EarHover: Mid-Air Gesture Recognition for Hearables Using Sound Leakage Signals

Shunta Suzuki
shunta130904@keio.jp
Keio University
Yokohama, Japan

Takashi Amesaka
amesaka@keio.jp
Keio University
Yokohama, Japan

Hiroki Watanabe
hwata@fun.ac.jp
Future University Hakodate
Hakodate, Japan

Buntarou Shizuki
shizuki@cs.tsukuba.ac.jp
University of Tsukuba
Tsukuba, Japan

Yuta Sugiura
sugiura@keio.jp
Keio University
Yokohama, Japan

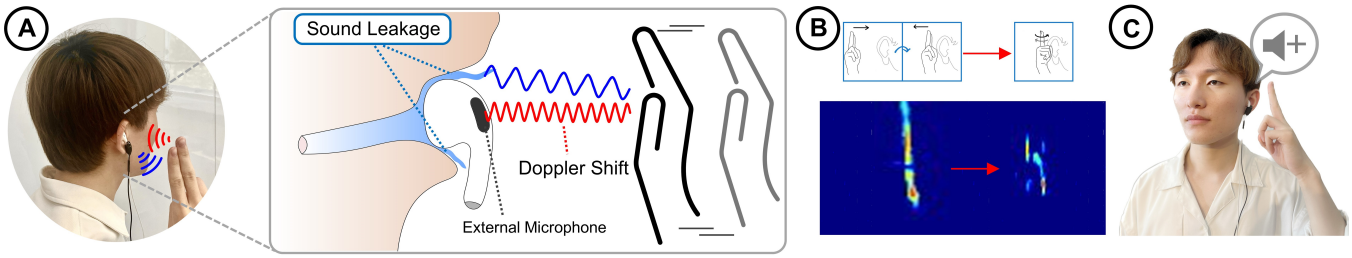


Figure 1: Overview and usage images of EarHover, A: EarHover recognizes mid-air gestures using sound leakage and Doppler shift, B: spectrogram for each mid-air gesture, C: operation image in a music application. EarHover can convert existing touch gesture input into user-demanded mid-air gestures. It also complements existing input methods and allows users to control more applications from their hearables.

ABSTRACT

We introduce EarHover, an innovative system that enables mid-air gesture input for hearables. Mid-air gesture input, which eliminates the need to touch the device and thus helps to keep hands and the device clean. However, existing mid-air gesture input methods for hearables have been limited to adding cameras or infrared sensors. By focusing on the sound leakage phenomenon unique to hearables, we have realized mid-air gesture recognition using a speaker and an external microphone that are highly compatible with hearables. The signal leaked to the outside of the device due to sound leakage can be measured by an external microphone, which detects the differences in reflection characteristics caused by the hand's speed and shape during mid-air gestures. Among 27 types of gestures, we determined the seven suitable gestures for EarHover in terms of signal discrimination and user acceptability. We then evaluated the gesture detection and classification performance of two prototype devices (in-ear type/open-ear type) for real-world application scenarios.

This work is licensed under a Creative Commons Attribution International 4.0 License.

UIST '24, October 13–16, 2024, Pittsburgh, PA, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0628-8/24/10
https://doi.org/10.1145/3654777.3676367

CCS CONCEPTS

• **Human-centered computing** → **Human-computer interaction (HCI)**; *Ubiquitous and mobile computing systems and tools.*

KEYWORDS

Hearables, mid-air gesture recognition, Doppler effect, sound leakage, deep learning, acoustic sensing

ACM Reference Format:
Shunta Suzuki, Takashi Amesaka, Hiroki Watanabe, Buntarou Shizuki, and Yuta Sugiura. 2024. EarHover: Mid-Air Gesture Recognition for Hearables Using Sound Leakage Signals. In *The 37th Annual ACM Symposium on User Interface Software and Technology (UIST '24), October 13–16, 2024, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3654777.3676367

1 INTRODUCTION

Hearables, ear-worn wearable devices, have become increasingly popular in recent years. Beyond standard earphone functions like calling and music playback, hearables now facilitate voice assistants, audio services, and health tracking [13, 33, 56]. Operation of hearables can be either indirect via a smartphone or direct on the device itself. Indirect operation via smartphones requires screen interaction, which reduces usability. For direct manipulation, many hearables offer hands-free methods such as voice control. They also offer manual operations using capacitive sensors, pressure-sensitive sensors, or physical buttons. Voice control allows operation even

with occupied hands, but there are issues such as hesitancy in speaking in public or inability to use the system during a conversation. In addition, recognition accuracy decreases in noisy environments. Hand input requires the user to press down on the sensor part with a finger, which places a physical strain on the ear and generates noise. In addition, while these sensors execute commands by varying the duration and frequency of presses, there is a limit to the number of variations; for instance, AirPods Pro [6] supports six commands: one to three taps, long press, and swipe up or down. Thus, using applications such as checking email or appointments while listening to music requires more input.

Many hands-input methods [1, 25, 27, 28, 31, 43, 54] have been proposed to address these issues, divided into touch and mid-air gesture inputs. Touch gestures leverage sound or ear deformation generated by touching the ear or face. While touch gestures provide tactile feedback and are easy to use, their need for direct contact makes them impractical when hands are dirty, such as when eating or working, or must be kept clean, such as in cooking or medical settings. Similarly to manual operations, touch gesture inputs also cannot prevent noise during operation. On the other hand, the mid-air gesture input method allows gesture input by hand movement and shape without touching the device. Therefore, input is possible regardless of the condition of the hand, and there is no physical strain on the ear or noise caused by pressing down on the device. In addition, Gesture Elicitation Studies (GES) [47], which assume hearables and hand input to the ear, highlight a significant use of mid-air gestures [16, 32]. Previous mid-air gesture recognition methods for hearables have utilized bone-conduction speaker/microphone [55] , cameras [43], or infrared sensors [31]. However, cameras and infrared sensors are not commonly used in hearables and require additional sensor and design costs for installation, making them less feasible. Furthermore, previous study has primarily focused on bone-conduction earphones, lacking comprehensive research that encompasses various types of hearables, such as in-ear and open-ear devices.

We propose EarHover (Figure 1), a system that minimizes the cost of additional sensors and enables mid-air gesture recognition. EarHover leverages sound leakage signals and the Doppler shift in hearables, allowing for implementation using speakers and external microphones. The sound leakage signal is the component of the signal played from hearables that leaks outward and has been used to measure ear shape [3]. In this study, we extend the measurement range to the area around the device and use it for user mid-air gesture recognition. Specifically, when the user performs a mid-air gesture near the device, the sound leakage signal is reflected off the hand and measured by the external microphone. Since each mid-air gesture is performed at a different speed, angle, and hand shape, the reflected signal is different and can be recognized by a deep-learning model. The speaker and external microphone used by EarHover are sensors widely installed in hearables for music playback, voice recording, and noise cancellation; therefore, the cost of implementing EarHover is very low compared to existing methods. In this study, we conducted evaluation experiments on two prototype devices (in-ear type/open-ear type) to demonstrate the feasibility of mid-air gesture recognition using sound leakage signals. The results of this study are as follows:

- We evaluated the effect of EarHover on the auditory experience when listening to music (N=20). The experimental results showed that the percentage of users who perceived a difference in music with and without signal playback was 50.0%. Even among users who perceived a difference, there was no significant difference in preference between the presence and absence of the signal.
- We conducted gesture detection experiments with in-ear and open-ear prototype devices. The experimental results confirmed that the false detection rate for 11 types of daily activities was 1.8% on average (in-ear: 2.5%, open-ear: 1.0%), and the false detection rate for seven types of gesture movements was 10.6% on average (in-ear: 10.6%, open-ear: 10.7%).
- We conducted gesture recognition experiments using a prototype device in five environments: Stable, Gloves, Neighbor, Walking, and Music. The experiments yielded a 79.7–88.8% gesture classification rate for the seven gestures.

We introduced EarHover, a method for mid-air gesture input in hearables, devices that are anticipated to be equipped with a wide range of functions soon. We evaluated the system in various assumed usage environments. This study's contributions are listed below:

- We introduced the first mid-air gesture input method for hearables based on sound leakage that can be implemented with a built-in speaker and an external microphone. This is a significant step toward practical mid-air gesture control in hearables.
- We designed 27 mid-air gestures for EarHover and selected seven based on their discriminability and user preference, marking the method's first application.
- We demonstrated that EarHover's impact on music listening is minimal, confirming its feasibility through the assessment of recognition performance across two device types in real-world application scenarios.

2 RELATED WORK

2.1 Input with Hearables

Input methods for hearables can be categorized as hand input, which uses hand movements and shapes, and hands-free input, which uses head-related movements.

2.1.1 Hand Input. Hand input methods [1, 18, 28, 36] can be divided into touch gesture input and mid-air gesture input. In the study on touch gesture input, Xu et al. [54] proposed a system that recognizes tapping and sliding gestures near the face and ear using the device's built-in microphone. In evaluation experiments, the system achieved a gesture recognition rate of 95.3% for eight types of gestures. Kikuchi et al. [25] used four photo reflectors attached to hearables to measure the deformation caused by pulling on the ear for gesture recognition. Lissermann et al. [27] developed open-ear devices with capacitive sensors to expand the device's input area.

In the study on mid-air gesture input, Yang et al. [55] proposed a gesture input system using bone-conduction earphone devices and also investigated mid-air gesture recognition using the phenomenon of sound leakage. Tamaki et al. [43] proposed a system that

estimates hand contours, fingernail positions, and finger joint angles from a camera mounted on the hearables. This system demonstrates an application for hover gesture recognition. Metzger et al. [31] proposed a system that uses an infrared proximity sensor to recognize mid-air swipes and other gestures around the device.

