SIP (2012), vol. 1, e6, page 1 of 15 © The Authors, 2012.

doi:10.1017/ATSIP.2012.7

ORIGINAL ARTICLE

Latent acoustic topic models for unstructured audio classification

SAMUEL KIM, PANAYIOTIS GEORGIOU, AND SHRIKANTH NARAYANAN

We propose the notion of latent acoustic topics to capture contextual information embedded within a collection of audio signals. The central idea is to learn a probability distribution over a set of latent topics of a given audio clip in an unsupervised manner, assuming that there exist latent acoustic topics and each audio clip can be described in terms of those latent acoustic topics. In this regard, we use the latent Dirichlet allocation (LDA) to implement the acoustic topic models over elemental acoustic units, referred as acoustic words, and perform text-like audio signal processing. Experiments on audio tag classification with the BBC sound effects library demonstrate the usefulness of the proposed latent audio context modeling schemes. In particular, the proposed method is shown to be superior to other latent structure analysis methods, such as latent semantic analysis and probabilistic latent semantic analysis. We also demonstrate that topic models can be used as complementary features to content-based features and offer about 9% relative improvement in audio classification when combined with the traditional Gaussian mixture model (GMM)–Support Vector Machine (SVM) technique.

Keywords: Unstructured Audio, Latent topic models, Audio information retrieval, text-like audio signal processing, Acoustic topic models

Received 3 February 2012; Revised 31 October 2012

I. INTRODUCTION

A perennial challenge in designing a content-based audio information retrieval system is linking the audio signals to linguistic descriptions of audio that are generated, and utilized, by end users. While methodologies to extract acoustic features from audio signals according to pre-defined descriptive categories have been studied intensely, various open challenges remain.

The challenges are often related to *sound ambiguity*. A key source of ambiguity arises from the potential heterogeneous nature of audio. A generic audio signal typically can contain a mixture of several sound sources; each sound source carries its own information in the mixture (e.g., an occasional phone ringing during an office chat or a distant siren in a cafe recording). This heterogeneity leads to the importance of context in the interpretation of sounds. The context dependency underscores the fact that perceptively similar acoustic content may lead to different semantic interpretation depending on the evidence provided by the co-occurring acoustic sources. For example, an engine sound can be interpreted as recorded either from a factory or a car (and can be labeled *machinery* or *automobiles* with the category labels used in the BBC sound effects library [1] considered in this work). However, when this sound co-occurs with a baby's crying in the same audio clip, the audio clip is more likely to be recorded in a car rather than a factory.

Considering the wide variation in the characteristics of generic audio, a rich variety of embedded contexts can be expected since each type of audio signal may have risen from a different generative process. In specific cases, knowledge about the generation can be advantageously used in their modeling. Music audio signals, for example, result from well-structured production rules that can be represented as a musical score, while speech audio signals are governed by a linguistic structure that defines their production. Our focus in this paper, however, is on generic unstructured audio signals whose generation rules are non-evident or hidden. Performing information retrieval from unstructured audio signals is a well-known (and an increasingly important) application and several promising algorithms have been proposed. For example, Slaney presented a framework to derive semantic descriptions of audio from signal features [2]. Turnbull *et al.* [3] applied their supervised multi-class labeling method, originally devised for music information retrieval, to a sound effects database. In research from Google, Chechik *et al.* [4] successfully performed a large-scale content-based audio retrieval from text queries for audio clips with multiple tags. Their method, based on a passive-aggressive model for image retrieval, is scalable to a very large number of audio data sources.

3710 S. McClintock Ave, RTH 320, Los Angeles, CA 90089, U.S.A

Corresponding author: Samuel Kim
E-mail: worshipersam@gmail.com

Applications like environment sound recognition which aim to characterize ambient sound conditions have also been investigated [5]. The modeling contributions proposed in the present work also aim at unstructured audio by focusing on the underlying contextual structure of the audio and without the need for explicit tagging.

To provide a useful representation of context-dependent information in an unstructured audio signal, we propose the notion of latent acoustic topic models that can be learned directly from co-occurring audio content in audio signals. The latent topic model was originally proposed for text-based information retrieval systems to tackle similar context dependency (e.g., the word 'bank' can be interpreted differently depending on context the word is used: related to a river or a financial institution) [6–8]. Drawing analogies between text documents and audio clips, we hypothesize that each audio clip consists of a number of latent acoustic topics and these latent acoustic topics, in turn, generate the acoustic segments that constitute the audio clip. In other words, assuming appropriately defined units of audio signals can play a similar role as words in text documents and we suppose that there exist latent acoustic topics in audio signals that can be mapped to latent topics in text.

Such ideas from text processing have also been successfully extended to content-based image retrieval applications [9–12]. Topic models have been used in image processing with the assumption that there exist hidden topics that generate image features. The image features are often quantized to provide discrete index numbers to resemble the linguistic words in the text topic modeling. A number of techniques have been gainfully used for audio processing problems as well. Smaragdis *et al.* [13, 14] introduced various audio applications using the topic models, such as source separation and music transcription. Sundaram and Narayan [15, 16] used the latent perceptual indexing (LPI) method for classifying audio descriptions inspired by the latent semantic analysis (LSA). Lee and Ellis [17] used the probabilistic latent semantic analysis (pLSA) in consumer video classification with a set of semantic concepts. Only with audio information from video clips, they decomposed the Gaussian mixture model (GMM) histograms of feature vectors using pLSA to remove redundant structure and demonstrated promising performance in classifying video clips. Levy and Sandler [18] used an aspect model, which is based on the pLSA, on music information retrieval. To build the aspect model, they proposed *muswords* extracted from music audio signals and words from social tags. Hu and Saul [19] used the Latent Dirichlet Allocation (LDA) method in a musical key-profiling application.

The first contribution of this paper is the introduction of a generative model to capture contextual information in generic audio signals, a model that is distinct from the well-known content-based methods which are based on modeling realizations of sound sources. These two approaches differ in the sense that the context-based approach seeks to model the latent generative rules that generate observations in audio content based on co-occurring acoustic properties. We also describe an approach to process audio signals in a text-like manner and how to interpret the topics in audio signals. The benefits of the proposed modeling of unstructured audio are evaluated on an audio tag classification task. Our goal is to consider the context dependency in classifying the audio tags using the latent acoustic topic models that are directly obtained from the audio signals. The basic idea is to model the distributions over latent acoustic topics with supervised classifiers assuming that audio signals in the same category (e.g., annotations, labels, and tags) would have similar latent acoustic topic distributions. We also report experiments using a hybrid method that utilizes the proposed acoustic topic model as complementary features to conventional content-based methods.

The paper is organized as follows. A description of the proposed latent acoustic topic model along with the background, implementation, and interpretation is provided in Section II. The audio tag classification task, which is the experimental framework of this paper, is described in Section III. The experimental setup and results are described in Section IV followed by conclusions in Section V.

II. LATENT ACOUSTIC TOPIC MODEL

In this section, we describe the proposed acoustic topic model and its realization in detail. First we provide a brief overview of the LDA method that is popularly used to implement latent topic models.

A) LDA

As discussed earlier, the latent topic model that was originally proposed for text signal processing assumes that a document consists of latent topics and each topic has a distribution over words in a dictionary [7]. This idea can be realized using a generative model such as LDA. Figure 1 illustrates the basic concept of LDA in a graphical representation as a three-level hierarchical Bayesian model.

Let V be the number of words in a dictionary and w be a V-dimensional vector whose elements are zero except for the corresponding word index in the dictionary. Assume that a document consists of N words, and be represented as $\mathbf{d} = \{w_1, w_2, \ldots, w_i, \ldots, w_N\}$, where w_i is the ith word in the document. Let the dataset consist of M documents and be represented as $S = \{\mathbf{d_1}, \mathbf{d_2}, \ldots, \mathbf{d_M}\}$.

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 $\mathbf{d}$ in dataset 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 $t_i \sim Multi(\theta)$, where t_i is the topic that corresponds with the word w_i and $Multi(\cdot)$ represents a multinomial distribution.

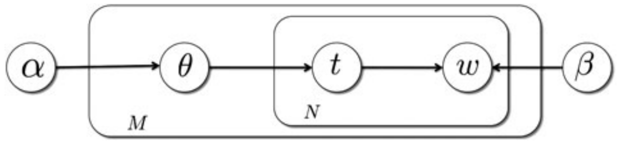


Fig. 1. Graphical representation of the topic model using LDA.

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., $\beta_{nm} = p(w_i = m|t_i = n)$.

