Energy-Based Multiple Source Localization with Blinkies

Daiki Horiike*, Robin Scheibler*, Yuma Kinoshita*, Yukoh Wakabayashi*, and Nobutaka Ono*
* Tokyo Metropolitan University, Tokyo, Japan
E-mail: horiike-daiki@ed.tmu.ac.jp, onono@tmu.ac.jp

***Abstract*—We propose energy-based multiple source localization using sound power sensors called Blinkies that we have recently developed. A Blinky consists of a microphone, LEDs, a microcontroller, and a battery. The intensity of the LED is varied by sound power. Namely, Blinkies work as sound-to-light conversion sensors. They are easy to distribute over a large area, and thus, sound power information can be obtained by capturing the Blinky signals with a video camera. When multiple sources are present, their sounds are mixed, and Blinky signals also reflect the power of the mixture. The idea of the proposed source localization is to decompose a multiple-sources localization problem into "single source localization" problems. More specifically, the Blinky signals can be factorized into transfer function gains and temporal activations by non-negative matrix factorization. Each obtained gain vector is used for estimating each source location. We conduct numerical simulations to evaluate the performance of this method in indoor space like a meeting room. The experimental results show that the proposed framework using Blinkies is effective.**

I. Introduction

Sound source localization is one of the most important tasks in array signal processing for audio/speech processing systems. A typical approach for the sound source localization is to use a microphone array. The microphone array techniques have been developed for a long time [1] and have become essential for source localization [2], speech enhancement [3] via beamforming, and source separation [4], [5]. In general, having more microphones over a large area gives us more spatial sound information. However, it causes more technical challenges, i.e., cable connection with wired communication or network bandwidth limitation through wireless communication. In this paper, we consider an alternative to traditional arrays providing a trade-off between ease of sensor distribution and spatial sound information.

In many practical situations, a video camera is available in addition to the microphones, e.g., in smartphones, teleconference rooms, etc. It is thus possible to purposefully embed sound side information in a video using specially made sensors. One way to embed acoustic information in video frames is sound-to-light conversion. This technique has a long history, i.e., visualization for acoustic holography [6] and communication for acoustic imaging [7]. Recently, it has also been used for the study of frog chorus in the field [8]. Inspired by these ideas to capture acoustic information with a video camera, we also developed sound-to-light conversion sensors called Blinkies [9], [10] (see Fig. 1 left). These sensors measure sound power by a microphone. The sound power is used to modulate the intensity of an on-board light-emitting diode (LED). Finally, a video camera is used to capture the measurements from all Blinkies synchronously. Because Blinkies are battery-powered and do not require cables or wireless networks, they can be easily distributed over a large area. This system is showed in Fig. 1 right.

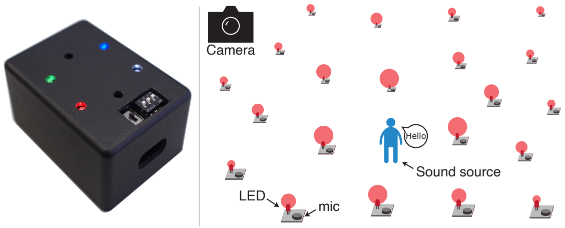

Fig. 1. Left: picture of an actual Blinky sensor. Right: illustration of acoustic sensing system with Blinkies. Blinkies closer to sound source are brighter due to larger sound power. Light signals from all Blinkies are captured by using video camera synchronously.

In previous work, we proposed energy-based single source localization to use a hundred-and-one Blinkies to provide sound power distribution of a single source. We showed how it could be used to perform a live source localization [9], [10]. This method results in precise localization in an environment with high reverberation time and shows the usefulness of sound-to-light conversion. However, when several sources are present, it is not easy to provide sound power distribution of individual target sources because the sound power of their mixture is measured.

In this paper, we propose the energy-based source localization based on the sound power separation in the presence of multiple sources. In this situation, we previously proposed a sound power separation algorithm [11], [12]. A sound power matrix is approximately low-rank and can be factorized into a transfer function matrix and a source activity matrix by non-negative matrix factorization (NMF) [13]. Because the transfer function matrix is expected to consist of the transfer function gains of individual sources, they can be used for estimating the location of each source. In other words, the sound power separation decomposes a multiple-source localization problem into single-source localization problems. A mapping from transfer function gains to a location of a single sound source is learned by a neural network.

