Classification of Overlapped Audio Events Based on AT, PLSA, and the Combination of Them

Yan LENG [1], Chengli SUN [2], Chuanfu CHENG [1], Xinyan XU [3], Si LI [4], Honglin WAN [1], Jing FANG [1], Dengwang LI [1]

[1] College of Physics and Electronics, Shandong Normal University, East Wenhua Road 88, 250014, Ji'nan, China
[2] School of Information, Nanchang Hangkong University, Nanchang, 330063, China
[3] Dept. of Computer Science and Technology, Shandong College of Electronic Technology, Wenhua Road 678, 250200, Zhangqiu, Ji'nan, China
[4] Dept. of Computer Science, Brandeis University, 415 South St, Waltham, MA 02453, USA

lyansdu@163.com, sun_chengli@163.com, chengchuanfu@sdnu.edu.cn, xuxinyan@sohu.com, ls198cf@gmail.com, visage1979@sdu.edu.cn, shanshifangjing@sina.com, lidengwang@sdnu.edu.cn

Abstract. *Audio event classification, as an important part of Computational Auditory Scene Analysis, has attracted much attention. Currently, the classification technology is mature enough to classify isolated audio events accurately, but for overlapped audio events, it performs much worse. While in real life, most audio documents would have certain percentage of overlaps, and so the overlap classification problem is an important part of audio classification. Nowadays, the work on overlapped audio event classification is still scarce, and most existing overlap classification systems can only recognize one audio event for an overlap. In this paper, in order to deal with overlaps, we innovatively introduce the author-topic (AT) model which was first proposed for text analysis into audio classification, and innovatively combine it with PLSA (Probabilistic Latent Semantic Analysis). We propose 4 systems, i.e. AT, PLSA, AT-PLSA and PLSA-AT, to classify overlaps. The 4 proposed systems have the ability to recognize two or more audio events for an overlap. The experimental results show that the 4 systems perform well in classifying overlapped audio events, whether it is the overlap in training set or the overlap out of training set. Also they perform well in classifying isolated audio events.*

Keywords

Audio event classification, author-topic model, PLSA, overlapped audio event, isolated audio event

1. Introduction

Audio information, as a manifestation of multimedia information, can carry rich information, and has been developed and applied extensively [1–5]. Recently, audio event classification technology which is an important part of Computational Auditory Scene Analysis (CASA) has attracted much attention. Unlike audio event detection, which means to determine the identity and the occurrence time of the sounds that may exist in an audio document, audio event classification is to identify the sounds in the given audio segments. Audio event classification is useful in a variety of applications, including multimedia retrieval [6], intelligent robots [7], and smart home project etc. [8]. For an audio document, there are two types of audio event which can be defined as follows:

***Definition 1 Isolated Audio Event*:** The audio event that does not have temporal overlap with other audio events. That is, at the time when the audio event occurs, no other audio events occur simultaneously.

***Definition 2 Overlapped Audio Event*:** The audio event that has temporal overlap with other audio events. That is, at the time when the audio event occurs, there are other audio events that occur simultaneously.

Nowadays, the audio classification technology is mature enough to classify the isolated audio events accurately, but when encounters with the overlapped ones, large performance decay would occur. In the international evaluation campaign of CLEAR 2007 [9], the overlapped segments (the segments that contain overlapped audio events) account for more than 70% of errors produced by every submitted system. Toni Heittola [10] pointed out that the overlapped audio events would make the automatic sound event recognition problem more difficult to handle. So dealing with the overlapped audio events is really a challenge. While in real life, most audio files would have certain percentage of overlapped audio events, and so overlapped audio event classification is an important part for audio file analysis. The overlapped audio events constitute a natural auditory scene. Most researches did the auditory scene recognition by modeling global acoustic characteristics of the auditory scene, and had neglected the classification of the overlapped audio events. In this paper, we propose several overlap classification systems based on two topic models, i.e. AT (author-topic model) [11] and PLSA (Probabilistic Latent Semantic Analysis) [12]. Both AT and

PLSA were first proposed in text analysis field. AT can extract the topic information of authors, and PLSA can extract the topic information of documents. The two topic models will be briefly introduced in Sec. 3. The related work will be described in Sec. 2. The problem of how to use the two topic models and the combination of them to classify the overlaps will be discussed in Sec. 3. The experimental results are presented in Sec. 4. Finally, conclusions and future work are given in Sec. 5.

2. Related Work

Both AT and PLSA are specific cases of topic models. AT is in fact an extension of the LDA (Latent Dirichlet Allocation) [13]. So far there has been no report on applying AT in audio field, but much work has been done on applying LDA in audio retrieval. For example, Samuel Kim [14] assumed that an audio clip was a mixture of some acoustic topics, and took LDA to extract the topic distribution information for each audio clip to realize audio retrieval. Pengfei Hu [4] overcame the shortage of LDA in processing continuous data, and proposed a new topic model named Gaussian-LDA for audio retrieval. In this paper we introduce AT into audio classification based on the idea that an audio document can be expressed as a combination of acoustic topics as well as a combination of acoustic events. A similar idea is proposed in [15], where a LATEA (Latent Acoustic Topic and Event Allocation) model was proposed for acoustic scene analyzing. The difference is that instead of expressing an audio document as a combination of acoustic events, LATEA expresses an acoustic topic as a combination of acoustic events. PLSA is a popular topic model in audio processing field. Yuxin Peng [16] employed audio PLSA model to do semantic annotation. Through PLSA, Keansub Lee [17] decomposed the soundtrack into separate descriptions of the specific sounds, and successfully applied it to classify consumer videos. With the latent topics learnt by PLSA, Timothy J. Hazen [18] proposed a method to automatically summarize the content of an audio corpus.

As that pointed out in [19], the overlap problem can be addressed at different system levels. At the signal level, the overlap problem is related to the source separation technology. For example, in order to detect sound events from everyday contexts, Toni Heittola [10] adopted the source separation technology to separate the audio signal into four individual signals, and then each individual signal was separately processed and classified. At the decision level, the overlap problem is dealt by assigning different weights to different microphones based on the assumption that the audio sources are well separated in space. At the model level, the overlap problem is resolved by modeling all types of overlap. For example, in [19], a SVM-based audio event detection system, called ISO-CLUSTER, was proposed to detect the non-speech events that were overlapped with speech in meeting-room environment. The ISO-CLUSTER system is a two-step approach. First, a [mp] class which contains all overlaps is defined. The [mp] class, along with the ISO system (the system that is constructed only by isolated audio events) is used to complete the set of 1 vs. 1 SVM classifiers. Then in order to further classify the detected overlapped segments, an optimal decision tree is generated based on a confusion matrix. At each node of the decision tree, the audio event classes are split into two clusters by minimizing the splitting criterion shown in formula (1), and then a SVM model is trained.

$$i(C_1, C_2) = \frac{1}{|C_1||C_2|} \sum_{i=1}^{|C_1|} \sum_{j=1}^{|C_2|} \left(\frac{e_{ij}}{e_{ii}} + \frac{e_{ji}}{e_{jj}} \right). \quad (1)$$

Here, e_{ij} is the i,j-th element of the confusion matrix, and $|C_1|$, $|C_2|$ denote cardinalities of the two clusters.

