IEEE Open Journal of the
Computer Society

Received 25 March 2021; accepted 19 April 2021. Date of publication 26 April 2021;
date of current version 28 May 2021. The review of this article was arranged by Associate Editor Donato Impedovo.

Digital Object Identifier 10.1109/OJCS.2021.3075469

Speaker Identification for Business-Card-Type Sensors

SHUNPEI YAMAGUCHI [1] (Member, IEEE), RITSUKO OSHIMA [2], JUN OSHIMA [2], RYOTA SHIINA [3], TAKUYA FUJIHASHI [1] (Member, IEEE), SHUNSUKE SARUWATARI [1] (Member, IEEE), AND TAKASHI WATANABE [1] (Member, IEEE)

[1] Graduate School of Information Science and Technology, Osaka University, Suita, Osaka 565-0871, Japan
[2] Graduate School of Integrated Science and Technology, Shizuoka University, Hamamatsu, Shizuoka 432-8011, Japan
[3] NTT Access Network Service Systems Laboratories, NTT Corporation, Musashino, Tokyo 180-8585, Japan

CORRESPONDING AUTHOR: SHUNPEI YAMAGUCHI (e-mail: yamaguchi.shunpei@ist.osaka-u.ac.jp)

This work was supported in part by JSPS KAKENHI under Grant 19H01714, NTT Corporation, and in part by JST, PRESTO under Grant JPMJPR2032, Japan.

ABSTRACT Human collaboration has a great impact on the performance of multi-person activities. The analysis of speaker information and speech timing can be used to extract human collaboration data in detail. Some studies have extracted human collaboration data by identifying a speaker with business-card-type sensors. However, it is difficult to realize speaker identification for business-card-type sensors at low cost and high accuracy because of spikes in the measured sound pressure data, ambient noise in the non-speaker sensor, and synchronization errors across each sensor. This study proposes a novel sound pressure sensor and speaker identification algorithm to realize speaker identification for business-card-type sensors. The sensor extracts the user's speech at low cost and high accuracy by employing a peak hold circuit and time synchronization module for spike mitigation and precise time synchronization. The algorithm identifies a speaker with high accuracy by removing ambient noise. The evaluations show that the algorithm accurately identifies a speaker in a multi-person activity considering varying numbers of users, environmental noises, and reverberation conditions as well as long or short utterances. In addition, the peak hold circuit enables accurate extraction of speech and the synchronization error between the sensors is always within $\pm 30\ \mu$s, that is, negligible error.

INDEX TERMS Human activity recognition, sensor networks, speaker identification, speaker recognition, time synchronization.

I. INTRODUCTION

Human collaboration has a great impact on the effectiveness of multi-person activities; examples include collaborative work and learning. For example, the literature [1] finds that in collaborative learning, learners using the same problem-solving methods tend to consistently produce higher learning outcomes. Some studies have used speaker information in multi-person activities to estimate human collaboration [2]–[5]. They found that the analysis of speaker information and speech timing can be used to extract detailed information regarding the collaboration, such as the most active group and the conversation patterns of the members who lead the activity.

Some studies have attempted to identify the speaker in multi-person activities using microphones, including methods such as speaker localization [6]–[19], speaker verification using voice features [20]–[29], speaker identification using voice features [20], [23], [30]–[47], and speaker recognition using a mobile device [2]–[5], [48], [49]. For example, speaker localization determines the location of the speakers from multiple sound sources using a microphone or microphone array. The abovementioned studies consider using microphones with a high sampling rate of several kHz or higher for speaker recognition. In this study, we focus on speaker identification using sound pressure sensors with a low sampling rate. Speaker identification based on a low sampling rate sensor can extract collaboration data in multi-person activities, even using low-price and low-power-consumption mobile devices.

One of the pioneer studies is Rhythm [5]. Rhythm uses a mobile device, that is, a business-card-type sensor with a 700 Hz sound pressure sensor. Each speaker employs the sensor, and Rhythm uses a series of integration circuits, voice activity detection (VAD) [50], and thresholding algorithms for speaker identification. However, it is difficult to accurately identify a speaker using a sound pressure sensor because of the following issues.

The first issue is spikes in the measured sound pressure data. The integration circuit in Rhythm experiences spikes in the measured sound pressure, and these spikes may cause incorrect speech detection. The second issue is the classification of speech and noise based on the measured sound pressure. Even when spikes can be removed from the sound pressure, each element of speech affects the sound pressure of the business-card-type sensors of the non-speakers. For accurate speaker identification, the sensor for each speaker needs to classify the sound pressure into his/her speech or ambient noise. The third issue is a synchronization error across the sensor for each speaker. For classification, one solution is to cooperatively use sound pressure across the sensor of the speaker. However, there is no time synchronization module used with the sensors of Rhythm. This causes synchronization errors across the sensors, and errors also cause an incorrect chronological order of the sound pressure data across the sensors. In this case, classification is difficult, even with the sound pressure data of the speakers.[1]

To resolve these three issues, we propose the following: 1) a sound pressure sensor for business-card-type sensors and 2) a high-accuracy speaker identification algorithm using sound pressure data with a low sampling rate. To realize the proposed scheme, we implement a novel business-card-type sensor, namely, the Sensor-based Regulation Profiler Badge. Here, the business-card-type sensor is assumed to be worn on the chest of the speaker. For accurate identification, our study contains the following elements:

- Our sound pressure sensor uses a peak hold circuit for spike mitigation.
- We propose three-step speaker identification for removal of the effects of ambient noise.
- We implement a flooding-based synchronization module on our business-card-type sensor for precise time synchronization.

From the evaluations, we found that 1) the peak hold circuit removes the spikes from the measured sound pressure data, 2) the experiments show the effectiveness of the proposed scheme under different numbers of users, environmental noises, and reverberation conditions as well as for long or short utterances, and 3) the synchronization error between the sensors is always within $\pm 30\,\mu$s.

[1]To prevent meaningless analysis resulting from time synchronization error, the time synchronization accuracy should be less than one-tenth of the maximum sampling rate of the sensor.

The remainder of this paper is organized as follows: Section II describes related studies. In Section III, we present the proposed system for identifying a speaker, including the sound pressure sensor design (Section III-B) and speaker identification algorithm (Section III-C). Section IV describes the implementation of a business-card-type sensor for speaker identification. Experimental and simulation evaluations are conducted in Section V. Finally, Section VI concludes our paper.

II. RELATED WORKS

Our study is related to studies on speaker recognition using stationary and mobile devices.

A. SPEAKER RECOGNITION USING STATIONARY DEVICE

Existing studies can be classified into speaker localization, speaker verification, and speaker identification using voice features. Speaker localization [6]–[14] finds the location of the speaker from multiple sound sources. There are various applications that use speaker localization, such as mobile robots [15]–[17], passive sonar [18], and hearing aids [19]. For example, in response to wideband noise, the literature [17] proposes a method for distinguishing the time difference of arrival (TDOA) of sources and noise to estimate the position of the speaker.

