, D$ is an acoustic component index. If we let $Z_{d,t} \in \mathbb{N}$ denote the spectral basis state of the d-th acoustic component that is activated at time t, NMF with deformable bases can be written as

$$Y_{\omega,t} = \sum_d C_{\omega,t,d}, \quad C_{\omega,t,d} \sim \text{Poisson}(H^{(Z_{d,t})}_{\omega,d} U_{d,t}). \quad (1)$$

This is defined as the following generative model (Fig. 2). First, $\boldsymbol{H}_d = (H^{(k)}_{\omega,d})_{\Omega \times K_d}$ denotes K_d spectral bases for the d-th component, and $\boldsymbol{U} = (U_{d,t})_{D \times T}$ denotes the activation matrix and consists of binary sequences that represent the on/off states of the acoustic components. The gains of the components are represented in the spectral bases. Then, $\boldsymbol{C}_d = (C_{\omega,t,d})_{\Omega \times T}$ corresponds to the mel-filter bank outputs for the d-th component, and $C_{\omega,t,d}$ is generated from a Poisson distribution. Finally, an audio mixture signal is represented as a sum of these components, where we assume that each audio event is represented by combining these components. Note that Poisson likelihood models suffer from a theoretical issue. That is, it is only applicable to discrete counts data. Recently, [27] proposed Poisson-Uniform NMF to overcome this problem. However, we use the classical Poisson likelihood model as a case study.

Furthermore, we leverage the mIBP and the CRP as prior distributions of $\boldsymbol{U}$ and $\boldsymbol{Z}$, respectively, and we employ Bayesian logistic regression to detect the events in a semi-supervised learning setup.

2.1. Activation matrix generated by mIBP

The nonparametric Bayesian factor model called the mIBP [24], which extends the IBP [28] to allow temporal dependencies, defines a distribution over binary matrices to model whether a component at frame t is on or off and satisfies the following properties: (1) the potential number of rows (representing latent components) should be capable of being arbitrarily large; (2) the columns (representing timesteps) should evolve according to a Markov process to represent the durations of the component. This construction allows us to learn a factorial representation for time series. Thus, we apply the mIBP to the prior distribution of activation matrix $\boldsymbol{U}$.

Let $U_{d,t}$ represent the on/off state at frame t for the d-th component. Each Markov chain evolves according to the transition matrix

$$\boldsymbol{W}^{(d)} = \begin{bmatrix} 1 - a_d & a_d \\ 1 - b_d & b_d \end{bmatrix}, \quad (2)$$

where $W^{(d)}_{i,j} = p(U_{d,t+1} = j - 1 | U_{d,t} = i - 1),\ i, j \in \{1, 2\}$. We give the parameters of $\boldsymbol{W}^{(d)}$ distributions $a_d \sim \text{Beta}(\theta_{\text{a}}/D, 1)$ and $b_d \sim \text{Beta}(\theta^{(0)}_{\text{b}}, \theta^{(1)}_{\text{b}})$. Each chain starts with a dummy zero state $U_{d,0} = 0$. The binary sequence for the d-th component is generated by sampling T steps from a Markov chain with a transition matrix $\boldsymbol{W}^{(d)}$. Next, we introduce the following notation. Let c^{00}_d, c^{01}_d, c^{10}_d, c^{11}_d be the number of $0 \to 0$, $0 \to 1$, $1 \to 0$, and $1 \to 1$ transitions respectively, in the d-th component. The probability distribution of a binary matrix $\boldsymbol{U}$ can be written as

$$p(\boldsymbol{U} | \boldsymbol{a}, \boldsymbol{b}) = \prod_d (1 - a_d)^{c^{00}_d} a_d^{c^{01}_d} (1 - b_d)^{c^{10}_d} b_d^{c^{11}_d}, \quad (3)$$

where $\boldsymbol{a} = \{a_1, \ldots, a_D\}$ and $\boldsymbol{b} = \{b_1, \ldots, b_D\}$. To compute the limit for $D \to \infty$ of the finite model in Eq. (3), we adapt the stick breaking construction [29] to the mIBP. Let $a_{(1)} > a_{(2)} > \ldots > a_{(D)}$ be a decreasing ordering of $\boldsymbol{a}$. We use the result in [29], which show that in the limit $D \to \infty$ the $a_{(d)}$'s obey the following law

$$\nu_{(d)} \sim \text{Beta}(\theta_{\text{a}}, 1), \quad a_{(d)} = \nu_{(d)} a_{(d-1)} = \prod_{d'=1}^{d} \nu_{(d')}. \quad (4)$$

