Human-Machine Cooperative Echolocation Using Ultrasound

Watanabe, Hiroki

Sumiya, Miwa

Terada, Tsutomu

(Citation)
IEEE Access,10:125264-125278

(Issue Date)
2022

(Resource Type)
journal article

(Version)
Version of Record

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 20 October 2022, accepted 15 November 2022, date of publication 24 November 2022,
date of current version 5 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3224468

RESEARCH ARTICLE

Human-Machine Cooperative Echolocation Using Ultrasound

HIROKI WATANABE[1], MIWA SUMIYA[2], AND TSUTOMU TERADA[3], (Member, IEEE)
[1]Graduate School of Information Science and Technology, Hokkaido University, Sapporo 060-0814, Japan
[2]Japan Society for the Promotion of Science, Tokyo 102-0083, Japan
[3]Graduate School of Engineering, Kobe University, Kobe 657-8501, Japan

Corresponding author: Hiroki Watanabe (hiroki.watanabe@ist.hokudai.ac.jp)

This work was supported in part by the Japan Society for the Promotion of Science (JSPS) Grants-in-Aid for Scientific Research (KAKENHI) under Grant JP21K11973, in part by the Japan Science and Technology Agency (JST) Precursory Research for Embryonic Science and Technology (PRESTO) under Grant JPMJPR2138, and in part by the JST Core Research for Evolutional Science and Technology (CREST) under Grant JPMJCR18A3.

This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by the Human Ethics Committee of Graduate School of Engineering, Kobe University under Permission No. 04-02, and performed in line with the Declaration of Helsinki.

ABSTRACT Echolocation has been shown to improve the independence of visually impaired people, and utilizing ultrasound in echolocation offers additional advantages, such as a higher resolution of object sensing and ease of extraction from background sounds. However, humans cannot innately make and hear ultrasound. A wearable device that enables ultrasonic echolocation, i.e., that transmits ultrasound through an ultrasonic speaker and converts the reflected ultrasound into audible sound, has therefore been attracting interest. Such a system can be utilized with machine learning (ML) to help visually impaired users recognize objects. We have therefore been developing a cooperative echolocation system that combines human recognition with ML recognition. As the first step toward cooperative echolocation, this paper presents the effectiveness of ML in echolocation. We implemented a prototype device and evaluated the performance of object detection with/without ML and found that the mental workload on the user was significantly decreased when ML was used. Based on the findings from the evaluation, we discussed the design of cooperative echolocation.

INDEX TERMS Assistive technology, echolocation, object recognition, ultrasound, wearable computing.

I. INTRODUCTION

According to the World Health Organization, there are currently 2.2 billion visually impaired people in the world [1]; therefore, creating a supportive environment for them is vital. Tactile paving and braille are two elements of the support environments that are already installed around many urban areas [2]. However, as visually impaired people are often unable to take advantage of these clues until they directly touch them, we consider there is considerable room for improvement. A number of researchers have been developing systems for supporting visually impaired people by utilizing the wearable/mobile computing devices that have recently become more readily available. Most of these methods utilize a camera to recognize the environment [3], [4], [5], [6]; however, this can be problematic because the cameras are adversely affected by light conditions and thus are difficult to use in highly dark or bright conditions. Although depth cameras can mitigate these challenges to some extent, they are still adversely affected by sunlight because most of them utilize infrared sensors [7], [8], [9], [10].

Alternatively, sound has also been used for recognizing the environment. For example, animals such as bats and dolphins use echolocation, a biological sonar in which they emit ultrasound to the environment and listen to the echoes that return from various nearby objects. These echoes are extremely useful for locating and identifying the objects. Some visually impaired people make similar use of echolocation with the auditory information from the sounds they make [11], which has been shown to enhance their

The associate editor coordinating the review of this manuscript and approving it for publication was Sandra Baldassarri.

independence [12], [13]. Therefore, we consider it can be beneficial for visually impaired people to use echolocation to recognize the surroundings with their own auditory perception.

Humans typically use the audible sounds they make (e.g., clicking with the mouth or tapping with canes) for echolocation; however, reports have shown that ultrasonic echolocation has several advantages, such as a higher resolution of object sensing, an inaudible clicking sound, and ease of extraction from background sounds [14]. However, this potential remains untapped because humans cannot innately make and hear ultrasound on their own.

In light of this background, we focus on a wearable computing environment that enables the transmission and receiving of ultrasound in real time. Earphone-type wearable computing devices (hearables) have become more common, and several studies have manipulated external sound by means of hearables [15], [16]. As such, an environment for always wearing the device to present converted sound is already in place. Moreover, we can utilize machine learning (ML) along with ultrasonic echolocation because the captured ultrasound is processed inside the device. One previous study on echolocation focused on interpreting the sensory perception of ''seeing by sound'' in bats and proposed an effective and practical manner of operation for adoption in human echolocation [14], [17]; however, they did not consider ML.

In the current work, we propose a cooperative echolocation system that combines innate human recognition with recognition by ML to reduce the burden on the user during echolocation. We assume our target users can use echolocation and are already wearing an ultrasonic speaker and microphone set to identify objects they want to find. The wearable speaker transmits ultrasound and the microphone captures the reflected sound, and our system converts the ultrasound into audible sound that the users can recognize. It also calculates the frequency spectrum and recognizes the objects by ML, and when the ML recognition result is the same as the user-specified object, the device vibrates. Users can search for and recognize objects by using the sounds and vibrations as clues in real time. In contrast to previous methods that are purely machine-based, our system is unique in that not only the machine but also the user senses the objects. This is important because it is more meaningful for visually impaired people to recognize the environment with their own ears, thereby forming a vital element in their sense of independence [12], [13]. In short, our motivation is to support users by means of ML without causing them to lose their sense of independence. Moreover, since we assume an environment in which users are already wearing an ultrasonic speaker and microphone set for ultrasonic human echolocation, there is no need for additional ML devices.

To the best of our knowledge, this is the first system that combines ML with echolocation. As a first step toward cooperative echolocation, this paper investigates the effectiveness of ML in echolocation and clarifies design principles for its adoption. We implemented a prototype to evaluate the recognition accuracy for six types of objects and found that the recognition accuracy of ML was 92.5%. We also evaluated the changes in object detection behavior with and without ML and found that the detection time was decreased from 71.5 s to 45.9 s by using ML. Additionally, the results of the NASA Task Load Index (NASA-TLX), an assessment tool for workload, showed that the mental workload on the user was significantly decreased thanks to ML. On the basis of our findings, we conclude the paper by discussing how ML should be incorporated into echolocation. Our main contributions are as follows.

- We propose a cooperative echolocation system that utilizes not only human but also ML recognition in ultrasonic echolocation.
- Evaluation results demonstrate that incorporating ML into echolocation is effective, especially in terms of significantly decreasing mental workload.
- On the basis of our findings, we present how ML should be incorporated into ultrasonic echolocation.

II. RELATED WORK

A. ECHOLOCATION

Some visually impaired people utilize echolocation, which is a ''seeing by sound'' technique that enables them to perceive their surroundings from the click sounds they make. Echolocation using audible sounds has been examined in a number of studies [11], [18], [19].

In other studies, inspired by echolocation in bats, researchers have investigated the application of ultrasonic echolocation to humans [14], [17], [20], [21]. Sohl-Dickstein et al. developed a system that transmits ultrasonic signals and converts the echoes into audible sound, thus enabling recognition of the distance from objects and the presence/absence of obstacles [20]. In a study by Sumiya et al., participants were able to discriminate object shapes and textures by using a system that converts ultrasonic echoes into audible sound from objects learned by training [17]. They found that the performance of object discrimination using ultrasound was higher than that using audible sound.

Although ultrasonic echolocation has been the focus of past research, most studies focused on interpreting the sensory perception of ''seeing by sound'' in bats and trying to come up with ways of harnessing this operation in human echolocation; in other words, a system that can operate in real time has rarely been considered. Sohl-Dickstein et al. [20] reported an ultrasonic echolocation system for use in real time; however, they focused on recognizing the distance to objects and the presence of objects, and did not consider ML. In the present work, we focus on recognizing what the object is and how to combine both human and ML recognition.

B. ASSISTIVE SYSTEMS FOR VISUALLY IMPAIRED PEOPLE

Many assistive systems for the visually impaired have been studied thanks to the increasing availability of

mobile/wearable devices. Bing et al. developed a navigation system using Google Tango [22]. Brock et al. proposed a method to detect an object by using Microsoft Kinect and present it to the user by sound [8]. Their system changes the pitch and volume of the presented sound depending on the location of the object and its distance from the user. Kayukawa et al. developed BBeep [23], a system that emits a warning sound to pedestrians who are likely to collide with the user by using an RGBD camera attached to a suitcase. They also developed another suitcase-type system that helps visually impaired people avoid pedestrians in their path [24]. Zhao et al. proposed an AR system for people with low vision [25] that displays the edges of stairs. Wang et al. proposed a system that uses a camera and vibration to guide the user [26]. Another system called Virtual Paving utilizes vibration and voice to navigate users in areas where tactile paving is not available [27]. Several systems for supporting supermarket shopping have been proposed [28], [29], [30], [31]. Acoussist is a system that assists visually impaired users in crossing the road by detecting ultrasound emitted from vehicles [32]. Mocanu et al. proposed a wearable assistive device designed to facilitate the autonomous navigation of visually impaired people in dynamic urban scenes [33]. Their system exploits ultrasonic sensors and the video camera embedded in a regular smartphone.

However, while many assistive technologies have been studied, these systems are based primarily on machine recognition. Our study differs in that users recognize an object through a combination of the sound obtained by their own auditory perception and the results of ML.

