MADRID
inter.noise 2019
June 16 - 19
NOISE CONTROL FOR A BETTER ENVIRONMENT

Three-dimensional sound source localization using microphone array and its display with mixed reality technology

Kusano, Tsubasa[1]
Matsumoto, Yasuhide[2]
Kataoka, Yuta[3]
Teraoka, Wataru[4]
Oikawa, Yasuhiro[5]
Waseda University
3-4-1 Ohkubo, Shinjuku-ku, Tokyo 169-8555, Japan

Yoshida, Kenichi[6]
Kitazumi, Yoshimi[7]
DENSO WAVE INCORPORTED
1 Yoshiike Kusagi Agui-cho, Chita-gun, Aichi 470-2297, Japan

ABSTRACT

Visualization of sound information has many interests for understanding sound fields. Especially, sound source localization using a microphone array is helpful for sound source separations, acoustical design, abnormality detection, etc. However, three-dimensional (3D) sound source localization result is difficult to present intuitively on ordinary displays. On the other hand, in recent years, mixed reality (MR) technology has rapidly developed and attracts many attentions. MR devices install many sensors, displays, and ICT technologies, which realize interaction between real environment and virtual spaces. In this paper, we propose an MR display system for 3D sound source localization results, which are obtained from a microphone array data. MR technology enables presentation of 3D sound source localization results which is difficult for ordinary displays. In addition, the user can observe both 3D sound source localization results and the environment simultaneously.

Keywords: Acoustic imaging, see-through head mounted display, simultaneous localization and mapping (SLAM), beamforming
I-INCE Classification of Subject Number: 74
(see http://i-ince.org/files/data/classification.pdf)

[1]tsubasa.k@suou.waseda.jp
[2]kikan.match1226@toki.waseda.jp
[3]kataoka8894@akane.waseda.jp
[4]tera-wata1129@fuji.waseda.jp
[5]yoikawa@waseda.jp
[6]kenichi.yoshida@denso-wave.co.jp
[7]yoshimi.kitazumi@denso-wave.co.jp

1. INTRODUCTION

Visualization of sound information has many interests for understanding sound fields. In previous studies, sound field visualization has been realized with various measurement methods, such as acoustical holography [1], optical methods [2, 3], and sound intensity [4–6]. Sound source localization using a microphone array is one of the visualization methods for sound information, which have been employed in many applications including sound source separations, acoustical design, and abnormality detection [7–12]. Three-dimensional (3D) sound source localization gives useful information for understanding sound fields, but observing and understanding the estimated result are difficult for ordinary displays because they cannot present the depth information.

On the other hand, in recent years, mixed reality (MR) technology has rapidly developed and attracts many attentions. MR devices install many sensors, displays, and ICT technologies, which realize interaction between real world and virtual spaces. MR technology has already applied to the visualization of sound information such as sound intensity [4–6], and closely located four-point microphone method [13].

In this paper, a display system of 3D sound source localization results using MR technology is proposed. The proposed system enables us to observe and understand the localization results intuitively by MR technology with the see-through head mounted display (STHMD).

2. METHODS

Figure 1 shows the state of observing the estimated sound source via the STHMD. First, the sound source localization is performed with signals obtained by the microphone array. The spherical 3D objects whose size corresponds to the power of sound sources are superimposed on the real world as the visualization of the sound localization result. In this paper, the two sound source localization methods are utilized, which is beamforming method [7–9] and sparsity-based localization method [10–12].

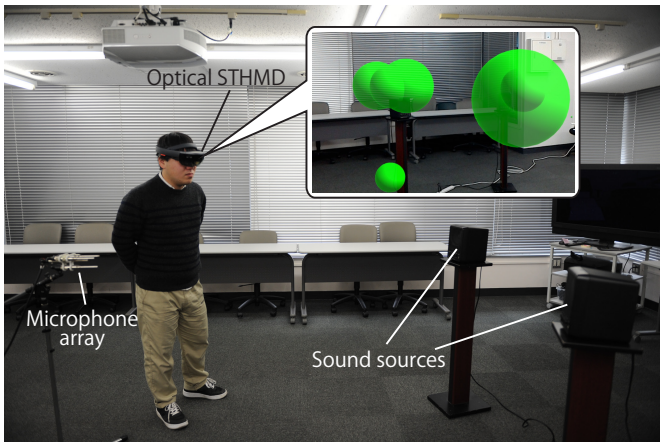

Figure 1: State of observing the estimated sound source via the STHMD.