There are also other model-level systems dealing with the audio overlap problem. Another system that was also designed to detect non-speech audio events was proposed by Miquel Espi [20]. In [20], some hidden features were learnt from spectrogram patches, and then were integrated within the deep neural network to detect audio events. The exemplar-based NMF approach for audio event detection in [21], the context-dependent sound event detection in [10], and the HMM based sound event detection in [22] have a similar idea. They all employed the Viterbi algorithm of HMM to detect the most likely event for each frame. Also, in order to detect more events in an overlapped frame, they all experimented with multiple Viterbi passes. At each pass, for each state, all events of the previous passes were forbidden. But the authors in [21] pointed out that the method of multiple Viterbi passes did not yield satisfactory results, because it would cause large numbers of insertion errors. Considering the success of the tandem connectionist-HMM in automatic speech recognition, Xiaodan Zhuang [23] introduced it into the real-world acoustic event detection. Other models, such as GMM [24], and the model constructed by NMF (Non-negative Matrix Factorization) [25] were also adopted to detect overlapped audio events.

Most of the model-level systems can only recognize one audio event for an overlap. In this paper, we aim to propose systems to recognize more than one audio event for an overlap. To do so, we adopt the topic models of AT, PLSA, and the combination of them. When combine AT with PLSA, one of them is used to find out the potential audio events in an overlap, and the other is used to determine the final audio events among them. The contributions of our work are as follows. To deal with the overlap classification problem, we innovatively introduce AT into audio classification, and innovatively combine it with PLSA. We design 4 systems, i.e. AT, PLSA, AT-PLSA and PLSA-AT to resolve the overlap classification problem. The proposed systems have the ability to recognize two or more audio events in an overlap, which cannot be done by most of the current audio overlap classification systems. Also we have tested the ability of AT, PLSA, AT-PLSA and PLSA-AT in classifying isolated audio event.

3. Classification of Overlapped Audio Events Based on AT, PLSA, and the Combination of Them

PLSA is first proposed by Thomas Hofmann for text analysis [12]. It can discover the latent topical structure of text documents, and so is very useful in disambiguating polysemes and in exploring synonyms. PLSA is also feasible in the audio field. There have been many studies that apply PLSA to do audio analysis.

AT is first proposed to extract the author and topic information of large text collections [11]. It is a generative model based on the idea that a document can be represented as a mixture of topics. AT takes the authorship information into consideration, and is in fact an extension of LDA [13]. For text documents, AT can be applied to rank authors by topic, or to rank topics by author, and to parse abstracts by topics and authors, etc. An audio document is comparable to a text document. The audio components are equivalent to words, and the audio events are equivalent to the authors of the text. Thus it is feasible to introduce AT into audio field.

In this section, we will first introduce PLSA and AT briefly. The symbols of AT are consistent with those used in [11]. Then the two topic models as well as the combination of them are tested to deal with the overlap classification problem.

3.1 PLSA

For a corpus with D documents, assume the words in the corpus are taken from a dictionary with W unique words, and there are totally T topics, denoted as $z_t \in \{z_1, \ldots, z_T\}$. Let $P(d_i)$ denote the probability that a word will be observed in document d_i, $P(w_j \mid z_t)$ the probability of word w_j conditioned on the latent topic z_t, $P(z_t \mid d_i)$ the probability of the latent topic z_t conditioned on document d_i. With these definitions, the generation process of the corpus can be described as follows:

(1) Pick a document d_i with probability $P(d_i)$;

(2) Choose a latent topic z_t with probability $P(z_t \mid d_i)$;

(3) Generate a word w_j with probability $P(w_j \mid z_t)$.

The goal of PLSA modeling is to maximize the following joint probability with the constraints $\sum_{j=1}^{W} P(w_j \mid z_t) = 1$ and $\sum_{t=1}^{T} P(z_t \mid d_i) = 1$:

$$L = \sum_{i=1}^{D}\sum_{j=1}^{W} n(d_i, w_j)\log P(d_i, w_j)$$
$$= \sum_{i=1}^{D} n(d_i)\left[\log P(d_i) + \sum_{j=1}^{W}\frac{n(d_i, w_j)}{n(d_i)}\log\sum_{t=1}^{T} P(w_j \mid z_t)P(z_t \mid d_i)\right] \quad (2)$$

Here $n(d_i, w_j)$ denotes the number of words w_j in document d_i, and $n(d_i) = \sum_j n(d_i, w_j)$ denotes the document length. EM (Expectation Maximization) is employed to resolve the above maximum likelihood estimation, and finally $P(w_j \mid z_t)$ and $P(z_t \mid d_i)$ could be obtained.

3.2 The Author-Topic Model

Assume there are T topics and A authors in the text corpus, and the words in the corpus are taken from a dictionary with W unique words. Θ stands for a $T \times A$ matrix whose element θ_{ta} denotes the probability of assigning topic t to a word generated by author a. The column θ_a in Θ indicates the multinomial distribution over topics for author a, and satisfies $\sum_{t=1}^{T}\theta_{ta} = 1$. Φ stands for a $W \times T$ matrix whose element ϕ_{wt} denotes the probability of generating word w from topic t. The column ϕ_t in Φ indicates the multinomial distribution over words for topic t, and satisfies $\sum_{w=1}^{W}\phi_{wt} = 1$. Take the A_d-dimensional vector $\boldsymbol{a}_d$ to represent the authors of document d, and take the N_d-dimensional vector $\boldsymbol{w}_d$ to represent the words in document d, then a corpus with D documents can be represented by a vector $\boldsymbol{w}$ obtained through concatenating all document vectors, and thus $\boldsymbol{w}$ has $N = \sum_{d=1}^{D} N_d$ entries. Each word in the corpus is associated with a latent author, x, and a latent topic, z, and then we use N-dimensional vectors $\mathbf{x}$ and $\mathbf{z}$ to represent the latent authors and the latent topics for the N words of the corpus. Assume the prior distributions of θ_a and ϕ_t are symmetric Dirichlet with hyperparameters α and β respectively, and the authors of each document are known in advance, then the generation process of the corpus can be described as follows:

(1) For each author a $(a = 1, \ldots, A)$, generate θ_a according to the Dirichlet distribution with hyperparameters α; for each topic t $(t = 1, \ldots, T)$, generate ϕ_t according to the Dirichlet distribution with hyperparameters β.

(2) For word i $(i = 1, \ldots, N_d)$ in document d $(d = 1, \ldots, D)$, given the authors $\boldsymbol{a}_d$, first, choose an author x_{di} uniformly at random; next, choose a topic z_{di} according to the multinomial distribution $\theta_{x_{di}}$; finally, choose a word w_{di} according to the multinomial distribution $\phi_{z_{di}}$.

The graphical model of the generation process is shown in Fig. 1.