C. OBJECT RECOGNITION

Image-based methods are often used for recognizing the environment. One example is the RGB camera, which is utilized for recognizing materials/objects [3], [4], [5], [6], [34]. Although image-based methods provide useful information similar to what we can pick up with our eyes, they are susceptible to light conditions (e.g., darkness and brightness), which can adversely affect the recognition.

The depth camera is also used to recognize the environment [7], [8], [9], [10]. However, as most depth cameras utilize infrared, they are adversely affected by sunlight.

Researchers have thus been developing other approaches for material/object recognition. SpecTrans is an image sensor-based material recognizer that utilizes reflected light patterns produced by multi-spectral lighting sources [35]. SpeCam is a lightweight surface color and material sensing system for mobile devices that uses a front-facing camera and display as a multi-spectral light source [36]. RadarCat is a radar-based system for material and object classification [37]. Harrison et al. developed a method for identifying materials in proximity to the device by means of a multispectral optical sensor [38]. Liu et al. proposed a method for recognizing surface material by utilizing the vibrations generated from contacted objects [39]. Although these methods can be quite effective, they need to be in direct contact with the materials for recognition. From the perspectives of hygiene and social acceptance, it is desirable to avoid touching objects more than necessary.

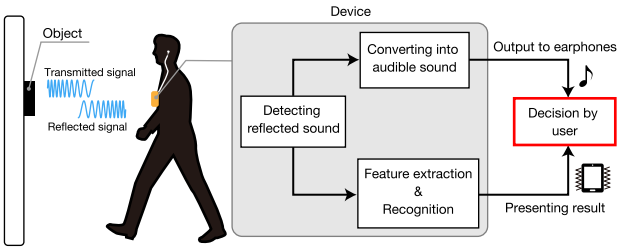


FIGURE 1. System configuration.

DeepThermalImaging is a system that recognizes the material of an object by using a thermal camera [40]. It is advantageous in that it can recognize material without contact and is robust against changes in ambient light. However, the thermal reflection of wet and glossy materials is difficult to recognize, and since the system uses a thermal camera, this would need to be added to most smartphones.

There are many studies on object detection/recognition using sound. Komatsu et al. discriminated the characteristics of a target object based on the frequency characteristics of the reflected ultrasonic waves [41]. DeepRange is a deep learning-based acoustic ranging system [42]. BatMapper is a system that uses acoustic ranging to detect walls and create floor plans [43]. BumpAlert is a system to detect obstacles using embedded sensors in smartphones [44]. It utilizes sound to detect objects in places with little reverberation (e.g., outdoors) and images to detect objects in places with large reverberation (e.g., corridors). ObstacleWatch is a system that detects objects using the reflected sound signal [45]. Remaggi et al. proposed a method to estimate the material of an object by using the attenuation rate of the reflected sound in the frequency domain [46]. Mao et al. developed an acoustic imaging method using an off-the-shelf smartphone [47]. The user moves the phone along a predefined trajectory to mimic a virtual sensor array, and the system reproduces the target object using the reflected sound.

Considering the characteristics of wearable computing environments, additional devices for object recognition (e.g., specialized cameras or depth sensors) are not desirable because of the limited power consumption and the increased cost. In our study, we assume an ultrasonic human echolocation environment in which the user is already wearing an ultrasonic speaker and microphone; thus, there is no need for additional devices. We also utilize recognition by both machines and humans.

III. PROPOSED METHOD

Fig. 1 shows the system configuration of the proposed method. It consists of two parts: processing for human recognition and processing for machine recognition. First, the speaker transmits ultrasound and stereo microphones capture the reflected sound. For human recognition, the system

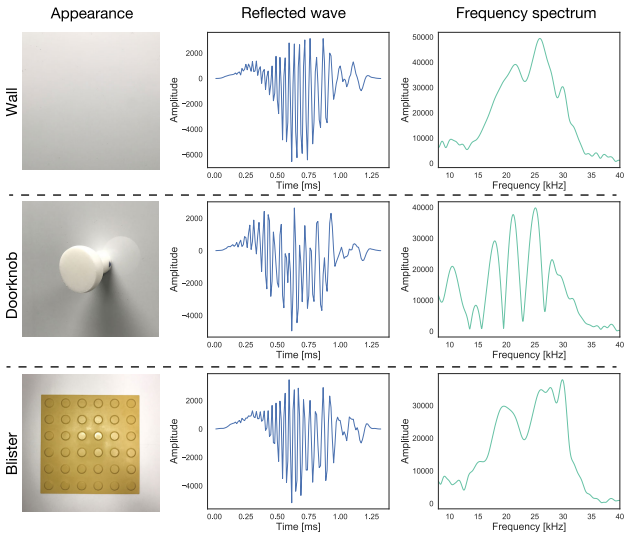


FIGURE 2. Object and corresponding reflected wave and frequency spectrum.

converts the ultrasound into audible sound by lowering the frequency and presents the converted sound to the user via earphones. For machine recognition, the system applies ML to the obtained signal and presents the recognition result to the user by vibration or sound. The device then notifies the user of the result by vibration when the recognition result is the same as the user-specified object. Fig. 2 shows examples of reflected waves and the corresponding frequency spectra. As we can see, there are several points where the magnitude of the amplitude of the reflected wave is different depending on the object, and it also appears in the frequency spectrum. Humans and machines recognize these differences by the timbre and by the differences in the frequency spectra, respectively. Finally, the user recognizes the object by combining the sound obtained from the ears and the vibration resulting from the ML recognition. We describe each process in detail in the following sections.

A. TRANSMITTING AND DETECTING SIGNAL

The speaker transmits the ultrasonic signal and the microphone captures the reflected signal from the target object. The sampling rate for transmitting/recording is 96 kHz. The signal is a downward frequency-modulated (FM) sweep signal that shifts from 40 kHz to 20 kHz in 1 ms. The FM sweep signal is used for target classification by bats [48], and a previous study using a similar signal reported that the downward FM signal was the most effective to discriminate target objects [17]. Although that study used the frequency range of 8 k–40 kHz, we selected 20 k–40 kHz here because the constant generation of audible sound (below 20 kHz) while using the system is unpleasant for both the user and surrounding people. Assuming that the speed of the sound in the air is 340 m/s, the transmitted signal does not overlap with the reflected signal when the distance between the object and the microphone/speaker is 17 cm or more. Therefore, we set

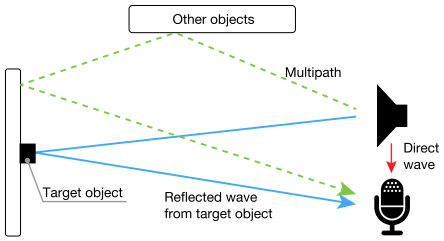


FIGURE 3. Multipath.

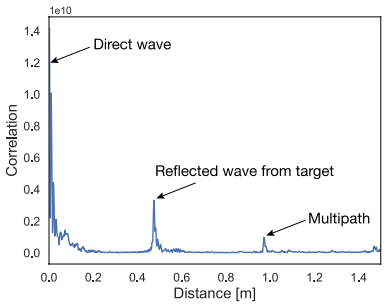


FIGURE 4. Cross-correlation.

the minimum distance that the object can be detected to 17 cm in this study.

As shown in Fig. 3, the received signal from the microphone includes the signal directly from the speaker, the reflected signal from the target object, and the reflected signal from other objects. To detect the reflected signal from the target object only, we calculate the cross-correlation value between the transmitted signal and the received signal. Fig. 4 shows the envelope of cross-correlation values obtained when the signal is transmitted 50 cm away from the object. As we can see, the cross-correlation shows a large value when the received signal has a similar waveform to the transmitted signal. The first large peak shows the direct wave from the speaker to the microphone. In this study, we define the largest cross-correlation after the direct wave as the reflected wave from the object since, as preliminary investigation confirmed that the cross-correlation of the object in front of the speaker was the largest. We extract 128 samples as the reflected wave from the above point. Although the number of transmitted signal samples is 96 (96,000 Hz $\times$ 1 ms), we extract 128 samples to provide a margin.

To enable real-time processing, the above process is performed after a certain amount of data is stored in a buffer. We set the buffer size to 4,096 and acquire non-overlapping buffers with a sliding size of 4,096. If direct and reflected waves do not exist in a single buffer, false detection of the reflected sound occurs; thus, we configure the system to not detect the reflected sound when the distance between the device and the object is more than a certain range. We calculate the distance to the object from the time of signal transmission and reception as

$$d = \frac{c(t_1 - t_0)}{2}, \tag{1}$$

where d is the distance between the device and object, c is the speed of sound, t_0 is the time of signal transmission, and t_1 is the time of signal reception. The detection range is set to 20–300 cm and when the detected distance is outside this range, the system does not perform any further processing. Note that it is possible for a false detection to be included in the above range. Assuming the human walking speed of 1.3 m/s [49] and the echolocation signal transmission interval of 600 ms [18], [19], the moving distance between the signal transmission is considered to be 78 cm (1.3 m/s × 0.6 s). Therefore, when the difference in distance from the previous detection is 80 cm or more, the system does not perform further processing even when the detected distance is in the range of 20–300 cm.

B. CONVERTING ULTRASOUND INTO AUDIBLE SOUND

1) CONVERSION METHOD

The ultrasound detected in the previous section is next converted into audible sound. Among the various conversion methods we could use (e.g., time expansion, heterodyning, and phase vocoder [50], [51]), we selected time expansion because it has been used in previous studies [17], [20] and can preserve all the characteristics of the original signal. The detected reflected signal is played to the user at $1/m$ of normal speed, where m is an adjustable magnification factor. This magnifies the signal linearly on the time axis by a factor of m and lowers the frequencies into the audible range. We set different sampling rates for recording and playing (e.g., when recording at 96 kHz and playing at 12 kHz, m becomes eight).

