



Condition and Position Estimation Using Spatial Ultrasound and Machine Learning

Amir Kamarulzaman, M. Shahrul ; Sato, Shun ; Yasuda, Yuto ; Ohara, Ryotaro ; Izumi, Shintaro ; Kawaguchi, Hiroshi

(Citation)

IEEE Open Journal of Ultrasonics, Ferroelectrics, and Frequency Control, 6:53-63

(Issue Date)

2026

(Resource Type)

journal article

(Version)

Version of Record

(Rights)

© 2026 The Authors.

This work is licensed under a Creative Commons Attribution 4.0 License

(URL)

<https://hdl.handle.net/20.500.14094/0100502601>



Condition and Position Estimation Using Spatial Ultrasound and Machine Learning

M. SHAHRUL AMIR KAMARULZAMAN¹ (Graduate Student Member, IEEE), SHUN SATO², YUTO YASUDA², RYOTARO OHARA², SHINTARO IZUMI¹ (Member, IEEE), AND HIROSHI KAWAGUCHI¹

¹Graduate School of Science, Technology and Innovation, Kobe University, Kobe 657-8501, Japan

²Graduate School of System Informatics, Kobe University, Kobe 657-8501, Japan

CORRESPONDING AUTHOR: M. S. AMIR KAMARULZAMAN (muhammad.shahrul@cs28.cs.kobe-u.ac.jp)

ABSTRACT This study addresses the significant risk of bathroom-related injuries, often exacerbated by slippery surfaces and delays in accident detection, particularly as individuals frequently bathe unattended. Traditional accident detection methods, including those based on radio waves, optical sensors, and cameras, have inherent limitations. Although previous research has explored beamforming with low-cost ultrasound array sensors for object detection, its considerable computational load necessitates alternative approaches, such as machine learning. This paper presents a novel position and condition estimation method employing a Transformer-based model coupled with ultrasound array sensors tailored for bathroom environments. The proposed model architecture incorporates an encoder and two distinct decoders to generate binary segmentation masks and depth estimations, complemented by a multilayer perceptron classification block for condition estimation. The model was trained on binary masks and depth images acquired from a depth camera. Input signals received at 227 kHz are processed through cross-correlation with 25 kHz transmitted signals, followed by downsampling. The performance evaluation was conducted in two areas: bathtub and general washroom. In the bathtub area, condition estimation achieved $97.5 \pm 2.0\%$ accuracy, whereas position estimation yielded a mean absolute error (MAE) of 11.9 ± 2.9 cm and an intersection over union (IoU) of 0.735 ± 0.022 . Similarly, for the washroom area, condition estimation accuracy was $96.5 \pm 0.8\%$, with position estimation resulting in an MAE of 14.1 ± 2.0 cm and an IoU of 0.656 ± 0.023 . Notably, In the context of critical incident detection, the recall rates for drowning and falls were exceptionally high, exceeding 99.3 ± 1.5 and $99.5 \pm 1.0\%$, respectively. These results demonstrate the significant potential and effectiveness of the proposed method for robust accident detection within bathroom environments.

INDEX TERMS Bathroom safety, condition estimation, machine learning, position estimation, signal processing, Transformer, ultrasonic sensor array.

I. INTRODUCTION

BATHROOMS are inherently hazardous environments owing to slippery surfaces and hard floors and walls, which significantly increase the risk of severe injury in the event of a fall. Additionally, incidents of drowning, potentially caused by transient loss of consciousness resulting from fluctuations in blood pressure, are alarmingly frequent. In Japan, for instance, approximately 19,000 deaths in bathtubs are reported annually [1]. Notably, individuals aged ≥ 65 years account for 90% of these drowning fatalities [2], [3]. Given the global trend of an aging population, the inci-

dence of such fatal accidents is projected to rise further. In the United States, approximately 81% of bathroom injuries are attributed to falls [4]. Although installing non-slip mats and grab bars is recommended to mitigate the risk of falls, these preventative measures do not entirely eliminate the possibility of accidents. Moreover, because bathrooms are typically used in private, delays in detecting falls or drowning incidents are common, potentially leading to severe health consequences or fatalities. To address these critical safety concerns, developing a system that can swiftly detect accidents, such as falls and drowning, within bathroom

environments and promptly alert appropriate emergency or rescue services is imperative [5]. Such systems must accurately determine the location and condition of the individual, as this information is crucial for timely and effective interventions.

II. RELATED WORKS

Research on positioning systems has remained particularly active [6]. Outdoor positioning systems, exemplified by the Global Positioning System, are widely employed in automobile navigation and smartphone mapping applications, relying on satellite networks for precise location information. By contrast, indoor positioning systems predominantly utilize radio waves, such as ultra-wideband (UWB) and Bluetooth Low Energy (BLE). UWB [7], [8] provides high precision and rapid data transmission, whereas BLE is favored for its minimal power consumption and straightforward integration with existing devices. However, these technologies encounter challenges in bathroom environments, as carrying a receiving terminal, such as a beacon, is often impractical. Alternative methods that do not rely on such receiving terminals have been explored, including camera-based systems employing machine learning and optical sensors like light detection and ranging (LiDAR) [9], [10], [11]. For instance, De-La-Llana-Calvo et al. [9] reported a positioning error of approximately 10 cm using optical sensors and infrastructure lighting. Camera-based systems use advanced image processing to analyze visual data, whereas LiDAR generates detailed 3D maps by employing laser pulses. Despite their potential effectiveness, these methods introduce privacy concerns and are associated with high costs. Accordingly, ultrasonic positioning systems have emerged as a viable alternative. These systems employ inexpensive microphones and offer a non-intrusive method for tracking movement. Beamforming techniques utilizing ultrasonic arrays facilitate precise localization by directing signal waves toward specific areas [12], [13], [14]. Notably, in the acoustic domain, Qi and Liu [13] achieved a high localization accuracy with a maximum error of 1.02 cm for moving robots using a wireless sensor network. Integrating machine learning with beamforming techniques enhances accuracy and efficiency, enabling real-time monitoring and accident detection [15], [16]. For condition estimation, machine-learning techniques can be applied to identify human joints and skeletal structures, frequently utilizing algorithms derived from video game technologies [17], [18], [19]. However, the use of cameras in bathroom environments presents significant privacy issues. While specific optical systems, such as the photodetector-based system [4], have achieved high fall discrimination rates of 97.8%, they remain dependent on clear line-of-sight conditions. Given that the primary objective is to detect accidents such as falls and drowning, precise skeletal estimation is deemed unnecessary. Instead, the primary focus is to rapidly identify critical events and initiate appropriate responses. By approaching this as a classification problem [20], [21], machine learning