Studies on speaker verification [20]–[26] compare the voice of a speaker with that of a pre-registered person for authentication. Speaker verification is used for Internet of things (IoT) device authentication [27], network security [28], and user authentication [29]. For example, the literature [26] combines mel-frequency cepstral coefficients (MFCC) and linear predictive coding (LPC) to improve the performance of speaker verification for low-quality input speech signals.

Some studies have realized speaker identification [20], [23], [30]–[44] by comparing the voice of a speaker with the voice of a pre-registered person. Speaker identification has been applied to video conferences [45], criminal investigations [46], and television programs [47]. For example, the literature [45] improves the robustness of speaker identification by identifying key speakers during a video conference, partially discarding information originating from inactive participants and reducing the interference caused by their temporary speech.

However, the abovementioned studies require high hardware and processing costs because a voice must be sampled at a high frequency of several kHz or more using a microphone. Our study uses a business-card-type sensor that samples sound pressure data at 100 Hz for speaker identification. It can reduce hardware and processing costs for identification, thus enabling the extraction of collaboration data in multi-person activities.

B. SPEAKER RECOGNITION USING MOBILE DEVICE

Some studies [48], [49] have realized speaker identification using a smartphone or a business-card-type sensor for the extraction of collaboration data in organizations [2]–[4] and human interaction [5]. For example, Hitachi's business microscope [2]–[4] uses a business-card-type sensor to identify a

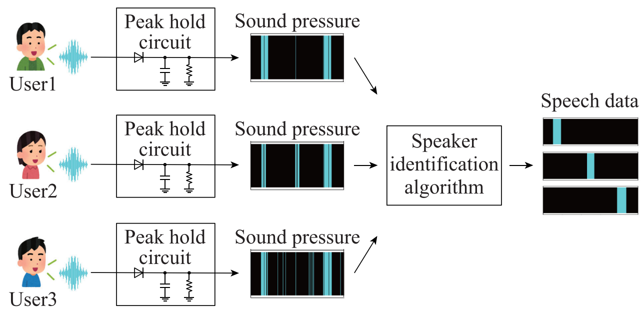


FIGURE 1. Overview of the proposed speaker identification technology.

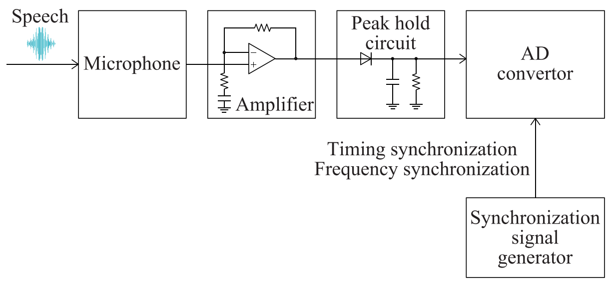


FIGURE 2. Sound pressure sensor.

speaker with an accuracy of 97.3 %. We note that the business microscope demonstrates high power consumption by the sound pressure sensor because of its high sampling rate of 8 kHz.

In addition, the MIT's Rhythm [5] also uses a business-card-type sensor called Rhythm Badge for speaker identification. Rhythm Badge consumes less power because it samples sound pressure at 700 Hz. It realizes speaker identification based on thresholding without the extraction of voice features; however, its identification accuracy is not high because of spikes in the measured sound pressure data using an integration circuit, the fixed threshold that allows ambient noise to cause errors, and the lack of time synchronization between sensors.

We design a novel business-card-type sensor to realize spike mitigation using a peak hold circuit and a speaker identification algorithm to remove the effect of ambient noise, and we incorporate high-precision time synchronization between sensors. Our experiments and simulations demonstrate that these steps can improve the accuracy of speech detection and speaker identification.

III. PROPOSED SYSTEM: SPEAKER IDENTIFICATION

A. OVERVIEW OF PROPOSED SYSTEM

To identify the speaker using sound pressure sensors on a business-card-type sensor with a low sampling rate, we propose a sound pressure sensor and a speaker identification algorithm. Fig. 1 shows the overview of our proposed system. We employ the following steps to identify the speaker from the sound pressure data.

1) We distribute our business-card-type sensors to users prior to multi-person activity.
2) The sensors acquire user speech through the sound pressure sensor with the peak hold circuit during multi-person activity.
3) We collect the distributed business-card-type sensors from the users.
4) We extract sound pressure data from the collected business-card-type sensors and feed them into the proposed speaker identification algorithm.
5) The proposed algorithm extracts and visualizes the identification results.

B. SOUND PRESSURE ACQUISITION

Fig. 2 shows a sound pressure sensor. The sensor samples sound pressure every 10 ms. A microphone converts user speech into electrical signals; because the converted signals are weak, the signals are amplified. The amplified signals are input into a peak hold circuit, which enables the detection of instantaneous signals using the discharge characteristics of an RC parallel circuit. An analog-to-digital (AD) converter converts the analog signal output from the peak hold circuit to digital signals. The digital signals are output every 10 ms with timing and frequency synchronization using a synchronization signal generator.

Our sound pressure sensor is simple and inexpensive. Specifically, the circuit consists of a microphone, an operational amplifier, a peak hold circuit, and an AD converter. This simple circuit allows the implementation of a sound pressure sensor at a low cost.

C. SPEAKER IDENTIFICATION ALGORITHM

Fig. 3 shows an overview of the proposed speaker identification algorithm. There are three steps for speaker identification: 1) pre-processing of sound pressure data, 2) speech section estimation, and 3) speaker identification.

1) Pre-Processing: The first step extracts the sound pressure detection for each user. The algorithm calculates the minimum sound pressure value for each user and subtracts the minimum value from all the sound pressure data to make a zero-point correction. The algorithm labels whether each user speaks with sliding windows for the sound pressure data of each user obtained by zero-point correction for each window. Algorithm 1 exhibits the labeling procedure in Fig. 3, and Table 1 lists the algorithm notation. Algorithm 1 outputs the array $\mathbb{A}$, which represents "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 $\mathbb{S} = \{S_1, S_2, \ldots, S_{|U|}\}$. We find the maximum of the sound pressure m for each user in each window W in line 6. If the maximum m in window W does not exceed the speech threshold η_s across all users, it is assumed that the speech of the user is not detected in window W, and the window slides in line 16. If the maximum m in window W exceeds the speech threshold η_s, the algorithm updates a threshold η_m as $m * 0.1$ in line 8. The algorithm compares the sound pressure of a user with the threshold η_m and assigns 1 if the sound pressure is higher than the threshold and 0 if the sound pressure is lower

