INVITED REVIEW

Introduction to acoustic event and scene analysis

Keisuke Imoto*

Ritsumeikan University, 1-1-1 Noji-higashi, Kusatsu, 525-8577 Japan

Abstract: Acoustic event and scene analysis has seen extensive development because it is valuable in applications such as monitoring of elderly people and infants, surveillance, life-logging, and advanced multimedia retrieval. This article reviews the basics of acoustic event and scene analysis, including its term and problem definitions, available public datasets, challenges, and recent research trends.

Keywords: Acoustic event analysis, Acoustic scene analysis, Abnormal event detection, Computational auditory scene analysis, Machine listening

PACS number: 43.60.+d [doi:10.1250/ast.39.182]

1. INTRODUCTION

Listen carefully to the sounds surrounding you. There are various sounds such as voices, music, footsteps, and the rumble of traffic. We use sound not only to communicate with others and enjoy listening to music but also to understand our surroundings: we use sound to sense danger, sense the presence of someone, and understand what someone is doing. Research about how humans understand the various sounds in their environment and on human-like listening systems for computers has been going on for the last several decades; however, for many of those years, the machine listening systems developed in those studies could deal with only a few kinds of sound, such as voices and music. In recent years, though, along with a spread of the sophisticated CPUs and GPUs and advances in machine learning theory, it has become possible to analyze any kind of sound, including voices, music, and environmental sounds, and extract useful information from it.

Researchers have attempted to use information extracted from various sounds to enrich various applications, such as monitoring systems for elderly people or infants [1,2], automatic surveillance systems [3–5], automatic life-logging systems [6,7], and advanced multimedia retrieval systems [8–10]. The techniques used in these applications include acoustic scene analysis (ASA) that analyzes scenes in terms of the places, situations, and user activities they depict and acoustic event analysis (AEA) that analyzes various sounds, such as footsteps, the rumble of traffic, and scream. This paper gives an introduction to acoustic event and scene analysis, by reviewing the terms and problem

definitions of acoustic event and scene analysis, the available open datasets, challenges to be solved, and recent research topics in this area.

2. WHAT IS ACOUSTIC EVENT AND SCENE ANALYSIS?

In this section, we review the technical terms and definitions used in acoustic event and scene analysis while referring to Plumley's [11] and Richard's [12] introductions to this research topic.

2.1. Technical Terms Used in Acoustic Event and Scene Analysis

Here, the terms used in acoustic event and scene analysis have been organized, as shown in Fig. 1. A **sound clip** (or audio clip) indicates a unit of analysis; in acoustic scene analysis, a clip has a length ranging from a few seconds to tens of seconds, while in acoustic event analysis, a clip typically lasts from a few milliseconds to several seconds. An **acoustic scene** denotes the label of the place where the sound was recorded (e.g., train, car, park, indoor), the situation (e.g., in a meeting, in an emergency), and the human activity involved (e.g., cooking, chatting, vacuuming). In many problems, acoustic scene labels are given to each sound clip. An **acoustic event** means a specific type of sound, such as footsteps, running water, exhaust fan noise, scream, or music. Many sound clips contain multiple acoustic events that overlap on the time axis.

2.2. Problem Definitions

2.2.1. Acoustic scene classification and acoustic event classification

Figure 2 shows the problem of **acoustic scene classification**. Acoustic scene classification is a task that takes as

*e-mail: keisuke.imoto@ieee.org

Fig. 1 Terms used in acoustic event and scene analysis.

Fig. 2 Problem of acoustic scene classification.

input a relatively long (several seconds to tens of seconds) sound clip of a single acoustic scene and outputs a single acoustic scene label representing the sound clip from a list of predefined acoustic scene classes. In acoustic scene classification, acoustic features are first extracted from a training dataset and an acoustic scene model is constructed using the acoustic features. After that, an acoustic feature extracted from a test sound clip is input to the acoustic scene classification system, and the system produces an acoustic scene label of the test sound clip.