It is apparent from the above that LDA assumes a large number of hidden or latent parameters (θ, $\mathbf{t}$, α, and β) and only one observable variable $\mathbf{w}$. In many estimation problems, parameters are often chosen to maximize the likelihood values of a given data $\mathbf{w}$. The likelihood can be defined as

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

Once α and β are estimated, the joint probability of θ and $\mathbf{t}$ with given $\mathbf{w}$ should be estimated as

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

These steps, however, are not computationally feasible because both inference and estimation require computing $p(\mathbf{w}|\alpha, \beta)$, which includes intractable integral operations as follows:

$$p(\mathbf{w}|\alpha, \beta) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \int \prod_{n=1}^{k} (\theta_n)^{\alpha_n - 1} \times \prod_{i=1}^{N} \sum_{n=1}^{k} \prod_{m=1}^{V} (\theta_n \beta_{nm})^{w_{im}} d\theta. \qquad (3)$$

To solve this problem, various approaches such as Markov Chain Monte Carlo (MCMC) [8], the gradient descent optimization method [20], and variational approximation [6] have been proposed. In this work, we use the variational approximation method. While the Gibbs sampling method is based on MCMC, which is an iterative process of obtaining samples by allowing a Markov chain to converge to the target distribution [8, 21], the rationale behind the variational approximation method is to minimize distance between the real distribution and the simplified distribution using Jensen's inequality [6, 22]. The simplified version consists of the Dirichlet parameter that determines θ and the multinomial parameter that generates topics, respectively (see [6] for more details).

B) Realization

To apply and extend the notion of latent topics to the proposed latent acoustic topics, here we introduce a couple of new ideas. Toward that, we need to define what documents and words are in the acoustic domain. Similar to the applications of latent topic models for text documents and images, we introduce the notion of acoustic words that play a similar role as words in text documents. An audio clip, intuitively, can be defined as an acoustic document that represents a sequence of acoustic words. With the extracted acoustic words, LDA can be used to model hidden acoustic topics. Figure 2 summarizes the proposed acoustic topic modeling procedure, and further details are given below.

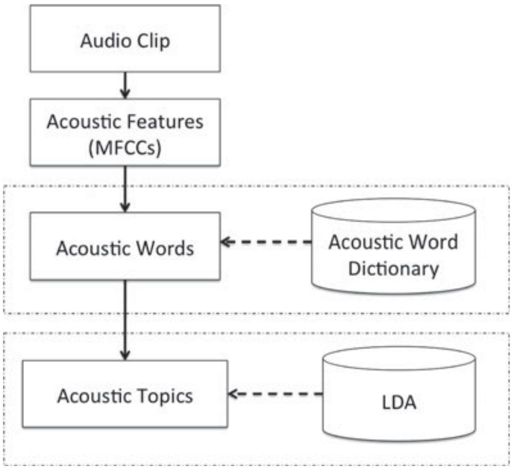


Fig. 2. Diagram of the proposed acoustic topic modeling procedure for unstructured audio signals.

To define acoustic words, one may come up with various methodologies to transform an audio signal to a sequence of word-like units that represent specific characteristics of the audio signal. There are several critical questions that arise in that regard. These include how to segment the audio, what to extract, how to discretize, etc. In this paper, for simplicity, we adopt conventional Mel Frequency Cepstral Coefficients (MFCCs) to parameterize the audio signal and use vector quantization (VQ) to derive the acoustic words. Note that, however, the VQ method might introduce quantization errors. Watanabe *et al.* [23] have introduced a probabilistic method instead.

Using fixed length frame-based analysis, we calculate MFCCs to represent the audio signal's time varying acoustic properties. The MFCCs provide spectral parameterization of the audio signal considering human auditory properties and have been widely used in many sound-related applications, such as speech recognition and audio classification [24]. In this work, we use 20 ms hamming windows with 50% overlap to extract 12-dimensional feature vectors.

With a given set of acoustic features, we derive an acoustic dictionary of codewords using the *Linde-Buzo-Gray Vector Quantization* (LBG-VQ) algorithm [25]. Similar ideas to create acoustic words can also be found in [4, 15, 26, 27]. The rationale is to cluster audio segments that have similar acoustic characteristics and to represent them as discrete indexing numbers. In this work, we empirically set the number of words in the dictionary, i.e., vocabulary size V. Specifically, we consider one of the values from $V \in \{200, 500, 1000, 2000, 4000\}$ for the number of words in the dictionary. Once the dictionary is built, the extracted

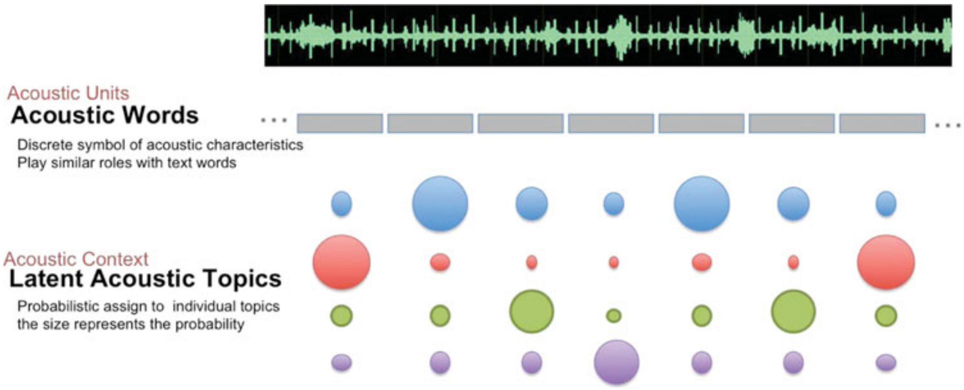

Acoustic Units
Acoustic Words
Discrete symbol of acoustic characteristics
Play similar roles with text words
Acoustic Context
Latent Acoustic Topics
Probabilistic assign to individual topics
the size represents the probability

Fig. 3. An example of interpretation of the acoustic topic models as a type of probabilistic clustering.

acoustic feature vectors from sound clips can be mapped to acoustic words by choosing the closest word in the dictionary. Further processing, such as n-gram modeling [28] and stopword detection [29], can be also used to deal with the text-like audio signals.

C) Interpretation of acoustic topic modeling

One of the ways to interpret the topic modeling procedure is as a dimension reduction process [30]. Instead of using bag-of-words approach with high dimensional yet sparse feature vectors, one can extract k-dimensional feature vectors, i.e., the topic distributions, from V-dimensional feature vectors, i.e., the word counts in documents [31].

On the other hand, the topics can be interpreted as clusters. Figure 3 visualizes an example of the acoustic topic model as a probabilistic clustering (also known as soft clustering). A given audio signal can be transformed into a sequence of acoustic words (depicted as gray boxes), and each acoustic word can be generated by different clusters (depicted as circles; different colors represent different clusters and size of the circles represent the probability of being generated by the corresponding cluster). Again, these clustering results are not based on geometrical similarities but based on contextual modeling using co-occurrence information.

Figure 4 illustrates an example of acoustic topic modeling results (where the number of acoustic words is 1000 and the number of latent topics is 100; we use a sound clip from the BBC sound effects library whose filename is 1-GOAT-MACHINE-MILKED-BB.wav). Figure 4(a) shows the topic distribution in the given audio clip, while Figure 4(b)–4(f) represent the five most probable acoustic words with their probabilities in the five most probable topics (the acoustic words were depicted as the centroids of MFCC codebooks and the probabilities of words are denoted in the legend). In Fig. 4(a), there are only a few topics strongly present among the 100 latent acoustic topics.

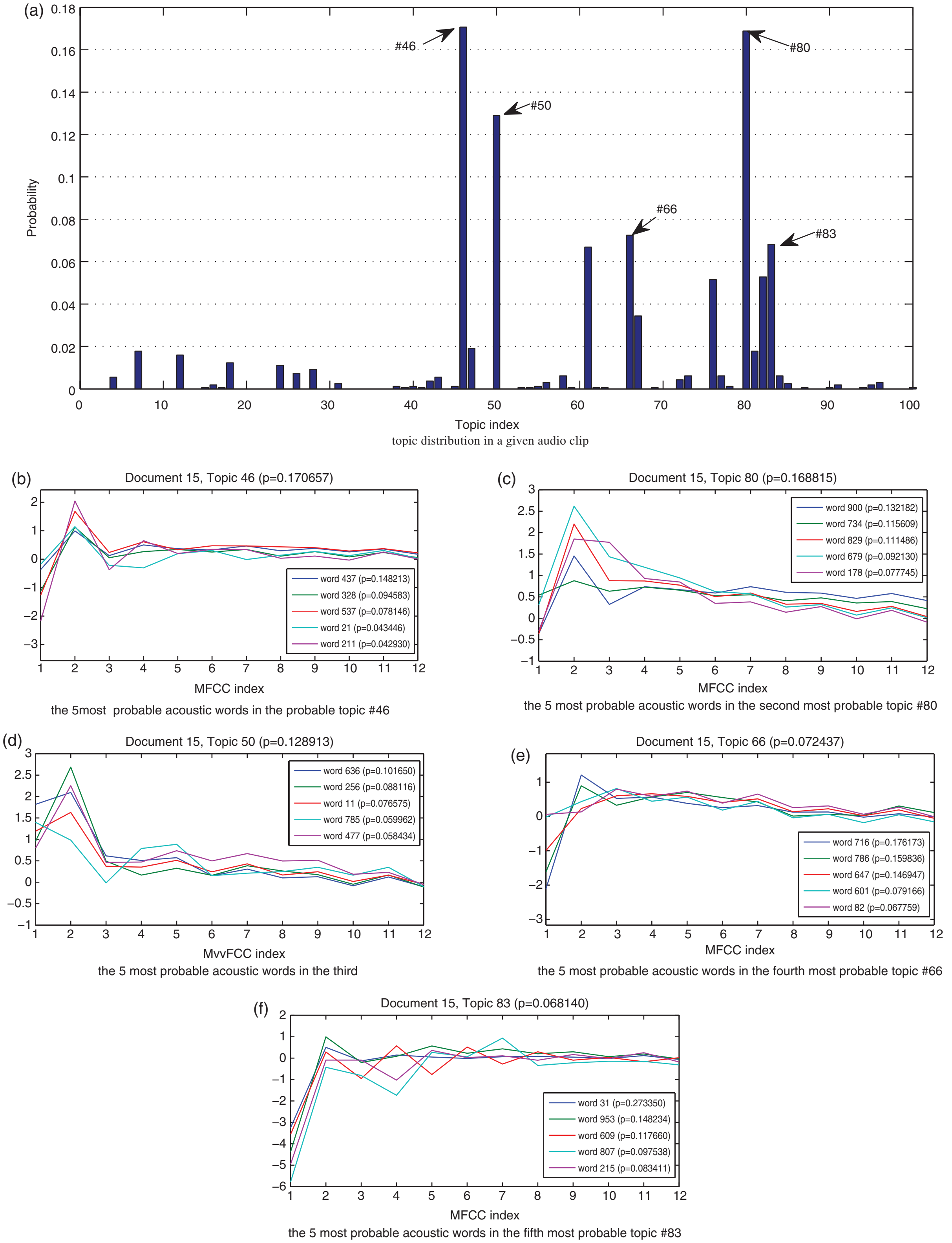
As illustrated in Figs. 4(b)–(f), each topic has a probability distribution over acoustic words (12-dimensional MFCC). Each topic can be interpreted as a cluster of acoustic features in terms of their co-occurrence probabilities instead of geometrical similarity measurements such as Euclidean distance or Mahalanobis distance. From the figure, it is remarkable that the highly probable acoustic features in an acoustic topic seem geometrically similar. This implies that the acoustic features that often co-occur may be close in the geometrical sense as well, although it is not guaranteed for the reverse to be true.

