2022 International Workshop on Acoustic Signal Enhancement (IWAENC) | 978-1-6654-6867-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/IWAENC53105.2022.9914786

JOINT LOCALIZATION AND SYNCHRONIZATION OF DISTRIBUTED CAMERA-ATTACHED MICROPHONE ARRAYS FOR INDOOR SCENE ANALYSIS

Yoshiaki Sumura[1], *Kouhei Sekiguchi*[2], *Yoshiaki Bando*[3], *Aditya Arie Nugraha*[2], *and Kazuyoshi Yoshii*[1,2]

[1]Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan
[2]Center for Advanced Intelligence Project (AIP), RIKEN, Tokyo 103-0027, Japan
[3]National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, 135-0064, Japan

ABSTRACT

This paper describes an automatic calibration method that localizes and synchronizes distributed camera-attached microphone arrays (*e.g.*, Microsoft Azure Kinect) used for audio-visual indoor scene analysis. Operating multiple audio-visual sensors as a large-scale array is a key to resolving object occlusions and sound overlaps by integrating audio-visual information obtained from multiple angles. A naive solution to the calibration problem is to synchronize microphone arrays after localizing them using only visual information. This cascading approach, however, would suffer from the error propagation problem. We thus propose a principled statistical method that fully uses audio-visual information at once. Our method only asks a user to make handclaps and jointly estimates the sensor positions and time offsets and the time-varying source position with the GraphSLAM algorithm based on a unified state-space model associating all the latent calibration targets with the audio-visual observations. The experiment using real recordings shows the stable behavior of the proposed method.

Index Terms— Audio-visual scene analysis, calibration, localization, synchronization, and microphone array.

1. INTRODUCTION

Computational audio-visual scene analysis forms the basis of machine intelligence. It covers a wide variety of research topics such as detection, localization, and classification of salient objects and events [1–4]. These tasks had typically been tackled by using either audio or visual information and the multi-modal approach has recently become popular at the intersection of the fields of audio signal processing and computer vision. The underlying common assumption is that audio-visual sensors (*e.g.*, RGB/depth cameras and microphones) are calibrated, *i.e.*, these sensors are synchronized and their positions are measured precisely, in advance of scene analysis.

In this paper we focus on the portability of an intelligent spoken dialogue system (*e.g.*, humanoid robot [5]) that can make multi-party conversations in indoor environments. Such a system should be capable of speaker diarization (detection and identification of active speakers) [6,7] and localization [8] and distant speech recognition [9–11] under acoustically and visually challenging conditions with speech overlaps and object occlusions. One thus may use multiple cameras and microphones distributed in the environment for observing participants from multiple angles. From a practical point of view, the system should be easy to install, *i.e.*, the full potential of the system should be drawn in any environment by just placing the sensors at arbitrary positions.

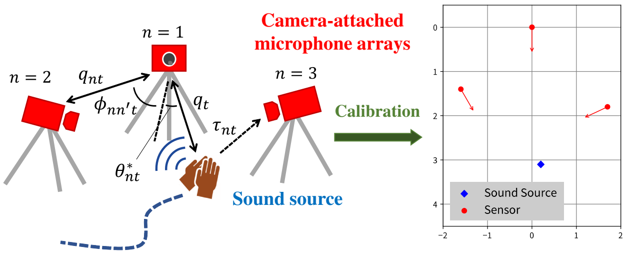


Fig. 1. Automatic calibration of camera-attached microphone arrays through observation of a movable sound source.

Accurate and efficient automatic calibration (localization and synchronization) of distributed audio and visual sensors is thus a key to better indoor scene analysis. For unimodal analysis, localization of distributed cameras [12, 13] and calibration of distributed microphones [14–17], which can be viewed as simultaneous localization and mapping (SLAM) problems, have separately been investigated. Several multi-modal methods were proposed to calibrate distributed microphone arrays by using multiple cameras whose positions are assumed to be known [18, 19]. A naive way of avoiding this assumption is to localize cameras using visual information [12, 13] and then calibrate microphone arrays using audio information [14, 15]. Such a cascading approach, however, would have a performance limitation due to the error propagation problem (suboptimality of the whole system). This calls for a new approach to joint calibration of audio and visual sensors.

In this paper we propose a statistical calibration method that can jointly estimate the positions, orientations, and time offsets of distributed asynchronous audio-visual sensors, each of which consists of a synchronous pair of an RGB and/or

This work was partially supported by JSPS KAKENHI Nos. 19H04137, 20K19833, and 20K21813.

