Chapter 13
Sound Analysis in Smart Cities

Juan Pablo Bello, Charlie Mydlarz, and Justin Salamon

Abstract This chapter introduces the concept of smart cities and discusses the importance of sound as a source of information about urban life. It describes a wide range of applications for the computational analysis of urban sounds and focuses on two high-impact areas, audio surveillance, and noise pollution monitoring, which sit at the intersection of dense sensor networks and machine listening. For sensor networks we focus on the pros and cons of mobile versus static sensing strategies, and the description of a low-cost solution to acoustic sensing that supports distributed machine listening. For sound event detection and classification we focus on the challenges presented by this task, solutions including feature design and learning strategies, and how a combination of convolutional networks and data augmentation result in the current state of the art. We close with a discussion about the potential and challenges of mobile sensing, the limitations imposed by the data currently available for research, and a few areas for future exploration.

Keywords Urban sound • Smart cities • Noise monitoring • Sensor network • Acoustic sensing • Internet of things (IOT) • MEMS microphone • Audio surveillance • Sound event detection • Sound classification • Machine listening • Machine learning • Deep learning • Convolutional neural networks • Data augmentation

J.P. Bello (✉)
Music and Audio Research Laboratory, New York University,
Suite 1077 35 W, 4th Street, New York, NY 10012, USA
e-mail: jpbello@nyu.edu

C. Mydlarz • J. Salamon
Center for Urban Science and Progress & Music and Audio Research Laboratory,
New York University, New York, NY, USA
e-mail: cmydlarz@nyu.edu; justin.salamon@nyu.edu

T. Virtanen et al. (eds.), *Computational Analysis of Sound Scenes and Events*,
DOI 10.1007/978-3-319-63450-0_13

13.1 Introduction

13.1.1 Smart Cities

Current estimates put the share of the world population living in urban environments at 50%, a number that is expected to grow to as much as 80% by 2050. In OECD member countries, for example, including most of Europe and North America, already 80% of the population lives in cities, with China seeing a net increase of 40% in its share of urban inhabitants during the last 50 years.[1] This rapid trend of urbanization creates massive opportunities for economic development, job diversification, and innovation, but also creates significant problems related to the environmental impact of human activity, the stress to systems and infrastructure, the difficulty of effectively policing and securing public spaces, and potential reductions in health and quality of living for city dwellers.

Unsurprisingly, there is a well-established and growing trend of leveraging technological systems and solutions towards addressing some of the most pressing issues facing urban communities. These *smart cities* initiatives benefit from recent advances in ubiquitous and intelligent sensing, widespread connectivity, and data science to collect, distribute and analyze the data needed to understand the situation on the ground, anticipate future behavior and drive effective action.

13.1.2 Urban Sound Sensing and Analysis

The term urban soundscape refers to the sound scenes and sound events commonly perceived in cities. While the specific characteristics of urban soundscapes vary between cities and even neighborhoods, they still share certain qualities that set them apart from other soundscapes. Perhaps most importantly, while rural soundscapes primarily contain geophony (naturally occurring non-biological sounds, such as the sound of wind or rain) and biophony (naturally occurring biological sounds, i.e., non-human animal sounds), urban soundscapes are dominated by anthrophony (sounds produced by humans), which consists not only of the human voice, but of all sounds generated by human-made artifacts including the sounds emitted by traffic, construction, signals, machines, musical instruments, and so on.

Sound is an important source of information about urban life, with great potential for smart city applications. The increase in smart phone penetration and the growing development of specialized acoustic sensor networks mean that urban sound monitoring is becoming an increasingly appealing alternative, or complement, to video cameras and other forms of environmental sensing. Microphones are generally smaller and less expensive than cameras and are robust to environmental conditions

[1] http://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS.

such as fog, pollution, rain, and daily changes in light conditions that negatively affect visibility. They are also less susceptible to occlusion and are capable of omni-directional sensing.

The automatic capture, analysis, and characterization of urban soundscapes can facilitate a wide range of novel applications including noise pollution mitigation, context-aware computing, and surveillance. The automatic analysis of urban soundscapes is also a first step towards studying their influence on and/or interaction with other quantifiable aspects of city-life, including public health, real estate, crime, and education.

However, there are also important challenges in urban sound monitoring. Urban environments are among the most acoustically rich sonic environments we could study—the number of possible sounds is unlimited and densely mixed. Furthermore, the production mechanisms and resulting acoustic characteristics of urban sounds are highly heterogeneous, ranging from impulse-like sounds such as gun shots to droning motors that run non-stop, from noise-like sources like air-conditioning units to harmonic sounds like voice. They include human, animal, natural, mechanical, and electric sources, spanning the entire spectrum of frequencies and temporal dynamics.

Furthermore, the complex interaction between this multiplicity of sources and the built environment, which is often dense, intricate and highly reflective, creates varying levels of "rumble" in the background. Therefore, it is not unusual for the sources of interest to overlap with other sounds and to present low signal-to-noise ratios (SNR) that change intermittently over time, tremendously complicating the analysis and understanding of these acoustic scenes.

Importantly, while some audio analysis tasks have a relatively clear delineation between what should be considered a "source of interest" and what should be considered "background" or "noise" (e.g., specific instruments versus accompaniment in music, or individual speakers against the background in speech), this distinction is far less clear in the case of urban soundscapes. Almost any sound source can be a source of interest, and many "noise-like" sources such as idling engines or HVAC units can have similar acoustic properties even though their type and function are very different.

Finally, urban soundscapes are not composed following top-down rules or hierarchical structures that can be exploited as in the case of speech and most music. However, natural patterns of activity resulting from our circadian, weekly, monthly, and yearly rhythms and cultural cycles abound.

13.1.3 Overview of this Chapter

The rest of this chapter is organized as follows. Section 13.2 will briefly discuss the range of applications of automatic sound event analysis and dense sensor networks in urban environments, with a focus on audio surveillance and noise pollution monitoring. Section 13.3 discusses existing solutions for large-scale urban acoustic

sensing and presents the design of a low-cost, scalable, and accurate acoustic sensor network. Section 13.4 provides an in-depth view of the problem of urban sound source identification, and the lessons learned from research efforts to date. Finally, Sect. 13.5 provides a summary of the chapter and some perspectives on future work in this field.

13.2 Smart City Applications

The intelligent and automated analysis of urban soundscapes has a number of valuable applications. For example, it can be used to enhance context-aware computing, particularly for robotic navigation in changing urban environments including for autonomous vehicles (private, public transport, cargo), drones, robotic assistants, wheelchairs, or even tour guides [24, 25, 88, 94]. In these applications, sound analysis can be used to recognize and focus attention on sources outside the field of vision of autonomous devices, e.g., incoming traffic, emergency vehicles, someone calling; or to shape the system's response to contextual variables such as the terrain in which a robotic wheelchair is operating, or the soundscape level and composition to which an intelligent hearing aid needs to adjust.

These technologies can also contribute to content-based retrieval applications dealing with urban data, such as personal audio archiving [34], highlight extraction [93], video summarization [55], and searching through CCTV or mobile phone data [82]. In these scenarios sound analysis can help characterize patterns of similarity, novelty, anomaly, and recurrence in audio and multimedia content that can facilitate search and navigation.

However, there are two application domains in particular that are driving increased interest in automatic urban sound analysis: audio surveillance and noise pollution monitoring.

Audio Surveillance The need for automatic or semi-automatic surveillance in urban areas has experienced progressive and rapid growth, particularly in the past three decades. This is due to the increased threat posed by crime and terrorism. Surveillance systems were originally operated solely by humans, who had to constantly monitor video streams coming from the large number of cameras required to cover wide and complex areas of interest. In order to guarantee safety, however, full coverage of such areas would often require an unreasonably large number of operators. In addition, while it is difficult to outperform human monitoring with machines, this is only true when human attention is at its peak, which cannot be guaranteed over a lengthy period of time.

As a result, much effort has been devoted to the development of high-end technologies capable of alerting humans of potential hazards before they turn into a full-blown threat or calamity. Examples include the detection of fights/brawls [29, 53, 80] and intrusion [36, 97]. Technology improvements mean that infrared cameras for night-time operation have become affordable and less noisy; video

resolution can now guarantee an interocular distance of tens of pixels even from afar (for face recognition), and the dynamic range has grown to withstand the most adverse outdoor/indoor conditions. At the same time, signal processing for hazard detection has become more sophisticated, accommodating advanced illumination models; complex machine intelligence algorithms for video analytics; and advanced multimodal sensor fusion techniques, making fully automated surveillance systems effective and reliable enough to be fruitfully employed.

