ICASSP 2010, Dallas TX

USING NAïVE TEXT QUERIES FOR ROBUST AUDIO INFORMATION RETRIEVAL

Samuel Kim, Panayiotis Georgiou, Shrikanth Narayanan

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

Shiva Sundaram

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

ABSTRACT

The goal of this work is to build an audio information retrieval system which provides users with flexibility in formulating their queries: from audio examples to naïve text. Specifically, the focus of this paper is on using naïve text to create input queries describing the desired information of the users. Using naïve text queries, however, raises interoperability issues between annotation and retrieval processes due to the wide variety of available audio descriptions. In this paper, we propose an intermediate audio description layer (iADL) to solve the interoperability issues between the annotation and retrieval processes. The iADL comprises two axes corresponding to semantic and onomatopoeic descriptions based on human-to-human communication experiments on how humans express sounds verbally. Various text modeling schemes, such as latent semantic analysis (LSA) and latent topic model, are utilized to transform the naïve text onto the proposd iADL.

***Index Terms*—** audio descriptions, naïve text query, audio information retrieval, out-of-vocabulary problem

1. INTRODUCTION

Audio information retrieval consists of two major steps: annotation and retrieval. The annotation process is used to describe the embedded information in given audio signals, while the retrieval process is used to extract an ordered list of audio clips that are relevant to users' input queries that describe their desired information. These two nearly-independent processes, however, are not necessarily interoperable because both are highly dependent on human descriptions that can span a wide range of possibilities; this is especially the case since to provide end use flexibility, users are typically not constrained as to their natural language query forms or structure. There is hence often a mismatch between descriptions of the desired information (query) and the embedded information which causes an *out-of-vocabulary* problem. This problem stems from users employing queries using descriptors with vocabulary not seen during the annotation process. The focus of this work is to bridge the gap between the annotation and retrieval processes by mitigating the mismatch between the two descriptors for robust query and retrieval.

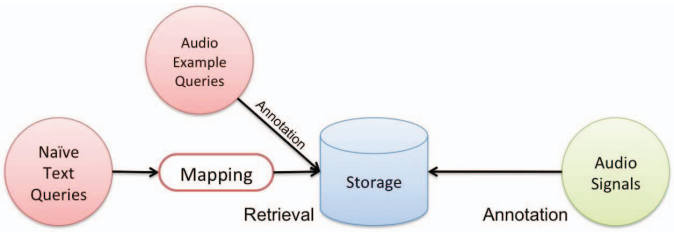


Fig. 1. A simple diagram of audio information retrieval system with various input queries: audio example and naïve text queries.

The output of the annotation process is usually stored in a database so that the retrieval process can access the storage instead of rerunning the annotation process for all the audio clips in a database. Although this strategy is reasonable for system efficiency, it becomes problematic when the systems face out-of-vocabulary queries. This issue was identified in [1] where the authors built an audio information retrieval system that deals with a large-scale database along with text queries. In their work, the authors observed that users tend to avoid "stating the obvious" when they are asked to annotate audio signals. They argued that this phenomenon might cause the out-of-vocabulary problem as this introduces noisy labeled data in the annotation phase. Although they pointed out this problem, their work focused on the annotation procedure and assumed that the problem is orthogonal to the annotation process. The out-of-vocabulary problem, therefore, is still left as an open challenge.

During the retrieval process, users should be able to describe their queries using various representations, such as sound categories, naïve text descriptions, and sound examples. In this work, we attempt to provide users with such a flexibility as illustrated in Fig. 1. In the case of using audio example queries, the system can annotate the input audio with the same annotation scheme as used for building the domain database and retrieve a list of audio clips related to the extracted information through existing database lookup techniques. In the case of naïve text queries, however, problems arise since the user expressions may be uncertain or ambiguous; naïve descriptions of users tend to be explanatory rather than categorical and are based on their own subjective annotation scheme. Furthermore, there are many types of polysemic and synonymic words in text descriptions which could further worsen the out-of-vocabulary problem. Similar issues exist in text information retrieval and spoken language understanding applications [2, 3, 4].

