First Demonstration of Acoustic Scene Classification Based on Trained Sound-to-Light Conversion

Shun Kotsugi, Takao Kawamura, and Nobutaka Ono*
* Tokyo Metropolitan University, Tokyo, Japan
E-mail: kotsugi-shun@ed.tmu.ac.jp, kawamura-takao@ed.tmu.ac.jp, onono@tmu.ac.jp

***Abstract*—In this paper, we present the first experimental demonstration of acoustic scene classification (ASC) using trained sound-to-light conversion with Blinkies, compact devices that convert acoustic signals into optical signals. In this study, we construct a small-scale real-world dataset by recording both acoustic and optical signals in a simulated living environment. To realize effective sound-to-light conversion under the bandwidth limitations of standard video cameras, we train frequency subband weights using recorded audio signals and evaluate classification performance using the resulting filtered short-time power features in simulation. The results show that these features achieve accuracy comparable to conventional spectral features. We then implement the trained conversion on Blinkies and successfully perform ASC using actual optical signals captured by a video camera. These results demonstrate the feasibility of ASC using trained sound-to-light conversion in real environments and provide a foundation for future development of Blinky-based sensing systems and datasets.**

***Index Terms*—Blinky, sound-to-light conversion, acoustic scene classification, distributed sensing**

I. Introduction

In recent years, the development of computing environments and machine learning techniques has driven significant progress in environmental sound analysis [1]. This field deals with various types of sounds, including speech, music, and ambient noise, with applications such as automatic life-logging [2] and monitoring systems for infants and seniors [3], [4]. One fundamental task in this field is acoustic scene classification (ASC), which involves classifying short audio clips into predefined scene categories (e.g., "Watching TV" and "Cleaning") [5], [6].

Conventional ASC systems rely on spectral features (e.g., Mel-frequency cepstral coefficients, Mel spectrograms) extracted from single or multiple microphones and classified using machine learning. While single-microphone systems are widely used, multi-microphone setups can incorporate spatial information, such as power ratio and time differences between microphones [7], to enhance classification accuracy. A distributed microphone array, where multiple microphones are placed across a large area, further extends this capability by covering a wider space and capturing sound sources [8]–[13]. However, obtaining spatial information requires synchronization between microphones. Without synchronization, differences in recording start times and sampling frequencies can degrade the accuracy of the obtained spatial information [14]–[16].

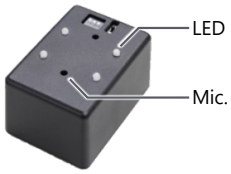


Fig. 1. The sound-to-light conversion device "Blinky"

To address these challenges, we have been developing a sound-to-light conversion device called "Blinky" [17] (Fig. 1) as a framework for distributed microphone arrays. Blinkies capture acoustic signals through built-in microphones and convert them into optical signals, which can be recorded using a video camera. By distributing multiple Blinkies in space and capturing their optical signals with a video camera, it becomes possible to obtain synchronized acoustic information over a large area without requiring direct synchronization between microphones. This approach has the potential to address the challenges of spatial information acquisition in ASC. The effectiveness of Blinkies for ASC has already been confirmed in simulations [18], [19].

However, once a specific sound-to-light conversion method is applied, the original acoustic signals are lost, making it difficult to optimize the conversion process. This presents a fundamental challenge in designing an effective conversion strategy for ASC. Previous ASC studies using Blinkies [18], [19] have relied on the DCASE 2018 Task 5 dataset [20], which was not recorded with actual Blinkies. Therefore, no prior study has validated ASC feasibility using real recorded signals and the features that Blinkies can actually transmit.

