Multimedia Systems (2015) 21:507–524
DOI 10.1007/s00530-014-0424-7

ORIGINAL ARTICLE

Context-based environmental audio event recognition for scene understanding

Tong Lu · Gongyou Wang · Feng Su

Received: 7 November 2013 / Accepted: 22 September 2014 / Published online: 9 October 2014

Abstract Automatic audio content recognition has attracted an increasing attention for developing multimedia systems, for which the most popular approaches combine frame-based features with statistic models or discriminative classifiers. The existing methods are effective for clean single-source event detection but may not perform well for unstructured environmental sounds, which have a broad noise-like flat spectrum and a diverse variety of compositions. We present an automatic acoustic scene understanding framework that detects audio events through two hierarchies, *acoustic scene recognition* and *audio event recognition*, in which the former is preceded by following dominant audio sources and in turn helps infer non-dominant audio events within the same scene through modeling their occurrence correlations. On the scene recognition hierarchy, we perform adaptive segmentation and feature extraction for every input acoustic scene stream through Eigen-audiospace and an optimized feature subspace, respectively. After filtering background, scene streams are recognized by modeling the observation density of dominant features using a two-level hidden Markov model. On the audio event recognition hierarchy, scene knowledge is characterized by an audio context model that essentially describes the occurrence correlations of dominant and non-dominant audio events within this scene. Monte Carlo integration and gradient descent techniques are employed to maximize the likelihood and correctly tag each audio event. To the best of our knowledge, this is the first work that models event correlations as scene context for robust audio event detection from complex and noisy environments. Note that according to the recent report, the mean accuracy for the acoustic scene classification task by human listeners is only around 71 % on the data collected in office environments from the DCASE dataset. None of the existing methods performs well on all scene categories and the average accuracy of the best performances of the recent 11 methods is 53.8 %. The proposed method averagely achieves an accuracy of 62.3 % on the same dataset. Additionally, we create a 10-CASE dataset by manually collecting 5,250 audio clips of 10 scene types and 21 event categories. Our experimental results on 10-CASE show that the proposed method averagely achieves the enhanced performance of 78.3 %, and the average accuracy of audio event recognition can be effectively improved by capturing dominant audio sources and reasoning non-dominant events from the dominant ones through acoustic context modeling. In the future work, exploring the interactions between acoustic scene recognition and audio event detection, and incorporating other modalities to improve the accuracy are required to further advance the proposed framework.

Keywords Acoustic scene · Audio event · Context modeling · Recognition · 10-CASE

1 Introduction

Recently, the rapid increase in speed and capacity of smart embedded devices equipped with acoustic sensors and powerful CPUs has allowed the inclusion of audio as a useful type of data in computing scene understanding tasks, especially in robotics, context-aware systems and

Communicated by M. Kankanhalli.

T. Lu (✉) · G. Wang · F. Su
National Key Laboratory for Novel Software Technology,
Department of Computer Science and Technology,
Nanjing University, Nanjing, China
e-mail: lutong@nju.edu.cn

various mobile applications. Comparing with vision-based approaches in these systems, audio signals are robust in situations from visual obstruction, weak lighting to multi-view observation. Moreover, when a vision-based application is employed to determine interior or exterior environments, its robustness or utility will be lost if visual information is compromised or totally absent.

Developing intelligent audio scene understanding systems in the same way that humans do has attracted an increasing attention in multimedia community over the past few years. For example, audio events such as *coughing*, *fall-down-stairs* and *collapsing* are of particular interest in developing daily healthcare systems [1] for the coming aged society, especially considering the fact that audio is relatively cheap in computing and can be exploited either independently or as an important component in audiovisual processing [2, 3]. Actually, the extraction of an accurate perception of everyday audio events presented in natural environments has revealed strong commercial and social demands in developing various multimedia systems such as content-based multimedia data indexing and retrieval [4], security surveillance [5], bioacoustic monitoring [6] and military applications [7]. Moreover, the knowledge about a surrounding environment does help in inspiring new forms of multimedia services or the augmentation to the existing ones, for example, a mobile phone can provide personalized music recommendations [8, 9] after the change of user's circumstance is automatically perceived.

A variety of "machine listening" systems has been developed [10] to yield audio events from speech [11, 12] or music [13, 14]. Over the past decade, the most popular approaches are frame-based features such as Mel-frequency cepstral coefficients (MFCC) and MPEG-7 descriptors, which can be modeled with support vector machine (SVM), hidden Markov models (HMMs) or Gaussian mixture models (GMMs) for discriminative classification and recognition. These methods are effective in recognizing events from a clean single source, but face difficulties in handling the challenging environmental sounds that often have unpredicted conditions [15]. Systems have been presented to overcome the problem to an extent [16]; however, there is still a major challenge in separating the structured signal that has formantic or harmonic characteristics from unstructured background noises [17]. That is, to recognize audio events from complex and noisy environments, most current systems face the following two problems. First, it is difficult to build models for an unstructured environmental sound stream, which has a broad noise-like flat spectrum and a diverse variety of signal compositions. Thus, no assumptions can be made in advance about the harmonic structure in the stream. Second, how to detect environmental audio events accurately and robustly is still admittedly a hard problem since acoustic scenes are composed of sounds from a variety of sources. Due to such unpredicted diverse nature of audio data, it is relatively difficult to describe and quantify environmental audio signals. Therefore, most methods are far from the level of human performance in recognizing scene events from a daily environment [18], and automatic acoustic scene content understanding is still at its early stage.

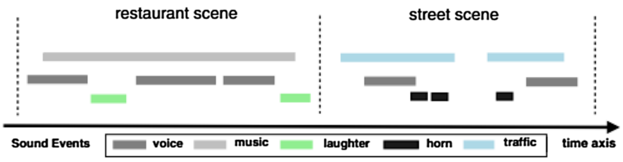


Fig. 1 Illustration of two acoustic scenes and the audio events within them

In this paper, we propose a novel audio event recognition framework for acoustic scene understanding based on our previous work on sound classification [3, 19], audio summarization [20, 21] and audio–visual correlation [4, 22]. The term auditory scene here refers to the acoustic modeling of a specific location or site such as *home*, *bus station*, *restaurant* and *shopping mall*, which is similar to what an image of the same location provides visually. Acoustic scene understanding is thus separated into two hierarchies consisting of acoustic scene recognition and audio event recognition. The two hierarchies are necessary due to the fact that each acoustic scene often contains or is characterized by various audio events occurring in it. That is, the recognition of meaningful sound events that have distinct acoustic patterns relies on the knowledge of a particular scene category where the events occur. For example, after knowing an input stream is from a *home* scene, an aged daily life assistant system can pay particular efforts on detecting the events such as *coughing* and *fall-down-stairs* that probably occur, simultaneously filtering the sounds like *horning* and *traffic* that are less important for this scene category. Theoretically, this kind of scene knowledge can be properly modeled as scene contexts that do help in detecting meaningful audio events from complex and noisy real-life environments. Figure 1 shows two acoustic scene examples and their audio events, which will be respectively recognized by the two hierarchies in sequence.

Inspired by the discussions, on the scene recognition hierarchy, we perform adaptive analysis for every input acoustic scene stream, respectively, through two spaces, namely, Eigen-audiospace and an optimized feature subspace. In Eigen-audiospace, all the audio change boundaries are detected to perform adaptive segmentations, aiming at avoiding the intermission of independent audio events in the acoustic scene. Then in the feature subspace, a binary wavelet packets tree is constructed for each audio segment, and the best local discriminatory base (LDB) feature subspace is maximized by a particular discriminant measure among classes to initialize the semantic, namely, non-stationary background or meaningful foreground, for every segment. After filtering background segments, acoustic scene streams are finally recognized by modeling the observation density of the LDB–MFCC features using the mixture of Gaussian in HMM.

On the audio event recognition hierarchy, scene knowledge is characterized by an audio context model that essentially describes the occurrence correlations of all the audio events that probably occur in this scene. That is, by knowing scene category through the scene recognition hierarchy, the coupled occurrences of all the training audio events appear in this scene can be characterized as a contextual model to help predict the semantic of each unknown audio event from the same scene. This is true due to that each audio event within a real-world acoustic scene tends to correlate to the others, namely, audio events within the same scene can take a complementary role in improving the accuracy of audio event recognition, and the presence of one audio event probably helps identifying the existence of others. Monte Carlo integration and gradient descent techniques are employed to maximize the likelihood and correctly recognize each audio event. Diverse variety of unstructured sound stream can thus be analyzed more accurately.

The two main contributions of the paper are:

- The introduction to acoustic scene context modeling, which successfully captures the occurrence correlations among the audio events within the scene, is proposed for more accurate audio event recognition. By this way, acoustic scene understanding is systematically explored through two hierarchies consisting of acoustic scene recognition and audio event recognition. Scene recognition is thus preceded by following dominant audio sources, and in turn helps reasoning non-dominant audio events within it through scene context modeling, which plays an important role in recognizing complex, noisy or even overlapped sound events by capturing the occurrence correlations between dominant and non-dominant sounds to infer the semantics of the latter. To the best of our knowledge, this is the first work that explores audio event recognition from unstructured environments by modeling event occurrence correlations as acoustic scene contexts.
- A benchmark dataset for evaluating acoustic scene understanding performance is created, which contains a manually selected list of totally 5,250 audio clips comprising 10 categories of indoor/outdoor scenes and 21 audio event types. The experimental results on our proposed dataset, another recent benchmark dataset DCASE [23] and public TV-Movies show the effectiveness of the proposed method.
- The rest of the paper is organized as follows. Section 2 discusses the related work. In Sect. 3, we describe the scene recognition method. Section 4 gives the details of the proposed audio context model for event detection. Experimental results are given in Sect. 5, and finally Sect. 6 highlights the discussions and concludes the method.

2 Related work

Research on general audio signal analysis has received long time interests in the past years and a variety of audio features or processing models have been proposed in temporal or spectral domain. However, researches on audio event recognition from unstructured environments are relatively less. The leading approaches have been investigated to extract environmental audio contents can be roughly classified into two categories: *acoustic feature selection* and *context-aware analysis*.