978-1-6654-6867-1/22/$31.00 ©2022 IEEE

depth camera and a microphone array. To do this, a mobile sound source (*e.g.*, system user) that emits reference signals (*e.g.*, handclaps) repeatedly is introduced. Using each sensor solely, the directions and distances to the other sensors and the sound source and the time differences of arrival (TDoAs) between the sensors are acoustically or visually estimated for each trial. We formulate a state-space model that represents the generative process of such noisy estimates (observed variables) from the time-invariant sensor positions, orientations, and time offsets and the time-varying source position (latent variables). Given the observed variables, the joint posterior distribution of the latent variables can be approximated with an iterative optimization method called GraphSLAM [20].

The main contribution of this study is to propose the principled approach to joint calibration of audio and visual sensors based on statistical inference of a unified probabilistic model. Our method could significantly improve the portability and generalization capabilities of audio-visual scene analysis systems without expert knowledge and experience, human labor, and expensive measurement devices (*e.g.*, laser rangefinders and motion capture systems) in initial installation and configuration in a target environment.

2. RELATED WORK

This section introduces existing studies on calibration of audio-only and audio-visual sensors.

2.1. Audio Sensor Calibration

A major approach for calibrating multiple microphones or microphone arrays is to solve a SLAM problem using only acoustic information. Miura *et al.* [14] proposed an online calibration method of asynchronous microphones based on extended Kalman filter (EKF)-SLAM. Using handclaps as reference sounds, the position and time offset of each microphone are estimated. Su *et al.* [15] proposed an offline calibration method of an asynchronous microphone array. They did not assume that the clock timing in each microphone was exactly the same. Based on a graph-based SLAM [21] algorithm, their method estimates clock difference of each microphone together with the positions of sound sources and microphones and time offsets in an offline manner. Sekiguchi *et al.* [17] addressed the online calibration of asynchronous multiple microphone arrays. Their method estimates the position and time offset of each microphone array and the sound source position using FastSLAM [22] algorithm.

2.2. Audio-Visual Sensor Calibration

Visual information is also used for the calibration of microphone arrays. Jacob *et al.* [18] estimated a common coordinate system for audio-visual sensor networks. In this method, a speaker is tracked by both microphone and camera networks. The microphone network should be calibrated by a self-localization algorithm using the speaker's voice tracking at first, and the camera network is assumed to be calibrated in advance. Then, the two modalities are embedded into a random sample consensus (RANSAC) framework to obtain a mapping. Plinge *et al.* [19] proposed a method to calibrate multiple microphone arrays utilizing cameras with known positions installed in a room. The calibration is performed by finding the position of the microphone arrays that minimizes the error between the sound source positions estimated from the microphone array observations and those obtained from camera observations. While these methods use separately distributed cameras and microphone arrays and assume that the camera positions are known, we remove this assumption using camera-attached microphone arrays.

3. PROPOSED METHOD

This section describes the proposed calibration method based on a probabilistic model for localizing and synchronizing distributed camera-attached microphone arrays.

3.1. Problem Specification

We tackle the sensor calibration problem that aims to localize and synchronize audio-visual sensors (*e.g.*, Microsoft Azure Kinect) used for indoor scene analysis. Suppose that N asynchronous sensors are located at arbitrary positions in a room. Each sensor consists of an RGB camera and M synchronous microphones (microphone array) with a known geometry and has a marker. One of the N sensors (indexed by 1 and called the reference sensor) is allowed to additionally have a depth camera to avoid interference of infrared rays. A movable sound source that emits reference signals T times (*e.g.*, a person who makes handclaps) is prepared in the same room. In practice, it is desirable that the sensors are located on the walls of a room such that each object is observed from multiple angles, *i.e.*, the sound source is surrounded by the sensors.

We aim to estimate the state vector $\mathbf{z}_n \triangleq [\mathbf{r}_n^\mathsf{T}, \omega_n, o_n]^\mathsf{T}$ of each sensor $n \in [1, N]$, where $\mathbf{r}_n \triangleq [r_n^x, r_n^y]^\mathsf{T}$ is the 2D position relative to the position of the reference sensor ($\mathbf{r}_1 = \mathbf{0}$), $\omega_n \in [-\pi, \pi)$ is the orientation, and o_n is the time offset from the reference sensor ($o_1 = 0$). We also aim to estimate the time-varying position $\mathbf{s}_t \triangleq [s_t^x, s_t^y]^\mathsf{T}$ of the sound source at each step $t \in [1, T]$. Let $\mathbf{s} \triangleq [\mathbf{s}_1^\mathsf{T}, \cdots, \mathbf{s}_T^\mathsf{T}]^\mathsf{T}$ and $\mathbf{z} \triangleq [\mathbf{z}_1^\mathsf{T}, \cdots, \mathbf{z}_N^\mathsf{T}]^\mathsf{T}$ be the sets of latent variables.

