Exploring User-Defined Gestures as Input for Hearables and Recognizing Ear-Level Gestures with IMUs

YUKINA SATO[*] and TAKASHI AMESAKA[*], Keio University, Japan
TAKUMI YAMAMOTO, Keio University, Japan
HIROKI WATANABE, Future University Hakodate, Japan
YUTA SUGIURA, Keio University, Japan

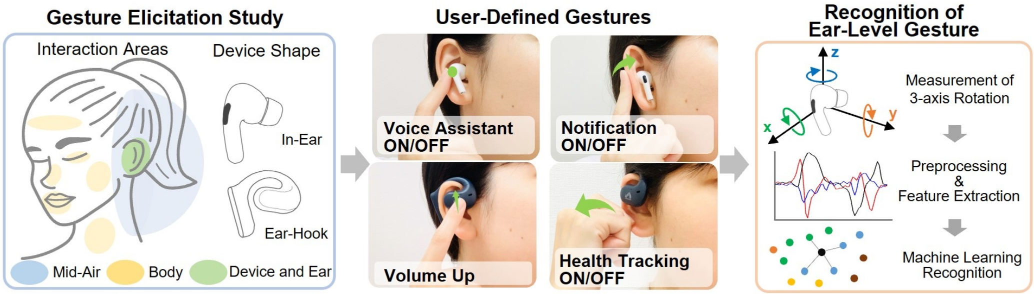

Fig. 1. This study explored hand gestures for hearables through a gesture elicitation study (GES) under six conditions (three interaction areas × two device shape types) and conducted gesture recognition experiments with ear-level gestures that IMU can recognize.

Hearables are highly functional earphone-type wearables; however, existing input methods using stand-alone hearables are limited in the number of commands, and there is a need to extend device operation through hand gestures. In previous research on hearables for hand input, user understanding and gesture recognition systems have been developed. However, in the realm of user understanding, investigation concerning hand input with hearables remains incomplete, and existing recognition systems have not demonstrated proficiency in discerning user-defined gestures. In this study, we conducted a gesture elicitation study (GES) assuming hand input using hearables under six conditions (three interaction areas × two device shapes). Then, we extracted ear-level gestures that the device's built-in IMU sensor could recognize from the user-defined gestures and investigated the recognition performance. The results of sitting experiments showed that the gesture recognition rate for in-ear devices was 91.0% and that for ear-hook devices was 74.7%.

CCS Concepts: • **Human-centered computing** → **Ubiquitous and mobile computing systems and tools**; **Gestural input**.

Additional Key Words and Phrases: Hearables; Hands Gesture Recognition; User-Defined Gesture; Gesture Elicitation Study; IMU

[*] Both authors contributed equally to this research.

Authors' addresses: Yukina Sato, sato-yukina@keio.jp; Takashi Amesaka, amesaka@keio.jp, Keio University, Yokohama, Japan; Takumi Yamamoto, Keio University, Yokohama, Japan; Hiroki Watanabe, Future University Hakodate, Hakodate, Japan; Yuta Sugiura, Keio University, Yokohama, Japan.

ACM 2573-0142/2024/9-ART258
https://doi.org/10.1145/3676503

ACM Reference Format:
Yukina Sato, Takashi Amesaka, Takumi Yamamoto, Hiroki Watanabe, and Yuta Sugiura. 2024. Exploring User-Defined Gestures as Input for Hearables and Recognizing Ear-Level Gestures with IMUs. *Proc. ACM Hum.-Comput. Interact.* 8, MHCI, Article 258 (September 2024), 23 pages. https://doi.org/10.1145/3676503

1 INTRODUCTION

Hearables are highly functional earphone-type devices with diverse applications, including lifelogging, voice assistance, and audio services, alongside traditional functions like music playback and communication [25, 49]. Development efforts are underway, and more applications are expected to be incorporated. Many commercially available hearables are designed for synergistic use with smartphones. However, the current operational paradigm necessitates physical interaction with the smartphone's screen, reducing usability. Consequently, a compelling demand exists for a device operation method that uses only hearables.

Manipulation techniques for hearables can be categorized into hands-free and hand input methods. Various hands-free input methods, including voice input [47], silent speech input [12, 34], and head gesture input [4, 5, 20, 36, 41] methods are useful when hands are otherwise engaged, such as when managing luggage. Conversely, voice input presents challenges, notably difficulties associated with speaking in public places, and reduced recognition accuracy in noisy environments. Additionally, conversational command input, encompassing silent speech input, necessitates the user to articulate a wake word, a process that becomes cumbersome for frequently used commands [1]. Head gestures, particularly those involving substantial head movements such as shaking, include the potential disruption of eye contact. Furthermore, gestures reliant on small head movements, such as jaw shifting, may risk confusion with everyday movements.

The hand input method necessitates the use of hands; however, it is user-friendly, and the facilitation of device operation is accompanied by tactile feedback [15]. Moreover, given that gestures are executed directly toward the device or the user's own body, this approach minimizes confusion with routine activities and has the additional benefit of swift input execution. Research on hand input for hearables encompasses two principal domains: user understanding [8, 16, 24, 30, 45], and the development of hand input methods [2, 14, 16, 18, 21, 35, 45]. In the realm of user understanding, gesture elicitation studies (GES) [42] have been undertaken, wherein multiple users have devised gestures to ascertain an optimal set of gestures. On the other front, efforts have been directed toward advancing the development of a hand gesture recognition system, leveraging the device's integrated sensors.

In their investigation of GES, Chen et al.[8] revealed user-defined gestures centered around the ear periphery as the designated interaction region. Similarly, Rateau et al. [24] revealed user-defined gestures in the context of wearing both a smartwatch and earphones. However, in their study, there were no constraints on the types of gestures that could be defined, and no specific survey was conducted exclusively for hearables. Consequently, their study exhibits the following two characteristics:

No interaction area restrictions

Previous studies have defined aerial gestures, which do not involve physical contact with the device or body. The concept of aerial gesture recognition has been explored [21, 35]. However, implementing these methods necessitates installing a camera or an outward-facing infrared sensor, rendering it a challenging prospect for swift integration into commercial hearables. In contrast, touch-based gesture recognition has been proposed, including methods using microphones or 9-axis sensors [2, 45]. Furthermore, some products available on the market employ gesture input by measuring vibrations generated through tapping around the tragus with an acceleration sensor [33]. These studies and products highlight the current limitations, primarily involving surface tapping and swiping as the predominant gestures. As technology advances, the scope of available gestures and interaction areas is anticipated to expand. Consequently, there is a need to systematically identify and classify gesture sets that align with various input technologies, particularly by exploring user-defined gestures under defined interaction area constraints.

No GES has been exclusively conducted for the singular use of hearables. Previous studies have often considered users either without any device [8] or simultaneously wearing a smartwatch [24]. By contrast, our study focuses on a scenario where the hearables operate independently without utilizing commands from another device. Consequently, it becomes important to investigate user-defined gestures that are suitable for scenarios where the user is only wearing the hearables. Moreover, the contemporary market has an array of hearables characterized by diverse shapes and sizes, each featuring distinct touch areas and points. The implications of these variations on user-defined gestures remain unclear.

Studies exploring hand input for hearables have proposed various methods utilizing infrared sensors [14], IMU sensors [2], and microphones [45]. Many commercial products now include inertial measurement units (IMUs) for spatial audio reproduction, which have also been used for motion tracking and user activity recognition [6, 22, 26, 41]. Consequently, harnessing IMUs for hand gesture recognition is efficient, especially for hearables with limited space. However, existing research [2] has not evaluated systems based on user-defined gestures and has conducted limited and preliminary experiments.

Our study focuses on scenarios where users exclusively wear hearables. Fig. 1 provides a concise summary of our study. We conducted a GES under six conditions with different interaction area restrictions (no-restriction, touch, and ear-touch) and device shapes (in-ear and ear-hook). We then introduced a hand gesture recognition method using the IMU sensor embedded in the hearables. The evaluation of its recognition performance was based on ear-level gestures derived from user-defined gestures obtained in our GES. The contributions of our research are outlined as follows:

- We conducted a GES involving 19 participants, assuming hand input using hearables. This study unveiled the impact of interaction area restrictions and differences in device shapes on user-defined gestures.
- We proposed a hand gesture recognition system utilizing an IMU sensor integrated into hearables. In an evaluative experiment, we selected ear-level gestures from the user-defined gesture sets to assess the recognition performance of our system. The results of the sitting condition experiment demonstrated a gesture recognition rate of 91.0% for in-ear devices (nine types of gestures) and 74.7% for ear-hook devices (six types of gestures) among 10 participants. Additionally, in the walking condition experiment, there was a gesture recognition rate of 79.6% for in-ear devices and 58.0% for ear-hook devices among five participants.

This study offers a comprehensive understanding of hand input for hearables, predicts diversification in input areas and device shapes, and contributes valuable insights to the design of future hearables. Furthermore, we demonstrate a highly compatible gesture recognition method tailored for hearables.

In this paper, Section 2 describes related works, Section 3 describes the GES for hearable input, Section 4 describes gesture recognition experiments using an IMU, Section 5 discusses the overall study, and Section 6 summarizes the study.

2 RELATED WORK

2.1 Elicitation Study for Defining Gesture Sets

In the context of gesture recognition experiments, the determination of gestures involves two distinct approaches: one utilizing gesture sets formulated by researchers and the other employing sets derived from user-defined gestures generated through a GES. Wobbrock et al. [44] defined gesture sets for surface computing. Their methodology involved conducting a GES in which researchers assigned tasks to participants, who then devised corresponding gestures. In a related study, Morris et al. [23] found that the gesture set defined through GES proved to be more intuitive than the one defined solely by researchers. GES, a versatile tool, has been employed across various body parts, encompassing the face [19], foot [10], hand [7, 29], head [48], and even the skin [43]. Additionally, GES has been implemented using a diverse array of devices, including televisions [37], mobile devices [27], head-mounted displays [28], smartwatches [11, 17], smart rings [11], hats [9], and masks [46].