Conventional approaches to acoustic scene classification use cepstrum coefficients, Mel-frequency cepstrum coefficients (MFCC), and MPEG-7 audio features [13] as acoustic features, and use Gaussian mixture models (GMM), hidden Markov models (HMM) and support vector machines (SVM) for constructing classification models [14–16]. More recently, deep-learning frameworks such as the convolutional neural network (CNN) [17,18] and recurrent neural network (RNN) as well as the Bayesian generative model of sound from an acoustic scene [19,20] have been applied to acoustic scene classification.

Acoustic event classification is a similar task to acoustic scene classification. In acoustic event classification, a

Fig. 3 Problem of acoustic sound detection.

relatively short (tens milliseconds to several seconds) sound clip including a single acoustic event is input and a single acoustic event label is output. However, because acoustic events often overlap temporally in actual situations, acoustic event classification is not a practical problem.

2.2.2. Acoustic event detection

Acoustic event detection, shown in Fig. 3, is a more practical problem setting. Acoustic event detection is a task in which a relatively long (several seconds to tens of seconds) sound clip including multiple acoustic events is input and acoustic event labels and their time stamps (start and end times) are output.

The whole process of acoustic event detection is similar to acoustic event classification, meanwhile, how to detect active durations of acoustic events is a more difficult problem than classifying acoustic events. For instance, Lee *et al.* proposed an acoustic event detection based on a convolutional neural network, which calculates a posterior for the existence of acoustic events time frame by time frame [21]. Ohishi *et al.* proposed a method combining non-negative matrix factorization (NMF) and a non-parametric Bayesian generative model, which detects the active durations of acoustic events by utilizing an activation matrix of NMF [10].

2.2.3. Abnormal sound detection

Abnormal sound detection (anomalous sound detection) is also being studied by many researchers. As shown in Fig. 4, abnormal sound detection is a task in which a relatively long (several seconds to tens of seconds) sound clip is input and abnormal sound labels and their time stamps (start and end times) are output. The difference between acoustic event detection and abnormal sound detection is in the size of the training data of “abnormal” sounds including “screaming,” “gun-shot sounds,” and “abnormal mechanical noise.” That is, we face this sort of problem when it is difficult to collect abnormal sounds to be detected. To detect abnormal sounds, some methods

Fig. 4 Problem definition of **abnormal event detection**.

use a criterion for how different input sounds are from “normal” sounds [22–24].

One-shot learning and few-shot learning [25] can be applied when we can collect a few samples of abnormal sounds. These techniques use normal sounds for training of common properties of abnormal and normal sounds; they construct an abnormal sound model by using a few abnormal sounds and the common properties.

2.2.4. Other problems of acoustic event and scene analysis

Some researchers have tried event and scene analysis using multimedia information including sounds, images, and text. For example, TRECVID MED (Media event detection)* [26] is a task in which a relatively long (several seconds to dozens of seconds) video clip including a single scene is input, and a single scene label representing the video clip is selected from a list of predefined scene classes and output. Some methods submitted to this competition utilize image and sound information simultaneously, and they outperform conventional methods that use only image information.

Much interest has been also expressed in bio-acoustic event detection, especially bird song detection [27].

3. WHICH DATASETS SHOULD WE USE FOR ACOUSTIC EVENT AND SCENE ANALYSIS?

Compared with speech or music, it is not easy to collect environmental sounds because there are numerous types of acoustic events and some of them occur infrequently. Therefore, having a publicly available sound dataset for acoustic event and scene analysis is very helpful when developing new methods for acoustic event and scene analysis. Moreover, it is important to use a common dataset and evaluation metric when comparing one method against another. In this section, public sound datasets that can be

used for our research on acoustic event and scene analysis are described.

3.1. RWCP Sound Scene Database in Real Acoustical Environments (RWCP-SSD) [28]

RWCP-SSD is a sound dataset comprised of dry sounds of 105 types of acoustic event such as “spray,” “clapping,” and “bell jingling.” It contains ten thousand samples of acoustic events. Nine impulse responses are also included, so that various sound recording conditions with reverberation can be simulated. Each sound clip ranges from 0.5 s to 2.0 s and includes a single acoustic event.