Many potentially dangerous events, however, can only be detected at an early stage through the analysis of an audio stream. Relevant examples range from the detection of specific sound sources such as gunshots, screams, and sirens; to actions like a car suddenly screeching to a stop; to scenes such as a brawl outside a night club, or a mugging. Audio surveillance is particularly beneficial in highly cluttered scenes, where visual events are likely to be occluded. Hence, the past decade has seen audio-based surveillance systems on the market, and new research focusing on the identification of dangerous events from the analysis of audio streams alone [27, 50, 69, 89] or from joint audio–video analysis [28, 96]. Crucially, sound event detection across dense sensor networks enables important surveillance capabilities such as localization and tracking of acoustic sources [9].

Noise Monitoring: Noise pollution is one of the topmost quality of life issues for urban residents worldwide [37]. In the United States alone, it has been estimated that over 70 million urban residents are exposed to harmful levels of noise [42, 62]. Such levels of exposure have proven effects on health such as sleep disruption, stress, hypertension, and hearing loss [8, 15, 43, 61, 90]. There is additional evidence of impact on learning and cognitive impairment in children [8, 14], productivity losses resulting from noise-related sleep disturbance [35, 92], and impact on real estate markets [63, 64].

Most major cities have ordinances that seek to regulate noise generation as a function of time of day/week and location. These codes define and measure noise in terms of overall sound pressure level (SPL) and its derivative metrics [87]. Such standards are in marked contrast with the emphasis on *sound sources* that is prevalent in noise surveys and complaints, as well as throughout the literature on the effect of noise pollution. The need for source-specific metrics is acknowledged by noise experts [87], especially in urban environments that are constantly reshaped by a large numbers of sources. As with audio surveillance, the benefits of applying sound classification technologies are evident and motivate recent efforts from the research community [68, 75–77].

The shortcomings of noise monitoring using SPL metrics are compounded by the difficulties of monitoring at scale. Site inspections by city officials are often few and far between and insufficient to capture the dynamics of noise across time and space. Alternatively, cities rely on civic complaint systems for noise monitoring such as New York City's 311, effectively the largest noise reporting system anywhere in the world [66]. However, research shows that noise information collected by such systems can be biased by location, socio-economic status, and source type, failing to accurately characterize noise exposure in cities [65]. Therefore, recent years have

seen a proliferation of work on using dense networks of mobile or fixed acoustic sensors as an alternative and complementary solution to noise monitoring. In this context, sound analysis can contribute to the identification of specific sources of noise and their characteristics (e.g., level, duration, intermittence, bandwidth). This can in turn empower novel insights in the social sciences and public policy regarding the relationship of urban sound to citizen complaints, reported levels of annoyance, stress, activity, as well as health, economic and educational outcomes.

13.3 Acoustic Sensor Networks

13.3.1 Mobile Sound Sensing

In recent years consumer mobile devices , namely smart phones have seen rapid improvements in processing power, storage capacity, embedded sensors, and network data rates. These advances coupled with their global ubiquity have paved the way for a new paradigm in large-scale remote urban sensing: participatory sensing [18, 21]. The idea behind this approach is to utilize the sensing, processing, and communication capabilities of consumer smart phones to enable members of the public to collect and upload environmental data from their surroundings. This approach benefits from the use of existing infrastructure (sensing platform and cellular networks) meaning that deployment costs are effectively zero, provides unrivaled spatial coverage and also allows for the gathering of the subjective response to these environments, in situ. The drawbacks of this approach mainly lie in the low temporal resolution of its data resulting from the submission of short term measurements and the quality of the gathered data, as the model, physical, and handling conditions of the smart phones may not be consistent, resulting in aggregated environmental data of variable accuracy. A number of initiatives have sought to crowdsource sound and noise monitoring using mobile devices [31, 45, 56, 72, 73, 79, 81]. Their apps are typically limited to logging geo-located instantaneous SPL measurements. The EveryAware project [4, 10, 11] is an EU project intending to integrate environmental monitoring, awareness enhancement and behavioral change by creating a new technological platform combining sensing technologies, networking applications, and data-processing tools. One of its subprojects is the WideNoise application, which allows for the compilation of noise pollution maps using participants' smart phones, including objective and subjective response data. In addition to this, they are examining the motivations for participation among their user base, as well as monitoring behavior change resulting from the access to personalized sound information. The OnoM@p project [41] follows some of the same goals and strategies of the above initiative. Notably, they attempt to address the issue of erroneous data through a cross-calibration technique between multiple device submissions, a welcome development for mobile noise sensing, with the caveat that it requires large-scale public adoption to be successful.

13.3.2 Static Sound Sensing

Static sound sensing solutions can take many forms with varying abilities and price points. Their main advantage over mobile sensing solutions is the ability to monitor continuously with increased levels of data quality. Highly accurate (± 0.7 dB), dedicated, commercially made networks such as the Bruel & Kjaer Noise Sentinel 3639-A/B/C [17] can produce legally enforceable acoustic data, but can cost upwards of $15,000 USD per node. The high cost means that deployments are spatially sparse with durations usually in the order of a few months. Lower cost commercial solutions include the $560 USD Libelium Waspmote Plug & Sense Smart Cities device [52] which, amongst other things, measures decibel (dB) values with a ± 3.0 dB accuracy. The reduced cost per sensor node brings with it new possibilities for larger network deployments but a trade-off on data accuracy may mean its suitability is limited for large-scale urban deployments. Other examples [60] make use of hybrid deployments of low-cost, low-accuracy sensors with higher-cost, higher-accuracy sensors in an attempt to strike a balance between accuracy and scalability. Networks utilizing even lower cost sensors at the $150 USD per sensor price point provide the potential for more network scalability, but make sacrifices in sensor capabilities. Examples of these [12, 46] make use of low-power computing cores that limit their ability to carry out any advanced in situ audio processing. With these acoustic sensor networks, it is desirable to have low-cost, powerful sensor nodes able to support the computational sound analysis techniques described in this book. The rest of this section will present the design and implementation of an acoustic sensor network capable of satisfying the cost, accuracy, and performance considerations described above.

13.3.3 Designing a Low-Cost Acoustic Sensing Device

In this section we describe the design of an acoustic sensing device developed in the context of the SONYC project,[2] a research initiative concerned with novel smart city solutions for urban noise monitoring, analysis, and mitigation. The device is based around the popular Raspberry Pi single-board computer (SBC) outfitted with a custom USB microelectromechanical systems (MEMS) microphone module where low-cost, acoustic accuracy, and high processing power are the primary considerations.

[2] https://wp.nyu.edu/sonyc/.

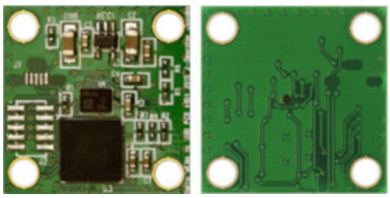

Fig. 13.1 Acoustic sensing module—back of board on *left* with MEMS microphone in center, front of board on *right* with microphone port in center

13.3.3.1 Microphone Module

In recent years, interest in microelectromechanical systems (MEMS) microphones has expanded due to their versatile design, greater immunity to radio frequency interference (RFI) and electromagnetic interference (EMI), low-cost and environmental resiliency [6, 7, 91]. Current MEMS models are generally 10× smaller than their more traditional electret counterparts. This miniaturization has allowed for additional circuitry to be included within the MEMS housing, such as a pre-amp stage and an ADC to output digitized audio in some models. The production process used to manufacture these devices also provides an extremely high level of part-to-part consistency, making them more amenable to multi-capsule and multi-sensor arrays. The sensing module shown in Fig. 13.1 uses an entirely digital design, utilizing a digital MEMS microphone (including a built-in ADC), and an onboard micro controller (MCU) enabling it to connect directly to the nodes computing device as a USB audio device. The digital MEMS microphone features a wide dynamic range of 32–120 dBA, ensuring all urban sound pressure levels can be effectively monitored. The use of an onboard MCU also allows for efficient, hardware level filtering of the incoming audio signal to compensate for the frequency response of the MEMS microphone before any further analysis is carried out. The standalone nature of this acoustic sensing module also means it is computing core agnostic, as it can be plugged into any computing device.

13.3.3.2 Form Factor, Cost, and Calibration

The sensor's prototype housing and form factor is shown in Fig. 13.2. The low-cost unfinished/unpainted aluminum housing was chosen to reduce radio frequency interference (RFI) from external sources, solar heat gain from direct sunlight and it also allows for ease of machining. All of the sensor's core components are housed within this rugged case except for the microphone and Wi-Fi antenna which is externalized for maximum signal gain.

