Investigation on Spatial and Frequency-Based Features for Asynchronous Acoustic Scene Analysis

Yuki Shiroma* Keisuke Imoto† Sayaka Shiota* Nobutaka Ono* and Hitoshi Kiya*
* Tokyo Metropolitan University, Tokyo, Japan
E-mail: shiroma-yuki@ed.tmu.ac.jp
† Doshisha University, Kyoto, Japan

***Abstract*—We investigate the characteristics of spatial and frequency-based features as an aspect of asynchronous acoustic-scene analysis. Recently, many tasks related to acoustic scene classification (ASC) have attracted a lot of interest. One of the most famous events, the challenges and workshops on detection and classification of acoustic scenes and events (DCASE) series, started in 2013. Here, we focus on DCASE 2018 Task 5, a multi-channel dataset, from among the many tasks of the challenges. This dataset consists of audio data acquired by multiple microphone arrays at different positions. It should provide rich spatial information on acoustic scenes. However, little research has exploited spatial features for the task. We focus on the spatial cepstrum, a robust spatial feature. First, we analyze DCASE 2018 Task 5 with cross-correlation to estimate synchronous and asynchronous data. Then, a convolutional neural network-based ASC model using the spatial cepstrum is used to evaluate performances under synchronous or asynchronous datasets. In our experiments, a system with spatial cepstrum and frequency-based features was evaluated under multi-channel datasets, and the results show that it achieved the highest F-score.**

I. Introduction

Recently, many tasks related to acoustic scene classification (ASC) have attracted a lot of interest. ASC contains many subtasks such as tagging [1], [2], localization [3], monitoring of domestic activities [4], [5], and anomalous detection [6], [7]. International challenges and workshops related to these subtasks started in 2013 and are called the challenges and workshops on detection and classification of acoustic scenes and events (DCASE). Recently, DCASE has become one of the most famous events involving ASC.

So far, many kinds of convolutional neural network (CNN)-based ASC methods with spectrogram features have been proposed [8], [9], and the baseline systems of DCASE also adopt CNN-based ASC systems [10], [11]. One of the important points of these systems is the selection of input features for the architectures. So far, frequency-based features, such as log-mel spectrograms [12], [13] and mel-frequency cepstral coefficients (MFCCs) [14], [15], and raw waveforms [16], [17] have been used as input features.

In this paper, we focus on DCASE 2018 Task 5 since the dataset of this task provides multi-channel audio segments acquired by multiple microphone arrays at different positions for monitoring domestic activities [18]. From the multi-channel audio dataset, not only frequency-based features but also spatial features can be extracted. Even though Task 5 initially focused on systems that can exploit spatial cues independent of sensor location using multi-channel audio, most of the systems proposed for DCASE 2018 Task 5 used only frequency-based features [19], [20]. One of the reasons is that the provided dataset contained asynchronous multi-channel data, which directly affected the performance of spatial features. However, there has been no discussion on the effects of asynchronous multi-channel data on spatial-based features.

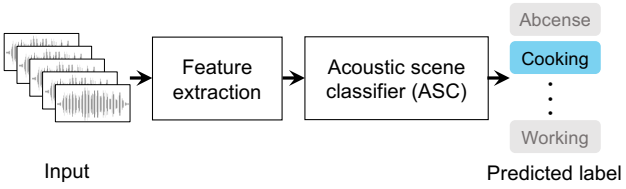

Fig. 1. Block diagram of acoustic scene classification

First, this paper analyzes the DCASE 2018 Task 5 dataset from the perspective of synchronous and asynchronous multi-channel data. For the analysis, we investigate the performances of CNN-based systems with spatial or frequency-based features. As a spatial feature, the spatial cepstrum is used since it was proposed for using spatial information obtained from a distributed microphone array [21]. In addition, as a frequency-based feature, the log-mel spectrogram is used. From our investigation, it was confirmed that almost half of the DCASE 2018 Task 5 dataset contained asynchronous multi-channel data, and the performance when using the spatial cepstrum was degraded by the asynchronous data. After arranging the dataset to contain synchronous data only, the experimental results show that synchronous data only improved the performance of a system using the spatial cepstrum. Additionally, since a system combining spatial and frequency-based features achieved the highest F1-score, it is indicated that these features compensated for each other and improved the performance.

