Environmental Audio Scene and Sound Event Recognition for Autonomous Surveillance: A Survey and Comparative Studies

S. CHANDRAKALA, Intelligent Systems Group, School of Computing, SASTRA University, India
S. L. JAYALAKSHMI, Velammal Engineering College, India

Monitoring of human and social activities is becoming increasingly pervasive in our living environment for public security and safety applications. The recognition of suspicious events is important in both indoor and outdoor environments, such as child-care centers, smart-homes, old-age homes, residential areas, office environments, elevators, and smart cities. Environmental audio scene and sound event recognition are the fundamental tasks involved in many audio surveillance applications. Although numerous approaches have been proposed, robust environmental audio surveillance remains a huge challenge due to various reasons, such as various types of overlapping audio sounds, background noises, and lack of universal and multi-modal datasets. The goal of this article is to review various features of representing audio scenes and sound events and provide appropriate machine learning algorithms for audio surveillance tasks. Benchmark datasets are categorized based on the real-world scenarios of audio surveillance applications. To have a quantitative understanding, some of the state-of-the-art approaches are evaluated based on two benchmark datasets for audio scenes and sound event recognition tasks. Finally, we outline the possible future directions for improving the recognition of environmental audio scenes and sound events.

CCS Concepts: • **Computing methodologies** → **Machine learning**; **Learning paradigms**; **Machine learning approaches**; • **Applied computing** → **Surveillance mechanisms**;

Additional Key Words and Phrases: Environmental audio surveillance, environmental audio scene recognition, sound event recognition, audio features, audio tagging, acoustic source localization

ACM Reference format:
S. Chandrakala and S. L. Jayalakshmi. 2019. Environmental Audio Scene and Sound Event Recognition for Autonomous Surveillance: A Survey and Comparative Studies. *ACM Comput. Surv.* 52, 3, Article 63 (June 2019), 34 pages.
https://doi.org/10.1145/3322240

1 INTRODUCTION

Automated surveillance systems that use multi-modal techniques with both video and audio information have recently gained importance (Aytar et al. 2016; Ben Mabrouk and Zagrouba 2018; Brun

The authors acknowledged the support of research grant from Department of Science and Technology, Government of India, under the scheme, Cognitive Science Research Initiative, grant number DST/CSRI/2017/131.
Authors' addresses: S. Chandrakala, Intelligent Systems Group, School of Computing, SASTRA University, Thanjavur-613401, Tamil Nadu, India; emails: sckala@gmail.com, chandrakala@cse.sastra.edu; S. L. Jayalakshmi, Department of Computer Science and Engineering, Velammal Engineering College, Chennai-600066, Tamil Nadu, India; email: sathishjayalakshmi02@gmail.com.

0360-0300/2019/06-ART63 $15.00
https://doi.org/10.1145/3322240

et al. 2014; Butko et al. 2011; Gomes et al. 2016; Lv et al. 2018; Takahashi et al. 2018). Suspicious events can be caused either in the environment (natural disasters) or among persons themselves (gunshots or screaming sounds) (Bello et al. 2018; Gerosa et al. 2007; Imoto 2018; Valenzise et al. 2007). Thus, monitoring of human and social activities and early detection of suspicious events are essential for public security and safety applications. Automated surveillance systems collect input data from neighboring environments using one or more video cameras. The overall performance can be affected because of the limited quality of video cameras in natural calamity conditions and their vulnerability to sudden light switching, reflections, and shadows. In addition, popular video cameras cannot capture useful information during nighttime because of restricted lighting effects and automobile lights (Crocco et al. 2016). To overcome these drawbacks, infrared (thermal) cameras can be an alternative to capture video inputs during nighttime. Infrared cameras offer a challenge in the separation of background and foreground items. When visual cues cannot recognize the activities (events) and environments (contexts), audio cues are supplementary to visual cues. For instance, when a suspicious activity/object is occluded; when an activity happens in the dark; and when an anomalous activity happens in an area very close to the site that is beyond the coverage of video cameras.

Audio sensors (microphones) offer the following advantages: (i) microphones can be easily deployed, as it supports the arrangement of maximum number of audio sensors in more challenging environments; (ii) omnidirectional coverage; and (iii) specular reflections of the audio signal can be another form of audio input. Information extracted from a semantic audio event is beneficial for audio surveillance and related applications such as analysis and forecast of patterns of events, classifying/searching audio records, customer alerts, and robot navigation (Baum et al. 2018; Bello et al. 2018; Ozer et al. 2018; Ren et al. 2017a). The Environmental Audio Scene Recognition (EASR) and Sound Event Recognition (SER) tasks in an uncontrolled environment are part of the research field called Computational Auditory Scene Analysis (CASA) (Temko et al. 2009; Wang and Brown 2006). CASA typically addresses the difficulty of segregating a few sound sources or segmenting sound streams into a small number of acoustically compact categories (speech or non-speech events) (Eghbal-zadeh et al. 2017; Eronen et al. 2006; Geiger et al. 2013; Stowell et al. 2015).

1.1 Overview of Audio Surveillance Systems

This review focuses on the use of audio data in EASR and SER tasks in surveillance applications. Figure 1 depicts an overview of a generic methodology involved in an audio surveillance system. Environmental audio scene recognition refers to the process of recognizing the environment of an audio stream, with applications in devices requiring contextual awareness. Some of the indoor audio scenes include the following: grocery shops, cafes/restaurants, homes, supermarkets, and traveling inside vehicles. Outdoor audio scenes include busy streets, open-air markets, parks, quiet streets, city centers, forest-paths, beaches, residential areas, metro stations, and vehicles on the road. In an EASR task, the duration of an environmental audio scene is long, and it is in the range of a few seconds to a few tens of seconds. EASR can be tailored to different applications such as forensic analysis (Marchi et al. 2016), audio surveillance (Barchiesi et al. 2015; Ntalampiras et al. 2011), music recommendations (Bisot et al. 2015), hearing aids (Agcaer et al. 2015), adaptive audio context recognition (Mafra et al. 2016), acoustic source localization (Saggese et al. 2017), audio tagging (Xu et al. 2017a, 2017b; Yun et al. 2016), and sound event recognition (Medhat et al. 2017; Stowell et al. 2015).

SER aims to locate and recognize each occurrence of a monophonic event or polyphonic event in a specific environment (Beltrán et al. 2015; McLoughlin et al. 2015). For instance, in a busy street, some of the specific sound events are conversational speech, a car horn, noise made by a

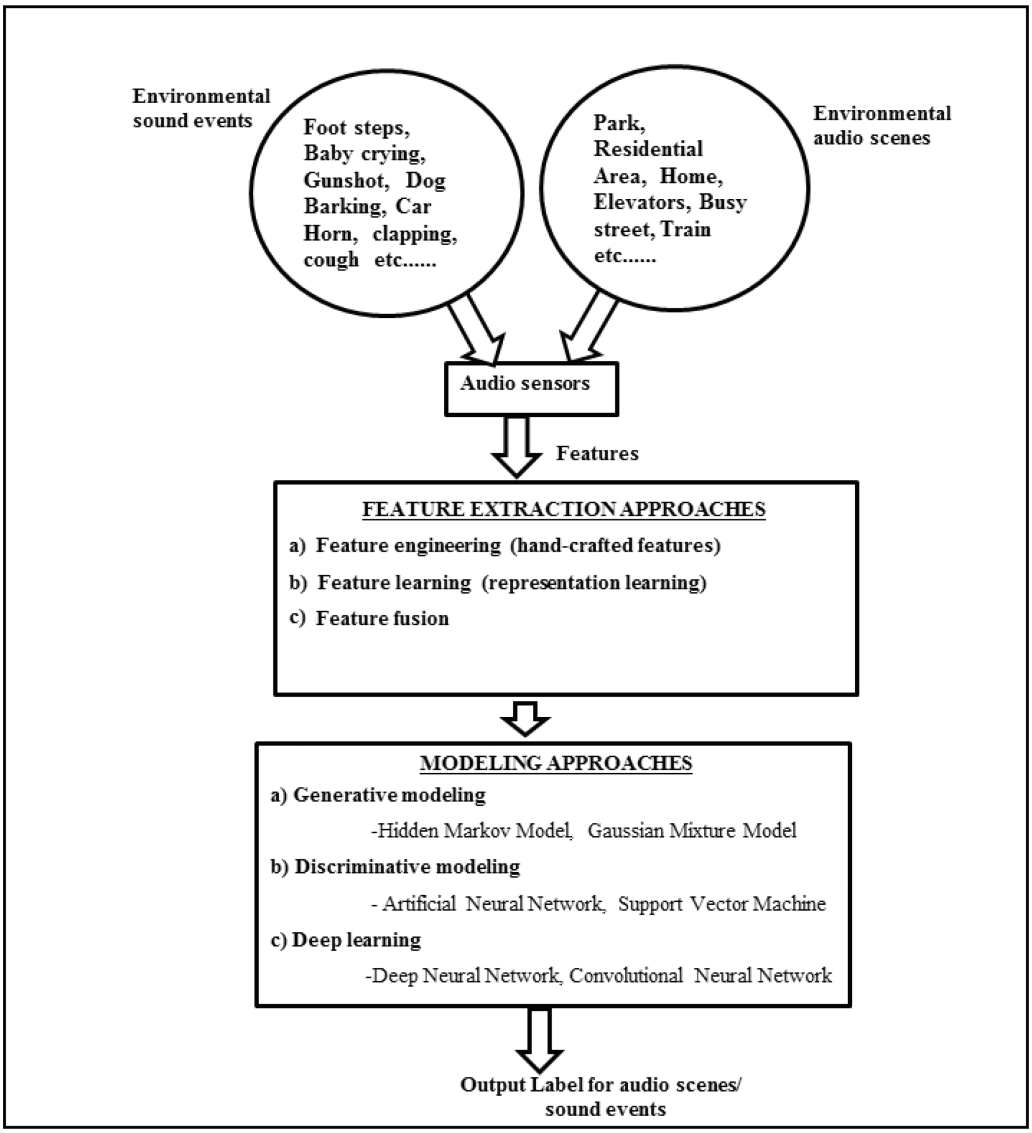

Fig. 1. Overview of environmental audio scene and sound event recognition tasks in an audio surveillance application.

motorcycle or bicycle, sudden brake sounds, and the sound of footsteps. Sound events can either be produced by humans (non-speech sounds) or objects. Some of the human non-speech events include the following: infant screams, adult screams, crying, coughing, clapping, whistling, sneezing, footsteps, and laughing. Sound events that are generated by objects include gunshots, explosions, breaking glass, the cheering of an audience at a sports event, alarm bells, door slams, a ball bouncing, keyboard typing, and washing of hands. Sound events take a shorter duration of about 100 to 500ms compared with the audio scenes of longer duration. Recognition of sound events as normal or abnormal events can be instrumental to various applications such as EASR (Bello

et al. 2018; Crocco et al. 2016; Lee et al. 2013; McLoughlin et al. 2015; Ng 2014; Ntalampiras et al. 2011; Ozer et al. 2018; Stowell et al. 2015), multimedia information retrieval (Cristani et al. 2007; Elizalde et al. 2016a; Evangelopoulos et al. 2013; Moeslund et al. 2014), road surveillance (Foggia et al. 2016), audio forensics (Malik 2013), monitoring anomalous sounds in healthcare (Cheffena 2016; Irtaza et al. 2017; Lozano et al. 2010), Ambient Assisted Living (AAL) tools (McLoughlin et al. 2015; Beltrán et al. 2015), and fall detection (Cheffena 2016; Irtaza et al. 2017) applications. This article is organized as follows: Section 2 presents the challenges involved in environmental audio surveillance applications. Section 3 provides a review of the various audio features used in EASR and SER tasks. Section 4 explores the datasets and methodologies used in an EASR task. Section 5 provides a review of the various datasets and methodologies used in sound event recognition tasks. Future research directions are presented in Section 6.

2 CHALLENGES

Some unique characteristics of environmental audio scenes and sound events include the following:

- —The signal-to-noise ratio (SNR) is typically very small in an audio signal, particularly if the microphone is not very near to the acoustic source (Crocco et al. 2016).
- —Discriminative information exists in low-frequency ranges (Chachada and Kuo 2014).
- —Environmental sounds/scenes do not have any specific structures such as phonemes or prosody (Cowling and Sitte 2003).

As opposed to music signals, the audio events/scenes do not show meaningful patterns and so it is hard to model their characteristics (Abdel-Hamid et al. 2014; Gold et al. 2011). For example, the speech recognition task often exploits the phonetic sequence. However, environmental scenes/sounds such as "typing" or "gunshots" do not have any phonetic structure, which makes the audio event/scene recognition task more challenging.

Some of the challenges related to EASR and SER include the following:

- —recognizing many events from a single environment, such as an office room, a residential area, or a busy street that may have multiple sound sources (Mesaros et al. 2010, 2015);
- —a dictionary of basic units is unidentifiable (Cowling and Sitte 2003);
- —the existence of overlapping or polyphonic events (Gemmeke et al. 2013; Mesaros et al. 2010);
- —recognition of confusing scenes, e.g., street traffic vs. restaurant (Ntalampiras et al. 2011);
- —lack of discrimination among scenes such as pedestrian street, market, quiet street, and shop (Rakotomamonjy and Gasso 2015);
- —identifying acoustic sound sources in the presence of background noise (Beltrán et al. 2015; Salamon et al. 2014);
- —the existence of certain audio events in multiple environments, e.g., "gunshot" sound present in environments such as street and home (Heittola et al. 2013);
- —the multimodal surveillance system for critical indoor environments (Moeslund et al. 2014);
- —lack of standard and multimodal datasets (Chachada and Kuo 2013);
- —lack of robust and compact representation learning techniques for audio scenes and sound events (Ozer et al. 2018; Phan et al. 2017).

In this article, we present an extensive review of EASR and SER tasks employed in audio surveillance applications. We review the following three important aspects: (1) the impact of audio

features on recognition performance, (2) the availability of various audio scenes and sound event datasets, and (3) the methodologies employed for recognition tasks.

3 AUDIO FEATURES

3.1 Basics of Audio Feature Extraction

The selection of robust audio features plays a critical role in environmental audio surveillance. Generally, audio features are intended to capture the discriminative information useful for the recognition or classification task while neglecting redundancies and background noises. The feature extraction approach based on frame-based processing involves dividing an audio signal into frames, often using a Hamming or Hanning window. Subsequently, features are extracted from each frame, and this sequence of feature vectors is used to represent an audio signal. Feature extraction can be divided into two broad categories: stationary and non-stationary feature extraction (Cowling and Sitte 2003). Stationary feature extraction produces detailed frequency contents from the whole signal. However, it cannot recognize where these frequencies are available in the signal. In stationary feature extraction, there are eight prominent features commonly used in non-speech sounds (Cowling and Sitte 2003): (i) frequency (spectral roll-off, spectral flatness, etc.); (ii) homomorphic cepstral coefficients; (iii) Mel Frequency Cepstral Coefficients (MFCC); (iv) Linear Prediction Cepstral (LPC) coefficients; (v) Mel frequency LPC coefficients; (vi) bark frequency cepstral coefficients; (vii) bark frequency LPC coefficients; and (viii) Perceptual Linear Prediction (PLP) features. Conversely, non-stationary feature extraction partitions the signals into discrete time units. This method helps to identify the occurrence of frequency components in a specific part of the signal to understand the nature of the signal. The fundamental features that are commonly referenced in general literature (Cowling and Sitte 2003) are as follows: (i) Short-time Fourier Transform (STFT); (ii) Fast (discrete) Wavelet Transform (FWT); (iii) Continuous Wavelet Transform (CWT); and (iv) Wigner-Ville Distribution (WVD). The aforementioned features use different algorithms to obtain a Time-frequency Representation (TFR) of a signal.

