Blinkies: Sound-to-light conversion sensors and their application to speech enhancement and sound source localization

Robin Scheibler, Daiki Horiike, and Nobutaka Ono
Tokyo Metropolitan University
E-mail: robin@tmu.ac.jp Tel: +81-42-585-8418

***Abstract*—We introduce the use of blinkies for acoustic sensing and audio processing applications. Blinkies are low-power sensors with a microphone and an LED that can be easily distributed over a large area. The power of the LED is modulated by the sound intensity and the signals from all devices can be captured by a regular video camera. We present our design for blinkies and characterize the transmission function from blinky to captured video signal. The usefulness of such a system is demonstrated with two applications. First, we evaluate beamforming informed by a high-quality voice activity signal obtained from a blinky. Second, we investigate sound source localization using several blinkies distributed in a room.**

I. Introduction

The effectiveness of microphone arrays has been largely demonstrated for speech enhancement via beamforming [1], [2], source separation [3], source localization [4] and tracking [5], room geometry inference [6], and dereverberation [7]. Microphone arrays come in many shapes and sizes, from the simple stereo microphones in smartphones and computers, to arrays with tens or hundreds of microphones such as the Pyramic array [8] or that of Perrodin et al., [9], to over a thousand channels in the LOUD array [10]. All these arrays perform synchronous sampling of all the channels, a pre-requisite of most of the above-mentioned techniques. Their design and construction is both costly and challenging. Recently, clever sampling frequency mismatch compensation algorithms have enabled the use of distributed, asynchronous microphone arrays [11], and alleviated somewhat the cost of multichannel processing. Yet, these arrays come with their own set of limitations, e.g., network bandwidth and latency issues.

In this paper, we explore a different paradigm of multi-channel acoustic sensing. We design a simple, inexpensive embedded device that records sound using a microphone and converts it into a luminous signal via a light emitting diode (LED). We call it a *blinky*. Blinkies can be spread over a large area and their signal recorded synchronously by a conventional video camera, as illustrated in Fig. 1. Such a system can easily scale to hundreds or thousands of channels without many technical hurdles and minimal setup time. However due to the low frequency of off-the-shelf video cameras – typically 30 to 60 Hz – the type of processing and problems it can tackle differs significantly from conventional array processing. Fig. 4b shows a frame extracted from a video where twenty blinkies can be seen reacting to four sound sources.

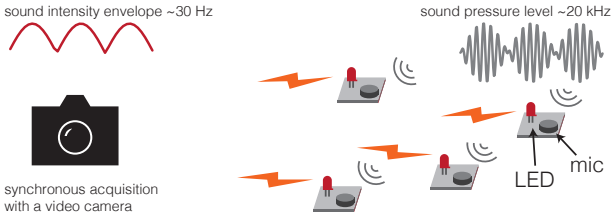


Fig. 1: Diagram of the blinky acoustic sensing paradigm.

Using light as a medium for acoustic sensing has been explored in the past for visualization [12] and communication [13]. More recently, an analog blinky design and various algorithms were proposed to study frog chorus [14]. Our intent in this work is to lay the groundwork to open the door to a wider range of applications of blinkies. We show that recent advances in low-cost embedded devices enable the design of a considerably more flexible platform for the blinkies. We believe this can be useful to the community and make the design openly available[1]. Finally, we showcase two applications of blinkies – first to speech enhancement in conjunction with a conventional microphone array, second to sound source localization. Some early results of this paper have been presented at a meeting of the Acoustical Society of Japan [15]. In this paper, we further extend them with a detailed description of our now completed blinky design, new and extended experiments on light-aided beamforming, and more details in general.

The rest of this paper is organized as follows. Section II describes the sensor design and the characterization of the channel. Section III and Section IV describe the two example applications: beamforming with VAD side-information, and energy-based localization, respectively. Section V concludes this paper.

This work was supported by a JSPS post-doctoral fellowship and grant-in-aid (№17F17049), and the SECOM Science and Technology Foundation.

In the spirit of reproducible research, the code used to produce the results in this paper is available at https://github.com/onolab-tmu/otohikari.