The rest of the paper is organized as follows. In Section II, the ASC task is introduced. Then, the spatial cepstrum is introduced in Section III. In Section IV, we analyze the asynchronous multi-channel dataset and show the experimental results. Finally, our conclusion is given in Section VI.

II. Acoustic scene classification with multi-channel audio

Acoustic scene classification (ASC) is a task that classifies sounds into predefined categories such as "cooking," "vacuuming," and "watching TV" or situations such as "being on the bus," "being in a park," and "meeting" [22]. Figure 1 illustrates a block diagram of ASC. There are two important modules. One is feature extraction, and the other is a classifier of acoustic scenes. So far, many features such as MFCCs or log-mel frequency spectrograms have been used for ASC. These features are categorized into two types: those extracted from single-channel or from multi-channel audio. In the single-channel case, frequency-based features are generally extracted, whereas in the multi-channel case, not only frequency information but also rich spatial information can be extracted. To use spatial information effectively, spatial information-based features are used for acoustic scene classification [21], [23].

III. Spatial features for acoustic scene classification

To make use of spatial cues that can be extracted from multi-channel audio for ASC, Tanabe et al. applied the preprocessing methods of blind dereverberation and blind source separation for scene analysis [24]. However, these methods require synchronized multi-channel observations. To extract spatial information from asynchronous multi-channel observations, Imoto and Ono proposed the spatial cepstrum [21]. The spatial cepstrum can extract spatial information from multi-channel audio. The positions of microphones are not required to calculate the spatial cepstrum, enabling convenient spatial feature extraction using a distributed microphone array. To calculate the spatial cepstrum, a channel-based log-amplitude vector is used:

$$\mathbf{q}_\tau = \begin{pmatrix} \log \tilde{a}_{\tau,1} \\ \log \tilde{a}_{\tau,2} \\ \vdots \\ \log \tilde{a}_{\tau,n} \\ \vdots \\ \log \tilde{a}_{\tau,N} \end{pmatrix}, \tag{1}$$

where τ, n, and N are the time frame, channel index, and number of microphones, and

$$\tilde{a}_{\tau,n} = \sqrt{\frac{1}{\Omega} \sum_{w} a_{w,\tau,n}^2} \tag{2}$$

is the multi-channel power observation at each time frame. $a_{w,\tau,n}$ and w represent the amplitude information in a short time Fourier transform (STFT) representation and the frequency index. Then, principal component analysis (PCA) is applied for basis transformation. To apply PCA, the covariance matrix R_q is calculated by

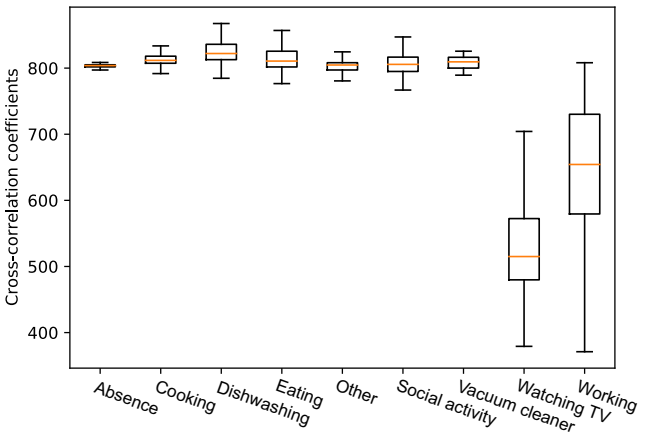

Fig. 2. Box plot of coefficients of cross-correlation of DCASE 2018 Task 5 evaluation fold

$$\mathbf{R}_q = \frac{1}{T} \sum_{\tau} \mathbf{q}_\tau \mathbf{q}_\tau^{\mathsf{T}}, \tag{3}$$