3.2. TUT Acoustic Scenes 2017/TUT Sound Events 2017 [29]

This dataset series includes i) TUT Acoustic Scenes 2017 for acoustic scene classification, ii) TUT Sound events 2017, and iii) TUT Rare Sound Event 2017 for acoustic event detection. All samples in TUT Acoustic Scenes 2017 and TUT Sound events 2017 were recorded with an in-ear stereo microphone. The dataset series were recorded and published for the Detection and Classification of Acoustic Scenes and Events (DCASE) 2017, which was a competition and workshop on computational acoustic scene and event analysis methods.

The TUT Acoustic Scenes 2017 dataset consists of sound clips including 15 classes of acoustic scenes such as “bus,” “park,” “home,” and “office.” For each acoustic scene, the development dataset contains 312 clips of 10 seconds, and the evaluation dataset contains 108 clips of 10 seconds. A fourfold cross-validation setup is provided, which enables us to readily develop systems with this dataset. This dataset was used for task 1 of the DCASE 2017 Challenge (Acoustic scene classification), and thus, we can compare the performance of a system under development with those of the systems submitted to the competition [30].

The TUT Sound events 2017 dataset consists of six classes of acoustic events, including “brakes squeaking,” “car,” “children,” “large vehicle,” “people speaking,” and “people walking.” It is a subset of TUT Acoustic Scenes 2017. Each sound clip runs 3 minutes to 5 minutes and includes multiple and temporally overlapping acoustic events. It has 659 acoustic events in the development sound clips and 272 acoustic events in the evaluation sound clips. Note that there is a difference of about seven times between the largest and smallest samples in terms of the acoustic events, so we need to use this imbalanced dataset with care. This dataset was used in task 3 of the DCASE 2017 Challenge (Sound event detection in real-life audio).

The TUT Rare Sound Event 2017 was used for task 2 of the DCASE 2017 Challenge (Detection of rare sound event). This dataset includes rare acoustic events of three

*TRECVID is a competition on automated understanding of video content held by the National Institute of Standards and Technology (NIST).

classes such as “baby crying,” “glass breaking,” and “gunshot,” as well as sound clips for background noise. Additionally, it provides a python script that generates mixture sounds of rare acoustic events and background noise with various signal-to-noise ratios. To make the dataset, sounds including three rare acoustic events are downloaded from Freesound [31], and the rare sounds are mixed with the background noise sound clips used for task 1 of DCASE 2016 Challenge (TUT Acoustic Scenes 2016 [32]). The resulting sound mixture includes a rare acoustic event at most once within 30 seconds.

3.3. AudioSet [33]

AudioSet is a sound dataset including sound clips of 623 classes of acoustic events such as “aircraft,” “fire alarm,” “harmonica,” “rain.” The classes of acoustic events are categorized based on an Is-a hierarchy in sound ontology [34], which is an abstraction level in an environmental sound. In total, this dataset has 2 million samples of acoustic events. Each sound clip is part of a ten-second video clip from Youtube™. A csv file including the URL of the Youtube™ video clip, the start and end times of the target acoustic event in the video clip, and acoustic event labels can be downloaded from the Web page of the AudioSet [35].

Part of this dataset was used for task 4 of the DCASE 2017 Challenge (Large-scale weakly supervised sound event detection for smart cars).

3.4. CHiME-Home [36]

The CHiME-Home dataset was originally recorded for the CHiME challenge [37], which aimed to develop speech separation and recognition methods for domestic environments. Afterwards, nine classes of acoustic event labels such as child speech, adult male speech, video game/TV, percussive sound, and unidentifiable sounds, were given to 6,100 sound clips, and these sound clips were used for task 4 of the DCASE 2016 Challenge (Domestic audio tagging). Each sound clip comprises a single acoustic event, four seconds in length. Isolated labels are given to “child speech,” “adult male speech,” and “adult female speech,” while acoustic events such as door knocks and footsteps are merged into events labeled “percussive sounds.” This dataset is aimed at the development of preprocessing methods for speech recognition.