[1]Design files and code available at https://github.com/onolab-tmu/blinky.

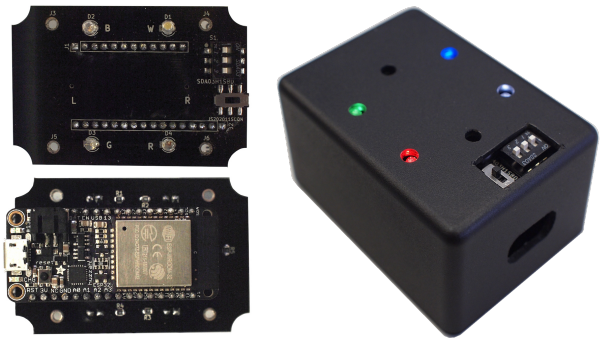


Fig. 2: A blinky sound-to-light conversion sensor. Left: top and bottom of the circuit board with the two microphones and four LEDs. Right: the assembled blinky with enclosure.

II. Sound-to-Light Sensor Design

In this section, we present the design and operation of the blinky sensor.

A. Hardware Design

Our design for the blinky sound-to-light conversion sensor benefits from the recent developments in affordable, yet powerful, embedded processing platforms. After reviewing several options, we settled on the ESP32 system-on-chip microcontroller from Espressif systems [16]. The following features make it an attractive platform for a low-cost distributed audio processing platform.

- Dual core RISC CPU at up to 240 MHz
- Floating-point processing unit
- Two I2S audio buses
- Low-power modes
- Wifi and Bluetooth
- A large online developer community and comprehensive documentation
- Unit price around USD 5 in small quantities

While Wifi and Bluetooth are not used by the blinkies, they allow the platform to be reused as a wireless microphone for asynchronous array processing.

To simplify the design process, we used the HUZZAH32 platform that adds to the ESP32 all the necessary power regulation, programming circuit, as well as a Lithium Ion/Polymer USB battery charger. An extension circuit board with two digital microphones, four LEDs, and a few switches for power and configuration, custom ordered for the blinky . The two digital MEMS microphones (ICS-43432) can connect directly to the ESP32 thanks to their integrated I2S interface, greatly simplifying the design. The LEDs are selected for their high brightness at low current consumption. The four LEDs are all chose of a different colors (red, green, blue, and white) to potentially exploit the different color channels of cameras. The extension board was designed to fit in a commercially available enclosure. The circuit as well as the assembled blinky can be seen in Fig. 2.

B. Operations

While different schemes are possible, we investigate the direct modulation of the LED with the power of the sound. To this end, the ESP32 is programmed to continuously acquire sound from the microphone. The variance of blocks of 64 consecutive samples is computed and mapped in a non-linear way to the range of the pulse width modulation (PWM) driving the LED. The non-linearity is necessary due to the large range of amplitudes in natural sounds. Using a linear mapping from the audio PCM range (24 bit) to the PWM range (12 bit) was empirically confirmed to discard too much useful information from the lower amplitude components of sound. In addition to this problem, the transfer function from PWM duty cycle to measured pixel intensity with a commercial camera was measured and found to be approximately logarithmic in the PWM duty cycle (Fig. 3, bottom left).

In this work, we use an empirically derived mapping that preserves information from small amplitudes components of speech and takes into account the non-linearity of the PWM-to-pixel transfer function of the system. The mapping is a composition of two functions. The first one is derived from the empirical cumulative distribution function (CDF) of the variance of blocks of 64 samples of natural speech. This CDF was estimated using the whole training set of the TIMIT corpus [17]. Applying the CDF, shown in Fig. 3, top, to the input data makes its distribution uniform, thus maximizing the entropy of the signal transmitted. This way, a larger range of values of the PWM duty cycles are allocated to amplitudes of speech that are most frequent. The second map applied is simply the inverse of the PWM-to-pixel transfer function. A measurement of the transfer function of the system using the non-linear mapping just presented is shown in Fig. 3, bottom right.

In the next two sections, we demonstrate the usefulness of the blinky paradigm though two applications. First, we show how using a single blinky together with a microphone array can dramatically enhance conventional beamforming in a challenging scenario. Second, we evaluate the potential of blinkies for sound source localization based on energy only in an indoor scenario.