In the prototype node shown in Fig. 13.2, the MEMS microphone is mounted externally via a repositionable metal goose-neck allowing the sensor node to be reconfigured for deployment in varying locations such as building sides, light poles, and building ledges. Figure 13.2 also shows the sensors bird spikes to ensure no damage is caused by perching birds. The total cost of the sensor excluding construction and deployment costs is $83 USD, as of December 2016.

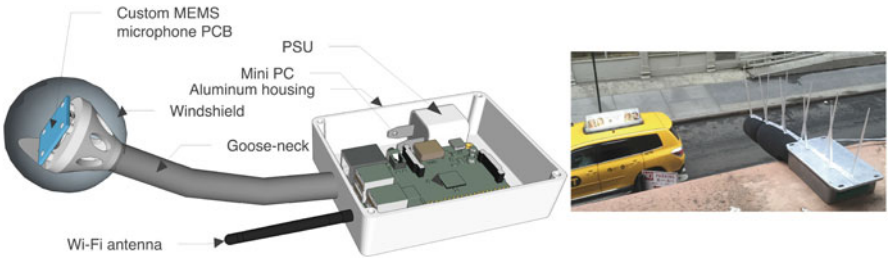

Fig. 13.2 Acoustic sensor node showing core components viewed from the underside (*left*) and a deployed node in NYC (*right*)

The sensing module was calibrated using a precision grade sound level meter as reference (Larson Davis 831 [49]), under low-noise, anechoic conditions. The sensor was then shown empirically to produce continuous decibel data at a level of accuracy used by the NYC city agencies tasked with enforcing the city's noise code.

13.3.4 Network Design & Infrastructure

The prototype node relies on a continuous power supply and wireless network connectivity so its deployment locations are mainly determined by these prerequisites. Security and wider localized spatial acoustic coverage of sensors is maintained by mounting at a height of ∼4 m above street level with a distance between sensors of around 2 city blocks or ∼150 m. Ideally acoustic sensors would be mounted on poles, rather than on or close to building sides to reduce variations in SPL response due to wall proximity. Partnering with infrastructure owners/managers is crucial when selecting and deploying sensor nodes, and it is worth noting that the cost of deploying a sensor on urban locations such as light poles can spiral when lifting equipment and professional personnel are involved. Selection of sites with the likelihood of high variation in sound sources is also prioritized in order to facilitate the collection of a wide variation of ground-truth audio data as discussed in Sect. 13.4. In order to maintain public privacy, audio data is captured, losslessly FLAC[3] compressed, and encrypted in 10 s snippets, interleaved with random durations of time. This data is transmitted from the sensor via Wi-Fi, directly to the project's control server, which in turn transfers the data to the storage servers, ready for further analysis. Each sensor also transmits its current state every minute via a small "status ping". This allows for near real-time remote telemetry display of all deployed sensors for fault diagnosis. Further in-depth control and maintenance of the deployed sensors is provided via a Virtual Private Network (VPN) that provides a method for remote Secure Shell (SSH) access to each node.

[3] https://xiph.org/flac/.

The VPN also enhances the wireless transmission security of the sensor as all data and control traffic is routed through this secure network. Future versions of the project's acoustic network will utilize multi-hop mesh networking approaches for sensor-server communications in order to increase the range of the network and reduce its power consumption to open up the possibility of battery powered, energy harvesting acoustic sensor nodes. Without the requirement of continuous power and pre-existing wireless network infrastructure, many more urban deployment possibilities become available.

13.4 Understanding Urban Soundscapes

Most prior work on understanding urban soundscapes has been focused on identifying acoustic scenes that are commonly found in urban environments such as parks, commercial streets, residential streets, construction sites, restaurants, or different modes of transportation (e.g., inside a taxi, train or bus). However, it is difficult to disambiguate work specific to urban environments from general acoustic scene classification (ASC) as described in Chap. 8. This is because the most widely used datasets for ASC research are largely or exclusively made from urban soundscapes. To make this clear, we provide a summary of those datasets in Table 13.1, where for each dataset we list the total number of audio recordings, the number of classes (acoustic scenes) and the number of these classes that can be considered urban sound scenes. As can be seen, all datasets contain a significant proportion of urban sound scenes.

While the focus of these datasets (and the approaches evaluated on them) is not necessarily urban sound scene analysis, they serve as a good proxy for it. Thus if we wish to understand the current state of the art in urban sound scene classification, we can refer to the DCASE 2016 acoustic scene classification challenge,[4] which was based on the TUT Acoustic Scenes 2016 dataset [59] listed in Table 13.1. The challenge received close to 50 submissions spanning a variety of techniques,

Table 13.1 Some commonly used datasets for acoustic scene classification

Dataset	Recordings	Total scenes	Urban scenes
UAE noise DB series 1 [84]	10	10	9
UAE noise DB series 2 [84]	35	12	11
DCASE 2013 [38]	100	10	10
DARES G1 [40]	123	28	25
TUT acoustic scenes 2016 [59]	1170	15	13
LITIS rouen [71]	3026	19	19

As seen from the table, all datasets contain a significant proportion of urban scenes

[4] http://www.cs.tut.fi/sgn/arg/dcase2016/.

ranging from a baseline system which uses MFCC features with a GMM classifier, to deep learning architectures including fully connected and convolutional neural networks trained on a variety of input representations. Since the general problem of scene classification and the DCASE challenge are discussed in detail in previous chapters, here we will only limit ourselves to point out that the maximum reported classification accuracy was of 0.897, with incremental differences of 1% to the second and third best performing systems, and that the best performing method in the challenge was based on the late fusion of a deep and a shallow feature learner [33]. For a detailed comparison of algorithmic performance and further details about all participating methods the reader is referred to the challenge's results page.[5]

The challenge supports the notion that current strategies are already capable of providing robust solutions to urban ASC. This is not new, since high performance in this task has been reported for close to a decade at the time of writing [5]. At the same time, practically all datasets used for ASC evaluation to date are closed-set, meaning the data are divided into a fixed, known number of scenes. In a real-world scenario (for instance, a robot operating in a new environment) it is possible to encounter previously unheard acoustic scenes, which a model would have to identify as "unknown". Existing models are not trained to perform this task, which requires open-set data for training, and it is quite possible that model performance on this (more challenging) scenario would be lower.

Next we turn our attention to the more challenging task of sound source identification, which has received less attention and has ample room for improvement. As was the case before, this task is covered in detail elsewhere in this book, which is why for the rest of this section we will focus on research specifically targeting urban environments.

13.4.1 Urban Sound Dataset

In Chap. 6 a number of annotated datasets for environmental sound event detection and classification were discussed. While some of these contain sound events from urban soundscapes, up to 2013 there was no dataset focusing specifically on urban sounds. Previous work has focused on audio from carefully produced movies or television tracks [19], from specific environments such as elevators or office spaces [39, 70], and on commercial or proprietary datasets [23, 44]. The large effort involved in manually annotating real-world data means datasets based on field recordings tend to be relatively small (e.g., the event detection dataset of the IEEE AASP Challenge [39] consists of 24 recordings per each of 17 classes). A second challenge faced by the research community was the lack of a common vocabulary when working with urban sounds. This meant the classification of sounds into semantic groups varied from study to study, making it hard to compare results.

[5] http://www.cs.tut.fi/sgn/arg/dcase2016/task-results-acoustic-scene-classification.

Specific efforts to describe urban sounds have often been limited to subsets of broader taxonomies of acoustic environments (e.g., [16]), and thus only partially fulfill the needs of systematic urban sound analysis. To address this, Salamon et al. proposed an urban sound taxonomy [77] based on the subset of the taxonomy proposed by Brown et al. [16] dedicated to the urban acoustic environment. This taxonomy defines four top-level groups: human, nature, mechanical, and music, which are common in the literature [67], and specifies that its leaves should be sufficiently low-level to be unambiguous—e.g., car "brakes," "engine," or "horn," instead of simply "car." Furthermore, it is built around the most frequently complained about sound categories and sources—e.g., construction (e.g., jackhammer), traffic noise (car and truck horns, idling engines), loud music, air conditioners and dog barks—according to 370,000 noise complaints filed through New York City's 311 service from 2010 to 2013.[6]

A subset of the resulting taxonomy, focused on mechanical sounds, is provided in Fig. 13.3. A scalable digital version of the complete taxonomy is available online.[7] Rounded rectangles represent high-level semantic classes (e.g., human, nature, mechanical, music). The leaves of the taxonomy (rectangles with sharp edges) correspond to classes of concrete sound sources (e.g., siren, footsteps). For conciseness, leaves can be shared by several high-level classes (indicated by an earmark).

