RU MULTICHANNEL DOMESTIC ACOUSTIC SCENES 2019: A MULTICHANNEL DATASET RECORDED BY DISTRIBUTED MICROPHONES WITH VARIOUS PROPERTIES

Keisuke Imoto

Ritsumeikan University, Japan
keisuke.imoto@ieee.org

Nobutaka Ono

Tokyo Metropolitan University, Japan
onono@tmu.ac.jp

ABSTRACT

Acoustic scene analysis has seen extensive development recently because it is used in applications such as monitoring, surveillance, life-logging, and advanced multimedia retrieval systems. Acoustic sensors, such as those used in smartphones, wearable devices, and surveillance cameras, have recently rapidly increased in number. The simultaneous use of these acoustic sensors will enable a more reliable analysis of acoustic scenes because they can be utilized for the extraction of spatial information or application of ensemble techniques. However, there are only a few datasets for acoustic scene analysis that make use of multichannel acoustic sensors, and to the best of our knowledge, no large-scale open datasets recorded with multichannel acoustic sensors composed of different devices. In this paper, we thus introduce a new publicly available dataset for acoustic scene analysis, which was recorded by distributed microphones with various characteristics. The dataset is freely available from `https://www.ksuke.net/dataset`.

Index Terms— Distributed microphone array, acoustic scene classification, publicly available dataset

1. INTRODUCTION

Acoustic scene classification (ASC), which associates a sound with a related scene, has recently attracted much attention because of its many useful applications such as those in monitoring systems for elderly people or infants [1, 2], automatic surveillance systems [3, 4, 5, 6], automatic life-logging systems [7, 8, 9], and advanced multimedia retrieval [10, 11, 12, 13].

Many approaches to ASC are based on machine learning techniques, especially deep neural network (DNN)-based methods [14, 15, 16, 17, 18, 19, 20, 21]. For instance, Valenti *et al.* have proposed a method based on convolutional neural networks (CNNs) [17], which allows robust feature extraction of acoustic scenes against time and frequency shifts in the spectrogram domain. More sophisticated models such as VGG [22], ResNet [23], and Xception [24], which achieve reasonable performance in image recognition, have also been applied to acoustic scene analysis [18, 19, 20]. Ren *et al.* have applied the attention mechanism to CNN-based acoustic scene classification [21]. These DNN-based approaches for ASC require a large-scale dataset; thus, the large-scale datasets that are publicly available have contributed to related research and development. Moreover, evaluation using a publicly available dataset is an impartial means of assessing a method under development. There are some open datasets for ASC, such as the LITIS dataset [25], TUT Acoustic Scenes 2016 [26] and 2017 [27], and TUT Urban Acoustic Scenes 2018 [28], which were recorded with a single or stereo microphone(s). There are also other publicly available datasets for detecting sound events that occur in a domestic environment, such as the CHiME-Home dataset [29].

On the other hand, acoustic sensors that are easily accessible, such as those in smartphones, smart speakers, IoT devices, and surveillance cameras, have rapidly increased in number. By making use of these microphones simultaneously, we obtain spatial information, which will help to recognize acoustic scenes [30, 31, 32, 33]. For instance, an acoustic scene "cooking" and related sounds tend to occur in a kitchen, whereas an acoustic scene "shaving" and related sounds are likely to occur in a powder room. There are also datasets for ASC or sound event classification based on multichannel observation, such as ITC-Irst AED Database [34], FINCA Multi-channel Acoustic Event Dataset [35], and SINS Database [36]. For example, ITC-Irst AED Database consists of sound recordings including 16 types of acoustic events, such as "door knock," "cough," and "keyboard." Eight T-shaped microphone arrays, each of which had four microphones, were used for the recording. SINS Database consists of sound recordings including 16 different activities in the home, such as "cooking," "vacuuming," and "phone call." The recording was conducted using 13 microphone arrays, all of which were composed of four Sonion N8AC03 MEMS microphones.