3.5. Other Sound Datasets

UrbanSound and UrbanSound8K [38] are sound datasets recorded in an outdoor environment; they each consist of ten classes of acoustic events. They contain 1,302 and 8,732 labeled sound clips, where each clip may contain multiple and overlapping acoustic events. The Environmental sound classification (ESC) [39] is another sound

dataset for acoustic event analysis; it consists of 50 classes of acoustic events. The ESC contains 2,000 labeled short sound clips and 250,000 unlabeled sound clips; the clips were collected from Freesound [31]. There are also multi-channel sound datasets that were recorded in indoor environments, such as the DIRHA Simulated Corpus [40] and TU Dortmund Multi-channel Acoustic Event Dataset [41].

4. WHAT IS THE CHALLENGE OF ACOUSTIC EVENT AND SCENE ANALYSIS?

4.1. Difficulty of Constructing Sound Corpus for Acoustic Event and Scene Analysis

Acoustic event and scene analysis based on machine learning makes use of an environmental sound corpus, which consists of sound clips with content labels. Compared with speech or music, there are numerous types of acoustic event, and thus, it is not easy to record environmental sounds for acoustic event and scene analysis. Moreover, acoustic events often overlap each other and the sound level of a recorded environmental sound is often very small. Because of these difficulties, giving acoustic event labels and time stamps to environmental sounds is much laborious than, say, labeling speech signals.

Another difficulty in constructing a sound corpus is the hierarchy based on abstraction level for environmental sounds. For example, in giving an acoustic event label to the sound of a saxophone, we could, depending on the circumstance, choose from various labels such as “musical instrument,” “saxophone,” or even “alto saxophone.” This hierarchical property of environmental sound is called as the Is-a hierarchy in sound ontology [34]. Thus, when we annotate environmental sounds with acoustic event labels, we need to preliminarily determine which abstraction level is to be used for each acoustic event.

4.2. Noise in the Environment

In acoustic event and scene analysis, we often analyze sounds recorded in outdoor environments, often in the presence of noise. In such cases, recorded sounds may be unclear because of wind noise, rustling sounds, or saturation of the sound pressure level, and thus, we need to analyze sounds affected by such noise. Moreover, when we analyze acoustic events and scenes using cloud computation, packet loss in data transmission over the network might cause some parts to be completely lost.

In Sect. 5, we review recent trends (see Fig. 5) in research aimed at meeting these challenges.

5. WHAT IS THE TREND OF ACOUSTIC EVENT AND SCENE ANALYSIS?

5.1. Acoustic Event and Scene Analysis for Poor-Quality Sound Data

The public sound datasets described in Sect. 3 include

Fig. 5 Map of acoustic event and scene analyzing methods in terms of sound quality and amount of training data.

relatively high quality sound clips. Meanwhile, in practical situations in which acoustic event and scene analysis would be deployed, recorded sounds often have intermittently unreliable parts because of wind noise, rustling sounds, or saturation of the sound pressure level. Thus, we need to analyze acoustic events and scenes affected by such noise. The conventional methods of acoustic event detection in noisy environments use noise suppression techniques as preprocessings for speech [42,43]. However, conventional noise suppression techniques focus only on the statistical properties of voices and background noise and are sometimes ineffective for acoustic event and scene analysis. Thus, we need to develop a noise suppression method especially for acoustic event and scene analysis.

If a client server system is used in acoustic event and scene analysis, data may get lost when it is transmitted over the network. In addition, continuous recording threatens privacy in some cases, so we may have to make do with incomplete recordings [44]. As an acoustic scene analysis for these situations, Imoto *et al.* have proposed an acoustic topic model (ATM) that considers temporal transitions of acoustic events; their method can analyze acoustic scenes and restore missing observations at the same time [45].