2) MINIATURE DUMMY HEAD

We utilize a dummy of a human head for recording in order to take into consideration the head-related transfer function (HRTF) [52]. The miniature dummy head (MDH) is a standard small-scale dummy head and is used to record ultrasound [14]. Humans can hear echoes in the ultrasonic range as the audible range in a stereophonic acoustic space with a realistic feeling. A previous study reported that MDH has similar characteristics to the standard size dummy head when the received ultrasound is converted into audible sounds [21]. In this study, we utilize 1/7 scale MDH, which was used in the previous studies [14], [17].

3) MAGNIFICATION OF CONVERSION

In the previous study [14], 1/7 magnification for MDH of 1/7 scale was effective for localization and out-of-head perception of sound; however, it is not clear whether this magnification is also suitable for the frequency range used in this study or for distinguishing the timbre of reflected waves from objects. For the frequency range here (20 k–40 kHz), the frequency is lowered to 2.9 k–5.7 kHz by 1/7 magnification. Since humans can better discriminate lower timbre than higher timbre [53], converting to lower frequencies may be more useful for distinguishing objects. Therefore, we conducted a preliminary experiment to determine the suitable magnification factor.

TABLE 1. Signal information after conversion.

Magnification	Frequency band [kHz]	Signal length [ms]
1/7	2.9–5.7	7
1/14	1.4–2.9	14
1/20	1–2	20

In this experiment, the reflected waves from no object and object were converted into an audible sound, and participants attempted to discriminate the two. As no object and object, we selected the wall and the doorknob in Fig. 2. The tested magnifications were 1/7, 1/14, and 1/20. The signal information after conversion is listed in Table 1. Magnification of 1/7 is the same as that used in the previous studies [14], [17]. Also, since one previous study [20] lowered the frequencies from 25 k–50 kHz to 1 k–2 kHz, we selected 1/20 so that the converted frequencies in this study are the same. Magnification of 1/14 was the midway point between the two others. To evaluate the performance of signal discrimination, we used three-interval two-alternative forced-choice tasks (3I-2AFC) in which participants listened to three consecutive sound stimuli consisting of the converted reflected sound of no object and object and identified whether the second sound stimulus was the same as the first or third. The five participants were 23–31-year-old men and women. The interval between sound stimuli was 300 ms and there were four possible combinations per magnification. As a result, we collected 60 data (five participants × three magnifications × four signal combinations) and calculated the percentage of correct answers as evaluation metrics.

Fig. 5 shows the correct answer rate for each magnification, where we can see that the magnifications of 1/7, 1/14, and 1/20 were 55%, 90%, and 95%, respectively. At the magnification of 1/7, which was used in the previous study [17], the correct answer rate was the lowest (in fact, almost the same level as chance), while the magnification of 1/20 was the best. There are several potential reasons for this result (e.g., frequency range and signal length after conversion); however, as we are focused here on a suitable magnification for time expansion, we simply selected 1/20. We leave a detailed investigation of the relationship between magnification and discrimination performance to future work.

C. MACHINE LEARNING

1) PREPROCESSING

We apply fast Fourier transform (FFT) to the extracted signal in Section III-A and obtain the frequency spectrum, as shown in Fig. 2. To resolve waveform discontinuities, we multiply Hann window to the extracted signal. The extracted wave is made to be 8,192 samples with zero padding. The frequency resolution becomes approximately 11.7 Hz (96,000 / 8,192) by zero padding, which should enable us to acquire more detailed characteristics of the signal.

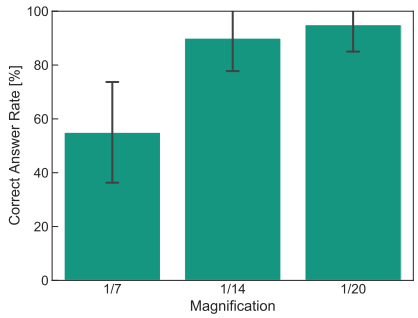


FIGURE 5. Correct answer rate of three conversion rates. Error bars show standard deviations.

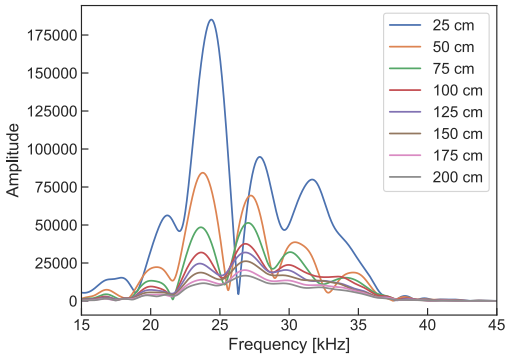


FIGURE 6. Change in frequency spectrum depending on distance.

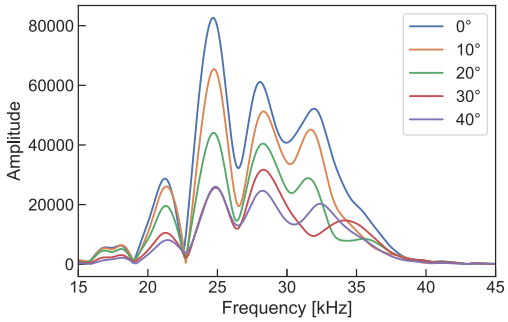


FIGURE 7. Frequency spectrum of doorknob at each angle.

2) FEATURE EXTRACTION AND RECOGNITION

As shown in Fig. 2, the characteristics of the reflected waves appear in frequency spectra; thus, we utilized several spectral features. Linear-frequency cepstral coefficients (LFCCs) [54], which are a linear version of mel-frequency cepstral coefficients (MFCCs), were selected as the feature values. In contrast to the MFCCs typically used for audio and speech recognition, LFCCs utilize a linear filterbank to reduce dimensions, and we opted to use them here because we want to equally extract features from the frequency spectrum. The number of used filterbanks was 20. After we removed the first LFCC, which is the DC component, we acquired 19 features.

We also calculated nine spectral features (mean, variance, spectral centroid, roll-off, flatness, skewness, kurtosis, bandwidth, and entropy) for both the left and right channels of the microphones. As a result, we obtained 56 feature values in total (28 features $\times$ two channels). Note that we utilized both channels even when both features were assumed to be similar, e.g., the device and the object were facing each other. This is because we confirmed in the preliminary experiment that the frequency spectra obtained in both channels were different due to the characteristics of MDH and subtle installation errors of the microphones.

The classifier needs to be computationally lightweight so that it can work in real time on wearable devices and have a high enough performance. Although we do not limit classifier algorithm, we selected the support vector machine (SVM) as the classifier in this study.

3) SWITCHING MACHINE LEARNING MODEL DEPENDING ON DISTANCE

Since sound pressure attenuates as the transmission distance increases, the frequency spectrum obtained for the same object may be different depending on the distance. Fig. 6 shows an example of the change in frequency spectrum depending on the distance. As we can see, the approximate peak/notch positions of the frequency spectrum are located at nearby frequencies at all distances; however, there is a slight shift and the amplitude ratio of the peak to the notch changes depending on the distance. This suggests that the recognition rate decreases if the same learning model is used to recognize all distances. We therefore switched the ML model depending on the detected distance so as to increase the recognition accuracy. Specifically, we acquired data at intervals of 25 cm from 25 cm to 200 cm and trained eight models. The system calculates the distance to the object based on (1) and switches to the learning model that is closest to the detected distance.

4) DETECTING ANGLE BETWEEN OBJECT AND DEVICE

When the angle between the object and the device is different, the obtained frequency spectrum may change due to the HRTF of the MDH and the characteristics of the speaker/microphone. Figures 7 and 8 shows the frequency spectrum at each angle for a doorknob and a blister tactile, respectively. In Fig. 7, we can see that the frequency spectra are similar for different angles except for amplitude, while in Fig. 8, the obtained frequency spectrum differs depending on the angle. This means that although the frequency spectrum obtained at each angle is different, the degree of change varies depending on the object.

Our system therefore performs converting ultrasound and ML only when the angle between the object and device is within a certain range depending on the user-specified object. Concretely, the system estimates the angle with the object using the arrival time difference between the stereo microphones (as shown in Fig. 9), which is calculated as

$$t_l - t_r = \frac{l \sin\theta}{c}, \tag{2}$$

where θ is the angle between the device and object, l is the distance between the stereo microphones, t_l and t_r are the

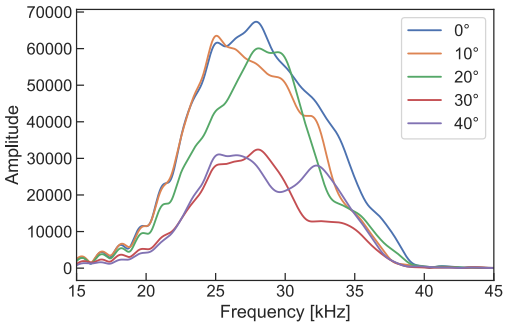


FIGURE 8. Frequency spectrum of blister tactile at each angle.

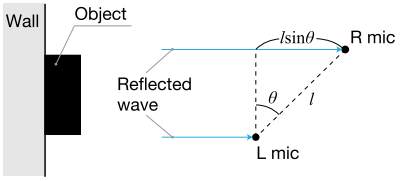


FIGURE 9. Overview of angle detection.

arrival times of the left and right microphones, respectively, and c is the speed of sound. On the basis of (2), the angle is then calculated as

$$\theta = \arcsin \frac{c(t_l - t_r)}{l}. \tag{3}$$

Note that since the sampling rate is 96 kHz and l is 2.2 cm, the resolution of the detected angle is approximately 9.26°.

