Applied Acoustics 117 (2017) 246–256

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Urban sound event classification based on local and global features aggregation ☆

Jiaxing Ye *, Takumi Kobayashi, Masahiro Murakawa

National Institute of Advanced Industrial Science and Technology, Tsukuba, Japan

ARTICLE INFO

Article history:
Received 30 August 2015
Received in revised form 19 July 2016
Accepted 1 August 2016
Available online 10 November 2016

Keywords:
Urban sound classification
Crowd-sourcing
Signal processing
Machine learning
Dictionary learning
Feature aggregation

ABSTRACT

The automatic content-based classification of complex and dynamic urban sound is an important aspect of various emerging applications, such as surveillance, urban soundscape understanding and noise source identification, therefore the research topic has gained a lot of attention in recent years. The aim of this paper is to develop efficient machine learning-based scheme for urban sound classification in real-life noise conditions. Unlike conventional sound event classification methods that mainly address local temporal-spectral patterns, we propose an aggregation scheme to combine both local and global acoustic features. For characterizing local patterns, we employ feature learning method to extract class-dependent temporal-spectral structures; on the other hand, long-term descriptive statistics are employed to exploit global features of sound events, e.g. variability and recurrence, which also carry rich discriminant information. In order to aggregate the heterogeneous acoustic information for classification, we introduce mixture of experts model (MoE) which effectively formulates relationship between local and global information. At validation stage, we conduct experiments on *UrbanSound8K* database which consists of 10 categories of urban sound events with 8732 real-world clips. It is noteworthy that the 10 classes of crowdsourced recordings, including air conditioner, car horn, children playing, dog bark, drilling, engine idling, gunshot, jackhammer, siren and street music, are most common urban sounds closely related to urban life. According to experimental results, the proposed scheme achieved superior performance compared with 3 other latest approaches and it can be a fundamental building block of various urban multimedia information processing systems that help to improve quality of life.

1. Introduction

Rapid worldwide urbanization poses serious challenge to human society for building highly liveable and sustainable cities to increase inhabitants [1]. In order to provide efficient urban services with low costs and reduced resource consumption, the concept of smart city has been proposed and it is still evolving to represent the interdisciplinary fields where information science meet conventional city-related issues, such as energy, environment and health care [2,3]. The advances in mobile computing technologies and Internet of Things (IoT) enable us to collect massive urban acoustic data related to both environment and dwellers activities [4], such as children playing, road traffic and even gun shot, etc. Through investigating the contents of dynamic sound perspectives together with time and location information, a better understanding of sound issues affecting daily lives of citizens can be obtained and this understanding could be the foundation for improvement of quality of life in cities.

Whenever urban sound is being discussed, the noise health effects and legislation to reduce noise are always major issues [5,6]; however, in the context of smart cities, urban sound, which can be generated by both environment and human activities, is regarded as valuable information source about city's need. On attempting to make a smarter life in cities, several pioneering applications based on sound content information retrieval have been recently evaluated. For instance, in the EU-FP7 project EAR-IT [7], in order to enhance city traffic management, novel audio analysis methods have been developed for estimating the number of cars passing in real time [8]. Such traffic density information is also vital for air quality monitoring as it is used in conjunction with pollution detectors [9]. Audio surveillance is one of the most extensively studied urban acoustic application [10,11], which typically employs sound content analysis techniques to detect specific emergency urban sound, such as gunshot and screaming, in living environment [12]. As a violence incident is detected, an alarm can

☆ Manuscript for Applied Acoustics: Special Issue on Acoustics in Smart Cities.

* Corresponding author.
E-mail addresses: jiaxing.you@aist.go.jp (J. Ye), takumi.kobayashi@aist.go.jp (T. Kobayashi), m.murakawa@aist.go.jp (M. Murakawa).

http://dx.doi.org/10.1016/j.apacoust.2016.08.002
0003-682X/

be immediately transmitted to local police officers and hence fast intervention can be performed by emergency personnel [13]. These research projects demonstrated the potential of using acoustic modality to provide useful services in the city. Content-based urban sound classification can also contribute to noise control. In nowadays, the mainstream methods to characterize annoying environmental sounds are grounded on measurement of time-averaged A-weighted sound pressure levels [14] and the noise source information is often neglected. However, latest research results revealed that people's acoustic comfort is determined not only by noise level, but also by noise source [15]. According to comparison study presented in [16], typical urban noises, such as traffic noise from rail, road and aircraft, the neighborhood and industry noises, present different impacts on health-related quality of life (HRQOL). Sound content classification technique can be employed to retrieve noise source information, which is a counterpart to intensity level. More efficient noise assessment can be performed by taking both noise level and type into account, which could greatly facilitate urban noise management. Besides, computational urban sound classification can facilitate soundscape assessment and modeling [17,18]. Since Truax coined the concept of soundscape in 1978 as 'an environment of sound (sonic environment) with emphasis on the way it is perceived and understood by the individual, or by a society' [19], lots of studies had been carried out in the acoustic research field in pursuit of "standardization for perceptual assessment of human sound preference" [20,21]. Significant progresses were made by 'Soundscapes of European Cities and Landscapes' project, which was performed from 2009 to 2013 and dedicated for urban soundscape research on various aspects, e.g. definition and categorization, (psycho-) acoustical assessment and modeling [15,22]. More recently, automatic sound recognition techniques have been employed for soundscape research [23,24]. The basic process is to make automatic classification of soundscapes by pattern recognition algorithms and then ascertain how certain type of soundscape is perceived by the population exposed to it, e.g. negative or positive effects [25,26]. Based on above review of current applications of sound classification techniques to provide various services for city inhabits, We present a general framework in Fig. 1 which describes usage of sound classification to meet information needs in cities.

In this paper, we propose novel sound classification framework which aggregates local temporal-spectral features and global structures to recognize urban sound events with higher accuracy. In most previous studies, e.g. the ones shown in last paragraph, specific sound analysis system had been developed for given task, such as gun shot detection or noise source recognition. We propose an universal classification scheme that can effectively classify multiple types of urban sounds and the system can work for various applications. Conventional sound event analysis mainly addresses local time-frequency dynamics only, e.g. pitch shifting and format patterns. However, the long-term structures, such as variability and recurrence properties, also carry rich discriminant acoustic information. For example, mechanical sounds, which are generated by reciprocating or rotational motions, usually exhibit homogeneous repetitive patterns that should be exploited for classification. To demonstrate the property, in Fig. 2, we present six types of urban sounds: three non-mechanical sounds (a–c) and three mechanical sounds (d–f). It is evident that recurrent structures are discriminant characteristics for identifying mechanical sounds (d–f). The proposed method takes advantage of both local patterns and global structural information for achieving better classification performance. The main contribution of this paper is threefold:

- This paper presents an effective urban sound classification framework which effectively combines local and global acoustic information through conditional fusion scheme. Grounded on the fact that importance of local and global features varies with respect to different urban sound classes, we employ product of experts model [27] to conditionally fuse two level information. According to extensive experiments on real-world data, proposed scheme outperformed state-of-the-art approaches.
- For characterizing local patterns in sounds, we adopt unsupervised feature learning model which have been successfully applied to objective recognition in computer vision [28]. The model encodes representative patterns of sounds events with compact basis set (dictionary) and subsequently applies a mapping (coding) to generate new feature representation with respect to the dictionary. By doing so, feature learning method favorably generates content-based representations to facilitate classification. Since coding scheme is deemed crucial for dic-

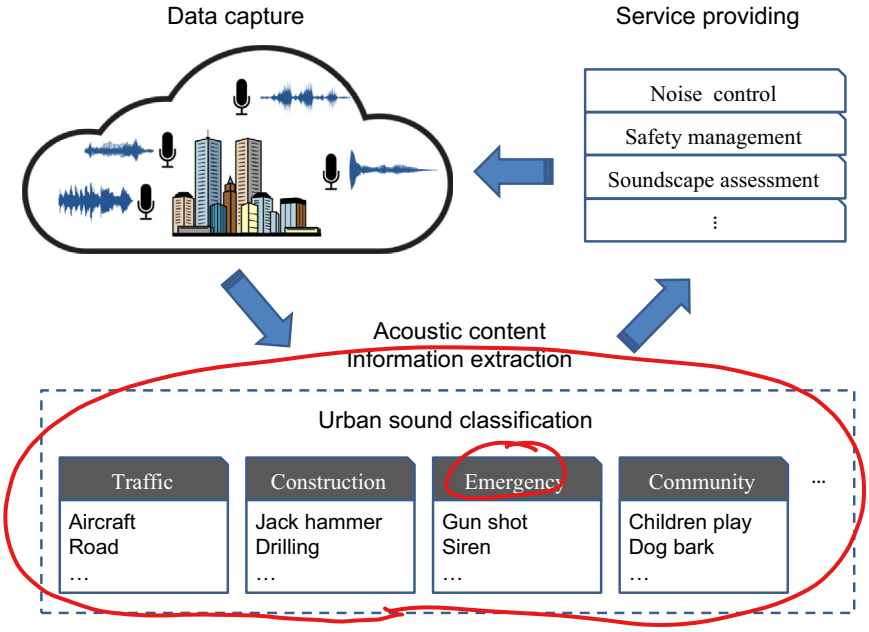

Fig. 1. General framework of content-based sound classification for city life service.