We assume that a set of *noisy* measurements and estimates $\mathbf{x}_t \triangleq [q_t, \{q_{nt}, \tau_{nt}, \theta_{nt}^\mathrm{v}, \theta_{nt}^\mathrm{a} \{\phi_{nn't}\}_{n' \neq n}\}_{n=1}^N]$ are obtained at each step t as observed data, where q_t is the measured distance from the reference sensor to the sound source, q_{nt} is the measured distance from the reference sensor to sensor n ($q_{1t} = 0$), τ_{nt} is the estimated time difference of arrival (TDoA) of sensor n from the reference sensor ($\tau_{1t} = 0$), $\theta_{nt}^\mathrm{v} \in [-\pi, \pi)$ and $\theta_{nt}^\mathrm{a} \in [-\pi, \pi)$ are the visually and acoustically estimated directions of the sound source from sensor n, respectively, and $\phi_{nn't} \in [-\pi, \pi)$ is the estimated direction of sensor n' from sensor n. Let $\mathbf{x} \triangleq [\mathbf{x}_1^\mathsf{T}, \cdots, \mathbf{x}_T^\mathsf{T}]^\mathsf{T}$ be the set of observed variables.

In practice, q_t and q_{nt} are measured with the depth camera of the reference sensor and τ_{nt} is estimated from the signals of the reference sensor and sensor n with generalized cross correlation with phase transform (GCC-PHAT) [23]. θ^{v}_{nt} is estimated by visually detecting the marker of the sound source from the image observed by sensor n and θ^{a}_{nt} is estimated by acoustically localizing the sound source from the multichannel signal of sensor n with multiple signal classification (MUSIC) [8]. $\phi_{nn't}$ is estimated by detecting the marker of sensor n' from the image observed by sensor n.

3.2. Probabilistic Modeling

We formulate a unified audio-visual state-space model (Fig. 2) that represents the generative process of the time-varying observed variables $\mathbf{x}$ from the time-varying latent variables $\mathbf{s}$ and the time-invariant latent variables $\mathbf{z}$ as follows:

$$p(\mathbf{x}, \mathbf{z}) = \prod_{t=1}^{T} p(\mathbf{x}_t \mid \mathbf{s}_t, \mathbf{z}) p(\mathbf{s}_t \mid \mathbf{s}_{t-1}) \prod_{n=1}^{N} p(\mathbf{z}_n), \quad (1)$$

where $\mathbf{s}_0$ is a dummy random variable introduced for mathematical convenience.

Assuming that the sound source moves randomly at each step, the source positions $\mathbf{s}$ can be represented with a Gaussian random-walk model as follows:

$$p(\mathbf{s}_t \mid \mathbf{s}_{t-1}) = \mathcal{N}(\mathbf{s}_t \mid \mathbf{s}_{t-1}, \mathbf{\Lambda}_{\mathbf{s}}^{-1}), \quad (2)$$

where $\mathbf{\Lambda}_{\mathbf{s}} \triangleq \mathrm{Diag}\big(\lambda_x^2, \lambda_y^2\big)$ is a precision matrix (hyperparameter). Using prior knowledge about the sensor information $\mathbf{z}$, we assume

$$p(\mathbf{z}_n) = \mathcal{N}(\mathbf{z}_n \mid \boldsymbol{\mu}_{\mathbf{z}_n}, \mathbf{\Lambda}_{\mathbf{z}}^{-1}), \quad (3)$$

where $\boldsymbol{\mu}_{\mathbf{z}_n} \triangleq [\boldsymbol{\mu}_{\mathbf{r}_n}^{\mathsf{T}}, \mu_{\omega_n}, \mu_{o_n}]^{\mathsf{T}}$ and $\mathbf{\Lambda}_{\mathbf{z}}$ are the mean vector and diagonal precision matrix of the prior Gaussian distribution (hyperparameters). In practice, the sensor position $\mathbf{r}_n$ and orientation ω_n can be roughly estimated and set to $\boldsymbol{\mu}_{\mathbf{r}_n}$ and μ_{ω_n}, respectively. The time offset o_n is assumed to be around zero, *i.e.*, $\mu_{o_n} = 0$.

