Compressed Sensing of Sparse Spectrum Using Distributed Sound-to-Light Conversion Device Blinkies

Satoshi Motoyama*, Natuki Ueno*, Yuma Kinoshita*†, and Nobutaka Ono*
* Tokyo Metropolitan University, Tokyo, Japan
E-mail: motoyama-satoshi@ed.tmu.ac.jp
† Tokai University, Kanagawa, Japan

***Abstract*—In this study, we propose a method for estimating a sparse spectrum of sound using multiple sound-to-light conversion devices called Blinkies. A Blinky is a compact device that converts sound information into the light with a programmable operation. By distributing multiple Blinkies and monitoring them with a video camera, we can obtain acoustic information from a wider area without wired or wireless communication. However, due to the limited frame rate of the video camera, the bandwidth of the observed video signals is significantly narrower than that of the original sound. To restore the spectrum of an original sound signal, we design the entire signal acquisition process with multiple Blinkies and a video camera to be regarded as compressed sensing. We show that a sparse sound spectrum can be restored with the proposed method by numerical experiments.**

I. Introduction

Sound is one of the most informative media in scene analysis and recognition [1]–[3]. Acoustic scene analysis using distributed microphones has been much studied in the literature with a focus on spatial information as well as temporal information of the sound [4]. For example, several competitions on acoustic scene analysis, such as the DCASE challenge, have been held in recent years, and various state-of-the-art methods have been proposed and evaluated [5]–[7].

One difficulty in acoustic scene analysis based on distributed microphones is communication and synchronization of the data observed by the multiple microphones. A *Blinky*, a compact, battery-powered sound-to-light conversion device proposed by Sheibler and Ono [8], is a promising solution for such problems. A Blinky is equipped with a microphone and a light emitting diode (LED). An acoustic signal measured by the microphone is converted to a light signal from the LED via a programmable sound-to-light conversion process. By distributing multiple Blinkies over space and observing them with a single video camera, it is easy to capture synchronized data emitted from distributed Blinkies without wired or wireless communication.

Various acoustic sensing frameworks using Blinkies have been proposed, and their practical effectiveness has been validated in numerical and real-world experiments [9]–[13]. In most methods, the intensity of the light signal is determined by the short-time power of the observed sound signal, which is mainly because the frame rate of a video camera is generally much lower than the sampling rate of a microphone (for example, 30 Hz against 16 kHz). Therefore, in contrast to spatial information, temporal information of the target sound is lost significantly in such an observation system. To improve the accuracy and applicability of the system, further investigation regarding the acquisition of temporal information has been desired.

In this study, we propose a method for estimating the sparse temporal spectrum of the target sound using distributed Blinkies by designing their suitable sound-to-light conversion processes, which is the first attempt in the literature. Spectra of artificial sounds, such as alarm sounds of household appliances, are expected to be sparse, and the proposed method can be used to detect such artificial sounds, for example. In the proposed framework, the Blinkies' sound-to-light conversion processes are designed so that the entire observation system corresponds to the compressed sensing [14]–[16] from the spectrum of the source signal to the observed light signal. Finally, the spectrum of the source signal is estimated by solving a sparse optimization problem. Here, we propose an iterative algorithm based on the proximal gradient method [17], which guarantees convergence to the global optimal solution. Numerical experiments were conducted to evaluate the proposed method, and their results indicated that the spectrum of various types of source signals was estimated successfully using the distributed Blinkies.

II. Problem Setting

Suppose multiple Blinkies are distributed in a certain environment and observed by a video camera, as in Fig. 1, where the Blinkies' positions need not be given. When a sound is generated from some sound source, each Blinky converts the captured sound to a light signal via a programmable conversion process. Here, the target sound is assumed to be sparse in the frequency domain, which is satisfied well for various artificial and natural sounds. Our objective is to estimate the normalized amplitude spectrum of the target sound from the observed video signal. The Blinkies are not synchronized or connected with a network, but their light signals are regarded to be synchronized since they are captured by a single video camera. In addition, the video camera and Blinkies are assumed to be calibrated so that the output light signal of each Blinky can be obtained directly from the video signal (see [11] for a calibration method).