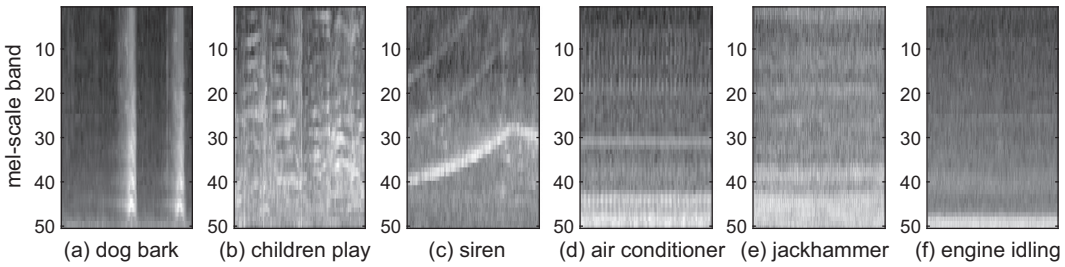

Fig. 2. Demonstration of recurrence patterns in mechanical sounds.

tionary learning [29], we carefully design a soft coding method which delivers better urban sound classification performance in this study.

- Global structures in sounds, such as repetition and variability, carry distinctive information for acoustic event classification. We employed entropy and recurrence quantification analysis (RQA) measures to characterize long-term characteristics of urban sound [30]. The contribution of global feature is empirically demonstrated in the experiments.

This paper is organized as follows: in Section 2, we present a literature review of current urban sound classification methods; in Section 3, we introduce proposed framework, including feature characterization methods for local and global patterns and an aggregation scheme to combine those two level information for urban sound classification; in Section 4, we further demonstrate effectiveness of proposed system by using *UrbanSound8K* [31] dataset based on extensive experiments and the comparison results are shown; a discussion on possible ways to improve the quality of life of city inhabits using urban sound classification techniques in presented in Section 5; finally, in Section 6, we make conclusion of this study.

2. Literature review of urban sound classification methods

Motivated by various innovative applications, the automatic content-based sound classification is regarded as the essential aspect of urban informatics for characterizing complex, dynamic and massive urban acoustic environment, and therefore has garnered increasing research attention in recent years [32–35]. Unlike speech and music processing, where appropriate frameworks of automatic speech recognition (ASR) and music information retrieval (MIR) have been well established [36,37], research towards sound event analysis is still at early stage. Since urban sound events include all kinds of heterogeneous acoustics arise from city sonic environment, the patterns can vary widely. In order to characterize dynamic signal, development of efficient acoustic feature representation is the primary problem. Audio signal processing techniques, e.g. time-frequency representations of spectrogram and Mel-frequency cepstral coefficients (MFCCs) are initially applied [38]. Some study suggested that Gammatone frequency cepstral coefficients (GFCCs) produced better results for describing sound events [39]. Sparse coding algorithms have been introduced to generate robust acoustic features to noise [40]. Bag-of-words model, which had been proved effective for natural language processing, had been investigated for soundscape classification, however, according to latest report, is not sufficient model for the task [41]. Unsupervised feature learning techniques, in the light of their success in computer vision applications, become popular for sound content analysis [28,42]. After extracting sound spectrogram, unsupervised feature learning further extracts distinctive temporal-spectral structures, which greatly improves classification performance. Image patch descriptor, such as Histogram of Gradients (HoG) [35] and Local Binary Patterns (LBP) [43], are also employed to characterize local sound textures in time-frequency domain. Based on acoustic features, classifiers are employed to perform content-based classification in the feature space. Generative Gaussian models, such as Naive Bayes Classifier (NBC) and Gaussian Mixture Models (GMM) are initially evaluated [38]. Other conventional classification schemes, such as K-Nearest Neighbors, Neural Networks (NN) and Decision Tree (DT) had also been evaluated for soundscape classification tasks [23]. However, these systems are vulnerable to the presence of noises due to property of Gaussian model. Recently, discriminant models, such as support vector machines (SVMs) [44,17] and random forests [33], are extensively used to deliver better classification precision. A key issue faced by the sound classification research community is the lack of labeled audio data, which hampered comparison and reproducibility of research results. To tackle this problem, some efforts have been made. For example, several evaluation campaigns, such as CLEAR 2007 and AASP CASA 2013, have been launched by the research community [45,38,44]. Some large annotated datasets, such as *UrbanSound8K* [31] and ESC50 [46], have been released to public. In this study, we evaluate our approach by using *UrbanSound8K* dataset, which consists large amount of labeled real world sound data with various single to noise (SNR) condition and it is suitable to evaluate the performance of sound classification method in practice.

3. Proposed framework

In this section, we introduce three major components of the proposed framework: 1. Local spectro-temporal patterns characterization by discriminant feature learning. 2. Global structural information extraction by using descriptive statistics. 3. Mixture of Experts model (MoE) [27] for two-way information aggregation. At first, we present general formulation of the proposed feature fusion scheme as follows:

$$l_c^{\text{all}} = \frac{1}{Z} l_c(\mathbf{f}_{\text{local}})^{\alpha_c} \times l_c(\mathbf{f}_{\text{global}})^{1-\alpha_c}, \quad Z = \int l_c(\mathbf{f}_{\text{local}})^{\alpha_c} \times l_c(\mathbf{f}_{\text{global}})^{1-\alpha_c} dc, \tag{1}$$

where $l_c(\mathbf{f}_{\text{local}})$ and $l_c(\mathbf{f}_{\text{global}})$ denote class membership scores estimated based on local acoustic features $\mathbf{f}_{\text{local}}$ and global acoustic features $\mathbf{f}_{\text{global}}$, respectively. Z is a partition function and subscript c is sound class index. Weighting factor α_c represents relative importance of two series of class scores for final class score l_c^{all} computation. Formula (1) interprets our fundamental hypothesis that both global and local patterns deliver critical information for sound event classification. Particularly, α_c governs the contributions of counterpart information, which can vary with respect to sound classes. We introduce process of computing class scores of $l_c(\mathbf{f}_{\text{local}}), l_c(\mathbf{f}_{\text{global}})$ as well as estimating fusion weights α_c in following three subsections.

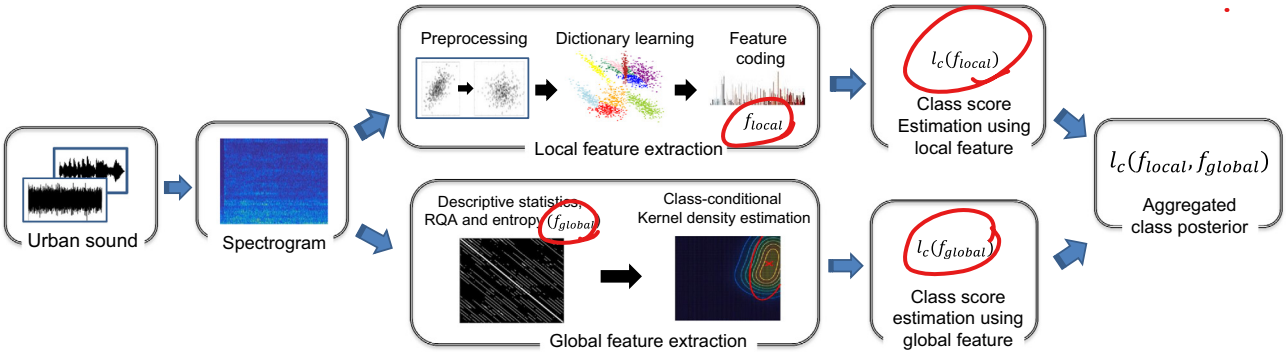

Fig. 3. General framework of urban sound analysis for city life service.