The observations $\mathbf{x}$ are assumed to be Gaussian distributed around the theoretical expectations as follows:

$$p(\mathbf{x}_t \mid \mathbf{s}_t, \mathbf{z}) = \mathcal{N}(\mathbf{x}_t \mid \boldsymbol{h}(\mathbf{s}_t, \mathbf{z}), \mathbf{\Lambda}_{\mathbf{x}}^{-1}), \quad (4)$$

where $\mathbf{\Lambda}_{\mathbf{x}}$ is a diagonal precision matrix and $\boldsymbol{h}(\mathbf{s}_t, \mathbf{z})$ is a nonlinear vector-output function. More specifically, each dimension of $\mathbf{x}_t$ is assumed to be distributed as follows:

$$q_t \sim \mathcal{N}\big(\|\mathbf{r}_1 - \mathbf{s}_t\|, \sigma_q^2\big), \quad (5)$$

$$q_{nt} \sim \mathcal{N}\big(\|\mathbf{r}_1 - \mathbf{r}_n\|, \sigma_{q_n}^2\big), \quad (6)$$

$$\tau_{nt} \sim \mathcal{N}\left(\frac{\|\mathbf{r}_n - \mathbf{s}_t\|}{v} + o_n - \frac{\|\mathbf{r}_1 - \mathbf{s}_t\|}{v}, \sigma_{\tau_n}^2\right), \quad (7)$$

$$\theta^{*}_{nt} \sim \mathcal{N}\left(\arctan\frac{s_t^y - r_n^y}{s_t^x - r_n^x} - m_n^{\theta}, \sigma_{\theta_n^*}^2\right) \; (*\in\{\mathrm{v}, \mathrm{a}\}), \quad (8)$$

$$\phi_{nn't} \sim \mathcal{N}\left(\arctan\frac{r_{n'}^y - r_n^y}{r_{n'}^x - r_n^x} - m_n^{\theta},\, \sigma_{\phi_n}^2\right), \quad (9)$$

where v is the sound speed, and σ_q^2, $\sigma_{q_n}^2$, $\sigma_{\tau_n}^2$, $\sigma_{\theta_n^*}^2$, and $\sigma_{\phi_n}^2$ are the hyperparameters that control the variance of measurement errors.

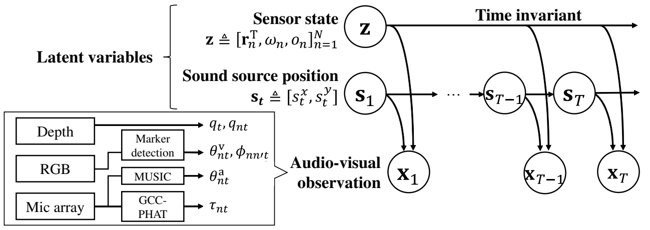


Fig. 2. Unified state-space model for audio-visual calibration.

3.3. Statistical Inference

We infer the latent variables $\mathbf{s}$ and $\mathbf{z}$ from the observed variables $\mathbf{x}$ using an iterative optimization method called GraphSLAM [20]. Let $\hat{\mathbf{s}}_t$ and $\hat{\mathbf{z}}_n$ be the current estimates of $\mathbf{s}_t$ and $\mathbf{z}_n$ and $\hat{\mathbf{z}} \triangleq [\hat{\mathbf{z}}_1^{\mathsf{T}}, \cdots, \hat{\mathbf{z}}_N^{\mathsf{T}}]^{\mathsf{T}}$. Since $\boldsymbol{h}(\mathbf{s}_t, \mathbf{z})$ is a nonlinear function that makes the posterior inference intractable, it is locally approximated as a linear function around $\hat{\mathbf{s}}_t$ and $\hat{\mathbf{z}}_n$ as follows:

$$\boldsymbol{h}(\mathbf{s}_t, \mathbf{z}) \approx \hat{\mathbf{x}}_t + \mathbf{H}_{\hat{\mathbf{s}}_t, \hat{\mathbf{z}}}([\mathbf{s}_t^{\mathsf{T}}, \mathbf{z}^{\mathsf{T}}]^{\mathsf{T}} - [\hat{\mathbf{s}}_t^{\mathsf{T}}, \hat{\mathbf{z}}^{\mathsf{T}}]^{\mathsf{T}}), \quad (10)$$

