Supervised acoustic topic model with a consequent classifier for unstructured audio classification

Samuel Kim[1,2], Panayiotis Georgiou[2], Shrikanth Narayanan[2]

[1] IDIAP Research Institute, Martigny, Switzerland
[2] Signal Analysis and Interpretation Lab. (SAIL), University of Southern California
samuel.kim@idiap.ch, {georgiou,shri}@sipi.usc.edu

Abstract

In the problem of classifying unstructured audio signals, we have reported promising results using acoustic topic models assuming that an audio signal consists of latent acoustic topics [1, 2]. In this paper, we introduce a two-step method that consists of performing supervised acoustic topic modeling on audio features followed by a classification process. Experimental results in classifying audio signals with respect to onomatopoeias and semantic labels using the BBC Sound Effects library show that the proposed method can improve the classification accuracy relatively $10 \sim 14\%$ *against the baseline supervised acoustic topic model. We also show that the proposed method is compatible with different labels so that the topic models can be trained with one set of labels and used to classify another set of labels.*

1 Introduction

Latent topic models have been widely used in various signal processing domains including text and image [3, 4, 5] due to their potential for offering meaningful insights into the underlying latent structure of a data set. Topic models were originally proposed for text-based information retrieval systems to tackle ambiguities due to context dependency (e.g. the word 'bank' can be interpreted differently depending on context in which the word is located; related to a river or a financial institution) and have been applied in various domains owing largely to their generic assumption of latent structure.

Our previous work focused on applying latent topic models in audio signal processing by drawing analogies between text documents and audio clips, specifically for characterizing unstructured audio [1, 2]. We hypothesized that each audio clip consists of a number of latent acoustic topics and these latent acoustic topics generate the acoustic segments that constitute an audio clip. Experimental results in classifying unstructured audio signals with respect to descriptive semantic and onomatopoeic labels supports such a modeling hypothesis. In [1], we proposed a two-step modeling strategy which first models the latent acoustic topics in an unsupervised manner and then learns the topic distributions in a supervised manner along with corresponding labels. In contrast, in [2], we used a supervised topic model which models the latent acoustic topics in a supervised manner along with corresponding labels so that it does not require a subsequent supervised classifier. By incorporating the corresponding labels in learning latent acoustic topics, we showed some improvements in the classification accuracy at the cost of complexity.

In this paper, we extend the work of [2] to further investigate supervised acoustic topic models. While the classification tasks in [2] utilize posterior probabilities embedded in the supervised acoustic topic model, here we introduce a two-layer strategy that consists of the supervised acoustic topic models as the first layer and a classifier (e.g., a support vector machine) that employes as features the parameters of the topic modeling layer as the second layer.

Another contribution of this paper comes from the study on multiple types of audio descriptors in the proposed two-step modeling framework. There are many ways for humans to describe what they hear [6], and we have focused on semantic descriptors and onomatopoeic descriptors assuming that these two categories would provide an intermediate descriptive layer which bridges naïve descriptions and audio classes [7]. Although we have demonstrated our classification results based on these two categories individually [1, 2], the interplay between individual categories has been hitherto neglected. In this work, we show the interplay by training the supervised acoustic topic models with one set of descrip-

978-1-4673-2369-7/12/$31.00 ©2012 IEEE

tors and using the supervised machine learning method with the other set of descriptors. The motivation behind is to resolve one of the disadvantages of the supervised acoustic topic models, i.e., the compatibility between labels; the supervised acoustic topic modeling method requires separate models for individual descriptor classes while the conventional acoustic topic model can share the topic models for different sets of descriptors [2]. The empirical results indicate that the supervised acoustic models trained with an existing set of descriptors can be also used for an unseen set of descriptors.

The organization of this paper is as follows. In the next section, we provide a brief review of acoustic topic models using Latent Dirichlet Allocation (LDA) and supervised LDA (sLDA) and classification strategies for individual acoustic topic models. The experimental setup and results are discussed in Section 3 and Section 4, respectively, followed by the conclusions in Section 5.

2 Acoustic Topic Models

In this section, we provide a brief review of the implementations of acoustic topic models using LDA and sLDA. We also provide a description of classification strategies according to the types of acoustic topic models in each subsection.

