The University of Osaka
Institutional Knowledge Archive

|  |  |
|---|---|
| Title | A Study on an IoT Sensing Platform for Multimodal Collaboration Analysis |
| Author(s) | 山口, 隼平 |
| Citation | 大阪大学, 2025, 博士論文 |
| Version Type | VoR |
| URL | https://doi.org/10.18910/101766 |
| rights |  |
| Note |  |

*The University of Osaka Institutional Knowledge Archive : OUKA*

https://ir.library.osaka-u.ac.jp/

The University of Osaka

# A Study on an IoT Sensing Platform for Multimodal Collaboration Analysis

Submitted to
Graduate School of Information Science and Technology
Osaka University

January 2025

Shunpei YAMAGUCHI

# List of Publications

## Related Journal Articles

1. Shunpei Yamaguchi, Motoki Nagano, Ritsuko Oshima, Jun Oshima, Takuya Fujihashi, Shunsuke Saruwatari, and Takashi Watanabe, “Multi-Speaker Identification with IoT Badges for Collaborative Learning Analysis,” *Journal of Information Processing*, vol. 31, pp. 375–386, 2023.

2. Shunpei Yamaguchi, Motoki Nagano, Shunpei Ohira, Ritsuko Oshima, Jun Oshima, Takuya Fujihashi, Shunsuke Saruwatari, and Takashi Watanabe, “Web Services for Collaboration Analysis With IoT Badges,” *IEEE Access*, vol. 10, pp. 121318–121328, 2022.

3. Shunpei Yamaguchi, Shusuke Ohtawa, Ritsuko Oshima, Jun Oshima, Takuya Fujihashi, Shunsuke Saruwatari, and Takashi Watanabe, “An IoT System with Business Card-Type Sensors for Collaborative Learning Analysis,” *Journal of Information Processing*, vol. 30, pp. 238–249, 2022.

4. Shunpei Yamaguchi, Ritsuko Oshima, Jun Oshima, Ryota Shiina, Takuya Fujihashi, Shunsuke Saruwatari, and Takashi Watanabe, “Speaker Identification for Business-Card-Type Sensors,” *IEEE Open Journal of the Computer Society*, vol. 2, pp. 216–226, 2021.

## Related Conference Papers

1. Shunpei Yamaguchi, Aditya Arun, Takuya Fujiwara, Misaki Sakuta, Ryotaro Hada, Takuya Fujihashi, Takashi Watanabe, Dinesh Bharadia, and Shunsuke Saruwatari, “Experience: Practical Challenges for Indoor AR Applications,” *The 30th Annual International Conference on Mobile Computing and Networking (ACM MobiCom ’24)*, pp. 1030–1044, 2024.

2. Jun Lu, Shunpei Yamaguchi, Jun Oshima, Shunsuke Saruwatari, and Ritsuko Oshima, “Multimodality in Transactive Discourse: Integration of MmLA and SSNA,” *International Conference on Quantitative Ethnography (Poster)*, pp. 1–4, 2023.

3. Daichi Yamaguchi, Shunpei Yamaguchi, Ritsuko Oshima, Jun Oshima, Takuya Fujihashi,

Shunsuke Saruwatari, and Takashi Watanabe, “A Web Application with Business Card-Type Sensors for Collaborative Learning Analysis,” *IEEE Global Conference on Consumer Electronics*, pp. 748–749, 2021.

4. Shunpei Yamaguchi, Shusuke Ohtawa, Ritsuko Oshima, Jun Oshima, Takuya Fujihashi, Shunsuke Saruwatari, and Takashi Watanabe, “Collaborative Learning Analysis Using Business Card-type Sensors,” *International Conference on Quantitative Ethnography*, pp. 319–333, 2021.

5. Shunpei Yamaguchi, Ritsuko Oshima, Jun Oshima, Takuya Fujihashi, Shunsuke Saruwatari, and Takashi Watanabe, “A Preliminary Study on Speaker Identification Using Business Card-Type Sensors,” *IEEE International Conference on Consumer Electronics*, pp. 1–3, 2021.

## Other Conference Papers

1. Zaibei Li, Shunpei Yamaguchi, and Daniel Spikol, “OpenMMLA: an IoT-based Multimodal Data Collection Toolkit for Learning Analytics,” *The 15th International Conference on Learning Analytics & Knowledge (LAK25)*, pp. 1–10, 2025 (accepted).

2. Shunpei Yamaguchi, “An IoT-based Multimedia System for Fine-grained Multimodal Learning Analytics,” *International Conference on Quantitative Ethnography (Doctoral Consortium)*, pp. 1–5, 2023.

3. Takuya Fujiwara, Shunpei Yamaguchi, Takumasa Ishioka, Ritsuko Oshima, Jun Oshima, Kazuhiro Kizaki, Takuya Fujihashi, Shunsuke Saruwatari, and Takashi Watanabe, “An IoT System for Collaboration Analytics in Hybrid Learning Environments,” *IEEE International Conference on Advanced Learning Technologies*, pp. 132–133, 2023.

4. Jun Oshima, Jun Lu, Ritsuko Oshima, Shunpei Yamaguchi, and Shunsuke Saruwatari, “Multimodal Analytics of Transactive Discourse,” *International Society of the Learning Sciences Annual Meeting (Pre-Conference Workshops & Tutorials)*, pp. 1–1, 2023.

5. Jun Lu, Shunpei Yamaguchi, Jun Oshima, Ritsuko Oshima, and Shunsuke Saruwatari, “Utterance Pattern Extraction during Idea Improvement Using Sound Pressure Modality,” *Poster presented at New Member’s Session in ISLS Annual Meeting*, pp. 1–1, 2022.

# Abstract

Collaboration plays a crucial role in accomplishing tasks and exerts significant influence across various domains, including workplaces and educational settings. In the field of learning sciences in particular, the mechanisms of collaboration have been extensively studied, providing valuable insights for boosting intellectual productivity through analyses of collaboration among learners. However, traditional qualitative analyses require considerable human and time resources, which makes it difficult to conduct analyses on a large number of participants or groups, or over extended periods.

Recent advancements in the Internet of Things (IoT) suggest a way to reduce the costs associated with qualitative analysis. By automatically providing quantitative data on collaboration to analysts via IoT systems, the labor and time costs of qualitative analysis can be reduced. This cost reduction expands the potential scope of collaboration analysis—extending to contexts previously out of reach—and helps generate new insights into the nature of collaboration.

To support qualitative collaboration analysis through IoT, three major requirements must be met. The first requirement is time synchronization among sensor devices. Since multiple devices may be deployed across many individuals and diverse environments, lack of time synchronization among devices will lead to data inconsistencies that hinder accurate collaboration analysis. The second requirement is multimodal extraction of collaboration. In qualitative analysis, multiple modalities—such as video and audio—are examined, so IoT systems likewise need to extract data from multiple modalities. In particular, it is necessary to quantitatively capture key factors such as face-to-face interactions among learners, learning phases, speakers, activity, and postures. The third requirement is a system design that prioritizes usability. Researchers and practitioners who analyze collaboration are not always Information Technology (IT) experts, so it is crucial to provide a system that is easy to operate, even for non-technical users.

In response to these requirements, this study proposes an IoT sensing platform for quantitative collaboration analysis. The proposed system comprises three components: a set of portable sensors that collect data with high-precision time synchronization, a suite of algorithms that extract collaboration data in a multimodal manner, and a web-based visualization tool that offers an intuitive interface for analysts. By adopting this system, it becomes possible to automate much

of the process that previously depended on manual labor, thereby substantially reducing the time and effort needed for collaboration analysis.

Chapter 2 describes the research on the architecture of the proposed system. Specifically, the study proposes and implement 1) a business-card-type sensor that collects each learner's data under precise time synchronization, 2) analytical algorithms that multimodally extract learners' face-to-face interactions, learning phases, speakers, and activity, and 3) a web application that visualizes the resulting data without requiring complex operations. Through both qualitative and quantitative evaluations, the study demonstrates that the proposed system meets its required specifications and supports qualitative analysis of collaboration.

Chapter 3 investigates a method for accurately identifying speakers—a crucial modality within the proposed system. Specifically, the study realizes high-accuracy speaker identification for collaboration analysis by using a sound pressure sensor equipped with a peak-hold circuit, achieving high-precision time synchronization among sensors, and employing an algorithm that reliably identifies speakers from noisy sound pressure data. Evaluation experiments show that the proposed method remains robust for varying numbers of participants, different types of noise, and diverse speaking durations.

Chapter 4 focuses on localization used for estimating posture, another modality in the proposed system. The study clarifies the issues and solutions involved in applying a simple yet highly accurate vision-based localization to real-world collaboration analysis. Through large-scale case studies and controlled experiments, the study identifies practical challenges in applying the vision-based localization to posture estimation. To address these challenges, this study proposes a prototype solution that integrates Ultra Wide Band (UWB) with visual data. Through the evaluation experiments, this study demonstrates the robustness of the prototype solution for localization.

Overall, this study suggests the potential to reduce the human and time costs required for traditional qualitative analysis while also expanding the scope of collaboration analysis. By enabling approaches in domains where collaboration analysis has previously been limited, this work is expected to further advance our understanding of collaboration.

# Contents

# List of Figures

# List of Tables

# List of Algorithms

# Chapter 1

# Introduction

## 1.1 Background

Human interaction, in which individuals influence one another through complex social relationships, plays a pivotal role in shaping the dynamics and outcomes of collaborative tasks carried out by multiple people. The nature and quality of these interactions can significantly affect the performance of teams, making their study and understanding critical in diverse settings such as education and the workplace. By examining these interactions in detail, we can uncover underlying patterns and develop strategies to foster more effective collaboration and enhance overall team performance.

Over the years, researchers have dedicated considerable effort to qualitatively analyzing human interaction to uncover strategies for enhancing productivity and improving collaborative outcomes. Collaboration analysis has traditionally been conducted using the ethnographic approach [1]. Ethnography, which originated in cultural anthropology, is a method for studying human behavior and social interactions. Specifically, this involves the collection of field notes, often supplemented by audio and visual recordings to provide detailed context. Ethnographic methods have been widely applied across various disciplines to analyze human behaviors in different settings. Similarly, in collaborative environments, ethnographic techniques have been the predominant approach for examining group interactions and dynamics.

In educational contexts, learning sciences have extensively examined the mechanisms of collaborative learning — a field that has garnered increasing attention for its potential to foster deeper understanding and engagement among learners. Researchers have investigated interaction patterns, communication strategies, and collaborative behaviors that emerge during group learning activities. Some studies in [2–5] have illuminated various interaction patterns, shedding light on how learners construct knowledge together. For example, the work presented in [4] uniquely combined social network analysis with in-depth dialogical analysis to study collegiate discourse recordings of collaborative reading activities. This study not only identified shared awareness patterns within the group but also highlighted the diverse contributions of individual students, thereby revealing nuanced dynamics of group learning. Another notable example is the study conducted by Chen and colleagues [5], which employed a randomized controlled trial alongside case studies to evaluate the impact of a year-long video-based professional development program utilizing the Classroom Discourse Analyzer. The findings demonstrated that such video-based programs could significantly enhance both classroom discourse and student learning. Moreover, this research provided valuable insights for designing effective visualization tools to enrich the professional development experience, enabling educators to better analyze and improve their teaching strategies.

Despite these advances, qualitative analysis of human interaction comes with significant challenges, particularly in terms of the time and effort required. Most of these analyses involve manual examination of large datasets, making it a labor-intensive process. This limitation becomes especially pronounced when dealing with collaborative tasks involving a large number of participants, where the sheer volume of data can render qualitative methods impractical. The bottleneck created by manual analysis not only hinders scalability but also limits the application of these methods to smaller, more controlled scenarios. Consequently, there is a growing need for innovative approaches and tools that can complement qualitative analysis, enabling researchers and practitioners to handle larger datasets more efficiently and effectively while still capturing the rich complexity of human interaction.

## 1.2 Internet of Things for Collaboration Analysis

To support the existing qualitative collaboration analysis, Quantitative Ethnography (QE) was proposed in 2017 by David Shaffer at the University of Wisconsin, a methodology that combines quantitative and qualitative analysis to overcome the limitations of traditional ethnographic approaches. This method aims to guide qualitative analysis by leveraging quantitative techniques to narrow the focus to specific, high-potential areas, thereby reducing the overall costs typically associated with ethnographic studies while enhancing analytical precision. Specifically, QE employs data mining and natural language processing to extract structured data from dialogue texts. These structured data are then interpreted through ethnographic frameworks, allowing for a seamless integration of computational analysis and qualitative contextual understanding.

Building on these advancements in QE, IoT-based collaboration analysis has gained increasing attention as a means to deepen the understanding of collaboration. IoT technologies enable multimodal data collection through wearable devices and environmental sensors, automating the observation of collaboration scenarios traditionally conducted qualitatively. By integrating IoT-based approaches, QE reduces the costs associated with narrowing the focus for qualitative analysis and, in turn, facilitates the discovery of deeper insights into group and individual dynamics across various contexts. This fusion of data-driven methodologies and ethnographic interpretations paves the way for innovative strategies to analyze and enhance human interactions.

To support such analyses in fact, IoT systems often employ compact, portable sensors like business card-type devices that integrate seamlessly into the daily activities of individuals and teams. These sensors have been widely used in studies to collect data on communication patterns, interactions, and activity levels — key metrics for understanding collaboration dynamics. Examples include Hitachi's Business Microscope [6,7], MIT's Sociometric Badge [8], and devices such as Open Badges and Rhythm, developed by Lederman et al. [9, 10].

MIT Media Lab initially developed the Sociometric Badge [8], a sociometric wearable device (SWD) for quantitatively investigating human behavior and interactions in collaborative environments. Subsequent advancements led to Open Badge [9], focusing on miniaturization, and Rhythm [10], which supports both on-site and online collaboration analysis. These devices monitor interactions using sound pressure and radio frequency (RF) signals for voice recognition and proximity detection. Similarly, Hitachi's Business Microscope [6] uses business card-type sensors to monitor workplace interactions and employee behaviors, while the Sensor-based Regulation Profiler (SRP) incorporates precise synchronization RF modules for fine-grained collaboration analysis. Additionally, MBox offers a low-cost, easy-to-use platform designed iteratively based on learning theories to investigate collaborative learning in diverse group work contexts.

## 1.3 System Requirement for Multimodal Collaboration Analysis

**Synchronization accuracy for collaborative sensor data**: Synchronization between devices is essential for accurately extracting data related to interpersonal collaboration. The study focuses on analyzing collaboration with mobile devices deployed across various targets, such as participants and environments. As described in Sec. 1.2 existing systems like Hitachi's Business Microscope, MIT's Sociometric Badge, Open Badges, and Rhythm, have enabled quantitative analysis of social interactions. However, a key limitation in these studies lies in their inability to achieve precise synchronization across multiple devices. These synchronization errors result in misleading analyses of collaborative activities. To ensure meaningful results, synchronization accuracy must be at least one-tenth of the sensor's maximum sampling rate. For example, sensors with a maximum sampling rate of 100 Hz require synchronization precision within 1 millisecond or less.

**Multimodal data extraction for qualitative analysis**: Extracting multimodal data is essential for supporting qualitative analysis and achieving a comprehensive understanding of collaboration. Quantitative methods provide objective metrics, but they cannot fully replicate the nuanced interpretations derived from qualitative approaches. To bridge this gap, the system identifies key multimodal data points, such as face-to-face interactions, learning phases, speakers, activity, and posture [11–13]. By focusing on these dimensions, the system enhances qualitative analysis, reducing manual observation costs while improving analytical accuracy.

**User-friendly system design for non-technical analysts**: Supporting analysts without advanced technical expertise is essential for the widespread adoption of collaborative analysis systems. Analysts often lack the technical knowledge required to install and operate complex software systems, which creates a barrier to their effective utilization. To address this, the system must prioritize ease of installation and usability, enabling analysts to focus on their work without struggling with technical difficulties. For example, implementing a web-based application allows for seamless access and quick deployment, eliminating the need for time-consuming setup processes. Such an intuitive interface ensures that non-technical users can efficiently perform quantitative analysis and integrate it into their workflow for collaborative analysis.

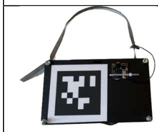
## 1.4 Related Work

Table 1.1 shows related studies on an IoT system for collaboration analysis. MIT launched the initial study called Sociometric Badge [8], which enables the measurement of collaboration by sensing individual activities and interactions through wearable sensors. Hitachi commercialized Sociometric Badge as Business Microscope [7], integrating additional features to analyze intra-organizational communication quantitatively. The company thus initiated the application of collaboration analysis in organizational contexts. MIT further expanded the capabilities of Sociometric Badge and developed Rhythm [10] in 2018, aiming to provide deeper insights into team dynamics in offline, online, and hybrid environment. In the context of learning analytics, MBoX [14] was developed as an IoT system specifically designed to support multimodal learning analysis by capturing and analyzing various learning behaviors.

However, these systems do not meet three requirements described in Sec. 1.3, leaving gaps in their ability to fully address collaboration analysis for collaborative learning. In the first requirement of time synchronization, they do not meet the precision targeted in this study. These analyses are useful for capturing long-term collaboration trends but are not applicable to fine-grained analyses that detect second-by-second changes. Sec. 2.5.1 discussed details about the precision requirements and synchronization accuracy of each method.

As for the second requirement of multimodal data extraction for collaboration analysis, none of these systems focus on all five modalities required for collaborative learning analysis. Sociometric Badge emphasizes activity levels, speech features, location, proximity, and face-to-face interaction, while Rhythm primarily targets speaker turn-taking, conversation time, and proximity. Business Microscope is designed to analyze face-to-face interactions and concentration levels. MBox captures

Table 1.1: Related work

| Image | Scheme | Abstract | Req. 1 | Req. 2 | Req. 3 |
|---|---|---|---|---|---|
| | Sociometric Badge [8] | Sensing user activity with business-card-type sensors | Unsatisfied | Partially satisfied | Unsatisfied |
| | Rhythm [10] | Sensing user activity with business-card-type sensors | Unsatisfied | Partially satisfied | Unsatisfied |
| | Business Microscope [7] | Quantitative analysis of intra-organizational communication | Unsatisfied | Partially satisfied | Unsatisfied |
|  | MBoX [14] | IoT system to support multimodal learning analysis | Unsatisfied | Partially satisfied | Unsatisfied |

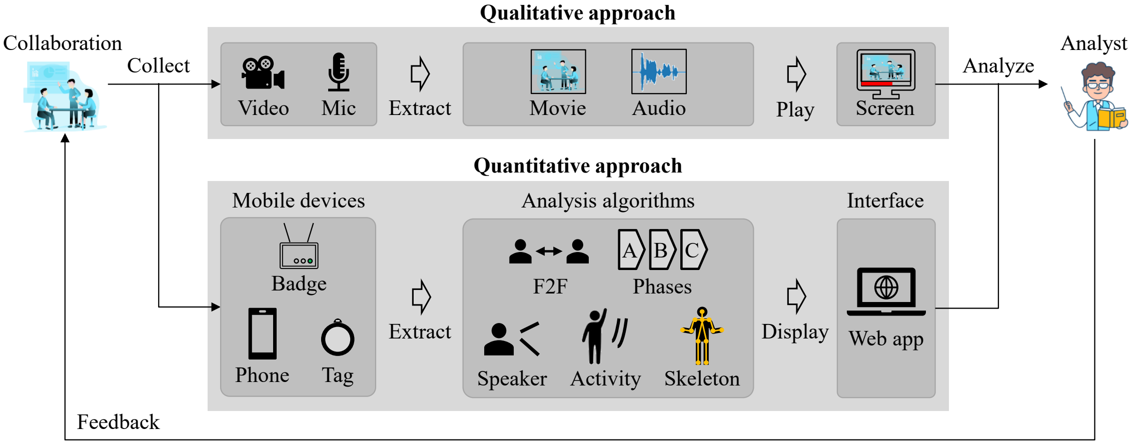


Figure 1.1: The concept of the IoT system for collaboration analysis.

face-to-face and speech between learners.

In the third requirement of usability, none of these systems support easy operation and installation of the system. Each method requires software installation and command-line operations. In the case of Business Microscope, collaboration analysis depends on outsourcing, making it difficult to consider it a highly user-friendly system.

## 1.5 Outline

Based on the requirements described in Sec. 1.3, this study focuses on an IoT platform for multimodal collaboration analysis. Figure 1.1 shows a structure of the IoT system.

The IoT system is composed of three major parts: data collection with mobile devices, data interpretation with multimodal analysis algorithms, and data visualization with a web-based application. The entire process of collaboration analysis with the IoT system is shown below.

1. The analyst installs video cameras and voice recorders in the collaboration environment.

2. Each user mounts a mobile device.

3. The users starts collaborative activity.

4. The analyst collects video and audio data from the video cameras and voice recorders.

5. The analyst collects sensor data from the devices worn by the users.

6. The proposed system quantitatively extracts key points for collaboration analysis from the data.

7. The system visualizes the key information for analysts on a web browser with the web application.

8. The analyst starts qualitative analysis using the relevant parts of the video and audio.

The collected sensor data includes a variety of modalities from mobile devices such as badges, smartphones, and smart tags. These modalities provide rich insights into the interactions and dynamics during collaboration. The collected data is sent to a central repository for further processing.

The next step involves interpreting the multimodal data using advanced algorithms designed to extract meaningful points for qualitative collaboration analysis. These algorithms identify patterns and relationships within the data to provide wide insights into the collaboration process. The system extracts key points for collaboration analysis, such as face-to-face interaction, learning phases, speakers, activity, and posture.

Once the analysis is complete, the results are visualized through an intuitive web-based application. This application offers an interactive platform where collaboration analysts can explore the data in detail. The visualizations provide a wide and detailed view of the collaboration with multimodal information. The web application also allows analysts to filter and compare data across different sessions, enabling them to identify trends and areas for improvement.

Analysts can conduct qualitative assessments and propose actionable recommendations to enhance collaboration efficiency by examining the video and audio corresponding to the obtained key points. For example, they may suggest adjustments to team structures, reconfigure spatial layouts, or recommend communication strategies based on data-driven findings. This iterative process ensures that the IoT system continually contributes to refining collaborative practices.

The contributions of this study are as follows:

- This study contributes to improving the efficiency of identifying key aspects of collaboration analysis, enabling the low-cost extraction of analytical points even for large groups or extended activities. An IoT system was designed and implemented to support collaboration analysis based on specific system requirements. Experimental evaluations in collaborative learning scenarios demonstrated that the system significantly reduced the human effort and time required for collaboration analysis.

- This study also contributes to establishing practical demonstrations for collaboration analysis in learning scenarios. Using the developed IoT system, collaboration patterns were quantitatively identified, offering new insights into promoting collaborative behaviors. This approach serves as a catalyst for advancing collaboration analysis.

- In the IoT system, this study advances collaboration analysis on mobile devices through precise speaker identification. The proposed scheme addresses challenges such as spike mitigation in a sound pressure sensor, precise synchronization across the sensors, and noise reduction for the sensor data. This improvement enables more precise collaboration analysis and facilitates fine-grained insights.

- Finally, this study reveals practical challenges and solutions for motion capture in collaboration analysis. Vision-based localization, a mainstream approach for indoor positioning, is applicable for motion capture with smart tags. This study comprehensively identifies practical limitations of the current vision-based localization for collaboration analysis. In addition, it proposes a novel solution that integrates vision-based and radio-based localization, presenting a robust and effective modality for IoT systems to enhance practical motion capture.

Chapter 2 focuses on the whole design and implementation of the IoT system with business-card-type mobile devices for collaboration analysis. Chapter 3 delves into precise speaker identification for such mobile devices. Finally, Chapter 4 finds that vision-based localization, which potentially contributes to posture recognition in the IoT system, has practical challenges to apply for collaboration analysis. In addition, the chapter proposes an prototype solution with radio frequency for robust localization.

# Chapter 2

# An IoT System for Collaboration Analysis

## 2.1 Introduction

Collaboration fosters our human ability to address complex problems in partnership with fellows. Many fields, such as workplace and education etc., adopt collaboration to their environment to exceed our personal ability. In the field of learning science, for example, collaborative learning has been featured as a learning method for future education. Collaborative learning promotes the learner's ability to solve complex problems through collaboration between learners.

To further enhance collaboration, the field of cognitive science has analyzed the patterns of collaboration types and their effectiveness. Especially, the field has explored specific patterns which promote our collaboration. However, the previous research often relies on substantial time to identify such patterns during collaborative activities. The process takes much time due to manual collaboration analysis with recorded videos and transcribed audio to evaluate the activities. This qualitative approach hinders collaboration analysts from applying the method in collaboration environments with a large number of users or real-time feedback.

To address the issue of time cost, Internet of Things (IoT) system has a potential to improve the efficiency of collaboration analysis. The system collects data from users and environment in collaboration with sensing devices. The system extracts key points for collaboration analysis from the acquired sensor data. Based on the extracted information, a facilitator analyzes collaboration and finally gives users feedback.

To develop an IoT system for collaboration analysis, there are three system requirements.

**Time synchronization across sensing devices**: The devices should precisely synchronize each other for fine-grained collaboration analysis. Collaboration analysis includes various range of duration from second-scale to month-scale. To adapt the second-scale collaboration analysis, the sensing devices should keep the consistency across sensor data. In detail, the synchronization error should be less than one tenth of the sampling rate.

**Multimodal data extraction**: Multimodal data should be extracted in the IoT system for collaboration analysis. Including the field of learning science, multimodal data from users and environments are focused for collaboration analysis. This chapter poses four key modalities for collaboration analysis: face-to-face between users, learning phases in a group, speakers, and activity of each user.

**Accessibility and usability**: The requirement is necessary for any users who conduct collaboration analysis. Collaboration analysts do not necessarily have skills of information technologies. The IoT system should be easily accessed and utilized by the analysts. To ensure the accessibility, the system should be accessible on a web browser for any users. In addition, the system should be operable with graphical user interfaces (GUI).

Following the three requirements, this chapter proposes Sensor-based Regulation Profiler (SRP) Web Services to quantitatively analyze collaboration. The system automates the extraction and visualization of key aspects of collaboration, thus supporting researchers in conducting qualitative analysis more efficiently. The proposed system consists of business card-type sensors called SRP Badges, multimodal analysis algorithms called SRP Analysis, and a web-based visualization tool called SRP Web. SRP badges precisely collect sensor data from users in collaboration under radio frequency (RF)-based time synchronization across the badges. SRP Analysis multimodally extracts key points of collaboration from the acquired sensor data: face-to-face, learning phases, speakers, and activity. SRP Web finally visualizes extracted information on a web application for the users.

To evaluate the proposed IoT system, this chapter conducted both qualitative and quantitative evaluation of the system. To conduct qualitative evaluation, the system was evaluated in collaborative learning analysis. The evaluation reveals that the IoT system supports reduce the existing time costs for collaboration analysts. In addition, both qualitative and quantitative evaluation shows that the IoT system satisfies three system requirements.

The rest of this chapter is organized as follows. Section 2.2 describes the proposed IoT system. Sections 2.3 and 2.4 describe the qualitative and quantitative evaluation of the proposed system. Section 2.5 describes related works of the IoT system for collaboration analysis. Finally, Sec. 2.6 concludes this chapter.

## 2.2 Proposed Scheme

This chapter proposes an IoT system for collaboration analysis called SRP Web Services. Figure. 2.1 shows the workflow with SRP Web Services in collaboration environment. The system consists of three parts: SRP Badge to collect sensor data from users and environment under precise time synchronization, SRP Analysis to extract key points for collaboration analysis from the ac-

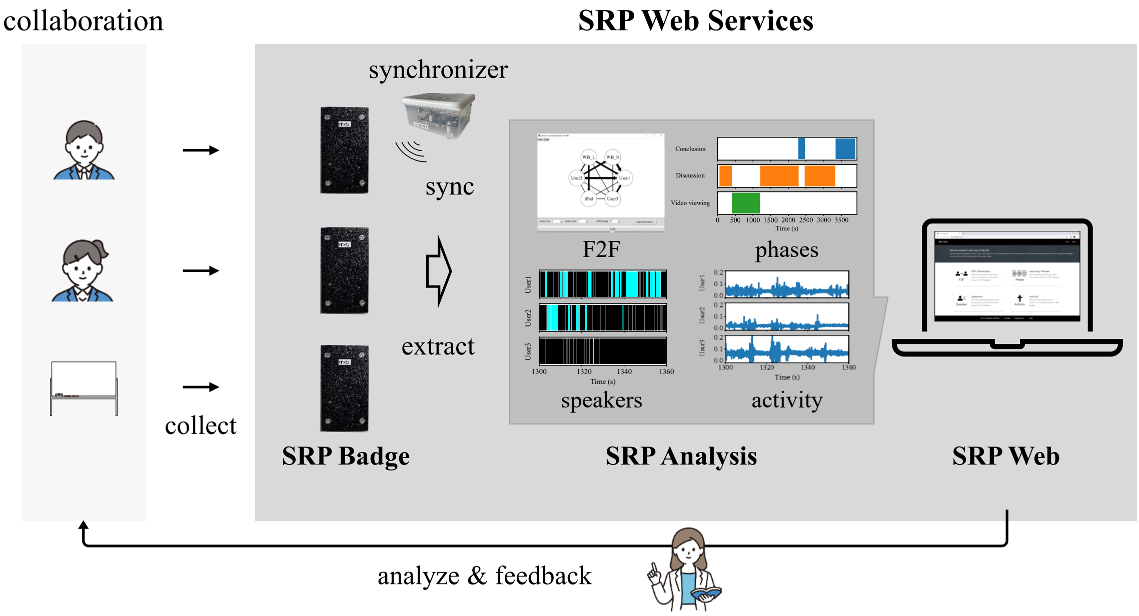


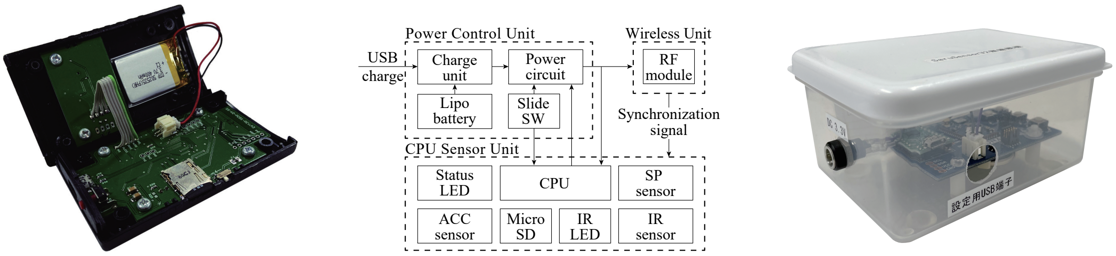
Figure 2.1: The overview of the proposed Sensor-based Regulation Profiler Web Services.

quired sensor data, and SRP Web to graphically provide the extracted information to collaboration analysts. The system is utilized along the procedure below.

1. Each user mounts SRP Badges on a chest
2. Start collaboration between the users
3. Collect all the badges from the users after collaboration
4. Extract sensor data from the badges with SRP Analysis
5. Visualize the extracted information on a web browser with SRP Web
6. Start qualitative analysis based on the extracted information by collaboration analysts

### 2.2.1 Sensor-based Regulation Profiler Badge

SRP Badge is a business-card-type sensor supposed to be worn on a user's chest. Figures 2.2 (a), (b), and (c) show the appearance, the block diagram, and the synchronizer of the SRP Badge. The badge is composed of three units: a power control unit, a CPU sensor unit, and a wireless unit.
**Power control unit**: The unit mounts a lithium-ion battery to run the badge. The battery supplies power to the power switch and Micro Controller Unit (MCU) in Fig. 2.2 (b). The badge can continuously run for 24 hours with the supplied power. The battery is also rechargeable via micro-USB adapter in the badge.
**CPU sensor unit**: The unit mounts STM32L476RGT6 from STMicroelectronics as the MCU, OSI5LAS1C1A infrared light emitting diode (LED) from OptoSupply, PIC79603 infrared receiver



(a) Appearance (b) Block diagram (c) Synchronizer

Figure 2.2: Sensor-based Regulation Profiler Badge.

