Audio Scene Understanding using Topic Models

Samuel Kim
Signal Analysis and Interpretation Lab (SAIL)
University of Southern California
kimsamue@usc.edu

Shiva Sundaram
Deutsche Telekom Laboratories
TU-Berlin, Berlin, Germany
Shiva.Sundaram@telekom.de

Panayiotis Georgiou
Signal Analysis and Interpretation Lab (SAIL)
University of Southern California
georgiou@sipi.usc.edu

Shrikanth Narayanan
Signal Analysis and Interpretation Lab (SAIL)
University of Southern California
shri@sipi.usc.edu

Abstract

This paper introduces a method to apply the topic models in an audio scene understanding framework. Assuming that an audio signal consists of latent topics that generate acoustic words describing an audio scene, we propose to use a vector quantization method to build an *acoustic word* dictionary. The classification experiments with semantic labels yield promising results of using the topic models, compared to the conventional GMM-based approach, in audio scene understanding tasks.

1 Introduction

The topic model has been shown to be valuable in various text modeling applications. It assumes that a document consists of a set of hidden topics and each topic can be interpreted as a distribution over words in a dictionary. This assumption enables the use of a generative model like Latent Dirichlet allocation (LDA) which is a three-level hierarchical Bayesian model [1]. One of the motivations of modeling hidden topics in a text document is to handle the ambiguities of interpretations of words [2]; individual words may have variety of meanings, and the interpretations of the words vary according to the context around the word and the topic of the document. This idea has been successfully extended to various applications beyond text processing including image information retrieval [3].

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 sound or audio related applications (our previous work in [4] has demonstrated promising preliminary results). In this paper, we introduce a method to use the topic model for audio signal processing applications, specifically audio scene understanding tasks. This is motivated by drawing analogy between audio clips and text documents. As in interpreting words in text documents, there are ambiguities in the understanding of acoustic contents in audio signals; perceptively similar acoustic contents may cause different understanding of sound depending on the co-occuring acoustic sources. For example, an audio signal which includes an engine sound can be considered as recordings either from a factory, a construction site, or a car. This ambiguity can be reduced when other surrounding sounds are also perceivable. Suppose a sound of baby crying exists in the same audio clip, it is likely to be recored in a car rather than a construction site.

Using the topic model, in this paper, we explore audio scene understanding tasks which deal with highly unstructured audio signals. This particular application has been studied previously using various approaches [5, 6, 7]. In this work, we attempt to classify audio clips according to their semantic context using the proposed topic model method.

2 Proposed Method

To apply the topic model for audio context understanding tasks, we hypothesize that an audio clip consists of *latent acoustic topics* that generate *acoustic words*. We assume that the latent acoustic topics can be modeled using the topic model. Defining the acoustic words, however, is not trivial. Since the topic model is originally proposed for text document modeling applications, emulating it (at least initially) requires word-like discrete indices to utilize the topic model approach.

Fig. 1 illustrates the proposed method. In defining the acoustic words, various factors, such as the length of audio segments, types of acoustic feature, and total number of acoustic words, should be considered. For simplicity, we focus on the spectral characteristics of fixed length short-time segments. Specifically, we use the mel frequency cepstral coefficients (MFCC) which represent spectral attributes of audio segments based on characteristics of the early auditory system of humans [8]. We apply 20 ms hamming windows with 50% overlap to extract 12-dimensional feature vectors. After extracting MFCC feature vectors, each vector is assigned to an acoustic word based on the closest word in a pre-trained acoustic word dictionary built with a vector quantization algorithm called LBG-VQ [9]. With the acquired acoustic words, we utilize *variational inference method* [1] and *Gibbs sampling method* [2] to estimate and infer the parameters of the topic model.

Using the topic model, we can estimate *posterior Dirichlet parameter* of each audio clip and use it as a feature vector of the corresponding audio clip. The feature vector represents the distribution over latent topics in the corresponding audio clip. We utilize a Support Vector Machine (SVM) with polynomial kernels as a machine learning algorithm for this application.

3 Experiments

3.1 Database

We have collected 2,140 audio clips from the BBC Sound Effects Library [10]. The semantic labels of individual audio clips are categorized as one of 21 predetermined different classes such as for example *transportation*, *military*, *ambiences*, *human*, and so on. The audio clips are originally recorded with 44.1 kHz (stereo) sampling rate but down-sampled to 16 kHz (mono) to extract acoustic features. The results reported below are based on a 10 fold cross-validation.

3.2 Results

Fig 2 shows the accuracy of classifying the semantic labels associated with audio clips using LDA (solid lines) according to the number of latent topics and GMM-based approaches (dashed lines) proposed by Turnbull *et. al.* [7]. The GMM-based approaches are used as a representative of conventional content-based approaches.