2.1 Unsupervised acoustic topic model

2.1.1 LDA

The unsupervised acoustic topic models adopted here utilize 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. Note that the words are discretized audio features in this work (more details are provided in Section 3.2). A document which consists of N words 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 which consists of M documents is represented as $S = \{\mathbf{d_1}, \mathbf{d_2}, \cdots, \mathbf{d_M}\}$. We also 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 $\mathbf{d}$ in data set S
   (a) Choose the topic distribution $\theta \sim Dir(\alpha)$ where $Dir(\cdot)$ and α represent a Dirichlet distribution and its Dirichlet coefficient, respectively.
2. For each word w_i in document $\mathbf{d}$,
   (a) Choose a topic $t_i \sim \text{Multi}(\theta)$ where t_i is the topic that corresponds with the word w_i and $\text{Multi}(\cdot)$ represents a multinomial distribution.
   (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., $\beta_{nm} = p(w_i = m|t_i = n)$.

Since estimating and inferring parameters involves intractable integral operations, computing exact values is not feasible. To solve this problem, in this work, we utilize the variational inference method which minimizes the distance between the real distribution and the simplified distribution using Jensen's inequality [8].

2.1.2 LDA-SVM

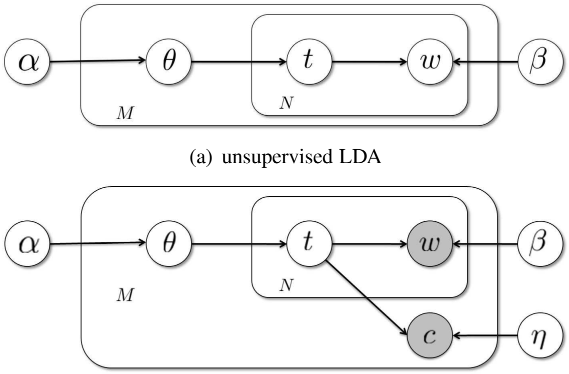
Note that LDA models the latent topics in an unsupervised manner so that it does not require any label information during the topic modeling process. One might need a consequent machine learning algorithm that learns corresponding labels for classification tasks. In our previous work, we had proposed a two-step learning procedure called LDA-SVM which first models the latent acoustic topics in an unsupervised manner and then learns the topic distribution with support vector machine (SVM) in a supervised manner along with corresponding labels [1]. For each training session, we estimate the LDA parameters and train the SVM with topic distributions θ as representative feature vectors of individual sound clips. For each test session, in turn, we infer the topic distributions based on the estimated parameters from the training session and perform classification tasks using the SVM classifier.

2.2 Supervised acoustic topic model

2.2.1 sLDA

The supervised acoustic topic models, unlike the unsupervised acoustic topic models, utilize 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 distributions t to the categories [9]. The generative process can be described as follows:

1. For each document $\mathbf{d}$ in data set S
   (a) Choose the topic distribution $\theta \sim \text{Dir}(\alpha)$



(a) unsupervised LDA

(b) supervised LDA

Figure 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 $t_i \sim \text{Multi}(\theta)$
   (b) Choose a word w_i with a probability $p(w_i|t_i, \beta)$
3. For each document $\mathbf{d}$
   (a) 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

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

To estimate and infer parameters, we also use the variational inference as done in section 2.1.1.

2.2.2 sLDA maximum likelihood

Since the models using sLDA are already trained with corresponding labels, we can classify test audio clips without consequent classifiers. The classification can be directly performed by estimating the probability of classes with given observations, i.e.,

$$\hat{c} = \arg\max p(c|w)\ . \tag{2}$$

The above posterior probability can be approximately estimated using a variational distribution q, i.e.

$$p(c|w) \approx \int p\left(c|t\right) q(t) dt \tag{3}$$

and can be further simplified using Jensen's inequality as

$$\begin{aligned}
&\int p\left(c|t\right) q(t) dt \\
&\quad = \int \frac{\exp({\eta_c}^T \bar{t})}{\sum_{c'=1}^{C} \exp({\eta_{c'}}^T \bar{t})} q(t) dt \\
&\quad \geq \exp\left(E_q\left[{\eta_c}^T \bar{t}\right] - E_q\left[\log\left(\sum_{c'=1}^{C} \exp\left({\eta_{c'}}^T \bar{t}\right)\right)\right]\right)\ .
\end{aligned} \tag{4}$$

Since the second term of (4) is common for all classes, we can infer the class which maximizes the first term, i.e.,

$$\begin{aligned}
\hat{c} &= \arg\max E_q\left[{\eta_c}^T \bar{t}\right] \\
&= \arg\max {\eta_c}^T E_q\left[\bar{t}\right]
\end{aligned} \tag{5}$$

2.2.3 sLDA-SVM

Instead of using the above approximated posterior probabilities for classification tasks, we can alternatively introduce a two-step learning procedure that uses the sLDA parameters as feature vectors of a consequent classifier (SVM in this work). Similar to the LDA-SVM method, for each training session, we estimate the sLDA parameters and train the SVM with topic distributions as representative feature vectors of individual sound clips. For each test session, we infer the topic distributions and perform classification tasks using the SVM classifier.