In this work, we propose an intermediate audio description layer (iADL) to mitigate the *out-of-vocabulary* problem caused by mismatches between descriptions in annotation and retrieval processes. It will provide interoperability between the annotation and retrieval processes by mapping descriptions from both processes onto the iADL. In this paper, we will focus on the retrieval process, specifically mapping the naïve descriptions of desired information to the proposed iADL. Various text modeling schemes, such as latent semantic analysis and latent topic modeling, are utilized to transform the naïve text queries onto the iADL.

The organization of this paper is as follows. In the next section, we provide the description of the proposed intermediate audio description layer (iADL). In Section 3, various text modeling schemes

to transform naïve text queries onto the proposed iADL are introduced. The experimental setup and results are discussed in Section 4 and Section 5 respectively, followed by the conclusions in Section 6.

2. INTERMEDIATE AUDIO DESCRIPTION LAYER

To mitigate the effects of mismatches in descriptions of embedded information and desired information, we introduce an intermediate audio description layer (iADL) which can carry abundant information about sounds in a set of pre-determinded categorical classes. This idea is motivated by previous research on human-to-human communication; Wake and Asahi found that people can successfully describe sounds with "sounding situation", "sound itself", and "sound impression" to other people [5]. These categories of descriptions are related to semantic information, onomatopoeias, and emotional information, respectively. Specifically, in this paper, we focus on two different aspects of audio data: semantic and onomatopoeic descriptions. These types of information 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 human sensory processing and semantic labels from the view point of cognitive processing [6]. We will address the emotional information in future work.

Fig. 2 illustrates the proposed iADL which consists of two axes: semantic words and onomatopoeias. As shown in the figure, the results of the annotation process are mapped to the two-dimensional layer. In other words, each audio clip is labeled with two different tags; one is based on onomatopoeia and the other is based on lexical (word) semantics. In the retrieval process, in turn, input queries are also transformed to the same layer. As a result, both the annotation and retrieval processes are represented in the same description layer that mitigates the possible mismatches.

As we mentioned earlier, we focus on the retrieval process assuming that we can utilize various available conventional methodologies for the annotation process to map audio signals to the iADL. This builds on our prior related work reported in [7] where the audio classification tasks were performed with onomatopoeias and semantic labels using the acoustic topic model.

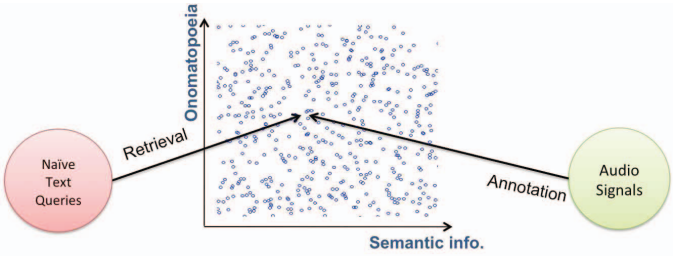


Fig. 2. Intermediate audio description layer (iADL) to mitigates possible mismatches between annotation and retrieval processes.

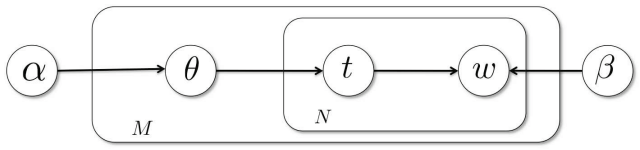


Fig. 3. Graphical representation of the topic model using Latent Dirichlet Allocation.