where T is the number of time frames, and $^{\mathsf{T}}$ represents the vector transpose. Because $\mathbf{R_q}$ is a symmetric matrix, the eigendecomposition of $\mathbf{R_q}$ can be expressed as

$$\mathbf{R}_q = \mathbf{E}\mathbf{D}\mathbf{E}^{\mathsf{T}}, \tag{4}$$

where $\mathbf{E}$ and $\mathbf{D}$ are the eigenvector matrix and the diagonal matrix. The spatial cepstrum is defined using $\mathbf{E}$ as

$$\mathbf{d}_\tau = \mathbf{E}^{\mathsf{T}} \mathbf{q}_\tau. \tag{5}$$

IV. Experiments

A. Analysis of DCASE2018 Task 5 dataset

The DCASE 2018 Task 5 dataset was released for monitoring domestic activities. The dataset consists of multi-channel audio segments acquired by multiple microphone arrays at different positions. From the official description of the dataset, "there is not a full time-wise overlap by all sensor nodes for a particular consecutive activity of those classes" [25]. This means the dataset contains some asynchronous data.

First, we assumed that the synchronous data had a high cross-correlation between two nodes, and we calculated the cross-correlation of the spectrograms between two nodes per each label. The procedure was as follows. In step 1, direct current components were removed from all files. In step 2, STFT was performed to extract spectrograms. In step 3, L2 normalization was applied. In step 4, a cross-correlation of the spectrograms between two nodes was calculated while shifting spectrograms along the time axis. In step 5, the maximum coefficient of the cross-correlation was set to the value for dataset analysis. Since there were four nodes, a round-robin

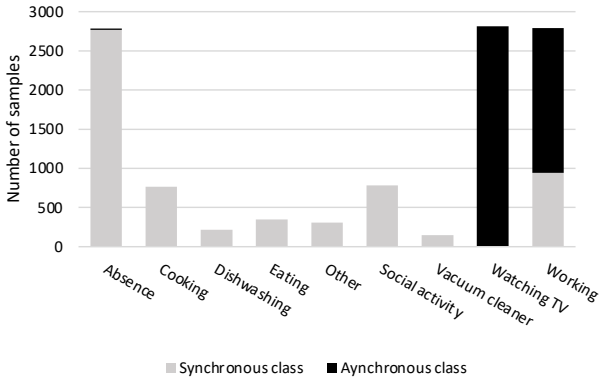

Fig. 3. Number of samples of K-means clustering results for each label in DCASE 2018 Task 5 evaluation fold

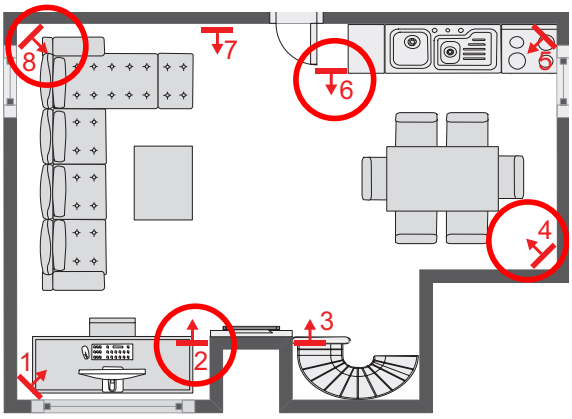

Fig. 4. 2D floorplan of combined kitchen and living room [26] with selected microphone numbers

TABLE I
PERCENTAGE OF MANUAL ANNOTATION FOR EACH LABEL [%]

Label	Synchronous	Asynchronous
Absence	70	30
Cooking	100	0
Dishwashing	100	0
Eating	100	0
Other	100	0
Social activity	100	0
Vacuum cleaner	100	0
Watching TV	4	96
Working	2	98