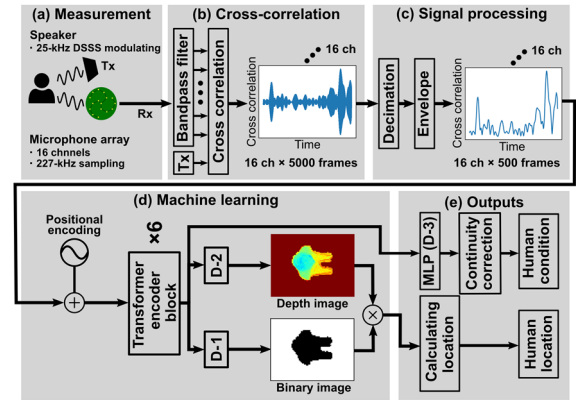


FIGURE 1. Proposed pipeline for human location and condition estimation. The system comprises five main stages: (a) measurement, (b) cross-correlation, (c) signal processing, (d) machine learning, and (e) outputs. The machine-learning stage uses a Transformer encoder and three distinct decoders: D-1 (depth decoder), D-2 (binary decoder), and D-3 (MLP-based condition classifier).

can be adopted to estimate conditions such as drowning and an unattended state, based on ultrasonic sensor data. This approach is designed to optimize the system for detecting specific conditions indicative of accidents, thereby enhancing its effectiveness and reliability in real-world applications.

III. ESTIMATION PIPELINE

The pipeline of the proposed method is illustrated in Fig. 1. Initially, ultrasonic waves emitted from a speaker reflect off the individual and are received by a 16-channel microphone array. The transmitted signal is modulated using a 25-kHz direct spreading method to enhance tolerance to interfering waves. The received signals, sampled at 227 kHz, undergo bandpass filtering, followed by the calculation of cross-correlation values between the transmitted and received signals. Subsequently, these signals are downsampled, and an envelope detection process is applied to reduce the input data size, primarily focusing on the signal peaks. The processed signals serve as input for a Transformer-based model. During the inference phase, the model generates two outputs: a binary segmentation mask (D-1 output) and a depth map (D-2 output). Crucially, the binary mask is employed to filter the depth map via element-wise multiplication. This masking process suppresses background noise and artifacts by zeroing out depth values in non-human regions. The final stages involve deriving the individual's position from this masked depth image by calculating its center of gravity. For condition estimation, a second training stage is employed, utilizing a classifier structured as a multilayer perceptron (MLP). In this stage, the pre-trained weights of the Transformer model are fixed, and only the classifier, comprising seven fully connected layers, is trained.

For the D-1 (depth decoder) and D-2 (binary decoder), each decoder block comprises a fully connected layer, followed by batch normalization and a rectified linear unit

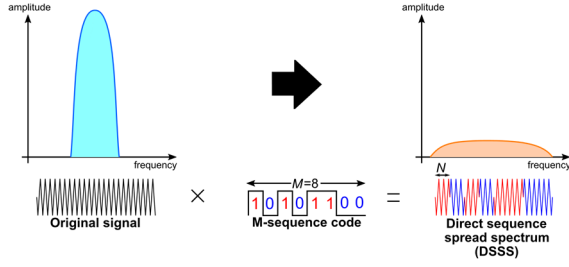


FIGURE 2. Principle of DSSS modulation, showing the original signal (left) spread across a wider bandwidth (right) after multiplication with an M-sequence code.

(ReLU) activation function. By contrast, the decoder block for the D-3 (Condition classifier) path includes only a fully connected layer and ReLU activation function. This architectural distinction reflects the differing output requirements: D-1 and D-2 generate image-like outputs (depth maps and binary masks, respectively), where batch normalization typically improves training convergence, whereas D-3 outputs categorical condition labels, for which such normalization is generally less critical.

A. SIGNAL PROCESSING

1) DIRECT SEQUENCE SPREAD SPECTRUM (DSSS)

DSSS is a modulation technique wherein the transmitted signal is multiplied by a spreading code, thereby dispersing the signal energy across a broader bandwidth than that of the original signal (Fig. 2). This approach significantly enhances the system's resilience to interference and obstructions.

In this study, transmitted signals are modulated using DSSS with an 8-bit M-sequence pseudo-random noise spreading code [22], [23]. The 8-bit M-sequence is produced by an $N = 8$ stage linear-feedback shift register (LFSR). Let $\alpha_i = (\alpha_i \alpha_{i+1} \dots \alpha_{i+7}) \in GF(2)^8$ represent the register state at the i -th clock. The next bit in the sequence is obtained through the following linear recurrence relation:

$$a_{i+8} = \sum_{j=0}^7 q_j a_{i+j} \pmod{2}, \quad (1)$$

where the feedback coefficients $\{q_i\}$ are determined by the *primitive* polynomial as follows:

$$f(x) = x^8 + q_7 x^7 + \dots + q_1 x + q_0 \quad (2)$$

Because $f(x)$ is a primitive polynomial over the binary field $GF(2)$, the register cycles through all $2^8 - 1 = 255$ nonzero states, yielding a maximal-length sequence (M-sequence). Such sequences possess a two-level autocorrelation ($R(0) = 255$ and $R(\tau) = -1$) for any nonzero delay ($\tau \neq 0$). They also exhibit minimal cross-correlation with shifted versions of themselves or independent M-sequences. These characteristics ensure that the energy of the spreading code is concentrated into a single sharp correlation peak while remaining nearly orthogonal to other codes, thereby minimizing mutual interference in cases of multiple transmitting elements operating simultaneously.

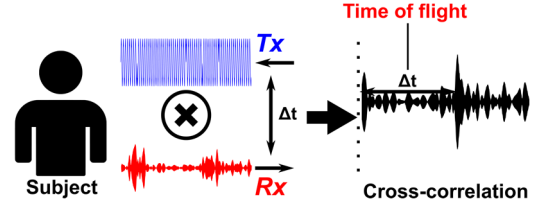


FIGURE 3. Cross-correlation process for ToF estimation.

2) CROSS-CORRELATION

Cross-correlation assesses the similarity between transmitted and received signals. It involves calculating the cross-correlation value between the signal emitted from the speaker and that received by each microphone channel (Fig. 3).

The prominent peak in the resulting cross-correlation function represents the arrival time (time of flight (ToF)) of the reflected wave, which enables the measurement of the distance to an object. This calculation is fundamental for determining the object's position within the environment. The cross-correlation function, denoted as $\hat{R}_{xy}$, is given by

$$\hat{R}_{xy}(m) = \begin{cases} \sum_{n=0}^{N-m-1} x_{n+m} y_n^*, & m \geq 0 \\ \hat{R}_{yx}^*(-m), & m < 0 \end{cases} \quad (3)$$

where x and y are the transmitted and received signals, respectively. Additionally, y_n^* denotes the Hermitian (complex) conjugate; however, because all the signals in this study are purely real, $y_n^* = y_n$. This function helps isolate relevant data points corresponding to the individual's location by identifying the peaks corresponding to the time delays of the reflected signals.