III. Compressed Sensing with Sound-to-Light Conversion Devices

This section describes the observation model relating the sound and light signals and the method for estimating the original sound signal from the observed light signals.

A. Observation Model

Suppose an L-sample short-time source signal $\mathbf{s} \in \mathbb{R}^L$ is observed by M Blinkies. Let $\mathbf{s}_m \in \mathbb{R}^L$ be the observed sound signal at the mth Blinky. The one-sided amplitude spectra of $\mathbf{s}$ and $\mathbf{s}_m$ obtained by the discrete Fourier transform (DFT) are denoted by $\mathbf{x} \in \mathbb{R}^N$ and $\mathbf{x}_m \in \mathbb{R}^N$, respectively ($N = \lfloor L/2+1 \rfloor$). By assuming the reverberation and sound traveling time are small enough, $\mathbf{x}_m$ is given by

$$\mathbf{x}_m = a_m \mathbf{x}, \tag{1}$$

where $a_m \geq 0$ is a coefficient representing the acoustic attenuation determined by the distance between the source to the mth Blinky.

Then, by defining $h_m(\cdot)$ as the mth Blinky's sound-to-light conversion, the light signal y_m captured by the video camera is given by

$$y_m = h_m(\mathbf{x}_m) + n_m, \tag{2}$$

where n_m denotes the observation noise. Here, we assume the video camera is calibrated so that we can directly obtain the output light signal $\mathbf{x}_m$.

B. Proposed Sound-to-Light Conversion in Blinky

We design the sound-to-light conversions in Blinkies so that the whole observation system corresponds to that of linear compressed sensing, whose reconstruction method is well established. Here, we consider the compressed sensing of the normalized amplitude spectra $\mathbf{x}' \in \mathbb{R}^N$ defined as

$$\mathbf{x}' = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}, \tag{3}$$

instead of $\mathbf{x}$ because the coefficient a_m is generally difficult to obtain.

For this purpose, we design the sound-to-light conversion function $h_m(\cdot)$ as

$$h_m(\mathbf{x}_m) = \boldsymbol{\psi}_m^\mathsf{T} \frac{\mathbf{x}_m}{\|\mathbf{x}_m\|_2} \tag{4}$$

with a certain weight vector $\boldsymbol{\psi}_m \in \mathbb{R}^N$. Then, from (1), (2), and (3), the output signal $y_m \in \mathbb{R}$ is represented as

$$y_m = \boldsymbol{\psi}_m^\mathsf{T} \mathbf{x}' + n_m. \tag{5}$$

Furthermore, by stacking (5) for $m = 1, \dots, M$, we obtain

$$\mathbf{y} = \boldsymbol{\Psi} \mathbf{x}' + \mathbf{n} \tag{6}$$

where $\mathbf{y} = [y_1, \dots, y_M]^\mathsf{T}$, $\boldsymbol{\Psi} = [\boldsymbol{\psi}_1, \dots, \boldsymbol{\psi}_M]^\mathsf{T}$, and $\mathbf{n} = [n_1, \dots, n_M]^\mathsf{T}$.

C. Amplitude Spectrum Estimation

Our objective is to estimate the normalized amplitude spectrum $\mathbf{x}'$ from the observed light signal $\mathbf{y}$ under the sparsity assumption on $\mathbf{x}'$. This problem can be formulated as

$$\underset{\mathbf{x}' \in R^N}{\text{minimize}} \ \frac{1}{2}\|\mathbf{y} - \boldsymbol{\Psi}\mathbf{x}'\|_2^2 + \lambda\|\mathbf{x}'\|_1 \ \text{s.t.}\ \mathbf{x}' \geq 0, \tag{7}$$

where $\lambda > 0$ is a regularization parameter and the constraint $\mathbf{x}' \geq 0$ means each element of $\mathbf{x}'$ is nonnegative. This is a convex optimization problem and can be solved by the proximal gradient method, which is derived as follows. First, the optimization problem (7) can be rewritten as

$$\underset{\mathbf{x}' \in R^N}{\text{minimize}} \ \frac{1}{2}\|\mathbf{y} - \boldsymbol{\Psi}\mathbf{x}'\|_2^2 + \lambda\|\mathbf{x}'\|_1 + \Gamma(\mathbf{x}'), \tag{8}$$

where

$$\Gamma(\mathbf{z}) = \begin{cases} 0 & \mathbf{z} \geq 0 \\ \infty & \text{otherwise} \end{cases}. \tag{9}$$