Note that these latent topics do not directly correspond to any semantic interpretations. Having said that the proposed acoustic topic models are learned in an unsupervised way without any class information, and the topics simply represent the clusters of acoustic words as we discussed above. As an alternative, one can think of using supervised LDA (sLDA), which includes class information within the latent topic modeling framework [32]. Indeed, in our previous work [33, 34], we had shown that using sLDA can help to cluster audio features into a sparse topic space and consequently improve the overall classification performance. Even in the sLDA framework, however, the topics are not directly mapped to class information. Therefore, in this work, we employ a two-step strategy to study the relationship between the topic distribution and semantic interpretations.

III. AUDIO TAG CLASSIFICATION

We now evaluate the potential of acoustic topic models in the context of audio tag classification, a task that exemplifies automatic audio annotation and example-based retrieval. The tags considered are semantic and onomatopoeic categories. The rationale behind this choice is that these two descriptive categories can provide an intermediate layer on which users' naïve text query can be mapped to prevent out-of-vocabulary problems [35]. In this work, the classification performance of audio clips for semantic and onomatopoeic labels are individually demonstrated.

As shown in Fig. 5, we adopt a two-step learning strategy for the audio tag classification tasks: an unsupervised acoustic modeling step and a supervised classifier step. For the unsupervised modeling step, the proposed acoustic topic model is used. This unsupervised modeling step can be also considered as a feature extraction step in the sense that its output can be fed into a classifier subsequently.



topic distribution in a given audio clip

the 5most probable acoustic words in the probable topic #46

the 5 most probable acoustic words in the second most probable topic #80

the 5 most probable acoustic words in the third

the 5 most probable acoustic words in the fourth most probable topic #66

the 5 most probable acoustic words in the fifth most probable topic #83

Fig. 4. Illustrative examples of acoustic topic modeling: (a) topic distribution in a given audio clip, (b) the 5 most probable acoustic words in the most probable topic #46, (c) the 5 most probable acoustic words in the second most probable topic #80, (d) the 5 most probable acoustic words in the third most probable topic #50, (e) the 5 most probable acoustic words in the fourth most probable topic #66, and (f) the 5 most probable acoustic words in the fifth most probable topic #83 (the number of acoustic words is 1000 and the number of latent topics is 100).

Specifically, we use the posterior Dirichlet parameter that represents the probability distribution over latent acoustic topics as the feature vector of the corresponding audio clip, assuming that audio signals under the same category would have similar latent acoustic topic distributions. For the supervised classifier step, we utilize a Support

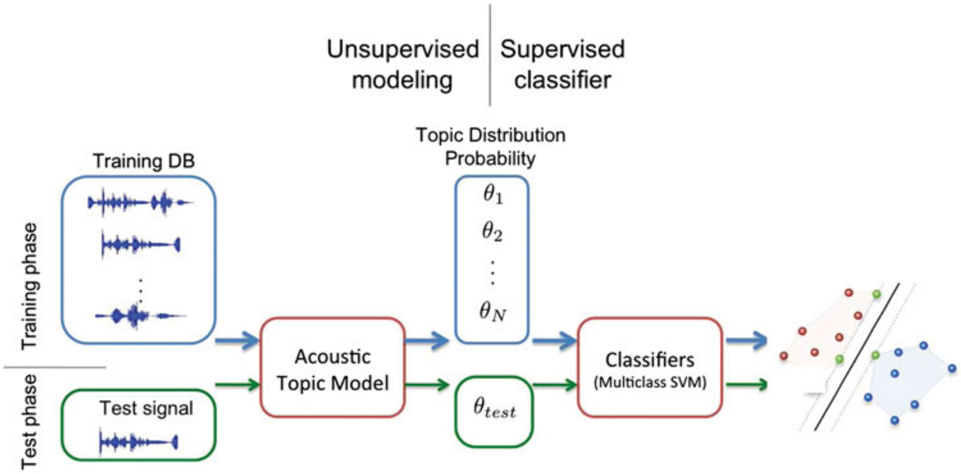


Fig. 5. A simple diagram of two-step learning strategy for audio tag classification task.

Vector Machine (SVM) with Bhattacharyya kernel [36] as the machine-learning algorithm. Since the SVM was originally designed for binary classification tasks, we use a one-against-one structure for a multi-class classifier which trains $C(C-1)/2$ binary classifiers, where C represents the number of classes.

For comparison purposes, we consider other latent variable approaches, i.e., LSA [15] and pLSA [17], as baseline systems. The fundamental difference between LSA, pLSA, and LDA is in the way of inferring the topic distribution. LSA estimates the topic distribution deterministically using Singular Value Decomposition (SVD) [37], whereas pLSA and LDA use statistical inference. On the other hand, LDA differs from pLSA in that it includes Dirichlet prior to the distribution [6, 38].

We also consider a GMM-based classifier to represent content-based classification tasks. Particularly, we use the GMM–SVM framework that is widely used in many pattern recognition tasks such as speaker verification [39]. Similar to the proposed acoustic topic model, it utilizes a two-step learning strategy that learns feature distributions with GMM in the first step and uses mean supervectors as features for the consequent SVM classifier in the second step. We use this methodology because it is convenient to compare performance and to examine complementary information since they share a similar two-step structure, whereas our preliminary experiments show comparable performance with other GMM-based methodologies.

We perform a 5-fold cross-validation by randomly partitioning the database into five equal-size but exclusive 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, and training SVM classifiers are done using the training subsets. The overall performance is obtained by the held out test subsets. Other baseline systems, i.e., LSA and GMM-SVM, are handled in the same way. To evaluate the performance of the proposed framework, we use the *F-measure* that is widely used for evaluating information retrieval systems [40]. The metric considers both *precision* and *recall*, and can be written as

$$F = 2\frac{precision \times recall}{precision + recall},$$

where

$$precision = \frac{\text{number of correctly classified trials in class C}}{\text{total number of test trials classified as class C}},$$

$$recall = \frac{\text{number of correctly classified trials in class C}}{\text{total number of test trials from class C}}.$$

Since there are multiple classes, we calculate the *F-measure* values separately over the different classes. The overall *F-measure* value can be computed as a weighted average of individual *F-measures* by the number of trials in corresponding classes.

IV. EXPERIMENTS AND DISCUSSION

A) Database

A selection of 2140 audio clips from the BBC Sound Effects Library [1] 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*, *ambience*, and *human*. Each linguistic description consists of, on average, 7.2 words after removing stop words and punctuation marks. 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 [15]. It is notable that the overlap between different sound sources within a clip is, if any, rarely present in the database. This property enables us not to wrestle with sound source separation problems that are very challenging as well.

Table 1. Summary of BBC sound effect library.

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

The audio clips are available in two-channel format with 44.1 kHz sampling rate and are down-sampled to 16 kHz (mono) for acoustic feature extraction. The average audio clip length is about 13 seconds and generates about 1300 acoustic words. A summary of the database is given in Table 1. Table 2 shows the distribution of onomatopoeic words and semantic labels for the database. For example, there are 349 audio clips whose semantic labels are 'animals'. In the category of 'animals', there exist various onomatopoeic words to represent the audio clips (e.g., 62 clips for 'growl' and 60 clips 'meow').