In this study, we conduct an initial investigation into ASC using Blinkies in a real-world environment. We first recorded acoustic signals for several scenes in a mock living room using Blinkies, in order to analyze the data and optimize the sound-to-light conversion. We used two types of features that can be transmitted from Blinky's LED within the limited bandwidth of a widely available video camera: short-time power and weighted short-time power across frequency subbands. We then compared ASC performance using the two Blinky-based features against that using conventional spectral features. We also installed the weighted short-time power feature on Blinkies and performed scene classification using the optical signals captured by a video camera. This study provides the first ASC results using real recorded signals from Blinkies,

979-8-3315-7206-8/25/$31.00

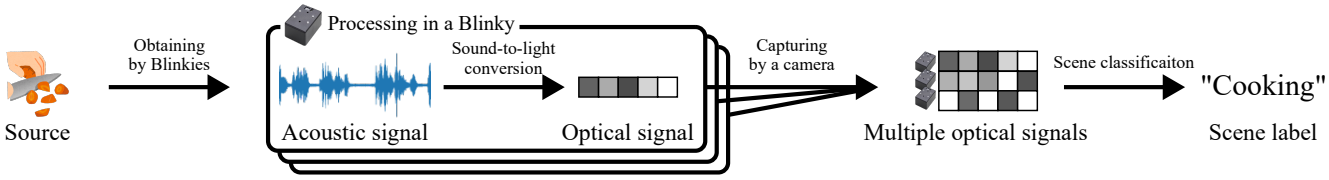


Fig. 2. Overview of the ASC pipeline using Blinkies, from acoustic signal acquisition to optical signal-based classification

laying the groundwork for future optimization of sound-to-light conversion and construction of datasets for Blinky-based ASC.

II. Recording Acoustic Scenes

Blinky is a device that converts acoustic signals through a microphone into light (Fig. 1). In ASC using Blinkies, a video camera captures optical signals emitted from Blinkies distributed over a large space, and then a system classifies scenes using the aggregated optical signals (Fig. 2). The system classifies scenes based on optical signals instead of acoustic features like the log-mel spectrogram, which is widely used in typical ASC. When constructing a dataset in ASC using Blinkies, we should obtain optical signals and determine the method of sound-to-light conversion. Previous studies have proposed obtaining optical signals [21] and considered various methods for Blinky's sound-to-light conversion [19], [22]. The simplest method is to convert the short-time power of the acoustic signal to the optical signal. Other studies have proposed a method based on a random projection of the spectrum [22] and a method that optimizes ASC performance in simulation [19]. However, the conversion method varies depending on the scene and situation we want to analyze. It is vital to optimize the sound-to-light conversion method for the problem to be solved. Thus, we have constructed an acoustic signal dataset captured by the microphone built on Blinky. It enables the investigation of sound-to-light conversion in real-world sound scenes.

A. Recording Conditions

We recorded the dataset in a space designed to replicate a typical living environment. The Blinkies were distributed as shown in Fig. 3) (from #1 to #7) and placed near electrical outlets so they could remain continuously charged, simulating realistic usage scenarios. Although each Blinky is equipped with two microphones, in this study we used only one of them.

We constructed an acoustic signal dataset using the microphones built into the Blinkies for the purpose of investigating sound-to-light conversion, as detailed in Sec. III-C. The sampling frequency and quantization bit rate were set to 16 kHz and 16 bits, respectively. The Blinkies were wirelessly connected to a computer, and their recorded signals were transmitted in real time.

We also recorded optical signals from Blinkies with the effective sound-to-light conversion method studied in Sec. III-D. The optical signals were captured from position ① in Fig. 3, and the Blinkies visible from the video camera's viewpoint are shown in the right photo of Fig. 3. A mirror was used to include Blinkies #1, #2, and #3 in the frame, as they were not directly visible from the camera's position. When recording the scene, we were careful to avoid occlusion of objects and people with Blinkies. The video was recorded at a frame rate of 52 fps.

TABLE I
Recorded scenes and durations of audio and video data

Scene	Duration (min.) Audio	Video
Vacuum cleaner	13.5	13.2
Watching TV	18.7	18.3
Chatting	13.7	13.0
Absence	18.5	11.8

B. Recorded Scenes

In this study, we recorded four scenes that exhibit distinct spatial characteristics of sound: "Vacuum cleaner" (the sound source moves around the room), "Watching TV" (the sound source remains fixed), "Chatting" (the sound sources switch between different positions), and "Absence" (no sound source is present). Figure 3 illustrates the approximate positions of the sound sources for each scene, except for "Vacuum cleaner," where the source moves throughout the room and thus is not depicted. The total recording durations for each scene are summarized in Table I.

