Capturing Sound by Light: Towards Massive Channel Audio Sensing via LEDs and Video Cameras

Gabriel Pablo Nava, Yoshifumi Shiraki, Hoang Duy Nguyen, Yutaka Kamamoto, Takashi G. Sato, Noboru Harada, and Takehiro Moriya

Abstract

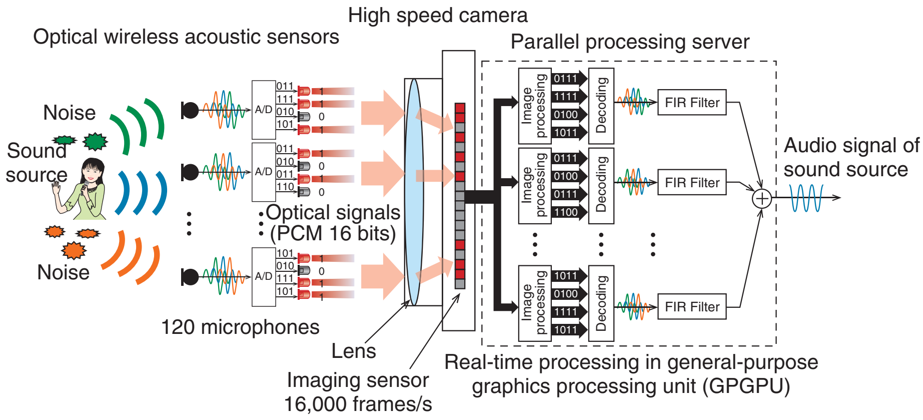
We envision the future of sound sensing as large acoustic sensor networks present in wide spaces providing highly accurate noise cancellation and ultra-realistic environmental sound recording. To achieve this, we developed a real-time system capable of recording the audio signals of large microphone arrays by exploiting the parallel-data transmission feature offered by free-space optical communication technology based on light-emitting diodes and a high speed camera. Typical audio capturing technologies face limitations in complexity, bandwidth, and the cost of deployment when aiming for large scalability. In this article, we introduce a prototype that can be easily scaled up to 120 audio channels, which is the world's first and largest real-time optical-wireless sound acquisition system to date.

Keywords: microphone array, free-space optical communication, beamforming

1. Introduction

Imagine the TV broadcast of a live event taking place in a noisy, wide open environment. At the user end, it is often desired to have not only high quality image reproduction but also highly realistic sound that gives a clear impression of the event [1]. This can be achieved by the use of microphone arrays. According to the theory of sensor array signal processing [2], it is possible to listen to a particular sound from a desired location, and also to suppress the noise from the surroundings, by properly aligning and mixing the audio signals recorded by a microphone array. The theory also indicates that large microphone arrays produce remarkable sound enhancement. Examples have been demonstrated with arrays of 100 microphones [3]. However, to accurately record a three-dimensional sound field at a rate of up to 4 kHz and impinging from every direction on a 2-m^2 wall of a room, an array of about 2500 microphones would be needed. Moreover, if the space under consideration is larger, for example a concert hall, tens of thousands of microphones might be necessary. Unfortunately, typical wired microphones and audio recording hardware have limitations in terms of complexity and cost of deployment when the objective is large scalability. Furthermore, the use of multiple wireless microphones is constrained by radio frequency (RF) bandwidth issues.

To overcome these difficulties, we developed a prototype that allows the simultaneous capture of multichannel audio signals from a large number of microphones (currently up to 120). In contrast with existing RF wireless audio interfaces, the proposed system relies on free-space optical transmission of digital signals. Such technology allows the parallel transmission of multiple data channels, each with full bandwidth capacity regardless of the number of channels transmitted.


High speed camera
Optical wireless acoustic sensors
Parallel processing server
Noise
Sound source
Noise
A/D
A/D
A/D
Optical signals (PCM 16 bits)
120 microphones
Lens
Imaging sensor 16,000 frames/s
Image processing
Decoding
FIR Filter
FIR Filter
FIR Filter
Audio signal of sound source
Real-time processing in general-purpose graphics processing unit (GPGPU)

A/D: analog to digital
FIR: finite impulse response
PCM: pulse code modulation

Fig. 1. Architecture of the multichannel audio acquisition system.