To summarize the flow of proposed method, an overview flowchart is presented in Fig. 3.

3.1. Class score estimation based on local feature

Finding good spectro-temporal feature representations for sound events has been long standing issue for sound classification [11,39,42]. In this paper, we employ feature learning scheme to characterize representative local time-frequency patterns. Dictionary learning was originated from Bag-of-Words model for document classification research, due to its superior performance in image classification [47], the method has been introduced for sound classification lately [42]. In this work, we modify BoF model with a well-designed coding function to characterize discriminative local structures in sound events. BoF model usually composes of 4 steps: local feature extraction with pre-processing, dictionary generation, encoding and pooling. The details are presented as follows.

3.1.1. Local feature extraction with pre-processing

Modern sound analysis is commonly based on time-frequency representations of audio signal, i.e. spectrogram. In this study, we denote spectrogram as $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_t, \ldots \mathbf{x}_T] \in \mathbb{R}^{F\times T}, t \in [1, \ldots, T]$, where F and T denote the number of Mel-scale frequency bands and time frames, respectively, and t-th frame spectrum is noted by $\mathbf{x}_t$. Notably, to reduce frequency dimension, Mel-filer bank is adopted. It is recognized that pre-processing is critical for dictionary learning approaches [29]. We introduce pre-processing method adopted for urban sound feature extraction. We first normalize $\mathbf{X}$ to zero mean and unit variance spectrogram by following formula:

$$\mathbf{x}'_t = \frac{\mathbf{x}_t - \text{mean}_x}{\sqrt{\text{var}_x + const}}, \tag{2}$$

where $\mathbf{x}_t$ and $\mathbf{x}'_t$ denote the spectrum before and after pre-processing, respectively. mean_x and var_x are the mean and variance of $\mathbf{X}$. Whitening transform is subsequently applied to decorrelate frequency bands, such operation can greatly help feature learning process. Let $\mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}} = \text{cov}(\mathbf{x}')$ is the eigenvalue decomposition of spectrogram covariance and $\mathbf{V}, \mathit{\Lambda}$ are eigen vectors and corresponding eigenvalues, respectively. The whitened spectrogram can be extracted through:

$$\hat{\mathbf{x}} = \mathbf{V}(\mathbf{\Lambda} + \epsilon\mathbf{I})^{-1/2}\mathbf{V}^{\mathrm{T}}\mathbf{x}', \tag{3}$$

where ϵ is an adjustable parameter to enhance numerical stability.

3.1.2. Class-conditional dictionary learning

Through pre-processing, whitened spectrogram can be obtained, which is denoted by $\hat{\mathbf{X}} \in \mathbb{R}^{F\times T}$. To learn a (small) set of basis functions for representing temporal-spectral patterns in audio spectrogram, we employ K-means method [28] and the objective function is defined as follows

$$\underset{\mathbf{D},\mathbf{s}}{\text{minimize}} \sum_t \|\mathbf{D}\mathbf{s}_t - \hat{\mathbf{x}}_t\|_2^2, \quad \text{subject to} \|\mathbf{s}_t\|_0 \leqslant 1, \quad \|\mathbf{D}_j\|_2 = 1 \tag{4}$$

The goal is to find a dictionary $\mathbf{D}$ and a new representation (code) $\mathbf{s}_t$ to reconstruct each input spectrum vector $\hat{\mathbf{x}}_t$ well. The first constraint tells that each $\mathbf{s}_t$ is forced to have at most one non-zero entry, and second constraint scaled dictionary entries to have unit length. The estimation of $\mathbf{D}$ and $\mathbf{s}_t$ can be done by running following iterations until convergence:

$$s_{t,j} = \begin{cases} \mathbf{D}_j^{\mathrm{T}}\hat{\mathbf{x}}_t & \text{if } j == \underset{k}{\text{argmax}} \left|\mathbf{D}_k^{\mathrm{T}}\hat{x}^i\right| \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$$\mathbf{D} = \hat{\mathbf{X}}\mathbf{S}^{\mathrm{T}} + \mathbf{D}, \quad \mathbf{D}_j = \mathbf{D}_j / ||\mathbf{D}_j||_2, \tag{6}$$

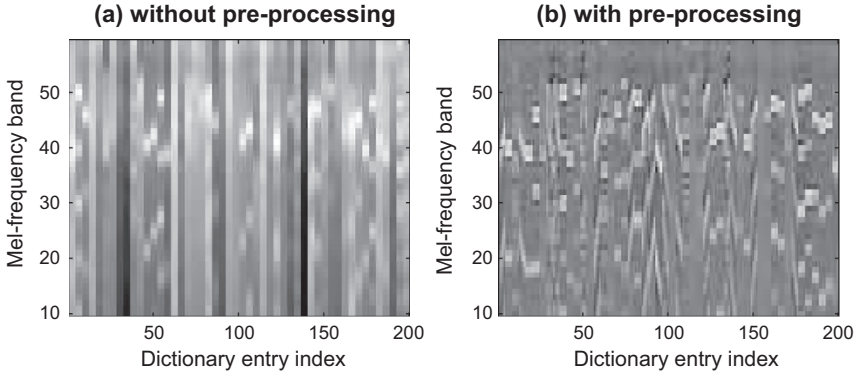

Fig. 4. (a) Dictionary learnt by spherical K-means from siren sound without whitening. (b) Dictionary learnt for siren sound from whitened spectrogram.

where K denotes number of entries in dictionary, which is required to be determined in advance. $\mathcal{D}_k$ denotes k-th ($k \in [1, K]$) entry in the dictionary. Notably, dictionary learning is performed in a class-conditional manner, in which $\mathbf{D}$ is extracted for every class in a supervised fashion and overall dictionary can be obtained by concatenating all dictionaries learnt for individual classes. In addition, size of dictionary K significantly affects the classification performance. We will revisit this parameter in experiments section.

To demonstrate effectiveness dictionary learning, we present the learnt dictionary for siren sounds in Fig. 4. Moreover, to clarify contribution of whitening pre-processing, we show the learnt dictionaries with and without pre-processing applied in Fig. 4(b) and (a), respectively. From Fig. 4(b), it is obvious that, with whitening process, K-means method favorably captures distinctive local structures of siren sounds.

3.1.3. Encoding by soft thresholding

According to experimental studies in [29], soft-threshold coding performed quite well for dictionary learning, which achieved competitive performance even compared to much complex methods, i.e. sparse coding. Therefore, we apply soft thresholding to encode input data (sound spectrogram) by using learnt dictionary. The coding function is defined as follows:

$$\tilde{s}_t^k = \max\left\{0,\ \mu_{\mathbf{x}}^{\mathbf{D}} - \tau - \left\|\hat{\mathbf{x}}_t - \mathbf{D}_k^{\mathrm{T}}\hat{\mathbf{x}}_t\right\|_2^2\right\} \tag{7}$$

where $\mu_{\mathbf{x}}^{\mathbf{D}}$ is mean of all $\left\|\hat{\mathbf{x}}_t - \mathbf{D}_k^{\mathrm{T}}\hat{\mathbf{x}}_t\right\|_2^2, t \in [1, T]$ over t, τ is gating intercept for soft coding. The formula interprets idea that activation function outputs zero value for any code $\tilde{s}_t^k$ where the distance to dictionary entry $\left\|\hat{\mathbf{x}}_t - \mathbf{D}_k^{\mathrm{T}}\hat{\mathbf{x}}_t\right\|_2^2$ exceeds threshold of average minus τ. Since urban sounds are always noisy, we introduce τ to enhance the sparsity in codes (more codes will be set to 0). It is noteworthy that our coding method is different from the one used in [42], in which a simple linear coding scheme of $\tilde{s}_t^k = \mathbf{D}_k^{\mathrm{T}}\hat{\mathbf{x}}_t$ is applied. By introducing such non-linear gating function, robust codes can be generated for classification.