Regarding GES related to ear interactions, Chen et al. [8] conducted a comprehensive GES centered around the ear, offering an in-depth analysis and discussion on user-defined gestures in the context of ear-based interactions. However, in their study, participants were not constrained in defining gestures, and they were not wearing earphone-type devices. While advantageous for investigating pure ear interaction, this approach does not align with the development of gesture recognition technology specifically tailored for hearables, where available interaction areas are expected to gradually expand. Moreover, when envisioning input for hearables, it is logical to consider input via the device itself. Rateau et al. [24] conducted a GES involving users wearing both earphones and a smartwatch. However, their study, similar to Chen et al.'s, did not exclusively focus on hearables, nor did it impose restrictions on the interaction area. In addressing this gap, our study emphasizes the necessity of operating multifunctional hearables with the devices themselves, conducting a GES specifically on users wearing only hearables. Furthermore, our study offers valuable insights into the incremental development of hearables under conditions where the interaction area is limited or the device shape varies.

2.2 Hand Input Method for Hearables

Manabe et al. [18] demonstrated that commercially available headphones could be augmented with a simple circuit to recognize taps. Roman et al. [16] proposed a touch input system utilizing an ear-hook device equipped with a capacitive sensor, while Kikuchi et al. [14] proposed a system to recognize ear deformation gestures using a reflective sensor attached to the rear of the earphone. Additionally, Xiu et al. [45] proposed a system capturing the sound generated by swiping gestures on the cheek or ear using the device's built-in microphone. In the realm of commercial products, SONY's LinkBuds [33] can recognize vibrations during tapping around the tragus, facilitated by an acceleration sensor integrated for device input.

In aerial gestures, Metzger et al. [21] proposed an aerial gesture recognition system with an equipped outward-facing infrared sensor. Tamaki et al. [35] advanced an in-air gesture input system on an earphone-type device equipped with a camera. The methods in these studies requiring additional sensors have the limitation that they are not available in commercially available hearables. Gesture recognition using microphones has found implementation in commercial devices [45]. Nevertheless, there are inherent limitations associated with this approach, such as certain gestures becoming impractical when wearing a mask. Additionally, challenges arise in terms of diminished gesture extraction and recognition rates in noisy environments.

In our study, we present a gesture recognition method utilizing an IMU, a component already integrated into numerous commercial hearables. This choice eliminates significant limitations concerning implementation costs. Moreover, the IMU offers distinct advantages, notably the absence of recognition rate degradation in noisy environments. Khaled et al. [2] explored hand gesture recognition using IMU sensors. Nevertheless, their study did not investigate user-defined gestures, and the experimental scale (N = 4 for in-ear type devices) remains preliminary. In contrast, our study endeavors to assess the recognition rate of a gesture set derived from a gesture elicitation study (GES) across two device types (in-ear and ear-hook) with a more extensive participant group (N = 10). This expansion enables a more comprehensive evaluation of IMU-based hand gesture recognition systems for hearables. Importantly, it contributes to the exploration of user-centered gestures and widens the spectrum of devices under consideration.

3 GESTURE ELICITATION STUDY FOR INPUT TO HEARABLES

We conducted the GES to address the following three research questions:

RQ1 What gestures do users prefer as input to hearables?
RQ2 How do interaction area restrictions and device shape differences affect user-defined gestures?
RQ3 What are the similarities and differences with previous studies on hearables and ears?

In this experiment, we endeavored to address the aforementioned research questions by soliciting users to define gestures for the same set of tasks across a total of six conditions (three interaction areas × two device shapes). The interaction area conditions include no-restriction, touch, and ear-touch, with reference to established gesture input methods for hearables. The no-restriction condition permits users to define gestures, such as hovering and touching around the ear, assuming the gestures are recognizable by existing methods utilizing a camera, microphone, IMU, or infrared sensor [2, 14, 21, 35, 45]. In the touch condition, the specified gesture involves a fingertip touch to either the ear or the surrounding area, assuming the gestures are recognizable by existing methods utilizing a microphone, IMU, or infrared sensor [2, 14, 45]. The ear-touch condition requires that the gesture induce device movement, assuming the gestures are recognizable by existing methods using an IMU and infrared sensor [2, 14]. Additionally, touch gestures directed to the device can be defined for all conditions. Taking into account the prevalence of commercial products and available models, we opted for the two most common device shapes: in-ear and ear-hook types, as depicted in Fig. 1.

3.1 Experiment Summary

3.1.1 Participants and Experimental Environment. Nineteen experimental participants (male: 11, female: 8) were recruited and surveyed for user-defined gestures. Their ages ranged from 21 to 54 years (average 25.8 years). All participants were right-handed, and sixteen of them used earphones at least once a week. Nine of them had experience with operations on the device itself, such as tapping on the device, and two of them usually used operations using the device itself. The experiment took one to two hours for each participant, and we paid approximately 20 US dollars as a reward. We conducted the experiments in an open laboratory environment and used AirPods Pro (Apple) for in-ear type devices and HA-NP35TBK (Victor) for ear-hook devices. This experiment was approved by the ethical board at the author's institution and informed consent was obtained from the participants.

3.1.2 Tasks. The tasks are enumerated in Table 1 and organized into four distinct groups: navigation, music player, phone, and application. The three groups (navigation, music player, and phone) align with those primarily surveyed in related study [8]. Additionally, we introduced an applications group to encompass tasks related to activating or deactivating specific functions, such as voice assistants or voice memos, which are anticipated for use with hearables [24]. The total number of tasks across all groups was 32.

3.1.3 Procedure. To enhance participants' comprehension of the functionality associated with each task, we provided them with an opportunity to observe the control screens while executing

Table 1. List of task groups.

Task Group	Task
Navigation	Scroll Right, Scroll Left, Scroll Up, Scroll Down, Zoom In, Zoom Out, Maximize / Minimize, Go to Home Screen, Next App, Previous App, Forward, Back
Music Player	Play / Stop, Volume Up, Volume Down, Next Song, Previous Song
Phone	Answer / Hang up, Ignore Call, Make a Call, Microphone on / off, Speaker on / off
Application	Voice Assistant, Voice Memo on / off, Calendar on / off, Health Tracking on / off, Notifications on / off

Table 2. Taxonomy of gestures.

Gesture Mapping

Nature	Metaphoric	Gesture is a metaphor of another object.
	Physical	Gesture acts physically on object.
	Symbolic	Gesture visually depicts a symbol.
	Abstract	Gesture Mapping is arbitrary.
Context	In-context	Gesture requires specific context.
	No-context	Gesture does not require specific context.
Flow	Continuous	Action occurs during the gesture.
	Discrete	Action occurs after the gesture completion.

Physical Characteristics

Locale	Device-level	Gesture involves contact with the device.
	Ear-level	Gesture involves contact with the ear.
	Body-level	Gesture involves contact with the upper body except ears.
	Mid-air-level	Gesture occurs in the air with no physical contact.
Complexity	Simple	Gesture consists of a single gesture.
	Compound	Gestures can be decomposed into simple gestures.
Form	Static Pose	Hand pose is held in only one locale.
	Static Pose and Path	Hand pose is held as hand moves.
	Dynamic Pose	Hand pose changes in one location.
	Dynamic Pose and Path	Hand pose changes as hand moves.
	Deformation	Hand pose makes the ear deformation.

navigation, music, and phone tasks. Additionally, we explained the specific control outcomes corresponding to each task. In the case of the application group, we verbally explained the application content. Following this, we outlined the rules for defining gestures, which were as follows:

- The same gesture cannot be assigned to the same task group.
- Gestures separated by“/” in Table 1 (e.g., maximize/minimize) can be assigned to the same gesture because they are state-switching tasks.
- Gestures can be changed at any time during the experiment.
- The same gesture can be assigned to different tasks with the right and left hands.
- Gestures performed with both hands can be defined.

Subsequent to clarifying the rules, we explained the constraints associated with the gesture definition location. Concerning aerial and body gestures, it was emphasized that if the gesture definition position is excessively distant from the hearables, the envisioned system in this study may face challenges in recognizing the gesture. Consequently, in the case of defining aerial gestures, participants were instructed that the hearables sense their movements, and they were guided to perform gestures directed toward the device. However, specific distance instructions were intentionally omitted. Regarding the definition of body gestures, excluding the ear, the designated gesture definition location was confined to the head, neck, and chest areas. After the explanations, each participant devised gestures for each of the tasks indicated in Table 1. During the experiment, video recording was made, and the hand combination used by the experimenter (right hand, left hand, or both hands), the location, type (e.g., tap or swipe), direction, and number of times (e.g., one tap) the gesture was performed were recorded in text format. For example, "Tap once on the driver part of the device with the right hand"or "Swipe once up on the cheek with the left hand" were recorded. The location and type of gesture were updated each time the subject devised a new gesture.

3.2 Taxonomy of Gestures

With 19 participants, 32 tasks, 3 conditions, and 2 devices, a total of 19 × 32 × 3 × 2 = 3648 gestures were made. This section summarizes the taxonomy of gestures. The entire taxonomy adopted from previous studies [8, 27, 38, 44] is shown in Table 2. In this experiment, the codebook model [40] was used to classify the elicited gestures. Gesture mapping describes the process of mapping gestures to various tasks, including nature, context, and flow. Conversely, physical features capture the characteristics of the gesture itself, such as locale, complexity, and form.