2. System description

Our system is composed of three main parts: 1) an optical wireless acoustic sensor (OWAS), 2) a high speed camera, and 3) a parallel processing server. The overall architecture of the system is illustrated in **Fig. 1**.

2.1 OWAS device

An OWAS device is shown in **Fig. 2**. The microphone picks up samples of the acoustic waves at a rate of 16 kHz and outputs a delta-sigma modulated digital stream. Then, a microcontroller converts that serial data into binary symbols of 16-bit pulse code modulation (PCM). The PCM symbols are used to light up an array of 16 light-emitting diodes (LEDs). An LED in the ON or OFF state means a binary 1 or 0, respectively. Because the camera can observe several OWASs simultaneously, the sound field can be sensed with a large array of OWASs such as the one shown in **Fig. 3**, where 200 OWAS devices have been arranged in a 5 × 40-node grid. With this array, our current experimental setup can acquire signals from 120 OWAS devices as allowed by the maximum image size of the camera. Each OWAS device is also equipped with an infrared photosensor that enables it to receive the master clock signal emitted by the pulse generator shown in **Fig. 4**. Therefore, the synchronization between the OWAS devices and the high speed camera is maintained through the master clock generator.

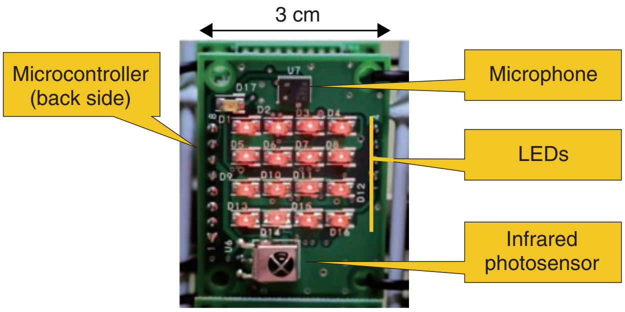


Fig. 2. Photo of OWAS device.

2.2 High speed camera

The imaging sensor of the high speed camera observes the optical signals from the OWAS and records them into intensity images at the rate of 16,000 frames per second (fps). An example of the actual images captured by the camera is shown in **Fig. 5**. To transfer the image data from the camera to the processing server, the camera is connected to a frame grabber card installed on the PCI (Peripheral Component Interconnect) bus of the server. Thus, the flow of the image data can be seen in **Fig. 6**.

2.3 Parallel processing server

The server is equipped with dual CPU (central processing unit) support and a standard general-purpose graphics processing unit (GPGPU), which provides enough massive parallel computing power to process 16,000 images and 120 audio channels within a

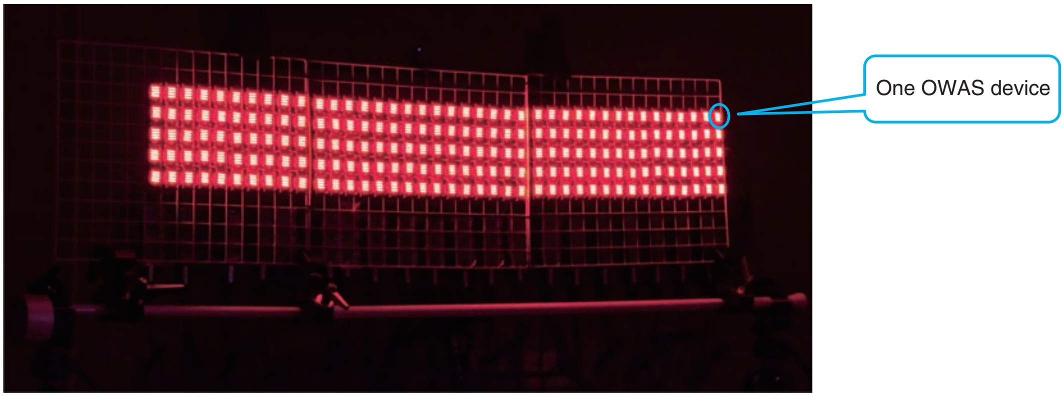


Fig. 3. Array of 200 acoustic sensors (5 × 40).

