An N-gram model for unstructured audio signals toward information retrieval

Samuel Kim, Shiva Sundaram†, Panayiotis Georgiou, and Shrikanth Narayanan

Signal Anlaysis and Interpretation Lab. (SAIL), University of Southern California, Los Angeles, USA.
†*Deutsche Telekom Laboratories, Quality and Usability Lab, TU-Berlin, Berlin, Germany.*
kimsamue@usc.edu

***Abstract*—An N-gram modeling approach for unstructured audio signals is introduced with applications to audio information retrieval. The proposed N-gram approach aims to capture local dynamic information in acoustic words within the acoustic topic model framework which assumes an audio signal consists of latent acoustic topics and each topic can be interpreted as a distribution over acoustic words. Experimental results on classifying audio clips from BBC Sound Effects Library according to both semantic and onomatopoeic labels indicate that the proposed N-gram approach performs better than using only a bag-of-words approach by providing complementary local dynamic information.**

I. Introduction

Extracting useful information from unstructured data is receiving significant attention from both research and industrial perspectives. Unstructured data can include various types of media: text, video, and audio that are typically user generated and poorly annotated in comparison to the rich information contained in them. In this work, we focus on unstructured audio data which can be present in various multimedia content, such as broadcasting [1], consumer videos [2], and personal sound logs [3]. The main challenge in retrieval of unstructured audio is that the context and the individual acoustic sources (for example human speech, laughter, or other environmental sounds such as car horns) are not known a-priori. To avoid making erroneously-biased a-priori assumptions, we index audio clips using a "bag-of-words" view of audio clips.

Researchers have been showing promising results by treating audio signals analogous to text documents [2], [4], [5], [6]. Many of them used mel-frequecy cepstral coefficients (MFCC) to capture acoustic properties and transformed the coefficients into discrete indices. Once the audio signals are represented with a sequence of discrete indices like text documents, many text modeling algorithms can be applied; Chechik *et al.* used the passive-aggressive model for image retrieval (PAMIR) [4]; Sundaram *et al.* introduced a latent perceptual indexing (LPI) method based on latent semantic analysis (LSA) [5], and Lee *et al.* applied probabilistic latent semantic analysis (pLSA) [2]. Recently, we have proposed the acoustic topic models using latent Dirichlet allocation (LDA) to characterize unstructured audio signals [6]. Assuming that there exist latent acoustic topics and each audio clip is a mixture of those latent topics, we could demonstrate promising results in audio classification tasks.

One of the drawbacks of these algorithms, including our acoustic topic modeling scheme, is that the bag-of-words approach which does not consider temporal dynamics of features. In this work, we introduce an N-gram approach to account for temporal dynamic information of audio features. The closest work to this idea has been done by Reed and Lee [7]. For music information retrieval applications, they proposed a new iterative segmentation method based on Viterbi decoding and Baum-Welch estimation. With the proposed segments, they apply the bi-gram approach to capture the temporal dynamic information in an LSA framework.

The organization of this paper is as follows. In the next section, we provide a brief review of acoustic topic models, the main framework of this work. In Section III, the proposed N-gram approach is described in detail. The experimental setup and results are discussed in Section IV and Section V, respectively, followed by the conclusions in Section VI.

II. Review - Acoustic Topic Model

The topic model was originally proposed for text processing applications, and it assumes that text documents consist of hidden topics and each topic can be interpreted as a distribution over words in a dictionary [8], [9]. This assumption enables the use of a generative model like Latent Dirichlet allocation (LDA). Our previous work had successfully adopted the topic model ideas into audio information retrieval applications by drawing analogies between audio signals and text documents [6]. In the following subsections, a brief review of the acoustic topic model is provided.

A. Latent topic 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.

Fig. 1(a) illustrates the basic concept of LDA in a graphical representation, a three-level hierarchical Bayesian model, and the generative process can be described as follows:

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

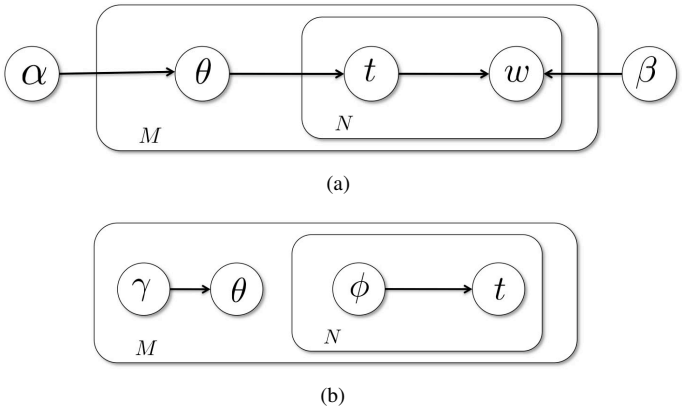


Fig. 1. Graphical representation of the topic model using Latent Dirichlet Allocation for (a) full representation and (b) approximated model for variational inference.