TABLE II
ACOUSTIC SCENE CLASSIFICATION F-SCORE OF SPATIAL CEPSTRUM FOR DIFFERENT DATASETS

Layer	Output size
Input	1 × 497 × 16
Conv (1 × 7, 64) + BN + ReLU + Dropout (0.2)	1 × 497 × 64
Conv (1 × 10, 128) + BN + ReLU + Dropout (0.2)	1 × 488 × 128
Conv (1 × 13, 256) + BN + ReLU + Dropout (0.2)	1 × 476 × 256
Global max pooling + Dropout (0.2)	256
Dense	128
Softmax output	9

check was carried out. Figure 2 shows a box plot of the coefficients of the cross-correlation of the DCASE 2018 Task 5 development-dataset evaluation fold. Most of the values for "watching TV" and "working" were lower than those of the other labels. This indicates that most of the asynchronous data was labeled as "watching TV" or "working." To estimate the synchronous or asynchronous data automatically, the K-means clustering method was applied to the coefficient distribution. A class that had higher coefficients was regarded as synchronous, and the other was regarded as asynchronous. Figure 3 show the K-means clustering results for each label. The majority of the asynchronous data belonged to "watching TV" or "working." Of the samples of the evaluation fold, 42% were classified as synchronous. Additionally, manual annotation was performed to verify this result. From each label, 50 samples were randomly selected and annotated by a single male annotator. Table I shows the manual annotation results. All data labeled "watching TV" or "working" were removed from the DCASE 2018 Task 5 dataset, and few pieces of "absence" data that was estimated to be asynchronous data by K-means clustering were also removed. The dataset with asynchronous data removed is referred to as "DCASE sync. only" data in this paper. The amounts of data selected for training, validation, and testing were 5,338, 1,325, 2,167 samples, respectively. Each amount was almost 50% reduced from the original DCASE 2018 Task 5 dataset.

B. Dataset

The DCASE 2018 Task 5 development dataset was used for our experiments. For training and validation, a training fold of the dataset was used and split into 8:2. For testing, an evaluation fold of the dataset was used. Additionally, the SINS dataset [18], [27] was used. It contained a continuous recording of one person living in a vacation home over a period of one week. For simplification, only nodes 2, 4, 6, and 8 were used for the experiments as shown in Figure 4. To bring labels in line with the DCASE dataset, the labels "visit" and "calling" were merged into "social activity." To set the same data size as the DCASE dataset, 11,672, 2,914, and 3,654 samples were selected for the training, validation, and testing data. In all datasets, each sample had four nodes, and each node had four-channel audio. Both datasets were sampled at 16 kHz.

C. Acoustic features

In our experiments, the spatial cepstrum and the log-mel spectrogram were used as a spatial feature and a frequency one, respectively. We followed the original procedure of the spatial cepstrum [21]. For log-mel spectrograms, two feature representations were prepared. The first one was a simple log-mel spectrogram that was extracted from single-channel audio, which was separated from 16-channel audio. The other one was a 16-channel combined log-mel spectrogram that was created by extracting 16 channels of audio and combining them in the channel dimension. The 16-channel combined log-