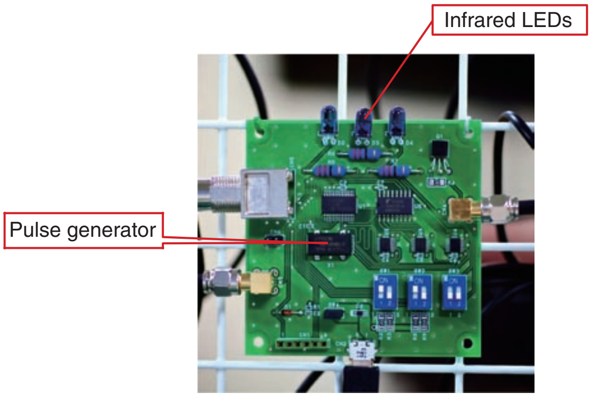


Fig. 4. Infrared transmitter for camera-sensor synchronization.

real-time factor of 0.75. The process to decode the audio signals from the images starts with the detection of the LEDs on the images. Several image processing algorithms have been proposed to accomplish this [4], and it was suggested that the optical transmission channel can be modeled as a MIMO (multiple input multiple output) port as shown in **Fig. 7**. With this model, the pixels of the images can be organized into clusters ***C*** by analyzing the spatiotemporal correlation of their intensity signals ***s*** across a block of captured frames, as shown in **Fig. 8**. Each cluster of pixels represents a detected LED on the images. Once the pixels for each LED have been identified, their intensity signals are optimally thresholded to convert them back to binary symbols (see Fig. 7). Finally, the binary data is decoded to obtain the originally transmitted audio signals from all the OWAS devices.

These audio signals are further processed with digital filters and mixed down to produce a single output channel (see Fig. 1). This multichannel signal processing, often known as *beamforming* [2], enhances the sound from the desired direction (with respect to the OWAS array), while the noise from other directions is reduced at the output audio signal. In other words, the OWAS array can be acoustically focused in any desired direction. In our preliminary experiments, we have been able to focus on the sound from targets placed as far as 10 m away from the OWAS array while suppressing the noise from the surroundings. An online demonstration video showing the experimental setup is available [6]. Furthermore, we have also achieved the optical transmission of multichannel signals at a distance of 30 m from the receiver camera [5]. Our system also has large scalability. Our numerical simulations indicated that our algorithms can receive and decode the optical signals of as many as 12,000 OWAS devices simultaneously within a single GPGPU card, therefore maintaining real-time processing, as can be seen in **Fig. 9**.

3. Future work

The current limitation we face in expanding our prototype involves the high speed camera. The resolution of existing commercial cameras must be considerably reduced in order to achieve high-speed frame rates (tens of thousands of fps). Nevertheless, the accelerated advances in imaging sensor and parallel computing technologies motivate us to carry out further development of our prototype. In the near future, we expect to build OWAS arrays over large