The key point of the author-topic model is to estimate the parameter Θ and Φ. This is done by estimating the posterior distribution $P\left(\Theta, \Phi \mid D^{train}, \alpha, \beta\right)$ through the following equation:

$$P\left(\Theta, \Phi \mid D^{train}, \alpha, \beta\right) = \sum_{\mathbf{z},\mathbf{x}} P\left(\Theta, \Phi \mid \mathbf{z}, \mathbf{x}, D^{train}, \alpha, \beta\right)P\left(\mathbf{z}, \mathbf{x} \mid D^{train}, \alpha, \beta\right) \quad (3)$$

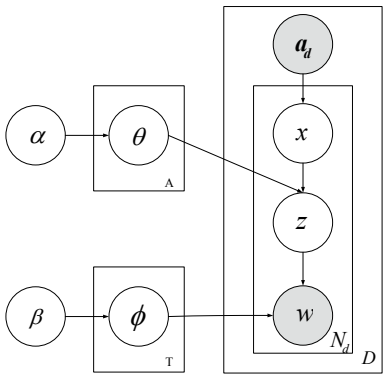


Fig. 1. Graphical model for the author-topic model [11].

The above equation is executed as follows: first, a sample-based estimation of $P\left(\mathbf{z},\mathbf{x}\middle|D^{train},\alpha,\beta\right)$ is obtained through Gibbs sampling; second, for any specific sample, $P\left(\Theta,\Phi\middle|\mathbf{z},\mathbf{x},D^{train},\alpha,\beta\right)$ can be computed directly.

To assess the convergency of the Markov chain, a perplexity score is proposed as follows:

$$\begin{aligned} &Perplexity\left(\boldsymbol{w}_d\middle|\boldsymbol{a}_d,\Theta,\Phi\right) \\ &\quad=\exp\left(-\frac{\log P\left(\boldsymbol{w}_d\middle|\boldsymbol{a}_d,\Theta,\Phi\right)}{N_d}\right) \end{aligned} \tag{4}$$

Here $P\left(\boldsymbol{w}_d\middle|\boldsymbol{a}_d,\Theta,\Phi\right)$ is the posterior probability of words $\boldsymbol{w}_d$ conditioned on $\boldsymbol{a}_d$, Θ and Φ, and can be calculated as follows:

$$\begin{aligned} &P\left(\boldsymbol{w}_d\middle|\boldsymbol{a}_d,\Theta,\Phi\right) \\ &=\prod_{i=1}^{N_d}P\left(w_{di}\middle|\boldsymbol{a}_d,\Theta,\Phi\right) \\ &=\prod_{i=1}^{N_d}\sum_{a=1}^{A}\sum_{t=1}^{T}P\left(w_{di},z_{di}=t,x_{di}=a\middle|\boldsymbol{a}_d,\Theta,\Phi\right) \\ &=\prod_{i=1}^{N_d}\sum_{a=1}^{A}\sum_{t=1}^{T}P\left(w_{di}\middle|z_{di}=t,\Phi\right)P\left(z_{di}=t\middle|x_{di}=a,\Theta\right)P\left(x_{di}=a\middle|\boldsymbol{a}_d\right) \\ &=\prod_{i=1}^{N_d}\frac{1}{A_d}\sum_{a\in\boldsymbol{a}_d}\sum_{t=1}^{T}\phi_{w_{di}t}\theta_{ta} \end{aligned} \tag{5}$$

From (4) it can be seen that lower perplexity value means larger posterior probability $P\left(\boldsymbol{w}_d\middle|\boldsymbol{a}_d,\Theta,\Phi\right)$, and therefore means better performance of the model.

3.3 Classification of Overlapped Audio Events

1) Classification through AT

In order to apply AT to classify the overlapped audio events, there are three key problems needed to be resolved.

(1) How to get the "words" of one audio document?
(2) What are the authors of one audio document?
(3) With AT, how to perform classification?

(1) To get the "words" of audio documents, here we adopt the vector quantization method. The training audio documents are first split into frames, and for each frame, some audio features are extracted. Assume there are totally L frames in the corpus, denoted as $\{f_1, f_2,\ldots,f_L\}$. The frames are clustered by k-means. Assume the frames are clustered into W clusters, and then the cluster centroids, denoted as $\{C_1, C_2,\ldots,C_W\}$, are taken as the dictionary of size W. With the dictionary, each frame would then get an index as follows:

$$IDX\left(f_i\right)=\underset{j,j\in\{1,2,\cdots,W\}}{arg\,min}\ Dis\left(f_i,C_j\right) \tag{6}$$

where $Dis\left(f_i,C_j\right)$ represents the distance between f_i and C_j, $IDX\left(f_i\right)$ represents the index of frame f_i, and then $\left\{IDX\left(f_i\right)\middle|i=1,2,\cdots,L\right\}$ are just the "words" of the audio documents.

(2) An audio document is comparable to a text document. The audio components are equivalent to words, and the audio event classes are equivalent to the authors. For a text document, it can be represented as a combination of latent topics, and its authors are the people who write it. AT can extract the topic distribution of each author and the word distribution of each topic. The combination of the author-topic distributions and the topic-word distributions then generates the text. Similarly and reasonably, an audio document can also be represented as a combination of latent topics which can be understood in the same way as that they are understood in the text document. For an audio document, we think that the audio event classes in it have generated it, and then we take the audio event classes as the authors. For example, if an audio segment is an overlap with speech and music in it, then we say that the authors of this audio segment are speech and music. With AT, the topic distribution of each audio event and the word distribution of each topic could be obtained, and then an audio document could be generated by combining the audio event-topic distributions and the topic-word distributions.

(3) In [11], the application of AT includes detecting unusual papers by authors and separating the combined documents into its component parts. In this paper we extend its application, and reform it to classify the overlaps. How to reform it to be fit for classification is the key problem.

From (4) it can be seen that conditioned on specific authors, perplexity can be used to estimate the posterior probability of a document. Inspired by it, we think that conditioned on a specific audio event, the perplexity value can be used to estimate the likelihood that the audio event is one of the real audio events in the document. But one problem is that the calculation of perplexity needs that the

authors should be known in advance, while the authors of the audio documents are just what we want to get. To overcome this problem, we design the following classification scheme. For an overlapped audio segment to be tested, conditioned on each author, one perplexity value could be got, and with the authors appearing in the training set, a series of perplexity values could be obtained. In Sec. 3.2 we have discussed that the lower the perplexity the better the performance of the model, and then if an author is one of the real authors of the segment, we have reason to believe that the perplexity value conditioned on it should be relatively small. Based on the above discussion, we take the authors with smaller perplexity values as the audio events or potential audio events of the segment.

To express the above idea more clearly, here we define some variables as follows. Assume there are totally A authors, that is, $a \in \{1,\ldots,A\}$. For an overlapped test segment d_{test}, conditioned on author a, a perplexity value, $Perplexity\,(d_{test} \mid a, \Theta, \Phi)$, could be obtained according to (4). Then the audio events or potential audio events of audio segment d_{test}, denoted as $AE(d_{test})$, can be expressed as follows:

$$AE(d_{test}) = \arg \underset{a}{F_M}\, min\{Perplexity\left(d_{test} \mid a, \Theta, \Phi\right), a = 1, \cdots, A\} \tag{7}$$

Here $F_M min$ denotes the first M minimum values.