Considering that a large microphone array is constructed by combining microphones that are mounted on smartphones, smart speakers, IoT devices, and surveillance cameras, some microphones often have a mismatch under acoustic conditions, such as the sampling rate, frequency response, sensitivity, and/or noise level. This condition mismatch often has a detrimental effect on the classification performance of acoustic scenes and needs to be addressed. However, there are no open datasets for ASC that were recorded in a home environment using multiple microphones with various properties. In this paper, we thus introduce a dataset for ASC named Ritsumeikan University (RU) Multichannel Domestic Acoustic Scenes 2019, which was recorded by distributed microphones with various properties. The characteristics of RU Multichannel Domestic Acoustic Scenes 2019 are as follows:

- The dataset consists of 21 kinds of acoustic scenes including an "absent" scene and high-privacy scenes such as "toilet," "sleeping," and "taking a bath/shower."
- The dataset was recorded using 42 distributed microphones with various characteristics.
- The dataset consists of a total of 1,995.8 h of sounds (47.5 h $\times$ 42 ch.), which can be divided into about 11,400 segments of 15 s sounds for each channel.
- The dataset can be utilized for evaluating ASC methods using spatial information, ensemble techniques, or domain adapta-

https://doi.org/10.33682/hy6p-g607

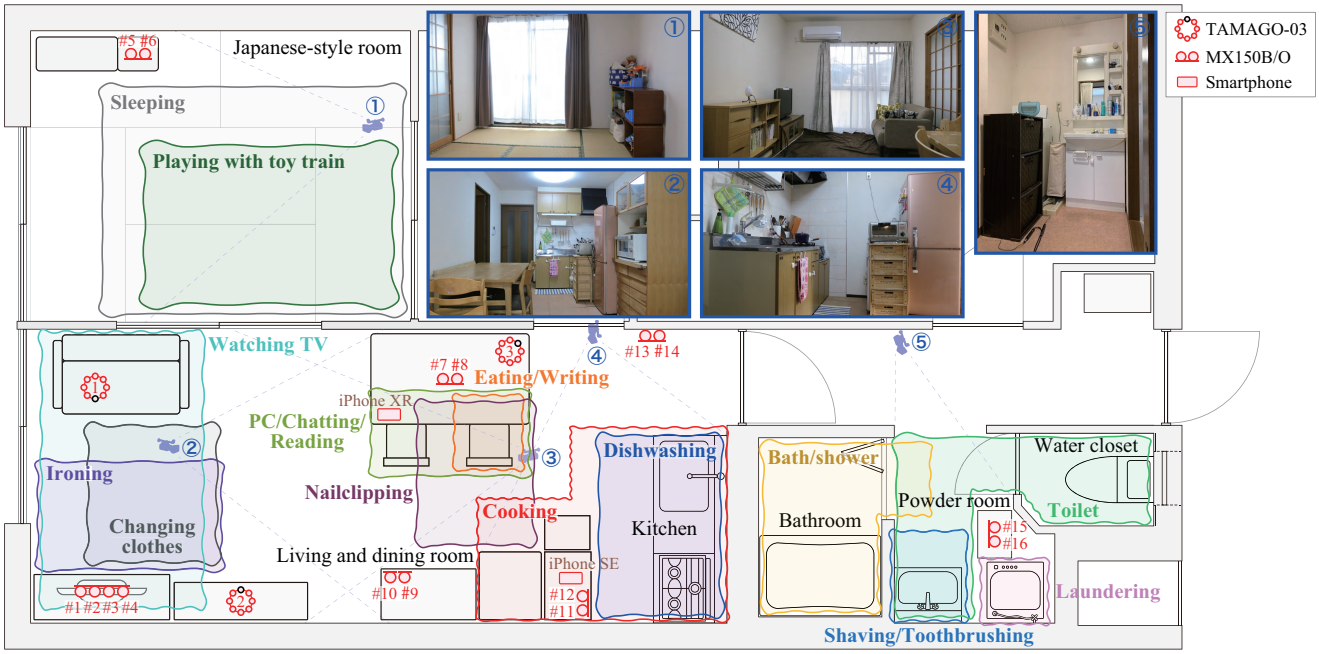


Figure 1: Floor plan of recording environment, microphone arrangement, and approximate positions of sound sources

tion techniques (by combining this and another multichannel dataset such as SINS Database [36]).