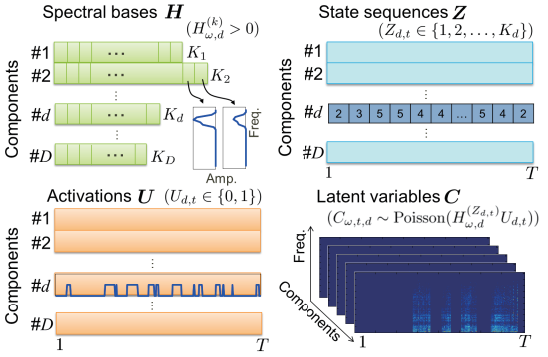

Fig. 2. Audio event transcription model based on NMF and mIBP

The variables b_d are all independent draws from a $\text{Beta}(\theta^{(0)}_{\text{b}}, \theta^{(1)}_{\text{b}})$ distribution which is independent of D. Hence, if we denote with $b_{(d)}$ the b variable corresponding to the d-th largest a value then it follows that $b_{(d)} \sim \text{Beta}(\theta^{(0)}_{\text{b}}, \theta^{(1)}_{\text{b}})$. In the mIBP, more frequent activated component is assigned to the smaller index d.

2.2. Time-varying spectral characteristics generated by CRP

The number of states needed to represent time-varying spectral characteristics varies depending on the acoustic components. For example, a piano note would be more accurately characterized by a succession of several spectral patterns such as attack, decay, sustain and release. The same is true for each phoneme in speech. The sounds made by a door opening and closing would be characterized by fewer spectral patterns than notes and phonemes. Thus, it is desirable to determine automatically the appropriate number of spectral bases for each acoustic component. Nakano *et al*. introduced the Dirichlet process (DP) [30] into the deformable bases and appropriately decomposed a polyphonic music signal into notes for each instrument [23]. We apply this approach to the audio event transcription. Several practical methods have been derived for the construction of DP [25, 31, 32, 33]. In this study, we represent the DP using the CRP [25, 34].

Let $Z_{d,1}, \ldots, Z_{d,T}$ represent a sequence of state indices for the d-th component, where each $Z_{d,t}$ can take on values $1, \ldots, K_d$ with proportions given by $\boldsymbol{\pi}_d = \{\pi_{d,1}, \ldots, \pi_{d,K_d}\}$. The joint distribution of the sequence is a multinomial distribution. Let us give the mixing proportions a symmetric Dirichlet prior, which is a conjugate prior of the multinomial distribution, with positive concentration hyperparameter $\theta^{(d)}_\beta : p(\boldsymbol{\pi}_d | \theta^{(d)}_\beta) = \text{Dirichlet}(\theta^{(d)}_\beta / K_d, \ldots, \theta^{(d)}_\beta / K_d)$. The probability over $Z_{d,t}$ conditioned on the state assignments of all other frames $\boldsymbol{Z}_{d,\setminus t}$ under $K_d \to \infty$:

$$p(Z_{d,t} = k | \boldsymbol{Z}_{d,\setminus t}, \theta^{(d)}_\beta) = \begin{cases} \dfrac{n^{(k)}_{d,\setminus t}}{(T - 1 + \theta^{(d)}_\beta)} & (n^{(k)}_{d,\setminus t} > 0) \\ \dfrac{\theta^{(d)}_\beta}{(T - 1 + \theta^{(d)}_\beta)} & (k = K_{\setminus t,+} + 1) \end{cases},$$

where $n^{(k)}_{d,\setminus t}$ is the number of frames assigned to state k, not including frame t, and $K_{\setminus t,+}$ is the number of states for which $n^{(k)}_{d,\setminus t} > 0$. As we can see, $Z_{d,t}$ tends to choose an already popular state. The concentration parameter $\theta^{(d)}_\beta$ controls the tendency to populate a previously unrepresented state. Each component thus tends to keep an adequate number of states depending on the observed signals.