2) Classification through PLSA

To reform PLSA to be fit for overlap classification, the concepts of word, topic and document should be redefined. The words and the construction of the dictionary are the same as that in AT. The topics refer to audio events, or in other words, refer to authors, and then $P(w_j \mid z_t)$ should be rewritten as $P(w_j \mid a)$, $a \in \{1,\ldots,A\}$, and $P(z_t \mid d_i)$ should be rewritten as $P(a \mid d_i)$, $a \in \{1,\ldots,A\}$. The document refers to the audio segment segmented from the original audio documents. That is to say, the original audio documents are segmented into a series of shorter segments, and these segments are taken as the classification units. In the training stage of PLSA, as that described in Sec. 3.1, $P(w_j \mid a)$, $a \in \{1,\ldots,A\}$, could be obtained through EM. Since here we refer topics to audio events, and in the training set, the audio events in each audio segment are known, then $P(w_j \mid a)$, $a \in \{1,\ldots,A\}$, does not need to be calculated through EM, but can simply be obtained through statistics. For an overlapped segment in the training set, it should participate the statistics of all the audio events contained in it. For example, if an overlapped segment contains the audio events of speech and music, then it should participate the statistics of $P(w_j \mid a)$ for speech and also for music. In the test stage, for an audio segment d_{test}, $P(a \mid d_{test})$ can be obtained through EM. In each M-step, only $P(a \mid d_{test})$ is updated, while the $P(w_j \mid a)$, $a \in \{1,\ldots,A\}$, obtained from the training set are kept fixed. $P(a \mid d_{test})$ reflects the test segment-specific probability distribution over audio events. The audio events with larger probability can be taken as the audio events or the potential audio events in the test segment. That is:

$$AE(d_{test}) = \arg \underset{a}{F_M}\, max\{P(a \mid d_{test}), a = 1, \cdots, A\} \tag{8}$$

Here $F_M max$ denotes the first M maximum values.

3) Classification through Combining AT with PLSA

From the above discussion it can be seen that both AT and PLSA can be used separately to classify overlaps, and both can recognize two or more audio events for an overlap. Also, we can combine AT with PLSA to classify overlaps. Here we design two combination strategies. One is that we use AT to find out the potential audio events for a test audio segment, and then within these potential events, PLSA is performed to find out the most likely audio events which are then taken as the classification result. This strategy will be denoted as AT-PLSA hereafter. The other is that we use PLSA to find out the potential audio events, and then within these potential events, AT is performed to find out the first several audio events with smaller perplexity values, and then such audio events are taken as the classification result. This strategy will be denoted as PLSA-AT hereafter. More details about AT-PLSA and PLSA-AT are explained as follows.

AT-PLSA: For a test segment d_{test}, first, $M1$ potential audio events, denoted as a_i, $i = 1,2,\ldots,M1$, are determined through (7); then, for a_i, $P(a_i \mid d_{test})$ $i = 1,2,\ldots,M1$ are obtained as that described in 2); finally, among these potential audio events, $M2$ $(1 \le M2 < M1)$ audio events, selected through (8), are taken as the classification result. That is, the classification result can be expressed as:

$$AE(d_{test}) = \arg \underset{a_i}{F_{M2}}\, max\{P(a_i \mid d_{test}), i = 1, \cdots, M1\} \tag{9}$$

PLSA-AT: For a test segment d_{trst}, first, $M1$ potential audio events, denoted as a_i, $i = 1, 2, \cdots, M1$, are determined through (8); then, for $a_i, i = 1, 2, \cdots, M1$, a series of perplexity values, $Perplexity\,(d_{test} \mid a_i, \Theta, \Phi)$, $i = 1, 2, \cdots, M1$, are obtained as that described in 1); finally, among these potential audio events, $M2$ $(1 \le M2 < M1)$ audio events, selected through (7), are taken as the classification result. That is, the classification result can be expressed as:

$$AE(d_{test}) = \arg \underset{a_i}{F_{M2}}\, min\{Perplexity\left(d_{test} \mid a_i, \Theta, \Phi\right), i = 1, \cdots, M1\} \tag{10}$$

The 4 proposed systems of AT, PLSA, AT-PLSA and PLSA-AT will all be tested to classify overlaps in the experimental section. Also we are interested in the classification performance of the 4 systems in classifying isolated

audio events and in classifying the complete test set (including overlapped audio events and isolated audio events), so these two aspects will also be tested.

4. Experimental Results

4.1 Dataset, Feature and Metric

The proposed systems are evaluated on two datasets. One is a dataset constructed by the first 5 episodes of drama "Band of Brothers", abbreviated as BOB dataset, and the other is a dataset constructed by 5 episodes of melodrama "Friends", abbreviated as Friends dataset. The average length of one episode is about 55 minutes in BOB dataset, and about 22 minutes in Friends dataset. For both datasets, the audio events are hand-labeled, and the labeling results are shown in Tab. 1. The time intervals for which the content is difficult to describe are labeled as unknown, and are not used. The audio events that occur rarely in the dataset are not labeled and not used. Since the silence class can be easily classified through a threshold of energy, it is also not used. The audio recordings are split into segments according to labels.

Dataset	Event Type	
BOB	Isolated Event	speech, airplane, machine, explosion, shot, shout, music, door, step, laugh, traffic, sigh
	Overlapped Event	speech&airplane, speech&airplane&music, speech&machine, speech&explosion, speech&explosion&shot, speech&shot, speech&shot&traffic, speech&music, speech&step, speech&traffic, airplane&explosion, airplane&shot, airplane&music, explosion&shot, explosion&shout, explosion&music, explosion&step, explosion&sigh, shot&shout, shot&music, shot&step, shot&traffic, music&door, music&step, music&traffic, step&traffic
Friends	Isolated Event	speech, music, laugh, applause, door, step, silence, unknown
	Overlapped Event	speech&music, speech&laugh, speech&door, speech&step, speech&applause, music&laugh, music&door, laugh&applause, laugh&door, speech&laugh&applause, speech&laugh&door, speech&music&laugh, music&step, music&applause, music&door, speech&music&door

Tab. 1. The labeling results of the two datasets.

The audio segments are set to be mono channel format, down-sampled to 16 kHz, and framed using a Hamming window. The frame length/shift is 32/16 ms. For each frame, some features are extracted. MFCCs, as the most efficient audio features, are first adopted. Some other features that are proposed in works about content-based audio analysis are also adopted, including energy entropy, signal energy, zero crossing rate, spectral rolloff, spectral centroid and spectral flux.

The evaluation metrics are: the segment-based version of audio event error rate (AEER), precision (Pre), recall (Rec), and F1-measure, which are defined as follows:

$$AEER = \frac{De + In + Su}{Num} , \tag{11}$$

$$Pre = \frac{ce}{es} , \tag{12}$$

$$Rec = \frac{ce}{gt} , \tag{13}$$

$$F1 = \frac{2 \cdot Pre \cdot Rec}{Pre + Rec} . \tag{14}$$

Here *Num*, *De*, *In*, and *Su* are the number of events to classify for a specific segment, the number of deletions, the number of insertions, and the number of substitutions respectively. *gt*, *es*, and *ce* denote respectively the number of ground truth, estimated and correctly estimated audio events for a given audio segment. Segment-level metrics are averaged throughout the segments in the test set.