- The dataset includes sample videos of most of the sound clips (for understanding of recording environments and situations)

This dataset is freely available and can be downloaded at [37].

The remainder of this paper is structured as follows. In section 2, we provide an overview of RU Multichannel Domestic Acoustic Scenes 2019. In section 3, the benchmark evaluation results are reported. Finally, a conclusion is given in section 4.

2. OVERVIEW OF RU MULTICHANNEL DOMESTIC ACOUSTIC SCENES 2019

2.1. Recording conditions

The dataset was recorded in an apartment where people actually live. As shown in Fig. 1, the recording was conducted in six different rooms: a Japanese-style room (washitsu), hall, powder room, bathroom, water closet, and a combined living room, dining room, and kitchen. As the recording equipment, three TAMAGO-03 microphone arrays [38] (8ch × 3), 16 Shure MX150B/O microphones (1ch × 16), one iPhone SE (1ch × 1), and one iPhone XR (1ch × 1) were used. Each TAMAGO-03 array consisted of eight microphones mounted on a circle of a 36.5 mm radius at 45° intervals, as shown in Fig. 2-(a). The sampling rate and bit depth of the TAMAGO-03 microphones were 16 kHz and 16, respectively. The Shure MX150B/O microphones were arranged in pairs with 50.0 mm intervals. As the microphone amplifier and AD converter for the MX150B/O microphones, we used two MOTU 8Ms [39]. The sampling rate and bit depth of the MX150B/O microphones [40], iPhone XR, and iPhone SE were 48 kHz and 16, respectively. The microphones were synchronized between microphones in each TAMAGO-03 array and 16ch MX150B/O microphones, respectively, but not between different devices. The recording conditions are given in detail in [37].

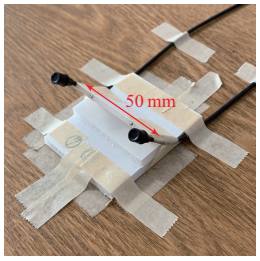
Table 1: Recorded acoustic scenes and their durations

Acoustic scene	# clips	Duration (min)
Absent	26	125.3
Changing clothes	67	119.8
Chatting	23	121.5
Cooking	14	228.0
Dishwashing	36	122.8
Eating	24	129.3
Ironing	25	129.6
Laundering	10	138.0
Moving	30	122.0
Nail clipping	37	121.1
Operating PC	22	123.3
Playing with toys	21	127.5
Reading newspaper/magazine	25	121.5
Shaving	59	146.5
Sleeping	23	144.0
Taking a bath/shower	18	181.5
Toilet	101	134.6
Toothbrushing	42	132.5
Vacuuming	29	122.8
Watching TV	28	128.4
Writing	18	131.2

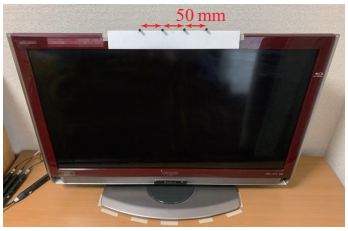
2.2. Recorded acoustic scenes and recording procedure

We recorded 21 acoustic scenes that frequently occur in daily activities at home. Table 1 lists the recorded acoustic scenes, which include “absent” and high-privacy scenes such as “toilet,” “chang-

(a) TAMAGO-03



(b) MX150B/O



(c) MX150B/O (mounted on TV)

Figure 2: Detailed microphone arrangements

ing clothes," "taking a bath/shower," and "sleeping." Each sound clip includes all the sounds derived from a series of actions in one scene, for instance, a sound clip of "toothbrushing" includes sounds derived from "picking up toothbrush," "putting toothpaste on toothbrush," "brushing teeth," and "rinsing mouth." The approximate position of the sound source in each acoustic scene is also shown in Fig. 1, except for the acoustic scenes "absent," "moving," and "vacuuming," in which the sound may occur over the entire apartment. Each recording was started with a cue, which was an impulsive sound, but detailed scenarios and recording times were not directed.