B) Audio tag classification results

There are two parameters that can be empirically tuned for obtaining a reasonable model for classification: the size of the acoustic dictionary and the number of latent components. Figure 6 shows the results of audio classification tasks using LSA (dashed lines), pLSA (dotted lines), and acoustic topic model (ATM, solid lines) according to the number of latent components. The size of the dictionary is set to 1000 for this experiment. Figure 6(a) and 6(b) represent the results with respect to onomatopoeic words and semantic labels, respectively. The number of latent components can be interpreted as the dimension of the feature vector extracted from an audio clip.

The results clearly show that classification using the proposed acoustic topic model outperforms LSA and pLSA for both onomatopoeia labels and semantic labels. This significant improvement is evident regardless of the number of latent components[1]. We argue that this benefit comes from utilizing LDA to model the latent topics. 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 a Euclidean space. With the proposed topic model, on the other hand, we are able to model the probabilities of acoustic topics and their priors that generate a specific acoustic word using a generative model.

In classifying onomatopoeia labels, the overall performance is lower than for the task of classifying semantic labels. This might be because the onomatopoeic words are related to context-free subjective (as opposed to context-dependent exact) interpretation of sound content that results in greater overlap between the categories.

Another significant trend that can be observed is that the performance increases as the number of latent components increase. This is reasonable in the sense of more information being captured for the classification task. It should be noted, however, that there is a trade-off between performance and complexity. Increasing the number of latent components to represent audio clips would also increase computing requirements.

Figures 7 and 8 show the classification results as a function of the number of latent components for different codebook. Figure 7 represents the results with respect to onomatopoeic words, while Fig. 8 represents the results with respect to semantic labels. As shown in the figures, the overall performance increases as the number of latent components increases regardless of codebook sizes and types of categories. It is consistent with the previous experimental results with 1000 codewords. Note that the performance variations with respect to the codebook size within a certain number of latent components are greater when the LSA is used. This indicates that the proposed ATM is less sensitive to the number of acoustic words defined, instead its performance depends on the number of latent topics.

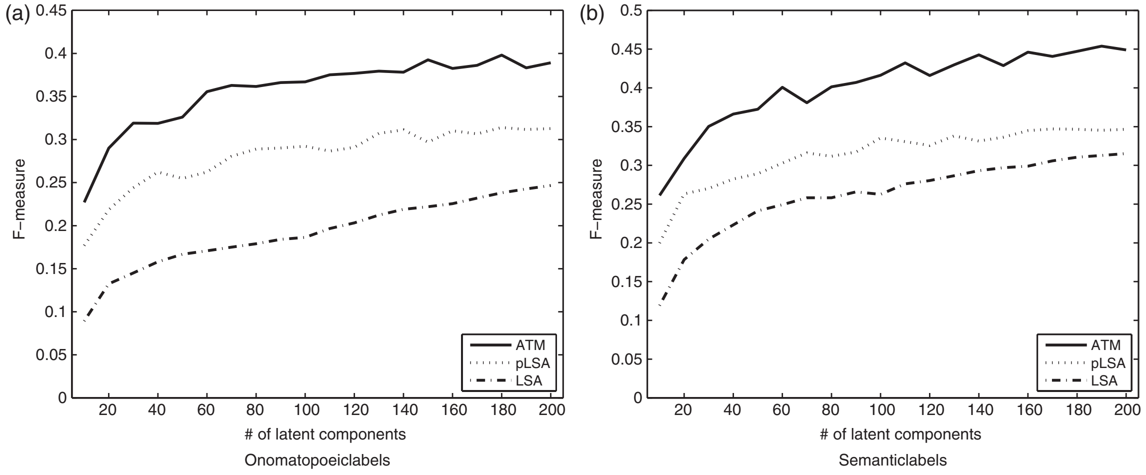
Interestingly, the performance with the same number of latent components in LSA decreases as the codebook size increases, although one might expect tasks with larger codebook should yield better performance since more codewords generally indicate higher resolution to describe given audio signals. It can be partially explained by the fact that the LSA can be considered as a dimension reduction process since it only considers part of eigenvectors, by the nature of SVD; in the case of using 100 latent components, 50% of eigenvectors are used after the decomposition with codebook size 200 while only 2.5% of eigenvectors are used with codebook size 4000. Figure 9 shows additional experiments regarding the codebook size. It illustrates the classification results as a function of the size of the acoustic dictionary while the percentage of latent components are fixed. In this experiment, we set the number of latent components as 5% of the size of the acoustic dictionary for simplicity (e.g., 10 latent components for 200 acoustic words and 200 latent components for 4000 acoustic words). The results again confirm that the proposed acoustic topic model outperforms LSA and pLSA for both onomatopoeia (Fig. 9(a)) and semantic labels (Fig. 9(b)) of audio clips. The performance does not seem to be monotonically improving as the size of the acoustic dictionary increases in the LSA cases. This is because these results are not directly comparable between the different sizes of the acoustic dictionary; the number of latent components required may also be different if we apply different sizes of acoustic dictionary.

We also perform the following experiments to use the proposed ATM as complementary features within conventional content-based audio tag classification task. The rationale is that the ATM can be used to disambiguate acoustic features that are common in several sounding situations while the content-based classification methods represent the overall distribution of the acoustic features. In this work, we use the GMM–SVM classifier for content-based audio tag classification.

[1] The Wilcoxon signed-rank test was used to show the statistical significance.

Table 2. Distribution of onomatopoeic words and semantic labels in the BBC sound library (22 onomatopoeic labels and 21 semantic labels).

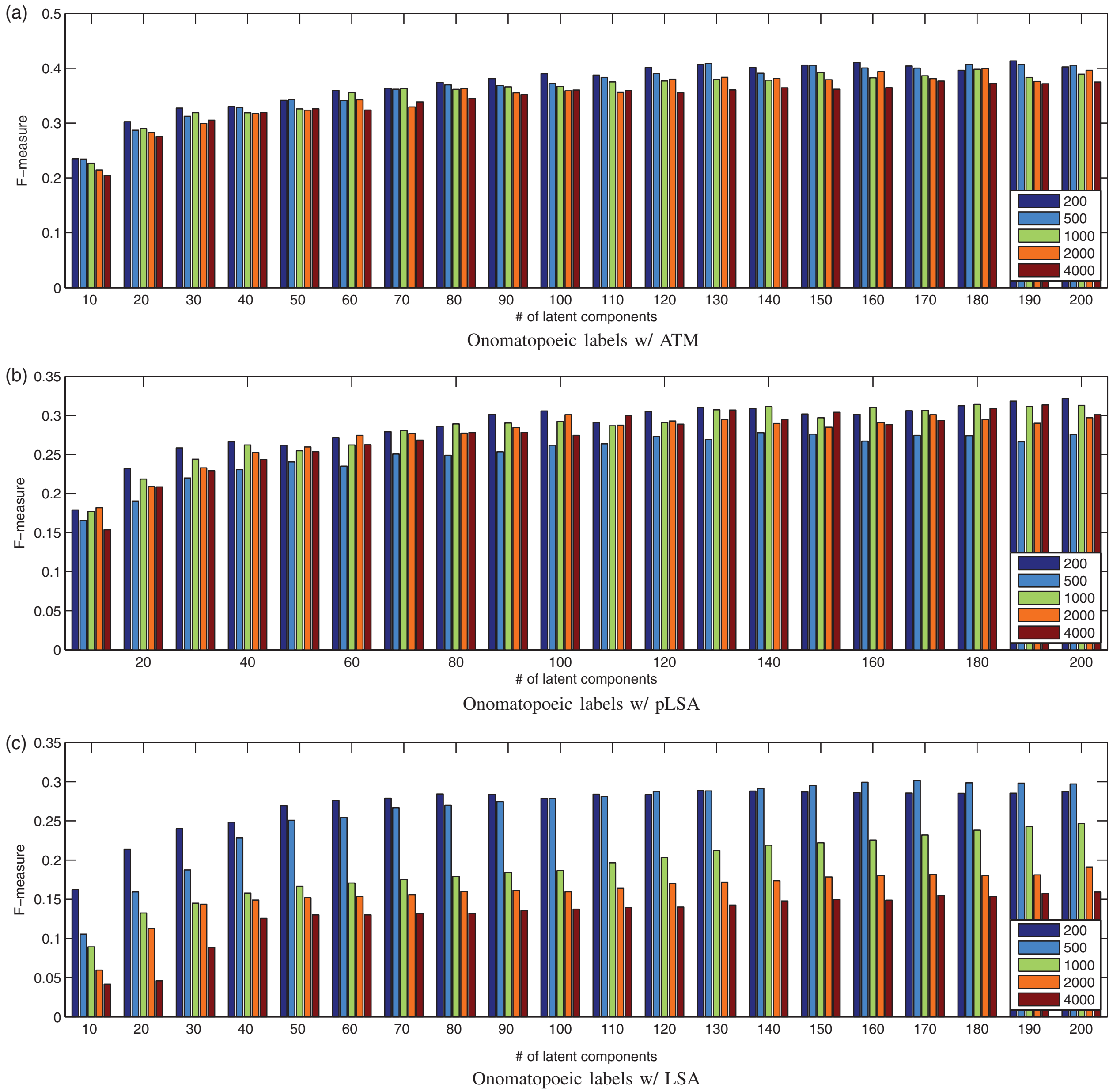
Semantic labels Onomatopoeic labels	ANIMALS	HUMAN	TRANSPORTATION	OFFICE	MACHINERY	ELECTRONICS	PUBLIC	POLICE	HORROR	MILITARY	AMBIENCES	NATURE	HOUSEHOLD	SCI-FI	SPORTS	DOORS	OPEN	IMPACT	MUSIC	AUTOMOBILES	EXPLOSIONS	SUM
TWEET	53	1	0	0	0	0	0	0	1	0	50	3	0	0	0	0	0	0	0	0	0	108
SQUEAK	54	10	6	7	2	1	1	1	12	0	8	4	1	1	0	4	2	0	0	9	1	124
CLATTER	26	4	47	20	13	8	0	1	0	17	7	3	1	0	8	0	1	1	0	0	0	157
GABBLE	0	33	0	15	0	0	9	0	2	0	56	0	0	0	13	0	0	0	0	0	0	128
BURR	0	4	43	5	24	0	0	8	0	17	17	0	0	0	1	0	0	0	0	24	0	143
DONG	3	6	8	4	3	5	3	5	6	0	7	1	9	14	6	0	0	4	23	1	0	108
BUZZ	13	3	26	18	18	3	2	1	10	5	17	11	5	0	11	0	0	1	0	3	4	151
BLEAT	14	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17
GROWL	62	1	1	5	0	0	0	0	1	1	0	0	0	1	1	0	0	0	0	0	0	73
HUM	2	1	72	16	44	6	4	5	6	22	15	3	7	20	12	0	0	0	0	4	1	240
TAP	33	177	1	10	0	3	1	0	0	0	0	0	0	0	13	0	2	0	0	0	0	240
BEEP	0	3	5	5	0	17	0	46	2	0	1	0	0	30	0	0	0	0	0	1	1	111
WHOOSH	2	3	2	0	0	0	0	0	11	4	2	14	1	37	9	0	0	0	0	0	1	86
BANG	1	0	1	3	0	1	0	3	4	26	0	14	0	9	2	0	1	0	0	0	4	69
HONK	2	6	15	0	2	0	1	21	0	3	3	4	0	0	1	0	0	1	0	5	0	64
TICK	1	0	2	0	0	3	0	0	1	1	2	0	4	0	0	0	0	0	1	2	0	17
THUD	3	17	14	1	7	1	0	1	11	2	0	1	0	1	12	0	2	2	0	2	0	77
CRACKLE	1	8	0	0	1	0	0	0	2	3	0	14	1	1	0	0	0	0	0	0	3	34
CRUNCH	10	4	5	5	2	0	0	1	7	0	1	1	0	1	3	0	0	7	0	0	1	48
SPLASH	3	3	26	0	0	0	2	0	2	0	1	10	8	0	10	0	0	0	0	0	1	66
MEOW	60	0	0	6	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	67
CROW	6	1	0	4	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	12
SUM	349	285	274	126	117	48	23	93	79	101	187	83	37	116	102	4	8	16	24	51	17	2140