where $\hat{\mathbf{x}}_t \triangleq \boldsymbol{h}(\hat{\mathbf{s}}_t, \hat{\mathbf{z}})$ is the predicted observation based on the current estimates $\hat{\mathbf{s}}_t$ and $\hat{\mathbf{z}}$ and $\mathbf{H}_{\hat{\mathbf{s}}_t} = \frac{\partial \boldsymbol{h}}{\partial \mathbf{s}_t}\big|_{\mathbf{s}_t = \hat{\mathbf{s}}_t}$, $\mathbf{H}_{\hat{\mathbf{z}}_n} = \frac{\partial \boldsymbol{h}}{\partial \mathbf{z}_n}\big|_{\mathbf{z}_n = \hat{\mathbf{z}}_n}$, and $\mathbf{H}_{\hat{\mathbf{s}}_t, \hat{\mathbf{z}}} = (\mathbf{H}_{\hat{\mathbf{s}}_t}, \mathbf{H}_{\hat{\mathbf{z}}_1}, \cdots, \mathbf{H}_{\hat{\mathbf{z}}_N})$ are the Jacobian matrices.

Since (2), (3), and (4) with (10) are Gaussian distributions, the posterior distribution $p(\mathbf{s}, \mathbf{z}|\mathbf{x})$ can be expressed as a Gaussian distribution and computed as follows:

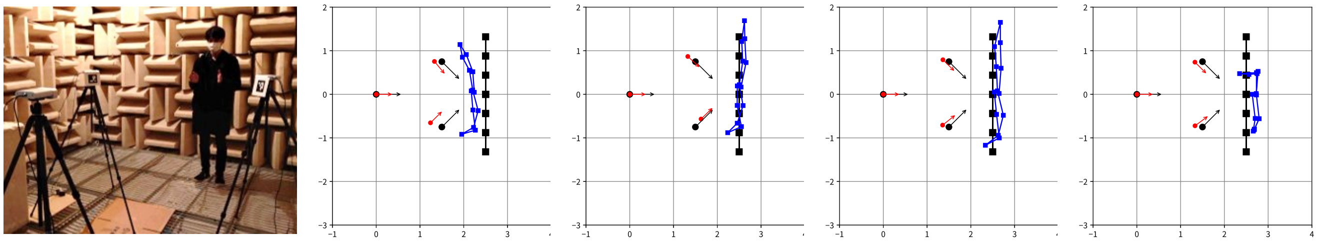
$$\begin{aligned}
&\log p(\mathbf{s}, \mathbf{z}|\mathbf{x}) \\
&= -\frac{1}{2}\sum_{t=1}^{T} [\mathbf{s}_{t-1}^{\mathsf{T}}, \mathbf{s}_t^{\mathsf{T}}] \begin{pmatrix} \mathbf{\Lambda}_{\mathbf{s}} & -\mathbf{\Lambda}_{\mathbf{s}} \\ -\mathbf{\Lambda}_{\mathbf{s}} & \mathbf{\Lambda}_{\mathbf{s}} \end{pmatrix} [\mathbf{s}_{t-1}^{\mathsf{T}}, \mathbf{s}_t^{\mathsf{T}}]^{\mathsf{T}} \\
&\quad - \frac{1}{2}\sum_{n=1}^{N} (\mathbf{z}_n - \boldsymbol{\mu}_{\mathbf{z}})^{\mathsf{T}} \mathbf{\Lambda}_{\mathbf{z}} (\mathbf{z}_n - \boldsymbol{\mu}_{\mathbf{z}}) \\
&\quad + \sum_{t=1}^{T} [\mathbf{s}_t^{\mathsf{T}}, \mathbf{z}^{\mathsf{T}}] \mathbf{H}_{\hat{\mathbf{s}}_t, \hat{\mathbf{z}}}^{\mathsf{T}} \mathbf{\Lambda}_{\mathbf{x}} \left(\mathbf{x}_t - \hat{\mathbf{x}}_t + \mathbf{H}_{\hat{\mathbf{s}}_t, \hat{\mathbf{z}}} [\hat{\mathbf{s}}_t^{\mathsf{T}}, \hat{\mathbf{z}}^{\mathsf{T}}]^{\mathsf{T}}\right) \\
&\quad - \frac{1}{2}\sum_{t=1}^{T} [\mathbf{s}_t^{\mathsf{T}}, \mathbf{z}^{\mathsf{T}}] \mathbf{H}_{\hat{\mathbf{s}}_t, \hat{\mathbf{z}}}^{\mathsf{T}} \mathbf{\Lambda}_{\mathbf{x}} \mathbf{H}_{\hat{\mathbf{s}}_t, \hat{\mathbf{z}}} [\mathbf{s}_t^{\mathsf{T}}, \mathbf{z}^{\mathsf{T}}]^{\mathsf{T}} + \text{ const} \\
&\triangleq \mathcal{N}([\mathbf{s}^{\mathsf{T}}, \mathbf{z}^{\mathsf{T}}]^{\mathsf{T}} \mid \boldsymbol{\mu}, \mathbf{\Omega}^{-1}),
\end{aligned} \quad (11)$$