3) DOWNSAMPLING

To manage data volume and focus on critical information, the signals are downsampled to one-tenth of their original sampling rate. This reduction in data size improves computational efficiency without a significant loss of crucial signal information. This downsampling results in an input comprising 16 channels of signal data, where the original data length of approximately 5,000 samples per channel is reduced to around 500 samples per channel.

4) ENVELOPE

After downsampling, an envelope detection process is applied to the signal to further enhance data-processing efficiency. The resulting envelope represents the magnitude of the signal over time. This representation is crucial for the subsequent machine-learning stages as it clearly depicts the reflected signal's structure. This processed, enveloped data are then input into the Transformer-based machine-learning model, which analyzes them to determine the individual's position and condition.

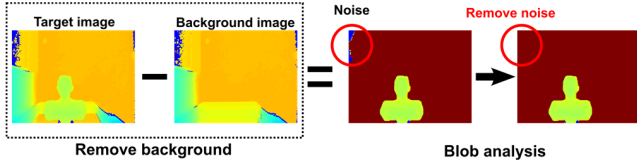


FIGURE 4. Process for generating ground-truth depth and binary mask images.

B. GROUND TRUTH

Ground-truth binary and depth images of 60×80 pixels were acquired using a depth camera. To generate a depth image that accurately reflects only the individual of interest, the difference between a background image (captured when no individual was present) and the individual's image (captured when they were present) is computed (Fig. 4). However, this depth image may contain noise attributable to reflections from surfaces such as walls and windows. To address this, blob analysis is employed to filter such noise, thereby retaining only the data corresponding to the relevant individual. This procedure enhances the accuracy of the ground-truth images for training the machine-learning model.

C. TRANSFORMER ARCHITECTURE

We employed a standard Transformer encoder architecture [24] adapted for time-series ultrasound data. To process the sequential input without recurrence, sinusoidal positional encodings are added to the input embeddings to preserve temporal order information.

The encoder consists of six stacked blocks. Each block contains a multi-head self-attention mechanism (with 10 heads) and a position-wise feed-forward network, employing residual connections and layer normalization. This architecture allows the model to capture long-range dependencies across the 16 input channels effectively. The encoder output is then fed into the task-specific decoders described previously.

IV. MEASUREMENT SYSTEM

This section details our measurement system.

A. SENSOR SYSTEM

The measurement system integrates hardware and software components developed in previous work. It uses a 25 kHz transducer, an array of micro-electro-mechanical systems (MEMS) microphones, and a field-programmable gate array (FPGA) to control signal transmission and reception. A local server is employed for data storage.

A PYNQ-Z2 board, featuring a Xilinx ZYNQ system-on-chip (SoC), was used. This SoC integrates an ARM CPU and programmable logic, enabling communication between the CPU and the FPGA. The board supports a Linux operating system, offering TCP/IP functionality and access to Linux file systems, thereby facilitating real-time data streaming and logging.

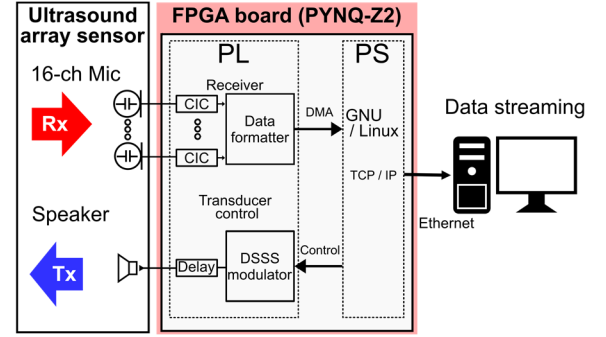


FIGURE 5. Block diagram of the ultrasound signal measurement system, detailing components from the ultrasound array sensor to the data streaming server.

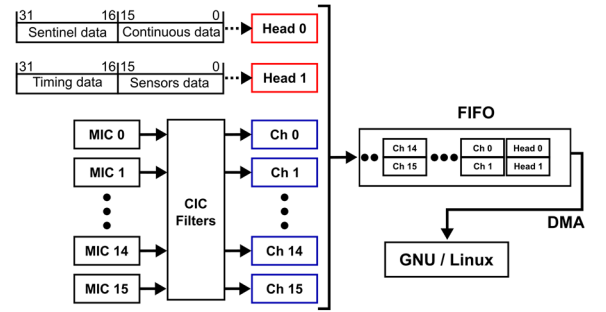


FIGURE 6. Data packet format structure.

The FPGA is divided into two primary functional units: speaker control and microphone array control (Fig. 5). The speaker control unit generates ultrasound signals employing DSSS modulation with a specified M-sequence at 25 kHz (8-bit, 8 waves per bit). The microphone array control unit receives pulse density modulation signals from the microphones; these signals are subsequently demodulated into 227-kHz pulse code modulation (PCM) signals using a cascaded integrator-comb (CIC) filter implemented on the FPGA. The resulting PCM signals are then packetized and transmitted to the storage server.

The data format, illustrated in Fig. 6, consists of ten 32-bit words. The first two words are employed in the header, whereas the subsequent eight words contain signal data. In the first header word, the upper 16 bits represent an initial sentinel node, and the lower 16 bits function as a continuity counter to ensure signal integrity. The first 16 bits of the second header word contain timing data, such as signal transmission or Bluetooth synchronization; the last 16 bits of this word are reserved for future use. The remaining data contain the PCM signals for 16 channels, with each channel's data typically divided between upper and lower 16-bit segments within these words.

The data-processing workflow, encompassing operations from the circuit level through software on both the FPGA and the local server, is as follows: To prevent data loss owing to processing delays, tasks are distributed across multiple

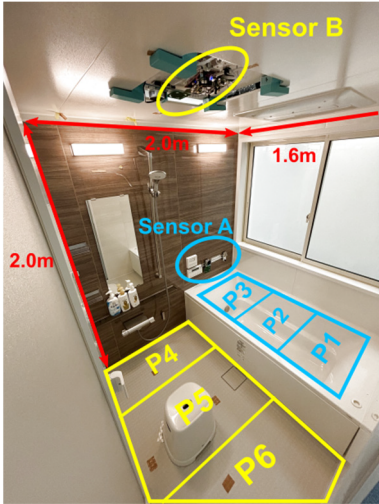


FIGURE 7. Experimental setup in the bathroom environment, showing the placement of Sensor A (monitoring bathtub areas P1–P3) and Sensor B (monitoring washroom areas P4–P6).

TABLE 1. Key signal characteristics and system parameters.