III. Experiment

We conducted several experiments to evaluate the applicability of Blinkies to ASC. In these experiments, we decided on short-time power as the method of sound-to-light conversion, which is considered the simplest method. First, we observed examples of the short-time power of each Blinky in each scene (described Sec. III-B). Next, we evaluated ASC performance using the short-time power (described in Sec. III-C). Finally, we evaluated ASC performance using the optical signals of the Blinkies, with weights trained as described in Sec. III-C (described Sec. III-B).

A. Weighted Short-Time Power as a Transmittable Feature from Blinkies to a Video Camera

A widely available commercial video camera typically operates at lower frame rate (e.g., 30Hz) than audio signals. Consequently, only information captured at this video frame rate can be transmitted from Blinkies. To align with this constraint, we consider short-time power as a transmittable

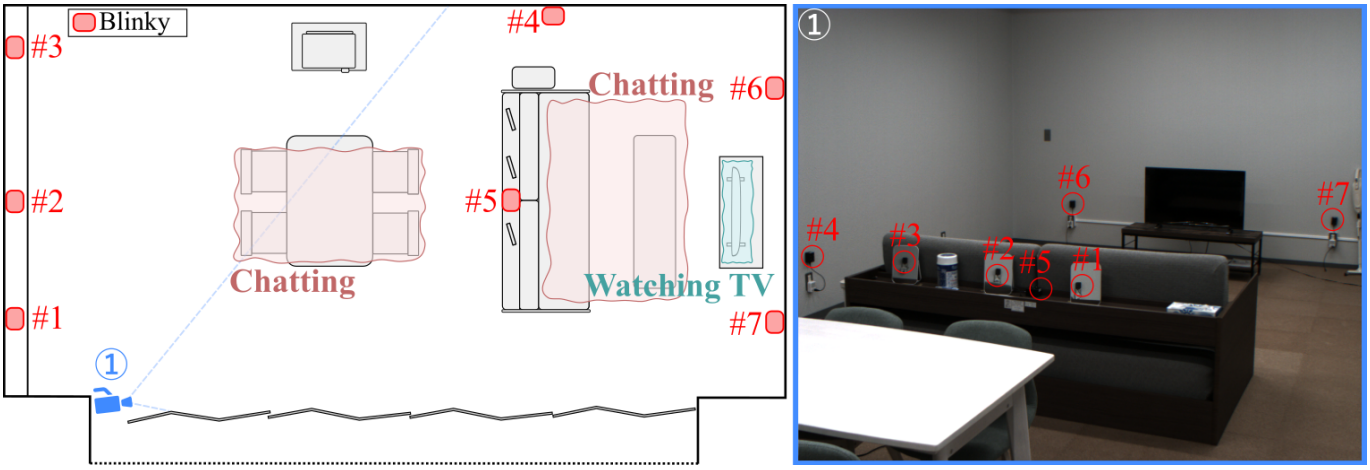


Fig. 3. Recording environment, arrangement of Blinkies, and approximate locations of sound sources. The Blinkies are placed near electrical outlets to allow continuous charging. Blinkies #1, #2, and #3 are located outside the camera's direct field of view, so mirrors are used to include them into the field of view.

feature from Blinkies to a video camera, defined as

$$u_i(t) = 10\log_{10}\left(\frac{1}{L}\sum_f \alpha_i(f)P_i(t,f)\right), \quad (1)$$

$$P_i(t,f) = |X_i(t,f)|^2, \quad (2)$$

where $X_i(t,f)$ is the STFT coefficients of the acoustic signal recorded by the i-th Blinky with frame length L and no overlap, i $(i = 1, \cdots, I)$ is the Blinky index, $\alpha_i(f) \geq 0$ is a learnable weight, respectively. In this study, $\alpha_i(f)$ is set to a constant value within each frequency band corresponding to the Mel scale. Specifically, we define $\alpha_i(f) = c_m$, where c_m is a learnable, band-specific constant. In this experiment, we set $L = 1024$ and $I = 7$. The number of mel bands in $\alpha_i(f)$ was set to 8. At a sampling rate of 16 kHz, this frame length corresponds to 64 ms, approximately 15.6 Hz, which can be transmitted. Note that short-time power is defined when $\forall i, f, \alpha_i(f) = 1$.