3.1.4. Pooling for concise feature vector

Through the trained dictionary $\mathbf{D} \in \mathbb{R}^{F\times K}$ and coding scheme, we can transform input (whitened) audio spectrogram $\hat{\mathbf{X}} \in \mathbb{R}^{F\times T}$ into codes – a content-based representation preferred for classification, which is denoted by $\tilde{\mathbf{S}} \in \mathbb{R}^{K\times T}$ and $\tilde{s}_t^k$ is t-th input frame's code corresponds to k-th dictionary entry. However, the dimension of code matrix ($K \times T$) is too high to handle directly. In order to reduce the dimension of audio codes, pooling is commonly piled. In this study, we adopt maximum, average and standard deviation pooling, which are most extensively applied [48]. The computation procedures are shown as follows:

$$\text{avg pooling}: \mathbf{f}_{\text{avg}} = \frac{1}{T}\sum_{t=1}^{T}\tilde{\mathbf{s}}_t; \quad \text{max. pooling}: \mathbf{f}_{max} = \max_{\mathrm{t}}\left\{\tilde{\mathbf{S}}\right\} \tag{8}$$

$$\text{std pooling}: \mathbf{f}_{\text{std}} = \left(\frac{1}{T-1}\sum_{t=1}^{T}\left|\tilde{\mathbf{s}}_t - \boldsymbol{\mu}_{\tilde{s}}\right|^2\right)^{1/2},$$

$$\text{where } \boldsymbol{\mu}_{\tilde{s}} = \frac{1}{T}\sum_{t=1}^{T}\tilde{\mathbf{s}}_t. \tag{9}$$

After pooling, we concatenate pooled feature to construct local feature vector for urban sound: $\mathbf{f}^{\text{local}} = [\mathbf{f}_{\text{avg}}^{\mathrm{T}}, \mathbf{f}_{\text{max}}^{\mathrm{T}}, \mathbf{f}_{\text{std}}^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{3K\times 1}$.

3.1.5. Class score estimation by ProbSVM

To estimate class membership scores based on local features, we adopt probabilistic support vector machine (probSVM). The probSVM converts the label outputs of SVM to class conditional probabilities through further establishing logistic regression model in (kernel) feature space. The objective is presented by following formula:

$$\min_{A,B}\frac{1}{N}\sum_{n=1}^{N}\log\left(1+\exp\left(-y_n(A(\mathbf{w}'_{svm}\Phi(\mathbf{f}_n^{\text{local}}) + b_{svm}) + B)\right)\right), \tag{10}$$

where $\mathbf{f}_n^{\text{local}}, y_n$ is the n-th training sample (local feature vector) and label pair. By fitting the probSVM model to training data, all parameters can be inferred and we add subscript of L to indicate that the model is constructed for local feature analysis, i.e. $\mathbf{w}_L, b_L, A_L, B_L$. Finally, class posterior probability can be derived for input feature by:

$$l_c(\mathbf{f}_{\text{local}}) = \text{sigmoid}(A_L(\mathbf{w}'_L\Phi_L(\mathbf{f}^{\text{local}}) + b_L) + B_L). \tag{11}$$

3.2. Class score estimation based on global feature

According to the result of recent IEEE AASP challenge on detection and classification of acoustic scenes and events, global features presented competitive results for sound classification [38]. Especially, the recurrence quantification analysis (RQA) measures achieved best result in sound scene classification [49]. To investigate repetitive pattern in urban sounds, we employ recurrence quantification analysis (RQA) measure and spectrogram entropy in this study.

3.2.1. Spectrogram entropy

Based on hypothesis that recurrent structures will make a sound more predictable (with lower uncertainty), we introduce spectrogram entropy as a one global feature. Computation of spectrogram entropy is introduced as follows. As in previous notation, $X \in \mathbb{R}^{F\times T}$ denotes sound spectrogram. We introduce principal component analysis approach to enhance the band-wise variance of audio signal, which can be solved via eigenvalue decomposition of spectrogram, that is $V\Lambda V^{\mathrm{T}} = \text{cov}(\mathbf{x})$ [50]. Eigenvalues, denoted by $\text{diag}\ (\Lambda) = [\lambda_1, \ldots, \lambda_F]$, indicate spectrogram variabilities after principal component projection. And we further define spectrogram entropy as:

$$H = -\sum_{d=1}^{F}\lambda'_d \ln\lambda'_d, \quad \lambda'_d = \frac{\lambda_d}{\sum_{d=1}^{F}\lambda_d}. \tag{12}$$

In general, a sound contains more repetitive patterns would generate lower H value. By contrast, dynamic sounds such as *children play* and *street music* are anticipated to exhibit bigger H.

3.2.2. Recurrence quantification analysis (RQA)

RQA is effective approach for describing recurrent patterns in signal, which was originated from the field of chaos and complex systems analysis [30]. The basic idea is to quantify recurrence behavior of the phase space trajectory of dynamic systems. RQA is based on recurrence plots (RP), which reveals all the times when the phase space trajectory of the dynamical system visits roughly the same area in the phase space. It can be mathematically expressed as

$$R_{t_1,t_2} = \Theta(\sigma - \|\mathbf{x}_{t1} - \mathbf{x}_{t2}\|), \quad t1, t2 \in [1, T], \tag{13}$$

where $\mathbf{x}_{t1}, \mathbf{x}_{t2}$ are spectrum vectors of $t1/t2$-th frame. σ is cutoff distance and Θ is heaviside function. The cutoff distance σ defines a sphere centered at x_{t1}, and if x_{t2} falls within this sphere, i.e., the state is close to x_{t1}, then $R_{t1,t2} = 1$; otherwise $R_{t1,t2} = 0$. We can

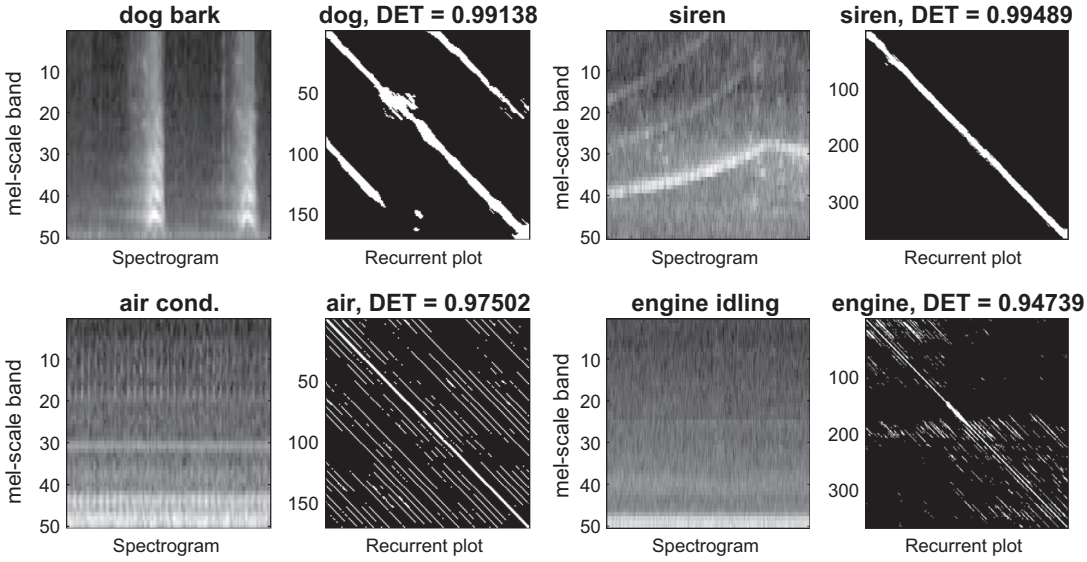

Fig. 5. Recurrent plots (RP) and RQA determinism measures for typical urban sounds.

examine basic recurrent patterns through focusing on diagonal lines in PR, which present evolution of states (trajectories in phase space) is similar at different times. Furthermore, RQA measures of determinism (DET) is designated to quantify periodicities through investigating ratio of recurrence points on the diagonal (of at least length $l_{\min}$) to all recurrence points. The formula of DET is:

$$\mathrm{DET} = \frac{\sum_{l=l_{\min}}^{N} l \cdot p(l)}{\sum_{t1,t2}^{N} R_{t1,t2}}, \tag{14}$$

where $p(l)$ is histogram of the lengths l of the diagonal lines. To demonstrate effectiveness of RQA measure, we present RP charts and DET values of several types of urban sounds in Fig. 5. From the charts, we can observe repetitive patterns in mechanical sounds (second row) can be effectively described by RQA.