3. TEXT QUERY TRANSFORMATION

In this section, we apply various text document modeling schemes to transform naïve text queries to predetermined audio descriptions on the intermediate audio description layer, i.e. onomatopoeias and semantic labels. As we will show, this enables users to use their naïve text descriptions to retrieve sounds they want. Our work is based on the latent semantic analysis (LSA) [8] and latent topic model [9, 10] algorithms. The latent topic model represents the probabilistic word distribution over latent topics, while the LSA yields the association between words in a semantic space. The following provides an overview of these methods:

3.1. Latent Semantic Analysis

Let V be the number of words in the dictionary and M be the number of audio clips. Then, we can build a $V \times M$ word-description co-occurrence matrix $\mathbf{F}$ whose element f_{ij} represents the frequency of word i in description j. The word-description co-occurrence can be decomposed into three matrices using singular value decomposition (SVD), i.e.

$$\mathbf{F} = \mathbf{U} \cdot \mathbf{S} \cdot \mathbf{V}^T \tag{1}$$

where $\mathbf{U}$, $\mathbf{S}$ and $\mathbf{V}$ represent a matrix of word vectors, a diagonal matrix with singular values and a matrix with description vectors. This procedure allows LSA to capture the association between a set of descriptions and words in a semantic space. A reduced rank approximation can be obtained by retaining only R greatest singular values, i.e.

$$\mathbf{F} \simeq \mathbf{U}_{V\times R} \cdot \mathbf{S}_{R\times R} \cdot {\mathbf{V}_{M\times R}}^T \tag{2}$$

where $R < min(V, M)$. The value of R is determined experimentally. In this work, we use the matrix with description vectors, $\mathbf{V}_{M\times R}$, to represent descriptions of audio clips. Each description of audio clip, therefore, will be represented with a single R-dimensional feature vector.

3.2. Latent Topic Model

The latent topic model assumes that text documents consist of unobservable topics[1] and each topic can be interpreted as a distribution over words in a dictionary [10]. Using a generative modeling method like latent Dirichlet allocation (LDA) is appropriate under this assumption. Fig. 3 illustrates the basic idea of LDA as a three-level hierarchical Bayesian model.

Let V be the number of words in the dictionary and w be a V-dimensional vector whose elements are zero except for the corresponding word index in the dictionary. A description consists of N words, and is represented as $\mathbf{d} = \{w_1, w_2, \cdots, w_i, \cdots, w_N\}$ where w_i is the ith word in the description. A data set consists of M audio clips and is represented as $S = \{\mathbf{d_1}, \mathbf{d_2}, \cdots, \mathbf{d_M}\}$. Suppose that there are k latent topics and 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 $\theta \sim Dir(\alpha)$
   where θ and α are k-dimensional Dirichlet random variable and Dirchlet parameter, respectively.

[1]The word *topic* here is used loosely (as an alternative to a cluster) and may or may not represent what a human annotator would consider a topic.

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. $\beta_{mn} = p(w^n = 1|t^m = 1)$. The superscripts represent element indices of individual vectors, while the subscripts represent vector indices.

In this work, each text description will be represented with the posterior Dirichlet random variable θ which represents the topic distribution over k latent topics. However, it is not computationally feasible to learn the model directly from a given database, because it requires computing $p(\mathbf{w}|\alpha, \beta)$ which includes intractable integral operations. To solve this problem, various approaches such as Markov Chain Monte Carlo (MCMC) [11], gradient descent optimization method [12], and variational approximation [9], have been proposed. In this work, we use a *variational approximation method* [9] and a *Gibbs sampling method* [11], a specific form of MCMC, to estimate and infer the parameters of the topic model.

4. EXPERIMENTAL SETUP

4.1. Database

We have collected 2,140 audio clips from the BBC Sound Effects Library [13] and labeled each file with onomatopoeic labels, semantic labels, and short descriptions. The semantic labels and short descriptions are provided with the database. The semantic labels are given as one of predetermined 21 different categories. They include *transportation*, *military*, *ambiences*, *human*, and so on. The short descriptions consist of a set of words that represent the audio clip. The size of vocabulary is 2,820 and the average number of words in each description is 7.2 words after removing stop words and punctuational marks. For deriving the onomatopoeic words, we performed subjective annotation to label individual audio clips. We asked subjects to label the corresponding audio clip choosing from among 22 onomatopoeia descriptions. See [14] for more details about collecting onomatopoeic words. The audio clips are originally recorded with 44.1kHz (stereo) sampling rate. Table 1 shows examples from the BBC sound library along with various labels: semantic labels, onomatopoeic words, and short text descriptions. Both examples include the sound of a goat; while the subjective annotation of onomatopoeic words are the same, the semantic categories are different. These examples show the ambiguity of information even when they include the same audio contents. A summary of the database is given in Table 2.