To ensure the diversity of recorded sounds, we used various household commodities and electronic devices such as four different kitchen sponges, two irons, three nail clippers, three PCs, four computer mouses, three electric shavers, five toothbrushes, and two vacuum cleaners. Figure 3 shows these household commodities and electronic devices.

2.3. Postprocessing

Since the microphones were not synchronized between different devices, after recordings, we simply synchronized the sound clips using the cross-correlation between the nearest microphone pair. The procedure for the synchronization and reshaping of recorded signals is shown in Fig. 4. We first selected the nearest microphone pair from the unsynchronized microphones, and we then synchronized the acoustic signals recorded by the microphone pair using the cross-correlation all over the signals. Since the sampling rates of the TAMAGO-03 microphones and the other microphones were 16 kHz and 48 kHz, respectively, the recorded sound at 48 kHz was downsampled to 16 kHz when synchronizing. After that, we cut the acoustic signals to remove cue sounds, which are irrelevant to recorded scenes. Note that we did not take an arrival time difference of sounds between channels, which is a significant cue for extracting spatial information, into account; thus, sound clips needs to be resynchronized accurately using blind compensation techniques for distributed microphone array [41, 42] if we extract spatial information using conventional methods of microphone array processing.

Figure 3: Household commodities and electronic devices used for recording

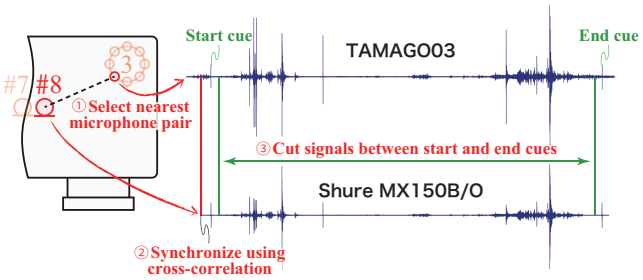


Figure 4: Synchronization procedure between unsynchronized microphones

Moreover, the different devices have sampling frequency mismatch; however we did not compensate the mismatch between devices.

Although the length of the sound differs from sound clip to sound clip, we suppose that each sound clip will be divided into 10 or 15 s segments, which are the units of analysis. A manipulation tool that divides each sound clip into shorter segments is also included in the dataset.

2.4. Contents of RU Multichannel Domestic Acoustic Scenes 2019

RU Multichannel Domestic Acoustic Scenes 2019 includes the following contents:

- Sound files in wav format (RIFF waveform audio format)
- Impulse responses at each microphone position (RIFF waveform audio format)
- Documents of recording conditions and postprocessing procedures
- Sample videos (for understanding of recording environments and situations)
- Tools for manipulating sound files

Each sound file is stored in the wave format, and 42-channel sound files obtained in each recording are stored in one directory. The dataset also contains impulse responses from some sound source

Table 2: Experimental conditions

# total microphones	42
Sound clip length	15 s
Frame length	40 ms
Frame shift	20 ms
Network structure	3 conv. & 3 FC layers
Pooling in CNN layers	3×3 max pooling
Activation function	ReLU, softmax (output layer)
# channels of CNN	42, 32, 16
# units of FC layers	128, 64, 32
Dropout ratio in FC layer	0.5
# epoch	150

locations to all microphone positions. Documents providing the details of recording conditions, postprocessing procedures, and photographs of recording environments are also included in the dataset. We provide some sample videos for understanding of the recording environments and useful tools for manipulating sound files (e.g., a tool for dividing sound clips into segments of 10 or 15 s length).