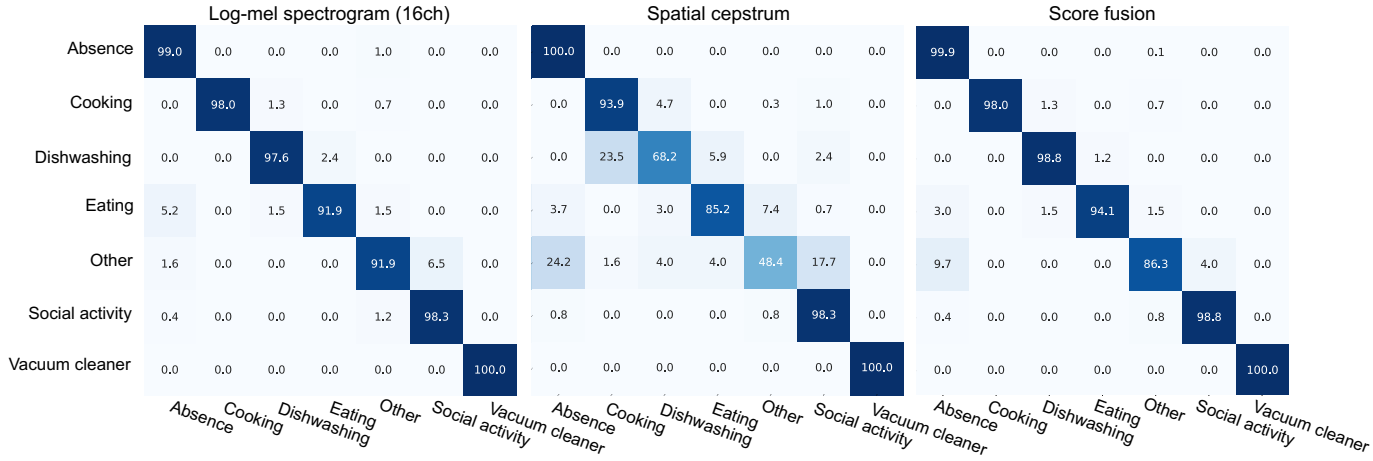

Fig. 5. Confusion matrix of classification accuracy [%] with DCASE sync. only dataset

mel spectrogram included frequency and spatial information simultaneously. Direct current components were removed from all audio waves before extracting acoustic features.

D. Experimental conditions

The duration of each file in the DCASE 2018 Task 5 dataset was ten seconds. Since the duration of each file in the SINS dataset was longer than ten seconds, each file was cut into ten seconds. For STFT, the frame length and frame shift were set to 1024 and 320, respectively. The number of bins for the log-mel spectrogram was 40. We prepared four systems for comparison as follows.

(A) Spatial cepstrum:
Spatial cepstrum extracted from 16-channel audio data.

(B) Log-mel spectrogram (1ch):
Log-mel spectrograms (1ch) extracted from single-channel audio data.

(C) Log-mel spectrogram (16ch):
Log-mel spectrograms (16ch) created by 16-channel audio data and combining each channel into the channel dimension.

(D) Score fusion:
The system with the spatial cepstrum and the system with log-mel spectrogram (16ch) were trained; then, the softmax outputs were averaged to predict a label.

To construct the system with the spatial cepstrum (A), a CNN-based classification network was used. The network for (A) consisted of convolution, batch normalization (BN) [28], ReLU activation, dropout, global max pooling, and dense and softmax layers. The modified network architecture and parameters are shown in Table II. For the network architectures of (B) and (C), the CNN-based classification network [29] proposed for the DCASE 2018 Task 5 Challenge was used. The training conditions for all systems were 500 epochs using the Adam optimizer [30], where the parameters of the optimizer were set at a learning rate of 0.0001, $\beta 1$ = 0.9, and $\beta 2$ = 0.999. The macro F1-score was used as an evaluation metric for ASC and was defined as follows,

$$\text{Macro F-score} = 2 * \frac{(\text{precision} * \text{recall})}{(\text{precision} + \text{recall})}. \quad (6)$$

Additionally, we compared the ASC performances of the spatial cepstrum (A) in simulated asynchronous cases. To simulate the asynchronous data, we prepared three cases:

- All: Asynchronous data in all nodes.
- Two: Asynchronous data in two nodes. The other two nodes had synchronous data.
- One: Asynchronous data in one node. The other two nodes had synchronous data.

We assumed the asynchronous data had 10 sec delay. For example, in the "All" case, the time of one node had 30 sec delay from one node. There was no overlapped segments in asynchronous node.

TABLE III
ACOUSTIC SCENE CLASSIFICATION F-SCORE OF COMPARISON SYSTEMS IN DIFFERENT DATASETS

System	DCASE (Sync. only)	SINS
(A) Spatial cepstrum	86.8%	79.0%
(B) Log-mel spectrogram (1ch)	86.4%	79.8%
(C) Log-mel spectrogram (16ch)	95.7%	88.9%
(D) Score fusion	**96.4%**	**89.6%**

E. Experimental results

Table III shows the classification results of different features under the synchronous datasets. Comparing (B) [log-mel spectrogram (1ch)] with (A) [spatial cepstrum], (B) obtained a higher score on both datasets. This indicates that the frequency-based feature more easily captured the characteristics for ASC than the spatial-based feature only. Comparing (C) [log-mel spectrogram (16ch)] with (A) and (B), (C) obtained a higher score. The reason could be that log-mel spectrogram (16ch) included both frequency and spatial

TABLE IV
ACOUSTIC SCENE CLASSIFICATION F-SCORE OF SPATIAL CEPSTRUM IN SIMULATED ASYNCHRONOUS DATA

	Simulated case	SINS
Async.	All	71.6%
	Two	73.4%
	One	74.0%
Sync.		**79.0**%

information. Finally, comparing (D) [score fusion] with (A), (B), and (C), (D) obtained the highest score on both datasets. This indicates that the spatial feature and frequency feature complemented each other.