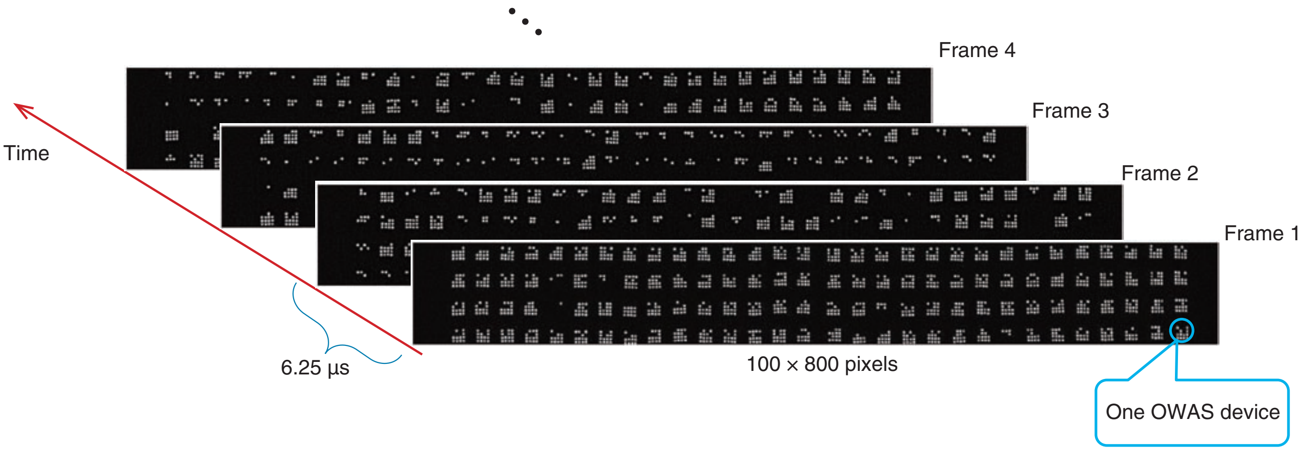
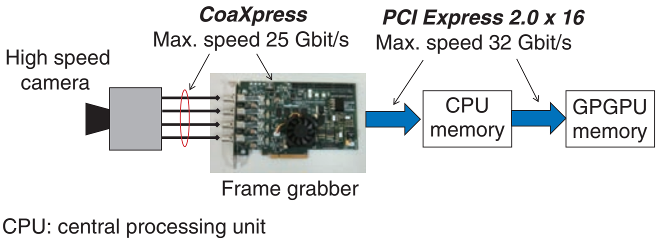


Fig. 5. Images streamed from the high speed camera.



CPU: central processing unit

Fig. 6. Data flow from camera to GPGPU.

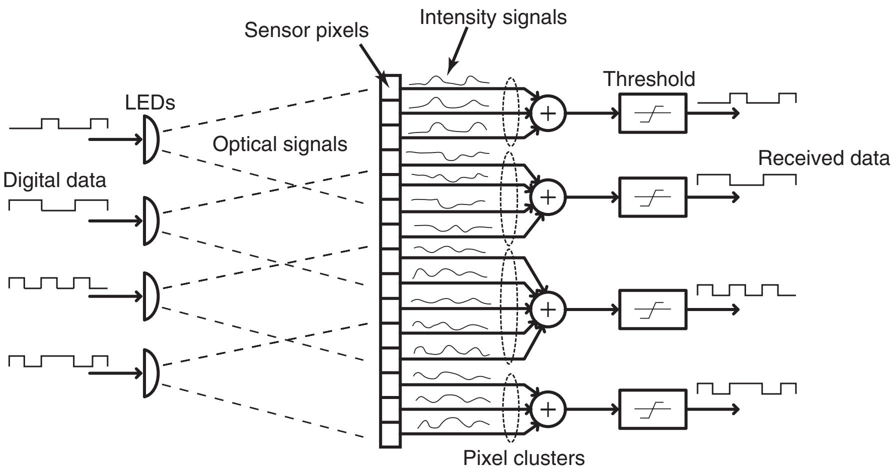
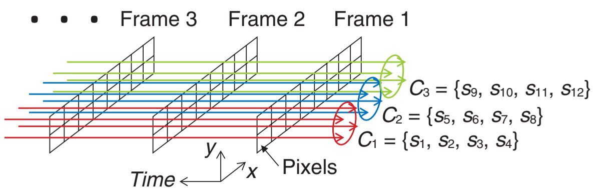
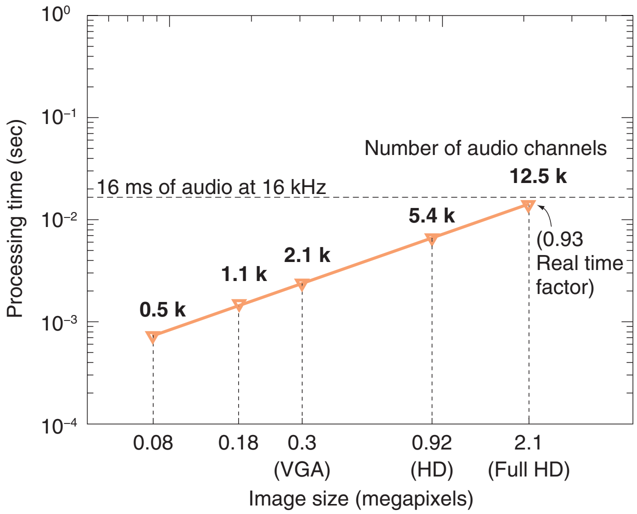