From this taxonomy, a dataset [77] was developed by focusing on ten low-level classes: air conditioner, car horn, children playing, dog bark, drilling, engine idling, gunshot, jackhammer, siren, and street music. With the exception of "children playing" and "gun shot" which were added for variety, all other classes were selected due to the high frequency in which they appear in NYC urban noise complaints.

The audio data was collected from Freesound,[8] an online sound repository containing over 160,000 user-uploaded recordings under a creative commons license. For each class, the authors downloaded all sounds returned by the Freesound search engine when using the class name as a query (e.g., "jackhammer"), manually inspected all recordings and kept only actual urban field recordings where the sound class of interest was present, and used Audacity[9] to label the start and end times of every occurrence of the sound in each recording, with an additional *salience* description indicating whether the occurrence was subjectively perceived to be in the foreground or background of the recording. This resulted in a total of 3075 labeled occurrences amounting to 18.5 h of labeled audio. The distribution of total occurrence duration per class and per salience is provided in Fig. 13.4a.

The resulting dataset of 1302 full and variable length recordings with corresponding sound occurrence and salience annotations, *UrbanSound*, is freely available

[6] https://nycopendata.socrata.com/data.

[7] http://serv.cusp.nyu.edu/projects/urbansounddataset/.

[8] http://www.freesound.org.

[9] http://audacity.sourceforge.net/.

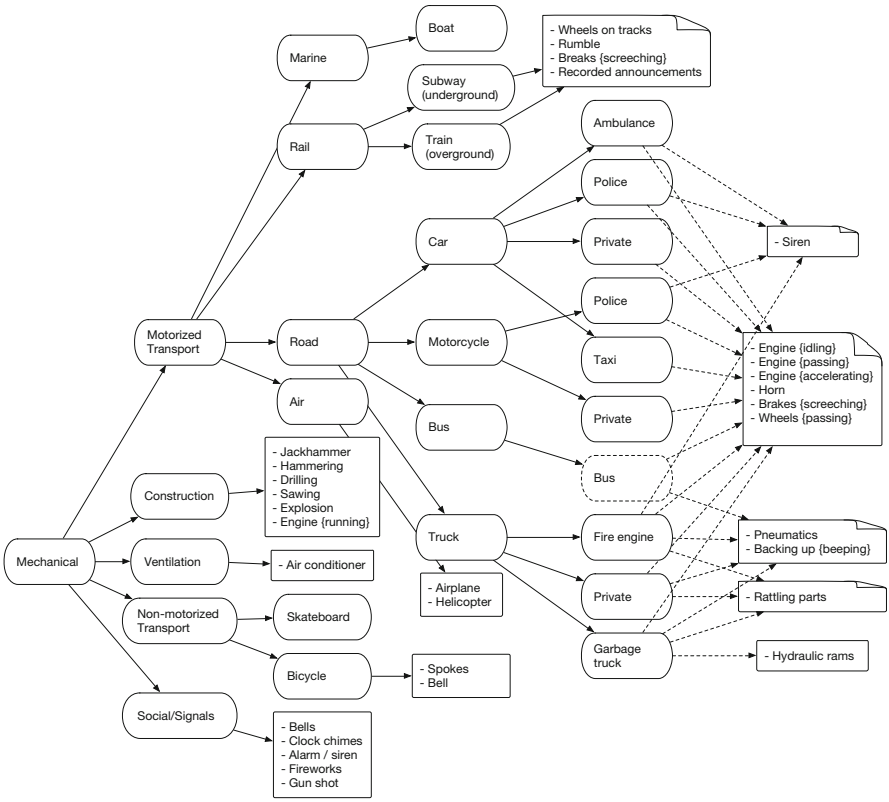

Fig. 13.3 Subset of the Urban Sound Taxonomy [77] focusing on mechanical sounds

online.[10] Moreover, for research on sound source classification the authors curated a subset of short audio snippets, the *UrbanSound8K* dataset (also available online at the same url). Following the findings in [25], these snippets are limited to a maximum duration of 4 s. Longer clips are segmented into 4 s clips using a sliding window with a hop size of 2 s. To avoid large differences in the class distribution, there is a limit of 1000 clips per class, resulting in a total of 8732 labeled clips (8.75 h). The distribution of clips per class in UrbanSound8K with a breakdown into salience is provided in Fig. 13.4b.

A number of signal processing techniques and machine learning models have been proposed to date for urban sound classification and evaluated on the UrbanSound8K dataset [68, 74–77]. In the following sections we will review and contrast these approaches, comparing their performance in terms of classification accuracy. A summary of the key characteristics of each approach is provided in Table 13.2.

[10] http://serv.cusp.nyu.edu/projects/urbansounddataset/.

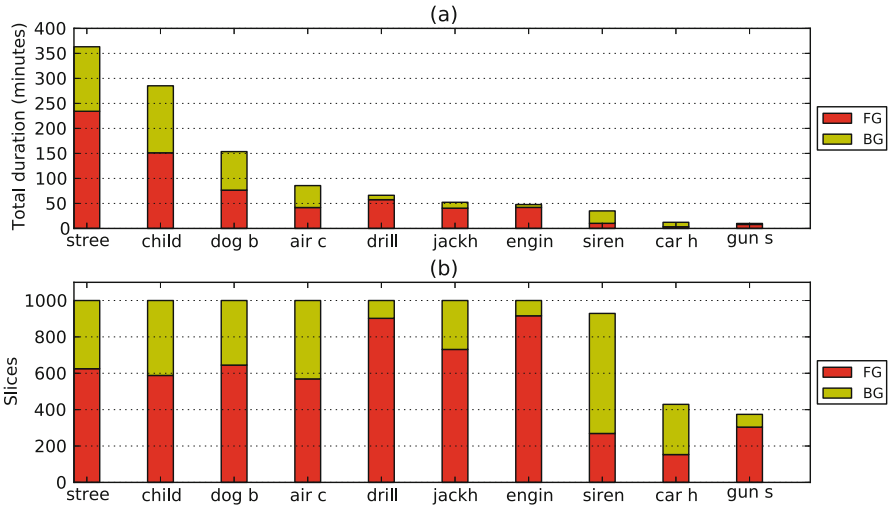

Fig. 13.4 (**a**) Total occurrence duration per class in UrbanSound. (**b**) Clips per class in UrbanSound8K. Breakdown by foreground (FG)/background (BG)

Table 13.2 Methods for urban sound classification

Method	Input features	Model
Baseline [77]	MFCC summary statistics	SVM
SKM-mel [75]	Dictionary encoded log-mel-spectrogram	Random forest
SKM-scattering [74]	Dictionary encoded deep scattering spectrum	SVM
Piczak-CNN [68]	Log-mel-spectrogram + delta	Deep CNN
SB-CNN [76]	Log-mel-spectrogram	Deep CNN + augmentation

13.4.2 Engineered vs Learned Features

The first step employed by all methods listed in Table 13.2 is feature extraction, i.e., transforming the raw audio signal into a feature space that is more amenable to machine learning. We can group audio feature spaces into two broad categories: designed (or engineered) features, and learned features. The former includes all features whose computation is independent of the input data, i.e., they are defined as the concatenation of operations whose goal is to capture a certain characteristic of the audio signal. The latter category includes features spaces that are learned directly from the data, including, for example, dictionary learning and deep learning methods.

Audio classification systems, including methods for environmental sound source classification, have traditionally relied on engineered features [19, 44, 70]. Thus the baseline system listed in Table 13.2 is a combination of a popular feature, the Mel-Frequency Cepstral Coefficients (MFCC), and a standard classification model

(Random Forest). However, most recent methods, including the remainder of the methods listed in the table, fall under the category of feature learning.

The first method listed following the baseline, SKM-mel [75], is based on unsupervised dictionary learning. The idea is to learn a dictionary of representative codewords directly from the audio signal in a data-driven fashion. The learned dictionary is then used to encode the samples in a dataset into feature vectors, which are then used to train/test a discriminative model of choice. The method employs the *spherical k-means* algorithm (SKM [26]) to learn the dictionary. Unlike the traditional k-means clustering algorithm [54], the codewords are constrained to have unit L2 norm (they must lie on the unit sphere, preventing them from becoming arbitrarily large or small), and represent the distribution of meaningful directions in the data. Compared to standard k-means, SKM is less susceptible to events carrying a significant amount of the total energy of the signal (e.g., background noise) dominating the dictionary. The algorithm is efficient and highly scalable, and it has been shown that the resulting set of vectors can be used as a dictionary for mapping new data into a feature space which reflects the discovered regularities [26, 30, 86]. The algorithm is competitive with more complex (and consequently slower) techniques such as sparse coding and has been used successfully to learn features from audio for music [32], and birdsongs [86]. After applying this clustering to the training data, the resulting cluster centroids can be used as the codewords of the learned dictionary. The number of codewords learned is typically much larger than the number of classes present in the data. It is also typically larger than the dimensionality of the input representation, i.e., the algorithm is used to learn an over-complete dictionary.