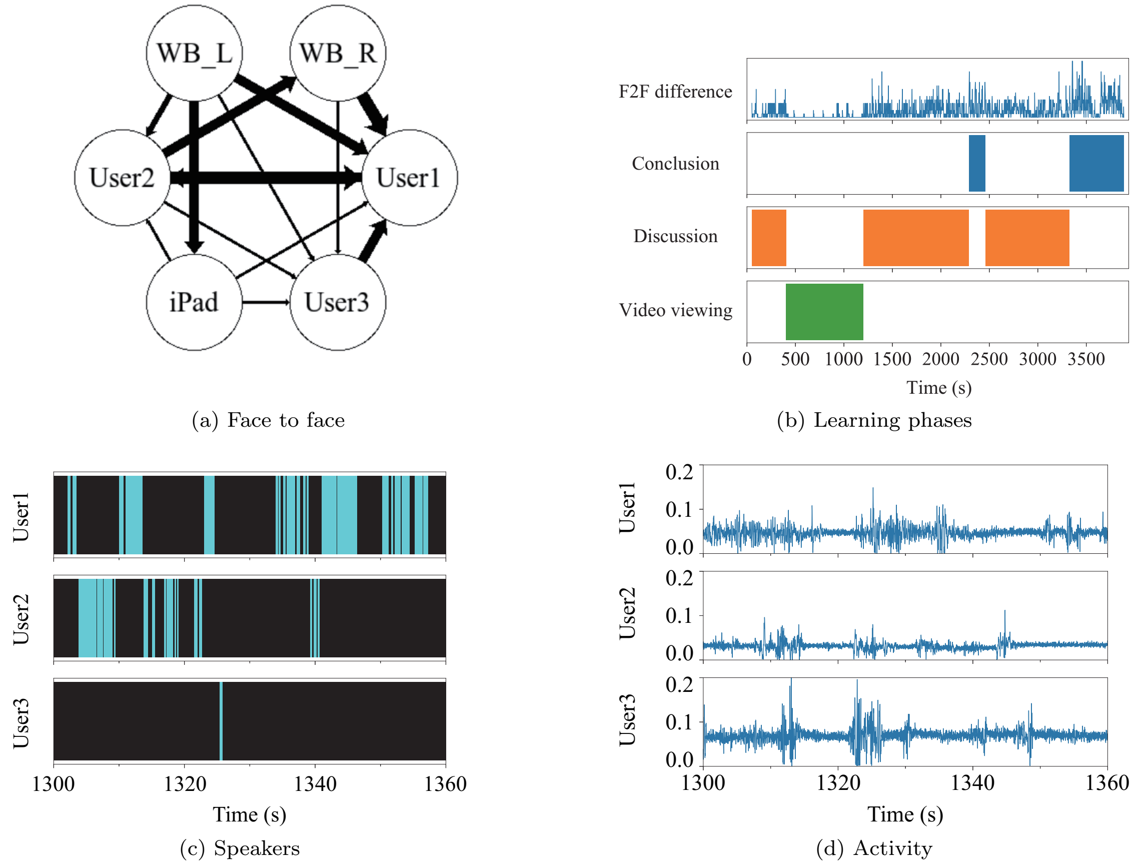
from KODENSHI CORP., INMP510 analog microphone from TDK, and ADXL362 accelerometer from ANALOG DEVICES. The MCU regulates sampling rates of data from each sensor: the infrared data at 12 bits and 34 Hz, the sound pressure data at 12 bits and 100 Hz, and the three-axes acceleration data at 100 Hz. The microSD card slot of DM3AT-SF-PEJM5 from Hirose Electric is equipped with the unit to record the sensor data.

**Wireless unit**: The unit mounts an RF module of CC2650 from Texas Instruments for wireless time synchronization across badges. The module sends synchronous packets every 10 seconds from other badges or its synchronizer. The CC2650 utilizes the protocol optimized for wireless synchronization across devices called UNISONet [15] to achieve precise synchronization across the badges. The synchronizer initially sends the synchronous packet for neighbor badges. Badges which receive the packet minimize and fix the processing time from reception to forwarding, enabling simultaneous reception of the same packet at neighboring nodes and triggering constructive interference [16]. By repeating reception and forward of synchronous packets across the devices, all badges in the environment keep the time consistency. Each badge can estimate the current time in the flooding-based system by combining the original timestamp from the synchronizer with the fixed delay per hop and the number of hops required for the packet to reach the badge.

### 2.2.2 Sensor-based Regulation Profiler Analysis

SRP Analysis consists of algorithms to extract key points for collaboration analysis with the acquired sensor data. Figures 2.3 (a), (b), (c), and (d) show the appearance of each algorithm for collaboration analysis. The algorithms extract face to face, learning phases, speakers, and activity from the sensor data.

**Face-to-face**: The algorithm extracts face-to-face across users based on the transmission and reception of the infrared data. Algorithm 1 and Table 2.1 show the procedure of the face-to-face graph extraction and its notation. The algorithm starts by initializing the face-to-face graph matrix $G$ with zeros, representing no initial interactions between any users. For each sensor $d$ in the set of sensors $U$, the algorithm collects the infrared data ($l_d$) that has been received within a specific time



(a) Face to face

(b) Learning phases

(c) Speakers

(d) Activity

Figure 2.3: Sensor-based Regulation Profiler Analysis.

window from $t_0$ to $t_0 + W$. The received data contains a list of sensor IDs $S$ that were detected by sensor $d$ during this period. For each detected sensor ID $s$ in the list $S$, the algorithm increments the corresponding element in the matrix $G[s][d]$. This increment represents the interaction between sensor $s$ and sensor $d$, essentially recording a face-to-face encounter. Finally, after processing all sensors and their received data, the face-to-face graph matrix $G$ is returned, summarizing all interactions between the sensors during the given time window.

**Learning phase**: The algorithm extracts learning phases in collaboration based on the time variation of face-to-face interaction across users. The time variation of the interaction is quantified from face-to-face graph matrix $G$ in the face-to-face algorithm. To reflect the context of face-to-face interaction, the algorithm creates social network matrices by applying sliding windows to the graph matrix $G$. The sliding windows consist of 3-seconds slide width and 60-seconds window size. To automatically classify learning phases, the algorithm adopts AutoPlait [17] to the social network matrices. AutoPlait quickly and automatically classifies similar patterns of the data based on hidden Markov models. The proposed algorithm finally classifies collaboration into several learning phases.

Table 2.1: Notation of the face-to-face graph extraction

| Variable / Function | Description |
|---|---|
| $U$ | Set of all the sensor IDs |
| $L$ | Set of the infrared data<br>obtained from all the sensors |
| $l_d$ | Infrared data of sensor $d$ |
| $t_0$ | Target time for social graph extraction |
| $G$ | Face-to-face graph matrix with the size of $\|U\| \times \|U\|$ |
| $W$ | Window size [s] |

**Algorithm 1** Face-to-face graph extraction

```
Require: $L$, $U$, $t_0$
Ensure: $G$
Insert zeros into all elements of $G$
for all $d \in U$ do
    $S \leftarrow$ all received IDs in $l_d \in L$ between $t_0$ to $t_0 + W$
    for all $s \in S$ do
        Increment $G[s][d]$
    end for
end for
return $G$
```

**Speaker**: The algorithm identifies a speaker in collaboration with sound pressure data acquired from SRP Badges. Figure 2.4 shows the overall process of the speaker identification algorithm. For the accurate speaker identification, there are deliberate three-step algorithm.

1) Pre-processing: In this step, the algorithm detects the rise of sound pressure for each user. The algorithm initially find the minimum sound pressure and subtract the entire sound pressure values with the minimum value for each user. Based on the above zero-correction, the algorithm distinguishes speech and non-speech in the data with sliding windows. Algorithm 2 exhibits the procedure to label speech in Figure 2.4, and Table 2.2 lists its notation. The algorithm outputs the array $\mathbb{A}$ called “the 1–0 data for each user” from the set of all sensor IDs $U$ and the set of the sound pressure data from all the sensors $\mathbb{S} = \{S_1, S_2, \ldots, S_{|U|}\}$.

2) Speech Section Estimation: This step estimates whether speech occurred in the collaboration from the 1–0 data for each user. The algorithm initially complements slight silence as speech and removes sudden pulse noise as non-speech on the 1–0 data for each user. Specifically, the algorithm regards the section between labels 0 within 90 ms as a part of consecutive speech. The algorithm also regards the section between labels 1 within 150 ms as a part of mis-detected speech duration due to noise. Finally, the algorithm extracts speech data where at least one user speaks or not called “the speech section data.”

3) Speaker Identification: In this step, the algorithm fuses the data: the 1–0 data for each user and the speech section data to extract a speaker in the collaboration. The algorithm focuses on

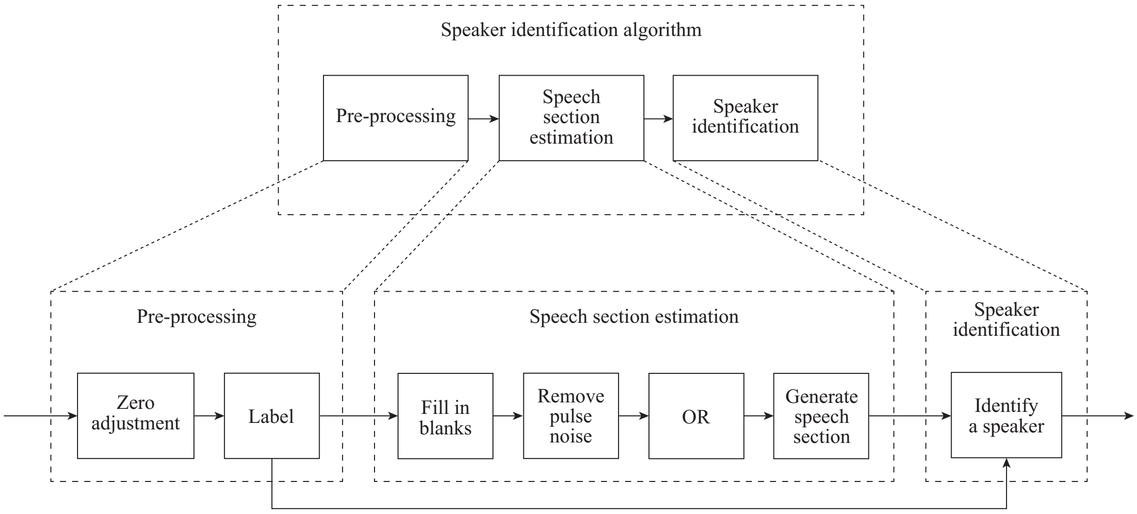


Figure 2.4: The procedure of speaker identification in SRP Analysis.

Table 2.2: Notation of labeling in pre-processing

| Variable / Function | Description |
|---|---|
| $U$ | Set of all sensor IDs |
| $d$ | Sensor ID |
| $\mathbb{S}$ | Set of the sound pressure data obtained from all the sensors |
| $S_d$ | Sound pressure data for sensor $d$ |
| $\mathbb{A}$ | Set of 1  arrays with speech labels |
| $A_d$ | 1 bit arrays with speech labels of sensor $d$ |
| $\xi$ | Top index of window |
| $D$ | Window size |
| $\eta_s$ | Speech threshold for all users |
| $\eta_m$ | Speech threshold based on maximum sound pressure in the window |
| $\max(X)$ | Calculate the maximum of all the elements of $X$ |

each speech section in the speech section data. The algorithm identifies a user whose labels 1 in the 1–0 data for each user are the most in the users as a speaker in the speech section.

**Activity**: The algorithm extracts activity of each user based on acceleration data acquired in each user's SRP Badge. The acceleration data is originally saved by three axes. The algorithm converts the acceleration to L2 norm to acquire the motion scale. The acceleration norm is converted to relative values from 0 to 1 for each user. The acquired data is used as a personal activity in collaboration.

### 2.2.3 Sensor-based Regulation Profiler Web

SRP Web enables collaboration analysts to easily analyze collaboration on a web browser. Figures 2.5 (a) and (b) show the architecture, the appearance, and the analysis view of the proposed

**Algorithm 2** Labeling in pre-processing

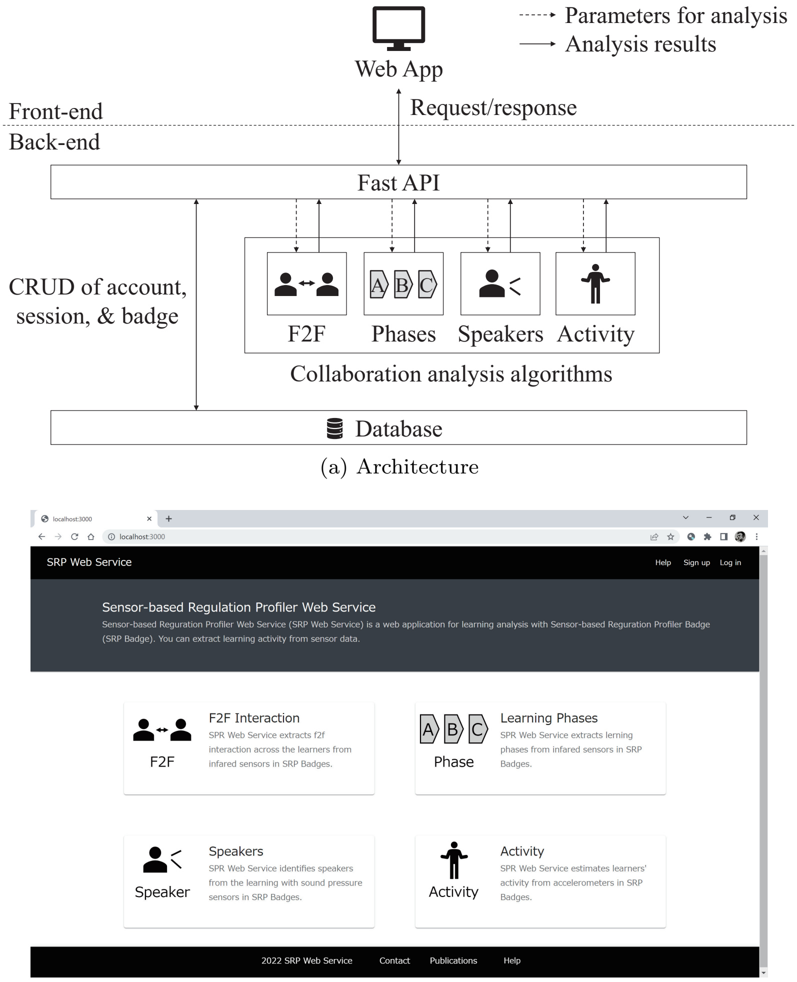
```
Require: $U, \mathbb{S}$
Ensure: $\mathbb{A}$
for all $d \in U$ do
    Insert zeros into all elements of $A_d$
    $\xi \Leftarrow 0$
    while $\xi <$ length of $A_d$ do
        $W \Leftarrow S_d \in \mathbb{S}$ between $\xi$ to $\xi + D$
        $m \Leftarrow \max(W)$
        if $m > \eta_s$ then
            $\eta_m \Leftarrow m * 0.1$
            if $w \in W > \eta_m$ then
                $w \Leftarrow 1$
            else
                $w \Leftarrow 0$
            end if
            Insert $w \in W$ into elements of $A_d$ with OR
        end if
        $\xi \Leftarrow \xi +$ slide width
    end while
    Insert $A_d$ into $\mathbb{A}$
end for
return $\mathbb{A}$
```

web application named Sensor-based Regulation Profiler Web (SRP Web). The application is composed of the front-end for the user interface and the back-end for data management. The front-end is structured with Next.js in the version of 12. The back-end is structured with FastAPI in the version of 0.72.0, SQLite, and Python 3.6. Requests from the user are sent to FastAPI in the back-end. FastAPI receives the requests and communicates with SRP Analysis described in 2.2.2 or the database. The requests include data operation such as creating, reading, updating, and deleting (CRUD) the user's information: accounts, projects, and sensor data acquired from SRP Badges described in 2.2.1. FastAPI then sends required data, including parameters such as the start and end time for collaboration analysis, to each collaboration analysis algorithm. The analyzed data in SRP Analysis are sent to FastAPI. FastAPI finally returns the response to the front-end and the user can start to analyze the collaboration on SRP Web.

SRP Web includes five main functions below.

**High accessibility**: The system is open for any collaboration analysts in terms of information technology skills thanks to a web platform for high accessibility. The analyst can easily access to the system on a web browser without any operation on command line interface (CLI) and software installation. For example, the analyst does not need to consider software versions, packages, or the operating system of the installation environment. Such users benefit from web services that operate solely with a web browser and an internet connection.

**Low performance dependence on end devices**: A web application depends little on computer



(a) Architecture

Sensor-based Regulation Profiler Web Service

Sensor-based Reguration Profiler Web Service (SRP Web Service) is a web application for learning analysis with Sensor-based Reguration Profiler Badge (SRP Badge). You can extract learning activity from sensor data.

F2F Interaction

SPR Web Service extracts f2f interaction across the learners from infared sensors in SRP Badges.

Learning Phases

SPR Web Service extracts lerning phases from infared sensors in SRP Badges.

Speakers

SPR Web Service identifies speakers from the learning with sound pressure sensors in SRP Badges.

Activity

SPR Web Service estimates learners' activity from accelerometers in SRP Badges.

(b) Appearance

Figure 2.5: Sensor-based Regulation Profiler Web.

performance, allowing any user to access the service on any device. Since the application runs on a server, users are not required to have a high-performance computer. Such low dependency on hardware allows any user to analyze collaboration with the system.

**Physical separation for easy maintenance**: The application is designed with separate front-end and back-end components for easy maintenance. System developers can independently manage the functions of each component. This separated structure enables developers to respond to user feedback and update the application instantly.

**Account management for multiple users' access**: Users can simultaneously and independently utilize the application with their personal accounts. The application requires users to register their own accounts and log in before using it. Multiple users can simultaneously analyze collaboration with the system, each separated by their individual accounts.

**Session management for multiprocessing of collaboration analysis**: The system enables

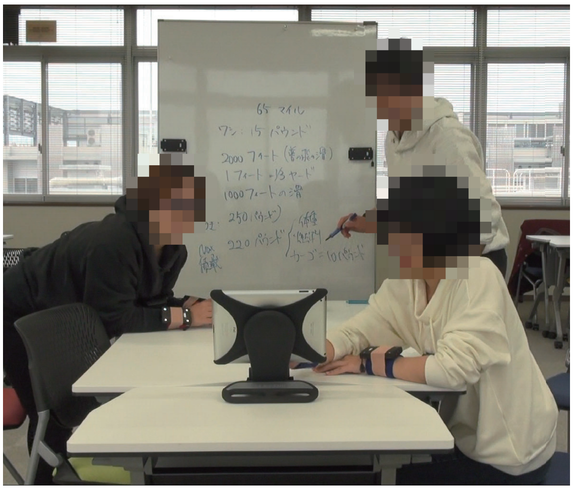


Figure 2.6: Experimental environment of collaborative learning.

each user to manage and analyze multiple session of collaboration in the account. The application provides sessions to hold sensor data corresponding to collaboration cases. The user can parallelly analyze multiple collaboration sessions in the same account.

## 2.3 Qualitative Evaluation

This section qualitatively evaluates the proposed system with the experiment of collaboration. As a collaboration, this study focuses on collaborative learning with three learners. Figure 2.6 shows a snapshot of the collaborative learning activity. The learning environment was composed of a table, chairs, an iPad to watch a video material, and a whiteboard to discuss. Collaborative learning was conducted five times in total and captured by SRP Badges. Each badge was attached to learners, the iPad, and the whiteboard. The learners wore the badge on their chest in case 1 and on their head in cases 2 to 5. Two badges were installed on both left and right sides of the whiteboard. To synchronize the badges, the synchronizer is installed on the table. The scenario of collaborative learning is composed of video viewing for 15 minutes, discussion for 30 minutes, and conclusion for 15 minutes based on a learning material for collaborative learning called the Adventures of Jasper Woodbury [18]. The learning material provides learners with interactive, narrative-based problem-solving challenges that integrate mathematical reasoning with real-world applications, encouraging critical thinking and collaboration. In the phases of video viewing and discussion, each learner can watch the learning material on the iPad.

### 2.3.1 Sensor Deployment

Automatic data collection using SRP Badges helped reduce costs in collaboration analysis compared to traditional methods. In conventional data collection, multiple video cameras were used to record users and the learning environment. This approach had the issue of high installation costs depending on the number of participants and the range of movement. In contrast, the proposed method improved scalability by adjusting the number of badges based on the number of users, reducing the overall installation costs for data collection devices.

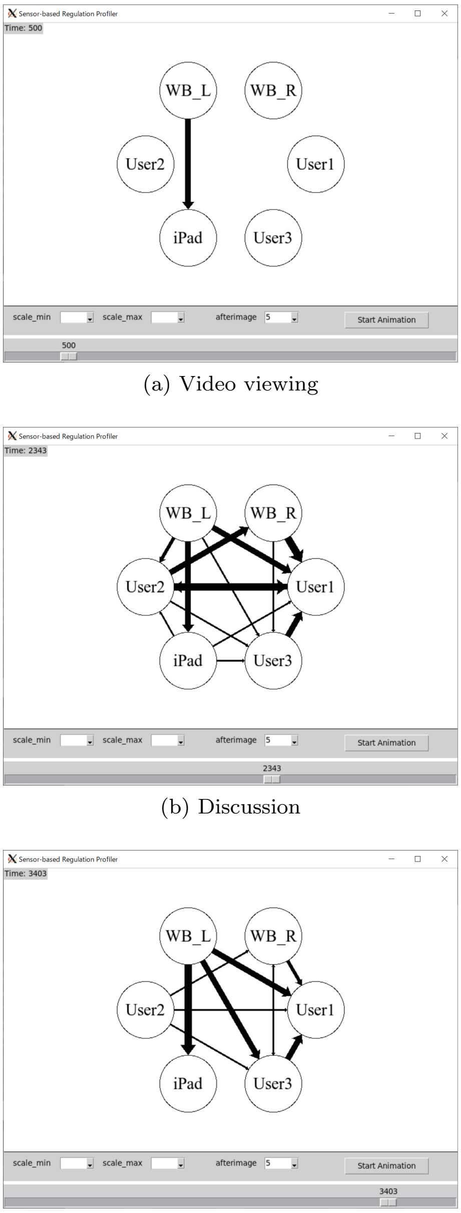
### 2.3.2 Face-to-Face Extraction

The proposed algorithm for face-to-face extraction sufficiently supported collaboration analysts, especially researchers in learning science, to reduce the cost of qualitative analysis for face-to-face detection in each experiment case. Figures 2.7 (a), (b), and (c) show the face-to-face relationship across learners in the phases of video viewing, discussion, and conclusion in the case 1 as an example. The horizontal axis in each figure shows the elapsed time [s] of collaborative learning for 60 minutes. Each figure shows face-to-face relationship across three users named User 1, User 2, and User 3, an iPad, and left and right sides of a whiteboard named WB_L and WB_R.

Figure 2.7 (a) shows face-to-face was scarce across the learners since they watched the learning material on the iPad. Figure 2.7 (b), the discussion phase, shows User 1, User 2, and the right side of the whiteboard faced. In addition, User 2 faced the left side of the whiteboard. Since the position of User 1 was closest to the right side of the whiteboard, User 1 used the whiteboard to leave clues to solve the problem. At the same time, User 2 saw User 1's writing. This interaction suggests that User 1 took on the role of leading the problem-solving effort, while User 2 acted as a collaborator by observing and interpreting User 1's input. Figure 2.7 (c), the conclusion phase, shows all the users faced the right side of the whiteboard. In addition, User 1 and User 2 faced each other. The figure indicates that User 1 wrote stuff to conclude the work and summarized the answer of the problem. User 2 and User 3 simultaneously saw User 1's writing. This interaction suggests that User 1 played a central role in synthesizing the group's ideas, while User 2 and User 3 acted as reviewers, validating and integrating the final solution. Including these results, the proposed algorithm enables efficient identification of key interactions, roles, and phases in collaborative learning, reducing the challenges of traditionally time-consuming qualitative video analysis.

### 2.3.3 Learning Phase Extraction

The algorithm for learning phase extraction supported collaboration analysts to reduce the cost of qualitative classification of learning phases. Figure 2.8 shows the result of learning phase extraction which the algorithm automatically output. The horizontal axis shows the elapsed time [s] of



Time: 500

afterimage 5

500

(a) Video viewing

Time: 2343

afterimage 5

2343

(b) Discussion

Time: 3403

afterimage 5

3403

(c) Conclusion

Figure 2.7: Extracted face-to-face relationship in each learning phase.

collaborative learning for 60 minutes. The top figure shows the normalized time variation of face-to-face difference across the learners extracted in the process of Sec. 2.3.2. The middle three figures show results automatically extracted by the proposed algorithm. The bottom figure shows the result of manual classification by researchers in learning science based on the recorded video of collaborative learning.

The results from the quantitative analysis indicate that: 1) learners rarely turned around during the video viewing phase due to their focus on the screen, 2) learners began to turn around more frequently during the discussion phase as they engaged in problem-solving conversations, and 3)

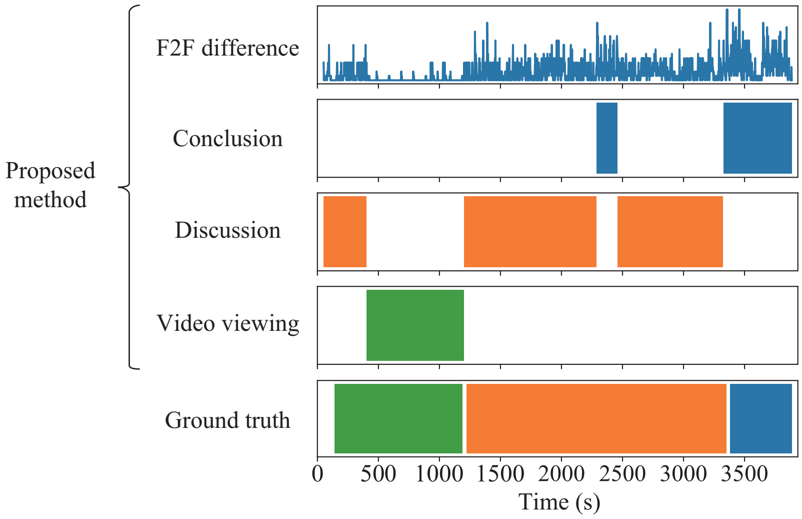


Figure 2.8: Automatic extraction results of the learning phases.

learners commonly turned around in the conclusion phase to finalize their solution. The transitions between these three phases were observed at 1,202 seconds and 3,326 seconds. In contrast, the qualitative analysis revealed transitions occurring between 1,173 seconds and 1,213 seconds, and between 3,335 seconds and 3,360 seconds. While there are some discrepancies between the two methods, specifically between 51 and 403 seconds, and between 2,289 and 2,459 seconds, the automatic extraction still reliably captured the transitions between the phases.

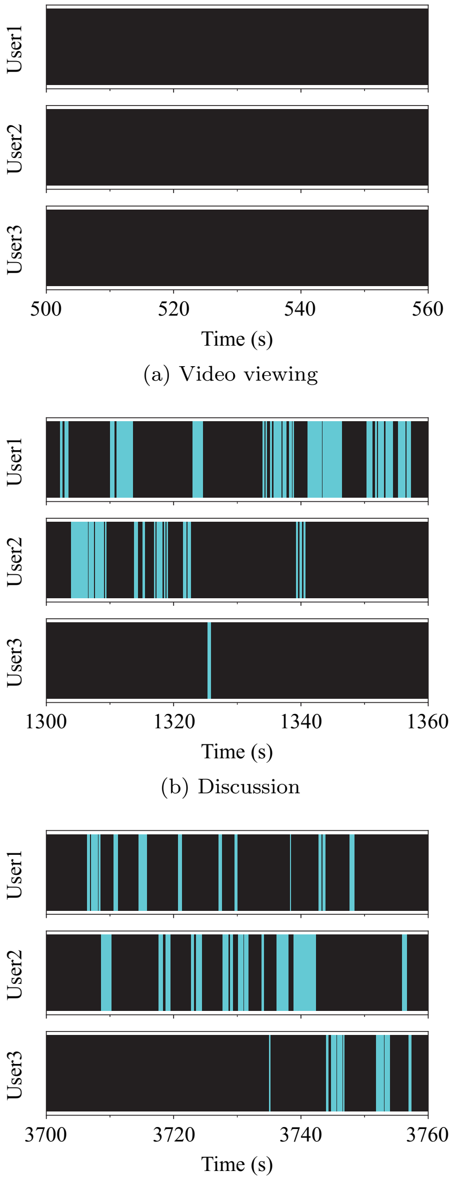
These quantitative results provide significant convenience to analysts by enabling quick and visual identification of phase transitions. For example, once the transition to the discussion phase is identified, analysts can focus on that period to efficiently investigate how problem-solving behaviors evolve. Similarly, pinpointing the start of the conclusion phase allows analysts to examine what aspects of the discussion prompted learners to begin synthesizing their ideas. By automatically highlighting clear phase transitions, the need to meticulously review the entire video is reduced, significantly saving time and effort in qualitative analysis.

### 2.3.4 Speaker Identification

The experimental evaluation shows the proposed algorithm for speaker identification supports collaboration analysis with automatic annotation of user's speech or non-speech. Figures 2.9 (a), (b), and (c) show the result of speaker identification in each learning phase in the case 1. The horizontal axis represents the elapsed time [s] and the blue bars indicate the speech of each user. For simplicity, we extracted 60 seconds of speaker identification results for each learning phase. To compare with the ground truth, the audio data was recorded and transcribed. Tables 2.3 (a) and (b) present the results of speech transcription in the same sections for 60 seconds shown in Figs. 2.9 (b) and (c). Figure 2.9 (a) shows an accurate detection of non-speech between 500 and 560 seconds during the video viewing phase, confirming that learners did not speak during this period. Figures 2.9 (b) and (c) identify the speech sections between 1,300 and 1,360 seconds in the discussion phase and between 3,700 and 3,760 seconds in the conclusion phase. In the discussion phase, the frequent alternation of speech turns between User1 and User2 suggests active engagement and collaborative exchange of ideas, a key indicator of productive group problem-solving. In the conclusion phase, the dense clustering of speech segments between User1 and User2 reflects their joint effort to synthesize and finalize the learning outcomes, while the minimal contributions from User3 suggest a more peripheral role in this stage. Qualitative analysis requires learning science researchers to repeatedly review the recorded video, noting the speech timing and identifying speakers, as shown in Tables 2.3 (a) and (b). In contrast, the proposed speaker identification method automates this process, significantly reducing the need for manual video review.

### 2.3.5 Activity Estimation

The proposed algorithm for activity supports collaboration analysis by reducing the cost of qualitative activity estimation. Figures 2.10 (a), (b), and (c) present the estimated activity results for each learner. The horizontal and vertical axes represent the elapsed time [s] and the relative acceleration. To compare with the ground truth, the video data was recorded and transcribed. Tables 2.4 (a) and (b) provide the qualitative records of the learners' activities corresponding to the sections in Figs. 2.10 (b) and (c). For consistency, we extracted 60 seconds from the same sections as the speaker identification for each learning phase. Figure 2.10 (a) accurately detects minimal activity between 500 and 560 seconds during the video viewing phase, indicating that the learners remained still while watching the video. Figures 2.10 (b) and (c) successfully capture specific activity between 1,300 and 1,360 seconds during the discussion phase, and between 3,700 and 3,760 seconds during the conclusion phase. In the discussion phase, the elevated activity levels visible in the graph allow analysts to quickly identify periods of heightened engagement, reducing the time needed to manually pinpoint moments of collaborative interaction for further analysis. In the conclusion phase, the sustained activity levels suggest a focus on finalizing the task, enabling analysts to concentrate on these periods to examine how learners consolidate their ideas and reach consensus, without manually reviewing less relevant sections. Qualitative analysis traditionally requires learning science researchers to replay the recorded video, closely observing the learners' activity as detailed in Tables 2.4 (a) and (b). The proposed activity estimation method automates this process, significantly reducing the need for manual observation by automatically extracting key behaviors.