where $\boldsymbol{\mu}$ and $\mathbf{\Omega}$ are the mean vector and precision matrix of the posterior Gaussian distribution. Let $\mathbf{\Omega}_{\mathbf{s}_{t-1}, \mathbf{s}_t}$ be the partial precision matrix over the dimensions of $\mathbf{\Omega}$ corresponding to $\mathbf{s}_{t-1}$ and $\mathbf{s}_t$ and $\mathbf{\Omega}_{\mathbf{s}_t, \mathbf{z}}$ be defined similarly. Let $\boldsymbol{\xi} \triangleq [\boldsymbol{\xi}_{\mathbf{s}_1}^{\mathsf{T}}, \cdots, \boldsymbol{\xi}_{\mathbf{s}_T}^{\mathsf{T}}, \boldsymbol{\xi}_{\mathbf{z}_1}^{\mathsf{T}}, \cdots, \boldsymbol{\xi}_{\mathbf{z}_N}^{\mathsf{T}}]^{\mathsf{T}}$ be an auxiliary vector given by $\boldsymbol{\xi} = \mathbf{\Omega}\boldsymbol{\mu}$, where $\boldsymbol{\xi}_{\mathbf{s}_t, \mathbf{z}_n}$ is the partial vector over the dimensions of $\boldsymbol{\xi}$ corresponding to $\mathbf{s}_t$ and $\mathbf{z}$, respectively.

Given all the observations $\mathbf{x}$, we can analytically update $\boldsymbol{\xi}$ (instead of $\boldsymbol{\mu}$) and $\mathbf{\Omega}$ and follows:

$$\mathbf{\Omega}_{\mathbf{s}_{t-1}, \mathbf{s}_t} \leftarrow \mathbf{\Omega}_{\mathbf{s}_{t-1}, \mathbf{s}_t} + \begin{pmatrix} \mathbf{\Lambda}_{\mathbf{s}} & -\mathbf{\Lambda}_{\mathbf{s}} \\ -\mathbf{\Lambda}_{\mathbf{s}} & \mathbf{\Lambda}_{\mathbf{s}} \end{pmatrix}, \quad (12)$$

$$\boldsymbol{\xi}_{\mathbf{s}_t, \mathbf{z}} \leftarrow \boldsymbol{\xi}_{\mathbf{s}_t, \mathbf{z}} + \mathbf{H}_{\hat{\mathbf{s}}_t, \hat{\mathbf{z}}}^{\mathsf{T}} \mathbf{\Lambda}_{\mathbf{x}} \left(\mathbf{x}_t - \hat{\mathbf{x}}_t + \mathbf{H}_{\hat{\mathbf{s}}_t, \hat{\mathbf{z}}} [\hat{\mathbf{s}}_t^{\mathsf{T}}, \hat{\mathbf{z}}^{\mathsf{T}}]^{\mathsf{T}}\right), \quad (13)$$



(a) Recording environment (b) Mic (c) Mic + RGB (d) Mic + RGB + D (e) RGB + D

Fig. 3. Recording environment (a) and calibration results (b)–(e). Black circles and squares show the ground truth of the sensors and moving sources, respectively. Red circles and blue squares show the estimated sensor and sound positions, respectively.

$$\boldsymbol{\Omega}_{\mathbf{s}_t,\mathbf{z}} \leftarrow \boldsymbol{\Omega}_{\mathbf{s}_t,\mathbf{z}} + \mathbf{H}^{\mathsf{T}}_{\hat{\mathbf{s}}_t,\hat{\mathbf{z}}}\boldsymbol{\Lambda}_{\mathbf{x}}\mathbf{H}^{\mathsf{T}}_{\hat{\mathbf{s}}_t,\hat{\mathbf{z}}}, \tag{14}$$

If the observation $\mathbf{x}_t$ is partially missing (*e.g.*, due to the out-of-view and audio-clipping problems), (13) and (14) can be performed by deleting the corresponding rows of $\mathbf{H}_{\hat{\mathbf{s}}_t}$, $\mathbf{H}_{\hat{\mathbf{z}}_n}$, and $\boldsymbol{\Lambda}_{\mathbf{x}}$. (12), (13), and (14) are iterated until convergence and $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}^{-1}$ are obtained from conclusive $\boldsymbol{\xi}$ and $\boldsymbol{\Omega}$. The converged $\boldsymbol{\mu}_{\mathbf{s},\mathbf{z}}$ corresponds to the latent state $\mathbf{s}, \mathbf{z}$ to be estimated.