Automatic recognition of auditory environment by an *acoustic feature selection* strategy is known from many earlier works. In general, the process of auditory scene understanding is very similar regardless of the sensors or the data sources used for recognition. The temporal or spectral domain feature vectors obtained from sensors are fed to classifiers that try to identify the location or the environment the particular feature vectors present. For example, Eronen et al. [24] propose an approach to recognize everyday scenes with the popular MFCC, which have been shown to work well for structured sounds. However, the performance of MFCC probably degrades in the existence of noise due to that it is not effective in analyzing noise-like signals that have a flat spectrum. Aleh et al. [25] thus classify five environmental noise classes (car, street, babble, factory and bus) using line spectral features and a Gaussian classifier. Couvreur et al. [26] use linear prediction cepstral coefficients and discrete HMMs to recognize five types of environmental events, namely, car, truck, moped, aircraft and train. The choice of proper signal streams is likely helpful to find the discriminations among different environments; however, due to the diverse nature of acoustic scenes, it is still difficult in selecting a proper feature to build robust scene understanding systems.

The combination of several acoustic features seems to be another choice. Scheirer and Slaney [27] use a combination of several features to describe an audio discrimination system. Eronen et al. [24] develops a system to evaluate the extraction of audio features (ZCR, MFCC, Band-energy, Spectral flux, etc.) and feature transforms (PCA, ICA, LDA) on 24 audio contexts. In [28], the performances of three new types of transforms, chirplet, curvelet, and Hilbert transforms, are investigated for environmental audio classification. Ghoraani and Krishnan [29] construct the time–frequency matrix (TFM) of audio signals and apply the non-negative matrix decomposition to decompose TFM into its significant components. As a result, they propose seven new features from spectral and temporal structures of the decomposed vectors in a way that they successfully represent joint time–frequency (TF) structure of the audio signal, and combine them with the MFCCs features. Unfortunately, adding more features is not always helpful as illustrated in [19], where the use of all features does not always produce good performance for the audio classification problem. It is probably due to the fact that as the feature dimension increases, data points become sparser and potentially irrelevant features could negatively impact the classification result.

It in turns leads to another strategy of choosing an optimal subset of features from a larger set of possible features to yield more accurate and the most effective subset. Umapathy et al. [30] propose a time–frequency approach for audio classification, which is considered the best way to analyze audio signals non-stationary in nature. In their method, audio signals are decomposed using an adaptive TF decomposition algorithm to generate a set of 42 features over three frequency bands within the auditory range. These features are analyzed using linear discriminant functions and classified into six groups. In their succeeding work [31], they further propose an audio feature extraction and a multigroup classification scheme that focuses on identifying discriminatory time–frequency subspaces using the LDB technique. Chu et al. [32] propose to use the matching pursuit algorithm to select effective time–frequency features as the supplementation to MFCC for audio environment characterization and general acoustic scene types are represented as a whole rather than collections of discrete audio events pre-extracted. Their process includes finding the decomposition of a signal from a dictionary of atoms, which would yield the best set of functions to form an approximate representation. More recently, Mäkinen et al. [33] propose an evolutionary feature synthesis technique to enhance common audio descriptors by using multi-dimensional particle swarm optimization to search for the optimal feature synthesis parameters. These works forward the auditory scene research; however, not all the unpredictable structures of auditory environments can be directly discriminated by a low-level feature subset.

Context-aware approach is still at its early stage comparing with the acoustic feature selection methods to the best of our knowledge. Generally, context in an auditory scene can be considered as two different levels towards incorporating hints for scene understanding. The first level characterizes the correlations of certain audio events inside the same acoustic scene such as gunfire and the accompanied cries in a war, or ball-hit and the corresponding applause in a basketball game. Niessen et al. [34] model knowledge and context in audio recognition by investigating the role of dynamic network model to improve automatic audio identification and simultaneously reduce the search space of low-level audio features. Heittola et al. [35] present each audio context using a histogram of audio events which are detected using a supervised classifier based on annotated recordings. In their further work [36], they provide context rule knowledge to better describe the search space. For example, it will determine excluding the footsteps class when the tested recording is from inside a car. However, the contexts are modeled by three-state left-to-right hidden Markov parameters and thereby still face difficulties because of the great number of possible event combinations and the transitions among them. The second context-aware level describes more general information about the surroundings around an audio device in spite of the location, such as time [37], weather [38], running pace [39] and even user-dependent states like emotion [40] or physiological state [37]. The advent of smart mobile phones with rich sensing capabilities is making real-time context information collecting and exploration a possibility, and new auditory scene recognition systems are expected in the next few years.

Both the two categories of methods have to consider their dependency on selecting a proper learning algorithm to obtain their recognition results. Mirikitani and Nikolaev [41] recursively train recurrent neural networks for improved time-series modeling. In [42], a set of key audio effects are modeled with HMMs and a Bayesian network-based approach is proposed to discover the high-level semantics of an auditory context embedded in key effect sequences. Recently, the combination of different classifiers has been proved effective in environmental audio understanding. For example, in [43], a hybrid SVM/k-nearest neighbor (kNN) classifier is used for environmental audio classification based on MPEG-7 audio low-level descriptors. Räsänen et al. [44] study the combination of classifier output distributions using a number of different classifiers instead of performing audio context fusion at a feature level. Kinnunen et al. [45] choose MFCC parameterization and present an extensive comparison of six different classifiers of kNN, vector quantization (VQ), Gaussian mixture model trained with both maximum likelihood and maximum mutual information (GMM-MMI) criteria, GMM supervector support vector machine and SVM with generalized linear discriminant sequence for auditory scene recognition. They find GMM-MMI and VQ classifiers perform the best identification rates. On the other side, multimodality algorithms such as probabilistic models to integrate voice and visual identification cues gained from microphones and cameras in a smart environment have also been proposed [4, 46], which do help in automatic acoustic

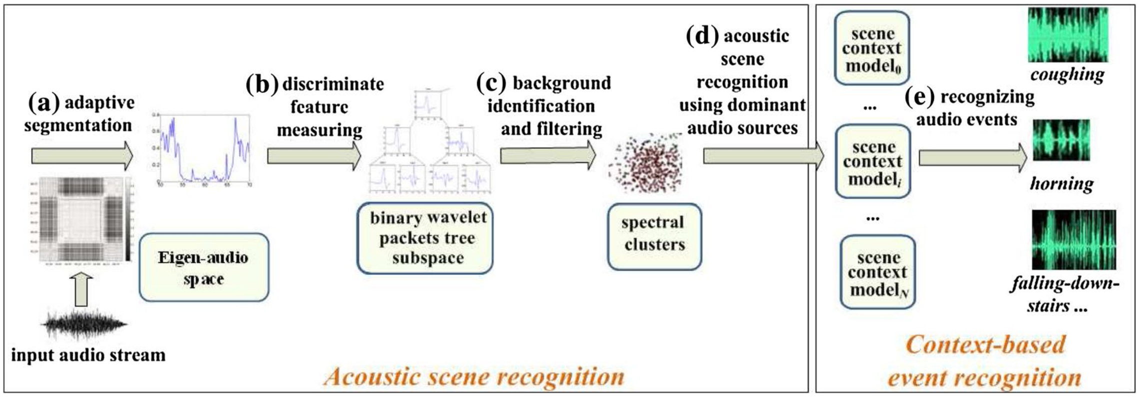

(a) adaptive segmentation
input audio stream
Eigen-audio space
(b) discriminate feature measuring
binary wavelet packets tree subspace
(c) background identification and filtering
spectral clusters
(d) acoustic scene recognition using dominant audio sources
Acoustic scene recognition
scene context model$_0$
...
scene context model$_i$
...
scene context model$_N$
(e) recognizing audio events
coughing
horning
falling-down-stairs ...
Context-based event recognition

Fig. 2 Overall framework of the presented approach for meaningful event recognition from an environmental audio stream: *a* we first locate audio changes in the Eigen-audiospace to perform adaptive segmentation on the input audio stream, *b* we maximize a particular discriminant measure among classes in the LDB feature subspace, *c* background audio noises are filtered after spectral clustering, *d* acoustic scenes are recognized using dominant audio sources using HMM and the discriminant features, and *e* audio events are detected using the proposed context model by reasoning non-dominant audio sources

understanding. For example, the latent semantic analysis derived from language processing is proposed as an interesting solution to learn overlapped audio events [47]. Lu et al. [22] propose a multimodal correlation network, in which audio-to-audio retrievals can be improved by incorporating visual image information. However, this area requires more studying to apply the techniques efficiently into auditory scene understanding.

3 Our approach

Our approach consists of two hierarchies: *acoustic scene classification* and *context-based audio event recognition*, which are shown in the overview framework of Fig. 2. On the scene recognition hierarchy, the main problem comes from how to describe the diverse nature of environmental audio data for identifying the category of the input acoustic scene. The following two stages are used to solve this problem:

- Locating audio changes in the Eigen-audiospace to perform adaptive segmentation (Fig. 2a, Sect. 3.1) on the input audio stream. This is necessary since a fixed-size slide window may cause intermission of independent audio events within the audio stream, and thereby probably increases the complexity of succeeding semantic analysis.
- By maximizing a particular discriminant measure among classes in the LDB feature subspace (Fig. 2b, Sect. 3.2.1) and filtering background audio noises after spectral clustering (Fig. 2c, Sect. 3.2.2), the acoustic scene stream is recognized by modeling HMM of the scene on the discriminant features (Fig. 2d, Sect. 3.2.3).
- After knowing the category of the acoustic scene, the meaningful audio events that appear in it will be recognized by an audio context model (Fig. 2e, Sect. 4), which captures the occurrence correlations among the audio events within the scene during the training stage. The context model is learned by employing Monte Carlo integration and gradient descent techniques to maximize the likelihood, thus allowing correctly tagging unknown audio events from the same scene category during testing.

3.1 Adaptive audio segmentation

Feature extraction through a fixed-size window may cause intermission of meaningful audio events. This is especially true for the variant durations of miscellaneous audio events in a complex audio environment. Considering the fact that for an input acoustic scene stream, the structured sounds in it most probably indicate meaningful audio events while unstructured compositions are relatively reentrant, we first project the input scene stream into the Eigen space and accordingly search for the unstructured sounds to adaptively segment it.