III. Application I: Blinky-informed Beamforming

In this application, we consider the use of blinkies (and camera) together with a conventional microphone array system. A single blinky is placed in the vicinity of the target sound source and can thus be used to provide reliable voice activity detection (VAD). The VAD is subsequently used to compute the optimum beamforming filters as explained next.

A. Maximum SINR Beamforming

As usual, we work in the time-frequency domain and suppose the microphone signal $\boldsymbol{x}$ is a mixture of the target source $\boldsymbol{s}$, a number of interferers $\{\boldsymbol{z}_q\}_{q=1}^{Q}$, and noise $\boldsymbol{b}$, i.e.,

$$\boldsymbol{x}(t,\omega) = \boldsymbol{s}(t,\omega) + \sum_{q=1}^{Q} \boldsymbol{z}_q(t,\omega) + \boldsymbol{b}(t,\omega), \qquad (1)$$

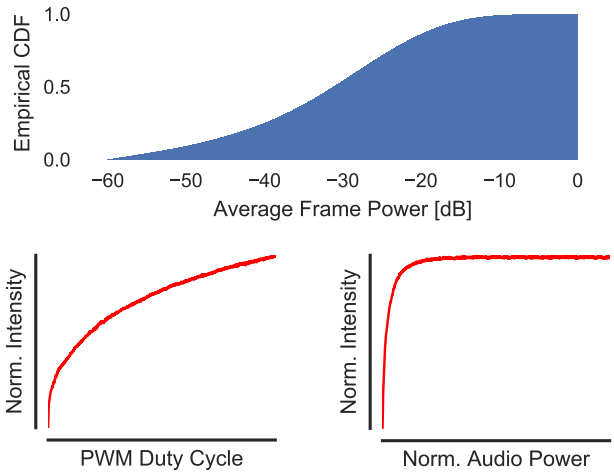


Fig. 3: Top: the empirical cumulative distribution function of the variance of short speech blocks. Bottom left: measured transfer function from PWM duty cycle to measured pixel intensity. Bottom right: measured transfer function between input sound variance to measured pixel intensity with the non-linear mapping applied in the blinky.

where t is the frame index, and ω the frequency. Given beamforming weight vector $\boldsymbol{w}(\omega)$, defining $\boldsymbol{z}(t,\omega) := \sum_q \boldsymbol{z}_q(t,\omega)$, and further omitting the frame index and frequency for clarity, the *signal-to-interference-and-noise* ratio (SINR) at the output of the beamformer is defined as

$$\mathsf{SINR}(\boldsymbol{w}) = \frac{\mathbb{E}|\boldsymbol{w}^H \boldsymbol{s}|^2}{\mathbb{E}|\boldsymbol{w}^H(\boldsymbol{z}+\boldsymbol{b})|^2} = \frac{\boldsymbol{w}^H \mathbf{R}_s \boldsymbol{w}}{\boldsymbol{w}^H \mathbf{R}_{zb} \boldsymbol{w}}, \tag{2}$$

where $\mathbf{R}_s$ and $\mathbf{R}_{zb}$ are the covariance matrices of signal and interference-and-noise, respectively. The aptly named maximum SINR (Max-SINR) beamformer is chosen so as to maximize this ratio,

$$\boldsymbol{w}_{\mathrm{M-SINR}} = \arg\max_{\boldsymbol{w}} \mathsf{SINR}(\boldsymbol{w}) = \arg\max_{\boldsymbol{w}} \frac{\boldsymbol{w}^H \mathbf{R}_x \boldsymbol{w}}{\boldsymbol{w}^H \mathbf{R}_{zb} \boldsymbol{w}}, \tag{3}$$

where in the last equality we replaced $\mathbf{R}_s$ by $\mathbf{R}_x = \mathbb{E}|\boldsymbol{x}|^2$ as this only changes the ratio up to a constant additive factor [1]. Provided $\mathbf{R}_x$ and $\mathbf{R}_{zb}$ are known, the $\boldsymbol{w}_{\mathrm{M-SINR}}$ is the eigenvector corresponding to the largest generalized eigenvalue for the generalized eigenvalue problem $\mathbf{R}_x \boldsymbol{w} = \lambda \mathbf{R}_z \boldsymbol{w}$.