IV. IMPLEMENTATION

We implemented the prototype device shown in Fig. 10. It consists of a tablet (Huawei MediaPad M5), audio interface (Zoom U-24), speaker (Peerless S0155), microcontroller (Sony Spresense), two microphones (Countryman B6), and headphones (Bose QC35). Note that we did not use the noise-canceling function that comes with these headphones. To mitigate the vibration from the speaker to the microphone through the case and MDH, elastomer resin was sandwiched between the case and MDH. The frequency response of the microphone was 30–20 kHz; however, we confirmed that it can capture the ultrasonic signal, and the previous study also used the same one [17]. The MDH was the 1/7 scale of the standard dummy head, which is the same scale used in the previous studies [14], [17]. Stereo microphones were placed at the eardrums of the MDH. The height, width, and depth of the MDH were 6.5 cm, 5.2 cm, and 2.5 cm, respectively. The MDH was constructed of a silicone rubber material (Shore A hardness: 35) to mimic the hardness of the human body. The microphones were connected to the tablet via the audio interface. The microcontroller was wired to the speaker and played the sound source that was stored in the microSD card. Since Spresense supports playing at 96 kHz or higher, it can transmit ultrasound, and the speaker can transmit a signal with sufficient sound pressure since Spresense has a built-in amplifier. Note that while this prototype device is in a form where the user grasps the speaker part with one hand, it is also possible to implement it as a hands-free device, e.g., hanging it from the neck like a pendant.

FIGURE 10. (a) Implemented prototype device and (b) side view of MDH and speaker.

We also implemented the proposed method in Section III using an Android application. Although we fixed the parameters of the proposed method in the experiment, users can change between with/without the presence of converted sound, with/without the presence of ML, and magnification of ultrasound conversion in this application. We utilized SVM as the classifier, as described in Section III-C2.

V. EVALUATION

We conducted two evaluations: object recognition by ML (Section V-A) and object detection using human recognition and machine recognition (Section V-B). Our primary objective is to determine whether ML can reduce the burden on the user during echolocation, i.e., the latter evaluation is the main experiment. However, as far as we know, there are no studies that recognize objects using only the acoustic characteristics of reflected sound from the objects assuming echolocation. Therefore, as a preliminary step to achieving the research purpose, we also conducted ML evaluation with several objects.

A. OBJECT RECOGNITION BY MACHINE LEARNING

1) INFLUENCE OF DISTANCE

We selected the target objects based on the idea that it is useful to know the objects before touching them. Braille and tactile paving are widely utilized as clues for visually impaired people; however, they cannot know where these clues are until they have directly touched them. Also, doorknobs and door handles are two objects that visually impaired people cannot know the location of until they have directly touched them, and although they can search for these objects by groping, it is desirable to avoid touching unknown objects from the viewpoints of hygiene and social acceptance. Therefore, it will be useful if they can identify whether these objects exist or not—and if they exist, what they are—without contact. On the basis of these considerations, we selected six target objects: wall, braille, door handle, doorknob, directional tactile, and blister tactile (Fig. 11). We used braille labels as braille. The door handle and doorknob were made of ABS resin constructed by a 3D printer (UP Plus 2). The directional/blister tactile was made of synthetic

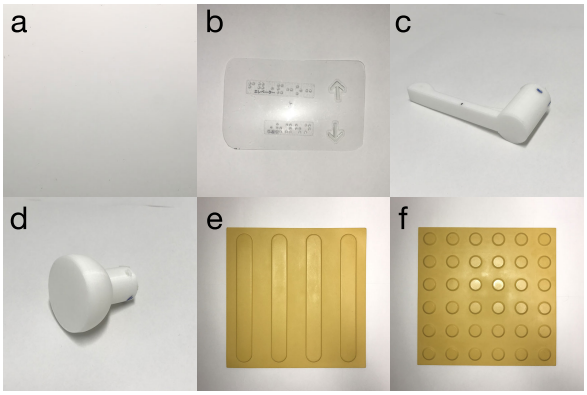


FIGURE 11. Tested objects: (a) wall, (b) braille, (c) door handle, (d) doorknob, (e) directional tactile, and (f) blister tactile.

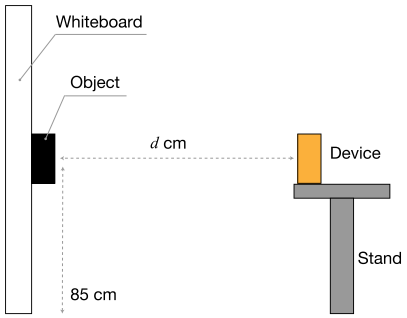


FIGURE 12. Experimental setup (side view).

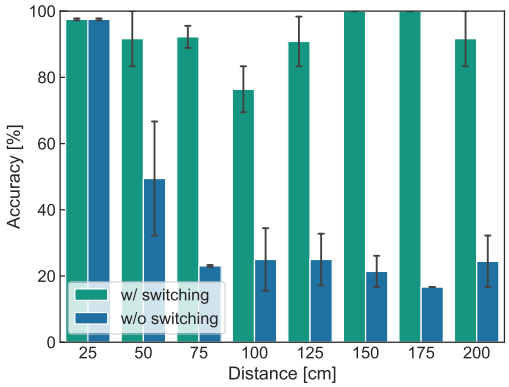


FIGURE 13. Recognition accuracy at each distance. Error bars show standard deviations between sets.

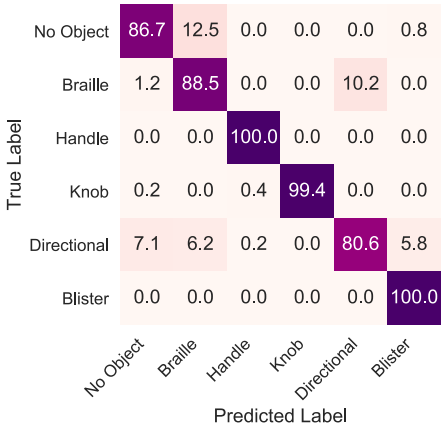


FIGURE 14. Confusion matrix of object recognition [%].

rubber. The heights of the braille, door handle, doorknob, and direction/blister tactile objects were 0.3 mm, 4 cm, 5 cm, and 3 mm, respectively.

A diagram of the evaluation setup is shown in Fig. 12. Each target object was attached to a whiteboard, and the prototype device was used to transmit the ultrasonic signal and record the reflected sound. The distances between the device and whiteboard were 25, 50, 75, 100, 125, 150, 175, and 200 cm. Thirty measurements were performed at each distance (one set). After detaching/attaching the objects and resetting the device, we measured another set. In total, we obtained 2,880 data (six objects $\times$ eight distances $\times$ 30 measurements $\times$ two sets). We calculated the feature values described in Section III-C2 and performed the evaluation by leave-one-set-out cross-validation.

Fig. 13 shows the recognition accuracy at each distance. In the case of without switching model, the system utilized a 25-cm model for all distances. As we can see in the figure, the recognition accuracies were over 90% except for 100 cm and the overall recognition accuracy was 92.5% by switching model while the accuracy significantly decreased in the case of without switching model. Even when switching model, the recognition accuracy at 100 cm (75.6%) was lower compared with other distances. This lower accuracy is presumably due to the multipath characteristic. Specifically, reflected waves from non-target objects located 100 cm away from the device and target objects were recorded at the same time, which means the obtained frequency spectrum was different from the original frequency spectrum.

Fig. 14 shows the confusion matrix of all distances. Each row is normalized to 100%. As we can see, all recognition accuracies were over 86% except for directional tactile. We also found that no object and braille were confused and that directional tactile was confused as no object, braille, or blister tactile. This is presumably due to the height of the objects: for example, the handle and doorknob used in this study were 4 and 5 cm, respectively, which is higher than that of the other objects. The previous study [17] also reported that the depth of the notch in the frequency spectrum increases with the height of the object; thus, the system could recognize these objects more correctly than other objects using these characteristic frequency spectra.

From the above results, we confirmed that the system can recognize objects based on the reflected sound, and that switching the ML model depending on the distance was effective.

2) INFLUENCE OF ANGLE

As discussed in Section III-C4, the obtained frequency spectrum changes depending on the angle between the device and the object. Therefore, in this section, we investigated the

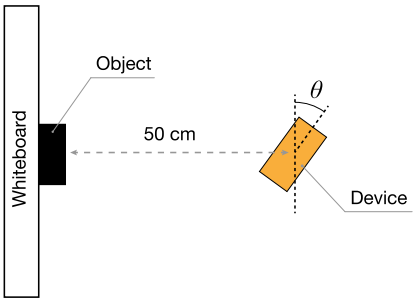


FIGURE 15. Angle experiment setup (top view).

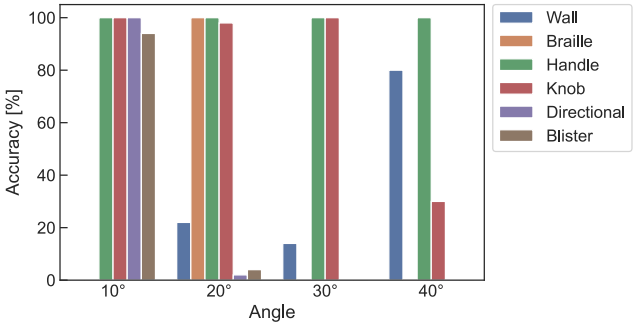


FIGURE 16. Recognition accuracy at each angle.

dependence of recognition accuracy on the angle. Fig. 15 shows the experimental setup. The angle between the device and the object θ was varied from 0° to 40° in 10° increments and the data was acquired 50 times at each angle. Since the system could not detect the angles of 50° or more, we used angles up to 40° for this evaluation. The data of 0° was utilized as training data and we then tested the data of each angle. The distance between the device and each object was set to 50 cm.

