Chapter 1
Introduction to Sound Scene and Event Analysis

Tuomas Virtanen, Mark D. Plumbley, and Dan Ellis

Abstract Sounds carry a great deal of information about our environments, from individual physical events to sound scenes as a whole. In recent years several novel methods have been proposed to analyze this information automatically, and several new applications have emerged. This chapter introduces the basic concepts and research problems and engineering challenges in computational environmental sound analysis. We motivate the field by briefly describing various applications where the methods can be used. We discuss the commonalities and differences of environmental sound analysis to other major audio content analysis fields such as automatic speech recognition and music information retrieval. We discuss the main challenges in the field, and give a short historical perspective of the development of the field. We also shortly summarize the role of each chapter in the book.

Keywords Sound event detection • Sound scene classification • Sound tagging • Acoustic event detection • Acoustic scene classification • Audio content analysis

1.1 Motivation

Imagine you are standing on a street corner in a city. Close your eyes: what do you hear? Perhaps some cars and buses driving on the road, footsteps of people on the pavement, beeps from a pedestrian crossing, rustling, and clunks from shopping bags and boxes, and the hubbub of talking shoppers. Your sense of hearing tells you

T. Virtanen (✉)
Laboratory of Signal Processing, Tampere University of Technology, Tampere, Finland
e-mail: tuomas.virtanen@tut.fi

M.D. Plumbley
Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, Surrey GU2 7XH, UK
e-mail: m.plumbley@surrey.ac.uk

D. Ellis
Google Inc, 111 8th Ave, New York, NY 10027, USA
e-mail: dpwe@google.com

T. Virtanen et al. (eds.), *Computational Analysis of Sound Scenes and Events*,
DOI 10.1007/978-3-319-63450-0_1

what is happening around you, without even needing to open your eyes, and you could do the same in a kitchen as someone is making breakfast, or listening to a tennis match on the radio.

To most people, this skill of listening to everyday events and scenes is so natural that it is taken for granted. However, this is a very challenging task for computers; the creation of "machine listening" algorithms that can automatically recognize sounds events remains an open problem.

Automatic recognition of sound events and scenes would have major impact in a wide range of applications where sound or sound sensing is—or could be—involved. For example, acoustic monitoring would allow the recognition of physical events such as glass breaking (from somebody breaking into a house), a gunshot, or a car crash. In comparison to video monitoring, acoustic monitoring can be advantageous in many scenarios, since sounds travel through obstacles, is not affected by lighting conditions, and capturing sound typically consumes less power.

There exist also large amounts of multimedia material either broadcast, uploaded via social media, or in personal collections. Current indexing methods are mostly based on textual descriptions provided by contributors or users of such media collections. Such descriptions are slow to produce manually and often quite inaccurate. Methods that automatically produce descriptions of multimedia items could lead to new, more accurate search methods that are based on the content of the materials.

Computational sound analysis can also be used to endow mobile devices with context awareness. Devices such as smartphones, tablets, robots, and cars include microphones that can be used to capture audio, as well as possessing the computational capacity to analyze the signals captured. Through audio analysis, they can recognize and react to their environment. For example, if a car "hears" children yelling from behind a corner, it can slow down to avoid a possible accident. A smartphone could automatically change its ringtone to be most appropriate for a romantic dinner, or an evening in a noisy pub.

Recent activity in the scientific community such as the DCASE challenges and related workshops—including significant commercial participation—shows a growing interest in sound scene and event analysis technologies that are discussed in this book.

1.2 What is Computational Analysis of Sound Scenes and Events?

Broadly speaking, the term *sound event* refers to a specific sound produced by a distinct physical sound source, such as a car passing by, a bird singing, or a doorbell. Sound events have a single source, although as shown by the contrast between a car and its wheels and engine, defining what counts as a single source is still subjective. Sound events typically have a well-defined, brief, duration in time. By contrast,

the term *sound scene* refers to the entirety of sound that is formed when sounds from various sources, typically from a real scenario, combine to form a mixture. For example, the sound scene of a street can contain cars passing by, footsteps, people talking, etc. The sound scene in a home might contain music from radio, a dishwasher humming, and children yelling.

The overarching goal of computational analysis of sound scenes and events is extracting information from audio by computational methods. The type of information to be extracted depends on the application. However, we can sort typical sound analysis tasks into a few high-level categories. In *classification*, the goal is to categorize an audio recording into one of a set of (predefined) categories. For example, a sound scene classification system might classify audio as one of a set of categories including home, street, and office. In (event) *detection*, the goal is to locate in time the occurrences of a specific type of sound or sounds, either by finding each instance when the sound(s) happen or by finding all the temporal positions when the sound(s) are active. There are also other more specific tasks, such as estimating whether two audio recordings are from the same sound scene.