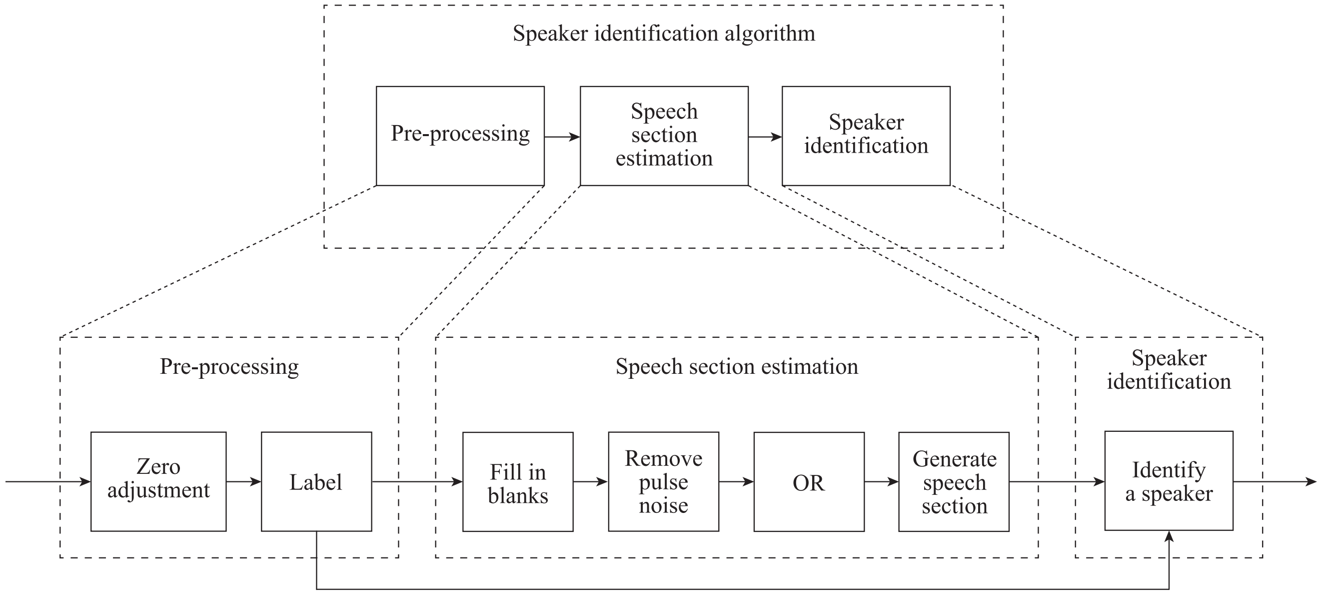


FIGURE 3. **Overview of the speaker identification algorithm.**

TABLE 1. **Notation**

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
ξ	Top index of window
D	Window size
η_s	Speech threshold for all users
η_m	Speech threshold based on maximum sound pressure in the window
$\max(X)$	Calculate the maximum of all the elements of X

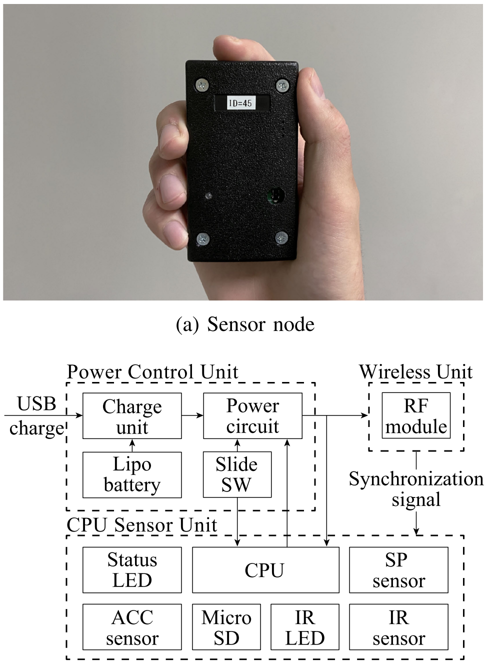
Algorithm 1: Labeling in pre-processing

Require: U, $\mathbb{S}$
Ensure: $\mathbb{A}$
1: **for all** $d \in U$ **do**
2: Insert zeros into all elements of A_d
3: $\xi \Leftarrow 0$
4: **while** $\xi <$ length of A_d **do**
5: $W \Leftarrow S_d \in \mathbb{S}$ between ξ to $\xi + D$
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: Insert $w \in W$ into elements of A_d with OR
15: **end if**
16: $\xi \Leftarrow \xi +$ slide width
17: **end while**
18: Insert A_d into $\mathbb{A}$
19: **end for**
20: **return** $\mathbb{A}$

than the threshold in lines 9–13. The labels w in window W overwrite the corresponding elements of array A_d in line 14. We call the data obtained through pre-processing "the 1–0 data for each user."

2) Speech Section Estimation: The second step extracts the presence or absence of a user's speech from the 1–0 data for each user. The algorithm fills the data using the 1–0 data for each user. The algorithm complements labels 1 in a section with consecutive labels 0 within 90 ms between labels 1 considered in the middle of speech in the 1–0 data for each user. The algorithm removes pulse noise using 1–0 data for each user with complements. The algorithm replaces a short interval with continuous labels 1 within 150 ms by labels 0, assuming that the section is where speech is falsely detected by ambient noise. The algorithm takes the logical summation of the 1–0 data for each user with pulse noise removal. We call the binary data obtained through the speech section estimation "the speech section data."

3) Speaker Identification: The third step determines who speaks in each speech section by combining the 1–0 data for each user and speech section data. The algorithm focuses on each section where a user is considered to speak based on the speech section data. The algorithm extracts a user with the most labels 1 in each speech section and regards the user as a speaker in the speech section on the basis of the 1–0 data for each user.



(a) Sensor node

(b) Block diagram

FIGURE 4. **Sensor-based Regulation Profiler Badge.**

IV. IMPLEMENTATION: BUSINESS-CARD-TYPE SENSOR

We implement a novel business-card-type sensor, namely, the Sensor-based Regulation Profiler Badge. Figs. 4 (a) and (b) show the proposed Sensor-based Regulation Profiler Badge (sensor node) and its block diagram. The sensor node consists of a power control unit, CPU sensor unit, and wireless unit.

The power control unit has a lithium-ion battery that drives the sensor node. The lithium-ion battery supplies power to the power switch and microcontroller unit (MCU). The sensor node can continuously run for 24 h.

The CPU sensor unit is equipped with a STM32L476RGT6 from STMicroelectronics as the MCU, an ADXL362 accelerometer (ACC sensor) from Analog Devices, an OSI5LAS1C1A infrared light emitting diode (IR LED) from OptoSupply, a PIC79603 infrared receiver (IR sensor) from Kodenshi Corp., and an INMP510 analog microphone in the sound pressure sensor (SP sensor) from TDK. The three-axis accelerometer samples 12 bits at 100 Hz, and the sound pressure sensor samples 12 bits at 100 Hz. The microSD card connector of a DM3AT-SF-PEJM5 from Hirose Electric is used to record the sensor data. The acceleration, infrared, and sound pressure data are recorded on a microSD card.