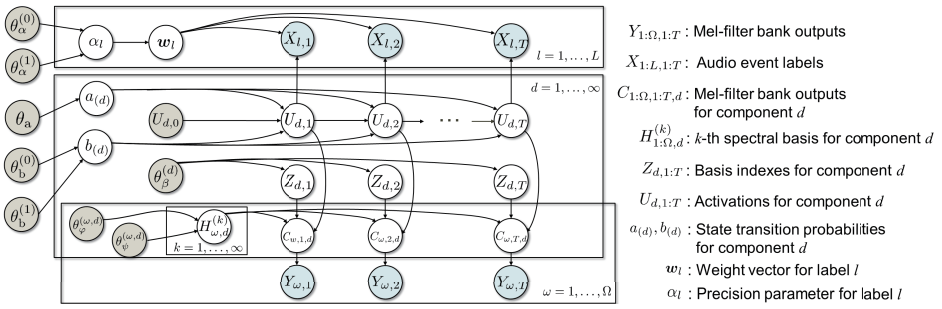

Fig. 3. Graphical representation of our audio event transcription model

2.3. Multiple event labeling using Bayesian logistic regression

We employ Bayesian logistic regression [35, 36] to estimate audio event labels from the activation matrix. This approach is similar to sLDA [37]. Let $\boldsymbol{U}_t = [U_{1,t}, U_{2,t}, \ldots, U_{D,t}]^{\mathrm{T}}$ represent the activations of the acoustic components at frame t. The likelihood function of labels $\boldsymbol{X}_l = \{X_{l,1}, X_{l,2}, \ldots, X_{l,T}\}$ $(X_{l,t} \in \{0, 1\})$ for audio event l can be written as

$$p(\boldsymbol{X}_l|\boldsymbol{U}, \boldsymbol{w}_l) = \textstyle\prod_{t=1}^{T} \exp(\boldsymbol{w}_l^{\mathrm{T}} \boldsymbol{U}_t X_{l,t}) \sigma(-\boldsymbol{w}_l^{\mathrm{T}} \boldsymbol{U}_t), \quad (5)$$

where $\boldsymbol{w}_l = [w_{l,1}, w_{l,2}, \ldots, w_{l,D}]^{\mathrm{T}}$ is the weight vector and $\sigma(\cdot)$ is the logistic sigmoid function. We consider a simple isotropic Gaussian prior distribution of the form $p(\boldsymbol{w}_l|\alpha_l) = \mathcal{N}(\boldsymbol{0}, \alpha_l^{-1}\boldsymbol{I}_D)$ and a conjugate hyperprior over α_l given by a Gamma distribution $p(\alpha_l) = \mathrm{Gamma}(\theta_\alpha^{(0)}, \theta_\alpha^{(1)})$ governed by $\theta_\alpha^{(0)}$ and $\theta_\alpha^{(1)}$. $\boldsymbol{I}_D$ denotes the $D \times D$ identity matrix.

Fig. 3 shows the graphical representation of our model which generates both acoustic feature matrix $\boldsymbol{Y}$ and label matrix $\boldsymbol{X}$. For the spectral bases $\boldsymbol{H}$, we use a Gamma distribution $H_{\omega,d}^{(k)} \sim \mathrm{Gamma}(\theta_\varphi^{(\omega,d)}, \theta_\psi^{(\omega,d)})$. $\theta_\alpha^{(0)}, \theta_\alpha^{(1)}, \theta_{\mathrm{a}}, \theta_{\mathrm{b}}^{(0)}, \theta_{\mathrm{b}}^{(1)}, \theta_\beta^{(d)}$, $\theta_\varphi^{(\omega,d)}$ and $\theta_\psi^{(\omega,d)}$ are constant hyperparameters.

3. INFERENCE

The stick breaking construction of the mIBP allows us to use a combination of slice sampling and dynamic programming to do inference [24, 38, 39]. A slice sampler adaptively truncates the infinite dimensional model after which dynamic programming performs exact inference. However, since our model includes annotation based on logistic regression, we use Gibbs sampling [26] in the truncated stick-breaking construction. Owing to space limitations, we only describe the inference of binary sequence $U_{d,1}, \ldots, U_{d,T}$ given other variables. We use a blocked Gibbs sampler that fixes all but one row of $\boldsymbol{U}$ and performs forward-filtering backward-sampling on the remaining row [24]. We compute $p(U_{d,t}|Y_{:,1:t}, C_{:,1:t,d}, Z_{d,:}, H_{:,d}^{(:)}, X_{:,1:t})$ for all t as follows:

$$\begin{aligned}
&p(U_{d,t}|Y_{:,1:t}, C_{:,1:t,d}, Z_{d,:}, H_{:,d}^{(:)}, X_{:,1:t}) \\
&\quad \propto p(Y_{:,t}, C_{:,t,d}|U_{d,t}, Z_{d,:}, H_{:,d}^{(:)}) \textstyle\prod_l p(X_{l,t}|U_{d,t}, \boldsymbol{U}_{\setminus d,t}, \boldsymbol{w}_l) \\
&\quad \textstyle\sum_{U_{d,t-1}} p(U_{d,t}|U_{d,t-1}) \\
&\qquad\qquad p(U_{d,t-1}|Y_{:,1:t-1}, C_{:,1:t-1,d}, Z_{d,:}, H_{:,d}^{(:)}, X_{:,1:t-1}).
\end{aligned}$$

The likelihood function of label $X_{l,t}$ is calculated only for labeled frames, which means a semi-supervised learning setup. Finally, to sample the trajectory $U_{d,1}, \ldots, U_{d,T}$, we sample $U_{d,T}$ from $p(U_{d,T}|Y_{:,1:T}, C_{:,1:T,d}, Z_{d,:}, H_{:,d}^{(:)}, X_{:,1:T})$ and perform a backward pass where we sample $U_{d,t}$ given the sample for $U_{d,t+1}$:

$$\begin{aligned}
&p(U_{d,t}|U_{d,t+1}, Y_{:,1:t}, C_{:,1:t,d}, Z_{d,:}, H_{:,d}^{(:)}, X_{:,1:t}) \\
&\qquad \propto p(U_{d,t}|Y_{:,1:t}, C_{:,1:t,d}, Z_{d,:}, H_{:,d}^{(:)}, X_{:,1:t}) p(U_{d,t+1}|U_{d,t}).
\end{aligned}$$

4. EVALUATION

We evaluate the performance of our model using the first five minutes of an audio podcast which consists of an English-learning program. The signal has a 16 kHz sampling rate with 16 bit resolution. The top of Fig. 4 shows the manually annotated "ground truth" event labels every 100 ms with "on" being black and "off" being white. This signal contains eight kinds of audio events consisting of music, sound effects, a telephone bell, and five speakers. A magnitude spectrogram is computed using a short time Fourier transform with a 100 ms long Hanning window and no overlap. We use 24 mel-filter bank outputs for the magnitude spectrum of each frame as a acoustic feature (i.e., $\boldsymbol{Y}$ corresponds to the 24×3000 matrix).

As shown in Fig. 4, we use the annotated labels of the first (1) 50 sec., (2) 100 sec., and (3) 150 sec. as labeled data and evaluate the performance for the last 150 sec. As an evaluation measure, we use an approximate predictive distribution for label $X_{l,t}$ based on a probit function [35, 36]:

$$\begin{aligned}
p(X_{l,t} = 1|\boldsymbol{U}_t, \boldsymbol{X}_{l_N}, \boldsymbol{U}_N) &\simeq \sigma\Big(\mu_{l,t} \Big/ \sqrt{1 + \pi\sigma_{l,t}^2/8}\Big), \\
\mu_{l,t} = \boldsymbol{w}_{l_{\mathrm{MAP}}}^{\mathrm{T}} \boldsymbol{U}_t, &\quad \sigma_{l,t}^2 = \boldsymbol{U}_t^{\mathrm{T}} \boldsymbol{\Sigma}_l \boldsymbol{U}_t, \\
\boldsymbol{\Sigma}_l^{-1} &= \alpha_l \boldsymbol{I}_D + \textstyle\sum_{t=t_1}^{t_N} X_{l,t}(1 - X_{l,t}) \boldsymbol{U}_t \boldsymbol{U}_t^{\mathrm{T}}, \quad (6)
\end{aligned}$$