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. Variational Approximation Method

Now, the question is how to estimate or infer the latent parameters $\mathbf{t}$ and θ while the only variable we can observe is $\mathbf{w}$. With given values of α and β, the joint probability of θ and $\mathbf{t}$ can be estimated as

$$p(\theta, \mathbf{t}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{t}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)} \tag{1}$$

where

$$p(\theta, \mathbf{t}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{i=1}^{N} p(t_i|\theta) p(w_i|t_i, \beta) \tag{2}$$

and

$$p(\mathbf{w}|\alpha, \beta) = \int \sum_{\mathbf{t}} \left\{ p(\theta|\alpha) \prod_{i=1}^{N} p(t_i|\theta) p(w_i|t_i, \beta) \right\} d\theta \ . \tag{3}$$

These steps, however, are computationally impossible because computing $p(\mathbf{w}|\alpha, \beta)$ includes 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 introduced in [8]. 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 γ and ϕ which, resepectively, are the Dirichlet parameter that determines θ and the multinomial parameter that generates topics, as depicted in Fig. 1(b). The joint probability of θ and $\mathbf{t}$ can be simplified as

$$\begin{aligned} q(\theta, \mathbf{t}|\gamma, \phi) &= q(\theta|\gamma) q(\mathbf{t}|\phi) \\ &= q(\theta|\gamma) \prod_{i=1}^{N} q(t_i|\phi_i) \end{aligned} \tag{4}$$

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

$$\arg\min_{\gamma, \phi} D(q(\theta, \mathbf{t}|\gamma, \phi)||p(\theta, \mathbf{t}|\mathbf{w}, \alpha, \beta)) \ . \tag{5}$$

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.

$$\gamma_n = \alpha_n + \sum_{i=1}^{N} \phi_{in} \tag{6}$$

$$\phi_{in} \propto \beta_{n\tau} \exp\left(\Psi(\gamma_n) - \Psi\left(\sum_{j=1}^{k} \gamma_j \right) \right) \ . \tag{7}$$

Therefore, in the variational inference method, an iterative procedure of (7) and (6) alternatively is required until it converges.

III. N-GRAM APPROACH

A. Acoustic word

To make text-like audio signals, we have introduced 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 words 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 [10]. 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 [11]. Similar ideas to create acoustic words can also be found in [4], [6], [12]. The rationale is to cluster audio segments which have similar acoustic characteristics and to represent them as discrete code words (indexed appropriately).

B. Uni-gram approach

Once the dictionary is built, the extracted acoustic feature vectors from the test sound clips can be mapped to acoustic words by choosing the closest word in the dictionary so that individual short time segments have their assigned indices,

acoustic word. In this work, we call this method uni-gram approach to contrast with the proposed N-gram approach. After extracting uni-gram words, we generate a *word-document co-occurrence matrix* which describes a histogram of acoustic words in individual audio clips. The word-document co-occurrence matrix is used with the LDA to model audio clips as a distribution of latent acoustic topics. After the LDA modeling, we use the Dirichlet parameter γ as the representative feature vector of a single sound clip to be used consequent classifiers.

C. N-gram approach

To model the dynamic information embedded in acoustic words, we introduce the N-gram approach which describes partial dynamics of acoustic words by considering consecutive words. In this work, without the loss of conceptual generality, we consider only one adjacent word to make bi-grams, since $N = 2$ case is a good starting point to explore the usefulness of local context and computational complexity (since dictionary size increases exponentially).

A new acoustic dictionary $\widetilde{\mathcal{W}}$ can be built based on the bi-grams whose elements are from the original acoustic dictionary $\mathcal{W}$. The i-th word in the new dictionary $\widetilde{w_i}$ is defined as follows:

$$\widetilde{w_i} = \{(w_n, w_m) \,|w_n, w_m \in \mathcal{W}\} \tag{8}$$

where

$$\begin{aligned} n &= \lfloor i/V \rfloor \\ m &= \text{mod}(i/V) \quad , \end{aligned} \tag{9}$$

$\lfloor \cdot \rfloor$ and mod$(\cdot)$ represent the maximum integer that does not exceed the value of the division and the modulus of the division, respectively. Note that the size of the new dictionary is V^2.

Once the new dictionary for bi-grams is built, the extracted acoustic feature vectors from the test sound clips should be first mapped to acoustic words by choosing the closest word in the dictionary and then consider the adjacent acoustic words to generate the bi-grams. After extracting bi-gram words, we follow the same procedure as uni-gram approach; we generate a word-document co-occurrence matrix and feed into the LDA framework.

In this work, we set the number of words in the dictionary 200 for simplicity. Consequently, the size of bi-gram dictionary is 40,000.

D. Hybrid approach

We also introduce a way of combining both uni-gram and bi-gram approaches. In this work, we propose to make a super-vector of Dirichlet parameters after the LDA inference process; γ_{unigram} and γ_{bigram} from using uni-gram and bi-gram approaches, respectively. The dimension of the features which are fed into classifiers is, therefore, $2 \times k$ where k represents the number of latent acoustic topics.

IV. Experiments

A. Database

A selection of 2,140 audio clips from the BBC Sound Effects Library [13] 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 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 onomatopoeia descriptions. For more details on the annotation process, please refer to [5]. 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.