For a scene stream S, let $\mathbf{M}_S$ denote its $N_p \times M_{N_p}$ sampling matrix whose columns are M_{N_p} zero-mean non-overlapping audioframes. In $\mathbf{M}_S$, each column x_i has $N_p = 2{,}205$ samples obtained by sampling an audioframe of duration 0.2 s at the rate of 11.025 kHz. The sound or the event intermission that is shorter than this threshold is generally meaningless to humans in real-life applications. Then, we apply the principal component analysis (PCA) transformation to $\mathbf{M}_S$ and correspondingly obtain a projection matrix by sequentially selecting the smallest K_S eigenvalues in the Eigen-audiospace, denoted by $\mathbf{E}_S$ of $N_p \times K_S$. The sampling matrix $\mathbf{M}_S$ can be projected to the Eigen space with the rank K_S by

$$\mathbf{E}_{\mathbf{M_S}} = \left\{ \left(E'_{M_S}\right)_1, \left(E'_{M_S}\right)_2, \dots \left(E'_{M_S}\right)_i, \dots \right\} = \mathbf{E}_S^T \mathbf{M_S} \quad (1)$$

Next, we search for the audioframe with the minimum L2 norm of sampling points through a sliding window along the audio stream S and denote it as x_S^b, which is zero-mean after subtracting the mean audioframe of S. x_S^b is then projected to the Eigen-audiospace by $x_S^{b'} = \mathbf{E_S^T} x_S^b$, which is used as the reference candidate audio in the Eigen-audiospace. Accordingly, we measure the normalized distance between $x_S^{b'}$ and any projected audioframe, and segment the input audio stream based on the distances by

$$\mathrm{NDis}(i) = \frac{\left\| (E'_{M_S})_i - x_S^{b'} \right\| - \min_S}{\max_S - \min_S} \quad (2)$$

where

$$\min_S = \arg\min_{j \in [1, M_{N_p}]} \left\| (E'_{M_S})_j - x_S^{b'} \right\| \quad (3)$$

$$\max_S = \arg\max_{j \in [1, M_{N_p}]} \left\| (E'_{M_S})_j - x_S^{b'} \right\|. \quad (4)$$

We finally search for all the valley points on the calculated $\mathrm{NDis}(i)$ to obtain all the adaptive segmentation results $\{s_1, s_2, \dots s_i, \dots, s_{N'}\}$ from the input audio stream S.

3.2 Acoustic scene recognition

Generally, there are three stages for effectively recognizing an acoustic scene stream:

- *Extracting discriminant features for every audio segment.* It requires a discriminant audio feature descriptor to quantify the sampling points that are probably from different audio sources such as *speech*, *laughter*, *music* and *applause*.
- *Selecting dominant feature vectors.* Inspired by [47], it would be easier to understand the semantics of every audio segment by following the dominant source while simultaneously ignoring the others in the segment.
- *Filtering background clusters from the scene stream.* This is necessary because an acoustic scene could be in most cases decided by the dominant source, while background noises usually have a negative effect on the judgment.

3.2.1 Searching for dominant sources in the scene

In this section, we realize the first two considerations as follows. First, for extracting discriminant features, we believe the *best-basis* strategy, specifically the LDB, is a good choice to select the features that minimize entropy or maximize a certain discriminant measure among classes, through obtaining an orthonormal optimum subspace from a large collection of bases. Once such a basis is selected, a small number of the most significant coordinates (features) can be used to enhance the performance of a classifier without losing important details of an environmental audio stream. Similar to [31] which has been successfully applied to various classification problems in signal and image processing, we adopt the following steps to extract our LDB+MFCC features for the audio segments in $S = \{s_1, s_2, \dots s_i, \dots, s_{N'}\}$. We decompose s_i into a binary wavelet packets tree $\left\{\Omega_{j,k}\right\}$, where j indicates the level of the subspace spanned by a set of wavelet packet basis vectors $\left\{\mathbf{w}_{j,k,l}\right\}_{l=0}^{l=2^{u-j}-1}$ (2^u corresponds to the length of s_i). As a result, s_i can be expressed as

$$s_i = \Sigma_{j,k,l}[a_{j,k,l}]_i \cdot \mathbf{w}_{j,k,l} \quad (5)$$

We search for the set of best feature subspaces to provide maximum dissimilarity information between different scene audio classes from all 2^J mutually orthogonal subspaces (J is the maximum number of decomposition levels). This is actually a pruning process on the wavelet packet tree based on the discriminative capability of a wavelet packet node (subspace) (j,k) on P signal classes. Repeating this process for each form of the dissimilarity measurement $\mathrm{DSV}^{(j,k)}$ and combining the resulting nodes, we finally obtain a set of discriminative hybrid descriptors for every segmented environmental audio signal. Theoretically, the combination of the two features would form a better feature fusion in recognizing complex scene audio, in which the MFCC features are performing better for natural sounds, whereas the LDB features perform slightly better for artificial sounds [31].

Second, to concentrate on the dominant audio source while ignoring the others in every audio segment, we apply the spectral clustering algorithm on the extracted features of each audio segment. Let $\mathbf{x}_{ij}$ denote the jth LDB–MFCC feature vector that is extracted from audio segment s_i, s_i can then be represented by the feature matrix of $\mathbf{F}M_{s_i} = \left\{\mathbf{x}_{ij}\right\}_{j=1}^{n_{s_i}}$, where n_{s_i} is the total number of features extracted from s_i. Next, s_i is processed as follows:

1. Construct the similarity matrix $\mathbf{W}_{s_i} \in R^{n_{s_i} \times n_{s_i}}$ from $\mathbf{F}M_{s_i}$ by

   $$\mathbf{W}_{s_i}^{jk} = \exp\left(- \frac{\|\mathbf{x}_{ij} - \mathbf{x}_{ik}\|}{2\mathrm{factor}_{ij}\mathrm{factor}_{ik}} \right), \quad j \neq k, \ \mathbf{W}_{s_i}^{jj} = 0 \quad (6)$$

   where $\mathrm{factor}_{i\cdot}$ is a scale factor to evaluate the average distance from $\mathbf{x}_{i\cdot}$ to its nearest N_P points (N_P is a constant) in the feature space, defined as

   $$\mathrm{factor}_{i\cdot} = \frac{\Sigma_{l|\mathbf{x}_{il} \in \mathrm{nearest}(\mathbf{x}_{i\cdot})} \|\mathbf{x}_{i\cdot} - \mathbf{x}_{il}\|}{\min(N_P, n_{s_i})}. \quad (7)$$

2. Calculate the diagonal matrix for $\mathbf{W}_{s_i}$ as $\mathbf{D}_{jj}(\mathbf{W}_{s_i}) = \sum_{k=1}^{n_{s_i}} \mathbf{W}_{\mathbf{s_i}}^{\mathbf{jk}}$.
3. Decide the clustering number k_{s_i} automatically by calculating the orthogonal Eigen vectors of the matrix $\mathbf{D}^{-1/2}\mathbf{W}_{s_i}\mathbf{D}^{-1/2}$ and its corresponding Eigen values. The clustering number for segment s_i will be calculated by

$$k_{s_i} = \underset{k\in[0,n_{s_i}-1]}{\arg\max} \left(1 - \frac{\lambda_{k+1}}{\lambda_k}\right) \tag{8}$$

where λ_k is an Eigen vector of $\mathbf{D}^{-1/2}\mathbf{W}_{s_i}\mathbf{D}^{-1/2}$.
4. Apply the spectral clustering algorithm to the feature vectors of $\mathbf{F}M_{s_i}$ and correspondingly the result clusters $C_{\mathbf{FM_{s_i}}} = \left\{C_{i_1}, C_{i_2}, \dots C_{i_{k_{s_i}}}\right\}$ can be obtained.

Finally, we calculate the gap of dominant clusters by

$$gap_{s_i} = \arg\max \sum_{j=1}^{k_{s_i}} n_{c_{ij}} \geq 0.8 \times n_{s_i} \tag{9}$$

where $n_{c_{ij}}$ is the total feature number in cluster c_{ij}. Note that the clusters have been sorted according to their feature numbers in $C_{\mathbf{FM_{s_i}}}$. As a result, segment s_i can be finally characterized by averaging the features from the dominant clusters by

$$\mathbf{x}'_{s_i} = \frac{\sum_{j=1}^{\text{gap}_{s_i}} \sum_{k=1}^{n_{c_{ij}}} \mathbf{x_{ik}}}{\sum_{j=1}^{\text{gap}_{s_i}} n_{c_{ij}}} \tag{10}$$

3.2.2 *Identifying and filtering background*

For context-aware processing that this work mainly targets, we want to simulate people's ability to reconstruct an accurate perception of meaningful events presented in a natural environment by computation modeling. Considering audio events in a real-life environment usually correspond to short clusters with distinct acoustic patterns, we further cluster the dominant features from all the audio scene segments and accordingly discriminate the clusters into audio event candidates.

Specifically, we again apply the spectral clustering algorithm on the dominant features $\left\{\mathbf{x}'_{s_i}\right\}$ generated from the whole audio stream and obtain

$$C_S = \left\{c_1, c_2, \dots c_{k_S}\right\} \tag{11}$$

where k_S is the clustering number to indicate event classes in S, which can be similarly decided as in (8). Accordingly, every resulting cluster c_i is considered as a set of audio segments as $c_i = \left\{s_{i1}, s_{i2}, \dots s_{ij}, \dots\right\}$. Then, we define the following affinity function to spot the background audio clusters which in general make up the largest majority of an audio scene stream as follows:

$$\text{aff}(c_i, S) = \frac{\exp(d_{c_i} - \mu_S)^2}{(2\sigma_S^2) \cdot \exp(\sigma_{c_i}/\mu_{c_i})} \tag{12}$$

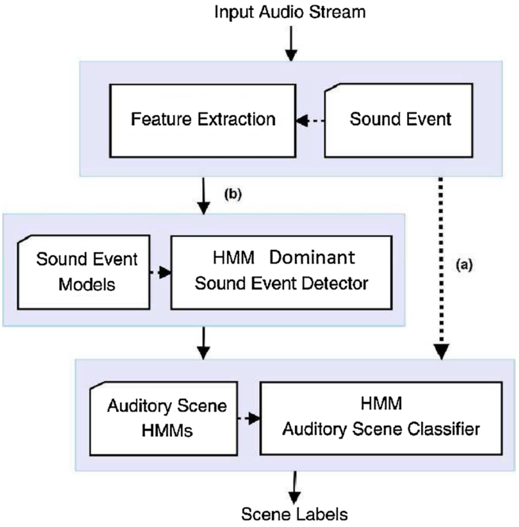


Fig. 3 Block diagram of the proposed auditory scene recognition algorithm, where *a* and *b* denote two alternative models in handling sound events

where d_{c_i} is the total length of all the audio segments in cluster c_i, μ_{c_i} and σ_{c_i} are the mean length and the standard deviation of the segments in c_i, while μ_S and σ_S are the mean length and the corresponding standard deviation of all the segments in the input signal S, respectively. Accordingly, background audio events can be identified by selecting those clusters that have the largest affinity value. The spotted clusters essentially obey the following two observations. First, there is in general a high affinity for a background cluster if its segment duration is longer than the others. Second, the larger a segment length varies in a cluster, the larger is the affinity.