Fig. 7. Optical transmission channel as a MIMO port.


• • • Frame 3
Frame 2
Frame 1
$C_3 = \{s_9, s_{10}, s_{11}, s_{12}\}$
$C_2 = \{s_5, s_6, s_7, s_8\}$
$C_1 = \{s_1, s_2, s_3, s_4\}$
y
x
Time
Pixels

Each cluster *C* represents a detected LED on the images.

Fig. 8. Pixels clustered according to their spatiotemporal correlation.



HD: high definition
VGA: video graphics array

Fig. 9. System scalability in terms of parallel processing power.

spaces such as stadiums or concert halls, and we may achieve real-time position tracking of portable/wearable OWAS devices. Such progress will pave the way for novel applications and high quality services.

References

[1] S. Koyama, Y. Hiwasaki, K. Furuya, and Y. Haneda, "Inverse Wave Propagation for Reproducing Virtual Sources in Front of Loudspeaker Array," Proc. of the European Signal Processing Conference (EUSIPCO) 2011, pp. 1322–1326, Barcelona, Spain, 2011.

[2] H. L. Van Trees, "Optimum Array Processing: Part VI of Detection, Estimation, and Modulation Theory," Wiley-Interscience, New York, 2002.

[3] K. Niwa, Y. Hioka, S. Sakauchi, K. Furuya, and Y. Haneda, "Sharp Directive Beamforming Using Microphone Array and Planar Reflector," Acoustical Science and Technology, Vol. 34, No. 4, pp. 253–262, 2013.

[4] G. P. Nava, Y. Kamamoto, T. G. Sato, Y. Shiraki, N. Harada, and T. Moriya, "Image Processing Techniques for High Speed Camera-based Free-field Optical Communication," Proc. of the IEEE International Conference on Signal and Image Processing Applications (ICSIPA) 2013, pp. 384–389, Melaka, Malaysia, October 2013.

[5] G. P. Nava, Y. Kamamoto, T. G. Sato, Y. Shiraki, N. Harada, and T. Moriya, "Simultaneous Acquisition of Massive Number of Audio Channels through Optical Means," Proc. of the 135th Audio Engineering Society (AES) Convention, New York, USA, October 2013.

[6] Demo video (media player required),
mms://csflash.kecl.ntt.co.jp/cslab/mrl/pablo/vasdemo1.wmv

Gabriel Pablo Nava

Research Scientist, Moriya Research Laboratory, NTT Communication Science Laboratories.

He completed his B.E. studies in Mexico City at the Instituto Politécnico Nacional in 1999. He received his M.S. and Ph.D. in 2004 and 2007, respectively, from the Department of Information Science and Technology of the University of Tokyo. He also held a postdoctoral position at the University of Tokyo from March 2008 to April 2014. He joined NTT Communication Science Laboratories in April 2008. He is currently conducting research and development on room acoustics, multichannel audio, microphone array signal processing, and digital image processing for next-generation ultra-realistic teleconferencing systems.

Yoshifumi Shiraki

Research Scientist, Moriya Research Laboratory, NTT Communication Science Laboratories.

He received his B.A. in natural science from International Christian University, Tokyo, in 2008 and his M.S. in computational intelligence and systems science from Tokyo Institute of Technology in 2010. Since joining NTT Communication Science Laboratories in 2010, he has been studying signal processing, particularly distributed compressed sensing. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) and the Institute of Electrical and Electronics Engineers (IEEE).