The *nature* dimension reflects the different levels of meaning contained in the gesture [8]. A figurative gesture acts on, with, or like something else. In other words, it is a metaphor for another physical object, such as tapping an imaginary button. A physical gesture acts on the device itself. A symbolic gesture visually depicts a symbol. For example, pointing right in the air to "forward" the screen controls or drawing a heart on the cheek to turn on/off the health tracking function. Finally, abstract gesture mapping is arbitrary, for example, tapping the device to stop the music or tapping the right earlobe in the Next App.

The *context* dimension indicates whether the gesture should be performed independently or within a specific context. For example, swiping the helix up when turning up the volume on music playback is an in-context gesture, whereas tapping the earpiece twice to answer the phone is a no-context gesture.

The *flow* dimension indicates whether the gesture action on the object occurs simultaneously with or after the gesture is performed. A gesture is considered a discrete gesture if the action occurs after the gesture is performed, such as tapping an ear to select an object. A gesture is considered continuous if a task and the gesture are performed simultaneously, such as scrolling the screen while swiping in the air.

The *locale* dimension, classified with reference to the previous study [8, 38], represents the gesture's location in relation to the ear. In this study, an additional category "Device-level" was added because a device is worn on the ear. Considering the interaction area, we classified the devices into four levels: device, ear, body, and mid-air. The touch condition is restricted in that mid-air-level cannot be defined as an interaction area, and the ear-touch condition is restricted in that mid-air-level and body-level cannot be defined as an interaction area.

The *complexity* dimension indicates whether the gesture is simple or complex. For example, an earlobe pinching gesture is simpler than an earlobe pulling (pinch + pull) gesture.

The *form* dimension indicates the movement of the hand when the gesture is made. A static pose is a hand pose that stays in one place, such as covering the ear with the hand. A static pose and path are hand postures that remain the same (fingers do not move) even if the hand position changes, such as swiping through the air with the index finger. The dynamic pose changes only

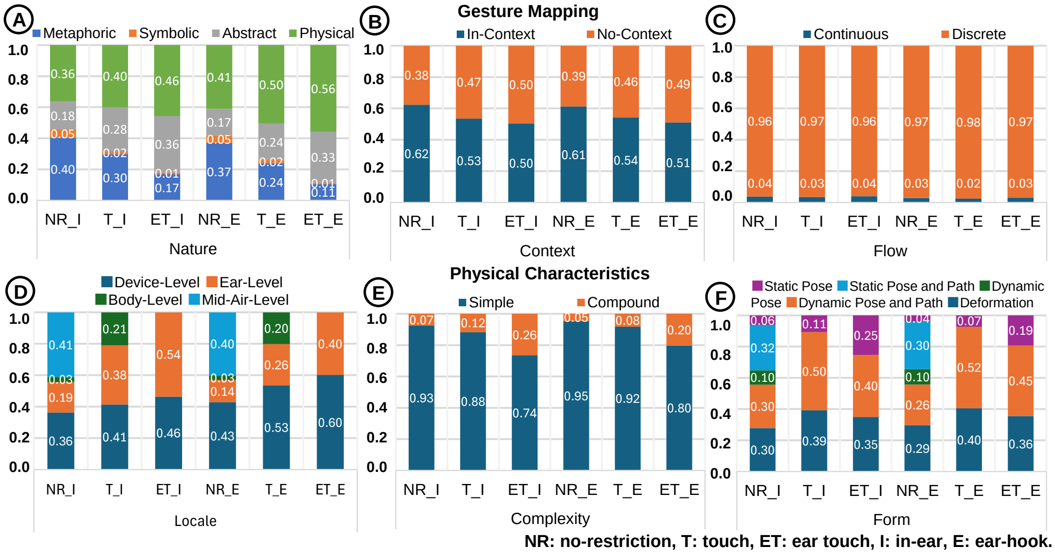

Fig. 2. Ratio of gestures within each category in the six-dimensional taxonomy; the y-axis represents the ratio. A: nature, B: context, C: flow, D: locale, E: complexity, F: form.

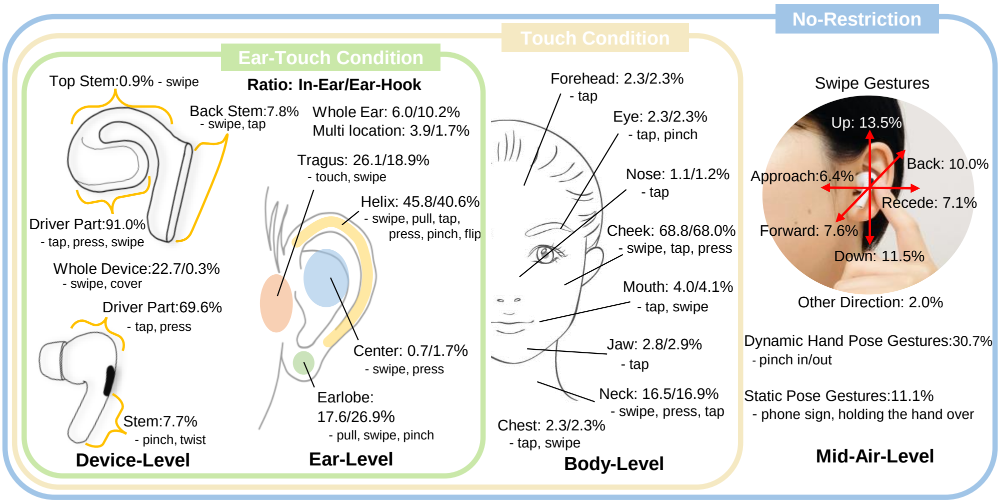

Fig. 3. Detailed definition ratio for each local dimension.

the hand pose, such as opening the palm and moving it away from the ear, but the hand position remains the same. Dynamic poses and paths change both the pose and position of the hand, such as opening the hand while moving it away from the ear. Gestures that transform the ear, such as folding the ear or pulling the earlobe, are classified as transformational gestures.

Fig. 2 shows the distribution of each dimension and illustrates the breakdown of our classifications.

Device-Level Gestures. Device-level gestures were categorized by location, as shown in Fig. 3. In the driver part of the in-ear device, tap and press gestures were specified, constituting a combined ratio of 69.6%. Within the stem part, pinch and twist gestures were defined, with a total ratio of

7.7%. Additionally, gestures involving the entire device, such as swiping down from the driver to the stem, were defined, accounting for a total ratio of 22.7%.

For the driver part of the ear-hook device, tap, press, and swipe gestures were defined, encompassing a total ratio of 91.0%. In the top stem part, only swipe gestures were defined, with a total ratio of 0.9%. In the back stem part, swipe and tap gestures were defined, resulting in a total ratio of 7.8%. Additionally, a gesture involving covering the entire device with the hand was defined, constituting a total ratio of 0.3%.

Ear-Level Gestures. Ear-level gestures were categorized by location, as shown in Fig. 3. In the tragus part, touch and swipe gestures were defined for 26.1%/18.9% for in-ear/ear-hook devices, respectively. Within the helix part, a broader array of gestures, including swipe, pull, tap, long press, pinch, and flip, were defined, accounting for 45.8%/40.6%. For the earlobe part, gestures such as pull, swipe, and pinch were also defined for 17.6%/26.9%. In the central part, swipe and long press gestures were defined, comprising 0.7%/1.7%. Additionally, a gesture involving covering the entire ear with the hand was defined, constituting 6.0%/10.2%.

Body-Level Gestures. Body-level gestures were categorized by location, as shown in Fig. 3. In the forehead part, tap gestures were defined for 2.3%/2.3% for the in-ear/ear-hook device, respectively. In the eye part, tap and pinch gestures were defined for 2.3%/2.3%. In the nose part, tap gestures were defined for 1.1%/1.2%. In the cheek part, swipe, tap, and press gestures were defined for 68.8%/68.0%. In the mouse part, tap and swipe gestures were defined for 4.0%/4.1%. In the jaw part, tap gestures were defined for 2.8%/2.9%. In the neck part, swipe, press, and tap gestures were defined for 16.5%/16.9%. In the chest part, tap and swipe gestures were defined for 2.3%/2.3%.

Mid-Air-Level Gestures. The types of mid-air-level gestures are shown in Fig. 3. For gestures of static pose, gestures such as signing for the phone and holding the hand over were defined (11.1%). For gestures of static pose and path, gestures such as swipe and long press were defined. The ratios of gestures defined for the seven directions (up, down, approach, recede, forward, back, and other) were 13.5%, 11.5%, 6.4%, 7.1%, 7.6%, 10.0%, and 2.0%, respectively, for a total of 58.1%. For gestures of dynamic hand pose, gestures such as pinch-in and pinch-out of the fingertip and opening and closing of the hand were defined (30.7%).