This two-step strategy allow us to use different categories of audio descriptors in each step, since they do not have to be identical. By utilizing this possibility, we train the supervised acoustic topic models with one set of labels and use the supervised machine learning method with the other set of labels (e.g., topic models with onomatopoeias and SVM models with semantic labels).

3 Experimental Setup

3.1 Database

A selection of 2,140 audio clips from the BBC Sound Effects Library [10] was used for the experiments. Each clip is annotated with both semantic labels and onomatopoeic labels. The semantic labels and short descriptions are made available as a part of the database and belong in 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 onomatopoeic words [11].

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. A summary of the database is given in Table 1.

3.2 Experimental setup

To define acoustic words, in this paper, we use conventional mel frequency cepstral coefficients (MFCCs) to parameterize the audio signal and use vector quantization (VQ) to derive the acoustic words. Using fixed length frame-based analysis (20 ms hamming windows with 50% overlap), we calculate 12-dimensional MFCCs to represent the audio signal's time varying acoustic properties. With a given set of acoustic features, we derived an acoustic dictionary of codewords using the *Linde-Buzo-Gray Vector Quantization* (LBG-VQ) algorithm [12]. In our experiments, we choose the size of acoustic dictionary to be 1,000.

We perform a 5-fold cross validation; randomly partitioning the database into five exclusive equal-size subsets and retain one subset for testing while using the rest for training. All the training procedures such as building an acoustic dictionary, modeling acoustic topics (either supervised or unsupervised), and training SVM classifiers (with Bhattacharyya kernel [13]) are done using the training subsets. Then, the classification performance is obtained on the held out test subsets.

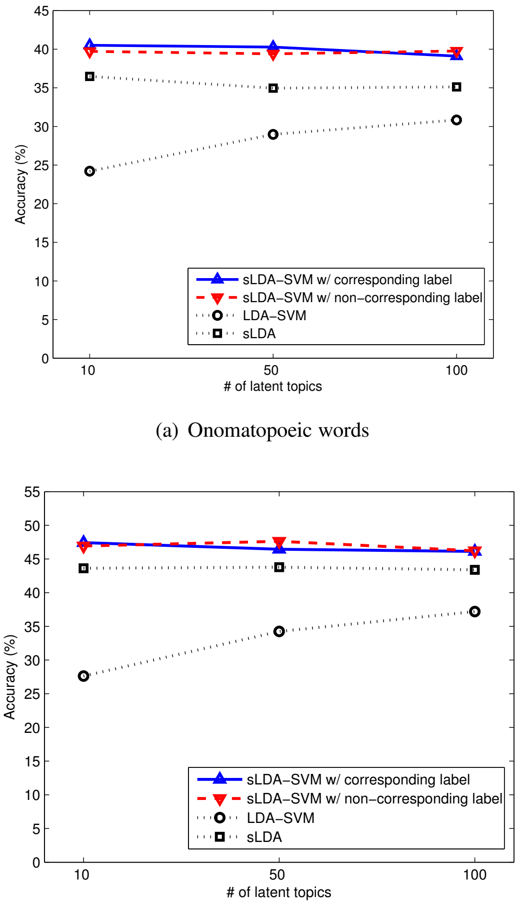
4 Results and Discussion

Fig. 2 (a) and (b) show the results of audio classification tasks using various acoustic latent topic models according to the number of topics for onomatopoeic words and semantic labels, respectively. Individual curves correspond to different approaches; dotted lines with circles for LDA-SVM, dotted lines with squares for sLDA, solid lines with upper triangles for sLDA-SVM with corresponding labels, and dashed lines with lower triangles for sLDA-SVM with non-corresponding labels. The sLDA-SVM with corresponding labels means that the classification tasks with SVM are performed with the same set of labels that sLDA is trained with, while the sLDA-SVM with non-corresponding labels means that the classification tasks with SVM are performed with the other set of labels that sLDA is not trained with. For instance, in classifying onomatopoeic words, results for sLDA-SVM with corresponding label mean that both supervised acoustic topic models and SVM are trained and tested with respect to onomatopoeic labels while results for sLDA-SVM with non-corresponding label mean that supervised acoustic topic models are learnt with semantic labels and SVM uses onomatopoeic labels.

Table 1. 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

From the results, we observe that the tasks with sLDA and sLDA-SVM outperform the ones with LDA-SVM regardless of number of latent topics (this is consistent with the results reported in [2]). Particularly, the clas-



(a) Onomatopoeic words

(b) Semantic labels

Figure 2. Audio classification results using various acoustic topic model approaches according to the number of latent topics: (a) onomatopoeic words and (b) semantic labels.