Since the first term is differentiable and the sum of the second and third terms is proximable, the update rule for the proximal gradient method is given by

$$\mathbf{x}' \leftarrow S_{\gamma\lambda}(\mathbf{x}' - \gamma\boldsymbol{\Psi}^\mathsf{T}(\boldsymbol{\Psi}\mathbf{x}' - \mathbf{y})), \tag{10}$$

where $S_{\gamma\lambda}(\cdot)$ is the proximal operator of the sum of the second and third terms of (8), which is given by the elementwise operation of the scalar function $T_{\gamma\lambda}(\cdot)$ defined as

$$T_{\gamma\lambda}(z) = \begin{cases} z - \gamma\lambda & z \geq \gamma\lambda \\ 0 & z < \gamma\lambda \end{cases}. \tag{11}$$

To guarantee convergence, the step size parameter γ has to be set as

$$0 < \gamma < \frac{2}{\lambda_{\max}(\boldsymbol{\Psi}^\mathsf{T}\boldsymbol{\Psi})}, \tag{12}$$

where $\lambda_{\max}(\cdot)$ denotes the maximum eigenvalue of the matrix. This algorithm can be interpreted as a slightly modified form of the Iterative Shrinkage Soft-thresholding Algorithm (ISTA) [17], whose modification is due to the nonnegative constraint (9).

IV. Experiment

To verify the performance of the proposed framework in estimating amplitude spectra, the accuracy was evaluated by simulation experiments.

A. Ideal Condition

First, the ideal condition where the observed signals were generated exactly in accordance with (5) was simulated. The number of Blinkies was $M = 30$, and several source signals with $L = 512$ ($N = 257$) samples whose sampling rate was 16 kHz were investigated. Each element of $\boldsymbol{\Psi}$ was determined randomly to follow the Gaussian distribution with mean 0 and variance 1. The noise n_m is sampled independently for each m from the Gaussian distribution with mean 0, where several variances were investigated. The noise level for the observed

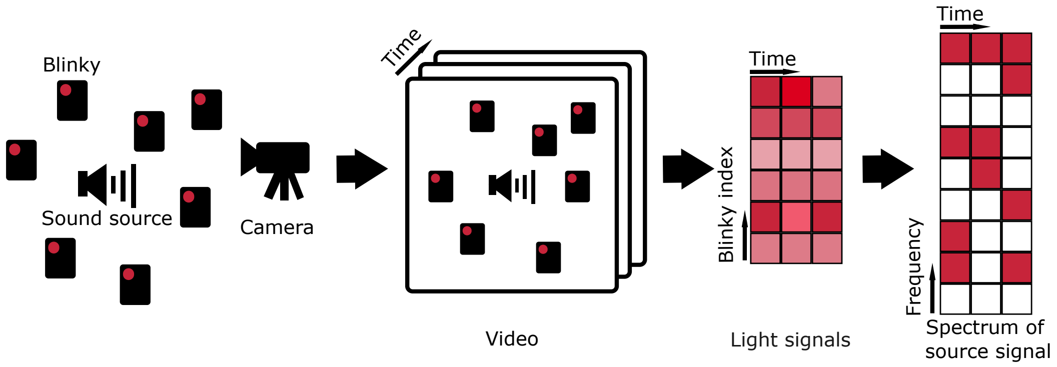


Fig. 1. Proposed sensing framework.

signal was evaluated by the Input Signal-to-Noise Ratio (Input SNR), defined as

$$\text{Input SNR} = 10\log_{10}\frac{\|\mathbf{y}-\mathbf{n}\|_2^2}{\|\mathbf{n}\|_2^2}. \tag{13}$$

