Audio Data Mining for Anthropogenic Disaster Identification: An Automatic Taxonomy Approach

Jiaxing Ye, *Member, IEEE,* Takumi Kobayashi, Xiaoyan Wang, *Member, IEEE,* Hiroshi Tsuda, and Masahiro Murakawa

Abstract—Disasters are undesirable and often sudden events causing human, material and economic losses, which exceed the coping capability of the affected community or society. In recent years, with significant advancement in information technology, various intelligent systems have been developed to support all aspects of disaster management, including emergency prediction, timely response and aftermath recovery. This paper addresses the anthropogenic disaster identification issue by exploiting audio big data mining. Specifically, a novel and efficient sound classification scheme is proposed, which is based on unsupervised acoustic feature learning and data-driven taxonomy. The proposed framework could accurately identify anthropogenic disaster events, e.g., gun shot, explosion, scream cry, etc. from dynamic audio data, and it consists of three major stages as follows. First, predominant acoustic patterns are characterized by dictionary learning algorithms, which can generate robust acoustic feature representations for recognition under noisy conditions. Second, hazard sound event taxonomy is created by exploiting probabilities distances between extracted sound dictionaries. Finally, taxonomy structure is embedded into hierarchical classification algorithm to improve classification. The Proposed approach is evaluated using real-world dataset with 10 emergency sound categories and 3275 clips. According to extensive experimental comparisons, proposed approach achieved state-of-the-art performance in anthropogenic disaster identification.

Index Terms—disaster management, audio surveillance, feature learning, taxonomy creation, data-driven approach.

1 Introduction

DISASTERS are the phenomenons that pose serious threat to people, economic assets or the functioning infrastructure of society. They are caused either by natural forces (known as natural disasters), or by human actions, negligence or errors (known as anthropogenic disasters). The natural disasters include tropical storms, floods, earthquakes, landslides, etc, While the anthropogenic disasters are generally classified into technological disasters, (e.g., engineering failures, transport disasters), and sociological disasters (e.g., criminal acts, riots) [1]. The damage of disasters can be significantly reduced by employing information technologies, such as developing and deploying geographic information systems, remote sensing and satellite data to predict natural disasters [2], [3] and building anomaly surveillance system using advanced information and communication technologies (ICTs) to accelerate emergency response [4], [5]. In this study, we focus on development of audio content retrieval approach to identify multiple types of man made disasters.

People living in both urban and rural areas are potentially exposed to the threats resulted of human intent or negligence. Literature statistics reveal a trend that the impact of man-caused disasters is increasing in recent years [6]. Concretely, there were nearly 7000 deaths in man-made disasters in 2015, compared to approximately 5900 in 2014. man made disasters can occur in every aspect of life. For instance, numerous transportation routes, including road, air, rail, and water have been regard as source of hazard. An accident could occur on any of these routes, and put lives, property and natural resources in danger [7]. Another example of man-made hazard garnering much attentions is terrorism movement, and it was reported that 23 countries recorded their highest number of deaths from terrorism in 2015 [8].

In order to reduce vulnerability to hazards, disaster/emergency management is commonly performed and the process has been coined into four key stages, which are warning phase ahead of disaster occurrence, immediate disaster detection and response, aftermath recovery phase and mitigating or preparedness phase that aiming at avoidance of future reoccurrences of same type disaster [1]. Among those four phases, early emergency prediction/detection is regarded as foremost step due to its significant effect on reduction of further loss from hazards in progress.

Modern information technologies, such as high-speed wireless networks, low-cost sensory device and efficient machine learning algorithms, enable us to establish intelligent surveillance system to detect emergence event efficiently. As for the sensory input, video and audio are most actively employed [9], [10] and we focus on use of audio information for hazard event recognition in this study. Audio-based surveillance exhibits couples of advantages over video-based approach and we summarize those metrics as follows:
• Audio data size is much smaller and hence we can collect multiple sensor data to increase detection coverage.
• Cameras are limited by angular field of view. On contrary, microphones can be omnidirectional and can collect acoustic

- *This study was partly supported by Cross-ministerial Strategic Innovation Promotion Program (SIP) / Technologies for maintenance, renewal, and management for infrastructure, and supported by the New Energy and Industrial Technology Development Organization (NEDO), Japan.*
- *J. YE, T.Kobayashi, H.Tsuda and M.Murakawa are with National Institute of Advanced Industrial Science and Technology, Tsukuba, Japan, 3058560 E-mail: {jiaxing.you, t.kobayashi, m.masahiro}@aist.go.jp*
- *X. Wang is with Department of Media and Telecommunications Engineering, Ibaraki University, Japan. E-mail: xiaoyan.wang.shawn@vc.ibaraki.ac.jp*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2017.2700843, IEEE Transactions on Emerging Topics in Computing

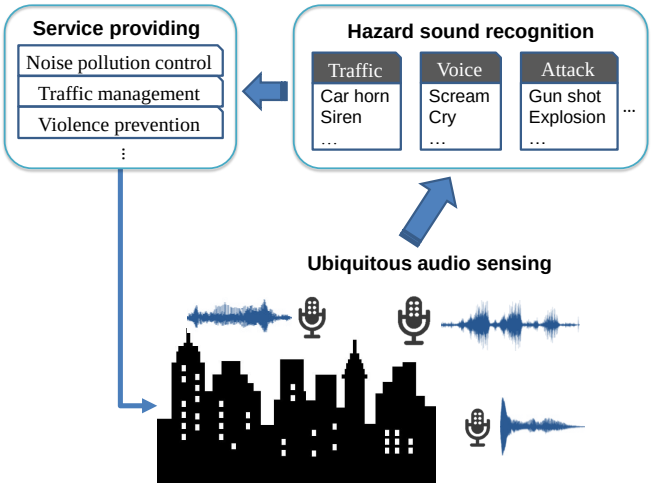


Fig. 1: Flowchart of the proposed acoustic scene classification approach

information with a spherical field of view.
• Image sensors suffer from illumination variations, while it is not an issues for acoustic sensory devices which can work constantly well in the day and night.
• For some specific hazard events, audio clues are more effective comparing with image, such as scream and gunshots.

The dynamic sound perspectives carry rich information in living environment, and therefore can be regarded as vital clues to indicate hazard events. On attempting to make safer and enjoyable lives for inhabitants, considerable research efforts have been devoted to the develop semantic audio data mining system for hazard detection, such as identification of scream and gunshot in urban area [11], [12] and recognition of different types of weapons' sound [13]. These audio-based disaster detection systems have been evaluated under varying environments, including offices [Kotus et al., 2013], elevators [Radhakrishnan et al., 2005b; Chua et al., 2014], and public transport vehicles [Pham et al., 2010; Rouas et al., 2006; Vu et al., 2006]. Besides research articles, several practical real projects had been carried out to promote real applications of audio-based disaster alert systems. For instance, in the European research project of EU-FP7-EAR-IT [17], which aims at minimizing risks in city traffic, a novel audio processing system have been developed for estimating the number of cars passing by in real time [18]. Meanwhile, the resultant traffic density measures had been exploited for air quality/urban noise monitoring [19]. In addition, a project developed an sound detector to identify sirens among street noises, well-designed schemes, such as changing traffic lights to help emergency vehicle pass through complex junctions quickly, can then be taken [20]. It is the case that exploring audio content information to accelerate emergence response. Audio surveillance is another major application field where acoustic signal is examined to discern man made disasters of violence conflicts or terrorism movements. Its general working process is as follows: once the incident is detected via audio content retrieval, an alarm can be immediately issued to notify local police officers and thus immediate intervention can be performed to control emergency event [21]. These research projects demonstrated acoustic modality can contribute to disaster management in practice. To summarize, we show a conceptual chart to demonstrate application of hazard sound recognition technique for man made disaster management in Fig.1.