Table IV shows the classification results of spatial cepstrum (A) in simulated asynchronous data. Compared asynchronous cases with synchronous case, the F-scores of asynchronous cases decreased from that of synchronous case. When the number of asynchronous microphones was increased, the performances were getting worse. This indicates that the performance of using spatial cepstrum was seriously affected by the asynchronous data.

Figure 5 shows a confusion matrix of the classification results with DCASE sync. only. From the spatial cepstrum results, "dishwashing" was often predicted as "cooking." Since, in the "dishwashing" and "cooking" cases, sound came from a kitchen, it was difficult to distinguish them from only spatial information. In comparison, from the log-mel spectrogram (16ch) results, "dishwashing" and "cooking" were distinguished with high accuracies. From the confusion matrix for the score fusion system, the accuracies of almost all labels improved. It was also demonstrated that the spatial feature and frequency-based feature compensated for each other.

V. CONCLUSION

In this paper, we investigated the characteristics of spatial and frequency-based features as an aspect of asynchronous acoustic-scene analysis. From an analysis of the DCASE 2018 Task 5 development dataset, we confirmed that 42% of data was classified as asynchronous data. In addition, almost all of the asynchronous data belonged to "watching TV" and "working." In experiments, we compared synchronous data with asynchronous data when using the spatial cepstrum, and it was shown that the spatial cepstrum required synchronous data to construct reliable systems. Additionally, comparing the score fusion system with a 16-ch log-mel spectrogram, 1-ch log-mel spectrogram, and the spatial cepstrum, the score fusion system obtained the highest F1-score. This shows that the spatial and frequency-based features of the system compensated for each other and improved the performance. As future work, we will investigate the robustness of the spatial cepstrum to variations of microphone positions.

VI. ACKNOWLEDGEMENT

This work was supported in part by JSPS KAKENHI Grant number JP19K20271, JP20H00613, ROIS DS-JOINT (030RP2021) to S. Shiota.

REFERENCES

[1] J. P. Bello, C. Silva, O. Nov, R. L. Dubois, A. Arora, J. Salamon, C. Mydlarz, and H. Doraiswamy, "SONYC: A system for monitoring, analyzing, and mitigating urban noise pollution," *Communications of the ACM*, vol. 62, no. 2, pp. 68–77, 2019.

[2] M. Cartwright, A. E. M. Mendez, J. Cramer, V. Lostanlen, G. Dove, H.-H. Wu, J. Salamon, O. Nov, and J. Bello, "SONYC urban sound tagging (SONYC-UST): A multilabel dataset from an urban acoustic sensor network," *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, pp. 35–39, 2019.

[3] A. Politis, A. Mesaros, S. Adavanne, T. Heittola, and T. Virtanen, "Overview and evaluation of sound event localization and detection in dcase 2019," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 684–698, 2020.

[4] G. Dekkers, L. Vuegen, T. van Waterschoot, B. Vanrumste, and P. Karsmakers, "DCASE 2018 Challenge - Task 5: Monitoring of domestic activities based on multi-channel acoustics," *arXiv:1807.11246*, 2018.

[5] N. Turpault, R. Serizel, A. Parag Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, pp. 253–257, 2019.

[6] Y. Koizumi, Y. Kawaguchi, K. Imoto, T. Nakamura, Y. Nikaido, R. Tanabe, H. Purohit, K. Suefusa, T. Endo, M. Yasuda, and N. Harada, "Description and discussion on DCASE2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring," *Proc. Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, pp. 81–85, 2020.