On the other hand, the estimation accuracy was evaluated by the Output Signal-to-Noise-Ratio (Output SNR), defined as

$$\text{Output SNR} = 10\log_{10}\frac{\|\mathbf{x}'\|_2^2}{\|\hat{\mathbf{x}}'-\mathbf{x}'\|_2^2}, \tag{14}$$

where $\hat{\mathbf{x}}' \in \mathbb{R}^N$ denotes the estimated value of $\mathbf{x}'$. Here, $\hat{\mathbf{x}}'$ was obtained using the algorithm given by (10) for 20000 iterations with $\gamma = \frac{1}{\lambda_{\max}(\boldsymbol{\Psi}^\top\boldsymbol{\Psi})}$.

Figure 2 shows the relationship between λ and the Output SNR for several different Input SNRs. In this figure, a sin wave whose frequency was 1 kHz was used as the source signal. Each value in Fig. 2 is the average of the Output SNR for 50 trials with respect to the random weight matrix $\boldsymbol{\Psi}$ and noise $\mathbf{n}$. We can see that $\lambda = 1$ achieved high Output SNR for each Input SNR. From this result, we fixed $\lambda = 1$ in the following experiments.

Table I shows the average of Output SNR against the Input SNR for seven different source signals. Each value in Table I is the average of the Output SNR for 50 trials with respect to the random weight matrix $\boldsymbol{\Psi}$ and noise $\mathbf{n}$. Here, instrumental sounds (monophonic sounds) from SMILE2004 [18] were used as the source signals. We can see that the Output SNR was improved significantly as the Input SNR increases to 20 dB, but the improvement becomes smaller from 20 dB to 30 dB.

The ℓ_1 norm for each instrumental sound was 4.2906 for Violin, 5.0590 for Cello, 5.0974 for Flute, 3.4606 for Piccolo, 3.2908 for Oboe, 4.3401 for Horn, and 5.1886 for Trumpet. The relationship between these ℓ_1 norms and the Output SNRs given in Table I indicates that the proposed method was suitable for estimating a sound with a sparse spectrum.

For further investigation, the true and estimated amplitude spectra for two different source signals at Input SNR = 20 dB were plotted in Figs. 3a and 3b. Also from these figures, we can see that the spectrum of the Piccolo sound, which was relatively sparse, was estimated more accurately than that of the Trumpet sound.

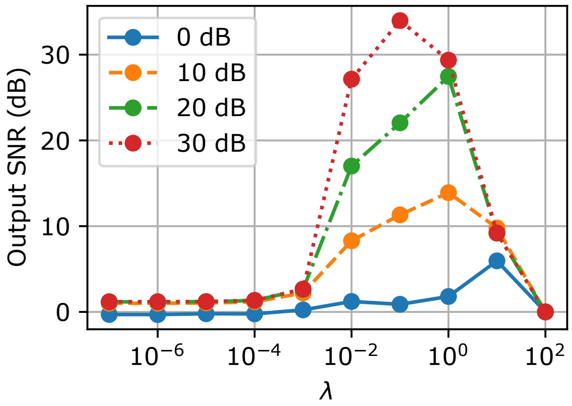


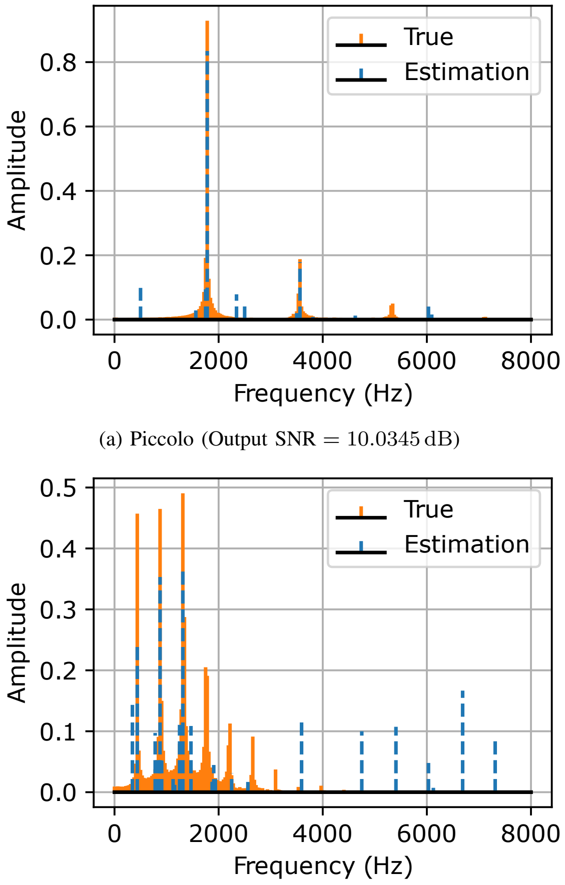
Fig. 2. Output SNR for each regularization parameter λ. The legend indicates Input SNR.