The clustering produces a dictionary matrix with k columns, where each column represents a codeword. Every sample in the dataset is encoded against the dictionary by taking the matrix product between each frame of its input representation, a mel-spectrogram, and the dictionary matrix. Every column i ($i = 1 \ldots k$) in the resulting encoded matrix can be viewed as a time series whose values represent the match scores between the input representation and the ith codeword in the dictionary: when the input is similar to the codeword the value in the time series will be higher, and when it is dissimilar the value will be lower.

To ensure that all samples in the dataset are represented by a feature vector of the same dimensionality, the time series are summarized over the time axis by computing the mean and standard deviation of each time series and using these as features. The resulting feature vectors are thus all of size $2k$ and are standardized across samples before being passed on to the classifier for training and testing.

Note that for learning, one can choose to learn features from individual frames of the input representation, or alternatively group the frames into 2D patches and apply the learning algorithm to the patches. In [75] the authors show that the latter approach facilitates the learning of features that capture short-term temporal dynamics, which proves to be important for urban sound classification. The best result reported by the authors was obtained using patches with a time duration of roughly 370 ms (16 frames). For training, patches are extracted from the mel-spectrogram using a sliding window with a hop size of 1 frame. This results in

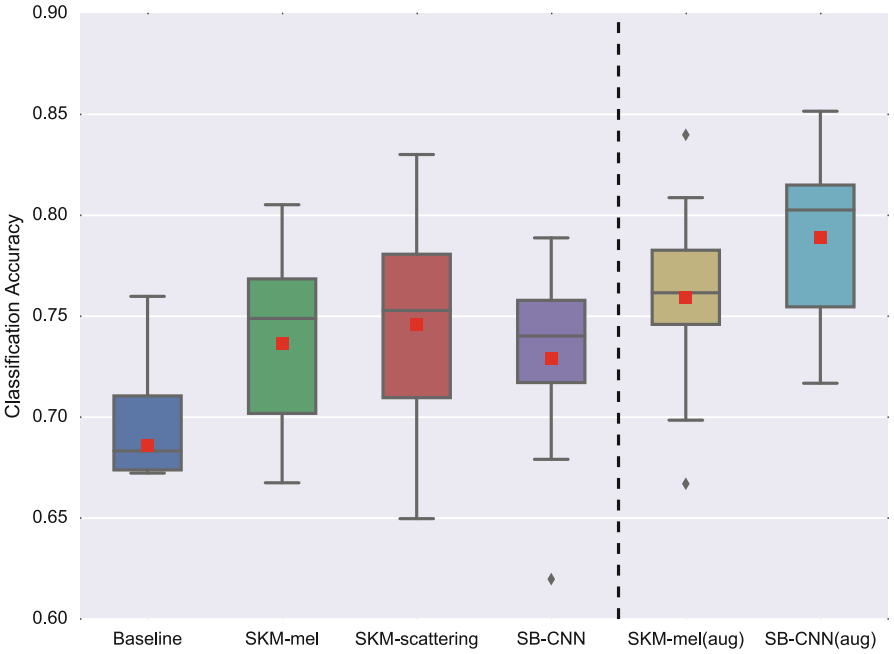

Fig. 13.5 Classification accuracy obtained on the UrbanSound8K dataset by different models: MFCC Baseline [77], spherical k-means dictionary learning from mel spectra [75] (SKM-mel), SKM learned from deep scattering spectra [74] (SKM-scattering) and the deep CNN proposed by Salamon and Bello [76] (SB-CNN). Models to the *left of the dashed line* were trained without data augmentation. To the *right of the dashed line* we present the results obtained by SKM-mel and SB-CNN when trained on an augmented training set: SKM-mel(aug) and SB-CNN(aug), respectively

significantly more training data for the unsupervised dictionary learning stage, and also ensures that the learned codewords account for different time-shifts of each sound source, hopefully increasing the robustness of the model to such shifts in the data.

While one could use the resulting patches directly as input for the feature learning, it has been shown that the learned features can be significantly improved by decorrelating the input dimensions using, e.g., Zero-phase Component Analysis (ZCA) whitening [47] or Principal Component Analysis (PCA) whitening [26].

Figure 13.5 presents classification accuracy results for UrbanSound8K in the form of a boxplot computed from the per-fold accuracies obtained by each model. Mean accuracies are indicated by the red squares. We will initially focus on the two left-most boxes and will discuss the remainder of the results in the following sections.

We clearly see that the SKM-mel model outperforms the MFCC baseline, with mean accuracies of 0.74 and 0.68, respectively. The difference is robust to the parameters of the mel-spectrogram, which are optimal for both reported results, but depends on the size of the dictionary for SKM, with best results for $k = 2000$ [75].

Such a significant improvement provides clear evidence of the advantage of feature learning compared to off-the-shelf engineered features, even when using a simple and shallow feature learning approach such as SKM.

13.4.3 Shift Invariance via Convolutions

The following method in Table 13.2, SKM-scattering [74], uses a different input representation altogether—the *scattering transform* [1–3]. This representation can be viewed as an extension of the mel-spectrogram that computes modulation spectrum coefficients of multiple orders through cascades of wavelet convolutions and modulus operators. Given a signal x, the first-order (or "layer") scattering coefficients are computed by convolving x with a wavelet filterbank ψ_{λ_1}, taking the modulus, and averaging the result in time by convolving it with a low-pass filter $\phi(t)$ of size T:

$$S_1 x(t, \lambda_1) = |x * \psi_{\lambda_1}| * \phi(t). \tag{13.1}$$

The wavelet filterbank ψ_{λ_1} has an octave frequency resolution Q_1. By setting $Q_1 = 8$ the filterbank has the same frequency resolution as the mel filterbank, and this layer is approximately equivalent to the mel-spectrogram. The second-order coefficients capture the high-frequency amplitude modulations occurring at each frequency band of the first layer and are obtained by:

$$S_2 x(t, \lambda_1, \lambda_2) = ||x * \psi_{\lambda_1}| * \psi_{\lambda_2}| * \phi(t). \tag{13.2}$$

In [74] $Q_1 = 8$ and $Q_2 = 1$, the filterbank is constructed of 1D Morlet wavelets, and T is set to the same duration covered by the 2D mel-spectrogram patches used for dictionary learning in [75], i.e., 370 ms (for a sampling rate of 44,100 Hz this implies $T = 1024 \times 16$). Higher order coefficients can be obtained by iterating over this process, but it has been shown that for the chosen value of T, most of the signal energy is captured by the first- and second-order coefficients [3].

For each frame the first-order coefficients are concatenated with all of the second-order coefficients into a single feature vector. The second order coefficients are normalized using the previous order coefficients as described in [3]. From this point the process replicates the method described in the previous section: PCA whitening, dictionary learning using SKM, projection into the feature space, summarization and classification, in this case using a support vector machine (although the difference with a random forest is minimal). Therefore, the main difference between [75] and [74] is the addition of a *phase-invariant convolutional layer*, which is able to capture amplitude modulations in the input representation, in a time-shift invariant manner.

Figure 13.5 shows that learning a dictionary from scattering coefficients as opposed to the mel-spectrogram results in a relatively marginal improvement in classification accuracy (0.75 vs 0.74). Notably, the authors did observe a 5

percentage-point absolute improvement in the classification accuracy (i.e., +0.05) of masked sounds, i.e., sounds that were labeled by the annotators of the dataset as being in the background of the acoustic scene. This fits with findings in the sound perception and cognition literature showing that modulation plays an important role in sound segregation and the formation of auditory images [22, 57, 95], further motivating the exploration of deep convolutional representations, such as the scattering transform, for machine listening.

However, the most important finding is that the scattering transform's inherent invariance to local time shifts allows for the comparable performance between SKM-scattering and SKM-mel, but using a dictionary that is an order of magnitude smaller ($k = 200$ versus $k = 2000$) while reducing the amount of 2D patches (samples) necessary for training by an order of magnitude too. In other words, shift invariance results in smaller machines trained with less data that are equally powerful, a finding that motivates further exploration using deep convolutional approaches.