The features are grouped into the following four categories: feature engineering (generic features) approaches, auditory-image based features, feature learning approaches, and feature fusion approaches.

3.2 Feature Engineering-based (Generic Features) Approaches

The feature engineering approach utilizes engineered/generic features such as temporal, spectral, cepstral-domain features, and perceptually driven features for representing the characteristics of an audio signal, as shown in Table 1.

3.2.1 Time-domain Features. Time-domain features such as Zero-Crossing Rate (ZCR) (Eronen et al. 2006; Salamon et al. 2014; Stowell and Clayton 2015), short-time energy (Chu et al. 2009), and waveform minimum and maximum (Ntalampiras et al. 2009) are frequently used along with other developed sets of features. The ZCR feature measures the rate of sign changes along a signal. It indicates the frequency that dominates in the frame and finds whether a segment of audio is normal or abnormal. These are strongly correlated with the Spectral Centroid (SC). The Short-Time Energy (STE) feature measures the total signal energy in a frame (Chu et al. 2009; Phan et al. 2016). The waveform minimum and maximum feature measures the maximum and the minimum value of a signal waveform (Ntalampiras et al. 2009). In particular, these features are directly extracted from the temporal information of an audio signal. It can be used in combination with more complex features (e.g., MFCC, wavelets) (Rabaoui et al. 2008).

Table 1. Generic Audio Features Employed in Environmental Audio Scene and Sound Event Recognition Tasks

Domain	Features	EASR task	SER task
Time	Zero-crossing rate.	(Eronen et al. 2006)	(Salamon et al. 2014)
	Short time energy		(Chu et al. 2009; Phan et al. 2016)
	Waveform minimum and maximum	(Ntalampiras et al. 2009)	(Ntalampiras et al. 2009)
Frequency	Spectral roll-off, spectral flatness, spectral centroid, spectral flux, and pitch ratio	(Petetin et al. 2015; Agcaer et al. 2015; Han and Lee 2016)	(Foggia et al. 2015; Rabaoui et al. 2008)
	Spectral Dynamic Features (SDF)		(Karbasi et al. 2011)
Cepstrum	MFCC, MFCC derivatives.	(Eronen et al. 2006; Jing et al. 2017; Mafra et al. 2016; Mesaros et al. 2016; Rabaoui et al. 2008; Salamon et al. 2014; Stowell and Clayton 2015; Stowell et al. 2015)	(Mafra et al. 2016)
	Linear Prediction Cepstral Coefficients (LPCC).	—	(Atrey et al. 2006)
	Linear Frequency Cepstral Coefficients (LFCC).	—	(Atrey et al. 2006)
Energy	Signal energy.	(Zieger et al. 2009; Li et al. 2009)	(Zieger and Omologo 2008)
	Log energy first and second derivatives.	(Rabaoui et al. 2008; Zieger and Omologo 2008)	
Biologically or perceptually driven	Gammatone filter bank features (GTCC).	—	(Valero and Alias 2012)
	Perceptual Linear Prediction (PLP) coefficients and derivatives.	—	(Rouas et al. 2006)
	Intonation and Teager Energy Operator (TEO) features	—	(Ntalampiras et al. 2011)

3.2.2 Frequency-domain Features. Spectral features are extracted by converting the time-domain signal into a frequency-domain signal using the Discrete Fourier Transform. Some of the features such as pitch-ratio, Spectral Centroid (SC), spectral flatness, spectral roll-off, spectral variation, band energy, bandwidth, fundamental frequency, and Spectral Flux (SF) (Agcaer et al. 2015; Han and Lee 2016; Petetin et al. 2015; Rabaoui et al. 2008) provide measures about spectral content properties of a sound event. Most of the spectral features used in the literature are used in combination with higher-dimensional features (e.g., MFCC). Sawhney and Maes (1997) analyzed various environments (people, voices, subway, traffic, and outdoor sounds) using types of features such as Perceptual Linear Predictive (PLP), Power Spectral Density (PSD), and frequency bands from a filter bank. The authors experimented these features with the Recurrent Neural Network (RNN) classifier and a simple nearest neighbor classifier. When compared with PSD, PLP is not

perfect for recognizing environmental sounds, as it was primarily intended for decreasing noise and enhancing recognition of human speech. PSD features worked well, since they were less computationally intensive and provided an approximate for first-level classification. Features computed from frequency bands of filter-banks with a simple nearest neighbor classifier outperformed the PLP and PSD features. The robust features of filter-banks discarded many of the fast-changing features, yet we are intrigued only about the slow-changing attributes of environmental sounds. Spectral Dynamic Features (SDF) were recently proposed for extracting the temporal difference among sub-frames (Karbasi et al. 2011). The experiment was carried out on several classical features such as ZCR, LPC, and MFCC using the generative model-based classifiers such as *k*-Nearest Neighbors (KNN), Gaussian Mixture Model (GMM), and Support Vector Machine (SVM). The SDF outperformed all the other features using all three classifiers.

3.2.3 Cepstral-domain Features. Cepstral features are normally computed on cepstrum. Cepstrum is the inverse Fourier transform of the log-magnitude of the power spectrum. Cepstrum can be taken from the information about the rate of change in spectrum bands. Some of the generic cepstral features used in the literature are MFCCs, MFCC derivatives, homomorphic cepstral coefficients, and Linear Predictive Cepstral Coefficients (LPCCs) (Atrey et al. 2006). MFCCs are the most widely used features for environmental audio scenes and sound events recognition (Eronen et al. 2006; Jing et al. 2017; Ma et al. 2006; Mafra et al. 2016; Mesaros et al. 2016; Rabaoui et al. 2008; Salamon et al. 2014; Stowell and Clayton 2015; Stowell et al. 2015). The Mel-frequency Cepstrum (MFC) is a result of taking the Discrete Cosine Transform (DCT) of the log-magnitude of the signal spectrum. MFC was made from a collection of MFCCs. It gives the statistical properties of cepstrum. Dynamic characteristics of the cepstrum were computed using Mel-frequency Delta Cepstral Coefficients (Delta MFCCs). These values are calculated from MFCC across neighboring frames. Homomorphic cepstral coefficients are calculated by applying the logarithm and DCT on spectrum coefficients without using Mel filter-bank (Cowling and Sitte 2003). Band-energy refers to the energies of subbands normalized with the total energy of the audio signal.

Linear Prediction Coefficients (LPCs) were extracted using the autocorrelation method. Linear Prediction Cepstral Coefficients (LPCCs) are obtained using a direct recursion from LPCs. In Tsau et al. (2011), the use of Code Excited Linear Prediction (CELP)-based features along with the LPC, pitch, and pitch gain features is proposed. CELP is more robust than LPC, since it uses a fixed codebook for excitation of a source-filter model. The authors reported that sound classes such as thunder, rain, and stream show improved performance for a combination of CELP and MFCC features. A wide set of frequency- and time-domain features (Eronen et al. 2006) were explored to recognize various indoor and outdoor environments (streets, vehicles, restaurants, offices/meetings, and homes). Some of the features extracted from the input signal are as follows: ZCR, short-time average, MFCCs, band-energy SC, bandwidth, spectral roll-off, spectral flux (SF), linear prediction coefficients (LPCs), and linear prediction cepstral coefficients (LPCCs). The authors evaluated linear feature transforms such as Linear Discriminant Analysis (LDA), Principal Component Analysis (PCA) or Independent Component Analysis (ICA) and discriminative training (maximum mutual information) to improve the accuracy obtained with low-order HMMs. MFCC has consistently shown better recognition accuracies compared with human recognition accuracy. However, autoregression-based features such as LPC and LPCC incorporate the source-filter model for speech and emotion data, and hence they are not useful for environmental sound events (Gold et al. 2011).

3.2.4 Energy-domain Features. Energy features are more widely involved in forepart sound feature extraction and cause an increase in intra-class variation. Signal energy and log energy derivatives are commonly used energy-based features (Li et al. 2009; Rabaoui et al. 2008; Zieger et al. 2009; Zieger and Omologo 2008).

Table 2. Auditory Image-based Features Employed in Environmental Audio Scene and Sound Event Recognition Tasks

Features	EASR task	SER task
Wavelet coefficients	—	(Rabaoui et al. 2008)
Matching Pursuit (MP) features	—	(Chu et al. 2009)
Gabor Filter Bank (GBFB) features	—	(Schroder et al. 2013)
Histogram of Oriented Gradients (HOG) features	(Rakotomamonjy and Gasso 2015)	
Spectrogram Image Features (SIF)	—	(McLoughlin et al. 2015)
MFCC+Local Binary Pattern (LBP)+SC	—	(Yang et al. 2017)
variable-Q transform (VQT) and Adjacent Evaluation Completed LBP (AECLBP)	—	(Abidin et al. 2018)

3.2.5 Biologically or Perceptually Driven Features. Perceptually driven features represent the non–stationarity of environmental audio scenes and sound events. They represent the values calculated based on time, frequency, time-frequency, or cepstral representations by considering the human auditory and/or vocal structure. Gammatone Cepstral Coefficients (GTCCs) (Agrawal et al. 2017; Valero and Alias 2012), Perceptual Linear Prediction (PLP) coefficients (Rouas et al. 2006), and Intonation and Teager Energy Operator (TEO) features (Ntalampiras et al. 2011) are commonly used in audio surveillance applications. GTCCs are a variation of MFCC, in which the triangular filter bank spaced on the Mel scale is substituted with a Gammatone filter bank, achieving better resolution at lowest frequencies and a tighter model of the human cochlear response. The PLP analysis is a combination of spectral analysis and linear prediction analysis. The cepstral features of MFCC and PLP are extremely similar. The cepstral coefficients are measured from the power spectrum. Because of the similarity between Mel frequency-based filters and bark frequency filters, the results of these filters are always similar. Intonation and TEO features are used to measure the energy of a stressed speech signal. While calculating the energy values, the TEO approach ignores the noisy part present in the audio signal (Jena and Singh 2018).

3.3 Auditory Image-based Features

Audio-visual descriptor plays a vital role in an environmental audio scene recognition task. In spectrogram-based features, a particular intensity region of the spectrogram was considered as an image. The time-frequency features are extracted from a spectrogram. Table 2 shows the commonly used time-frequency-based features, such as Gabor Filter Bank (GBFB) features (Schroder et al. 2013), wavelet packets (Rabaoui et al. 2008), and Spectrogram Image Features (SIFs) (McLoughlin et al. 2015). GBFB features (Schroder et al. 2013) represent the spectro-temporal modulation patterns of the signal using the Gabor filter bank. It works better for noise-robust SER tasks. The wavelet-coefficient-based approach (Rabaoui et al. 2008) replaces the classical Fourier analysis method by capturing the time and frequency information from the audio signal. For a given sound, basis functions are used to measure the high-frequency and low-frequency contents. In wavelet transform, the frequency dependent temporal information was identified to aid the understanding of signals. Some of the wavelet-based methods reported in literature are Discrete Wavelet Transform (DWT) and Continuous Wavelet Transform (CWT) (Cowling and Sitte 2003). Spectrogram image features (McLoughlin et al. 2015) provided a better representation of auditory signal using

either a linear or log-scaled spectrogram. This spectrogram is represented by using a pseudo colormap that is partitioned into primary color intensities to highlight the particular intensity region of the spectrogram image.

The authors in Yang et al. (2017) proposed the Local Binary Pattern (LBP) descriptor approach for capturing the temporal dynamics of extracted MFCC features. LBP descriptor approach results in a normalized histogram vector that contains supplemental information on the temporal evaluation. Further, complementary spectral features such as Spectral Centroid (SC) were added to both MFCC and LBP features to slightly improve the recognition accuracy. In the classification phase, the SVM and the ensemble classifier called as Dynamic Selection and Circulating Combination-based Clustering (D3C) were used. The proposed LBP features with ensemble classifier D3C captured better characteristics than the Recurrence Quantification Analysis (RQA) features in the temporal domain.

Another variant of LBP descriptor is a unique combination of Variable-Q Transform (VQT) and Adjacent Evaluation Completed LBP (AECLBP) (Abidin et al. 2018). The VQT captures the time resolution at low frequencies better than the Constant-Q Transform (CQT). The AECLBP threshold values are computed distinctively for each zone for a better capture of local intensity information. The VQT Time-frequency Representation (TFR) preserves the important spectral and temporal structures as texture images. AELBP is a variant of LBP in which the extracted micro-structure of image texture features is different from that of a TFR. Thus, a unique combination of VQT and AECLBP provides better discriminative performance over the CQT and LBP with better improvement in recognition accuracy. Finally, the results are calculated from a simple feature-level fusion of Histogram of Oriented Gradients (HOG) and AECLBP histogram features. The class-wise recognition accuracy of park, residential area, and home is improved using an SVM classifier.

Another type of feature extraction is mostly based on local features and global features. Some of the local features such as ZCR, MFCC, and Log Frequency Cepstral Coefficient (LFCC) are extracted from every frame of an audio signal (Chachada and Kuo 2013). The global features such as pitch, energy, duration, and formants are extracted from the complete length of an audio signal (El Ayadi et al. 2011). Some of the wavelet-based methods reported in literature are Discrete Wavelet Transform (DWT) and Continuous Wavelet Transform (CWT).

A summary of the set of feature extraction methods explored in literature is shown in Tables 1 and 2. The selection of an optimal window length is one of the major challenges of the feature extraction method (Gold et al. 2011). Temporal and spectral features are the simplest features used frequently along with other refined sets of features. The short-term spectral features are easy to compute, most discriminative, and so it is widely used for EASR and SER tasks. LPC, LPCC, and PLP features are not useful for environmental sound events, and hence they incorporate the source-filter model for speech and emotion data (Dhanalakshmi et al. 2011). Spectral features provide basic measures about time and frequency-domain properties. Cepstral features such as MFCCs are used by experts for baseline studies. When compared with MFCC features, the performance of wavelet features is not better at a comparable level. The gammatone features replaced the capability of MFCC by showing slightly improved recognition accuracy for impulsive sounds, such as footsteps, explosions, screams, and gunshots (Agrawal et al. 2017; Valero and Alias 2012). A common approach is to combine MFCC features with newly developed features to improve the performance of both EASR and SER tasks (Chachada and Kuo 2013).