(a) Video viewing

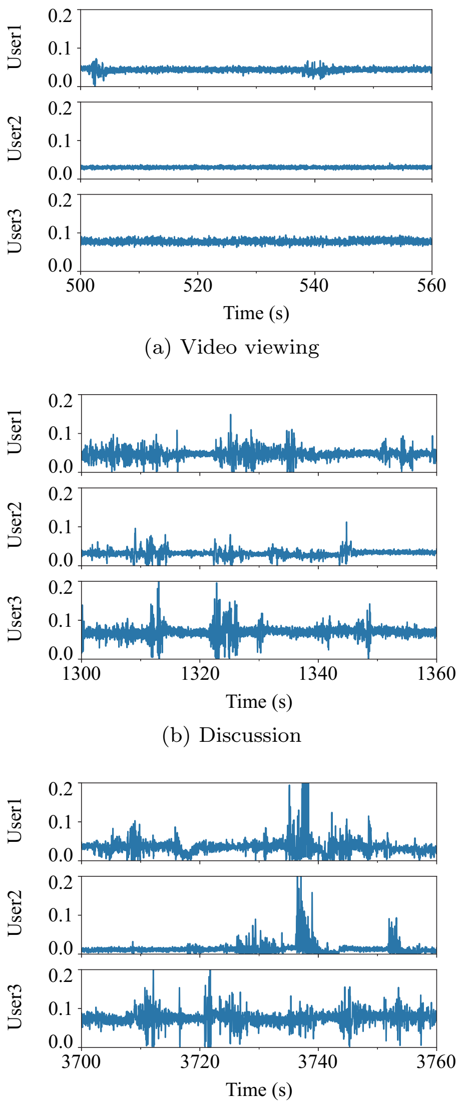
(b) Discussion

(c) Conclusion

Figure 2.9: Speaker identification results in each learning phase.

### 2.3.6 Web User Interface

The web application enables collaboration analysts to use the system without specialized operation for CLI operation and complex software installation. Figure 2.11 shows each step for collaboration analysis with SRP Web. The analyst just accessed to the webpage and operate the system on GUI. Concretely, the analysts initially registered their accounts or login with the accounts in Fig. 2.11 (a). The analysts created each project and prepared each session to summarize the acquired sensor data in Fig. 2.11 (b). Based on the sensor data saved in the session, the analysts executed each algorithm for collaboration analysis. Therefore, the proposed web application showed accessibility



(a) Video viewing

(b) Discussion

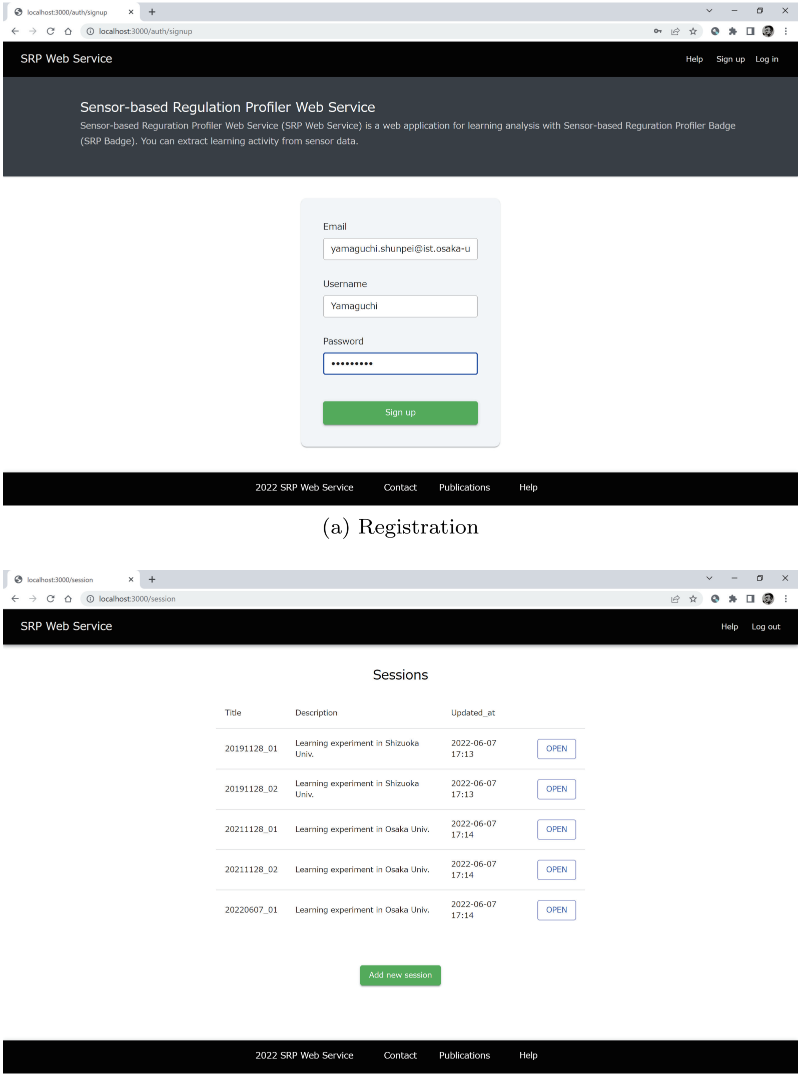
(c) Conclusion

Figure 2.10: Activity estimation results in each learning phase.

and usability for learning analysts who are unfamiliar with information technology.

### 2.3.7 Collaboration Elucidation

This section presents the application of the proposed IoT system in uncovering novel collaboration patterns through an integrated approach combining quantitative and qualitative analysis. The study integrated the proposed system with Socio-Semantic Network Analysis (SSNA), which quantitatively evaluates interactions, relationships, and communication patterns by analyzing the structure of social networks based on verbal information. SSNA analyzed word co-occurrence networks to calculate degree centralities, providing insights into group communication trajectories and individual transactive contributions. The proposed IoT system examined multimodal patterns to



SRP Web Service

Help Sign up Log in

Sensor-based Regulation Profiler Web Service

Sensor-based Reguration Profiler Web Service (SRP Web Service) is a web application for learning analysis with Sensor-based Reguration Profiler Badge (SRP Badge). You can extract learning activity from sensor data.

Email: yamaguchi.shunpei@ist.osaka-u

Username: Yamaguchi

Password: •••••••••

Sign up

2022 SRP Web Service Contact Publications Help

(a) Registration

SRP Web Service

Help Log out

Sessions

| Title | Description | Updated_at | |
|---|---|---|---|
| 20191128_01 | Learning experiment in Shizuoka Univ. | 2022-06-07 17:13 | OPEN |
| 20191128_02 | Learning experiment in Shizuoka Univ. | 2022-06-07 17:13 | OPEN |
| 20211128_01 | Learning experiment in Osaka Univ. | 2022-06-07 17:14 | OPEN |
| 20211128_02 | Learning experiment in Osaka Univ. | 2022-06-07 17:14 | OPEN |
| 20220607_01 | Learning experiment in Osaka Univ. | 2022-06-07 17:14 | OPEN |

Add new session

2022 SRP Web Service Contact Publications Help

(b) Session management

Figure 2.11: Procedure of collaboration analysis on SRP Web.

identify nonverbal cues that signaled transitions in collaboration phases. These combined analyses narrowed the focus to specific segments of video and audio data, enabling a targeted qualitative examination.

The analysis revealed that leadership transitions in transactive discourse were closely linked to distinct patterns of nonverbal behavior, including activity dynamics. Figure 2.12 shows the result of SSNA in a case of collaborative learning. The x-axis and y-axis show conversation turns and sum of the degree centralities calculated from word co-occurrence networks. By analyzing the trajectory of the group's idea development, two key segments were extracted: conversation turns 108–111 and 294–299. During the first segment, P2 and P3 played a prominent leadership role, while in the second segment, P1 and P2 took the lead. Figures 2.13 (a) and (b) present

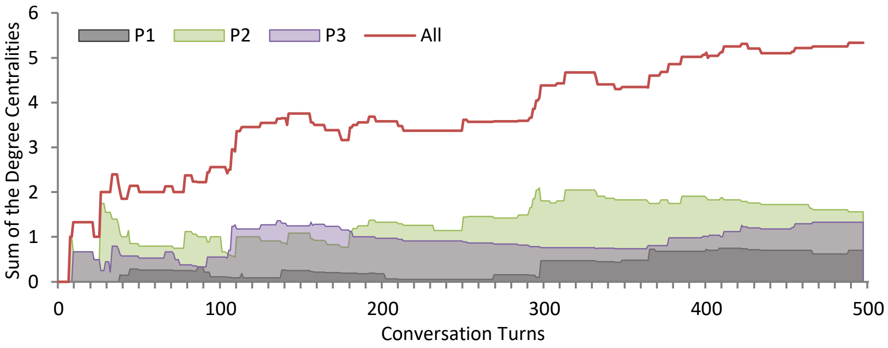
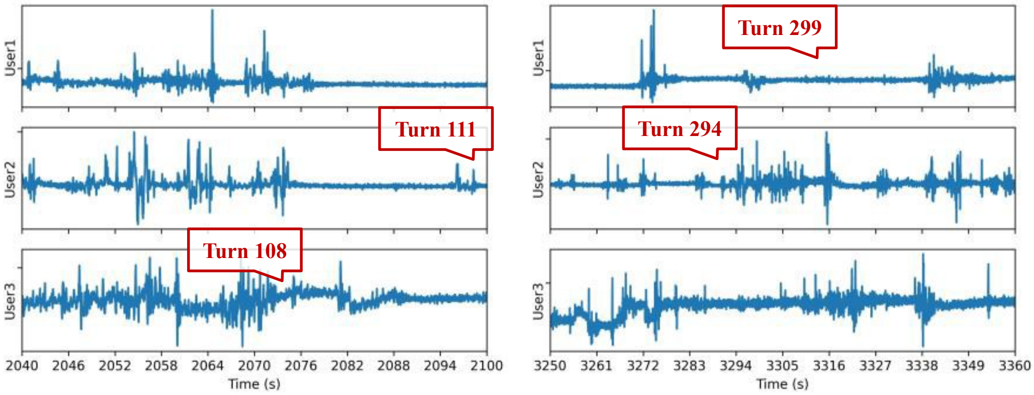


Figure 2.12: The trajectory of the group's idea improvement based on SSNA.



(a) Turn 108–111

(b) Turn 294–299

Figure 2.13: The activity dynamics during the idea improvement.

results extracted from the proposed system, emphasizing two segments that showcase partial insights into the extracted activity. The activity data shows consistent patterns through the entire collaborative process, although there were some partial variations. For instance, when considering the findings on leadership dynamics, Figures 2.13 demonstrate that the intensity of activity varies across problem-solving phases, while the relationships among participants remain consistent regardless of leadership changes. Finally, qualitative analysis in these two segments revealed that the first segment focused on gathering information, whereas the second segment centered on discussing solutions to the task. These findings indicate that leadership roles shifted among participants, adapting to the specific phases of problem-solving.

Traditional approaches to collaboration analysis relied on manual observation of video and audio data, requiring researchers to qualitatively identify key points for analysis. Quantitative approaches based on SSNA and the proposed system enable the identification of critical elements within the data, significantly reducing the analytical cost associated with its multimodal complexity. Focusing on these identified segments enabled efficient qualitative analysis, providing new insights into the

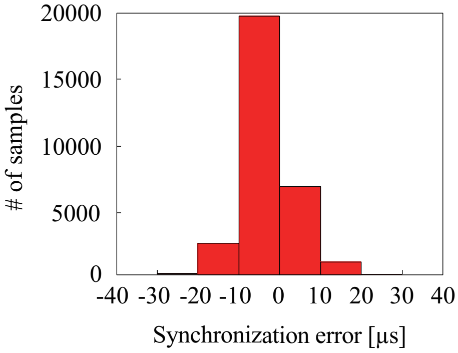


Figure 2.14: Time synchronization accuracy between SRP Badge and its synchronizer.

dynamics of collaboration.

## 2.4 Quantitative Evaluation

This section quantitatively evaluates the proposed system with the experiment.

### 2.4.1 Synchronization Accuracy of Sensors

We conducted an experimental assessment of the time synchronization accuracy between SRP Badge and its synchronizer. Each device was positioned close to each other on a desk and started synchronization. The time difference between the devices was measured based on the synchronization signals transmitted from the synchronizer. An oscilloscope was used to precisely measure the clock rise time at both devices to determine the time deviation accurately. In this setup, the synchronization error was calculated 30,003 times.

Figure 2.14 illustrates the synchronization accuracy between the devices. The horizontal and vertical axes represent the time deviation and the number of samples corresponding to each deviation. As shown in Fig. 2.14, the time synchronization error was confined to within ± 30 µs. The mean and maximum synchronization errors recorded were $-7.7$ µs and 30 µs, respectively. Given that both the sound pressure sensor and the acceleration sensor on SRP Badge operate at a sampling rate of 100 Hz, the synchronization error was allowable within the required threshold of less than 1 ms. In addition, wireless sensor networks generally hinder high-precision synchronization due to the complex interplay of factors such as changes in network topology, hardware resource constraints, and environmental factors [19]. Achieving such µs-level synchronization demonstrates precise synchronization within this wireless sensor network. The proposed synchronization method thus ensures accurate sensor data analysis for collaboration analysis.

### 2.4.2 Accuracy of Face-to-Face Extraction

The accuracy of face-to-face extraction in SRP Analysis was evaluated through an experiment using infrared sensors embedded in SRP Badges. The experiment took place in a room with dimensions of 10.6 m × 7.05 m × 2.65 m, equipped with multiple LED recessed ceiling lights. Three subjects wore SRP Badges on the chest and were positioned 1.50 m apart, with two of the three engaging in a face-to-face conversation for 60 s, while the non-speaking user faced the midpoint between the speakers. All combinations of speakers were tested, and the accuracy of face-to-face detection was calculated. Results showed that the infrared sensors detected face-to-face interactions with accuracies of 75.3 %, 78.0 %, and 78.0 % across the different speaker combinations. These findings suggest that the proposed face-to-face detection method is effective in supporting researchers in learning science by reducing the qualitative analysis costs associated with face-to-face interaction tracking in experimental scenarios.

### 2.4.3 Accuracy of Learning Phase Extraction

The accuracy of learning phase extraction was assessed with the five experimental cases of collaborative learning in Sec. 2.3. A simulation of all combinations of window size and slide width for sliding windows in learning phase extraction was conducted, with the optimal parameters selected to calculate face-to-face differences across users. The accuracy of learning phase extraction was determined using the qualitative analysis results as the ground truth. Based on the learning phase design detailed in Sec. 2.2.2, the best combinations of parameters for sliding windows in learning phase extraction were identified from all possible combinations, which resulted in three phases using AutoPlait.

Table 2.5 presents the best parameter combinations and the qualitative/quantitative phase transitions for learning phase extraction. In cases 1 through 5, the learning phases were extracted with accuracies of 86.9 %, 100 %, 99.8 %, 91.1 %, and 90.9 %, respectively, and phase transitions were predicted within an average of 1 min. These results indicate that this approach effectively supports researchers in learning science by reducing the costs associated with qualitative analysis of learning phases.

### 2.4.4 Accuracy of Speaker Identification

The accuracy of speaker identification was assessed with three subjects. Each subject wore SRP Badge on the chest and was seated 1.5 meters apart from the others. To ensure time synchronization between the badges, the synchronizer was positioned at the center of the desk used by the subjects. For the experiment, each subject was provided with a printed speech manuscript. The speeches on the manuscripts were designed to take approximately 6 to 8 seconds to deliver in Japanese. The subjects took turns reading their speech. Each subject spoke at regular intervals to avoid overlapping with others.

The algorithm correctly identified user 1's speech with 100 % accuracy across 15 samples, while the speeches of users 2 and 3 were also identified with 100 % accuracy for 14 samples. The results demonstrate that the speaker identification algorithm successfully distinguished all speakers.

### 2.4.5 Processing Time

The processing time for each function of the proposed web services deployed on EC2 instances was evaluated. A dataset of collaborative learning activity using SRP Badges was selected for analysis. Sensor data of different durations (15 min, 30 min, 45 min, and 60 min) were extracted from three participants' chest-mounted sensors during a one-hour learning activity. Each processing time was recorded for three functions without sensor data—sign up, log in, and create a session—and five functions with sensor data—importing sensor data, extracting face-to-face interaction, extracting learning phases, identifying speakers, and estimating activity. Each processing time was calculated as the average of ten measurements.

Tables 2.6 and 2.7 present the results of processing time for functions without and with sensor data, respectively. Table 2.6 illustrates the processing time for signing up, logging in, and creating a session on t3.large, m6i.large, and m6i.2xlarge instances. Table 2.7 provides the processing time for importing sensor data, extracting face-to-face interaction, extracting learning phases, identifying speakers, and estimating activity for durations of 15 min, 30 min, 45 min, and 60 min on the same instances.

Three key observations were made from the results presented in Tables 2.6 and 2.7:

- The m6i.large and m6i.2xlarge instances processed each function faster than the t3.large instance.
- Differences in processing time between m6i.large and m6i.2xlarge were minimal.
- Speaker identification emerged as the most computationally intensive function.

The first observation indicates that t3.large might delay the analysis of collaborative learning activities when using the proposed web services. This delay can be attributed to differences in network bandwidth: t3.large supports up to 5 Gbps, whereas m6i.large and m6i.2xlarge support up to 12.5 Gbps. For improved performance, using m6i.large or m6i.2xlarge is recommended.

The second observation suggests that m6i.large provides sufficient CPU performance and memory for handling the functions in the web services. The m6i.large instance is equipped with four CPUs and 16 GiB of memory, whereas m6i.2xlarge features eight CPUs and 32 GiB of memory. Despite the higher specifications of m6i.2xlarge, the processing speed of m6i.large was adequate for the tested functions.

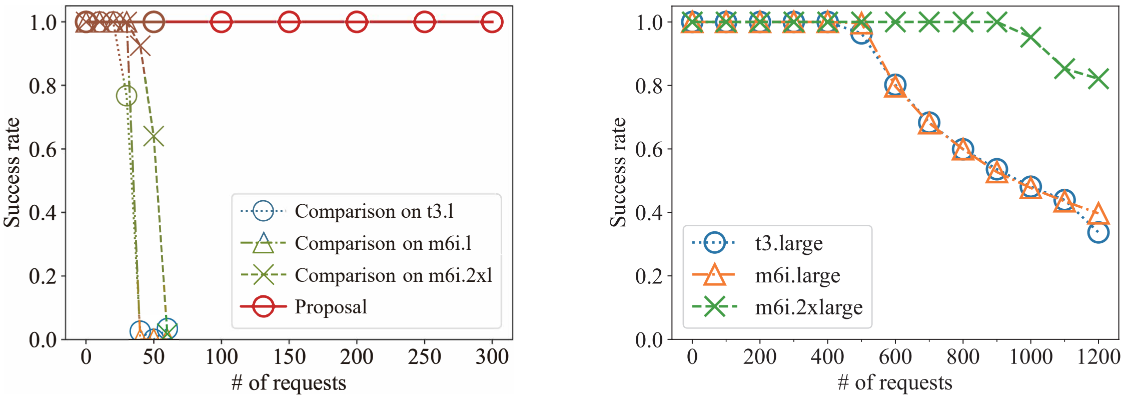
The third observation highlights potential for optimization in the speaker identification process. For example, on m6i.large, providing speaker information from 60 min of sensor data took 61.8 s, as shown in Tab. 2.7 (b). While the web services successfully provide this information, the function requires over 1 min to complete, which may hinder efficiency. Addressing this issue and accelerating the speaker identification process remain areas for future development.

### 2.4.6 Scalability

This section shows the scalability of the proposed web services deployed on Amazon EC2 instances, a service that provides scalable virtual server instances in the cloud. The scalability was compared to standard implementation of the application in Django. Multiple access requests were generated from one server to another using Apache's JMeter [20], with each server hosting the respective services. Speaker identification was performed using 1 min of sensor data collected from three users' chest-mounted SRP Badges during a collaborative learning session [1]. The number of access requests on the proposed web services ranged from 0 to 300, increasing in increments of 50, while the earlier web application handled requests ranging from 0 to 60, increasing in increments of 5. Furthermore, scalability testing with 60 min of sensor data was carried out. The proposed web services were deployed on EC2 instances, and multiple access requests were made under similar conditions. Speaker identification was requested using 60 min of sensor data collected in the same learning environment, with access requests ranging from 0 to 1200, increasing in increments of 100.

Figure 2.15 (a) depicts the scalability for handling multiple requests with 1 min of sensor data on both the standard web application (labeled Comparison) and the proposed web application (labeled Proposal). The x-axis shows the number of simultaneous requests sent to each service, while the y-axis represents the response success rate. The legend outlines the combinations of each service and the type of EC2 instance used: the standard application running on t3.large (Comparison on t3.l), m6i.large (Comparison on m6i.l), m6i.2xlarge (Comparison on m6i.2xl), and the proposed application running on any instance (Proposal). For the standard web application running on t3.large, the success rate remained at 1.0 for request numbers between 0 and 20, increasing by 5 each time. However, for 25 and higher request counts (25, 30, 35, 40, 45, 50, 55, and 60), the success rate declined to 0.960, 0.767, 0.029, 0.025, 0.000, 0.000, 0.000, and 0.033, respectively. On m6i.large, the success rate stayed at 1.0 for 0 to 30 requests, but dropped to 0.000 for requests between 35 and 60. On m6i.2xlarge, the success rate remained at 1.0 for requests from 0 to 35, but decreased to 0.925, 0.844, 0.640, 0.036, and 0.017 for 40, 45, 50, 55, and 60 requests, respectively. The proposed web application, in contrast, maintained a success rate of 1.0 for all requests ranging from 0 to 300, increasing in increments of 50. This shows that the proposed web application offered improved scalability compared to the standard web application.

[1] Speaker identification was the most resource-intensive function in the proposed web services.



(a) 1-minute data

(b) 60-minutes data

Figure 2.15: Scalability of SRP Web Services for multiple requests.

Figure 2.15 (b) illustrates the scalability of the proposed web services for multiple requests. The x-axis again represents the number of simultaneous requests, while the y-axis shows the success rate of responses. On t3.large, the success rate was 1.0 for request counts between 0 and 400 (in increments of 100), but then dropped to 0.964, 0.802, 0.683, 0.599, 0.536, 0.481, 0.439, and 0.337 for 500 to 1200 requests. On m6i.large, the success rate stayed at 1.0 for requests up to 400, but then fell to 0.996, 0.797, 0.681, 0.599, 0.527, 0.477, 0.435, and 0.397 for 500 to 1200 requests. On m6i.2xlarge, the success rate was 1.0 for up to 900 requests, but declined to 0.952, 0.853, and 0.821 for 1000, 1100, and 1200 requests. This figure demonstrates that the proposed web application could handle speaker identification, the most demanding function, with 400 simultaneous requests on t3.large and m6i.large, and up to 900 requests on m6i.2xlarge.

### 2.4.7 Running Cost

The cost of running the proposed web services was estimated based on the service fees on AWS. The deployment was assumed to take place on EC2 instances located in the Ohio region, USA, with an hourly rate of 0.0832 USD for t3.large, 0.096 USD for m6i.large, and 0.384 USD for m6i.2xlarge. The running cost was calculated as the product of the running time and the hourly rate.

The number of users was assumed to be 400 for t3.large, 400 for m6i.large, and 900 for m6i.2xlarge, representing the maximum number of users the proposed web services can handle simultaneously without rejecting requests, as discussed in Sec. 2.4.6. Each user was assumed to analyze five sessions of collaborative learning activity per month, with three learners per session and a session duration of 60 min. The running time included time for signing up, logging in, creating sessions, importing sensor data, extracting face-to-face interaction, extracting learning phases, identifying speakers, and estimating activity. These times were summed for each user based on the average processing times reported in Sec. 2.4.5.

Based on these assumptions, the total running cost for one month was approximately 5.658 USD on t3.large for 400 users, 4.320 USD on m6i.large for 400 users, and 18.816 USD on m6i.2xlarge for 900 users.

## 2.5 Related Work

### 2.5.1 Collaborative Extraction Using Business-Card-Type Sensors

Previous research has investigated the detection of collaboration between individuals using business-card-type sensors worn by users. One such example is Hitachi's Business Microscope [6, 7], which features an infrared sensor. Business Microscope captures face-to-face interaction and suggests that the frequency of meetings influences work efficiency. Similarly, MIT developed the Sociometric Badge [8] with accelerometers, sound pressure sensors, position sensors, Bluetooth, and infrared sensors. Sociometric Badge collects data on face-to-face interactions, conversational tone shifts, and proximity. The study in [8] indicates that these interactions correlate with workplace productivity and efficiency. Furthermore, MIT introduced a compact, energy-efficient variant called Open Badges [9], which includes sound pressure sensors and Bluetooth, and is worn around the neck. Open Badges enable visualization of face-to-face interactions using sound pressure and Bluetooth received signal strength indicator (RSSI) data. MIT later integrated Open Badges into a hybrid environment platform named Rhythm [10], designed to track face-to-face interactions in physical settings and facilitate interaction tracking in distributed settings through online applications.

Despite these advancements, there are challenges in achieving precise synchronization of sensor data, which limits the accuracy of collaboration analysis. Existing approaches typically rely on software-based synchronization methods. For instance, one study [6] attempts synchronization by identifying similar sound pressure patterns, aligning data sampled at 8 kHz within a 100 ms window. However, this method becomes less effective with sensors that operate at a lower sampling rate of 100 Hz to conserve power, introducing errors that can lead to inaccurate analysis of collaborative activities.

To address these limitations, this study proposes a new type of business card-type sensor, building on the earlier work [21], which focuses on achieving precise synchronization across multiple sensors. The proposed sensor incorporates hardware that enables high-precision synchronization through the transmission of synchronization packets between devices. This setup allows for the accurate capture of sound pressure, acceleration, and infrared data across all sensors.

### 2.5.2 Sensor-based Activity Recognition

Several studies have explored methods for recognizing user behavior using multiple sensors attached to the user [22–27]. In one such study [22], accelerometers were placed on the user's wrist, ankle, and chest, and the collected sensor data was transmitted to the cloud. The cloud then utilized decision tree analysis to classify six activities: lying down, sitting, standing, walking, running, and cycling. Another study [23] employed a wristwatch-style wearable device with integrated sensors such as an accelerometer, light sensor, thermometer, and sound sensor, enabling real-time classification of six activities: sitting, standing, walking, climbing stairs, descending stairs, and running, achieving an accuracy of 92.5 % using decision tree analysis. Additionally, literature [24] leveraged Zephyr BioHarness Bluetooth to gather acceleration and biometric data, classifying three activities — running, walking, and sitting — again using decision tree analysis. This study also demonstrated the ability to handle new users without requiring re-training by utilizing data from a diverse group of users. Another example [25] used fuzzy basis functions to analyze data from a 3-axis accelerometer worn on the user's dominant wrist, successfully classifying seven activities: brushing teeth, tapping a person, tapping a desk, working on a computer, running, waving, and walking.

Building on this body of work, this study proposes SRP leverages data from SRP Badges to identify and visualize key moments in collaborative activities. For instance, the system can automatically detect shifts in learning phases by analyzing variations in network activity among participants, as captured by infrared sensors mounted on the badges. This automated phase detection has the potential to significantly reduce the qualitative analysis workload for researchers studying collaborative activities, while also providing useful insights to guide the collaboration process.

### 2.5.3 Web Services for Sensor Data Analysis

Several studies have focused on creating user-friendly web services designed for sensor data analysis [28–39]. For instance, the work presented in [33] developed a model for smart agriculture, enabling real-time monitoring of soil conditions and remote control of field operations via mobile and web applications. This model offered users a convenient way to monitor data processed by the system through a web browser from any location at any time. Similarly, the study in [39] introduced a new SaaS platform called motch, designed to simplify the operation of IoT systems for end users via a web interface, allowing users to easily check the status of IoT devices directly from a browser.

This paper introduces a web application named SRP Web, which enhances the usability of sensor data analysis in collaboration. SRP Web aims to make it easier for analysts, even those with limited technical expertise, to begin conducting analysis by providing improved scalability and access to analysis algorithms.

## 2.6 Conclusion

This study introduces an innovative IoT system utilizing business-card-type sensors designed to facilitate the analysis of collaboration. The system is composed of three main components: compact business-card-type sensors called SRP Badge for data collection, an interaction analysis algorithm called SRP Analysis to interpret the collected data, and a web-based application called SRP Web for visualizing the analysis results in a browser. SRP Badge, worn individually by users, accurately collects data such as sound pressure, acceleration, and infrared signals while maintaining precise synchronization between devices. SRP Analysis processes this synchronized data to identify interactions, including face-to-face communication, learning phases, speakers, and activities. The results are then presented through SRP Web that allows for easy visualization and interpretation. To assess the system's effectiveness, experiments were carried out focusing on sensor synchronization accuracy, the performance and reliability of the interaction analysis algorithm, and the usability of the web application. The findings highlighted several advantages for researchers analyzing collaborative learning. First, the sensors achieved high precision in data acquisition, with synchronization errors between devices kept within ±30 µs. Second, the interaction analysis algorithm successfully identified collaborative behaviors, such as face-to-face interactions, task phases, speech, and activities, providing valuable insights for qualitative analysis. Finally, the web application enabled intuitive visualization of key data points, significantly streamlining the process of human interaction analysis due to its web-based design and usability.