B. Observing Short-Time Power at Each Blinky

Figure 4 shows examples of short-time power patterns for each scene. In Fig. 4, the vertical and horizontal axes indicate the Blinky's index i and the time [s], respectively. It was expected that the spatio-temporal patterns would be distinct for each scene. Figure 4 (a) shows that short-time power was distributed throughout the space, and distribution transitions appeared due to the movement of the vacuum cleaner. On the other hand, in Fig. 4 (b), the short-time power is concentrated at Blinkies #6 and #7, which are located near the TV. Figure 4 (c) shows a short-time pattern with pauses like a sequence of sparse and dense areas. Therefore, we confirmed differences in the short-time power patterns between acoustic scenes.

C. Comparison of Short-Time Power-Based Features and Spectral Features

We evaluated ASC performance using features derived from the acoustic signals recorded by Blinkies. In this experiment,

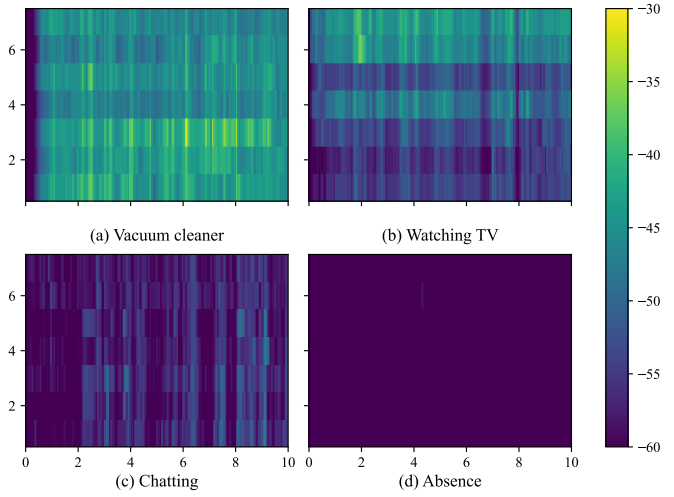


Fig. 4. Spatio-temporal patterns of short-time power across Blinkies for each recorded scene

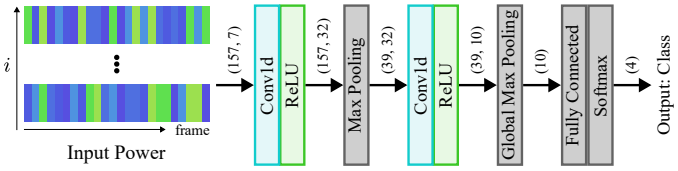


Fig. 5. Model architecture for ASC using short-time power as input features. The CNN treats short-time power at each Blinky as a separate channel. Numbers in parentheses indicate data dimensions.

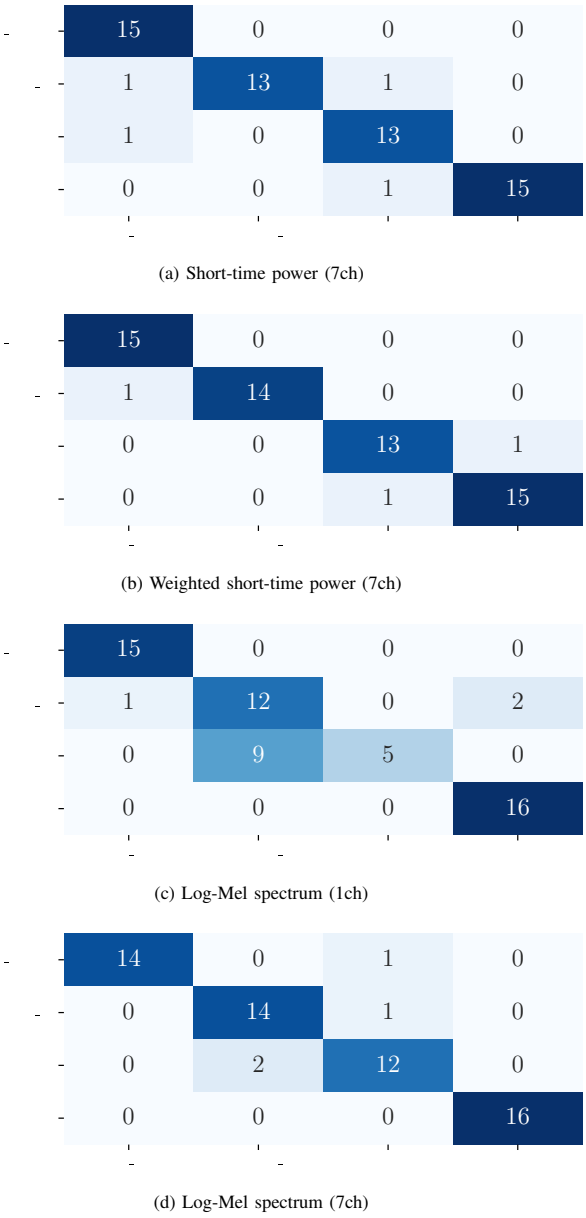
we simulated the sound-to-light conversion by computing the short-time power features from the acoustic signals recorded by Blinkies, without performing the actual sound-to-light conversion. We randomly split the training set (43.7 minutes), validation set (10.7 minutes), and evaluation set (10.0 minutes), where the proportion of each label is equal. We divided each data into 10-second audio clips. The ASC network is shown in Fig. 5. CNN's kernel size was 4. We compared the ASC performance with a model that uses spectral features.