In practice, the covariance matrices are unknown and are replaced by their sample estimates $\widehat{\mathbf{R}}_x$ and $\widehat{\mathbf{R}}_{zb}$. While the former can be computed from the input signal, there is usually no good estimate for the latter, which is why the Max-SINR beamformer is seldom used. This is where the blinky enters the stage. The blinky placed close to the target source detects when it is active, and $\mathbf{R}_{zb}$ can be estimated from frames where it is not. Let the VAD function be $\mathsf{VAD}(t) = 1$ if the target source is active in the t-th frame, and zero otherwise. Then, the covariance matrices estimators are

$$\widehat{\mathbf{R}}_x(\omega) = \sum_{t=1}^{N} \boldsymbol{x}(t,\omega)\boldsymbol{x}(t,\omega)^H, \tag{4}$$

$$\widehat{\mathbf{R}}_{zb}(\omega) = \sum_{t=1}^{N} (1-\mathsf{VAD}(t))\boldsymbol{x}(t,\omega)\boldsymbol{x}(t,\omega)^H. \tag{5}$$

These are then used in place of their expected values in (3). Note that this scheme is completely data-driven and doesn't require any extra information about the location of sources or microphones.

B. Experiment Setup

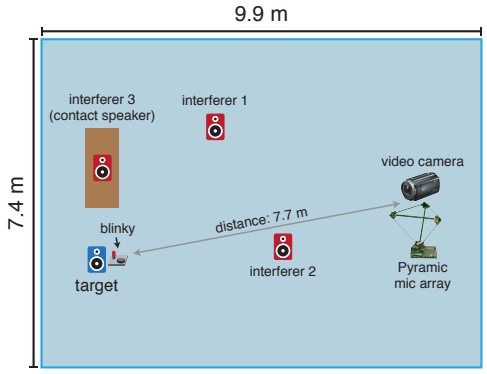
We evaluated the scheme just described in a practical experiment. We placed four loudspeakers in an office $9.9\,\mathrm{m}$ by $7.4\,\mathrm{m}$ with a T60 of $0.3\,\mathrm{s}$, each playing a different sound. The first three sources are natural speech extracted from the CMU ARCTIC corpus [18]. The fourth one is a contact speaker on a $0.75\,\mathrm{m}$ by $1.8\,\mathrm{m}$ table playing factory noise extracted from the BBC Sound Effects archive [19], thus acting as an extended sound source. All the sound samples have a duration of around $15\,\mathrm{s}$. The distance between the target sound source, placed at the first loudspeaker, and the microphones was approximately $7.7\,\mathrm{m}$. The blinky was placed directly on top of the target source. A diagram of the setup as well as a picture taken during the experiment are shown in Fig. 4.

A calibration step was used to measure the gain of each source, which was subsequently compensated to play the signal at a pre-determined *signal-to-interference-ratio* (SIR)

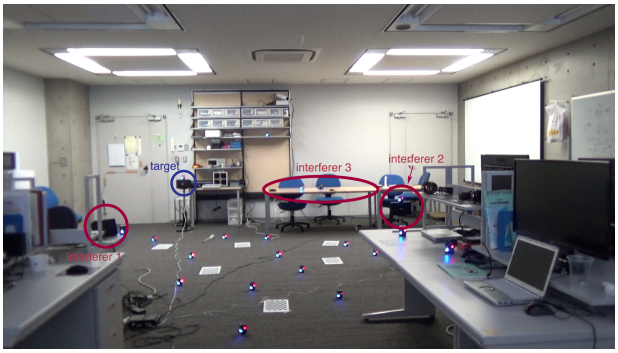
$$\mathsf{SIR} = \frac{\sigma_s^2}{3\sigma_z^2} \tag{6}$$

where σ_s^2 is the power of the target source and all three interferers have the same power σ_z^2. The power of the interferers σ_z^2 is obtained by fixing the target source power to $\sigma_s^2 = 1$. At record time, both the mix of all sound sources and each source alone were recorded, resulting in five segments per target SIR. To avoid synchronization mismatch between the mix and reference samples, they were first all concatenated in one audio file and all recorded in a single session. A known sequence of white noise was played at the beginning of the recording session to mark precisely the beginning. The signal from the blinky was captured at 60 frames-per-second with a Sony HDR-CX535 camera. The pixels in an 11×11 patch around the location of the LED in the video frame were averaged to obtain a more reliable blinky signal. The sound was recorded using the Pyramic 48-channel microphone array [8].