Table 2.3: Qualitative transcription in each phase for case 1

(a) Discussion

| Number | Start [s] | End [s] | Speaker | Speech content (in Japanese) |
|---|---|---|---|---|
| 1 | 1302 | 1303 | User 1 | Then two thousand feet are... Ah, I see. |
| 2 | 1303 | 1309 | User 2 | One foot is one-third yard so three feet are two thousand-third yards. |
| 3 | 1310 | 1314 | User 1 | Really... I learn something new. |
| 4 | 1310 | 1311 | User 2 | Ha ha. |
| 5 | 1310 | 1311 | User 3 | Ha ha. |
| 6 | 1314 | 1316 | User 2 | I'm not confident... |
| 7 | 1314 | 1315 | User 3 | Ha ha. |
| 8 | 1317 | 1319 | User 2 | Six pounds... |
| 9 | 1322 | 1323 | User 2 | Fifteen pounds. |
| 10 | 1323 | 1325 | User 1 | Fifteen pounds. |
| 11 | 1325 | 1326 | User 3 | Pound... |
| 12 | 1332 | 1333 | User 2 | Ten... |
| 13 | 1334 | 1339 | User 1 | I know that the normal plane is two thousand feet long, but... |
| 14 | 1339 | 1441 | User 2 | They used this plane? |
| 15 | 1441 | 1347 | User 1 | Didn't the video say that the fuel is half? |
| 16 | 1342 | 1343 | User 2 | Yes, the video said. |
| 17 | 1350 | 1357 | User 1 | At the end of the video... Well, as I said before, the part of the normal plane is two thousand feet long... |
| 18 | 1352 | 1353 | User 3 | At the end? |

(b) Conclusion

| Number | Start [s] | End [s] | Speaker | Speech content (in Japanese) |
|---|---|---|---|---|
| 1 | 3706 | 3711 | User 1 | Yes, yes, yes, fifteen plus sixty, the fuel is loaded here and fully used... |
| 2 | 3708 | 3710 | User 2 | Ah, I see. |
| 3 | 3710 | 3711 | User 3 | (Whispered) |
| 4 | 3714 | 3716 | User 1 | About six gallons. |
| 5 | 3717 | 3719 | User 2 | One gallon is six pounds, right? |
| 6 | 3720 | 3721 | User 1 | Yes, yes, yes, yes. |
| 7 | 3722 | 3724 | User 2 | Then eight gallons are... |
| 8 | 3727 | 3728 | User 1 | Forty eight? |
| 9 | 3728 | 3729 | User 2 | Forty eight pounds. |
| 10 | 3729 | 3730 | User 1 | I see. |
| 11 | 3730 | 3731 | User 2 | Can they load the fuel of forty eight pounds? |
| 12 | 3731 | 3732 | User 3 | Forty eight pounds are bad. |
| 13 | 3732 | 3733 | User 2 | Bad? |
| 14 | 3733 | 3734 | User 3 | Less than forty five. |
| 15 | 3736 | 3737 | User 2 | Oh my! |
| 16 | 3737 | 3738 | User 1 | Ah... |
| 17 | 3738 | 3742 | User 2 | Ha ha ha, and they also have to load the eagle. |
| 18 | 3742 | 3743 | User 1 | The eagle, guy. |
| 19 | 3743 | 3754 | User 3 | But they use fifteen so reduce one gallon when the eagle, the eagle arrives. |
| 20 | 3755 | 3756 | User 2 | Hmm... ha ha ha. |
| 21 | 3756 | 3757 | User 3 | So... |

Table 2.4: Qualitative recodes of activity in each phase for case 1

(a) Discussion

| Number | Start [s] | End [s] | Learner | Activity |
|---|---|---|---|---|
| 1 | 1300 | 1316 | User 1 | He wrote on the whiteboard. |
| 2 | 1303 | 1314 | User 3 | She watched the iPad and whiteboard in turn. |
| 3 | 1308 | 1314 | User 2 | She spoke moving the chair back and forth. |
| 4 | 1323 | 1336 | User 1 | He wrote on the whiteboard. |
| 5 | 1323 | 1326 | User 2 | She watched the iPad and whiteboard in turn. |
| 6 | 1323 | 1326 | User 3 | She manually replayed the video on the iPad. |
| 7 | 1329 | 1330 | User 3 | She turned her head toward the whiteboard from the iPad. |
| 8 | 1339 | 1342 | User 3 | She manually replayed the video on the iPad. |
| 9 | 1343 | 1344 | User 2 | She pulled away from the desk. |
| 10 | 1346 | 1348 | User 3 | She manually replayed the video on the iPad. |
| 11 | 1350 | 1356 | User 1 | He turned his head toward the whiteboard from the iPad. |

(b) Conclusion

| Number | Start [s] | End [s] | Learner | Activity |
|---|---|---|---|---|
| 1 | 3705 | 3717 | User 1 | He wrote on the whiteboard. |
| 2 | 3708 | 3713 | User 3 | She pointed out to the whiteboard. |
| 3 | 3720 | 3722 | User 3 | She scratched the side of her nose. |
| 4 | 3723 | 3730 | User 3 | She nodded repeatedly. |
| 5 | 3726 | 3734 | User 2 | She gestured in thinking. |
| 6 | 3734 | 3740 | User 1 | He swang the body with laughing. |
| 7 | 3735 | 3739 | User 3 | She laughed. |
| 8 | 3736 | 3740 | User 2 | She swang the body with laughing. |
| 9 | 3742 | 3749 | User 1 | He wrote on the whiteboard. |
| 10 | 3744 | 3749 | User 3 | She pointed out to the whiteboard. |
| 11 | 3750 | 3758 | User 3 | She swang the body with putting hand on her hip. |
| 12 | 3752 | 3754 | User 2 | She wondered scratching her head. |

Table 2.5: Best combinations of window size and slide width, accuracy, and phase transitions in learning phase extraction

| Case | Window size [s] | Slide width [s] | Accuracy | Transit from video viewing to discussion [s] | | Transit from discussion to conclusion [s] | |
|---|---|---|---|---|---|---|---|
| | | | | Qualitative | Quantitative | Qualitative | Quantitative |
| 1 | 86 | 1 | 100 % | 1,356 to 1,445 | 1,361 | 3,166 to 3,167 | 3,167 |
| 2 | 571 | 1 | 99.8 % | 1,386 to 1,502 | 1,386 | 3,011 to 3,012 | 3,003 |
| 3 | 554 | 2 | 91.1 % | 1,283 to 1,334 | 1,403 | 2,609 to 2,610 | 2,483 |
| 4 | 127 | 1 | 90.9 % | 1,275 to 1,343 | 1,262 | 2,541 to 2,542 | 2,259 |

Table 2.6: Processing time [s] for each function without sensor data

| Process | Instance type | | |
|---|---|---|---|
| | t3.large | m6i.large | m6i.2xlarge |
| Sign up | 0.72 | 0.53 | 0.54 |
| Log in | 0.42 | 0.28 | 0.27 |
| Create a session | 0.071 | 0.040 | 0.036 |

Table 2.7: Processing time [s] for each function with sensor data

(a) t3.large

| Process | Length of sensor data | | | |
|---|---|---|---|---|
| | 15 min | 30 min | 45 min | 60 min |
| Import a sensor datum | 1.76 | 2.97 | 4.19 | 5.89 |
| Extract F2F interaction | 2.89 | 5.29 | 7.93 | 9.98 |
| Extract learning phases | 2.20 | 2.98 | 4.08 | 5.07 |
| Identify speakers | 19.9 | 40.0 | 61.9 | 82.9 |
| Estimate activity | 2.55 | 3.21 | 4.31 | 5.39 |

(b) m6i.large

| Process | Length of sensor data | | | |
|---|---|---|---|---|
| | 15 min | 30 min | 45 min | 60 min |
| Import a sensor datum | 0.866 | 1.66 | 2.40 | 3.25 |
| Extract F2F interaction | 0.275 | 0.499 | 0.754 | 1.14 |
| Extract learning phases | 1.09 | 1.78 | 2.48 | 3.14 |
| Identify speakers | 15.1 | 31.1 | 46.3 | 61.8 |
| Estimate activity | 1.24 | 1.95 | 2.67 | 3.54 |

(c) m6i.2xlarge

| Process | Length of sensor data | | | |
|---|---|---|---|---|
| | 15 min | 30 min | 45 min | 60 min |
| Import a sensor datum | 1.00 | 1.97 | 2.82 | 4.25 |
| Extract F2F interaction | 0.284 | 0.552 | 0.894 | 1.27 |
| Extract learning phases | 1.40 | 1.97 | 2.62 | 3.29 |
| Identify speakers | 14.6 | 30.1 | 47.0 | 66.1 |
| Estimate activity | 1.28 | 2.00 | 2.80 | 3.54 |

# Chapter 3

# Speaker Identification for Mobile Devices

## 3.1 Introduction

Collaboration plays a significant role in the success of activities involving multiple participants, such as teamwork and collaborative learning. It enables individuals to integrate diverse perspectives and enhance social skills through interaction with others. The field of cognitive science has explored collaboration extensively, leading to various insights into how collaborative processes can improve performance, particularly in learning environments. For instance, researchers have conducted qualitative analyses of collaborative activities, uncovering patterns that contribute to enhanced learning outcomes [3, 40–46]. As highlighted in [43], learners who approach problem-solving in a unified manner often achieve better results. Analyzing collaboration often requires transcription, which is critical for accurately capturing and understanding the interactions. However, this process is both time-consuming and labor-intensive, as researchers must repeatedly watch recorded sessions to manually note the timing of each speaker's contributions.

One approach to reduce the challenges of transcription is speaker identification, which has been explored in several studies. These studies have investigated methods such as speaker localization using microphones [47–60], speaker verification using voice features [61–70], speaker identification using voice features [61, 64, 71–87], and speaker recognition using a mobile device [10, 88–92]. For example, speaker localization determines the positions of multiple speakers by analyzing audio data captured through microphones or microphone arrays. While many of these studies rely on high sampling rates (several kHz or more) for speaker recognition, this study focuses on identifying speakers using sound pressure sensors operating at a lower sampling rate. This approach offers a cost-effective solution for collecting collaboration data using low-power mobile devices.

One noteworthy contribution to this field is Rhythm [10], which employs a mobile device with a sound pressure sensor operating at 700 Hz. Rhythm uses integration circuits, voice activity detection (VAD) [93], and thresholding algorithms to identify speakers. Despite its innovation, several challenges remain in achieving accurate speaker identification using sound pressure sensors.

The first challenge is handling spikes in the recorded sound pressure data. Rhythm's integration circuit is susceptible to spikes, which can result in incorrect speech detection. The second challenge involves distinguishing between speech and ambient noise. Even if spikes are mitigated, non-speaking individuals' sensors still pick up elements of speech, complicating the classification of sound pressure data as either speech or noise. The third challenge is the lack of time synchronization among the sensors, which can lead to errors in aligning sound pressure data across devices. This misalignment makes it difficult to accurately classify the sound pressure data [1].

To address these challenges, this paper proposes a new speaker identification system specifically designed for business-card-type sensors. The system includes: 1) a sound pressure sensor designed to mitigate spikes, 2) a wireless synchronization framework to ensure data consistency, and 3) a high-accuracy speaker identification algorithm optimized for low sampling rate data. The key innovations of the approach are as follows:

- This study uses a peak hold circuit to reduce spikes in the sound pressure data.
- The system incorporates a flooding-based synchronization module for precise time alignment across devices.
- The study introduces a three-step process for distinguishing speech from ambient noise, improving the accuracy of speaker identification.

Evaluations demonstrate that 1) the peak hold circuit effectively removes spikes from the sound pressure data, 2) the synchronization error between sensors is consistently within ±30 µs, and 3) the proposed system performs well under various conditions, including different user numbers, noise levels, and utterance lengths.

The structure of the rest of this paper is as follows: Section 3.2 reviews the related works. Section 3.3 introduces the proposed algorithm for identifying speakers. The experimental results are presented in Sec. 3.4. Finally, Sec. 3.5 concludes the paper.

## 3.2 Related Work

### 3.2.1 Speaker Recognition Using Stationary Devices

Previous research in the field of speaker recognition can generally be divided into three categories: speaker localization, speaker verification, and speaker identification using voice characteristics. Speaker localization [47–55] focuses on determining a speaker's position by analyzing multiple audio

[1] To avoid errors caused by synchronization issues, the time synchronization accuracy must be less than one-tenth of the sensor's maximum sampling rate.

signals. Applications of this technique include mobile robotics [56–58], passive sonar systems [59], and hearing aids [60]. For instance, in environments with wideband noise, [58] introduces a method to differentiate the time difference of arrival (TDoA) between the sound source and the noise, which helps estimate the speaker's location.

Research on speaker verification [61–67] focuses on comparing a speaker's voice with a pre-registered voice sample to verify identity. This technology has been used for authenticating IoT devices [68], securing networks [69], and user authentication [70]. For example, [67] enhances the performance of speaker verification for low-quality voice inputs by integrating mel-frequency cepstral coefficients (MFCC) with linear predictive coding (LPC).

Speaker identification, another area of research, involves matching a speaker's voice to that of a pre-registered individual [61,64,71–84]. Applications of speaker identification include video conferencing [85], criminal investigations [86], and television broadcasts [87]. For instance, [85] improves speaker identification robustness by focusing on key speakers in video conferences, reducing noise from inactive participants and minimizing the interference of brief speech interruptions.

Despite the progress, many of these approaches involve substantial hardware and processing requirements, as they rely on microphones to capture voice samples at high frequencies, often exceeding several kHz. In contrast, this study uses a business-card-sized sensor that captures sound pressure at 100 Hz to identify speakers. This setup significantly reduces both hardware and processing costs, facilitating the extraction of collaborative data during multi-person activities.

### 3.2.2 Speaker Recognition Using Mobile Devices

Several studies [88,89] have implemented speaker identification using smartphones or business-card-type sensors to extract collaboration data in organizational settings [90–92] and human interaction contexts [10]. For instance, Hitachi's business microscope [90–92] utilizes a business-card-type sensor to achieve 97.3 % accuracy in speaker identification. However, it is important to note that this system exhibits high power consumption due to the sensor's high sound pressure sampling rate of 8 kHz.

Additionally, MIT's Rhythm project [10] employs a business-card-sized sensor known as the Rhythm Badge for speaker identification. The Rhythm Badge operates at a lower power consumption rate as it samples sound pressure at 700 Hz. It performs speaker identification using a threshold-based approach without extracting specific voice features. Nevertheless, its accuracy is somewhat limited due to the spikes observed in the sound pressure measurements (due to the integration circuit), the fixed threshold which makes it susceptible to ambient noise, and the lack of time synchronization across sensors.

This study introduces an innovative business-card-sized sensor designed to reduce spikes by incorporating a peak hold circuit. The study also develop a speaker identification algorithm that

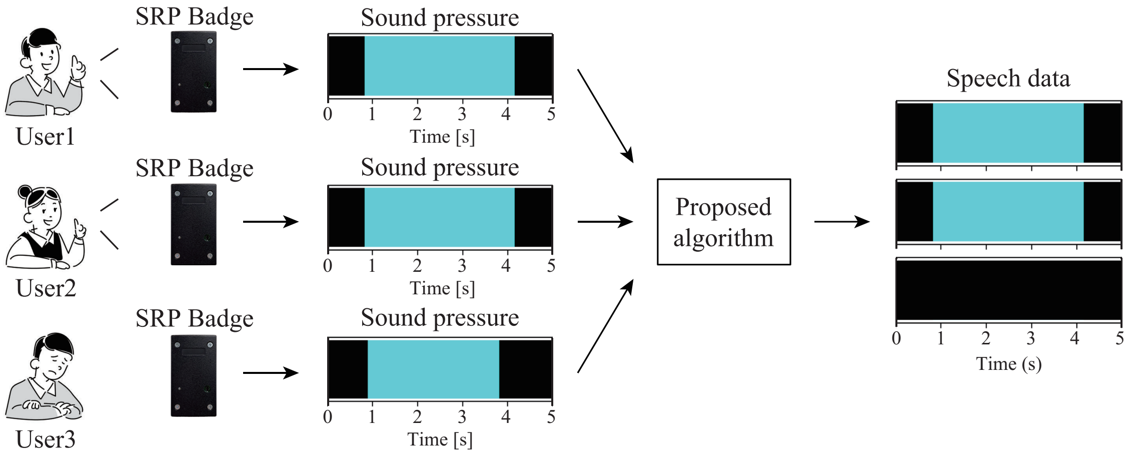


Figure 3.1: Overview of the proposed speaker identification system.

minimizes the impact of ambient noise and include precise time synchronization between sensors. Through simulations and experiments, the study demonstrate that these enhancements significantly improve the accuracy of both speech detection and speaker identification.

## 3.3 Proposed Scheme

### 3.3.1 Overview of Proposed System

To identify speakers using sound pressure sensors embedded in a business-card-sized sensor with a low sampling rate, this paper proposes a novel speaker identification system. Figure 3.1 provides an overview of the system we have developed. The speaker identification process follows these steps:

1. Before a multi-person activity, the business-card-sized sensors are distributed to participants.
2. During the activity, the sensors capture user speech using a sound pressure sensor equipped with a peak hold circuit.
3. After the activity, the sensors are collected from the participants.
4. The sound pressure data from the collected sensors are extracted and processed through the proposed speaker identification algorithm.
5. Finally, the algorithm generates and visualizes the speaker identification results.

### 3.3.2 Sound Pressure Acquisition

Figure 3.2 illustrates the design of the sound pressure sensor. This sensor samples sound pressure at intervals of 10 ms. The microphone converts spoken audio into electrical signals, which are weak

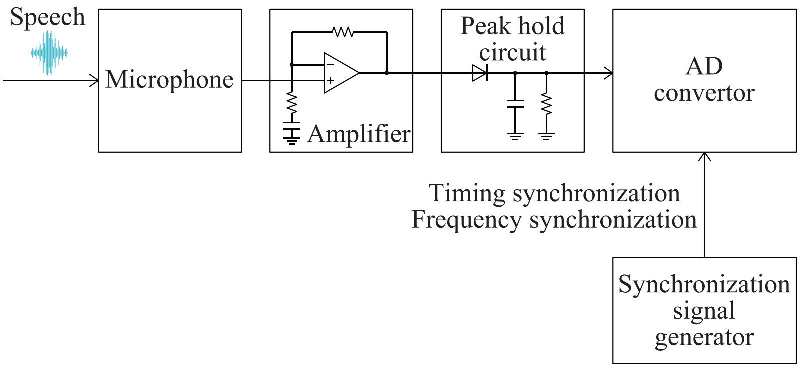


Figure 3.2: Sound pressure sensor.

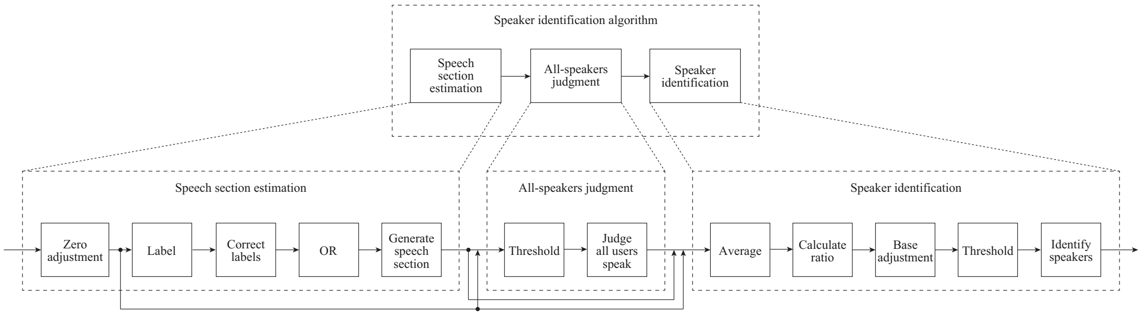


Figure 3.3: Overview of the speaker identification algorithm.

and therefore require amplification. The amplified signals are then passed through a peak hold circuit that detects rapid signal peaks by utilizing the discharge properties of an RC parallel circuit. The analog signal produced by the peak hold circuit is converted into a digital format through an analog-to-digital (AD) converter. The digital output is provided every 10 ms, with both timing and frequency synchronized using a synchronization signal generator. The sensor is designed to be both cost-effective and simple. The circuit is composed of a microphone, an operational amplifier, a peak hold circuit, and an AD converter, making it affordable and straightforward to implement.

Based on this hardware design, this study implemented a sound pressure sensor in the SRP Badges described in Sec. 2.2.1. As outlined in Sec. 2.2.1, the badge, which is the size of a small business card, operates continuously for 24 hours and is robust enough to endure extended collaboration sessions lasting several hours. The low-sampling sound pressure acquisition contributes to both low power consumption and compact design.

### 3.3.3 Speaker Identification Algorithm

Figure 3.3 illustrates the overall structure of the speaker identification algorithm. This algorithm operates in three stages: 1) estimating the speech segments, 2) evaluating all speakers, and 3) identifying the target speaker.

**Speech section estimation**: The initial step involves determining whether users are speaking by

**Algorithm 3** Labeling in speech section estimation

```
**Require:** $U, \mathbb{P}$
**Ensure:** $\mathbb{L}$
 1: **for all** $d \in U$ **do**
 2:   Insert zeros into all elements in $L_d$
 3:   $\xi \Leftarrow 0$
 4:   **while** $\xi <$ length of $L_d$ **do**
 5:     $W \Leftarrow P_d \in \mathbb{P}$ between $\xi$ to $\xi + 1\,\mathrm{s}$
 6:     $m \Leftarrow \max(W)$
 7:     **if** $m > \eta_s$ **then**
 8:       $\eta_m \Leftarrow m * 0.1$
 9:       **if** $w \in W > \eta_m$ **then**
10:         $w \Leftarrow 1$
11:       **else**
12:         $w \Leftarrow 0$
13:       **end if**
14:       Replace elements in $L_d$ with $w \in W$
15:     **end if**
16:     $\xi \Leftarrow \xi + 0.5\,\mathrm{s}$
17:   **end while**
18:   Insert $L_d$ into $\mathbb{L}$
19: **end for**
20: **return** $\mathbb{L}$
```

analyzing sound pressure signals collected from the sensors of all users. The algorithm performs a zero-point correction by identifying the minimum sound pressure value across all sensors and subtracting this value from each sensor's respective sound pressure readings. Using sliding windows, the algorithm then labels whether multiple users are speaking based on the corrected sound pressure values for each user within each window.

Algorithm 6 illustrates the labeling procedure, which is depicted in Figure 3.3, and the notation used is summarized in Table 3.1. The labeling process produces an array, $\mathbb{L}$, which represents "1-0 data" for each user, using both the set of all sensor IDs $U$ and the sound pressure data $\mathbb{P} = P_1, P_2, \ldots, P_{|U|}$ from each sensor. For each window $W$, the algorithm identifies the maximum sound pressure value $m$ for each sensor, as seen in line 6.

If $m$ does not exceed the speech threshold $\eta_s$ in any sensor for window $W$, the algorithm considers that no users are speaking and moves to the next window (line 16). If $m$ does surpass $\eta_s$, the algorithm updates the threshold $\eta_m$ to $m * 0.1$ in line 8. It then compares the sound pressure values from each sensor with $\eta_m$, assigning a label of 1 or 0 depending on whether the sound pressure is higher or lower than $\eta_m$ (lines 9–13).

The corresponding element in array $L_d$ is updated with the label for window $W$ in line 14. The resulting pre-processed data, referred to as "1-0 data for each user," is used to refine the speech labels for each sensor. Labels of 1 are filled in sections with consecutive labels of 0 if these zeros occur within 90 ms between ones, treating them as part of the speech in the 1-0 data. Additionally,

Table 3.1: Notation

| Variable / Function | Description |
|---|---|
| $A$ | 1 bit array with labels for all users' speech |
| $d$ | Sensor ID |
| $f$ | Flag for user utterances |
| $\mathbb{J}$ | Set of 1 bit arrays with speech judgment labels for all the sensors |
| $J_d$ | 1 bit array with speech judgment labels for sensor $d$ |
| $\mathbb{L}$ | Set of 1 bit arrays with speech labels |
| $L_d$ | 1 bit array with speech labels of sensor $d$ |
| $\mathbb{P}$ | Set of sound pressure data acquired from all sensors |
| $P_d$ | Sound pressure data for sensor $d$ |
| $\mathbb{P}_{avg}$ | Set of averaged sound pressure data acquired from all sensors |
| $P_{d_{avg}}$ | Averaged sound pressure data for sensor $d$ |
| $\mathbb{S}$ | Set of arrays with start and end times for speech sections |
| $S$ | Array with start and end times for a speech section |
| $U$ | Set of all sensor IDs |
| $\eta_m$ | Speech threshold based on the maximum sound pressure in the window |
| $\eta_r$ | Speech threshold for sound pressure ratio |
| $\eta_S$ | Threshold for all users' speech in the speech section $S$ |
| $\eta_s$ | Speech threshold for all sensors |
| $\xi$ | Top index of window |
| average($X$) | Calculate the average of all the elements in $X$ |
| max($X$) | Calculate the maximum of all the elements in $X$ |
| min($X$) | Calculate the minimum of all the elements in $X$ |
| size($X$) | Count the number of all elements in $X$ |

continuous labels of 1 lasting less than 150 ms are replaced with 0s, assuming the section contains false positives caused by background noise.

The final labels for each user are logically combined and output as scalar binary data. This data, derived from the speech section estimation process, is referred to as "speech section data."

**All-speakers judgment**: In the second step, the algorithm determines whether all users are speaking within each speech section by combining the 1-0 data for each user with the speech section data. The focus is placed on sections where the speech section data indicates that a user is speaking. For each speech section, the algorithm calculates a threshold based on the maximum and minimum sound pressure values across all sensors. If the sound pressure for all sensors surpasses

**Algorithm 4** All-speakers judgment

```
Require: $U, \mathbb{S}, \mathbb{P}$
Ensure: $A$
 1: Insert zeros into all elements in $A$
 2: for all $S \in \mathbb{S}$ do
 3:     $p_{min} \Leftarrow \min(\mathbb{P})$ in $S$ before and after 100 ms
 4:     $p_{max} \Leftarrow \max(\mathbb{P})$ in $S$
 5:     $\eta_S \Leftarrow p_{min} + (p_{max} - p_{min}) * 0.95$
 6:     for all $d \in U$ do
 7:         $p_{d_{max}} \Leftarrow \max(P_d)$ in $S$
 8:         if $p_{d_{max}} > \eta_S$ then
 9:             Replace $a \in A$ in $S$ with 1
10:         end if
11:     end for
12: end for
13: return $A$
```

this threshold, it concludes that all users are speaking.

The process for making this all-speakers determination is detailed in Algorithm 4 and illustrated in Figure 3.3, while Table 3.1 outlines the algorithm's notation. The output of this step is an array $A$, representing the speech activity of all users. This array is generated using the set of all sensor IDs $U$, the speech sections $\mathbb{S}$ identified from the speech section estimation, and the sound pressure data $\mathbb{P} = P_1, P_2, \ldots, P_{|U|}$ collected from all sensors.

To estimate the noise floor, the algorithm calculates the minimum sound pressure $p_{min}$ by examining a 100 ms interval before and after each speech section across all sensors (line 3). This margin is added to ensure that the minimum sound pressure is accurately captured. Next, it identifies the maximum sound pressure $p_{max}$ within each speech section for all sensors (line 4). The threshold for determining all-users speech $\eta_S$ is then set as $p_{min} + (p_{max} - p_{min}) * 0.95$ (line 5). The value of 0.95 was chosen as it maintained high accuracy while reducing the likelihood of overly lenient judgments.

If the sound pressure within the speech section exceeds $\eta_S$ for all sensors, the algorithm classifies the section $S$ as one where all users are speaking, assigning a label of 1 for that section (lines 6–11). The algorithm then returns the speech labels for all users across the identified sections in array $A$.

**Speaker identification**: In the third step, the algorithm identifies which user is speaking during each speech section by utilizing sound pressure values that are averaged, relativized, and adjusted based on a baseline. Each speech section is analyzed to estimate where a user is speaking, relying on the previously extracted speech section data. The speech of individual users is determined by comparing their sound pressure values with the established speech threshold.

The sound pressure for each sensor is averaged using sliding windows, with a window size of 0.5 s and a slide interval of 0.01 s, allowing for fine-grained detection of simultaneous speech from multiple speakers. The averaged sound pressures across all users $\mathbb{P}_{avg}$ are then used to identify

**Algorithm 5** Speaker identification with averaged sound pressure

```
Require: $U, \mathbb{S}, \mathbb{P}_{avg}$
Ensure: $\mathbb{J}$
 1: for all $d \in U$ do
 2:   Insert zeros into all elements in $J_d$
 3: end for
 4: $\eta_r \Leftarrow 1/\text{size}(U)$
 5: for all $S \in \mathbb{S}$ do
 6:   for all $t_i \in S$ do
 7:     $f \Leftarrow 0$
 8:     for all $d \in U$ do
 9:       $r \Leftarrow$ ratio of $P_{d_{avg}}$ to $\mathbb{P}_{d_{avg}}$ at $t_i$
10:       $\delta \Leftarrow \text{average}(\forall r \text{ in } \neg S) - \eta_r$
11:       $r_{base} \Leftarrow r - \delta$
12:       if $r_{base} > \eta_r + 0.01$ then
13:         Replace $j \in J_d$ at $t_i$ with 1
14:         $f \Leftarrow 1$
15:       end if
16:     end for
17:     if $f = 0$ then
18:       for all $d \in U$ do
19:         if $r_{base} > \eta_r - 0.001$ then
20:           Replace $j \in J_d$ at $t_i$ with 1
21:         end if
22:       end for
23:     end if
24:   end for
25: end for
26: Insert $J_d$ into $\mathbb{J}$
27: return $\mathbb{J}$
```

the active speakers. The steps for speaker identification, based on these averaged sound pressures, are detailed in Algorithm 5, illustrated in Figure 3.3, and the relevant notation is explained in Table 3.1.

The output of the identification process is the array $\mathbb{J}$, which represents the labeled speech data for each user. This array is derived from the set of all sensor IDs $U$, the identified speech sections $\mathbb{S}$ from speech section estimation, and the averaged sound pressure data for all sensors $\mathbb{P}avg = P1_{avg}, P_{2_{avg}}, \ldots, P_{|U|_{avg}}$.

The speech threshold $\eta_r$ is determined based on the number of sensors, as indicated in line 4. For each time $t_i$ within a speech section $S \in \mathbb{S}$, the algorithm calculates the sound pressure ratio $r$ for each sensor using the averaged sound pressure $p$ of each sensor $d$ (line 9). A baseline adjustment $\delta$ is then computed by comparing the sound pressure ratio during non-speech sections $\neg S$ to the threshold $\eta_r$ (line 10), and this offset is subtracted from the sound pressure ratio for each sensor (line 11).

The algorithm identifies speakers within each speech section $S$ using a two-step process that incorporates both the averaged and baseline-adjusted ratio $r_{base}$. If the ratio $r_{base}$ for sensor $d$ exceeds the sum of the threshold $\eta_r$ and a margin of error of 0.01 in section $S$, the algorithm classifies the user associated with sensor $d$ as speaking in that section (lines 12–15). This initial step detects clear speech when only a few people are talking.

The threshold $\eta_r$ was optimized to 0.01, balancing detection accuracy without misclassifying noise as speech. If the first step does not detect any speakers in section $S$, a second step is applied with a slightly reduced margin of error of -0.001 (lines 18–22), which helps identify cases where multiple users may be speaking simultaneously. This second pass detects more ambiguous speech where multiple people are involved. Finally, the algorithm outputs the speech labels for each sensor in the array $\mathbb{J}$.

## 3.4 Evaluation

### 3.4.1 Speaker Identification Accuracy

An experimental evaluation was conducted to assess the accuracy of the proposed algorithm for detecting speech using sound pressure data acquired from SRP Badges. The experiments took place in a conference room, considering different numbers of participants, environmental noises, and short utterances from the users. This study assumes a situation where environmental noise is generated by video materials [18] played during collaborative learning activities.

The subjects were male university students in their early twenties. The room's dimensions were 10.6 m, 7.05 m, and 2.65 m. The influence of reverberation was taken into account, as the room was intended for collaborative learning. Each participant wore an SRP Badge on the chest and was seated 1.50 m away from adjacent participants. A time synchronizer was placed on a table at the center of the participants to ensure sensor synchronization.

For the experiments, both long and short utterances were tested using two types of speech scripts provided to each participant. Table 3.2 shows the speech script prepared for the experiments. Each script included 15 sentences in English. Participants took turns speaking one sentence from the script, with a two-second interval between speakers. After completing a sentence, all participants moved on to the next. The combinations of participants who spoke simultaneously were varied for each sentence. For example, in an experiment involving three participants, the combinations were as follows:

- One participant speaks while the other two remain silent
- Two participants speak simultaneously while the other remains silent
- All three participants speak simultaneously

Table 3.2: Speech script prepared for the experiments

| Order | Long speech | Short utterance |
|---|---|---|
| 1 | Nice to meet you everyone. | Oh. |
| 2 | What do you study at the university? | Hmmm. |
| 3 | Do you know where the library is? | Huh? |
| 4 | I have a friend who speaks Chinese. | What? |
| 5 | Please don't keep the door open. | Hey. |
| 6 | I can hardly believe your story. | Hello. |
| 7 | I don't know what you want to do. | Pardon? |
| 8 | Shall we go hiking if it is sunny tomorrow? | OK. |
| 9 | What should I do in order to improve my English? | Thanks. |
| 10 | It is said that English is an international language. | Good. |
| 11 | Without your help, we could not finish this job. | Really? |
| 12 | It is dangerous for children to play here. | Me, too. |
| 13 | Walking to the station, I met my father. | Yes. |
| 14 | It takes five minutes to walk to the station. | No. |
| 15 | I got up early so that I could make lunch. | Nice. |

In each case, all possible combinations of speakers and non-speakers were tested, taking into account the variations in participants' voice characteristics.