Touch gesture input, which requires contact with the face or ear, can be difficult to use when hands are wet or dirty or need to be kept clean. Mid-air gesture input, using cameras or photo reflectors, presents challenges with sensor costs and privacy issues. EarHover's built-in speaker and external microphone enable music playback and voice recording, increase compatibility with hearables, and keep implementation costs low. Yang et al.'s method [55] uses a similar principle to EarHover. However, our study focuses on in-ear/open-ear devices that generate air-conducted sound. Additionally, we analyzed the sensitivity to synthetic music signals, the impact of hand conditions, and the detection of gestures in everyday activities.

2.1.2 Hands-Free Input. Many hands-free input methods sensing head-related movements have been explored [2, 4, 12, 23, 26, 41, 44, 45]. Manabe et al. [29] detected eye gestures by inserting conductive rubber into the ear tips of hearables to measure eye movements. Denys et al. [30] proposed a facial gesture recognition system using myoelectric sensors attached to the ear tip. Tobias et al. [34] created a prototype device with a speaker and a pressure sensor with the ear tip and proposed a system to realize hands-free input by measuring the tinnitus sound in the eardrum. Sun et al. [42] proposed a system that monitors ear canal deformation with a speaker and an internal microphone, facilitating silent speech interaction.

Compared to hand input, hands-free input does not require the use of the hands, making it suitable for situations where hands are occupied, like carrying luggage. However, subtle facial movements can be confused with normal daily activities. In addition, oral command input requires wake words or mouth movements. This can be time-consuming for continuous input and challenging during conversations. Therefore, both input methods are likely to be used depending on the user's situation, which requires further study.

2.2 Acoustic-Based Recognition

Like EarHover, many studies use acoustic signals to capture reflection and reverberation, and recognize user gesture [35, 48, 51, 58], biometric features [21, 50], and motion [19, 53]. Yongpan et al. [57] proposed a finger-tracking system that recognizes numbers and alphabets drawn mid-air using a smartwatch and a smartphone speaker-microphone. Mingshi et al. [15] proposed a system that recognizes number-drawing gestures in mid-air around smartwatches. Yang et al. [20] proposed a word recognition system using tone signals from a smartphone to measure the Doppler shift of mouth and tongue movements. Yingcheng et al. [24] proposed a system that plays ultrasonic tone signals from a loudspeaker mounted outside the headphones to recognize sign language actions of the person in front of the user. Amesaka et al. [3] proposed a personal authentication system that acquires and analyzes ear shape information through sound leakage of a sweep signal.

Gesture recognition methods using smartphones/smartwatches or hearables with external speakers can send the measurement signal directly to the target, making it easy to acquire the necessary signal strength. In contrast, EarHover uses the leakage signal from hearables, which requires careful sound pressure adjustment to avoid hearing loss. The signal is also attenuated by reflection from the outer ear. In personal authentication systems that use acoustic leakage signals, the target is the shape of the ear, typically using in-ear devices. In this study, we expanded the sensing target of the sound leakage signal to the space above the device and demonstrated its ability to detect hand movements during mid-air gestures with prototype devices of two different shapes. Although EarHover is not yet compatible with current commercial hearables, its design principle offers a clear theoretical path for integration, highlighting the potential for adaptation to existing hardware.

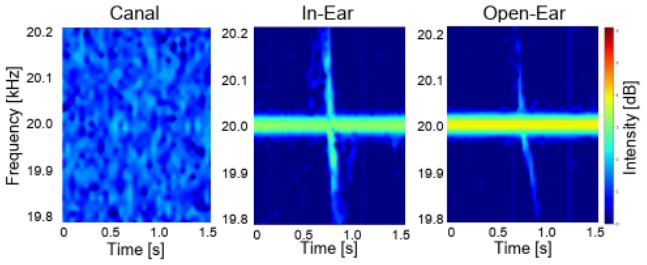


Figure 2: Sound leakage and Doppler shift phenomena during mid-air gesture for each device type.

3 EARHOVER

3.1 Device Insights

Hearables come in three main types: canal-type, in-ear, and open-ear. Canal-type devices [6, 38] feature ear tips that seal the ear canal, providing high airtightness and often including noise cancellation capabilities. In-ear devices [5, 7], although inserted into the ear canal opening, do not have ear tips, resulting in less airtightness and reduced ear blockage compared to canal-type devices. On the other hand, open-ear devices [37, 46] do not block the ear canal opening, offering the lowest level of airtightness and allowing ambient sounds to be easily heard.

We examined their sound leakage characteristics to determine if the devices were compatible with EarHover. We played a 20 kHz tone signal at approximately 45 dB from a device worn on the ear and measured the sound leakage with an additionally worn microphone. Using a separate microphone for measurement serves two purposes: to standardize conditions and to address the limitation that the built-in external microphones of the devices do not support recording at a sampling rate sufficient for 20 kHz signals. The tested devices were Apple's AirPods Pro [6] for canal-type, Apple's EarPods [5] for in-ear, and Victor's HA-NP35 [46] for open-ear, recorded using an audio-technica microphone [10]. Figure 2 shows the sound leakage for each device type when moving a hand back and forth above the device. Sound leakage and Doppler shifts were absent in the canal type, while in-ear and open-ear types had detectable sound leakage and Doppler shifts. Based on these findings, this study targets in-ear and open-ear types for EarHover utilization.

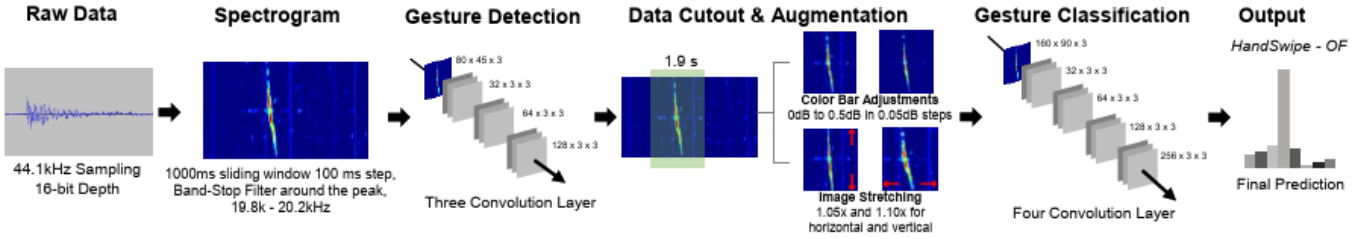


Figure 3: Overview of the EarHover pipeline, which includes converting signals into spectrogram images, detecting gestures using a 3-layer CNN model, applying data cutout and augmentation, and classifying gestures using a 4-layer CNN model.

The actual usage of different types of hearables among users has yet to be fully understood, and the impact of EarHover's potential commercialization remains uncertain. To estimate the future usage of EarHover, we conducted a crowdsourced survey of 332 participants who use earphones or headphones at least once a week. The survey revealed that 115 (33.9%) users used in-ear type, and 61 (18.0%) used open-ear type devices. These results indicate that EarHover technology has the potential for widespread market adoption and underscore its value as a new interaction method for users of in-ear and open-ear devices.

3.2 System

Figure 3 shows an overview of the EarHover system. Generally, commercially available hearables have a reproducible frequency bandwidth of 20 Hz–22.05 kHz. This study uses a 20 kHz sine wave for the sound leakage signal, as commonly used in related work. The system continuously plays the signal and records the sound leakage with an external microphone. The recognition model consists of two stages: gesture detection and gesture classification. The system uses a gesture detection model with 80×45 spectrogram images as input to detect gestures in daily life. When a gesture is detected, a 160×90 spectrogram image is created and fed into the gesture classification model to determine the gesture type.

3.3 Gesture Detection Model

3.3.1 Preprocessing and Spectrogram Image Generation. The detection model operates by extracting signals in 1000 ms, sliding in steps of 100 ms, and repeatedly activating the detection model on the extracted signals. Bandstop filtering is applied to the extracted sound leakage signal to eliminate the 20 kHz frequency component before calculating the spectrogram. The FFT window size is 100 ms, using a Hamming window for the window function to calculate the spectrogram. Window overlap was set at 100 ms. The spectrogram data is converted to a spectrogram image, trimmed to only the band where the Doppler shift occurs, and 80×45 image data is generated. Since the gesture detection model operates continuously, we opted for a smaller image size to reduce computational costs. Binary classification between everyday actions and gestures is generally not highly challenging because everyday actions and specific gestures exhibit distinct patterns and characteristics. Consequently, we decided to use a smaller image size compared to the gesture classification model. For the color bar, the minimum and maximum values were set at 0.5 dB and 2.0 dB, respectively, during image generation.

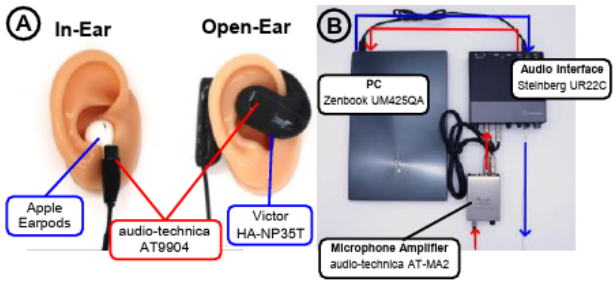


Figure 4: A: prototype devices, B: device configurations (blue line: playback signal path, red line: recording signal path).