3.4 Feature Learning-based Approaches

Recently, the outcome of machine learning techniques has been enhanced by representing the data using feature learning approaches (Bengio et al. 2013; Hinton et al. 2012). To overcome the

Table 3. Feature Learning-based (CNN-based) Approaches for Environmental Audio Scene and Sound Event Recognition Tasks

Features	EASR task	SER task
i-vector features using MFCCs		(Dehak et al. 2011)
Exemplar coding		(Gemmeke et al. 2013)
Sparse coding		(Lu et al. 2014)
Bag of aural words		(Plinge et al. 2014)
Non-negative matrix factorization (NMF)		(Stowell et al. 2015)
Optimized AMS features + Covariance Matrix Adaption Evolutionary Strategy (CMA-ES)		(Agcaer et al. 2015)
Phone triplets approach		(Phan et al. 2016)
L1-penalized sparse coding and spherical k-means dictionary learning (using MFCC features).		(Ye et al. 2019)
Label tree embedding images using CNN.	(Phan et al. 2017)	
Perceptually weighted spectrograms and i-vector system (extracted from MFCC features).	(Dorfer et al. 2018)	

drawbacks such as overlapping scenes and noises in generic features, the feature learning methods were used. Table 3 summarizes the feature-learning-based approaches used in EASR and SER tasks.

The feature learning approaches include encoding with a pre-calculated codebook method and a stage of pooling to get lower-dimensional representations. Widely used approaches are sparse code representation (Lu et al. 2014), Bag of Aural wWords (BoAW), Non-negative Matrix Factorization (NMF) (Stowell et al. 2015), and exemplar-based coding (Stowell and Clayton 2015; Gemmeke et al. 2013). In Plinge et al. (2014), the multinomial maximum-likelihood classifier with Bag of Super Features-Pyramid (BOSF-P) approach is experimented for sound event recognition. The Mel and gammatone frequency cepstral coefficients are used for building the bag-of-features representations. In the smart room environment, the codebook approach using supervised learning with temporal encoding showed improved performance in recognizing sound events.

In previous studies, frame-based classification of extracted features using the bag-of-frames approach over a majority voting algorithm was popularly employed. However, for realistic audio data, it did not score better than a mere one-point average of the audio features. This approach was replaced by learning an internal acoustic representation from spectral information or similar representation of an audio signal by optimizing their parameters. More recent studies combine the feature engineering approach with feature learning approach for audio scene recognition. The learned features (high-level reduced acoustic representation) such as i-vectors and supervectors were most likely constructed on top of engineered features such as MFCC and log Mel-band energies, probably giving better robustness or better recognition accuracy. For instance, a low-dimensional feature space was created for short audio segments with the support of i-vector (identity vector) features (Dehak et al. 2011) using MFCCs. With the help of a factor analysis method, the i-vector features are estimated from the adapted mean supervectors. Supervectors give more features with huge dimension. The size of the i-vector is larger than the underlying MFCC feature vector but much smaller than the supervector. Similarly, the authors in Agcaer et al. (2015) optimized Amplitude Modulation Spectrogram (AMS) features using Covariance Matrix Adaption Evolutionary

Strategy (CMA-ES). An interesting alternative approach is proposed by Phan et al. (2016), where a phone triplet approach is used to convert non-speech data into speech data. In Eghbal-zadeh et al. (2017), the parametrization of MFCCs is proposed with i-vectors approach. A more detailed approach is proposed (Rakotomamonjy and Gasso 2015) for comparing MFCC with RQA to the HOG features for capturing the spectral characteristics present within the spectro-temporal representation. HOG-marginalized features outperformed the MFCC with the RQA approach, because the MFCC cannot encode the direction of local variations of the signal power spectrum. In Phan et al. (2017), the features were represented in the form of Label Tree Embedding (LTE) images. Other kinds of representation techniques used for representing sound events are Deep Neural Network (DNN)-based approach. The Deep Belief Network (DBN) with stack pre-trained Restricted Boltzmann Machines (RBMs) is used for the intermediate representation of sound events (Bengio et al. 2013).

In Shuyang et al. (2017), a sound event is represented using MFCC features with K-medoid clustering-based active learning method to eliminate the annotation procedure. The work reported by Dorfer et al. (2018) describes the environmental scene recognition approach for Detection and Classification of Acoustic Scenes and Event (DCASE 2018) task1 called Audio Scene classification. In the proposed approach, three different deep Fully Convolutional Neural Networks (FCNN) are trained on perceptually weighted spectrograms and i-vector systems based on MFCCs and fused via linear logistic regression. Better accuracy is achieved by fusing this neural network system with i-vectors approach. Recently, the work in Ye et al. (2019) proposed a novel approach based on unsupervised acoustic feature learning (L1-penalized sparse coding and spherical k-means dictionary learning) and construction of data-driven taxonomy (data-driven clustering analysis). Finally, the Hierarchical Regularized Logistic Regression (HR-LR) model-based classification algorithm is combined with the constructed taxonomy structure to improve multi-class hazard sound event recognition. Thus, the discriminative characteristics of non-speech events can be learned with the help of new feature spaces using effective feature learning methods (Dorfer et al. 2018; Plinge et al. 2014).

3.5 Feature Fusion-based Approaches

The main problem to be faced in the EASR task is the fusion of knowledge from multiple devices, which can be carried out in feature extraction level or decision level. The feature fusion-based approach will ensure capture of the complete characteristics of signal. The system typically achieves enhanced system performance or robustness, which receives much attention from research groups in audio forensics, acoustic localization, and multimedia information retrieval. Table 4 shows some of the methods explored in the literature for EASR and SER tasks.

In Rabaoui et al. (2008), a discriminative method with a dissimilarity measure was proposed to recognize an acoustic event from the trained class. The efficiency of various acoustic features, such as discrete wavelet coefficients, MFCCs, energy, log energy, spectral roll-off, spectral centroid, ZCR, as well as the influence of a combination of features were studied. In Piczak (2015b), an analysis for classifying sound events was provided using the fusion of MFCCs and ZCR features. The proposed approach has given better performance when compared with human recognition performance. A common approach based on the conversion of Bag of Aural-Words (BoAw) approach with the feature set using spectral, temporal, and energy features to construct the short-time descriptor was proposed in Foggia et al. (2015).

The method presented in Chu et al. (2009) is aimed to propose a novel feature extraction method that utilizes Matching Pursuit (MP) algorithm to select a small set of time-frequency features. The MP-based feature supplemented the MFCC features to obtain maximum performance for

Table 4. Feature Fusion-based Approaches for Environmental Audio Scene and Sound Event Recognition Tasks

Features	EASR task	SER task
MFCCs and zero crossing rate		(Piczak 2015b)
Fusion of wavelet and Mel-Frequency Cepstral Coefficient (MFCC)	(Eghbal-zadeh et al. 2017)	(Li et al. 2013)
MP + MFCC features	(Zieger and Omologo 2008)	(Chu et al. 2009)
Mel + Gammatone frequency cepstral coefficients (BOF approach)		(Grzeszick et al. 2017)

environmental audio sounds. MP features with generative classifier are used to recognize environmental sounds compared with pure frequency-domain features. Another work, presented in Li et al. (2013), used the fusion of wavelet and MFCC features for classifying various audio scenes. In Grzeszick et al. (2017), Mel and Gammatone frequency cepstral coefficients are used as a generic feature set for the Bag-of-Features (BoF) concept. In a BOF approach, acoustic features calculated from frames are quantized with respect to the learned codebook. Subsequently, the histogram representation is calculated and given as input to a maximum likelihood classifier. Furthermore, the extensions of various BoF approaches, such as soft quantization, supervised codebook learning, and temporal modeling are explored. The supervised codebook learning with the maximum likelihood classifier outperformed the other state-of-the-art approaches in the literature. In contrast to the deep learning approach in the DCASE 2013 challenge, BoF approaches performed slightly well in real-time scenarios where training data is limited. To conclude, feature fusion-based approach will take more testing time and may affect real-time surveillance performance (Foggia et al. 2015; Rabaoui et al. 2008). In this section, reviews of feature engineering (generic features) approaches, auditory-image based approaches, feature learning approaches, and feature fusion approaches are discussed.

Some of the tasks related to audio surveillance applications are as follows: Environmental Audio Scene Recognition (EASR) (Mafra et al. 2016), Sound Event Recognition (SER) (Stowell et al. 2015), audio tagging (Xu et al. 2017a), Ambient Assisted Living (AAL) (Beltrán et al. 2015; McLoughlin et al. 2015), and fall detection (Cheffena 2016). We focus on the two primary tasks; namely, Environmental Audio Scene Recognition (EASR) and Sound Event Recognition (SER) tasks.

4 ENVIRONMENTAL AUDIO SCENE RECOGNITION

4.1 Methodologies

Generally, environmental audio scene recognition task consists of the following two steps: (1) feature extraction and (2) a modeling method that decides the underlying environment of the given test example. In Section 3, we have reviewed various audio features involved in the literature. A simple categorization divides model-based approaches into generative-model-based, discriminative-model-based, deep-learning-based, and hybrid-model-based approaches.

4.1.1 Generative Model-based Approaches. In this category, for each environment class, a model is built using examples belonging to that class alone. The class label for a test example is assigned based on the class for which the log-likelihood score is maximum. However, among various generative models, Hidden Markov Models (HMM), Gaussian Mixture Models (GMM), and their

Table 5. Generative Model-based Approaches Used in the Environmental Audio Scene Recognition (EASR) Task

Features	Methodologies	Datasets	Ref
MFCC	HMM	Own dataset	(Ma et al. 2003)
MFCC	HMM	RWCP scene dataset	(Ma et al. 2006)
Gabor features	HMM	Own dataset	(Rabaoui et al. 2009)
MFCC	HMM	Own dataset	(Dargie 2009)
Spectral features	GMM	Noise scenes	(El-Maleh et al. 1999)
MFCC (Bag-of-Frames)	GMM	DCASE2013 dataset	(Stowell et al. 2015)
MFCC	GMM	TUT-DCASE 2016	(Mesaros et al. 2016)

variances are the most widely used models in an EASR task (Chandrakala and Chandra Sekhar 2010; Dufaux et al. 2000; Stowell et al. 2015). A comprehensive exposition of generative-model-based approaches employed in EASR can be found in Table 5.

A Hidden Markov Model (HMM) represents stochastic sequences as hidden states (Markov chains), but are incorporated with a probability density function (pdf). The random sequence is generated by emission of visit in each state. The observation probability density for a state is modeled using a Gaussian Mixture Model (GMM-HMM) or Deep Neural Network (DNN-HMM). The methods reported in Table 5 used the GMM-HMM approach for modeling. If HMM allows the transitions from any emitting state to any other emitting state, then it is known as an ergodic HMM. However, if HMM allows the transitions only to go from one state to itself or to a unique successor, it is called a left-right HMM. In Ma et al. (2003), HMMs on MFCCs are employed to recognize various auditory noise scenes. The hidden states of the HMMs were varied to achieve better context recognition accuracy. Notably, the authors achieved a low accuracy for a higher number of HMM states. Another work of Ma et al. (2006) proposed the MFCC features with HMM classifier to discriminate various environmental scenes and sound events. The experimental results showed that the sound events are perfectly modeled using five-state HMMs. Similarly, the environmental audio scenes are modeled using nine-state HMMs. The HMM classifier performed well for recognizing short duration examples in acoustic scenes when compared with human listening tests. HMMs are suited for accurate modeling of the temporal variations in the sequence of feature vectors over consecutive frames. Impulsive sounds, such as gunshots or screams, have typical temporal signatures that can be captured with left-right HMMs (Rabaoui et al. 2009), whereas stationary sounds such as alarms, dogs barking, and baby crying can be efficiently modeled by ergodic HMMs. The HMM classifier with Gabor features provides better recognition accuracy for impulsive sounds. Another work in Dargie (2009) employed the MFCC features with left-to-right HMM for adaptive context recognition. The proposed approach was deployed in the iBadge device to monitor individual children activities in nursery schools to check the student's nature (social or aggressive).

Gaussian Mixture Model (GMM) is another generative model-based classifier based on the linear combination of Gaussian components. It is used to model the pdf of a multi-dimensional feature vector. The GMM–Universal Background Model (GMM-UBM) is considered as a variation of the GMM approach. In El-Maleh et al. (1999), the authors employed GMMs on spectral features to classify the following five environmental noise scenes: street, car, babble, bus, and factory. They achieved various recognition accuracies by changing the number of Gaussian components of GMMs.

In Stowell et al. (2015), the GMM with MFCC features was proposed to recognize 10 environmental audio scenes. The authors used the standard approach called the bag-of-frames model to

Table 6. Discriminative-Model-based Approaches Used in the Environmental Audio Scene Recognition (EASR) Task

Features	Methodologies	Datasets	Ref
MFCC	Discriminative training of HMM	Own dataset	(Eronen et al. 2006)
MFCC	Discriminative training of GMM	TUT-DCASE 2016	(Yun et al. 2016)
Various features	SVM	Own dataset	(Lu et al. 2001)
RQA+MFCC	SVM	Own dataset	(Roma et al. 2013)
HOG marginalized features	SVM	Own dataset	(Rakotomamonjy and Gasso 2015)
Various features	ANN	Own dataset	(Dhanalakshmi et al. 2011)
log Mel-band energies	Multilayer Perceptron architecture (MLP)	DCASE 2017	(Mesaros et al. 2017)

ignore the concept of frames with sequence order as input. In Mesaros et al. (2016), the baseline system is proposed with a classical Gaussian Mixture Model (GMM)-based classifier using the Mel Frequency Cepstral Coefficient (MFCC) features. A GMM is built for each acoustic scene with 32 components using an expectation maximization algorithm. The maximum likelihood decision is calculated for labeling all-acoustic scene models. The methods in Table 5 reviewed the various feature extraction processes and experiment evaluations used in the literature using both HMM and GMM classifiers.

To conclude, when the data of different classes are more confusable due to similar characteristics among different environmental audio scenes, generative model-based classifiers such as the Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM) are not suitable, since a model for each class is built using the samples of that class alone. It gives an idea of a discriminative-model-based approach to classify the examples of environmental audio scenes.

4.1.2 Discriminative Model-based Approaches. Discriminative-model-based classifiers such as SVM and Artificial Neural Network (ANN) focus on constructing a hyper-plane between environmental audio scene classes. SVM is a kernel-based discriminative classifier that focuses on modeling the decision boundaries between classes (Vapnik 1998). The SVM-based classifier gives a good generalization performance to classify the unseen data. It performs well, not only for linearly separable data, but also for non-linear data. ANNs may be used as non-parametric discriminant functions or as universal function approximators in recognition tasks. Thanks to their characteristics, ANNs have been applied to audio scene recognition over HMM- and GMM-based approaches. Multilayer Perceptron (MLP) is one of the popular neural network architectures used for recognition.

In Table 6, a summary of several discriminative-model-based approaches for environmental audio scenes is reported. In Eronen et al. (2006), the experiment is carried out on spectro-temporal domain features to recognize various indoor and outdoor environments (streets, vehicles, restaurants, offices/meetings, and homes). The discriminative training of low-order HMMs with MFCC has consistently shown better recognition accuracies compared with human listening accuracy. An experimentation of a discriminative training algorithm to increase the performance of the conventional GMM model was proposed in Yun et al. (2016). The conventional GMM parameters were updated by maximizing the margin between classes. The hierarchical GMM classifier was applied for the classification task.