[7] Y. Kawaguchi, K. Imoto, Y. Koizumi, N. Harada, D. Niizumi, K. Dohi, R. Tanabe, H. Purohit, and T. Endo, "Description and discussion on DCASE 2021 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring under domain shifted conditions," *In arXiv e-prints: 2106.04492, 1–5*, 2021.

[8] H. Chen, Z. Liu, Z. Liu, P. Zhang, and Y. Yan, "Integrating the data augmentation scheme with various classifiers for acoustic scene modeling," *Tech. Rep. DCASE2019 Challenge Task1*, 2019.

[9] S. Suh, S. Park, Y. Jeong, and T. Lee, "Designing acoustic scene classification models with CNN variants," *Tech. Rep. DCASE2020 Challenge Task1*, 2020.

[10] S. Adavanne, A. Politis, and T. Virtanen, "A multi-room reverberant dataset for sound event localization and detection," *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, pp. 10–14, 2019.

[11] E. Fonseca, M. Plakal, F. Font, D. P. W. Ellis, and X. Serra, "Audio tagging with noisy labels and minimal supervision," *Proc. Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, pp. 69–73, 2019.

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

[13] T. Heittola, A. Mesaros, and T. Virtanen, "Acoustic scene classification in DCASE 2020 challenge: generalization across devices and low complexity solutions," *Proc. Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, pp. 56–60, 2020.

[14] Z. Huang and D. Jiang, "Acoustic scene classification based on deep convolutional neuralnetwork with spatial-temporal attention pooling," *Tech. Rep. DCASE2019 Challenge Task1*, 2019.

[15] H. Wang, D. Chong, and Y. Zou, "Acoustic scene classification with multiple decision schemes," *Tech. Rep. DCASE2020 Challenge Task1*, 2020.

[16] J. Huang, P. Lopez Meyer, H. Lu, H. Cordourier Maruri, and J. Del Hoyo, "Acoustic scene classification using deep learning-based ensemble averaging," *Tech. Rep. DCASE2019 Challenge Task1*, 2019.

[17] P. Lopez-Meyer, J. A. Del Hoyo Ontiveros, G. Stemmer, L. Nachman, and J. Huang, "Ensemble of convolutional neural networks for the DCASE 2020 acoustic scene classification challenge," *Tech. Rep. DCASE2020 Challenge Task1*, 2020.

[18] G. Dekkers, S. Lauwereins, B. Thoen, M. W. Adhana, H. Brouckxon, T. van 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. the Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, pp. 32–36, 2017.

[19] H. Liu, F. Wang, X. Liu, and D. Guo, "An ensemble system for domestic activity recognition," *Tech. Rep. DCASE2018 Challenge Task5*, 2018.

[20] Y. Shen, K. He, and W. Zhang, "Home activity monitoring based on gated convolutional neural networks and system fusion," *Tech. Rep. DCASE2018 Challenge Task5*, 2018.
[21] K. Imoto and N. Ono, "Spatial cepstrum as a spatial feature using a distributed microphone array for acoustic scene analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1335–1343, 2017.
[22] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, 2015.
[23] M. C. Green and D. Murphy, "Acoustic scene classification using spatial features," pp. 42–45, November 2017.
[24] 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. DCASE2018 Challenge Task5*, 2018.
[25] "DCASE2018 task5." http://dcase.community/challenge2018/task-monitoring-domestic-activities. Accessed:2021-07-23.
[26] G. Dekkers, S. Lauwereins, B. Thoen, M. W. Adhana, H. Brouckxon, T. van 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," pp. 32–36, November 2017.
[27] "SINS database." http://github.com/KULeuvenADVISE/SINS_database. Accessed:2021-07-23.
[28] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," vol. 37, pp. 448–456, 2015.
[29] T. Inoue, P. Vinayavekhin, S. Wang, D. Wood, N. Greco, and R. Tachibana, "Domestic activities classification based on CNN using shuffling and mixing data augmentation," *Tech. Rep. DCASE2018 Challenge Task5*, 2018.
[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Proc. International Conference on Learning Representations (ICLR2015)*, 2015.