Onomatopoeiclabels

Semanticlabels

Fig. 6. Audio tag classification results using LSA, pLSA and ATM according to the number of latent components: (a) onomatopoeic labels and (b) semantic labels.

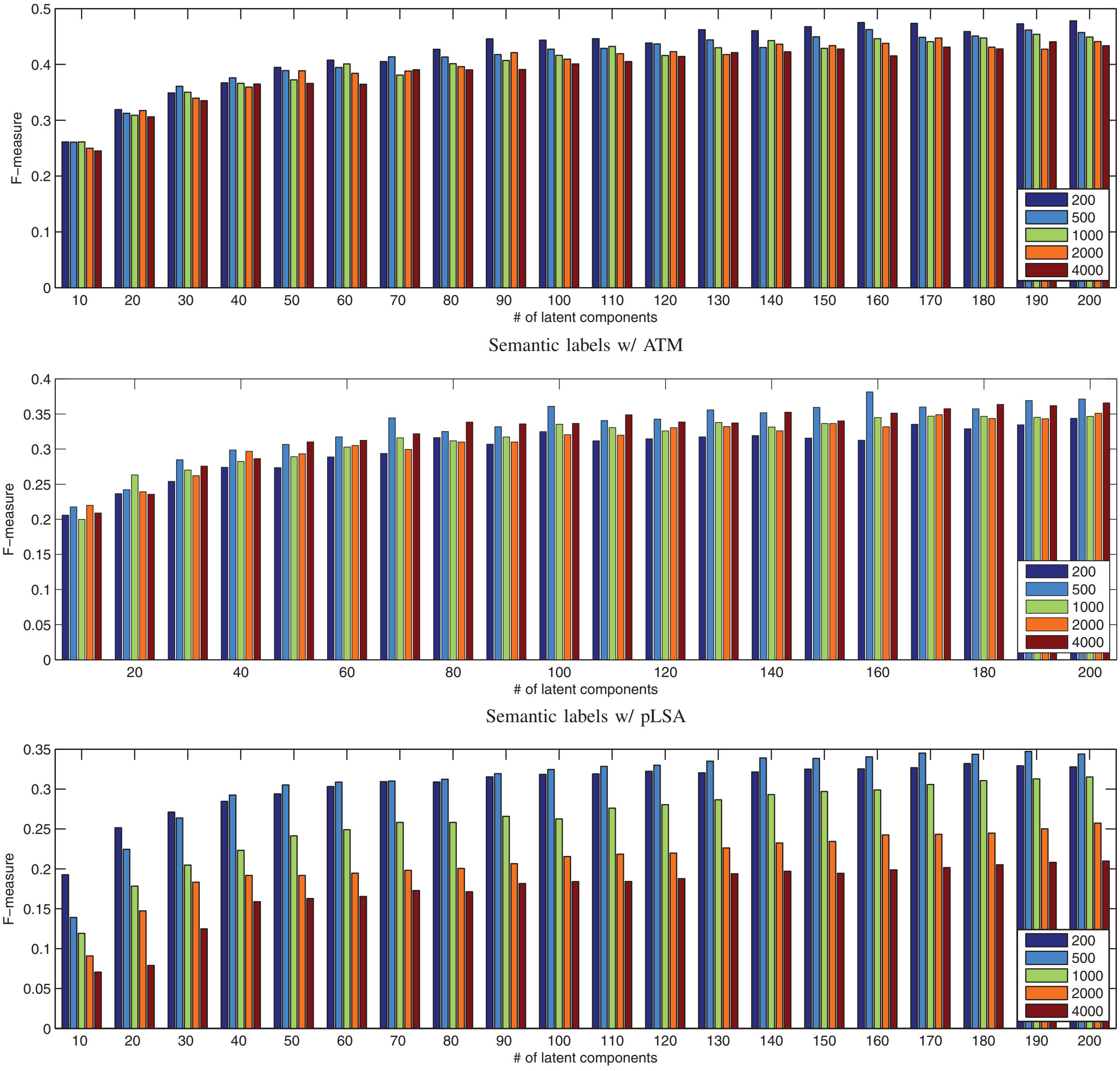


Onomatopoeic labels w/ ATM

Onomatopoeic labels w/ pLSA

Onomatopoeic labels w/ LSA

Fig. 7. Audio tag classification results with respect to onomatopoeic labels using (a) ATM, (b) pLSA, and (c) LSA according to the number of latent components and the size of acoustic word dictionary.



Semantic labels w/ ATM

Semantic labels w/ pLSA

Semantic labels w/ LSA

Fig. 8. Audio tag classification results with respect to semantic labels using (a) ATM, (b) pLSA, and (c) LSA according to the number of latent components and the size of acoustic word dictionary.

Figures 10(a) and 11(a) show the classification results as a function of the number of latent components when we use the proposed ATM (colored bars) and the GMM–SVM (solid lines) separately on onomatopoeic labels and semantic labels. The number of latent components represents the number of latent topics for the ATM and the number of Gaussian mixtures for the GMM–SVM. The sizes of feature vectors are k and $12 \times k$ for the ATM and GMM–SVM, respectively, where k is the number of latent components. Note that the size of acoustic dictionary does not apply for the GMM–SVM cases since there is no quantization process. The results show that the ATM itself may not provide much information toward audio tag information compared to the content-based GMM–SVM method.