Item	Value
Code length ($M \times N$)	64
Carrier Frequency	25 kHz
Sampling rate	227 kHz
Array type	Fermat Spiral
Frame rate	2 fps
Interval	0.25 s

components. The FPGA utilizes direct memory access to read the packetized data, which are subsequently transmitted to the local server via TCP. For enhanced processing speed, the FPGA software is implemented in C++. On the local server, data reception, storage, and visualization are managed by Python-based scripts. The received data are concurrently forwarded to both the writing and visualization processes. The writing process stores the data on the local server, whereas the visualization process displays them in real-time, enabling the monitoring of any anomalies during the experiments.

B. MEASUREMENT ENVIRONMENT

To evaluate the proposed method, two sensor units were installed in a bathroom environment: Sensor A was positioned to monitor activities in the bathtub area from a frontal perspective, whereas Sensor B was mounted on the ceiling for surveillance of the general washroom area (Fig. 7). Although both sensors utilize different DSSS modulation codes, they operate within the same frequency band; this design results in correlation values that significantly exceed typical environmental noise levels. To prevent interference between the sensors, their transmissions are time-staggered by 0.25 s, with each sensor transmitting once every 0.5 s. Detailed signal specifications are provided in TABLE 1. As indicated in Fig. 7, P1–P6 represent distinct positions of

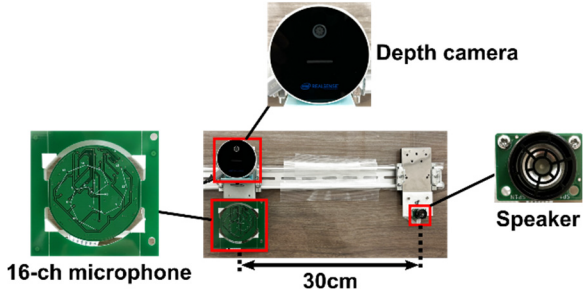


FIGURE 8. Components of the experimental sensor setup.

TABLE 2. Specifications of the intel RealSense L515 depth camera.

Item	Value
Field of view	$70^{\circ} \times 55^{\circ}$
Operating range (min-max)	0.25–9.00 m
Depth resolution	1024×768
Frame rate	30 fps

TABLE 3. Description of experimental scenarios.

Scenario ID	Condition	Description
A-0 / B-0	Background (empty) in bathtub / washroom	No participant in the bathroom (to differentiate subject’s presences)
A-1 to A-6	Static poses in bathtub	Subjects maintaining specific postures (standing, sitting, drowning) at fixed locations (P1 – P3)
B-1 to B-6	Static poses in washroom	Subjects maintaining specific postures (standing, sitting, lie down) at fixed locations (P4 – P6)
A-7 / B-7	Sequential activities in bathroom / washroom	Subjects moving freely and transitioning between postures (used for testing robustness in real-life activities)

the individual, which are further detailed in the subsequent subsection.

The measurement setup is shown in Fig. 8. An analog MEMS microphone (SPU410-338, SPU0410LR5H-QB; Knowles Electronics), with dimensions of 3.76×3.00 mm, was used. The transducer element employed was the CUSP-T80-12-2600-TH (Same Sky (formerly CUI Devices)). The microphone array comprised 16 such microphone elements arranged in a spiral configuration. The Intel RealSense L515 depth camera was used to acquire ground-truth data. Its specifications are listed in TABLE 2.

C. MEASUREMENT DETAILS

This section describes the experimental setup and procedure. The experiments were categorized into eight scenarios per area (A-0 to A-7 for bathtub area and B-0 to B-7 for washroom area), as detailed in TABLE 3. Six predefined points

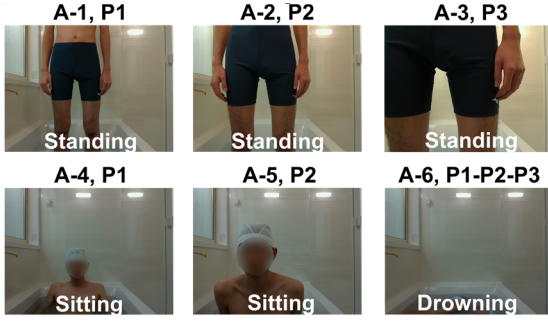


FIGURE 9. Examples of participants' conditions and positions for Experiment A scenarios in the bathtub (water level 33 cm).

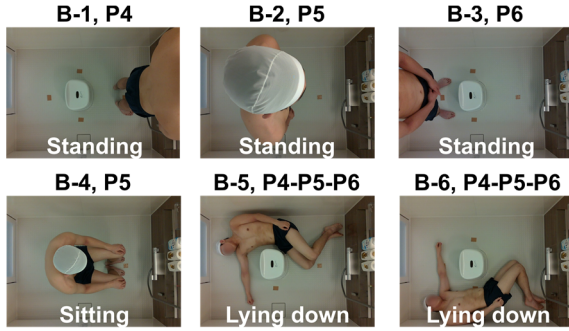


FIGURE 10. Examples of participants' conditions and positions for Experiment B scenarios in the washing area.

(Fig. 7) were used as reference locations. Four participants, denoted as P, Q, R, and S, participated in the experiment. A cross-validation methodology was implemented, wherein data from three participants were used to train the model, and those from the remaining participant were used for testing. The height difference among the participants ranged from 11 to 13 cm, and the weight difference ranged from 11 to 15 kg.

Although a full combination of locations (three positions) and conditions (three postures) would yield nine scenarios, due to physical constraints only six were chosen. For instance, “sitting” posture at P3 is excluded due to the presence of the faucet on the bathtub. Additionally, since the “drowning” postures across P1 to P3 are geometrically identical; only one combination (A-6) is included in the scenarios. Similarly, “sitting” in the washroom is excluded at P4 due to another faucet on the wall and at P6 due to space limitation. The “lying down” in the middle are excluded because of the stool. See Fig. 9 and Fig. 10

1) BATHTUB (WATER LEVEL = 33 CM)

This set of experiments involved Sensor A monitoring the P1–P3 bathtub areas, with an initial water level of 33 cm. the experimental scenarios are detailed in TABLE 3 and illustrated in Fig. 9. Data collection time varied by scenario complexity: static poses lasted approximately 3 min (2 min training, 1 min testing), while drowning (A-6) and sequential

(A-7) were extended to 5 min and 7 min respectively to capture sufficient variation. Notably, owing to the acoustic similarity between a fully submerged individual and an empty bathtub, these states were consolidated into a single class. Consequently, conditions were classified into three categories: standing, sitting, and unattended/drowning.

2) BATHTUB (WATER LEVEL = 38 CM)

This experiment was designed to test the effect of changes in water level. For this specific test, test data were collected exclusively from subject P. The model used for this evaluation was the same as that trained on data from the previously described 33-cm water level experiments (Subsection C.1). Data from this participant were not included in the training set.