In Lu et al. (2001), SVMs for classifying music, background sound, silence, pure speech, and non-pure speech audio classes were presented. The feature sets used for representing the various classes include brightness, bandwidth, MFCC, ZCR, STE, sub-band power distribution, spectrum flux, band periodicity, and Noise Frame Ratio (NFR). Some of the newly introduced features in this set are SF, band periodicity, and NFR. It also showed better accuracy for the SVM-based approach compared with the approaches based on KNN and GMM. The embedding of the RQA method into MFCC features was proposed to achieve a robust recognition with SVM as a classifier (Roma et al. 2013). In another work, a comparative analysis of MFCC with RQA to the Histogram of Oriented Gradients (HOG) features was presented for capturing the spectral characteristics present within the TFR (Rakotomamonjy and Gasso 2015). HOG-marginalized features with SVM outperformed the MFCC with RQA approach, because the MFCC cannot encode the direction of local variations of the signal power spectrum.

An approach using LPC, LPCC, and MFCC acoustic features to characterize various audio contents was proposed in Dhanalakshmi et al. (2011). The Autoassociative Neural Network model (AANN) is used to model the distribution of acoustic feature vectors of a class using the Backpropagation (BP) learning algorithm. The proposed AANN-based method outperformed the GMM approach. In Mesaros et al. (2017), a system using log Mel-band energies with Multilayer Perceptron architecture (MLP) is presented for robust environmental audio scene recognition. However, in the case of ANNs, as the learning is based on the principle of empirical risk minimization, it does not guarantee a good performance on test data. It needs a huge amount of training data and takes more time for training. SVMs are found based on the principle of structural risk minimization that guarantees good generalization performance (Vapnik 1998).

To summarize, in the discriminative approach, a model is built for examples of that class with positive examples only and examples of other classes are considered as negative examples. The SVM approach constructs a hyper-plane with maximal margin to provide better generalization for non-linearly separable data using kernel function. Hence, SVM is still proved effective in most of the complex recognition task. The challenges addressed by the SVM approach are difficult in hyper-parameter tuning and it can handle data only in the form of fixed dimensional representations (Chandrakala and Chandra Sekhar 2010). In the ANN approach, the estimation of training parameters (weight and bias vector) depends on the supervised BP algorithm. However, these methods are not efficient when the number of neurons and hidden layers increases. The parameter tuning is also difficult in ANN. To address the issues in discriminative classifiers, there are some recent works proposed towards a representation learning framework that maps sequences to global vector representations (Eghbal-zadeh et al. 2017; Phan et al. 2017).

4.1.3 Deep Learning Model-based Approaches. The complex recognition task with more data can be effectively handled by deep learning methods where conventional machine learning methods do not guarantee better performance. Table 7 shows different deep-learning-model-based approaches employed for the audio scene recognition task. DNN works on an unsupervised pre-training step using probabilistic graphical models such as RBM and DBN to initialize the parameters. Convolutional Neural Network (CNN) is one of the widespread architectures used in deep learning approaches. The DNN-based approach for acoustic scene recognition has been proposed in Petetin et al. (2015) using MFCC, spectral centroid, and spectral flatness features. DNN outperformed the classical classifiers such as GMM and SVM with the same features. The results have been exceptionally good for DNN with cepstral and frequency features compared with well-known features such as HOG classified by the SVM approach.

In Han and Lee (2016), multi-width frequency-delta data augmentation was applied on input features for training using the convolutional neural network. The frequency-delta features and

Table 7. Deep Learning–based Approaches Used in the Environmental Audio Scene Recognition (EASR) Task

Features	Methodologies	Datasets	Ref
Cepstral (MFCC) features, spectral centroid and spectral flatness	DNN	LITIS Rouen dataset	(Petetin et al. 2015)
Multi-width frequency-delta data augmentation (Melspectrogram and frequency-delta features)	CNN	TUT-DCASE 2016	(Han and Lee 2016)
Temporal averaged Mel-log spectrograms	SVM	DCASE 2013	(Mafra et al. 2016)
Parametrized MFCCs using i-vectors	CNN	TUT-DCASE 2016	(Eghbal-zadeh et al. 2017)
Multi-channel i-vectors (extracted from MFCC features)	Visual Geometry Group (VGG)-net CNN architecture	TUT-DCASE 2016	(Phan et al. 2017)

Melspectrogram are used as input features for data augmentation to represent examples with same labels. Another work in Mafra et al. (2016) reviewed different time granularities for combining the features using SVM, MLP, and CNN classifiers. The proposed compact representation with temporal averaged Mel-log spectrograms using SVM achieved better recognition accuracy. However, DNN with a log spectrogram approach has not performed well compared with many of the classical SVM-based approaches. The authors in Phan et al. (2017) proposed an approach called Convolutional Neural Network–Label Tree Embeddings (CNN-LTE) strategy. Using the CNN-LTE approach, the features were represented in the form of label tree embedding images. Then these features were learned using the simple 1-X pooling CNNs. In Eghbal-zadeh et al. (2017), the parametrization of MFCCs using the i-vector approach is investigated. To display the effectiveness of audio scene classification, a hybrid system using multi-channel i-vectors and CNN was proposed. This hybrid approach utilized score fusion techniques to capture the complementary information from indoor and outdoor scenes. To conclude, the DNN needs a substantial amount of labelled training data and takes more time for training. The deeper structures make the training process more tedious, and hence the estimation of large numbers of parameter values (Mesaros et al. 2018b).

4.1.4 Hybrid Model-based Approaches. More recent studies combine the discriminative approach with generative learning approach for audio scene recognition. The hybrid frameworks are used for recognition to achieve better generalization with a smaller amount of training data. The authors in Zieger and Omologo (2008) presented a hybrid model using GMM combined with SVM by utilizing the log energy first and second derivatives as features. Finally, the recognition is performed on the combined score with the maximum value. In Chit and Lin (2013), Zero-crossing Rate (ZCR), Short-time Energy (STE), Volume Root Mean Square (VRMS), Volume Dynamic Range (VDR), and MFCC are extracted from given audio clips. Then the extracted features are used by multiple HMMs to train the SVM classifier. The hybrid-model-based approaches employed in the EASR task can be found in Table 8.

In this section, reviews of generative-model-based, discriminative-model-based, deep-learning-based, and hybrid-model-based approaches are discussed for EASR tasks.

Table 8. Hybrid Model-based Approaches Used in the Environmental Audio Scene Recognition (EASR) Task

Features	Methodologies	Datasets	Ref
Log energy first and second derivatives	GMM combined with SVM	Own dataset	(Zieger and Omologo 2008)
Zero-Crossing Rate (ZCR), Short-time Energy (STE), Volume Root Mean Square (VRMS), Volume Dynamic Range (VDR) and MFCC	Multiple HMMs and SVM	Own dataset	(Chit and Lin 2013)

Table 9. Datasets Used in Environmental Audio Scene Recognition (EASR) Task

Datasets	References	Environmental audio scene classes
DARES-G1 (10 contexts)	(Heittola et al. 2013)	Basketball game, beach, inside an office facility, inside a bus, grocery shop, inside a car, street, hallways, restaurant, and stadium with track and field events.
LITIS-Rouen audio scene dataset (19 contexts)	(Rakotomamonjy and Gasso 2015)	Busy street, bus, cafe, car, train station hall , kid game hall, market, metro-Paris, metro-Rouen, high-speed train, billiard pool hall, quiet street, plane, student hall, restaurant, pedestrian street, shop, train, and tube station.
TUT-CASR (18 contexts)	(Eronen et al. 2006)	Six high-level classes are: (1) vehicles, (2) outdoors, (3) public/social, (4) offices/meetings/quiet, (5) home, and (6) reverberant places.
IEEE/AASP DCASE 2013 Scene dataset (10 contexts)	(Phan et al. 2017; Agcaer et al. 2015; Stowell et al. 2015; Rakotomamonjy and Gasso 2015)	Bus, busy street, restaurant, office, open-air market, park, quiet street, supermarket, tube (subway train), and tubestation (subway station).
TUT-2016 dataset (15 contexts)	(Mesaros et al. 2016)	Beach, bus, cafe, car, city center, forest path, grocery store, home, library, metro station, office, Residential area, train, tram, and urban park.
TUT-2017 dataset (15 contexts)	(Mesaros et al. 2017; Ren et al. 2017b; Mun et al. 2017; Xu et al. 2017b)	Bus, cafe/restaurant, car, city center, forest path, grocery store, home, lakeside beach, library, metro station, office, residential area, train (traveling, vehicle), tram (traveling, vehicle), and urban park.

4.2 Datasets

Most of the developed audio scene datasets are not publicly accessible (Classification of Events, Activities and Relationships 2007 evaluation). Table 9 summarizes some of the publicly available datasets commonly used in audio scene recognition. The design criteria considered for developing the scene datasets are as follows: (i) Most of the environmental scenes are collected from real-life situations to give a more realistic view; (ii) A balanced number of examples are used for controlled experimental analysis. In Heittola et al. (2013), the DARES-G1 dataset contains scenes from typical scenarios, such as conveyances (car and bus), public space (grocery shop and restaurant), and leisure time (beach, basketball game, and fields). The performance of the system is good for all

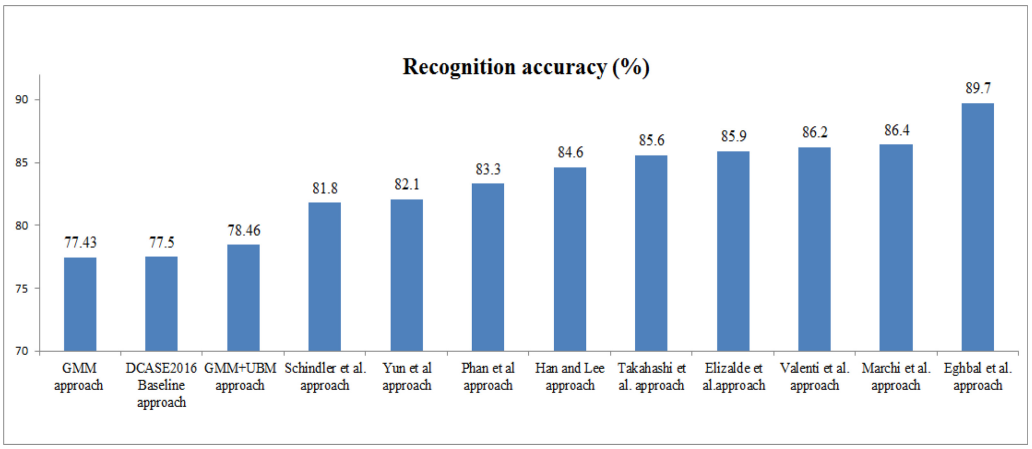

Fig. 2. Experimental studies on the DCASE 2016 dataset for EASR task.

scenes, since the recordings for the same scene were done around the same geographical location, e.g., along the same street. Similarly, in the LITIS-Rouen dataset, modes of transportation such as plane, bus, car, train, and high-speed train are exactly recognized when compared with metro-Rouen and metro-Paris. More confusions have occurred in the group of crowd scenes in which people are walking; namely, pedestrian street, market, train station hall, quiet street, and shop. The main confusion between a pedestrian street and quiet street would be the number of people walking in the scene.

IEEE AASP Challenge dataset (DCASE 2013) (Stowell et al. 2015) shows 10 different urban scenes with 5 indoor and 5 outdoor locations. These audio scenes were recorded in the London field. The dataset has two disjoint datasets: public (development) and private (evaluation) datasets, each containing 30s long, 10 classes with 10 instances per class. Another publicly released TUT-DCASE2016 (Mesaros et al. 2016) dataset is similar to the previous DCASE 2013 challenge, but with a higher number of classes and range of examples. It consists of 15 acoustic scenes recorded in different locations. The confusion made by automated systems in Mesaros et al. (2018b) for a pair of classes such as {*Park, Residential area*}, {*Home, Library*}, and {*Train, Cafe*} are similar even for a human recognition test. Recently, the TUT Acoustic Scenes 2017 dataset was released with two subsets: development dataset and evaluation dataset. The development dataset consists of both development and evaluation data of the TUT-2016 Challenge, and the evaluation dataset contains recordings of similar audio scenes but from different locations. The TUT-Challenge datasets have the following real-life scenes in common: beach, bus, grocery shop, office, park, restaurant, street, and train.

4.3 Studies on Environmental Audio Scene Recognition

4.3.1 Dataset Description. The publicly released TUT-DCASE 2016 (Mesaros et al. 2016) Challenge dataset consists of two types of datasets. The development dataset consists of 15 acoustic scenes. Each audio scene has 78 examples with a summary of 39mins of audio. It was used for training the system. The evaluation dataset contains 26 unknown examples per class that were used only for testing the system.

4.3.2 Result Analysis. We have studied the conventional GMM and GMM-UBM approaches on the DCASE 2016 dataset. Comparison of various approaches reported in the literature is shown in Figure 2. In our experiments, MFCC features are used as the basic features to learn representation. We consider diagonal covariance matrices for both approaches. The first method used is the

conventional GMM approach in which the GMM model is used where the parameters are estimated using the maximum likelihood method. The Gaussian mixture components are used to build the GMM for each class. The amount of Gaussian components was experimentally chosen from the set {8, 16, 32, 64, 128, 256} considering diagonal covariance matrix. The best performance is obtained for GMM with 32 components. In case of GMM-UBM-based approach, the best performance is obtained for 64 components. In the GMM-UBM approach, we build the class model by adapting the means, variances, and mixture coefficients of UBM using the class's training example and a form of Bayesian adaptation with the relevance factor (Reynolds et al. 2000) taken as 16.

Figure 2 shows the recognition accuracy of a conventional GMM, adapted Gaussian Mixture Model (GMM-UBM) and some of the state-of-the-art approaches used in the DCASE 2016 dataset. In the DCASE 2016 Challenge, most of the submitted systems outperformed the baseline system (Mesaros et al. 2016). A large number of submissions in DCASE 2016 used MFCCs (Elizalde et al. 2016b) or log Mel energies (Valenti et al. 2016) as audio features. These features provide better characterization of the spectral features and also give high discrimination among different environmental audio scenes. Other features include time-frequency domain-based Constant-Q-Transform (CQT) features (Lidy and Schindler 2016) and a fusion of various features (Eghbal-Zadeh et al. 2016; Marchi et al. 2016). Similarly, the majority of systems used deep learning (Valenti et al. 2016) and SVM (Elizalde et al. 2016b) classifiers for the recognition task. Mesaros et al. (2016) provided the baseline system using MFCC with GMM-based classifier for 15 environmental audio scenes. The CNN (Lidy and Schindler 2016) was trained using the Mel scale with Constant-Q-Transform (CQT) representations to capture low and mid-to-low frequencies. In Yun et al. (2016), a discriminative training algorithm to increase the performance of the conventional GMM model is explored. The conventional GMM parameters were updated by maximizing the margin between classes. The hierarchical GMM classifier was applied for the classification task. A novel approach called CNN-LTE strategy using the simple 1-X pooling CNNs is proposed in Phan et al. (2017).

