Supervised acoustic topic model for unstructured audio information retrieval

Samuel Kim, Panayiotis Georgiou, and Shrikanth Narayanan

Signal Anlaysis and Interpretation Lab. (SAIL)
University of Southern California, Los Angeles, USA.

kimsamue@usc.edu

***Abstract*—We introduce a modified version of the acoustic topic model, which assumes an audio signal consists of latent acoustic topics and each topic can be interpreted as a distribution over acoustic words, for unstructured audio information retrieval applications. The proposed supervised acoustic topic model is based on supervised latent Dirichlet allocation (sLDA) while the conventional acoustic topic model is built upon latent Dirichlet allocation (LDA) which learns its parameters in an unsupervised manner. The experimental results with BBC Sound Effects Library indicate that the supervised acoustic model brings benefits in terms of classification accuracy by learning parameters with respect to corresponding categories of audio clips, i.e., semantic and onomatopoeic labels.**

***Index Terms* — audio information retrieval, acoustic topic model, unstructured audio, supervised LDA**

I. Introduction

A generic audio signal is heterogeneous in the sense that there may exist more than one distinct sound sources mixed together in a sound clip. Since how the different sound sources are combined in a generic audio signal is typically not known and difficult to estimate, retrieving desired information from generic audio signals is very challenging.

Researchers have been showing promising results in classifying generic audio clips with pre-defined descriptive categories using various machine learning approaches, such as with Gaussian mixture model (GMM) [1] and hidden Markov model (HMM) [2], [3]. These types of machine learning algorithms are usually trained in a supervised manner which requires corresponding labels at the training phase. On the other hand, various unsupervised learning methods based on latent variables have also been proposed [4], [5], [6]. Sundaram *et al.* introduced a latent perceptual indexing (LPI) method based on latent semantic analysis (LSA) [4]. Lee *et al.* [5] and Zeng *et al.* [6] applied a modified version of LSA, probabilistic latent semantic analysis (pLSA), for generic audio categorization and consumer video classification using sound track, respectively.

Recently, we have proposed the acoustic topic model to characterize unstructured audio signal for information retrieval tasks [7]. The acoustic topic model is based on the topic model that was originally developed for text processing applications. It assumes that text documents consist of hidden topics and each topic in turn can be interpreted as a distribution over words in a dictionary [8], [9]. This assumption enables the use of generative model like latent Dirichlet allocation (LDA). Our previous work [7], [10] had successfully adopted the topic model ideas into audio information retrieval applications by drawing analogies between audio signals and text documents.

These unsupervised learning methods usually require consequent classifiers, such as k-nearest neighborhood (kNN) or support vector machines (SVM), to perform pattern recognition. Therefore, the classification performance also depends on the specific classifiers rather than audio modeling procedure itself.

In this work, we propose the supervised version of acoustic topic model to associate the categorical labels of sound clips with latent acoustic topics; specifically we apply the supervised LDA (sLDA) method introduced in [11], [12]. The rationale behind this is that considering categorical labels in learning latent variables might endow discriminant power rather than treating the acoustic topic modeling and the classification processes separately and independently.

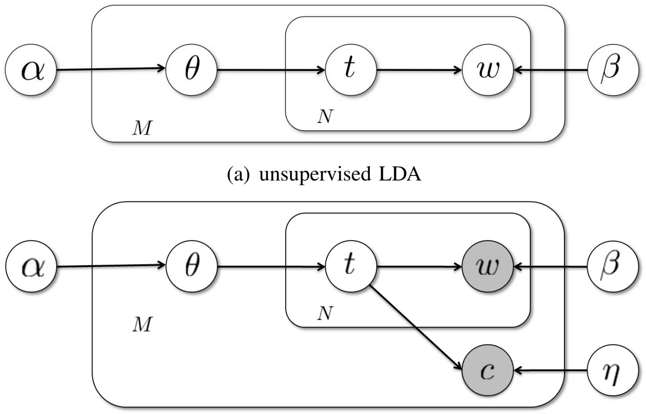
The organization of this paper is as follows. In the next section, we provide a brief review of acoustic topic models using LDA and supervised LDA methods: the commonalities and differences. The experimental setup and results are discussed in Section III and Section IV, respectively, followed by the conclusions in Section V.