3.2.3. Class score estimation by Kernel Density Estimation (KDE)

This section presents our path to estimate class membership probability of input global features. The general idea is to use posterior probability as class score for classification and to estimate posterior based on global features, Bayes' law is introduced [50]:

$$l_c(\mathbf{f}_{\text{global}}) = p(c|\mathbf{f}_{\text{global}}) = \frac{p(\mathbf{f}_{\text{global}}|c) \cdot p(c)}{p(\mathbf{f}_{\text{global}})}. \tag{15}$$

The posterior $p(c|\mathbf{f}_{\text{global}})$ can be computed based on likelihood $p(\mathbf{f}_{\text{global}}|c)$ and prior $p(c)$. It is noteworthy that the prior $p(c)$ should be carefully inferred onsite for real applications, because occurrence of sound has strong relation to tempo-spatial contexts. To estimate $l_c(\mathbf{f}_{\text{global}})$, we apply Kernel Density Estimation (KDE) method a supervised manner.

Let us denote global feature vector as $\mathbf{f}^n_{\text{global}} = [\mathrm{H}^n, \mathrm{DET}^n]^{\mathrm{T}}$, where H^n and DET^n are spectrogram entropy and RQA measures, respectively. n denotes sound sample index. The likelihood inference through KDE can be expressed as

$$\begin{aligned} p(\mathbf{f}_{\text{global}}|c) &= \frac{1}{N}\sum_{n=1}^{N} Ker_h(\mathbf{f}^c_{\text{global}} - \mathbf{f}^{c,n}_{\text{global}}) \\ &= \frac{1}{Nh}\sum_{n=1}^{N} Ker\left(\frac{\mathbf{f}^c_{\text{global}} - \mathbf{f}^{c,n}_{\text{global}}}{h}\right), \end{aligned} \tag{16}$$

where $Ker(\cdot)$ is a kernel function, $h > 0$ is a smoothing parameter commonly called bandwidth. The subscript h denotes it is scaled kernel that is defined as $Ker_h(f) = 1/h \cdot Ker(f/h)$. Based on prior and class-conditional probability density functions (PDF) inferred by MLE using training data, we can estimate $l_c(\mathbf{f}_{\text{global}})$ (posterior) for input.

3.3. Conditional aggregation of local and global information for classification

Through local and global feature investigation, we derive two series of class scores: $l_c(\mathbf{f}_{\text{local}})$ and $l_c(\mathbf{f}_{\text{global}})$. According to formula (1), we need to further estimate class-conditional fusion weights α_c to compute final class score. To this end, we can rewrite (1) by using log-sum-exp trick and obtain following expression:

$$l^{\text{all}}_c = \frac{1}{Z}\exp\left(\alpha \cdot \log(l_c(\mathbf{f}_{\text{local}})) + (1-\alpha)\cdot\log(l_c(\mathbf{f}_{\text{global}}))\right). \tag{17}$$

The fusion weighting factor α_c reveals relative importance between local and global features corresponds to certain type of urban sound denoted by subscript c. Parameter α_c can be effectively estimated at cross-validation stage through mining following objective function:

$$\min_{\alpha_c}\sum_{i=1}^{C}\sum_{j=1}^{N_c}\left(y_{ij} - (\alpha_c \cdot -\log(l_{c_i}(\mathbf{x}^{ij}_{\text{local}})) + (1-\alpha_c)\cdot -\log(l_{c_i}(\mathbf{x}^{ij}_{\text{global}})))\right)^2, \tag{18}$$

where y_{ij} denotes label of j-th sample in i-th sound class, $l_{c_i}(\mathbf{x}^{ij}_{\text{local}})$ and $l_{c_i}(\mathbf{x}^{ij}_{\text{global}})$ represent probabilities of successfully classifying j-th sample from i-th sound category through characterizing local and global information, respectively. It is obvious that formula (18) can be solved analytically. By substituting the inferred optimal α into (17), final class membership probability can be computed. Class label can be derived through $\underset{c}{\mathrm{argmin}}(-l^{\text{all}}_c)$.

4. Experiments

4.1. Dataset and parameters

In order to validate the proposed approach, we conducted experiments using *UrbanSound8K* dataset [31] which includes ten

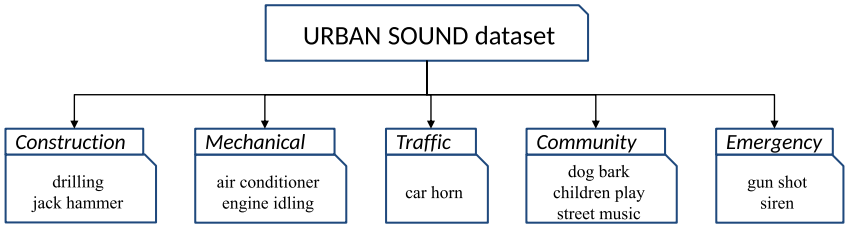

Fig. 6. Taxonomy of sound classes in URBAN SOUND dataset.

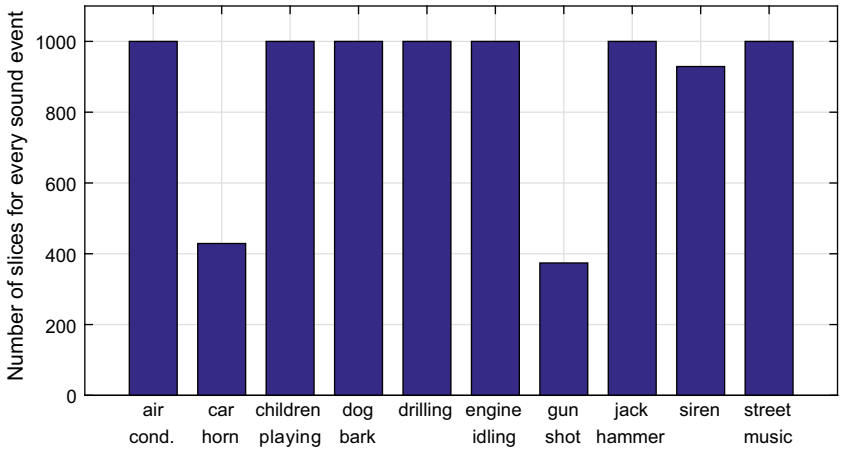

Fig. 7. Number of samples for every sound class.

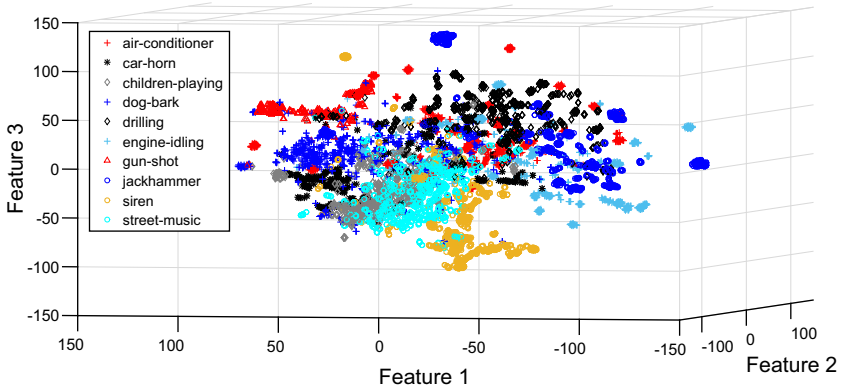

Fig. 8. Visualization of *UrbanSound8K* database in local feature space by t-SNE.

classes of urban sounds with 8732 real-world clips. A brief taxonomy of the ten sound categories is shown in Fig. 6. It is evident that the dataset includes typical urban noises and emergency sounds, which are highly related to city life, and thus the dataset is suitable for testing urban sound classification algorithms. In addition, we also presented Fig. 7 to show number of samples with respect to event classes. The dataset also gives 10 cross validation sets, where ensures there is no overlapping between training and testing splits. We measure classification accuracy to evaluate the methods.