In this study, we present novel hazard sound recognition framework with data-driven taxonomy. The fundamental idea is to create taxonomy to organize unstructured hazard sound data. The hierarchical formation can significantly facilitate browse, search and classification of acoustic patterns.In order to characterize predominant patterns in emergency sounds, unsupervised acoustic feature learning algorithms are employed. The methods can effectively extract effective feature that invariant to background noise. On creation of disaster sound taxonomy, we introduce probabilistic distance metrics in both Euclidean and Grassmannian spaces to quantize difference between hazard sound categories. A taxonomy can be subsequently built using well-defined categorical distance measures in an agglomerative fashion. At multi-class emergency sound recognition stage, we devise method to embed hierarchical dependencies in acoustic data into classification algorithm. We show through experimental comparisons that state-of-the-art hazard sound event recognition accuracy can be achieved with public acoustic dataset.

The contribution of this study can be summarized in three-fold:
† This study presents a novel framework emergency sound classification. The classification engine can be used to manage multiple man made disasters, such as traffic load assessment, violence conflict and terrorism movement detection.
† Hierarchies render efficient way to organize and retrieve unstructured data at multiple levels of granularity. In this study, we develop novel scheme to create taxonomy of acoustic events so as to improve hazard sound recognition.
† We propose improved formulation for hierarchical regularized logistic regression (HR-LR), to alleviate unbalanced class issue in classification model construction. The proposed formulation exhibits favourable performance in real-data evaluation.

Remainder of this paper is organized as follows. Section 2 delivers brief overview of current audio surveillance systems. Section 3 introduces the proposed hazard sound recognition system framework. The three key components: unsupervised acoustic feature learning, data-driven sound class taxonomy construction and hierarchical classification scheme are explained explicitly. Section 4 shows our evaluation dataset, validation protocol, experimental results and comparison analysis. Finally, in Section 5, we summarize our findings in this work.

2 Related Work

Audio content classification has been long standing research topic through decades, and the major objectives are speech and music [22], [23]. More recently, audio surveillance garnered increasing interests and plenty of research results have been reported [13], [24]. Those systems aim at

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2017.2700843, IEEE Transactions on Emerging Topics in Computing

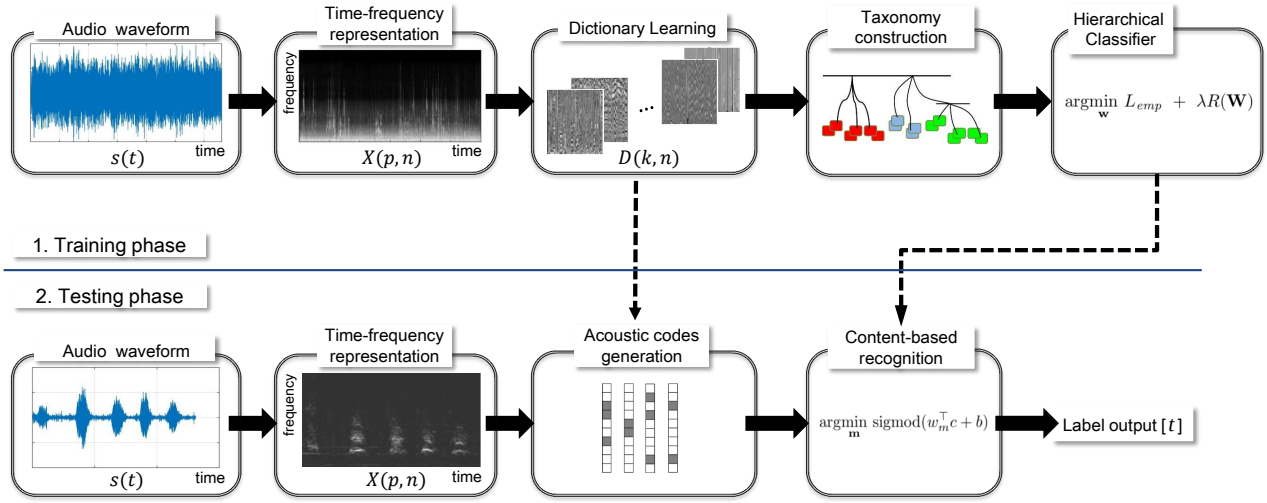


Fig. 2: Flowchart of the proposed audio-based hazard event recognition scheme

detecting potentially hazardous situations through audio-based monitoring in pubic space. Initial audio surveillance systems imitates frameworks of modern automatic speech recognition (ASR) or musical information retrieval (MIR) systems, such as applying standard MFCC feature to represent input sound and adopting GMM models to conduct content-based classification [10]. However, hazard sounds, such as crying or explosion, are inherently different from patterns in speech and music where existed predominant harmonicity or phonemic structures. As a result, standard ASR or MIR systems performed badly in recognizing hazard sound events. In addition, a throughout comparison of applying conventional audio processing techniques for sound event recognition has been conducted [25], in which extensively investigated acoustic features of spectrograms of Short-Time Fourier Transform (STFT), Continuous/Discrete Wavelet Transform (CWT/DWT) and MFCCs together with conventional classifiers, such as Artificial Neural Network (ANN) and Learning Vector Quantization (LVQ). Advanced classification schemes have been investigated for various sound content recognition lately, such as using Support Vector Machines (SVM) [26] to exploit non-linear distributions of acoustic events in projected feature space. Moreover, biological research reveals that local time-frequency information contributes greatly to human auditory, recent studies drew more attentions for audio representation development, and various novel acoustic features have been proposed, such as spikegram [27], a neural-spike-like representation of sound event, likewise, sparse audio features are presented in [28], which is robust to noise interferences. A key issue faced by the sound event classification research community is the lack of labeled data, which hampered comparison and reproducibility of research results. Lately, some efforts have been made to tackle such issue, and open datasets are released in regard to facilitate reproducibility of results, such as UrbanSound8K [29], ESC dataset [30] and DCASE2016 [31]. Inspired by the success of deep neural networks (DNN) in numerical application fields, e.g. computer vision and speech recognition applications, DNN and its variants of convolution neural networks (CNN) and recurrent neural networks (RNN) have been employed for content analysis of ambient sounds [32], [33]. However, lack of large-scale labelled sound event data is the practical issue that deteriorates DNN-based sound event recognition performance. The approaches to augment current dataset or develop novel data-efficient DNN models are critical concerns in application DNN to sound event identification [34].