4.2. Experimental setup

We design a text query transformation task that can be used in the retrieval process using naïve text queries. This experimental setting is realistic in practice because the annotation process and retrieval process are nearly-independent other than their joint dependence on the database being created or queried as illustrated in Fig. 1. Using various text document modeling methods, i.e. LSA and latent topic model, we can extract a single feature vector for a description of an audio clip. With the feature vectors, we utilize a multi-class support vector machine (SVM) with polynomial kernels as a machine learning algorithm for this application. Results are obtained by averaging 10-fold cross validation trials.

5. RESULTS AND DISCUSSION

Fig. 4 illustrates the results of classification tasks of text descriptions of audio signals using latent semantic analysis (LSA, dashed lines) and latent topic model with latent Dirichlet allocation (LDA, solid lines) according to the number of latent components used. The number of latent components can be interpreted as the dimension of feature vector extracted from each description of an audio clip. However, the interpretation differs according to the analysis methods. In LSA, the number of latent components indicates a reduced rank after singular value decomposition (SVD), while in LDA it represents the number of hidden topics used. Fig. 4 (a) and 4 (b) represent the results using onomatopoeic words and semantic labels, respectively. We utilize both variational inference and Gibbs sampling methods for LDA approximation (red line for variational approximation (VA) and green line for Gibbs sampling (Gibbs)).

The results clearly show that the SVD-based LSA method outperforms LDA method in classifying both onomatopoeic and semantic labels. These significant differences are evident regardless of the number of latent components and approximation methods for LDA.

The results may seem to contradict prior results in text document modeling that have shown the advantages of latent topic model over latent semantic analysis [9, 10]. We argue that this is because only small number of words are available for each description of an audio clip, while previous reports were performed on text documents which include many words for a single document. As given in Table 2, the average number of words in a description is 7.2 which might be too small to train the topic models which utilize a probabilistic approach; LSA utilizes a deterministic method.

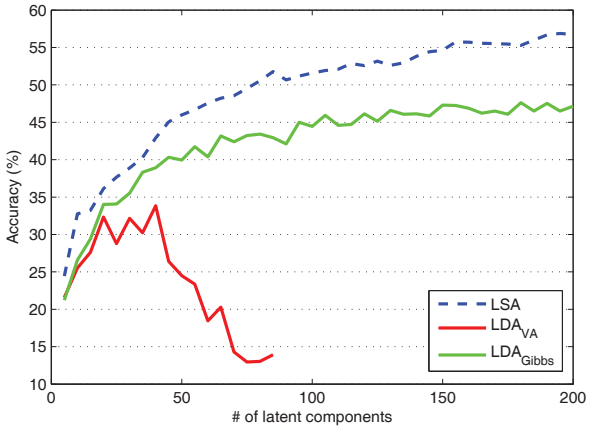
Furthermore, in the cases of the variational approximation scheme, the topic model method cannot be even trained as the number of latent topics increases. However, reasonable results are obtained with the Gibbs sampling scheme. We argue that this is be-

Table 1. Examples of BBC sound library along with its various descriptions.

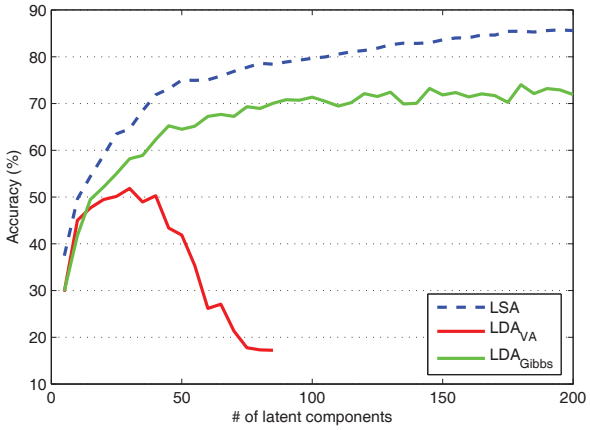
Ex 1.	Filename	1-GOAT-MACHINE-MILKED-BB
	Semantic label	MACHINERY/TOOLS
	Onomatopoeia	BLEATING
	Description	animals: goats one goat milked by machine other goats bleating occasionally - interior - abrupt end
Ex 2.	Filename	1-ENGLISH-GOAT-BLEATING-BB
	Semantic label	ANIMALS
	Onomatopoeia	BLEATING
	Description	animals: goats one old english goat bleating - occasional wind noise - interior