In Han and Lee (2016), the multi-width frequency-delta data augmentation method for the convolutional neural network is proposed. The data augmentation uses frequency-delta features and Mel-spectrogram for single examples with same labels. The DNN-GMM approach (Takahashi et al. 2016) was applied using the novel high-dimensional features for acoustic scene classification. The high-dimensional features were constructed by concatenating the acoustic features in temporally adjacent frames. In Elizalde et al. (2016b), SVM classifier with MFCC distribution was proposed for recognition of real-life recordings. CNN with log-Mel spectrogram approach was proposed to recognize short sequences of an audio signal (Valenti et al. 2016). The pairwise decomposition with binary deep multilayer perceptron neural networks (Marchi et al. 2016) was proposed to recognize various environmental audio scenes. Subsequently, the dimension of audio features was reduced using multiscale Gaussian kernel subspace learning. In Elizalde et al. (2016b), a novel i-vector-based approach was proposed using a Deep Convolutional Neural Network (DCNN) architecture. In summary, the cepstral features with CNN approach show an improved performance for the EASR task.

5 SOUND EVENT RECOGNITION

5.1 Methodologies

Information from the environmental audio scene recognition system can provide additional information about an environment to the sound event recognition system. SER in the real-world environment (Lee et al. 2013; Ntalampiras et al. 2011) differs from the classification of isolated events in a silent environment (Crocco et al. 2016; Lozano et al. 2010; McLoughlin et al. 2015). A summary of generative- and discriminative-model-based approaches is shown in Tables 10

Table 10. Generative Model-based Approaches Used in Sound Event Recognition (SER) Task

Features	Methodology	Datasets	Ref
MFCC	GMM	Own dataset	(Radhakrishnan et al. 2005)
MFCC, Perceptual wavelet packet integration analysis, Intonation and Teager Energy Operator (TEO)-based features, and MPEG	GMM	Own dataset	(Ntalampiras et al. 2009)
MFCC	HMM	CLEAR 2007	(Mesaros et al. 2010)
MFCC, Intonation and Teager Energy Operator (TEO) based features, MPEG-7 audio protocol and Perceptual wavelet packet integration analysis	Universal HMM (general purpose security and ATM scenarios) and GMM Clustering (smart-home)	PROMETHEUS	(Ntalampiras et al. 2011)
MFCC	GMM background model	IEEE/AASP DCASE-2013	(Vuegen et al. 2013)
log-mel-spectrogram features	k-means algorithm	UrbanSound8k	(Salamon and Bello 2015)
MFCC	GMM	TUT-DCASE 2016	(Mesaros et al. 2016)
GTCC + TEO-GTCC feature sets	GMM	UrbanSound8k	(Agrawal et al. 2017)

and 11. Similarly, Tables 12 and 13 tabulate the deep learning and hybrid model-based approaches for an SER task. Over the past few years, several studies (Phan et al. 2015; Stowell et al. 2015) were conducted for SER in several applications such as audio surveillance, audio forensics, AAL, and audio indexing for video retrieval. Various sound features such as ZCR (Salamon et al. 2014), MFCC (Mafra et al. 2016), GBFB features (Schroder et al. 2013), wavelet packets (Rabaoui et al. 2008), and SIFs (McLoughlin et al. 2015) have been studied for the SER task.

5.1.1 Generative Model-based Approaches. In this category, HMM is used for modeling the audio data of varying lengths in a class. In the case of SER, approaches based on statistical models use HMMs for modeling feature vectors of sound events. A comprehensive analysis of generative-model-based approaches employed in SER can be found in Table 10. A hybrid framework for monitoring elevators is proposed in Radhakrishnan et al. (2005). The adaptive background modeling technique used in the hybrid framework first builds GMM from MFCC features and then updates the parameters of components in the background GMM. In Mesaros et al. (2010), the MFCC features with HMM-based classifier is proposed to recognize the acoustic events from real-time recordings with different background noises.

The recognition of human vocal sounds (i.e., screams, expressions of pain) and non-vocal abnormal sounds related to harmful circumstances (gunshot and explosions) was explored in Ntalampiras et al. (2009). The discrimination between vocal and non-vocal sounds was performed by using various features such as MFCC, autocorrelation envelope area, pitch, Harmonic to Noise Ratio (HNR), Teager Energy Operator (TEO), and MPEG-7 audio standards. Gaussian Mixture Models (GMM) were used for modeling the distribution of various sound events. Another work in

Table 11. Discriminative Model-based Approaches Used in Sound Event Recognition (SER) Task

Features	Methodology	Datasets	Ref
MFCC, Linear Prediction Coefficients (LPC) and the Perceptual Linear Prediction (PLP) Coefficients	SVM	SAMSIT project	(Rouas et al. 2006)
Wavelet coefficients	SVM	RWCP	(Rabaoui et al. 2008)
MFCC and deltas coefficients	Multiclass AdaBoost based acoustic context classifier	Own dataset(abnormal events or normal events)	(Lee et al. 2013)
MFCC	SVM	UrbanSound8K	(Salamon et al. 2014)
MELMBSES (MEL-Multiband spectral entropy signature)-H1DH2D	SVM	CICESE	(Beltrán et al. 2015)
MFCC and ZCR	Random forest ensemble	ESC-50 and ESC-10 dataset	(Piczak 2015b)
Acoustic superframes	Random regression forest	ITC-IRST and UPC-TALP datasets	(Phan et al. 2015)
Phone Triplets and Low-Level Acoustic Features	SVM	UPC-TALP, Freiburg-106, and NAR datasets	(Phan et al. 2016)
MFCC with GMM	Kernel Ridge Regression (KRR)	YLI-MED	(Jing et al. 2017)

Table 12. Deep Learning Model-based (CNN-based) Approaches Used in Sound Event Recognition (SER) Task

Features	Methodology	Datasets	Ref
Spectrogram Image Features (SIF)	DNN+Multi-Condition (MC)	RWCP dataset	(McLoughlin et al. 2015)
Log-scaled mel-spectrograms	CNN	ESC-50, ESC-10 and UrbanSound8K dataset	(Piczak 2015a)
log mel-band energies	CRNN	TUT-SED Synthetic 2016 dataset	(Cakir et al. 2017)
Log-scale Mel-spectrogram features	Fully CNN	UrbanSound8K	(Su et al. 2017)
Spectrogram Image Features (SIF)	CNN	RWCP dataset	(Ozer et al. 2018)
Auditory Receptive-Field Binary Pattern (ARFBP)	Hierarchical-diving Deep Belief Network (HDDBN)	TUT-SED 2016 dataset	(Wang et al. 2018)

Ntalampiras et al. (2011) extended the work in Ntalampiras et al. (2009) by employing a wide variety of acoustic parameters such as MFCC, perceptual wavelet packet integration analysis, intonation and TEO-based features, and MPEG-7 audio features for recognizing various sound events. These features were used to form a multi-domain feature vector to represent the sounds. Subsequently, the extracted feature coefficients are applied to three probabilistic novelty recognition methodologies such as universal HMM, universal GMM, and GMM clustering. The GMM

Table 13. Hybrid Model-based Approaches Used in Sound Event Recognition (SER) Task

Features	Methodology	Datasets	Ref
ZCR, LFCC, LPC, and LPCC	Multiple GMMs	Own dataset	(Atrey et al. 2006)
Gabor Filter-bank Feature (GBFB)	2-layer HMMs	DCASE-2013	(Stowell et al. 2015)
Bag of aural-words (spectral, temporal and energy features)	Pool of SVM classifiers	MIVIA	(Foggia et al. 2015)

clustering approach provided the best performance for a smart-home scenario, and the universal HMM method provided a slightly better outcome for the general purpose security and Automatic Teller Machine (ATM) scenarios. The authors in Vuegen et al. (2013) explored GMM using MFCCs for acoustic event recognition. To bound the impact of silence, a shared GMM background model is used. In Salamon and Bello (2015), the log-mel-spectrogram features with K-means algorithm is proposed to recognize various urban sound events with different background noises. To summarize, these generative models are not efficient for overlapping/noisy sounds and partial sounds such as gunshot, baby cry, and dog barking sounds. In Mesaros et al. (2016), the baseline system is proposed with a classical GMM-based classifier using MFCC features. A binary classifier was used for modeling each event class. In testing, likelihood estimate between the positive and negative models for each individual class is considered for taking decision. The authors in Agrawal et al. (2017) proposed the score level fusion of TEO-based GTCCs (TEO-GTCC) with MFCC and GTCC features. With GMM, GTCC + TEO-GTCC feature sets achieved a good classification accuracy for various sound events.

5.1.2 Discriminative Model-based Approaches. Various discriminative-model-based approaches employed in SER can be found in Table 11. The work in Rouas et al. (2006) proposed a system with SVM classifier using the fusion of features such as Linear Prediction Coefficients (LPC), MFCC, and PLP for recognizing various sounds. Similarly, in Rabaoui et al. (2008), the efficiency of a combination of features for various sound events is reviewed. The proposed optimized one-class SVMs with the set of wavelet coefficients approach performed better than the state-of-the-approaches reported in the work.

SER in real-life environments reported in Lee et al. (2013) and Ntalampiras et al. (2011) differs from the classification of isolated events in a silent environment (McLoughlin et al. 2015; Crocco et al. 2016; Lozano et al. 2010). The work in Lee et al. (2013) recognized an event by analyzing the acoustic event as either normal or abnormal from extracted MFCC and delta coefficients. A novel concept was proposed to boost the binary weak classifiers by adopting an exponential criterion with the weighted least-square solution. In Salamon et al. (2014), the system using MFCC features with an SVM classifier is demonstrated for urban environments. It involved challenging sound events from real-life environments with more confusable classes. The authors in Beltrán et al. (2015) proposed a representation based on spectral band-filtered frame level feature to produce a standard time-based representation. Then, a time-independent representation MELMBSES (MEL-Multiband Spectral Entropy Signature) is computed using the normalized histogram for each band. After the histogram representation, the class models are built using a linear SVM classifier.

In Piczak (2015b), a baseline system using MFCC and ZCR features is proposed to explore the potential drawbacks of the ESC dataset. The three types of classifiers such as k-nearest neighbors (k-NN), random forest ensemble, and support vector machine were explored for analysis of the dataset, in which the random forest ensemble classifier outperformed the remaining two classifiers.

In Phan et al. (2015), an approach using acoustic superframes and a random-forest regression model for learning is proposed. Recently, the phone triplets approach was proposed in Phan et al. (2016), and it was different from approaches used in Chin and Burred (2012). They attempted to create a general descriptor using the phone triplets approach with low-level acoustic features such as zero crossing rate, log-frequency filter bank coefficients, their first and second derivatives, spectral centroid, spectral bandwidth, short-time energy, and sub-band energies. This approach converts the unstructured non-speech data into structured data similar to speech information. These feature learning techniques construct a discriminative feature space on top of generic features to learn the heterogeneity among various classes using the SVM classifier.

The authors in Jing et al. (2017) presented a novel two-phase Discriminative and Compact Audio Representation (DCAR) method for sound events. In the DCAR approach, each example is modeled using a GMM to capture the variability within that example. Additionally, the global and the local structure of an example are taken into consideration. The Kernel Ridge Regression (KRR) method was used to build the audio scenes and sound events.

5.1.3 Deep Learning Model-based Approaches. Table 12 shows different deep-learning-model-based approaches employed for sound event recognition. The first approach using a DNN classifier with the time-frequency feature called Spectrogram Image Features (SIF) for sound event recognition was proposed in McLoughlin et al. (2015). The proposed system was evaluated with both SVM and DNN classifiers and results showed that the DNN classifier with simple de-noising performed well for the recognition task. In Piczak (2015a), CNNs are trained on manually engineered spectrogram features to achieve a similar level of output as other deep learning methods (DNN and RNN). The use of multi-label Convolutional Recurrent Neural Network (CRNN) for polyphonic, scene-independent SED in real-life recordings was proposed in Cakir et al. (2017). The time-frequency representation (log-Mel band energies) with CNN was demonstrated. In CRNN, learning local translation invariant filters of CNN's and modeling capability of short- and long-term temporal dependencies in RNN's are gathered in a single classifier. CRNN performance is slightly better compared with CNN and RNN for gunshots, baby crying, thunder, birds singing, and a mix of sound events. However, it is hard to make any generalizations on the acoustic characteristics of these events that can explain superior performance. Another work in Su et al. (2017) demonstrated weakly supervised learning to identify audio events using the fully convolutional network with log-scale Mel-spectrogram features.

Recently, the authors in Ozer et al. (2018) studied the Spectrogram Image Features (SIF) for noisy environments. In this study, the highly overlapped spectrograms were converted into linear quantized images. Then, these dimensions were reduced by applying various image resizing methods. The feature learning and recognition was performed with the CNN approach. The work in Wang et al. (2018) presented an audio-visual descriptor, called the Auditory Receptive-field Binary Pattern (ARFBP). The ARFBP is constructed based on the SIF, the cepstral features, and the Human Auditory Receptive Field model. The extracted features are then fed into a proposed classifier called the Hierarchical-diving Deep Belief Network (HDDBN). The proposed HDDBN classifier is a DNN system that hierarchically learns the discriminative characteristics from physical feature representation to the abstract concept. Using the TUT-2016 sound event dataset, the proposed system achieves very few error rates in sound event detection of home and residential area scenes.

5.1.4 Other Deep Learning Based Approaches for Challenging Environments. The authors in Berger et al. (2018) presented the deep learning techniques for acoustic bird detection. Deep Convolutional Neural Networks (DCNNs), originally designed for image classification, are adapted and fine-tuned to detect the presence of birds in audio recordings. Various data augmentation techniques such as adding noise/content from random files, piecewise time and frequency stretching,

and time interval dropout, are applied to increase model performance and improve generalization to unknown recording conditions and new habitats. The proposed approach is evaluated in the Bird Audio Detection task that is part of the IEEE AASP Challenge on DCASE 2018. It provides the best system for the task and surpasses previous state-of-the-art approaches.

In JiaKai (2018), the neural network for the DCASE 2018 Challenge's large-scale weakly labeled semi-supervised sound event detection in domestic environments is presented. The proposed mean-teacher model with a context-gating Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) was used to maximize the use of unbalanced and unlabeled data together.

The authors in Inoue et al. (2018) illustrated the way to apply deep CNN with different data augmentation techniques such as shuffling and mixing two sounds in the same class for monitoring domestic activities in the DCASE 2018 Challenge. Thus, the new variations generated on both sequence and density of sound events have significantly improved performance than that of the baseline system.

Another method in Vesperini et al. (2018) for DCASE 2018 Challenge's large-scale weakly labeled semi-supervised Sound Event Detection (SED) in domestic environments was proposed to use weakly labeled training data in a semi-supervised way. Initially, an event activity detection technique is applied to convert weak labels to strong labels before training. At last, event activity probabilities of a capsule-based method and gated CNNs are fused to obtain the final SED estimation.

Another work for monitoring of domestic activities based on multi-channel acoustics proposed for multichannel acoustic scene classification was discussed in Tanabe et al. (2018). The system is a combination of blind signal processing in front-end and back-end modules based on machine learning. The modules in the front-end part perform de-reverberation, source separation, and noise reduction. The modules in the back-end part perform feature extraction, classification, and ensemble-based decision. The back-end modules employ one-dimensional CNN (1DCNN)-based architectures and VGG16-based architectures for individual front-end modules, and all the 89 probability outputs are ensembled.