This paper is an extension based on our previous works presented at 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) [35] and 2015 ACM international conference on Multimedia [36]. The key technical update is two-fold: first, new acoustic feature learning method of Spherical k-means is introduced to characterize distinctive patterns with respect to each emergency sound class. Second, we proposed novel approach to create categorical taxonomy to organize and classify unstructured hazard sound events. All details will be explained explicitly in following section.

3 Proposed System Architecture

The main goal of proposed system is to perform content recognition on emergency sounds. With the help of Internet Communication Technologies (ICTs), ubiquitous acoustic sensors can be deployed to multiple position for collecting audio data. The critical issue turns out to be content retrieval for incremental audio data. In this work, we propose novel scheme to understand acoustic contents with high accuracy and efficiency and the details are demonstrated in this section. First of all, we show a flow chart of propose hazard recognition framework in Fig.2.

3.1 Acoustic feature extraction

We first convert hazard sound waveform $s(n)$ to frequency domain via Discrete Fourier Transform (DFT), and the re-

sulting spectrogram is denoted as $X \in \mathbb{R}^{P\times N}$ with N frames and P frequency bins. Then, high pass filter with cut-off frequency of 200Hz is employed to eliminate low band noises. In order to smooth the spectrum as well as to reduce dimension, Mel filter bank is applied, which is identical to the one used in speech recognition [22]. The further process are based on bank-scale audio spectrogram representation.

3.2 Unsupervised feature learning for acoustic events

This section demonstrates our approach to hazard sound event feature learning based on sparse coding. To clarify the process, we begin with fundamental mathematical formulation of sparse coding as follows. given column-wise hazard event spectrogrm $X = [x_1, ..., x_N] \in \mathbb{R}^{P\times N}$ and we learn representative dictionary, which is denoted as $D = [d_1, ..., d_K] \in \mathbb{R}^{P\times K}$ and K columns referred to as dictionary *atoms*. At meantime, a series of sparse (audio) codes $C = [c_1, ..., c_N] \in \mathbb{R}^{K\times N}$ can be obtained such that input signal x_n can be well-approximated by several dictionary atoms, i.e. $x_n \approx \sum c_n d_n$. The values of sparse codes indicate significance of corresponding dictionary entries in signal reconstruction and "sparse" implies there are many zeros in codes c_n. To realize sparse coding from given data, a constraint matrix factorization problem is always presented as follows:

$$\min_{\rho\leq 1} \frac{1}{N}\sum_{n=1}^{N}\left[\underbrace{\frac{1}{2}||X-DC||_2^2}_{fitting\ term} + \underbrace{\gamma||C||_\rho}_{sparsity-inducing\ term}\right]. \quad (1)$$

In a nutshell, "fitting" terms assures D is good at representing input data X, meanwhile, "sparsity-inducing term", which is a l_ρ-norm, emphasizes that only few entries in D will be activated in recovery of X. It is a joint optimization with respect to dictionary D and the coefficients (codes) C of the sparse decomposition and standard optimizer can be employed to solve the problem. In this study, we evaluate two types of feature learning schemes, which are l_1-penalized sparse coding and Spherical k-means method.

3.2.1 Dictionary learning (DL) with l_1-penalty term

l_1-penalized sparse coding, which is also called Lasso estimator or basis pursuit, is most conventional formulation. In this paper, we first introduce Lasso sparse coding approach to characterize representative hazard sound patterns using compact dictionaries, at the mean time, input spectrogram is converted to sparse audio codes, which are anticipated to be more robust to noise. In the formation of Lasso estimation, that is, we solve optimization problem [37].

$$\min_{D,c(i)} \sum_i ||Dc^{(i)} - x^{(i)}||_2^2 + \gamma_1||c^{(i)}||_1, \quad \text{subject to } ||D^{(j)}||_2^2 = 1, \forall j \quad (2)$$

There are several well developed algorithms to minimize the objective function and we choose coordinate descend method [38].

3.2.2 Spherical k-means dictionary learning

More recently, a feature representation learning method called "spherical K-means" has been proposed and it achieved favourable performance in computer vision tasks [39]. We add this method in this evaluation due to its superior efficiency. This algorithm consists mainly 3 parts:
step 1: Input standardization:

$$x^{(i)} = \frac{x^i - \mu_i}{\sqrt{\delta_i + \epsilon_{norm}}} \quad \text{where } \mu_i = \frac{1}{N}\sum_n x_n^{(i)},\ \ \delta_i = \frac{1}{N}\sum_n (x_n^{(i)} - \mu_i)^2 \quad (3)$$

By subtracting sample mean (μ_i) and dividing by standard deviation (δ_i), feature variables now have zeros means and unit deviations (close to 1).

step 2: Whiten features to enhance subtle variations:

$$[V, D] = \text{eig}(\text{cov}(x)), \text{where } VDV^\top = \text{cov}(x) \quad x^{(i)} = V(D + \epsilon_{zca} I)^{-1/2} V^\top x^{(i)}, \forall i \quad (4)$$

Above process is coined as 'zero-phase component analysis' or ZCA whitening transform. Through eigenvalue decomposition of the data covariance matrix, the high-frequency noise can be suppressed, and thus, significant discriminant patterns can be characterized.

step 3: Building representative acoustic pattern dictionary

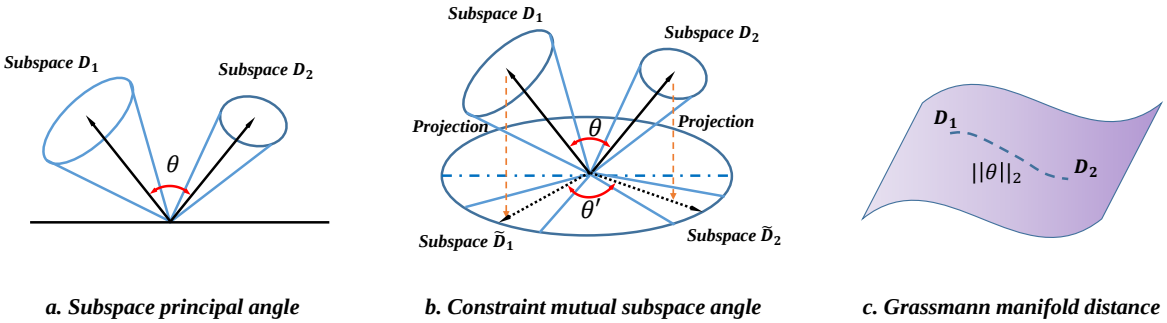
$$c_j^{(i)} = \begin{cases} D^{(j)\top} x^{(i)}, \text{ if } j == \underset{l}{\text{argmax}} |D^{(l)\top} x^{(i)}| \\ 0, \text{otherwise} \end{cases} \quad D = XC^\top + D \quad D^j = D^{(j)}/||D^{(j)}||_2, \ \ \forall j \quad (5)$$

Parameters of ϵ_{norm} and ϵ_{zca} are determined experimentally. Through performing above steps, compact dictionary for hazard sounds and a series of sparse codes can be obtained. Iterate until convergence (usually 10 iterations will be enough) to build statistically stable dictionary for one sound class. In sec. 4, we conducted extensive experiments on real data to compare two dictionary learning approaches for hazard sound events characterization.