After spotting background audio events, we consider all the rest clusters as audio event candidates for further analysis.

3.2.3 *Scene recognition using two-level HMM*

Based on the detected audio event candidates, we propose a two-level HMM scene recognition scheme, which has two alternative models as shown in Fig. 3. The main difference lies in whether or not having a separate explicit model for the discrete sound events.

For the former, we train one HMM model $\text{HMM}_e^{(j)}$ for each sound event class $j \in [1 \dots k_S]$, capturing the timbre (by HMM states) and the rhythm (by state transitions) of the audio event and modeling the observation density of the LDB–MFCC features by the mixture of Gaussian. The model size and initial probability of HMM are decided by a clustering algorithm on the training samples. For an input audio segment s_i, the extracted LDB–MFCC feature vector $\mathbf{x}'_{s_i}$ is given to each $\text{HMM}_e^{(j)}$, and the corresponding log-likelihood values $[ll_1, ll_2, \ldots, ll_{k_S}]$ are computed using the Forward algorithm.

The likelihood $[ll_1, ll_2, \ldots, ll_{k_S}]$ of a dominant audio segment belonging to each specific sound event is used as the feature vector fed into the succeeding scene recognition model. For more robustness, we can further build a *pesudo-semantic feature* as in [48] based on *likelihood ratio test*, which derives an inner-class distribution $f_i(x|\theta_1)$ and out-class distribution $f_i(x|\theta_0)$ of the likelihood value x for the ith event class, based on training samples from the ith event class and all other classes, respectively. Then, $[f_i(x|\theta_1)/f_i x|\theta_0]_{i=1\ldots k_S}$ is considered the pseudo-semantic feature vector and used in place of the likelihood in acoustic scene recognition.

An acoustic scene is thus characterized by the occurrences of dominant sound events. For each scene class $j \in [1 \ldots C]$ (C is the number of scene classes considered), we construct and train an HMM model $\mathrm{HMM}_s^{(j)}$ to capture its compositions (i.e., sound events) and temporal variation characteristics through HMM states and their transitions. In the training stage, the specification of the HMM including the model size (state number), the number of observation Gaussian mixtures and the initial probability in each state are determined experimentally based on the training acoustic scene samples collected, reflecting the complexity and variations of different scenes. The transition and emission probabilities of HMM are estimated with the Baum–Welch algorithm. The extracted dominant features of an audio segment belonging to each specific recognized sound event are fed into each $\mathrm{HMM}_s^{(j)}$ and the model with the maximum output likelihood is considered the recognized scene class. Note that an audio segment probably comprising multiple overlapping events is actually represented as a set of likelihoods (soft labels) belonging to each event category, which make their respective contributions to the characterization of the scene.

Since scene recognition is preceded by following dominant audio events while simultaneously ignoring the others in the segment, event context of the recognized scene class will in turn be used to detect all the rest audio events and update the recognized candidates iteratively in the next section.

4 Context-based audio event recognition

Essentially, audio events are considered in a correlated way to describe an audio scene [36, 49], e.g., a crowded restaurant scene probably contains the sounds of *the clatter of cutlery*, *opening/closing doors*, and *talks* simultaneously. That is, in the real world, audio events tend to correlate to other events inside a particular environment which provides a rich collection of contextual event associations. We refer such audio event correlations that happen in the same category audio environment as our audio event context. We, thereby, propose an audio context model to recognize audio events. In this model, each candidate event is respectively described by scene context and its own inherent characterizations.

Modeling audio event contexts. Each candidate event e_i is given a contextual descriptor $P_{\mathrm{con}}(o_i, e_i)$ that is defined as the average of every event pairwise probability $p(\langle e_i, e_j\rangle)$ in the same audio scene, where o_i is the desired semantic tag of e_i. For each audio event pairwise relationship $R^\triangle(\triangle = 1, 2, \ldots, N_S)$ from the same scene, we define a relationship matrix Ψ^i which captures the probability distribution of every audio event pairwise. The probability of an audio event pairwise $\langle s, t\rangle (s, t = 1, \ldots, K_S)$ in the relationship matrix $\Psi^\triangle$ is as follows:

$$p(\langle s, t\rangle; \Psi^\triangle) = \frac{1}{Z(\Psi^\triangle)} \exp\left\{\psi_{st}^\triangle\right\} \tag{13}$$

where $\psi_{st}^\triangle$ is the entry (s, t) in the relationship matrix $\Psi^\triangle$, while $Z(\Psi^\triangle)$ is a partition function. Essentially, an event pairwise $\langle s, t\rangle \in R^\triangle$ implies that two audio event classes of s and t satisfy the $\triangle$th relationship according to the training data. Thereby, given an audio scene data set D, we define the probability of relationship $R^\triangle$ as:

$$p(D, \Psi^\triangle) = \frac{1}{Z(\Psi^\triangle)^{M^\triangle}} \exp\left\{\sum_{s=1}^{k_S}\sum_{t=1}^{k_S} l_{st}^\triangle \psi_{st}^\triangle\right\} \tag{14}$$

where $l_{st}^\triangle$ is the entry (s, t) of a frequency matrix for relationship $R^\triangle$ in D, which counts the times of audio event pairwise $\langle s, t\rangle$ that appears in the training audio scene dataset, and $M^\triangle$ is the total number of the event pairwise which has relationship $R^\triangle$.

Characterizing each candidate event. Besides the contextual characterizations, we further define the following measure $P_{\mathrm{sco}}(o_i, e_i)$ to evaluate the saliency of each candidate event e_i, which essentially reveals the inherent characteristics for conveying semantic contents in an audio scene. We consider three score functions to evaluate the inherent properties of every candidate event, namely, the occurrence frequency of $f_{of}(c_i, S)$, the total duration of $f_{td}(c_i, S)$ and the average length of $f_{al}(c_i, S)$ as follows:

$$\begin{aligned} P_{\mathrm{sco}}(o_j, e_j \in c_i) &= \mathrm{salient}(c_i, S) \\ &= f_{of}(c_i, S) \cdot f_{td}(c_i, S) \cdot f_{al}(c_i, S) \end{aligned} \tag{15}$$

in which:

$$f_{of}(c_i, S) = \exp\left(-\frac{(n_{c_i} - \alpha \cdot \mu_{n_{c_i}})^2}{2\delta_{n_{c_i}}^2}\right) \tag{16}$$

$$f_{td}(c_i, S) = \exp\left(-\frac{(d_{c_i} - \beta \cdot \mu_{d_{c_i}})^2}{2\delta_{d_{c_i}}^2}\right) \tag{17}$$

$$f_{as}(c_i, S) = \exp\left(-\frac{(\overline{d}_{c_i} - \gamma \cdot \mu_{l_{c_i}})^2}{2\delta^2_{\overline{d}_{c_i}}}\right) \tag{18}$$

where n_{c_i}, d_{c_i} and $\overline{d}_{c_i}$ are the occurrence times, the total duration and the average segment length of c_i, respectively. μ and δ are the mean and the standard deviations of a corresponding measure, while α, β and γ are the co-efficiencies of μ.

We then integrate the contexts and the inherent characteristics of each audio candidate event e_i into a new representation of $P(o_i, e_i)$ to characterize the probability whether a candidate audio event e_i can be calibrated by tag o_i. The details to calculate $P(o_i, e_i)$ and the context-based audio event recognition model are shown in Algorithm 1, in which audio event recognition is considered as the problem of simultaneously maximizing the possibilities of scene context co-occurrences and the inherent characteristics of being a particular audio event for every audio candidate by

$$P(o_i, e_i) = \lambda P_{\text{sco}}(o_i, e_i) + (1 - \lambda) P_{\text{con}}(o_i, e_i). \tag{19}$$

Accordingly, audio event candidates will be iteratively updated through the context model.

Algorithm 1 Iterative probability-based audio event recognition using context model.

```
Input:
    Initially labeled audio cluster candidates: $(o_1, e_1)$, $\cdots$,$(o_{N'}, e_{N'})$ with the importance measure of $P_{sco}(o_1, e_1)$, $\cdots$,$P_{sco}(o_{N'}, e_{N'})$;
1: Set $R = \emptyset$ and initialize the context probability $P_{con}(o_1, e_1)$,$\cdots$,$P_{con}(o_{N'}, e_{N';})$ to be zero;
2: $P(o_i, e_i) = \lambda P_{sco}(o_i, e_i) + (1 - \lambda) P_{con}(o_i, e_i)$, where $\lambda$ is a weight factor. Search for $P(o_*, e_*) = \arg\max_{(o_i,e_i)\notin R} P(o_i, e_i)$.
    For $i = 1$ to $k_S$
        If $P(o_*, e_*) > P_{sco}(o_j, e_j)$
            $o_* \rightarrow o_i$, $R = R \cup \{(o_i, e_*)\}$; break;
    EndFor
3: For the remaining candidates, update the context probability as follows:
    $P_{con}(o_j, e_j) = \frac{1}{M}\left[(M-1)P_{con}(o_j, e_j) + \frac{exp\{\psi^{\triangle}_{e_j e_*}\}}{Z(\Psi^{\triangle})}\right]$
    where $M$ is the number of detected audio events and $\triangle \in \{1, \cdots, N_S\}$ . Go to 2.
Output:
    The detected audio events with updated labels $R = \{(o_i, e_i)\}$.
```

To solve Algorithm 1, we maximize the log likelihood of each observed audio event pairwise by

$$L(\Psi^{\triangle}) = \log p(D, \Psi^{\triangle}) = \sum_{s=1}^{k_S}\sum_{t=1}^{k_S} l^{\triangle}_{st}\psi^{\triangle}_{st} - M^{\triangle} \times \log Z(\Psi^{\triangle}). \tag{20}$$

We approximate the partition function using Monte Carlo integration. The importance sampling is employed and the distribution is equal to their observed frequency. Thus, we can use the gradient descent to find $\Psi^{\triangle}$, which approximately optimizes the likelihood, and accordingly the gradient is as follows:

$$\nabla_{\Psi^{\triangle}} L(\Psi^{\triangle}) = \begin{bmatrix} l^{\triangle}_{11} & \cdots & l^{\triangle}_{1k_S} \\ \vdots & \ddots & \vdots \\ l^{\triangle}_{k_S 1} & \cdots & l^{\triangle}_{k_S k_S} \end{bmatrix}. \tag{21}$$