5.2. Acoustic Event and Scene Analysis Utilizing Spatial Information

Many methods of acoustic event and scene analysis utilize spectral information such as spectrum, cepstrum, and Mel-frequency cepstral coefficients (MFCCs) as acoustic features. The sounds of a window breaking and a glass breaking sound from a TV have very different meanings; thus, spatial information is also important in acoustic event and scene analysis. Kwon *et al.* [46], and Giannoulis *et al.* [47] proposed acoustic event analyses that use spatial information extracted with a synchronized microphone array.

Acoustic sensors on smartphones, wearable devices, and IoT devices are everywhere these days. Combining

such sensors into an array is a promising way to extract spatial information. However, such microphones are usually not accurately synchronized and their positions and geometry cannot be used for acoustic event and scene analysis. This makes it hard to apply the conventional sort of microphone array processing to them. For extracting spatial information using an unsynchronized and distributed microphone array, Kürby *et al.* [41] and Phan *et al.* [48] devised scene classification methods based on late fusion of scene classification results in each microphone. Imoto *et al.* [49–51] proposed a spatial cepstrum that can extract spatial information robustly when microphones are not synchronized.

5.3. Acoustic Event and Scene Analysis for Small-scale Sound Dataset

Recent machine learning techniques, including deep-learning frameworks, require a large-scale sound dataset to achieve high performance. However, since a lot of sound clips cannot always be collected, developing acoustic event and scene analyzing methods using a small-scale sound dataset is still an important problem. To analyze acoustic events and scenes from a small sound dataset, Bisot *et al.* [52] and Komatsu *et al.* [53] proposed methods based on non-negative matrix factorization (NMF). Kim *et al.* [19] and Imoto *et al.* [54] proposed acoustic scene analyzing methods based on acoustic topic models, which are Bayesian generative models of acoustic events from acoustic scenes. These methods focus on the sparsity of sound structures and extract acoustic features of lower dimension to train an acoustic event or scene model with a small number of sound clips.

5.4. Automatic Data Annotation and Data Augmentation for Acoustic Event and Scene Analysis

Some researchers have tried to construct acoustic event corpora automatically. One such a technique is acoustic event and scene analysis using weakly labeled data, which can reduce labor of annotating large numbers of samples [55–57]. As shown in Fig. 6, acoustic event detection using weakly labeled data is a task in which a relatively long (several seconds to tens of seconds) sound clip and acoustic event labels are input, and acoustic event labels and their time stamps (strong labels) are output. The time stamps are not manually assigned, and the resulting strong labels can be used in other acoustic event detection methods. Task 4 of the DCASE 2017 Challenge (Large-scale weakly supervised sound event detection for smart cars) was this sort of task, and many algorithms were submitted to it [29].

Data augmentation, which composes new data artificially from existing sound data, has also been investigated. For example, Takahashi *et al.* [58] and Mun *et al.* [59] proposed data augmentation methods based on machine

Fig. 6 Problem of acoustic event detection using weakly labeled data.

learning techniques such as the convolutional neural network and the generative adversarial network (GAN).

6. CONCLUSION

In this article, we reviewed acoustic event and scene analysis, which analyzes various sounds and extracts useful information such as types of acoustic events, locations where sounds were recorded, and situations. In particular, we defined the terms and problems of acoustic event and scene analysis, reviewed the public datasets that are available for this sort of analysis, and recent research trends.

Acoustic event and scene analysis will be very valuable in applications such as monitoring of elderly people and infants, surveillance, life-logging, and advanced multimedia retrieval. This topic has received a lot of attention from researchers for several decades. However, many challenges remain to be met. I hope this review will help readers address them.

