# ACOUSTIC TOPIC MODEL FOR AUDIO INFORMATION RETRIEVAL

*Samuel Kim, Shrikanth Narayanan*

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

kimsamue@usc.edu, shri@sipi.usc.edu

*Shiva Sundaram*

Deutsche Telekom Laboratories,
Quality and Usability Lab,
TU-Berlin, Berlin, Germany.

shiva.sundaram@telekom.de

## ABSTRACT

A new algorithm for content-based audio information retrieval is introduced in this work. Assuming that there exist hidden acoustic topics and each audio clip is a mixture of those acoustic topics, we proposed a topic model that learns a probability distribution over a set of hidden topics of a given audio clip in an unsupervised manner. We use the *Latent Dirichlet Allocation* (LDA) method for the topic model, and introduce the notion of *acoustic words* for supporting modeling within this framework. In audio description classification tasks using Support Vector Machine (SVM) on the BBC database, the proposed acoustic topic model shows promising results by outperforming the *Latent Perceptual Indexing* (LPI) method in classifying onomatopoeia descriptions and semantic descriptions.

## 1. INTRODUCTION

A significant growing body of interest exists in content-based information retrieval of multimedia data including video, music, speech, and audio. Specifically, information retrieval from audio has been studied intensively and has yielded promising results that have been reported widely. In a recent work from Google, Chechik *et. al.* successfully performed a large-scale content-based audio retrieval from text queries [1]. Their method is scalable to very large number of audio data based on a passive-aggressive model for image retrieval (PAMIR).

The definition of audio information varies according to the intended end use. One of the classic examples is speech recognition whose interest lies on only text transcription of the speech signal and suppresses any other acoustic information as noise. On the other hand, applications like environment sound recognition aim to decode ambient noise rather than evident signals [2]. In another approach, Slaney presented a framework to derive semantic descriptions of audio to signal features [3].

In this work, we focus on two different information aspects of audio data: semantic and onomatopoeia descriptions. The semantic descriptions focus on what makes sounds, while the onomatopoeia descriptions focus on how people describe what they hear. These labels are particularly interesting because they are highly related to psychoacoustic processes which connect physical properties and human experience of sounds; onomatopoeia labels can be considered from the perspective of the *sensation* process, and semantic labels from *cognition* or *perception* process [4].

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

To model these labels with content-based features directly obtained from audio signals, we utilize the topic model that was originally proposed in the context of text information retrieval [5, 6]. It models each document as a distribution over a fixed number of unobservable hidden topics. Each topic, in turn, can be modeled as a distribution over a fixed number of observable words. One of the motivations of modeling hidden topics in a document is to handle the ambiguities of interpretations of words. Although individual words have their own meanings, the interpretations of the words vary according to the context around the word and topic.

This idea has been successfully extended to content-based image information retrieval applications [7, 8, 9]. Assuming that there exist hidden "topics" behind image features, many researchers have been using the topic modeling approach in their applications. The image features are often quantized to provide discrete index numbers to resemble words in text topic modeling approach. Despite the advantages of the latent topic model, to the best of our knowledge, there have been only few efforts that have applied topic modeling to content-based sound or audio information retrieval applications. One of the first steps can be found in [10, 11]. Sundaram *et. al.* used the Latent Perceptual Index (LPI) method for classifying audio descriptions inspired by Latent Semantic Indexing (LSI) [12]. In two categories of audio descriptions, i.e., onomatopoeia and semantic descriptions, they demonstrated promising performance using latent structure in audio information retrieval applications.

In this work, we propose an *acoustic topic model* motivated by drawing analogies between text and sound. We hypothesize that short segments of audio signals play similar roles as words in text and that there are latent topics in audio signals which would be determined by the context of audio signals. In other words, each audio clip is viewed to consist of latent acoustic topics that generate acoustic words. We use Latent Dirichlet Allocation (LDA) method [13] to model the acoustic latent topics, and perform audio description classification tasks using onomatopoeia labels and semantic labels.

The paper is organized as follows. A brief overview of Latent Dirichlet Allocation (LDA) is given in Section 2 followed by the description of the proposed acoustic topic model in Section 3. Experimental set up description and results are provided in Section 4. The concluding remarks are given in Section 5.

## 2. LATENT DIRICHLET ALLOCATION

The topic model assumes that documents consist of hidden topics and each topic can be interpreted as a distribution over words in a

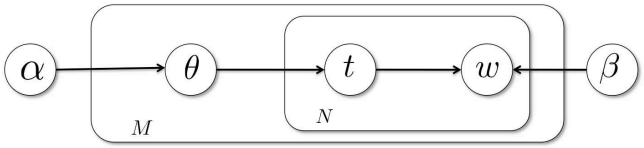



Figure 1: Graphical representation of the topic model using Latent Dirichlet Allocation.

