End-to-End Training for Acoustic Scene Analysis with Distributed Sound-to-Light Conversion Devices

Yuma Kinoshita and Nobutaka Ono
Tokyo Metropolitan University, Tokyo, Japan

***Abstract*—We propose an end-to-end acoustic scene analysis framework with distributed sound-to-light conversion devices called *Blinkies*. Blinkies transmit sound information as the intensity of an on-board light-emitting diode (LED). A video camera can then easily collect acoustic information by capturing the LED intensities from multiple Blinkies distributed over a large area. However, the transmitted signal is band-limited owing to a video camera's frame rate, typically 30 frames per second. We aim to optimize the sound-to-light conversion process for acoustic scene analysis under this bandwidth constraint. In light-signal propagation in air, signal degradation due to physical constraints such as light attenuation and noise will also occur. We model the physical constraints as differentiable physical layers, which enable us to train two deep neural networks (DNNs) for sound-to-light conversion and acoustic scene analysis in an end-to-end manner. Simulation experiments of acoustic scene analysis using a DCASE 2018 dataset show that the proposed framework can produce a higher accuracy than the previous framework with Blinkies. This result suggests the suitability of Blinkies for acoustic scene analysis.**

***Index Terms*—Blinky, Sound-to-light conversion, Deep learning, Differentiable physical layer**

I. Introduction

Interest in acoustic scene analysis has recently increased, and many workshops and competitions have been held [1], [2]. Acoustic scene analysis is aimed at recognizing activities, such as "cooking," "vacuuming," and "watching TV," or determining what is going on, such as "being on a bus," "being in a park," and "meeting with people," from acoustic information [3]. To analyze acoustic scenes with high performance, spatial information is also important in addition to spectral information. The spatial information is obtained using multiple microphones at the same time, that is, a distributed microphone array [4], [5]. A simple way of using spatial information for acoustic scene analysis is to localize sound sources. However, even when there is a single source, source localization may be difficult in a real environment because of background noise, reverberation, and reflection. Furthermore, an acoustic signal generally includes multiple sound sources, making it necessary to carry out more complex operations, such as estimating the number of sound sources.

To overcome these difficulties, Imoto and Ono proposed the use of the spatial cepstrum [6]. The spatial cepstrum is a robust and efficient feature for extracting spatial information with a distributed microphone array. In the calculation of the spatial cepstrum, the average sound power for each microphone channel at a given time is used rather than the amplitude for each frequency. On the basis of the idea to employ sound power as a feature for acoustic scene analysis, we previously developed a sound-to-light conversion device called a *Blinky* [7]–[10]. The use of Blinkies can solve technical challenges in real-time acoustic sensing by using a distributed microphone array, i.e., cable connection with wired communication, network bandwidth limitation through wireless communication, or the synchronization of signals recorded using microphones [11]. In a previous framework using Blinkies, a Blinky measured a sound signal using a microphone and calculated its power on the device. In accordance with the sound power, the Blinky modulated the intensity of an on-board light-emitting diode (LED). Finally, a video camera was used to synchronously capture LED intensities from multiple Blinkies distributed over a large area. However, signal degradation due to light attenuation and noise will occur in the light-signal propagation in air. Also, captured signals will be strongly band-limited because of the limited frame rate of cameras, typically 30 frames per second. Furthermore, important acoustic features will vary depending on the scene and situation that we want to analyze.

Because of such a situation, our aim is to learn the optimal sound-to-light conversion process in Blinkies as an alternative to transmitting sound power information. To realize this aim, in this paper, we propose an end-to-end acoustic scene analysis framework with Blinkies. In the proposed framework, the light-signal propagation in air and camera responses are modeled as differentiable physical layers. These physical layers enable us to obtain appropriate signal transformations in Blinkies by using a data-driven approach while considering the physical constraints and the accuracy of acoustic scene analysis. Namely, we can train two deep neural networks (DNNs) with an end-to-end approach: an encoding network that transforms a sound signal measured via a microphone into a signal to be transmitted by an LED and a scene analysis network that estimates the acoustic scene using captured LED intensities.

We performed a simulation experiment of acoustic scene classification using the DCASE 2018 Challenge Task 5 dataset to evaluate the effectiveness of the proposed framework. Experimental results show that the proposed framework enables us to obtain a higher classification accuracy than a previous framework with Blinkies.