5.1.5 Hybrid Model-based Approaches. The summary of the hybrid-model-based approach is shown in Table 13. In Atrey et al. (2006), a top-down event recognition approach is proposed using multiple GMM classifiers with four different audio features such as ZCR, LFCC, LPC, and LPCC. Among all the features, the LFCC feature provided a slightly better performance for the vocal sounds compared with non-vocal sounds. In Stowell et al. (2015), a hybrid approach is proposed using two-layer HMMs with Gabor Filter-bank feature (GBFB). Another work in Foggia et al. (2015) employed the spectral, temporal, and energy features to construct BoAw where a pool of SVM classifiers was used for classification.

In this section, reviews of generative-model-based, discriminative-model-based, deep-learning-based, and hybrid-model-based approaches are discussed for SER tasks.

5.2 Datasets

In machine learning, public evaluation and benchmark datasets help for studying the performance of various proposed systems. Various sound events for specific environmental audio scenes are listed in Table 14. It summarizes some of the publicly available datasets commonly used in sound event recognition and the corresponding sound event classes.

In Environmental Sound Classification (ESC)-50 dataset, 50 sound classes are grouped into 5 major categories (10 classes per category), such as animal sounds, natural soundscapes and water sounds, human (non-speech) sounds, interior/domestic sounds, and exterior/urban noises. The dataset provides exposure to a variety of sound sources, such as common sounds (laughing,

Table 14. Datasets Used in Sound Event Recognition (SER) Task

Datasets	References	Environment	Sound event classes
CICESE dataset (7 sound event classes)	(Beltrán et al. 2015; Jayalakshmi et al. 2018; Zhang et al. 2017; Mesaros et al. 2018a)	Smart home environment	Baby crying, bouncing ball, cricket, keys, tooth brushing, typing, and washing hands.
ESC-10 (10 sound event classes) datasets	(Piczak 2015b; Baelde et al. 2017; Agrawal et al. 2017; Badlani et al. 2017; Zhu et al. 2018; Medhat et al. 2017; Boddapati 2017; Park et al. 2017; Wang 2017)	Pedestrian escalator, elevator	Sneezing, dog barking, clock ticking, crying baby, crowing rooster, rain, sea waves, fire crackling, helicopter, and chainsaw.
ESC-50 (50 sound event classes) dataset	(Piczak 2015b; Aytar et al. 2016; Baelde et al. 2017; Tax et al. 2017; Zhu et al. 2018; Park et al. 2017)	City identification, pedestrian escalator	Animal sounds, interior/domestic sounds, exterior/urban noises, human (non-speech) sounds, natural soundscapes, and water sounds.
Urbansound8k dataset (10 sound event classes)	(Salamon et al. 2014; Badlani et al. 2017; Salamon and Bello 2017; Bello et al. 2018; Baum et al. 2018)	Smart city and smart-home analysis	Air-conditioner, car horn, children play, dog bark, drilling, engine idling, gunshot, jackhammer, siren, and street music.
Real World Computing Partnership (RWCP) dataset (50 sound event classes)	(McLoughlin et al. 2015; Su et al. 2017; Wang et al. 2017; Imoto 2018; Takahashi et al. 2018)	Road traffic, old-age home, and home	Coin drop, bell ring, lock, case strike, clapping, cough, dice drop, firecracker, saw sound, and real speech
IEEE/AASP-DCASE 2013 dataset (16 sound event sound event classes)	(Stowell et al. 2015; Park et al. 2017; Kuccukbay and Sert 2015; Plinge et al. 2014)	Office	Alert, clear throat, cough, door slam, drawer, keyboard, keys, knock, laughter, mouse, page turn, pen drop, phone, printer, speech, and switch
TUT-DCASE 2016 dataset (15 sound event classes)	(Mesaros et al. 2016; Badlani et al. 2017; Park et al. 2017; Wang et al. 2017)	Residential area and home environment	Object-rustling, object-snapping, cutlery, cupboard, dishes, drawer, glass jingling, object impact, people walking, washing dishes, and water tap running.
DCASE 2017 dataset (18 sound event classes)	(Mesaros et al. 2017; Salamon et al. 2017; Hershey et al. 2017; Adavanne and Virtanen 2017)	Street	Brakes squeaking, car, children, large vehicle, people speaking, and people walking.
Mivia dataset (3 sound event classes)	(Foggia et al. 2015; Brun et al. 2014; Foggia et al. 2014; Saggese et al. 2016, 2017; Strisciuglio 2018; Strisciuglio et al. 2015)	Road	Scream, glass breaking, and gunshot.
UPC-TALP dataset (14 sound event classes)	(Butko et al. 2011; Phan et al. 2016)	Meeting-room environment	Door knock, door slam, spoon cup jingle, steps, paper wrapping, chair moving, keyboard typing, key jingle, applause, phone ring, cough, door open, laugh, and unknown

mewing, barking), distinct sounds (breaking of a glass, brushing of teeth), and noisy sounds (helicopter and airplane noise). Similarly, the ESC-10 dataset is a subset of the ESC-50 dataset. The examples are equally distributed among the 10 classes with 40 examples per class. CICESE dataset consists of 392 examples from 7 environment sound classes. In this dataset, the training set was composed of 4 different subjects with 10 examples per subject for each class. The test dataset was created with the same audio source with 4 samples per class from 4 different subjects. The DCASE 2013 dataset used in Plinge et al. (2014) for event detection consists of 3 subsets called development, training, and testing datasets. The training set will contain examples of individual events for every class. The audio events for this training dataset are composed of 320 examples from 16 classes. The extension of the DCASE 2013 dataset is given in DCASE 2016 and DCASE 2017 with a large number of classes and diversity of data. The UPC-TALP dataset was recorded in a meeting room location. This dataset is multimodal (i.e., audio and video) and contains recordings of both isolated and spontaneous audio events. However, the recordings of isolated events were taken from 8 recording sessions with 6 different participants performed 10 times for each event. Totally, there are 1,418 instances of 11 event categories. The UrbanSound8K dataset is composed of 8,732 slices (examples) varying for a duration of 4s. Each slice consists of 10 different sound events. The MIVIA dataset contains highly noisy environmental sounds with events of interest superimposed at different values of the SNR (in our case, 6 different values), making the detection and classification of events very challenging tasks. The intensity of the background sound is modulated to obtain low levels of the SNR and simulate events that occur at various distances from the microphone.

In a Real World Computing Partnership (RWCP) Sound Scene dataset, a total of 50 sound classes are chosen from the real acoustic environments. In the RWCP database, every class contains 80 recordings and contains a single example sound per recording. The sounds were captured with high SNR. A total of 2,500 files are available for training and 1,500 files for testing. All evaluations apart from the multi-condition tests use classifiers that are trained with exclusively clean sounds with no pre-processing or noise removal applied.

5.3 Studies on Sound Event Recognition

5.3.1 Dataset Description. We have used isolated sound events from the UrbanSound8K dataset (Salamon et al. 2014). These sounds were recorded in different environments with different subjects and noise levels. The UrbanSound8K dataset is composed of 8,732 slices (examples) varying for a duration of 4s. Each slice consists of 10 different sound events. Slices in UrbanSound8K is arranged in 10 folds to ensure that slices from the same recording will not be used for training and testing.

5.3.2 Result Analysis. Figure 3 shows the recognition accuracies of the conventional HMM, conventional SVM, and some of the state-of-the-art approaches reported in the recent literature for UrbanSound8K dataset. In our experiments, MFCC features are used as basic features to learn representation. An HMM is built for each class with a varying number of states chosen from the following values { 3 5 6 7 }. The Gaussian mixture components were experimentally chosen from the set { 2 3 4 } for each HMM state. LibSVM (Chang and Lin 2011) was used for implementing the SVM classifier to classify the sound events (Jayalakshmi et al. 2018). Su et al. (2017) demonstrated weakly supervised learning using fully convolutional network with log-scale Mel-spectrogram features to identify audio events.

Agrawal et al. (2017) proposed a score level fusion of TEO-based Gammatone Cepstral Coefficients (TEO-GTCC) with MFCC and GTCC features. With GMM, 59.72% is achieved as the maximum classification accuracy for GTCC + TEO-GTCC feature sets. Piczak (2015a) trained CNN on manually engineered spectrogram features. Salamon and Bello (2015) explored unsupervised feature learning using the spherical k-means algorithm for classifying the extracted

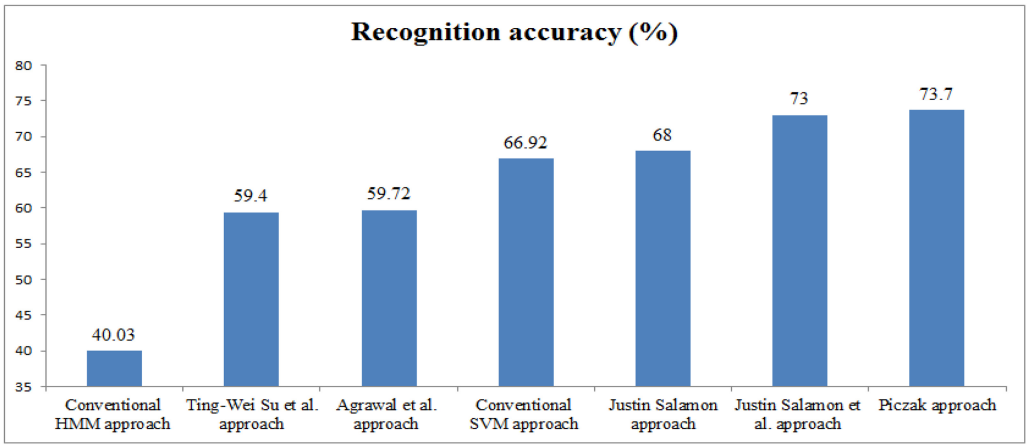

Fig. 3. Experimental studies on the Urbansound8K dataset for SER task.

log-Mel-spectrogram features from urban audio events. Salamon et al. (2014) conducted experimental studies on the challenges of UrbanSound8K dataset using the SVM classifier with statistical features derived from MFCC features as the baseline method. The conventional MFCC-SVM approach leads to approximately 10% improvement when compared with the conventional HMM-based approach.

6 FUTURE DIRECTIONS AND CONCLUSION

In this article, we have presented a detailed survey on recent developments in environmental audio surveillance tasks. First, we reviewed various categories of audio features such as feature-engineering-based approaches, feature-learning-based approaches, and feature-fusion approaches used for both environmental scene and sound event recognition tasks. Next, we presented a detailed review on various categories of modeling techniques used in environmental scene and sound event recognition tasks. The categories include generative-model-based, discriminative-model-based, deep-learning-based, and hybrid-model-based approaches. Finally, we carried out extensive studies on benchmark datasets; namely, DCASE2016 and Urbansound8K for both environmental audio scene and sound event recognition tasks. We have also presented and analyzed results of our own studies with other state-of-the-art approaches. A few important directions for further research are summarized below:

—In the case of real-time surveillance systems, there is a need to develop robust techniques to identify any kind of anomalous sound in the presence of background noise, polyphonic sounds, or multiple sound sources.
—Developing robust and compact deep representation techniques with better discrimination that can avoid overlap among different audio scenes and sound events.
—Fusion of audio and video features for effective recognition of environmental audio scenes and sound events in real-time surveillance applications.
—Developing standard universal and multi-modal benchmark datasets.
—Extension of various representation learning techniques for audio tagging and acoustic source localization tasks in audio surveillance applications.
—There is a need to explore new deep learning architectures to improve performance for real-time audio surveillance applications.

—Exploring auditory image-based features such as Gabor Filter Bank (GBFB), Histogram of Oriented Gradients (HOG), and Local Binary Pattern (LBP) with hand-crafted cepstral features will increase the performance of both Environmental Audio Scene Recognition (EASR) and Sound Event Recognition (SER) tasks.
—Rare sound event recognition with different background noises and multisource conditions can also be explored for real-time audio surveillance systems.

Deep features and deep-learning-based approaches proposed in the literature performed better than the other conventional machine-learning-based approaches. However, in case of deep features extraction and modeling, the need for huge amounts of training data and parameter and hyper-parameter tuning are issues that influence the performance of recognition in both EASR and SER tasks.

REFERENCES

Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. 2014. Convolutional neural networks for speech recognition. *IEEE/ACM Trans. Aud., Speech, Lang. Proc.* 22, 10 (2014), 1533–1545.

Shamsiah Abidin, Roberto Togneri, and Ferdous Sohel. 2018. Spectrotemporal analysis using local binary pattern variants for acoustic scene classification. *IEEE/ACM Trans. Aud., Speech, Lang. Proc.* 26, 11 (2018), 2112–2121.

Sharath Adavanne and Tuomas Virtanen. 2017. A report on sound event detection with different binaural features. Retrieved from: arXiv preprint arXiv:1710.02997.

Semih Agcaer, Anton Schlesinger, Falk-Martin Hoffmann, and Rainer Martin. 2015. Optimization of amplitude modulation features for low-resource acoustic scene classification. In *Proceedings of the 23rd European Signal Processing Conference (EUSIPCO'15)*. IEEE, 2556–2560.

Dharmesh M. Agrawal, Hardik B. Sailor, Meet H. Soni, and Hemant A. Patil. 2017. Novel TEO-based Gammatone features for environmental sound classification. In *Proceedings of the 25th European Signal Processing Conference (EUSIPCO'17)*. IEEE, 1809–1813.

Pradeep K. Atrey, Namunu C. Maddage, and Mohan S. Kankanhalli. 2006. Audio based event detection for multimedia surveillance. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'06)*, Vol. 5. IEEE.

Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2016. Soundnet: Learning sound representations from unlabeled video. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16)*. 892–900.

Rohan Badlani, Ankit Shah, Benjamin Elizalde, Anurag Kumar, and Bhiksha Raj. 2017. Framework for evaluation of sound event detection in web videos. Retrieved from: arXiv preprint arXiv:1711.00804.

Maxime Baelde, Christophe Biernacki, and Raphaël Greff. 2017. A mixture model-based real-time audio sources classification method. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'17)*. 2427–2431.

Daniele Barchiesi, Dimitrios Giannoulis, Dan Stowell, and Mark D. Plumbley. 2015. Acoustic scene classification: Classifying environments from the sounds they produce. *IEEE Sig. Proc. Mag.* 32, 3 (2015), 16–34.

Elizabeth Baum, Mario Harper, Ryan Alicea, and Camilo Ordonez. 2018. Sound identification for fire-fighting mobile robots. In *Proceedings of the 2nd IEEE International Conference on Robotic Computing (IRC'18)*. IEEE, 79–86.

Juan Pablo Bello, Charlie Mydlarz, and Justin Salamon. 2018. Sound analysis in smart cities. In *Computational Analysis of Sound Scenes and Events*. Springer, 373–397.

Jessica Beltrán, Edgar Chávez, and Jesús Favela. 2015. Scalable identification of mixed environmental sounds, recorded from heterogeneous sources. *Pattern Recog. Lett.* 68 (2015), 153–160.