The accuracy of speaker identification was evaluated by comparing the proposed scheme with three alternatives: "Scheme with absolute sound pressure" (absolute scheme), "Scheme with relative sound pressure" (relative scheme), and "An extended version of the method presented in [10]" (Rhythm scheme). Both the absolute and relative schemes incorporate speech-section estimation from parts of the proposed algorithm described in Sec. 3.3.3. In the absolute scheme, speaker identification relied on a speech threshold applied to each speech section, similar to the all-speakers judgment approach in Sec. 3.3.3. For each user, the threshold $\eta_S$ was determined in each speech section $S$ to identify individual speech segments. On the other hand, the relative scheme identified multiple speakers using averaged and base-adjusted sound pressure, employing the thresholding approach found in the speaker identification method of Sec. 3.3.3. The optimal speech detection threshold for both schemes was dependent on the evaluation conditions. The Rhythm scheme, based on the work in [10], originally focused on identifying a single speaker using IoT devices called Rhythm Badges. This method was extended to support the identification of multiple speakers. The original scheme employed the VAD (Voice Activity Detection) algorithm [93] and a thresholding algorithm to identify a single speaker for organizational management purposes. The VAD algorithm used sliding windows to process sound pressure power, mitigating noise. The thresholding algorithm then identified the speaker by selecting the user with the longest detected speech segment in each section. To extend this approach for multiple speakers, modifications were made to the thresholding algorithm, allowing it to detect simultaneous speakers by evaluating speech activity for each user. The sliding window parameters for the VAD algorithm were empirically set to a window size of 2 s and a slide width of 0.01 s. The optimal threshold for speech detection in the

**Algorithm 6** Labeling in speech section estimation

```
Require: $U, \mathbb{P}$
Ensure: $\mathbb{L}$
for all $d \in U$ do
    Insert zeros into all elements in $L_d$
    $\xi \Leftarrow 0$
    while $\xi <$ length of $L_d$ do
        $W \Leftarrow P_d \in \mathbb{P}$ between $\xi$ to $\xi + 1\,\mathrm{s}$
        $m \Leftarrow \max(W)$
        if $m > \eta_s$ then
            $\eta_m \Leftarrow m * 0.1$
            if $w \in W > \eta_m$ then
                $w \Leftarrow 1$
            else
                $w \Leftarrow 0$
            end if
            Replace elements in $L_d$ with $w \in W$
        end if
        $\xi \Leftarrow \xi + 0.5\,\mathrm{s}$
    end while
    Insert $L_d$ into $\mathbb{L}$
end for
return $\mathbb{L}$
```

thresholding algorithm varied depending on the specific evaluation settings.

**The number of users**: The speaker identification accuracy was evaluated with varying numbers of users, using a script for long speech utterances. The number of users ranged from two to five. The speech threshold $\eta_s$ of Algorithm 6 was empirically set to 75 dB in the absolute, relative, and proposed algorithms for speech section estimation. For the Rhythm scheme, the speech threshold was set to 84 dB. This threshold was maintained consistently across different numbers of users.

Tables 3.3 and 3.4 present the F1-scores of each scheme and the corresponding confusion matrices for two to five users. In Table 3.4, symbols indicate whether speech was present (T) or absent (F), and whether the proposed algorithm estimated speech (P) or non-speech (N). Compared to the absolute and relative schemes, the proposed algorithm combined the strengths of both comparative schemes. It effectively identified all speakers in situations where all users spoke, leveraging a combination of techniques from the two schemes. In cases involving fewer speakers, the proposed scheme achieved high F1-scores by utilizing the relative scheme's advantages. Additionally, for detecting a single speaker, the proposed scheme performed well by benefiting from both comparative schemes. However, Table 3.4 shows that the F1-scores were slightly lower in the cases of one and four speakers out of five users, where the threshold incorrectly identified a non-speaker as a speaker (False Positive). When compared with the Rhythm scheme, the proposed algorithm accurately detected intermediate numbers of speakers, such as two out of three users, and similarly for four and five users. As indicated in Table 3.4, the proposed scheme successfully avoided false

Table 3.3: F1-scores under the different number of users

| Case | | Scheme | | | |
|---|---|---|---|---|---|
| # of users | # of speakers | Absolute | Relative | Rhythm | Proposed |
| 2 | 1 | **1.00** | **1.00** | **1.00** | **1.00** |
| | 2 | 0.881 | 0.706 | **1.00** | 0.881 |
| 3 | 1 | 0.978 | 0.978 | **0.988** | 0.978 |
| | 2 | 0.876 | 0.810 | 0.942 | **0.963** |
| | 3 | 0.876 | 0.810 | **1.00** | 0.913 |
| 4 | 1 | 0.945 | 0.960 | **0.974** | 0.960 |
| | 2 | 0.846 | **0.960** | 0.957 | **0.960** |
| | 3 | 0.893 | **0.960** | 0.874 | **0.960** |
| | 4 | 0.879 | 0.852 | **1.00** | 0.912 |
| 5 | 1 | **1.00** | 0.993 | 0.951 | 0.993 |
| | 2 | 0.821 | **0.976** | 0.913 | **0.976** |
| | 3 | 0.779 | **0.962** | 0.938 | **0.962** |
| | 4 | 0.857 | **0.938** | 0.905 | 0.937 |
| | 5 | 0.894 | 0.772 | **1.00** | 0.909 |

positives in most cases.

**Environmental noise**: The influence of environmental noise on speaker identification accuracy was also assessed, involving three participants. A noise source was positioned 2 m away from the table, generating five types of ambient noise: recorded sounds from trains, offices, streets, cars, and rain. The other settings were consistent with the experiments under the different number of users. Noise levels were set at 75 dB for trains, 70 dB for offices and streets, and 60 dB for cars and rain, on average. Speech thresholds for the absolute, relative, and proposed algorithms were empirically adjusted based on the environment, ranging from 80 dB to 85 dB. The Rhythm scheme had thresholds between 84 dB and 89 dB, depending on the noise type.

Tables 3.5 and 3.6 display the F1-scores and confusion matrices for various environmental noises. In these tables, the same symbol conventions were used as in previous evaluations. The proposed scheme generally outperformed the absolute and relative schemes by accurately detecting speakers and combining the strengths of both methods. However, Table 3.4 highlights slightly lower F1-scores for the proposed scheme in certain scenarios, such as one or two speakers out of three users under train or office noise, where false positives occurred. Compared to the Rhythm scheme, the proposed algorithm performed better under street noise, showing a higher tolerance for low-frequency noise, particularly between 10 Hz and 20 Hz. However, the Rhythm scheme excelled in identifying speakers in rain noise, which involved uniform frequencies ranging between 0 Hz and 50 Hz, indicating some limitations in the proposed algorithm when handling such noise.

**Short utterance**: The impact of short utterances, defined as speeches lasting less than one second [94], was also examined using a dedicated script. The other settings were consistent with the experiments under the different number of users. The speech threshold $\eta_s$ was empirically set to 73 dB for the absolute, relative, and proposed algorithms, while the Rhythm scheme used a threshold of 78 dB.

Table 3.4: Confusion matrices under the different number of users

(a) Two users

| | One speaker out of two users | | | | | | | | Two speakers out of two users | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Absolute | | Relative | | Rhythm | | Proposed | | Absolute | | Relative | | Rhythm | | Proposed | |
| | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N |
| T | 30 | 0 | 30 | 0 | 30 | 0 | 30 | 0 | 26 | 4 | 18 | 12 | 30 | 0 | 26 | 4 |
| F | 0 | 90 | 0 | 90 | 0 | 90 | 0 | 90 | 3 | 27 | 3 | 27 | 0 | 30 | 3 | 27 |

(b) Three users

| | One speaker out of three users | | | | | | | | Two speakers out of three users | | | | | | | | Three speakers out of three users | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Absolute | | Relative | | Rhythm | | Proposed | | Absolute | | Relative | | Rhythm | | Proposed | | Absolute | | Relative | | Rhythm | | Proposed | |
| | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N |
| T | 44 | 1 | 44 | 1 | 44 | 1 | 44 | 1 | 85 | 5 | 90 | 0 | 90 | 0 | 90 | 0 | 39 | 6 | 34 | 11 | 45 | 0 | 42 | 3 |
| F | 1 | 224 | 1 | 224 | 0 | 225 | 1 | 224 | 9 | 171 | 8 | 172 | 11 | 169 | 8 | 172 | 5 | 40 | 5 | 40 | 0 | 45 | 5 | 40 |

(c) Four users

| | One speaker out of four users | | | | | | | | Two speakers out of four users | | | | | | | | Three speakers out of four users | | | | | | | | Four speakers out of four users | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Absolute | | Relative | | Rhythm | | Proposed | | Absolute | | Relative | | Rhythm | | Proposed | | Absolute | | Relative | | Rhythm | | Proposed | | Absolute | | Relative | | Rhythm | | Proposed | |
| | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N |
| T | 60 | 0 | 60 | 0 | 57 | 3 | 60 | 0 | 143 | 37 | 179 | 1 | 178 | 2 | 179 | 1 | 146 | 34 | 166 | 14 | 180 | 0 | 166 | 14 | 51 | 9 | 46 | 14 | 60 | 0 | 52 | 8 |
| F | 7 | 413 | 5 | 415 | 0 | 420 | 5 | 415 | 15 | 525 | 14 | 526 | 14 | 526 | 14 | 526 | 1 | 299 | 0 | 300 | 52 | 248 | 0 | 300 | 5 | 55 | 2 | 58 | 0 | 60 | 2 | 58 |

(d) Five users

| | One speaker out of five users | | | | | | | | Two speakers out of five users | | | | | | | | Three speakers out of five users | | | | | | | | Four speakers out of five users | | | | | | | | Five speakers out of five users | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Absolute | | Relative | | Rhythm | | Proposed | | Absolute | | Relative | | Rhythm | | Proposed | | Absolute | | Relative | | Rhythm | | Proposed | | Absolute | | Relative | | Rhythm | | Proposed | | Absolute | | Relative | | Rhythm | | Proposed | |
| | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N |
| T | 75 | 0 | 75 | 0 | 68 | 7 | 75 | 0 | 213 | 87 | 300 | 0 | 258 | 42 | 300 | 0 | 316 | 133 | 447 | 3 | 437 | 13 | 447 | 3 | 237 | 63 | 274 | 26 | 300 | 0 | 275 | 26 | 63 | 12 | 49 | 26 | 75 | 0 | 65 | 10 |
| F | 0 | 675 | 1 | 674 | 0 | 675 | 1 | 674 | 6 | 1194 | 15 | 1185 | 7 | 1193 | 15 | 1183 | 46 | 1005 | 32 | 1018 | 45 | 1005 | 32 | 1018 | 16 | 434 | 10 | 440 | 63 | 387 | 11 | 438 | 3 | 72 | 3 | 72 | 0 | 75 | 3 | 72 |

Table 3.5: F1-scores under the different environmental noise conditions

| Case | | | Scheme | | | |
|---|---|---|---|---|---|---|
| Noise | # of users | # of speakers | Absolute | Relative | Rhythm | Proposed |
| Train | 3 | 1 | **0.891** | 0.763 | 0.889 | 0.738 |
| | | 2 | 0.914 | 0.878 | **0.935** | 0.857 |
| | | 3 | 0.936 | 0.814 | 0.966 | **0.968** |
| Office | 3 | 1 | **0.938** | 0.918 | 0.875 | 0.928 |
| | | 2 | 0.865 | **0.888** | 0.878 | 0.845 |
| | | 3 | 0.918 | 0.725 | **0.989** | 0.938 |
| Street | 3 | 1 | **0.938** | 0.849 | 0.706 | 0.849 |
| | | 2 | 0.920 | **0.973** | 0.767 | **0.973** |
| | | 3 | 0.933 | 0.769 | 0.889 | **0.945** |
| Car | 3 | 1 | 0.865 | 0.882 | **0.989** | 0.900 |
| | | 2 | 0.867 | **0.938** | 0.927 | **0.938** |
| | | 3 | 0.839 | 0.587 | **1.00** | 0.795 |
| Rain | 3 | 1 | 0.928 | 0.938 | **0.989** | 0.947 |
| | | 2 | 0.853 | 0.899 | **0.942** | 0.899 |
| | | 3 | 0.938 | 0.824 | **1.00** | 0.947 |

Tables 3.7 and 3.8 provide the F1-scores and confusion matrices for short utterances. As with the previous evaluations, symbols in Table 3.8 indicate whether speech occurred (T) or not (F), and whether the proposed algorithm estimated speech (P) or non-speech (N). The proposed scheme demonstrated strong performance by leveraging the advantages of both comparative schemes. It accurately identified all speakers in most cases and performed well in single-speaker scenarios by using techniques from the relative scheme. However, in the case of two speakers out of three users, the proposed algorithm's F1-score was slightly lower due to false positives. In comparison to the Rhythm scheme, the proposed algorithm showed better accuracy for cases with one or two speakers

Table 3.6: Confusion matrices under the different environmental noise conditions

(a) Train

| | One speaker out of three users | | | | | | | | Two speakers out of three users | | | | | | | | Three speakers out of three users | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Absolute | | Relative | | Rhythm | | Proposed | | Absolute | | Relative | | Rhythm | | Proposed | | Absolute | | Relative | | Rhythm | | Proposed | |
| | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N |
| T | 45 | 0 | 45 | 0 | 36 | 9 | 45 | 0 | 87 | 3 | 89 | 1 | 79 | 11 | 90 | 0 | 37 | 8 | 28 | 17 | 42 | 3 | 45 | 0 |
| F | 34 | 191 | 5 | 220 | 0 | 225 | 55 | 170 | 33 | 147 | 9 | 171 | 0 | 180 | 47 | 133 | 5 | 40 | 5 | 40 | 0 | 45 | 6 | 39 |

(b) Office

| | One speaker out of three users | | | | | | | | Two speakers out of three users | | | | | | | | Three speakers out of three users | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Absolute | | Relative | | Rhythm | | Proposed | | Absolute | | Relative | | Rhythm | | Proposed | | Absolute | | Relative | | Rhythm | | Proposed | |
| | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N |
| T | 45 | 0 | 45 | 0 | 42 | 3 | 45 | 0 | 77 | 13 | 82 | 8 | 90 | 0 | 90 | 0 | 29 | 16 | 24 | 21 | 45 | 0 | 45 | 0 |
| F | 16 | 209 | 3 | 222 | 9 | 216 | 26 | 199 | 12 | 168 | 7 | 173 | 25 | 155 | 30 | 150 | 0 | 45 | 0 | 45 | 1 | 44 | 2 | 43 |

(c) Street

| | One speaker out of three users | | | | | | | | Two speakers out of three users | | | | | | | | Three speakers out of three users | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Absolute | | Relative | | Rhythm | | Proposed | | Absolute | | Relative | | Rhythm | | Proposed | | Absolute | | Relative | | Rhythm | | Proposed | |
| | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N |
| T | 45 | 0 | 45 | 0 | 24 | 19 | 45 | 0 | 82 | 8 | 90 | 0 | 56 | 34 | 89 | 1 | 39 | 6 | 26 | 19 | 36 | 9 | 45 | 0 |
| F | 30 | 195 | 4 | 221 | 1 | 226 | 29 | 196 | 15 | 165 | 4 | 176 | 0 | 180 | 16 | 164 | 0 | 45 | 0 | 45 | 0 | 45 | 0 | 45 |

(d) Car

| | One speaker out of three users | | | | | | | | Two speakers out of three users | | | | | | | | Three speakers out of three users | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Absolute | | Relative | | Rhythm | | Proposed | | Absolute | | Relative | | Rhythm | | Proposed | | Absolute | | Relative | | Rhythm | | Proposed | |
| | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N |
| T | 45 | 0 | 45 | 0 | 44 | 1 | 45 | 0 | 83 | 7 | 86 | 4 | 89 | 1 | 89 | 1 | 24 | 21 | 18 | 27 | 45 | 0 | 35 | 10 |
| F | 24 | 201 | 9 | 216 | 0 | 225 | 16 | 209 | 15 | 165 | 4 | 176 | 13 | 167 | 7 | 173 | 0 | 45 | 1 | 44 | 0 | 45 | 8 | 37 |

(e) Rain

| | One speaker out of three users | | | | | | | | Two speakers out of three users | | | | | | | | Three speakers out of three users | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Absolute | | Relative | | Rhythm | | Proposed | | Absolute | | Relative | | Rhythm | | Proposed | | Absolute | | Relative | | Rhythm | | Proposed | |
| | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N |
| T | 45 | 0 | 45 | 0 | 44 | 1 | 45 | 0 | 72 | 18 | 81 | 9 | 90 | 0 | 87 | 3 | 35 | 10 | 25 | 20 | 45 | 0 | 45 | 0 |
| F | 24 | 201 | 3 | 222 | 0 | 225 | 4 | 221 | 14 | 166 | 9 | 171 | 11 | 169 | 13 | 167 | 1 | 44 | 2 | 43 | 0 | 45 | 1 | 44 |

Table 3.7: F1-scores of short utterances

| Case | | Scheme | | | |
|---|---|---|---|---|---|
| # of users | # of speakers | Absolute | Relative | Rhythm | Proposed |
| 3 | 1 | 0.916 | **0.929** | 0.878 | **0.929** |
| | 2 | 0.775 | **0.960** | 0.857 | 0.878 |
| | 3 | 0.767 | 0.800 | **0.989** | 0.846 |

out of three users, though slightly lower F1-scores were observed in the all-speakers case, where the threshold occasionally misidentified a speaker as a non-speaker (True Negative).

### 3.4.2 Impact of Sound Pressure Sensors

The accuracy of the proposed scheme derives from two components: the architecture of the sound pressure sensor and its synchronization. In Sec. 2.4.1, the synchronization accuracy of SRP Badges is less than 30 µs, which seems one of the contribution for precise speaker identification. This section shows the impact of sound pressure sensor for precise speaker identification.

Figures 3.4 (a) and (b) display the circuit diagrams of a sound pressure sensor in the Rhythm

Table 3.8: Confusion matrices of short utterances

| | One speaker out of three users | | | | | | | | Two speakers out of three users | | | | | | | | Three speakers out of three users | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Absolute | | Relative | | Rhythm | | Proposed | | Absolute | | Relative | | Rhythm | | Proposed | | Absolute | | Relative | | Rhythm | | Proposed | |
| | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N | P | N |
| T | 16 | 29 | 22 | 23 | 43 | 2 | 22 | 23 | 81 | 9 | 82 | 8 | 90 | 0 | 86 | 4 | 13 | 32 | 20 | 25 | 45 | 0 | 21 | 24 |
| F | 0 | 225 | 0 | 225 | 10 | 215 | 0 | 225 | 15 | 165 | 0 | 180 | 30 | 150 | 20 | 160 | 1 | 44 | 0 | 45 | 1 | 44 | 0 | 45 |

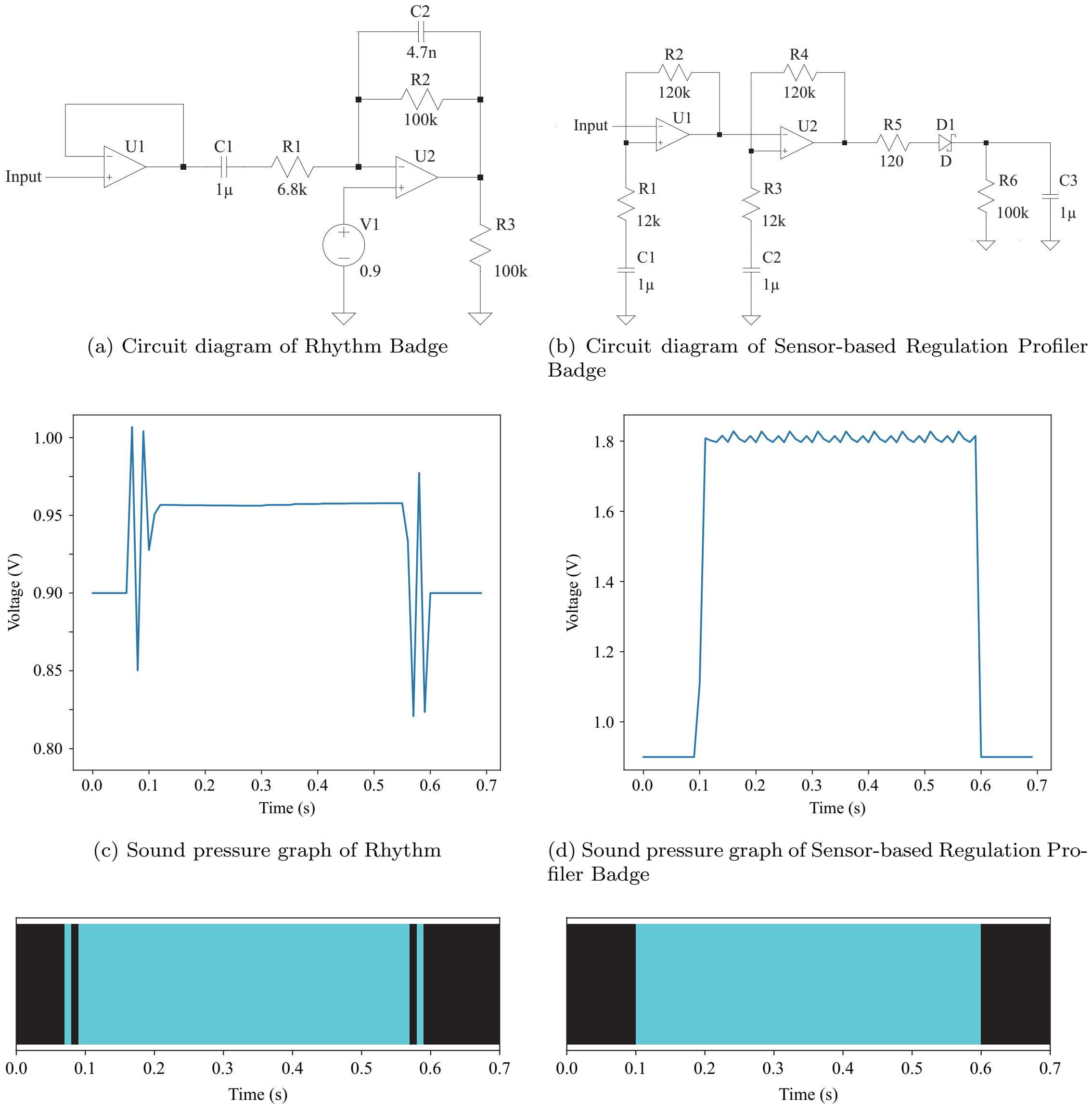
Badge and the Sensor-based Regulation Profiler Badge. The Rhythm Badge, based on Open Badge [9], utilizes an integration circuit, while the Sensor-based Regulation Profiler Badge employs a peak hold circuit for sound pressure acquisition. The parameters of the circuit in the Sensor-based Regulation Profiler Badge were chosen to achieve three objectives:

Eliminate low-frequency noise, specifically frequencies below 20 Hz, as they are unrelated to speech. Amplify the sound pressure data 100 times to capture detailed changes in voice volume. Precisely extract the beginning and end of each speech segment by adjusting the discharge slope of the resistor-capacitor (RC) circuit. Simulations were performed for each circuit. A sinusoidal wave was used to represent speech, with an amplitude of 0.8 V at 340 Hz and a duration of 500 ms. Additionally, a direct current (DC) signal with an amplitude of 0.9 V and a length of 100 ms was applied to simulate silence, placed before and after the sinusoidal wave.

Figures 3.4 (c) and (d) show the measured sound pressure as a function of time for both the Rhythm Badge and the Sensor-based Regulation Profiler Badge. The Rhythm Badge exhibits spikes at the start and end of the sound pressure measurement due to the integration circuit. In contrast, the Sensor-based Regulation Profiler Badge, using a peak hold circuit, avoids spikes in the measured sound pressure.

To evaluate the effect of the measured sound pressure data, a threshold-based speech detection algorithm was applied to both the Rhythm Badge and the business-card-type sensor. A threshold was set to detect the edges of speech segments. As shown in Fig.3.4 (c), the sound pressure in the Rhythm Badge before and after the spikes was approximately 0.90 V and 0.95 V. The threshold was set at 0.92 V to mitigate the effect of the spikes. In Fig.3.4 (d), the measured sound pressure in the proposed sensor scheme was 0.9 V before speech and 1.8 V after. A threshold between 0.9 V and 1.8 V yielded similar performance, so a threshold of 0.92 V was used in the proposed scheme as well.

Figures 3.4 (e) and (f) show the results of the threshold-based speech detection for the Rhythm Badge and the Sensor-based Regulation Profiler Badge. The Rhythm Badge struggled to accurately extract speech using threshold-based detection due to the presence of spikes in the measured sound pressure data. However, the Sensor-based Regulation Profiler Badge accurately detected speech since its measured sound pressure lacked spikes. The results in Figs.3.4 (e) and (f) indicate that the peak hold circuit detects speech more accurately than the integration circuit used in the Rhythm Badge.



(a) Circuit diagram of Rhythm Badge

(b) Circuit diagram of Sensor-based Regulation Profiler Badge

(c) Sound pressure graph of Rhythm

(d) Sound pressure graph of Sensor-based Regulation Profiler Badge

(e) Speech section of Rhythm

(f) Speech section of Sensor-based Regulation Profiler Badge

Figure 3.4: Simulation results for speech detection in Rhythm and Sensor-based Regulation Profiler Badge.

### 3.4.3 Influence of Synchronization Accuracy

An evaluation experiment using the SRP Badge was conducted to assess the relationship between its time synchronization accuracy and speaker detection algorithm performance. The experimental environment was identical to that described in Sec. 3.4.1. For this evaluation, pseudo data with varying levels of time synchronization accuracy were generated by adding random values to the timestamps of sound pressure data acquired by the SRP Badge. Since the distribution of time synchronization errors shown in Fig. 2.14 resembles a normal distribution, normally distributed

Table 3.9: Confusion matrix of speaker detection accuracy at each time synchronization error

| | 0 ms | | 0.1 ms | | 1 ms | | 10 ms | |
|---|---|---|---|---|---|---|---|---|
| | P | N | P | N | P | N | P | N |
| T | 440 | 10 | 440 | 10 | 440 | 10 | 440 | 10 |
| F | 0 | 450 | 0 | 450 | 98 | 352 | 192 | 258 |

random values were used for this evaluation. The range of random values was determined by adjusting the standard deviation of the normal distribution. For example, when the standard deviation $\sigma = 0.5\,\text{ms}$, approximately 95 % of the random values fall within the range of $\pm 2\sigma = \pm 1.0\,\text{ms}$. In this evaluation, $\pm 2\sigma$ was treated as the time synchronization error.

Table 3.9 shows the confusion matrices of speaker detection accuracy for each synchronization error. From Table 3.9, it can be observed that as the time synchronization error increases, the misdetection of speech during non-speech intervals also increases. To verify the statistical significance of these results, a t-test was conducted to compare the mean differences for each time synchronization error. Given a sample size of 10, degrees of freedom of 18, and a significance level of 5 %, no significant difference was observed between time synchronization errors of 0 ms and 0.1 ms. However, a significant difference was observed between 0 ms and 1 ms. Based on these results, it can be concluded that time synchronization accuracy within 1 ms between SRP Badges is necessary to ensure the high speaker detection accuracy of the SRP Badge.

## 3.5 Conclusion

This chapter introduced an innovative sound pressure sensor along with a speaker identification algorithm designed for business-card-sized sensors, aimed at analyzing collaborative dynamics in multi-person activities. The sound pressure sensor incorporates a peak hold circuit and a time synchronization module, which help reduce spikes and ensure precise synchronization between sensors, enabling accurate and low-cost detection of user speech. The algorithm effectively filters out background noise from non-speaker sensors, achieving high accuracy in identifying the speaker. The evaluation demonstrated the proposed method's efficiency across various conditions, including different user numbers, background noise levels, and both long and short speech durations. Furthermore, the peak hold circuit reliably captures user speech, and the synchronization error between sensors consistently remains within ±30 µs.

# Chapter 4

# Indoor Localization on Mobile Devices

## 4.1 Introduction

The adoption of IoT systems for multimodal collaboration analysis is expanding. Multimodal analysis has become critical for detailed descriptions of individuals and their environments. In traditional multimodal collaboration analysis, experts often analyzed collaboration environments by placing video cameras and microphones to record activities, then reviewing the recordings for insights. This approach has historically incurred significant human and time costs due to its reliance on manual effort. With the proliferation of IoT systems, these costs are expected to decrease, potentially accelerating the adoption of multimodal collaboration analysis.

One modality for collaboration analysis is the posture of each user [11,95,96]. For example, the literature [96] captures learners' posture data to analyze its impact on the quality of collaborative learning in augmented reality environments. To collect posture data, static cameras are typically installed in the collaboration environment. Using the captured visual data, posture information is extracted through joint detection using computer vision algorithms. However, such technologies involve high setup costs in practical collaboration scenarios to accommodate user mobility and address occlusions, rendering them unsuitable for effectively supporting collaboration analysis.

To address these limitations in practical scenarios, tag-based motion capture offers a promising solution. Unlike traditional systems, this approach utilizes several small and lightweight tags worn by each user, offering robust performance in practical collaboration scenarios. Tags provide a more robust and practical solution for motion capture in multimodal collaboration analysis, particularly in dynamic or crowded scenarios.

In the context of collaboration analysis, tag-based motion capture requires indoor localization for each tag to meet two key requirements: centimeter-level accuracy and low-cost anchor setup. Centimeter-level accuracy is essential for accurately capturing and analyzing learners' postures.

Low-cost anchor setup is required to support collaboration analysis without introducing additional costs for the environment. Although various indoor localization methods such as acoustic, infrared, WiFi/BLE, and RFID are available as discussed in Sec. 4.5, none of these methods simultaneously satisfy both requirements.

Visual-Inertial Odometry (VIO) is a potential localization scheme to meet these requirements. VIO complementarily integrates sensor data from cameras, LiDAR, and inertial measurement units (IMUs), enabling centimeter-level localization in indoor scenarios through a straightforward setup process. By utilizing miniaturized visual and inertial sensors, each tag is expected to achieve precise localization for motion capture in multimodal collaboration analysis.

However, the challenges of VIO-based localization in practical environments remain unclear. Various situations can arise in the context of real-world collaboration scenarios. For example, some collaboration environments may be monochromatic with limited visual features. In addition, collaboration environments utilizing projectors may involve dim lighting conditions or flickering light sources. Whether VIO consistently achieves precise localization in practical environments, including the aforementioned examples, has not been sufficiently evaluated.

This study aims to comprehensively evaluate the effectiveness of current VIO systems in practical environments, addressing the research question of their applicability to real-world collaboration analysis. Specifically, the research involves conducting case studies simulating practical environments to verify the positioning accuracy of VIO. Controlled experiments are then performed to identify the practical challenges of VIO observed in the case studies.

This study also poses a prototype solution with Ultra Wide Band (UWB) to address these challenges. The study demonstrates how the proposed approach enhances the robustness of indoor positioning in practical environments.

## 4.2 Case Study

### 4.2.1 Experiment Workflow

To evaluate the practicality of VIO, this study conducted a user-driven case study with a mobile device. As a mobile device, iPhone 12 Pro was adopted in this study. A commercial Augmented Reality (AR) application, which highly relies on VIO to overlay virtual elements onto the real world, was also adopted for the evaluation. Figure 4.1 shows the overview of the application. The application is called AR Visual [97], which virtually simulates furniture on the environment. This study set a task: classroom setup with the application. The primary objective is to find a table and chair arrangement, which maximizes the student capacity while ensuring the accessibility and clear screen visibility from all the seats. Each user deployed virtual tables and chairs on an empty classroom and checked the whole layout from different vantages. This study gathered 17

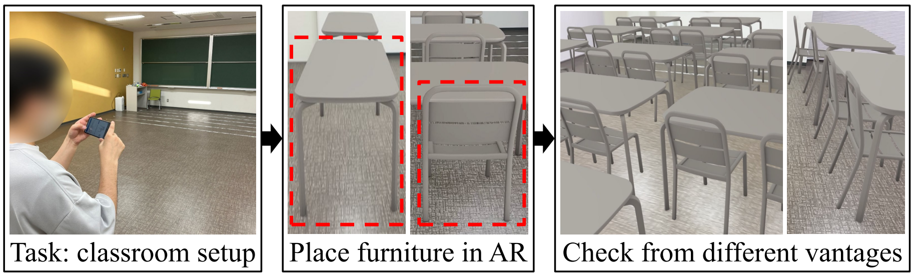


Figure 4.1: The experimental procedure of case study.