Note that in Algorithm 1, we assign the label for each candidate event that is not identified as a dominant source in Sect. 3.2 by searching for the nearest distance with the training events in the feature space. Assuming that such a candidate event c_i contains M_{c_i} audio segments and each segment is characterized by the LDB+MFCC feature vector with the dimension of N, an audio candidate event can essentially be represented by an $N \times M_{c_i} (N < M_{c_i})$ matrix $\mathbf{E}_{c_i}$. Then, the SVD algorithm is employed again to extract the dominant features of an audio event candidate in the feature space by decomposing $\mathbf{E}_{c_i}$ as

$$\mathbf{E}_{c_i} = \mathbf{USV}^T \tag{22}$$

where $\mathbf{U} = \{u_1, \ldots, u_N\}$ is an $N \times N$ orthogonal matrix, $\mathbf{S} = diag\{\lambda_1, \ldots, \lambda_N\}$ is an $N \times M_{c_i}$ diagonal matrix of singular values, for which $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$, and $\mathbf{V}$ is an $M_{c_i} \times M_{c_i}$ matrix. Since the principal components associated with large singular values represent the primary distribution of the audio element in the feature space, the desirable principal component number to describe a candidate event is chosen by:

$$m = \arg\min_k \left\{ \sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{N} \lambda_i > \eta \right\} \tag{23}$$

η is a threshold to initialize an event label. As a result for a test audio event c_j, its L2 distances of m largest principal components between the features of c_j and all the training audio events are calculated and accordingly the minimum is selected to initialize c_j with the same label. The initialized labels for non-dominant candidate events will then be recognized in Algorithm 1 by further considering their audio scene contexts and co-occurrence correlations to improve the accuracy.

Our audio context model is essentially inspired by the fact that different audio events in the same audio scene can take a complementary role on event recognition by meeting the following two requirements. First, like a visual object that can help recognize others in the same scene [50], the presence of one audio event in general helps to identify the existence of others. For example, a rain environment recognition system that exploits both *rain* and *thunder* sounds

Table 1 Confusion matrix for scene recognition on the DCASE benchmark using the proposed method

	Recognized scenes % (in the same order as rows)									
	1	2	3	4	5	6	7	8	9	10
Bus	81			4				15		
Busstreet		90				6				4
Office			81		9		8	2		
Openair				62		6	1	31		
Park		10	31		57	2				
Quietstreet			2	10	4	74		10		
Restaurant				23			34	43		
Supermarket	3	10	14	11				62		
Tube	10	3		16				14	44	13
Tubestation	3	2	1		2	16	3	9	26	38

may achieve better performance in both accuracy and efficiency than one which exploits either one or the other. As a result, an audio event context allows people to understand the semantic behind it and more importantly, inspires people to associate this event with other known events for more accurate audio event recognition. Theoretically, the acoustic scene context model is also useful in recognizing overlapped sound events since it captures the occurrence correlations between dominant and non-dominant sounds to infer the semantics of the latter, rather than directly separating the overlapped sounds which has been proved difficult and sometimes inaccurate. Second, audio events tend to be correlated and audio context modeling can take a complementary role on solving environmental audio content analysis tasks by following a specific audio source while simultaneously ignoring or simply acknowledging the other noises. Accordingly, audio events in the same environmental scene can be inferred and recognized using the importance measure evaluated by every audio event itself and the concurrence measure characterized by associated audio contexts in a relatively accurate and robust way.

5 Experimental results

In this section, we evaluate the performance of the proposed approach on several audio datasets, including the recent DCASE benchmark [23], our proposed 10-Category Audio Scene and Event dataset (10-CASE dataset), and a public TV-Movie dataset with nearly altogether 24 h audio data.

5.1 Evaluations on the benchmark and the 10-CASE dataset

We first evaluate the proposed method on the recent DCASE benchmark dataset, the scene types in which are pre-selected with an equal balance of indoor/outdoor scenes in London area, including: *bus*, *busytreet*, *office*, *openairmarket*, *park*, *quietstreet*, *restaurant*, *supermarket*, *tube* and *tubestation*. Binaural stereo format using a Soundman OKM II microphone is used to record the audio data. Table 1 gives the confusion matrix using the proposed method on this dataset. The rows of the matrix denote the scene classes we attempt to classify, and the columns depict the classified results. It can be found that our method averagely achieves an accuracy of 62.3 %, outperforming 52.1 % of the baseline method in [23] on the same dataset. This is probably due to our processing steps on the scene recognition hierarchy consisting of adaptive audio segmentation and background filtering. However, in this experiment, both the proposed scene recognition method and the other recent 11 methods that adopt either SVM, GMM, random forest or likelihood ratio test only focus on selecting proper statistic models or discriminative classifiers, but not contextual content analysis of environmental sounds themselves. This is because the types and the number of the audio events in each scene category in the DCASE dataset are limited. Ideally, for better understanding noisy environmental sounds, sufficient representative samples of audio events from different types within each scene category are required. Note that the audio events in each scene are not segmented and labeled in this dataset, making the training and the selection of the optimized LDB features for specific audio events impossible. Thereby, only MFCC features are used in this experiment.

We thus create a 10-CASE dataset consisting of 10 scene types and 21 event categories, which are distinct enough to be perceived, by collecting audio data from Internet. The 10-CASE dataset has 5,250 audio clips, in which each clip corresponds to an individual short audio event that has been well segmented. Altogether six outdoor scenes and four indoor scenes are considered, namely, *restaurant*, *street with traffic*, *playground*, *train station*, *inside*

Table 2 Confusion matrix for scene recognition on 10-CASE dataset using LDB+MFCC features

	Recognized scenes % (in the same order as rows)									
	1	2	3	4	5	6	7	8	9	10
Vehicle	90		1	6	2			1		
Beach	4	39	2	30		8		2	13	2
Station	17	2	69	3	6				3	
Street	18	8	7	57				1	8	1
Restau.	1			4	93				2	
Audito.	3		1	4	17	73			2	
Forest							100			
Raining	1	2		1	1	1	1	89	3	1
Playgnd	2			4	1	6			87	
War	3	1	2	8						86

moving vehicles, *auditorium*, *battle field*, *forest*, *beach*, and *raining with thundering*. Totally 21 audio event types are classified, including *engine*, *car-braking*, *siren*, *horn*, *gunshoot/explosion*, *tableware*, *running water*, *bird*, *thundering*, *people talking*, *applause*, *laughter*, *cheer*, *traffic* and *crowd background*, etc. The clip length ranges 1–3 s for audio events and 15 s–2 min for scenes. To extract audio features, we learn a 40-dimensional LDB subspace (20 for node energy and 20 for coefficient variance) based on the audio event training set. Then, for each input audio stream, we apply a sliding event detection window of around 1 s length on it, which is further divided into overlapping 200 ms frames, and extract LDB, MFCC, LDB+MFCC feature vectors from each frame.

Tables 2 and 3 give the confusion matrices constructed by applying the audio scene recognizer to the test set in a single arbitrary trial using the hybrid LDB+MFCC features and the single LDB features, respectively. The average accuracies of these two ways are 78.3 and 74.6 %. The average accuracy is higher than that on the DCASE dataset, which is probably because there are sufficient sound events of different types in each scene for better selecting features and training the model. The model size of the HMM for each acoustic scene type listed in it is experimentally chosen as 4, while the number of observation mixtures is set as 4 in this work, which can also be learned from samples. We further compare the scene recognition method with the recent baseline algorithm in [23] on our 10-CASE dataset. The confusion matrix of the baseline algorithm is given in Table 4 with the average accuracy of 75.2 %, which is slightly higher than the single LDB features but lower than the hybrid MFCC+LDB features. This verifies our hypothesis that the combination of two different features may form a better fused feature in recognizing scene audio that is complex and has various compositions.

To test audio event recognition, we additionally select five real audio scene streams consisting of *raining*, *restaurant*, *sport*, *street* and *war* from the BBC Sound Effects Library [51], which is for real systems like movie, TV or music production. Altogether 1,163 individual event sound segments (e.g., moving chair, closing door, and stepping) are detected. The final event recognition results from noisy acoustic scene streams are shown in the confusion matrix of Fig. 4, in which the mean accuracy is 62.2 %. This accuracy is similar to that of the DCASE dataset but lower than that on the 10-CASE dataset, illustrating that the number and the types of sound events, namely, the characteristics of acoustic scene compositions, may greatly affect the performance of acoustic scene understanding algorithms.

To evaluate the LDBs inherent discriminability resulting from the adaptive learning (node selection) process on the audio event training samples together with LDBs capability of depicting multiscale temporal signatures and audio spectral shape, in another experiment, the LDB feature vectors extracted from the input audio steam are directly used as the observations of the HMM-based scene model for training and classification instead of using the proposed audio context model. The worse case is for crowd background (28.6 %), followed by running water (52 %) and traffic (66.7 %), since they are usually covered by other events and noise.