3.3 Analysis and Discussion

3.3.1 Trends by Interaction Area Conditions. In the nature dimension, there was a trend toward fewer metaphoric gestures and more abstract gestures as conditions became more restrictive. In the context dimension, there was a trend toward slightly fewer in-context gestures as conditions became more narrow, but there were no significant differences in any of the conditions. We found that the number of gestures that deviate from the context does not increase noticeably just because the conditions have become more strict. In the flow dimension, there was little difference among the conditions. In the locale dimension, a considerable proportion of gestures was defined at the mid-air-level and at the device-level for both device shapes in the no-restriction condition. Conversely, very few users chose to define gestures for other body parts in this condition. However, in the touch condition, the ratio of gestures defined for body-level increased to over 20%, as many gestures initially defined in the air were redirected to other body parts. Notably, cheek and neck gestures emerged as preferred alternatives. For device-level and ear-level gestures, which could be defined up to the most restrictive ear-touch condition, there was an observable rise in the ratio of definitions as the condition became more restrictive. However, the rate of increase was more pronounced for ear-level gestures than for device-level gestures. This can be attributed to the higher initial definition rate of device-level gestures, coupled with limitations on the types of

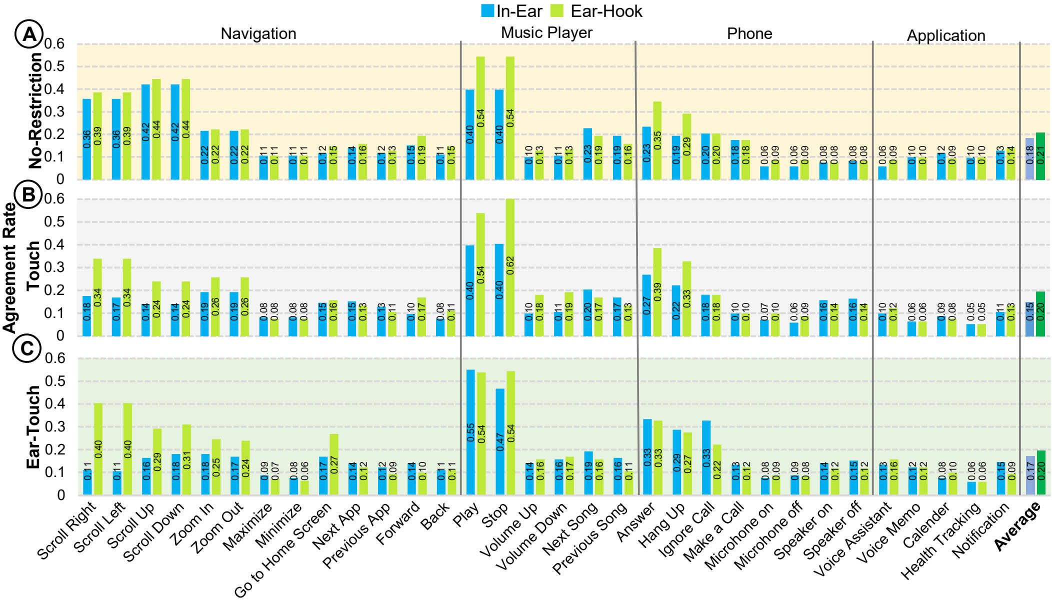

Fig. 4. AR scores for each task. A: no-restriction condition, B: touch condition, C: ear-touch condition.

gesture actions that could be defined. In the complexity dimension, compound gestures became more frequent as conditions became more restrictive. In the form dimension, static pose and path gestures accounted for the highest percentage of unrestricted conditions. This may be due to the fact that many swiping actions were defined with the air as the virtual screen. Static pose and path gestures were not defined in the touch and ear touch conditions. The percentage of deformation gestures and dynamic pose and path gestures was highest in the touch condition.

3.3.2 *Trends by Device Shape Conditions.*

In the nature dimension, ear-hook devices had slightly more physical and metaphoric gestures. Symbolic and abstract gestures did not differ much, and there were no major differences between devices in the context and flow dimensions. In the complexity dimension, there were fewer compound gestures for the ear-hook devices. This is thought to be due to the fact that the ear-mounted device has a larger driver part and a larger flat area that can be touched with a finger.

In contrast to in-ear devices, the ratio of gestures directed at the device was notably higher for ear-hook devices across all interaction area conditions. For instance, while in-ear devices lacked user-defined swipe gestures in various locations, ear-hook devices featured users defining swipe gestures in the driver and top/back stem parts. Consequently, the substantial housing size of ear-hook devices contributed to an increased ratio of definitions on the device body.

3.3.3 *Agreement Rate.*

Fig. 4 shows the agreement rate (AR) [39] for each condition, which is the evaluation index used to determine the user-defined gesture for each task.

$$AR(r) = \frac{|P|}{|P|-1} \sum_{P_i \subseteq P} \left(\frac{|P_i|}{|P|} \right)^2 - \frac{1}{|P|-1},$$

where P is the set of all proposals for referent r, $|P|$ is the number of the set, and P_i is the subsets of the same proposals from P. The AR is classified into four types: very high agreement ($AR \geq 0.5$), high agreement ($0.5 > AR \geq 0.3$), medium agreement ($0.3 \geq AR > 0.1$), and low agreement($0.1 \geq AR$).

The ARs for each condition are summarized in Fig 4. The average, highest, and lowest ARs for each condition were 0.18/0.42/0.06 (in-ear/no-restriction condition), 0.21/0.54/0.08 (ear-hook/no-restriction condition), 0.15/0.40/0.05 (in-ear/touch condition), 0.20/0.62/0.05 (open-ear/touch condition), 0.17/0.55/0.06 (in-ear/ear-touch condition), and 0.20/0.54/0.06 (open-ear/ear-touch condition). The stop task in the music player group had the highest average score across all conditions, with an average of 0.50. The health tracking task in the application group had the lowest average score across all conditions, with an average of 0.07. The ratio distribution of AR scores was as follows: 2.7% ($AR \geq 0.5$), 22.5% ($0.5 > AR \geq 0.3$), 50.9% ($0.3 \geq AR > 0.1$), and 23.9% ($0.1 \geq AR$).

To investigate the effect of device shape difference on AR, significance tests were conducted for each condition. The Shapiro-Wilk test showed a violation of the normality assumption in all conditions (no-restriction in-ear/ear-hook: $p = 0.00025/0.0000078 < .05$, touch in-ear/ear-hook: $p = 0.00019/0.00015 < .05$, ear-touch in-ear/ear-hook: $p=0.0000045/p = 0.00027 < .05$). The homogeneity of variance assumption was not violated (F test results: no-restriction $p = 0.42$, touch $p = 0.11$, ear-touch $p = 0.43$; all $> .05$). The Wilcoxon rank-sum test revealed significant differences between devices in the no-restriction ($p = 0.0019 < .05$) and touch conditions ($p = 0.0015 < .05$), but not in the ear-touch condition ($p = 0.50 > .05$). Additionally, the Friedman test was used to examine the effect of interaction area restriction for each device shape, revealing no significant differences (in-ear: $p = 0.201$, open-ear: $p = 0.339$; both $> .05$).

3.4 Finalized User-Defined Gesture

After categorizing similar gestures based on the taxonomy in Section 3.2, the gesture that occurs most frequently in each task is referred to as the representative gesture. We refer to the collection of these representative gestures as our user-defined gestures. The user-defined gestures for the ear-touch condition are listed in Table 3. The user-defined gestures for the no-restriction and touch conditions are summarized in Appendix A. In Section 4, we select ear-level gestures that can be recognized by the IMU from user-defined gestures and conduct gesture recognition experiments.

3.5 Design Implications

3.5.1 *RQ1: What gestures do users prefer as input to hearables?*

The gesture with the highest AR and the most frequent definition was the tap gesture. This gesture was defined across all task groups except for the application task group in the no-restriction condition. The ubiquity of this gesture can be attributed to its intuitive nature and the fact that the tap feature is already a prevalent functionality on many hearables. The preference for tap gestures has also been confirmed in a study that evaluated the usability of the presented gesture sets, further reinforcing this finding [45]. Symmetrical or directional tasks, such as zoom in/out or scroll right/left/up/down, were consistently translated into corresponding symmetrical or directional gestures, such as pinching in/out and swiping. This trend persisted across all conditions.

3.5.2 *RQ2: How do interaction area limitations and device shape differences affect user-defined gestures?*

The impact of interaction area conditions on user-defined gestures was particularly pronounced for directional tasks, specifically scroll right/left/up/down (navigation). The average AR for these tasks exhibited variation: 0.42 (no-restriction condition), 0.18 (touch condition), and 0.21 (ear-touch condition). Directional tasks tended to be defined in expansive spaces, often involving hand or finger swiping gestures. Consequently, the no-restriction condition likely witnessed the utilization of the most versatile aerial definitions for directional tasks. However, when mid-air-level gestures were not available, gesture definitions were dispersed across devices, ears (primarily tragus and

Table 3. User-defined gestures for ear-touch conditions. Pairs with different user-defined gestures for different devices are shown in bold.