3.3 Acoustic events taxonomy construction

Clustering analysis seeks efficient way to browse and organize unstructured data with tree hierarchy and the approach can improve recognition accuracy in contrast to flat classification fashion [40]. In this study, we conduct data-driven clustering analysis to build taxonomy of various types of emergency sounds. The brief process is demonstrated as follows. Based on aforementioned process of acoustic feature learning, a set of dictionaries can be extracted from all types of acoustic events. On data-driven taxonomy creation, critical issue is to select appropriate distance metric between representative dictionaries of hazard sound events. We evaluate three effective set-to-set distance measures that are derived from Euclidean and Grassmannian geometry to investigate between-dictionary distances. Grounded on the optimal distance measure selection, we adopt agglomerative approach to build up taxonomy in an bottom-up manner. Since the dictionary-to-dictionary distance metric played

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2017.2700843, IEEE Transactions on Emerging Topics in Computing



a. Subspace principal angle *b. Constraint mutual subspace angle* *c. Grassmann manifold distance*

Fig. 3: Metrics for dictionary-to-dictionary distance measure

central role in creating taxonomy, we show details of those metrics in subsections.

3.3.1 Principle angles between mutual subspaces

There are several metrics to investigate similarities between subspace and the most fundamental one is principle angles. It had been successfully applied for various application, including face recognition [41] and video tracking. We introduce the measure to evaluation similarity between acoustic event dictionaries. Let $D_1,\ D_2 \in \mathbb{R}^{P\times K}$ be two subspaces with identical dimensionality. The principal angles, or canonical angles, $0 \leq \theta_1 \leq ..... \leq \theta_k \leq \pi/2$ between D_1 and D_2 are recursively defined for $k = 1, ..., K$ by:

$$\begin{aligned} \cos\theta_k = \max_{\mathbf{u_k}\in \mathrm{span}(D_1)} \max_{\mathbf{v_k}\in \mathrm{span}(D_2)} \mathbf{u_k}'\mathbf{v_k}, \quad \text{subject to} \\ \mathbf{u_k}'\mathbf{u_k} = 1, \quad \mathbf{v_k}'\mathbf{v_k} = 1, \\ \mathbf{u_k}'\mathbf{u_j} = 0, \quad \mathbf{v_k}'\mathbf{v_j} = 0, \quad k \neq j \end{aligned} \tag{6}$$

where $\mathbf{u_k}'\mathbf{v_k}$ are called the k-th pair of canonical vectors. The standard approach to compute principle angles between subspaces is to introduce SVD, in which

$$\begin{aligned} D_1'D_2 = USV', \quad \text{where } U'U = I, V'V = I \\ S = \mathrm{diag}(\cos^2\theta_1, ..., \cos^2\theta_m) \end{aligned} \tag{7}$$

$\cos\theta_i$ is the cosine of the ith principal angle. $\cos\theta_1, ..., \cos\theta_d$ are known as canonical correlations and the maximum eigenvalue of decomposition represented by $\cos\theta_1$ indicates the minimum canonical angle θ_1. We denote $\Theta = [\theta_1, ..., \theta_d]$. Finally, the distance between mutual subspaces is defined as

$$l_{MSM} = \frac{1}{K}\sum_{p=1}\cos^2\theta_p \tag{8}$$

The $l_{MSM} \in [0, 1]$ score exhibits following characteristics, if two subspaces coincide perfectly, l_{MSM} is 1; on the other hand, in the case of two orthogonal subspaces, l_{MSM} will be 0. The value reflects structural similarities.

3.3.2 Distance defined as mutual constraint subspaces angle

The main drawback of principle angle metric is that it is vulnerable to within-class variations. Since environmental sound always merged with background noises, it is necessary to eliminate effect of variations that are irrelevant to between subspace difference. In pursuit of between-subspace discriminant power, in a sense analogous to linear discriminant analysis (LDA), constraint mutual subspace method (CMSM) had been proposed [42] and the major process is presented as follows.

Let $D_1, ..., D_M$ denote M acoustic dictionaries (subspaces) with K entries of P dimension, and we define the concatenated matrix as $D_{all} = [D_1, ..., D_2]$. CMSM seeks a joint projection that subspace-to-subspace difference can be revealed. We denote the projection vectors as V, which are composed of the eigenvectors of the smaller eigenvalues in eigendecomposition:

$$D_{all}D_{all}^{\top}V_{CMSM} = V_{CMSM}\Sigma, \quad s.t.,\ V_{CMSM}^{\top}V_{CMSM} = I, \tag{9}$$

where $\Sigma = \mathrm{diag}(\{\delta\})$ is the eigenvalue diagonal matrix. According to [42], it is shown that project vectors of smaller eigenvalues are based on the differential vectors between subspace and therefore the projection can greatly facilitate between-subspace discrimination. Through V_{CMSM}, we can obtain new representation of input subspaces

$$\tilde{D_m} = V_{CMSM}^{\top}D_m \tag{10}$$

Subsequently, principle angles can be measured between projected categorical dictionaries $[\tilde{D_1}, ..., \tilde{D_M}]$ in constraint subspace using eq(8).

3.3.3 Grassmann manifold distance metrics

Grassmann manifold $\mathcal{G}(K, P)$, which is defined as a set of K-dimensional linear subspaces in $\mathbb{R}^P$, is another formulation to conduct subspace-based learning. Grounded on intrinsic geometry of Grassmann manifold, the geodesic distance is introduced to measure the length of the geodesic curve connecting two subspaces along the Grassmannian surface [43] and several effective distance metrics have been further developed [44], such as geodesic distance (arc length):

$$l_{geodesic}(D_1, D_2) = ||\Theta||_2, \tag{11}$$

where Θ represents canonical angles, and Chordal distance (projection F-norm):

$$l_{proj}(U_1, U_2) = ||U_1U_1^{\top} - U_2U_2^{\top}||_2 \tag{12}$$

According to comparison study reported in [45], projection F-norm Grassmann distance delivered favourable performance across multiple tests, and therefore we select this metric in this work. It is noteworthy that Grassmannian distances are closely related to principal angles, which are discussed in previous section. However, the conceptual formulations are completely different. Grassmannian distances manifests distances of two subspaces in the embedding space (Grassmannian manifold), while principal angles exploit subspace-similarity in each individual dimension.

3.4 Hierarchical classification scheme

In order to take advantage of learnt acoustic events taxonomy for efficient hazard sound identification, we employ hierarchical regularized logistic regression (HR-LR) model to performance classification. The method can leverage the dependencies in class hierarchy to boost performance. In addition, to tackle the imbalanced class issue, we propose improved formulation for HR-LR and the details are presented as follows.