3.3.2 Model Training. We collected spectrogram image data of gesture movements and daily activities and developed a deep learning-based binary classifier. The model architecture consists of three convolutional layers with 32, 64, and 128 filters, respectively, followed by batch normalization and max pooling layers. These layers lead to fully connected layers with 128 and 64 units and culminate in a softmax layer for two-class classification. We set the training parameters with a batch size of 16 for 20 epochs and implemented early stopping with a patience of five epochs. The detection algorithm retains the predictions from the last three instances, and if more than two of these are for a gesture, the gesture is determined to have occurred.

3.4 Gesture Classification Model

When a gesture is detected, the system extracts a 1.9 s sound leakage signal and generates a spectrogram image following the procedure described in Section 3.3.1. The cutout length was set to cover 98% of the annotated mid-air gesture data. Since gesture classification is a multi-label task, higher resolution 160×90 size images were used as input. Spectrogram images of the target gesture were collected to create a multi-classification model that predicts the gesture type.

3.4.1 Data Augmentation. EarHover users lack tactile feedback and visual confirmation of their actions, leading to hand position and angle variations during gesture execution. We applied two data augmentation methods to spectrogram images to generate a robust CNN model for this motion noise.

First, we adjusted the color bar of the spectrogram images to reflect the intensity changes due to varying distances between the hand and the device. This simulates the hand position variability.

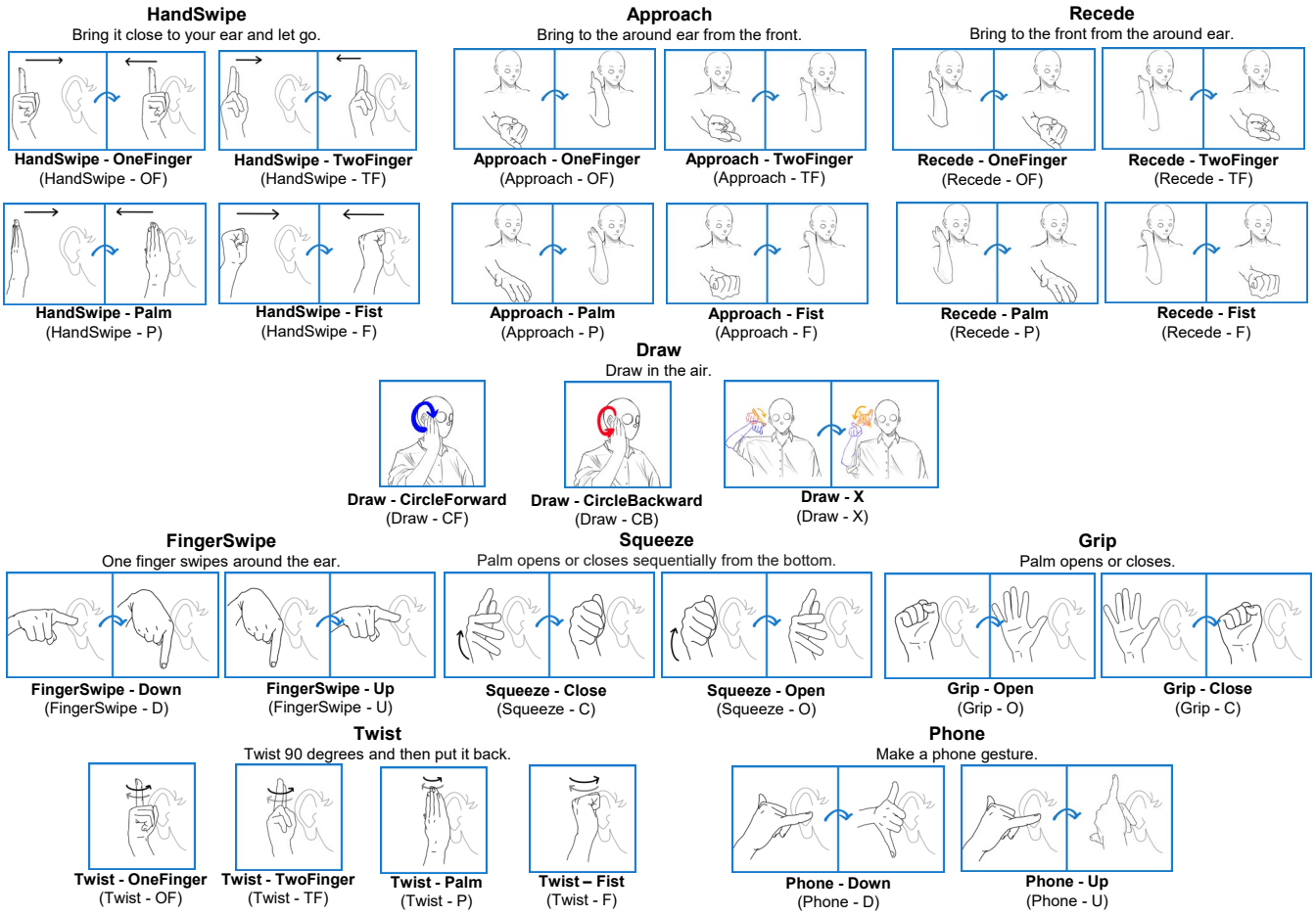


Figure 5: Tested 27 mid-air gestures.

In this study, we varied the minimum sound level from 0.0 dB to 0.5 dB in 0.05 dB increments, resulting in an 11-fold augmentation.

Second, we applied stretching to the spectrogram images: vertical stretching to represent frequency changes due to speed variations and horizontal stretching to mimic different gesture durations. We scaled each image horizontally and vertically by 1.00x (original), 1.05x, and 1.10x, resulting in three variations for each axis. This approach produced nine versions per original image, resulting in a nine-fold data augmentation.

3.4.2 Model Training. We collected gesture spectrogram images and performed data augmentation before building a deep learning-based multi-class classification model. The model architecture includes four convolutional layers of 32, 64, 128, and 256 filters each, followed by batch normalization and max pooling layers. After flattening, the model includes fully connected layers of 128 and 64 units, ending with a softmax layer for multi-class classification. We set the training parameters with a batch size of 16 for 200 epochs and implemented early stopping with patience of 20 epochs.

4 IMPLEMENTATION

Figure 4 shows two experimental prototype devices and device configurations. Commercial in-ear/open-ear devices generally do not allow recording in the high-frequency range (8–10 kHz>) to minimize power consumption and avoid capturing unnecessary noise. This is because the microphone in these devices is primarily designed to record human voices and low-frequency noise (40–1500 Hz) for noise canceling [11]. Therefore, we attached a separate microphone capable of high-frequency bandwidth recording to hearables to investigate the gesture recognition performance of EarHover. We used Apple's EarPods [5] as an in-ear type device, Victor's HA-NP35T [46] as an open-ear type, and audio-technica's AT9904 [10] (omnidirectional) as an additional microphone. A microphone amplifier (audio-technica: AT-MA2 [9]) was used to amplify the measured acoustic data. The microphone and hearables were connected to a PC (ASUS Zenbook 14 UM425QA [8]) for recording and playback, and an audio interface (Steinberg: UR22C [39]) was used for AD/DA conversion during playback and recording of the EarPods and AT9904. The HA-NP35T was connected to the PC via Bluetooth standard Ver. 5.1.

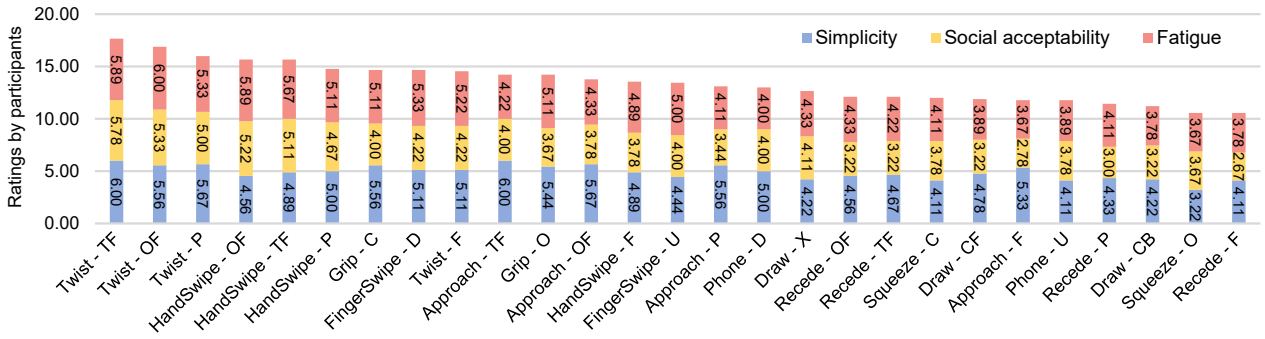


Figure 6: Ratings by all participants in terms of simplicity, social acceptability, and fatigue (reversed).

The World Health Organization recommends a volume limit of 80 dB for 40 hours per week for adults as a guide for volume settings in using hearables [52]. In this study, we considered the health risk and set the signal reaching the eardrum to approximately 45 dB, which does not cause health problems. The sound leakage signal is approximately 40 dB at the external microphone for both devices and is amplified by approximately 35 dB through the amplifier. The signal recording program was implemented in Python 3.10, and the sampling rate was set to 44.1 kHz and the quantization bit rate to 16 bits.