All recorded signals were downsampled at $16\,\mathrm{kHz}$ to match the sampling frequency of the speech samples. The VAD signal used to compute the Max-SINR beamformer was obtained by thresholding the blinky signal using an empirically determined threshold. The same value of the threshold was used at all SIR. To make certain no target signal was mixed in the estimation of the interference-and-noise covariance matrix, the voice-active intervals were extended by 3000 samples on both sides. Blind source separation using independent vector analysis



(a) Source and receiver locations



(b) Frame extracted from the video

Fig. 4: (a) Illustration of the four sound sources, video camera, and microphone array in the room. (b) A frame extracted from the video recorded with source locations highlighted. The blinkies can be spotted thanks to their LEDs.

(AuxIVA) [3] was also applied as a baseline algorithm. Both Max-SINR and AuxIVA were performed in the short time Fourier transform (STFT) domain with frame size of 2048 samples, half overlap, a Hann analysis window, and matched synthesis window. For both algorithms, the scale ambiguity was resolved by the so-called projection-back method [20]. Max-SINR was tested on subsets of 2, 4, 24, and 48 channels of the Pyramic array. AuxIVA is specialized for the determined source separation case (as many microphones as sources) and only was thus only tested on 2 and 4 channels, as the evaluation with more channels is ambiguous.

C. Results

The performance of the algorithms is evaluated using the source separation metrics of *signal-to-distortion* ratio (SDR) and SIR [21]. The metrics are computed using their implementation in the `mir_eval` Python package [22].

Fig. 5 shows the results from the evaluation. Because of the ambient noise in the reference recording used for SDR and SIR computations, there is a small discrepancy between the target and actual SIR. To account for this, we evaluate the SDR and SIR of the input *mix* signal as well. We first describe the results in terms of SIR which characterizes the level of separation achieved. For two channels, AuxIVA improves the SIR between 0 and 3 dB, while Max-SINR achieves 4 to 5 dB. Note that at very low input SIR (-5 and 0 dB), Max-SINR manages 3-4 dB improvement whereas AuxIVA completely fails. For four channels, we have 3-4 dB and 8-10 dB improvements using AuxIVA and Max-SINR, respectively. Here, AuxIVA performs slightly better than Max-SINR at -5 and 0 dB input SIR. At all other cases, however, Max-SINR does around 5 dB better than AuxIVA. With 24 and 48 channels, performance is shockingly good with over 15 and 25 dB improvements, respectively, at the lowest input SIR values. When increasing the input SIR, the output SIR tends to plateau around 30 dB.

Regarding the SDR, at low input SIR, the SDR is improved by both AuxIVA and Max-SINR compared to the input mix. Going to higher input SIR, the output SDR tends to saturate around 7.5-8 dB for all algorithms. The exception is AuxIVA with 4 channels that saturates around 5 dB. Something surprising is that increasing the number of channels leads to worse SDR. Informal listening to the output signals[2] revealed that using 24 and 48 channels reduces ambient noise so dramatically that a mismatch appears with the reference recording of the target source (which also contains ambient noise). While further investigation is required, we believe this might be the cause of this discrepancy.

IV. Application II: Energy-based Localization

The second application we investigate is sound source localization. We consider a number of blinkies spread in an indoor location and a camera recording the scene is used to obtain the amplitude of the sound at each blinky via the intensity of its LED (Fig. 4b shows a picture of such a setup). In scenarios where the area is not so large, time of flight methods cannot be used due to the low sampling frequency of the camera. Instead, we adapt an energy-only algorithm from Chen et al. [23]. The modification is needed because we will assume the locations of the blinkies to be known. This is justified since they could be recovered from the camera recording using computer vision techniques [24]. We first describe the modified algorithm, then the simulation setup, and finally discuss the result of the evaluation.

