EchoWrist: Continuous Hand Pose Tracking and Hand-Object Interaction Recognition Using Low-Power Active Acoustic Sensing On a Wristband

Chi-Jung Lee*, Ruidong Zhang*, Devansh Agarwal, Tianhong Catherine Yu, Vipin Gunda
Oliver Lopez, James Kim, Sicheng Yin, Boao Dong, Ke Li, Mose Sakashita
François Guimbretière and Cheng Zhang
{cl2358,rz379,da398,ty274,vg245,ojl23,jjk297,sy594,bd324,kl975,ms3522,fvg3,chengzhang}@cornell.edu
Cornell University
Ithaca, New York, USA

Figure 1: EchoWrist is a wristband that can understand 3D hand poses as well as hand-object interactions. (a) The EchoWrist prototype worn by a user. EchoWrist adopts a minimally-obtrusive design that keeps the device compact and low-profile. EchoWrist is able to (b) continuously track hand poses and (c) recognize various hand-object interactions.

ABSTRACT

Our hands serve as a fundamental means of interaction with the world around us. Therefore, understanding hand poses and interaction contexts is critical for human-computer interaction (HCI). We present EchoWrist, a low-power wristband that continuously estimates 3D hand poses and recognizes hand-object interactions using active acoustic sensing. EchoWrist is equipped with two speakers emitting inaudible sound waves toward the hand. These sound waves interact with the hand and its surroundings through reflections and diffractions, carrying rich information about the hand's shape and the objects it interacts with. The information captured by the two microphones goes through a deep learning inference system that recovers hand poses and identifies various everyday hand activities. Results from the two 12-participant user studies show that EchoWrist is effective and efficient at tracking 3D hand poses and recognizing hand-object interactions. Operating at 57.9 mW, EchoWrist can continuously reconstruct 20 3D hand joints

*Both authors contributed equally to this research.

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642910

with MJEDE of 4.81 mm and recognize 12 naturalistic hand-object interactions with 97.6% accuracy.

CCS CONCEPTS

• **Human-centered computing → Interaction devices**.

KEYWORDS

Acoustic Sensing, Wearable, Smartwatch, Hand Pose, Hand-Object Interaction

ACM Reference Format:

Chi-Jung Lee*, Ruidong Zhang*, Devansh Agarwal, Tianhong Catherine Yu, Vipin Gunda, Oliver Lopez, James Kim, Sicheng Yin, Boao Dong, Ke Li, Mose Sakashita, François Guimbretière and Cheng Zhang. 2024. EchoWrist: Continuous Hand Pose Tracking and Hand-Object Interaction Recognition Using Low-Power Active Acoustic Sensing On a Wristband. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3613904.3642910

1 INTRODUCTION

Human hands play an essential role in our daily lives. From non-verbal communication through gestures (e.g., sign language) to exploring our surroundings through touch and even grasping and manipulating objects, our hands serve as indispensable tools. Appreciating the significance of these activities involving our hands not only allows us to better understand our daily lives but also defines its practical applications, particularly in the context of human-computer interaction (HCI), which spans from passive context awareness to active input methods.

Building systems to track hand activities has been a long-standing challenge for the research community. This includes continuously capturing the 3D poses of hands and understanding the context of their interactions, e.g., what they are interacting with. The challenge arises due to 1) the highly flexible nature of hands and their numerous joints, 2) the potential occlusion of the fingers, and 3) the interaction with other objects that can change the shape of hands and obscure them from view. Consequently, understanding hand-object interactions poses a significant challenge since the system must possess knowledge of both the hand and the object. As a result, traditional hand-tracking solutions rely on external cameras to "see" the entire hand [13, 21, 41, 46, 48, 53, 57, 74, 90]. However, these methods are often intrusive, power-hungry, or require pre-setup, making them inconvenient for deployment in everyday life.