TABLE I
OUTPUT SNR (dB) FOR EACH INSTRUMENTAL SOUND IN IDEAL CONDITION

Input SNR	0 dB	10 dB	20 dB	30 dB
Violin	0.6445	5.1774	6.5376	6.7998
Cello	-0.3978	4.1895	4.9960	5.0392
Flute	-0.5325	3.7114	4.9322	5.0168
Piccolo	0.9665	7.4542	9.2871	9.7334
Oboe	-0.3774	6.7760	9.8458	10.4685
Horn	-0.4375	4.8808	6.7914	7.1207
Trumpet	-1.1470	1.8885	2.7205	2.8437
Average	-0.1768	4.8683	6.4444	6.7174

B. Simulated Room Environment

Next, similar evaluations were conducted in a simulated room environment. In this experiment, the sound transfer from the source to the Blinkies was simulated using Pyroomacoustics [19]. In this case, because of the time difference between Blinkies and the room reverberation, the observed signal $\mathbf{y}$ does not follow (5) exactly. A two-dimensional model with a size of $5\,\mathrm{m}\times 5\,\mathrm{m}$ was simulated. The reflection coefficient and the maximum number of reflections were determined so that the reverberant time was RT60 = 300 ms. A signal source was located in the center of the room, and the $M = 30$ Blinkies were located randomly as shown in Fig. 4. Other experimental



(a) Piccolo (Output SNR $= 10.0345\,\mathrm{dB}$)

(b) Trumpet (Output SNR $= 3.2932\,\mathrm{dB}$)

Fig. 3. True and estimated spectra in ideal condition for Input SNR $= 20\,\mathrm{dB}$.

conditions, such as Blinky's sampling rates, were the same as in an ideal condition.

Table II shows the Output SNR against the Input SNR for seven different source signals. Each value in the Table II is the average of the Output SNR for 50 trials with respect to the random weight matrix $\mathbf{\Psi}$ and noise $\mathbf{n}$. By comparing Table I and Table II, we can see that the Output SNR generally became lower in the presence of reverberation and time delay. However, the proposed method achieved the Output SNR of around $5\,\mathrm{dB}$ when the Input SNR was $20\,\mathrm{dB}$ or $30\,\mathrm{dB}$ even under the influence of reverberation and time delay.

The true and estimated amplitude spectra for two different source signals at Input SNR $= 20\,\mathrm{dB}$ were plotted in Figs. 5a and 5b. Output SNR generally decreased in a simulated room environment compared to the ideal conditions, but roughly similar trends were observed between the different instruments.

Finally, a linear chirp signal whose frequency varied from $1\,\mathrm{Hz}$ to $8\,\mathrm{kHz}$ over five seconds was estimated by using the proposed method for each 512 sample without overlap. The true and estimated spectrograms are shown in Fig. 6 and 7. We can see the linear chirp in Fig. 6 and 7, which indicates the proposed method was able to reconstruct major features of the target spectrogram.