REFERENCES

- [1] Y. Peng, C. Lin, M. Sun and K. Tsai, "Healthcare audio event classification using hidden Markov models and hierarchical hidden Markov models," *Proc. ICME*, pp. 1218–1221 (2012).
- [2] P. Guyot, J. Pinquier and R. André-Obrecht, "Water sound recognition based on physical models," *Proc. ICASSP 2013*, pp. 793–797 (2013).
- [3] A. Harma, M. F. McKinney and J. Skowronek, "Automatic surveillance of the acoustic activity in our living environment," *Proc. ICME* (2005).
- [4] R. Radhakrishnan, A. Divakaran and P. Smaragdis, "Audio analysis for surveillance applications," *Proc. WASPAA*, pp. 158–161 (2005).
- [5] S. Ntalampiras, I. Potamitis and N. Fakotakis, "On acoustic surveillance of hazardous situations," *Proc. ICASSP 2009*, pp. 165–168 (2009).
- [6] A. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho and J. Huopaniemi, "Audio-based context recognition," *IEEE Trans. Audio Speech Lang. Process.*, **14**, 321–329 (2006).
- [7] K. Imoto, S. Shimauchi, H. Uematsu and H. Ohmuro, "User activity estimation method based on probabilistic generative model of acoustic event sequence with user activity and its subordinate categories," *Proc. Interspeech 2013*, pp. 2609–2613 (2013).
- [8] T. Zhang and C. J. Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *IEEE Trans. Audio Speech Lang. Process.*, **9**, 441–457 (2001).
- [9] Q. Jin, P. F. Schulam, S. Rawat, S. Burger, D. Ding and F. Metze, "Event-based video retrieval using audio," *Proc. Interspeech 2012*, pp. 2085–2088 (2012).
- [10] Y. Ohishi, D. Mochihashi, T. Matsui, M. Nakano, H. Kameoka, T. Izumitani and K. Kashino, "Bayesian semi-supervised audio event transcription based on Markov Indian buffet process," *Proc. ICASSP 2013*, pp. 3163–3167 (2013).
- [11] M. D. Plumbley and T. Virtanen, "DCASE challenge: Philosophy, tasks and results," *Proc. DCASE* (2016).
- [12] G. Richard, "Acoustic scene and events recognition: How similar is it to speech recognition and music genre recognition?" *Proc. DCASE* (2016).
- [13] H. G. Kim, N. Moreau and T. Sikora, *MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval* (John Wiley & Sons, Hoboken, NJ, 2006).
- [14] A. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho and J. Huopaniemi, "Audio-based context recognition," *IEEE Trans. Audio Speech Lang. Process.*, **14**, 321–329 (2006).
- [15] J. T. Geiger, B. Schuller and R. Gerhard, "Large-scale audio feature extraction and SVM for acoustic scene classification," *Proc. WASPAA* (2013).
- [16] M. Chum, A. Habshush, A. Rahman and C. Sang, "IEEE AASP scene classification challenge using hidden Markov models and frame based classification," *Proc. WASPAA* (2013).
- [17] M. Valenti, A. Diment, G. Parascandolo, S. Squartini and T. Virtanen, "DCASE 2016 acoustic scene classification using convolutional neural networks," *Proc. DCASE*, pp. 95–99 (2016).
- [18] S. H. Bae, I. Choi and N. S. Kim, "Acoustic scene classification using parallel combination of LSTM and CNN," *Proc. DCASE*, pp. 11–15 (2016).
- [19] S. Kim, S. Narayanan and S. Sundaram, "Acoustic topic models for audio information retrieval," *Proc. WASPAA*, pp. 37–40 (2009).
- [20] K. Imoto, Y. Ohishi, H. Uematsu and H. Ohmuro, "Acoustic scene analysis based on latent acoustic topic and event allocation," *Proc. MLSP* (2013).
- [21] D. Lee, S. Lee, Y. Han and K. Lee, "Ensemble of convolutional neural networks for weakly-supervised sound event detection using multiple scale input," *Proc. DCASE*, pp. 74–79 (2016).
- [22] E. Marchi, F. Vesperini, F. Eyben, S. Squartini and B. Schuller, "A novel approach for automatic acoustic novelty detection using a denoising autoencoder with bidirectional LSTM neural networks," *Proc. ICASSP 2015*, pp. 1996–2000 (2015).
- [23] Y. Koizumi, S. Saito, H. Uematsu and N. Harada, "Optimizing acoustic feature extractor for anomalous sound detection based on Neyman-Pearson lemma," *Proc. EUSIPCO*, pp. 698–702 (2017).
- [24] Y. Kawaguchi and T. Endo, "How can we detect anomalies from subsampled audio signals?" *Proc. MLSP* (2017).
- [25] B. M. Lake, R. R. Salakhutdinov and J. Tenenbaum, "One-shot learning by inverting a compositional causal process," *Proc. NIPS*, pp. 2526–2534 (2013).