II. Acoustic Topic Models versus Supervised Acoustic Topic Models

A. Acoustic Topic Model

The unsupervised acoustic topic model utilizes the LDA method. Fig. 1(a) illustrates the basic concept of LDA in a graphical representation, a three-level hierarchical Bayesian model. Let V be the number of words in dictionary $\mathcal{W}$ and w be a V-dimensional vector whose elements are zero except the corresponding word index in the dictionary. A document consists of N words, and it is represented as $\mathbf{d} = \{w_1, w_2, \cdots, w_i, \cdots, w_N\}$ where w_i is the ith word in the document. A data set consists of M documents and it is represented as $S = \{\mathbf{d_1}, \mathbf{d_2}, \cdots, \mathbf{d_M}\}$. In this work, we define k latent topics and assume that each word w_i is generated by its corresponding topic. The generative process can be described as follows:

1) For each document $\mathbf{d}$, choose $\theta \sim Dir(\alpha)$



(a) unsupervised LDA

(b) supervised LDA

Fig. 1. Graphical representation of topic models: (a) unsupervised LDA and (b) supervised LDA.

2) For each word w_i in document $\mathbf{d}$,
   a) Choose a topic $t_i \sim Multinomial(\theta)$
   b) Choose a word w_i with a probability $p(w_i|t_i, \beta)$, where β denotes a $k \times V$ matrix whose elements represent the probability of a word with a given topic, i.e. $\beta_{ij} = p(w^j = 1|t^i = 1)$. The superscripts represent element indices of individual vectors, while the subscripts represent vector indices.

B. Supervised Acoustic Topic Model

As pointed out in the previous section, the LDA-based acoustic topic model is trained in an unsupervised manner which does not require any labels during learning phase. The proposed supervised acoustic topic model utilizes a modified version of LDA as shown in Fig. 1(b) which shares most of properties with unsupervised LDA except it includes a node c that represents the category of a document and a kernel function η that transfers the topic distribution t to the categories. The generative process can be described as follows:

1) For each document $\mathbf{d}$, choose $\theta \sim \text{Dir}(\alpha)$
2) For each word w_i in document $\mathbf{d}$,
   a) Choose a topic $t_i \sim \text{Multinomial}(\theta)$
   b) Choose a word w_i with a probability $p(w_i|t_i, \beta)$
3) Choose class label $c|t \sim softmax(\bar{t}, \eta)$,
   where $\bar{t}$ represents the topic frequency of a document, i.e., $\bar{t} = \frac{1}{N}\sum_{n=1}^{N} t_n$. The probability of a certain class with given $\bar{t}$ and η can be represented as

$$p\left(c|\bar{t}, \eta\right) = \frac{\exp({\eta_c}^T \bar{t})}{\sum_{c'=1}^{C} \exp({\eta_{c'}}^T \bar{t})} \tag{1}$$

C. Variational Approximate Inference

Computing exact values is not computationally feasible because it involves intractable integral operations. To solve this problem, various approaches such as Laplace approximation and Gibbs sampling method, have been proposed. In this work, we utilize the variational inference method. The rationale behind the method is to minimize the distance between the real distribution and the simplified distribution using Jensen's inequality.

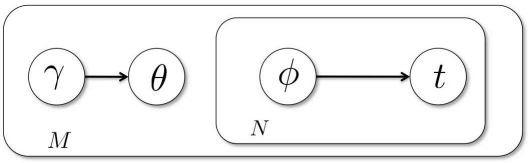


Fig. 2. Graphical representation of simplified version of topic models for variational approximation.

The simplified version has γ and ϕ which, respectively, are the Dirichlet parameter that determines θ and the multinomial parameter that generates topics, as depicted in Fig. 2. Note that this variational approximate method is valid for both unsupervised LDA and supervised LDA, since the node c in supervised LDA is not associated with any latent variable.