TABLE II
COMPARISON OF FEATURE TYPES FOR ACOUSTIC SCENE CLASSIFICATION

Feature	#ch	#feat. per sec.	macro-F1	Accuracy
Short-time power	7 (All)	112	90.5 ± 9.1	91.0 ± 8.0
Weighted short-time power	7 (All)	112	96.5 ± 4.1	96.7 ± 3.8
Log-Mel spectrogram	1	2000	82.3 ± 3.9	83.5 ± 3.4
	7 (All)	14000	93.5 ± 1.3	93.7 ± 1.2

Specifically, we decided on log-Mel spectrogram as spectral features. The model using log-Mel spectrogram achieved the top score in the DCASE 2018 Challenge Task 5 [20], [23]. The detailed network architecture is described in [23]. In the case of 1ch case, we used Blinky #5 as a channel (located at the center of the room) because it was considered to capture information from all scenes equally compared with other Blinkies. For reference, we also compared the case where spectral features of all channels were used. In training, the activation function was the ReLU function, the optimization method was Adam [24], the learning rate was 1.0×10^{-4}, the loss function was cross-entropy, and the batch size was 8, respectively. The number of epochs was 200 when using short-time power as the input feature and 50 when using spectral features. The evaluation metric is the macro F1-score, which is the average of the F1-scores for each scene.

We compared with the cases using short-time power, weighted short-time power, and log-Mel spectrogram as input features. Table II presents the ASC performance for these three feature types. A comparison between short-time power and weighted short-time power in Table II shows that the latter achieves better performance. This suggests that the weights learned for each mel frequency bin contribute effectively to scene classification. Table II shows that the performance with short-time power was higher than that with 1ch log-Mel spectrogram and comparable to that with 7ch log-Mel spectrogram. For a detailed comparison, we observed the confusion matrix. Figure 6 shows the confusion matrix for the features shown in Table II. In Fig. 6 (a), we confirmed that the model using the short-time power as input features could classify all scenes with high accuracy, while the model of 1ch log-Mel spectrogram often misclassified "Chatting" as "Watching TV." It was considered challenging to distinguish between conversation and speech in TV programs based on spectral information alone. In Fig. 4, we observed the spatial-temporal pattern of 7ch short-time powers, which suggests that their spatial information is effective for ASC. The result for the CNN using the 7ch log-Mel spectrogram suggests that the CNN improved ASC performance by learning spatial information such as spectral power ratio. Despite having significantly fewer features, the short-time power achieved performance comparable to log-Mel spectrogram in the case of the same number of channels. This suggests that the limited number of scenes and the fact that these scenes could be distinguished sufficiently using short-time power. Since we treated simple scenes, we will consider more complex scenes in future work.