13.4.4 Deep Learning and Data Augmentation

The last two methods in Table 13.2, Piczak-CNN [68] and SB-CNN [76], are based on deep (feature) learning [51]. This means that unlike the methods described above, here there are multiple feature learning layers including both fully-connected (like SKM) and convolutional (like scattering) layers, the feature learning is fully integrated with the classifier, and the machine is trained using supervised methods and a discriminative objective.

Since the two CNNs perform comparably when trained on the original UrbanSound8K dataset (and only SB-CNN was evaluated both with and without data augmentation as discussed further below), for the remainder of the discussion we shall focus on SB-CNN as an instance of a deep learning model. SB-CNN takes log-scaled mel-spectrograms with 128 bands and a duration of 3 s as input to the network. Each 3 s spectrogram "patch" is Z-score normalized. The model is comprised of three convolutional layers interleaved with two pooling operations, followed by two fully connected (dense) layers. Notably, the convolutional layers of SB-CNN use a comparatively small receptive field of $(5, 5)$ compared to the input dimensions of $(128, 128)$. This is intended to allow the network to learn small, localized patterns, or cues, that can progressively build-up evidence for the presence/absence of specific sources even when there is spectro-temporal masking by interfering sources.

During training the model optimizes cross-entropy loss via mini-batch stochastic gradient descent [13]. Each batch consists of 100 patches randomly selected from the training data (without repetition). The model is trained using a constant learning rate of 0.01 and dropout [85] with probability 0.5 is applied to the input of the last two layers. L2-regularization is applied to the weights of the last two layers with a penalty factor of 0.001. The model is trained for 50 epochs with

a validation set used to identify the parameter setting (epoch) that achieves the highest classification accuracy. Prediction is performed by slicing the test sample into overlapping patches, making a prediction for each patch and finally choosing the sample-level prediction as the class with the highest mean output activation over all patches.

From Fig. 13.5 we see that SB-CNN, while outperforming the baseline, does not outperform its "shallow" SKM counterpart. This suggests that the UrbanSound8K dataset, despite being the largest dataset publicly available for urban sound classification, is not sufficiently large for the benefits of high-capacity, deep learning models to become apparent.

To address this limitation and increase the model's robustness to intra-class variance, the authors also trained SB-CNN using *data augmentation*, that is, the application of one or more deformations to the training set which result in new, additional training data [48, 58, 83]. Assuming the deformations do not change the validity of the labels, augmentation aims to increase the model's invariance to said transformations and thus generalize better to unseen data.

The authors applied four types of audio deformations: time stretching, pitch shifting, dynamic range compression, and the addition of background noise at different SNR, resulting in a training set an order of magnitude larger than the original UrbanSound8K. Augmentation was performed using the MUDA library [58]. After training SB-CNN with augmentation [Fig. 13.5: SB-CNN(aug)], the model significantly outperforms the SKM approach. Furthermore, we see that this improvement is not independent of the use of deep learning—training the SKM approach with augmentation [Fig. 13.5: SKM-mel(aug)] failed to improve as much. Increasing the capacity of the SKM model by increasing the dictionary size from $k = 2000$ to $k = 4000$ did not yield any further improvement either, even with the augmented training set. Instead, it is the combination of an augmented training set and the increased capacity and representational power of the deep learning model that results in this state-of-the-art performance.

13.5 Conclusion and Future Perspectives

In this chapter we have discussed intelligent acoustic sensing and analysis in the context of urban environments, particularly as one component of a larger trend towards smart city solutions. While we discuss a range of potential applications, we focus on two, audio surveillance and noise monitoring, that motivate new and exciting developments at the intersection of ubiquitous sensing and machine listening capabilities such as sound event detection, classification, localization, and tracking. These new technologies have the potential to improve the public safety and quality of life of urban residents.

In our discussion of acoustic sensor networks, we clearly favored the use of static over mobile sensing, and presented an example of a low-cost, high-quality solution intended for noise monitoring. However, the intended application greatly influences that choice: precise source localization and tracking is desirable but not necessary for noise monitoring, and the cyclic and seasonal nature of noise patterns means that off-network responses can be estimated by exploiting spatial correlations with other data types encoding information about, e.g., traffic, zoning, nightlife, construction, and tourist activity. On the other hand, audio surveillance requires relatively-dense arrays of sensors, something that is prohibitively expensive for static sensor networks, even for low-cost solutions such as the one presented in Sect. 13.3. One possibility is to deploy selectively and densely, as it is done for specific applications such as gunshot detection in neighborhoods with high gun crime incidences.[11] However, this is not applicable to surveillance scenarios (e.g., emergencies or terrorism) which are less predictable in space. Therefore, future developments will most likely require leveraging sensing from smart phones and other consumer-grade mobile devices, which in turn requires finding robust solutions to on-the-fly calibration, synchronization, and embedded computing that work well for acoustic data.

We devoted significant attention to the tasks of sound event detection and classification in cities. While the results are promising and much improvement has been accrued in a short period of time, there is still significant room for improvement and important challenges ahead. For example, one of the challenges of urban sound analysis is the heterogeneity of source types, a problem for which large-capacity models and ensemble methods might prove beneficial, as has been shown in acoustic scene classification [33] and bioacoustic classification [78]. However, current annotated datasets are small, include only a handful out of hundreds of possible sources, and are weakly labeled, meaning that comprehensive multi-source annotations are the exception rather than the norm. This hinders the ability to test such solutions.

Furthermore, real-world applications are intended to work on continuous audio streams, but many of the datasets discussed only contain snippets and thus fail to characterize the complex temporal dynamics of urban soundscapes. This scenario calls for the exploitation of longer temporal relationships, making the combination of convolutional and recurrent models an attractive direction for future research. These problems and solutions have been studied in the context of general environmental sound analysis (e.g., [20]), but remains to be explored for urban applications.

Finally, these sets only contain a small and arbitrary sample of the full range of acoustic conditions one might encounter in urban outdoor environments, and to which these systems are supposed to generalize. While data augmentation can help to a certain extent, future developments will be dependent on significant data collection from large-scale acoustic sensor networks, whether mobile or fixed. Encouraging developments include the recent launch of the YouTube-8M dataset

[11] http://www.shotspotter.com/.

of tagged videos,[12] which contain a sizable and diverse sample of urban acoustic environments from mobile devices, and the ongoing deployment of audio sensor networks by various smart cities initiatives such as SONYC.[13]

References

1. Andén, J., Mallat, S.: Multiscale scattering for audio classification. In: 12th International Society for Music Information Retrieval Conference, Miami, pp. 657–662 (2011)
2. Andén, J., Mallat, S.: Scattering representation of modulated sounds. In: 15th DAFx, York (2012)
3. Andén, J., Mallat, S.: Deep scattering spectrum. IEEE Trans. Signal Process. **62**(16), 4114–4128 (2014)
4. Atzmueller, M., Becker, M., Doerfel, S., Hotho, A., Kibanov, M., Macek, B., Mitzlaff, F., Mueller, J., Scholz, C., Stumme, G.: Ubicon: observing physical and social activities. In: 2012 IEEE International Conference on Green Computing and Communications (GreenCom), pp. 317–324. IEEE, New York (2012)
5. Aucouturier, J., Defreville, B., Pachet, F.: The bag-of-frames approach to audio pattern recognition: a sufficient model for urban soundscapes but not for polyphonic music. J. Acoust. Soc. Am. **122**(2), 881–891 (2007)
6. Barham, R., Goldsmith, M., Chan, M., Simmons, D., Trowsdale, L., Bull, S.: Development and performance of a multi-point distributed environmental noise measurement system using mems microphones. In: Proceedings of the 8th European Conference on Noise Control (Euronoise 2009) (2009)
7. Barham, R., Chan, M., Cand, M.: Practical experience in noise mapping with a MEMS microphone based distributed noise measurement system. In: 39th International Congress and Exposition on Noise Control Engineering (Internoise 2010) (2010)
8. Basner, M., Babisch, W., Davis, A., Brink, M., Clark, C., Janssen, S., Stansfeld, S.: Auditory and non-auditory effects of noise on health. The Lancet **383**(9925), 1325–1332 (2014)
9. Baxter, K.C., Fisher, K.: Gunshot detection sensor with display. US Patent 7,266,045, 2007
10. Becker, M., Caminiti, S., Fiorella, D., Francis, L., Gravino, P., Haklay, M.M., Hotho, A., Loreto, V., Mueller, J., Ricchiuti, F., et al.: Awareness and learning in participatory noise sensing. PLoS One **8**(12), e81638 (2013)
11. Becker, M., Mueller, J., Hotho, A., Stumme, G.: A generic platform for ubiquitous and subjective data. In: Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing Adjunct Publication, pp. 1175–1182. ACM, New York (2013)
12. Bell, M.C., Galatioto, F.: Novel wireless pervasive sensor network to improve the understanding of noise in street canyons. Appl. Acoust. **74**(1), 169–180 (2013)
13. Bottou, L.: Large-scale machine learning with stochastic gradient descent. In: 19th International Conference on Computational Statistics (COMPSTAT), Paris, pp. 177–186 (2010)
14. Bronzaft, A.L.: The effect of a noise abatement program on reading ability. J. Environ. Psychol. **1**(3), 215–222 (1981)
15. Bronzaft, A.: Neighborhood noise and its consequences. Survey Research Unit, School of Public Affairs, Baruch College, New York (2007)
16. Brown, A.L., Kang, J., Gjestland, T.: Towards standardization in soundscape preference assessment. Appl. Acoust. **72**(6), 387–392 (2011)

[12] https://research.googleblog.com/2016/09/announcing-youtube-8m-large-and-diverse.html.