B. Audio classification

The average classification performance using support vector machines (SVM) in terms of onomatopoeic and semantic labels was estimated by 5-fold cross-validation. In each fold, the LDA and the classifier parameters were estimated from the train set and tested on the corresponding test set.

To evaluate the performance of the proposed framework, we use the *F-measure* which is widely used for evaluating information retrieval systems. The metric considers both *precision* and *recall* and can be written as

$$F = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{10}$$

and is evaluated for individual descriptive categories, i.e., onomatopoeic and semantic labels.

V. Results and Discussion

Fig. 2 shows the performance of audio classification tasks for both onomatopoeic labels and semantic labels. These two types of labels are chosen based on our previous work in [14] where the intermediate audio descriptive layer (iADL) was proposed to provide interoperability between the annotation and retrieval processes in an audio retrieval framework.

The 5-fold cross validation performance is shown as a function of number of latent components on the figure. As shown in the figure, the accuracy increases as the number of latent acoustic topics increases across various types of

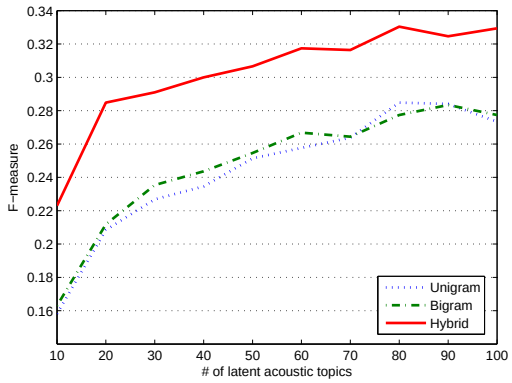
TABLE I
Summary of BBC Sound Effects Library.

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

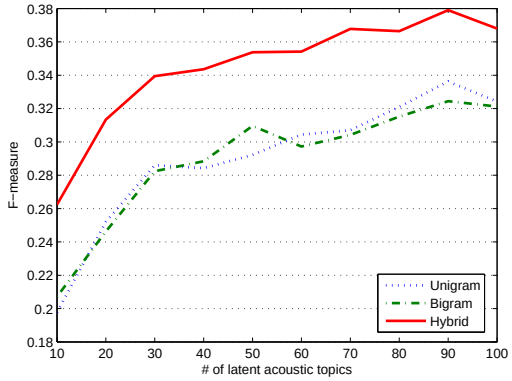
experimental settings. It is consistent with our previous work reported in [6] where we argued that this trend is reasonable in the sense of feature dimension reduction.

The direct comparison of performance with respect to the same number of latent acoustic topics is fair in the sense that the feature dimensions are the same which are fed into the classifier. In that sense, there is no significant performance differences by using the bi-gram modeling approach (dash-dot lines) compared to the uni-gram approach (dashed lines). However, the direct comparison may not be fair if we consider the latent Dirichlet allocation algorithm as a dimension reduction process. For example, in the case that the number of latent acoustic topics is 100 (the feature dimension of audio clips is 100), the system only uses 0.25% of original feature vector dimension in bi-gram cases while it uses 50% in unigram cases. It is also related to the sparseness of data; for bi-gram modeling, 40,000 acoustic words are used to represent audio signals while 200 acoustic words are used for the uni-gram approach.

The solid lines show the performance using the hybrid method which makes super-vectors of feature vectors from uni-gram and bi-gram approaches. For simplicity, we use the feature vectors which are extracted using the same number of latent acoustic topics. Since we concatenate two feature vectors to make a super-vector, the dimension of a super-vector is greater than the one of original feature vectors (twice greater in this experiment). The results clearly shows the significant performance improvement by using the hybrid method which indicate that the uni-gram and bi-gram approaches represent complementary information.



(a) Onomatopoeic words



(b) Semantic labels

Fig. 2. Classification results of audio clips using unigram and bigram acoustic words in the acoustic topic model framework according to the number of latent acoustic topics: (a) onomatopoeic words and (b) semantic labels.

VI. Concluding Remarks

We proposed the N-gram approach to model dynamic information within the text-like audio modeling scenario for information retrieval applications. Specifically, we have used the bi-gram model to consider adjacent acoustic words and built a new acoustic word dictionary for the bi-grams. Experimental results showed that the proposed N-gram approach brought significant improvements in the performance by providing complementary local dynamic information.

In the future, we will explore various types of dynamic modeling method to bring forth the dynamic information to the bag-of-words strategies.

Acknowledgement

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

References

[1] G. Friedland, L. Gottlieb, and A. Janin, "Joke-o-mat: Browsing sitcoms punchline by punchline," in *Proceedings of ACM Multimedia*, 2009.

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

[3] D. Ellis and K. Lee, "Minimal-impact audio-based personal archives," in *ACM workshop on Continuous Archiving and Recording of Personal Experiences CARPE-04*, 2004.

[4] 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.

[5] 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.

[6] 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.

[7] J. Reed and C.-H. Lee, "On the importance of modeling temporal information in music tag annotation," in *IEEE International Conference of Acoustics, Speech, and Signal Processing*, 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] L. R. Rabiner and B. Juang, *Fundamentals of speech recognition*. Prentice Hall, 1993.

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

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

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

[14] 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.