3) WASHING AREA

In this set of experiments, Sensor B monitored the P4–P6 washing areas (Fig. 10), while the bathtub was maintained at a water level of 33 cm. Four distinct condition classes were considered for this area: unattended, standing, sitting, and lying (prone). Eight experimental scenarios (B-0–B-7) were implemented, adhering to the same time constraints for data acquisition detailed in Subsection C.1. Similar to the bathtub experiments, scenarios involving random movements by the participant within the washing area (B-7) and an unattended state were included. Furthermore, cross-validation was also performed for the data collected in these washing area experiments.

V. TRAINING AND EVALUATION

A. TRAINING PROCEDURE

The proposed model was implemented using PyTorch. The training process utilized the Adam optimizer with a learning rate of 0.0001. The model was trained for 300 epochs with a mini-batch size of 128 ($n = 128$). To balance the multi-task learning objectives, the loss function for position estimation, presented in (4), combines binary cross-entropy for binary-image outputs (b_i : ground-truth, $\hat{b}_i$: inference) and mean squared error (MSE) for depth-image outputs (d_i : ground-truth, $\hat{d}_i$: inference), weighted by factors $\lambda = 3$ and $\nu = 20,000$ respectively (as determined by preliminary tuning).

$$Loss = \lambda \bullet \left\{ - \sum_{i=1}^n b_i \log \hat{b}_i + (1 - b_i) \log (1 - \hat{b}_i) \right\} + \nu \bullet \left\{ \frac{1}{n} \sum_{i=1}^n (\hat{d}_i - (d_i * \hat{b}_i))^2 \right\} \quad (4)$$

B. EVALUATION METRICS

To ensure the robustness of the results, we employed a leave-one-subject-out cross-validation scheme. Data from 3 subjects were used for training, while data from the fourth subject served as the test set. This process was rotated so that every subject served as the test subject once.

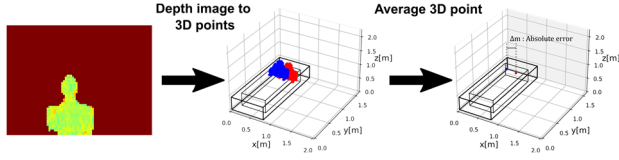


FIGURE 11. Depth image to 3D median point transformation for MAE calculation.

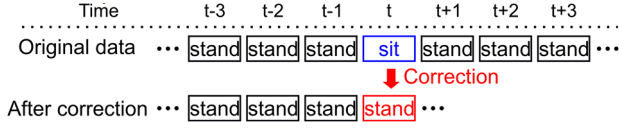


FIGURE 12. Application of continuity correction to classified states over time.

1) LOCATION ESTIMATION

Location-estimation performance was evaluated using two metrics: intersection over union (IoU) and mean absolute error (MAE). IoU quantifies the overlap between two regions and is calculated by dividing the area of intersection by that of their union; a maximum value of 1 signifies perfect overlap.

The participant's position was determined from the three-dimensional center of gravity coordinates derived from depth images. These images were first converted into 3D point clouds. The MAE was then calculated by comparing the center of gravity of the ground-truth point cloud with that of the inferred point cloud (as depicted in the transformation method shown in Fig. 11).

2) CONDITION ESTIMATION

The condition-estimation performance was assessed using accuracy (5), recall (6), and precision (7). Given the safety-critical nature of this application, recall was prioritized as the most significant metric, as minimizing false negatives (i.e., missing a drowning or fall incident) is paramount for preventing fatalities. Recall, calculated per class, indicates the model's ability to correctly identify true instances of that class. Precision, also calculated per class, measures the proportion of correct positive predictions for that class and was included to evaluate the system's robustness against false alarms (alarm fatigue). Accuracy measures the overall proportion of correct predictions among the total cases, serving as a general benchmark for system stability.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (5)$$

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

Additionally, the raw classification results were refined using a continuity correction method (Fig. 12). This involved considering the majority class over a sliding window of

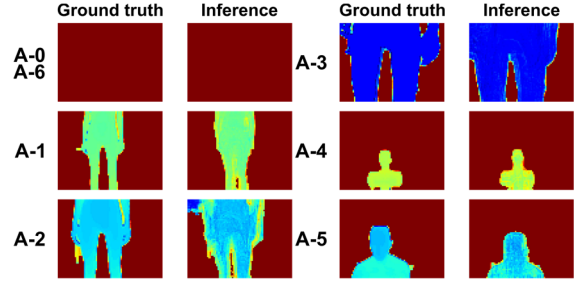


FIGURE 13. Qualitative comparison of ground-truth and inferred depth images from Sensor A for various experiment A scenarios.

TABLE 4. Average cross-validation evaluation results for position and condition estimation in bathtub area with 33-cm water level.

Metric	Results			
	Position estimation			
IoU	0.735 ± 0.022			
MAE [cm]	11.9 ± 2.9			
	Condition estimation			
	Accuracy [%]	Stand	Sit	Drown
		97.5 ± 2.0		
Recall [%]	98.1 ± 1.7	96.2 ± 3.8	99.3 ± 1.5	
Precision [%]	99.3 ± 0.8	98.3 ± 0.6	94.5 ± 6.8	

seven consecutive frames (three frames before and three after the target frame, plus the target frame itself). All evaluation metrics reported in the condition estimation are calculated after the continuity correction.

VI. EXPERIMENTAL RESULTS

This chapter presents the evaluation metrics and corresponding results for the aforementioned experiments.

The results for each experiment are presented in the following subsections. These results were averaged across all participants. To provide realistic insights while maintaining conciseness, we focus our quantitative evaluation on Scenarios A-7 and B-7 (Sequential activities). These scenarios represent the most challenging conditions where the participant transitions between states. The results presented below are the average ± standard deviation across the cross-validation folds.

1) BATHTUB (WATER LEVEL 33 CM):

TABLE 4 presents the average estimation results for the bathtub area experiments conducted with a water level of 33 cm. The model achieved an IoU of 0.735 ± 0.022 and an MAE of 11.9 ± 2.88 cm, indicating its ability to accurately determine the participant's position, even in instances of random movements. Fig. 13 illustrates that the model provided clear inference results.

By applying continuity correction, the condition-estimation accuracy reached 97.5 ± 2.0%, demonstrating reliable detection of standing, sitting, and unattended/drowning states. Some misclassifications were observed

TABLE 5. Confusion matrix for condition estimation in experiment A-7 (bathtub area, 33-cm water level).

		Prediction		
True		Stand	Sit	Drown
	Stand	81	2	0
	Sit	0	115	5
	Drown	0	0	76

TABLE 6. Evaluation results of subject p for position and condition estimation in bathtub area with 38-cm water level.

Metric	Results		
Position estimation			
IoU	0.751		
MAE [cm]	9.1		
Condition estimation			
Accuracy [%]	99.6		
	Stand	Sit	Drown
Recall [%]	100.0	98.8	100.0
Precision [%]	100.0	100.0	98.8

TABLE 7. Confusion matrix for condition estimation in experiment A-7 (bathtub area, 38-cm water level).