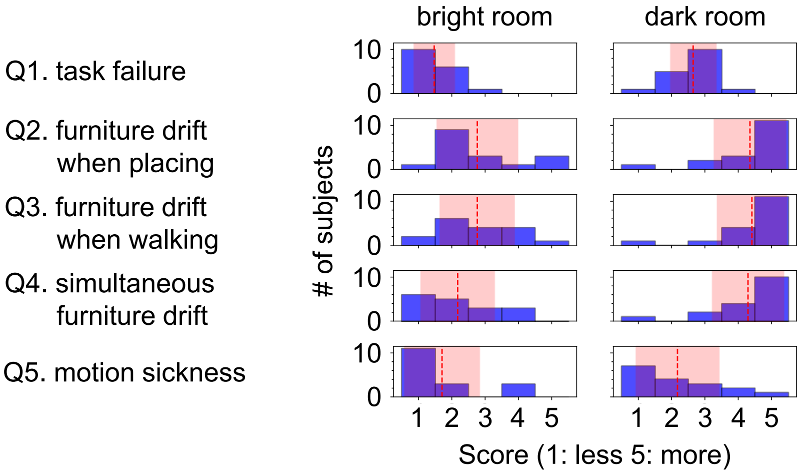


Figure 4.2: The user responses in the case study.

subjects aged between teens and forties in Osaka University and University of California, San Diego. To further examine the effect of lighting conditions on localization performance, this study also conducted the experiment under two lighting conditions: normal room (500 lux) and dark room (2.5 lux). After the simulation, each subject was interviews with a questionnaire shown in Table 1.

### 4.2.2 Case Study Analysis

Fig. 4.2 presents the participants' responses, displaying average scores and interquartile ranges for each question. The results show that task completion was rated lower in the dark room, which is likely due to increased drift, as indicated by responses related to furniture misalignment and global drift. The dark room, with fewer visual features, led to more tracking failures. Moreover, some participants reported motion sickness in the dark room, likely due to greater drift of the virtual elements.

From a quantitative perspective, the application reduced setup time. Arranging physical tables and chairs in a bright room took 16 minutes and 13 seconds, while using AR shortened the time to 12 minutes and 31 seconds in the bright room and 9 minutes and 58 seconds in the dark room.

However, the drifting error for the front table averaged 30 cm in the bright room and 50 cm in the dark room, which impacted task completion and comfort.

### 4.2.3 Takeaways

The case study revealed several barriers to the VIO-based localization in practical scenarios, particularly in complex environments. The first is global drift of virtual elements. Virtual tables and chairs exhibited a tendency to drift over time, with average drift distances of 30 cm in the bright room and 50 cm in the dark room. The second is tracking failures. AR systems experienced difficulty tracking virtual objects in low-light conditions or environments lacking visual features, resulting in glitches. Finally, such localization and tracking failures cause motion sickness for each subject, especially in darker environments, due to inconsistencies in user perception.

Similar challenges were observed in other AR applications, such as AnywheRe [98] and ARvid [99], Augment [100], COCOAR [101], Measure [102], and Monster Park [103]. These issues are largely due to reliance on visual-based sensing systems (monocular cameras and LiDARs). Further analysis of these failure modes is provided in Sec. 4.3, where detailed experiments quantify the extent of errors and provide a root-cause analysis of the problems encountered.

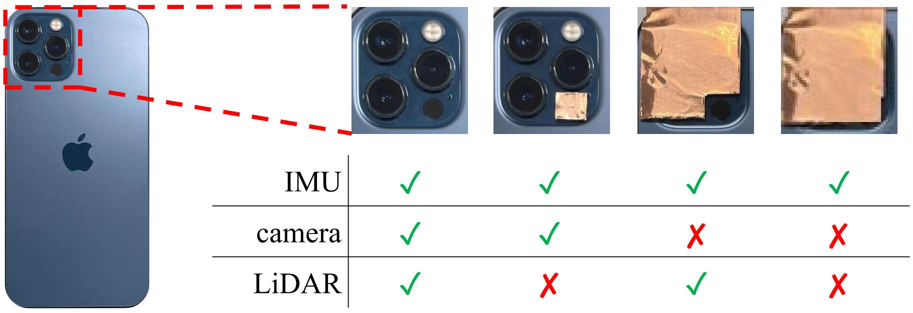
## 4.3 Controlled Experiment

This section conducts controlled experiments to quantitatively identify failure factors of VIO-based localization and tracking in Sec. 4.2. To delve into the factors, this section extracted variables in the case study environment: when using a smartphone integrated with an IMU, cameras, and LiDAR, across environments with varying complexity, under diverse lighting conditions, and at different motion speeds.

### 4.3.1 Control Variables

**Sensors**: Most smartphones utilize monocular cameras to enable VIO for tracking and localization, while some modern devices also include time-of-flight sensors or LiDAR for enhanced functionality. This study analyzed how different sensor combinations influence tracking and localization accuracy. For this purpose, this study selected the iPhone 12 Pro, which features these advanced sensors. As shown in Fig. 4.3, this study applied copper foil tape to create four configurations: *IMU + camera + LiDAR*, *IMU + camera*, *IMU + LiDAR*, and *IMU* only.

**Environment complexity**: The number of feature points captured by the camera and LiDAR varies with the complexity of the environment. To evaluate the influence, this study tested three types of environments: *wall*, *shelf corner*, and *crowded*, as depicted in Fig. 4.6. Figures 4.5 (a, b) illustrate the changes in the number of feature points detected by the camera, using OpenCV [104], and by LiDAR, using LOAM [105], as the smartphone moves back and forth on an xy-stage. In the



| | | | | |
|---|---|---|---|---|
| IMU | ✓ | ✓ | ✓ | ✓ |
| camera | ✓ | ✓ | ✗ | ✗ |
| LiDAR | ✓ | ✗ | ✓ | ✗ |

Figure 4.3: Sensor combination.

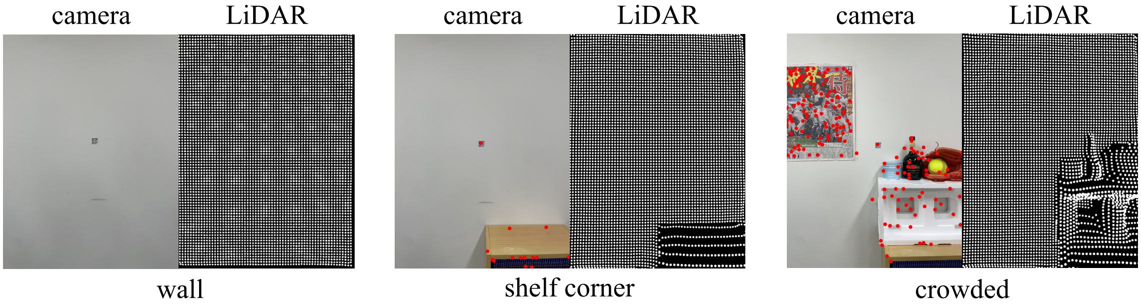
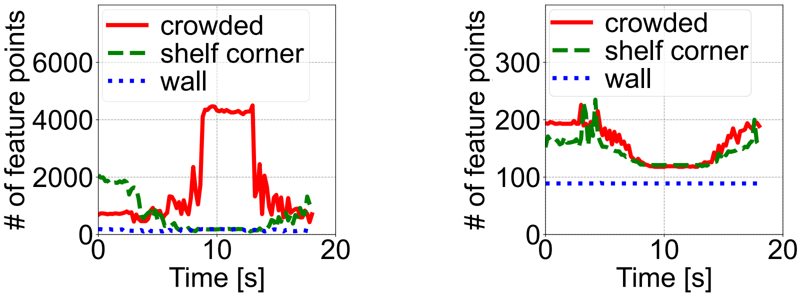


Figure 4.4: Visual complexity.



(a) Camera feature (b) LiDAR feature

Figure 4.5: The extracted feature points.

*wall* environment, the number of feature points remains stable for both the camera and LiDAR. However, in the *shelf corner* and *crowded* environments, significant fluctuations occur based on the objects visible in the camera's field of view. Additionally, the trends in feature point counts between the camera and LiDAR do not always correspond. For example, at around the 10-second mark in a *crowded* environment, a wall covered with newspapers leads to a reduction in LiDAR-detected feature points, while the camera simultaneously detects an increased number of feature points. This highlights discrepancies between the two sensors in capturing feature points.

**Brightness**: Localization accuracy with a camera is influenced by the level of brightness. To examine the influence, the study conducted experiments under three distinct lighting conditions. Using an Urceri MT-912 light meter, which offers an accuracy of ±3 % of the reading plus ±8

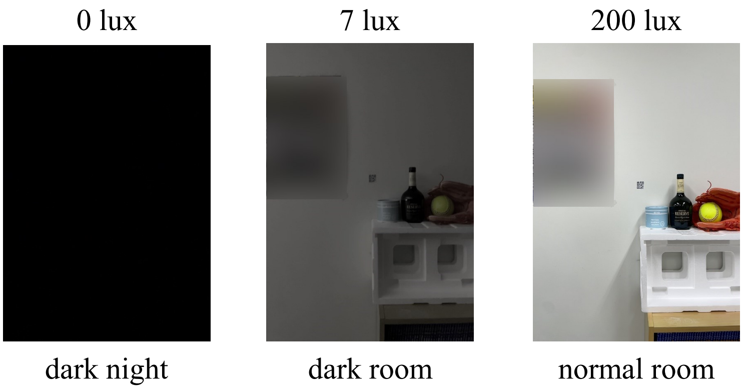
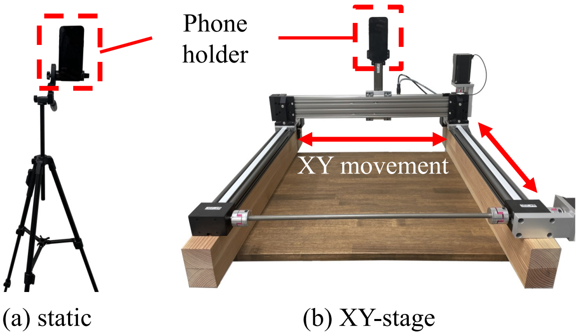


Figure 4.6: Lighting conditions.



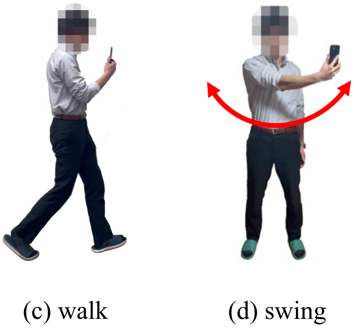


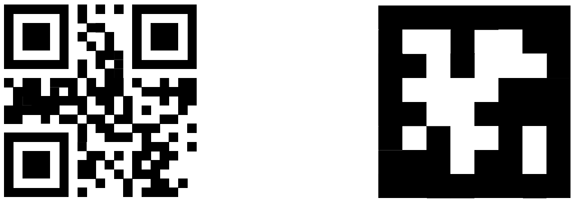
Figure 4.7: Movement types.

digits on the least significant digit and a resolution of 0.1 lux up to 1000 lux, this study measured the illuminance. This study created three scenarios: *dark night* (0 lux), *dark room* (7 lux), and *normal room* (200 lux), as illustrated in Fig. 4.6. In the *crowded* environment, the camera detected 0 feature points under *dark night* conditions, 167 feature points in the *dark room*, and 222 feature points in the *normal room*.

**Movement**: This study evaluated four types of motion patterns, as depicted in Figures 4.7. In *static*, the smartphone remains completely still, as shown in Fig. 4.7 (a). In *xy-stage*, the device moves back and forth over a distance of 40 cm at speeds of up to 0.08 m/s, as illustrated in Fig. 4.7 (b). In *walk*, a person walks around the room, carrying the smartphone, at speeds of up to 1 m/s, as shown in Fig. 4.7 (c). In *swing*, the smartphone is waved back and forth with a motion speed of up to 3 m/s, as depicted in Fig. 4.7 (d).

### 4.3.2 Localization: Distance

**Experimental settings**: As outlined in 4.1, smartphones can utilize landmark QR codes or AprilTags to determine their position within global coordinates in environments lacking GPS. To evaluate the accuracy and reliability of this approach, this study conducted a quantitative analysis focusing on two key metrics: the detection distance for QR codes and AprilTags and the initial

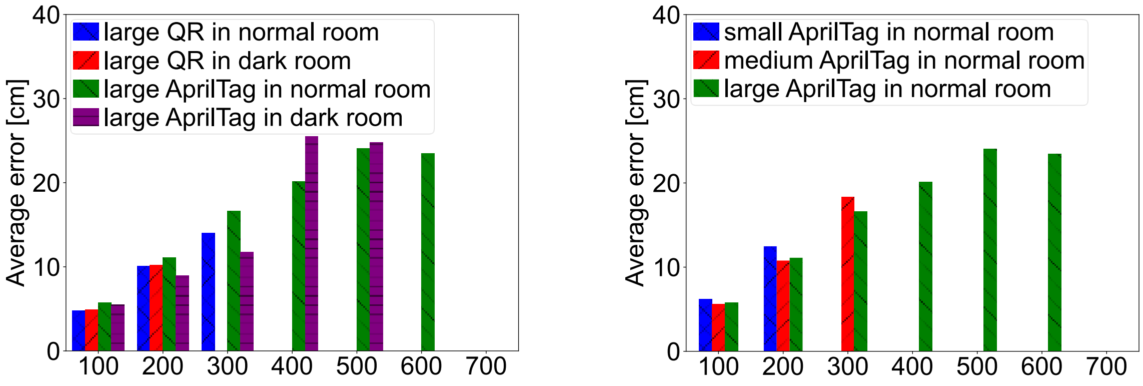


(a) QR code (b) AprilTag

Figure 4.8: Landmark types.

positional error when these landmarks are detected. Fig. 4.8 illustrates the landmark patterns used in the study, featuring QR codes and AprilTags in three sizes: *small* (3 cm × 3 cm), *medium* (6 cm × 6 cm), and *large* (9 cm × 9 cm). The evaluation procedure involves two device movements: first, rotating the device along the z-axis by moving it left and right, and second, rotating it along the x-axis by moving it up and down. These actions enable the smartphone's VIO to effectively map the surrounding space and integrate the physical environment with the virtual AR environment. After spatial mapping, the smartphone attempts to detect and localize itself using a QR code or AprilTag from a specific distance. Finally, the perceived location in the AR coordinate system is compared with the smartphone's actual physical location to assess accuracy.

**Results**: Fig. 4.9 illustrates the relationship between the distance from a QR code or AprilTag to a smartphone (x-axis) and the localization error, represented as the discrepancy between the actual and perceived AR space locations (y-axis). Notably, bars are absent in instances where the landmark was undetectable. A key finding emerged: the localization error depends solely on the distance between the smartphone and the landmark, irrespective of ambient lighting, the type of landmark, or its size. In Fig. 4.9 (a), the error for both QR codes and AprilTags increases linearly with distance, unaffected by brightness or landmark type. Similarly, Fig. 4.9 (b) demonstrates that localization error trends remain consistent across different sizes of AprilTags, underscoring that factors like lighting, type, and size have no impact on the error at a given distance. However, these factors did influence the detectable range of the landmarks. For instance, in Fig. 4.9 (a), a *large* QR code was readable up to 300 cm in a *normal room*, but this range decreased to 200 cm in a *dark room.* QR codes also exhibited shorter detectable distances compared to equivalently sized AprilTags, as their smaller pixel structure accommodates more data but reduces readability. Additionally, Fig. 4.9 (b) shows that larger landmarks allowed greater detectable ranges, with readability improving as the landmark size increased. This significant observation highlights that localization error is solely dependent on distance, providing valuable insights for optimizing AR systems. It emphasizes the critical limitation of proximity in achieving accuracy with optical AR systems.



(a) Landmark & brightness

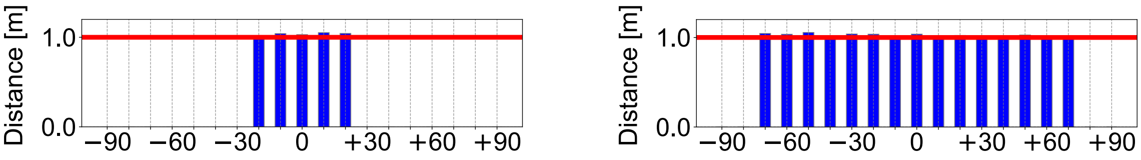
(b) Landmark size

Figure 4.9: Localization accuracy with different detection distance.

### 4.3.3 Localization: Angle

**Experimental settings**: In exploring the constraints and strengths of QR code-based and AprilTag-based landmark detection for smartphone VIO, this study hypothesized that a landmark's readability depends not only on the distance from the scanner but also on the angle of approach. To test the influence, this study mounted 3 cm × 3 cm QR code and AprilTag landmarks on a wall and conducted evaluations at a fixed distance of 1 m. The angle of approach was varied from $-90°$ to $+90°$ in 10-degree increments, with 0° representing a direct, perpendicular approach to the landmark. The experimental conditions were consistent with those outlined in Sec. 4.3.1, ensuring that other variables remained constant and did not influence the results. This setup allowed for a focused assessment of how the angle of approach affects landmark detection.

**Results**: The experimental results, depicted in Fig. 4.10, highlight a significant limitation regarding the angle at which a smartphone's camera can reliably detect QR codes and AprilTags in AR applications. In Figures 4.10 (a, b), the horizontal axis represents the angle of the smartphone relative to the landmark, while the vertical axis shows the recognition distance, with accurate readings indicated at 1 m. Areas without bars denote instances where the landmark was undetectable. The findings reveal that QR codes have a notably narrow detection angle, limited to approximately 50° from the front, due to their high pixel density. Within this range, the detection distance remains consistent at 1 m. However, this restricted angle poses challenges, particularly for dynamic AR interactions where users may not consistently approach landmarks from ideal perspectives. These limitations have significant implications for AR applications. A narrow detection angle reduces the flexibility and usability of landmark-based systems, potentially diminishing user engagement and immersion. Aligning a smartphone precisely within the required detection angle can be challenging for users, leading to frustration and interruptions in the experience. This issue is particularly critical in scenarios like navigation, education, or interactive environments with dense or complex



(a) QR code

(b) AprilTag

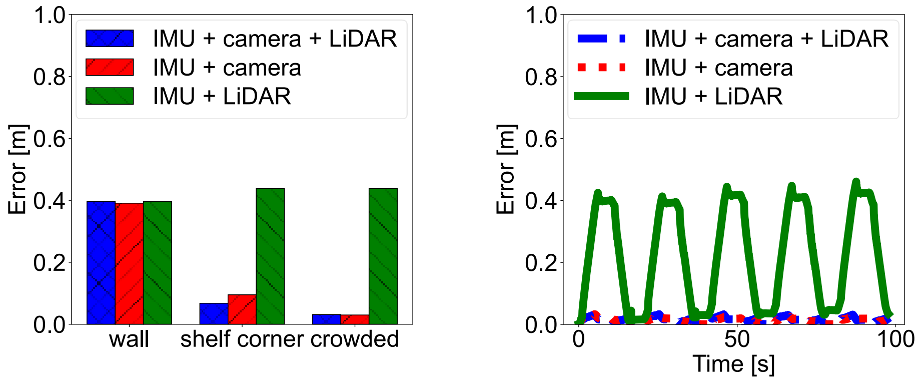
Figure 4.10: Localization accuracy with different detection angle.

layouts, where ease of interaction is essential for a seamless user experience.

### 4.3.4 Tracking: Environment Complexity

**Experimental settings**: This section explores the challenges environmental complexity presents to tracking performance in VIO systems. The intricacy of an environment plays a crucial role in determining the effectiveness of different sensing technologies used for spatial recognition and mapping. For example, environments with distinct color variations benefit camera-based systems by providing numerous feature points essential for accurate tracking. On the other hand, environments with pronounced physical irregularities, such as bumps and ridges, are better suited for LiDAR systems, which depend on surface variations for precise mapping. To investigate these effects, this experiments systematically varied environmental complexity and sensor configurations to assess their combined impact on tracking accuracy. The tests were conducted under controlled conditions simulating *normal room* lighting to ensure practical applicability. Movement was simulated using an xy-stage, allowing for consistent and reproducible evaluations of how environmental features influence VIO system performance. This methodology provided valuable insights into optimizing VIO systems by examining the interactions between different sensors and environmental characteristics, paving the way for enhanced performance across diverse application scenarios.

**Results**: Fig. 4.11 (a) presents the 99th percentile localization error observed in our experiments. While LiDAR can enhance localization accuracy when combined with cameras, it is unable to independently track the environment when using the iPhone 12 Pro. Specifically, as shown in Fig. 4.11 (a), LiDAR alone struggles to track effectively across environments such as *wall*, *shelf corner*, and *crowded* scenes. In feature-rich environments like *crowded* settings, the *IMU + camera* configuration achieves high localization accuracy, and the addition of LiDAR does not degrade performance. This suggests that LiDAR complements camera data under suitable conditions. However, in feature-sparse environments such as *shelf corner*, integrating LiDAR with the IMU and camera improves localization accuracy compared to using only *IMU + camera*. These results demonstrate LiDAR's ability to enhance depth perception and feature detection where camera-based systems are less effective. Despite its benefits, the *IMU + LiDAR* configuration performed poorly, highlighting the critical importance of sensor fusion in achieving precise localization. The findings underline the nuanced role of LiDAR in VIO systems, showcasing its potential to enhance



(a) Environment

(b) *Crowded* timeline

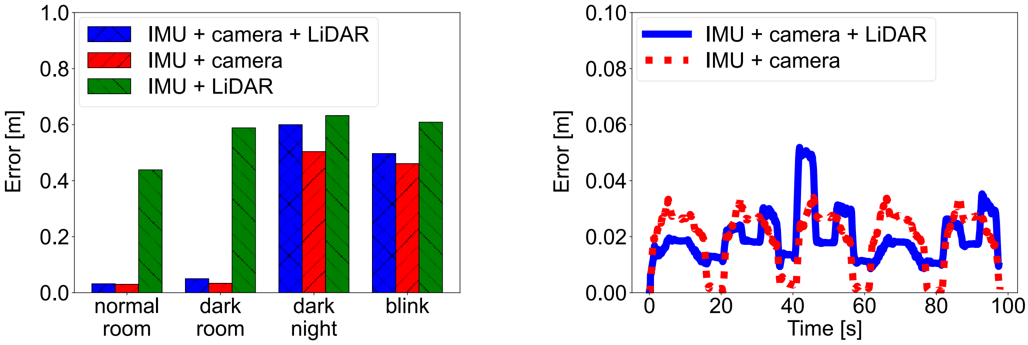
Figure 4.11: Tracking accuracy in different visual complexity.

localization accuracy when used alongside cameras, while also emphasizing its limitations as a standalone input. To further explore the results, Fig. 4.11 (b) examines location tracking errors under the *crowded* scene. The figure plots localization errors over time, with the x-axis representing time in seconds and the y-axis denoting localization error. In this experiment, a smartphone mounted on an xy-stage moved back and forth over a distance of 40 cm for 20 seconds. At the starting point, all sensor configurations began with zero localization error. As the smartphone moved away, both the *IMU + camera* and *IMU + camera + LiDAR* configurations showed increasing errors, which diminished upon returning to the starting point, indicating effective error correction. In contrast, the *IMU + LiDAR* setup exhibited a gradual increase in error over time, even as the device returned to the start. This drift was caused by IMU inaccuracies, leading to a slight misalignment over time and a failure to accurately track movement. The *IMU + LiDAR* configuration notably perceived the device as stationary despite actual motion. This phenomenon, driven by IMU drift, is further analyzed in Sec. 4.3.7, emphasizing the limitations of LiDAR-IMU configurations without the aid of a camera for reliable tracking.

### 4.3.5 Tracking: Brightness

**Experimental settings**: Fig. 4.11 (a) revealed that in complex environments, the combination of a camera and IMU enables high-precision tracking without relying on LiDAR. This finding highlights the adaptability of VIO systems in intricate settings by leveraging the complementary strengths of cameras and IMUs. However, a well-known limitation of camera-based systems is their sensitivity to changes in lighting conditions, which can significantly impact the detection and reliability of feature points, posing challenges for consistent tracking accuracy across varying lighting environments. On the other hand, LiDAR offers a distinct advantage with its immunity to ambient light variations. This stability ensures that LiDAR can provide reliable data even in conditions where camera-based systems falter. To thoroughly evaluate these differences, this study extended our experiments using the *crowded* complexity environment described in Sec. 4.3.4, introducing deliberate variations in lighting to assess their effects. The lighting conditions tested included *dark night* (0 lux), *dark room* (7 lux), and *normal room* (200 lux), as shown in Fig. 4.6. Additionally, a *blink* condition was introduced, where lighting alternated between 0 lux and 200 lux every 3 seconds. This approach allowed to systematically evaluate the adaptability and limitations of camera and IMU configurations under varying lighting conditions, while also assessing the robustness of LiDAR-based data acquisition in scenarios that challenge camera-based systems.

**Results**: The investigation under varying lighting conditions uncovered a surprising and significant finding regarding performance under *blink* conditions, where illumination alternates between darkness and normal room lighting. As shown in Fig. 4.12 (a), which presents the 99th percentile positional error across experiments, all sensor combinations experienced increased error rates as the environment became darker. Notably, the degradation in accuracy under *blink* conditions was as severe as that observed in the *dark night* scenario. Although *blink* conditions periodically provided the same illumination level as the *normal room*, even brief periods of darkness significantly impaired positional accuracy. This effect was evident even when using the combined *IMU + camera + LiDAR* configuration, highlighting that the inability to consistently capture feature points, even momentarily, can substantially compromise tracking precision. Interestingly, further analysis revealed that incorporating LiDAR could, in some cases, result in poorer accuracy than using only *IMU + camera*. For instance, in the *dark room* scenario, the *IMU + camera + LiDAR* setup performed worse than *IMU + camera* alone. This counterintuitive result suggests that under certain complex conditions, LiDAR's additional data does not always enhance performance and may even degrade it. The degradation likely stems from both hardware and software factors: low light reduces the number of camera-detectable feature points, making it difficult to integrate LiDAR point clouds effectively with visual data. To better understand these limitations, this study conducted a detailed time-series analysis of positional accuracy in the *dark room* scenario. As illustrated in 4.12 (b), with time on the x-axis and positional error on the y-axis, configurations using *IMU + camera + LiDAR* exhibited greater fluctuations (or jitter) compared to those using *IMU + camera*. This increased jitter indicates challenges in matching camera-detected feature points with LiDAR-acquired depth data under low-light conditions. Insufficient lighting introduces errors in aligning visual feature points with corresponding LiDAR point clouds, exacerbating positional inaccuracies. This finding underscores a critical challenge in fusing LiDAR and visual data in sub-optimal lighting. It highlights the need for improved algorithms or methodologies capable of more robustly integrating disparate sensor inputs, particularly in environments with limited or variable lighting.



(a) Brightness

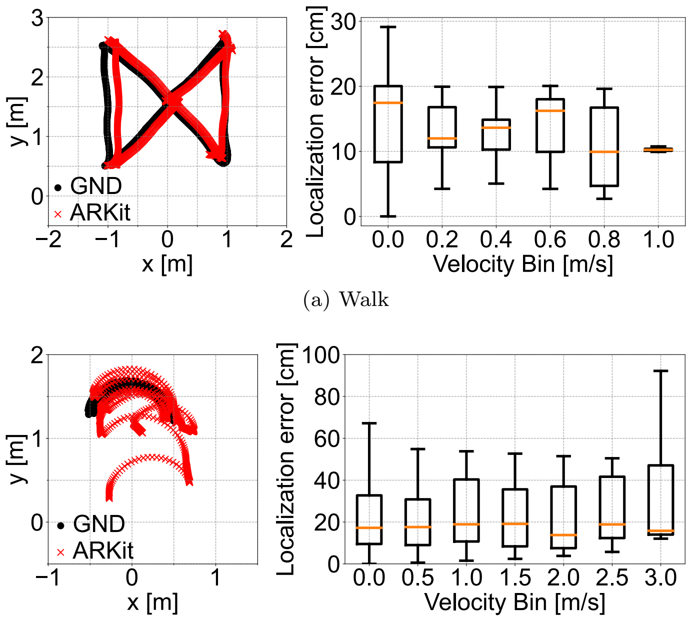
(b) *Dark room* timeline

Figure 4.12: Tracking accuracy under different lighting conditions.

### 4.3.6 Tracking: Movement

**Experimental settings**: In Sec. 4.3.5, this study observed that the system incorrectly inferred the smartphone to be stationary in the *dark night* environment. This raised questions about the IMU's sensitivity to detect movement, prompting us to hypothesize that the speed of the xy-stage, set at 8 cm/s, may have been too low to trigger motion detection by the IMU. To investigate further, this study conducted additional experiments using two distinct motion patterns: a walking motion at speeds of up to 1 m/s and a dynamic swinging motion reaching up to 3 m/s, both relying solely on the IMU for movement detection. Ground truth data for these experiments was obtained using the HTC VIVE Tracker, a well-established tracking system [106]. This approach enabled us to evaluate the IMU's performance across different motion intensities and address potential limitations in its ability to detect movement at lower speeds.

**Results**: One of the key findings from this research on smartphone-based VIO systems is the limited capability of the IMU to accurately track movement across various speeds. As shown in Fig. 4.13 (a), which plots speed against positional tracking error, the IMU performs well for walking movements (*walk*), capturing general trends despite minor errors. However, as illustrated in Fig. 4.13 (b), the IMU struggles significantly with more dynamic movements such as swinging (*swing*), where errors are much larger compared to *walk*. Additionally, Sections 4.3.4 and 4.3.5 revealed that the IMU failed to detect slower movements, such as those performed on the xy-stage at a speed of 8 cm/s. Despite these limitations, the iPhone 12 Pro's VIO system was capable of achieving positional tracking at walking speeds of approximately 1 m/s, even without the support of a camera or LiDAR. These findings underscore the nuanced strengths and weaknesses of IMU-based tracking in VIO systems, highlighting its ability to handle moderate-speed movements while revealing challenges with both slow and highly dynamic motions.



(a) Walk

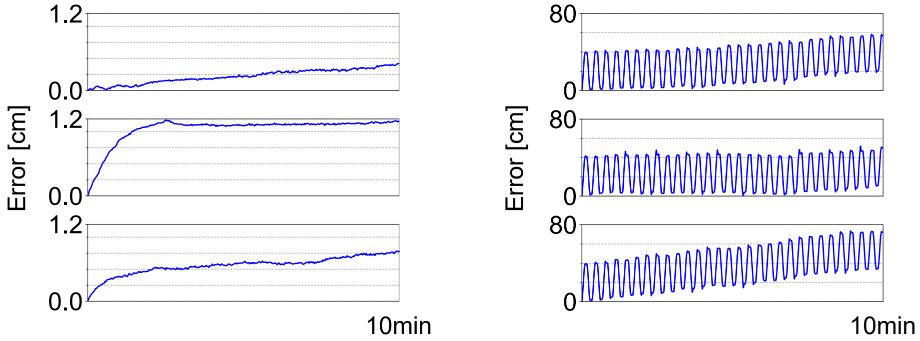
(b) Swing

Figure 4.13: The tracking results under different movement.

### 4.3.7 Microbenchmark: IMU's Drift

**Experimental settings**: It is well-documented in inertial navigation and AR technologies that IMUs are prone to drift [107, 108], a phenomenon that can significantly degrade location-tracking accuracy over time. In Sec. 4.3.4, this study attributed the widening error over time in the *IMU + LiDAR* configuration, as observed in Fig. 4.11 (b), to IMU drift. This section systematically investigated the occurrence and impact of IMU drift within our experimental setup, focusing on its effect on positional accuracy under varying motion conditions. This study conducted evaluations in two distinct scenarios. In *static*, the device was kept stationary to assess drift in the absence of movement. In *xy-stage*, the device underwent continuous back-and-forth motion at a controlled speed of 8 cm/s. Each scenario was evaluated over a duration of 10 minutes to capture the progression of drift and its cumulative effects. To ensure the reliability of the results, each condition was repeated five times. This thorough analysis provides a quantitative understanding of IMU drift in both static and dynamic contexts, offering valuable insights into its impact on VIO system performance for smartphone applications. These findings are instrumental in identifying strategies to mitigate drift, thereby improving the robustness of location tracking in real-world scenarios.