Hierarchical classification problems admit a general optimization objective which consists of empirical loss and model penalty:

$$\underset{\mathbf{w}}{\operatorname{argmin}}\ L_{emp}\ +\ \lambda \times R(\mathbf{W}). \tag{13}$$

Concretely, in the hierarchical regularized logistic regression (HR-LR) model [46], empirical loss L_{emp} was defined as the loss incurred by the instances at every leaf-nodes:

$$L_{emp} = \sum_{n\in T}\sum_{i=1}^{M} L(y_{in}, c_i, w_n) \tag{14}$$

On the other hand, the learnt hazard acoustic events taxonomy was embedded into regularization term $R(\mathbf{W})$, which incorporates hierarchy structure of acoustic data as well as enhancing generalization power of the model for unseen data

$$R(\mathbf{W}) = \sum_{n\in N}\frac{1}{2}||w_n - w_{\pi(n)}||^2. \tag{15}$$

Intuitively, such setting enhances the hierarchical dependencies in the sense that it encourages node and its parent to adopt similar weights (close to each other in euclidean norm).

Putting two parts together, we have HR-LR formulation:

$$\min_{\mathbf{w}} \sum_{n\in T}\sum_{i=1}^{M} \text{sigmod}(y_{in} w_n^\top c_i)\ +\ \lambda \sum_{n\in N}\frac{1}{2}||w_n - w_{\pi(n)}||^2 \tag{16}$$

However, there is one critical drawback in current formulation, that is, if sample numbers varies widely for each sound category, the model will be biased towarding majority class because objective function endeavours to minimize quantity of sample-wise error rate, not taking overall data distribution into account. To alleviate imbalanced classes matter, we revise empirical loss part by assigning class weights denominator. It can be regarded as that now we focus on class-wise recognition accuracy and thus majority class no long overwhelms the minority. The newly proposed formulation is as follows:

$$\min_{\mathbf{w}} \sum_{n\in T}\frac{1}{M}\sum_{i=1}^{M} \text{sigmod}(y_{in} w_n^\top c_i)\ +\ \lambda \sum_{n\in N}\frac{1}{2}||w_n - w_{\pi(n)}||^2 \tag{17}$$

Since the objective function of HR-LR is convex and differentiable, second order methods are applicable to perform optimization. In concern of dealing with large scale audio data, we employ LBFGS algorithm in this study.
The corresponding gradient Gd can be computed in closed-form as

$$Gd = \mathbf{w_n} - \mathbf{w}_{\pi,\mathbf{n}} - \frac{1}{M}\sum_{i=1}^{M}\frac{1}{1+\exp(y_{in} w_n^\top c_i)} y_{in} c_i \tag{18}$$

We summarize optimization routine in algorithm 1.

Algorithm 1 HR-LR optimization

Input: **C** λ, π, T, M, N
Result: model parameter $\tilde{W}$

```
procedure HR-LR
    Initialize (W_0)
    while not converaged do
        Solve eq.(18) using LBFGS
        W_i ← W_{i-1}.
    end while
    W̃ ← W_i
end procedure
```

TABLE 1: Audio corpus for evaluation

Category	Number of clips	Average duration (sec)
car horn	429	3.36
dog bark	1000	3.40
gun shot	374	3.29
siren	929	3.31
cough	80	8.35
cry	60	6.94
alarm	60	5.50
explosion	118	4.12
scream	125	6.41
swords	100	3.27

4 Experimental validation

4.1 Dataset and settings

We validate proposed emergency sound identification system through extensive experiments using real corpus composed of various types of emergency event sounds. The evaluation dataset was constructed from the following compilations: (i) BBC Sound Effects Library [47], (ii) UrbanSound8K datasets [29], (iii) ESC: Dataset for Environmental Sound Classification [30], (iv) sound effects from internet sources[1]. By incorporating multiple datasets, we obtain various types of hazard sounds with high variation and wide diversity. In total, we extracted 10 classes of emergency sound events with 3275 audio clips. For all sound clips, the sampling rate was set to 16 kbps with 16 bit resolution.

1. http://sound.natix.org/

TABLE 2: Parameter settings

name of parameter	presence equation	value
Fourier window length	T (1)	23.2ms
Balance weight	γ_1 (4)	0.2
Dictionary size	K (5)	30
normalization intercept	ϵ_{norm}(6)	0.01
whitening intercept	ϵ_{zca}(7)	0.05
regularization coeff.	λ (16)	0.2

Details of samples numbers of each event category and averaged clip length are given in Table 1. It is noteworthy that the UrbanSound8K dataset contributed as major part of our evaluation corpus and original excerpts are taken from website of Freesound[2] — an online sound repository containing over 160,000 user-uploaded recordings. Those crowd sourcing audio clips were mixed with all kinds of background noises with varying intensities and thus the dataset is well-suited to assess robustness of proposed audio content retrieval scheme. In the experimental validation, we set Fourier analysis window length to 23.2ms (1024 points) with half overlapping. 60 Mel-filters were applied to convert spectrogram to mel-scale. In tab.2, we summarize all parameters used in our processing.

In our algorithmic settings, several parameters were involved in both acoustic feature extraction and multiclass emergency sound classification stages. Those parameters were critical to systematic performance. For example, during feature learning, we prefer to construct class-wise dictionary to express predominant patterns in hazard sounds, meanwhile a series of robust sparse codes were expected. The trade-off between two demands were manipulated by a balancing parameter γ_1 in eq.(4), which propagates our preference of model and can be turned through experiments. Likewise, regularization coefficient of λ in eq.(16) controls the model characteristic leaning to fewer loss in fitting to evaluation corpus or to smaller generalization error in coping with out-of-sample data. In summary, we listed out all hyperparemeters applied in our model in Table 2. We exhibit classification rates using boxplots of 10-folder cross-validation in all experiments, making sure there is no overlapping between training and test data in each evaluation iteration.

4.2 Demonstration of learnt acoustic features and taxonomy

In this section we show intermediate results of our audio-based disaster system. Through examining those outputs some insights can be derived in acoustic pattern analysis of hazard sound events. In Fig.4, we first exhibit the example of learnt representative acoustic pattern dictionaries with respect to each type of hazard sounds using Spherical k-means method. Class-wise dictionary can be regarded as a concise representation of acoustic data since significant patterns are effectively encoded. In Fig. 4, we can observe with-in class similarities for one dictionary, meanwhile, between class discrimination can also be seen. For instance, car horn sounds looks more similar to scream class and multiple impulsive feature patterns can be found in both gun shot and cough dictionaries due to their short-time characteristic. Grounded on those compact representations (dictionary matrix), we further investigated between-dictionary distance metrics using methods demonstrated in sec. 3.2 and further construct hazard sound class hierarchy in an agglomerative fashion. We show the extracted emergency sound taxonomy using Grassmann distance metric in Fig. 5. According to the bottom-up combinational structure, gun shot and explosion sounds exhibit highest similarities compared to other classes, while car horn is distinctly dispersed from other acoustic categories. Based on such data-driven dependency, we further establish fine classification hyperplane to classify multiple disaster sound events using the algorithm discussed in sec. 3.4. Notably, by changing between-dictionary distance metric, different taxonomy formulation can be extracted. To ensure the metric inherently reflect class dependencies, we conducted extensive experiments to validate optimal distance metric for taxonomy construction and the results were shown in following section.