4. EVALUATION

This section reports the experimental evaluation of the proposed method by using a real audio-visual recording.

4.1. Experimental Conditions

Three audio-visual sensors (Microsoft Azure Kinect), each of which had seven microphones ($N = 3$ and $M = 7$), and a moving sound source (human) were placed in an anechoic room (Fig. 3-(a)). Only the depth camera of sensor 1 was available and the other sensors were placed inside the field of view of sensor 1. The sensor positions, orientations, and time offsets and the time-varying sound source position were estimated jointly using audio information only (Mic) or audio-visual information with/without depth information (Mic+RGB+D/Mic+RGB). The human made handclaps 13 times ($T = 13$) while moving straight as shown in Fig. 3. When the sound source was not observed visually, only audio information was used. The number of iterations was 100.

The hyperparameters were experimentally determined as $\sigma_x^2 = \sigma_y^2 = 1.5$ [m], $\sigma_q = \sigma_{q_n} = 0.2$ [m], $\sigma_{\theta_n^{\mathrm{a}}} = \sigma_{\theta_n^{\mathrm{v}}} = \sigma_{\phi_1} = 5$ [deg], and $\sigma_{\tau_1}^2 = 0.004$ [ms]. The diagonal elements of $\boldsymbol{\Lambda}_{\mathbf{z}}$ corresponding to the sensor orientation and time offset were set to 0.0001. The partial matrix corresponding to the sensor position was set to $(\mathbf{a}_n\mathbf{a}_n^{\mathsf{T}} + 0.1\mathbf{I})^{-1}$, where $\mathbf{a}_n \triangleq [\cos(\phi_{1n1}), \sin(\phi_{1n1})]^{\mathsf{T}}$ and $\mathbf{I}$ is an identity matrix.

4.2. Experimental Results

Table 1 shows the estimation errors obtained with the three variants of the proposed method. The use of visual information obtained from the RGB cameras significantly improved the performance of automatic sensor calibration. The use of the depth camera further slightly improved the performance in this experimental setting. The lack use of audio information, i.e., using only visual information (RGB+D), degraded the performance, especially of the sensor time offset estimation. When the cameras failed to detect the marker of the sound source, the visual information could naturally be treated as missing data. This is one of the main advantages of the proposed statistical approach based on the unified state-space model. Our calibration method is thus considered to make an audio-visual scene analysis system compact, portable, and easy-to-use, because it works well even when the depth camera is unavailable.

Table 1. Estimation errors with different observations.

Observations	Source pos. [m]	Sensor pos. [m]	Sensor orien. [deg]	Sensor time offset [ms]
Mic	0.83	0.36	27.7	0.44
Mic + RGB	0.38	0.22	2.88	0.55
Mic + RGB + D	0.20	0.14	3.50	0.27
RGB + D	0.34	0.17	4.07	3.68

5. CONCLUSION

This paper described an automatic calibration method for distributed camera-attached microphone arrays using a moving sound source as a reference. Our method is based on a unified state-space model that represents the generative process of the audio-visual observations from the latent sensor and source positions. The iterative GraphSLAM algorithm is used for estimating the latent variables that maximize the posterior probability. The experimental evaluation with recorded data demonstrated that the proposed audio-visual integrated method outperformed an audio-only method.

Future work includes online sensor calibration under dynamically changing conditions. The sensor arrangements and time offsets might be gradually changed in long-term recordings, *e.g.*, in a one-day-long demonstration of a robotic system. It is thus necessary to continuously monitor and calibrate the sensors using surrounding objects and sound events as references, *i.e.*, integrate sensor calibration with scene analysis into a unified statistical framework. We also plan to demonstrate the effectiveness of the proposed method with more realistic conditions, such as highly reverberant meeting rooms and noisy exhibition halls.

6. REFERENCES

[1] S. Wang, A. Mesaros, T. Heittola, T. Virtanen, I. Martin, A. Mesaros, I. Martin, T. Heittola, A. Mesaros, T. Virtanen, et al., "Audio-visual scene classification: Analysis of DCASE 2021 challenge submissions," in *DCASE 2021 Workshop*, 2021, pp. 45–49.