This work was supported by JST CREST Grant Number JPMJCR19A3 and JSPS KAKENHI Grant Number JP20H00613.

II. Sound-to-Light Conversion Device Blinky

The use of Blinkies enables us to avoid complicated processing, such as synchronization, in the signal acquisition using a distributed microphone array. In this section, we briefly summarize an acoustic sensing procedure with Blinkies and the aim of this study.

A. Acoustic Sensing with Blinkies

Acoustic sensing with Blinkies and a video camera consists of three parts: sound-to-light conversion in each Blinky, signal transmission by light, and capturing the LED light of Blinkies by a video camera (see Fig. 1).

1) Sound-to-Light Conversion: Let n, $F_{s,n}$, $x[n]$, and B be the discrete-time index, the sampling frequency corresponding to n, the microphone signal, and the audio buffer size, respectively. From the signal $x[n]$, the sound power measurement $u[n]$ is computed as

$$u[n] = \begin{cases} \frac{1}{B}\sum_{i=1}^{B} x[n-B+i]^2 & n \bmod B = 0 \\ u[n-1] & otherwise \end{cases}. \quad (1)$$

To efficiently encode sound power measurements $u[n]$ as LED intensities, we map $u[n]$ using a nonlinear function $\varphi(\cdot)$. The function $\varphi(\cdot)$ was designed so that it maximizes the entropy of $\varphi(u[n])$ [10]. Then, the actual emitted light intensity $I(t)$ at continuous time t is given by

$$I(t) = \varphi(u[\lfloor tF_{s,n} \rfloor]), \quad (2)$$

where $\lfloor \cdot \rfloor$ indicates the floor function.

2) Signal Transmission by Light: After the sound-to-light conversion, LED light from Blinkies propagates in air, and a video camera captures it. The LED light intensity at the camera is affected by attenuation a depending on the angle and distance between each LED and the video camera. In addition to this attenuation, ambient light is added to the light intensity as a positive bias b. For these reasons, the radiant power density at an imaging sensor on the camera, i.e., irradiance $E(t)$, is calculated using attenuation a, bias b, and noise ϵ as

$$E(t) = aI(t) + b + \epsilon. \quad (3)$$

3) Capturing Light of Blinkies by Video Camera: An imaging sensor on a camera captures irradiance E. The camera then encodes it as a video file. Irradiance E is integrated over the time, which depends on the frame rate $F_{s,m}$ of the camera. This process can be written as

$$X[m] = \int_{(m-1)/F_{s,m}}^{m/F_{s,m}} E(t)dt, \quad (4)$$

where m is the discrete-time index for video frames and $X[m]$ is the energy density. Finally, the captured pixel value p is given by

$$p[m] = f(X[m]), \quad (5)$$

where f is a function combining the sensor saturation $s(\cdot)$ and the camera response function (CRF) $h(\cdot)$. The CRF represents the processing in each camera that makes the final image appear better. One of the typical CRFs is the *Gamma correction*. It converts sensor output $v[m] = s(X[m])$ so that $p[m] = (v[m])^{1/\gamma}$ with $\gamma = 2.2$. Because industrial cameras usually provide raw video frames that directly store sensor output $v[m]$, the nonlinear transform by the CRF can be avoided and we can assume $p[m] = v[m]$.

B. Scenario

In this paper, we assume that Blinkies placed at fixed locations record acoustic signals and a video camera located at a fixed location captures their LED intensities. Here, we assume that their spatial positions are given. We will discuss how to estimate the spatial positions of Blinkies in another paper. Because of the nonlinear mapping in eq. (2), the propagation in eq. (3), and the camera response in eqs. (4) and (5) such as the *Gamma correction*, the captured pixel value p differs from the actual sound power u measured by Blinkies. Furthermore, important acoustic features for acoustic scene analysis will vary in accordance with the scene labels we want to attach to sounds or the ambient sound type and volume. Therefore, the sound-to-light conversion based on the sound power in the previous framework with Blinkies might not be optimal for transmitting sound information by light or for acoustic scene analysis.

To overcome these issues with a data-driven approach, we propose an end-to-end acoustic scene analysis framework with Blinkies in the next section.