Amira Ben Mabrouk and Ezzeddine Zagrouba. 2018. Abnormal behavior recognition for intelligent video surveillance systems. *Expert Syst. Applic.: Int. J.* 91, C (2018), 480–491.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Machine Intell.* 35, 8 (2013), 1798–1828.

Franz Berger, William Freillinger, Paul Primus, and Wolfgang Reisinger. 2018. *Bird Audio Detection—DCASE 2018*. Technical Report. DCASE2018 Challenge.

Victor Bisot, Slim Essid, and Gaël Richard. 2015. Hog and subband power distribution image features for acoustic scene classification. In *Proceedings of the 23rd European Signal Processing Conference (EUSIPCO'15)*. IEEE, 719–723.

V. Boddapati, A. Petef, J. Rasmusson, and L. Lundberg. 2017. Classifying environmental sounds using image recognition networks. *Procedia Computer Science* 112 (2017), 2048–2056.

Luc Brun, Gennaro Percannella, Alessia Saggese, and Mario Vento. 2014. HAck: A system for the recognition of human actions by kernels of visual strings. In *Proceedings of the 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS'14)*. IEEE, 142–147.

Taras Butko, Cristian Canton-Ferrer, Carlos Segura, Xavier Giró, Climent Nadeu, Javier Hernando, and Josep R. Casas. 2011. Acoustic event detection based on feature-level fusion of audio and video modalities. *EURASIP J. Adv. Sig. Proc.* 2011, 1 (2011), 485738.

Emre Cakir, Giambattista Parascandolo, Toni Heittola, Heikki Huttunen, Tuomas Virtanen, Emre Cakir, Giambattista Parascandolo, Toni Heittola, Heikki Huttunen, and Tuomas Virtanen. 2017. Convolutional recurrent neural networks for polyphonic sound event detection. *IEEE/ACM Trans. Aud., Speech, Lang. Proc.* 25, 6 (2017), 1291–1303.

Sachin Chachada and C.-C. Jay Kuo. 2013. Environmental sound recognition: A survey. In *Proceedings of the Asia-Pacific Signal and Information Processing Association Summit and Conference (APSIPA'13)*. 1–9.

Sachin Chachada and C.-C. Jay Kuo. 2014. Environmental sound recognition: A survey. *APSIPA Trans. Sig. Inform. Proc.* 3 (2014).

S. Chandrakala and C. Chandra Sekhar. 2010. Classification of varying length multivariate time series using Gaussian mixture models and support vector machines. *Int. J. Data Mining, Modell. Manag.* 2, 3 (2010), 268–287.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2, 3 (2011), 27.

Michael Cheffena. 2016. Fall detection using smartphone audio features. *IEEE J. Biomed. Health Inform.* 20, 4 (2016), 1073–1080.

Michele Lai Chin and Juan José Burred. 2012. Audio event detection based on layered symbolic sequence representations. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'12)*. IEEE, 1953–1956.

Khin Myo Chit and K. Zin Lin. 2013. Audio-based action scene classification using HMM-SVM algorithm. *Int. J. Adv. Res. Comput. Eng. Technol.* 2, 4 (2013), 1347–1351.

Selina Chu, Shrikanth Narayanan, and C.-C. Jay Kuo. 2009. Environmental sound recognition with time–frequency audio features. *IEEE Trans. Aud., Speech, Lang. Proc.* 17, 6 (2009), 1142–1158.

Michael Cowling and Renate Sitte. 2003. Comparison of techniques for environmental sound recognition. *Pattern Recog. Lett.* 24, 15 (2003), 2895–2907.

Marco Cristani, Manuele Bicego, and Vittorio Murino. 2007. Audio-visual event recognition in surveillance video sequences. *IEEE Trans. Multimed.* 9, 2 (2007), 257–267.

Marco Crocco, Marco Cristani, Andrea Trucco, and Vittorio Murino. 2016. Audio surveillance: A systematic review. *ACM Comput. Surv.* 48, 4 (2016), 52.

Waltenegus Dargie. 2009. Adaptive audio-based context recognition. *IEEE Trans. Syst., Man, Cyber.-Part A: Syst. Hum.* 39, 4 (2009), 715–725.

Najim Dehak, Patrick J. Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. 2011. Front-end factor analysis for speaker verification. *IEEE Trans. Aud., Speech, Lang. Proc.* 19, 4 (2011), 788–798.

P. Dhanalakshmi, S. Palanivel, and Vennila Ramalingam. 2011. Classification of audio signals using AANN and GMM. *Appl. Soft Comput.* 11, 1 (2011), 716–723.

Matthias Dorfer, Bernhard Lehner, Hamid Eghbal-zadeh, Heindl Christop, Paischer Fabian, and Widmer Gerhard. 2018. *Acoustic Scene Classification with Fully Convolutional Neural Networks and I-vectors*. Technical Report. DCASE2018 Challenge.

Alain Dufaux, Laurent Besacier, Michael Ansorge, and Fausto Pellandini. 2000. Automatic sound detection and recognition for noisy environment. In *Proceedings of the 10th European Signal Processing Conference (EUSIPCO'00)*. 1–4.

Hamid Eghbal-zadeh, Bernhard Lehner, Matthias Dorfer, and Gerhard Widmer. 2016. CP-JKU submissions for DCASE-2016: A hybrid approach using binaural i-vectors and deep convolutional neural networks. In *Proceedings of the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE'16)*.

Hamid Eghbal-zadeh, Bernhard Lehner, Matthias Dorfer, and Gerhard Widmer. 2017. A hybrid approach with multi-channel I-vectors and convolutional neural networks for acoustic scene classification. Retrieved from: arXiv preprint arXiv:1706.06525.

Moataz El Ayadi, Mohamed S. Kamel, and Fakhri Karray. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Patt. Recog.* 44, 3 (2011), 572–587.

Khaled El-Maleh, Ara Samouelian, and Peter Kabal. 1999. Frame level noise classification in mobile environments. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1. 237–240.

Benjamin Elizalde, Guan-Lin Chao, Ming Zeng, and Ian Lane. 2016a. City-identification of Flickr videos using semantic acoustic features. In *Proceedings of the 2nd IEEE International Conference on Multimedia Big Data (BigMM'16)*. IEEE, 303–306.

Benjamin Elizalde, Anurag Kumar, Ankit Shah, Rohan Badlani, Emmanuel Vincent, Bhiksha Raj, and Ian Lane. 2016b. Experiments on the DCASE challenge 2016: Acoustic scene classification and sound event detection in real life recording. Retrieved from: arXiv preprint arXiv:1607.06706.

Antti J. Eronen, Vesa T. Peltonen, Juha T. Tuomi, Anssi P. Klapuri, Seppo Fagerlund, Timo Sorsa, Gaëtan Lorho, and Jyri Huopaniemi. 2006. Audio-based context recognition. *IEEE Trans. Aud., Speech, Lang. Proc.* 14, 1 (2006), 321–329.

Georgios Evangelopoulos, Athanasia Zlatintsi, Alexandros Potamianos, Petros Maragos, Konstantinos Rapantzikos, Georgios Skoumas, and Yannis Avrithis. 2013. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Trans. Multimed.* 15, 7 (2013), 1553–1568.

Pasquale Foggia, Nicolai Petkov, Alessia Saggese, Nicola Strisciuglio, and Mario Vento. 2015. Reliable detection of audio events in highly noisy environments. *Pattern Recog. Lett.* 65 (2015), 22–28.

Pasquale Foggia, Nicolai Petkov, Alessia Saggese, Nicola Strisciuglio, and Mario Vento. 2016. Audio surveillance of roads: A system for detecting anomalous sounds. *IEEE Trans. Intell. Transport. Syst.* 17, 1 (2016), 279–288.

Pasquale Foggia, Alessia Saggese, Nicola Strisciuglio, and Mario Vento. 2014. Exploiting the deep learning paradigm for recognizing human actions. In *Proceedings of the 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS'14)*. IEEE, 93–98.

Jurgen T. Geiger, Bjorn Schuller, and Gerhard Rigoll. 2013. Large-scale audio feature extraction and SVM for acoustic scene classification. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'13)*. IEEE, 1–4.

Jort F. Gemmeke, Lode Vuegen, Peter Karsmakers, Bart Vanrumste, et al. 2013. An exemplar-based NMF approach to audio event detection. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'13)*. IEEE, 1–4.

Luigi Gerosa, Giuseppe Valenzise, Marco Tagliasacchi, Fabio Antonacci, and Augusto Sarti. 2007. Scream and gunshot detection in noisy environments. In *Proceedings of the 15th European Signal Processing Conference (EUSIPCO'07)*. 1216–1220.

Ben Gold, Nelson Morgan, and Dan Ellis. 2011. *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*. John Wiley & Sons.

Elsa Ferreira Gomes, Fábio Batista, and Alípio M. Jorge. 2016. Using smartphones to classify urban sounds. In *Proceedings of the 9th International C* Conference on Computer Science & Software Engineering*. ACM, 67–72.

René Grzeszick, Axel Plinge, and Gernot A. Fink. 2017. Bag-of-features methods for acoustic event detection and classification. *IEEE/ACM Trans. Aud., Speech, Lang. Proc.* 25, 6 (2017), 1242–1252.

Yoonchang Han and Kyogu Lee. 2016. Convolutional neural network with multiple-width frequency-delta data augmentation for acoustic scene classification. In *Proceedings of the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events*.

Toni Heittola, Annamaria Mesaros, Antti Eronen, and Tuomas Virtanen. 2013. Context-dependent sound event detection. *EURASIP J. Aud., Speech, Music Proc.* 2013, 1 (2013), 1–13.

Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, et al. 2017. CNN architectures for large-scale audio classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'17)*. IEEE, 131–135.

Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, et al. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Sig. Proc. Mag.* 29, 6 (2012), 82–97.

Keisuke Imoto. 2018. Introduction to acoustic event and scene analysis. *Acoustic. Sci. Technol.* 39, 3 (2018), 182–188.

Tadanobu Inoue, Phongtharin Vinayavekhin, Shiqiang Wang, David Wood, Nancy Greco, and Ryuki Tachibana. 2018. *Domestic Activities Classification Based on CNN Using Shuffling and Mixing Data Augmentation*. Technical Report. DCASE2018 Challenge.

Aun Irtaza, Syed M. Adnan, Sumair Aziz, Ali Javed, M. Obaid Ullah, and Muhammad Tariq Mahmood. 2017. A framework for fall detection of elderly people by analyzing environmental sounds through acoustic local ternary patterns. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC'17)*. IEEE, 1558–1563.

S. L. Jayalakshmi, S. Chandrakala, and R. Nedunchelian. 2018. Global statistical features-based approach for acoustic event detection. *Appl. Acoust.* 139 (2018), 113–118.

Bhagyalaxmi Jena and Sudhansu Sekhar Singh. 2018. Analysis of stressed speech on Teager energy operator (TEO). *International Journal of Pure and Applied Mathematics* 118, 16 (2018), 667–680.

Lu JiaKai. 2018. *Mean Teacher Convolution System for DCASE 2018 Task 4*. Technical Report. DCASE2018 Challenge.

Liping Jing, Bo Liu, Jaeyoung Choi, Adam Janin, Julia Bernd, Michael W. Mahoney, and Gerald Friedland. 2017. DCAR: A discriminative and compact audio representation for audio processing. *IEEE Trans. Multimed.* 19, 12 (2017), 2637–2650.

M. Karbasi, S. M. Ahadi, and M. Bahmanian. 2011. Environmental sound classification using spectral dynamic features. In *Proceedings of the 8th International Conference on Information, Communications and Signal Processing (ICICS'11)*. 1–5.

Selver Ezgi Kuccukbay and Mustafa Sert. 2015. Audio-based event detection in office live environments using optimized MFCC-SVM approach. In *Proceedings of the IEEE International Conference on Semantic Computing (ICSC'15)*. 475–480.

Younghyun Lee, David K. Han, and Hanseok Ko. 2013. Acoustic signal based abnormal event detection in indoor environment using multiclass adaboost. *IEEE Trans. Consum. Electron.* 59, 3 (2013), 615–622.

David Li, Jason Tam, and Derek Toub. 2013. Auditory scene classification using machine learning techniques. In *Proceedings of the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events*.

Qi Li, Huadong Ma, and Dong Zhao. 2009. A neural network based framework for audio scene analysis in audio sensor networks. In *Proceedings of the Pacific-Rim Conference on Multimedia*. Springer, 480–490.

Thomas Lidy and Alexander Schindler. 2016. CQT-based convolutional neural networks for audio scene classification. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE'16)*, Vol. 90. DCASE2016 Challenge, 1032–1048.

Héctor Lozano, Inmaculada Hernáez, Artzai Picón, Javier Camarena, and Eva Navas. 2010. Audio classification techniques in home environments for elderly/dependant people. In *Proceedings of the International Conference on Computers for Handicapped Persons*. Springer, 320–323.

Lie Lu, Stan Z. Li, and Hong-Jiang Zhang. 2001. Content-based audio segmentation using support vector machines. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'01)*, Vol. 1. 749–752.

Xugang Lu, Yu Tsao, Shigeki Matsuda, and Chiori Hori. 2014. Sparse representation based on a bag of spectral exemplars for acoustic event detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'14)*. IEEE, 6255–6259.

Tao Lv, He-yong Zhang, and Chun-hui Yan. 2018. Double mode surveillance system based on remote audio/video signals acquisition. *Appl. Acoust.* 129 (2018), 316–321.

Ling Ma, Ben Milner, and Dan Smith. 2006. Acoustic environment classification. *ACM Trans. Speech, Lang. Proc.* 3, 2 (2006), 1–22.

Ling Ma, D. J. Smith, and Ben P. Milner. 2003. Context awareness using environmental noise classification. In *Proceedings of the 8th European Conference on Speech Communication and Technology*.

Gustavo Mafra, Ngoc Duong, Alexey Ozerov, and Patrick Pérez. 2016. Acoustic scene classification: An evaluation of an extremely compact feature representation. In *Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE'16)*.

Hafiz Malik. 2013. Acoustic environment identification and its applications to audio forensics. *IEEE Trans. Inform. Forens. Secur.* 8, 11 (2013), 1827–1837.

Erik Marchi, Dario Tonelli, Xinzhou Xu, Fabien Ringeval, Jun Deng, Stefano Squartini, and Björn Schuller. 2016. Pairwise decomposition with deep neural networks and multiscale kernel subspace learning for acoustic scene classification. In *Proceedings of the 24th Acoustic Scene Classification Workshop at the European Signal Processing Conference (EUSIPCO'16)*.

Ian McLoughlin, Haomin Zhang, Zhipeng Xie, Yan Song, and Wei Xiao. 2015. Robust sound event classification using deep neural networks. *IEEE/ACM Trans. Aud., Speech, Lang. Proc.* 23, 3 (2015), 540–552.

Fady Medhat, David Chesmore, and John Robinson. 2017. Recognition of acoustic events using masked conditional neural networks. In *Proceedings of the 16th IEEE International Conference on Machine Learning and Applications (ICMLA'17)*. IEEE, 199–206.