- [26] G. Awad, A. Butt, J. Fiscus, D. Joy, A. Delgado, M. Michel, A. F. Smeaton, Y. Graham, W. Kraaij, G. Quénot, M. Eskevich, R. Ordelman, G. J. F. Jones and B. Huet, "TRECVID 2017: Evaluating ad-hoc and instance video search, events detection, video captioning and hyperlinking," *Proc. TRECVID* (2017).
- [27] D. Stowell, "Computational bioacoustic scene analysis," in *Computational Analysis of Sound Scenes and Events* (Springer, Berlin/Heidelberg, 2017), pp. 303–333.
- [28] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," *Proc. LREC*, pp. 965–968 (2000).
- [29] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," *Proc. DCASE*, pp. 85–92 (2017).
- [30] <http://www.cs.tut.fi/sgn/arg/dcase2017/challenge/> (accessed 2018-02-17).
- [31] <https://freesound.org/> (accessed 2018-02-17).
- [32] A. Mesaros, T. Heittola and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," *Proc. EUSIPCO*, pp. 1128–1132 (2016).
- [33] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," *Proc. ICASSP 2017*, pp. 776–780 (2017).
- [34] T. Nakatani and H. G. Okuno, "Sound ontology for computational auditory scene analysis," *Proc. AAAI/IAAI*, pp. 1004–1010 (1998).
- [35] <https://research.google.com/audioset/download.html> (accessed 2018-02-17).
- [36] P. Foster, S. Sigtia, S. Krstulovic, J. Barker and M. D. Plumbley, "CHiME-Home: A dataset for sound source recognition in a domestic environment," *Proc. WASPAA*, pp. 1–5 (2015).
- [37] H. Christensen, J. Barker, N. Ma and P. D. Green, "The CHiME corpus: A resource and a challenge for computational hearing in multisource environments," *Proc. Interspeech 2010*, pp. 1918–1921 (2010).
- [38] J. Salamon, C. Jacoby and J. P. Bello, "A dataset and taxonomy for urban sound research," *Proc. ACM MM*, pp. 1041–1044 (2014).
- [39] K. J. Piczak, "ESC: Dataset for environmental sound classification," *Proc. ACM MM*, pp. 1015–1018 (2015).
- [40] E. Zwyssig, M. Ravanelli, P. Svaizer and M. Omologo, "A multi-channel corpus for distant-speech interaction in presence of known interferences," *Proc. ICASSP 2015*, pp. 4480–4484 (2015).
- [41] J. Kürby, R. Grzeszick, A. Plinge and G. A. Fink, "Bag-of-features acoustic event detection for sensor networks," *Proc. DCASE*, pp. 55–59 (2016).
- [42] N. Waldo, R. Gerard and H. Perfecto, "Automatic event classification using front end single channel noise reduction, MFCC features and a support vector machine classifier," *Proc. DCASE* (2013).
- [43] J. Schröder, B. Cauchi, M. R. Schädler, N. Moritz, K. Adiloglu, J. Anemüller, S. Doclo, B. Kollmeier and S. Goetze, "Acoustic event detection using signal enhancement and spectro-temporal feature extraction," *Proc. DCASE* (2013).
- [44] S. Krstulović, "Audio event recognition in the smart home," in *Computational Analysis of Sound Scenes and Events* (Springer, Berlin/Heidelberg, 2017), pp. 335–371.
- [45] K. Imoto and N. Ono, "Acoustic scene analysis from acoustic event sequence with intermittent missing event," *Proc. ICASSP 2015*, pp. 156–160 (2015).