III. Proposed Framework

Figure 2 shows the proposed end-to-end acoustic scene analysis framework. In the proposed framework, we have two DNNs: an encoding network that converts recorded signals into signals that can be effectively transmitted and are appropriate for scene analysis, and a scene analysis network that performs scene analysis. To train these DNNs in an end-to-end manner, we model the light propagation between Blinkies and a camera, and camera responses as differentiable physical layers.

A. Differentiable Physical Layers

Differentiable physical layers are differentiable models of physical phenomena that can be incorporated into DNNs. They enable DNNs to consider physical phenomena.

1) Light Propagation Layer: A light propagation layer is a model of the signal transmission between a Blinky and a camera (see Sec. II-A2). Since eq. (3) is differentiable, we calculate the following equation in this layer:

$$y[n] = ax[n] + b + \epsilon, \quad (6)$$

where $x[n]$ and $y[n]$ are 1D signals input to this layer and output from this layer, respectively. We assume that attenuation a is inversely proportional to the square of the distance between a Blinky and a camera, and ϵ follows a normal distribution. b can be calculated from the pixel value p when the corresponding LED is not lit.

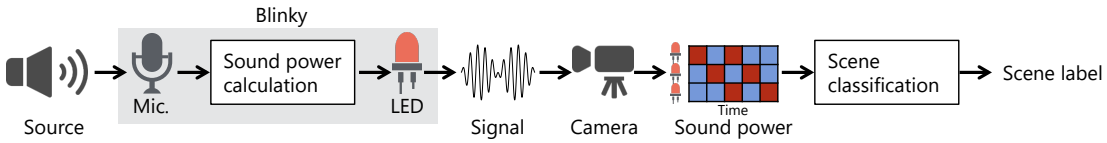


Fig. 1: Process of acoustic sensing with Blinkies

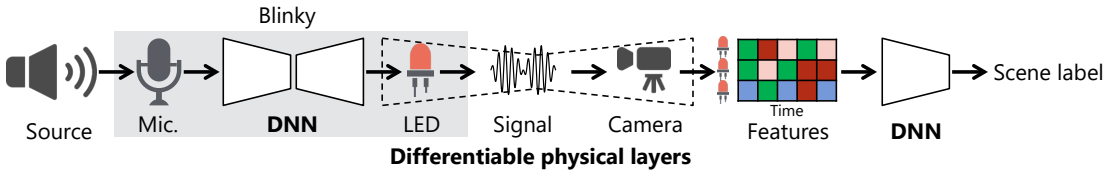


Fig. 2: Proposed end-to-end acoustic scene analysis framework

2) Camera Response Layer: A camera response layer is a model of the integration in eq. (4) on a camera sensor (see Sec. II-A3). This integration can be interpreted as a sampling operation with low-pass filtering. For this reason, the camera response layer resamples an input signal $x[n]$ to the camera frame rate $F_{s,m}$ using

$$y[m] = \mathrm{resample}(x[n]), \tag{7}$$

where $\mathrm{resample}(\cdot)$ indicates the resample operation. Since most cameras have a frame rate of 30 fps, we set $F_{s,m}$ to 30 Hz in this work. Note that the nonlinear transform by a CRF can be avoided by using raw video frames. Hence, we do not consider CRFs in the camera response layer.

B. Network Architecture

As shown in Fig. 2, there are two subnetworks in the proposed framework: an encoding network that transforms a sound signal into a signal transmitted by LED light and a scene analysis network that performs acoustic scene analysis. For the training of these networks, the light propagation and camera response layers are located between these two networks.

For the encoding network, we employed a 1D convolutional neural network (CNN) [12], where we did not consider hardware limitations of Blinkies in this paper. A 1D CNN is also adopted in Wave-U-Net, which transforms acoustic signals into other signals, and its effectiveness has been confirmed [13]. In the encoding network, we downsample microphone signals using six 1D strided convolution layers with a kernel size of 3, a stride of 2, and a padding of 1. In addition, two 1D convolution layers with a kernel size of 3 and a padding of 1 are inserted before each strided convolutional layer.

We adopted a simple VGG-like architecture with 1D convolution layers for the scene analysis network [14]. Similarly to the encoding network, downsampling layers in this network are replaced with 1D strided convolution layers with a kernel size of 3, a stride of 2, and a padding of 1. The depth of the network is 4, and the resulting feature map is transformed by a global average pooling layer into a 1D vector. The vector is fed into a linear layer to obtain the final scene analysis results.