2.1 Mixed reality technology

In MR technology, the real and virtual worlds of visual information interact with each other as if these two worlds were merged. As the STHMD, Microsoft HoloLens shown in Figure 2 is used in the proposed system. HoloLens is a self-contained holographic computer with an optical STHMD and simultaneous localization and mapping (SLAM) technology [14]. An optical STHMD is a stereo transparent display which is able to superimpose 3D computer graphics on the real world. The SLAM technology enables to acquire user's position and the surrounding environment using only the sensors installed in the wearable devices. By using this, 3D objects can be appropriately superimposed on the user's view without other markers or sensors. The proposed system was developed with Unity 2018, which contains the Unity3D engine of Unity Technologies.

2.2 Sound source localization

2.2.1 Beamforming method

Beamforming method is fundamental array signal processing technique, which is widely employed in sound source enhancement. In addition, Beamforming can also be applied to the sound source localization by finding peaks of the power of output signals [7–9].

Let us considered that uncorrelated narrowband signals radiated at the positions $\mathbf{p}_n \in \mathbb{R}^3\,(n = 0, \ldots, N-1)$, are observed microphones at $\mathbf{q}_m \in \mathbb{R}^3\,(m = 0, \ldots, M-1)$. The observed signals can be written as

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{e}(t), \tag{1}$$

where $\mathbf{e} \in \mathbb{C}^M$ is the additive noise, $\mathbf{s} \in \mathbb{C}^N$ is the source signal vector,

$$\mathbf{A} = [\mathbf{a}(\mathbf{p}_0), \mathbf{a}(\mathbf{p}_1), \ldots, \mathbf{a}(\mathbf{p}_{N-1})], \tag{2}$$

$$\mathbf{a}(\boldsymbol{\xi}) = [a_0(\boldsymbol{\xi}), a_1(\boldsymbol{\xi}), \ldots, a_{M-1}(\boldsymbol{\xi})]^{\mathrm{T}}, \tag{3}$$

$$a_m(\boldsymbol{\xi}) = \exp(\mathrm{i}\pi\omega\|\boldsymbol{\xi} - \mathbf{q}_m\|_2/c), \tag{4}$$

Figure 2: Microsoft HoloLens.

ω is the angular frequency, c is the sound speed, $\mathrm{i} = \sqrt{-1}$, $\mathbf{x}^{\mathrm{T}}$ is the transpose of $\mathbf{x}$, and $\|\cdot\|_p$ is the ℓ_p-norm (for $p \geq 1$) defined as $\|\mathbf{x}\|_p = (\sum_k |x_k|^p)^{1/p}$.

In the beamforming, a signal arriving from the position $\boldsymbol{\xi}$ is enhanced at by multiplying the weighting vector $\mathbf{w}(\boldsymbol{\xi}) \in \mathbb{C}^M$:

$$y(t) = \mathbf{w}^*(\boldsymbol{\xi})\mathbf{x}(t), \tag{5}$$

where $\mathbf{w}^*$ is the complex-conjugate transpose of $\mathbf{w}$. The power of the output signal $y(t)$ is obtained by

$$P(\boldsymbol{\xi}) = E\left\{|y(t)|^2\right\}, \tag{6}$$

where $E\{\cdot\}$ is the expected value. When samples $\mathbf{y} = [y(t_0), y(t_1), \ldots, y(t_{L-1})]^{\mathrm{T}} \in \mathbb{C}^L$ are given, $P(\boldsymbol{\xi})$ is estimated by

$$\hat{P}(\boldsymbol{\xi}) = \frac{1}{L}\sum_{l=0}^{L-1} |y(t_l)|^2 = \frac{1}{L}\sum_{l=0}^{L-1} |\mathbf{w}^*\mathbf{x}(t_l)|^2 = \mathbf{w}^*\hat{\mathbf{R}}\mathbf{w},$$

$\hat{\mathbf{R}} \in \mathbb{C}^{M\times M}$ is the array covariance matrix represented as

$$\hat{\mathbf{R}} = \sum_{l=0}^{L-1} \mathbf{x}(t_l)\mathbf{x}^*(t_l). \tag{7}$$