The performance of the proposed source-localization method is evaluated by an experiment on simulated signals. We compare this method to a baseline algorithm. The results show that the performance of the proposed method for two-source localization is on par with that for single-source localization. This indicates that the proposed multiple-source localization using Blinkies is effective.

The rest of the paper is organized as follows. Section II describes the details of acoustic sensing with Blinkies and the video camera and how to recover sound power from the video camera measurements. Section III describes the sound power separation algorithm and its application to the source location. Experiments are discussed in Section IV. Section V concludes this paper.

II. Acoustic sensing with Blinkies and video camera

Acoustic sensing with Blinkies and a video camera consists of three parts: sound-to-light conversion in each Blinky, capturing Blinkies' LED light by a video camera, and sound power estimation from video camera measurements. In this section, we first briefly summarise the acoustic sensing procedure. After that, we explain a problem statement in this work.

A. Sound-to-light Conversion

We now explain how the measured sound power is transformed to light intensity by Blinkies. Let n be a discrete time index. A sound power measurement $u[n]$ is computed from a microphone signal $x[n]$ and limited to a range including ambient sound levels. The measured value $u[n]$ is subsequently mapped in a non-linear way to a 12-bit duty cycle of the pulse width modulation (PWM) driving the LEDs. This non-linearity is necessary because of the discrepancy between the dynamic range of natural sounds, over 60 dB, and 8-bit pixel values measured by a video camera. In addition, the function mapping PWM duty cycle to pixel values measured by the video camera was found to be approximately logarithmic. Taking all this into account, we designed a non-linear function $\varphi(\cdot)$ which preserves information including small amplitude components of speech [10].

The sound power measurement $u[n]$ is converted into the B-bit PWM duty cycle, $\ell[n] \in 0, \cdots, 2^B - 1$, by the non-linear function $\varphi(\cdot)$. Then, the actual emitted light intensity $I[n]$ is given by

$$I[n] = \frac{\ell[n]}{2^B - 1} I_{\max}, \quad \ell[n] = \varphi(u[n]), \tag{1}$$

where $I_{\max}$ is the intensity of the LED driven continuously.

B. Capturing Blinkies' Light by Video Camera

After the sound-to-light conversion, LED light from Blinkies propagates in air and it will be captured by a video camera. The LED light intensity at the camera is affected by attenuation α depending on the angle and distance between the LED and the video camera. In addition to this attenuation, ambient light is added to the light intensity as a positive bias β. For these reason, the light intensity $v[n]$ at the video camera is calculated by using attenuation α and bias β as

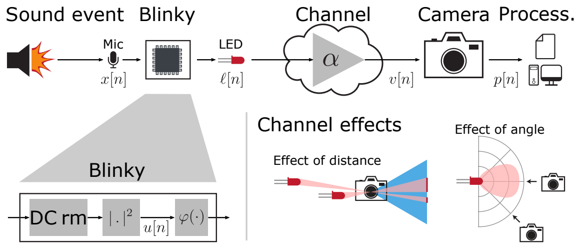

Fig. 2. Top: model of transmission channel from sound event to video file. Bottom left: processing internal to Blinky. Bottom right: channel effects impacting light transmission from Blinkies' LED to video camera.

$$v[n] = \alpha I[n] + \beta. \tag{2}$$

Note that, we assume the attenuation and bias are constants. These assumptions mean that Blinkies and the camera are located at fixed positions and the ambient light intensity is time invariant, respectively.

An imaging sensor on the camera capture the light intensity and the camera encodes it as a video file. Consumer cameras generally process sensor outputs into pixel values, but industrial cameras usually provide raw video frames that directly store sensor output. One of the typical processing is *Gamma correction*. It converts sensor output $v[n]$ so that $p[n] = (v[n])^{1/\gamma}$ with $\gamma = 2.2$. To avoid the non-linear transform, we utilize an industrial camera for capturing signal $v[n]$. Hence, we can assume $p[n] = v[n]$.