4.2 Experimental Setting

In order to determine the parameters in the topic models, one episode is chosen from each dataset to construct an experiment dataset. For the rest episodes in each dataset, the leave-one-out cross validation is adopted. Each time one episode is chosen as the test set, and the rest as the training set. The average performance of all the combinations of training-test set is taken as the final result.

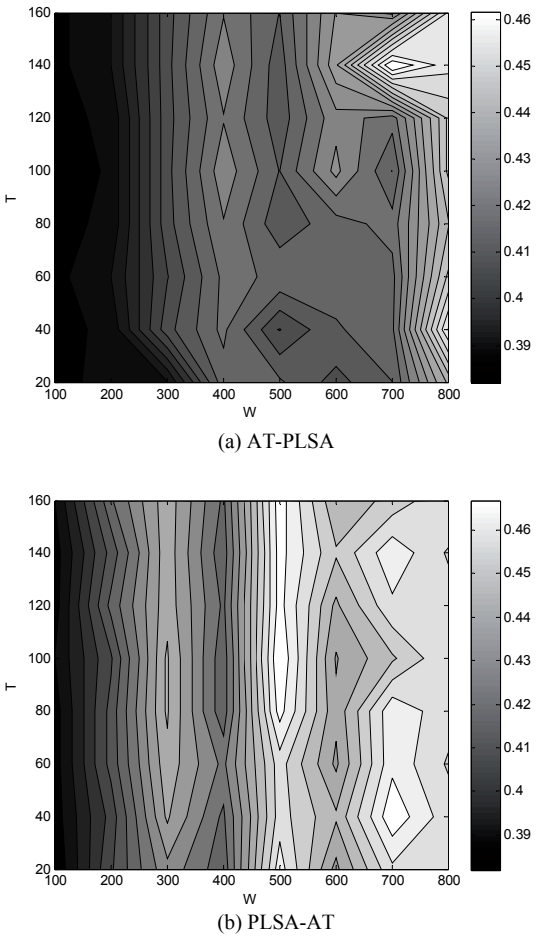
The proposed system is compared with the baseline system and the ISO-CLUSTER system both proposed in [19]. In [19], the authors only considered the overlapped segments in which one non-speech audio event is overlapped with speech, and the other overlaps of two or more audio events, either with or without speech are not used. In this paper, the overlapped segments of two or more audio events, whether the audio event is speech or non-speech, are all considered. The baseline system is constructed by several SVM classifiers, and the 1 vs. 1 multi-class classification strategy is used. Both the segments of isolated audio events and the segments of overlapped audio events in the training set are used to train the SVM classifiers. The overlapped segments are averagely assigned to the corresponding classes. For example, for the overlapped segments in which there are two audio events of A and B, 50% of the segments are included in class A, and the other 50% in class B. The ISO-CLUSTER system is trained by segments of isolated and overlapped audio events as that proposed in [19]. To be consistent with that in [19], all SVM classifiers use RBF kernel function. The parameter of the kernel function and the penalty factor of SVM are determined by 5-fold cross validation.

The baseline and the ISO-CLUSTER systems classify an overlapped segment as a certain audio event. In other words, they cannot recognize two or more audio events in an overlapped segment. Obviously, for an overlapped segment, we wish to recognize as many audio events in it as possible. Recognizing two or more audio events in an overlapped segment can help people to analyze the audio scenes, which cannot be well done by recognizing only one audio event, and also it is useful in other applications. For example, for an overlapped audio segment of the type speech&bus, if both audio events have been correctly recognized, then it can help us to infer that it is an outdoor scene, but this cannot be done by recognizing only speech. From the labeling results of the two datasets it can be seen that most of the overlaps are the overlapping of two audio events, so in order to classify an overlapped segment, we design the systems as follows: For AT, the 2 audio events determined through (7) are taken as the classification result; For PLSA, the 2 audio events determined through (8) are taken as the classification result; For AT-PLSA, 5 potential audio events are first found out by AT, and then among the potential audio events, the first 2 most likely audio events are determined by PLSA, and are taken as the classification result; For PLSA-AT, 5 potential audio events are first found out by PLSA, and then among the potential audio events, the 2 with the first two minimum perplexity values are taken as the classification result. Also we hope to test the performance of the proposed systems in classifying isolated audio events. A SVM classifier trained with some overlapped segments and some isolated segments (the segments that contain isolated audio events) is used to determine whether a test segment is an overlapped one or an isolated one. For an isolated segment, the classification strategy is similar to that of an overlapped segment. For AT/PLSA, the most likely audio event determined through (7)/(8) is taken as the classification result, and for AT-PLSA and PLSA-AT, among the 5 potential audio events, the most likely one determined by PLSA (for AT-PLSA), or by AT (for PLSA-AT) is taken as the classification result. With the classification results of isolated audio events and overlapped audio events, the overall classification performance of the systems is also tested.

The burn-in time of the Gibbs sampler is set to be 1000, and the parameters α and β are set to be $200/W$ and $50/T$ respectively, as that suggested in [11].

4.3 Determine W and T

To run the proposed systems, the size of the dictionary, W, and the number of topics, T, should be determined in advance. In our experiments, the optimal W and T are found by full grid searching of the F1-measure surface obtained on the experiment dataset. From each dataset, except for the episode chosen to construct the experiment dataset, from the rest episodes, 3 are randomly chosen, and then the 6 episodes from the two datasets are used to construct a training set. With this training set, the AT-PLSA model and the PLSA-AT model are trained respectively,



(a) AT-PLSA

(b) PLSA-AT

Fig. 2. The F1-measure contours at different W and different T, obtained on the experiment dataset: (a) for AT-PLSA, and (b) for PLSA-AT.

and are then tested respectively on the experiment dataset with different W and different T, as that shown in Fig. 2.

It can be seen that for AT-PLSA, generally, W should be no less than 700, and for PLSA-AT, generally, W should be no less than 500. When W is big enough, T can decrease appropriately. An appropriate W can well describe the content of the audio corpus, and an appropriate T can well discover the latent semantic structure of the audio corpus. For AT and AT-PLSA, W and T are set to be 700 and 140. For PLSA and PLSA-AT, they are set to be 500 and 100.

4.4 Classification Results

In this section, the proposed systems are compared with the baseline system and the ISO-CLUSTER system. Table 2 and Table 3 show the classification results of overlapped segments on dataset BOB and Friends respectively. Table 4 and Table 5 show the classification results of isolated segments on dataset BOB and Friends respectively. Table 6 and Table 7 show the overall classification results (the classification results on the complete test set, including isolated segments and overlapped segments) on dataset BOB and Friends respectively.