Some approaches exist for the choice of the weighting vector $\mathbf{w}$. In this paper, the minimum variance distortionless responses (MVDR) beamformer is used. In the MVDR beamformer, the weighting vector is chosen to minimize the output power with the constraint that the gain in the desired response position equals unity in order to reduce the effect of signals arriving from except the desired response position. The weighting vector of the MVDR beamformer is chosen as the optimal solution of

$$\begin{aligned} &\underset{\mathbf{w}}{\text{minimize}} \;\; \mathbf{w}^*\mathbf{R}\mathbf{w} \\ &\text{subject to} \;\; \mathbf{w}^*\mathbf{a}(\boldsymbol{\xi}) = 1. \end{aligned} \tag{8}$$

The solution of Equation 8 is analytically given by

$$\hat{\mathbf{w}}_{\mathrm{MVDR}} = \frac{\mathbf{R}^{-1}\mathbf{a}(\boldsymbol{\xi})}{\mathbf{a}^*(\boldsymbol{\xi})\mathbf{R}^{-1}\mathbf{a}(\boldsymbol{\xi})},$$

and the output power of the MVDR beamformer is expressed as

$$\hat{P}_{\mathrm{MVDR}}(\boldsymbol{\xi}) = \sum_{l=0}^{L-1} |\hat{\mathbf{w}}^*_{\mathrm{MVDR}}\mathbf{x}(t_l)|^2 = \frac{1}{\mathbf{a}^*(\boldsymbol{\xi})\mathbf{R}^{-1}\mathbf{a}(\boldsymbol{\xi})}.$$

When the multiple frequencies are used, the output power is calculated by the summation of each frequency ω_k $(k = 0, 1, \ldots, K-1)$ as

$$\sum_{k=0}^{K-1} \frac{1}{\mathbf{a}^*(\boldsymbol{\xi}, \omega_k)\mathbf{R}^{-1}(\omega_k)\mathbf{a}(\boldsymbol{\xi}, \omega_k)}. \tag{9}$$

2.2.2 Sparsity-based localization method

Sparsity-based localization method is a method based on the assumption of the sparseness of the sound sources [10–12]. Some models have been considered in the sparsity-based localization method [10, 11]. In this paper, a set of monopoles are considered as the model in order to estimate the 3D positions of sound sources [12].

Let consider approximating a sound field $b(\boldsymbol{\xi}, \omega) \in \mathbb{C}$ by linear combination as

$$b(\boldsymbol{\xi}, \omega) \simeq \sum_j \varphi_j(\boldsymbol{\xi}, \omega)\alpha_j, \tag{10}$$

where $\varphi_j(\boldsymbol{\xi}, \omega)$ is the monopole dictionary

$$\varphi_j(\boldsymbol{\xi}, \omega) = \frac{\exp(\mathrm{i}\pi\omega\|\boldsymbol{\xi} - \mathbf{q}_m\|_2/c)}{4\pi\|\boldsymbol{\xi} - \mathbf{q}_m\|_2}. \tag{11}$$

Then, observed signals of M microphones at $\mathbf{q}_m$ ($m = 0, \ldots, M-1$) is written as the matrix form:

$$\mathbf{b}_k = \boldsymbol{\Phi}_k\boldsymbol{\alpha}_k, \tag{12}$$

where

$$\mathbf{b}_k = [b(\mathbf{q}_0, \omega_k), b(\mathbf{q}_1, \omega_k), \ldots, b(\mathbf{q}_{M-1}, \omega_k)]^{\mathrm{T}}, \tag{13}$$
$$\boldsymbol{\Phi}_k = [\boldsymbol{\varphi}_j(\omega_k), \boldsymbol{\varphi}_j(\omega_k), \ldots, \boldsymbol{\varphi}_j(\omega_k)], \tag{14}$$
$$\boldsymbol{\varphi}_j(\omega) = [\varphi_j(\mathbf{q}_0, \omega), \varphi_j(\mathbf{q}_1, \omega), \ldots, \varphi_j(\mathbf{q}_{M-1}, \omega)]^{\mathrm{T}}. \tag{15}$$