2. http://www.freesound.org/

4.3 Evaluation of feature learning and dictionary-to-dictionary distance metric

This section covers our evaluation result on dictionary learning and taxonomy construction for hazard sound events. We tested two dictionary learning algorithms, which are Lasso sparse coding and Spherical k-means method. Together with three set-to-set distance metrics including subspace angle, constraint mutual subspace angle and Grassmann manifold distance. In order to evaluate performance of those methods, we conducted extensive experiments using test dataset. At classification stage, hierarchical regularized logistic regression (HR-LR) was employed and results were demonstrated with mean averaged precision (mAP) across all hazard sound categories. The evaluation results were exhibited in Fig. 6, including all six combinational methods' performance. By results ranking, we can see distance metric defined on Grassmann manifold is most effective to model dependencies between multiple sound classes. Set-to-set distance measures defined in Euclidean space, i.e. principle angles and constrained subspace angles, generated inferior performance due to their vulnerability to wide variation and high noise in sound events. In addition to precision comparison, efficiency is another critical concern for real applications, especially under the context of big-data era. In our system, highest computation load is consumed at feature dictionary learning stage since it is necessary to process millions frame of acoustic spectrum. We compared the two methods: Lasso sparse coding and Spherical k-means. Although incremental improvements had been made to improve efficiency of Lasso sparse coding, it is still much slower compared to latter method, because L1 regularized data fitting is inherently complex. In contrast, Spherical k-means is rather simple which involves only several steps of matrix multiplication, and with the help of whitening processing, Spherical k-means approach outperformed Lasso sparse coding method in both classification accuracies and computation efficiency.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2017.2700843, IEEE Transactions on Emerging Topics in Computing

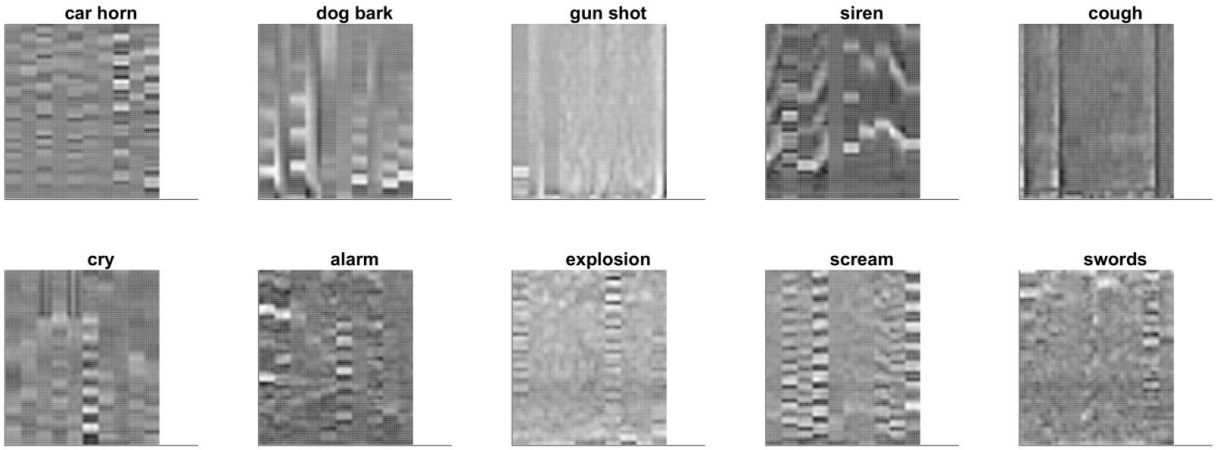


Fig. 4: Learnt representative pattern dictionaries for 10 classes of emergency sound events

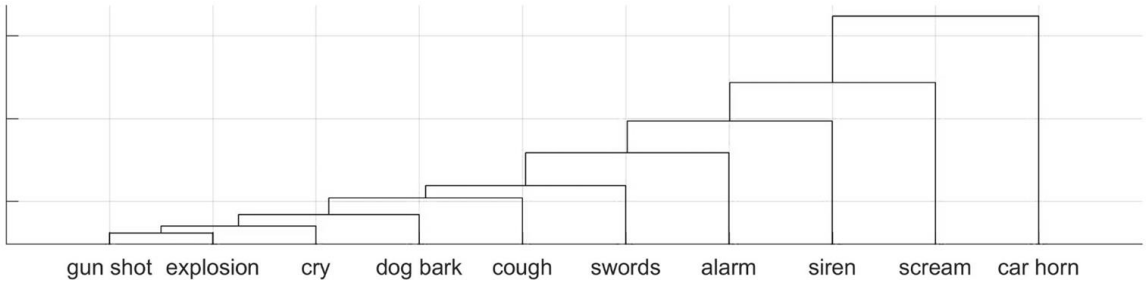


Fig. 5: hazard event taxonomy created by using spherical k-means dictionary learning and Grassmann distance metric

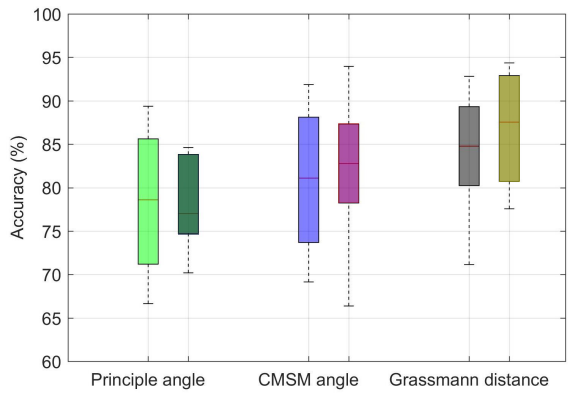


Fig. 6: Evaluation results of feature learning and between-dictionary distance metrics for taxonomy construction (results were organized by distance metrics and in each group, first and second results corresponds to using Lasso sparse coding and Spherical k-means dictionary learning methods, respectively)

4.4 Performance analysis proposed man made hazard sound identification approach

In our last experiment, we validated the performance of proposed framework through comparison with reference methods [48]. In Fig. 7, we presented comparison results on class-wise hazard sound identification accuracies. Proposed achieved superior performance in recognizing all categories of emergency sounds according to experiments. Particularly, for 6 cases of acoustic events out of total 10, we obtain identification accuracies exceeded 90%, they are *car horn*, *dog bark*, *gun shot*, *siren*, *cough* and *explosion*. The improvement is derived from mixture of key components in proposed framework, including acoustic feature dictionary learning, set-based distance measurement for sound event taxonomy creation as well as the taxonomy-embedded hierarchical classification algorithm. It is noteworthy that the two hazard event classes are *cry* and *alarm*, with both categories consists only 60 samples. The limited data collection may lead to high variance issue in classification model training, and therefore it can be anticipated that by providing more clips, the identification precision of the two classes can be further improved. In summary, proposed framework achieved 87.34% of mean averaged precision (mAP) for disaster sound identification and outperformed the reference method with large margin that delivered mAP as 82.80%.