5 GESTURE SELECTION EXPERIMENT

In this experiment, we formulated a set of gestures for EarHover based on three criteria: Doppler Shift Magnitude (DSM), Doppler Shift Similarity (DSS), and usability. DSM is the magnitude of the Doppler shift portion in the acoustic signal, and DSS is the similarity of the Doppler shift distribution among multiple acoustic signals. Our selection process, inspired by the methodology of Xu et al. [54], was adapted to meet our specific needs and context.

5.1 Mid-Air Gesture Design

Yu-Chun et al. [16] conducted a Gesture Elicitation Study (GES) around the ear to explore user interaction, gesture classification, and design space. Several trends were observed in the defined mid-air gestures. Hand shapes were categorized into five types: one finger, two fingers, palm, fist, and phone sign. Notably, apart from the phone sign, many gestures were defined by transitions between these shapes or the opening and closing of fingers. Gestures were also defined by the direction of hand movement and the trajectory of the fingertips. Our gesture design was informed by these trends and the necessity of inducing Doppler shifts, which is essential for our method. Figure 5 shows the 27 types of mid-air gestures conceived. The designed gestures include movements starting from the device in different directions (hand swipe, approach, recede, draw) and changing the shape or orientation of the hand around the device (finger swipe, squeeze, grip, twist, phone). From these gestures, we determined the mid-air gestures to be used in the recognition experiments.

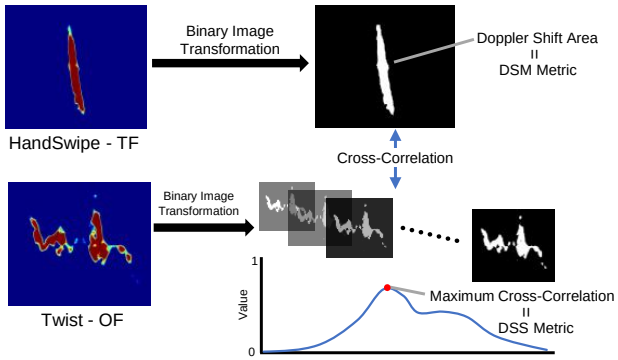


Figure 7: Flow of derivation of Doppler shift magnitude (DSM) and Doppler shift similarity (DSS).

5.2 Procedure

In a quiet laboratory, we measured each mid-air gesture five times with participants seated in a chair, ensuring no reflective objects were around. The 11 participants (male: 8, female: 3) were all right-handed, with a mean age of 27.9 years (standard deviation: 10.7 years). Participants used their hearables a mean of 4.13 days per week. The devices that participants used daily were either canal, in-ear, or open-ear type.

First, participants put the device on their right ear. The experimenter illustrated each mid-air gesture and verbally explained it to the participants. Participants practiced the gestures for several minutes. For the measurement, the experimenter ensured the participant's hand was on their lap before starting the signal playback and sound leakage measurement, then displayed the mid-air gesture illustration. The participant was asked to perform the displayed mid-air gesture and then return the hand to the lap. This process was repeated five times for each mid-air gesture. After completing the data measurements, participants completed three questionnaires. A scale of 1 to 7 was used, with 7 being the highest rating. (1: very bad, 7: very good)

- Simplicity: How easy is it to perform the gesture accurately?
- Social acceptability: How socially acceptable do you feel this gesture is?
- Fatigue: How tiring is the gesture? (Note: scale has been inverted for analysis)

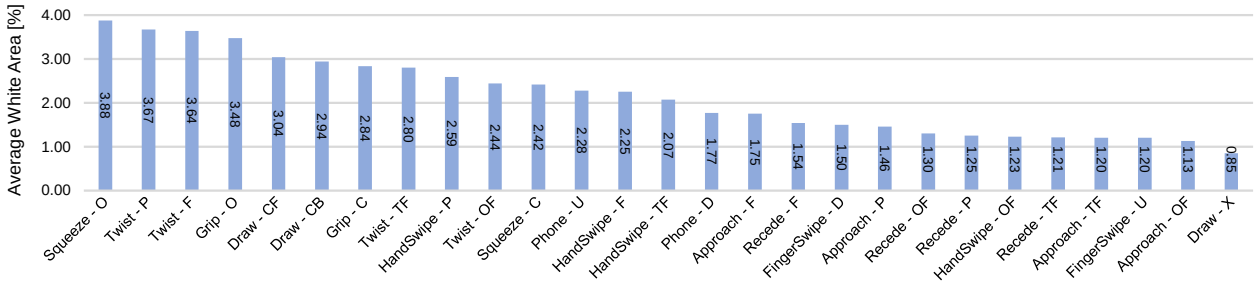


Figure 8: Average ratio of the white area for 27 gestures.

	HandSwipe - TF	HandSwipe - P	HandSwipe - F	Twist - OF	Twist - TF	Twist - P	Twist - F	Approach - F	Recede - F	FingerSwipe - D	Grip - O	Grip - C	Squeeze - C	Squeeze - O	Draw - CF	Draw - CB	Phone - D	Phone - U
HandSwipe - TF																		
HandSwipe - P	0.60																	
HandSwipe - F	0.57	0.62																
Twist - OF	0.39	0.36	0.40															
Twist - TF	0.41	0.39	0.41	0.53														
Twist - P	0.43	0.43	0.43	0.57	0.55													
Twist - F	0.43	0.43	0.44	0.57	0.54	0.61												
Approach - F	0.50	0.48	0.43	0.35	0.36	0.36	0.36											
Recede - F	0.56	0.63	0.56	0.42	0.42	0.44	0.44	0.44										
FingerSwipe - D	0.40	0.34	0.41	0.35	0.36	0.35	0.36	0.37	0.37									
Grip - O	0.45	0.51	0.46	0.42	0.43	0.52	0.51	0.42	0.50	0.34								
Grip - C	0.47	0.51	0.47	0.48	0.47	0.56	0.55	0.42	0.50	0.35	0.68							
Squeeze - C	0.43	0.47	0.42	0.41	0.41	0.45	0.45	0.39	0.46	0.32	0.52	0.50						
Squeeze - O	0.41	0.46	0.41	0.42	0.40	0.47	0.46	0.37	0.44	0.29	0.59	0.55	0.46					
Draw - CF	0.43	0.39	0.43	0.46	0.43	0.45	0.47	0.33	0.42	0.36	0.37	0.41	0.35	0.38				
Draw - CB	0.42	0.40	0.42	0.52	0.46	0.52	0.51	0.37	0.42	0.38	0.42	0.46	0.40	0.41	0.51			
Phone - D	0.57	0.55	0.58	0.46	0.47	0.51	0.52	0.42	0.67	0.46	0.48	0.49	0.43	0.40	0.50	0.48		
Phone - U	0.46	0.47	0.48	0.42	0.44	0.47	0.49	0.43	0.48	0.37	0.44	0.48	0.40	0.40	0.39	0.44	0.51	

Figure 9: Average cross-correlation values of the white area for 27 gestures.

The experiment took about 2 hours, and participants were compensated with about 15 US dollars. The experiment was approved by the ethics committee of the authors' institution. Figure 6 shows the average score of the questionnaire for each gesture. The average of the three indices was 4.89 for simplicity, 3.96 for social acceptability, and 4.63 for fatigue.

5.3 Gesture Selection

5.3.1 Doppler Shift Magnitude. Mid-air gestures with low DSMs are more likely to be difficult to distinguish from daily activities, resulting in a higher false detection rate. Therefore, we first selected mid-air gestures with high DSMs. For comparison, we need to quantify the DSM. In this study, we performed a binary image transformation on each pixel of the spectrogram image to make the area where the Doppler shift occurs white and the rest black based on a threshold value (Figure 7). We used the ratio of the white area to the binary image's total size as a measure of DSM. Figure 8 shows the average ratio for each mid-air gesture. Those with a value less than 1.5 were targeted for deletion, and the following nine gestures were deleted: Approach - P, Recede - OF, Recede - P, HandSwipe - OF, Recede - TF, Approach - TF, FingerSwipe - U, Approach - OF, Draw - X.

5.3.2 Doppler Shift Similarity. Discriminating between gestures with high DSS is likely to be challenging, and removing one from the recognition target is desirable to improve the performance. The comparison method calculates the cross-correlation between different gestures of the binary images generated in Section 5.3.1. The maximum value of the cross-correlation is used as the DSS evaluation index to compare the similarity between each gesture (Figure 7). In this study, we randomly extracted five images for the binary image set of the remaining 18 candidate gestures. We calculated the maximum cross-correlation value between the five images for all gesture pairs and obtained the average. Figure 9 shows the average of the maximum cross-correlation values of the gesture pairs. The average cross-correlation value among all pairs was 0.45. Among the pairs with a value of more than 0.60, one is removed from the pair with the highest value in order. In this study, the gestures with the lower scores on the questionnaire were removed, and the following six gestures were removed: HandSwipe - P, HandSwipe - F, Twist - P, Twist - F, Recede - F, Grip - O.

5.3.3 Usability. To improve usability, gestures with low participant ratings were excluded; the average of the three indices was 4.89 for simplicity, 3.96 for social acceptability, and 4.63 for fatigue. We removed from the target gestures those with simplicity lower than 4, social acceptability lower than 3.5, or fatigue lower than 4, and the following five gestures were removed: Approach - F, Squeeze - O, Draw - CF, Draw - CB, Phone - U.

The final selection included seven types of mid-air gestures: HandSwipe - TF, Twist - OF, Twist - TF, FingerSwipe - D, Grip - C, Squeeze - C, and Phone - D. Figure 10 shows the Doppler shift of the seven selected gestures for each device. Section 6 evaluates these mid-air gestures' detection and classification performance.