[13] https://wp.nyu.edu/sonyc/.

17. Bruel & Kjaer Noise Monitoring Terminal Type 3639 (2015). http://www.bksv.com/Products/EnvironmentManagementSolutions/UrbanEnvironmentManagement/NoiseInstrumentation/NoiseMonitoringTerminalFamily
18. Burke, J.A., Estrin, D., Hansen, M., Parker, A., Ramanathan, N., Reddy, S., Srivastava, M.B.: Participatory sensing. Center for Embedded Network Sensing (2006)
19. Cai, L.H., Lu, L., Hanjalic, A., Zhang, H.J., Cai, L.H.: A flexible framework for key audio effects detection and auditory context inference. IEEE Trans. Audio Speech Lang. Process. **14**(3), 1026–1039 (2006). doi:10.1109/TSA.2005.857575
20. Cakir, E., Heittola, T., Huttunen, H., Virtanen, T.: Polyphonic sound event detection using multi label deep neural networks. In: 2015 International Joint Conference on Neural Networks (IJCNN), pp. 1–7 (2015)
21. Campbell, A.T., Eisenman, S.B., Lane, N.D., Miluzzo, E., Peterson, R.A.: People-centric urban sensing. In: Proceedings of the 2nd Annual International Workshop on Wireless Internet, p. 18. ACM, New York (2006)
22. Carlyon, R.: How the brain separates sounds. Trends Cogn. Sci. **8**(10), 465–471 (2004)
23. Chaudhuri, S., Raj, B.: Unsupervised hierarchical structure induction for deeper semantic analysis of audio. In: IEEE ICASSP, pp. 833–837 (2013). doi:10.1109/ICASSP.2013.6637765
24. Chu, S., Narayanan, S., Kuo, C.C.J., Mataric, M.J.: Where am I? scene recognition for mobile robots using audio features. In: 2006 IEEE International Conference on Multimedia and Expo, pp. 885–888. IEEE, New York (2006)
25. Chu, S., Narayanan, S., Kuo, C.C.: Environmental sound recognition with time-frequency audio features. IEEE Trans. Audio Speech Lang. Process. **17**(6), 1142–1158 (2009). doi:10.1109/TASL.2009.2017438
26. Coates, A., Ng, A.Y.: Learning feature representations with K-means. In: Neural Networks: Tricks of the Trade, pp. 561–580. Springer, Berlin, Heidelberg (2012)
27. Cristani, M., Bicego, M., Murino, V.: On-line adaptive background modelling for audio surveillance. In: Proceedings of the 17th International Conference on Pattern Recognition, 2004 (ICPR 2004), vol. 2, pp. 399–402. IEEE, New York (2004)
28. Cristani, M., Bicego, M., Murino, V.: Audio-visual event recognition in surveillance video sequences. IEEE Trans. Multimedia **9**(2), 257–267 (2007)
29. Cristani, M., Raghavendra, R., Bue, A.D., Murino, V.: Human behavior analysis in video surveillance: a social signal processing perspective. Neurocomputing **100**, 86–97 (2013)
30. Dhillon, I., Modha, D.: Concept decompositions for large sparse text data using clustering. Mach. Learn. **42**(1), 143–175 (2001)
31. D'Hondt, E., Stevens, M., Jacobs, A.: Participatory noise mapping works! an evaluation of participatory sensing as an alternative to standard techniques for environmental monitoring. Pervasive Mob. Comput. **9**(5), 681–694 (2013)
32. Dieleman, S., Schrauwen, B.: Multiscale approaches to music audio feature learning. In: 14th ISMIR, Curitiba (2013)
33. Eghbal-Zadeh, H., Lehner, B., Dorfer, M., Widmer, G.: CP-JKU submissions for DCASE-2016: a hybrid approach using binaural i-vectors and deep convolutional neural networks. Technical report, DCASE2016 Challenge (2016)
34. Ellis, D.P.W., Lee, K.: Minimal-impact audio-based personal archives. In: 1st ACM workshop on Continuous Archival and Retrieval of Personal Experiences, New York, NY, pp. 39–47 (2004)
35. First report of the Interdepartmental Group on Costs and Benefits, Noise Subject Group: An economic valuation of noise pollution – developing a tool for policy appraisal. Department for Environment, Food and Rural Affairs (2008)
36. Foresti, G.: A real-time system for video surveillance of unattended outdoor environments. IEEE Trans. Circuits Syst. Video Technol. **8**(6), 697–704 (1998)
37. García, A.: Environmental Urban Noise. Wentworth Institute of Technology Press, Boston, MA (2001)

38. Giannoulis, D., Benetos, E., Stowell, D., Plumbley, M.D.: IEEE AASP challenge on detection and classification of acoustic scenes and events - public dataset for scene classification task. Technical report, Queen Mary University of London (2012)
39. Giannoulis, D., Stowell, D., Benetos, E., Rossignol, M., Lagrange, M., Plumbley, M.D.: A database and challenge for acoustic scene classification and event detection. In: 21st EUSIPCO (2013)
40. Grootel, M., Andringa, T., Krijnders, J.: DARES-G1: Database of annotated real-world everyday sounds. In: Proceedings of the NAG/DAGA Meeting 2009, Rotterdam (2009)
41. Guillaume, G., Can, A., Petit, G., Fortin, N., Palominos, S., Gauvreau, B., Bocher, E., Picaut, J.: Noise mapping based on participative measurements. Noise Mapp. **3**(1), 140–156 (2016)
42. Hammer, M.S., Swinburn, T.K., Neitzel, R.L.: Environmental noise pollution in the United States: developing an effective public health response. Environ. Health Perspect. **122**(2), 115–119 (2014)
43. Heinrich, U.R., Feltens, R.: Mechanisms underlying noise-induced hearing loss. Drug Discov. Today Dis. Mech. **3**(1), 131–135 (2006)
44. Heittola, T., Mesaros, A., Eronen, A., Virtanen, T.: Context-dependent sound event detection. EURASIP J. Audio Speech Music Process. **2013**, 1 (2013)
45. Kanjo, E.: Noisespy: a real-time mobile phone platform for urban noise monitoring and mapping. Mob. Netw. Appl. **15**(4), 562–574 (2010)
46. Kivelä, I., Gao, C., Luomala, J., Ihalainen, J., Hakala, I.: Design of networked low-cost wireless noise measurement sensors. Sensors Transducers **10**, 171 (2011)
47. Krizhevsky, A.: The ZCA whitening transformation. Appendix A of learning multiple layers of features from tiny images, Technical Report, University of Toronto (2009)
48. Krizhevsky, A., Sutskever, I., Hinton, G.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NIPS), pp. 1097–1105 (2012)
49. Larson Davis Model 831-NMS permanent noise monitoring system (2015). http://www.larsondavis.com/Products/NoiseMonitoringSystems/PermanentNoiseMonitoringSystem
50. Lecomte, S., Lengellé, R., C. Richard, C., Capman, F., Ravera, B.: Abnormal events detection using unsupervised one-class svm-application to audio surveillance and evaluation. In: 2011 8th IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS), pp. 124–129. IEEE, New York (2011)
51. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
52. Libelium smart cities board technical guide (2015). http://www.libelium.com/development/waspmote/documentation/smart-cities-board-technical-guide/
53. Lin, W., Sun, M., Poovendran, R., Zhang, Z.: Group event detection for video surveillance. In: 2009 IEEE International Symposium on Circuits and Systems, pp. 2830–2833. IEEE, New York (2009)
54. Lloyd, S.: Least squares quantization in PCM. IEEE Trans. Inf. Theory **28**(2), 129–137 (1982)
55. Ma, Y.F., Lu, L., Zhang, H.J., Li, M.: A user attention model for video summarization. In: 10th ACM International Conference on Multimedia, pp. 533–542 (2002)
56. Maisonneuve, N., Stevens, M., Ochab, B.: Participatory noise pollution monitoring using mobile phones. Inf. Polity **15**(1), 51–71 (2010)
57. McAdams, S.: Spectral fusion, spectral parsing and the formation of auditory images. Ph.D. thesis, Stanford University, Stanford (1984)
58. McFee, B., Humphrey, E., Bello, J.: A software framework for musical data augmentation. In: 16th International Society for Music Information Retrieval Conference, pp. 248–254. Malaga, Spain (2015)
59. Mesaros, A., Heittola, T., Virtanen, T.: TUT database for acoustic scene classification and sound event detection. In: 24th European Signal Processing Conference (EUSIPCO), Budapest (2016)
60. Mietlicki, F., Mietlicki, C., Sineau, M.: An innovative approach for long-term environmental noise measurement: Rumeur network. In: 10th European Congress and Exposition on Noise Control Engineering (EuroNoise), Maastricht (2015)