[2] A. Tsiami, P. P. Filntisis, N. Efthymiou, P. Koutras, G. Potamianos, and P. Maragos, "Far-field audio-visual scene perception of multi-party human-robot interaction for children and adults," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 6568–6572.

[3] H. Zhao, C. Gan, W.-C. Ma, and A. Torralba, "The sound of motions," in *Proc. Int. Conf. Comput. Vision*, 2019, pp. 1735–1744.

[4] J. Donley, V. Tourbabin, J.-S. Lee, M. Broyles, H. Jiang, J. Shen, M. Pantic, V. K. Ithapu, and R. Mehra, "Easycom: An augmented reality dataset to support algorithms for easy communication in noisy environments," in *arXiv*, 2021.

[5] D. F. Glas, T. Minato, C. T. Ishi, T. Kawahara, and H. Ishiguro, "ERICA: The ERATO intelligent conversational android," in *Proc. IEEE Int. Symp. Robot Hum. Interactive Commun.*, 2016, pp. 22–29.

[6] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 5, pp. 1557–1565, 2006.

[7] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, "A review of speaker diarization: Recent advances with deep learning," *Comput. Speech Lang.*, vol. 72, 2022.

[8] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.

[9] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *Proc. IEEE Workshop Autom. Speech Recognit Understanding*, 2011.

[10] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-end speech processing toolkit," in *Proc. INTERSPEECH*, 2018, pp. 2207–2211.

[11] T. Gburrek, C. Boeddeker, T. von Neumann, T. Cord-Landwehr, J. Schmalenstroeer, and R. Haeb-Umbach, "A meeting transcription system for an ad-hoc acoustic sensor network," in *Proc. INTERSPEECH*, 2022.

[12] A. J. Yang, C. Cui, I. A. Bârsan, R. Urtasun, and S. Wang, "Asynchronous multi-view SLAM," in *Proc. IEEE Int. Conf. Robotics Autom.*, 2021, pp. 5669–5676.

[13] S. Urban and S. Hinz, "Multicol-SLAM - a modular real-time multi-camera SLAM system," in *arXiv*, 2016.

[14] H. Miura, T. Yoshida, K. Nakamura, and K. Nakadai, "SLAM-based online calibration of asynchronous microphone array for robot audition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2011, pp. 524–529.

[15] D. Su, T. Vidal-Calleja, and J. V. Miro, "Simultaneous asynchronous microphone array calibration and sound source localisation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2015, pp. 5561–5567.

[16] F. Jacob, J. Schmalenstroeer, and R. Haeb-Umbach, "Microphone array position self-calibration from reverberant speech input," in *Proc. Int. Workshop Acoust. Signal Enhancement*, 2012, pp. 1–4.

[17] K. Sekiguchi, Y. Bando, K. Nakamura, K. Nakadai, K. Itoyama, and K. Yoshii, "Online estimation of asynchronous multiple microphone array positions using directions and time differences of arrivals," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2016, pp. 1973–1979.

[18] F. Jacob and R. Haeb-Umbach, "Coordinate mapping between an acoustic and visual sensor network in the shape domain for a joint self-calibrating speaker tracking," in *Proc. ITG Symp. Speech Comm.*, 2014, pp. 1–4.

[19] A. Plinge and G. A. Fink, "Geometry calibration of distributed microphone arrays exploiting audio-visual correspondences," in *Proc. Eur. Signal Process. Conf.*, 2014, pp. 116–120.

[20] S. Thrun and M. Montemerlo, "The GraphSLAM algorithm with applications to large-scale mapping of urban structures," *Int. J. Robot. Res.*, vol. 25, pp. 403–429, 2005.

[21] G. Grisetti, R. Kummerle, C. Stachniss, and W. Burgard, "A tutorial on graph-based SLAM," *IEEE Intell. Transp. Syst. Mag.*, vol. 2, no. 4, pp. 31–43, 2010.

[22] M. Montemerlo, S. Thrun, D. Koller, and B. Wegbreit, "FastSLAM: A factored solution to the simultaneous localization and mapping problem," in *Proc. AAAI Nat. Conf. Artif. Intell.*, 2002, pp. 593–598.

[23] C. Zhang, D. Florencio, and Z. Zhang, "Why does PHAT work well in low noise, reverberative environments?," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2008, pp. 2565–2568.