We present the parameter settings of the proposed method. For Short-time Fourier transform (STFT), we set analysis window length to 10 ms with half size overlapping. 50 filters are used to construct Mel-filter bank ($F = 50$). In (2) and (3), we set both parameters *const* and ϵ to be 0.01, and let $\tau = 0.1$ in (7). In ProbSVM model, we applied radial basis function (10). For generating recurrent plots (RP), we experimentally determined cutoff distance $\sigma = 0.05$ and minimum line length $l_{\min} = 2$ in (14). For kernel density estimation in section 3.2.3. we apply Gaussian kernel with bandwidth $h = 2.5$. It is noteworthy that the size of dictionary K in (5) is a key parameter for dictionary learning and we made it as variable in experimental validations (Section 4.3.1). For global feature analysis, prior probabilities $p(c)$ in (15) is inferred from sample numbers in each event class that is shown in Fig. 7.

4.2. Study on features

4.2.1. Local feature space

We employed t-Distributed Stochastic Neighbour Embedding (t-SNE) approach [51] to visualize the *UrbanSound8K* dataset in the local feature space ($\mathbf{f}_{\text{local}}$) and the result is presented in Fig. 8. According to the chart, six categories of sounds, including gun shot, children playing, siren, street music, car horn and dog bark, were

distributed in (relatively) compact clusters in the feature space such that higher classification accuracy can be anticipated. By contrast, remaining 4 types of urban sounds of air conditioner, drilling, engine idling and jack hammer exhibited much wider within-class variability with heavy overlapping, and therefore it is difficult to achieve high classification precision. They are related to motor/mechanical sounds and their acoustic patterns vary significantly due to the specification variations, e.g. power and revolutions per minute (RPM). Based on Fig. 8, for improving classification performance, we are required to add another types of features, global features in this study.

4.2.2. Global feature space

In Fig. 9, we presented distributions of 8 kinds of urban sounds in global feature space. The two coordinates are spectrogram entropy (H) and RQA determinism (DET). The contour lines (blue to yellow colors denoted low to high likelihoods) represent probability density estimated via KDE. In addition, we also conducted likelihood inference using Gaussian estimator which was outlined by red ellipse (sample covariance coverage) and red cross (sample mean). Compared to Gaussian estimates, KDE likelihood exhibited more accurate class-conditional distribution estimation, especially at boundary areas. Such comparison demonstrate the reason why KDE is selected. By examining all distributions of 8 types of urban sounds, it is evident that discriminative information was encoded by global features and thus global features can further contribution to urban sound classification as complementary information to local features.

4.3. Results

4.3.1. Performance study on dictionary size

To explore how the dictionary size K affects systematic performance, we tested $K = 200, 500, 1000, 1500, 2000$ in dictionary learning. The results were shown in Fig. 10, which suggested that applying bigger dictionary is generally helpful, though, the improvement was not very significant. Because increasing dictionary size will require more computation cost and large memory, K should be determined with practical system resource considerations. In this study, we set $K = 2000$.

4.3.2. Validation of the feature aggregation framework

In order to demonstrate effectiveness of proposed framework, we conducted extensive experiments on *UrbanSound8K* dataset. Comparison results were summarized using box plots in Fig. 11, from which three main conclusions can be drawn: (1). We first focus on using local acoustic features only and we compared our feature learning model presented in Section 2.1 with previous

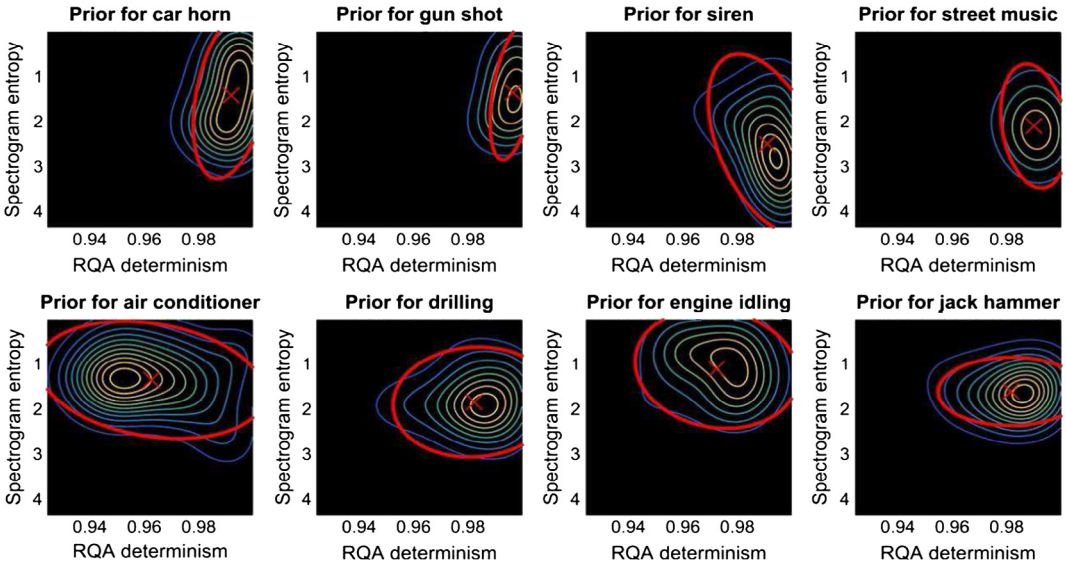

Fig. 9. Visualization of *UrbanSound8K* database in global feature space.

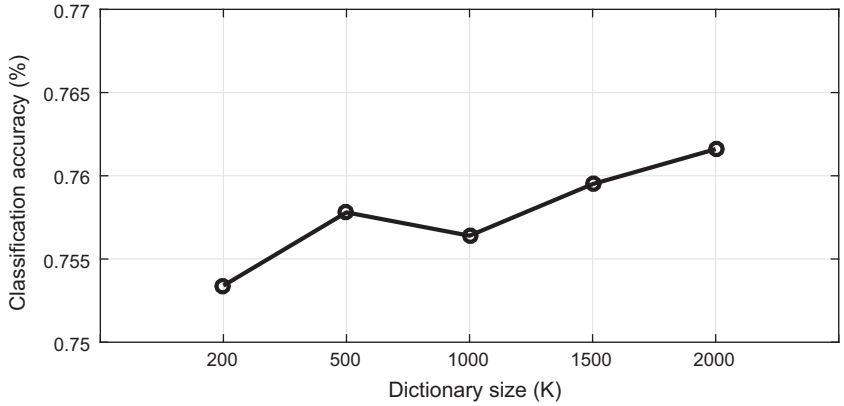

Fig. 10. Class mean average precision by changing dictionary size (K).

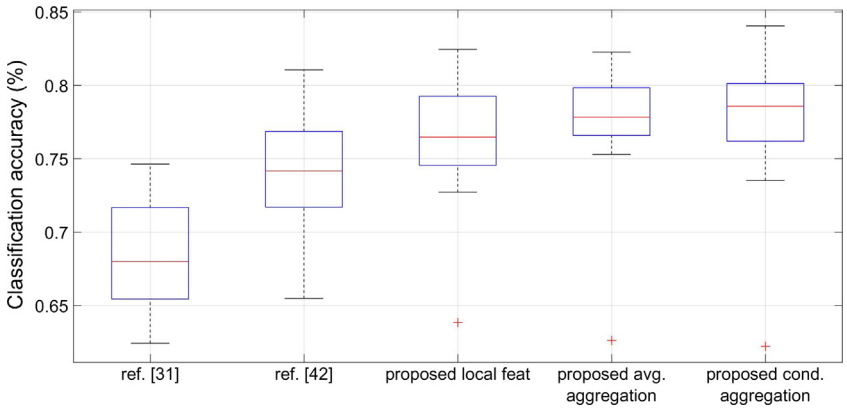

Fig. 11. Comparison of classification accuracy.