In response to these challenges, researchers from the wearable community have proposed various solutions. As skin-contacting sensors [1, 10, 12, 16, 24, 31, 39, 40, 45, 60, 65, 66, 86, 87] suffer from wearing discomfort, other solutions [3, 7, 22, 38, 61, 64, 71, 72, 79, 80] are limited to recognizing a pre-defined set of gestures without continuous pose tracking capability. In contrast, while wearable camera-based solutions [18, 25, 44, 78, 81] support continuous tracking, they experience challenges in power consumption, e.g., 3.6 W from DiscoBand [9]. This prevents the systems from full-day usage. Also, privacy concerns from bystanders are raised on these methods [8, 28, 51]. Besides, most continuous solutions require an obtrusive device [25, 39, 40, 42, 52, 78, 81] and typically focus on hand gestures without the ability to recognize hand-object interactions. In summary, existing solutions exhibit at least one of the following limitations: 1) discomfort from obtrusive form factor, 2) high power consumption, 3) lack of continuous tracking capabilities, 4) insufficient consideration of hand-object interactions, and 5) privacy concerns.

To address these challenges, we introduce EchoWrist, a minimally obtrusive, low-power wristband designed to provide both continuous 3D hand shape tracking and a nuanced understanding of various hand-object interactions. EchoWrist utilizes active acoustic sensing, incorporating two pairs of compact speakers and microphones positioned in close proximity to the skin on each side of the wrist. The speakers emit inaudible frequency-modulated continuous waves (FMCW) directed toward the hand, and the resulting sound wave reflections and diffractions are captured by the wristband's microphones, creating distinct patterns corresponding to different hand poses. We then use a customized deep convolutional neural network (CNN) to continuously deduce the 3D hand poses represented by the 3D positions of 20 finger joints while also classifying various hand-object interactions.

To evaluate EchoWrist's performance in continuous hand pose tracking and hand-object interaction recognition, we conducted two user studies, each involving 12 participants. The results indicate that EchoWrist can continuously track 20 finger joints with a mean joint Euclidean distance error (MJEDE) of 4.81mm or mean joint angular error (MJAE) of 3.79°with a user-dependent (UD) model. Furthermore, EchoWrist achieves a recognition rate of 97.6% across 12 diverse hand-object interactions, spanning static scenarios, such as firmly holding a cup, to dynamic actions involving movement, such as chopping. In addition, EchoWrist operates at a significantly lower power consumption compared to prior works [9] of just 57.9 mW, with the sensing modules (speakers and microphones) consuming only 10.0 mW, enabling full-day usage on standard smartwatches (e.g., 19 hours with 300mAh battery on an Apple Watch). Compared with previous work with continuous tracking capabilities [9, 39, 40], EchoWrist adopts a much smaller size and less obtrusive form factor.

This paper presents the following contributions:

- We propose EchoWrist, a wireless, low-power, and low-profile wristband that can continuously track 3D hand poses and recognize hand-object interactions using active acoustic sensing.
- To our knowledge, EchoWrist is the first low-power and low-profile wristband that can both track 3D hand poses continuously and recognize hand-object interactions.
- We evaluated EchoWrist with 3 user studies with 36 participants in total to demonstrate promising continuous hand pose tracking and hand-object interaction recognition capabilities.
- We presented the design considerations and iterations and further discussed the opportunities and challenges of deploying EchoWrist at scale.

2 RELATED WORK

The wrist represents an advantageous location for hand sensing due to its inherent benefits, such as minimal interference with intricate finger dexterity and reduced susceptibility to external

object occlusions. We discuss previous wristbands on hand-pose tracking, gesture recognition, and hand-object interaction sensing. In addition, we provide a brief overview of sensing techniques using acoustic signals, which constitute our core sensing method.

2.1 Hand Pose Tracking and Gesture Recognition

Form factors other than wristbands exist in hand tracking and gesture recognition, such as gloves [58] and rings [20, 55, 73, 75, 76, 88]. However, wristbands usually create less interference with daily activities. With the increasing prevalence of smartwatches and smart bands, wristbands have more advantages for large-scale deployment. Therefore, we focus on wristband-based methods for the following discussion.