where $t = t_1, \ldots, t_N$ is a labeled frame, $\boldsymbol{U}_N$ denotes the set of $\{\boldsymbol{U}_{t_1}, \boldsymbol{U}_{t_2}, \ldots, \boldsymbol{U}_{t_N}\}$, and $\boldsymbol{X}_{l_N}$ denotes the set of $\{X_{l,t_1}, X_{l,t_2}, \ldots, X_{l,t_N}\}$. $\boldsymbol{w}_{l_{\mathrm{MAP}}}$ is the MAP (maximum posterior) solution and is estimated using labeled data. Then, we calculate the area under the ROC curve (AUC) by comparing $p(X_{l,1501} = 1|\cdot), p(X_{l,1502} = 1|\cdot), \ldots, p(X_{l,3000} = 1|\cdot)$ for each event with the correct labels. The truncated number of acoustic components is set at $D = 100$. The initial values of $\boldsymbol{C}$, $\boldsymbol{H}$, and $\boldsymbol{U}$ are determined by the Bayesian NMF [22]. The $H_{1:\Omega,d}$ is set at the d-th initial spectral basis (the number of initial states is set at $K_d = 1$). The $\boldsymbol{U}$ is binarized using the median value. To ensure the numerical stability of the algorithm, we replace $U_{d,t} = 0$ with $U_{d,t} = 10^{-6}$. The hyperparameters are set at $\theta_{\mathrm{a}} = 1, \theta_{\mathrm{b}}^{(0)} = 10, \theta_{\mathrm{b}}^{(1)} = 1, \theta_\beta^{(d)} = 1, \theta_\varphi^{(\omega,d)} = 1, \theta_\psi^{(\omega,d)} = 1, \theta_\alpha^{(0)} = 1$, and $\theta_\alpha^{(1)} = 1000$. The parameter inference was run for 1000 iterations.

Fig. 4 shows the activation matrix $\boldsymbol{U}$ and annotation results estimated using the training data (3). The audio frames are annotated

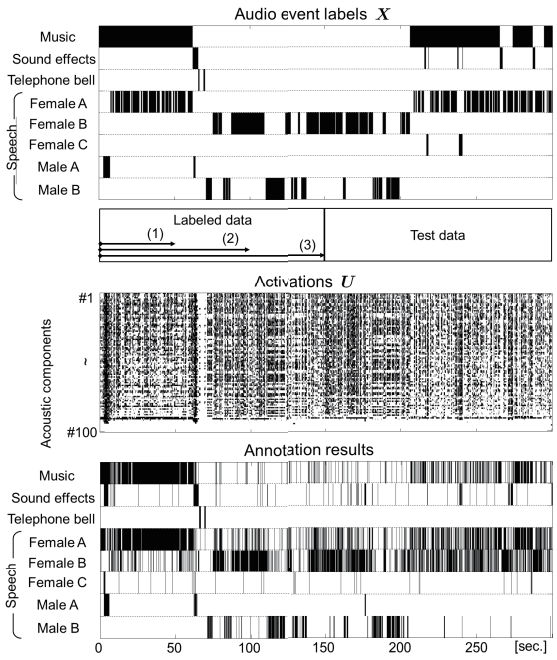

Fig. 4. Manually annotated "ground truth" event labels (top), activation matrix $\boldsymbol{U}$ (middle), and annotation results based on the predictive distribution (bottom)

according to $\hat{X}_{l,t} = \mathbb{I}(p(X_{l,t} = 1|\cdot) > 0.5)$, where $\mathbb{I}(A)$ is the indicator function for a condition A: $\mathbb{I}(A) = 1$ if A is true, and 0 otherwise. We found that the annotation results are totally close to the ground truth, but it is difficult to distinguish between female speakers. Since the labels "Telephone bell" and "Male A" are not included in the test data, we do not evaluate the performance for these events in Tab. 1. We found that the AUCs for the event labels improved as the labeled data increased.

In Tab. 2, we compare the performance of our model with that of a baseline method and determine whether the mIBP (i.e., the Markov chains and the stick breaking construction in $\boldsymbol{U}$) in our model is effective. We show the AUCs obtained by using labeled data (3). In the baseline method, we trained the distribution of the 24 mel-filter bank outputs for each event using GMM and calculated the posterior probability for each event label. We then smoothed the posterior probabilities using a 5-point moving average filter and calculated the average AUC over all audio events. Model (a) denotes the finite factor model, which assumes that the columns of $\boldsymbol{U}$ evolve according to a Markov process and that the variables $a_{(d)}$ are all independent draws from a $\mathrm{Gamma}(\theta_{\mathrm{a}}, 1)$ distribution. Model (b) denotes the infinite factor model, which assumes that the columns of $\boldsymbol{U}$ evolve independently and that the variables $a_{(d)}$ obey the stick breaking construction. Model (c) denotes the finite factor model, which assumes that the columns of $\boldsymbol{U}$ evolve independently and that the variables $a_{(d)}$ are all independent draws from a $\mathrm{Gamma}(\theta_{\mathrm{a}}, 1)$.