3. BENCHMARK OF ACOUSTIC SCENE CLASSIFICATION TASK

3.1. Experimental conditions

As the benchmark system in ASC, we evaluated the performance of a CNN-based method using RU Multichannel Domestic Acoustic Scenes 2019. In this experiment, we cut sound files into 15 s sounds. We then resampled the sound files to 44.1 kHz and extracted the 64-dimensional mel-band energies, which were calculated for each 40 ms time frame with 50% overlap. The implemented system was based on [17]; the detailed network structure and the parameter settings of the networks were determined with reference to [32]. Forty-two acoustic feature maps extracted from 42-channel recordings were input to different channels in the first CNN layer. The network was trained using the Adam optimizer with a learning rate of 0.001. The other experimental conditions are listed in Table 2. The evaluation was conducted using a four-fold cross-validation setup, where each fold had roughly the same number of sound clips with respect to each acoustic scene.

3.2. Experimental results

The performance of ASC using the CNN-based method was 58.3% the average F-score for all acoustic scenes. This result indicates that the ASC task using RU Multichannel Domestic Acoustic Scenes 2019 is still difficult even using the CNN architecture, which enables scene classification with reasonable performance. Thus, we consider that this dataset is suitable for evaluating ASC performance with more sophisticated acoustic features based on spatial information and/or models based on neural networks. More detailed experimental results are given in [37].

4. CONCLUSION

In this paper, we introduced the RU Multichannel Domestic Acoustic Scenes 2019 dataset, which was recorded by multichannel distributed microphones with various devices. This dataset consists of over 45 h $\times$ 42 channels of sounds recorded in a home environment in which people actually live. We hope that RU Multichannel Domestic Acoustic Scenes 2019 will be widely used for evaluating methods of ASC utilizing spatial information, ensemble techniques, and domain adaptation techniques.

5. ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Numbers JP16H01735 and JP19K20304, KDDI Foundation, and Support Center for Advanced Telecommunications Technology Research.

6. 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. IEEE International Conference on Multimedia and Expo* (*ICME*), pp. 1218–1221, 2009.

[2] P. Guyot, J. Pinquier, and R. André-Obrecht, "Water sound recognition based on physical models," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 793–797, 2013.

[3] A. Harma, M. F. McKinney, and J. Skowronek, "Automatic surveillance of the acoustic activity in our living environment," *Proc. IEEE International Conference on Multimedia and Expo* (*ICME*), 2005.

[4] R. Radhakrishnan, A. Divakaran, and P. Smaragdis, "Audio analysis for surveillance applications," *Proc. 2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (*WASPAA*), pp. 158–161, 2005.

[5] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "On acoustic surveillance of hazardous situations," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 165–168, 2009.

[6] T. Komatsu and R. Kondo, "Detection of anomaly acoustic scenes based on a temporal dissimilarity model," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 376–380, 2017.

[7] 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/ACM Trans. Audio Speech Lang. Process.*, vol. 14, no. 1, pp. 321–329, 2005.

[8] K. Imoto and S. Shimauchi, "Acoustic scene analysis based on hierarchical generative model of acoustic event sequence," *IEICE Trans. Inf. Syst.*, vol. E99-D, no. 10, pp. 2539–2549, 2016.

[9] J. Schröder, J. Anemiiller, and S. Goetze, "Classification of human cough signals using spectro-temporal Gabor filterbank features," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 6455–6459, 2016.

[10] T. Zhang and C. J. Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *IEEE Trans. Audio Speech Lang. Process.*, vol. 9, no. 4, pp. 441–457, 2001.

[11] Q. Jin, P. F. Schulam, S. Rawat, S. Burger, D. Ding, and F. Metze, "Event-based video retrieval using audio," *Proc. INTERSPEECH*, 2012.

[12] 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. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 3163–3167, 2013.

[13] J. Liang, L. Jiang, and A. Hauptmann, "Temporal localization of audio events for conflict monitoring in social media," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 1597–1601, 2017.

[14] K. Imoto, "Introduction to acoustic event and scene analysis," *Acoustical Science and Technology*, vol. 39, no. 3, pp. 182–188, 2018.

[15] Y. Han, J. Park, and K. Lee, "Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification," *the Detection and Classification of Acoustic Scenes and Events* (*DCASE*), pp. 1–5, 2017.

[16] H. Jallet, E. Çakır, and T. Virtanen, "Acoustic scene classification using convolutional recurrent neural networks," *the Detection and Classification of Acoustic Scenes and Events* (*DCASE*), pp. 1–5, 2017.