Table 2. Summary of BBC Sound Effect Library.

Number of sound clips	2,140
Number of semantic categories	21
Number of onomatopoeic words	22
Size of vocabulary for descriptions	2,820
Average number of words in a description	7.2
Average length of an audio clip	13 sec.



(a) Onomatopoeic words



(b) Semantic labels

Fig. 4. Classification results of text descriptions using Latent Semantic Analysis (LSA, dashed line) and Latent Dirichlet Allocation (LDA, solid lines) according to the number of latent components: (a) onomatopoeic words and (b) semantic labels.

cause of the characteristics of approximation methods; while the Gibbs sampling method uses an iterative process of sampling and updates, the variational approximation method requires sufficient number of training data to learn parameters via an expectation-maximization (EM) procedure.

In classifying onomatopoeia labels, the overall accuracy is lower than for the task of classifying semantic labels. This indicates that sound descriptions are highly related to the semantic labels rather than the onomatopoeias. It can be observed that the accuracy increases as the number of latent components increases. This is reasonable in the sense of feature dimension reduction; a larger feature vector usually may capture more discriminatory information. It should be noted that there is a trade-off between accuracy and complexity. Increasing the feature vector size would also increase computing requirements exponentially.

Although the proposed approach yields promising results, several open challenges are still remain. First, the ambiguity in the descriptions of sounds has not been completely solved yet; an audio clip can be labeled with multiple semantic labels and onomatopoeias, while we have considered only one semantic label and onomatopoeia for each audio clip. Therefore, in the future, we will extend our work so that an audio signal can have multiple tags. Real user annotations, which may include emotional information, will be included as well. Second, two axes in the proposed iADL are not necessarily orthogonal; there exist some degree of redundancy in representing a sound on the iADL, with which was not dealt in this work. In future work, we will investigate the dependency between the two axes of the iADL: semantic labels and onomatopoeias.

6. CONCLUDING REMARKS

Toward providing users with flexibility in their queries for audio information retrieval, we proposed an intermediate audio description layer (iADL) so that annotation and retrieval procedures are interoperable across a wide variety of user audio descriptions. We focused on transforming user generated naïve text queries onto the iADL by introducing various text modeling strategies, such as latent semantic analysis and latent topic model. The results of our classification tasks suggest that latent semantic analysis is more suitable than the latent topic model when only small number of words are available in each description.

7. 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] R. Jones, B. Rey, O. Madani, and W. Greiner, "Generating query substitutions," in *WWW '06: Proceedings of the 15th international conference on World Wide Web*. New York, NY, USA: ACM, 2006, pp. 387–396.

[3] J. Bai, J. Y. Nie, and G. Cao, "Context-dependent term relations for information retrieval," in *Conference of Emprical Methods in Natural Language Processing*, 2006, pp. 551–559.

[4] J. Bai, J. Y. Nie, G. Cao, and H. Bouchard, "Using query contexts in information retrieval," in *Proc. 30th Annu. Int. ACM SIGIR Conf. Reseach and Development in Information Retrieval*, 2007, pp. 15–22.

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

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

[7] 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.

[8] J. R. Bellegarda, "Latent semantic mapping," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 70–80, 2005.

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

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

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

[12] R. E. Madsen and D. Kuchak, "Modeling word burstiness using the dirichlet distribution," in *International Conference on Machine Learning*, 2005.

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

[14] 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.