Hoang Duy Nguyen

Intern, Moriya Research Laboratory, NTT Communication Science Laboratories.

Mr. Nguyen was born in Brussels, Belgium, in 1990. He graduated summa cum laude with an M.S. in electronics and telecommunications engineering from the University of Brussels, Belgium, in 2013. From 2012 to 2013 he worked as a research engineer in the Smart Systems and Energy Technology (SSET) department of the Interuniversity Micro-Electronics Centre (IMEC), Leuven, Belgium. Since 2014 he has been in the Special Research Group of NTT Communication Science Laboratories, Atsugi City, Japan, where he carries out experimental research and development involving free-field optical communication. His research interests include MIMO detection algorithms and implementation aspects of acoustic beamformers on GPGPUs.

Yutaka Kamamoto

Research Scientist, Moriya Research Laboratory, NTT Communication Science Laboratories.

He received his B.S. in applied physics and physico-informatics from Keio University, Kanagawa, in 2003 and his M.S. and Ph.D. in information physics and computing from the University of Tokyo in 2005 and 2012, respectively. Since joining NTT Communication Science Laboratories in 2005, he has been studying signal processing and information theory, particularly lossless coding of time-domain signals. He also joined NTT Network Innovation Laboratories, where he worked on the development of the audio-visual codec for online digital stuff (ODS) from 2009 to 2011. He has contributed to the standardization of coding schemes for MPEG-4 Audio lossless coding (ALS) and ITU-T Recommendation G.711.0 Lossless compression of G.711 pulse code modulation. He received the Telecom System Student Award from the Telecommunications Advancement Foundation (TAF) in 2006, the IPSJ Best Paper Award from the Information Processing Society of Japan (IPSJ) in 2006, the Telecom System Encouragement Award from TAF in 2007, and the Awaya Prize Young Researcher's Award from the Acoustical Society of Japan (ASJ) in 2011. He is a member of IPSJ, ASJ, IEICE, and IEEE.

Takashi G. Sato

Research Scientist, Moriya Research Laboratory, NTT Communication Science Laboratories.

He received his M.S. and Ph.D. in information science and technology from the University of Tokyo in 2005 and 2008, respectively. He is currently a researcher at NTT Communication Science Laboratories. His research interests include bioengineering and tactile and audio interfaces using psychophysiological and neural measurement as feedback information. He is a member of IEEE.

Noboru Harada

Senior Research Scientist, Moriya Research Laboratory, NTT Communication Science Laboratories.

He received his B.S. and M.S. from the Department of Computer Science and Systems Engineering of Kyushu Institute of Technology in 1995 and 1997, respectively. He joined NTT in 1997. His main research area is lossless audio coding, high-efficiency coding of speech and audio, and their applications. He was also with NTT Network Innovation Laboratories, where he worked on the development of the audio-visual codec for ODS, from 2009 to 2011. He is an editor of ISO/IEC 23000-6:2009 Professional Archival Application Format, ISO/IEC 14496-5:2001/Amd.10:2007 reference software MPEG-4 ALS, and ITU-T G.711.0. He is a member of IEICE, ASJ, the Audio Engineering Society (AES), and IEEE.

Takehiro Moriya

NTT Fellow, Moriya Research Laboratory, NTT Communication Science Laboratories.

He received his B.S., M.S., and Ph.D. in mathematical engineering and instrumentation physics from the University of Tokyo in 1978, 1980, and 1989, respectively. Since joining NTT laboratories in 1980, he has been engaged in research on medium- to low-bit-rate speech and audio coding. In 1989, he worked at AT&T Bell Laboratories, NJ, USA, as a Visiting Researcher. Since 1990, he has contributed to the standardization of coding schemes for the Japanese public digital cellular system, ITU-T, ISO/IEC MPEG, and 3GPP. He is a member of the Senior Editorial Board of the IEEE Journal of Selected Topics in Signal Processing. He is a Fellow member of IEEE and a member of IPSJ, IEICE, AES, and ASJ.