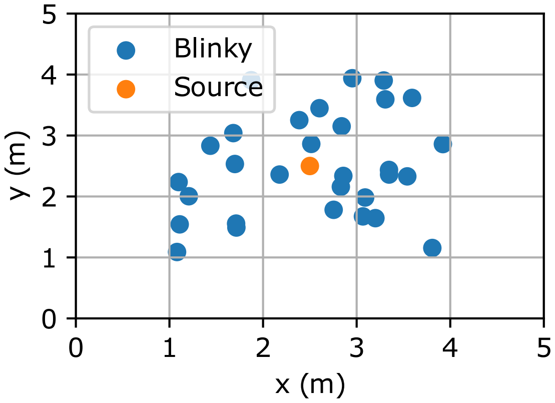


Fig. 4. Blinky and source location in simulated room environment.

TABLE II
OUTPUT SNR (dB) FOR EACH INSTRUMENTAL SOUND IN SIMULATED ROOM ENVIRONMENT

Input SNR	0 dB	10 dB	20 dB	30 dB
Violin	0.1344	4.9662	6.0851	6.0363
Cello	-0.9870	3.4329	4.3210	4.4029
Flute	-0.2573	3.6423	4.5894	4.7577
Piccolo	0.6508	6.4194	8.3880	8.5764
Oboe	-0.9431	4.3936	5.7117	6.0426
Horn	-0.6805	4.0543	5.6573	5.6959
Trumpet	-1.0468	1.6835	2.5766	2.6313
Average	-0.4471	4.0846	5.3327	5.4490

V. Conclusion

In this paper, we proposed a method for estimating the sparse spectrogram of sound using multiple Blinkies. In the proposed method, we designed the entire signal acquisition process with the Blinkies to be regarded as compressed sensing. Experimental results showed that the proposed method was able to estimate major spectrum components in an ideal condition and a simulated room environment. We will investigate the effect of the number and placement of Blinkies and evaluate the performance of the proposed framework in a real-world environment in future work.

Acknowledgment

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

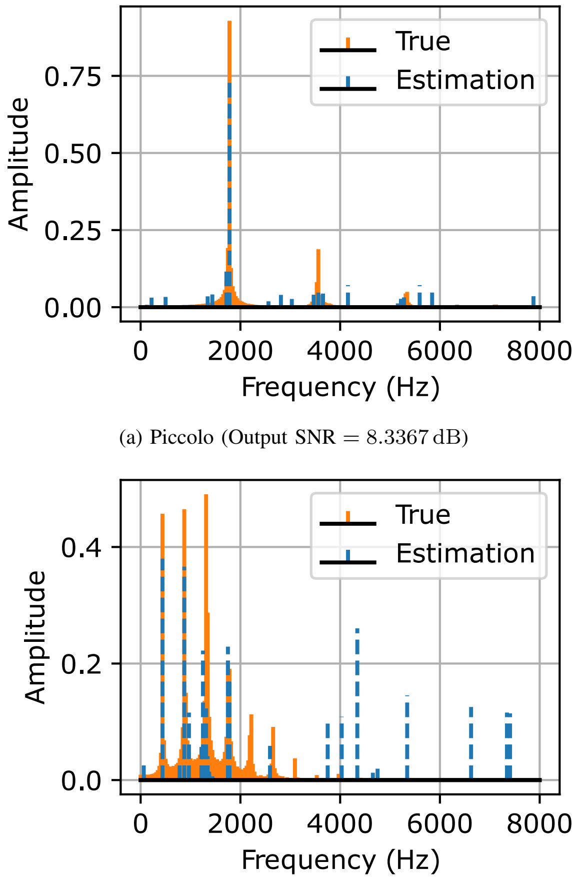
References

[1] D. Barchiesi, D. Giannoulis, Dan S., 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.

[2] J. Kürby, R. Grzeszick, A. Plinge, and G. A. Fink, "Bag-of-features acoustic event detection for sensor networks.," in *Proc. the Detection and Classification of Acoustic Scenes and Events 2016 Workshop*, Sep. 2016, pp. 55–59.

[3] A. Mesaros, T. Heittola, T. Virtanen, and M. D. Plumbley, "Sound event detection: A tutorial," *IEEE Signal Processing Magazine*, vol. 38, no. 5, pp. 67–83, 2021.

[4] 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.