**Results**: Fig. 4.14 illustrates positional error over time, with the horizontal axis representing time and the vertical axis indicating the magnitude of localization error. Figures 4.14 (a, b) reveal that drift is present under both *static* and *xy-stage* (dynamic) conditions, highlighting a persistent challenge for VIO systems. A key observation is the variability of drift rates across different trials, emphasizing the unpredictable nature of this phenomenon. Drift is notably more pronounced in the



(a) *static* (b) *xy-stage*

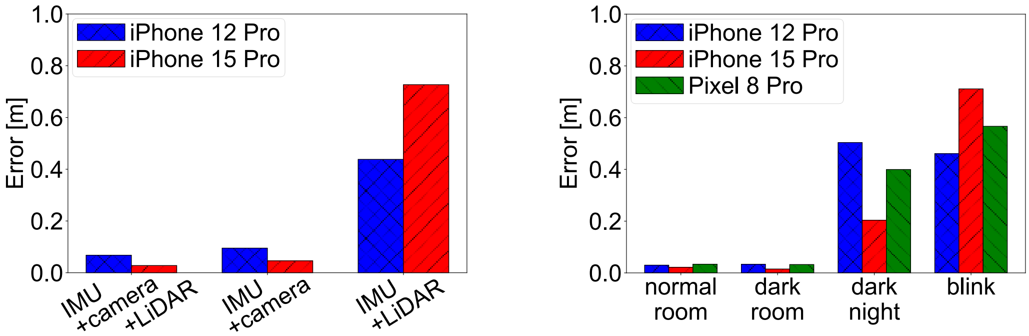
Figure 4.14: Experimental results of IMU's drift.

dynamic *xy-stage* condition compared to the *static* setup. In the *static* scenario, the largest error recorded over 10 minutes was approximately 1 cm—a relatively minor issue but still potentially impactful for applications requiring high precision. In contrast, the *xy-stage* scenario exhibited errors as large as 20 cm over the same period, posing a significant challenge for accurate positional tracking in dynamic environments. These findings have important implications for tracking. In *static* scenarios, a 1 cm error is unlikely to disrupt most navigation or interaction tasks. However, in dynamic conditions akin to the *xy-stage* setup, drifts of up to 20 cm can severely impair user experiences by misaligning virtual objects and disrupting spatial interactions. Such inaccuracies undermine the immersive quality of tracking and could limit their effectiveness in use cases requiring precise spatial awareness. Addressing IMU drift is therefore critical to improving tracking reliability and user satisfaction.

### 4.3.8 Microbenchmark: Various Smartphones

**Experimental settings**: To examine the dependency of tracking performance on different devices, this study conducted experiments using various smartphones in environments with the visual complexity and brightness levels described in Sections 4.3.4 and 4.3.5. In addition to the iPhone 12 Pro used in previous sections, this study included the iPhone 15 Pro and Google Pixel 8 Pro as representatives of the latest models from iPhone and Android (as of August 2024). Since the Pixel 8 Pro lacks LiDAR, its tracking accuracy was evaluated using only the *IMU + camera* configuration. For the iPhone 15 Pro, this study performed tracking experiments in the *shelf corner* environment to compare its performance trends with the iPhone 12 Pro. Additionally, tracking was tested with both the iPhone 15 Pro and Pixel 8 Pro under varying brightness conditions. The experimental settings and procedures for these tests adhered to the methodologies outlined in Sections 4.3.4 and 4.3.5, ensuring consistency and reliability across all device comparisons.

**Results**: Fig. 4.15 (a) illustrates the 99th percentile positional error for various smartphones in the *shelf corner* environment. The results indicate that the iPhone 15 Pro exhibits the same



(a) *Shelf corner*

(b) Brightness

Figure 4.15: Tracking accuracy with different smartphones.

general tracking trend as the iPhone 12 Pro. As highlighted in Sec. 4.3.4, the figure reaffirms that LiDAR alone is insufficient for effective tracking but demonstrates its capability to enhance tracking accuracy when integrated with camera and IMU data. A notable finding is the hardware advantage of the iPhone 15 Pro. Its higher-spec camera and LiDAR contribute to improved tracking accuracy compared to the iPhone 12 Pro, particularly in the *IMU + camera + LiDAR* and *IMU + camera* configurations. Fig. 4.15 (b) presents the 99th percentile positional error under varying brightness levels across the iPhone 12 Pro, iPhone 15 Pro, and Pixel 8 Pro. The results show a consistent trend among all devices, with tracking accuracy deteriorating as brightness decreases from 200 lux (*normal room*) to 0 lux (*dark night*). As noted in Sec. 4.3.5, the *blink* condition — where lighting alternates between 0 lux and 200 lux — failed to provide sufficient visual cues to improve tracking accuracy for any device, underscoring the limitations of momentary visual information in enhancing positional precision.

## 4.4 Potential Solution

In Sec. 4.3, controlled experiments allowed to uncover various failure modes, shedding light on the challenges associated with using current smartphone technologies for accurate localization and tracking. These experiments highlighted fundamental shortcomings in existing frameworks, emphasizing the need for more reliable and precise solutions to improve both user experience and application performance. This section shifts the focus towards addressing these challenges by exploring the use of UWB technology, which is now a standard feature in many modern smartphones. UWB stands out as an ideal candidate for precise location tracking due to its exceptional accuracy and minimal latency, making it a promising solution to the limitations discussed earlier.

### 4.4.1 Smartphone with UWB

UWB technology has recently found its way into smartphones, representing a major step forward in wireless communication and device interaction. Known for its precise location tracking capabilities,

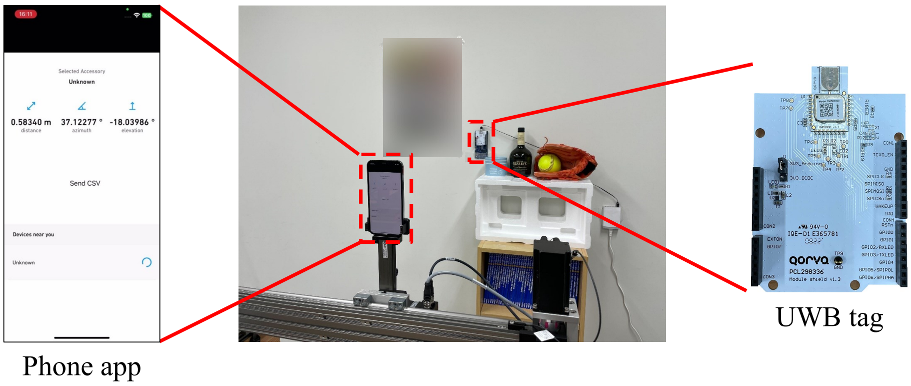


Figure 4.16: Experimental setup with UWB.

UWB has been predominantly employed in asset tracking solutions, as demonstrated by devices like Apple's AirTag [109] and Samsung's SmartTag+ [110]. These innovations highlight UWB's ability to enhance interactions with everyday items by delivering exceptional accuracy in locating them.

To further investigate its potential, this study used the DW3000 chip in conjunction with an iPhone 12 Pro to replicate the experimental setup outlined in Sec. 4.3. This included deploying an application on the iPhone to measure distance, bearing, and elevation angles, as illustrated in Fig. 4.16. This study customized Qorvo's iOS application, Qorvo Nearby Interaction [111, 112], which facilitates the localization of DW3000 devices using iOS-based UWB systems, to record localization data during the experiments.

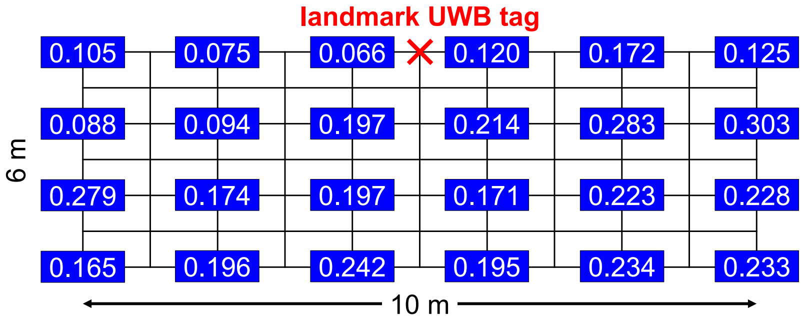
### 4.4.2 Global Coordinate Detection with UWB

As discussed in Sections 4.3.2 and 4.3.3, the detection range and angle of QR codes and AprilTags were found to be limited, particularly when approached from varying angles. In contrast, our experiments confirmed the reliability of using UWB tags as landmarks for localization.

In the setup illustrated in Fig. 4.16, we conducted tests by receiving signals from a UWB landmark tag while systematically placing a smartphone on a 2 m grid within a 6 m × 10 m room. The UWB landmark tag was fixed at the location indicated by the red "x" in the figure.

From the frames transmitted by the landmark UWB tag, which included distance, azimuth, and elevation information, we derived location estimates. Each blue square and corresponding number in Fig. 4.17 represents a measurement point and the associated positional error. The results demonstrate that communication with the UWB tag was possible from all areas of the room.

By combining UWB technology with vision-based approaches or leveraging advanced localization techniques capable of achieving errors within just a few centimeters [113, 114], we propose the possibility of attaining high-accuracy global coordinate detection over a broader area than what is



| | | | | | |
|---|---|---|---|---|---|
| 0.105 | 0.075 | 0.066 | 0.120 | 0.172 | 0.125 |
| 0.088 | 0.094 | 0.197 | 0.214 | 0.283 | 0.303 |
| 0.279 | 0.174 | 0.197 | 0.171 | 0.223 | 0.228 |
| 0.165 | 0.196 | 0.242 | 0.195 | 0.234 | 0.233 |

Figure 4.17: Localization error [m] with a UWB transceiver.

feasible with visual landmarks alone. This advancement has the potential to significantly enhance location-based services, paving the way for future innovations in wireless communication.

### 4.4.3 Global Coordinate Tracking with UWB

UWB signals exhibit minimal sensitivity to variations in lighting conditions. Moreover, the Nearby Interaction API [115] enables seamless extraction of UWB ranging and angle-of-arrival data between an iPhone and commercially available UWB tags. This makes the integration of UWB-based sensing with VIO-based localization, which can be prone to errors, an appealing solution to address the challenges identified in this paper.

To demonstrate this approach, we developed a prototype that combines these technologies. A single landmark UWB tag was placed in the environment to serve as a static physical anchor for the iPhone. As the iPhone moved through space, we measured its distance to the tag and the angle of the UWB signal's arrival. Simultaneously, we gathered VIO-based location estimates derived from VIO, depending on environmental factors, can experience drift or inaccuracies.

These datasets were coupled within a factor graph framework implemented using the open-source GTSAM [116] optimization library. The state space to be estimated includes the phone's position over time and the location of the static UWB tag. ARKit's VIO provided relative coordinates between time steps, while UWB measurements supplied range and bearing data relative to the tag. These inputs defined ”between factors” and ”range-bearing factors” to constrain the optimization.

The factor graph, illustrated in Fig. 4.18, was optimized using the Levenberg-Marquardt algorithm [117]. The optimized phone trajectory, alongside ARKit's VIO-only estimates and UWB-only measurements, is shown in Fig. 4.18. In challenging scenarios such as *dark night* or *blink* conditions, where VIO struggles to localize the phone (Sec. 4.3.5), UWB signal coupling significantly improves the estimates, as evidenced by Fig. 4.19 (c, d).

Notably, the optimized cumulative distribution functions (CDFs) in Figures 4.19 (a, b) outperform both standalone VIO and UWB sensing. This improvement demonstrates that combining

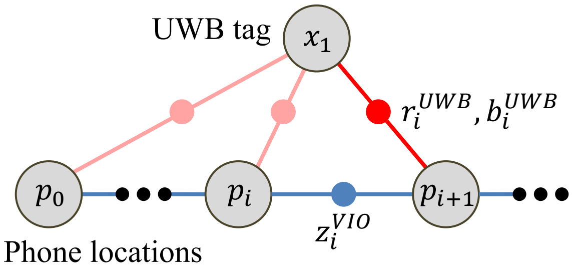
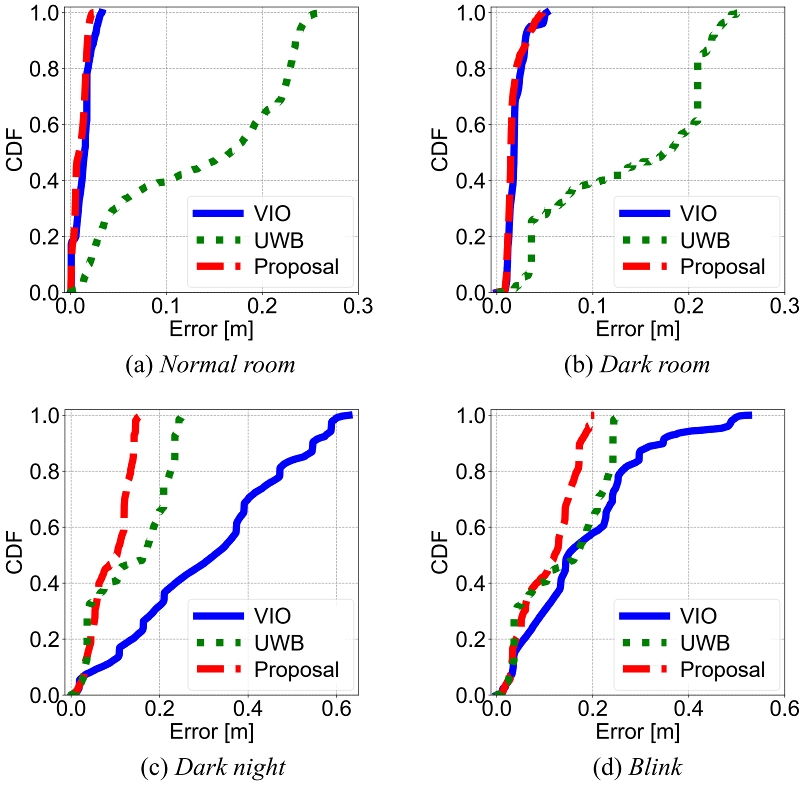


Figure 4.18: Factor graph for UWB + VIO coupling.



(a) *Normal room*

(b) *Dark room*

(c) *Dark night*

(d) *Blink*

Figure 4.19: CDF of localization error.

two independent sensing modalities—VIO and UWB—through an effective coupling algorithm like GTSAM can substantially enhance localization performance, even in scenarios where VIO fails.

## 4.5 Related Work

Many studies have explored alternative sensing methods to enhance vision-based localization systems, aiming to address the inherent limitations of such approaches. Vision-based systems, relying on cameras or LiDAR, often struggle in low-light conditions, featureless environments, or dynamic scenarios. Despite their widespread use, there has been limited systematic analysis of the failure modes associated with these systems. This study addresses that gap by identifying and analyzing key failure scenarios and advocating for the integration of complementary sensing technologies, such as UWB, to improve robustness and accuracy.

*Characterizing vision-based localization systems*: Precision in vision-based localization hinges on two aspects: accurate ego-localization and reliable placement of virtual anchors within an environment. Systems such as ARKit [118] for Apple devices and ARCore [119] for Android devices utilize computer vision techniques to deliver these capabilities for AR.

Previous research [120–122] has evaluated the accuracy of virtual anchor placement in controlled indoor environments, while others [123–126] have experimentally measured the effectiveness of ARKit's indoor localization. These studies provide insights into the general performance of vision-based systems but often lack detailed investigations into failure scenarios.

For instance, Nowacki et al. [127] analyzed the accuracy of ARKit's plane detection under varying lighting conditions but did not assess the end-to-end performance of localization and anchor placement. Similarly, UbiPose [128] highlighted ARKit's limitations in GPS-denied environments but focused on outdoor settings, leaving critical gaps in understanding indoor-specific challenges.

LiDAR-equipped devices, such as recent iPhones, have improved localization performance in certain scenarios, particularly in low-light conditions. Studies on LiDAR-based depth estimation and mapping [129, 130] have enhanced SLAM systems, yet they often fail to quantify system performance in challenging environments, such as featureless or highly reflective spaces. This lack of comprehensive data hinders progress in overcoming these limitations.

*Alternative sensing schemes*: The limitations of vision-based systems have motivated research into alternative sensing modalities, including acoustic, infrared, RFID, WiFi/BLE, and UWB. Recent advancements in acoustic sensing [131–133] have achieved high-accuracy localization, while systems such as X-AR [134] leverage RFID to improve anchor placement precision. Similarly, WiFi/BLE-based systems [135–137] have demonstrated robust tracking capabilities in indoor environments. UWB-based approaches, including XRLoc [113] and Garg et al. [114], achieve centimeter-level accuracy and provide a promising alternative to vision-based systems.

Despite these advancements, many studies do not adequately address the specific weaknesses of vision-based localization. This work contributes to the field by systematically evaluating these gaps, emphasizing the potential of alternative sensing technologies to complement and enhance vision-based localization systems.

*Integrating UWB with vision-based systems*: UWB technology, increasingly integrated into consumer devices, offers a complementary modality to vision-based systems. Recent studies [138–140] have explored UWB-enhanced localization for anchors and ego-localization in vision-based systems. These approaches integrate UWB into AR frameworks, bridging gaps in scenarios where vision-based methods falter. This study builds upon these findings, providing comprehensive measurements and identifying failure modes to guide future advancements in hybrid localization systems.

## 4.6 Conclusion

This chapter identified key challenges in vision-based indoor localization for motion capture of collaboration analysis based on exhaustive cases studies and controlled experiments. Errors in landmark-based localization were observed, with visual markers such as QR codes and AprilTags showing decreased localization accuracy at greater distances. Enlarging the size of landmarks did not improve accuracy, and lighting conditions had minimal impact. Angular constraints for landmark detection were also identified, with cameras detecting visual landmarks only within limited angular ranges: ±25° for QR codes and ±75° for AprilTags. Visual and LiDAR features for localization were examined, revealing that reduced visual features degrade localization accuracy. LiDAR sensors in modern iPhones provided additional depth information but were limited by low resolution. Coupling LiDAR with visual data improved tracking accuracy by 28.8 %. Tracking failures under low-light conditions were significant, with poor lighting reducing localization accuracy by 59.1 %, and dynamic lighting complicating tracking due to exposure issues. Speed limitations in IMU-based localization were evident, with accuracy decreasing at speeds over 2 m/s or below 0.2 m/s due to errors and drift.

A prototype leveraging UWB-based measurements was developed and demonstrated its potential to address these challenges. VIO and UWB were integrated using factor graphs in the prototype solution to preserve the unique characteristics of each localization method. Preliminary evaluations demonstrated that the prototype achieved superior accuracy under various lighting conditions.

# Chapter 5

# Conclusion

## 5.1 Conclusion

This study designed and implemented an IoT-based platform tailored for multimodal collaboration analysis, addressing the critical demands of modern collaborative environments. The system requires to meet three essential requirements for the system: synchronization accuracy of sensing devices, multimodal extraction of collaboration, and a user-friendly design for collaboration analysts.

Chapter 2 presented an innovative IoT system designed to support collaboration analysis, featuring three key components: the SRP Badge for data collection, the SRP Analysis for processing interaction data, and the SRP Web for visualizing results in a browser. The SRP Badge, a compact business-card-type sensor worn by individuals, captures data such as sound pressure, acceleration, and infrared signals with high precision, while ensuring accurate synchronization across devices. The SRP Analysis processes this synchronized data to identify collaboration, including face-to-face interaction, learning phases, speakers, and activity. These results are then visualized using SRP Web, providing a user-friendly interface for interpretation.

To evaluate the system's performance, experiments were conducted focusing on sensor synchronization accuracy, the reliability and effectiveness of the interaction analysis algorithm, and the usability of the web application. The findings demonstrated several key advantages for researchers analyzing collaborative learning. First, the sensors achieved precise data collection, maintaining synchronization errors within $\pm 30\,\mu s$. Second, the interaction analysis algorithm effectively identified collaborative behaviors such as face-to-face interactions, learning phases, speakers, and activity, offering valuable insights for qualitative studies. Lastly, the web application facilitated intuitive visualization of critical data points, significantly enhancing the efficiency of human interaction analysis through its web-based design and ease of use.

Chapter 3 presented a novel sound pressure sensor and a speaker identification algorithm specifically designed for compact, business-card-type sensors, aiming to analyze collaborative dynamics in multi-person settings. The sound pressure sensor integrates a peak hold circuit and a time synchronization module, which minimize signal spikes and maintain precise synchronization across devices. This ensures accurate and cost-effective detection of user speech. The proposed algorithm effectively suppresses background noise from non-speaker sensors, achieving reliable speaker identification.

The evaluation highlighted the system's robustness under diverse conditions, including varying numbers of users, levels of background noise, and durations of speech. The proposed algorithm demonstrated superior speaker identification accuracy compared to the comparative algorithm across all tested scenarios. Additionally, the peak hold circuit consistently captured user speech with high reliability, while synchronization errors between sensors were kept within ±30 µs. These two innovations are also considered to significantly contribute to the high accuracy of speaker identification.

Chapter 4 investigated practical challenges in vision-based indoor localization for motion capture in collaboration analysis, with three key findings emerging from the case studies and controlled experiments. First, visual markers such as QR codes and AprilTags exhibited limitations in localization accuracy as the distance from the camera increased, and enlarging the marker size failed to address this issue. Second, LiDAR integration in VIO rather disturbed localization accuracy in different lighting conditions. Third, IMU-based localization showed speed-related constraints, with performance deteriorating at speeds above 2 m/s or below 0.2 m/s due to drift and cumulative errors.

To address these challenges, a prototype leveraging UWB-based measurements was developed. By integrating VIO and UWB data through factor graphs, the prototype effectively combined the unique advantages of both localization methods. Initial evaluations demonstrated that the system achieved significantly higher accuracy under diverse lighting conditions.

## 5.2 Future Work

The development of this IoT-based platform for multimodal collaboration analysis represents a significant step forward in bridging the gap between qualitative and quantitative analytical approaches. However, to fully realize its potential and address the diverse needs of real-world applications, several areas for future improvement and expansion have been identified. These enhancements are categorized into four primary sections: an IoT system for collaboration analysis, speaker identification for mobile devices, indoor localization on mobile devices, and the discovery of collaboration patterns with the IoT system.

### 5.2.1 An IoT System for Collaboration Analysis

In Chapter 2, the IoT system at the core of this platform has demonstrated impressive capabilities in synchronizing multimodal data streams, ensuring precision, and providing a user-friendly interface. Nonetheless, further efforts are required to improve its usability and scalability.

**Balancing precision and modality expansion**: One of the platform's defining strengths is its high synchronization accuracy, achieving precision within ±30 µs. Maintaining this accuracy while expanding the range of modalities will be a critical challenge. The inclusion of additional data types, such as environmental sensing, physiological metrics, or contextual information, can enhance the depth of collaboration analysis. For environmental sensing, sensors monitoring light, temperature, humidity, or noise levels can offer valuable insights into environmental factors affecting collaborative dynamics. For physiological data, integrating devices like heart rate monitors, skin conductivity sensors, or other biometric tools can uncover how individual stress levels or emotional states influence group interactions. For contextual integration, tools that track task-specific contexts, such as digital tool usage or shared document activity, can provide a more comprehensive understanding of collaboration. The challenge lies in ensuring these additional modalities do not compromise the synchronization accuracy or overwhelm the system's processing capabilities. Advanced data fusion algorithms and optimized hardware architectures will be necessary to manage the increased complexity.

**Miniaturization and energy efficiency**: The current badge-type sensors provide robust performance for data collection, but their size and energy requirements may limit their application in some scenarios. Future efforts will focus on miniaturization, which involves reducing the size of sensors to make them less obtrusive and more wearable. Advances in micro-electromechanical systems (MEMS) and printed electronics could play a crucial role in achieving this goal. Additionally, enhancing energy efficiency is critical for long-term deployments by prolonging battery life without sacrificing functionality. Strategies for achieving this include incorporating energy-harvesting technologies and optimizing communication protocols to minimize power consumption. The combination of smaller, more energy-efficient sensors will expand the platform's usability across diverse settings.

**Hybrid environments**: As collaboration increasingly occurs across hybrid physical-virtual environments, the IoT system must adapt to these evolving contexts. Hybrid environments involve participants interacting both in person and remotely, often using a combination of physical tools and digital platforms. To support these scenarios, the platform will need to extend its capabilities to capture and analyze virtual interactions, such as screen sharing, video conferencing, and digital whiteboarding. It must also ensure seamless cross-platform compatibility, allowing smooth operation across different hardware and software ecosystems, including integration with popular

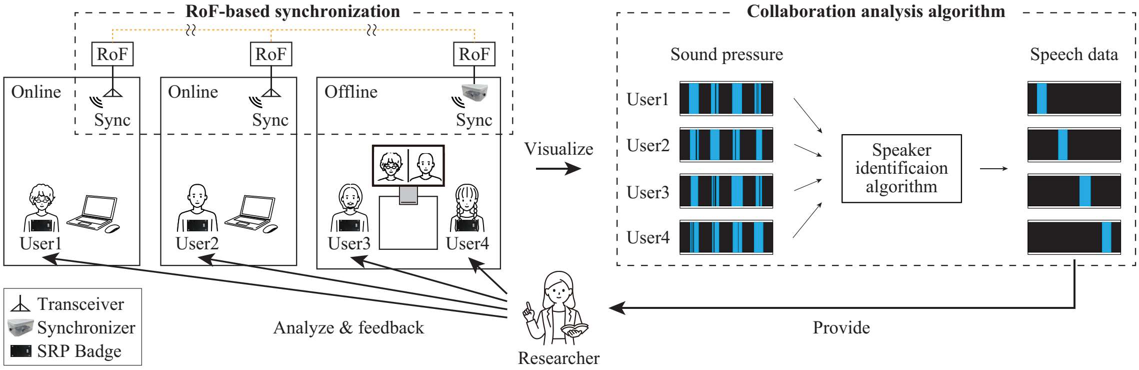


Figure 5.1: The overview of the prototype for hybrid collaboration analysis.

collaboration tools like Microsoft Teams and Zoom. Furthermore, the development of algorithms capable of dynamically adapting to the specific needs of hybrid environments will be essential, enabling the platform to balance data collection priorities between physical and virtual interactions. These enhancements will make the IoT platform indispensable for organizations navigating the complexities of modern collaboration.

Based on this motivation, this study steps into proposing a hybrid system designed for collaboration analysis in diverse environments. Figure 5.1 shows the overview of the system prototype. This system integrates key components: SRP Badges described in Chapter 2, precise synchronization to connect distant environments, and a representative analytical algorithm of speaker identification. For synchronization, Radio-over-Fiber (RoF) extends wireless networks to align clocks across badges located in separate rooms, ensuring consistent and accurate data integration. The analytical algorithm processes the collected data, focusing on multimodal aspects of collaboration. The algorithm includes speaker identification, a critical feature for interpreting interactions and dynamics within group activities. This concept expands the scope of collaboration analysis discussed in this study, providing a more comprehensive framework for understanding and evaluating interactions and dynamics in diverse environments.

### 5.2.2 Speaker Identification for Mobile Devices

In Chapter 3, effective speaker identification remains a cornerstone of collaboration analysis, particularly in settings with multiple participants. Although the current system has shown robust performance, real-world environments introduce a host of challenges that necessitate further refinement.

**Accuracy improvement in actual collaboration environments**: Speaker identification accuracy can be influenced by numerous factors in collaboration, including background noise, overlapping speech, and varying acoustic properties of the environment. Future work will focus on improving noise robustness by developing advanced noise reduction techniques to enhance detection accuracy in noisy settings such as busy offices or classrooms. Additionally, the system's ability to handle overlapping speech will be enhanced by leveraging advanced machine learning models and spatial audio processing to differentiate between simultaneous speakers. Furthermore, algorithms will be fine-tuned to adapt to the unique characteristics of different environments, such as open-plan offices that include mixed noise. Field testing in diverse real-world scenarios will be essential to identify and address specific challenges, ensuring the system's reliability across varied contexts.

### 5.2.3 Indoor Localization on Mobile Devices

Chapter 4 revealed practical challenges in using vision-based indoor localization for motion capture in collaboration analysis. The prototype solution with VIO and UWB was proposed and improved localization accuracy in environments where vision-based methods have struggled. However, there is still future work to apply this scheme to practical scenarios of collaboration analysis.

**Comprehensive evaluation and accuracy enhancement**: This study evaluated tracking performance of the prototype solution in a single scenario where VIO struggles, namely, varying lighting conditions. To demonstrate the improvement across a range of different environments, comprehensive controlled experiments as shown in Sec. 4.3 for VIO evaluation are required to further assess the solution's accuracy. Moreover, when cases of degraded accuracy emerge, it becomes crucial to identify and develop methods to improve performance in those specific situations.

In addition to controlled experiments, it is also important to evaluate the proposed scheme in real collaborative environments for a more practical assessment. In settings where multiple environmental factors interact simultaneously, performance equivalent to that achieved under controlled conditions cannot be guaranteed. Actual case studies in collaborative environments are needed to investigate how localization accuracy varies in more complex scenarios.

**Implementation of the motion capture tag**: As discussed in Sec. 4.1, implementing a dedicated tag is vital for using VIO and UWB in motion capture. Since the tag needs to be only a few centimeters in size, it must incorporate a camera, an IMU, and a UWB transceiver while maintaining operational power. Recent advancements in MEMS technology have made this increasingly feasible. However, the actual implementation requires careful consideration of various factors, including the tag's processing performance and continuous operating time.

**Localization accuracy improvement through tag collaboration**: By exchanging positional information among multiple tags worn by the user, tracking accuracy can be enhanced. Since these motion capture tags are attached to the user's body, each tag is subject to certain mechanical constraints. Leveraging these constraints can help prevent drift from accumulating and causing erroneous tracking directions. The communication protocol for tag collaboration remains a future challenge.

### 5.2.4 The Discovery of Collaboration Patterns with the IoT System

As described in Sec. 2.3.7, this study presented the application of the proposed IoT system in uncovering novel collaboration patterns. In this application, the proposed system was utilized with SSNA. SSNA revealed shifts in leadership roles by analyzing word co-occurrence networks and degree centralities, uncovering key moments of group interaction. The proposed IoT system highlighted nonverbal cues signaling transitions in collaboration phases. Intensive qualitative analysis showed that leadership adapted to task phases, with early efforts focused on gathering information and later efforts on problem-solving, demonstrating dynamic role shifts in collaboration.

Learning analytics already incorporates various quantitative methods, such as SSNA. Combining these established techniques with the proposed method will be key to uncovering new patterns of collaboration in future research.

# Acknowledgement

This thesis would not have been possible without the assistance and support of many individuals, and I would like to express my sincere gratitude to all of them.

First and foremost, I would like to extend my deepest appreciation to my supervisor, Professor Takashi Watanabe, for his invaluable guidance and insightful comments throughout my Ph.D. His creative suggestions and patient encouragement were indispensable to my research endeavors.

I am profoundly grateful to the members of my thesis committee, Professor Masayuki Murata and Professor Hirozumi Yamaguchi of the Graduate School of Information Science and Technology, Osaka University, and Professor Hideyuki Shimonishi of the Cyber Media Center, Osaka University, for their thorough reviews and perceptive feedback.

In addition, I would like to thank Professor Jun Oshima and Professor Ritsuko Oshima in Graduate School of Integrated Science and Technology, Shizuoka University for their support and teaching in the field of learning science.

I would also like to express my heartfelt gratitude to Associate Professor Shunsuke Saruwatari of Osaka University. He dedicated significant time and effort to guiding me and offering valuable advice that greatly enhanced my research.

Furthermore, I am deeply thankful to Assistant Professor Takuya Fujihashi of Osaka University for his thoughtful comments and encouragement. His kindness and assistance have been invaluable, and I remain truly indebted to him.

My appreciation also extends to the international researchers who provided me with opportunities to broaden my horizons. I am especially grateful to Associate Professor Daniel Spikol of the University of Copenhagen, who hosted me as a Guest Ph.D. from April to September 2024 and facilitated numerous discussions on the research and development of IoT systems for collaboration analysis.

I am equally grateful to Associate Professor Dinesh Bharadia and Dr. Aditya Arun of the University of California, San Diego. Our insightful discussions on identifying practical challenges and solutions for current vision-based indoor localization were incredibly enriching.