C. Sound Power Estimation

Due to the non-linear mapping in (1), the propagation in (2), and the video processing such as *Gamma correction*, captured pixel value $p[n]$ differs from the actual sound power $u[n]$ measured by a Blinky. To reconstruct $u[n]$ from $p[n]$, it is necessary to estimate both α and β. Here, we explain how to reconstruct $u[n]$ from $p[n]$.

There are several ways to attain the calibration depending on having the Blinky transmit a known pilot signal. We now describe a general way to use a second auxiliary LED for calibration. Fig. 2 summarizes the propagation model from sound power to pixel values.

Let ℓ_{sig} and ℓ_{ref} are the PWM duty cycles of the signal and calibration LEDs, respectively. We assume the two LEDs are sufficiently close so that the values of α and β are the same. Then, from (1) and (2), the light intensities obtained by the video camera are

$$p_{\text{sig}}[n] = v_{\text{sig}}[n] = \alpha \frac{\ell_{\text{sig}}[n]}{2^B - 1} I_{\max}^{(\text{sig})} + \beta, \tag{3}$$

$$p_{\text{ref-lo}} = v_{\text{ref-lo}} = \beta, \tag{4}$$

$$p_{\text{ref-hi}} = v_{\text{ref-hi}} = \alpha \frac{\ell_{\text{ref}}}{2^B - 1} I_{\max}^{(\text{ref})} + \beta, \tag{5}$$

where v_{sig} is the signal LED, and $v_{\text{ref-lo}}$ and $v_{\text{ref-hi}}$ are low and high levels of the calibration LED, respectively. From (1)–(5), the estimated α, β, and sound power $\hat{u}_{\mathrm{sig}}$ are

$$\alpha = (p_{\text{ref-hi}} - p_{\text{ref-lo}}) \frac{2^B - 1}{I_{\max}^{(\mathrm{ref})}} \frac{1}{\ell_{\mathrm{ref}}}, \tag{6}$$

$$\beta = p_{\text{ref-lo}}, \tag{7}$$

$$\begin{aligned} \hat{u}_{\mathrm{sig}}[n] &= \varphi^{-1}(\ell_{\mathrm{sig}}[n]) \\ &= \varphi^{-1}\left(\frac{p_{\mathrm{sig}}[n] - p_{\text{ref-lo}}}{p_{\text{ref-hi}} - p_{\text{ref-lo}}} \frac{I_{\max}^{(\mathrm{ref})}}{I_{\max}^{(\mathrm{sig})}} \ell_{\mathrm{ref}} \right). \end{aligned} \tag{8}$$

Note that, due to frame rate limitation of the video camera, frequancy information on the original signal cannot be recoverd.

D. Problem Statement

In this paper, we aim to localize K target sound sources by using M Blinkies for $M > K$. In other words, our aim is to obtain location $\mathbf{r}_k^{(s)} \in \mathbb{R}^3$, $k = 1, 2, \cdots, K$, of the k-th sound source from estimated sound power measurement $\hat{u}_m[n]$ of the m-th Blinky at $\mathbf{r}_m^{(b)} \in \mathbb{R}^3$, $m = 1, 2, \cdots, M$. We consider a scenario where Blinkies are distributed in the room, and a video camera records a scene. We assume that Blinkies are placed far enough apart, and their LED lights are independently captured at different pixels.

For sound source localization, an energy-based localization method, e.g., Chen et al. [2] has been proposed. The sound power is approximately inversely proportional to the distance from a sound source. This is a basis in the energy-based localization. However, when several sources are active at the same time, it is not easy to obtain the sound power of each target source because the sound of sources are mixed and the sound power $u_m[n]$ is then the power of the mixture. In this situation, we cannot directly apply the energy-based localization method to multiple source localization. For solving this problem, we propose using NMF for separating the sound power of each source.