Metrics / Systems	AEER	Pre	Rec	F1
baseline	1.077	0.620	0.300	0.403
ISO-CLUSTER	1.180	0.551	0.267	0.359
AT	1.574	0.473	0.461	0.465
PLSA	1.810	0.396	0.382	0.387
AT-PLSA	1.574	0.473	0.461	0.465
PLSA-AT	1.554	0.481	0.467	0.473

Tab. 2. The classification results of overlapped segments on dataset BOB.

Metrics / Systems	AEER	Pre	Rec	F1
baseline	0.682	0.884	0.433	0.581
ISO-CLUSTER	0.710	0.865	0.424	0.569
AT	0.465	0.917	0.669	0.778
PLSA	0.479	0.926	0.643	0.759
AT-PLSA	0.469	0.917	0.665	0.774
PLSA-AT	0.448	0.926	0.674	0.780

Tab. 3. The classification results of overlapped segments on dataset Friends.

Metrics / Systems	AEER	Pre	Rec	F1
baseline	1.805	0.398	0.398	0.398
ISO-CLUSTER	1.580	0.474	0.474	0.474
AT	1.500	0.500	0.500	0.500
PLSA	1.832	0.389	0.389	0.389
AT-PLSA	1.526	0.491	0.491	0.491
PLSA-AT	1.539	0.487	0.487	0.487

Tab. 4. The classification results of isolated segments on dataset BOB.

Metrics / Systems	AEER	Pre	Rec	F1
baseline	0.845	0.718	0.718	0.718
ISO-CLUSTER	0.803	0.732	0.732	0.732
AT	0.688	0.771	0.771	0.771
PLSA	0.815	0.728	0.728	0.728
AT-PLSA	0.706	0.765	0.765	0.765
PLSA-AT	0.730	0.757	0.757	0.757

Tab. 5. The classification results of isolated segments on dataset Friends.

Metrics / Systems	AEER	Pre	Rec	F1
baseline	1.480	0.497	0.354	0.401
ISO-CLUSTER	1.397	0.509	0.382	0.423
AT	1.533	0.488	0.482	0.485
PLSA	1.822	0.392	0.386	0.389
AT-PLSA	1.547	0.482	0.478	0.480
PLSA-AT	1.546	0.484	0.478	0.480

Tab. 6. The overall classification results on dataset BOB.

Metrics / Systems	AEER	Pre	Rec	F1
baseline	0.804	0.759	0.647	0.684
ISO-CLUSTER	0.780	0.765	0.656	0.692
AT	0.633	0.807	0.746	0.773
PLSA	0.732	0.777	0.707	0.736
AT-PLSA	0.647	0.803	0.740	0.767
PLSA-AT	0.660	0.799	0.737	0.763

Tab. 7. The overall classification results on dataset Friends.

From Tab. 2 and Tab. 3 it can be seen that when classify overlapped audio events, for the Friends dataset, the 4 proposed systems perform much better than the baseline and the ISO-CLUSTER systems; for the BOB dataset, except for PLSA, the other 3 systems perform better than the baseline and ISO-CLUSTER systems from the perspective of Rec and F1-measure, but a little worse from the perspective of AEER and Pre. Compare AT with AT-PLSA, it can be seen that they perform similarly on both datasets, which means that the classification performance of AT is not enhanced after combining it with PLSA, so when classify overlapped audio events, AT alone is enough. Compare PLSA with PLSA-AT, it can be seen that PLSA-AT performs better than PLSA on both datasets, which means that when classify overlapped audio events, the classification performance of PLSA can be enhanced after combining it with AT. In summary, AT has the ability to well explore the authorship information of overlapped audio segments, and then has the biggest advantage in classifying overlaps. On both datasets, the performance of ISO-CLUSTER is worse than that of the baseline system, which does not agree with the experimental results in [19]. It is maybe because that in our experiments, the overlaps of two or more audio events, whether the audio event is speech or non-speech, have all been used, and so the classification situation is more complex than that in [19]. Moreover, the unbalance problem of ISO-CLUSTER in constructing the decision tree would also cause performance degradation.

From Tab. 4 and Tab. 5 it can be seen that on both datasets, except for PLSA, the other 3 proposed systems perform better than the baseline and the ISO-CLUSTER systems from the perspective of all evaluation metrics, and among them, AT performs best. This means that the 3 systems (AT, AT-PLSA, PLSA-AT) proposed to classify overlaps can also better classify isolated audio events, and AT not only has the biggest advantage in classifying overlapped audio events, but also has the biggest advantage in classifying isolated audio events. PLSA performs a little worse, but its performance can be enhanced by combining it with AT (see the performance of PLSA-AT).

From Tab. 6 and Tab. 7 it can be seen that for dataset BOB, comparing the baseline and the ISO-CLUSTER systems with our proposed 4 systems, the baseline and the ISO-CLUSTER systems perform much better from the perspective of AEER and Pre, and AT, AT-PLSA and PLSA-AT perform better from the perspective of Rec and F1. For dataset Friends, the proposed 4 systems all perform better than the baseline and the ISO-CLUSTER systems from the perspective of all evaluation metrics. In summary, except for PLSA on dataset BOB, the overall classification performances of our proposed systems are much better than that of the baseline and the ISO-CLUSTER systems. Among the 4 proposed systems, AT performs best; PLSA performs worst, but its performance can be enhanced through combining it with AT (see the performance of PLSA-AT), and after combination, the resulting PLSA-AT performs similarly to AT-PLSA.

4.5 Testing on the Overlaps in Training Set

Testing on the overlaps in training set means that the types of overlap being tested have ever appeared in the training set. In practical application, we would try to collect for the training set as many types of overlap as possible, in case that they would appear in the test set. If a type of overlap has been collected for training, we hope that once it appears in the test set, the system would recognize it. In this section we will test the ability of the proposed systems in recognizing the types of overlap that have ever appeared in the training set. For the leave-one-out cross validation in test stage, each time, from the test set, the types of overlap that have ever appeared in the training set are chosen for testing. The classification results are shown in Tab. 8 and Tab. 9.

Metrics / Systems	AEER	Pre	Rec	F1
baseline	1.079	0.614	0.307	0.409
ISO-CLUSTER	1.175	0.550	0.275	0.367
AT	1.618	0.461	0.461	0.461
PLSA	1.961	0.346	0.346	0.346
AT-PLSA	1.628	0.457	0.457	0.457
PLSA-AT	1.628	0.457	0.457	0.457

Tab. 8. The classification performances of the systems on dataset BOB when classify the overlaps in training set.

Metrics / Systems	AEER	Pre	Rec	F1
baseline	0.584	0.982	0.482	0.646
ISO-CLUSTER	0.527	0.987	0.485	0.650
AT	0.355	0.954	0.706	0.815
PLSA	0.424	0.944	0.661	0.777
AT-PLSA	0.358	0.954	0.702	0.811
PLSA-AT	0.374	0.951	0.699	0.805

Tab. 9. The classification performances of the systems on dataset Friends when classify the overlaps in training set.