(a) Short-time power (7ch)

(b) Weighted short-time power (7ch)

(c) Log-Mel spectrum (1ch)

(d) Log-Mel spectrum (7ch)

Fig. 6. Confusion matrices for each input feature type used in ASC

D. Acoustic Scene Classification Using Optical Signals Captured from Blinkies

In order to actually confirm the validity of the sound-to-light conversion obtained in the aforementioned experiment, we evaluated ASC performance using optical signals taken with the weights installed on the Blinkies when evaluated in Sec. III-C. Figure 7 shows the weights installed Blinkies. We randomly split the signals into the training set (41.3 minutes) and evaluation set (15.0 minutes) , where the proportion of each label is equal. We divided each data into 10-second audio clips. The network architecture and training hyper-parameters were the same as those used in Sec. III-C. Because of calibration problem of sound-to-light conversion, the classifier was trained to investigate whether the obtained Blinky's signals

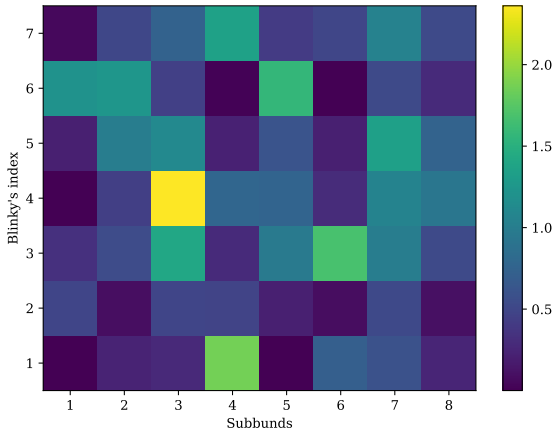


Fig. 7. Weights installed on Blinkies as described in Sec. III-D

are effective for scene classification. The evaluation metric is the macro-F1 score, which is the average of the F1-scores for each scene. The results of scene classification showed that accuracy was 0.97 and macro-F1 score was 0.97. Based on these results, we consider that scene classification is feasible with actual optical signals.

IV. Conclusion

In this study, we presented the first experimental demonstration of acoustic scene classification (ASC) based on trained sound-to-light conversion using Blinkies. To meet the bandwidth constraints of typical video cameras, we focused on short-time power as a transmittable feature and optimized frequency subband weights to enhance classification performance.

We first evaluated ASC performance using simulated conversion from recorded acoustic signals, and confirmed that the optimized short-time power features achieve accuracy comparable to conventional spectral features. We then implemented the trained sound-to-light conversion on physical Blinkies and demonstrated that ASC is feasible using actual optical signals captured by a video camera.

Although this study is limited in terms of the number of scenes and dataset size, it provides the first experimental validation that acoustic scene recognition is achievable through trained sound-to-light conversion and optical signal transmission. These results lay a solid foundation for future development of Blinky-based sensing systems and datasets in real-world environments. Future work will extend beyond frame-wise sound-to-light conversion to methods that also capture temporal dependencies across consecutive frames, which may further enhance recognition performance.

Acknowledgment

This work was supported by JST SICORP Grant Number JPMJSC2306 Japan.

References

[1] K. Imoto, "Introduction to acoustic event and scene analysis," *Acoustical Science and Technology*, vol. 39, no. 3, pp. 182–188, 2018. DOI: 10.1250/ast.39.182.

[2] M. A. M. Shaikh, M. K. I. Molla, and K. Hirose, "Automatic life-logging: A novel approach to sense real-world activities by environmental sound cues and common sense," in *Proc. International Conference on Computer and Information Technology (ICCIT)*, 2008, pp. 294–299. DOI: 10.1109/ICCITECHN.2008.4803018.

[3] Y.-T. Peng, C.-Y. Lin, M.-T. Sun, and K.-C. Tsai, "Healthcare audio event classification using hidden Markov models and hierarchical hidden Markov models," in *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, 2009, pp. 1218–1221. DOI: 10.1109/ICME.2009.5202720.

[4] K. K. B. Peetoom, M. A. S. Lexis, M. Joore, C. D. Dirksen, and L. P. D. Witte, "Literature review on monitoring technologies and their outcomes in independently living elderly people," *Disability and Rehabilitation: Assistive Technology*, vol. 10, pp. 271–294, 4 Jul. 2015, ISSN: 1748-3107. DOI: 10.3109/17483107.2014.961179.