Ear-Touch Condition			
Task Group	**Task**	**In-Ear Device [AR]**	**Ear-Hook Device [AR]**
Navigation	Scroll Right/Left	Swipe forward/back device [0.11/0.11]	Swipe forward/back device [0.40/0.40]
	Scroll Up/Down	Swipe up/down device [0.16/0.18]	Swipe up/down device [0.29/0.31]
	Zoom In	**Fold top helix and earlobe** [0.18]	**Pinch in device** [0.25]
	Zoom Out	**Pinch out helix** [0.17]	**Pinch out device** [0.24]
	Maximize	Pull up top helix [0.09]	Pull up top helix [0.07]
	Minimize	Pull down earlobe [0.08]	Pull down earlobe [0.06]
	Go to Home Screen	Tap device [0.17]	Tap device [0.27]
	Next App	**Swipe down helix** [0.14]	**Double tap device** [0.12]
	Previous App	**Swipe up helix** [0.12]	**Double tap device (left ear)** [0.09]
	Forward	**Swipe back tragus** [0.14]	**Pull back earlobe** [0.10]
	Back	**Swipe forward tragus** [0.11]	**Pull forward earlobe** [0.11]
Music Player	Play / Stop	Tap device [0.55/0.47]	Tap device [0.54/0.54]
	Volume Up/Down	**Swipe up/down helix** [0.14/0.16]	**Swipe up/down device** [0.16/0.17]
	Next Song	Double tap device [0.19]	Double tap device [0.16]
	Previous Song	Double tap device (left ear) [0.16]	Double tap device (left ear) [0.11]
Phone	Answer / Hang up	Tap device [0.33/0.29]	Tap device [0.33/0.27]
	Ignore Call	Long press device [0.33]	Long press device [0.22]
	Make a Call	Swipe back device [0.13]	Swipe back device [0.12]
	Microphone on/off	Pull down earlobe [0.08/0.09]	Pull down earlobe [0.09/0.08]
	Speaker on/off	Tap device (left ear) [0.14/0.15]	Tap device (left ear) [0.12/0.12]
Application	Voice Assistant	Long press device [0.13]	Long press device [0.16]
	Voice Memo	Tap device [0.12]	Tap device [0.12]
	Calendar	**Swipe up helix** [0.08]	**Swipe down helix** [0.10]
	Health Tracking	**Swipe back tragus** [0.06]	**Pull back middle helix** [0.06]
	Notifications	**Fold forward ear** [0.15]	**Pull down earlobe** [0.09]

helix), and other body parts (mainly cheeks and neck), leading to a markedly lower AR. This implies a lack of consensus on a specific body part for gesture definition, except for mid-air-level gestures.

Regarding differences in device shape, the AR for the ear-hook device surpassed that for the in-ear device. This discrepancy can be attributed to the larger device area of ear-hook devices, prompting more users to opt for defining gestures directly to the devices. Consequently, this choice diminishes the dispersion of definitions to other areas, encompassing mid-air-level gestures, ear touches, and interactions with various parts of the device. This observation aligns with the trend of pairs exhibiting different gestures for each task in Table 3. Most pairs with distinct gestures are defined as gestures to the ear for in-ear devices and gestures to the device for ear-hook devices. Conversely, user-defined gestures to the ear were evenly distributed concerning interaction area and deformation patterns, including tragus, helix, earlobe, and whole ear deformation (ear fold/fold top helix and earlobe) for in-ear devices. By contrast, for ear-hook devices, gestures were solely defined for the helix and earlobe.

In the ear-touch condition, identical pairs of user-defined gestures were observed for both device shapes, including swipe, tap, and long press on the device, along with pulling down the earlobe and pulling up the helix. These gestures appear to be universal user-defined actions, demonstrating independence from device shape. However, further exploration across various forms of hearables is warranted to validate their universality and applicability.

3.5.3 RQ3: What are the similarities and differences with previous studies on hearables and ears? In terms of taxonomy comparison, the results for the nature dimension showed the ratio of symbolic gestures was lower than in the previous study [27, 44]. Symbolic gestures were less common than in the previous study, possibly because users avoided using symbolic expressions, which require a certain amount of space because the gestures were performed on small earphone-type devices.

The results regarding the context dimension showed the ratio of in-context gestures was lower than in the previous study [8]. We surmise this is because users prefer simple gestures to devices rather than contextual gestures. The results regarding the flow dimension were similar to the smartphone-based and ear-related GES [8, 27], but differed from surface-related GES [44]. This suggests that the definition of discrete gestures increases significantly when the object being focused on is small, such as earphones or smartphones.

For the locale dimension, the distribution of gesture definitions within the ear also varied significantly. Notably, gestures defined behind the back of the ear, as identified in existing research, were not observed in this study, possibly for the same reason mentioned above. Shaikh et al. [30] conducted motion analysis of gestures for body parts above the neck, excluding the ears, and confirmed the preference for gestures to the cheeks and neck. This same preference trend was confirmed by the high ratio of cheeks and necks in the ratio of body-level definitions in our results. The results of the complexity dimension showed a higher ratio of simple gesture definitions than the ear-based GES [8]. We surmise that this was because users did not need compound gestures as they were able to define gestures to the device.

In terms of AR comparison, the results showed that tap and swipe gestures on the device had the highest AR for play/stop and directional tasks, consistent with a GES involving the device being worn [24]. However, this differs from a GES conducted without the device being worn [8], indicating that device presence significantly impacts user-defined gestures. The mean AR in this study was 0.19 (SD = 0.03), aligning closely with related studies [8, 24] (0.21/0.21). These consistent results suggest that users define gestures with similar intuitiveness for ear-related interactions, regardless of device shape or wearing status.

4 GESTURE RECOGNITION USING IMUS

We introduce a gesture recognition method employing an IMU incorporated into hearables, examining the recognition performance of ear-level gestures derived from the user-defined gestures established through the GES. Our approach utilizes an IMU that harmonizes effectively with hearables, minimizing implementation costs and optimizing IMU efficiency. In the evaluation experiment, we conducted a sitting and walking experiment utilizing both in-ear and ear-hook devices.

4.1 Recognition Systems

Fig. 5 presents a comprehensive recognition system overview. The user wears hearables and performs ear-level gestures. These gestures cause movements in the device due to ear deformation and pressure. Since each gesture affects the direction and intensity of the load differently, the IMU data vary accordingly. We use machine learning to create models that classify these gestures based on the data differences. Our system employs a k-NN (k-nearest neighbor) algorithm with dynamic time warping (DTW) as a metric for gesture classification, setting the parameter k to 3. Preliminary results showed that using only rotation data led to the best recognition performance; therefore, we

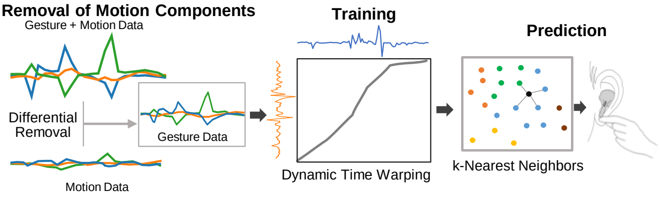

Fig. 5. System overview of gesture recognition system.

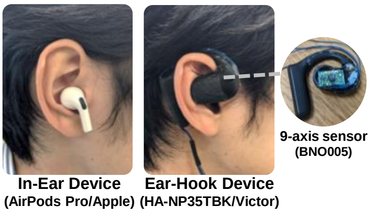
9-axis sensor (BNO005)
In-Ear Device (AirPods Pro/Apple)
Ear-Hook Device (HA-NP35TBK/Victor)

Fig. 6. In-ear devices and ear-hook devices. Location of the 9-axis sensor.

did not use acceleration and gravitational data. To address irrelevant motion data, we capture IMU data from both ears and compute differences to isolate gesture components. Due to the instability sampling rates, we employed timestamps and linear interpolation at 30 ms intervals to synchronize the data from both devices.

4.2 Implementation

In our investigation, gesture recognition experiments were conducted with two distinct device types: in-ear and ear-hook (Fig. 6). The AirPods Pro (Apple) served as the in-ear device, capable of capturing 9-axis IMU data. Data collection was facilitated through an Apple device connected via Bluetooth to the AirPods Pro. Owing to the unavailability of an ear-hook device capable of acquiring IMU data, we utilized a prototype device housing a 9-axis sensor, BNO005, enclosed in the HA-NP35TBK case (Victor). The sensor data was transmitted to a laptop (ASUS: ROG FLOW) through serial communication via an Arduino Uno connected by wire for data collection. Each gesture was recorded for a duration of 5 s, and the sampling rate for data measurement for both devices was approximately 30 Hz. The programs for machine learning and data collection on the laptop side of the ear-hook device were implemented in Python 3.7.

4.3 Evaluation

In this study, we investigated the recognition performance of our system by testing it in sitting and walking conditions.

4.3.1 Ear-Level Gestures. We selected ear-level gestures from the user-defined gestures determined in Section 3.4 and investigated the recognition rate. Gesture recognition with IMU sensors, which we focus on in this study, requires device displacement. Ear-level and device-level gestures cause this displacement. Certain commercial devices are equipped with built-in pressure-sensitive sensors for recognizing device-level gestures. Moreover, gestures presently unrecognizable, such as swipe forward/backward on devices, are anticipated to become recognizable in the near future through the built-in sensors. Consequently, our study concentrates on the recognition of ear-level gestures, specifically those expected to be recognized by IMU sensors. Fig. 7 shows a compilation of ear-level gestures. There were nine types of ear-level gestures identified for in-ear devices and six types for ear-hook devices.

4.3.2 Data Collection. In this procedure, 10 participants participated in a sitting experiment using both in-ear and ear-hook devices. Five participants participated in experiments for both devices, while the remaining participants were distinct, resulting in a total of 15 participants (male: 10, female: 5, average age: 26.8 years). Twelve out of the 15 participants were also involved in the GES. Five participants, three of whom were participants in the sitting experiment and two new participants, participated in the walking experiment (male: 3, female: 2, average age: 29.4 years). The duration of the experiment for each device was approximately one hour, with participants receiving

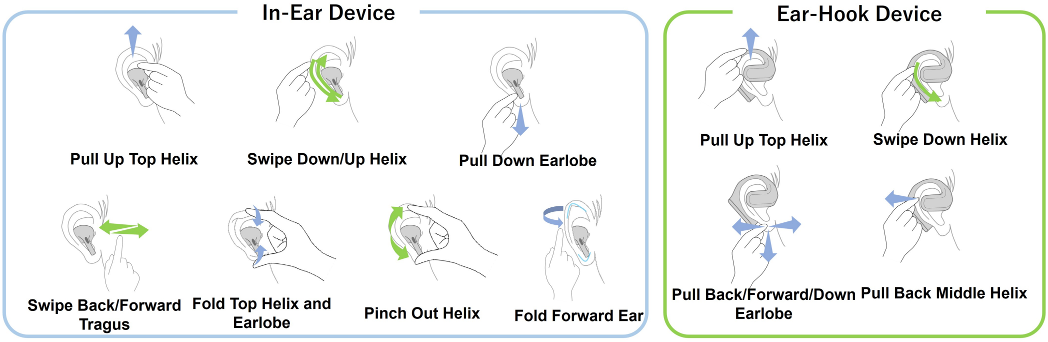

Fig. 7. Ear-level gestures for each device type.

a reward of approximately 10 US dollars. All participants in the experiment were right-handed, and 14 of them used earphones at least once a week. This experiment was approved by the ethical board at the author's institution, and informed consent was obtained from the participants.