Assuming the sparseness of the sound sources, it is expected that the observed signals can be well approximated by the coefficients $\boldsymbol{\alpha}_k$ with the most elements being zero. To find such coefficients $\boldsymbol{\alpha}_k$, the following Lasso problem,

$$\underset{\boldsymbol{\alpha}_k}{\text{minimize}} \quad \frac{1}{2}\|\boldsymbol{\Phi}_k\boldsymbol{\alpha}_k - \mathbf{b}_k\|_2^2 + \lambda\|\boldsymbol{\alpha}_k\|_1, \tag{16}$$

is considered [12], where $\lambda > 0$ is a regularization parameter. The left term of Equation 16 is the square error between the observed signals and the approximation with the dictionary. The right term is the ℓ_1-norm, which induces sparsity of $\boldsymbol{\alpha}_k$.

In the case of considering multiple frequencies at the same time, Equation 16 is modified to the group sparse problem

$$\underset{\boldsymbol{\alpha}}{\text{minimize}} \quad \frac{1}{2}\|\boldsymbol{\Phi}\boldsymbol{\alpha} - \mathbf{b}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_{2,1}, \tag{17}$$

where

$$\mathbf{b} = [\mathbf{b}_0^{\mathrm{T}}, \mathbf{b}_1^{\mathrm{T}}, \ldots, \mathbf{b}_{K-1}^{\mathrm{T}}]^{\mathrm{T}}, \tag{18}$$
$$\boldsymbol{\alpha} = [\boldsymbol{\alpha}_0^{\mathrm{T}}, \boldsymbol{\alpha}_1^{\mathrm{T}}, \ldots, \boldsymbol{\alpha}_{K-1}^{\mathrm{T}}]^{\mathrm{T}}, \tag{19}$$
$$\boldsymbol{\Phi} = \begin{bmatrix} \tilde{\boldsymbol{\Phi}}_0 & & & \\ & \tilde{\boldsymbol{\Phi}}_1 & & \\ & & \ddots & \\ & & & \tilde{\boldsymbol{\Phi}}_{K-1} \end{bmatrix}, \tag{20}$$

$\tilde{\mathbf{\Phi}}_k$ is the column-wise normalized version of $\mathbf{\Phi}_k$, $\|\boldsymbol{\alpha}\|_{2,1}$ is the $\ell_{2,1}$-norm defined by

$$\|\boldsymbol{\alpha}\|_{2,1} = \sum_{\mathcal{G}\in\mathscr{G}} \|\boldsymbol{\alpha}_{\mathcal{G}}\|_2. \tag{21}$$

$\mathscr{G}$ is the set of all monopoles for every frequency, and $\mathcal{G}$ is the group of all frequencies in a monopole. $\ell_{2,1}$-norm induces the same sparsity patterns within each frequency component.

3. EXPERIMENTS

In this section, a measuremental experiment of sound source localization using a microphone array was performed. For a microphone array, we used four microphones arranged in a regular tetrahedron of edge length 0.05 cm, as shown in Figure 3. Two sound signals (white noise) emitted from loudspeakers were measured in an ordinary meeting room at Waseda University. The measurement conditions are shown in Table 1.

The visualization result of the MVDR beamformer is shown in Figure 4. 3D objects corresponding high power output of the beamformer were lined up between the microphone array and the loudspeakers. It can be seen from Figure 4 that the visualization system using the beamformer can present the directions of sound the sources.

Then, the visualization result of the sparsity-based localization method is shown in Figure 5. The spherical objects in Figure 5 correspond non-zero coefficients $\boldsymbol{\alpha}$. Note that the large objects do not necessarily point to the sound source position since one sound source is approximated by a linear combination of some dictionaries.

Table 1: Measurement conditions.

Measured room	59-402 Meeting room, Nishi-Waseda campus, Waseda University
Equipment	Four microphones (AUDIX TM1) MacBook Pro (2.7 GHz Intel Core i7, 16 GB 1600 MHz DDR 3) Microsoft HoloLens Audio interface (MOTU 8M) Two Loudspeakers (YAMAHA MS101Ⅲ)
Sampling frequency	48 kHz

Figure 3: The microphone array utilized in the experiment.

These results indicate that the proposed system assists the user to grasp the sound source positions by superimposing the 3D objects on the real world, while both the methods do not point to the sound source positions directly.