6 SYSTEM EVALUATION

We evaluated EarHover's auditory impact on the user, gesture detection performance, and gesture classification performance.

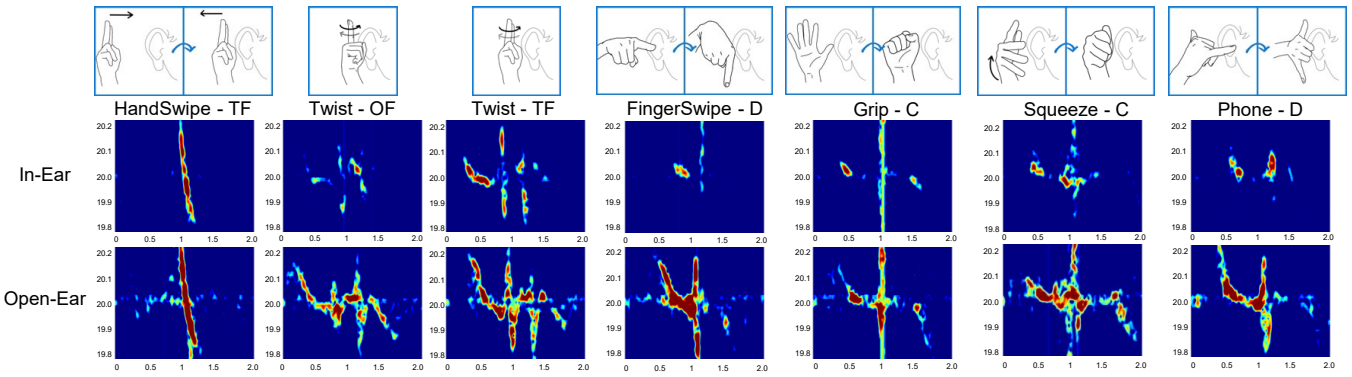


Figure 10: Spectrogram of selected seven gestures for each device.

6.1 Auditory Impact on User

Hearables is primarily used for listening to music and talking, and the sound quality is crucial. We investigated the auditory impact of using EarHover with music playback on the user. We created a "test music" by applying a filter to remove the frequency band (19.8–20.2 kHz) affected by the Doppler shift and then synthesizing it with the measurement signal (20 kHz). In this experiment, 20 participants (male: 13, female: 7) randomly listened to the original and test music without prior knowledge and answered a questionnaire. First, participants were asked whether they thought the music differed on a Likert scale (S: Same, MS: Mostly the same, SD: Somewhat different, and VD: Very different). Participants who answered that the music was not the same were further asked whether they preferred the former or the latter. We tested three types of music of different genres (orchestral style/ rock style/ ambient style).

Figure 11A shows the results of the audio impact for each device and each music genre. The average percentage of participants who perceived a difference was 10 (50.0%). Figure 11B summarizes the original and test music preferences for participants who noticed a difference.

We conducted statistical tests to determine whether there are significant differences in preferences for original and test music across devices and genres. We used Fisher's exact test for the In-Ear's Rock and Open-Ear's Orchestra, where no one responded to one scale, and a chi-squared test with Yates' continuity correction for the others. The statistical test results at the 5% significance level indicated no significant difference in preference trends under any condition. These results suggest that the influence of the EarHover system on music appreciation is minimal. In terms of user feedback, participants who preferred the original music tended to emphasize the clarity and balance of the sound, with some noting the presence or absence of high frequencies. Conversely, those who preferred the test music were inclined to seek the impact of low frequencies in their musical experience, with some commenting on the depth of the sound. Therefore, it can be inferred that although EarHover alters the auditory experience to some extent, this alteration does not significantly affect the listener's overall preference for music, indicating that EarHover can be integrated into the listening experience without detracting from the enjoyment of music.

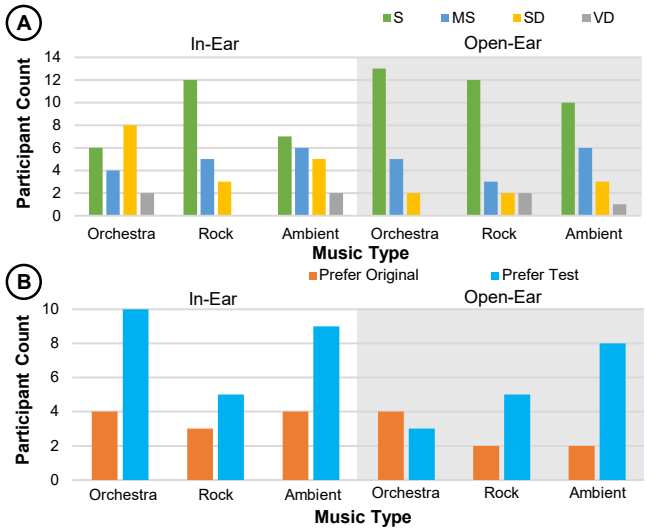


Figure 11: Results of sensory evaluation for each piece of music. A: impact on music differences, B: impact on preferences.

6.2 Gesture Detection Performance

We evaluated the detection performance of seven selected mid-air gestures in practical environments.

6.2.1 Experimental Setup. We collected data on gestures and daily activities from 20 participants (male: 13, female: 7). All participants were right-handed, with a mean age of 27.5 years (standard deviation of 9.93 years). Participants used their hearables a mean of 4.27 days per week. The experiment took about 2 hours, and participants were compensated with about 15 US dollars.

6.2.2 Data Collection. Participants first measured data on mid-air gestures. In this study, 20 rounds were measured in a sitting position (Stable), with a glove on the right hand (Gloves), in an environment with a person sitting next to the device side (Neighbor), while the participant was walking (Walking), and with music synthesized into the measurement signal (Orchestra/Rock/Ambient) were each measured for four rounds. For the Neighbor measurement, the participant and another person sat at a distance of approximately

50 cm with respect to the head position. We used the three test music pieces utilized in Section 5 for the music synthesized measurement. There were 20 participants in the Stable measurement, 12 of whom participated in the Gloves, Neighbor, and Walking measurements, and the remaining 8 in the Music measurement. We performed these measurements with two prototype devices and collected 4480 gesture data per device. The measurement procedure followed the same steps as in Section 5.2. However, for the walking measurement, a gesture illustration was first displayed, the walking state was established as the measurement start state, and the participants were asked to perform mid-air gestures while walking.

We then measured the daily activities of the same 20 participants. We set up 11 different tasks and movements that are frequently performed in daily life: passing each other (passing), standing up and sitting down (up down), spoon-to-mouth, glasses on/off, smartphone tap, typing, reading aloud, head shaking, raising arms, jacket on/off, and picking up an object (picking up). We performed these measurements with two prototype devices and collected 187 minutes of motion data for the in-ear device and 178 minutes for the open-ear device. No participants reported hearing the 20 kHz signal.

6.2.3 Detection Performance. First, we manually annotated and extracted the gesture portion of the data for training model generation and accurate performance evaluation. In the model generation, we used gesture and daily activity data, excluding that of the test user, as training data. Figure 12A and B show the false detection rate for each daily activity and mid-air gesture. The average false detection rate for daily activities was 2.5% for the in-ear device and 1.0% for the open-ear device. The false detection rate was high for glasses on/off and a spoon-to-mouth, while the false detection rate was low for smartphone tap and typing. The movements with the higher false detection rate were hand movements, which may have occurred in reflections and Doppler shifts of the sound leakage signal, similar to the mid-air gestures.

The average false detection rate per gesture was 10.6% for in-ear devices and 10.7% for open-ear devices. The false detection rate for Twist - OF/TF was low. These mid-air gestures are performed by wrist rotation. Therefore, compared to other mid-air gestures, they could be executed without concern for distance from the device and were considered to be more stable. In terms of the false detection rate for each measurement environment, the false detection rate for Neighbor was the lowest for both devices (Figure 12C); since the Neighbor condition involved a person by the hand performing the gesture, it is thought that the hand position was unconsciously closer to the device, resulting in a stable Doppler shift. When we checked the false positive gesture data, many of the data showed a weak occurrence of Doppler shift. The results suggest the importance of stable gesture motion in gesture detection.

6.3 Gesture Classification Performance

We investigated classification performance on Stable data as a baseline. The classification rate for each participant was calculated using 10-fold cross-validation, with two rounds as test data and the remaining 18 rounds as training data. This study used the F-score, the harmonic mean of precision and recall, as an evaluation index for the classification rate. Figure 13 shows the classification rate for

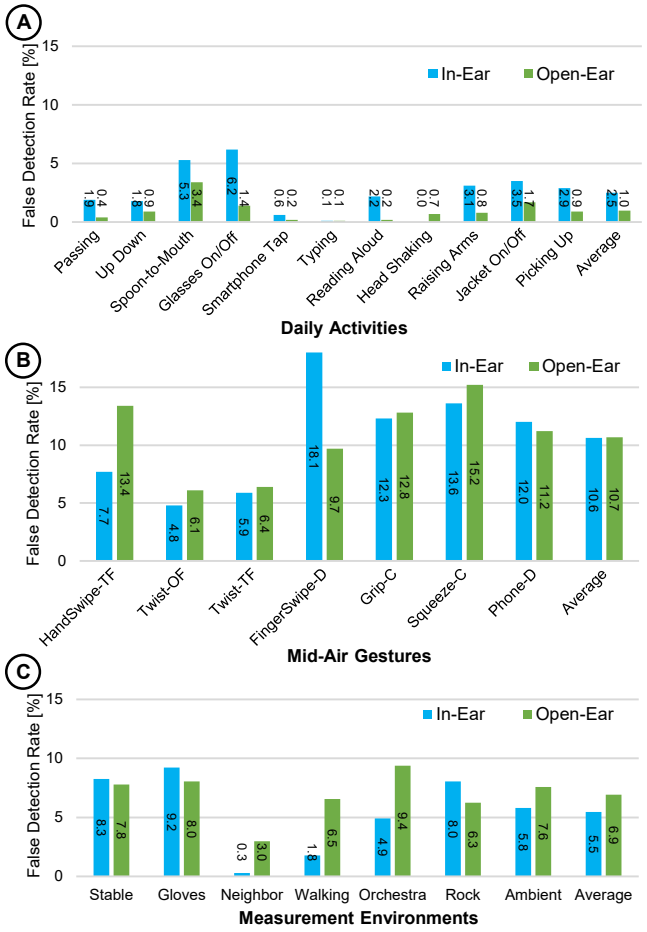


Figure 12: False detection rate, A: daily activities, B: mid-air gestures, C: false detection rate of mid-air gestures aggregated in the measurement environment.

each participant. The overall classification rate was 79.7% for the in-ear device and 85.9% for the open-ear device.