61. Muzet, A., et al.: The need for a specific noise measurement for population exposed to aircraft noise during night-time. Noise Health **4**(15), 61 (2002)
62. Neitzel, R.L., Gershon, R.R., McAlexander, T.P., Magda, L.A., Pearson, J.M.: Exposures to transit and other sources of noise among New York City residents. Environ. Sci. Technol. **46**(1), 500–508 (2011)
63. Nelson, J.P.: Airports and property values: a survey of recent evidence. J. Transp. Econ. Policy **14**, 37–52 (1980)
64. Nelson, J.P.: Highway noise and property values: a survey of recent evidence. J. Trans. Econ. Policy **16**, 117–138 (1982)
65. New York City Department of Health and Mental Hygiene: Ambient Noise Disruption in New York City, Data brief 45. New York City Department of Health and Mental Hygiene, NY (2014)
66. NYC 311 Website. http://www1.nyc.gov/311/
67. Payne, S.R., Davies, W.J., Adams, M.D.: Research into the Practical and Policy Applications of Soundscape Concepts and Techniques in Urban Areas. DEFRA, HMSO, London (2009)
68. Piczak, K.J.: Environmental sound classification with convolutional neural networks. In: 25th International Workshop on Machine Learning for Signal Processing (MLSP), Boston, MA, pp. 1–6 (2015). doi:10.1109/MLSP.2015.7324337
69. Rabaoui, A., Davy, M., Rossignol, S., Ellouze, N.: Using one-class svms and wavelets for audio surveillance. IEEE Trans. Inf. Forensics Secur. **3**(4), 763–775 (2008)
70. Radhakrishnan, R., Divakaran, A., Smaragdis, P.: Audio analysis for surveillance applications. In: IEEE WASPAA'05, pp. 158–161 (2005). doi:10.1109/ASPAA.2005.1540194
71. Rakotomamonjy, A., Gasso, G.: Histogram of gradients of time-frequency representations for audio scene classification. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(1), 142–153 (2015). doi:10.1109/TASLP.2014.2375575
72. Rana, R.K., Chou, C.T., Kanhere, S.S., Bulusu, N., Hu, W.: Ear-phone: an end-to-end participatory urban noise mapping system. In: Proceedings of the 9th ACM/IEEE International Conference on Information Processing in Sensor Networks, pp. 105–116. ACM (2010)
73. Ruge, L., Altakrouri, B., Schrader, A.: Soundofthecity-continuous noise monitoring for a healthy city. In: 2013 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops), pp. 670–675. IEEE, New York (2013)
74. Salamon, J., Bello, J.P.: Feature learning with deep scattering for urban sound analysis. In: 2015 European Signal Processing Conference, Nice (2015)
75. Salamon, J., Bello, J.P.: Unsupervised feature learning for urban sound classification. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane (2015)
76. Salamon, J., Bello, J.P.: Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Process. Lett. **24**(3), 279–283 (2017)
77. Salamon, J., Jacoby, C., Bello, J.P.: A dataset and taxonomy for urban sound research. In: 22nd ACM International Conference on Multimedia (ACM-MM'14), Orlando, FL, pp. 1041–1044 (2014)
78. Salamon, J., Bello, J.P., Farnsworth, A., Kelling, S.: Fusing shallow and deep learning for bioacoustic bird species classification. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, pp. 141–145 (2017)
79. Santini, S., Ostermaier, B., Adelmann, R.: On the use of sensor nodes and mobile phones for the assessment of noise pollution levels in urban environments. In: 2009 6th International Conference on Networked Sensing Systems (INSS), pp. 1–8. IEEE, New York (2009)
80. Saxena, S., Brémond, F., Thonnat, M., Ma, R.: Crowd behavior recognition for video surveillance. In: International Conference on Advanced Concepts for Intelligent Vision Systems, pp. 970–981. Springer, Berlin, Heidelberg (2008)
81. Schweizer, I., Meurisch, C., Gedeon, J., Bärtl, R., Mühlhäuser, M.: Noisemap: multi-tier incentive mechanisms for participative urban sensing. In: Proceedings of the 3rd International Workshop on Sensing Applications on Mobile Phones, p. 9. ACM, New York (2012)

82. Serizel, R., Bisot, V., Essid, S., Richard, G.: Machine listening techniques as a complement to video image analysis in forensics. In: 2016 IEEE International Conference on Image Processing (ICIP), pp. 948–952. IEEE, New York (2016)
83. Simard, P.Y., Steinkraus, D., Platt, J.C.: Best practices for convolutional neural networks applied to visual document analysis. In: International Conference on Document Analysis and Recognition, vol. 3, Edinburgh, Scottland, pp. 958–962 (2003)
84. Smith, D., Ma, L., Ryan, N.: Acoustic environment as an indicator of social and physical context. Pers. Ubiquit. Comput. **10**(4), 241–254 (2006). doi:10.1007/s00779-005-0045-4. http://dx.doi.org/10.1007/s00779-005-0045-4
85. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
86. Stowell, D., Plumbley, M.D.: Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning. PeerJ **2**, e488 (2014). doi:10.7717/peerj.488. http://dx.doi.org/10.7717/peerj.488
87. Taber, R.: Technology for a quieter america, national academy of engineering. Technical report, NAEPR-06-01-A (2007)
88. Thrun, S., Bennewitz, M., Burgard, W., Cremers, A., Dellaert, F., Fox, D., Haehnel, D., Rosenberg, C., Roy, N., Schulte, J., et al.: Minerva: a second geration mobile tour-guide robot. In: IEEE International Conference on Robotics and Automation, pp. 3136–3141 (1999)
89. Valenzise, G., Gerosa, L., Tagliasacchi, M., Antonacci, F., Sarti, A.: Scream and gunshot detection and localization for audio-surveillance systems. In: IEEE Conference on Advanced Video and Signal Based Surveillance, 2007 (AVSS 2007), pp. 21–26 (2007)
90. Van Kempen, E., Babisch, W.: The quantitative relationship between road traffic noise and hypertension: a meta-analysis. J. Hypertens. **30**(6), 1075–1086 (2012)
91. Van Renterghem, T., Thomas, P., Dominguez, F., Dauwe, S., Touhafi, A., Dhoedt, B., Botteldooren, D.: On the ability of consumer electronics microphones for environmental noise monitoring. J. Environ. Monit. **13**(3), 544–552 (2011)
92. Wicke, L.: Die ökologischen Milliarden: das kostet die zerstörte Umwelt-so können wir sie retten. Kösel, Munich (1986)
93. Xu, M., Xu, C., Duan, L., Jin, J.S., Luo, S.: Audio keywords generation for sports video analysis. ACM Trans. Multimed. Comput. Commun. Appl. **4**(2), 1–23 (2008)
94. Yanco, H.A.: Wheelesley: a robotic wheelchair system: Indoor navigation and user interface. In: Assistive Technology and Artificial Intelligence, pp. 256–268. Springer, Berlin, Heidelberg (1998)
95. Yost, W.: Auditory image perception and analysis: the basis for hearing. Hear. Res. **56**(1), 8–18 (1991)
96. Zajdel, W., Krijnders, J., Andringa, T., Gavrila, D.: Cassandra: audio-video sensor fusion for aggression detection. In: IEEE Conference on Advanced Video and Signal Based Surveillance, 2007. AVSS 2007, pp. 200–205. IEEE, New York (2007)
97. Ziliani, F., Cavallaro, A.: Image analysis for video surveillance based on spatial regularization of a statistical model-based change detection. In: Proceedings of IEEE International Conference on Image Analysis and Processing, pp. 1108–1111. IEEE, New York (1999)