When the classes being recognized and/or detected have associated textual descriptions, the above techniques (classification and detection) can be used to construct a verbal description of an audio signal that is understandable by humans. The number of sound events or scene classes can be arbitrarily high and in principle it is possible to train classifiers or detectors for any type of sounds that might be present in an environment. In practice the number of classes or the types of sounds that can be classified is constrained by the availability of data that is used to train classifiers, and by the accuracy of the systems. The accuracy that can be achieved is affected by many factors, such as the similarity of classes to be distinguished from each other, the diversity of each class, external factors such as interfering noises, the quality and amount of training data, and the actual computational methods used.

The above vision of automatic systems producing abstract, textual descriptions is quite different from the mainstream research on computational analysis methods of a decade ago [21], where the main focus was on lower-level processing techniques such as source separation, dereverberation, and fundamental frequency estimation. Such low-level techniques are important building blocks in classification and detection systems, but they do not yet produce information that can be naturally interpreted by humans. The number of distinct sound classes handled by current classification and detection technologies is still limited, and their analysis accuracy is to be improved, but the capability of these methods to produce human-interpretable information gives them a significantly broader potential impact than more low-level processing techniques.

The core tasks of detection and classification require using several techniques related to audio signal processing and machine learning. For example, typical computational analysis systems first extract some acoustic features from the input signal, and supervised classifiers such as neural networks can be used for classification and detection. Therefore acoustic features and classifiers, as well as more complex statistical techniques for integrating evidence, and mechanisms for representing complex world knowledge, are all core tools in the computational analysis of sound scenes and events, and hence are covered in this book.

We refer to the domain of these sound analysis techniques as "everyday sounds," by which we mean combinations of sound sources of the number and complexity typically encountered in our daily lives. Some sound events may be quite rare (it is not every day that one encounters a snake hissing, at least for most of us), but when it does occur, it is more likely to be in the context of several other simultaneous sources than in isolation.

1.3 Related Fields

While computational analysis of non-speech, non-music sound scenes and events has only recently received widespread interest, work in analysis of speech and music signals has been around for some time. For speech signals, key tasks include recognizing the sequence of words in speech (automatic speech recognition), and recognizing the identity of the person talking (speaker recognition), or which of several people may be talking at different times (speaker diarization). For music audio, key tasks include recognizing the sequence of notes being played by one or more musical instruments (automatic music transcription), identifying the *genre* (style or category) of a musical piece (genre recognition), or identifying the instruments that are being played in a musical piece (instrument recognition): these music tasks are explored in the field of *music information retrieval* (MIR).

There are parallels between the tasks that we want to achieve for general everyday sounds, and these existing tasks. For example, the task of *sound scene classification* aims to assign a single label such as "restaurant" or "park" to an audio scene, and is related to the tasks of speaker recognition (for a speech signal with a single speaker) and musical genre recognition. Similarly, the task of *audio tagging*, which aims to assign a set of tags to a clip, perhaps naming audible objects, is related to the music task of instrument recognition in a multi-instrument musical piece. Perhaps most challenging, the task of *audio event detection*, which aims to identify the audio events—and their times—within an audio signal, is related to the speech tasks of automatic speech recognition and speaker diarization, as well as the task of automatic music transcription.

Since the analysis of everyday sounds can be related to speech and music tasks, it is not surprising to find that researchers have borrowed features and methods from speech and music, just as MIR researchers borrowed methods from the speech field. For example, features based on mel-frequency cepstral coefficients (MFCCs) [3], originally developed for speech, have also been used for MIR tasks such as genre recognition [20], and subsequently for sound scene recognition. Similarly, non-negative matrix factorization (NMF), which has been used for automatic music transcription, has also been applied to sound event recognition [4].

Nevertheless, there are differences between these domains that we should be aware of. Much of the classical work in speech recognition has focused on a single speaker, with a "source-filter" model that can separate excitation from the vocal tract: the cepstral transform at the heart of MFCCs follows directly from this

assumption, but although music and speech do not fit this model, MFCCs continue to be useful in these domains. Also, music signals often consist of sounds from instruments that have been designed to have a harmonic structure, and a particular set of "notes" (frequencies), tuned, for instance, to a western 12-semitone scale; everyday sounds will not have such carefully constructed properties. So, while existing work on speech and music can provide inspiration for everyday sound analysis, we must bear in mind that speech and music processing may not have all the answers we need.