Fig. 16 shows the recognition accuracy at each angle. As we can see, the accuracy of all objects except handle and doorknob decreased significantly when angles occurred, and the handle and doorknob could be recognized with high accuracy up to the angle of 30°. This is presumably due to the height of the objects. For example, the height of braille/tactile paving and handle/doorknob was several millimeters and 4–5 cm, respectively, and since we know that the height affects the obtained frequency spectrum, the handle and doorknob could be more accurately recognized compared to other objects.

These results confirm that when an angle occurs between objects, some objects can be recognized while others cannot. Our system assumes that the user selects the object to find and then the machine sends a notification when it finds it. This suggests that we need to change the acceptable object-device angle depending on the user-specified object: i.e., for objects like the handle and doorknob, which have high recognition accuracy even when an angle occurs, a device-object angle up to 30° is acceptable, while for objects like braille, which have low recognition accuracy when an angle occurs, the system should not perform recognition when an angle is detected.

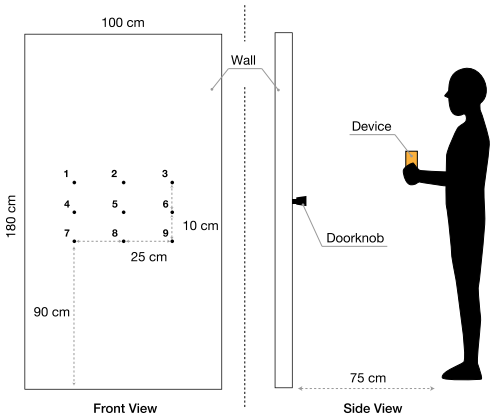


FIGURE 17. Object detection experiment setup.

B. OBJECT DETECTION USING HUMAN RECOGNITION AND MACHINE RECOGNITION

We next investigated the effect of combining ultrasonic echolocation with ML. In this evaluation, we focus on the effects of adding ML to conventional echolocation; thus, we compare the performance of the ultrasonic echolocation with and without ML. Fig. 17 shows a diagram of the experimental setup. We assumed a scene in which the user enters/exits a room via a door, and selected the doorknob as the object to be detected. Note that our objective in this work is not to investigate whether it is possible to search for an object among multiple objects but rather to determine whether the introduction of ML to echolocation reduces the burden on the user. Therefore, we tested only one type of object here. We placed the doorknob at any one of nine locations within a height range of 90–110 cm. This range was determined by considering where the doorknob was expected to exist. We used the same used doorknob as the previous evaluation and trained the ML model using the data obtained in Section V-A1. Also, considering the results of Section V-A2, we set the acceptable angle as −30° to 30°. The following experiments were approved by the Human Ethics Committee of Graduate School of Engineering, Kobe University (Permission Number: 04-02).

1) SIGNAL TRANSMISSION INTERVAL

We first investigated the suitable interval of signal transmission. Although the signal interval of expert human echolocation is approximately 600 ms [18], [19], it is not clear whether this interval is also suitable for object detection using ultrasonic echolocation. In human echolocation, the clicking sound is generated by clicks of the tongue, which limits the interval of signal transmission; however, the system is capable of transmitting signals at shorter intervals than human echolocation. Therefore, we investigated the suitable interval of signal transmission.

Ten sighted echolocation novices aged 21–23 year-old (all males) participated in this experiment. As shown in Fig. 17, participants stood 75 cm away from the wall and moved the device as if scanning the wall to detect the object.

First, we asked them to practice this detection for 5–10 minutes with their eyes open. The practice was finished when they felt they could find the object with their eyes closed. The doorknob was then set at a random location and participants explored it with their eyes closed. They reported to the experimenter when they thought they had found it. The experimenter checked the distance between the positions of the doorknob and the projected position of the device on the wall, which was visualized by a laser pointer placed at the bottom of the device. When the distance was within approximately 5 cm, the experiment was finished; otherwise, they continued the experiment until they could find it. Participants tried this experiment for three intervals of signal transmission:

- 200 ms: the shortest interval of signal transmission that was possible in the implemented device
- 400 ms: midway point between 200 ms and 600 ms
- 600 ms: interval of human echolocation [18], [19]

The order of tested signal intervals was set randomly. Participants performed the above three trials under two conditions: using only the audible sound converted from the ultrasound (*audio*) and using converted sound and the vibration that indicates the ML result (*audio + vibration*). Since the proposed method is based on ultrasonic echolocation, all participants first tested the *audio* condition to get familiar with ultrasonic echolocation and then tested the *audio + vibration* condition. At the end of each test, participants were asked which intervals they preferred.

Table 2 lists the preferred signal intervals. For *audio*, the most preferred signal interval was 200 ms (selected by seven out of ten participants). They commented that the reason for this was "*shorter signal intervals increased the scanning speed of the object*". On the other hand, participants A and D commented that "*too short a signal interval was annoying*". Participants C, E, and G, who selected 400 ms or 600 ms, commented that "*when the signal interval was too short, it was difficult to distinguish the sound*".

For *audio + vibration*, the most preferred interval was also 200 ms (selected by nine out of ten participants). More participants here preferred 200 ms than under the *audio* condition, commenting that "*vibration and sound were presented in short intervals, making it easier to find the doorknob when moving the device little by little.*".

Considering these results, we set the interval to 200 ms in the following experiment. Note that, in the actual use of the system, users can set their preferred signal interval. Also, as described above, the short signal interval increased the obtained information while it was uncomfortable for the user. We discuss this in detail in Section VI-B.

2) OBJECT DETECTION

Next, we investigated the effect of combining ML with echolocation. The experimental environment was the same as in the previous section (see Fig. 17), and the participants were also the same (i.e., all had already trained with the proposed method). We investigated the same two methods (*audio* and *audio + vibration*) as in the previous section. Note that *audio* is the baseline, as our method is based on echolocation. The tested order of *audio* and *audio + vibration* was set to provide a counterbalance across participants. Participants tested three times for each method. To reduce the influence of the doorknob position, the three positions were set to include all three vertical lines (left, middle, and right) and horizontal lines (top, middle, and bottom) of Fig. 17 (e.g., for participant A, the positions of the doorknob were 1, 5, and 9 in the case of *audio* and 2, 4, and 9 in the case of *audio + vibration*). The start position of each experiment was set to the middle of the area (position 5). A total of 60 measurement data items (three trials × two methods × ten participants) were obtained. To investigate how ML affects user behavior, we placed a laser pointer at the bottom of the device (5 cm below the center of the speaker), recorded the trajectory of the laser pointer projected on the wall by video, and obtained the trajectory of the device by image processing. The participants reported to the experimenter when they thought they had found the doorknob. The experimenter checked the position of the device in the same way as in the previous section. When the distance was within approximately 5 cm, the experiment was finished; otherwise, it was counted as an incorrect answer and the experiment was continued. When the time exceeded 300 s, the experimenter finished the experiment. After each experiment, the participants were asked to answer the NASA-TLX as well as to answer open-ended questions regarding how they felt about the system at the end of both methods. Evaluation metrics here were the number of incorrect answers, the required time for detecting the object, NASA-TLX, and the exploratory behavior of the user (trajectory of the laser pointer). The obtained results are shown below.

TABLE 2. Preferred interval of signal transmission [ms]. 200/400 means participants felt there was no difference between the two.

Participant	Audio	Audio + Vib.
A	200	200
B	200	200
C	600	200
D	200	200
E	400	200/400
F	200	200/400
G	400	400
H	200	200
I	200/400	200/400
J	200	200

a: NUMBER OF INCORRECT ANSWERS

Table 3 lists the number of incorrect answers and timeouts for each participant. As we can see, the total number was the same for both methods. The number of incorrect answers was low regardless of ML except for participant G, who stated that "*I felt the sound changed even in places where the object did not exist*", which resulted in a higher number of incorrect answers compared to the other participants.

TABLE 3. Number of incorrect answers and (timeouts) for each participant.

Participant	Audio	Audio + Vib.
A	0	0
B	0	0
C	1	0
D	0	0
E	0	0
F	0	0
G	5 (1)	5
H	0	1
I	0	0
J	0	0 (1)

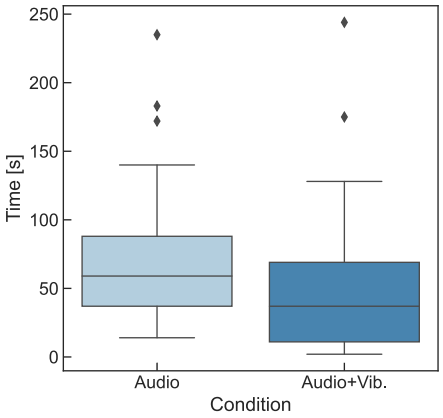


FIGURE 18. Required time for detecting the object.

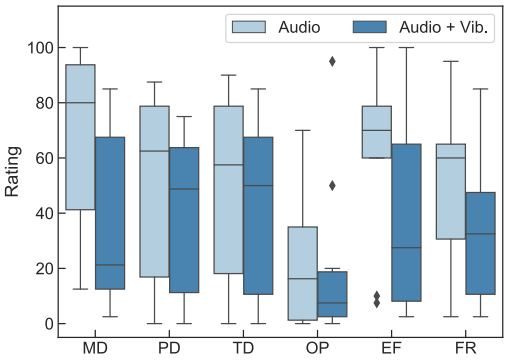


FIGURE 19. Subscale ratings of NASA-TLX: mental demand (MD), physical demand (PD), temporal demand (TD), own performance (OP), effort (EF), and frustration (FR).