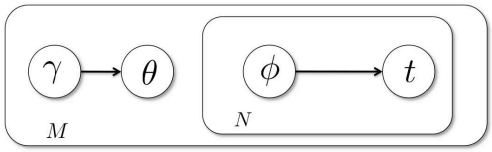



Figure 2: Graphical representation of the approximated topic model for variational inference method to estimate and infer the Latent Dirichlet Allocation parameters.

dictionary [5]. This assumption enables the generative model like Latent Dirichlet allocation (LDA). Fig. 1 illustrates a basic concept of the LDA in a graphical representation, a three-level hierarchical Bayesian model.

Let $V$ be the number of words in dictionary 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 $i$ th 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 $\theta \sim Dir(\alpha)$
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 $\beta$ 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.

Now, the question is how to estimate or infer parameters like $\mathbf{t}$, $\alpha$, and $\beta$ while the only variable we can observe is $w$. In many estimation processes, parameters are often chosen to maximize the likelihood values of a given data $\mathbf{w}$. The likelihood can be defined as

$$l(\alpha, \beta) = \sum_{w \in \mathbf{w}} \log p(w|\alpha, \beta) \, . \tag{1}$$

Once $\alpha$ and $\beta$ are estimated, the joint probability of $\theta$ and $\mathbf{t}$ with given $\mathbf{w}$ should be estimated as

$$p(\theta, \mathbf{t}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{t}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)} \, . \tag{2}$$

These processes, however, are computationally impossible because both inference and estimation requires computing $p(\mathbf{w}|\alpha, \beta)$ which includes intractable integral operations. To solve this problem, various approaches such as Laplace approximation and variational approximation, have been proposed. In this work, we utilize the variational inference method introduced in [13]. Blei *et al* have shown that this approximation works reasonably well in various applications, such as document modeling and document classification.

The rationale behind the method is to minimize distance between the real distribution and the simplified distribution using Jensen's inequality. The simplified version has $\gamma$ and $\phi$ which are the Dirichlet parameter that determines $\theta$ and the multinomial parameter that generates topics respectively, as depicted in Fig. 2. The joint probability of $\theta$ and $\mathbf{t}$ in (2) can be simplified as

$$q(\theta, \mathbf{t}|\gamma, \phi) \tag{3}$$

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

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

## 3. ACOUSTIC TOPIC MODEL

Since the topic model is originally proposed for text document modeling applications, it requires word-like discrete indexing numbers to apply the topic model as it is done in image retrieval applications. In this work, we introduce the notion of *acoustic words* to tackle this problem. 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 words dictionary. With the extracted acoustic words, we perform the Latent Drichlet Allocation (LDA) to model hidden acoustic topics in an unsupervised way. Then, we use the posterior Dirichlet parameter which describes the distribution over the hidden topics of each audio clip as a feature vector of the corresponding audio clip. Fig. 3 illustrates a simple notion of the proposed acoustic topic model, and the detailed descriptions are given below:

- Acoustic features: We used mel frequency cepstral coefficients (MFCC) to extract acoustic properties in a given audio signal. The MFCCs provide spectral information considering human auditory characteristics, and they have been

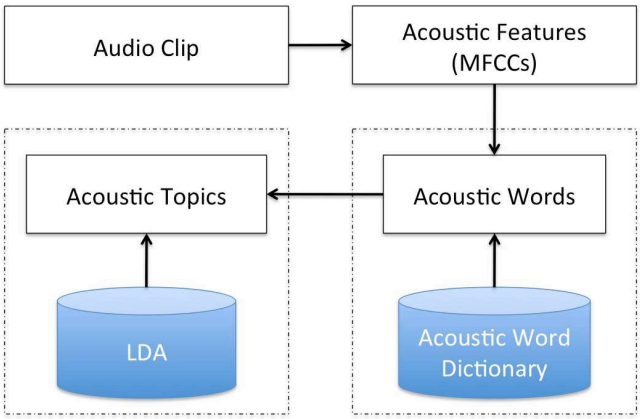



Figure 3: Diagram of the proposed acousic topic model algorithm

widely used in many sound related applications, such as speech recognition and audio classification. The reason we have chosen MFCC is to investigate the effects of spectral characteristics in topic model approach. In this work, we applied 20 ms hamming windows with 50% overlap to extract 12-dimensional feature vectors.