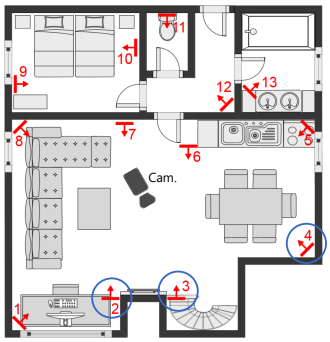


Fig. 3: Arrangement of microphone arrays [17]

IV. Simulation

We evaluated the effectiveness of the proposed framework by an acoustic scene analysis experiment with the DCASE 2018 Challenge Task 5 development dataset [15], [16].

A. Simulation Conditions

The DCASE 2018 Challenge Task 5 dataset is a derivative of the SINS dataset [17]. It contains a continuous recording of one person living in a vacation home for one week. Figure 3 shows the arrangement of the 13 microphone arrays used to construct the SINS dataset. Although the DCASE 2018 Challenge Task 5 dataset consists of a development dataset and an evaluation dataset, we used only the development dataset. This is because the evaluation dataset has no information on which microphone recorded each clip in the evaluation dataset. For this reason, we divided the development dataset into three subsets for training, validation, and testing. This partitioning was performed in accordance with a list for cross-validation provided with the dataset.

Sound clips in the DCASE 2018 Challenge Task 5 development dataset were recorded with four microphones called Nodes 1–4 (see Fig. 3). Their length and sampling frequency are unified to 10 s and 16 kHz, respectively. In these sound clips, we utilized clips recorded by Nodes 2, 3, and 4 for this simulation, because the number of clips recorded by Node 1 is different from those recorded by the other nodes.

We prepared three encoding networks and fed clips recorded by Nodes 2, 3, and 4 into the networks. Signals transformed by the networks and propagated through the differentiable physical layers were concatenated and fed into the scene analysis network, where we assumed that a camera was located and fixed at the center of the living room, as shown in Fig. 3. Under this assumption, the distances between the camera and Nodes 2, 3, and 4 were set to 1.13, 1, and 1.62, respectively, with the distance between the camera and Node 3 being 1. These networks were trained with 200 epochs using the training subset and well-known cross-entropy loss. Here, the Adam optimizer [18] was utilized for optimization, where the parameters in Adam were set as $\alpha = 0.001, \beta_1 = 0.9$, and $\beta_2 = 0.999$. The learning rate α was multiplied by 1/10 when the number of epochs reached 100 and 150. The method by He et al. [19] was used for initializing the network. The validation subset was used to check for the overlearning of the networks.

We compared the following four frameworks:

(a) VGG 2D with log-mel spectrogram calculation and lossless transmission (Raw signal / Log-mel energy + VGG 2D in Table I),

(b) VGG 1D without any preprocessing and lossless transmission (Raw signal / VGG 1D in Table I),

(c) Blinky's power calculation in II-A1 + physical layers + VGG 1D (Power / VGG 1D in Table I),

(d) The proposed end-to-end framework, i.e., the CNN-based encoding network + physical layers + VGG 1D (CNN / VGG 1D in Table I).

Note that frameworks (a) and (b) require a wide bandwidth for transmitting raw signals, but frameworks (c) and (d) do not.