- [46] H. Kwon, H. Krishnamoorthi, V. Berisha and A. Spanias, "A sensor network for real-time acoustic scene analysis," *Proc. ISCS*, pp. 169–172 (2009).
- [47] P. Giannoulis, A. Brutti, M. Matassoni, A. Abad, A. Katsamanis, M. Matos, G. Potamianos and P. Maragos, "Multi-room speech activity detection using a distributed microphone network in domestic environments," *Proc. EUSIPCO*, pp. 1271–1275 (2015).
- [48] H. Phan, M. Maass, L. Hertel, R. Mazur and A. Mertins, "A multi-channel fusion framework for audio event detection," *Proc. WASPAA*, pp. 1–5 (2015).
- [49] K. Imoto and N. Ono, "Spatial cepstrum as a spatial feature using distributed microphone array for acoustic scene analysis," *IEEE/ACM Trans. Audio Speech Lang. Process.*, **25**, 1335–1343 (2017).
- [50] K. Imoto and N. Ono, "Acoustic scene classification based on generative model of acoustic spatial words for distributed microphone array," *Proc. EUSIPCO*, pp. 2343–2347 (2017).
- [51] K. Imoto and N. Ono, "Acoustic scene classification using asynchronous multichannel observations with different lengths," *Proc. MMSP*, pp. 1–5 (2017).
- [52] V. Bisot, R. Serizel, S. Essid and G. Richard, "Acoustic scene classification with matrix factorization for unsupervised feature learning," *Proc. ICASSP 2016*, pp. 6445–6449 (2016).
- [53] T. Komatsu, Y. Senda and R. Kondo, "Acoustic event detection based on non-negative matrix factorization with mixtures of local dictionaries and activation aggregation," *Proc. ICASSP 2016*, pp. 2259–2263 (2016).
- [54] K. Imoto and S. Shimauchi, "Acoustic scene analysis based on hierarchical generative model of acoustic event sequence," *IEICE Trans. Inf. Syst.*, **E99-D**, 2539–2549 (2016).
- [55] A. Kumar and B. Raj, "Audio event detection using weakly labeled data," *Proc. ACM MM*, pp. 1038–1047 (2016).
- [56] Q. Kong, Y. Xu, W. Wang and M. D. Plumbley, "A joint detection-classification model for audio tagging of weakly labelled data," *Proc. ICASSP 2017*, pp. 641–645 (2017).
- [57] S. Adavanne and T. Virtanen, "Sound event detection using weakly labeled dataset with stacked convolutional and recurrent neural network," *Proc. DCASE*, pp. 12–16 (2017).
- [58] N. Takahashi, M. Gygli, B. Pfister and L. V. Gool, "Deep convolutional neural networks and data augmentation for acoustic event recognition," *Proc. Interspeech*, pp. 2982–2986 (2016).
- [59] S. Mun, S. Park, D. K. Han and H. Ko, "Generative adversarial network based acoustic scene training set augmentation and selection using SVM hyper-plane," *Proc. DCASE*, pp. 93–97 (2017).

Keisuke Imoto received the B.E. and M.E. degrees from Kyoto University in 2008 and 2010, respectively. He received the Ph.D. degree from SOKENDAI (The Graduate University for Advanced Studies) in 2017. He joined the Nippon Telegraph and Telephone Corporation (NTT) in 2010. He joined to Ritsumeikan University as an Assistant Professor in 2017. He has been engaged in research on acoustic event detection, acoustic scene analysis, and microphone array signal processing. He is a member of the IEEE Signal Processing Society and the Acoustical Society of Japan.