- Acoustic words: With a given set of acoustic features, we trained a dictionary using a vector quantization algorithm called LBG-VQ [14]. Similar ideas can be also found in [1, 10, 11]. Once the dictionary is built, the extracted acoustic feature vectors from sound clips can be mapped to acoustic words by choosing the closest word in the dictionary. In this work, we set the number of words in the dictionary to 1000, i.e. $V = 1000$. After extracting acoustic words, we generate a *word-document co-occurance matrix* which describes a histogram of acoustic words in individual audio clips. The word-document co-occurance matrix is fed in to the Latent Dirichlet Allociation (LDA) algorithm to model the acoustic topics.
- Acoustic topics: Each sound clip is assumed to be a mixture of acoustic topics. Since the acoustic topics are hidden variables, they are learned in an unsupervised manner (the number of latent topics are set manually). As described in the previous section, we use the *variational inference method* to estimate and infer the parameters of the topic model. We use the *posterior Dirichlet parameter* $\gamma$ of individual audio clips as the feature vectors of the corresponding sound clips.

For comparison, we use the Latent Perceptual Indexing (LPI) scheme proposed in our previous work [11] as a baseline. It is based on Singular Value Decomposition (SVD) that would reduce the feature vector dimension. The procedure is identical up to the step of generating the word-document co-occurance matrix. In the sense that the dimensions of feature vectors are reduced after the process, the topic model can be also considered as a feature dimension reduction. However, they differ in interpretations of feature vectors that represent the corresponding audio clip; the topic model uses statistical inference, while Latent Perceptual Indexing (LPI) extracts feature vectors deterministically using Singular Value Decomposition (SVD). The differences in experimental results will be described in the following section.

## 4. EXPERIMENTS

### 4.1. Database

We have collected 2,140 audio clips from the BBC Sound Effects Library [15], and annotated each file with both onomatopoeia and semantic labels. The semantic descriptions focus on what makes the sound, while the onomatopoeia descriptions focus on how people describe what they hear. These labels are annotated by following methods;

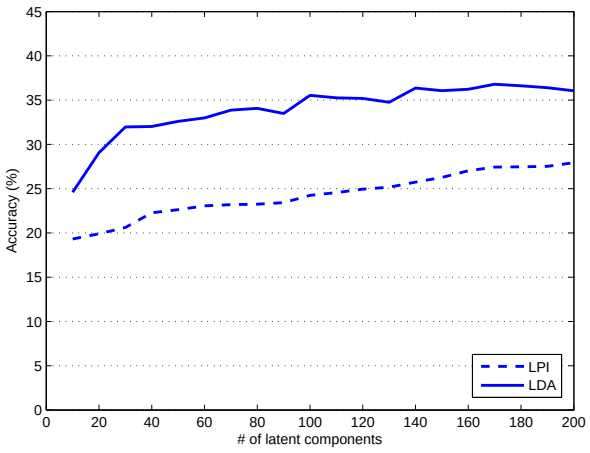
- Semantic labels: Based on the file name of the audio clip, it is labeled as one of predetermined 21 different categories. They include *transportation*, *military*, *ambiences*, *human*, and so on.
- Onomatopoeia labels: We performed subjective annotation to label individual audio clips. We asked subjects to label the corresponding audio clip among 39 onomatopoeia descriptions.

The audio clips are originally recorded with 44.1kHz (stereo) sampling rate, and down-sampled to 16kHz (mono) for acoustic feature extraction. See [10] for more details.

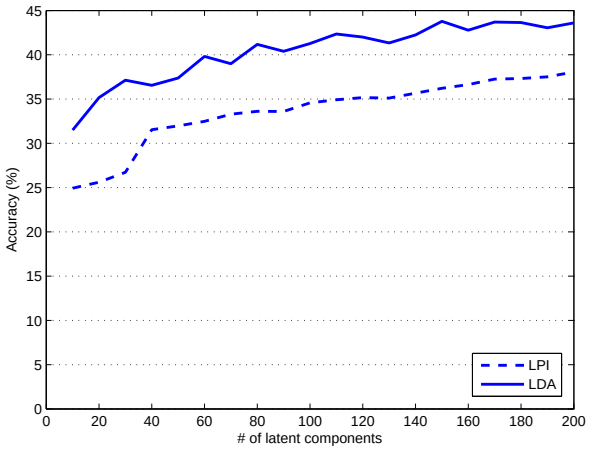
### 4.2. Results and discussion

Using the acoustic topic model, we can extract a single feature vector from an audio clip. The feature vector, i.e. a posterior Dirichlet parameter of the corresponding audio clip, represents the distribution over latent topics in the corresponding audio clip. With the feature vectors, we utilize a Support Vector Machine (SVM) with polynomial kernels as a machine learning algorithm for this application.