The performance of our model is the same or superior to that of the baseline method. Note that the baseline method cannot train the GMMs effectively when the amount of labeled data is small, such as (1) and (2). However, our model can estimate the event labels while utilizing the labeled and unlabeled data even if the amount of labeled data is small. Furthermore, our model outperforms models (a), (b), and (c). We confirmed the effectiveness of the Markov property of activations and the stick breaking construction which assumes that there are an infinite number of acoustic components in the signal.

Table 1. AUCs for event labeling with the amount of labeled data

Time length of labeled data		50 sec.	100 sec.	150 sec.
Music		0.542	0.735	**0.769**
Sound effects		0.298	0.382	**0.683**
Speech	Female A	0.647	0.766	**0.769**
	Female B	0.605	0.647	**0.744**
	Female C	0.437	0.466	**0.813**
	Male B	0.560	0.935	**0.962**
Average		0.514	0.655	**0.790**

Table 2. Comparison of our model with baseline GMM based method and degenerate models (a), (b), and (c) (see text for details)

Model	Our model	GMM	(a)	(b)	(c)
Average	**0.790**	0.734	0.722	0.686	0.693

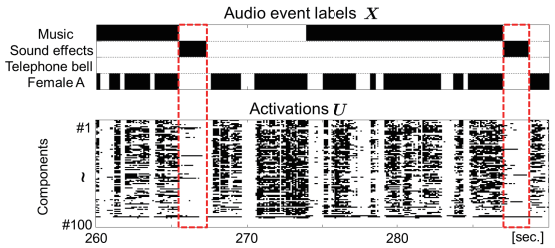

Fig. 5. Unknown acoustic events detected in activation matrix

5. DISCUSSION

It is difficult for the discriminative model to detect unknown acoustic events that are not included in the training data. We verify that our generative model is able to detect those events. Here we focus on the event "Sound effects": actually, since this event consists of different types of sounds, the performance is poor (Tab. 1). However, it is interesting to see the estimated activation matrix in Fig. 5. In the "Sound effects" segments, specific components are activated. This means that a set of the components is detected as new audio events. In the future, we plan to leverage these results effectively. However, since many acoustic components are activated in speech segments, it is necessary to improve the prior distribution of $\boldsymbol{H}$ to integrate similar acoustic features into a single acoustic component. In addition, we plan to model the gains of spectral bases using the beta-negative binomial process [40] to improve the performance.

6. CONCLUSION

To detect and locate "events" on audio signals including a variety of overlapping sounds, we proposed a new infinite generative model based on NMF into which two Bayesian nonparametric approaches and Bayesian logistic regression for the multiple labeling were incorporated. We found that our model can better annotate an mixture signal with the multiple event labels in comparison with a baseline and confirmed that the mIBP is effective as the prior distribution of an activation matrix $\boldsymbol{U}$. In addition, we confirmed that our generative model can suggest the potential for detecting unknown audio events that are not included in the training data. The future challenge is to evaluate model validity using larger data sets and reduce the computational cost involved in the parameter inference.

7. REFERENCES

[1] J. Saunders, "Real-time discrimination of broadcast speech/music," in *Proc. ICASSP 1996*.

[2] C. Clavel et al., "Events detection for an audio-based surveillance system," in *Proc. ICME 2005*.

[3] A. Temko et al., "CLEAR evaluation of acoustic event detection and classification systems," in *Multimodal Technologies for Perception of Humans*, 2007.

[4] A. Temko et al., "Acoustic event detection in meeting-room environments," *Pattern Recogn. Lett.*, vol. 30, no. 14, pp. 1281–1288, 2009.

[5] T. Butko et al., "Audio segmentation of broadcast news: A hierarchical system with feature selection for the Albayzin-2010 evaluation," in *Proc. ICASSP 2011*.

[6] K. Sumi et al., "Acoustic event detection for spotting "Hot spots" in podcasts," in *Proc. INTERSPEECH 2009*.

[7] A. Mesaros et al., "Acoustic event detection in real life recordings," in *Proc. EUSIPCO 2010*.

[8] X. Zhuang et al., "Compact audio representation for event detection in consumer media," in *Proc. INTERSPEECH 2012*.

[9] Q. Jin et al., "Event-based video restrieval using audio," in *Proc. INTERSPEECH 2012*.