5 Discussion

The results we showed above demonstrate that supervised hierarchical classification scheme is highly competitive when it comes to detect hazard sounds in ambient

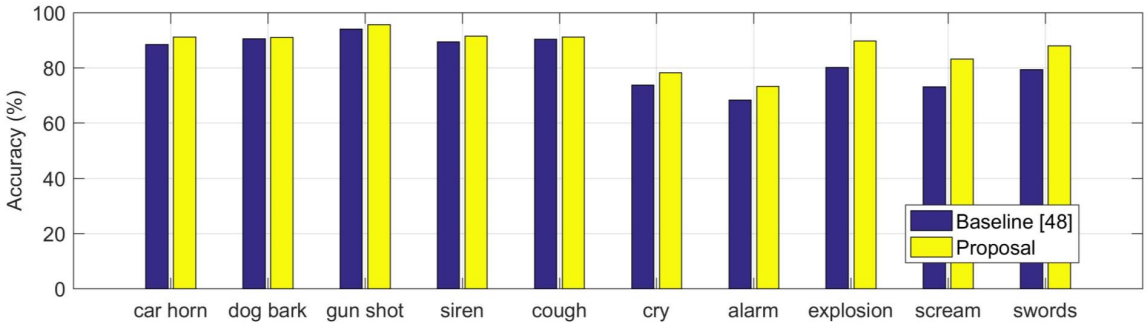


Fig. 7: Comparison of class-wise recognition accuracies for hazard sound recognition

environment. However, it is small-scale validation using the current datasets combination. For instance, once a piece sound recording of unseen event enters the content retrieval engine, a mis-classification maybe issued. It is a challenge for all supervised pattern analysis systems and a rejection option is one reasonable measure to deal with unseen pattern matter. In addition, it is worth noting that Deep Neural Networks (DNN) approaches have garnered much of interests in machine learning field and the method had been extensively applied for acoustic scene analysis and sound event classification [32], [33]. Nevertheless, according to latest evaluations, the DNN methods has not generated significant higher results than its predecessors, i.e. the systems using hand-crafted acoustic features (MFCC) with classical SVM or random forest classifier so far. One crucial reason is that size of current sound event dataset is rather small compared with the ones used for imaging parsing/natural language processing. To tackle the limited audio data issue, one possible solution might be employing transfer learning strategies for DNN based systems where part of the DNN can be initially learned by a large amount of external audio data. It will also be long-lasting research topic cross multiple fields.

6 Conclusion

This paper presented novel approach to investigate ambient sound for protecting lives, property and the economy from anthropogenic disasters. Specific sounds, e.g. screaming, shouting, gun shout and explosion, are high related to anthropogenic disasters of violence conflict, accident or even terrorism movements. Compared to video surveillance, audio information can be quite effective to indicate occurrence of those hazard events and therefore win time for immediate response and reduce the damage of disaster. To effectively characterize acoustic clues for early disaster detection, we developed emergency sound recognition framework based on acoustic feature learning and data-driven taxonomy creation. Feature learning methods were adopted to extract distinctive feature from emergency sound events. Subsequently, we developed an automatic taxonomy construction approach to facilitate multi-class hazard sound event identification. Our scheme was validated by using crowd souring audio dataset. Experimental results demonstrated the effectiveness of the proposed hazard sound recognition method.

Acknowledgments

The authors would like to thank the anonymous reviewers for their constructive comments and suggestions on previous versions of this article.

References

[1] Centers for Disease Control and Prevention (CDC), *A Primer for Understanding the Principles and Practices of Disaster Surveillance in the United States: First edition*, Atlanta (GA): CDC, 2016.

[2] Dan Chen, Zhixin Liu, Lizhe Wang, Minggang Dou, Jingying Chen, and Hui Li, "Natural disaster monitoring with wireless sensor networks: a case study of data-intensive applications upon low-cost scalable systems," *Mobile Networks and Applications*, vol. 18, no. 5, pp. 651–663, 2013.

[3] Jochen Zschau and Andreas N Küppers, *Early warning systems for natural disaster reduction*, Springer Science & Business Media, 2013.

[4] Neeraj Kumar, Jong-Hyouk Lee, and Joel JPC Rodrigues, "Intelligent mobile video surveillance system as a bayesian coalition game in vehicular sensor networks: learning automata approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 3, pp. 1148–1161, 2015.

[5] Sarvesh Vishwakarma and Anupam Agrawal, "A survey on activity recognition and behavior understanding in video surveillance," *The Visual Computer*, vol. 29, no. 10, pp. 983–1009, 2013.

[6] Christopher M. Bishop, *Natural catastrophe and man-made disasters in 2015: Asia suffers substantial losses*, SWISS RE, Secaucus, NJ, USA, 2016.

[7] Reza Faturechi and Elise Miller-Hooks, "Measuring the performance of transportation infrastructure systems in disasters: a comprehensive review," *Journal of infrastructure systems*, vol. 21, no. 1, pp. 04014025, 2014.

[8] The Institute for Economics and Peace, *2016 Global Terrorism Index*, New York, NY 10022,USA, 2016.

[9] Ying-Li Tian, Max Lu, and Arun Hampapur, "Robust and efficient foreground analysis for real-time video surveillance," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. IEEE, 2005, vol. 1, pp. 1182–1187.

[10] Marco Crocco, Marco Cristani, Andrea Trucco, and Vittorio Murino, "Audio surveillance: A systematic review," *ACM Comput. Surv.*, vol. 48, no. 4, pp. 52:1–52:46, Feb. 2016.

[11] Luigi Gerosa, G Valenzise, M Tagliasacchi, F Antonacci, and A Sarti, "Scream and gunshot detection in noisy environments," in *Signal Processing Conference, 2007 15th European*. IEEE, 2007, pp. 1216–1220.

[12] Mahesh Kumar Nandwana, Ali Ziaei, and John HL Hansen, "Robust unsupervised detection of human screams in noisy acoustic environments," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 161–165.

[13] Chloé Clavel, Thibaut Ehrette, and Gaël Richard, "Events detection for an audio-based surveillance system," in *2005 IEEE International Conference on Multimedia and Expo*. IEEE, 2005, pp. 1306–1309.

[14] Aki Harma, Martin F McKinney, and Janto Skowronek, "Automatic surveillance of the acoustic activity in our living environment," in *2005 IEEE International Conference on Multimedia and Expo*. IEEE, 2005, pp. 4–pp.

[15] J. Kotus, P. Dalka, M. Szczodrak, G. Szwoch, P. Szczuko, and A. Czyzewski, "Multimodal surveillance based personal protection system," in *2013 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, Sept 2013, pp. 100–105.

[16] Teck Wee Chua, Karianto Leman, and Feng Gao, "Hierarchical audio-visual surveillance for passenger elevators," in *Proceedings of the 20th Anniversary International Conference on MultiMedia Modeling - Volume 8326*, New York, NY, USA, 2014, MMM 2014, pp. 44–55, Springer-Verlag New York, Inc.

[17] Congduc Pham and Philippe Cousin, "Streaming the sound of smart cities: Experimentations on the smartsantander test-bed," in *Proceedings of the 2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing*, Washington, DC, USA, 2013, GREENCOM-ITHINGS-CPSCOM '13, pp. 611–618, IEEE Computer Society.

[18] Gyrgy Nagy, Rene Rodigast, and Danilo Hollosi, "Energy based traffic density estimation using embedded audio processing unit," in *Audio Engineering Society Convention 136*, Apr 2014.