		Prediction		
True		Stand	Sit	Drown
	Stand	67	0	0
	Sit	0	81	1
	Drown	0	0	81

between the sitting and drowning states, likely attributable to head movements occurring while participants held their breath. The confusion matrix for these experiments is provided in TABLE 5. Misclassifications frequently involved transitions between conditions, such as from standing to sitting, which are inherently challenging to label with perfect consistency. The recall rate for the drowning class, a crucial metric for anomaly detection, reached an average of $99.3 \pm 1.5\%$.

2) BATHTUB (WATER LEVEL = 38 CM):

This experiment evaluated the model's performance with the water level increased to 38 cm, utilizing the model trained on 33-cm water-level data. as specified in Subsection IV-C.2, data for this specific evaluation were collected exclusively from Participant P. Consequently, the results displayed in TABLE 6 represent the performance for this single subject and are reported without standard deviations. An IoU of 0.751 and an MAE of 9.1 cm were achieved in Experiment A-7, indicating high position-estimation accuracy. The condition-estimation accuracy was consistent with that obtained for the 33-cm water level, suggesting that water level changes within 5 cm did not significantly affect performance.

TABLE 7 shows the confusion matrix for this experiment.

3) WASHING AREA:

TABLE 8 presents the position-estimation results for the washing area. The overall accuracy is lower than that in

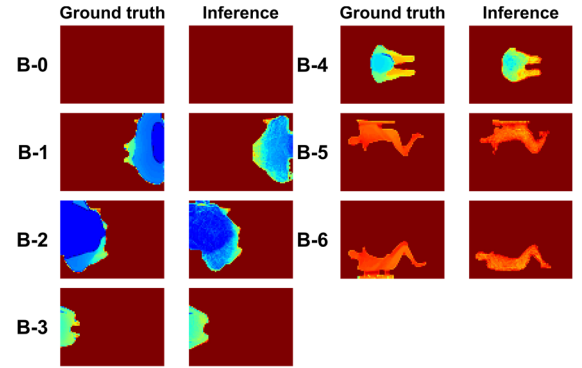


FIGURE 14. Qualitative comparison of ground-truth and inferred depth images from Sensor B for selected experiment B scenarios in the washing area.

TABLE 8. Average cross-validation evaluation results for position and condition estimation in the washing area.

Metric	Results		
Position estimation			
IoU	0.656 ± 0.023		
MAE [cm]	14.1 ± 2.0		
Condition estimation			
Accuracy [%]	96.5 ± 0.8		
	Stand	Sit	Lie down
Recall [%]	93.9 ± 3.4	97.0 ± 3.7	99.5 ± 1.0
Precision [%]	98.2 ± 3.0	96.7 ± 4.4	94.8 ± 2.6

TABLE 9. Confusion Matrix for condition estimation in experiment B-7 (washing area).

		Prediction		
True		Stand	Sit	Lie down
	Stand	110	1	5
	Sit	2	44	0
	Lie down	1	0	102

the bathtub area. This decrease was likely attributable to the sensor's overhead perspective and the greater variability in participants' postures and conditions within this area. Both IoU and MAE varied significantly among participants, potentially owing to differences in their specific standing and sitting positions. However, the condition-estimation results revealed a high average accuracy. We also noted that environmental factors, such as temperature and humidity, affected the sensor's performance, leading to observable accuracy fluctuations. Fig. 14 shows an example of the inference results obtained with the proposed method for this area.

The recall and precision for Experiment B-7 are detailed in TABLE 8, and the corresponding confusion matrix is provided in TABLE 9. The "lie down" class, which is critical for detecting fallen individuals, achieved a recall of $99.5 \pm 1.0\%$ with continuity correction applied. These results demonstrate the model's high accuracy for detecting fallen individuals. Nevertheless, further improvements are

warranted to enhance robustness against various environmental conditions.

VII. DISCUSSION

A. COMPARISON WITH EXISTING METHODS

In this section we contrast the proposed method with existing positions and conditions/postures solutions. While camera-based systems offer high resolution, they suffer from significant privacy concerns and performance degradation in steamy environments. Wearable sensors (e.g., accelerometers) avoid privacy issues but require user compliance, which is impractical.

In contrast, our acoustic method preserves privacy and requires no contact. Furthermore, compared to traditional ultrasound beamforming, which struggles with the complex multipath reverberations typical of tiled bathrooms, our Transformer-based approach effectively learns to interpret these spatial features, achieving high recall rates (>99%) for critical events.

Quantitatively, our system's performance is competitive with existing solutions. Regarding positioning, our MAE of 11.9 cm is similar to the 10 cm error reported in privacy-preserving optical systems [9]. Although this is less precise than the 1.02 cm accuracy of active ultrasound beacons used for robotics [13], sub-decimeter precision is not required for bathroom safety monitoring, where the primary goal is distinguishing between broad zones. Regarding condition estimation, our Recall of 99.3% for drowning events is comparable to the 97.8% accuracy reported in photodetector-based systems [4], but achieves this without the need for line-of-sight optical paths.

B. PRACTICAL IMPLEMENTABILITY AND LATENCY

For real-world deployment, the system must process data in real-time. The proposed Transformer model is computationally lightweight compared to vision-based deep learning models. On our experimental server (RTX 4090), the average inference time per frame was approximately less than 50 ms.

Given that the current system operates at a sampling interval of 0.5 seconds (2 fps), the inference time is well within the margin required for real-time processing. While falls occur instantaneously, the critical requirement for a safety system is the rapid detection of the post-fall state or drowning state to trigger an alert. A latency of under one second is sufficient for timely intervention in these scenarios. However, since this method still relied on depth camera for inference, a few changes are needed for public uses.

VIII. CONCLUSION

This study proposed a method for estimating an individual's position and condition within a bathroom environment using ultrasonic sensors integrated with machine-learning techniques. A 16-channel spiral microphone array was

employed to capture reflected ultrasonic signals, and the cross-correlation data derived from transmitted and received signals served as input for a Transformer-based model.

IoU and MAE were employed to evaluate position-estimation performance. For assessing condition-estimation performance, accuracy, recall, and precision were used as primary metrics, with a continuity correction technique subsequently applied to the classification outputs.

Experiments were conducted in two distinct locations within the bathroom. For the bathtub area (with a water level of 33 cm), Sensor A (positioned on the wall) achieved an average IoU of 0.735 and an MAE of 11.9 cm for position estimation, and a condition-estimation accuracy of 97.5%, which included a 99.3% recall rate for detecting drowning incidents. For the washing area, Sensor B (mounted on the ceiling) achieved an IoU of 0.655 and an MAE of 14.1 cm for position estimation, and a condition-estimation accuracy of 96.5%, with a 99.5% recall rate for detecting fallen individuals.

The effect of increasing the bathtub water level to 38 cm was also tested, and the estimation accuracy remained stable under this condition.