To examine the complementary information embedded in the ATM, we apply a feature-level hybrid method to make a super-vector of the GMM mean super-vector and the topic distribution. For simplicity, we choose to perform the hybrid method for those cases that have the same number of latent components so that the size of feature vector should be $k + 12 \times k = 13 \times k$. The classification results as a function of the number of latent components are shown in Figs 10(b) and 11(b) for the hybrid method (colored bars) and the GMM-SVM (solid lines). In the figures we can observe the feature-level hybrid method can improve overall performance by providing complementary information in the topic distribution to content-based GMM method in audio tag classification tasks: $8.6 \pm 0.7\%$ relative improvement for each setting. We argue that these improvements

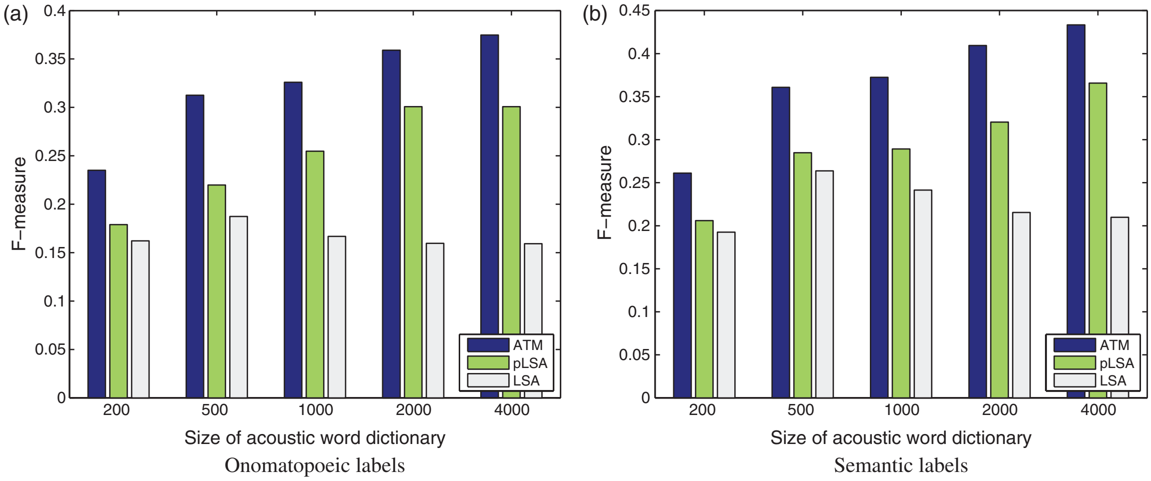


Fig. 9. Audio tag classification results with respect to (a) onomatopoeic labels and (b) semantic labels using ATM, pLSA and LSA according to the size of acoustic word dictionary; when the number of topics are 5% of size of the dictionary.

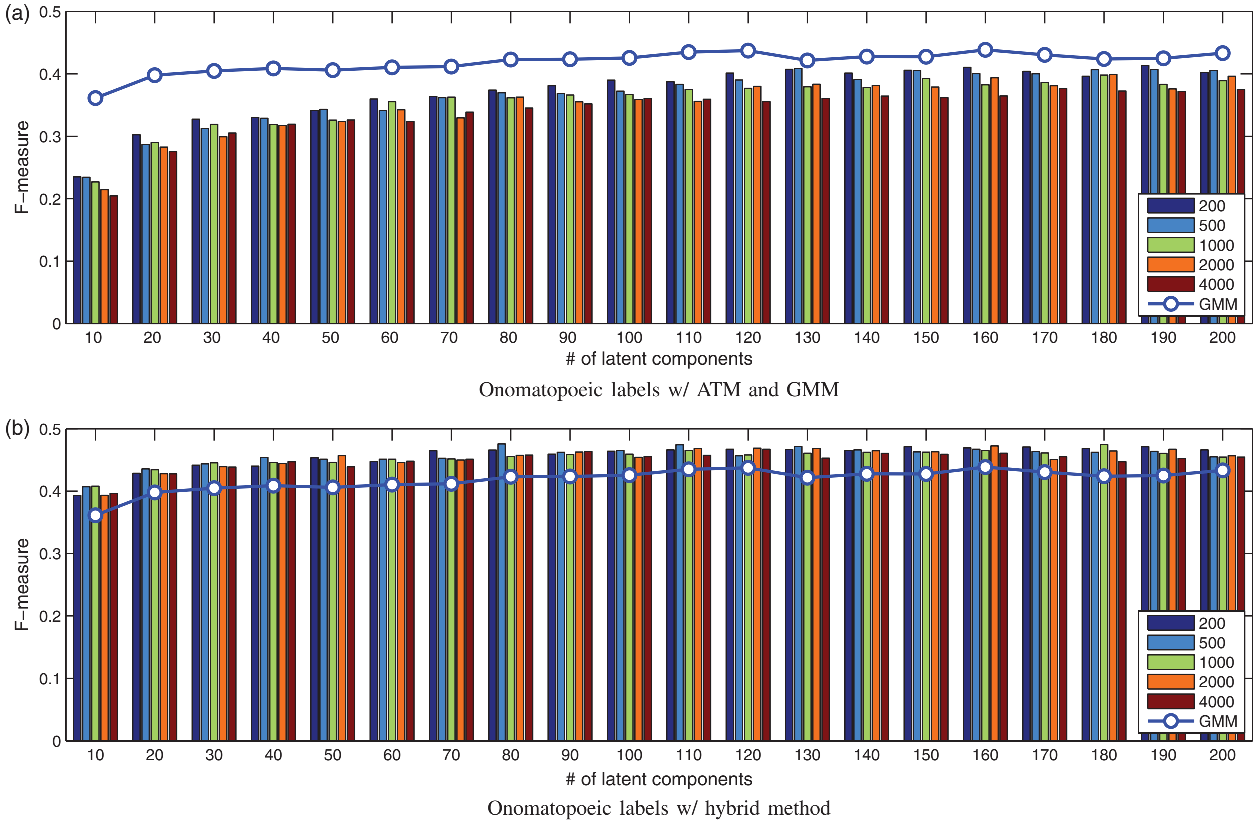
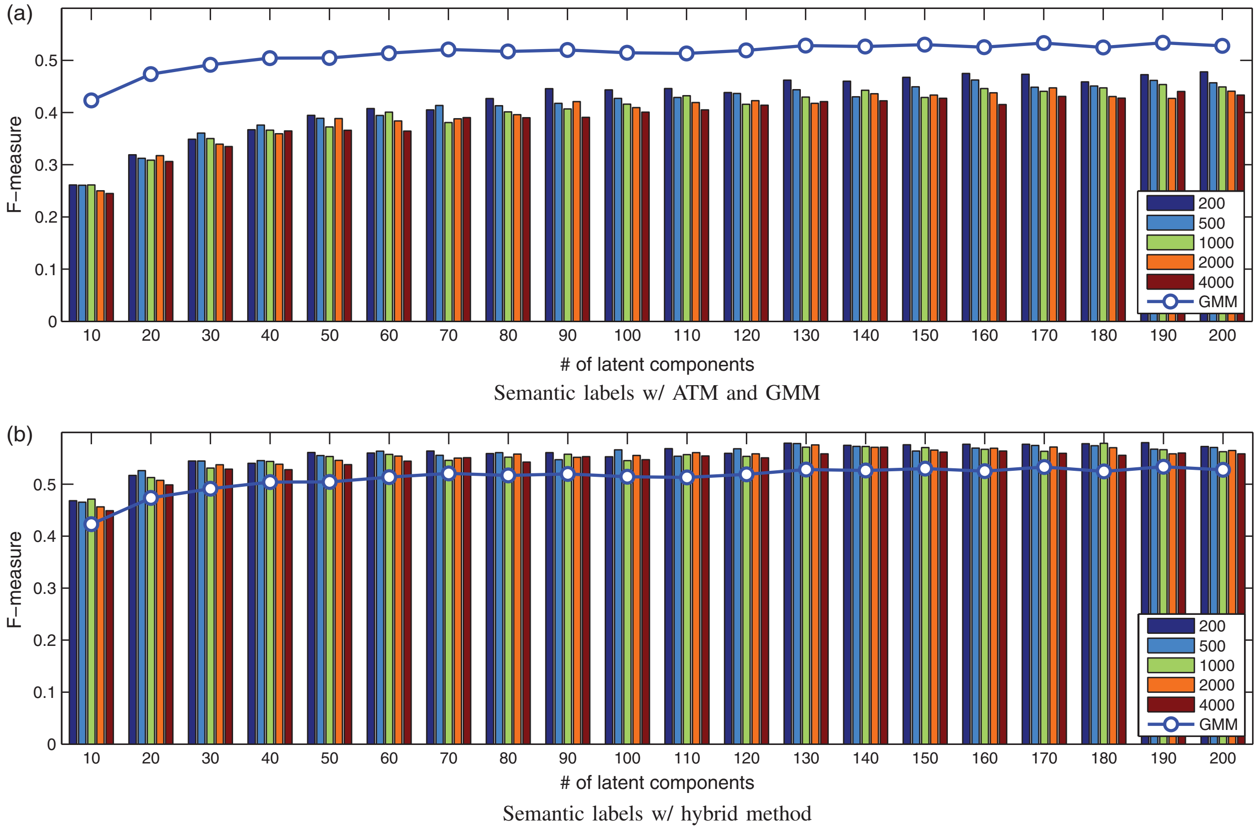


Fig. 10. Audio tag classification results with respect to onomatopoeic labels using (a) ATM, GMM, and (b) their hybrid according to the number of latent clusters.

come from modeling the inherent structural context that underscores the fact that perceptually similar acoustic content may lead to different semantic interpretation depending on the evidence provided by the co-occurring acoustic sources.