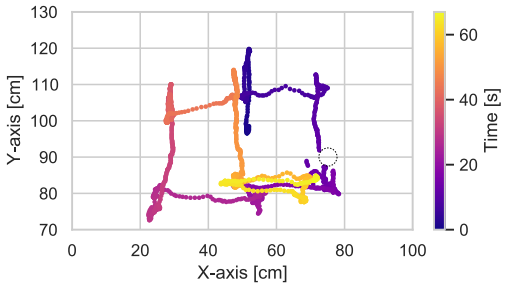


FIGURE 20. Trajectory under audio condition (participant C). Dotted circle shows location of the doorknob.

b: REQUIRED TIME FOR DETECTION

Fig. 18 shows the required time for detecting the object. By combining ML with echolocation, the required time decreased from 71.5 s to 45.9 s on average. We conducted Wilcoxon signed rank tests and while we confirmed that there was no significant difference between the methods, it was close to significance ($p = .054$).

c: NASA-TLX

As overall workload, we selected Raw TLX (RTLX), which is the average of each subscale [55]. The RTLX of *audio* and *audio + vibration* was 50.6 and 35.6, respectively. We conducted a paired t-test and found a significant difference between the methods ($t(9) = 2.93, p < .05$). Fig. 19 shows the rating of each subscale of NASA-TLX, where lower values correspond to a lower workload. As we can see, the overall rating decreased when ML was combined. In particular, the means of mental demand (67), effort (61.5), and frustration (50.5) decreased by 43.3% (38), 36.6% (39), and 30.7% (35), respectively. On the other hand, the means of physical demand (49.5), temporal demand (50.5), and own performance (24.8) decreased by 20.2% (39.5), 16.8% (42), and 19.2% (20), respectively, all of which are smaller than the other three subscales. It is worth noting that the rating of own performance was low regardless of the method, i.e., even when the participants used only the sound, the sense of accomplishment was high.

d: EXPLORATORY BEHAVIOR

Figures 20 and 21 show an example of exploratory behavior by using the trajectory of the third trial of participant C. Note that, as described above, since the laser pointer was placed 5 cm below the speaker, the actual search trajectory of the speaker was 5 cm above the trajectory shown in the figure. In the case of *audio* (Fig. 20), the participant scanned the whole area once and returned to the point where he felt suspicious. Then, he moved the device left and right to be sure of the difference in sound. In contrast, in the case of *audio + vibration* (Fig. 21), the participant stopped at the point where he felt suspicious before scanning the whole area and when he felt the vibration from the machine, he answered the location of the object with confidence.

VI. DISCUSSION & LIMITATIONS

A. EFFECTIVENESS OF MACHINE LEARNING

The results of our evaluation confirmed that ML can recognize an object using the reflected sound and that the required time for object detection was decreased, although the difference was not statistically significant. Also, the result of NASA-TLX showed that the mental workload was significantly different between *audio* and *audio + vibration*, as the participants could detect the object with greater

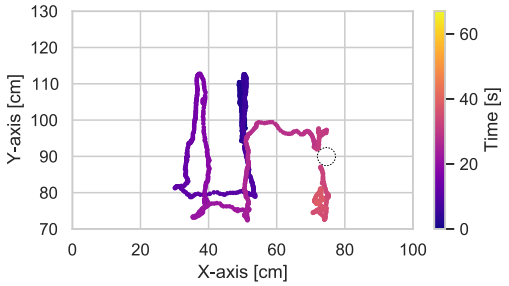


FIGURE 21. Trajectory under audio + vib. condition (participant C). Dotted circle shows location of the doorknob.

confidence by utilizing the ML recognition. As shown in Fig. 20, with *audio*, the participant searched the same area several times to be sure of the difference of sound, whereas with *audio* + *vibration*, he could detect the object much faster thanks to being supported by the machine. Thus, we conclude that the ML can both support the decisions of participants and increase their confidence in responding.

In the questionnaire, nine out of ten participants answered that they preferred *audio* + *vibration* to *audio*, stating that they had more confidence compared to when relying on sound alone. Participant A, who preferred *audio*, answered that "*the vibration from the ML distracted me and I could not concentrate on the sound*". A possible solution is to use sound as the main method and offer vibration as an option, where users turn on ML only when they want it. Users should also be able to switch the sound presentation on/off when needed, since unintended sound presentation may interfere with their perception of acoustic clues from the environment.

In this study, we set the signal interval to 200 ms; however, the preferred interval was different depending on the participant (see Table 2). To examine this in greater detail, we compared the required time for object detection under the *audio* condition between participants who preferred 200 ms and those who preferred another interval. Note that participant I was excluded from this comparison because he felt both 200 and 400 ms were suitable. The average required time for participants who selected 200 ms was 58.6 s and for those who selected another interval was 82.7 s. We presume that an unsuitable signal interval may have caused extra time for object detection.

In our evaluation, we controlled the system settings to observe the changes with and without ML; however, the proposed system allows users to change various parameters (interval of signal transmission, with/without ML, with/without sound, and magnification of conversion) in accordance with their preferences. We therefore consider that the above problem can be solved by adjusting the settings depending on the preference of each user.

Although there was no significant difference in the number of incorrect answers or in the required time for detecting the object with/without ML, our goal is to reduce the burden of conventional echolocation by introducing ML. Therefore, we consider that the proposed method has a positive effect by reducing NASA-TLX, which indicates mental workload.

B. COMBINATION OF HUMAN RECOGNITION AND MACHINE RECOGNITION

As described in Section V-B1, the system transmitted the signal every 200 ms, and the converted sound and ML result were presented for each detected signal. Several participants were annoyed by the signals presented at short intervals, while in *audio* + *vibration*, most of the participants preferred the ML to respond at 200-ms intervals. On the basis of these findings, we felt that the presentation intervals of audio and vibration should be different, i.e., the system should transmit the signal every 200 ms. However, the converted sound is presented to the user at intervals of 200–600 ms depending on user preference, and the machine recognizes and presents the result to the user at intervals of 200 ms.

As shown in Fig. 19, the ratings of own performance were low (high sense of accomplishment) for most participants regardless of ML. Subject E commented that "*I felt a stronger sense of accomplishment when I used only sound because I felt that I did it on my own*". Prior studies have indicated that understanding the surroundings by sound is valuable [12], [13]. Moreover, in the experiment of participant J, the required times for detection with and without ML were 69 s and 300 s (timeout), respectively, even though the position of the object was the same for each trial. He commented that "*in the case of audio + vibration, I relied only on the machine (vibration) without noticing and did not use the sound*". Therefore, we conclude that both human and ML recognition is essential for our system.

C. STRATEGY OF MACHINE LEARNING

Participants I and J commented that "*ML misrecognition happened in an area where there was no object, and it took a long time to search for the true position of the object*". Considering actual use in a real environment, misrecognition will probably happen even more, and ML may confuse human judgment. Therefore, we consider that it is desirable to make the system a precision priority, i.e., the probability that the machine judges that the object is the user-specified object should be high even if the machine makes some mistakes in detecting objects. Moreover, although we used all of the ML recognition results in this study, we can utilize the majority voting of the latest few recognition results.

Finally, our system used a monotonous vibration of 100 ms for presenting the detection result in this study; however, participant B commented that "*it may be easier to understand if there were patterns to the vibrations*". Therefore, we consider it is effective to provide a pattern to the vibration depending on the probability of the machine result (e.g., strong and long vibration when the probability is high and weak vibration when the probability is low).

D. LIMITATIONS

1) TOWARD COOPERATIVE ECHOLOCATION

In this paper, as a first step toward cooperative echolocation, we have focused on clarifying the effectiveness of ML in

echolocation and how best to incorporate it. To the best of our knowledge, this is the first study to specifically examine cooperative echolocation. As the next steps, we need to conduct further investigation related to the effects of using visually impaired people as participants, long-term effects, and evaluations across different objects and conditions (e.g., wider/more crowded environments). We consider the findings in this paper to provide a solid basis for implementing the next steps.

In this study, we set six objects as the targets. It is preferable if the user remembers the sounds of all objects; however, of course this becomes more difficult when the number of objects increases. In such cases, the user should try to identify a place that sounds different from the surroundings, rather than to memorize the sound of every object. By obtaining ML results at the location where the user perceives a difference in sound, the user should be able to detect the object with confidence.

2) MAGNIFICATION OF TIME EXPANSION

When 1/20 magnification is used, it is desirable to use MDH of 1/20 scale; however, this was too small to insert the microphones in this study. If we utilize a smaller microphone, it will be possible to use MDH of 1/20 scale, which will make both discrimination and sound source localization easier. Also, since our system can change the magnification of conversion, users can change the magnification depending on the situation: i.e., if they want to grasp the positional relationship of surrounding objects, they can use 1/7 magnification, and if they want to recognize the kind of object, they can use 1/20 magnification. Further investigation into the relationship between magnification and ease of recognition will be needed in the future.

3) DYNAMIC CHANGE OF TRANSMITTED SIGNAL

Bats are known to dynamically change the type and interval of the signals they transmit depending on the distance to the target and the information obtained from the echoes, after which they select the optimal signal for the information they want [48], [56]. Using the ecology of bats as a reference, the proposed method can also dynamically change the transmitting signal according to the situation, which may lead to more accurate recognition for humans/machines. We leave the adaptive change of transmitting signal to future work.