[5] B. Ding, T. Zhang, C. Wang, *et al.*, "Acoustic scene classification: A comprehensive survey," *Expert Systems with Applications*, vol. 238, p. 121 902, 2024, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2023.121902.

[6] K. Zaman, M. Sah, C. Direkoglu, and M. Unoki, "A survey of audio classification using deep learning," *IEEE Access*, vol. 11, pp. 106 620–106 649, 2023. DOI: 10.1109/ACCESS.2023.3318015.

[7] P. Giannoulis, G. Potamianos, and P. Maragos, "Room-localized speech activity detection in multi-microphone smart homes," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2019, p. 15, 1 2019, ISSN: 1687-4722. DOI: 10.1186/s13636-019-0158-8.

[8] G. Dekkers, S. Lauwereins, B. Thoen, *et al.*, "The SINS database for detection of daily activities in a home environment using an acoustic sensor network," in *Proc. Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, 2017, pp. 32–36.

[9] K. Imoto and N. Ono, "RU multichannel domestic acoustic scenes 2019: A multichannel dataset recorded by distributed microphones with various properties," in *Proc. Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, 2019, pp. 104–108.

[10] M. Yasuda, Y. Ohishi, S. Saito, and N. Harada, "Multi-view and multi-modal event detection utilizing transformer-based multi-sensor fusion," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 4638–4642, ISBN: 978-1-6654-0540-9. DOI: 10.1109/ICASSP43922.2022.9746006.

[11] J. Kürby, R. Grzeszick, A. Plinge, and G. A. Fink, "Bag-of-Features acoustic event detection for sensor

networks," in *Proc. Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, 2016, pp. 55–59.

[12] K. Imoto, "Acoustic scene classification using multichannel observation with partially missing channels," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2021, pp. 875–879. DOI: 10.23919/EUSIPCO54536.2021.9616170.

[13] 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. DOI: 10.1109/TASLP.2017.2690559.

[14] S. Miyabe, N. Ono, and S. Makino, "Blind compensation of interchannel sampling frequency mismatch for ad hoc microphone array based on maximum likelihood estimation," *Signal Processing*, vol. 107, pp. 185–196, 2015, ISSN: 0165-1684. DOI: https://doi.org/10.1016/j.sigpro.2014.09.015.

[15] D. Hu, H. Zhang, F. Bao, and R. Wang, "Distributed sampling rate offset estimation over acoustic sensor networks based on asynchronous network newton optimization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 301–312, 2023. DOI: 10.1109/TASLP.2022.3224256.

[16] Y. Masuyama, K. Yamaoka, T. Kawamura, and N. Ono, "Efficient joint optimization of sampling rate offsets using entire multichannel signal," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1816–1828, 2024. DOI: 10.1109/TASLP.2024.3369532.

[17] R. Scheibler and N. Ono, "Blinkies: Open source sound-to-light conversion sensors for large-scale acoustic sensing and applications," *IEEE Access*, vol. 8, pp. 67 603–67 616, 2020. DOI: 10.1109/ACCESS.2020.2985281.

[18] Y. Kinoshita and N. Ono, "Analysis on roles of DNNs in end-to-end acoustic scene analysis framework with distributed sound-to-light conversion devices," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2021, pp. 1167–1172.

[19] Y. Kinoshita and N. Ono, "End-to-end training of acoustic scene classification using distributed sound-to-light conversion devices: Verification through simulation experiments," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, Sep. 2024. DOI: 10.1186/s13636-024-00369-z.

[20] 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 preprint arXiv:1807.11246*, 2018. arXiv: 1807.11246 [eess.AS].

[21] K. Nishida, N. Ueno, Y. Kinoshita, and N. Ono, "Estimation of transfer coefficients and signals of sound-to-light conversion device Blinky under saturation," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, pp. 717–722. DOI: 10.23919/APSIPAASC55919.2022.9980090.

[22] S. Motoyama, N. Ueno, Y. Kinoshita, and N. Ono, "Compressed sensing of sparse spectrum using distributed sound-to-light conversion device Blinkies," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Sep. 2022, pp. 12–16.

[23] 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," Detection, Classification of Acoustic Scenes, and Events Challenge (DCASE), Tech. Rep., 2018.

[24] D. P. Kingma, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.