4. CONCLUSIONS

In this paper, we propose a display system the localization results of sound sources with MR technology. The measurement experiment shows that the proposed system will assists users to grasp the sound source positions by superimposing the localization results

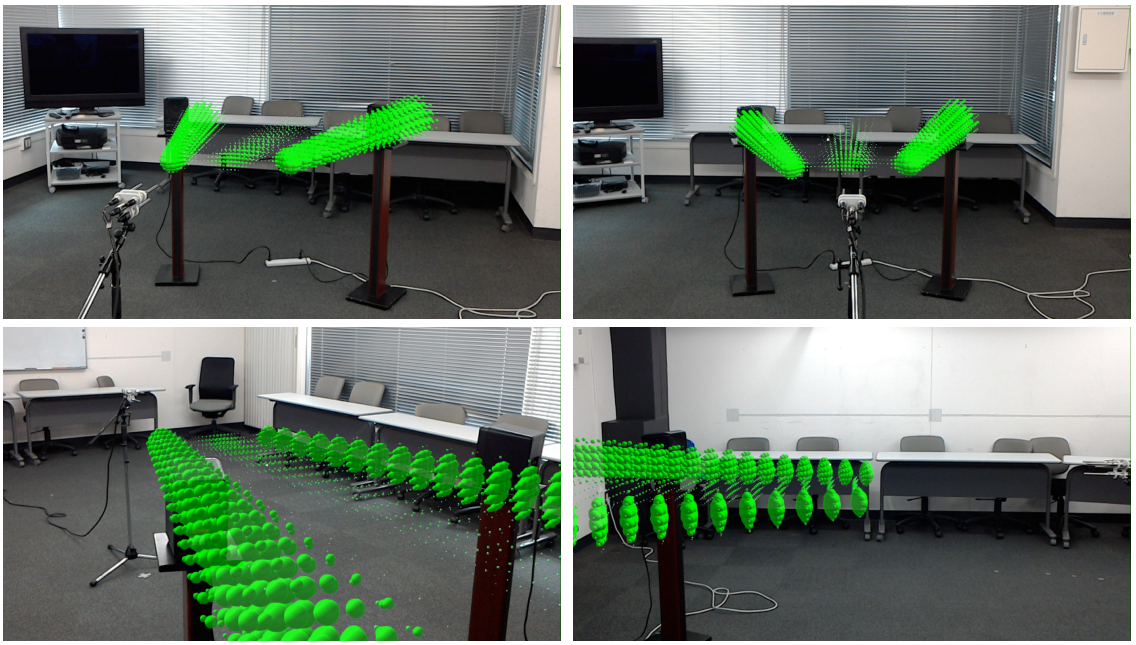

Figure 4: Visualization of the localization result using the MVDR beamformer.