4) EFFECT OF MULTIPATH AND ANGLE

In places with many reverberations, it is difficult to recognize objects because when the system simultaneously captures the reflected sound of the target object and that of another object, the frequency spectrum changes, as discussed in Section V-A1. When the user is close to a certain object, the distance from the other object presumably increases, which means the system is less susceptible to the multipath effect. Therefore, in a reverberant environment, the system should recognize objects within a close distance.

By setting the acceptable angle depending on objects, the system can detect the correct objects, as shown in Section V-B2. Since the signal transmission interval is 200 ms (5 signals/s), we consider there is a certain timing at which the user should search for the object within the acceptable angle.

5) VISUALLY IMPAIRED PARTICIPANTS

Even when sound and ML were combined, the participants required 45.9 s on average to detect an object, which is too long for use in a real environment. However, the participants were all sighted and were novices when it came to echolocation; therefore, echolocation experts will presumably be able to locate target objects faster, especially when ML is utilized. However, it is difficult to perform evaluations with echolocation experts because relatively few visually impaired people can perform echolocation. We expect that the proposed system will encourage visually impaired people who are not familiar with echolocation to try it. In future work, we plan to conduct long-term experiments with such individuals and include the training of echolocation.

Since the experiments in this study were conducted with sighted people, the participants were asked to match the sounds they heard with the actual environment by using their own eyes during the training phase. When visually impaired people use the proposed method, they will need someone to teach them the correct answer during the training phase, which is also a necessary procedure for learning the current echolocation techniques. However, the proposed method will be able to teach the correct answer through ML, thus eliminating the need for someone to teach the correct answer during the training phase.

We should also point out that the participants in this paper were young while many visually impaired people are elderly. Since our work here represents the first step toward cooperative echolocation, our focus was limited to determining whether combining echolocation with ML is effective; thus, we tested its effectiveness on participants who are easy to recruit. We will improve the system based on our findings here and conduct a long-term evaluation of the system with actual visually impaired people as the next step.

VII. CONCLUSION

In this paper, we proposed a cooperative echolocation method that combines human recognition of audible sound converted from ultrasound with recognition by ML. We implemented a prototype device and experimentally demonstrated that the recognition accuracy for six objects was 92.5% on average. We also found that the detection time was decreased from 71.5 s to 45.9 s when using ML compared to when it was not used. The results also confirmed that mental workload was significantly decreased by ML. On the basis of these findings, we were able to clarify how ML should be incorporated into echolocation.

REFERENCES

[1] World Health Organization. (2019). *World Report on Vision*. [Online]. Available: https://www.who.int/publications/i/item/9789241516570

[2] T. Mizuno, A. Nishidate, K. Tokuda, and K. Arai, "Installation errors and corrections in tactile ground surface indicators in Europe, America, Oceania and Asia," *IATSS Res.*, vol. 32, no. 2, pp. 68–80, 2008.

[3] S. Bell, P. Upchurch, N. Snavely, and K. Bala, "Material recognition in the wild with the materials in context database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3479–3487.

[4] A. Dimitrov and M. Golparvar-Fard, "Vision-based material recognition for automated monitoring of construction progress and generating building information modeling from unordered site image collections," *Adv. Eng. Informat.*, vol. 28, no. 1, pp. 37–49, Jan. 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1474034613000943

[5] C. Liu, L. Sharan, E. H. Adelson, and R. Rosenholtz, "Exploring features in a Bayesian framework for material recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 239–246.

[6] T.-C. Wang, J.-Y. Zhu, E. Hiroaki, M. Chandraker, A. A. Efros, and R. Ramamoorthi, "A 4D light-field dataset and CNN architectures for material recognition," in *Computer Vision–ECCV*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 121–138.

[7] H.-H. Pham, T.-L. Le, and N. Vuillerme, "Real-time obstacle detection system in indoor environment for the visually impaired using Microsoft Kinect sensor," *J. Sensors*, vol. 2016, pp. 1–13, Jul. 2016.

[8] M. Brock and P. O. Kristensson, "Supporting blind navigation using depth sensing and sonification," in *Proc. ACM Conf. Pervasive Ubiquitous Comput. Adjunct Publication*, Sep. 2013, pp. 255–258.

[9] V. Filipe, F. Fernandes, H. Fernandes, A. Sousa, H. Paredes, and J. Barroso, "Blind navigation support system based on Microsoft Kinect," *Proc. Comput. Sci.*, vol. 14, pp. 94–101, Jan. 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050912007739

[10] H.-C. Huang, C.-T. Hsieh, and C.-H. Yeh, "An indoor obstacle detection system using depth information and region growth," *Sensors*, vol. 15, no. 10, pp. 27116–27141, Oct. 2015. [Online]. Available: https://www.mdpi.com/1424-8220/15/10/27116

[11] L. Thaler, H. P. J. C. De Vos, D. Kish, M. Antoniou, C. J. Baker, and M. C. J. Hornikx, "Human click-based echolocation of distance: Superfine acuity and dynamic clicking behaviour," *J. Assoc. Res. Otolaryngol.*, vol. 20, no. 5, pp. 499–510, Oct. 2019.

[12] (2021). *We Teach Blind People to See With Sonarvision*. [Online]. Available: https://visioneers.org/

[13] L. J. Norman, C. Dodsworth, D. Foresteire, and L. Thaler, "Human click-based echolocation: Effects of blindness and age, and real-life implications in a 10-week training program," *PLoS ONE*, vol. 16, no. 6, pp. 1–34, Jun. 2021.

[14] S. Uchibori, Y. Sarumaru, K. Ashihara, T. Ohta, and S. Hiryu, "Experimental evaluation of binaural recording system using a miniature dummy head," *Acoust. Sci. Technol.*, vol. 36, no. 1, pp. 42–45, 2015.

[15] D. Storek, J. Stuchlik, and F. Rund, "Modifications of the surrounding auditory space by augmented reality audio: Introduction to warped acoustic reality," in *Proc. Int. Conf. Auditory Display*, 2015, pp. 225–230.

[16] H. Watanabe and T. Terada, "Manipulatable auditory perception in wearable computing," in *Proc. Augmented Humans Int. Conf.*, Mar. 2020, pp. 1–12.

[17] M. Sumiya, K. Ashihara, K. Yoshino, M. Gogami, Y. Nagatani, K. I. Kobayasi, Y. Watanabe, and S. Hiryu, "Bat-inspired signal design for target discrimination in human echolocation," *J. Acoust. Soc. Amer.*, vol. 145, no. 4, pp. 2221–2236, Apr. 2019.

[18] L. Thaler, R. De Vos, D. Kish, M. Antoniou, C. Baker, and M. Hornikx, "Human echolocators adjust loudness and number of clicks for detection of reflectors at various azimuth angles," *Proc. Roy. Soc. B, Biol. Sci.*, vol. 285, no. 1873, Feb. 2018, Art. no. 20172735.

[19] L. Thaler, X. Zhang, M. Antoniou, D. C. Kish, and D. Cowie, "The flexible action system: Click-based echolocation may replace certain visual functionality for adaptive walking," *J. Experim. Psychology, Hum. Perception Perform.*, vol. 46, no. 1, pp. 21–35, 2020.

[20] J. Sohl-Dickstein, S. Teng, B. M. Gaub, C. C. Rodgers, C. Li, M. R. DeWeese, and N. S. Harper, "A device for human ultrasonic echolocation," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 6, pp. 1526–1534, Jun. 2015.

[21] Y. Sarumaru, M. Sumiya, T. Banda, K. Ashihara, K. Kobayashi, and S. Hiryu, "Human echo perception by using miniature dummy head," in *Proc. Auditory Res. Meeting*. Japan: Acoustical Society of Japan, 2015, pp. 57–62.

[22] B. Li, J. P. Muñoz, X. Rong, J. Xiao, Y. Tian, and A. Arditi, "ISANA: Wearable context-aware indoor assistive navigation with obstacle avoidance for the blind," in *Proc. Comput. Vis. Workshops (ECCV)*, G. Hua and H. Jégou, Eds. Cham, Switzerland: Springer, 2016, pp. 448–462.

[23] S. Kayukawa, K. Higuchi, J. A. Guerreiro, S. Morishima, Y. Sato, K. Kitani, and C. Asakawa, "BBeep: A sonic collision avoidance system for blind travellers and nearby pedestrians," in *Proc. CHI Conf. Hum. Factors Comput. Syst.* New York, NY, USA: ACM, 2019, p. 52.

[24] S. Kayukawa, T. Ishihara, H. Takagi, S. Morishima, and C. Asakawa, "Guiding blind pedestrians in public spaces by understanding walking behavior of nearby pedestrians," in *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 4, no. 3, Sep. 2020, pp. 1–22.

[25] Y. Zhao, E. Kupferstein, B. V. Castro, S. Feiner, and S. Azenkot, "Designing AR visualizations to facilitate stair navigation for people with low vision," in *Proc. 32nd Annu. ACM Symp. User Interface Softw. Technol.* New York, NY, USA: Association for Computing Machinery, 2019, pp. 387–402, doi: 10.1145/3332165.3347906.

[26] H.-C. Wang, R. K. Katzschmann, S. Teng, B. Araki, L. Giarre, and D. Rus, "Enabling independent navigation for visually impaired people through a wearable vision-based feedback system," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 6533–6540.

[27] S. Xu, C. Yang, W. Ge, C. Yu, and Y. Shi, "Virtual paving: Rendering a smooth path for people with visual impairment through vibrotactile and audio feedback," in *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, Sep. 2020, vol. 4, no. 3, pp. 1–25.