The wireless unit uses a CC2650 from Texas Instruments, which contains a wireless synchronization module. The wireless synchronization module transfers a synchronization signal sent every 10 ms from a synchronizer (sync node) to other sensor nodes to synchronize the time between the sensor nodes. CC2650 uses UNISONet, which is also known as Choco [51], [52], to realize precise time synchronization between the sensor nodes. In Choco, an arbitrary sensor node forwards a time-synchronous packet to the neighboring sensor nodes and then propagates the received time-synchronous packet to the destination node. When a sensor node receives a new time-synchronous packet from a neighboring sensor node, it immediately forwards the packet to all neighboring sensor nodes. Each sensor node repeatedly receives and forwards time-synchronous packets by flooding, resulting in the fast propagation of time-synchronous packets throughout the sensor nodes.

V. EVALUATION

A. SPEAKER IDENTIFICATION ACCURACY

We experimentally evaluated the accuracy of the speaker identification algorithm using the sound pressure data obtained from existing and proposed business-card-type sensors. We carried out the experiment in a conference room considering different numbers of users, environmental noises, and reverberation conditions as well as long or short utterances. All users were male university students in their early 20 s. The dimensions of the conference room were 10.6 m × 7.05 m × 2.65 m. In each experiment, each user wore a sensor node on his chest and sat on a chair 1.50 m away from adjacent users around the table. We set a sync node at the center of the table for time synchronization between the sensor nodes.

For the experiments of long and short utterances, we prepared two types of speech scripts for each user. Table 2 shows the prepared script for the experiments of long and short utterances. Specifically, all the users spoke a sentence in Table 2 in order with a two-second interval to avoid speech overlapping. After all users spoke a sentence, they started to speak the next sentence.

We compared the speaker identification accuracy of our proposed scheme with that of Rhythm [5]. Rhythm identifies a speaker using only sound pressure. Both algorithms use sliding window-based speech detection for speaker identification. Rhythm uses the VAD and thresholding algorithms for speaker identification. Here, the identification accuracy depends on the window size, slide width, and speech detection threshold. Because each user began his speech after a two-second interval from the speech of the former user, we set the window size to two seconds in both algorithms to include at most one speech in each window. We set the slide width to 0.01 seconds and one second for Rhythm and the proposed algorithm. The slide width in Rhythm is to the same as in the literature [5]. The slide width in the proposed algorithm is the best parameter, which achieved the most accurate speaker identification in the preliminary evaluations using varying slide widths. The optimal value of the speech detection threshold for Rhythm and the proposed algorithm depends on the evaluation settings.

1) THE NUMBER OF USERS

We show the speaker identification accuracy considering different numbers of users from two to five using the script of long utterances in Table 2. We set the speech thresholds of

TABLE 2. Speech Script Prepared for the Experiments

Order	Sentence for experiments of long utterances	Sentence for experiments of short utterances
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.

TABLE 3. Confusion Matrices Under the Different Numbers of Users

	Two users				Three users				Four users				Five users			
	Rhythm		Proposed		Rhythm		Proposed		Rhythm		Proposed		Rhythm		Proposed	
	P	N	P	N	P	N	P	N	P	N	P	N	P	N	P	N
T	30	0	30	0	44	1	44	1	58	2	60	0	73	2	75	0
F	0	30	0	30	0	45	0	45	0	60	0	60	0	75	0	75

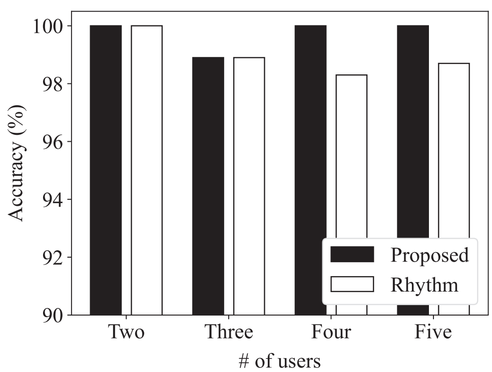


FIGURE 5. Speaker identification accuracy under the different numbers of users.

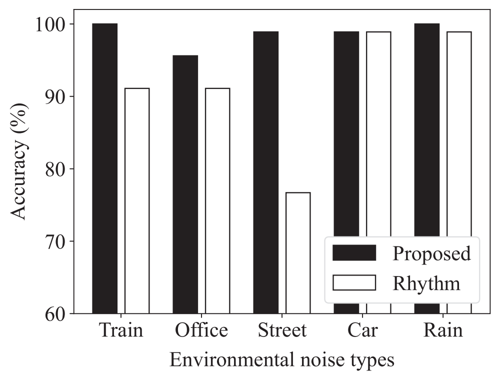


FIGURE 6. Speaker identification accuracy under the different environmental noises.

82 dB and 75 dB to Rhythm and the proposed algorithm. The thresholds were appropriately met irrespective of the number of users.

Fig. 5 shows the speaker identification accuracy of the proposed scheme and Rhythm and Table 3 shows the corresponding confusion matrices for two through five users. In this case, the F1-scores of Rhythm are 0.667, 0.662, 0.659, and 0.661, whereas the F1-scores of the proposed scheme are 0.667, 0.662, 0.667, and 0.667. We can see that the F1-score of Rhythm is lower than that of the proposed scheme for four and five users. Because Rhythm uses a single speech threshold across users, the threshold does not detect speech in some users.

2) ENVIRONMENTAL NOISE

To evaluate the effect of environmental noise on the speaker identification accuracy, we prepared a noise source in our environment. The experiments were conducted for three users. The noise source was set on the ceiling of the room 2 m away from the center of the table. The noise source used five types of ambient noise recorded in trains, offices, streets, cars, and rain. Other considerations were the same as the experiments in Fig. 5. We set the average sound volume of each noise as 75 dB in the train, 70 dB in the office and street, and 60 dB for cars and rain. We set the speech thresholds of Rhythm to 89 dB, 86 dB, 89 dB, 84 dB, and 85 dB for train, office, street, car, and rain noises. We also set the speech thresholds of the proposed algorithm to 84 dB, 85 dB, 84 dB, 83 dB, and 80 dB for train, office, street, car, and rain noises.

Fig. 6 shows the speaker identification accuracy under the different environmental noises, and Table 4 shows the corresponding confusion matrices. In this case, the F1-scores of Rhythm for the ambient noises of train, office, street, car, and rain are 0.622, 0.651, 0.536, 0.662, and 0.662 whereas those of the proposed scheme are 0.667, 0.651, 0.662, 0.662, and 0.667. We can see that the proposed scheme achieves a better F1-score compared with Rhythm irrespective of the environmental noise type.