Initially, we provided participants with an explanation of the gestures. Subsequently, the experimenter guided the participant through a brief practice phase for each gesture to confirm their ability to perform it accurately. Following this, participants were instructed to execute the gestures displayed on the screen, and sensor data were recorded during the gestures. The sequence of gestures performed was randomized, and 12 measurement rounds were undertaken, involving the attachment and detachment of the device for each round. For the walking experiment, data were collected as participants executed gestures while walking freely within an approximate space of 3 m × 3 m.

4.3.3 Results. In this study, we first conducted preliminary system design experiments comparing recognition performance within a combination of data or multiple training models. We then examined the recognition performance of both per-user and general models in the sitting experiment, as well as the performance of the per-user model in the walking experiment.

Data Selection. The 9-axis inertial sensor measures acceleration, rotation, and gravitational acceleration, each offering unique motion characteristics. We explored how different data types and their combinations affect gesture recognition rates. Training data were obtained through a sitting experiment with the in-ear device to construct individualized gesture classification models for each participant. Test and training data sets are separated, coupled with a 12-leave-one-round-out cross-validation. Fig. 8A revealed that rotation exhibited the highest recognition rate, averaging 91.03%. Hence, rotation was chosen as the data pattern for this study.

Model Selection. We compared various machine learning models to determine the best performer for our system. Our investigation encompassed the evaluation of recognition rates for DTW-kNN, conventional machine learning algorithms (SVM: support vector machine, RF: random forest, k-NN, MLP: multilayer perceptron, GB: gradient boosting) utilizing basic statistical features (mean, variance, standard deviation, median, maximum, minimum, root-mean-square), and deep learning (DNN: deep neural network) applied to raw sensor data. The learning method is the same as in the previous paragraph. The outcomes depicted in Fig. 8B revealed that DTW-kNN exhibited the highest recognition rate, averaging 91.0%. Hence, DTW-kNN was chosen as the machine learning model for this study.

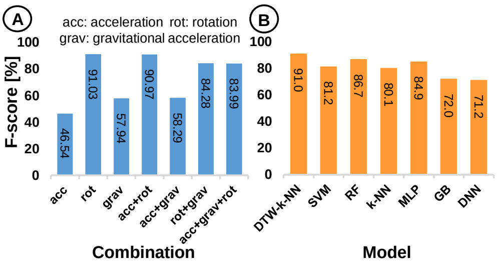

Fig. 8. Results of the preliminary system design experiments. A: recognition rate per data combination, B: recognition rate per model.

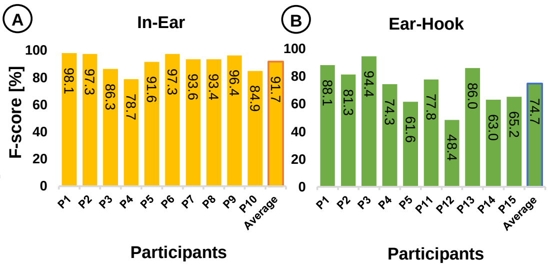

Fig. 9. Recognition rate for each participant. A: in-ear type, B: ear-hook type.

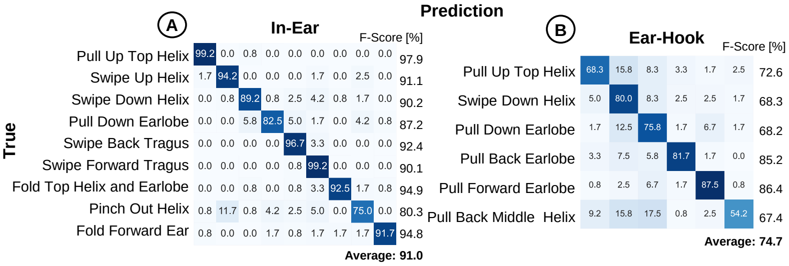

Fig. 10. Confusion matrix for the per-user model. A: in-ear type, B: ear-hook type.

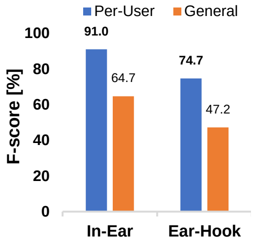

Fig. 11. Average recognition rate for per-user/general model.

Recognition Performance of Per-User Model. We provide a more detailed breakdown of the recognition rates for per-user models using DTW-kNN, shown in Fig. 9. The recognition rates per user are displayed for each device. For in-ear devices, recognition rates ranged from 98.1% (highest, participant P1) to 78.7% (lowest, participant P4). For ear-hook devices, rates varied from 94.4% (highest, participant P3) to 48.4% (lowest, participant P12). We conducted statistical tests to compare recognition rates between devices. The Shapiro-Wilk test showed no normality violation for in-ear and ear-hook device accuracies ($p = 0.117$ and $p = 0.879$, respectively, both $> .05$). However, the F test indicated unequal variances ($p = 0.00213 < .05$), leading us to perform a Welch's t-test, which revealed a significant difference between the devices ($p = 0.00353 < .05$). Fig. 10 shows the confusion matrix, with precision percentages for each gesture. The highest recognized gesture for in-ear devices was "pull up top helix" (97.9%), and for ear-hook devices, it was "pull forward earlobe" (86.4%). The lowest recognized gestures were "pinch out helix" (80.3%) for in-ear and "pull back middle helix" (67.4%) for ear-hook devices.

Recognition Performance of General Model. The general model, which doesn't require initial user data, offers usability benefits over the per-user model if the recognition rate is acceptable. We evaluated the recognition rate using leave-one-user-out cross-validation, where each user's data is omitted from training. Fig. 11 shows the recognition rates: 64.7% for in-ear devices and 47.2% for ear-hook devices. These rates are lower than those of the per-user model. This is likely caused by differences in gesture execution and device fit. Future work will focus on improving the general model's performance by collecting more gesture data or using transfer learning techniques.

Recognition Performance of Walking Experiment. We investigated the recognition rates for walking data and assessed the effect of the motion component removal algorithm. Fig. 12 shows the

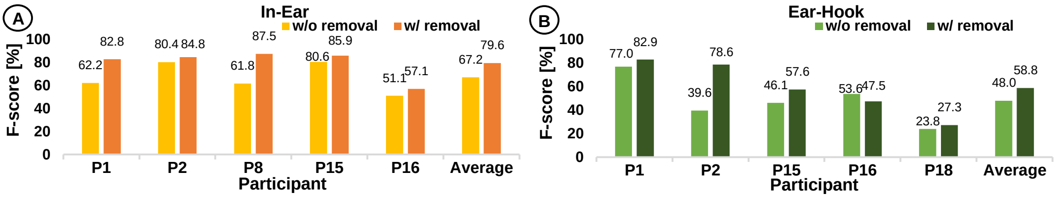

Fig. 12. Recognition rate per participant in the walking experiment. A: in-ear type, B: ear-hook type.

results with and without motion component removal for each device. The average recognition rate improved from 67.2% to 79.6% for in-ear devices and from 48.0% to 58.8% for ear-hook devices after removing motion components. For in-ear devices, the Shapiro-Wilk test indicated that data followed a normal distribution without motion component removal ($p = 0.266 > .05$), but not with it ($p = 0.00585 < .05$). Since the F test showed homogeneity of variance ($p = 0.164 > .05$), a Wilcoxon signed-rank test was performed, which revealed no significant difference ($p = 0.0625 > .05$). For ear-hook devices, the Shapiro-Wilk results showed normality (without/with $p = 0.932/0.690 > .05$) and homogeneity of variance (F-test, $p = 0.447 > .05$). The paired t-test showed no significant difference between groups ($p = 0.230 > .05$). However, removing motion components improved recognition rates by 12.4 points for in-ear and 10.8 points for ear-hook devices. These rates were lower than in the sitting experiment, likely due to reduced gesture stability from walking. Various methods [13, 32] have been proposed to eliminate motion noise and prevent recognition performance degradation. In future work, we will explore the integration of these methods into our approach.

5 DISCUSSION, LIMITATIONS, AND FUTURE WORK

5.1 Gesture Elicitation Study

5.1.1 Influence of Familiarity with the Device. Several participants in our study were already familiar with the tap gesture associated with earphones. In the music player task group, where many gesture-based operations are common in existing products, many participants preferred the same gestures. Of the 19 participants, nine had experience with device operations, such as tapping, and two frequently used these operations. Additionally, many inexperienced participants also assigned a single tap gesture to the earphones for play/stop tasks, resulting in a very high AR, especially for play/stop tasks.

Because of the varying implementation of gestures across different earphones and some participants' experiences with multiple operations, it was challenging to analyze the influence of past experience on user-defined gestures strictly. Future research should investigate whether experience with device operations affects user-defined gestures by comparing groups with and without specific operational experience.