2.1.1 Discrete Gesture Recognition. Discrete gesture recognition is relatively less challenging compared with continuous 3D hand tracking. Researchers have explored various sensors. A heavily explored direction is using electromyography (EMG), which captures electric signals from muscle movements [11, 23, 59, 60, 65, 66]. Similar sensors include electric impedance sensing [24, 86, 87] and ultrasonic imaging [45]. These technologies detect gestures through internal changes but usually require skin contact and calibration. Another approach utilizes piezoelectric sensors [1, 10, 16], surface transducer [82], or high-frequency motion sensors [31], which capture bio-acoustic signals from wrist and finger movements. Technologies based on monitoring local shape changes use less invasive modalities such as pressure/flexors [7, 38] and capacitive sensors [61, 64, 70–72], with some achieving ultra-low-power performance [72].

Overall, skin-contacting technologies require sensors tightly attached to the skin, potentially causing discomfort over time. Therefore, contact-free hand-tracking technologies have garnered interest due to their comfort for extended use and promising performance potential. For instance, IMUs have been used on the palm/wrist [56, 73] or exclusively on smart devices [80] to recognize gestures. Other methods include proximity sensor arrays positioned on the wrist [15, 26] or thumb [69], as well as vision sensors like RGB [78] and IR [37, 44, 81]. Still, due to their limited precision in capturing hand-pose data, these methods usually recognize only a few gestures, preventing many potential applications.

2.1.2 Continuous Hand Pose Tracking. Recent advancements make it possible to track hand poses continuously with a wristband. Recent studies on EMG have shown promising results in recording continuous finger movements [39, 40]. However, EMG requires skin-contacting electrodes, which may not be comfortable for long-term wearing. In addition, many EMG-based methods usually place sensors at mid-forearm rather than a wristband, which may compromise comfort and convenience. Another promising direction is using wearable cameras, such as IR cameras [25, 81], thermal cameras [18] and depth cameras [9]. However, they usually have significant power requirements and spacing constraints [9, 25, 81], making them difficult to integrate into wearables such as smartwatches. For instance, DiscoBand [9] operates at 3.6 W while Digits [25] requires a bulky camera on the palm's side. Compared with previous work, EchoWrist operates at 57.9mW, and the highest point of the sensors is 5mm from the skin. EchoWrist provides a fully contact-free low-power solution that has a minimally obtrusive form factor with low-profile commercial speakers and microphones and provides continuous tracking capability.

2.2 Hand-Object Interactions

2.2.1 Camera-Based Methods. As cameras offer a wealth of data, including contextual information, the use of a wrist-mounted camera can simultaneously capture both hand postures and the surrounding environment. As a result, wrist-worn camera methods have been proposed to recognize daily activities [36, 42, 52]. However, to optimize data quality and minimize occlusion, the camera must be positioned at a certain distance from the skin, resulting in increased device thickness and potential discomfort. Additionally, there are concerns about the adequacy of privacy protection. To address these issues, DiscoBand [9] utilized multiple depth cameras while consuming high power and being bulky.

2.2.2 Other Sensing Signals. To minimize the occlusion issue, other signals were analyzed to understand hand activities. Fan et al. [12] recognized in-hand objects via EMG. Rudolph et al. [64] analyzed wrist topography to understand hand activities. However, these methods are sensitive to the grasping postures.

Since hands move to interact with the surroundings, motion-based methods were proposed. Using inertial sensors, EatingTrak [85] detected the eating action. Using the multimodel method, Mollyn et al. [47] and Bhattacharya et al. [2] sensed hand activities with lower power consumption. However, while the research demonstrates promising results, it remains challenging to track static interactions with minimal movements.

On the other hand, some research has explored recognizing the activities based on the vibration profile from the contacting objects. Surface acoustic waves [14] were used to sense the gestures against objects. ViBand [31] passively utilized accelerometers to capture bio-acoustic signals, while VibEye [50] actively propagated the vibration. In addition, Laput et al. [30] leveraged commodity smartwatches to capture passive bio-acoustic signals. These methods achieve plausible results in recognizing the in-hand objects. However, passive vibration methods suffer from recognizing objects with minimal vibrations, while active vibration methods constrain the hand postures.

In summary, EchoWrist excels in achieving hand-object interaction recognition through a low-power and low-profile wristband design while minimizing constraints related to interaction types, objects, and postures.