[17] M. Valenti, S. Squartini, A. Diment, G. Parascandolo, and T. Virtanen, "A convolutional neural network approach for acoustic scene classification," *Proc. International Joint Conference on Neural Networks* (*IJCNN*), pp. 547–1554, 2017.

[18] R. Tanabe, T. Endo, Y. Nikaido, T. Ichige, P. Nguyen, Y. Kawaguchi, and K. Hamada, "Multichannel acoustic scene classification by blind dereverberation, blind source separation, data augmentation, and model ensembling," *Tech. Rep. DCASE*, 2018.

[19] A. Raveh and A. Amar, "Multi-channel audio classification with neural network using scattering transform," *Tech. Rep. DCASE*, 2018.

[20] Y. Liping, C. Xinxing, and T. Lianjie, "Acoustic scene classification using multi-scale features," *Tech. Rep. DCASE*, 2018.

[21] Z. Ren, Q. Kong, K. Qian, M. D. Plumbley, and B. W. Schuller, "Attention-based convolutional neural networks for acoustic scene classification," *Proc. Detection and Classification of Acoustic Scenes and Events* (*DCASE*) *Workshop*, pp. 39–43, 2018.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv, arXiv:1409.1556*, 2014.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv, arXiv:1512.03385*, 2015.

[24] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), pp. 1251–1258, 2017.

[25] A. Rakotomamonjy and G. Gasso, "Histogram of gradients of time-frequency representations for audio scene classification," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 1, pp. 142–153, 2015.

[26] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," *Proc. European Signal Processing Conference* (*EUSIPCO*), pp. 1128–1132, 2016.

[27] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events* (*DCASE*), pp. 85–92, 2017.

[28] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events* (*DCASE*), pp. 9–13, 2018.

[29] 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. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (*WASPAA*), pp. 1–5, 2015.

[30] H. Kwon, H. Krishnamoorthi, V. Berisha, and A. Spanias, "A sensor network for real-time acoustic scene analysis," *Proc. IEEE International Symposium on Circuits and Systems*, pp. 169–172, 2009.

[31] 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.*, vol. 25, no. 6, pp. 1335–1343, 2017.

[32] K. Imoto, "Acoustic scene analysis using partially connected microphones based on graph cepstrum," *Proc. European Signal Processing Conference* (*EUSIPCO*), pp. 2453–2457, 2018.

[33] K. Nakadai and D. R. Onishi, "Partially-shared convolutional neural network for classification of multi-channel recorded audio signals," *Tech. Rep. DCASE*, 2018.

[34] C. Zieger and M. Omologo, "Acoustic event detection - itcirst aed database," *Internal ITC report, Tech. Rep.*, 2005.

[35] J. Kürby, R. Grzeszick, A. Plinge, and G. A. Fink, "Bag-of-features acoustic event detection for sensor networks," *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events* (*DCASE*), pp. 1–5, 2016.

[36] G. Dekkers, S. Lauwereins, B. Thoen, M. W. Adhana, H. Brouckxon, B. V. Bergh, T. Waterschoot, B. Vanrumste, M. Verhelst, and P. Karsmakers, "The SINS database for detection of daily activities in a home environment using an acoustic sensor network," *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events* (*DCASE*), pp. 1–5, 2017.

[37] https://www.ksuke.net/dataset.

[38] http://www.sifi.co.jp/system/modules/pico/index.php?content_id=39&ml_lang=en.

[39] https://motu.com/products/avb/8m.

[40] https://pubs.shure.com/guide/MX150/en-US.

[41] N. Ono, H. Kohno, and S. Sagayama, "Blind alignment of asynchronously recorded signals for distributed microphone array," *Proc. Applications of Signal Processing to Audio and Acoustics* (*WASPAA*), pp. 161–164, 2009.

[42] Z. Liu, "Sound source separation with distributed microphone arrays in the presence of clock synchronization errors," *Proc. International Workshop on Acoustic Echo and Noise Control* (*IWAENC*), pp. 1–4, 2008.