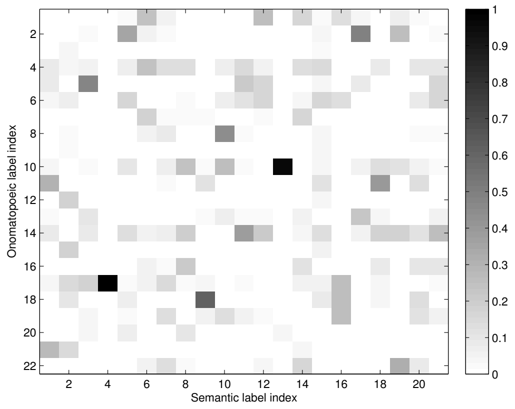


Figure 3. Co-occurrence of onomatopoeias and semantic labels in terms of conditional probabilities of onomatopoeias with given semantic labels (the values are normalized column-wise). For instance, the semantic/onomatopoeic label pair for (4,17) and (13,10) are "door/squeak" and "music/dong", respectively.

sification accuracies in the sLDA cases are not highly dependent on the number of latent topics as much as in the LDA-SVM cases. This indicates that having class information during the topic modeling procedure helps to capture sparse topics in audio signals even with small number of topics. We also observe that deployment of two-stage strategy in the sLDA framework provides about $10 \sim 14\%$ relative improvement in classifying both sets of descriptors.

It is remarkable that there is no significant difference between corresponding labels and non-corresponding labels using the sLDA-SVM in terms of accuracy. Recall that this compatibility issue between different categories was acknowledged as one of drawbacks of the supervised acoustic topic models; the supervised acoustic topic modeling method requires separate models for individual descriptor classes while the conventional acoustic topic model can share the topic models for different sets of descriptors [2]. One possible reason is that, as we discussed earlier, having class information during the topic modeling procedure helps to cluster audio signals into a topic space so that it can be used for a different set of classes. One might also want to note that those two sets of descriptors are highly related as seen in Fig. 3 which depicts the co-occurrence of onomatopoeias and semantic labels; the values are normalized column-wise so that they can be represented as conditional probabilities of onomatopoeias with given semantic labels. Although analyzing the interplay between those two sets of descriptors is beyond the scope of this paper, the results suggest that we can train the supervised acoustic topics with one set of labels and use the acoustic topic models for another set of labels. This is promising to use the supervised acoustic models trained with an existing set of descriptors for an unseen set of descriptors.

5 Conclusion

In this work, we introduced a modified version of the supervised acoustic topic models that combines supervised acoustic topic models with the support vector machine (SVM) as a consequent classifier. The proposed two-step method is shown to outperform the supervised acoustic topic models as wBell as the conventional acoustic topic models in classifying audio using both onomatopoeia and semantic labels. We also empirically showed that the compatibility of the proposed model with onomatopoeic and semantic descriptors. The results suggest that we can use a topic model trained with one set of descriptors to train/classify unseen set of descriptors.

In the future, we will investigate other labels which describe some other audio property, such as an affective description (either categorical or dimensional) in this supervised acoustic topic modeling framework.

6 Acknowledgement

This study was partially funded by NSF and DoD while the first author was with Signal Analysis and Interpretation Lab. (SAIL), University of Southern California.

References

[1] 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.

[2] S. Kim, P. G. Georgiou, and S. Narayanan, "Supervised acoustic topic model for unstructured audio information retrieval," in *Asia Pacific Signal and Information Processing Association (APSIPA) annual summit and conference*, 2010.

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

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

[5] M. Steyvers and T. Griffiths, *Probabilistic Topic Models*. Laurence Erlbaum, 2006.

[6] S. Wake and T. Asahi, "Sound retrieval with intuitive verbal expressions," in *International Conference on Auditory Display*, 1998.

[7] S. Kim, P. Georgiou, S. Narayanan, and S. Sundaram, "Using naive text queries for robust audio information retrieval system," in *IEEE International Conference on Acoustic, Speech, and Signal Processing*, 2010.

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

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

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

[11] S. Sundaram and S. Narayanan, "Audio retrieval by latent perceptual indexing," in *IEEE International Conference of Acoustics, Speech, and Signal Processing*, 2008.

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

[13] T. Jebara, R. Kondor, and A. Howard, "Probability product kernels," *J. Mach. Learn. Res.*, vol. 5, pp. 819–844, December 2004. [Online]. Available: http://portal.acm.org/citation.cfm?id=1005332.1016786