Research on systematic classification of real-world sounds stretches back to the 1990s. One of the earliest systems was the SoundFisher of Wold et al. [22] which sought to provide similarity-based access to databases of isolated sound effects by representing each clip by a fixed-size feature vector comprising perceptual features such as loudness, pitch, and brightness. Other work grew out of the needs of the fragile speech recognizers of the time to avoid being fed non-speech signals such as music [18, 19], or to provide coarse segmentation of broadcast content [24]. The rise of cheap and ubiquitous recording devices led to interest in automatic analysis of unconstrained environmental recordings such as audio life-logs [5]. The growth of online media sharing sites such as YouTube poses enormous multimedia retrieval challenges which has fueled the current wave of interest in audio content information, including formal evaluations such as TRECVID [12, 16] which pose problems such as finding all videos relevant to "Birthday Party" or "Repairing an Appliance" among hundreds of thousands of items using both audio and visual information. While image features have proven most useful, incorporating audio features gives a consistent advantage, showing their complementary value.

Image content analysis provides an interesting comparison with the challenge of everyday sound analysis. For decades, computer vision struggled with making hard classifications of things like edges and regions even in relatively constrained images. But in the past few years, tasks such as ImageNet [17], a database of 1000 images for each of 1000 object categories, have seen dramatic jumps in performance, thanks to the development of very large "deep" neural network classifiers able to take advantage of huge training sets. We are now in an era when consumer photo services can reliably provide content-based search for a seemingly unlimited vocabulary of objects from "cake" to "sunset" within unconstrained collections of user-provided photos. This raises the question: Can we do the same thing with content-based search for specific sound events within unconstrained audio recordings?

1.4 Scientific and Technical Challenges in Computational Analysis of Sound Scenes and Events

In controlled laboratory conditions where the data used to develop computational sound scene and event analysis methods matches well with the test data, it is possible to achieve relatively high accuracies in the detection and classification of sounds

[2]. There also exist commercial products that can recognize certain specific sound categories in realistic environments [10]. However, current technologies are not able to recognize a large variety of different types of sounds in realistic environments. There are several challenges in computational sound analysis.

Many of these challenges are related to the acoustics of sound scenes and events. First, the acoustic characteristics of even a single class of sounds can be highly diverse. For example in the case of class "person yelling," the acoustics can vary enormously depending on the person who is yelling and the way in which they yell. Second, in realistic environments there can be many different types of sounds, some of whose acoustic characteristics may be very close to the target sounds. For example, the acoustics of a person yelling can be close to vocals in some background music that is present in many environments. Thirdly, an audio signal captured by a microphone is affected by the channel coupling (impulse response) between the source and microphone, which may alter the signal sufficiently to prevent matching of models developed to recognize the sound. Finally, in realistic environments there are almost always multiple sources producing sound simultaneously. The captured audio is a superposition of all the sources present, which again distorts the signal captured. In several applications of sound scene and event analysis, microphones that are used to capture audio are often significantly further away from target sources, which increases the effect of impulse responses from source to microphone as well as other sources in the environment. This situation is quite different from speech applications, where close-talk microphones are still predominantly used.

In addition to these complications related to the acoustics of sound scenes and events, there are also several fundamental challenges related to the development of computational methods. For example, if we are aiming at the development of methods able to classify and detect a large number of sounds, there is need for a taxonomy that defines the classes to be used. However, to date there is no established taxonomy for environmental sound events or scenes.

The computational methods used are heavily based on machine learning, where the parameters of a system are automatically obtained by using examples of the target (and non-target) sounds. In contrast to the situation in image classification, currently available datasets that can be used to develop computational scene and event scene analysis systems are more limited in size, diversity, and number of event instances, even though recent contributions such as AudioSet [6] have significantly reduced this gap.

1.5 About This Book

This book will provide a comprehensive description of the whole procedure for developing computational methods for sound scene and event analysis, ranging from data acquisition and labeling, designing the taxonomy used in the system, to signal processing methods for feature extraction and machine learning methods for sound recognition. The book will discuss commonalities as well as differences between

various analysis tasks, such as scene or event classification, detection, and tagging. It will also discuss advanced techniques that can take advantage of multiple microphones or other modalities. In addition to covering this kind of general methodology, the most important application domains, including multimedia information retrieval, bioacoustic scene analysis, smart homes, and smart cities, will also be covered. The book mainly focuses on presenting the computational algorithms and mathematical models behind the methods, and does not discuss specific software or hardware implementations (even though Chap. 13 discusses some possible hardware options). The methods present in the book are meant for the analysis of any everyday sounds in general. We will not discuss highly specific types of sounds such as speech or music, since analysis problems in their case are also more specific, and there already exist literature to address them [7, 13, 23].