From Tab. 8 it can be seen that when classify the overlaps in training set for dataset BOB, except for PLSA, the other 3 proposed systems perform better than the baseline and the ISO-CLUSTER systems from the perspective of Rec and F1, but much worse from the perspective of AEER and Pre. Among the 4 proposed systems, PLSA performs worst, while the other 3 systems perform similarly. From Tab. 9 it can be seen that when classify the overlaps in training set for dataset Friends, the 4 proposed systems perform better than the baseline and the ISO-CLUSTER systems from the perspective of AEER, Rec and F1, but a little worse from the perspective of Pre. Once again, among the 4 proposed systems, PLSA performs worst, while the other 3 systems perform similarly. In summary, AT, AT-PLSA and PLSA-AT have the similar ability in classifying the overlaps in training set, while PLSA is relatively not good at classifying the overlaps in training set.

4.6 Testing on the Overlaps Out of Training Set

Testing on the overlaps out of training set means that the types of overlap being tested have never appeared in the training set, but the audio events in such overlaps have ever appeared in the training set in the form of isolated ones or in the form of other types of overlap. In practical application, though we would try our best to collect for the training set as many types of overlap as possible, there is always the case that the type of overlap being tested has never appeared in the training set. Because in real life, there would be many audio events in an audio document, and the number of combinations of them, that is, the number of types of overlap, would be very large, and then collecting all types of overlap for the training set would be unrealistic. In such case, it is very important for the system to have the ability of recognizing the overlaps out of training set.

In this section we will test the ability of the proposed systems in recognizing the types of overlap that have never appeared in the training set. For the leave-one-out cross validation in test stage, each time, from the test set, the types of overlap that have never appeared in the training set are chosen for testing. The classification results are shown in Tab. 10 and Tab. 11.

Metrics / Systems	AEER	Pre	Rec	F1
baseline	1.071	0.643	0.277	0.384
ISO-CLUSTER	1.155	0.571	0.250	0.345
AT	1.429	0.512	0.460	0.481
PLSA	1.309	0.560	0.500	0.524
AT-PLSA	1.392	0.524	0.472	0.493
PLSA-AT	1.309	0.560	0.500	0.524

Tab. 10. The classification performances of the systems on dataset BOB when classify the overlaps out of training set.

Metrics / Systems	AEER	Pre	Rec	F1
baseline	0.828	0.786	0.384	0.516
ISO-CLUSTER	0.856	0.758	0.375	0.504
AT	0.861	0.833	0.472	0.583
PLSA	0.611	0.917	0.556	0.667
AT-PLSA	0.861	0.833	0.472	0.583
PLSA-AT	0.694	0.917	0.472	0.611

Tab. 11. The classification performances of the systems on dataset Friends when classify the overlaps out of training set.

From Tab. 10 and Tab. 11 it can be seen that generally the 4 proposed systems perform better than the baseline and the ISO-CLUSTER systems when classify the overlaps out of training set. Among the 4 proposed systems, PLSA performs best, which means that though PLSA is relatively not good at classifying the overlaps in training set (considering its performance in Tab. 8 and Tab. 9), it is expert in classifying the overlaps out of training set.

In real life, if the number of audio events in an audio document is large, then the case of overlaps would be very complex. There would be many types of overlap, and it is very likely that one type of overlap to be classified has never appeared in the training set, and so the classification performance of overlaps out of training set is an important

performance indicator to evaluate the classification system. Our proposed systems can well classify the overlaps out of training set as long as the audio events in such overlaps have ever appeared in the training set, no matter in the form of isolated ones or in the form of other types of overlap. This indicates that based on the AT model and the PLSA model, the proposed systems can well discover the latent semantic structure of the audio corpus, and so when a new type of overlap appears, though it has never appeared in the training set, based on the latent semantic similarity, the proposed systems can still recognize it.

5. Conclusions and Future Work

In this paper, we focus on the audio overlap classification problem which is a big challenge in audio classification field. Inspired by AT and PLSA, both of which are first proposed for text analysis, we propose 4 systems, i.e. AT, PLSA, AT-PLSA and PLSA-AT, to resolve it. Compared with the baseline and the ISO-CLUSTER systems, the proposed systems have the following advantages: generally, they work better not only in classifying overlapped and isolated audio events, but also in classifying the types of overlap in and out of training set; they have the ability to recognize two or more audio events in an overlap, which cannot be done by the baseline and the ISO-CLUSTER systems.

Audio event classification is a more controlled task, while audio event detection is more realistic in real applications. In the future, more work will be done to try to expand the proposed systems into the general audio event detection problem. One direction is that for each frame of the audio document, one of the two models AT and PLSA is used to find out the active audio events, and the other is used to confirm the result; then some post-processing, such as smoothing, will be performed to improve the detection result.

Acknowledgments

This work has been jointly supported by the Project of National Natural Science Foundation of China (No. 61401259, No. 61362031, No. 61471226, No. 61201441, No. 61401258), Research Fund for Excellent Young and Middle-aged Scientists of Shandong Province (No. BS2013DX035, No. BS2012DX038,), Science and Technology Development Plan of Shandong Province (No. 2013GGX10113), and Natural Science Foundation of Shandong Province (ZR2013FQ019, ZR2012FQ015).

References

[1] LENG, Y., QI, G. H., XU, X. Y. A BIC based initial training set selection algorithm for active learning and its application in audio detection. *Radioengineering*, 2013, vol. 22, no. 2, p. 638–649. ISSN:1210-2512

[2] SCHIMMEL, J. Audible aliasing distortion in digital audio synthesis. *Radioengineering*, 2012, vol. 21, no. 1, p. 56–62.

[3] GREZL, F., CERNOCKY, J. Audio surveillance through known event classification. *Radioengineering*, 2009, vol. 18, no. 4, p. 671 to 675. ISSN:1210-2512

[4] HU, P., LIU, W., JIANG, W., YANG, Z. Latent topic model for audio retrieval. *Pattern Recognition*, 2014, vol. 47, no. 3, p. 1138 to 1143. DOI: 10.1016/j.patcog.2013.06.010

[5] GHORAANI B., KRISHNAN, S. Time-frequency matrix feature extraction and classification of environmental audio signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 2011, vol. 19, no. 7, p. 2197–2209. DOI: 10.1109/TASL.2011.2118753

[6] DUAN, S., ZHANG, J., ROE, P., ET AL. A survey of tagging techniques for music, speech and environmental sound. *Artificial Intelligence Review*, 2012, 25 p. DOI: 10.1007/s10462-012-9362-y

[7] YAMAKAWA, N., TAKAHASHI, T., KITAHARA, T., ET AL. Environmental sound recognition for robot audition using matching-pursuit. In *Modern Approaches in Applied Intelligence*. Springer, Berlin Heidelberg, 2011, p. 1-10. DOI: 10.1007/978-3-642-21827-9_1

[8] VACHER, M., FLEURY, A. PORTET, F., ET AL. Complete sound and speech recognition system for health smart homes: Application to the recognition of activities of daily living. *New Developments in Biomedical Engineering*, 2010. p. 645–673.

[9] CLEAR, 2007. Classification of events, activities and relationships. Evaluation and Workshop. Available at: http://www.clear-evaluation.org/. Acoustic event detection evaluation plan, Available at: http://isl.ira.uka.de/clear07/?download=CLEAR_2007_AED_EvaluationPlan.pdf.