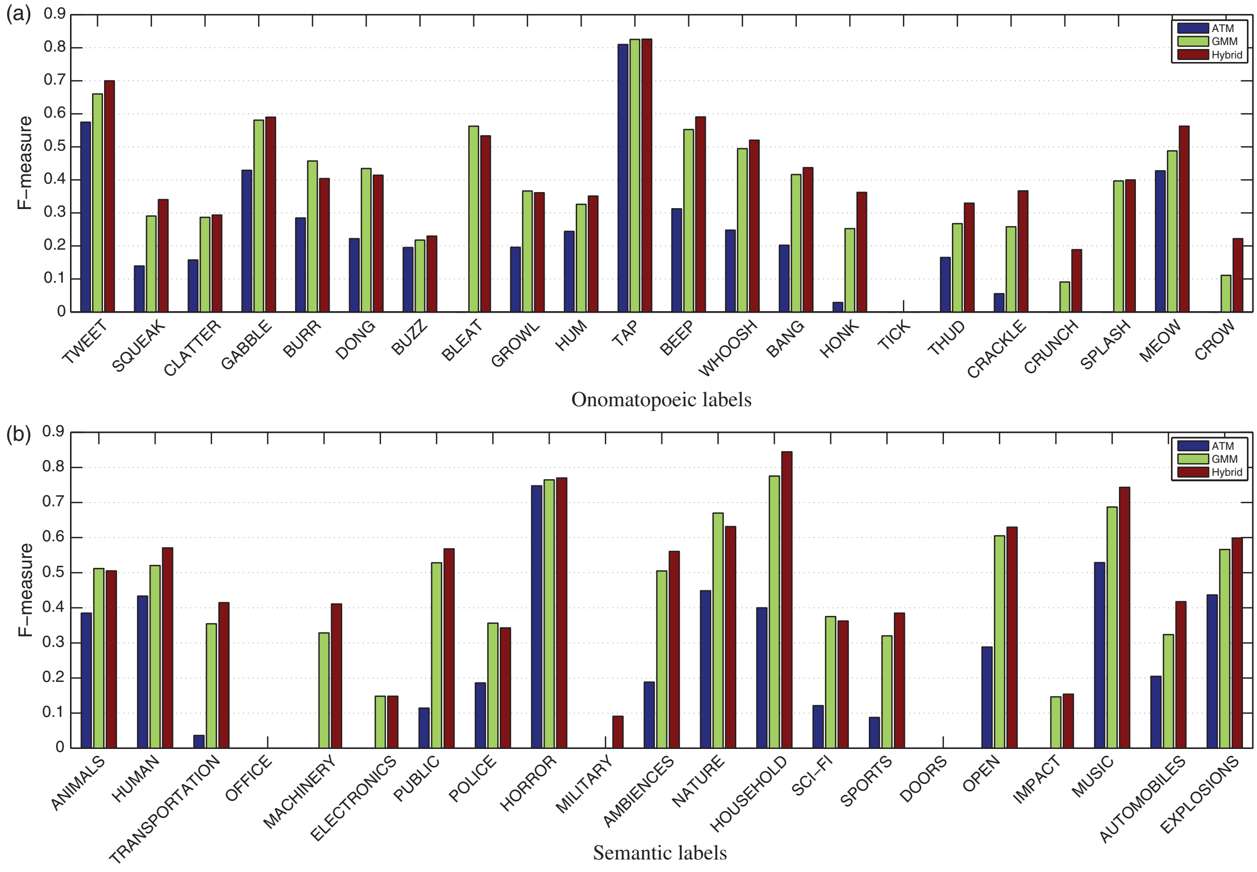
Figure 12 provides further details of the classification tasks, i.e., per-class F-measure using ATM, GMM, and their hybrid method, particularly for the case that the number of latent components is 100 and the size of acoustic word dictionary is 1000. The per-class F-measure is computed by collecting all the classification results of the 5-fold cross-validation. This reveals that the feature-level hybrid method can improve the classification performance in most of the classes as well as the overall performance and supports our argument that the proposed topic models provide complementary information to the content-based method in audio tag classification tasks. Note that, however, there are several classes that cannot be recognized at all in any classification strategy, e.g., *tick* in onomatopoeia and *office, doors* in semantic labels. Although the number of instances for those classes might affect the performance, it is not always true. For example, there only four instances for the



Semantic labels w/ ATM and GMM

Semantic labels w/ hybrid method

Fig. 11. Audio tag classification results with respect to semantic labels using (a) ATM, GMM, and (b) their hybrid according to the number of latent clusters.



Onomatopoeic labels

Semantic labels

Fig. 12. Per-class F-measure of audio tag classification results with respect to (a) onomatopoeic labels and (b) semantic labels, using ATM, GMM, and their hybrid method. The number of latent components is 100, and the size of acoustic word dictionary is 1000.

semantic label *doors* while there are 126 instances for the semantic label *office*.

V. CONCLUDING REMARKS

In this paper, a novel approach to incorporate context information for audio classification using acoustic topic models was presented. The proposed approach was tested using unstructured audio clips from the BBC Sound Effects Library. The framework discussed here supports both example-based and text-based query for audio information retrieval. In this regard, the proposed work can be viewed as contributing to a further generalization of the content-based audio retrieval problems.

We proposed an acoustic topic model based on LDA, which learns hidden acoustic topics in a given audio signal in an unsupervised manner. We adopted the variational inference method to train the topic model and used the posterior Dirichlet parameters as a representative feature vector for the audio clip. Due to the rich acoustic information present in audio clips, they can be categorized based on the intermediate audio description layer which includes semantic and onomatopoeic categories, and considered to represent the cognition of the acoustic realization of a scene and its perceptual experience, respectively. The classification results for the two descriptions showed that the acoustic topic model significantly outperforms the conventional latent structure analysis methods considered, such as LSA and pLSA, and offer promising results in providing complementary information to improve content-based modeling methods. Finally, experimental results show that the proposed context-based classification method can be advantageously combined with content-based classification. Specifically, the combinations of the proposed acoustic topic model with a GMM–SVM system was shown to yield significant F-score performance improvements of about 9%.

Our future work plans to test various refinements of the LDA model to associate with descriptions of audio clips and various fusion strategies with content-based approaches. We will also perform similar tasks with even more heterogeneous audio clips that include overlaps between different sound sources, such as audio signals from TV programs and movies.

VI. ACKNOWLEDGEMENT

This research was supported in part by the fund from the National Science Foundation (NSF) and Department of Defense (DoD).

REFERENCES

[1] The BBC Sound Effects Library–Original Series. [Online]. Available: http://www.sound-ideas.com

[2] Slaney, M.: Semantic-audio retrieval, in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 4, Orlando, Florida, USA, 2002, pp. 4108–4111.

[3] Turnbull, D.; Barrington, L.; Torres, D.; Lanckriet, G.: Semantic annotation and retrieval of music and sound effects, IEEE Trans. Audio, Speech, Language Process., **16**(2), (2008) 467–476.

[4] Chechik, G.; Ie, E.; Rehn, M.; Bengio, S.; Lyon, R. F.: Large-scale content-based audio retrieval from text queries, in *ACM Int. Conf. on Multimedia Information Retrieval (MIR)*, 2008, 105–112.

[5] Chu, S.; Narayanan, S.; Kuo, C.-C. J.: Environmental sound recognition with joint time- and frequency-domain audio features, *IEEE Trans. Speech, Audio Language Process.*, **17**(6), (2009) 1142–1158.

[6] Blei, D. M.; Ng, A. Y.; Jordan, M. I.: Latent Dirichlet Allocation, J. Mach. Learn. Res., **3**, (2003) 993–1022.

[7] Griffiths, T. L.; Steyvers, M.; Tenenbaum, J. B.: Topics in semantic representation, Psychol. Rev., **114**(2), (2007) 211–244.

[8] Steyvers, M.; Griffiths, T.: Probabilistic Topic Models, *Laurence Erlbaum*, Hillsdale, NJ, 2006.

[9] Barnard, K.; Duygulu, P.; Forsyth, D.; Freitas, N.; Blei, D. M.; Jordan, M. I.: Matching words and pitctures, J. Mach. Learn. Res., **3**(2003) 1107–1135.

[10] Blei, D. M.; Jordan, M. I.: Modeling annotated data, in *Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, Toronto, Canada, 2003, 127–134.

[11] Yakhnenko, O.; Honavar, V.: Multi-modal hierarchical dirichlet process model for predicting image annotation and image-object label correpsondence, in *SIAM Conf. on Data Mining*, Sparks, Nevada, USA, 2009, 283–293.

[12] Wang, C.; Blei, D. M.; Fei-Fei, L.: Simultaneous image classification and annotation, in *IEEE Conf. on Computer Vision and Pattern Recognition*, Miami, Florida, USA, 2009, 1903–1910.

[13] Smaragdis, P.; Shashanka, M. V.; Raj, B.: Topic models for audio mixture analysis, in *Neural Information Processing System (NIPS) Workshop (Applications for Topic Models: Text and Beyond)*, Whistler, B.C., Canada, 2009.

[14] Smaragdis, P.; Raj, B.; Shashanka, M. V.: Sparse and shift-invariant feature extraction from non-negative data, in *IEEE Int. Conf. on Audio and Speech Signal Processing*, Las Vegas, Nevada, USA, 2008, 2069–2072.

[15] Sundaram, S.; Narayanan, S.: Audio retrieval by latent perceptual indexing, in *IEEE Int. Conf. on Acoustic, Speech, and Signal Processing (ICASSP)*, Las Vegas, Nevada, USA, 2008, 49–52.

[16] Sundaram, S.; Narayanan, S.: A divide-and-conquer approach to latent perceptual indexing of audio for large web 2.0 application, in *IEEE International Conf. of Multimedia and Expo*, New York, NY, USA, 2009, 466–469.