Figure 14 shows the confusion matrix for each device. Both devices showed significant confusion between Twist - OF and Twist - TF, with low classification rates. The similarity of these gestures was 0.53 (Figure 9), which was below the threshold but higher than the average of 0.45, indicating similar Doppler shifts.

6.3.1 Classification Performance in Other Measurement Environments. The classification performance of other measured data was examined. In this experiment, we investigated the classification rate when only stable data was used as training data (non-data-augmentation learning: NDA) and when a portion of the test environment data was incorporated into the training data (test-data-augmentation learning: TDA). In TDA, we added data for three of the four rounds measured to the training data and evaluated the remaining one round. This was done in four different ways so that all rounds were evaluated. Figure 15 summarizes the classification rates for each measurement environment. The average for NDA was 64.3% for the in-ear device and 67.6% for the open-ear device;

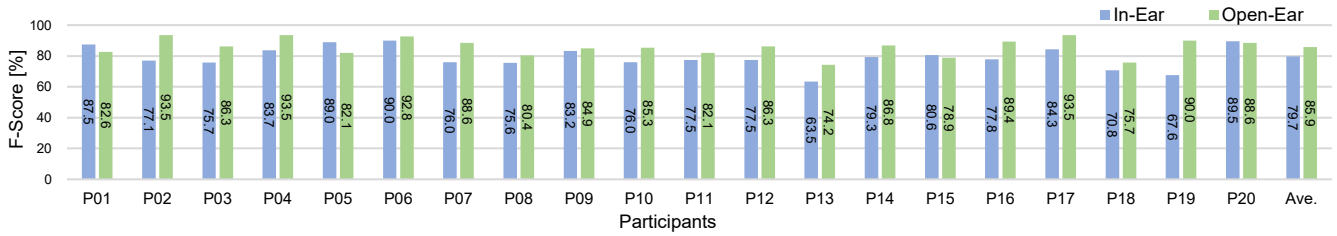


Figure 13: Mid-air gesture classification rate using each test user's own stable data as training data.



Figure 14: Confusion matrix for each device.

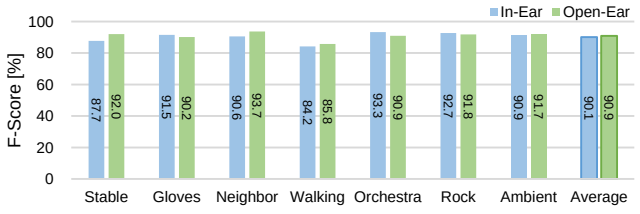


Figure 16: Classification rate of five types of mid-air gestures.

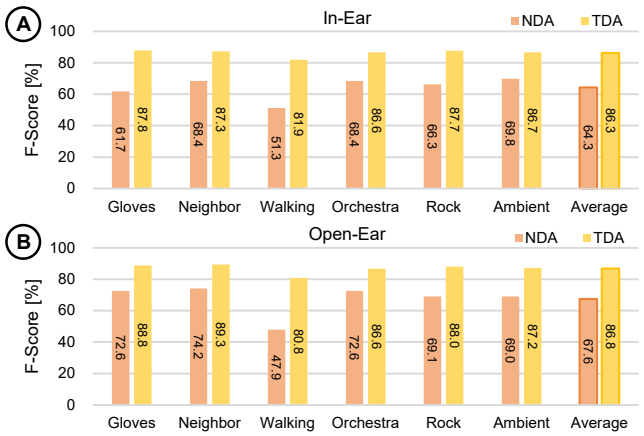


Figure 15: Mid-air gesture classification rate per measurement environment using each test user's own data as training data, A: in-ear device, B: open-ear device.

the average for TDA was 86.3% for the in-ear device and 86.8% for the open-ear device. TDA improved the classification rate by 22.0 and 19.2 points for the in-ear and open-ear devices, respectively.

6.3.2 Five types of gesture classification performance. We investigated how much the classification performance would improve if the number of gestures were reduced to five. With five types of gesture input available, basic operations such as music listening (play/stop, next song, previous song, volume up, volume down) and video calling (mute, video switch, volume up, volume down, exit) become possible on hearables. Due to significant confusion between "Twist - OF" and "Twist - TF," and "Grip - C" and "Squeeze - C" as shown in Figure 14, we excluded "Twist - OF" and "Squeeze - C," which had smaller white areas in Figure 8 and reevaluated the classification rates.

Figure 16 shows the classification rate for each measurement environment. The overall average for the in-ear device was 90.1%, and the overall average for the open-ear device was 90.9%. An improvement of 4.7 points for the in-ear device and 4.2 points for the open-ear device was observed in the classification rate compared to the seven-gesture classification.

7 DISCUSSION

We discuss the performance and limitations of the EarHover system, paving the way for future improvements and potential for sound leakage signal.

7.1 System Performance

7.1.1 Gesture Detection Performance. Gesture detection performance was approximately 2% for false detection of daily activity to gesture and approximately 10% for false detection of gesture to daily activity. False detection of daily activities can cause unintended device operation, reducing usability and increasing the risk of missing important sounds like cars or footsteps due to sudden sound playback. Thus, a low false detection rate is crucial. On the other hand, false detection of a gesture requires the user to repeat the gesture, which reduces usability but poses little safety risk. Future research should investigate the impact of unintended device operations and repeated gesture action requests on users. Furthermore, to improve detection performance, it is necessary to investigate whether increasing the image size can capture the fine Doppler shifts during gestures that were previously undetected. This should be done while considering the balance between computational cost and performance. Also, IMUs and microphones in hearables to monitor user activity have been studied extensively [22, 40, 49]. We

will use these methods to develop a detection model that dynamically changes the detection model and threshold settings depending on the user context.

7.1.2 Gesture Classification Performance. Our gesture selection process resulted in the inclusion of both Twist-OF/TF gestures, leading to frequent confusion between them. This indicates a flaw in our selection process. Moving forward, it is essential to comprehensively evaluate the actual recognition performance and confusion tendencies of candidate gestures to determine the optimal gesture set.

Our classification performance was lower compared to Xiu et al. [54]'s touch gesture input method, achieving an 8-gesture classification rate of 95.3%. This is due to the instability of gesture execution in EarHover, as users could not verify their actions through touch feedback or visual observation. Repeated practice and use are effective in achieving stable gesture operations. We plan to incorporate systems that provide auditory and vibratory feedback during practice sessions to facilitate efficient learning. We also plan to conduct long-term research to understand how recognition performance improves with continued use. In addition, improvements are needed where the recognition system adapts to stable user actions, such as through online learning that incorporates high-quality gesture operation data into the recognition model.

In this study, we evaluated gesture classification using individual models. The results of the Leave-One-Subject-Out Cross-Validation model (LOSOCV model) showed a significantly low classification rate of about 48% for seven mid-air gestures. This is believed to be due to the variability in users' reproduction during gesture execution and the fact that the shape of the hands and ears differs among users [17], resulting in different reflection characteristics of the sound leakage signal for each individual. Even when adding each user's data to the LOSOCV model, the individual models still performed better given the same amount of data. In the future, we aim to enhance the individual model training by effectively utilizing data from other users. This could involve identifying and extracting data points that closely resemble a specific user's data, thereby enriching the training dataset and potentially improving model accuracy and personalization.

7.2 Limitations

This study utilized a prototype device equipped with an additional external microphone to validate the EarHover concept. In the future, we aim to conduct evaluation experiments using the built-in microphones of commercial hearables. This transition will enable a more robust demonstration of EarHover's compatibility and functionality with standard hearables. Platforms proposed by Yetong et al. [14], which allow direct access to these built-in microphones, will be instrumental in further refining EarHover's integration and effectiveness. We also used an external power source, which is not feasible for mobile systems. We will focus on using MEMS microphones and amplification circuits that can be embedded in hearables.

In the auditory impact experiment, most participants were between 21 and 23 years old. To further validate the practicality of EarHover, it is crucial to conduct studies involving teenagers, who generally have a broader auditory range.

The resulting gesture set does not necessarily result in the gesture set with semantic and understandable mappings, instead it is a more random set of gestures that work well for the system and are rated as comfortable. We have to investigate a balance between the gesture set with semantic and the EarHover's performance.