Future work will focus on expanding the training set to incorporate a wider variety of bathroom layouts, additional activities such as showering or washing, and diverse environmental conditions, including varying temperature and humidity levels. These efforts are expected to further improve the overall robustness of the proposed system.

ACKNOWLEDGMENT

Shun Sato, Yuto Yasuda, and Ryotaro Ohara were with the Graduate School of System Informatics, Kobe University, Kobe, Japan. An earlier version of this paper was presented in part at the 2024 IEEE Ultrasonics, Ferroelectrics, and Frequency Control Joint Symposium [DOI: 10.1109/UFFC-JS60046.2024.10793655]. This study involved only voluntary participation by the authors and laboratory members. No personal or identifiable data were collected, and no formal ethical approval was required under the policies of Kobe University.

REFERENCES

- [1] Y. Kanawaku, "Deaths in bathtubs in Japan: Forensic and clinical implications," *J. Nippon Med. School*, vol. 89, no. 1, pp. 33–39, 2022, doi: [10.1272/jnms.jnms.2022_89-119](https://doi.org/10.1272/jnms.jnms.2022_89-119).
- [2] M. Katsuyama et al., "Development of prevention strategies against bath-related deaths based on epidemiological surveys of inquest records in Kagoshima prefecture," *Sci. Rep.*, vol. 13, no. 1, p. 2277, Feb. 2023, doi: [10.1038/s41598-023-29400-7](https://doi.org/10.1038/s41598-023-29400-7).
- [3] J. A. Stevens, E. N. Haas, *Nonfatal Bathroom Injuries Among Persons Aged ≥15 Years—United States, 2008*. Centers for Disease Control and Prevention. Accessed: Jun. 20, 2024. [Online]. Available: https://www.cdc.gov/mmwr/preview/mmwrhtml/mm6022a1.htm?s_cid=mm6022a1_w
- [4] Y. Kajiwaru, Y. Kobayashi, M. Nakamura, K. Hayashi, and H. Kimura, "Development of a fall detection system in bathroom using photodetectors," *IEEJ Trans. Sensors Micromachines*, vol. 132, no. 11, pp. 387–396, 2012, doi: [10.1541/ieejsmas.132.387](https://doi.org/10.1541/ieejsmas.132.387).

- [5] S. M. Asaad and H. S. Maghddid, "A comprehensive review of indoor/outdoor localization solutions in IoT era: Research challenges and future perspectives," *Comput. Netw.*, vol. 212, Jul. 2022, Art. no. 109041, doi: [10.1016/j.comnet.2022.109041](https://doi.org/10.1016/j.comnet.2022.109041).
- [6] N. Zhou, L. Lau, R. Bai, and T. Moore, "A robust detection and optimization approach for delayed measurements in UWB particle-filter-based indoor positioning," *NAVIGATION, J. Inst. Navigat.*, vol. 69, no. 2, Jun. 2022, Art. no. navi.514, doi: [10.33012/navi.514](https://doi.org/10.33012/navi.514).
- [7] S. Mazuelas et al., "Robust indoor positioning provided by real-time RSSI values in unmodified WLAN networks," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 5, pp. 821–831, Oct. 2009, doi: [10.1109/JSTSP.2009.2029191](https://doi.org/10.1109/JSTSP.2009.2029191).
- [8] T. Yokota and Y. Kuroda, "3P1-K04 human recognition by extracting shape features from LiDAR data(3D measurement/sensor fusion)," in *Proc. JSME Annu. Conf. Robot. Mechatron. Robomec*, 2014, p. 3, doi: [10.1299/jsmermd.2014_3P1-K04_1](https://doi.org/10.1299/jsmermd.2014_3P1-K04_1).
- [9] Á. De-La-Llana-Calvo, J. L. Lázaro-Galilea, A. Alcázar-Fernández, A. Gardel-Vicente, I. Bravo-Muñoz, and A. Iamnitchi, "Accuracy and precision of agents orientation in an indoor positioning system using multiple infrastructure lighting spotlights and a PSD sensor," *Sensors*, vol. 22, no. 8, p. 2882, Apr. 2022, doi: [10.3390/s22082882](https://doi.org/10.3390/s22082882).
- [10] S. Sato et al., "Estimating person location in bathroom with spatial ultrasound and variational autoencoder," in *Proc. IEEE Int. Ultrason. Symp. (IUS)*, Sep. 2023, pp. 1–4, doi: [10.1109/IUS51837.2023.10306421](https://doi.org/10.1109/IUS51837.2023.10306421).
- [11] Y. Yoshikawa, Y. Yasuda, I. Toru, S. Izumi, and H. Kawaguchi, "12.5-m distance measurement in high-interference environment using ultrasonic array sensors," in *Proc. IEEE Int. Instrum. Meas. Technol. Conf. (I2MTC)*, May 2021, pp. 1–6, doi: [10.1109/I2MTC50364.2021.9459822](https://doi.org/10.1109/I2MTC50364.2021.9459822).
- [12] G. Allevato et al., "Air-coupled ultrasonic spiral phased array for high-precision beamforming and imaging," *IEEE Open J. Ultrason., Ferroelectr., Freq. Control*, vol. 2, pp. 40–54, 2022, doi: [10.1109/OJUFFC.2022.3142710](https://doi.org/10.1109/OJUFFC.2022.3142710).
- [13] J. Qi and G.-P. Liu, "A robust high-accuracy ultrasound indoor positioning system based on a wireless sensor network," *Sensors*, vol. 17, no. 11, p. 2554, Nov. 2017, doi: [10.3390/s17112554](https://doi.org/10.3390/s17112554).
- [14] X. Wu, J. Luo, G. Li, S. Zhang, and W. Sheng, "Fast wideband beamforming using convolutional neural network," *Remote Sens.*, vol. 15, no. 3, p. 712, Jan. 2023, doi: [10.3390/rs15030712](https://doi.org/10.3390/rs15030712).
- [15] A. A. Nair, M. R. Gubbi, T. Duy Tran, A. Reiter, and M. A. L. Bell, "A fully convolutional neural network for beamforming ultrasound images," in *Proc. IEEE Int. Ultrason. Symp. (IUS)*, Oct. 2018, pp. 1–4, doi: [10.1109/ULTSYM.2018.8579960](https://doi.org/10.1109/ULTSYM.2018.8579960).
- [16] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1653–1660, doi: [10.1109/CVPR.2014.214](https://doi.org/10.1109/CVPR.2014.214).
- [17] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4724–4732, doi: [10.1109/CVPR.2016.511](https://doi.org/10.1109/CVPR.2016.511).
- [18] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 648–656, doi: [10.1109/CVPR.2015.7298664](https://doi.org/10.1109/CVPR.2015.7298664).
- [19] K. Leetang, S. Hirata, and H. Hachiya, "Selection on the combination of M-sequence codes in alternate transmission for extension of measurable distance," *Jpn. J. Appl. Phys.*, vol. 58, no. 7, Jun. 2019, Art. no. 076503, doi: [10.7567/1347-4065/ab25d5](https://doi.org/10.7567/1347-4065/ab25d5).
- [20] S. Sato, R. Ohara, M. S. A. Kamarulzaman, Y. Yasuda, S. Izumi, and H. Kawaguchi, "An ultrasound-based surveillance system for bathroom posture and location estimation," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, Las Vegas, NV, USA, Jan. 2024, pp. 1–5, doi: [10.1109/ICCE59016.2024.10444407](https://doi.org/10.1109/ICCE59016.2024.10444407).
- [21] M. S. A. Kamarulzaman, S. Sato, R. Ohara, Y. Yasuda, I. Shitaro, and H. Kawaguchi, "Position(-) and posture-estimation method in bathroom using spatial ultrasound," in *Proc. IEEE Ultrason., Ferroelectr., Freq. Control Joint Symp. (UFFC-JS)*, Sep. 2024, pp. 1–4, doi: [10.1109/UFFC-JS60046.2024.10793655](https://doi.org/10.1109/UFFC-JS60046.2024.10793655).
- [22] W. Kim, "Signal-to-noise ratio and bandwidth for pseudo-random codes in an ultrasonic imaging system," *Ultrason. Imag.*, vol. 6, no. 3, pp. 313–323, Jul. 1984, doi: [10.1016/0161-7346\(84\)90016-6](https://doi.org/10.1016/0161-7346(84)90016-6).
- [23] V. Ashish, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 3–6. Accessed: May 6, 2025. [Online]. Available: <https://arxiv.org/abs/13708.4994v2>
- [24] Y. Liu et al., "ITransformer: Inverted transformers are effective for time series forecasting," 2023, *arXiv:2310.06625*.