III. MULTIPLE SOURCE LOCALIZATION WITH BLINKIES

A. Overview

The proposed source-localization method based on sound power separation treats a multiple source localization problem as single source localization problems. A sound power matrix that has measurements $\hat{u}[n]$ as elements can be separated into a non-negative matrix that denotes a room transfer function gain and another non-negative matrix that means source activity by NMF. The transfer function gain has features of distance from sound source to Blinkies and provides a sound power distribution in space. Hence, using the transfer function gains obtained via NMF estimates each source location with energy-based source localization algorithm.

B. Sound Power Separation Using NMF

In the short-time Fourier transform (STFT) domain, we suppose that the m-th Blinky measurement $u_m[n]$ is the sum of the sound power over frequency at its location

$$u_m[n] = \sum_{f=1}^{F} \left| \sum_{k=1}^{K} a_{mk}[f] s_k[f, n] \right|^2, \tag{9}$$

where a_{mk} is the room transfer function from k-th source to m-th Blinky, s_k is the k-th source signal, and $f \in \{1, 2, \cdots, F\}$ denotes a frequency. Note that, in practice, this computation is carried out in the time domain.

Assuming that the room transfer function is frequency flat, sources are statistically uncorrelated, and the STFT frame length is longer than the reverberation time. We obtain the following approximation from (9):

$$\begin{aligned} u_m[n] &= \sum_{f=1}^{F} \left| \sum_{k=1}^{K} a_{mk}[f] s_k[f, n] \right|^2 \\ &\approx \sum_{k=1}^{K} \sum_{f=1}^{F} |a_{mk}[f]|^2 \, |s_k[f, n]|^2 \\ &= \sum_{k=1}^{K} g_{mk} \sum_{f=1}^{F} |s_k[f, n]|^2 = \sum_{k=1}^{K} g_{mk} h_k[n], \end{aligned} \tag{10}$$

where $g_{mk} = |a_{mk}[f]|^2$ and $h_k[n] = \sum_{f=1}^{F} |s_k[f, n]|^2$ are the transfer function gain and source activity, respectively. Therefore, in according with this model, the sound power matrix $\boldsymbol{U}$ having rank K can be separated into the non-negative transfer function matrix $\boldsymbol{G}$ and source activity matrix $\boldsymbol{H}$ as a NMF model

$$\boldsymbol{U} \approx \boldsymbol{G}\boldsymbol{H}, \tag{11}$$

where $(\boldsymbol{U})_{mn} = u_m[n]$, $(\boldsymbol{G})_{mk} = g_{mk}$, $(\boldsymbol{H})_{kn} = h_k[n]$.

Since $\boldsymbol{U}$ is only approximately low-rank, we find the factorization minimizing a well-chosen distance function. In general, Euclidean (EUC) distance, Kulback-Leibler (KL), and Itakura-Saito (IS) divergences are used and multiplicative update rules striking a good compromise between speed and ease of implementation are applied to solve these optimization problems [13]–[15]. The update rules of $\boldsymbol{G}$ and $\boldsymbol{H}$ for the EUC distance are given by

$$(\boldsymbol{H})_{kn} \leftarrow (\boldsymbol{H})_{kn} \frac{(\boldsymbol{G}^\top \boldsymbol{U})_{kn}}{(\boldsymbol{G}^\top \boldsymbol{G}\boldsymbol{H})_{kn}}, \tag{12}$$

$$(\boldsymbol{G})_{bk} \leftarrow (\boldsymbol{G})_{bk} \frac{(\boldsymbol{U}\boldsymbol{H}^\top)_{bk}}{(\boldsymbol{G}\boldsymbol{H}\boldsymbol{H}^\top)_{bk}}. \tag{13}$$

Similar update rules exist for the KL and IS divergences.

Note that, additivity of sources in the power domain will be required for NMF. Due to the non-linearity described in II-A, the source signals are not additive when measured as pixel values by the video camera. Hence, to reconstruct correct sound power measurement, sound power estimation in (8) is needed. A similar technique has been previously suggested for far noise reduction in microphone arrays [16].