The book is targeted for researchers, engineers, or graduate students in computer science and electrical engineering. We assume that readers will have basic knowledge of acoustics, signal processing, machine learning, linear algebra, and probability theory—although Chaps. 2 to 5 will give some background about some of the most important concepts. For those that are not yet familiar with the above topics, we recommend the following textbooks as sources of information: [9, 15], and [11] on signal processing, [14] on psychoacoustics, [1] on machine learning, and [8] on deep neural networks.

The book is divided into five parts. Part I presents the foundations of computational sound analysis systems. Chapter 2 introduces the supervised machine learning approach to sound scene and event analysis, which is the mainstream and typically the most efficient and generic approach in developing such systems. It will discuss the commonalities and differences between sound classification, detection, and tagging, and presents an example approach based on deep neural networks that can be used in all the above tasks.

Chapter 3 gives an overview of acoustics and human perception of sound events and scenes. When designing sound analysis systems it is important to have an understanding of the acoustic properties of target sounds, to support the development of the analysis methods. Knowledge about how the human auditory system processes everyday sounds is useful, and can be used to get ideas for the development of computational methods.

Part II of the book presents in detail the signal processing and machine learning methods as well as the data required for the development of computational sound analysis systems. Chapter 4 gives an overview of acoustic features that are used to represent audio signals analysis systems. Starting from representations of sound in general, it then moves from features based on signal processing towards learning features automatically from the data. The chapter also describes how to select relevant features for an analysis task, and how to temporally integrate and pool typical features extracted from short time frames.

Chapter 5 presents various pattern classification techniques that are used to map acoustic features to information about presence of each sound event or scene class. It first discusses basic concepts of supervised learning that are used in the development of such methods, and then discusses the most common discriminative and generative

classification models, including temporal modeling with hidden Markov models. The chapter also covers various models based on deep neural networks, which are currently popular in many analysis tasks. The chapter also discusses how the robustness of classifiers can be improved by various augmentation, domain adaptation, and ensemble methods.

Chapter 6 describes what kind of data—audio recordings and their annotations—are required in the development of sound analysis systems. It discusses possible ways of obtaining such material either from existing sources or by doing new recordings and annotations. It also discusses the procedures used to evaluate analysis systems as well as objective metrics used in such evaluations.

Part III of the book presents advanced topics related to categorization of sounds, analysis of complex scenes, and use of information from multiple sources. In the supervised learning approach for sound analysis which is the most typical and most powerful approach, some categorization of sounds is needed that will be used as the basis of the analysis. Chapter 7 presents various ways to categorize everyday sounds. It first discusses various theories of classification, and how new categorizations can be obtained. Then it discusses in more detail the categorization of everyday sounds, and their taxonomies and ontologies.

Chapter 8 presents approaches for the analysis of complex sound scenes consisting of multiple sound sources. It first presents a categorization of various sound analysis tasks, from scene classification to event detection, classification, and tagging. It discusses monophonic approaches that are able to estimate only one sound class at a time, as well as polyphonic approaches that enable analysis of multiple co-occurring sounds. It also discusses how contextual information can be used in sound scene and event analysis.

Chapter 9 presents multiview approaches, where data from multiple sensors are used in the analysis. These can include, for example, visual information or multiple microphones. The chapter first discusses general system architectures used in multiview analysis, and then presents how information can be fused at various system levels (features vs. classifier level). Then it discusses in detail two particularly interesting multiview cases for sound analysis: use of visual information in addition to audio and use of multiple microphones.

Part IV of the book covers selected computational sound scene and event analysis applications. Chapter 10 focuses on sound sharing and retrieval. It describes what kind of information (e.g., audio formats, licenses, metadata, features) should be taken into account when creating an audio database for this purpose. It then presents how sound retrieval can be done based on metadata, using freesound.org as an example. Finally, it presents how retrieval can be done using audio itself.

Chapter 11 presents the computational sound analysis approach to bioacoustic scene analysis. It first introduces the possible analysis tasks addressed in bioacoustics. Then it presents computational methods used in the field, including core methods such as segmentation, detection, and classification that share similarities to other fields, advanced methods such as source separation, measuring the similarity of sounds, analysis of sounds sequences, and methods for visualization and holistic soundscape analysis. The chapter also discusses how the methods can be employed at large scale, taking into account the computational complexity of the methods.