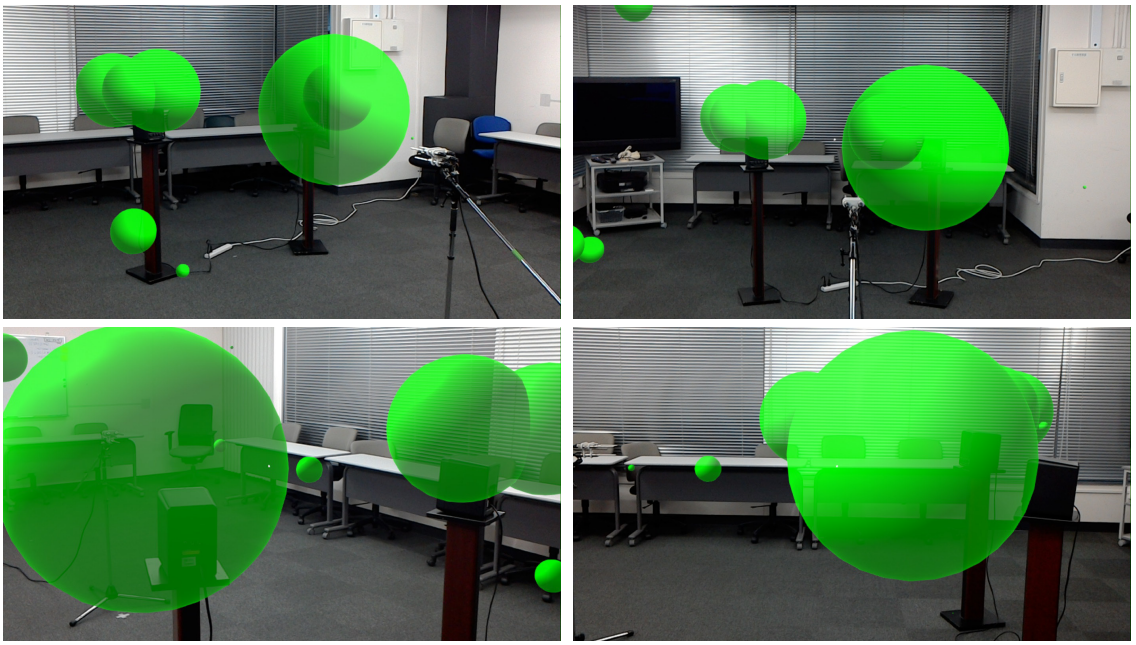

Figure 5: Visualization of the localization result using sparsity-based source localization.

on the real world. Future work includes the construction of the system for updating the 3D objects represented as the estimated results.

5. REFERENCES

[1] J. D. Maynard, E. G. Williams, and Y. Lee, "Nearfield acoustic holography: I. Theory of generalized holography and the development of NAH," *J. Acoust. Soc. Am.*, vol. 78, no. 4, pp. 1395–1413, Oct. 1985.

[2] K. Ishikawa, K. Yatabe, N. Chitanont, Y. Ikeda, Y. Oikawa, T. Onuma, H. Niwa, and M. Yoshii, "High-speed imaging of sound using parallel phase-shifting interferometry," *Opt. Express*, vol. 24, no. 12, pp. 12 922–12 932, Jun. 2016.

[3] Y. Oikawa, M. Goto, Y. Ikeda, T. Takizawa, and Y. Yamasaki, "Sound field measurements based on reconstruction from laser projections," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 4, Mar. 2005, pp. 661–664.

[4] A. Inoue, K. Yatabe, Y. Oikawa, and Y. Ikeda, "Visualization of 3D sound field using see-through head mounted display," in *Proc. ACM SIGGRAPH*. New York, NY, USA: ACM, 2017, pp. 34:1–34:2.

[5] Y. Kataoka, W. Teraoka, Y. Oikawa, and Y. Ikeda, "Real-time measurement and display system of 3D sound intensity map using optical see-through head mounted display," in *Proc. SIGGRAPH Asia*. New York, NY, USA: ACM, 2018, pp. 71:1–71:2.

[6] A. Inoue, Y. Ikeda, K. Yatabe, and Y. Oikawa, "Visualization system for sound field using see-through head-mounted display," *Acoust. Sci. Technol.*, vol. 40, no. 1, pp. 1–11, Jan. 2019.

[7] 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.

[8] J. C. Chen, K. Yao, and R. E. Hudson, "Source localization and beamforming," *IEEE Signal Process. Mag.*, vol. 19, no. 2, pp. 30–39, 2002.

[9] P. Chiariotti, M. Martarelli, and P. Castellini, "Acoustic beamforming for noise source localization – Reviews, methodology and applications," *Mech. Syst. Sig. Process.*, vol. 120, pp. 422–448, Apr. 2019.

[10] D. Malioutov, M. Cetin, and A. S. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 3010–3022, Aug. 2005.

[11] H. Jamali-Rad and G. Leus, "Sparsity-aware multi-source TDOA localization," *IEEE Trans. Signal Process.*, vol. 61, no. 19, pp. 4874–4887, Oct. 2013.

[12] T. Tachikawa, K. Yatabe, and Y. Oikawa, "3d sound source localization based on coherence-adjusted monopole dictionary and modified convex clustering," *Appl. Acoust.*, vol. 139, pp. 267–281, 2018.

[13] W. Teraoka, Y. Kataoka, Y. Oikawa, and Y. Ikeda, “Display system for distribution of virtual image sources by using mixed reality technology,” *Inter-Noise*, 2018.

[14] H. Durrant-Whyte and T. Bailey, “Simultaneous localization and mapping: Part I,” *IEEE Robot. Autom. Mag.*, vol. 13, no. 2, pp. 99–108, 2006.