C. Multiple Source Localization

We will now describe how to use the transfer function gain g_{mk} obtained from NMF for sound source localization. Our approach is to learn a mapping ϕ from transfer function gain g_{mk} to sound source location $\mathbf{r}_m^{(b)}$. We trained neural networks so that they model ϕ. We compare a performance of fully connected neural network (FCNN) and FCNN with residual

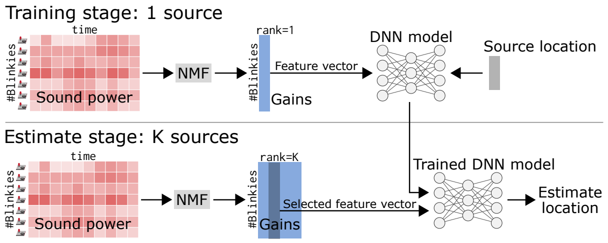

Fig. 3. System flow of proposed source-localization method based on sound power separation. Top: training stage with a single source. A neural network model is trained to learn a mapping from transfer function gains to a source location. Bottom: estimate stage with multiple sources. Source locations are estimated by inputting each transfer function gain into the trained model.

connection (FCNN w/ RC) [17]. FCNN is simply consists of five layers. FCNN w/ RC consists of three bottleneck architecture linked by residual connection sandwiched between input and output fully connected layers. The input layer is followed by 10% dropout for regularization [18]. Rectified linear units (ReLU) are used for activations. The optimizer is the Adam [19]. The loss function is the mean absolute error (MAE) defined as

$$\mathrm{MAE} = \frac{1}{K}\sum_{k}\left\|\hat{\mathbf{r}}_k^{(s)} - \mathbf{r}_k^{(s)}\right\|_1, \tag{14}$$

where $\|\cdot\|_1$ denotes $\ell 1$-norm, and $\hat{\mathbf{r}}_k^{(s)}$ is the estimate of the k-th source location. A flow diagram of the proposed multiple source localization is shown in Fig. 3.

IV. Experiments

A. Simulation setup

We simulated a 5 m × 6 m × 2.5 m room with reverberation time of about 300 ms by using Pyroomacoustics [20]. Twenty Blinkies were simulated by placing microphones on an approximate 4 × 5 grid filling the 4 m × 5 m rectangular area with lower left corner at [1.0, 1.0]. Their height was 1.0 m. We assumed Blinky measurements could be perfectly estimated from pixel values captured by a video camera.

We placed one or two target sources inside this grid at least 0.1 m away from any Blinkies. In the training stage and one of the test conditions, an interfering source was also placed in the same way. Their height was 1.2 m. An illustration of the setup is shown in Fig 4.

The simulation was conducted at a sampling frequency of 16 kHz. Before simulating propagation, the variances of target sources were fixed to $\sigma_k^2 = 1$ at these sources. The signal-to-interference-and-noise ratio (SINR) was defined as

$$\mathrm{SINR} = \frac{\sum_{k=1}^{K}\sigma_k^2}{\sigma_i^2 + \sigma_n^2}, \tag{15}$$

where σ_i^2 and σ_n^2 are the variances of the interfering source and uncorrelated white noise, respectively. We set them so that $\mathrm{SINR} = 60, 10, 5$ dB in training data and $\mathrm{SINR} = 60$ dB in evaluation data. We expect that adding an interference to training data affects the robustness of errors of NMF. Here, we used speech signals as sound source signals. All the speech samples of approximately 20 s were made by concatenating samples from the CMU ARCTIC [21]. The experiment was repeated 1100 times for different attributions of number of sources, speech samples and source locations.

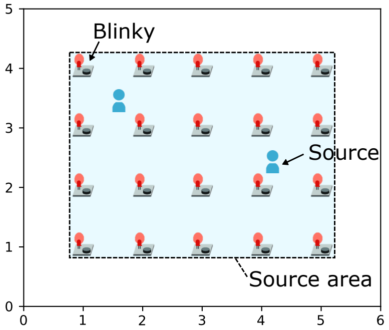

Fig. 4. Illustration of a room geometry and locations of sources and Blinkies in the experiments.

TABLE I
The network structures of FCNN and FCNN w/ RC, respectively.