(a) Piccolo (Output SNR $= 8.3367\,\mathrm{dB}$)

(b) Trumpet (Output SNR $= 2.6180\,\mathrm{dB}$)

Fig. 5. True and estimated spectra in simulated room environment for Input SNR $= 20\,\mathrm{dB}$.

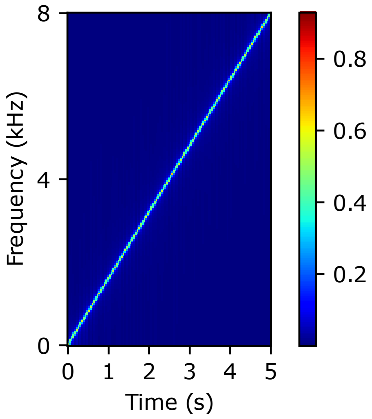


Fig. 6. True amplitude spectrogram of linear chirp signal in simulated room environment.

[5] A. Temko, R. Malkin, C. Zieger, D. Macho, Climent Nadeu, and Maurizio Omologo, “CLEAR evaluation of acoustic event detection and classification systems,” in *Proc. International Evaluation Workshop on Classification of Events, Activities and Relationships*, 2007, pp. 311–322.

[6] 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 *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 2013, pp. 1–4.

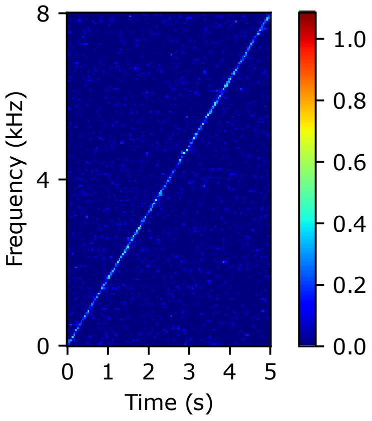


Fig. 7. Estimated amplitude spectrogram of linear chirp signal in simulated room environment.

[7] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, “Detection and classification of acoustic scenes and events,” *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, 2015.

[8] 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.

[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. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, Nov. 2018, pp. 1899–1904.

[10] R. Scheibler and N. Ono, “Multi-modal blind source separation with microphones and blinkies,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2019, pp. 366–370.

[11] D. Horiike, R. Scheibler, Y. Kinoshita, Y. Wakabayashi, and N. Ono, “Energy-based multiple source localization with Blinkies,” in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, Dec. 2020, pp. 443–448.

[12] Y. Kinoshita and N. Ono, “End-to-end training for acoustic scene analysis with distributed sound-to-light conversion devices,” in *Proc. European Signal Processing Conference*, Aug. 2021, pp. 1010–1014.

[13] K. Ishii, Y. Kinoshita, Y. Wakabayashi, and N. Ono, “Real-time pitch visualization with “Blinky” sound-to-light conversion device,” *Journal of Signal Processing*, vol. 25, no. 6, pp. 213–220, 2021.

[14] R. G. Baraniuk, “Compressive sensing [lecture notes],” *IEEE Signal Processing Magazine*, vol. 24, no. 4, pp. 118–121, 2007.

[15] A. Griffin and P. Tsakalides, “Compressed sensing of audio signals using multiple sensors,” in *Proc. European Signal Processing Conference*, Aug. 2008, pp. 1–5.

[16] M. Rani, S. B. Dhok, and R. B. Deshmukh, “A systematic review of compressive sensing: Concepts, implementations and applications,” *IEEE Access*, vol. 6, pp. 4875–4894, 2018.

[17] A. Beck and M. Teboulle, “A fast iterative shrinkage-thresholding algorithm for linear inverse problems,” *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[18] K. Kawai, K. Fujimoto, T. Iwase, H. Yasuoka, T. Sakuma, and Y. Hidaka, “Development of a sound source database for environmental/architectural acoustics: Introduction of SMILE 2004 (Sound Material in Living Environment 2004),” in *Proc. International Congress of Acoustics*, Apr. 2004, pp. 1561–1564.

[19] R. Scheibler, E. Bezzam, and I. Dokmanić, “Pyroomacoustics: A Python package for audio room simulation and array processing algorithms,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Apr. 2018, pp. 351–355.