I would also like to thank Ms. Ami Takahashi for their invaluable contributions to managing laboratory tasks and supporting my work.

Although space limitations prevent me from naming everyone, I am deeply grateful to all members of the Intelligent Networking Laboratory for their feedback, encouragement, and support. In particular, I extend my heartfelt thanks to Mr. Kazuhiro Kizaki, Specially Appointed Researcher at Osaka University, for sharing his extensive and profound knowledge of electronic circuits and communication, which was invaluable to my work. Additionally, I will never forget the laboratory members who shared the highs and lows of research with me: Mr. Motoki Nagano, Mr. Shunpei Ohira, Mr. Daichi Yamaguchi, Mr. Yutaro Osako, Mr. Takuya Fujiwara, Mr. Misaki Sakuta, Mr. Ryotaro Hada, Mr. Gaku Uchida, and Mr. Haruki Tanaka.

Finally, I would like to express my gratitude to my friends and family for their unwavering understanding and support throughout my Ph.D. journey.

# Bibliography

[1] B. Jordan and A. Henderson, “Interaction Analysis: Foundations and Practice,” *Journal of the Learning Sciences*, vol. 4, no. 1, pp. 39–103, 1995.

[2] J. Oshima, R. Oshima, and Y. Matsuzawa, “Knowledge Building Discourse Explorer: A Social Network Analysis Application for Knowledge Building Discourse,” *Educational Technology Research and Development*, vol. 60, no. 5, pp. 903–921, 2012.

[3] R. K. Sawyer, *Cambridge Handbook of the Learning Sciences, Second Edition.* Cambridge University Press, 2014.

[4] J. Oshima, R. Oshima, and W. Fujita, “A Multivocality Approach to Epistemic Agency in Collaborative Learning,” in *The Computer Supported Collaborative Learning (CSCL) Conference*, 2015, pp. 62–69.

[5] G. Chen, C. K. K. Chan, K. K. H. Chan, S. N. Clarke, and L. B. Resnick, “Efficacy of Video-based Teacher Professional Development for Increasing Classroom Discourse and Student Learning,” *Journal of Learning Analytics*, vol. 29, no. 4–5, pp. 642–680, 2020.

[6] Y. Wakisaka, N. Ohkubo, K. Ara, N. Sato, M. Hayakawa, S. Tsuji, Y. Horry, K. Yano, and N. Moriwaki, “Beam-Scan Sensor Node: Reliable Sensing of Human Interactions in Organization,” in *International Conference on Networked Sensing Systems (INSS)*, 2009, pp. 1–4.

[7] S. Tsuji, N. Sato, K. Yano, J. Broad, and F. Luthans, “Employees’ Wearable Measure of Face-to-Face Communication Relates to Their Positive Psychological Capital, Well-Being,” in *IEEE/WIC/ACM International Conference on Web Intelligence - Companion Volume (WI)*, 2019, pp. 14–20.

[8] L. Wu, B. Waber, S. Aral, E. Brynjolfsson, and A. Pentland, “Mining Face-to-Face Interaction Networks Using Sociometric Badges: Predicting Productivity in an IT Configuration Task,” in *International Conference on Information Systems (ICIS)*, 2008, pp. 1–19.

[9] O. Lederman, D. Calacci, A. MacMullen, D. Fehder, F. Murray, and A. Pentland, "Open Badges: A Low-Cost Toolkit for Measuring Team Communication and Dynamics," in *International Conference on Social Computing, Behavioral-Cultural Modeling & Prediction and Behavior Representation in Modeling and Simulation (SBP-BRiMS)*, 2016, pp. 1–7.

[10] O. Lederman, A. Mohan, D. Calacci, and A. S. Pentland, "Rhythm: A Unified Measurement Platform for Human Organizations," *IEEE MultiMedia*, vol. 25, no. 1, pp. 26–38, 2018.

[11] X. Ochoa, A. C. Lang, and G. Siemens, "Multimodal learning analytics," in *The Handbook of Learning Analytics*. SOLAR - Society for Learning Analytics Research, 2017, vol. 1, pp. 129–141.

[12] Q. Zhou, W. Suraworachet, and M. Cukurova, "Detecting non-verbal speech and gaze behaviours with multimodal data and computer vision to interpret effective collaborative learning interactions," *Education and Information Technologies*, vol. 29, pp. 1071–1098, 2024.

[13] O. Noroozi, H. J. Pijeira-Díaz, M. Sobocinski, M. Dindar, S. Järvelä, and P. A. Kirschner, "Multimodal data indicators for capturing cognitive, motivational, and emotional learning processes: A systematic literature review," *Education and Information Technologies*, vol. 25, pp. 5499–5547, 2020.

[14] Z. Li, M. T. Jensen, A. Nolte, and D. Spikol, "Field report for Platform mBox: Designing an Open MMLA Platform," in *Learning Analytics and Knowledge Conference*, 2024, pp. 785–791.

[15] M. Suzuki, C.-H. Liao, S. Ohara, K. Jinno, and H. Morikawa, "Wireless-Transparent Sensing," in *Proceedings of the International Conference on Embedded Wireless Systems and Networks*, 2017, pp. 66–77.

[16] F. Ferrari, M. Zimmerling, L. Thiele, and O. Saukh, "Efficient network flooding and time synchronization with Glossy," in *ACM/IEEE International Conference on Information Processing in Sensor Networks*, 2011, pp. 73–84.

[17] Y. Matsubara, Y. Sakurai, and C. Faloutsos, "AutoPlait: Automatic Mining of Co-Evolving Time Sequences," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2014, pp. 193–204.

[18] Cognition and T. G. at Vanderbilt, "The Jasper Series as an Example of Anchored Instruction: Theory, Program Description, and Assessment Data," *Educational Psychologist*, vol. 27, no. 3, pp. 291–315, 1992.

[19] J. Liu, Z. Jiang, and K. Lin, “A Robust Reliable Low-Power High-Throughput Data Collection Wireless Sensor Network,” *IEEE Sensors Journal*, vol. 24, no. 17, pp. 28 210–28 221, 2024.

[20] E. H. Halili, *Apache JMeter.* Packt Publishing Birmingham, 2008.

[21] Y. Akimoto, S. Ohtawa, R. Oshima, S. Saruwatari, and J. Oshima, “A Preliminary Evaluation of Collaborative Learning Analysis Tool with IoT (in Japanese),” in *National Convention of Information Processing Society of Japan*, 2019, pp. 1–2.

[22] M. Ermes, J. Pärkkä, and L. Cluitmans, “Advancing from Offline to Online Activity Recognition with Wearable Sensors,” in *International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, 2008, pp. 4451–4454.

[23] U. Maurer, A. Smailagic, D. P. Siewiorek, and M. Deisher, “Activity Recognition and Monitoring Using Multiple Sensors on Different Body Positions,” in *International Workshop on Wearable and Implantable Body Sensor Networks (BSN)*, 2006, pp. 113–116.

[24] Ó. D. Lara and M. A. Labrador, “A Mobile Platform for Real-Time Human Activity Recognition,” in *IEEE Consumer Communications and Networking Conference (CCNC)*, 2012, pp. 667–671.

[25] T.-P. Kao, C.-W. Lin, and J.-S. Wang, “Development of a Portable Activity Detector for Daily Activity Recognition,” in *IEEE International Symposium on Industrial Electronics (ISIE)*, 2009, pp. 115–120.

[26] A. Nandy, J. Saha, C. Chowdhury, and K. P. Singh, “Detailed Human Activity Recognition using Wearable Sensor and Smartphones,” in *International Conference on Opto-Electronics and Applied Optics (Optronix)*, 2019, pp. 1–6.

[27] B. Barshan and A. Yurtman, “Classifying Daily and Sports Activities Invariantly to the Positioning of Wearable Motion Sensor Units,” *IEEE Internet of Things Journal*, vol. 7, no. 6, pp. 4801–4815, 2020.

[28] J. chun Zhao, J. feng Zhang, Y. Feng, and J. xin Guo, “The study and application of the IOT technology in agriculture,” in *International Conference on Computer Science and Information Technology*, 2010, pp. 462–465.

[29] T. Fukatsu, T. Kiura, and M. Hirafuji, “A web-based sensor network system with distributed data processing approach via web application,” *Computer Standards & Interfaces*, vol. 33, no. 6, pp. 565–573, 2011.

[30] M. A. G. Maureira, D. Oldenhof, and L. Teernstra, “ThingSpeak–an API and Web Service for the Internet of Things,” *World Wide Web*, pp. 1–8, 2011.

[31] R. T. Hameed, O. A. Mohamad, O. T. Hamid, and N. Ţăpuş, “Patient Monitoring System Based on e-health Sensors and Web Services,” in *International Conference on Electronics, Computers and Artificial Intelligence*, 2016, pp. 1–6.

[32] J. Jo and I. Jang, “Applying sensor web enablement to retrieve and visualize sensor observations across the Web,” in *International Conference on Information and Communication Technology Convergence*, 2016, pp. 852–854.

[33] K. A. Patil and N. R. Kale, “A model for smart agriculture using IoT,” in *International Conference on Global Trends in Signal Processing, Information Computing and Communication*, 2016, pp. 543–545.

[34] L. Pescosolido, R. Berta, L. Scalise, G. M. Revel, A. D. Gloria, and G. Orlandi, “An IoT-inspired cloud-based web service architecture for e-Health applications,” in *IEEE International Smart Cities Conference*, 2016, pp. 1–4.

[35] K.-C. Kao, W.-H. Chieng, and S.-L. Jeng, “Design and development of an IoT-based web application for an intelligent remote SCADA system,” in *IOP Conference Series: Materials Science and Engineering*, 2018, pp. 1–7.

[36] A. Krishna, M. L. Pallec, R. Mateescu, L. Noirie, and G. Salaün, “IoT Composer: Composition and Deployment of IoT Applications,” in *IEEE/ACM International Conference on Software Engineering: Companion Proceedings*, 2019, pp. 19–22.

[37] J. Muangprathub, N. Boonnam, S. Kajornkasirat, N. Lekbangpong, A. Wanichsombat, and P. Nillaor, “IoT and agriculture data analysis for smart farm,” *Computers and Electronics in Agriculture*, vol. 156, pp. 467–474, 2019.

[38] R. K. Jain, B. J. Saikia, N. P. Rai, and P. P. Ray, “Development of Web-based Application for Mobile Robot using IOT Platform,” in *International Conference on Computing, Communication and Networking Technologies*, 2020, pp. 1–6.

[39] M. Nagano, Y. Arai, T. Fujihashi, T. Watanabe, and S. Saruwatari, “Design and Implementation of Device Monitoring SaaS for DIY-IoT Systems,” in *IEEE International Conference on Consumer Electronics*, 2021, pp. 1–4.

[40] C. R. Haller, V. J. Gallagher, T. L. Weldon, and R. M. Felder, “Dynamics of Peer Education in Cooperative Learning Workgroups,” *Journal of Engineering Education*, vol. 89, no. 3, pp. 286–293, 2000.

[41] E. Vass, K. Littleton, D. Miell, and A. Jones, "The discourse of collaborative creative writing: Peer collaboration as a context for mutual inspiration," *Thinking Skills and Creativity*, vol. 3, no. 3, pp. 192–202, 2008.

[42] M. A. Evans, E. Feenstra, E. Ryon, and D. McNeill, "A multimodal approach to coding discourse: Collaboration, distributed cognition, and geometric reasoning," *International Journal of Computer-Supported Collaborative Learning*, vol. 6, pp. 253–278, 2011.

[43] J. Oshima, R. Oshima, and K. Fujii, "Student Regulation of Collaborative Learning in Multiple Document Integration," *The Proceedings of the International Conference of the Learning Science (ICLS)*, vol. 2, pp. 967–971, 2014.

[44] E. Haataja, J. Malmberg, and S. Järvelä, "Monitoring in collaborative learning: Co-occurrence of observed behavior and physiological synchrony explored," *Computers in Human Behavior*, vol. 87, pp. 337–347, 2018.

[45] J. Oshima, R. Oshima, and W. Fujita, "A Mixed-Methods Approach to Analyze Shared Epistemic Agency in Jigsaw Instruction at Multiple Scales of Temporality," *Journal of Learning Analytics*, vol. 5, no. 1, pp. 10–24, 2018.

[46] L. D. Backer, H. V. Keer, F. D. Smedt, E. Merchie, and M. Valcke, "Identifying regulation profiles during computer-supported collaborative learning and examining their relation with students' performance, motivation, and self-efficacy for learning," *Computers & Education*, vol. 179, p. 104421, 2022.

[47] C. H. Knapp and G. C. Carter, "The Generalized Correlation Method for Estimation of Time Delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[48] R. O. Schmidt, "Multiple Emitter Location and Signal Parameter Estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[49] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, *Robust Localization in Reverberant Rooms.* Microphone Arrays, 2001, ch. 8, pp. 157–180.

[50] S. Mohan, M. E. Lockwood, M. L. Kramer, and D. L. Jones, "Localization of Multiple Acoustic Sources with Small Arrays Using a Coherence Test," *The Journal of the Acoustical Society of America*, vol. 123, no. 4, pp. 2136–2147, 2008.

[51] W. Zhang and B. D. Rao, "A Two Microphone-Based Approach for Source Localization of Multiple Speech Sources," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 1913–1928, 2010.

[52] N. Ma, G. J. Brown, and T. May, “Exploiting Deep Neural Networks and Head Movements for Binaural Localisation of Multiple Speakers in Reverberant Conditions,” in *Proceedings of the Annual Conference of the International Speech Communication Association*, 2015, pp. 3302–3306.

[53] N. Ma, T. May, and G. J. Brown, “Exploiting Deep Neural Networks and Head Movements for Robust Binaural Localization of Multiple Sources in Reverberant Environments,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2444–2453, 2017.

[54] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, “A Consolidated Perspective on Multimicrophone Speech Enhancement and Source Separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[55] Z.-Q. Wang, X. Zhang, and D. Wang, “Robust Speaker Localization Guided by Deep Learning-Based Time-Frequency Masking,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 178–188, 2019.

[56] J.-M. Valin, F. Michaud, J. Rouat, and D. Létourneau, “Robust Sound Source Localization Using a Microphone Array on a Mobile Robot,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2003, pp. 1228–1233.

[57] J.-M. Valin, F. Michaud, and J. Rouat, “Robust Localization and Tracking of Simultaneous Moving Sound Sources Using Beamforming and Particle Filtering,” *Robotics and Autonomous Systems*, vol. 55, no. 3, pp. 216–228, 2007.

[58] F. Grondin and F. Michaud, “Time Difference of Arrival Estimation Based on Binary Frequency Mask for Sound Source Localization on Mobile Robots,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2015, pp. 6149–6154.

[59] E. L. Ferguson, S. B. Williams, and C. T. Jin, “Sound Source Localization in a Multipath Environment Using Convolutional Neural Networks,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 2386–2390.

[60] M. Zohourian, G. Enzner, and R. Martin, “Binaural Speaker Localization Integrated Into an Adaptive Beamformer for Hearing Aids,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 515–528, 2018.

[61] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, “Robust Speaker Recognition in Noisy Conditions,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1711–1723, 2007.

[62] Y. Yang, S. Wang, M. Sun, Y. Qian, and K. Yu, "Generative Adversarial Networks based X-vector Augmentation for Robust Probabilistic Linear Discriminant Analysis in Speaker Verification," in *International Symposium on Chinese Spoken Language Processing*, 2018, pp. 205–209.

[63] S. Pandiaraj, H. N. R. Keziah, D. S. Vinothini, L. Gloria, and K. R. S. Kumar, "A Confidence Measure based — Score Fusion Technique to Integrate MFCC and Pitch for Speaker Verification," in *International Conference on Electronics Computer Technology*, 2011, pp. 317–320.

[64] S. Nakagawa, L. Wang, and S. Ohtsuka, "Speaker Identification and Verification by Combining MFCC and Phase Information," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1085–1095, 2012.

[65] A. Roy, M. Magimai.-Doss, and S. Marcel, "A Fast Parts-Based Approach to Speaker Verification Using Boosted Slice Classifiers," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 1, pp. 241–254, 2012.

[66] H. Taherian, Z.-Q. Wang, J. Chang, and D. Wang, "Robust Speaker Recognition Based on Single-Channel and Multi-Channel Speech Enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1293–1302, 2020.

[67] A. Chowdhury and A. Ross, "Fusing MFCC and LPC Features Using 1D Triplet CNN for Speaker Recognition in Severely Degraded Audio Signals," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1616–1629, 2020.

[68] D.-G. Shin and M.-S. Jun, "Home IoT Device Certification through Speaker Recognition," in *International Conference on Advanced Communication Technology*, 2015, pp. 600–603.

[69] Z. Wu, P. L. D. Leon, C. Demiroglu, A. Khodabakhsh, S. King, Z.-H. Ling, D. Saito, B. Stewart, T. Toda, M. Wester, and J. Yamagishi, "Anti-Spoofing for Text-Independent Speaker Verification: An Initial Database, Comparison of Countermeasures, and Human Performance," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 768–783, 2016.

[70] L. Yang, Z. Zhao, and G. Min, "User Verification Based On Customized Sentence Reading," in *IEEE International Conference on Cyber Science and Technology Congress*, 2018, pp. 353–356.

[71] N. McLaughlin, J. Ming, and D. Crookes, "Speaker Recognition in Noisy Conditions with Limited Training Data," in *European Signal Processing Conference*, 2011, pp. 1294–1298.

[72] G. Biagetti, P. Crippa, L. Falaschetti, S. Orcioni, and C. Turchetti, “Speaker Identification in Noisy Conditions Using Short Sequences of Speech Frames,” in *Smart Innovation, Systems and Technologies*, 2018, pp. 43–52.

[73] Mangesh S. Deshpande and Raghunath S. Holambe, “Speaker Identification Based on Robust AM-FM Features,” in *International Conference on Emerging Trends in Engineering & Technology*, 2009, pp. 880–884.

[74] D. A. Reynolds, “Experimental Evaluation of Features for Robust Speaker Identification,” *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 639–643, 1994.

[75] Z. Wu and Z. Cao, “Improved MFCC-Based Feature for Robust Speaker Identification,” *Tsinghua Science and Technology*, vol. 10, no. 2, pp. 158–161, 2005.

[76] S. Chakroborty, A. Roy, and G. Saha, “Fusion of a Complementary Feature Set with MFCC for Improved Closed Set Text-Independent Speaker Identification,” in *IEEE International Conference on Industrial Technology*, vol. 387–390, 2006.

[77] A. Maesa, F. Garzia, M. Scarpiniti, and R. Cusani, “Text Independent Automatic Speaker Recognition System Using Mel-Frequency Cepstrum Coefficient and Gaussian Mixture Models,” *Journal of Information Security*, vol. 3, no. 4, pp. 335–340, 2012.

[78] B. G. Nagaraja and H. S. Jayanna, “Efficient Window for Monolingual and Crosslingual Speaker Identification using MFCC,” in *International Conference on Advanced Computing and Communication Systems*, 2013, pp. 1–4.

[79] R. Ajgou, S. Sbaa, S. Ghendir, A. Chamsa, and A. Taleb-Ahmed, “Robust Remote Speaker Recognition System Based on AR-MFCC Features and Efficient Speech Activity Detection Algorithm,” in *International Symposium on Wireless Communications Systems*, 2014, pp. 722–727.

[80] A. Bakshi, S. K. Kopparapu, S. Pawar, and S. Nema, “Novel Windowing Technique of MFCC for Speaker Identification with Modified Polynomial Classifiers,” in *International Conference on Confluence The Next Generation Information Technology Summit*, 2014, pp. 292–297.

[81] P. M. Chauhan and N. P. Desai, “Mel Frequency Cepstral Coefficients (MFCC) Based Speaker Identification in Noisy Environment Using Wiener Filter,” in *International Conference on Green Computing Communication and Electrical Engineering*, 2014, pp. 1–5.

[82] K. Matsumoto, N. Hayasaka, and Y. Iiguni, “Noise Robust Speaker Identification by Dividing MFCC,” in *International Symposium on Communications, Control and Signal Processing*, 2014, pp. 652–655.

[83] S. S. Wali, S. M. Hatture, and S. Nandyal, “MFCC Based Text-Dependent Speaker Identification Using BPNN,” *International Journal of Signal Processing Systems*, vol. 3, no. 1, pp. 30–34, 2015.

[84] B. Ayoub, K. Jamal, and Z. Arsalane, “An Analysis and Comparative Evaluation of MFCC Variants for Speaker Identification over VoIP Networks,” in *World Congress on Information Technology and Computer Applications*, 2015, pp. 1–6.

[85] I. Volfin and I. Cohen, “Dominant Speaker Identification for Multipoint Videoconferencing,” in *IEEE Convention of Electrical and Electronics Engineers in Israel*, 2012, pp. 1–4.

[86] R. Karadaghi, H. Hertlein, and A. Ariyaeeinia, “Effectiveness in Open-Set Speaker Identification,” in *International Carnahan Conference on Security Technology*, 2014, pp. 1–6.

[87] J. Poignant, L. Besacier, and G. Quénot, “Unsupervised Speaker Identification in TV Broadcast Based on Written Names,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 57–68, 2015.

[88] K. Brunet, K. Taam, E. Cherrier, N. Faye, and C. Rosenberger, “Speaker Recognition for Mobile User Authentication: An Android Solution,” in *Conférence sur la Sécurité des Architectures Réseaux et Systèmes d’Information*, 2013, pp. 1–10.

[89] A. Larcher, K. A. Lee, B. Ma, and H. Li, “Text-Dependent Speaker Verification: Classifiers, Databases and RSR2015,” *Speech Communication*, vol. 60, pp. 56–77, 2014.

[90] J. Nishimura, N. Sato, and T. Kuroda, “Speech ”Siglet” Detection for Business Microscope,” in *IEEE International Conference on Pervasive Computing and Communications*, 2008, pp. 147–152.

[91] J. Nishimura and T. Kuroda, “Speaker Recognition using Speaker-independent Universal Acoustic Model and Synchronous Sensing for Business Microscope,” in *International Symposium on Wireless Pervasive Computing*, 2009, pp. 1–5.

[92] ——, “Hybrid Speaker Recognition Using Universal Acoustic Model,” *SICE Journal of Control, Measurement, and System Integration*, vol. 4, no. 6, pp. 410–416, 2011.

[93] A. Sangwan, M. C. Chiranth, H. S. Jamadagni, R. Sah, R. V. Prasad, and V. Gaurav, “VAD Techniques for Real-Time Speech Transmission on the Internet,” in *IEEE International Conference on High Speed Networks and Multimedia Communication*, 2002, pp. 46–50.

[94] G. Biagetti, P. Crippa, A. Curzi, S. Orcioni, and C. Turchetti, “Speaker Identification with Short Sequences of Speech Frames,” in *International Conference on Pattern Recognition Applications and Methods*, 2015, pp. 178–185.

[95] V. Echeverría, A. A. no, K. Chiluiza, A. Vásquez, and X. Ochoa, “Presentation skills estimation based on video and kinect data analysis,” in *ACM Workshop on Multimodal Learning Analytics Workshop and Grand Challenge*, 2014, pp. 53–60.

[96] I. Radu, E. Tu, and B. Schneider, “Relationships between body postures and collaborative learning states in an augmented reality study,” in *Artificial Intelligence in Education*, 2020, pp. 257–262.

[97] (2021) Ar visual. [Online]. Available: https://apps.apple.com/us/app/ar-visual/id1415771396

[98] (2023) Anywhere:cube. [Online]. Available: https://apps.apple.com/us/app/anywhere-cube/id1402127283

[99] (2024) Arvid augmented reality. [Online]. Available: https://apps.apple.com/us/app/arvid-augmented-reality/id1276546297

[100] (2023) Augment - 3d augmented reality. [Online]. Available: https://apps.apple.com/us/app/augment-3d-augmented-reality/id506463171

[101] (2024) Cocoar - ar app. [Online]. Available: https://apps.apple.com/us/app/cocoar-ar-app/id867328953

[102] (2024) Measure. [Online]. Available: https://apps.apple.com/us/app/measure/id1383426740

[103] (2019) Monster park - ar dino world. [Online]. Available: https://apps.apple.com/us/app/monster-park-ar-dino-world/id1259767702

[104] G. Bradski, “The OpenCV Library,” *Dr. Dobb’s Journal of Software Tools*, vol. 25, no. 11, pp. 122–125, 2000.

[105] J. Zhang and S. Singh, “LOAM: Lidar Odometry and Mapping in Real-time,” in *Robotics: Science and Systems*, 2014, pp. 1–9.

[106] M. Borges, A. Symington, B. Coltin, T. Smith, and R. Ventura, “HTC Vive: Analysis and Accuracy Improvement,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 2610–2615.

[107] D. Engelsman and I. Klein, “Information-Aided Inertial Navigation: A Review,” *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–18, 2023.

[108] Z. Yuan, D. Zhu, C. Chi, J. Tang, C. Liao, and X. Yang, “Visual-Inertial State Estimation with Pre-integration Correction for Robust Mobile Augmented Reality,” in *ACM International Conference on Multimedia*, 2019, pp. 1410–1418.

[109] (2022) Find your keys, backpack, and more with airtag. [Online]. Available: https://support.apple.com/en-us/HT210967

[110] (2021) Introducing the new galaxy smarttag+: The smart way to find lost items. [Online]. Available: https://news.samsung.com/us/introducing-the-new-galaxy-smarttag-plus/

[111] (2024) Qorvo Nearby Interaction. [Online]. Available: https://apps.apple.com/ml/app/qorvo-nearby-interaction/id1615369084

[112] (2024) DWM3001CDK: Ultra-Wideband (UWB) Module Development Kit. [Online]. Available: https://www.qorvo.com/products/p/DWM3001CDK#evaluation-tools

[113] A. Arun, S. Saruwatari, S. Shah, and D. Bharadia, "XRLoc: Accurate UWB Localization to Realize XR Deployments," in *ACM Conference on Embedded Networked Sensor Systems (SenSys)*, 2023, pp. 459–473.

[114] N. Garg, I. Shahid, K. Sankar, M. Dasari, R. K. Sheshadri, K. Sundaresan, and N. Roy, "Bringing AR/VR to Everyday Life - a Wireless Localization Perspective," in *International Workshop on Mobile Computing Systems and Applications (HotMobile)*, 2023, pp. 142–142.

[115] "Nearby interactions," 2023. [Online]. Available: https://developer.apple.com/design/human-interface-guidelines/nearby-interactions

[116] F. Dellaert, "Factor Graphs and GTSAM: A Hands-on Introduction," Georgia Institute of Technology, Tech. Rep., Tech. Rep., 2012.

[117] A. Ranganathan, "The Levenberg-Marquardt Algorithm," 2004. [Online]. Available: https://www.phy.olemiss.edu/~jgladden/sci_comp/resources/Levenber-Marquardt_Tutorial1.pdf

[118] (2023) Understanding world tracking. [Online]. Available: https://developer.apple.com/documentation/arkit/arkit_in_ios/configuration_objects/understanding_world_tracking

[119] (2023) Ar core. [Online]. Available: https://developers.google.com/ar

[120] Z. Oufqir, A. E. Abderrahmani, and K. Satori, "Arkit and arcore in serve to augmented reality," in *International Conference on Intelligent Systems and Computer Vision (ISCV)*, 2020, pp. 1–7.

[121] T. Scargill, S. Hurli, J. Chen, and M. Gorlatova, "Will it Move?: Indoor Scene Characterization for Hologram Stability in Mobile AR," in *International Workshop on Mobile Computing Systems and Applications (HotMobile)*, 2021, pp. 174–176.

[122] T. Scargill, G. Premsankar, J. Chen, and M. Gorlatova, "Here To Stay: A Quantitative Comparison of Virtual Object Stability in Markerless Mobile AR," in *International Workshop on Cyber-Physical-Human System Design and Implementation*, 2022, pp. 24–29.

[123] C. Yoon, R. Louie, J. Ryan, M. Vu, H. Bang, W. Derksen, and P. Ruvolo, "Leveraging augmented reality to create apps for people with visual disabilities: A case study in indoor navigation," in *International ACM SIGACCESS Conference on Computers and Accessibility*, 2019, pp. 210–221.

[124] W. Zhang, B. Han, and P. Hui, "Jaguar: Low Latency Mobile Augmented Reality with Flexible Tracking," in *ACM International Conference on Multimedia*, 2018, pp. 355–363.

[125] R. Cervenak and P. Masek, "Arkit as indoor positioning system," in *International Congress on Ultra Modern Telecommunications and Control Systems and Workshops*, 2019, pp. 1–5.

[126] G. Zhang, J. Yuan, H. Liu, Z. Peng, C. Li, Z. Wang, and H. Bao, "100-Phones: A Large VI-SLAM Dataset for Augmented Reality Towards Mass Deployment on Mobile Phones," *IEEE Transactions on Visualization and Computer Graphics*, pp. 2098–2108, 2024.

[127] P. Nowacki and M. Woda, "Capabilities of ARCore and ARKit Platforms for AR/VR Applications," in *Engineering in Dependability of Computer Systems and Networks*, 2020, pp. 358–370.

[128] W. Pang, C. Xia, B. Leong, F. Ahmad, J. Paek, and R. Govindan, "UbiPose: Towards Ubiquitous Outdoor AR Pose Tracking using Aerial Meshes," in *ACM Annual International Conference on Mobile Computing and Networking (MobiCom)*, 2023, pp. 736–751.

[129] T. Feigl, A. Porada, S. Steiner, C. Löffler, C. Mutschler, and M. Philippsen, "Localization Limitations of ARCore, ARKit, and Hololens in Dynamic Large-scale Industry Environments," in *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2020, pp. 307–318.

[130] U. Dilek and M. Erol, "Detecting position using arkit ii: generating position-time graphs in real-time and further information on limitations of arkit," *Physics Education*, vol. 53, no. 3, pp. 1–6, 2018.

[131] S. Yun, Y.-C. Chen, H. Zheng, L. Qiu, and W. Mao, "Strata: Fine-Grained Acoustic-based Device-Free Tracking," in *ACM International Conference on Mobile Systems, Applications, and Services (MobiSys)*, 2017, pp. 15–28.

[132] A. Wang and S. Gollakota, "MilliSonic: Pushing the Limits of Acoustic Motion Tracking," in *ACM Conference on Human Factors in Computing Systems (CHI)*, 2019, pp. 1–11.

[133] L. Wang, H. Wan, T. Zhao, K. Sun, S. Shi, H. Dai, G. Chen, H. Liu, and W. Wang, "SCALAR: Self-Calibrated Acoustic Ranging for Distributed Mobile Devices," *IEEE Transactions on Mobile Computing*, vol. 23, no. 2, pp. 1701–1716, 2023.

[134] T. Boroushaki, M. Lam, L. Dodds, A. Eid, and F. Adib, "Augmenting Augmented Reality with Non-Line-of-Sight Perception," in *USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, 2023, pp. 1341–1358.

[135] M. Kotaru and S. Katti, "Position Tracking for Virtual Reality Using Commodity WiFi," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2671–2681.

[136] Z. Han, Z. Lu, X. Wen, W. Zheng, J. Zhao, and L. Guo, "CentiTrack: Toward Centimeter-Level Passive Gesture Tracking With Commodity WiFi," *IEEE Internet of Things Journal*, vol. 10, no. 14, pp. 13 012–13 027, 2023.

[137] Z. Gu, T. He, J. Yin, Y. Xu, and J. Wu, "TyrLoc: a low-cost multi-technology MIMO localization system with a single RF chain," in *ACM International Conference on Mobile Systems, Applications, and Services (MobiSys)*, 2021, pp. 228–240.

[138] Y. Schröder and L. Wolf, "InPhase: Phase-based Ranging and Localization," *ACM Transactions on Sensor Networks*, vol. 18, no. 2, pp. 1–39, 2022.

[139] V. D. Pietra and P. Dabove, "Recent advances for uwb ranging from android smartphone," in *IEEE/ION Position, Location and Navigation Symposium*, 2023, pp. 1226–1233.

[140] A. Heinrich, S. Krollmann, F. Putz, and M. Hollick, "Smartphones with UWB: Evaluating the Accuracy and Reliability of UWB Ranging," 2023.