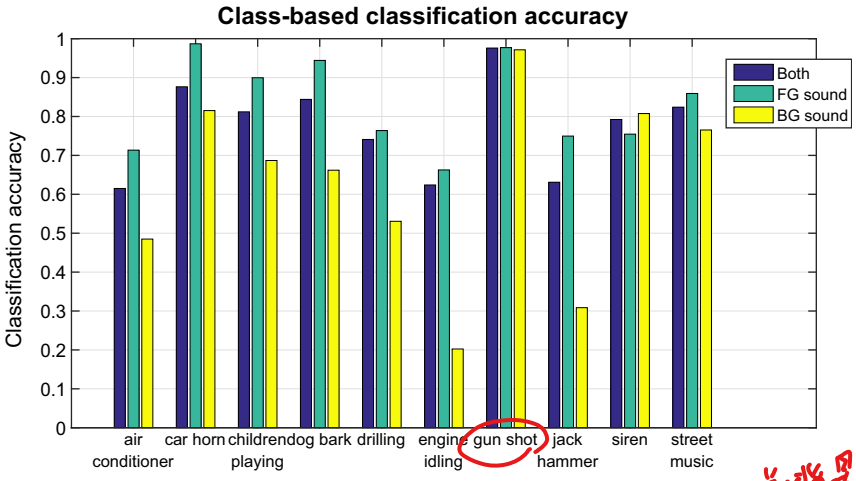

Fig. 12. Foreground and background sound classification comparison.

methods in [31,42]. Experimental results demonstrated that the proposed local feature learning model with soft feature coding scheme is favorable for urban sound events classification. 2. Furthermore, we compared the rules to aggregate local and global acoustic features and uniform feature fusion ($\alpha_c = 1/2$ in (17)) was compared with class-conditional fusion method. According to experimental results, conditional feature aggregation scheme obtained superior sound classification accuracy. 3. In total, proposed framework achieved 77.36% of mean averaged precision (mAP) for classification. Besides the two reference works of [31,42], deep convolution neural network (DCNN) had also been evaluated on *UrbanSound8K* dataset in a latest study [52] and it reported a classification result comparable to best performing model in [42] that achieved 73.69% mAp. Through comparison with 3 results from [31,42,52], the effectiveness of our proposal can be validated.

4.3.3. Foreground sound vs. background sound classification results

UrbanSound8K dataset also includes a series of salience labels indicating whether sound event was subjectively perceived to be in the foreground or background of the recording. In the case of foreground sounds, objective sound event is clear and environment is quite, while for background sounds, sound events are mixed with noisy environmental sounds. To demonstrate effectiveness of proposed classification scheme under distinct noise conditions, we conducted experiments on classifying foreground and background urban sounds separately. Classification performance is presented in Fig. 12. As expected, foreground sounds are easier to recognize, since the signal-to-noise ratio (SNR) was high. Particularly, the proposed scheme achieved ideal results for identifying foreground sounds of car horn (98.69%), dog barks (95.81%) and gun shot (96.71%), which are quite satisfying. On the other hand, in background sounds classification, performances were degraded because acoustic features were severely corrupted by noises. It, however, should be noted that our method can be more robust to noise by additionally introducing some processes specific to noise reduction [53].

5. Discussion

City planners and decision makers are trying to improve as much as possible the living quality of citizens and urban sound can contribute a lot for the purpose since it delivers rich information from both inhabit activities and environment. After presenting technical details in previous sections, now we discuss how can urban sound event classification techniques be used to provide services to urban people so as to improve their life quality. Proposed urban sound classification engine can be regarded as extractor which distills content information from raw sound wave. After classifying a number of sound events, we collect batch of information related what is happening inside the city and these contents, combing with location and time information, enable city agencies to take effective, information-driven actions accordingly. For instance, if gun shot is recognized by the system, such information

should be immediately reported to police office nearby, and if siren is detected somewhere, then proper measures, e.g. changing traffic lights, can be taken to help emergency vehicle pass through complex junctions quickly. Construction sound is regarded as the most common causes of noise complaints. After recognizing such sounds, e.g. drilling or jack hammer, better site monitoring can be realized, such as whether construction work has been performed within restricted hours or not. Moreover, sound classification technique can be useful tool for urban planners to perform acoustic environment evaluation. Through sound classification, we can promote a better understanding towards soundscape by knowing key elements in the sonic or acoustic environment, including animal and sounds from weather and sounds created by humans through musical composition, sounds of mechanical devices and other human activities. According to previous research results in hearing system physiology and auditory cognition [25,21,54], some sound events are intrusive and undesirable, while others can enhance people's mood [55]. By combining automatic sound classification techniques with human perception preference, categorization of positive/negative soundscapes can be effective performed, which is fundamental stage for design and management of urban sound environments.

6. Conclusion

A better understanding of urban sounds could greatly contribute to elaborate urban liveability, such as more efficient noise control, reliable safety surveillance and better acoustic environment planning. This paper focus on urban sound content retrieval which is regarded as essential building block of smart cities. An effective framework had been proposed in this work, which mapped urban sounds into two feature spaces, which are local feature space which characterizes temporal-spectral acoustic patterns and global feature space that describes long-term structures in sounds, e.g. variability and recurrence. To aggregate discriminative power in both feature spaces for classification, we adopted class-conditional fusion scheme which estimated relative contributions of the two level features with respect to each type of urban sounds in a supervised manner. At experimental stage, we conducted extensive experiments on *UrbanSound8K* dataset, which consists of 10 classes of common sounds in city with 8732 real-world recordings. We evaluated the proposed framework, contribution of key component and critical parameters. The experimental results demonstrated the proposed approach achieved superior classification accuracy compared with 3 other state-of-the-art results and the method presented enormous potential to facilitate people's urban life with various aspects.

Acknowledgment

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. The authors would like to thank the editor and the anonymous referees for their constructive comments to this study.