Chapter 12 focuses on sound event detection for smart home applications. It first discusses what kind of information sound can provide for these applications, and challenges such as the diversity of non-target sounds encountered and effect of audio channel. Then it discusses the user expectations of such systems, and how it affects the metrics that should be used in the development. Finally, it discusses the privacy and data protection issues of sound analysis systems.

Chapter 13 discusses the use of sound analysis in smart city applications. It first presents what kind of possibilities there are for computational sound analysis in applications such as surveillance and noise monitoring. It then discusses sound capture options based on mobile or static sensors, and the infrastructure of sound sensing networks. Then it presents various computational sound analysis results from studies focusing on urban sound environments.

Chapter 14 presents some future perspectives related to the research topic, for example, how to automatically obtain training data (both audio and labels) for the development of automatic systems. We also discuss how unlabeled data can be used in combination with active learning to improve classifiers and label data by querying users for labels. We discuss how weakly labeled data without temporal annotations can be used for developing sound event detection systems. The book concludes with a discussion of some potential future applications of the technologies.

Accompanying website of the book http://cassebook.github.io includes supplementary material and software implementations which facilitates practical interaction with the methods presented.

References

1. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, New York (2007)
2. Çakır, E., Parascandolo, G., Heittola, T., Huttunen, H., Virtanen, T.: Convolutional recurrent neural networks for polyphonic sound event detection. IEEE/ACM Transactions on Audio, Speech and Language Processing, **25**(6), (2017)
3. Davis, S.B., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process. **28**, 357–366 (1980)
4. Dikmen, O., Mesaros, A.: Sound event detection using non-negative dictionaries learned from annotated overlapping events. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) (2013)
5. Ellis, D.P., Lee, K.: Accessing minimal-impact personal audio archives. IEEE MultiMedia **13**(4), 30–38 (2006)
6. Gemmeke, J.F., Ellis, D.P.W., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: an ontology and human-labeled dataset for audio events. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (2017)
7. Gold, B., Morgan, N., Ellis, D.: Speech and Audio Signal Processing: Processing and Perception of Speech and Music. Wiley, New York (2011)
8. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge (2016)
9. Ifeachor, E., Jervis, B.: Digital Signal Processing: A Practical Approach, 2nd edn. Prentice Hall, Upper Saddle River (2011)

10. Krstulović, S., et al.: AudioAnalytic – Intelligent sound detection (2016). http://www.audioanalytic.com
11. Lyons, R.G.: Understanding Digital Signal Processing, 3rd edn. Pearson India, Harlow (2011)
12. Metze, F., Rawat, S., Wang, Y.: Improved audio features for large-scale multimedia event detection. In: Proceedings of IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6. IEEE, New York (2014)
13. Müller, M.: Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications. Springer, Cham (2015)
14. Moore, B.: An Introduction to the Psychology of Hearing, 6th edn. BRILL, Leiden (2013)
15. Oppenheim, A.V., Schafer, R.W.: Discrete-Time Signal Processing, 3rd edn. Pearson Education Limited, Harlow (2013)
16. Pancoast, S., Akbacak, M.: Bag-of-audio-words approach for multimedia event classification. In: Proceedings of Interspeech, pp. 2105–2108 (2012)
17. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
18. Saunders, J.: Real-time discrimination of broadcast speech/music. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 2, pp. 993–996. IEEE, New York (1996)
19. Scheirer, E., Slaney, M.: Construction and evaluation of a robust multifeature speech/music discriminator. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 2, pp. 1331–1334. IEEE, New York (1997)
20. Tzanetakis, G., Cook, P.: Musical genre classification of audio signals. IEEE Trans. Speech Audio Process. **10**(5), 293–302 (2002)
21. Wang, D., Brown, G.J.: Computational Auditory Scene Analysis. Wiley, Hoboken, NJ (2006)
22. Wold, E., Blum, T., Keislar, D., Wheaten, J.: Content-based classification, search, and retrieval of audio. IEEE MultiMedia **3**(3), 27–36 (1996)
23. Yu, D., Deng, L.: Automatic Speech Recognition: A Deep Learning Approach. Signals and Communication Technology. Springer, London (2014)
24. Zhang, T., Kuo, C.C.J.: Audio content analysis for online audiovisual data segmentation and classification. IEEE Trans. Speech Audio Process. **9**(4), 441–457 (2001)

http://www.springer.com/978-3-319-63449-4

Computational Analysis of Sound Scenes and Events
Virtanen, T.; Plumbley, M.D.; Ellis, D. (Eds.)
2018, X, 422 p. 81 illus., 54 illus. in color., Hardcover
ISBN: 978-3-319-63449-4