2.3 Active Acoustic Sensing

Active acoustic sensing emits sound waves using speakers and receives the reflected acoustic waves using microphones. These received reflected acoustic signals contain rich information, e.g., position and shape, about the object that reflected the signals. This sensing principle has been widely used as "sonar" in the past. Acoustic sensors are widely available on modern computing devices, including smartphones, wearables, and smart speakers. Therefore, much prior research has explored using active acoustic sensing on these form factors to recognize human activities. Some researchers explored using active acoustic sensing through surface propagation,

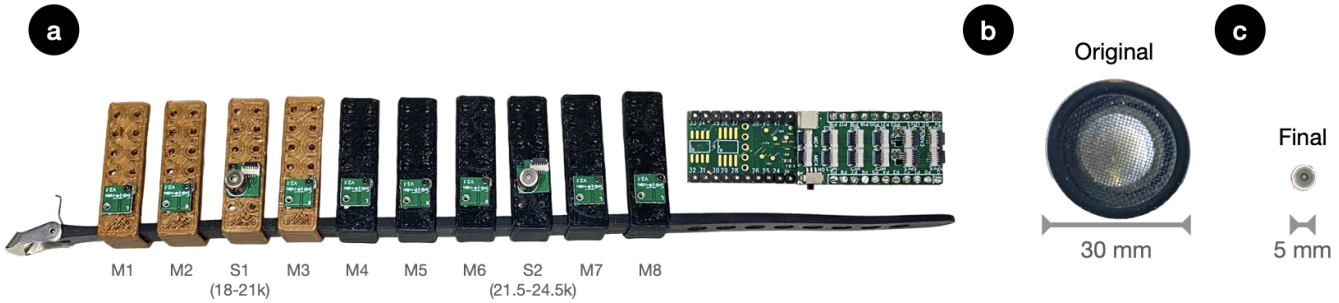

Figure 2: Pilot studies were conducted with (a) an experimental prototype. (b) An ultrasonic transducer was used for the initial design, while (c) a commodity speaker was chosen for the later prototype.

which can be used to recognize body contact on surfaces [54] or touch gestures if combined with air-borne signals [14, 67]. Recently, researchers have demonstrated the use of active acoustic sensing for various applications. These include tracking hand gestures through microphone and speaker arrays on a smart speaker [32], monitoring facial expressions through earphones [35], tracking eye movements [33], silent speech [84], mouth activities [68], facial movements [34], and upper body poses [43] on glasses, understanding silent speech with headphones [83], recognizing which finger is interacting with a smartwatch [27], and introducing novel attachments on phones for innovative interactions [29].

The work aligned most with our work is the active acoustic sensing used to sense hand gestures. FingerIO [49] tracks 2D fine-grained finger movements around a smartphone or smartwatch. WristAcoustic [19] achieves gesture recognition for authentication on smartwatches. AudioGest [63] enables gesture recognition on a laptop, tablet, or smartwatch. BeamBand [22] achieves gesture recognition on a smartwatch both within-session and across-session.

In contrast, EchoWrist is the first wristband to use active acoustic sensing to continuously track hand poses and recognize hand-object interactions. It demonstrates a unique sensing principle that uses acoustic sensing to capture the shape and position of the wrist contour and surrounding objects, which machine learning (ML) models can learn to infer hand poses and recognize hand-object interactions.

3 SENSING PRINCIPLES AND DESIGN CONSIDERATIONS

EchoWrist employs active acoustic sensing as its primary sensing method. In this approach, we use speakers to emit inaudible sound waves and microphones to receive the reflections of these emitted waves. These sound waves propagate from the wrist toward the palm, fingers, and the surrounding environment. The skin and surfaces of nearby objects act as the reflection medium for these sound waves. The waves undergo reflection and diffraction, eventually reaching the microphones. Distinct hand shapes and the varying characteristics of the surrounding environment lead to different signal paths, resulting in complex multipath echo patterns, which could be distinguished with customized echo profile analysis and deep learning pipelines.