Annamaria Mesaros, Toni Heittola, Emmanouil Benetos, Peter Foster, Mathieu Lagrange, Tuomas Virtanen, and Mark D. Plumbley. 2018b. Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge. *IEEE/ACM Trans. Aud., Speech, Lang. Proc.* 26, 2 (2018), 379–393.

Annamaria Mesaros, Toni Heittola, Onur Dikmen, and Tuomas Virtanen. 2015. Sound event detection in real life recordings using coupled matrix factorization of spectral representations and class activity annotations. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'15)*. 151–155.

Annamaria Mesaros, Toni Heittola, Aleksandr Diment, Benjamin Elizalde, Ankit Shah, Emmanuel Vincent, Bhiksha Raj, and Tuomas Virtanen. 2017. DCASE 2017 challenge setup: Tasks, datasets and baseline system. In *Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE'17)*.

Annamaria Mesaros, Toni Heittola, and Dan Ellis. 2018a. Datasets and evaluation. In *Computational Analysis of Sound Scenes and Events*. Springer, 147–179.

Annamaria Mesaros, Toni Heittola, Antti Eronen, and Tuomas Virtanen. 2010. Acoustic event detection in real life recordings. In *Proceedings of the 18th European Signal Processing Conference (EUSIPCO'10)*. 1267–1271.

Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. 2016. TUT database for acoustic scene classification and sound event detection. In *Proceedings of the 24th European Signal Processing Conference (EUSIPCO'16)*. IEEE, 1128–1132.

Thomas B. Moeslund, Omar Javed, Yu-Gang Jiang, and R. Manmatha. 2014. Special issue on multimedia event detection. *Machine Vision & Applications* 25, 1 (2014), 1–4.

Seongkyu Mun, Sangwook Park, David K. Han, and Hanseok Ko. 2017. *Generative Adversarial Network Based Acoustic Scene Training Set Augmentation and Selection Using SVM Hyper-plane*. Technical Report. DCASE2017 Challenge.

Terence Wen Zheng Ng. 2014. *Sound Event Recognition in Home Environments*. Ph.D. Dissertation. Nanyang Technological University, Singapore.

Stavros Ntalampiras, Ilyas Potamitis, and Nikos Fakotakis. 2009. An adaptive framework for acoustic monitoring of potential hazards. *EURASIP J. Aud., Speech, Music Proc.* 2009, 1 (2009), 594103.

Stavros Ntalampiras, Ilyas Potamitis, and Nikos Fakotakis. 2011. Probabilistic novelty detection for acoustic surveillance under real-world conditions. *IEEE Trans. Multimed.* 13, 4 (2011), 713–719.

Ilyas Ozer, Zeynep Ozer, and Oguz Findik. 2018. Noise robust sound event classification with convolutional neural network. *Neurocomputing* 272 (2018), 505–512.

Tae Hong Park, Minjoon Yoo, Chris Dye, Jaeseong You, Varatep Buranintu, Dima Rekesh, and Isaac Leonard. 2017. Urban soundmapping at the edge. In *Proceedings of the INTER-NOISE and NOISE-CON Congress and Conference*, Vol. 255. Institute of Noise Control Engineering, 5389–5400.

Yohan Petetin, Cyrille Laroche, and Aurélien Mayoue. 2015. Deep neural networks for audio scene recognition. In *Proceedings of the 23rd European Signal Processing Conference (EUSIPCO'15)*. 125–129.

Huy Phan, Lars Hertel, Marco Maass, Radoslaw Mazur, and Alfred Mertins. 2016. Learning representations for nonspeech audio events through their similarities to speech patterns. *IEEE/ACM Trans. Aud., Speech, Lang. Proc.* 24, 4 (2016), 807–822.

Huy Phan, Philipp Koch, Lars Hertel, Marco Maass, Radoslaw Mazur, and Alfred Mertins. 2017. CNN-LTE: A class of 1-X pooling convolutional neural networks on label tree embeddings for audio scene classification. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'17)*.

Huy Phan, Marco Maaß, Radoslaw Mazur, and Alfred Mertins. 2015. Random regression forests for acoustic event detection and classification. *IEEE/ACM Trans. Aud., Speech, Lang. Proc.* 23, 1 (2015), 20–31.

Karol J. Piczak. 2015a. Environmental sound classification with convolutional neural networks. In *Proceedings of the 25th IEEE International Workshop on Machine Learning for Signal Processing (MLSP'15)*. 1–6.

Karol J. Piczak. 2015b. ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*. ACM, 1015–1018.

Axel Plinge, Rene Grzeszick, and Gernot A. Fink. 2014. A bag-of-features approach to acoustic event detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'14)*. 3704–3708.

Asma Rabaoui, Manuel Davy, Stéphane Rossignol, and Noureddine Ellouze. 2008. Using one-class SVMs and wavelets for audio surveillance. *IEEE Trans. Inform. Forens. Sec.* 3, 4 (2008), 763–775.

Asma Rabaoui, Zied Lachiri, and Noureddine Ellouze. 2009. Using HMM-based classifier adapted to background noises with improved sounds features for audio surveillance application. *Int. J. Signal Process* 3 (2009), 535–545.

Regunathan Radhakrishnan, Ajay Divakaran, and A. Smaragdis. 2005. Audio analysis for surveillance applications. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. 158–161.

Alain Rakotomamonjy and Gilles Gasso. 2015. Histogram of gradients of time-frequency representations for audio scene classification. *IEEE/ACM Trans. Aud., Speech, Lang. Proc.* 23, 1 (2015), 142–153.

Jianfeng Ren, Xudong Jiang, Junsong Yuan, and Nadia Magnenat-Thalmann. 2017a. Sound-event classification using robust texture features for robot hearing. *IEEE Trans. Multimed.* 19, 3 (2017), 447–458.

Zhao Ren, Vedhas Pandit, Kun Qian, Zijiang Yang, Zixing Zhang, and Björn Schuller. 2017b. Deep sequential image features on acoustic scene classification. In *Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE'17)*.

Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. 2000. Speaker verification using adapted Gaussian mixture models. *Dig. Sig. Proc.* 10, 1 (2000), 19–41.

Gerard Roma, Waldo Nogueira, Perfecto Herrera, and Roc de Boronat. 2013. Recurrence quantification analysis features for auditory scene classification. In *Proceedings of the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events* 2 (2013).

J.-L. Rouas, Jérôme Louradour, and Sébastien Ambellouis. 2006. Audio events detection in public transport vehicle. In *Proceedings of the IEEE Intelligent Transportation Systems Conference (ITSC'06)*. IEEE, 733–738.

Alessia Saggese, Nicola Strisciuglio, Mario Vento, and Nicolai Petkov. 2016. Time-frequency analysis for audio event detection in real scenarios. In *Proceedings of the 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS'16)*. IEEE, 438–443.

Alessia Saggese, Nicola Strisciuglio, Mario Vento, and Nicolai Petkov. 2017. A real-time system for audio source localization with cheap sensor device. In *Proceedings of the 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS'17)*. IEEE, 1–7.

Justin Salamon and Juan Pablo Bello. 2015. Unsupervised feature learning for urban sound classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'15)*. 171–175.

Justin Salamon and Juan Pablo Bello. 2017. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Sig. Proc. Lett.* 24, 3 (2017), 279–283.

Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. 2014. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM International Conference on Multimedia*. ACM, 1041–1044.

Justin Salamon, Brian McFee, Peter Li, and Juan Pablo Bello. 2017. *DCASE 2017 Submission: Multiple Instance Learning for Sound Event Detection*. Technical Report. DCASE2017 Challenge.

Nitin Sawhney and Pattie Maes. 1997. *Situational Awareness from Environmental Sounds*. Project Report. MIT, Cambridge, MA.

Jens Schroder, Niko Moritz, Marc Rene Schadler, Benjamin Cauchi, Kamil Adiloglu, Jorn Anemuller, Simon Doclo, Birger Kollmeier, and Stefan Goetze. 2013. On the use of spectro-temporal features for the IEEE AASP Challenge on "Detection and Classification of Acoustic Scenes and Events." In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'13)*. IEEE, 1–4.

Zhao Shuyang, Toni Heittola, and Tuomas Virtanen. 2017. Active learning for sound event classification by clustering unlabeled data. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'17)*. 751–755.

Dan Stowell and David Clayton. 2015. Acoustic event detection for multiple overlapping similar sources. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'15)*. 1–5.

Dan Stowell, Dimitrios Giannoulis, Emmanouil Benetos, Mathieu Lagrange, and Mark D. Plumbley. 2015. Detection and classification of acoustic scenes and events. *IEEE Trans. Multimed.* 17, 10 (2015), 1733–1746.

Nicola Strisciuglio. 2018. Learning audio and image representations with bio-inspired trainable feature extractors. Retrieved from: arXiv preprint arXiv:1801.00688.

Nicola Strisciuglio, Mario Vento, and Nicolai Petkov. 2015. Bio-inspired filters for audio analysis. In *Proceedings of the International Workshop on Brain-Inspired Computing*. Springer, 101–115.

Ting-Wei Su, Jen-Yu Liu, and Yi-Hsuan Yang. 2017. Weakly supervised audio event detection using event-specific Gaussian filters and fully convolutional networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'17)*. IEEE, 791–795.

Gen Takahashi, Takeshi Yamada, Shoji Makino, and Nobutaka Ono. 2016. Acoustic scene classification using deep neural network and frame-concatenated acoustic feature. In *Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE'16)*.

Naoya Takahashi, Michael Gygli, and Luc Van Gool. 2018. Aenet: Learning deep audio features for video analysis. *IEEE Trans. Multimed.* 20, 3 (2018), 513–524.

Ryo Tanabe, Takashi Endo, Yuki Nikaido, Takeshi Ichige, Phong Nguyen, Yohei Kawaguchi, and Koichi Hamada. 2018. *Multichannel Acoustic Scene Classification by Blind Dereverberation, Blind Source Separation, Data Augmentation, and Model Ensembling*. Technical Report. DCASE2018 Challenge.

Tycho Max Sylvester Tax, Jose Luis Diez Antich, Hendrik Purwins, and Lars Maaløe. 2017. Utilizing domain knowledge in end-to-end audio processing. Retrieved from: arXiv preprint arXiv:1712.00254.

Andrey Temko, Climent Nadeu, Dušan Macho, Robert Malkin, Christian Zieger, and Maurizio Omologo. 2009. Acoustic event detection and classification. *Computers in the Human Interaction Loop*. Springer, 61–73.

EnShuo Tsau, Seung-Hwan Kim, and C.-C. Jay Kuo. 2011. Environmental sound recognition with CELP-based features. In *Proceedings of the 10th International Symposium on Signals, Circuits and Systems (ISSCS'11)*. IEEE, 1–4.

Michele Valenti, Aleksandr Diment, Giambattista Parascandolo, Stefano Squartini, and Tuomas Virtanen. 2016. DCASE 2016 acoustic scene classification using convolutional neural networks. In *Proceedings of the Workshop on Detection Classification of Acoustic Scenes and Events*. 95–99.

Giuseppe Valenzise, Luigi Gerosa, Marco Tagliasacchi, Fabio Antonacci, and Augusto Sarti. 2007. Scream and gunshot detection and localization for audio-surveillance systems. In *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS'17)*. 21–26.

Xavier Valero and Francesc Alias. 2012. Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification. *IEEE Trans. Multimed.* 14, 6 (2012), 1684–1689.

Vladimir Vapnik. 1998. *Statistical Learning Theory,* Vol. 3. Wiley, New York.

Fabio Vesperini, Leonardo Gabrielli, Emanuele Principi, and Stefano Squartini. 2018. *A Capsule Neural Networks Based Approach for Bird Audio Detection*. Technical Report, DCASE2018 Challenge.

Lode Vuegen, B. V. D. Broeck, Peter Karsmakers, J. F. Gemmeke, Bart Vanrumste, and H. V. Hamme. 2013. An MFCC-GMM approach for event detection and classification. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'13)*. 1–3.

Chien-Yao Wang, Jia-Ching Wang, Andri Santoso, Chin-Chin Chiang, and Chung-Hsien Wu. 2017. Sound event recognition using auditory-receptive-field binary pattern and hierarchical-diving deep belief network. *IEEE/ACM Trans. Aud., Speech, Lang. Proc.* 26, 8 (2017), 1336–1351.

Chien-Yao Wang, Jia-Ching Wang, Andri Santoso, Chin-Chin Chiang, and Chung-Hsien Wu. 2018. Sound event recognition using auditory-receptive-field binary pattern and hierarchical-diving deep belief network. *IEEE/ACM Trans. Aud., Speech, Lang. Proc.* 26, 8 (2018), 1336–1351.

DeLiang Wang and Guy J. Brown. 2006. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press.

Yun Wang. 2017. *Polyphonic Sound Event Detection with Weak Labeling*. Ph.D. Dissertation. Google Inc.

Yong Xu, Qiang Huang, Wenwu Wang, Peter Foster, Siddharth Sigtia, Philip J. B. Jackson, and Mark D. Plumbley. 2017a. Unsupervised feature learning based on deep models for environmental audio tagging. *IEEE/ACM Trans. Aud., Speech, Lang. Proc.* 25, 6 (2017), 1230–1241.

Yong Xu, Qiuqiang Kong, Qiang Huang, Wenwu Wang, and Mark D. Plumbley. 2017b. Attention and localization based on a deep convolutional recurrent model for weakly supervised audio tagging. Retrieved from: arXiv preprint arXiv:1703.06052.

Wenjun Yang, Sridhar Krishnan, Wenjun Yang, and Sridhar Krishnan. 2017. Combining temporal features by local binary pattern for acoustic scene classification. *IEEE/ACM Trans. Aud., Speech, Lang. Proc.* 25, 6 (2017), 1315–1321.

Jiaxing Ye, Takumi Kobayashi, Xiaoyan Wang, Hiroshi Tsuda, and Murakawa Masahiro. 2019. Audio data mining for anthropogenic disaster identification: An automatic taxonomy approach. *IEEE Trans. Emerg. Topics Comput.* (In Press). DOI: 10.1109/TETC.2017.2700843

Sungrack Yun, Sungwoong Kim, Sunkuk Moon, Juncheol Cho, and Taesu Kim. 2016. Discriminative training of GMM parameters for audio scene classification and audio tagging. In *Proceedings of the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE'16)*.

Xiaohu Zhang, Yuexian Zou, and Wei Shi. 2017. Dilated convolution neural network with LeakyReLU for environmental sound classification. In *Proceedings of the 22nd International Conference on Digital Signal Processing (DSP'17)*. IEEE, 1–5.

Boqing Zhu, Changjian Wang, Feng Liu, Jin Lei, Zengquan Lu, and Yuxing Peng. 2018. Learning environmental sounds with multi-scale convolutional neural network. Retrieved from: arXiv preprint arXiv:1803.10219.

Christian Zieger, Alessio Brutti, and Piergiorgio Svaizer. 2009. Acoustic based surveillance system for intrusion detection. In *Proceedings of the 6th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS'09)*. 314–319.

Christian Zieger and Maurizio Omologo. 2008. Acoustic event classification using a distributed microphone network with a GMM/SVM combined algorithm. In *Proceedings of the 9th Conference of the International Speech Communication Association*.

Received August 2018; revised January 2019; accepted March 2019