	FCNN	**FCNN w/ RC**
layers	20-15-10-5-3	20-20-{bottleneck}-20-3
bottleneck architecture	–	20-3-20
number of bottleneck	–	3
dropout	–	0.1

According to the results from previous work [12], we used the EUC distance cost function for NMF. In this experiment, the number of NMF basis vectors was set to the exact number of sources. When the number of sources is unknown, we could determine the number of basis vectors by the number of larger singular values of the sound power matrix $\boldsymbol{U}$. The NMF updates ran for 100 iterations. The number of training and validation data were 1000 and 100 examples, respectively. The neural network was trained with Pytorch [22] for 1000 epochs and with a mini-batch size of 16. Table. I shows details of neural network structure.

The localization algorithms using FCNN and FCNN w/ RC were compared with a baseline method that estimates the k-th source location $\hat{\mathbf{r}}_k^{(s)}$ as the location $\mathbf{r}_\mu^{(b)}$ of the brightest Blinky, namely,

$$\hat{\mathbf{r}}_k^{(s)} = \mathbf{r}_\mu^{(b)}, \tag{16}$$

where

$$\mu = \arg\max_m g_{mk}. \tag{17}$$

The ground-truth signals of separation were obtained by simulating each source separately. We tested three conditions: 1 source, 1 source plus 1 interference with same SINR for training (for example, when SINR at training was 10 dB, that at test was also 10 dB.), and 2 sources, and compared them in order to verify whether the NMF works appropriately.

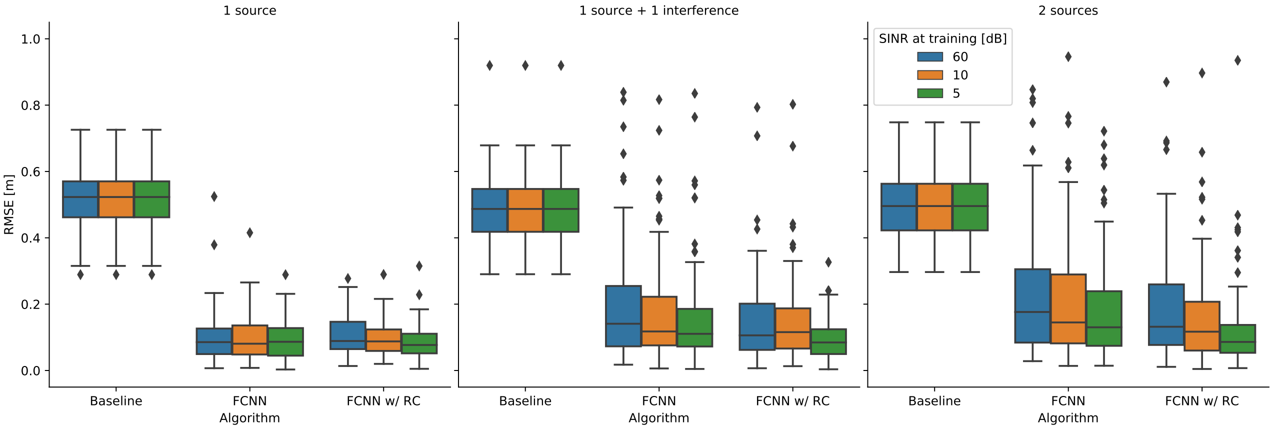

Fig. 5. Box-plot of root mean square error (RMSE) of estimated source locations for 1 source, 1 source plus 1 interference with same SINR for training, and 2 sources, respectively, from left to right. Source localization algorithms are shown on horizontal axis. The signal-to-interference-and-noise ratio (SINR) in the training stage is shown in legend. Boxes span from first to third quartile, referred to as Q_1 and Q_3, and whiskers show maximum and minimum values in range of $[Q_1 - 1.5(Q_3 - Q_1), Q_3 + 1.5(Q_3 - Q_1)]$. Band inside boxes indicates median.

B. Results

We evaluated the localization error respect to the root mean square error (RMSE). The lower the RMSE, the closer the estimated location are to ground-truth. The distribution of RMSE of validation dataset is illustrated as box-plots in Fig. 5.