A. Energy-based Localization Algorithm

We consider a scenario with K sources and M blinkies distributed in a room. As mentioned earlier, the locations $\{\boldsymbol{r}_m\}_{m=1}^{M}$ of the blinkies are assumed to be known. Following the original algorithm [23], we use a simple attenuation model for the energy received from source k, located at $\boldsymbol{s}_k$, by blinky m

$$a_{mk} = \frac{g_m p_k}{\|\boldsymbol{r}_m - \boldsymbol{s}_k\|^{2\alpha}}, \tag{7}$$

where g_m is the gain of the sensor, and p_k the power of the source. The exponent α characterizes the decay due to

[2] Available at: http://www.robinscheibler.org/apsipa2018.

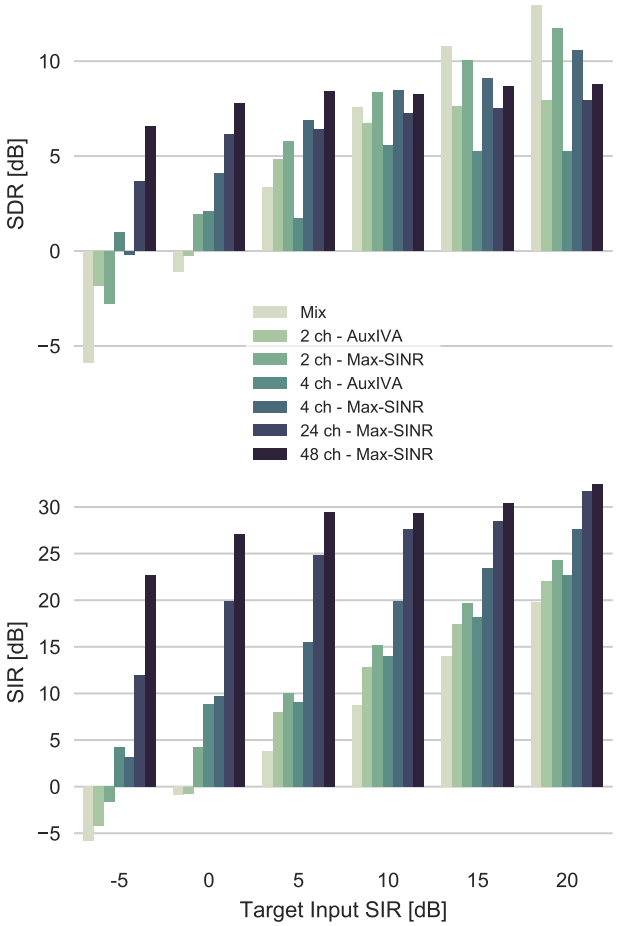


Fig. 5: The evaluation in terms of SDR and SIR of the input mix, the output of Max-SINR beamforming, and AuxIVA blind source separation, for 2, 4, 24, and 48 channels.

propagation, in anechoic conditions $\alpha = 1$. The original work considered a typical office to have approximately $\alpha \approx 0.5$ [23]. In contrast, we will try to estimate α together with the rest of the unknowns.

Assuming noise is normally distributed in the log domain, the maximum likelihood estimator is obtained by solving the following minimization problem

$$\min_{\alpha,\tilde{g}_m,\tilde{p}_k,\boldsymbol{s}_k} \sum_{m=1}^{M}\sum_{k=1}^{K}(\tilde{a}_{mk} - \tilde{g}_m + \alpha \log\|\boldsymbol{r}_m - \boldsymbol{s}_k\|^2 - \tilde{p}_k)^2 \quad (8)$$

where variables with a tilde are the log of their counterpart without it (e.g., $\tilde{a}_{mk} = \log a_{mk}$). This is a non-linear least-squares problem with a potentially large number of local minima. A good initialization is thus crucial. We follow Chen et. al., [23] for the initialization of $\tilde{g}_m$, $\tilde{p}_k$, and $d_{mk} = \|\boldsymbol{r}_m - \boldsymbol{s}_k\|^2$, for all m, k. Note that their initialization scheme assumes that each sound source is somewhat in the vicinity of one of the sensors. Whereas multidimensional scaling was