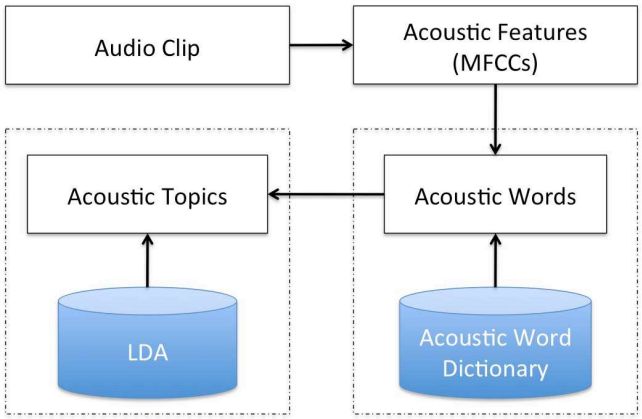

Figure 1: Using the topic model in audio scene understanding

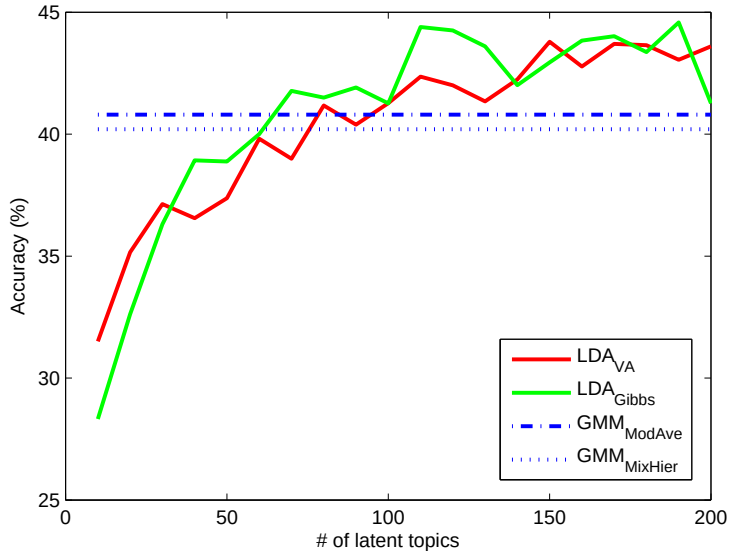

Figure 2: Classification results of semantic labels of audio clips using LDA (solid lines) according to the number of latent components and GMM-based approaches (dashed lines).

As mentioned earlier, we utilize both variational inference and Gibbs sampling methods for LDA approximation (red line for variational inference and green line for Gibbs sampling). The size of dictionary was set as 1,000 for this experiment, and the number of latent components can be interpreted as the dimension of feature vector extracted from an audio clip. For GMM-based approaches, we implemented two types of training methods which were proposed to deal with large database in [7] (dash-dotted line for ModAve and dotted line for MixHier)[1]. The accuracies using the GMM-based approaches are represented as horizontal lines because the number of latent topics cannot affect the performance.

The results show that the proposed method using the topic model outperforms the conventional GMM-based approaches when the proposed method consists of enough number of latent topics (no significant accuracy differences are found according to approximation methods in LDA or training GMM-based approaches). Note that the GMM-based approaches are powerful to cluster features in an unsupervised manner. 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.

4 Concluding Remarks

We introduced a method to apply the topic model for audio scene understanding tasks, specifically classification of audio clips by their semantic labels. To create word-like acoustic features, we utilized a vector quantization method with mel frequency cepstral coefficients (MFCC). The results suggest that the proposed method can outperform the conventional GMM-based approaches if it is equipped with enough number of latent topics.

We have limited the definition of acoustic word as the symbolized MFCC in this work. In the future, we will explore other alternative representations for acoustic words to model the context of audio signals. For examples, we will use dynamic length of audio segments to extract acoustic words instead of using fixed-length segments. We will also explore various topic model methods,

[1]Each audio clip is trained with 4 Gaussian mixtures, and the Gaussian mixture models for individual categories are obtained from the Gaussian mixtures of corresponding audio clips; the ModAve method keeps all the mixtures of individual audio clips, while the MixHier method merges the mixtures into the fixed number (16 mixtures in this work). See [7] for details about GMM-based approaches.

such as correspondence LDA (Corr-LDA) and supervised topic model (sLDA), to associate with descriptions of audio clips.

References

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

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

[3] C. Wang, D. M. Blei, and L. Fei-Fei, "Simultaneous image classification and annotation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

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

[5] 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. 17(6), pp. 1142–1158, 2009.

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

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

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

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

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