7.3 Future Works

Environmental noise and ambient light noise have affected existing research on touch gesture methods. The sound leakage signal used in our method can also detect hand motion during touch gestures. For future work, we would like to investigate multimodal applications such as a noise-robust touch gesture recognition system using sound leakage signals.

We found that daily activities like raising the hand can reflect sound leakage signals, leading to false detection in mid-air gesture recognition systems. This challenge also hints at the potential for employing sound leakage to recognize user movements. Future work will explore leveraging this signal for detailed hand motion capture and a broader contextual understanding of user posture and movement patterns.

8 CONCLUSION

In conclusion, this study introduces EarHover, the first mid-air gesture input system for hearables using sound leakage, integrated with built-in speakers and external microphones. This pioneering system offers contactless, noise-free interactions, enhancing the usability of hearables by overcoming the limitations of touch-based inputs. We developed 27 distinct mid-air gestures for EarHover, selecting seven for their discriminability and user preference. Our experiments in various real-world scenarios confirmed EarHover's minimal impact on music listening and its effectiveness in gesture recognition, showcasing the practicality and adaptability of the system across different device types. By integrating with standard hearables components, EarHover minimizes cost while maximizing compatibility, marking an advancement in hearables technology.

ACKNOWLEDGMENTS

This work was supported in part by JSPS KAKENHI (Grant Number JP23KJ1884, JP21H03485) and JST PRESTO (Grant Number JPMJPR2134, JPMJPR2138).

REFERENCES

[1] Khaled Alkiek, Moustafa Youssef, and Khaled A. Harras. 2023. EarBender: Enabling Rich IMU-based Natural Hand-to-Ear Interaction in Commodity Earables. In *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing (UbiComp/ISWC '23 Adjunct)*. Association for Computing Machinery, 333–338. https://doi.org/10.1145/3594739.3610671

[2] Takashi Amesaka, Hiroki Watanabe, and Masanori Sugimoto. 2019. Facial expression recognition using ear canal transfer function. In *Proceedings of the 2019 ACM International Symposium on Wearable Computers (ISWC '19)*. Association for Computing Machinery, 1–9. https://doi.org/10.1145/3341163.3347747

[3] Takashi Amesaka, Hiroki Watanabe, Masanori Sugimoto, Yuta Sugiura, and Buntarou Shizuki. 2023. User Authentication Method for Hearables Using Sound Leakage Signals. In *Proceedings of the 2023 ACM International Symposium on Wearable Computers (ISWC '23)*. Association for Computing Machinery, 119–123. https://doi.org/10.1145/3594738.3611376

[4] Toshiyuki Ando, Yuki Kubo, Buntarou Shizuki, and Shin Takahashi. 2017. CanalSense:Face-Related Movement Recognition System Based on Sensing Air Pressure in Ear Canals. In *Proceedings of the 30th Annual ACM Symposium on*

User Interface Software and Technology (UIST '17). Association for Computing Machinery, 679–689. https://doi.org/10.1145/3126594.3126649

[5] Apple. 2012. EarPods. https://www.apple.com/shop/product/MWU53AM/A/earpods-35mm-headphone-plug. (accessed on 2024-07-19).

[6] Apple. 2019. AirPods Pro. https://www.apple.com/airpods-pro/. (accessed on 2024-07-19).

[7] Apple. 2021. AirPods. https://www.apple.com/airpods-3rd-generation/. (accessed on 2024-07-19).

[8] ASUS. 2022. Zenbook 14 UM425QA. https://www.asus.com/laptops/for-home/zenbook/zenbook-14-um425-qa/. (accessed on 2024-07-19).

[9] Audio-Technica. 2002. AT-MA2. https://www.audio-technica.co.jp/product/AT-MA2. (accessed on 2024-07-19).

[10] Audio-Technica. 2008. AT9904. https://www.audio-technica.com.hk/index.php?op=productdetails&pid=461&lang=eng. (accessed on 2024-07-19).

[11] R.J. Baken and R.F. Orlikoff. 2000. *Clinical Measurement of Speech and Voice*. Singular Thomson Learning. https://books.google.co.jp/books?id=ElPyvaJbDiwC

[12] Abdelkareem Bedri, David Byrd, Peter Presti, Himanshu Sahni, Zehua Gue, and Thad Starner. 2015. Stick It in Your Ear: Building an In-ear Jaw Movement Sensor. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers (UbiComp/ISWC '15 Adjunct)*. Association for Computing Machinery, 1333–1338.

[13] Bosch. 2024. Sensing solutions for hearables. https://www.bosch-sensortec.com/applications-solutions/hearables/. (accessed on 2024-07-19).

[14] Yetong Cao, Chao Cai, Anbo Yu, Fan Li, and Jun Luo. 2023. EarAcE: Empowering Versatile Acoustic Sensing via Earable Active Noise Cancellation Platform. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 2, Article 47 (jun 2023), 23 pages. https://doi.org/10.1145/3596242

[15] Mingshi Chen, Panlong Yang, Shumin Cao, Maotian Zhang, and Ping Li. 2018. WritePad: Consecutive Number Writing on Your Hand With Smart Acoustic Sensing. *IEEE Access* 6 (2018), 77240–77249. https://doi.org/10.1109/ACCESS.2018.2880980

[16] Yu-Chun Chen, Chia-Ying Liao, Shuo-wen Hsu, Da-Yuan Huang, and Bing-Yu Chen. 2020. Exploring User Defined Gestures for Ear-Based Interactions. *Proc. ACM Hum.-Comput. Interact.* 4, ISS, Article 186 (nov 2020), 20 pages. https://doi.org/10.1145/3427314

[17] Ziga Emersic, Vitomir Struc, and Peter Peer. 2016. Ear Recognition: More Than a Survey. *CoRR* abs/1611.06203 (2016). arXiv:1611.06203 http://arxiv.org/abs/1611.06203

[18] Xiaoran Fan, Longfei Shangguan, Siddharth Rupavatharam, Yanyong Zhang, Jie Xiong, Yunfei Ma, and Richard Howard. 2021. HeadFi: bringing intelligence to all headphones. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking (MobiCom '21)*. Association for Computing Machinery, 147–159. https://doi.org/10.1145/3447993.3448624

[19] Biying Fu, Dinesh Vaithyalingam Gangatharan, Arjan Kuijper, Florian Kirchbuchner, and Andreas Braun. 2017. Exercise Monitoring On Consumer Smart Phones Using Ultrasonic Sensing. In *Proceedings of the 4th International Workshop on Sensor-Based Activity Recognition and Interaction (iWOAR '17)*. Association for Computing Machinery, Article 9, 6 pages. https://doi.org/10.1145/3134230.3134238

[20] Yang Gao, Yincheng Jin, Jiyang Li, Seokmin Choi, and Zhanpeng Jin. 2020. EchoWhisper: Exploring an Acoustic-Based Silent Speech Interface for Smartphone Users. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 3, Article 80 (sep 2020), 27 pages. https://doi.org/10.1145/3411830

[21] Yang Gao, Wei Wang, Vir V. Phoha, Wei Sun, and Zhanpeng Jin. 2019. EarEcho: Using Ear Canal Echo for Wearable Authentication. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3, Article 81 (sep 2019), 24 pages. https://doi.org/10.1145/3351239

[22] Jingyang Hu, Hongbo Jiang, Daibo Liu, Zhu Xiao, Qibo Zhang, Jiangchuan Liu, and Schahram Dustdar. 2023. Combining IMU With Acoustics for Head Motion Tracking Leveraging Wireless Earphone. *IEEE Transactions on Mobile Computing* (2023), 1–13. https://doi.org/10.1109/TMC.2023.3325826

[23] Yincheng Jin, Yang Gao, Xuhai Xu, Seokmin Choi, Jiyang Li, Feng Liu, Zhengxiong Li, and Zhanpeng Jin. 2022. EarCommand: "Hearing" Your Silent Speech Commands In Ear. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 57 (jul 2022), 28 pages. https://doi.org/10.1145/3534613

[24] Yincheng Jin, Yang Gao, Yanjun Zhu, Wei Wang, Jiyang Li, Seokmin Choi, Zhangyu Li, Jagmohan Chauhan, Anind K. Dey, and Zhanpeng Jin. 2021. SonicASL: An Acoustic-based Sign Language Gesture Recognizer Using Earphones. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 2, Article 67 (jun 2021), 30 pages. https://doi.org/10.1145/3463519

[25] Takashi Kikuchi, Yuta Sugiura, Katsutoshi Masai, Maki Sugimoto, and Bruce H. Thomas. 2017. EarTouch: Turning the Ear into an Input Surface. In *Proc. of the 19th Int'l Conf. on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '17)*. Association for Computing Machinery, Article 27, 6 pages. https://doi.org/10.1145/3098279.3098538

[26] Ke Li, Ruidong Zhang, Bo Liang, François Guimbretière, and Cheng Zhang. 2022. EarIO: A Low-power Acoustic Sensing Earable for Continuously Tracking Detailed Facial Movements. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 62 (jul 2022), 24 pages. https://doi.org/10.1145/3534621

[27] Roman Lissermann, Jochen Huber, Aristotelis Hadjakos, Suranga Nanayakkara, and Max Mühlhäuser. 2014. EarPut: Augmenting Ear-Worn Devices for Ear-Based Interaction. In *Proc. of the 26th Australian CHI Conf. on Designing Futures: The Future of Design (OzCHI '14)*. Association for Computing Machinery, 300–307. https://doi.org/10.1145/2686612.2686655