The joint probability of θ and $\mathbf{t}$ can be represented as

$$\begin{aligned} q\left(\theta, \mathbf{t}|\gamma, \phi\right) &= q(\theta|\gamma)q(\mathbf{t}|\phi) \\ &= q(\theta|\gamma)\prod_{i=1}^{N} q(t_i|\phi_i) \end{aligned} \tag{2}$$

and tries to minimize the difference between real and approximated joint probabilities using Kullback-Leibler (KL) divergence, i.e.

$$\arg\min_{\gamma,\phi} D(q(\theta, \mathbf{t}|\gamma, \phi)||p(\theta, \mathbf{t}|\mathbf{w}, \alpha, \beta)) \tag{3}$$

for unsupervised LDA and

$$\arg\min_{\gamma,\phi} D(q(\theta, \mathbf{t}|\gamma, \phi)||p(\theta, \mathbf{t}|\mathbf{w}, c, \alpha, \beta)) \tag{4}$$

for supervised LDA.

If we take a partial derivative with respect to γ_n and ϕ_{in}, we can obtain the following iterative process to minimize the difference between real and approximated joint probability:

- Unsupervised LDA

$$\gamma_n = \alpha_n + \sum_{i=1}^{N} \phi_{in} \tag{5}$$

$$\phi_{in} \propto \beta_{n\tau} \exp\left(\Psi(\gamma_n) - \Psi\left(\sum_{j=1}^{k} \gamma_j\right)\right) \tag{6}$$

- Supervised LDA

$$\gamma_n = \alpha_n + \sum_{i=1}^{N} \phi_{in} \tag{7}$$

$$\begin{aligned} \phi_{in} \propto \beta_{nm} \exp\left(\Psi(\gamma_n) - \Psi\left(\sum_{j=1}^{k} \gamma_j\right)\right) \\ \cdot \exp\left(\frac{1}{N}\eta_{cn} - (h^T {\phi_i}^{old})^{-1} h_n\right) \end{aligned} \tag{8}$$

where ϕ_i^{old} represents the value of ϕ_i at the previous iteration and h represents a simplified linear function of ϕ_i (see [12] for more details).

Note that both approaches share the same update for Dirichlet parameter γ while update for ϕ_{in} is scaled according to the kernel function η and the previous ϕ_i. The main difference between LDA and sLDA lies in this update.

D. Classification

With the BBC Sound Effects Library (details are given in Section III-B), we perform a 5-fold classification task with the onomatopoeic and semantic labels of audio clips.

1) Unsupervised acoustic topic model: For each training session, we estimate the LDA parameters and train the linear-kernel support vector machine (SVM) with Dirichlet parameters γ as representative feature vectors of individual sound clips. For each test session, in turn, we infer Dirichlet parameters γ based on the estimated parameters from the training session and perform classification tasks using the SVM classifier.

2) Supervised acoustic topic model: Since the models using sLDA are trained with corresponding labels, we can classify test audio clips without extra consequent classifiers. Inferring a class category from sLDA-based models requires some approximation processes as well, such as variational approximation and Jensen's inequality [12]. The inference can be written as follows.

$$\hat{c} = \arg\max p(c|w) \tag{9}$$

where

$$\begin{aligned} p(c|w) &\approx \int p\left(c|t\right) q(t)dt \\ &= \int \frac{\exp({\eta_c}^T \bar{t})}{\sum_{c'=1}^{C} \exp({\eta_{c'}}^T \bar{t})} q(t)dt \\ &\geq \exp\left(E_q\left[{\eta_c}^T \bar{t}\right] - E_q\left[\log\left(\sum_{c'=1}^{C} \exp\left({\eta_{c'}}^T \bar{t}\right)\right)\right]\right) \end{aligned} \tag{10}$$

Since the second term is common for all classes, we can infer the class which maximizes the first term, i.e.,

$$\begin{aligned} \hat{c} &= \arg\max E_q\left[{\eta_c}^T \bar{t}\right] \\ &= \arg\max {\eta_c}^T \bar{\phi} \end{aligned} \tag{11}$$

where $\bar{\phi} = \frac{1}{N}\sum_{n=1}^{N} \phi_n$.