Overall, the error range of the training-based algorithm is much more compact than that of the baseline algorithm in all conditions. In the case of one-source localization, the difference in SINR at training did not affect the performance. However, the results in the one-source plus one interference and two sources show that the training with adding interference, especially at low SINR, much reduces the errors. Among them, FCNN w/ RC shows the best performance, and in this case, the performance of the proposed method for two-sources localization is on par with that for single-source localization. This means that the proposed source-localization can treat multiple source localization as single-source localization. From these results, we confirm that multiple source localization with Blinkies is effective.

V. Conclusion

We proposed energy-based multiple source localization to estimate target sources using Blinkies and a video camera. Since Blinkies can be distributed over a large area, they provide spatial information of sound. The idea of the proposed source localization is to decompose a multiple-source localization problem into single-source localization problems based on the NMF. Using transfer function gains of individual sources obtained via NMF, we estimated each source location with a neural network. We evaluated the performance of the source localization in simulations and confirmed the effectiveness of the proposed method, particularly when the neural network is trained in a low SINR condition.

Our future work will focus on evaluating the performance of the proposed method for real-recorded data in reverberant conditions.

Acknowledgment

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

References

[1] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*, vol. 1. Springer Science & Business Media, 2008.

[2] M. Chen, Z. Liu, L.-W. He, P. Chou, and Z. Zhang, "Energy-based position estimation of microphones and speakers for ad hoc microphone arrays," in *Proc. IEEE WASPAA*, pp. 22–25, 2007.

[3] H. L. Van Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. John Wiley & Sons, 2004.

[4] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. IEEE WASPAA*, pp. 189–192, 2011.

[5] N. Ono, "Fast stereo independent vector analysis and its implementation on mobile phone," in *Proc. IEEE IWAENC*, pp. 1–4, 2012.

[6] W. E. Kock, *Seeing Sound*. Wiley, 1971.

[7] G. P. Nava, H. D. Nguyen, Y. Kamamoto, T. G. Sato, Y. Shiraki, N. Harada, and T. Moriya, "A high-speed camera-based approach to massive sound sensing with optical wireless acoustic sensors," *IEEE Trans. Comp. Imaging*, vol. 1, no. 2, pp. 126–139, 2015.

[8] I. Aihara, T. Mizumoto, T. Otsuka, H. Awano, K. Nagira, H. G. Okuno, and K. Aihara, "Spatio-temporal dynamics in collective frog choruses examined by mathematical modeling and field observations," *Scientific reports*, vol. 4, no. 3891, 2014.

[9] R. Scheibler, D. Horiike, and N. Ono, "Blinkies: Sound-to-light conversion sensors and their application to speech enhancement and sound source localization," in *Proc. APSIPA*, pp. 1899–1904, 2018.

[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. 67603–67616, 2020.

[11] R. Scheibler and N. Ono, "Multi-modal blind source separation with microphones and blinkies," in *Proc. IEEE ICASSP*, pp. 366–370, 2019.

[12] D. Horiike, R. Scheibler, Y. Wakabayashi, and N. Ono, "Blink-former: Light-aided beamforming for multiple targets enhancement," in *Proc. IEEE MMSP*, pp. 1–6, 2019.

[13] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. NIPS*, pp. 556–562, 1999.

[14] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[15] M. Nakano, H. Kameoka, J. Le Roux, Y. Kitano, N. Ono, and S. Sagayama, "Convergence-guaranteed multiplicative algorithms for nonnegative matrix factorization with β-divergence," in *Proc. IEEE MLSP*, pp. 283–288, 2010.

[16] Y. Matsui, S. Makino, N. Ono, and T. Yamada, "Multiple far noise suppression in a real environment using transfer-function-gain NMF," in *Proc. IEEE EUSIPCO*, pp. 2314–2318, 2017.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," pp. 770–778, June 2016.

[18] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, June 2014.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv.org*, Dec. 2014.

[20] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. IEEE ICASSP*, pp. 351–355, 2018.

[21] J. Kominek and A. W. Black, "CMU ARCTIC databases for speech synthesis," Tech. Rep. CMU-LTI-03-177, Language Technologies Institute, School of Computer Science, Carnegie Mellon University, 2003.

[22] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in neural information processing systems*, pp. 8026–8037, 2019.