TABLE 4. Confusion Matrices Under the Different Environmental Noises

	Train				Office				Street				Car				Rain			
	Rhythm		Proposed		Rhythm		Proposed		Rhythm		Proposed		Rhythm		Proposed		Rhythm		Proposed	
	P	N	P	N	P	N	P	N	P	N	P	N	P	N	P	N	P	N	P	N
T	37	8	45	0	42	3	42	3	26	19	44	1	44	1	44	1	44	1	45	0
F	0	45	0	45	5	40	1	44	2	43	0	45	0	45	0	45	0	45	0	45

TABLE 5. Simulation Environments for Reverberation Conditions

	Room dimensions (x, y, z)	Reverberation time (s)
Small room	(5, 4, 3)	0.3
Medium room	(7, 6, 4)	0.6
Large room	(9, 8, 7)	0.9
Actual room	(10.6, 7.05, 2.65)	0.9

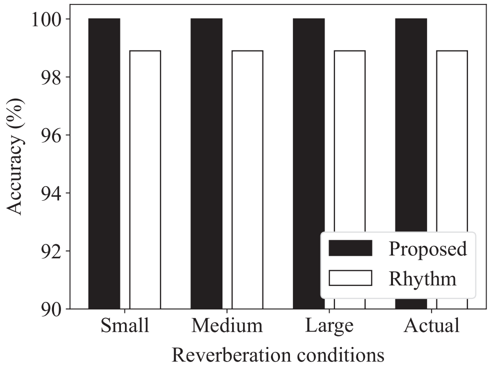


FIGURE 7. Speaker identification accuracy under the different reverberation conditions.

3) REVERBERATION CONDITIONS

We used trace-driven simulation to evaluate the effect of reverberation conditions. We first recorded the sound pressure signals of three users from the experiments in Fig. 5, and then added the effect of the reverberation conditions on the signals using the room impulse response generator [53]. The room impulse response generator simulates the impulse response considering the reverberation conditions of a room including the room dimensions, source position, receiver position, and reverberation time. We considered four room conditions based on [54]: small, medium, and large and the same room dimensions as in our experiment. Table 5 shows the assumed room conditions. We set thresholds of 83 dB and 80 dB to Rhythm and the proposed algorithm irrespective of the room dimensions.

Fig. 7 shows the speaker identification accuracy under the different reverberation conditions and Table 6 shows the corresponding confusion matrices. In this case, the F1-scores of Rhythm for the four room conditions of the small, medium, large, and actual rooms are 0.662, whereas those of the proposed scheme are 0.667. We can see that Rhythm and the proposed scheme achieves almost the same accuracy irrespective of the reverberation conditions.

4) SHORT UTTERANCES

We evaluated the effect of short utterances, speeches of less than one second [33], using the script of short utterances in Table 2. The experiments were conducted for three users. Other considerations were the same as the experiments in Fig. 5. We set the thresholds of 78 dB and 73 dB to Rhythm and the proposed algorithm.

Table 7 shows the corresponding confusion matrices. In this case, the accuracy of Rhythm and the proposed scheme for short utterances are 88.9 % and 97.8 % and their F1-scores are 0.651 and 0.657, respectively. Even in a short utterance, the proposed scheme achieves a better F1-score than Rhythm.

B. IMPACT OF SOUND PRESSURE SENSOR

Figs. 8 (a) and (b) show the circuit diagrams of a sound pressure sensor in the Rhythm Badge and our Sensor-based Regulation Profiler Badge. Rhythm Badge, which is based on Open Badge [55], uses an integration circuit, whereas the Sensor-based Regulation Profiler Badge uses a peak hold circuit for sound pressure acquisition. Each circuit parameter in the proposed Sensor-based Regulation Profiler Badge is determined to achieve the following three purposes in the proposed circuits.

- Noises in low-frequency components, i.e., less than 20 Hz, should be removed from the sound pressure data because they are not related to users' speech.
- The sound pressure data should be amplified 100 times to detect detailed changes in each user's voice volume.
- The beginning and end of each speech section should be accurately extracted from the sound pressure data by adjusting the discharging slope of a resistor capacitor (RC) circuit.

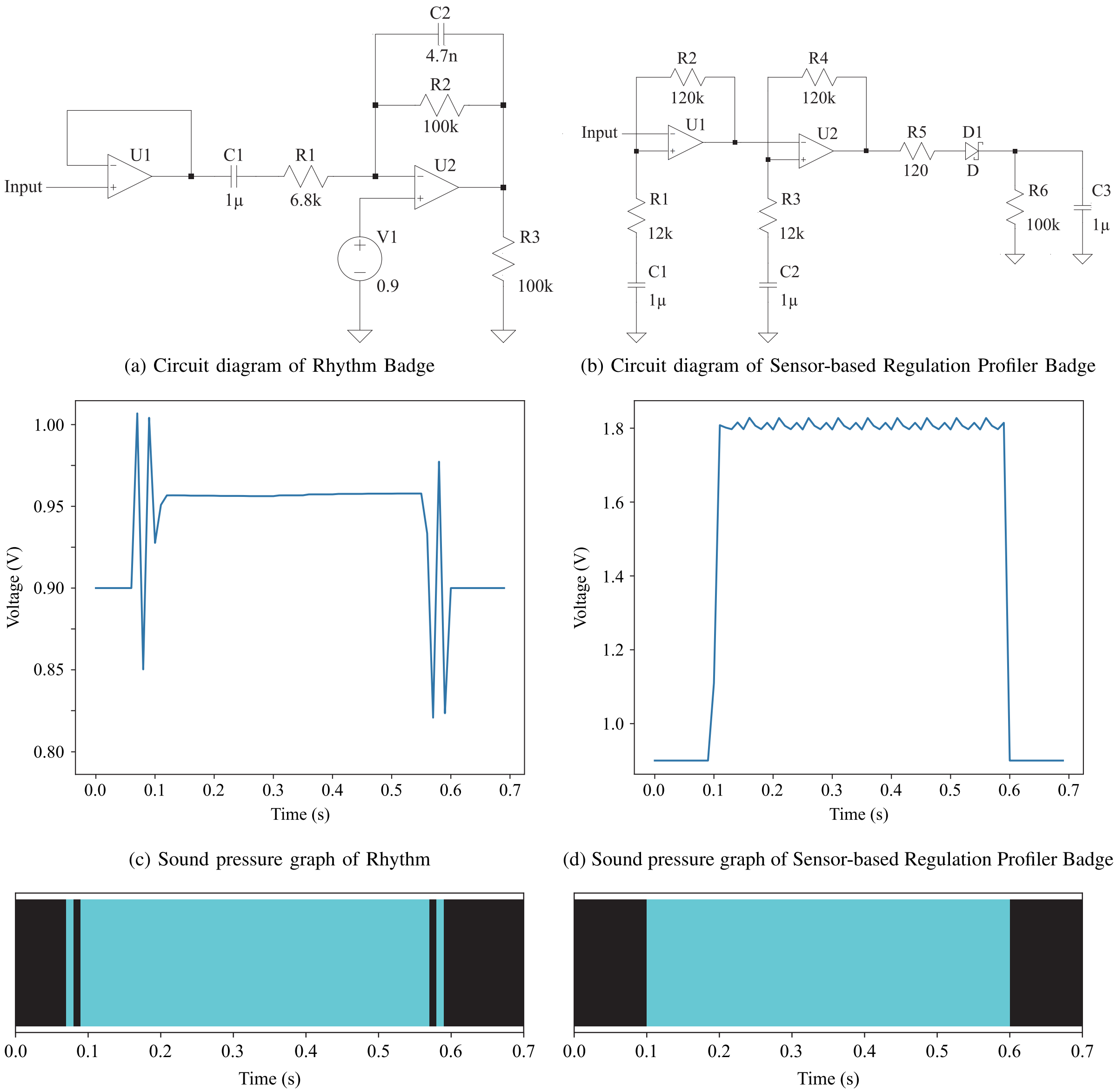
Circuit simulations were conducted for each circuit. Here, we regarded a sinusoidal wave as the speech of a user. The amplitude of the sinusoidal wave was 0.8 V at a frequency of 340 Hz, with a length of 500 ms. In addition, we used a direct current (DC) signal with an amplitude of 0.9 V and a length of 100 ms to represent silence. The DC signal was inserted before and after the sinusoidal wave.