References

[1] Burden of disease from environmental noise – quantification of healthy life years lost in europe. Tech. rep., World Health Organization Regional Office for Europe; 2011.
[2] Zheng Y, Capra L, Wolfson O, Yang H. Urban computing: concepts, methodologies, and applications. ACM Trans Intell Syst Technol 2014;5(3):38:1–38:55.
[3] Urban life: open-air computers. The Economist; Oct. 2012.
[4] Steele D, Krijnders JD, Guastavino C. The sensor city initiative: cognitive sensors for soundscape transformations. In: GIS ostrava 2013. p. 8.
[5] Brown A, Lam K. Urban noise surveys. Appl Acoust 1987;20(1):23–39.
[6] Stephen Stansfeld BB, Haines Mary. Noise and health in the urban environment. Rev Environ Health 2000;15:43–82.
[7] Pham C, Cousin P. 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, GREENCOM-ITHINGS-CPSCOM '13. Washington, DC, USA: IEEE Computer Society; 2013. p. 611–8.
[8] Nagy G, Rodigast R, Hollosi D. Energy based traffic density estimation using embedded audio processing unit. In: Audio engineering society convention 136.
[9] Steinle S, Reis S, Sabel CE. Quantifying human exposure to air pollution moving from static monitoring to spatio-temporally resolved personal exposure assessment. Sci Total Environ 2013;443:184–93.
[10] Crocco M, Cristani M, Trucco A, Murino V. Audio surveillance: a systematic review. ACM Comput Surv 2016;48(4):52:1–52:46.
[11] Valero X, Alas F. Gammatone wavelet features for sound classification in surveillance applications. In: 2012 Proceedings of the 20th European signal processing conference (EUSIPCO). p. 1658–62.
[12] Chacon-Rodriguez A, Julian P, Castro L, Alvarado P, Hernandez N. Evaluation of gunshot detection algorithms. IEEE transactions on circuits and systems I: Regular papers 2011;58(2):363–73.
[13] Foggia P, Petkov N, Saggese A, Strisciuglio N, Vento M. Reliable detection of audio events in highly noisy environments. Pattern Recogn Lett 2015;65:22–8.
[14] Liu T, Zheng Y, Liu L, Liu Y, Zhu Y. Methods for sensing urban noises. Tech. Rep. MSR-TR-2014-66; May 2014.
[15] Sakantamis K, Wang B, Hao Y, Kang J, Chourmouziadou K. Soundscapes of European cities and landscapes, Oxford; 2013.
[16] Hritier H, Vienneau D, Frei P, Eze IC, Brink M, Probst-Hensch N, Rsli M. The association between road traffic noise exposure, annoyance and health-related quality of life (hrqol). Int J Environ Res Public Health 2014;11(12):12652.
[17] Torija AJ, Ruiz DP, Ramos-Ridao ngel F. A tool for urban soundscape evaluation applying support vector machines for developing a soundscape classification model. Sci Total Environ 2014;482483:440–51.
[18] Romero VP, Maffei L, Brambilla G, Ciaburro G. Modelling the soundscape quality of urban waterfronts by artificial neural networks. Appl Acoust 2016;111:121–8.
[19] Truax B. Handbook for acoustic ecology, World soundscape project. Vancouver: ARC Publications; 1978.
[20] Park TH, Turner J, Musick M, Lee JH, Jacoby C, Mydlarz C, et al. Sensing urban soundscapes. In: EDBT/ICDT workshops'14. p. 375–82.
[21] Raimbault M, Dubois D. Urban soundscapes: experiences and knowledge. Cities 2005;22(5):339–50.
[22] Botteldooren D, Andringa T, Aspuru I, Brown L, Dubois D, Guastavino C, Lavandier C, Nilsson M, Preis A, et al. Soundscape for european cities and landscape: understanding and exchanging. In: COST TD0804 final conference: soundscape of European cities and landscapes, Soundscape-COST. p. 36–43.
[23] Valero X, Farré P, Alías F. Comparison of machine learning techniques for the automatic recognition of soundscapes. In: Forum acusticum. European Acoustics Association; 2011. p. 2037–42.
[24] Jeon JY, Hong JY. Classification of urban park soundscapes through perceptions of the acoustical environments. Landscape Urban Plann 2015;141:100–11.
[25] Rychtrikov M, Vermeir G. Soundscape categorization on the basis of objective acoustical parameters. Appl Acoust 2013;74(2):240–7.
[26] Niessen M, Cance C, Dubois D. Categories for soundscape: toward a hybrid classification. Inter-noise and noise-con congress and conference proceedings, vol. 2010. Institute of Noise Control Engineering; 2010. p. 5816–29.
[27] Hinton GE. Training products of experts by minimizing contrastive divergence. Neural Comput. 2002;14(8):1771–800.
[28] Coates A, Ng A. Neural networks: tricks of the trade. Berlin, Heidelberg: Springer; 2012. chapter 2. Learning feature representations with K-means.
[29] Coates A, Ng AY. The importance of encoding versus training with sparse coding and vector quantization. In: Getoor L, Scheffer T, editors. Proceedings of the 28th international conference on machine learning (ICML-11). New York, NY, USA: ACM; 2011. p. 921–8.
[30] Charles Webber JZ. Tutorials in contemporary nonlinear methods for the behavioral sciences Web Book. Chapter 2: Recurrence quantification analysis of nonlinear dynamical systems; 2005.
[31] Salamon J, Jacoby C, Bello JP. A dataset and taxonomy for urban sound research. In: 22th ACM int. conf. on multimedia.
[32] Hancke GP, Silva BdCe, Hancke Jr GP. The role of advanced sensing in smart cities. Sensors 2013;13(1):393.
[33] Phan H, Maa M, Mazur R, Mertins A. Random regression forests for acoustic event detection and classification. IEEE/ACM Trans Audio Speech Language Process 2015;23(1):20–31.
[34] Ye J, Kobayashi T, Murakawa M, Higuchi T. Robust acoustic feature extraction for sound classification based on noise reduction. In: 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP). p. 5944–8. http://dx.doi.org/10.1109/ICASSP.2015.7177954.
[35] Rakotomamonjy A, Gasso G. Histogram of gradients of time-frequency representations for audio scene classification. IEEE TASLP 2015;23(1):142–53.

[36] Juang BH, Rabiner LR. Automatic speech recognition – a brief history of the technology development. Elsevier encyclopedia of language and linguistics; 2005.
[37] Klapuri A, Davy M. Signal processing methods for music transcription. Springer Science & Business Media; 2007.
[38] Giannoulis D, Benetos E, Stowell D, Rossignol M, Lagrange M, Plumbley MD. Detection and classification of acoustic scenes and events: an ieee aasp challenge. In: 2013 IEEE workshop on applications of signal processing to audio and acoustics (WASPAA). p. 1–4.
[39] Chen J, Wang Y, Wang D. A feature study for classification-based speech separation at low signal-to-noise ratios. IEEE TASLP 2014;22(12):1993–2002.
[40] Ness SR, Walters T, Lyon RF. Auditory sparse coding. Music data mining, Boca Raton, FL 2011. 33487–2742.
[41] Lagrange M, Lafay G, Dfrville B, Aucouturier J-J. The bag-of-frames approach: a not so sufficient model for urban soundscapes. J Acoust Soc Am 2015;138(5).
[42] Salamon J, Bello JP. Unsupervised feature learning for urban sound classification. In: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). p. 171–5.
[43] Kobayashi T, Ye J. Acoustic feature extraction by statistics based local binary pattern for environmental sound classification. In: IEEE international conference on acoustics, speech and signal processing, ICASSP 2014, Florence, Italy, May 4–9, 2014. p. 3052–6.
[44] Barchiesi D, Giannoulis D, Stowell D, Plumbley M. Acoustic scene classification: classifying environments from the sounds they produce. Signal Process Mag, IEEE 2015;32(3):16–34.
[45] Temko A, Malkin R, Zieger C, Macho D, Nadeu C, Omologo M. Multimodal technologies for perception of humans: first international evaluation workshop on classification of events, activities and relationships, CLEAR 2006, Southampton, UK, April 6–7, 2006, revised selected papers. Berlin, Heidelberg: Springer; 2007. p. 311–22. http://dx.doi.org/10.1007/978-3-540-69568-4_29. chapter. CLEAR evaluation of acoustic event detection and classification systems.
[46] Piczak KJ. Esc: dataset for environmental sound classification. In: Proceedings of the 23rd ACM international conference on multimedia, MM '15. New York, NY, USA: ACM; 2015. p. 1015–8.
[47] Fei-Fei L, Perona P. A bayesian hierarchical model for learning natural scene categories. IEEE computer society conference on computer vision and pattern recognition, 2005 (CVPR 2005), vol. 2. p. 524–31.
[48] Boureau Y-L, Ponce J, Lecun Y. A theoretical analysis of feature pooling in visual recognition. In: 27th International conference on machine learning, Haifa, Israel.
[49] Gerard Roma PH, Nogueira Waldo. Recurrence quantification analysis for auditory scene classification. Tech. rep.; 2005.
[50] Bishop CM. Pattern recognition and machine learning (information science and statistics). Secaucus, NJ, USA: Springer-Verlag New York, Inc.; 2006.
[51] Van der Maaten L, Hinton G. Visualizing data using t-sne. J Machine Learn Res 2008;9(2579–2605):85.
[52] Piczak KJ. Environmental sound classification with convolutional neural networks. In: 2015 IEEE 25th international workshop on machine learning for signal processing (MLSP). p. 1–6.
[53] Vaseghi SV. Advanced digital signal processing and noise reduction. John Wiley & Sons; 2008.
[54] Warren PS, Katti M, Ermann M, Brazel A. Urban bioacoustics: it's not just noise. Animal Behav 2006;71(3):491–502.
[55] Payne SR. The production of a perceived restorativeness soundscape scale. Appl Acoust 2013;74(2):255–63. http://dx.doi.org/10.1016/j.apacoust.2011.11.005. Applied soundscapes: recent advances in soundscape research.