III. Experimental Setup

A. Acoustic word

We use the notion of acoustic words so that an audio signal can be represented with word-like discrete indices. After extracting feature vectors that describe acoustic properties of a given segment, we assign acoustic words based on the closest word in the pre-trained acoustic dictionary.

1) Acoustic features: Using frame-based analysis, we calculate mel frequency cepstral coefficients (MFCC) to represent the audio signal's acoustic properties. The MFCCs provide spectral information considering human auditory properties and have been widely used in many sound related applications, such as speech recognition and audio classification [13]. In this work, we used 20 ms Hamming windows with 50% overlap to extract 12-dimensional feature vectors.

2) Acoustic Dictionary: With a given set of acoustic features, we derived an acoustic dictionary of codewords using the *Linde-Buzo-Gray Vector Quantization* (LBG-VQ) algorithm [14]. The rationale is to cluster audio segments which have similar acoustic characteristics and to represent them as discrete code words (indexed appropriately). In this experiment, we set the number of acoustic words to be 1,000 and the number of latent acoustic topics be 100.

B. Database

A selection of 2,140 audio clips from the BBC Sound Effects Library [15] was used for the experiments. Each clip is annotated in three different ways: single-word semantic labels, onomatopoeic labels, and short multi-word descriptions. The semantic labels and short descriptions are made available as a part of the database and belong one of 21 predetermined categories. They include general categories such as *transportation*, *military*, *ambiences*, and *human*. There was no existing annotation in terms of onomatopoeic words; therefore we undertook this task through subjective annotation of all audio clips. We asked subjects to label the audio clip by choosing from among 22 onomatopoeia descriptions. For more details on the annotation process, please refer to [4]. The audio clips were available in two-channel format with 44.1kHz sampling rate and were down-sampled to 16kHz (mono) for acoustic feature extraction. The average audio clip length is about 13 seconds and generates about 1,300 acoustic words. A summary of the database is given in Table I.

IV. Results and Discussion

Fig. 3 illustrates the classification results of audio clips using latent acoustic topics with both LDA and sLDA along with their standard deviations in error bars (Table II shows the performance in numbers along with relative improvements).

As shown in the figure, the accuracy rates using the supervised acoustic topic model are higher than the ones using conventional acoustic topic model for both onomatopoeic and semantic labels (11.9% and 9.3 % relative improvements for semantic labels and onomatopoeic labels, respectively). This significant improvement is from using sLDA instead of LDA;

TABLE I
Summary of BBC Sound Effect Library.

Number of sound clips	2,140
Number of semantic categories	21
Number of onomatopoeic words	22
Average length of an audio clip	13 sec

TABLE II
CLASSIFICATION RESULTS OF AUDIO CLIPS USING LATENT ACOUSTIC TOPICS WITH LDA AND SUPERVISED LDA.

Accuracy (%)	LDA	sLDA	Relative Improvement
Semantic labels	38.8	**43.4**	11.9
Onomatopoeic labels	32.1	**35.1**	9.3

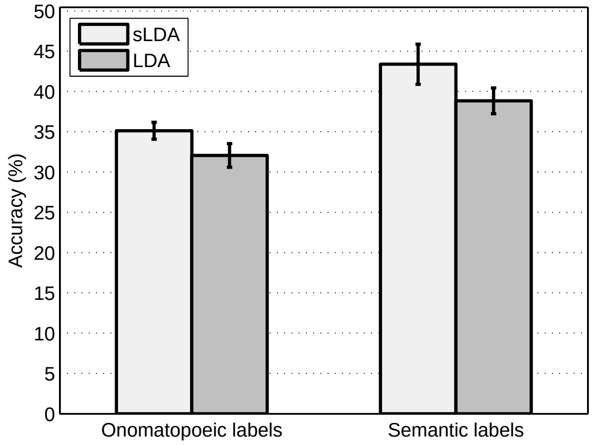


Fig. 3. Classification results of audio clips using latent acoustic topics with LDA and supervised LDA.

sLDA learns its parameters according to categories of training data, while LDA does not consider the categories. Instead, LDA uses a consequence classifier (SVM, in this work) for classification tasks so that the acoustic topic modeling process is independent of descriptive categories.