Figs. 8 (c) and (d) show the measured sound pressure obtained by Rhythm and the Sensor-based Regulation Profiler Badge as a function of elapsed time. Rhythm leaves spikes at the beginning and end of the measured sound pressure by inputting the sinusoidal wave into an integration circuit. Conversely, the measured sound pressure from the Sensor-based Regulation Profiler Badge has no spikes at the beginning nor end of the sound pressure data because it uses the peak hold circuit.

To discuss the effect of the measured sound pressure data, we adopted a threshold-based speech detection algorithm for both Rhythm and our business-card-type sensor. We set the sound pressure threshold to detect the edges of the section

TABLE 6. Confusion Matrices Under the Different Reverberation Conditions

	Small room				Medium room				Large room				Actual room			
	Rhythm		Proposed		Rhythm		Proposed		Rhythm		Proposed		Rhythm		Proposed	
	P	N	P	N	P	N	P	N	P	N	P	N	P	N	P	N
T	44	1	45	0	44	1	45	0	44	1	45	0	44	1	45	0
F	0	45	0	45	0	45	0	45	0	45	0	45	0	45	0	45



(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 8. Simulation results for speech detection in Rhythm and Sensor-based Regulation Profiler Badge.

of user speech. As shown in Fig. 8 (c), the power of the measured sound pressure in Rhythm before and after spikes was approximately 0.90 V and 0.95 V. In this case, we set the threshold to 0.92 V to reduce the effect of the spikes. As shown in Fig. 8 (d), the power of the measured sound pressure in the proposed scheme before and after the speech of the user was 0.9 V and 1.8 V. In this case, the threshold between 0.9 V and 1.8 V achieved almost the same performance; thus, we considered the same speech threshold of 0.92 V for the proposed scheme.

Figs. 8 (e) and (f) show the results of the threshold-based speech detection using Rhythm and the proposed Sensor-based Regulation Profiler Badge. It is difficult for Rhythm to extract the speech of a user accurately using threshold-based

TABLE 7. Confusion Matrices of Short Utterances

	Rhythm		Proposed	
	Positive	Negative	Positive	Negative
True	42	3	43	2
False	7	38	0	45

speech detection because spikes remain in the measured sound pressure data. Our Sensor-based Regulation Profiler Badge can accurately detect the speech of a user because the measured sound pressure has no spikes. The results in Figs. 8 (e) and (f) suggest that the peak hold circuit may detect the speech of a user more accurately than the integration circuit, that is, Rhythm.

C. TIME SYNCHRONIZATION PRECISION

We experimentally evaluated the time synchronization accuracy between the sync and sensor nodes in the proposed Sensor-based Regulation Profiler Badge. We set up a sync node and a sensor node at a short distance from each other on a desk and measured the time deviation between the nodes based on the synchronization signals sent from the sync node. An oscilloscope was used to measure the clock rise time at each node to accurately obtain the time deviation between the nodes. We assumed that the number of samples was 30 003, and the wireless synchronization module of each Sensor-based Regulation Profiler Badge transmitted a synchronization signal every 10 ms.

The results show that the time synchronization error is maintained within $\pm 30\,\mu$s. Here, the mean and maximum synchronization errors are $-7.7\ \mu$s and 30 μs. The obtained synchronization error is well below the required synchronization accuracy of 1 ms because the sampling rate of the pressure sensor in the sensor node is 100 Hz. Because the sensors realize accurate synchronization, they maintain the time series of the speech of each user. Accurate time-series data can realize speaker identification with high accuracy. Periodic correction of synchronization frequencies between the sync node and the sensor node maintains synchronization errors within $\pm 30\,\mu$s, suggesting accurate speaker identification by combining sensor data from multiple sensors.

VI. CONCLUSION

In this study, we proposed a novel sound pressure sensor and speaker identification algorithm for business-card-type sensors to extract collaboration characteristics in multi-person activities. The sound pressure sensor employs a peak hold circuit and time synchronization module for spike mitigation and precise time synchronization between sensors to detect the speech of a user at low cost and high accuracy. The algorithm removes ambient noise from non-speaker sensors to identify a speaker with high accuracy. We found that the evaluations show the effectiveness of the proposed scheme under different numbers of users, environmental noises, and reverberation conditions as well as for long or short utterances. In addition, the peak hold circuit accurately extracts the speech of a user and the synchronization error between the sensors is always within $\pm 30\,\mu$s.

REFERENCES

[1] J. Oshima, R. Oshima, and K. Fujii, "Student regulation of collaborative learning in multiple document integration," in *Proc. Int. Conf. Learn. Sci.*, 2014, pp. 967–971.

[2] J. Nishimura, N. Sato, and T. Kuroda, "Speech "Siglet" detection for business microscope," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun.*, 2008, pp. 147–152.

[3] J. Nishimura and T. Kuroda, "Speaker recognition using speaker-independent universal acoustic model and synchronous sensing for business microscope," in *Proc. Int. Symp. Wireless Pervasive Comput.*, 2009, pp. 1–5.