Finally, for better evaluating the influences brought by different audio features or their combinations, in Fig. 5, we compare the overall recognition accuracies for considered scene classes using LDB features, MFCC features, and LDB+MFCC features, respectively. It also shows the complementary aspect of LDB and MFCC features in characterizing variant audio events in audio scenes. Similar to the finding in [52] that MFCC is weak to characterize scenes composed of sounds with narrow spectral band structure, it performs insufficiently in our experiments for the low-frequency wave background and high-frequency bird chirps

Table 3 Confusion matrix for scene recognition on 10-CASE dataset using LDB features

	Recognized scenes % (in the same order as rows)									
	1	2	3	4	5	6	7	8	9	10
Vehicle	58				2			40		
Beach		77				2		17	4	
Station	3		81		4				12	
Street	1	1	30	46	19			1	2	
Restau.		7	1		82				10	
Audito.		6	3	1		68		7	15	
Forest		1					97	2		
Raining		2	7					90	1	
Playgnd		9	2		9	4	1		75	
War		4		4	3			16	1	72

Table 4 Confusion matrix for scene recognition on 10-CASE dataset using the baseline method

	Recognized scenes % (in the same order as rows)									
	1	2	3	4	5	6	7	8	9	10
Vehicle	90		3					4	3	
Beach	5	29		2	21	1	11		7	24
Station	1		83	1	15					
Street	18		1	49	3				14	15
Restau.	9		1		84					6
Audito.		13	1		13	69	1		2	1
Forest							100			
Raining	7						1	89		3
Playgnd			8	1	6	4	1		80	
War	5		12					4		79

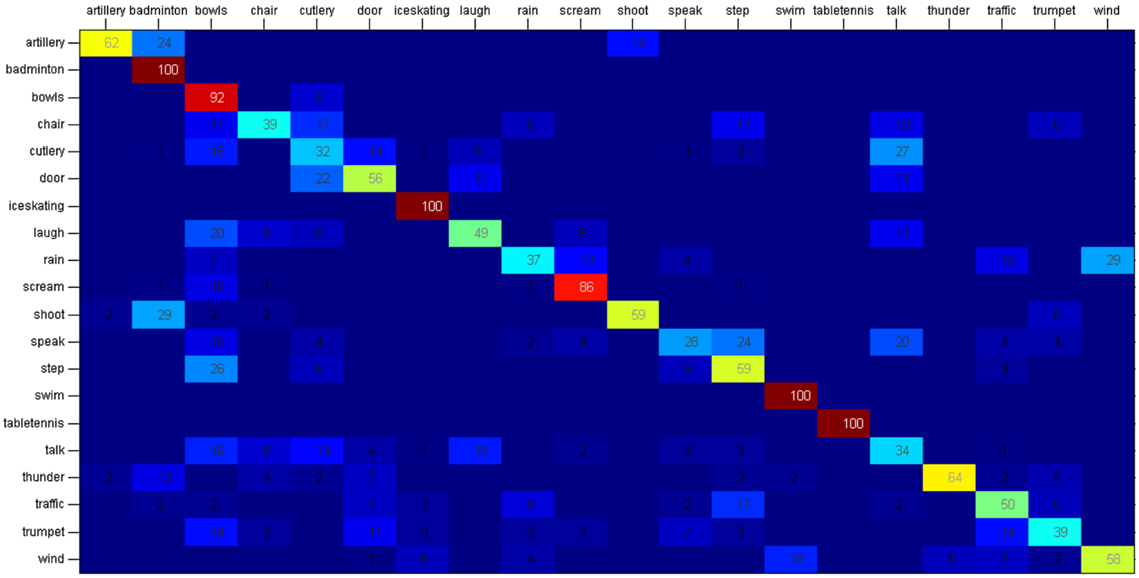


Fig. 4 Altogether 1,163 individual audio events from 5 real acoustic scene streams are detected, and the confusion matrix of all the detected results is shown

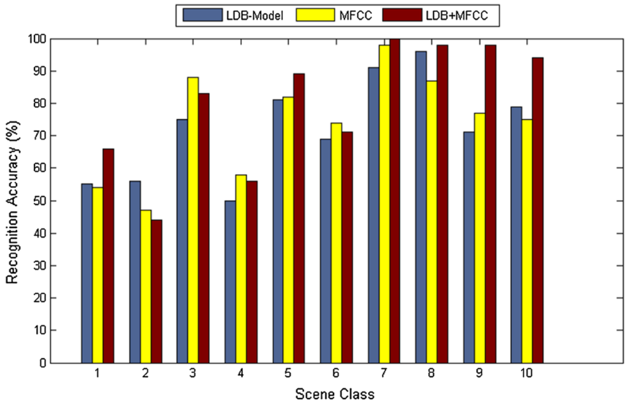


Fig. 5 Overall recognition rates comparing ten classes using LDB features, MFCC features, and LDB+MFCC features, respectively. Scene classes are arranged in the same order as Table 3

in the beach scene, for which LDB features are effective. On the other side, for some sounds with mixed and wide-spreading spectrum like cluttered voices, music and effects in auditorium, MFCC can be more efficient than LDB. By combining LDB and MFCC features, we achieve an averagely enhanced accuracy rate.

5.2 Experiments on public TV-Movies

Finally, we test our method on TVs and movies that have long durations and a lot of audio event changes to imitate real-life environments.

TV-Movie dataset. The public TV-Movie dataset is composed of continuous audio streams for evaluating our proposed event recognition method. The total duration of the dataset nearly reaches 24 h as shown in Table 5, covering four TV categories of *basketball*, *tennis*, *table tennis*, *awards party* and three movie categories of *action*, *comedy* and *war*. For simplification, each TV or movie category here is considered as a scene class. All the audio samples are in 44.1 KHz and mono channel. Since we achieve an improved accuracy after combining LDB and MFCC features as shown in Fig. 5, we still extract the 20-dim LDB features together with the 21-dim MFCC features from every audioframe which contains 1,024 sampling points without overlaps.

Adaptive segmentation. We evaluate our adaptive segmentation method from long duration audio streams. Figure 6 shows an example of the segmentation results on a 20-s scene audio stream. The curve in Fig. 6a represents the normalized distance between every audioframe and $x_S^{b'}$, where the two red valley points essentially represent the spotted boundaries of scene audio segments. Figure 6b shows the corresponding similarity matrix for the example audio stream S, which consists the sounds of *applause* (50–57 s), *silence* (57–64 s) and *speech* (64–70 s). Note that, there generally exists a potential audio change if the normalized distance decreases sharply, which can also be correspondingly found in the grayscale image as shown in Fig. 6b.

Table 6 illustrates the audio event recognition results by respectively employing adaptive segmentation and fixed-window segmentation on environmental audio streams. In Table 6, we find the accuracy of audio event recognition by using adaptive segmentation is averagely improved against using a fixed-size window by avoiding intermission of independent audio events.

Audio context modeling. We evaluate our context-based model for audio event recognition from audio scenes. Figure 7 shows the influence of audio context modeling in event detection. Similarly, we find that the average accuracy of audio event recognition can be effectively improved by audio scene context modeling. Table 7 shows some detected audio event examples together with their occurrence numbers in the test audio tracks from Table 5. We also notice that the true positive rate of specific audio events may be decreased like *noise*. The reason is potentially that the proposed greedy search algorithm is based on the local maxima in the hypothesis space but some unstructured audio events are relatively difficult to define from the impure training data.

Note that overlapped audio events can be handled if the audio features from one event are dominant enough. This is true because after feature extraction and spectral clustering, the audio features from overlapped events will be successfully separated into different clusters. For example, suppose there is an audio segment essentially composed of two overlapped events A and B in the form A_1-(A_2, B)-A_3, namely, the A_2 part in event A overlaps with another event B. After spectral clustering, A_1 and A_2 are probably clustered into the same candidate cluster indicating event A, and (A_2, B) will probably be clustered into another candidate cluster indicating B if event B is dominant in (A_2, B), or still clustered into candidate cluster A if A_2 is dominant. Otherwise, this overlapped part will probably be clustered into a new overlapped event candidate, or directly considered as background if no such overlapped event is defined. For example in Table 7 the overlapped events such as "music + laughter" and "laughter + speech" are detected.

Feature selection and parameter setting. In Table 8, we further compare the overall recognition accuracy for the audio scenes by respectively employing LDB, MFCC and LDB–MFCC feature vectors. It can be seen that the hybrid feature still remains valid in interpreting long-time audio scene streams.

During spectral clustering, we need set two empirical values of *AUDI_MIN* and *AUDI_MAX* as the cluster number range [*AUDI_MIN*, *AUDI_MAX*]. Such an empirical

Table 5 The experimental datasets

Category	Audio source	Dur. (s)
Action	A_1 The Fast and the Furious	21,020
	A_2 State of the Union	
	A_3 Sword fish	
	A_4 The rock	
Award	B_1 83rd academy awards	21,224
	B_2 Muchmusic video music	
	B_3 CMT music awards 2012	
	B_4 39th annual music awards	
Comedy	C_1 3rd rock from the sun	5,151
	C_2 Friends	
	C_3 The Big Bang Theory	
Sports	D_1 World table tennis game	17,717
	D_2 2012 Olympic table tennis	
	D_3 NBA 2012 finals	
	D_4 ATP 2012 Wimbledon SF1	
War	E_1 Enemy at the gates	19,496
	E_2 First blood	
	E_3 Pearl Harbor	
	E_4 Saving Private Ryan	

Table 6 Performance evaluation on audio event recognition by respectively using adaptive segmentation and fixed-window segmentation of environmental audio streams from action movies in the dataset

Source	Adaptive seg. (%)	Fixed seg. (%)
A_1	76.74	64.12
A_2	42.22	21.57
A_3	51.12	53.44
A_4	60.62	57.56
Average	57.68	49.17

cluster range will potentially affect the computational efficiency and accordingly decide the final accuracy. We, therefore, perform experiments on different audio sources to search for the best parameters. In Fig. 8, we find that when we set *AUDI_MIN* to 2 and *AUDI_MAX* is in the range from 8 to 14, *AUDI_MAX*=12 gives an averagely high accuracy for event recognition according to our experimental results (see Fig. 8a). Similarly, when we set *AUDI_MAX* to 12 and *AUDI_MIN* is in the range from 2 to 6, *AUDI_MIN* = 5 obtains the best accuracy (see Fig. 8b). Accordingly, the empirical range for spectral clustering is finally set as [5, 12] for an unknown audio signal.

Post-processing. For a long-time complex acoustic scene stream, we rank the detected audio events according to their importance probability calculated by (15), and merge neighboring events in a descending order based on the following two assumptions: (1) whether two neighboring audio events are considered from the same audio scene can be measured by the correlation coefficient (CC) of their audio features, and (2) the longer the time interval between two adjacent audio events is, the lower possibility in the same scene they will be. Thereby, we define the correlation function for two adjacent audio events ith and jth as follows:

$$S_{ij} = \frac{\delta}{d_{ij}} \cdot \exp(-(d_i - d_j)^2/(d_i + d_j)) \cdot \exp(\mathrm{corr}_{ij}) \qquad (24)$$

where d_{ij} and corr_{ij}, respectively, represent the time interval and the correlation coefficient between the ith and jth audio events, while d_i and d_j are the durations of the ith and jth audio events, δ denotes the harmonic factor. The final results from the test TV and movie streams are shown in Table 9.