The conventional acoustic model using LDA, however, has some advantages over the supervised acoustic model using sLDA, besides the fact that sLDA requires significant greater computational power than LDA does. Since LDA learns the latent variables in an unsupervised manner without considering the labels, one can apply various types of categories and classifiers without re-learning the parameters. For example, in this work, we trained two separate supervised acoustic topic models for two different descriptive categories, i.e., onomatopoeic and semantic labels, while we could train only one acoustic topic model for both descriptive categories and use consequent SVM for classification tasks.

V. CONCLUDING REMARKS

In this work, we investigated the effects of supervised acoustic topic models over the conventional acoustic topic model within the unstructured audio information retrieval framework. While the conventional acoustic topic model utilizes a latent Dirichlet allocation (LDA) method, we adopted a modified version, supervised LDA (sLDA), which considers categorical labels during learning latent variables. The experimental results with BBC Sound Effects Library showed that the supervised acoustic topic model using sLDA outperforms the conventional acoustic topic model with LDA; it indicates that the supervised acoustic model brings benefits in terms of classification accuracy by learning parameters considering corresponding descriptive categories of audio clips rather than unsupervised learning.

In the future, we plan to investigate various types of supervised topic models within audio information retrieval framework, such as discLDA [16] and labeled LDA [17]. We will also investigate different types of kernel functions in the sLDA to improve the classification performance.

ACKNOWLEDGEMENT

This research was supported in part by the fund from the National Science Foundation (NSF).

REFERENCES

[1] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 2, pp. 467–476, Feb. 2008.

[2] L. Ma, B. Milner, and D. Smith, "Acoustic environment classification," *ACM Transactions on Speech and Language Processing*, 2006.

[3] G. Wichern, J. Xue, H. Thornburg, B. Mechtley, and A. Spanias, "Segmentation, indexing, and retrieval for environmental and natural sounds," *IEEE Transactions on Audio, Speech, and Language Processing*, 2010.

[4] S. Sundaram and S. Narayanan, "Classification of sound clips by two schemes: using onomatopoeia and semantic labels," in *IEEE International Conference of Multimedia and Expo*, 2008.

[5] K. Lee and D. Ellis, "Audio-based semantic concept classification for consumer video," *IEEE Transactions on Audio, Speech, and Language Processing*, 2010.

[6] Z. Zeng, H. Li, W. Liang, and S. Zhang, "A hierarchical generative model for generic audio document categorization," in *IEEE International Conference of Acoustics, Speech, and Signal Processing*, 2010.

[7] S. Kim, S. Narayanan, and S. Sundaram, "Acoustic topic models for audio information retrieval," in *Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2009.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, 2003.

[9] T. L. Griffiths, M. Steyvers, and J. B. Tenenbaum, "Topics in semantic representation," *Psychological Review*, vol. 114, no. 2, pp. 211–244, 2007.

[10] S. Kim, S. Sundaram, P. Georgiou, and S. Narayanan, "Audio scene understanding using topic models," in *Neural Information Processing System (NIPS) Workshop (Applications for Topic Models: Text and Beyond)*, 2009.

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

[12] C. Wang, D. M. Blei, and L. Fie-Fei, "Simulateneous image classification and annotation," in *CVPR*, 2009.

[13] L. R. Rabiner and B. Juang, *Fundamentals of speech recognition*. Prentice Hall, 1993.

[14] A. Gersho and R. M. Gray, *Vector quantization and signal compression*. Norwell, MA, USA: Kluwer Academic Publishers, 1991.

[15] The BBC sound effects library - original series. [Online]. Available: http://www.sound-ideas.com

[16] S. LaCoste-Jullien, F. Sha, and M. I. Jordan, "Disclda: Discriminative learning for dimensionality reduction and classification," in *Neural Information Processing Systems*, 2008.

[17] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, August 2009, pp. 248–256.