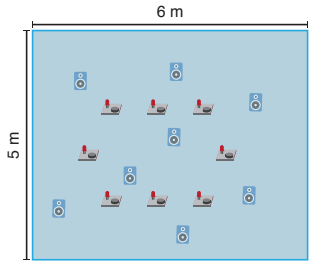


Fig. 6: The setup of the localization simulation.

originally used to recover $\boldsymbol{r}_m$, $\boldsymbol{s}_k$ from d_{mk}, knowing $\boldsymbol{r}_m$ lets us use the more powerful squared ranged least-squares (SRLS) method [25]. The only problem is the scale mismatch between the $\boldsymbol{r}_m$ (e.g., given in meters) and d_{mk} whose unit is unknown. This is addressed by modifying SRLS to also solve for the unknown scaling factor:

$$\boldsymbol{s}_k = \arg\min_{\boldsymbol{s},\rho} \sum_{m=1}^{M} \left(\|\boldsymbol{s} - \boldsymbol{r}_m\|^2 - \rho d_{mk}^2\right)^2. \quad (9)$$

Just like the original SRLS, this problem can be solved globally despite its non-convexity.

Starting at this initial estimate, the problem (8) is solved with the Levenberg-Marquardt algorithm through its implementation in the `least_squares` function from the `scipy` package [26]. Because it was empirically noticed that we do not always converge to a good solution, the obtained solution is perturbed with a small quantity of noise and the solver is restarted from the new position. This process is repeated a hundred times. The solution with smallest cost is chosen.

B. Simulation Setup

We evaluate the localization algorithm just described through numerical experiments. Simulations of indoor sound propagation are carried out using the `pyroomacoustics` Python package [27]. A 6 m by 5 m two-dimensional virtual room is created with eight blinkies configured as shown in Fig. 6. Eight sound sources are placed at random, but each in the vicinity of one of the blinkies. Namely, the distance of source k to blinky m is normally distributed with unit mean and standard deviation $\sigma = 0.2$. The experiment is repeated one thousand times with different source placements.

C. Results

The localization errors for all source placements are aggregated into the histogram of Fig. 7. We obtain that over all source placements, the median localization is just 6 cm, that is 1% of the room width. The 90-th percentile is slightly under 30 cm. In 7.45% of all cases, the method fails and the error is larger than 50 cm. Since there are eight sources, this can be interpreted as localizing seven of them on average.

V. Conclusion

We investigated the use of blinkies, sound-to-light conversion sensors, for acoustic sensing. We proposed a versatile

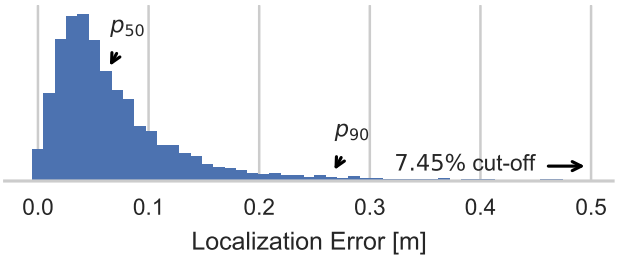


Fig. 7: Distribution of the energy-based localization error. The median is 6 cm, and 7.45% of all samples are outside the plot, spread between 50 cm and 25 m.

blinky architecture that is low-cost, low-power, versatile, and can potentially be reused in the context of asynchronous wireless microphone arrays. We demonstrated that the blinky sensing paradigm is suitable for a wide range of applications. We presented the result of a practical experiment where a blinky is used to obtain high-quality voice activity information. The resulting beamforming yields state-of-the-art performance in terms of SDR and SIR. Next, a preliminary simulation-based experiment suggests that blinkies can be successfully used for sound source localization.