Fig. 4 shows the results of content-based audio description classification tasks using Latent Perceptual Indexing (LPI, dashed line) and Latent Dirichlet Allocation (LDA, solid line) . The performances are obtained by averaging five times of 10-fold cross




(a) Onomatopoeia labels




(b) Semantic labels

Figure 4: Classification results using using Latent Perceptual Indexing (LPI, dashed line) and Latent Dirichlet Allocation (LDA, solid line).

validation tasks, and depicted according to number of latent components on the figure. The number of latent components can be interpreted as the dimension of feature vector extracted from an audio clip. However, the interpretation differs in Latent Perceptual Indexing (LPI) and Latent Dirichlet Allocation (LDA). The number of latent components indicates a reduced rank after Singular Value Decomposition (SVD) in LPI, while it represents the number of hidden topics used in LDA. In our previous work, we paid less attention to the number of latent components. Instead, we had set a threshold that determines the number of retaining singular values: the largest singular values that contain greater than $90\%$ of the total variance [11]. In this work, however, the number of latent components becomes crucial since we are interested in the latent topics. Therefore, we evaluate the classification performance with respect to the number of latent components.

The results clearly show that the proposed acoustic topic model outperforms the conventional SVD-based latent analysis method in both onomatopoeia labels and semantic labels. This significant improvement is evident regardless of the number of latent components. We argue that this comes from utilizing LDA to analyze the hidden topics in audio clips. Note that the LPI analysis uses a deterministic approach based on SVD to map each word to the semantic space [12]. Although the semantic space is powerful to cluster the words that are highly related, the capability to predict the clusters from which the words are generated is somewhat limited in an euclidean space. With the proposed topic model, on the other hand, we are able to model the probabilities of acoustic topics that generate a specific acoustic word using a generative model.

In classifying onomatopoeia labels, the overall accuracy is lower than for the task of classifying semantic labels. This might be because the onomatopoeic words are for local sound contents rather than global sound contents; while the onomatopoeic words are local representations of sound clips, the topic model utilizes all the acoustic words from sound clips. Saliency detection algorithms might be necessary to improve the accuracy. It can be also observed that the accuracy increases as the number of latent components increase. This is reasonable in the sense of feature dimension reduction; a larger feature vector usually may capture more information. It should be noted, however, that there is a trade-off between accuracy and complexity. Increasing the feature vector size would also increase computing power requirements exponentially.

## 5. CONCLUSION

We proposed an acoustic topic model based on Latent Dirichlet Allocation (LDA) which learns hidden acoustic topics in a given audio signal in an unsupervised way. We adopted the variational inference method to train the topic model, and use the posterior Dirichlet parameters as a representative feature vector for an audio clip. Due to the rich acoustic information present in audio clips, they can be categorized based on two schemes: semantic and onomatopoeic categories which represent the cognition of the acoustic realization of a scene and onomatopoeic categories represent its perceptual experience, respectively. The results of classifying theses two descriptions showed that the proposed acoustic topic model significantly outperforms the conventional SVD-based latent structure analysis method.

Our future research will include investigating various ways of approximating Latent Dirichlet Allocation (LDA) such as Gibbs Sampling and different probabilistic latent analysis methods such as probabilistic Latent Semantic Analysis (pLSA) to model acoustic topics. Performance changes according to the number of acoustic words and the scalability of the proposed method will be also studied.

## 6. REFERENCES

[1] G. Chechik, E. Ie, M. Rehn, S. Bengio, and R. F. Lyon, "Large-scale content-based audio retrieval from text queries," in *ACM International Conference on Multimedia Information Retrieval (MIR)*, 2008.

[2] S. Chu, S. Narayanan, and C.-C. J. Kuo, "Environmental sound recognition with joint time- and frequency-domain audio features," *IEEE Transactions on Speech, Audio and Language Processing*, vol. in press, 2009.

[3] M. Slaney, "Semantic-audio retrieval," in *IEEE International Conference of Acoustics, Speech, and Signal Processing*, vol. 4, 2002, pp. 4108–4111.

[4] R. Parncut, *Harmony: A psychoacoustical approach*. Berlin: Springer-Verlag, 1989.

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

[6] D. M. Blei and J. D. Lafferty, "A correlated topic model of *science*," *The annals of Applied Statistics*, vol. 1, no. 1, pp. 17–35, 2007.

[7] K. Barnard, P. Duygulu, D. Forsyth, N. Freitas, D. M. Blei, and M. I. Jordan, "Matching words and pitctures," *Journal of Machine Learning Research*, vol. 3, pp. 1107–1135, 2003.

[8] D. M. Blei and M. I. Jordan, "Modeling annotated data," in *Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003, pp. 127–134.

[9] O. Yakhnenko and V. Honavar, "Multi-modal hierarchical dirichlet process model for predicting image annotation and image-object label correpsondence," in *the SIAM Conference on Data Mining*, 2009.

[10] 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.

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

[12] J. R. Bellegarda, "Latent semantic mapping," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 70–8–, 2005.

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

[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