[10] A. Misra, "Speech/nonspeech segmentation in web videos," in *Proc. INTERSPEECH 2012*.

[11] S. Pancoast et al., "Bag-of-Audio-Words approach for multimedia event classification," in *Proc. INTERSPEECH 2012*.

[12] T. Heittola et al., "Sound event detection in multisource environments using source separation," in *Proc. CHiME 2011*.

[13] J. Geiger et al., "Convolutive non-negative sparse coding and new features for speech overlap handling in speaker diarization," in *Proc. INTERSPEECH 2012*.

[14] M. Espi et al., "A TANDEM connectionist model using combination of multi-scale spectro-temporal features for acoustic event detection," in *Proc. ICASSP 2012*.

[15] V. Y. F. Tan and C. Fevotte, "Automatic relevance determination in nonnegative matrix factorization with the β-divergence," *IEEE PAMI*, vol. 99, 2012.

[16] Z. Zhang et al., "Semi-supervised learning helps in sound event classification," in *Proc. ICASSP 2012*.

[17] B. Byun et al., "Consumer-level multimedia event detection through unsupervised audio signal modeling," in *Proc. INTERSPEECH 2012*.

[18] A. Kumar et al., "Audio event detection from acoustic unit occurrence patterns," in *Proc. ICASSP 2012*.

[19] S. Chaudhuri et al., "Exploiting temporal sequence structure for semantic analysis of multimedia," in *Proc. INTERSPEECH 2012*.

[20] T. Virtanen et al., "Bayesian extensions to non-negative matrix factorization for audio signal modeling," in *Proc. ICASSP 2008*.

[21] A. Ozerov et al., "Factorial scaled hidden Markov model for polyphonic audio representation and source separation," in *Proc. WASPAA 2009*.

[22] A. T. Cegmil, "Bayesian inference in non-negative matrix factorization models," in *University of Cambridge*, 2008.

[23] M. Nakano et al., "Infinite-state spectrum model for music signal analysis," in *Proc. ICASSP 2011*.

[24] J. V. Gael et al., "The infinite factorial hidden Markov model," in *Proc. NIPS 2008*.

[25] D. J. Aldous, "Representations for partially exchangeable arrays of random variables," *Journal of Multivariate Analysis*, vol. 11, pp. 581–598, 1981.

[26] J. S. Liu, "The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem," *JASA*, vol. 89, no. 427, pp. 958–966, 1994.

[27] M. Hoffman, "Poisson-uniform nonnegative matrix factorization," in *Proc. ICASSP 2012*.

[28] T. L. Griffiths et al., "Infinite latent feature models and the Indian buffet process," in *Proc. NIPS 2006*.

[29] Y. W. Teh et al., "Stick-breaking construction for the Indian buffet process," in *Proc. NIPS 2007*.

[30] T. S. Ferguson, "A Bayesian analysis of some nonparametric problems," *Annals of Statistics*, vol. 1, pp. 209–230, 1973.

[31] J. Sethuraman, "A constructive definition of Dirichlet priors," *Statistica Sinica*, vol. 4, pp. 639–650, 1994.

[32] H. Ishwaran and M. Zarepour, "Exact and approximate sum-representations for the Dirichlet process," *Canadian Journal of Statistics*, vol. 30, pp. 269–283, 2002.

[33] D. M. Roy and Y. W. Teh, "The Mondrian process," in *Proc. NIPS 2009*.

[34] Y. W. Teh et al., "Hierarchical Bayesian nonparametric models with applications," Cambridge University Press, 2010.

[35] D. Spiegelhalter et al., "Sequential updating of conditional probabilities on directed graphical structures," *Networks*, vol. 20, pp. 579–605, 1990.

[36] D. J. C. MacKay, "The evidence framework applied to classification networks," *Neural Computation*, vol. 4, pp. 448–472, 1992.

[37] D. M. Blei and J. D. McAuliffe, "Supervised topic models," in *Proc. NIPS 2007*.

[38] S. L. Scott, "Bayesian methods for hidden Markov models: Recursive computing in the 21st century," *JASA*, vol. 97, pp. 337–351, 2002.

[39] R. M. Neal, "Slice sampling," *Annals of Statistics*, vol. 31, pp. 705–767, 2003.

[40] M. Zhou et al., "Beta-negative binomial process and poisson factor analysis," in *Proc. AISTATS 2012*.