6 Discussion and conclusion

A novel audio scene understanding framework is presented, which consists of two hierarchies of acoustic scene recognition and audio event recognition. On the scene recognition hierarchy, we detect audio changes in the Eigen-audiospace adaptively segment an audio scene stream and then

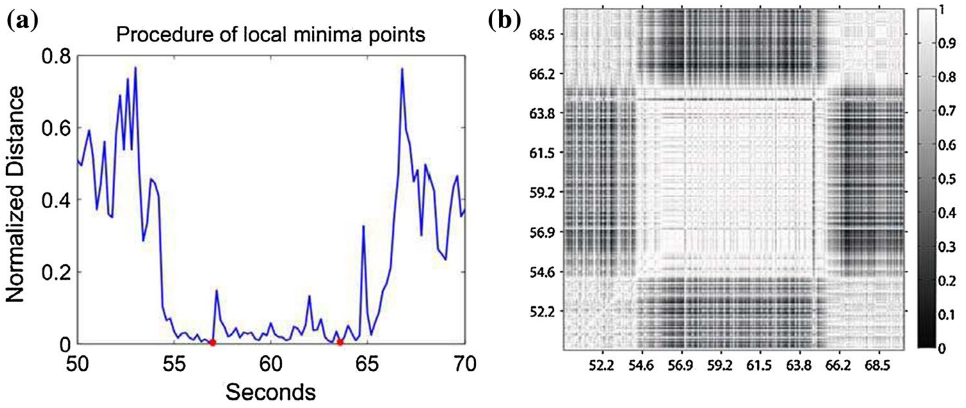


Fig. 6 Segmentation of a 20-s table tennis game scene stream consists of *applause* (50–57 s), *silence* (57–64 s) and *speech* (64–70 s): **a** local minima calculation of the normalized distance, and **b** the corresponding grayscale image of similarity matrix of the audio stream

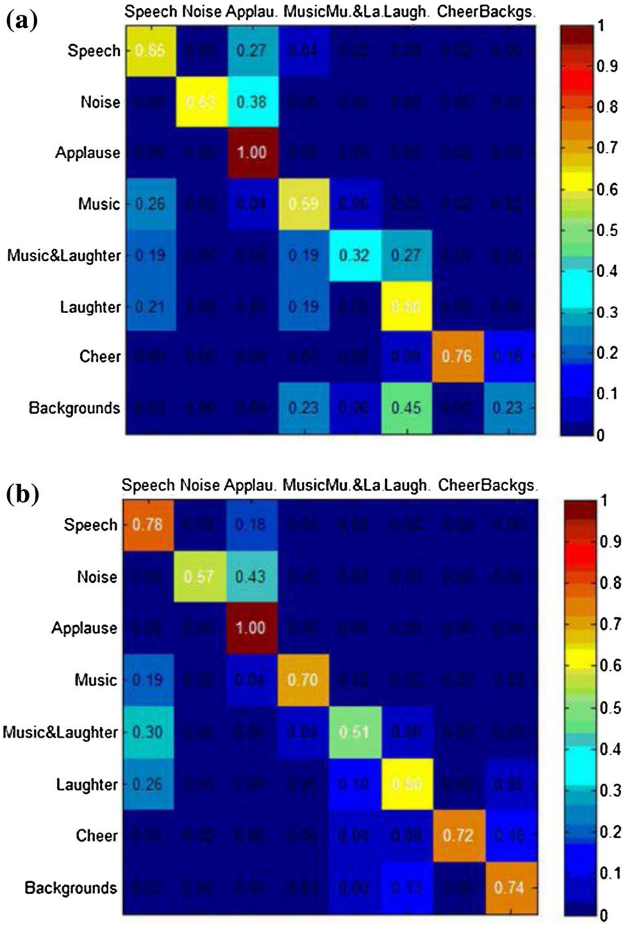


Fig. 7 The confusion matrices for audio event detection from five kinds of audio scenes. **a** Without audio context modeling; **b** with audio context modeling

Table 7 Audio event detection results from the TV-Movie dataset

No.	Audio events
A_1	*Fighting (114), speech+backgrounds (75), engine (28)* *Speech+noise (53), gunshot (70), siren (19)* *Backgrounds (141)*
B_1	*Applause+music (85), laughter+speech (102)* *Cheer (67), song (24), speech (83), music (42)* *Backgrounds (94)*
C_1	*Laughter (17), music+laughter (19), applause (25)* *Cheer (18), speech (38), music (38), backgrounds (23)* *Noise (4)*
D_1	*Applause (77), cheer+appluse (74), ball-hit (133),* *Speech (33), speech+music (92), music (84)* *Backgrounds (58)*
E_1	*Gunshot (37), explosion (38), speech+music (115),* *Speech+noise (121), music (104), noise (87)* *Silene (24), speech (107), backgrounds (90)*

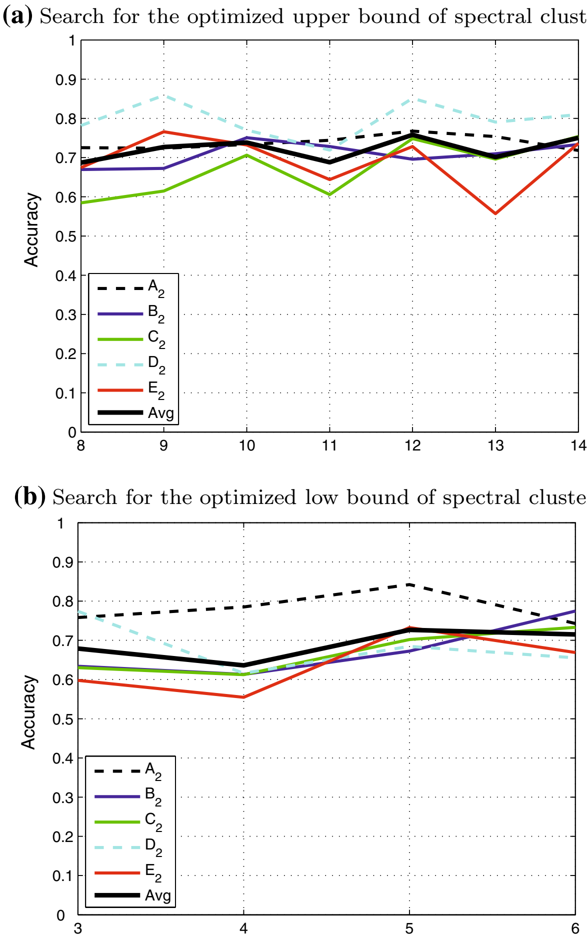
extract the LDB–MFCC features to represent every audio segment. After filtering background audio segments, acoustic scene streams are recognized by two-level HMM. Then on the audio event recognition hierarchy, a context-based model which essentially characterizes the occurrence correlations of the audio events within the same scene category is presented to correctly tag meaningful audio events from various sources.

Essentially, recognizing acoustic scenes that contain non-speech and non-music environmental sounds and detecting various events within each scene are still in their early stages as introduced, thus the researchers in this community are now establishing datasets and developing methods as benchmarks. As reported in [53], the mean accuracy by human listeners is only around 71 % on the data collected in office environments from the DCASE dataset. The proposed method averagely achieves an accuracy of 62.3 % on the same dataset, outperforming the average accuracy 53.8 % of the recent 11 methods. It still can be improved due to the following fact: the data in DCASE are directly recorded in London area, in which the types and the number of the audio events are very limited for training discriminative features. We then create a 10-CASE dataset by manually collecting 5,250 audio clips of 10 scene types and 21 event categories. Our experimental results on this dataset show that the proposed method averagely achieves the enhanced performance of 78.3 %. This result also illustrates our hypothesis that the inherent discriminability of the best-basis strategy greatly relies on whether there are sufficient types of different audio events for training. Finally, by comparing the confusion matrices for audio event detection from five kinds of audio scenes, we find that the average accuracy of audio event recognition can be effectively improved by capturing dominant audio sources and reasoning non-dominant events from the dominant ones through acoustic context modeling.

However, the current methods of audio event detection from acoustic natural scenes still face the following difficulties: (1) the definition of a particular real-life acoustic environment sometimes is difficult even for human listeners (e.g., an office environment can be either very quiet with only slight sounds of tapping keyboards or very noisy with loud discussions and laughers), (2) it is difficult to decide whether an audio event is dominant for a particular real-life acoustic scene while the others are not, and (3) the occurrence correlations among the audio events that appear in the same acoustic scene are statistically, which may be inaccurate to infer various audio events from multiple sources in a real-life acoustic scene as humans do. In the future work, we will improve the computational efficiency of the method, and enrich audio contexts from event level to local or user level for more accurate event detection. Another interesting research topic is reversely exploring

Table 8 Overall recognition rates by using LDB, MFCC and LDB+MFCC features

No.	LDB	MFCC	LDB+MFCC
A_2	0.6100	0.7140	0.7674
B_2	0.5920	0.6630	0.6959
C_2	0.6300	0.6783	0.7483
D_2	0.7450	0.7880	0.8512
E_2	0.6451	0.6732	0.7282
Overall	0.6466	0.7034	0.7582



(a) Search for the optimized upper bound of spectral clusters.

(b) Search for the optimized low bound of spectral clusters.

Fig. 8 Search for an optimized spectral cluster range for audio event recognition

the influence of the detected audio events on more accurate scene recognition in the post processing stage. Moreover, the incorporation of other modalities such as video and sensor information during collecting sound signals using mobile phones will be considered to further advance the performance of the proposed framework.

Table 9 The detected audio scene examples from the test audio streams

No.	Audio scenes
A_1	*Fighting (112): {fighting, speech+backgrounds}*
	Shootout (58): {gunshot, speech+noise}
	Chase (54): {engine, siren}
	Speech (156): {speech+backgrounds, music, speech+noise, silence}
	Backgrounds (126): {theme music, backgrounds}
B_1	*Awarding (180): {cheer, laughter+speech, applause+music}*
	Speech (172): {speech, music, applause+music}
	Song (34): {song,cheer,backgrounds}
	Music (41): {music, applause+music, backgrounds}
	Backgrounds (94): {backgrounds,music,speech}
C_1	*Comedy (60): {laughter, applause, cheer}*
	Speech (76): {speech, noise, music}
	Backgrounds (32): {music, backgrounds, noise}
D_1	*Match (148): {ball-hit, applause, backgrounds}*
	Speech (183): {speech, speech+music, applause}
	Cheer (56): {cheer+appluse, applause}
	Backgrounds (143): {music, backgrounds}
E_1	*War (117): {gunshot, explosion, noise}*
	Speech (158): {speech+noise, noise}
	Music (51): {music, noise}
	Backgrounds (110): {backgrounds, noise, silene}

Acknowledgments The work described in this paper was supported by the Natural Science Foundation of China under Grant Nos. 61272218 and 61321491, the 973 Program of China under Grant No. 2010CB327903, and the Program for New Century Excellent Talents under NCET-11-0232. We thank Yangbin Wen, Like Xue and Yirui Wu for their working on collecting data and experiments. The authors thank the anonymous reviewers for their constructive comments, which helped to improve the paper.