M. SHAHRUL AMIR KAMARULZAMAN

(Graduate Student Member, IEEE) received the B.Eng. degree in computer science and systems engineering and the M.Eng. degree in science, technology and innovation from Kobe University, Kobe, Japan, in 2022 and 2024, respectively. He is currently engaged in research focusing on the development of a robust signal processing method for a bather monitoring system in bathroom environments using ultrasonic array sensors.



SHUN SATO received the B.Eng. degree in computer science and systems engineering and the M.Eng. degree in system informatics from Kobe University, Kobe, Japan, in 2022 and 2024, respectively. His previous research interest was the development of a bather monitoring system for bathroom environments using ultrasonic array sensors and machine learning.



YUTO YASUDA received the B.Eng. degree in computer science and systems engineering and the M.Eng. degree in system informatics from Kobe University, Kobe, Japan, in 2021 and 2023, respectively. His previous research focused on the development of a bather monitoring system for bathroom environments using ultrasonic array sensors and machine learning.



RYOTARO OHARA received the Foundation degree from Tokyo Metropolitan College of Industrial Technology, Tokyo, Japan, in 2016, the B.Eng. degree in environmental and life science from Toyohashi University, Toyohashi, Japan, in 2018, and the master's degree in science, technology, and innovation and the Ph.D. degree in system informatics from Kobe University, Hyogo, Japan, in 2021 and 2024, respectively. He currently focuses his research on 3-D environment monitoring using ultrasonic technology.



SHINTARO IZUMI (Member, IEEE) received the B.Eng. and M.Eng. degrees in computer science and systems engineering and the Ph.D. degree in engineering from Kobe University, Hyogo, Japan, in 2007, 2008, and 2011, respectively. From 2009 to 2011, he was a JSPS Research Fellow with Kobe University. He was an Assistant Professor with the Organization of Advanced Science and Technology, Kobe University, from 2011 to 2018; and an Associate Professor with the Institute of Scientific and Industrial Research, Osaka University, Osaka, Japan, from 2018 to 2019. Since 2019, he has been an Associate Professor with the Graduate School of System Informatics, Kobe University. His current research interests include biomedical signal processing, communication protocols, low-power VLSI design, and sensor networks. He received the 2010 IEEE SSCS Japan Chapter Young Researchers Award. He has served as a Technical Committee Member for IEEE Biomedical and Life Science Circuits and Systems, a Student Activity Committee Member for the IEEE Kansai Section, and a Program Committee Member for the IEEE Symposium on Low-Power and High-Speed Chips (COOL Chips). Previously, he chaired the IEEE Kansai Section Young Professionals Affinity Group.



HIROSHI KAWAGUCHI received the B.Eng. and M.Eng. degrees in electronic engineering from Chiba University, Chiba, Japan, in 1991 and 1993, respectively, and the Ph.D. degree in electronic engineering from The University of Tokyo, Tokyo, Japan, in 2006. He began his career with Konami Corporation, Kobe, Japan, in 1993, where he developed arcade entertainment systems. He transitioned to the Institute of Industrial Science, The University of Tokyo, as a Technical Associate, in 1996, and was later appointed as a Research Associate in 2003. In 2005, he joined the Graduate School of Engineering, Kobe University, Hyogo, Japan, as a Research Associate. From 2015 to 2016, he was a Visiting Researcher with the Politecnico di Milano, Milan, Italy. Since 2016, he has been a Full Professor with the Graduate School of Science, Technology, and Innovation, Kobe University. He is also a Collaborative Researcher with the Institute of Industrial Science, The University of Tokyo. His current research interests encompass a wide range of topics, including low-voltage operating circuits, soft error characterization and mitigation, ubiquitous sensor networks, organic semiconductor circuits, data converters, healthcare devices, and neuro computer architecture. He is a member of the Association for Computing Machinery (ACM) and the Institute of Electronics, Information and Communication Engineers (IEICE). He has received several awards, including the IEEE ISSCC 2004 Takuo Sugano Outstanding Paper Award, the ACM/IEEE ASP-DAC 2013 University Design Contest Best Design Award, and the IEEE ICECS 2016 Best Paper Award. His work for IEEE includes, serving as a member for the Design and Implementation of Signal Processing Systems (DISPS) Technical Committee, IEEE Signal Processing Society; and as a Technical Program Committee Member for several conferences, such as IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE Global Conference on Signal and Information Processing (GlobalSIP), IEEE Workshop on Signal Processing Systems (SiPS), IEEE Custom Integrated Circuits Conference (CICC), and IEEE Symposium on Low-Power and High-Speed Chips (COOL Chips). He has also been involved in organizing committees for conferences, including IEEE Asian Solid-State Circuits Conference (A-SSCC) and ACM/IEEE Asia and South Pacific Design Automation Conference (ASP-DAC). He serves as an Associate Editor for journals, such as *Journal of Signal Processing Systems* (Springer), *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, *IEICE Transactions on Electronics*, and *IPSJ Transactions on System LSI Design Methodology* (TSLDM).