5.1.2 Influence of Age. For the gesture to answer the phone, many participants in the no-restriction condition used the phone gesture, which involves bending the index, middle, and ring fingers. Younger participants often defined this gesture by placing their hand in a phone pose near the ear, mimicking a smartphone, while older participants tended to mimic turning a dial or picking up an imaginary receiver as if using a rotary phone. This indicates that user-defined gestures vary with age, especially in metaphorical gestures. Future work should include experiments with various age groups to investigate the impact of age on user-defined gestures.

5.1.3 Social Acceptability and Comfort of Ear-based Gestures. A few participants disliked the gesture of significantly deforming their ears or pushing in the earphones because of social acceptability and comfort when wearing the device. We believe that further research is needed to determine how the severity of the restriction affects social acceptability and comfort when wearing the device.

5.1.4 Limitations and Future Work. We observed that the shape of the device affected the user-defined gestures. Beyond the in-ear and ear-hook devices investigated in this study, various other device configurations exist, including integrated left-and-right types [31] and clip-type devices [3]. By aggregating insights into user-defined gestures across diverse device shapes, we may provide design guidelines for device forms in the future. To deepen our understanding of our results, it is essential to compare them not only with the GES but also with the usability or design space revealed in input studies related to hearables.

5.2 Gesture Recognition

5.2.1 Discussion. The in-ear device demonstrated a commendable recognition rate of 91.0%, affirming the feasibility of gesture recognition through IMU technology. By contrast, the ear-hook devices exhibited a lower recognition rate of 74.7%. This discrepancy can be attributed to the similarity between several gestures designed for ear-hook devices, specifically the two-directional pulling gesture for the ear helix and the three-directional pulling gesture for the earlobe, leading to increased misrecognition rates among these gestures. Roman et al. [16] also reported that when the helix part is divided into three sections, the touch accuracy of the middle region is low (63%), indicating that there are issues with the stability of gesture movements around these areas. To improve gesture recognition accuracy for similar gesture sets, our future endeavors will aim to elevate the IMU's sampling rate, capturing more detailed motion data for precise differentiation. Additionally, we plan to develop a multimodal recognition system that integrates IMU data with other sensory inputs, such as acoustic signals, to enhance the robustness of gesture recognition across diverse conditions.

In the GES, the AR was observed to be higher for ear-hook devices, which have a larger device area. However, this larger device area also introduced a challenge, as users defined similar gestures, contributing to a lower recognition rate for ear-level gestures. This highlights a finding concerning the influence of user-defined gestures on the recognition system. Conversely, we anticipate that the system's recognition performance will impact usability in real-world scenarios. Moving forward, we aim to delve into the interplay between these two aspects, gaining insights into user behavior and refining the system accordingly.

5.2.2 Limitations and Future Work. This study lacks gesture detection experiments. Our future plans include implementing a gesture detection algorithm and assessing its accuracy during daily activities such as walking or exercise.

We focused on recognizing ear-level gestures using the IMU sensor. In future works, we will aim to explore user-defined gestures that can be recognized by other types of sensors, such as microphones [45] and capacitance sensors [16].

6 CONCLUSION

In our GES, we explored hand inputs for hearables, revealing users' preferred gestures and variations in definition tendencies based on interaction area restrictions and device shapes. Subsequently, we conducted gesture recognition experiments utilizing an IMU for ear-level gestures among the user-defined gestures identified. The outcomes revealed a recognition rate of 91.0% for nine gesture types on in-ear devices and 74.7% for six gesture types on ear-hook devices. In the future, we aim to synergize both studies to deepen our understanding of user behavior and enhance system performance.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Shashank Ahire and Michael Rohs. 2020. Tired of Wake Words? Moving Towards Seamless Conversations with Intelligent Personal Assistants. In *Proceedings of the 2nd Conference on Conversational User Interfaces* (Bilbao, Spain) *(CUI '20)*. ACM, Article 20, 3 pages. https://doi.org/10.1145/3405755.3406141

[2] 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)*. ACM, 333–338. https://doi.org/10.1145/3594739.3610671

[3] Ambie. 2017. Ambie. https://ambie.co.jp/soundearcuffs/. (2024.02.15. accessed).

[4] Takashi Amesaka, Hiroki Watanabe, and Masanori Sugimoto. 2019. Facial Expression Recognition Using Ear Canal Transfer Function. In *Proceedings of the 23rd International Symposium on Wearable Computers (ISWC '19)*. ACM, 1–9. https://doi.org/10.1145/3341163.3347747

[5] 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)*. ACM, 679–689. https://doi.org/10.1145/3126594.3126649

[6] Hongliang Bi, Yuanyuan Sun, Jiajia Liu, and Lihao Cao. 2022. SmartEar: Rhythm-Based Tap Authentication Using Earphone in Information-Centric Wireless Sensor Network. *IEEE Internet of Things Journal* 9, 2 (2022), 885–896. https://doi.org/10.1109/JIOT.2021.3063479

[7] Edwin Chan, Teddy Seyed, Wolfgang Stuerzlinger, Xing-Dong Yang, and Frank Maurer. 2016. User Elicitation on Single-Hand Microgestures. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, 3403–3414. https://doi.org/10.1145/2858036.2858589

[8] 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. *Proceedings of the ACM on Human-Computer Interaction* 4, ISS, Article 186 (2020), 20 pages. https://doi.org/10.1145/3427314

[9] Christine Dierk, Scott Carter, Patrick Chiu, Tony Dunnigan, and Don Kimber. 2019. Use Your Head! Exploring Interaction Modalities for Hat Technologies. In *Proceedings of the 2019 on Designing Interactive Systems Conference (DIS '19)*. ACM, 1033–1045. https://doi.org/10.1145/3322276.3322356

[10] Koumei Fukahori, Daisuke Sakamoto, and Takeo Igarashi. 2015. Exploring Subtle Foot Plantar-Based Gestures with Sock-Placed Pressure Sensors. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, 3019–3028. https://doi.org/10.1145/2702123.2702308

[11] Bogdan-Florin Gheran, Jean Vanderdonckt, and Radu-Daniel Vatavu. 2018. Gestures for Smart Rings: Empirical Results, Insights, and Design Implications. In *Proceedings of the 2018 Designing Interactive Systems Conference (DIS '18)*. ACM, 623–635. https://doi.org/10.1145/3196709.3196741

[12] 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. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2, Article 57 (2022), 28 pages. https://doi.org/10.1145/3534613

[13] Peiqi Kang, Jinxuan Li, Bingfei Fan, Shuo Jiang, and Peter B. Shull. 2022. Wrist-Worn Hand Gesture Recognition While Walking via Transfer Learning. *IEEE Journal of Biomedical and Health Informatics* 26, 3 (2022), 952–961. https://doi.org/10.1109/JBHI.2021.3100099

[14] Takashi Kikuchi, Yuta Sugiura, Katsutoshi Masai, Maki Sugimoto, and Bruce H. Thomas. 2017. EarTouch: Turning the Ear into an Input Surface. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '17)*. ACM, Article 27, 6 pages. https://doi.org/10.1145/3098279.3098538

[15] DoYoung Lee, Youryang Lee, Yonghwan Shin, and Ian Oakley. 2018. Designing Socially Acceptable Hand-to-Face Input. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology (UIST '18)*. ACM, 711–723. https://doi.org/10.1145/3242587.3242642

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

[17] Meethu Malu, Pramod Chundury, and Leah Findlater. 2018. Exploring Accessible Smartwatch Interactions for People with Upper Body Motor Impairments. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, 1–12. https://doi.org/10.1145/3173574.3174062

[18] 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

[19] Katsutoshi Masai, Kai Kunze, Daisuke Sakamoto, Yuta Sugiura, and Maki Sugimoto. 2020. Face Commands - User-Defined Facial Gestures for Smart Glasses. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 374–386. https://doi.org/10.1109/ISMAR50242.2020.00064

[20] 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 *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, 1911–1922. https://doi.org/10.1145/3025453.3025692

[21] 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

[22] Vimal Mollyn, Riku Arakawa, Mayank Goel, Chris Harrison, and Karan Ahuja. 2023. IMUPoser: Full-Body Pose Estimation Using IMUs in Phones, Watches, and Earbuds. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM, Article 529, 12 pages. https://doi.org/10.1145/3544548.3581392

[23] Meredith Ringel Morris, Jacob O. Wobbrock, and Andrew D. Wilson. 2010. Understanding Users' Preferences for Surface Gestures. In *Proceedings of Graphics Interface 2010 (GI '10)*. Canadian Information Processing Society, 261–268. https://www.microsoft.com/en-us/research/publication/understanding-users-preferences-surface-gestures/

[24] 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

[25] 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

[26] Tobias Röddiger, Daniel Wolffram, David Laubenstein, Matthias Budde, and Michael Beigl. 2020. Towards Respiration Rate Monitoring Using an In-Ear Headphone Inertial Measurement Unit. In *Proceedings of the 1st International Workshop on Earable Computing (EarComp '19)*. ACM, 48–53. https://doi.org/10.1145/3345615.3361130

[27] Jaime Ruiz, Yang Li, and Edward Lank. 2011. User-Defined Motion Gestures for Mobile Interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, 197–206. https://doi.org/10.1145/1978942.1978971

[28] Marcos Serrano, Barrett M. Ens, and Pourang P. Irani. 2014. Exploring the Use of Hand-to-Face Input for Interacting with Head-Worn Displays. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, 3181–3190. https://doi.org/10.1145/2556288.2556984

[29] Adwait Sharma, Joan Sol Roo, and Jürgen Steimle. 2019. Grasping Microgestures: Eliciting Single-Hand Microgestures for Handheld Objects. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, 1–13. https://doi.org/10.1145/3290605.3300632

[30] Shaikh Shawon Arefin Shimon, Ali Neshati, Junwei Sun, Qiang Xu, and Jian Zhao. 2024. Exploring Uni-manual Around Ear Off-Device Gestures for Earables. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 1, Article 3 (mar 2024), 29 pages. https://doi.org/10.1145/3643513

[31] SHOKZ. 2023. OPENRUN. https://jp.shokz.com/products/openrun. (2024.02.15. accessed).