[4] J. Nishimura and T. Kuroda, "Hybrid speaker recognition using universal acoustic model," *SICE J. Control, Measurement, System Integration*, vol. 4, no. 6, pp. 410–416, 2011.

[5] 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, Jan.–Mar. 2018.

[6] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 4, pp. 320–327, Aug. 1976.

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

[8] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*, 2001, ch. 8, pp. 157–180.

[9] S. Mohan, M. E. Lockwood, M. L. Kramer, and D. L. Jones, "Localization of multiple acoustic sources with small arrays using a coherence test," *J. Acoustical Soc. Amer.*, vol. 123, no. 4, pp. 2136–2147, 2008.

[10] W. Zhang and B. D. Rao, "A two microphone-based approach for source localization of multiple speech sources," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 1913–1928, Nov. 2010.

[11] 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 *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 3302–3306.

[12] 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 Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 12, pp. 2444–2453, Dec. 2017.

[13] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[14] 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, Language Processing*, vol. 27, no. 1, pp. 178–188, Jan. 2019.

[15] J.-M. Valin, F. Michaud, J. Rouat, and D. Létourneau, "Robust sound source localization using a microphone array on a mobile robot," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2003, pp. 1228–1233.

[16] J.-M. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *Robot. Auton. Syst.*, vol. 55, no. 3, pp. 216–228, 2007.

[17] F. Grondin and F. Michaud, "Time difference of arrival estimation based on binary frequency mask for sound source localization on mobile robots," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2015, pp. 6149–6154.

[18] E. L. Ferguson, S. B. Williams, and C. T. Jin, "Sound source localization in a multipath environment using convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 2386–2390.

[19] M. Zohourian, G. Enzner, and R. Martin, "Binaural speaker localization integrated into an adaptive beamformer for hearing aids," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 3, pp. 515–528, Mar. 2018.

[20] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1711–1723, Jul. 2007.

[21] 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 *Proc. Int. Symp. Chin. Spoken Lang. Process.*, 2018, pp. 205–209.

[22] 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 *Proc. Int. Conf. Electron. Comput. Technol.*, 2011, pp. 317–320.

[23] S. Nakagawa, L. Wang, and S. Ohtsuka, "Speaker identification and verification by combining MFCC and phase information," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1085–1095, May 2012.

[24] A. Roy, M. Magimai.-Doss, and S.Marcel, "A fast parts-based approach to speaker verification using boosted slice classifiers," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 1, pp. 241–254, Feb. 2012.

[25] H. Taherian, Z.-Q. Wang, J. Chang, and D. Wang, "Robust speaker recognition based on single-channel and multi-channel speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1293–1302, 2020, doi: 10.1109/TASLP.2020.2986896.

[26] A. Chowdhury and A. Ross, "Fusing MFCC and LPC features using 1D triplet CNN for speaker recognition in severely degraded audio signals," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 1616–1629, 2020, doi: 10.1109/TIFS.2019.2941773.

[27] D.-G. Shin and M.-S. Jun, "Home IoT device certification through speaker recognition," in *Proc. Int. Conf. Adv. Commun. Technol.*, 2015, pp. 600–603.

[28] Z. Wu *et al.*, "Anti-spoofing for text-independent speaker verification: An initial database, comparison of countermeasures, and human performance," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 4, pp. 768–783, Apr. 2016.

[29] L. Yang, Z. Zhao, and G. Min, "User verification based on customized sentence reading," in *Proc. IEEE Int. Conf. Cyber Sci. Technol. Congr.*, 2018, pp. 353–356.

[30] N. McLaughlin, J. Ming, and D. Crookes, "Speaker recognition in noisy conditions with limited training data," in *Proc. Eur. Signal Process. Conf.*, 2011, pp. 1294–1298.

[31] G. Biagetti, P. Crippa, L. Falaschetti, S. Orcioni, and C. Turchetti, "Speaker identification in noisy conditions using short sequences of speech frames," in *Proc. Smart Innovation, Syst. Technol.*, 2018, pp. 43–52.

[32] S. MangeshDeshpande and Raghunath S.Holambe, "Speaker identification based on robust AM-FM features," in *Proc. Int. Conf. Emerg. Trends Eng. Technol.*, 2009, pp. 880–884.

[33] G. Biagetti, P. Crippa, A. Curzi, S. Orcioni, and C. Turchetti, "Speaker identification with short sequences of speech frames," in *Proc. Int. Conf. Pattern Recognit. Appl. Methods*, 2015, pp. 178–185.

[34] D. A. Reynolds, "Experimental evaluation of features for robust speaker identification," *IEEE Speech Audio Process.*, vol. 2, no. 4, pp. 639–643, Oct. 1994.

[35] Z. Wu and Z. Cao, "Improved MFCC-based feature for robust speaker identification," *Tsinghua Sci. Technol.*, vol. 10, no. 2, pp. 158–161, 2005.

[36] S. Chakroborty, A. Roy, and G. Saha, "Fusion of a complementary feature set with MFCC for improved closed set text-independent speaker identification," in *Proc. IEEE Int. Conf. Ind. Technol.*, 2006, pp. 387–390.

[37] A. Maesa, F. Garzia, M. Scarpiniti, and R. Cusani, "Text independent automatic speaker recognition system using mel-frequency cepstrum coefficient and gaussian mixture models," *J. Inf. Secur.*, vol. 3, no. 4, pp. 335–340, 2012.

[38] B. G. Nagaraja and H. S. Jayanna, "Efficient window for monolingual and crosslingual speaker identification using MFCC," in *Proc. Int. Conf. Adv. Comput. Commun. Syst.*, 2013, pp. 1–4.

[39] 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 *Proc. Int. Symp. Wireless Commun. Syst.*, 2014, pp. 722–727.

[40] A. Bakshi, S. K. Kopparapu, S. Pawar, and S. Nema, "Novel windowing technique of MFCC for speaker identification with modified polynomial classifiers," in *Proc. Int. Conf. Confluence The Next Gener. Inf. Technol. Summit*, 2014, pp. 292–297.

[41] P. M. Chauhan and N. P. Desai, "Mel frequency cepstral coefficients (MFCC) based speaker identification in noisy environment using wiener filter," in *Proc. Int. Conf. Green Comput. Commun. Elect. Eng.*, 2014, pp. 1–5.

[42] K. Matsumoto, N. Hayasaka, and Y. Iiguni, "Noise robust speaker identification by dividing MFCC," in *Proc. Int. Symp. Commun., Control Signal Process.*, 2014, pp. 652–655.

[43] S. S. Wali, S. M. Hatture, and S. Nandyal, "MFCC based text-dependent speaker identification using BPNN," *Int. J. Signal Process. Syst.*, vol. 3, no. 1, pp. 30–34, 2015.