[10] HEITTOLA, T., MESAROS, A., ERONEN, A., ET AL. Context-dependent sound event detection. *EURASIP Journal on Audio, Speech, and Music Processing*, 2013, no. 1, p. 1–13. DOI: 10.1186/1687-4722-2013-1

[11] ROSEN-ZVI, M., CHEMUDUGUNTA, C., GRIFFITHS, T., ET AL. Learning author-topic models from text corpora. *ACM Transactions on Information Systems*, 2010, vol. 28, no. 1, p. 1–38. DOI: 10.1145/1658377.1658381

[12] HOFMANN, T. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 2001, vol. 42, no. 1-2, p. 177–196. DOI: 10.1023/A:1007617005950

[13] BLEI, D. M., NG, A. Y., JORDAN, M. I. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 2003, vol. 3, p. 993–1022.

[14] KIM, S., NARAYANAN, S., SUNDARAM, S. Acoustic topic model for audio information retrieval. In *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. 2009, p. 37–40. DOI: 10.1109/ASPAA.2009.5346483

[15] IMOTO, K., OHISHI, Y., UEMATSU, H., OHMURO, H. Acoustic scene analysis based on latent acoustic topic and event allocation. In *Proceedings of IEEE Workshop on Machine Learning for Signal Processing (MLSP)*. 2013, p. 1–6. DOI: 10.1109/MLSP.2013.6661957

[16] PENG, Y. X., LU, Z. W., XIAO, J. G. Semantic concept annotation based on audio PLSA model. In *Proceedings of ACM Multimedia Conf.* 2009, p. 841–844. DOI: 10.1145/1631272.1631428

[17] LEE, K., ELLIS, D. P. W. Audio-based semantic concept classification for consumer video. *IEEE Transactions on Audio, Speech and Language Processing*, 2010, vol. 18, no. 6, p. 1406-1416. DOI: 10.1109/TASL.2009.2034776

[18] HAZEN, T. J. Latent topic modeling for audio corpus summarization. In *Proceedings of INTERSPEECH*. 2011, p. 913–916. ISBN:978-1-61839-270-1

[19] TEMKO, A., NADEU, C. Acoustic event detection in meeting-room environments. *Pattern Recognition Letters*, 2009, vol. 30, no. 14, p. 1281–1288. DOI: 10.1016/j.patrec.2009.06.009

[20] ESPI, M., FUJIMOTO, M., KUBO, Y., NAKATANI, T. Spectrogram patch based acoustic event detection and classification in speech overlapping conditions. In *Proceedings of 4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*. 2014, p. 117–121. DOI: 10.1109/ HSCMA. 2014.6843263

[21] GEMMEKE, J. F., VUEGEN, L., VANRUMSTE, B., VAN-HAMME, H. An exemplar-based NMF approach for audio event detection. In *Proceedings of IEEE Workshop on Application of Signal Processing to Audio and Acoustics (WASPAA)*, 2013, p. 1–4. DOI: 10.1109/WASPAA.2013.6701847

[22] DIMENT, A., HEITTOLA, T., VIRTANEN, T. Sound event detection for office live and office synthetic AASP challenge. *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events,* 2013.

[23] ZHUANG, X., ZHOU, X., HASEGAWA-JOHNSON, M. A., HUANG, T. S. Real-world acoustic event detection. *Pattern Recognition Letters*, 2010, vol. 31, no. 12, p. 1543–1551. DOI: 10.1016/j.patrec.2010.02.005

[24] VUEGEN, L., VAN DEN BROECK, B., KARSMAKERS, P., ET AL. An MFCC-GMM approach for event detection and classification. *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events,* 2013.

[25] DIKMEN, O., MESAROS, A. Sound event detection using non-negative dictionaries learned from annotated overlapping events, In *Proceedings of IEEE Workshop on Application of Signal Processing to Audio and Acoustics (WASPAA)*, 2013, p. 1–4. DOI: 10.1109/WASPAA.2013.6701861

About the Authors ...

Yan LENG was born in Yantai, China, in 1981. She received both the B.S. degree and the M.S. degree from Shandong University (SDU), Ji'nan, China in 2003 and 2006 respectively, and received the Ph.D. degree from Beijing University of Posts and Telecommunications (BUPT), Beijing, China in 2012. Now she works as a lecturer at the College of Physics and Electronics, Shandong Normal University in Ji'nan, China. Her research interests include audio classification, audio detection, audio retrieval and medical image processing.

Chengli SUN was born in Hebei, China, in 1975. He received the B.S. degree in electronics engineering from Zhongbei University, Taiyuan, China, in 1999, and the Ph.D. degree in signal and information processing from Beijing University of Posts and Telecommunication (BUPT), Beijing, China, in 2008. Now he works as an associate professor at the Information School of Nanchang Aeronautical University in Nanchang, China. His current research interests include acoustic event detection, speech recognition, speech enhancement, and biomedical image analysis.

Chuanfu CHENG was born in Ji'nan, China, in 1962. He received the B.S. degree from Shandong Normal Unversity in 1982, and the M. S. degree in Suzhou University in 1988. In 2004, he received his Ph. D. degree from Shanghai Institute of Optics and Fine Mechanics, Chinese Academy of Sciences in Shanghai, China. Now he is a professor in the College of Physics and Electronics, Shandong Normal University in Ji'nan, China. His research interests are laser speckles, light scattering of random surfaces, near-field optics and plasmonics of metal film nano-structures.

Xinyan XU was born in Ji'nan, China, in 1962. She received the M.E. degree from Shandong University (SDU), Ji'nan, China in 2006. Her research interests include audio classification and medical image processing.

Si LI was born in Chifeng, China, in 1985. She received both the B.S. degree and the Ph.D. degree from Beijing University of Posts and Telecommunications (BUPT), Beijing, China in 2007 and 2012 respectively. Now she works as a postdoctoral fellow at the Dept. of Computer Science, Brandeis University in Waltham, Massachusetts, USA. Her research interests include natural language processing, information retrieval and information extraction.

Honglin WAN was born in Dezhou, China, in 1979. He received the Ph.D. degree from Shandong University (SDU),Ji'nan, China in 2008. Now he works as a lecturer at the College of Physics and Electronics, Shandong Normal University in Ji'nan, China. Her research interests include signal processing and biometrics.

Jing FANG was born in Liaocheng, China, in 1980. She received the B.S. degree from Shandong Normal University, Ji'nan, China in 2002, and the M.S. degree from Beijing Jiaotong University (BJTU), Beijing, China in 2005. Now she works as a lecturer at the College of Physics and Electronics, Shandong Normal University in Ji'nan, China. Her research interests include image enhancement and restoration, pattern recognition, and medical image processing.

Dengwang LI (corresponding author) was born in Shanxi, China, in 1983. He received both B.S. degree and Ph.D. degree from Shandong University, Ji'nan, China in 2006 and 2011 respectively. Now he works as an associate professor at the College of Physics and Electronics, Shandong Normal University in Ji'nan, China. His research interests include signal processing and medical image processing.