So far, we have focused on single-source scenarios for simplicity. In the future, we will investigate ways of demixing blinky signals created by multiple sources so that the algorithms developed here can be applied. Another benefit would be to alleviate the requirement that a blinky be present in the vicinity of the source for the blinky-informed beamforming.

References

[1] H. L. Van Trees, *Optimum Array Processing*. New York, USA: John Wiley & Sons, Inc., Mar. 2002.

[2] I. Dokmanić, R. Scheibler, and M. Vetterli, "Raking the cocktail party," *IEEE J. Sel. Top. Signal Process.*, vol. 9, no. 5, pp. 825–836, 2015.

[3] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," 2011.

[4] H. Krim and M. Viberg, "Two decades of array signal processing research: the parametric approach," *IEEE Signal Process. Mag.*, vol. 13, no. 4, pp. 67–94, Jul. 1996.

[5] E. Weinstein, "Optimal source localization and tracking from passive array measurements," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 30, no. 1, pp. 69–76, 1982.

[6] I. Dokmanić, R. Parhizkar, A. Walther, Y. M. Lu, and M. Vetterli, "Acoustic echoes reveal room shape," *Proc. Natl. Acad. Sci.*, vol. 110, no. 30, pp. 12 186–12 191, Jun. 2013.

[7] E. A. P. Habets and J. Benesty, "A two-stage beamforming approach for noise reduction and dereverberation." *IEEE Trans. Audio, Speech & Language Processing*, 2013.

[8] R. Scheibler, J. Azcarreta, R. Beuchat, and C. Ferry, "Pyramic: Full stack open microphone array architecture and dataset," in *IWAENC*, 2018, submitted.

[9] F. Perrodin, J. Nikolic, J. Busset, and R. Y. Siegwart, "Design and calibration of large microphone arrays for robotic applications," in *Proceedings of the IEEE/RSJ IROS*. 03737 - Siegwart, Roland Y., 2012, pp. 4596 – 4601.

[10] E. Weinstein, K. Steele, A. Agarwal, and J. Glass, "LOUD: a 1020-Node Microphone Array and Acoustic Beamformer," in *ICSV*, Cairns, Australia, Jul. 2007.

[11] 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, Feb. 2015.

[12] W. E. Kock, *Seeing sound*. John Wiley & Sons, Inc., 1971.

[13] G. Pablo Nava, H. Duy 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, Jun. 2015.

[14] 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," *Sci. Rep.*, vol. 4, no. 3891, 2014.

[15] R. Scheibler and N. Ono, "Audio processing with ad-hoc array of blinkies and a camera," in *Proc. of the Acoustical Society of Japan Spring Meeting*, 2018.

[16] Espressif Systems, "ESP32 datasheet," 2018. [Online]. Available: https://www.espressif.com/sites/default/files/documentation/esp32_datasheet_en.pdf

[17] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

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

[19] "BBC sound effects," http://bbcsfx.acropolis.org.uk/, last accessed 21 May 2018.

[20] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1-4, pp. 1–24, Oct. 2001.

[21] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, "First stereo audio source separation evaluation campaign: data, algorithms and results," in *International Conference on Independent Component Analysis and Signal Separation*. Springer, 2007, pp. 552–559.

[22] C. Raffel, B. McFee, E. J. Humphrey, J. Salomon, O. Nieto, D. Liang, D. P. W. Ellis, C. C. Raffel, B. Mcfee, and E. J. Humphrey, "mir_eval: A transparent implementation of common MIR metrics," in *Proc. ISMIR*, 2014.

[23] 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*, 2007.

[24] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*, 2nd ed. Cambridge: Cambridge Univ. Press, 2003.

[25] A. Beck, P. Stoica, and J. Li, "Exact and approximate solutions of source localization problems," *IEEE Trans. Signal Process.*, vol. 56, no. 5, pp. 1770–1778, Apr. 2008.

[26] E. Jones, T. Oliphant, P. Peterson *et al.*, "SciPy: Open source scientific tools for Python," 2001–, [Online; accessed September 10, 2018]. [Online]. Available: http://www.scipy.org/

[27] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A Python package for audio room simulations and array processing algorithms," in *Proc. IEEE ICASSP*, Calgary, CAN, 2018.