References

1. Peng, Y.T., Lin, C.Y., Sun, M.T., Tsai, K.C.: Healthcare audio event classification using hidden Markov models and hierarchical hidden Markov models. In: ICME, pp. 1218–1221 (2009)
2. Heittola, T., Mesaros, A., Virtanen, T., Eronen, A.: Sound event detection in multisource environments using source separation. In: CHiME, pp. 36–40 (2011)
3. Yang, L., Su, F.: Auditory context classification using random forests. In: ICASSP, pp. 25–30 (2012)
4. Lin, W., Lu, T., Su, F.: A novel multi-modal integration and propagation model for cross-media information retrieval. In: MMM, pp. 740–749 (2012)
5. Gerosa, L., Valenzise, G., Antonacci, F., Tagliasacchi, M., Sarti, A.: Scream and gunshot detection in noisy environments. In: EUSIPCO (2007)
6. Bardeli, R., Wolff, D., Kurth, F., Koch, M., Tauchert, K., Frommolt, K.: Detecting bird sounds in a complex acoustic

environment and application to bioacoustic monitoring. Pattern Recognit. Lett. **31**(12), 1524–1534 (2010)

7. Ntalampiras, S., Potamitis, I., Fakotakis, N.: On acoustic surveillance of hazardous situations. In: ICASSP, pp. 165–168 (2009)
8. Wang, X., Rosenblum, D., Wang, Y.: A daily, activity-aware, mobile music recommemder system. In: ACM Multimedia, pp. 1313–1314 (2011)
9. Rho, S., jun Han, B., Hwang, E.: SVR-based music mood classification and context-based music recommendation. In: ACM Multimedia, pp. 713–716 (2009)
10. Cowling, M., Sitte, R.: Comparison of techniques for environmental sound recognition. Pattern Recognit. Lett. **24**(15), 2895–2907 (2011)
11. Wang, Y., Li, B., Jiang, X., Liu, F., Wang, L.: Speaker recognition based on dynamic mfcc parameters. In: IASP, pp. 406–409 (2009)
12. Harsha, Y.S., Vasudeva, V., Kishore, P.: Significance of anchor speaker segments for constructing extractive audio summaries of broadcast news. In: SLT, pp. 12–18 (2010)
13. Shiu, Y., Jeong, H., kuo, C.-C.J.: Similarity matrix processing for music structure analysis. In: ACM Multimedia, pp. 69–76 (2006)
14. jun Han, B., Rho, S., Jun, S., Hwang, E.: Music emotion classification and context-based music recommendation. Multimed. Tools Appl. **47**(3), 433–460 (2010)
15. O'Shaughnessy, D.: Automatic speech recognition: history, methods and challenges. Pattern Recognit. **41**(10), 2965–2979 (2011)
16. Raj, B., Stern, R.: Missing-feature approaches in speech recognition. In: IEEE Signal Process, pp. 101–116. (2005)
17. Lyon, R.: Machine hearing. In: IEEE, Signal Process, pp. 131–139 (2010)
18. Heittola, T., Mesaros, A., Virtanen, T., Eronen, A.: Audio event detection in multisource environments using source separation. In: Machine Listening in Multisource Environments (2011)
19. Su, F., Yang, L., Lu, T., Wang, G.: Environmental sound classification for scene recognition using local discriminant bases and HMM. In: ACM Multimedia, pp. 1389–1392 (2011)
20. Lu, T., Wang, G.Y., Wen, Y.B.: Auditory movie summarization by detecting audio events and scene changes. In: ICPR (2014). (To appaer)
21. Jin, Y., Lu, T., Su, F.: Movie keyframe retrieval based on cross-media correlation detection and context model. In: IEA/AIE, pp. 816–825 (2012)
22. Lu, T., Jin, Y.K., Su, F., Shivakumara, P., Tan, C.L.: Content-oriented multimedia document understanding through cross-media correlation. In: Multimedia Tools and Applciations (2014). (To appear)
23. Giannoulis, D., Stowell, D., Benetos, E., Rossignol, M., Lagrange, M., Plumbley, M.D.: A database and challenge for acoustic scene classification and event detection. In: EUSIPCO (2013)
24. Eronen, A.J., Peltonen, V.T., Tuomi, J.T., Klapuri, A.P., Fagerlund, S., Sorsa, T., Lorho, G., Huopaniemi, J.: Audiobased context recogniton. IEEE Trans. Audio Speech Lang. Process. **14**(1), 321–329 (2006)
25. Aleh, K.I., Elian, A.A., Kabal, P.: Frame level noise classification in mobile environments. In: ICASSP, pp. 237–240 (1999)
26. Gaunard, P., Mubikangiey, C.G., Couvreur, C., Fontaine, V.: Automatic classification of environmental noise events by hidden markov models. In: ICASSP, pp. 3609–3612 (1998)
27. Scheirer, E.D., Slaney, M.: Construction and evaluation of a robust multifeature speech/music discriminator. In: ICASSP, pp. 1331–1334 (1997)
28. Han, B.-J., Hwang, E.: Environmental sound classification based on feature collaboration. In: ICME, pp. 542–545 (2009)
29. Ghoraani, B., Krishnan, S.S.: Time-frequency matrix feature extraction and classification of environmental audio signals. IEEE Trans. Audio Speech Lang. Process. **19**(7), 2197–2209 (2011)
30. Umapathy, K., Krishnan, S.S., Jimaa, S.A.: Multigroup classification of audio signals using time–frequency parameters. IEEE Trans. Multimed. **7**(2), 308–315 (2005)
31. Umapathy, K., Krishnan, S.S., Rao, R.K.: Audio signal feature extraction and classification using local discrimininant bases. IEEE Trans. Audio Speech Lang. Process. **15**(4), 1236–1246 (2006)
32. Chu, S., Narayanan, S., Kuo, C.-C.J.: Environmental sound recognition with timecfrequency audio features. IEEE Trans. Audio Speech Lang. Process. **17**(6), 1142–1158 (2009)
33. Mäkinen, T., Kiranyaz, S., Pulkkinen, J., Gabbouj, M.: Evolutionary feature generation for content-based audio classification and retrieval. In: EUSIPCO, pp. 27–31 (2012)
34. Niessen, M.E., Leendert, V.M., Andringa, T.C.: Disambiguating sounds through context. In: IEEE International Conference on Semantic Computing, pp. 88–95 (2008)
35. Heittola, T., Mesaros, A., Eronen, A., Virtanen, T.: Audio context recognition using audio event histograms. In: European Signal Processing Conference, pp. 23–27 (2010)
36. Heittola, T., Heittola, T., Mesaros, A., Eronen, A., Virtanen, T.: Context-dependent sound event detection. EURASIP J. Audio Speech Music Process. (2013). doi:10.1186/1687-4722-2013-1
37. Su, J.-H., Yeh, H.-H., Yu, P.S., Tseng, V.S.-M.: Music recommendation using content and context information mining. IEEE Intell. Syst. **25**(1), 16–26 (2010)
38. Park, H.-S., Yoo, J.-O., Cho, S.-B.: A context-aware music recommendation system using fuzzy bayesian networks with utility theory. In: FSKD, pp. 970–979 (2006)
39. Elliott, G.T., Tomlinson, B.: Personalsoundtrack: contextaware playlists that adapt to user pace. In: SIGCHI, pp. 736–741 (2006)
40. Rho, S., jun Han, B., Hwang, E.: Svr-based music mood classification and context-based music recommendation. In: ACM MM, pp. 713–716 (2009)
41. Mirikitani, D.T., Nikolaev, N.: Recursive bayesian recurrent neural networks for time-series modeling. IEEE Trans. Neural Netw. **21**(2), 262–274 (2010)
42. Cai, L.-H., Lu, L., Hanjalic, A., Zhang, H.J.: A flexible framework for key audio effects detection and auditory context inference. IEEE Trans. Audio Speech Lang. Process. **14**(3), 1026–1039 (2006)
43. Wang, J.-C. C., Wang, J.-F.-F., Kuok, W.,Hsu, C.-S.: Environmental sound classification using hybrid SVM/KNN classifier and MPEG-7 audio low-level descritpor. In: IJCNN, pp. 1731–1735 (2006)
44. Räsänen, O., Leppänen, J., Laine, U.K., Saarinen, J.P.: Comparison of classifiers in audio and acceleration based context classification in mobile phones. In: EUSIPCO, pp. 946–950 (2011)
45. Kinnunen, T., Saeidi, R., Leppanen, J., Saarinen, J.P.: Audio context recognition in variable mobile environments from short segments using speaker and language recognizers. In: The Speaker and Language Recognition Workshop, pp. 301–311 (2012)
46. Bernardin, K., Stiefelhagen, R., Waibel, A.: Probabilisitic intergration of sparse audio-visual cues for identify tracking. In: ACM Multimedia, pp. 151–158 (2008)
47. Mesaros, A., Heittola, T., Klapuri, A.P..: Latent semantic analysis in sound event detection. In: EUSIPCO, pp. 1307–1311 (2011)
48. Chu, W.-T., Cheng, W.-H., Wu, J.-L.L.: Generative and discriminative modeling toward semantic context detection in audio tracks. In: MMM, pp. 38–45 (2005)
49. Cai, R., Lu, L., Hanjalic, A.: Unsupervised content discovery in composite audio. In: ACM Multimedia, pp. 628–637 (2005)
50. Galleguillos, C., Rabinovich, A., Belongie, S.: Object categorization using co-occurrence, location and appearance. In: CVPR, pp. 1–8 (2008)

51. http://www.sound-ideas.com/sound-effects/bbc-1-60-hd-sound-effects-library.html
52. Selina, C., Shri, N.S., Jay, K.C.-C.-C.: Environmental sound recognition with time-frequency audio features. IEEE Trans. Audio Speech Lang. Process. **17**(6), 1142–1158 (2009)
53. Giannoulis, D., Benetos, E., Stowell, D., Rossignol, M., Lagrange, M., Plumbley, M.D.: Detection and classification of acoustic scenes and events: an IEEE AASP challenge. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 1–4 (2013)