[19] Susanne Steinle, Stefan Reis, and Clive Eric Sabel, "Quantifying human exposure to air pollutionmoving from static monitoring to spatio-temporally resolved personal exposure assessment," *Science of The Total Environment*, vol. 443, pp. 184 – 193, 2013.

[20] Congduc Pham, "Ear-it, acoustic sensing in smart environment: a case for audio streaming with low-resource iot devices," in *9th Senzations Summer School on IoT Applications*, Sept 2014.

[21] Pasquale Foggia, Nicolai Petkov, Alessia Saggese, Nicola Strisciuglio, and Mario Vento, "Reliable detection of audio events in highly noisy environments," *Pattern Recognition Letters*, vol. 65, pp. 22 – 28, 2015.

[22] Sven L Mattys, Matthew H Davis, Ann R Bradlow, and Sophie K Scott, "Speech recognition in adverse conditions: A review," *Language and Cognitive Processes*, vol. 27, no. 7-8, pp. 953–978, 2012.

[23] Yi-Hsuan Yang and Homer H Chen, "Machine recognition of music emotion: A review," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, no. 3, pp. 40, 2012.

[24] Regunathan Radhakrishnan, Ajay Divakaran, and A Smaragdis, "Audio analysis for surveillance applications," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005.* IEEE, 2005, pp. 158–161.

[25] M. Cowling and R. Sitte, "Comparison of techniques for environmental sound recognition," *Pattern Recognition Letters*, vol. 24, no. 15, pp. 2895–2907, 2003.

[26] X. Zhuang, X. Zhou, M. A. Hasegawa-Johnson, and Thomas S. Huang, "Systematic acquisition of audio classes for elevator surveillance," *Real-world acoustic event detection*, vol. 31, no. 12, pp. 1543–1551, 2010.

[27] X. Valero and F. Alas, "Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification," *Multimedia, IEEE Transactions on*, vol. 14, no. 6, pp. 1684–1689, 2012.

[28] S. Chu, S. Narayanan, and C.C.Jay Kuo, "Environmental sound recognition with time-frequency audio features," *Speech and Audio Processing, IEEE Transactions on*, vol. 17, no. 6, pp. 1142–1158, 2009.

[29] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *22th ACM Int. Conf. on Multimedia*, Nov 2014.

[30] Karol J. Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM International Conference on Multimedia*, New York, NY, USA, 2015, MM '15, pp. 1015–1018, ACM.

[31] Tuomas Virtanen, Annamaria Mesaros, Toni Heittola, Mark D. Plumbley, Peter Foster, Emmanouil Benetos, and Mathieu Lagrange, *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, Tampere University of Technology. Department of Signal Processing, 2016.

[32] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, Sept 2015, pp. 1–6.

[33] Giambattista Parascandolo, Heikki Huttunen, and Tuomas Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 6440–6444.

[34] Justin Salamon and Juan Pablo Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, 2016.

[35] Jiaxing Ye, Takumi Kobayashi, Masahiro Murakawa, and Tetsuya Higuchi, "Robust acoustic feature extraction for sound classification based on noise reduction," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 5944–5948.

[36] Jiaxing Ye, Takumi Kobayashi, Masahiro Murakawa, and Tetsuya Higuchi, "Acoustic scene classification based on sound textures and events," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 1291–1294.

[37] Adam Coates and Andrew Y. Ng, "The importance of encoding versus training with sparse coding and vector quantization," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. 2011, pp. 921–928, ACM.

[38] Tong Tong Wu and Kenneth Lange, "Coordinate descent algorithms for lasso penalized regression," *The Annals of Applied Statistics*, pp. 224–244, 2008.

[39] Adam Coates and Andrew Y Ng, "Learning feature representations with k-means," in *Neural Networks: Tricks of the Trade*, pp. 561–580. Springer, 2012.

[40] Carlos N Silla Jr and Alex A Freitas, "A survey of hierarchical classification across different application domains," *Data Mining and Knowledge Discovery*, vol. 22, no. 1-2, pp. 31–72, 2011.

[41] Osamu Yamaguchi, Kazuhiro Fukui, and K-i Maeda, "Face recognition using temporal image sequence," in *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*. IEEE, 1998, pp. 318–323.

[42] Kazuhiro Fukui and Osamu Yamaguchi, "Face recognition using multi-viewpoint patterns for robot vision," in *Robotics Research. The Eleventh International Symposium*. Springer, 2005, pp. 192–201.

[43] Alan Edelman, Tomás A Arias, and Steven T Smith, "The geometry of algorithms with orthogonality constraints," *SIAM journal on Matrix Analysis and Applications*, vol. 20, no. 2, pp. 303–353, 1998.

[44] Jihun Hamm and Daniel D Lee, "Extended grassmann kernels for subspace-based learning," in *Advances in neural information processing systems*, 2009, pp. 601–608.

[45] Jihun Hamm and Daniel D Lee, "Grassmann discriminant analysis: a unifying view on subspace-based learning," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 376–383.

[46] Siddharth Gopal and Yiming Yang, "Recursive regularization for large-scale classification with hierarchical and graphical dependencies," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pp. 257–265.

[47] British Broadcasting Corporation. and Films for the Humanities, *BBC Sound Effects Library*, 1977.

[48] J. Salamon and J. P. Bello, "Unsupervised feature learning for urban sound classification," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 171–175.

PLACE PHOTO HERE

Jiaxing Ye received the B.E and M.E. degrees in communication engineering from Harbin Institute of Technology, China and Ph.D in computer science from the University of Tsukuba in 2012. He is a research scientist at National Institute of Advanced Industrial Science and Technology (AIST), Japan. His research interest includes pattern recognition and signal processing, with focus on audio analysis, and time-series processing. He is a member of the IEEE.

PLACE PHOTO HERE

Takumi Kobayashi received master of engineering from University of Tokyo in 2005, and doctor of engineering from University of Tsukuba in 2009. He is currently a research scientist at the National Institute of Advanced Industrial Science and Technology (AIST), Japan. His research interest includes pattern recognition, multivariate analysis, and their applications to image processing and computer vision.

PLACE PHOTO HERE

Xiaoyan Wang received the B.E. degree in Automation from Beihang University, China, and the M.E. and Ph.D. degrees in Computer Science from University of Tsukuba, Japan. Currently he is working as the assistant professor at the Department of Media and Telecommunications Engineering, Ibaraki University, Japan. His research interest covers cognitive radio networks, network security and privacy, and cooperative communications.

PLACE PHOTO HERE

Hiroshi Tsuda received the Ph.D. degrees in Engineering from University of Tokyo, in 1994. He has been worked at National Institute of Advanced Industrial Science and Technology (AIST), Japan and has been leader of non-destructive measurement group since 2010. His major research interest is development of the optical measurement systems.

PLACE PHOTO HERE

Masahiro Murakawa received the B.E., M.E. and Ph.D. degrees from University of Tokyo, in 1994, 1996, and 1999, respectively. He is group leader of artificial intelligence applications research team, National Institute of Advanced Industrial Science and Technology (AIST), Japan. His research interests include evolutionary algorithms, neural networks, and reinforcement learning. Concurrently, he is a professor at cooperative graduate school, University of Tsukuba.