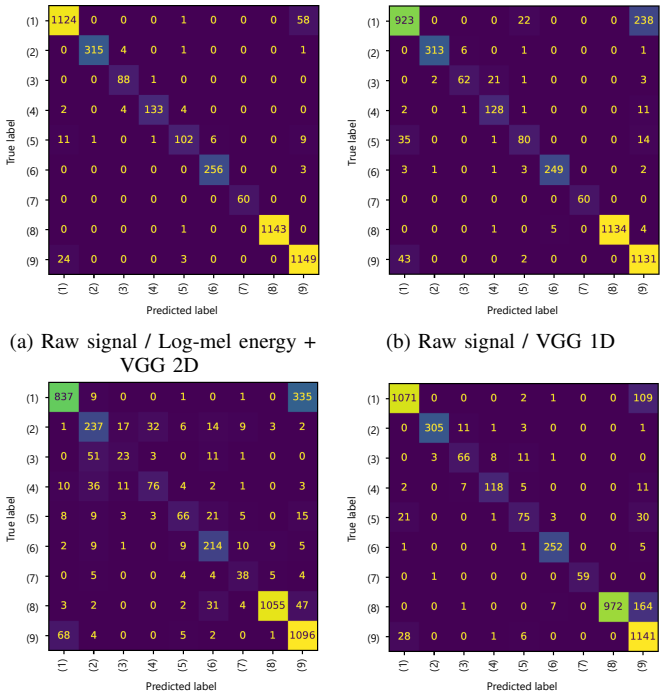
B. Results

Table I shows the classification accuracy for the test subset of the four frameworks. From the table, we can confirm that the proposed framework can achieve a higher accuracy than a non-end-to-end framework considering the same physical phenomena (i.e., "Power / VGG 1D"). In addition, the accuracy of the proposed method was comparable to that of "Raw signal / VGG 1D", while the use of the typical DNN-based approach for acoustic scene analysis, i.e., "Raw signal / Log-mel energy + VGG 2D," provided the highest accuracy among the four frameworks. Note that both "Raw signal / VGG 1D" and "Raw signal / Log-mel energy + VGG 2D" assume an unrealistic situation where distributed microphone arrays are synchronized and high-capacity lossless transmission is possible. Hence, this result suggests the suitability of the proposed end-to-end framework with Blinkies for acoustic scene analysis in practical situations.

More detailed classification results are shown in Fig. 4 as confusion matrices, where each element represents the number of sound clips whose true label is shown on the vertical axis and the predicted label is shown on the horizontal axis. By comparing Figs. 4 (b) and 4 (c), we can see that the number of misclassifications, especially of the cooking, dishwashing, and eating classes, increased owing to the signal propagation process. The proposed end-to-end framework can prevent this performance degradation, as shown in Fig. 4 (d). This figure illustrates that the proposed framework classified 9 out of 10 classes with a higher accuracy than "Power / VGG 1D."

TABLE I: Transmission bandwidth and accuracy for each framework. "Raw signal / Log-mel energy + VGG 2D" and "Raw signal / VGG 1D" require a wide bandwidth for transmitting raw signals, but conventional and proposed frameworks do not.

Encoder / Classifier	Transmission Bandwidth per Mic.	Accuracy
Raw signal / Log-mel energy + VGG 2D	16000 Hz	97.00%
Raw signal / VGG 1D	16000 Hz	90.57%
Power / VGG 1D (Conventional)	30 Hz	81.18%
CNN / VGG 1D (Proposed)	30 Hz	90.01%



(a) Raw signal / Log-mel energy + VGG 2D

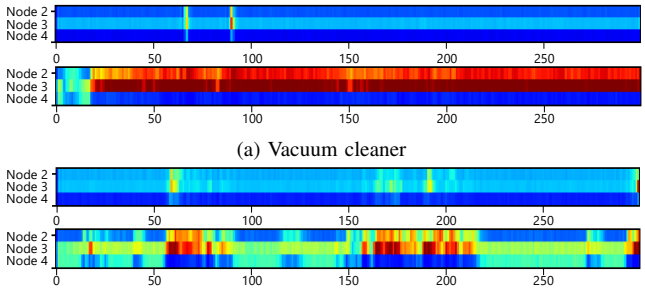
(b) Raw signal / VGG 1D

(c) Power / VGG 1D

(d) CNN / VGG 1D

Fig. 4: Confusion matrices for acoustic scene classification. Class labels are (1) absence, (2) cooking, (3) dishwashing, (4) eating, (5) other, (6) social activity, (7) vacuum cleaner, (8) watching TV, and (9) working.