[28] Hiroyuki Manabe and Masaaki Fukumoto. 2014. Headphone Taps: Tap Control for Regular Headphones. *Journal of Information Processing* 55, 4 (2014), 1334–1343. https://api.semanticscholar.org/CorpusID:236784614

[29] Hiroyuki Manabe, Masaaki Fukumoto, and Tohru Yagi. 2013. Conductive Rubber Electrodes for Earphone-Based Eye Gesture Input Interface. In *Proc. of the 2013 Int'l Symposium on Wearable Computers (ISWC '13)*. Association for Computing Machinery, 33–40. https://doi.org/10.1145/2493988.2494329

[30] Denys J. C. Matthies, Bernhard A. Strecker, and Bodo Urban. 2017. EarFieldSensing: A Novel In-Ear Electric Field Sensing to Enrich Wearable Gesture Input through Facial Expressions. In *Proc. of the 2017 CHI Conf. on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, 1911–1922. https://doi.org/10.1145/3025453.3025692

[31] C. Metzger, M. Anderson, and T. Starner. 2004. FreeDigiter: A Contact-Free Device for Gesture Control. In *Eighth International Symposium on Wearable Computers*, Vol. 1. 18–21. https://doi.org/10.1109/ISWC.2004.23

[32] Hanae Rateau, Edward Lank, and Zhe Liu. 2022. Leveraging Smartwatch and Earbuds Gesture Capture to Support Wearable Interaction. *Proceedings of the ACM on Human-Computer Interaction* 6, ISS, Article 557 (2022), 20 pages. https://doi.org/10.1145/3567710

[33] Tobias Röddiger, Christopher Clarke, Paula Breitling, Tim Schneegans, Haibin Zhao, Hans Gellersen, and Michael Beigl. 2022. Sensing with Earables: A Systematic Literature Review and Taxonomy of Phenomena. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3, Article 135 (sep 2022), 57 pages. https://doi.org/10.1145/3550314

[34] Tobias Röddiger, Christopher Clarke, Daniel Wolffram, Matthias Budde, and Michael Beigl. 2021. EarRumble: Discreet Hands- and Eyes-Free Input by Voluntary Tensor Tympani Muscle Contraction. In *Proc. of the 2021 CHI Conf. on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, Article 743, 14 pages. https://doi.org/10.1145/3411764.3445205

[35] Wenjie Ruan, Quan Z. Sheng, Lei Yang, Tao Gu, Peipei Xu, and Longfei Shangguan. 2016. AudioGest: Enabling Fine-Grained Hand Gesture Detection by Decoding Echo Signal. In *Proc. of the 2016 ACM Int'l Joint Conf. on Pervasive and Ubiquitous Computing (UbiComp '16)*. Association for Computing Machinery, 474–485. https://doi.org/10.1145/2971648.2971736

[36] Julia Schwarz, Chris Harri son, Scott Hudson, and Jennifer Mankoff. 2010. Cord Input: An Intuitive, High-Accuracy, Multi-Degree-of-Freedom Input Method for Mobile Devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. Association for Computing Machinery, 1657–1660. https://doi.org/10.1145/1753326.1753573

[37] SHOKZ. 2023. OpenFit. https://shokz.com/products/openfit. (accessed on 2024-07-19).

[38] Sony. 2021. wf-1000xm4. https://www.sony.jp/headphone/special/WF-1000XM4/. (accessed on 2024-07-19).

[39] steinberg. 2019. UR22C. https://www.steinberg.net/audio-interfaces/ur22c/. (accessed on 2024-07-19).

[40] David Strömbäck, Sangxia Huang, and Valentin Radu. 2020. MM-Fit: Multimodal Deep Learning for Automatic Exercise Logging across Sensing Devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 4, Article 168 (dec 2020), 22 pages. https://doi.org/10.1145/3432701

[41] Wei Sun, Franklin Mingzhe Li, Benjamin Steeper, Songlin Xu, Feng Tian, and Cheng Zhang. 2021. TeethTap: Recognizing Discrete Teeth Gestures Using Motion and Acoustic Sensing on an Earpiece. In *26th Int'l Conf. on Intelligent User Interfaces (IUI '21)*. Association for Computing Machinery, 161—-169. https://doi.org/10.1145/3397481.3450645

[42] Xue Sun, Jie Xiong, Chao Feng, Haoyu Li, Yuli Wu, Dingyi Fang, and Xiaojiang Chen. 2024. EarSSR: Silent Speech Recognition via Earphones. *IEEE Transactions on Mobile Computing* (2024), 1–17. https://doi.org/10.1109/TMC.2024.3356719

[43] Emi Tamaki, Takashi Miyaki, and Jun Rekimoto. 2009. Brainy Hand: An Ear-Worn Hand Gesture Interaction Device. In *CHI '09 Extended Abstracts on Human Factors in Computing Systems (CHI EA '09)*. Association for Computing Machinery, 4255–4260. https://doi.org/10.1145/1520340.1520649

[44] Kazuhiro Taniguchi, Hisashi Kondo, Mami Kurosawa, and Atsushi Nishikawa. 2018. Earable TEMPO: A Novel, Hands-Free Input Device that Uses the Movement of the Tongue Measured with a Wearable Ear Sensor. *Sensors* 18, 3 (2018). https://doi.org/10.3390/s18030733

[45] Dhruv Verma, Sejal Bhalla, Dhruv Sahnan, Jainendra Shukla, and Aman Parnami. 2021. ExpressEar: Sensing Fine-Grained Facial Expressions with Earables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3, Article 129 (2021), 28 pages. https://doi.org/10.1145/3478085

[46] Victor. 2022. HA-NP35T. https://www.victor.jp/headphones/lineup/ha-np35t/. (accessed on 2024-07-19).

[47] Santiago Villarreal-Narvaez, Jean Vanderdonckt, Radu-Daniel Vatavu, and Jacob O. Wobbrock. 2020. A Systematic Review of Gesture Elicitation Studies: What Can We Learn from 216 Studies?. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference (DIS '20)*. Association for Computing Machinery, 855–872. https://doi.org/10.1145/3357236.3395511

[48] Wei Wang, Alex X. Liu, and Ke Sun. 2016. Device-Free Gesture Tracking Using Acoustic Signals. In *Proc. of the 22nd Annual Int'l Conf. on Mobile Computing and Networking (MobiCom '16)*. Association for Computing Machinery, 82–94. https://doi.org/10.1145/2973750.2973764

[49] Yang Wang, Feng Hong, Yufei Jiang, Chenyu Bao, Chao Liu, and Zhongwen Guo. 2024. ToothFairy: Real-time Tooth-by-tooth Brushing Monitor Using Earphone Reversed Signals. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 4, Article 185 (jan 2024), 19 pages. https://doi.org/10.1145/3631412

[50] Hiroki Watanabe, Hiroaki Kakizawa, and Masanori Sugimoto. 2021. User Authentication Method Using Active Acoustic Sensing. *Journal of Information Processing* 29 (2021), 370–379. https://doi.org/10.2197/ipsjjip.29.370

[51] Hiroki Watanabe and Tsutomu Terada. 2018. Improving Ultrasound-Based Gesture Recognition Using a Partially Shielded Single Microphone. In *Proc. of the 2018 ACM Int'l Symposium on Wearable Computers (ISWC '18)*. Association for Computing Machinery, 9–16. https://doi.org/10.1145/3267242.3267274

[52] WHO. 2019. Safe listening devices and systems: a WHO-ITU standard. https://www.who.int/publications/i/item/9789241515276/. (accessed on 2024-07-19).

[53] Yadong Xie, Fan Li, Yue Wu, and Yu Wang. 2021. HearFit: Fitness Monitoring on Smart Speakers via Active Acoustic Sensing. In *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*. 1–10. https://doi.org/10.1109/INFOCOM42981.2021.9488811

[54] Xuhai Xu, Haitian Shi, Xin Yi, WenJia Liu, Yukang Yan, Yuanchun Shi, Alex Mariakakis, Jennifer Mankoff, and Anind K. Dey. 2020. EarBuddy: Enabling On-Face Interaction via Wireless Earbuds. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, 1–14. https://doi.org/10.1145/3313831.3376836

[55] Yongjie Yang, Tao Chen, Yujing Huang, Xiuzhen Guo, and Longfei Shangguan. 2024. MAF: Exploring Mobile Acoustic Field for Hand-to-Face Gesture Interactions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, Article 638, 20 pages. https://doi.org/10.1145/3613904.3642437

[56] Xiyuxing Zhang, Yuntao Wang, Jingru Zhang, Yaqing Yang, Shwetak Patel, and Yuanchun Shi. 2023. EarCough: Enabling Continuous Subject Cough Event Detection on Hearables. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23)*. Association for Computing Machinery, Article 94. https://doi.org/10.1145/3544549.3585903

[57] Yongpan Zou, Qiang Yang, Yetong Han, Dan Wang, Jiannong Cao, and Kaishun Wu. 2019. AcouDigits: Enabling Users to Input Digits in the Air. In *2019 IEEE Int'l Conf. on Pervasive Computing and Communications (PerCom)*. 1–9. https://doi.org/10.1109/PERCOM.2019.8767415

[58] Yongpan Zou, Qiang Yang, Rukhsana Ruby, Yetong Han, Sicheng Wu, Mo Li, and Kaishun Wu. 2019. EchoWrite: An Acoustic-based Finger Input System Without Training. In *2019 IEEE 39th Int'l Conf. on Distributed Computing Systems (ICDCS)*. 778–787. https://doi.org/10.1109/ICDCS.2019.00082