[17] Lee, K.; Ellis, D.: Audio-based semantic concept classification for consumer video, IEEE Trans. Audio, Speech, Language Process., **18**(6), (2010) 1406–1416.

[18] Levy, M.; Sandler, M.: Music information retrieval using social tags and audio, IEEE Trans. Multimedia, **11**(3), (2009) 383–395. [Online]. Available: http://dx.doi.org/10.1109/TMM.2009.2012913

[19] Hu, D. J.; Saul, L. K.: A probabilistic topic model for unsupervised learning of musical key-profiles, in *Int. Conf. on Music information Retrieval*, Kobe, Japan, 2009, 441–446.

[20] Madsen, R. E.; Kuchak, D.: Modeling word burstiness using the dirichlet distribution, in *Int. Conf. on Machine Learning*, Bonn, Germany, 2005, 545–552.

[21] Griffiths, T. L.; Steyvers, M.: Finding scientific topics, in Proc. Nat. Acad. Sci. U.S.A., **101**, 2004, 5228–5235.

[22] Cover, T.; Thomas, J.: Elements of Information Theory, *Wiley and Sons*, New York, 1991.

[23] Watanabe, S.; Mochihashi, D.; Hori, T.; Nakamura, A.: Gibbs sampling based multi-scale mixture model for speaker clustering, in *IEEE Int. Conf. on Acoustic, Speech, and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011, 4524–4527.

[24] Rabiner, L. R.; Juang, B.: Fundamentals of Speech recognition, *Prentice Hall*, 1993.

[25] Gersho, A.; Gray, R. M.: Vector Quantization and Signal Compression, *Kluwer Academic Publishers*, Norwell, MA, USA, 1991.

[26] Kim, S.; Narayanan, S.; Sundaram, S.: Acoustic topic models for audio information retrieval, in *Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, October 2009, 37–40.

[27] Kim, S.; Sundaram, S.; Georgiou, P.; Narayanan, S.: Audio scene understanding using topic models, in *Neural Information Processing System (NIPS) Workshop (Applications for Topic Models: Text and Beyond)*, Whistler, B.C., Canada, 2009.

[28] Kim, S.; Sundaram, S.; Georgiou, P.; Narayanan, S.: An n-gram model for unstructured audio signals toward information retrieval, in *IEEE Int. Workshop on Multimedia Signal Processing (MMSP)*, Saint-Malo, France, 2010, 477–480.

[29] Kim, S.; Georgiou, P.; Sundaram, S.; Narayanan, S.: Acoustic stopwords for unstructured audio information retrieval, in *European Signal Processing Conf. (EUSIPCO)*, Aalborg, Denmark, 2010, 1277–1280.

[30] Kakkonen, T.; Myller, N.; Sutinen, E.; Timonen, J.: Comparison of dimension reduction methods for automated essay grading, J. Educational Technol. Soc., **11**(3), (2008) pp. 275–288.

[31] Wallach, H. M.: Topic modeling: beyond bag-of-words, in *Proc. 23rd Int. Conf. on Machine Learning*, ser. ICML'06. *ACM*, New York, NY, USA, 2006, 977–984. [Online]. Available: http://doi.acm.org/10.1145/1143844.1143967

[32] Blei, D. M.; McAuliffe, J. D.: Supervised topic models, Adv. Neural Inform. Process. Syst., **20**(2007) 121–128.

[33] Kim, S.; Georgiou, P. G.; Narayanan, S.: Supervised acoustic topic model for unstructured audio information retrieval, in *Asia Pacific Signal and Information Processing Association (APSIPA) Annual Summit and Conf.*, Singapore, 2010, 243–246.

[34] Kim, S.; Georgiou, P.; Narayanan, S.: Supervised acoustic topic model with a consequent classifier for unstructured audio classification, in *Workshop on Content-Based Multimedia Indexing (CBMI)*, Annecy, France, June 2012, 121–126.

[35] Kim, S.; Georgiou, P.; Narayanan, S.; Sundaram, S.: Using naive text queries for robust audio information retrieval system, in *IEEE Int. Conf. on Acoustic, Speech, and Signal Processing*, 2010, 2046–2049.

[36] Jebara, T.; Kondor, R.; Howard, A.: Probability product kernels, J. Mach. Learn. Res., **5**(2004), 819–844. [Online]. Available: http://portal.acm.org/citation.cfm?id=1005332.1016786

[37] Bellegarda, J. R.: Latent semantic mapping, IEEE Signal Process. Mag., **22**(5), (2005) 70–80.

[38] Hofmann, T.: Probabilistic latent semantic indexing, in *Proc. 22nd Annual Int. SIGIR Conf. on Research and Development in Information Retrieval (SIGIR-99)*, 1999, 50–57.

[39] Campbell, W.; Sturim, D.; R. D.; S. A.: SVM based speaker verification using a gmm supervector kernel and svm based speaker verification using a GMM supervector kernel and nap variability compensation, in *IEEE Int. Conf. on Acoustic, Speech, and Signal Processing (ICASSP)*, Toulouse, France, 2006, 97–100.

[40] van Rijsbergen, C. J.: Information Retrieval, *Butterworths*, London, 1979.

Samuel Kim received his Ph.D. on Electrical Engineering from University of Southern California, Los Angeles, California in 2010 and B.S. (with a high honor) and M.S. on Electrical and Electronic Engineering from Yonsei University, Seoul, Korea in 2003 and 2005, respectively. Currently, he is a postdoctoral researcher at the Idiap Research Institute, Martigny, Switzerland working on the social signal processing project (SSPNet). Before he came to Switzerland in June 2011, he was with University of Southern California as a postdoctoral researcher working on speech-to-speech translation system (SpeechLinks), analyzing physiologic data (Knowme), and audio/music information retrieval. He has published over 35 research papers in the peer reviewed scientific literature. His research focuses on machine learning and digital signal processing algorithms for multimedia data, specifically audio, speech, music and text, with applications to data mining, information retrieval, classification, compression, and enhancement.

Panayiotis G. Georgiou received the B.A. and M.Eng. degrees (with Honors) from Cambridge University (Pembroke College), Cambridge, U.K., in 1996 and the M.Sc. and Ph.D. degrees from the University of Southern California (USC), Los Angeles, in 1998 and 2002, respectively. During the period 1992 to 1996, he was awarded a Commonwealth scholarship from Cambridge-Commonwealth Trust. Since 2003, he has been a member of the Signal Analysis and Interpretation Lab at USC, where he is currently a Research Assistant Professor. His interests span the fields of multimodal and behavioral signal processing. He has worked on and published over 100 papers in the fields of behavioral signal processing, statistical signal processing, alpha stable distributions, speech and multimodal signal processing and interfaces, speech translation, language modeling, immersive sound processing, sound source localization, and speaker identification. He has been a PI and co-PI on federally funded projects notably including the DARPA Transtac SpeechLinks and currently the NSF An Integrated Approach to Creating Enriched Speech Translation Systems and Quantitative Observational Practice in Family Studies: The case of reactivity. He is currently serving as a guest editor of the Computer Speech And Language journal and as a member of the Speech and Language Technical Committee. His current focus is on behavioral signal processing, multimodal environments, and speech-to-speech translation. Dr Georgiou received best paper awards for analyzing the multimodal behaviors of users in speech-to-speech translation in International Workshop on Multimedia Signal Processing (MMSP) 2006 and for automatic classification of married couplesÕ behavior using audio features in Interspeech 2010.

Shrikanth (Shri) Narayanan is Andrew J. Viterbi Professor of Engineering at the University of Southern California (USC), and holds appointments as Professor of Electrical Engineering, Computer Science, Linguistics and Psychology and as the founding director of the Ming Hsieh Institute. Prior to USC he was with AT&T Bell Labs and AT&T Research from 1995 to 2000. At USC, he directs the Signal Analysis and Interpretation Laboratory (SAIL). His research focuses on human-centered information processing and communication technologies with a special emphasis on behavioral signal processing and informatics. [http://sail.usc.edu] Prof. Narayanan is a Fellow of the Acoustical Society of America and the American Association for the Advancement of Science (AAAS) and a member of

Tau Beta Pi, Phi Kappa Phi, and Eta Kappa Nu. He is also an Editor for the Computer Speech and Language Journal and an Associate Editor for the IEEE Transactions on Multimedia, IEEE Transactions on Affective Computing, and the Journal of the Acoustical Society of America. He was also previously an Associate Editor of the IEEE Transactions of Speech and Audio Processing (2000–2004) and the IEEE Signal Processing Magazine (2005–2008). He is a recipient of a number of honors including Best Paper awards from the IEEE Signal Processing Society in 2005 (with A. Potamianos) and in 2009 (with C. M. Lee) and selection as an IEEE Signal Processing Society Distinguished Lecturer for 2010–2011. Papers co-authored with his students have won awards at Interspeech 2012 Speaker Trait Challenge, Interspeech 2011 Speaker State Challenge, InterSpeech 2010, InterSpeech 2009-Emotion Challenge, IEEE DCOSS 2009, IEEE MMSP 2007, IEEE MMSP 2006, ICASSP 2005, and ICSLP 2002. He has published over 500 papers and has 13 granted U.S. patents.