Figure 5 shows examples of feature maps, i.e., outputs from camera response layers, obtained by "Power / VGG 1D" and "CNN / VGG 1D," where Fig. 5 (a) shows feature maps for a sound clip labeled "vacuum cleaner" and Fig. 5 (b) shows those for a sound clip labeled "social activity." As shown in this figure, the feature maps obtained by the proposed framework were different from the sound power obtained by the conventional framework. In the case of vacuum cleaner, the sound of a vacuum cleaner was heard continuously and the proposed framework almost always produced high feature values for Nodes 2 and 3. By contrast, in the case of



(a) Vacuum cleaner

(b) Social activily

Fig. 5: Examples of feature maps. Top of each subfigure shows feature map of "Power / VGG 1D," and bottom of each subfigure shows feature map of "CNN / VGG 1D." The horizontal axis shows the discrete time index m (video frame index).

social activity, a man talked with a woman and the proposed framework produced high feature values for Nodes 2 and 3 when their voices were loud. In addition, the proposed framework yielded a lower feature value for Node 4 when it provided a higher feature value for Nodes 2 and 3. For these reasons, it is considered that the proposed framework trained encoding networks yielding a high feature value for Nodes 2 and 3 and a low feature value for Node 4 when a meaningful sound for classification was given.

V. Conclusion

In this paper, we proposed an end-to-end acoustic scene analysis framework considering the physical signal propagation process between Blinkies and a camera. In the proposed framework, the use of differentiable physical layers that model physical phenomena as differentiable equations enables us to consider physical constraints in DNNs. As a result, we can train DNNs by an end-to-end approach and can obtain appropriate signal transformations in a data-driven manner. Experimental results showed that the proposed framework provided a higher accuracy than the previous framework with Blinkies for the DCASE 2018 Challenge Task 5 development dataset. The accuracy of the proposed method was comparable to that of a framework that does not consider physical constraints.

This result indicates that the end-to-end training can give us more effective encoding in sound-to-light conversion and estimating the acoustic scene, even with the limitation of camera frame bandwidth. In future work, we will consider more practical conditions, e.g., occlusion of Blinky signals and hardware limitation of Blinkies. We will also collect data using Blinkies in real environments and conduct experiments for acoustic scene analysis in future work.

References

[1] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, "CLEAR Evaluation of Acoustic Event Detection and Classification Systems," in *Multimodal Technologies for Perception of Humans*, Springer Berlin Heidelberg, 2007, pp. 311–322.

[2] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. D. Plumbley, "Detection and Classification of Acoustic Scenes and Events: An IEEE AASP Challenge," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 2013, pp. 1–4.

[3] 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, May 2015.

[4] 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," in *Proceedings of European Signal Processing Conference*, Aug. 2015, pp. 1271–1275.

[5] J. Kürby, R. Grzeszick, A. Plinge, and G. A. Fink, "Bag-of-Features Acoustic Event Detection for Sensor Networks," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop*, Sep. 2016, pp. 55–59.

[6] 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, Jun. 2017.

[7] R. Scheibler, D. Horiike, and N. Ono, "Blinkies: Sound-to-Light Conversion Sensors and Their Application to Speech Enhancement and Sound Source Localization," in *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, Nov. 2018, pp. 1899–1904.

[8] R. Scheibler and N. Ono, "Multi-modal Blind Source Separation with Microphones and Blinkies," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2019, pp. 366–370.

[9] D. Horiike, R. Scheibler, Y. Wakabayashi, and N. Ono, "Blink-Former: Light-Aided Beamforming for Multiple Targets Enhancement," in *Proceedings of IEEE International Workshop on Multimedia Signal Processing*, Sep. 2019, pp. 1–6.

[10] 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.

[11] D. Cherkassky and S. Gannot, "Blind Synchronization in Wireless Acoustic Sensor Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 651–661, Mar. 2017.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. LNCS, vol. 9351, Nov. 2015, pp. 234–241.

[13] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation," in *Proceedings of International Society for Music Information Retrieval Conference*, Jun. 2018.

[14] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv:1409.1556*, Sep. 2014. [Online]. Available: http://arxiv.org/abs/1409.1556

[15] G. Dekkers, P. Karsmakers, and L. Vuegen, "Monitoring of Domestic Activities Based on Multi-Channel Acoustics," 2018. [Online]. Available: http://dcase.community/challenge2018/task-monitoring-domestic-activities

[16] G. Dekkers and P. Karsmakers, "DCASE 2018, Task 5: Monitoring of Domestic Activities Based on Multi-Channel Acoustics - Development Dataset," 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1247102

[17] 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," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop*, 2017, pp. 32–36.

[18] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv:1412.6980*, pp. 1–15, Dec. 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," in *Proceedings of IEEE International Conference on Computer Vision*, Dec. 2015, pp. 1026–1034.