[32] Zhipeng Song, Zhichao Cao, Zhenjiang Li, Jiliang Wang, and Yunhao Liu. 2021. Inertial motion tracking on mobile and wearable devices: Recent advancements and challenges. *Tsinghua Science and Technology* 26, 5 (2021), 692–705. https://doi.org/10.26599/TST.2021.9010017

[33] SONY. 2022. LinkBuds. https://www.sony.jp/headphone/products/LinkBuds/. (2024.02.15. accessed).

[34] 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

[35] 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)*. ACM, 4255–4260. https://doi.org/10.1145/1520340.1520649

[36] 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

[37] Radu-Daniel Vatavu. 2012. User-Defined Gestures for Free-Hand TV Control. In *Proceedings of the 10th European Conference on Interactive TV and Video (EuroITV '12)*. ACM, 45–48. https://doi.org/10.1145/2325616.2325626

[38] Radu-Daniel Vatavu. 2023. iFAD Gestures: Understanding Users' Gesture Input Performance with Index-Finger Augmentation Devices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, Article 576, 17 pages. https://doi.org/10.1145/3544548.3580928

[39] Radu-Daniel Vatavu and Jacob O. Wobbrock. 2015. Formalizing Agreement Analysis for Elicitation Studies: New Measures, Significance Test, and Toolkit. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, 1325–1334. https://doi.org/10.1145/2702123.2702223

[40] Radu-Daniel Vatavu and Jacob O. Wobbrock. 2022. Clarifying Agreement Calculations and Analysis for End-User Elicitation Studies. 29, 1, Article 5 (jan 2022), 70 pages. https://doi.org/10.1145/3476101

[41] 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

[42] Santiago Villarreal-Narvaez, Arthur Sluÿters, Jean Vanderdonckt, and Radu-Daniel Vatavu. 2024. Brave New GES World: A Systematic Literature Review of Gestures and Referents in Gesture Elicitation Studies. *Comput. Surveys* 56, 5, Article 128 (jan 2024), 55 pages. https://doi.org/10.1145/3636458

[43] Martin Weigel, Vikram Mehta, and Jürgen Steimle. 2014. More than Touch: Understanding How People Use Skin as an Input Surface for Mobile Computing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, 179–188. https://doi.org/10.1145/2556288.2557239

[44] Jacob O. Wobbrock, Meredith Ringel Morris, and Andrew D. Wilson. 2009. User-Defined Gestures for Surface Computing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. ACM, 1083–1092. https://doi.org/10.1145/1518701.1518866

[45] 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 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. ACM, 1–14. https://doi.org/10.1145/3313831.3376836

[46] Takumi Yamamoto, Katsutoshi Masai, Anusha Withana, and Yuta Sugiura. 2023. Masktrap: Designing and Identifying Gestures to Transform Mask Strap into an Input Interface. In *Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23)*. ACM, 762–775. https://doi.org/10.1145/3581641.3584062

[47] Yukang Yan, Chun Yu, Yingtian Shi, and Minxing Xie. 2019. PrivateTalk: Activating Voice Input with Hand-On-Mouth Gesture Detected by Bluetooth Earphones. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology (UIST '19)*. ACM, 1013–1020. https://doi.org/10.1145/3332165.3347950

[48] Yukang Yan, Chun Yu, Xin Yi, and Yuanchun Shi. 2018. HeadGesture: Hands-Free Input Approach Leveraging Head Movements for HMD Devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 4, Article 198 (2018), 23 pages. https://doi.org/10.1145/3287076

[49] 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)*. ACM, Article 94, 6 pages. https://doi.org/10.1145/3544549.3585903

A USER-DEFINED GESTURES

Table 4 and Table 5 list the user-defined gestures for the no restriction and touch conditions, respectively.

Received February 2024; revised May 2024; accepted June 2024

Table 4. User-defined gestures for no restriction conditions.

No Restriction Condition			
Task Group	**Task**	**In-Ear Device [AR]**	**Ear-Hook Device [AR]**
Navigation	Scroll Right	Swipe recede in the air [0.36]	Swipe recede in the air [0.39]
	Scroll Left	Swipe approach in the air [0.36]	Swipe approach in the air [0.39]
	Scroll Up	Swipe up in the air [0.42]	Swipe up in the air [0.44]
	Scroll Down	Swipe down in the air [0.42]	Swipe down in the air [0.44]
	Zoom In	Pinch in in the air [0.22]	Pinch in in the air [0.22]
	Zoom Out	Pinch out in the air [0.22]	Pinch out in the air [0.22]
	Maximize	Pinch in in the air with all fingers [0.11]	Pinch in in the air with all fingers [0.11]
	Minimize	Pinch out in the air with all fingers [0.11]	Pinch out in the air with all fingers [0.11]
	Go to Home Screen	Tap device [0.12]	Tap device [0.15]
	Next App	Swipe forward in the air [0.14]	Swipe forward in the air [0.16]
	Previous App	Swipe back in the air [0.12]	Swipe back in the air [0.13]
	Forward	Double tap device (right) [0.15]	Double tap device (right) [0.19]
	Back	Double tap device (left) [0.11]	Double tap device (left) [0.15]
Music Player	Play / Stop	Tap device [0.40]	Tap device [0.54]
	Volume Up	Swipe up in the air [0.40]	Swipe up in the air [0.54]
	Volume Down	Swipe down in the air [0.10]	Swipe down in the air [0.13]
	Next Song	Double tap device (right) [0.23]	Double tap device (right) [0.19]
	Previous Song	Double tap device (left) [0.19]	Double tap device (left) [0.16]
Phone	Answer	Tap device [0.23]	Tap device [0.35]
	Hang up	Tap device [0.19]	Tap device [0.29]
	Ignore Call	Long press device [0.20]	Long press device [0.20]
	Make a Call	Phone sign [0.18]	Phone sign [0.18]
	Microphone on / off	Double tap device [0.06]	Double tap device [0.09]
	Speaker on / off	Double tap device (left) [0.08]	Double tap device (left) [0.08]
Application	Voice Assistant	Long press device [0.06]	Long press device [0.09]
	Voice Memo	Put a hand close to a mouth [0.10]	Put a hand close to a mouth [0.10]
	Calendar	Draw a rectangle in the air [0.12]	Draw a rectangle in the air [0.09]
	Health Tracking	Pulse-taking hand sign [0.10]	Pulse-taking hand sign [0.10]
	Notifications	Put a hand close to an ear [0.13]	Put a hand close to an ear [0.14]

Table 5. User-defined gestures for touch conditions.

Touch Condition			
Task Group	**Task**	**In-Ear Device [AR]**	**Ear-Hook Device [AR]**
Navigation	Scroll Right	Swipe right cheek [0.18]	Swipe right cheek [0.34]
	Scroll Left	Swipe left cheek [0.17]	Swipe left cheek [0.34]
	Scroll Up	Swipe up cheek [0.14]	Swipe up cheek [0.24]
	Scroll Down	Swipe down cheek [0.14]	Swipe down cheek [0.24]
	Zoom In	Pinch in cheek [0.19]	Pinch in cheek [0.26]
	Zoom Out	Pinch out cheek [0.19]	Pinch out cheek [0.26]
	Maximize	Pinch in cheek (two times) [0.08]	Pinch in cheek (two times) [0.08]
	Minimize	Pinch out cheek (two times) [0.08]	Pinch out cheek (two times) [0.08]
	Go to Home Screen	Tap device [0.15]	Tap device [0.16]
	Next App	Swipe down helix [0.15]	Double tap device [0.13]
	Previous App	Swipe up helix [0.13]	Double tap device (left) [0.11]
	Forward	Tap device (right) [0.10]	Double tap device (right) [0.17]
	Back	Tap device (left) [0.08]	Double tap device (left) [0.11]
Music Player	Play	Tap device [0.40]	Tap device [0.54]
	Stop	Tap device [0.40]	Tap device [0.62]
	Volume Up	Swipe up helix [0.10]	Swipe up device [0.18]
	Volume Down	Swipe down helix [0.11]	Swipe down device [0.19]
	Next Song	Double tap device [0.20]	Double tap device [0.17]
	Previous Song	Double tap device (left) [0.17]	Double tap device (left) [0.13]
Phone	Answer	Tap device [0.27]	Tap device [0.39]
	Hang up	Tap device [0.22]	Tap device [0.33]
	Ignore Call	Long press device (left) [0.18]	Long press device (left) [0.18]
	Make a Call	Long press device [0.10]	Long press device [0.10]
	Microphone on	Double tap device [0.07]	Double tap device [0.10]
	Microphone off	Double tap device [0.06]	Double tap device [0.09]
	Speaker on	Double tap device (left) [0.16]	Double tap device (left) [0.14]
	Speaker off	Double tap device (left) [0.16]	Double tap device (left) [0.14]
Application	Voice Assistant	Long press device [0.10]	Long press device [0.12]
	Voice Memo	Tap device [0.06]	Tap device [0.06]
	Calendar	Swipe up helix [0.09]	Swipe down helix [0.08]
	Health Tracking	put two fingers on neck [0.05]	put two fingers on neck [0.05]
	Notifications	cover ear with hand [0.11]	cover ear with hand [0.13]