[28] R. Boldu, D. J. C. Matthies, H. Zhang, and S. Nanayakkara, "AiSee: An assistive wearable device to support visually impaired grocery shoppers," in *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, Dec. 2020, vol. 4, no. 4, pp. 1–25.

[29] V. Kulyukin, C. Gharpure, and J. Nicholson, "RoboCart: Toward robot-assisted navigation of grocery stores by the visually impaired," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Aug. 2005, pp. 2845–2850.

[30] P. A. Zientara, S. Lee, G. H. Smith, R. Brenner, L. Itti, M. B. Rosson, J. M. Carroll, K. M. Irick, and V. Narayanan, "Third eye: A shopping assistant for the visually impaired," *Computer*, vol. 50, no. 2, pp. 16–24, Feb. 2017.

[31] J. Nicholson, V. Kulyukin, and D. Coster, "ShopTalk: Independent blind shopping through verbal route directions and barcode scans," *Open Rehabil. J.*, vol. 2, no. 1, pp. 11–23, Mar. 2009.

[32] W. Jin, M. Xiao, H. Zhu, S. Deb, C. Kan, and M. Li, "Acoussist: An acoustic assisting tool for people with visual impairments to cross uncontrolled streets," in *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 4, no. 4, pp. 1–30, Dec. 2020.

[33] B. Mocanu, R. Tapu, and T. Zaharia, "When ultrasonic sensors and computer vision join forces for efficient obstacle detection and recognition," *Sensors*, vol. 16, no. 11, p. 1807, 2016. [Online]. Available: https://www.mdpi.com/1424-8220/16/11/1807

[34] A. Davis, K. L. Bouman, J. G. Chen, M. Rubinstein, O. Büyüköztürk, F. Durand, and W. T. Freeman, "Visual vibrometry: Estimating material properties from small motions in video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 732–745, Apr. 2017.

[35] M. Sato, S. Yoshida, A. Olwal, B. Shi, A. Hiyama, T. Tanikawa, M. Hirose, and R. Raskar, "SpecTrans: Versatile material classification for interaction with textureless, specular and transparent surfaces," in *Proc. 33rd Annu. ACM Conf. Hum. Factors Comput. Syst.* New York, NY, USA: Association for Computing Machinery, Apr. 2015, pp. 2191–2200, doi: 10.1145/2702123.2702169.

[36] H.-S. Yeo, J. Lee, A. Bianchi, D. Harris-Birtill, and A. Quigley, "SpeCam: Sensing surface color and material with the front-facing camera of a mobile device," in *Proc. 19th Int. Conf. Hum.-Comput. Interact. Mobile Devices Services*. New York, NY, USA: Association for Computing Machinery, Sep. 2017, pp. 1–9.

[37] H.-S. Yeo, G. Flamich, P. Schrempf, D. Harris-Birtill, and A. Quigley, "RadarCat: Radar categorization for input & interaction," in *Proc. 29th Annu. Symp. User Interface Softw. Technol.*, Oct. 2016, pp. 833–841, doi: 10.1145/2984511.2984515.

[38] C. Harrison and S. E. Hudson, "Lightweight material detection for placement-aware mobile computing," in *Proc. 21st Annu. ACM Symp. User interface Softw. Technol. (UIST)*, 2008, pp. 279–282, doi: 10.1145/1449715.1449761.

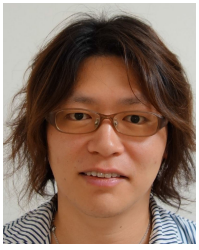
[39] H. Liu, X. Song, J. Bimbo, L. Seneviratne, and K. Althoefer, "Surface material recognition through haptic exploration using an intelligent contact sensing finger," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, May 2012, pp. 52–57.
[40] Y. Cho, N. Bianchi-Berthouze, N. Marquardt, and S. J. Julier, "Deep thermal imaging: Proximate material type recognition in the wild through deep learning of spatial surface temperature patterns," in *Proc. CHI Conf. Hum. Factors Comput. Syst.* New York, NY, USA: Association for Computing Machinery, Apr. 2018, pp. 1–13, doi: 10.1145/3173574.3173576.
[41] T. Komatsu and J.-I. Akita, "Power spectrum analysis of reflected waves with ultrasonic sensors indicates 'what target is,'" in *Proc. 13th ACM Conf. Embedded Networked Sensor Syst.*, 2015, pp. 387–388.
[42] W. Mao, W. Sun, M. Wang, and L. Qiu, "DeepRange: Acoustic ranging via deep learning," in *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, Dec. 2020, vol. 4, no. 4, pp. 1–23.
[43] B. Zhou, M. Elbadry, R. Gao, and F. Ye, "BatMapper: Acoustic sensing based indoor floor plan construction using smartphones," in *Proc. 15th Annu. Int. Conf. Mobile Syst., Appl., Services*. New York, NY, USA: Association for Computing Machinery, Jun. 2017, pp. 42–55, doi: 10.1145/3081333.3081363.
[44] Y.-C. Tung and K. G. Shin, "Use of phone sensors to enhance distracted Pedestrians' safety," *IEEE Trans. Mobile Comput.*, vol. 17, no. 6, pp. 1469–1482, Jun. 2018.
[45] Z. Wang, S. Tan, L. Zhang, and J. Yang, "ObstacleWatch: Acoustic-based obstacle collision detection for pedestrian using smartphone," in *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 4, p. 194, Dec. 2018.
[46] L. Remaggi, H. Kim, P. J. B. Jackson, and A. Hilton, "An audio-visual method for room boundary estimation and material recognition," in *Proc. Workshop Audio-Visual Scene Understand Immersive Multimedia*, 2018, pp. 3–9.
[47] W. Mao, M. Wang, and L. Qiu, "AIM: Acoustic imaging on a mobile," in *Proc. 16th Annu. Int. Conf. Mobile Syst., Appl., Services*, Jun. 2018, pp. 468–481.
[48] H.-U. Schnitzler and E. K. V. Kalko, "Echolocation by insect-eating bats: We define four distinct functional groups of bats and find differences in signal structure that correlate with the typical echolocation tasks faced by each group," *BioScience*, vol. 51, no. 7, pp. 557–569, Jul. 2001.
[49] S. Fritz and M. Lusardi, "White paper: Walking speed: The 6th vital sign," *J. Geriatric Phys. Therapy*, vol. 32, no. 2, pp. 2–5, 2009. [Online]. Available: https://journals.lww.com/jgpt/Fulltext/2009/32020/White_Paper___Walking_Speed__the_Sixth_Vital_Sign_.2.aspx
[50] P. Elektronik. (2019). *Bat Detectors and Sound Analysis Software*. [Online]. Available: http://www.batsound.com/
[51] J. Laroche and M. Dolson, "New phase-vocoder techniques for pitch-shifting, harmonizing and other exotic effects," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoustics.*, Oct. 1999, pp. 91–94.
[52] K. Watanabe, Y. Iwaya, Y. Suzuki, S. Takane, and S. Sato, "Dataset of head-related transfer functions measured with a circular loudspeaker array," *Acoust. Sci. Technol.*, vol. 35, no. 3, pp. 159–165, 2014.
[53] S. S. Stevens, J. Volkmann, and E. B. Newman, "A Scale for the measurement of the psychological magnitude pitch," *J. Acoust. Soc. Amer.*, vol. 8, no. 3, pp. 185–190, Jan. 1937.
[54] H. Lei and E. Lopez, "Mel, linear, and antimel frequency cepstral coefficients in broad phonetic regions for telephone speaker recognition," in *Proc. Interspeech*, Sep. 2009, pp. 2323–2326.
[55] S. G. Hart, "Nasa-task load index (NASA-TLX); 20 years later," in *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, 2006, vol. 50, no. 9, pp. 904–908.
[56] E. Fujioka, S. Mantani, S. Hiryu, H. Riquimaroux, and Y. Watanabe, "Echolocation and flight strategy of Japanese house bats during natural foraging, revealed by a microphone array system," *J. Acoust. Soc. Amer.*, vol. 129, no. 2, pp. 1081–1088, Feb. 2011.

HIROKI WATANABE received the B.E., M.E., and Ph.D. degrees in engineering from Kobe University, Hyogo, Japan, in 2012, 2014, and 2017, respectively.

From 2015 to 2017, he was a Research Fellow of the Japan Society for the Promotion of Science (DC2), Kobe University. Since 2017, he has been an Assistant Professor at the Graduate School of Information Science and Technology, Hokkaido University, Hokkaido, Japan. His research interests include wearable computing and ubiquitous computing.

MIWA SUMIYA received the B.E., M.E., and Ph.D. degrees in engineering from Doshisha University, Kyoto, Japan, in 2013, 2015, and 2018, respectively.

From 2015 to 2018, she was a Research Fellow of the Japan Society for the Promotion of Science (DC1), Doshisha University. From 2018 to 2021, she was a Research Fellow of the Japan Society for the Promotion of Science (PD), National Institute of Information and Communications Technology, Kyoto. Since 2021, she has been an Overseas Research Fellow of the Japan Society for the Promotion of Science, Durham University, U.K. Her research interest includes echolocation.



TSUTOMU TERADA (Member, IEEE) received the B.E., M.E., and Ph.D. degrees in engineering from Osaka University, Osaka, Japan, in 1997, 1999, and 2003, respectively.

From 2000 to 2004, he was an Assistant Professor at the Cybermedia Center, Osaka University, where he was a Lecturer, from 2005 to 2007. From 2007 to 2018, he was an Associate Professor at the Graduate School of Engineering, Kobe University, Hyogo, Japan, where he has been a Professor, since 2018. His research interests include wearable computing, ubiquitous computing, and entertainment computing.