[44] B. Ayoub, K. Jamal, and Z. Arsalane, "An analysis and comparative evaluation of MFCC variants for speaker identification over VoIP networks," in *Proc. World Congr. Inf. Technol. Comput. Appl.*, 2015, pp. 1–6.

[45] I. Volfin and I. Cohen, "Dominant speaker identification for multipoint videoconferencing," in *Proc. IEEE Conv. Elect. Electron. Engineers Isr.*, 2012, pp. 1–4.

[46] R. Karadaghi, H. Hertlein, and A. Ariyaeeinia, "Effectiveness in open-set speaker identification," in *Proc. Int. Carnahan Conf. Secur. Technol.*, 2014, pp. 1–6.

[47] J. Poignant, L. Besacier, and G. Quénot, "Unsupervised speaker identification in TV broadcast based on written names," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 1, pp. 57–68, Jan. 2015.

[48] K. Brunet, K. Taam, E. Cherrier, N. Faye, and C. Rosenberger, "Speaker recognition for mobile user authentication: An android solution," in *Proc. Conf. sur la Sécurité des Architectures Réseaux et Systémes d'Information*, 2013, pp. 1–10.

[49] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and RSR2015," *Speech Commun.*, vol. 60, pp. 56–77, 2014.

[50] 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 *Proc. IEEE Int. Conf. High Speed Netw. Multimedia Commun.*, 2002, pp. 46–50.

[51] M. Suzuki, C.-H. Liao, S. Ohara, K. Jinno, and H. Morikawa, "Wireless-transparent sensing," in *Proc. Int. Conf. Embedded Wireless Syst. Netw.*, 2017, pp. 66–77.

[52] F. Ferrari, M. Zimmerling, L. Thiele, and O. Saukh, "Efficient network flooding and time synchronization with glossy," in *Proc. ACM/IEEE Int. Conf. Inf. Process. Sensor Netw.*, 2011, pp. 73–84.

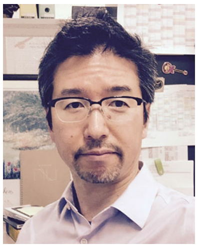
[53] E. A. P. Habets, "Room impulse response generator," 2010. [Online]. Available: https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator

[54] X. Zhao, Y. Wang, and D. Wang, "Robust speaker identification in noisy and reverberant conditions," *IEEE/ACM Trans. Audio,Speech, Lang. Process.*, vol. 22, no. 4, pp. 836–845, Apr. 2014.

[55] O. Lederman, D. Calacci, A. MacMullen, D. C. Fehder, F. E. Murray, and A. S. Pentland, "Open badges: A. low-cost toolkit for measuring team communication and dynamics," in *Proc. Int. Conf. Social Computing, Behav.-Cultural Model. Prediction Behav. Representation Model. Simul.*, 2016, pp. 1–7.

SHUNPEI YAMAGUCHI (Member, IEEE) received the Bachelor of Engineering degree from Osaka University, Japan, in 2020. He joined the Information Processing Society of Japan in 2018. He is currently a student with the Graduate School of Information Science and Technology, Osaka University. His research interest focuses on wearable sensors.

RITSUKO OSHIMA is a Professor with the Faculty of Informatics, Shizuoka University, Japan. She has been involved in a research project to develop a project-based learning curriculum at an engineering department for several years. Her current interest includes the development of a scenario-based questionnaire to evaluate students' collaboration skills.



JUN OSHIMA received Ph.D. for applied cognitive science at University of Toronto in 1995. In his recent work, he developed a social network analysis of discourse from the perspective of knowledge-building. His research interest includes the development of new methodologies to evaluate students' collective knowledge advancement. He is a Professor with the Faculty of Informatics, Shizuoka University, Japan.

RYOTA SHIINA received the M.E. degree in material science and engineering from the Tokyo Institute of Technology, Tokyo, Japan, in 2014. In 2014, he joined the NTT Access Network Service Systems Laboratories where he has been engaged in research on optical access systems mainly related to optical video distribution systems, optical radio-over-fiber transmission systems, and optical wireless communication systems. He was the recipient of the Young Researcher's Award from the Institute of Electronics, Information, and Communication Engineers (IEICE) of Japan in 2018, the Encouraging Award from IEICE Technical Committee on Communication Systems (CS) in 2017. He is a Member of IEICE.

TAKUYA FUJIHASHI (Member, IEEE) received the B.E. and M.S. degrees from Shizuoka University, Japan, in 2012 and 2013, respectively. In 2016, he received Ph.D. degree from the Graduate School of Information Science and Technology, Osaka University, Japan. He is currently an Assistant Professor with the Graduate School of Information Science and Technology, Osaka University since April 2019. He was an Assistant Professor with the Graduate School of Science and Engineering, Ehime University, from January 2017 to March 2019. He was Research Fellow (PD) of the Japan Society for the Promotion of Science in 2016. From 2014 to 2016, he was Research Fellow (DC1) of the Japan Society for the Promotion of Science. From 2014 to 2015, he was an Intern with Mitsubishi Electric Research Labs. (MERL) working with the Electronics and Communications Group. His research interests include the area of video compression and communications, with a focus on immersive video coding and streaming.

SHUNSUKE SARUWATARI (Member, IEEE) received the B.E. degree from The University of Electro-Communications, Japan, in 2002, and the M.S. and Ph.D. degrees from The University of Tokyo, Japan, in 2004 and 2007, respectively. In 2007, he was a Visiting Researcher with the Illinois Genetic Algorithms Laboratory, University of Illinois at Urbana-Champaign. From 2008 to 2011, he was a Research Associate with the Research Center for Advanced Science and Technology, The University of Tokyo. From 2012 to 2015, he was a Tenure-Track Assistant Professor with the Graduate School of Informatics, Shizuoka University, Japan. He is currently an Associate Professor with the Graduate School of Information Science and Technology, Osaka University, Japan. His research interests include the areas of wireless networks, sensor networks, and system software. He is a member of ACM, IPSJ, and IEICE.

TAKASHI WATANABE (Member, IEEE) received the B.E., M.E., and Ph.D. degrees from Osaka University, Japan, in 1982, 1984, and 1987, respectively. He is a Professor of Graduate School of Information Science and Technology, Osaka University, Japan. He joined Faculty of Engineering, Tokushima University as an Assistant Professor in 1987 and moved to Faculty of Engineering, Shizuoka University in 1990. He was a Visiting Researcher with University of California, Irvine from 1995 through 1996. He has served on many program committees for networking conferences, IEEE, ACM, IPSJ, IEICE (The Institute of Electronics, Information and Communication Engineers, Japan). His research interests include mobile networking, ad hoc networks, sensor networks, ubiquitous networks, intelligent transport systems, specially MAC and routing. He is a Member of IEEE Communications Society, IEEE Computer Society as well as IPSJ and IEICE.