In order to determine the optimal design for EchoWrist in efficiently tracking 3D hand poses and recognizing hand-object interactions, a series of pilot studies were conducted. We aimed to explore the limitations of the sensing principle and identify the optimal setup that balances device obtrusiveness, power consumption, and performance. Overall, two design considerations were proposed:

(1) Minimally-Obtrusive: Given that wearable devices are typically worn throughout the day and come into prolonged contact with the user, ensuring comfort is a prioritized consideration in the design. Additionally, as our aim is to introduce a wristband, it's essential to prevent the device from interfering with daily hand-related activities. It is also vital to consider the social acceptability of the wearable device and how it can be blended with current wearables. In summary, to optimize comfort and enhance the potential for integration with commercial smartwatches and wristbands, maintaining a minimally-obtrusive design that sits close to the skin is important. This includes using small and non-skin-contacting sensors and keeping the system small and low-profile.

(2) Low-Power: Wearable devices are generally worn for extended periods. Moreover, in the case of continuous tracking, the devices need to remain operational throughout. Therefore, it is crucial to consider power consumption during design.

Our goal was to find the balance between these design considerations and the sensing performance. To achieve the goal, the sensor type, number, and layout were examined in the studies.

3.1 Speaker Type and Position

An experimental prototype based on Teensy 4.1 [1] with 3D printed sensor mounts was built to explore sensor configurations (Fig. 2 (a)). The prototype allowed us to manipulate the sensor layout easily. We connected 2 speakers and 8 microphones to the prototype.

We started with ultrasonic transducers as the signal emitter (Fig. 2 (b)) for their excellent ultrasonic acoustic characteristics. We first tried to identify the optimal position for the speaker. We performed a grid search on the speaker position while doing single-finger movements (bending one of the five fingers at a time). One of the researchers collected the data and trained the model as described in Sec. 4.4.2, and the reconstruction mean joint Euclidean distance error (MJEDE) was calculated according to the method outlined in Sec. 4.5. Results (Fig. 3 (a)) indicate that speaker positions under the

[1] https://www.pjrc.com/store/teensy41.html

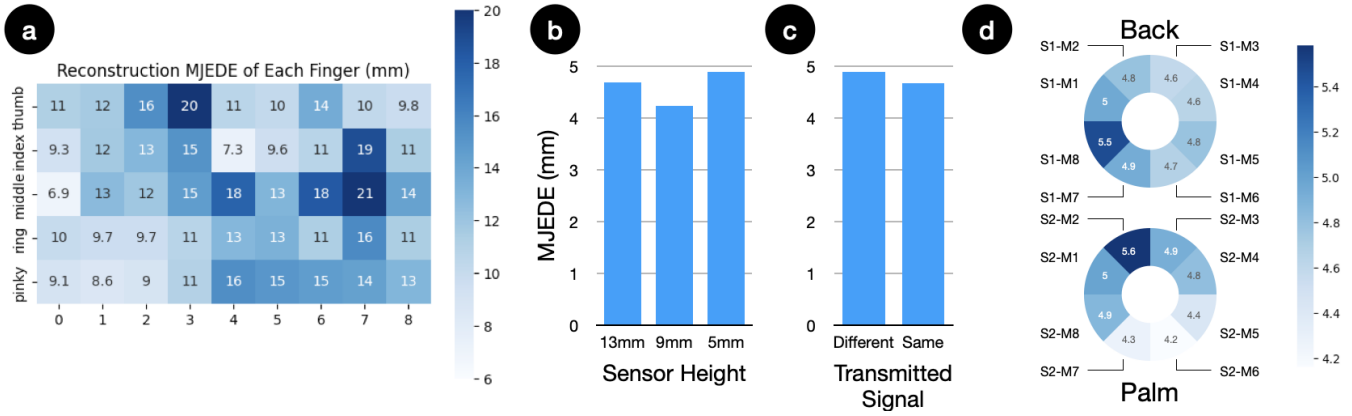

Figure 3: We explored the performance of (a) each finger in reconstruction mean joint Euclidean distance error (MJEDE) when the speaker was placed at different positions. Note that the numbers on the x-axis represent positions evenly distributed around the wrist. The center of the palm side is represented as 0, the one to its right as 1, and so forth. The MJEDE, while adjusting (b) the height of the sensors, (c) the transmitted signal on the speakers, and (d) the combination of speaker and microphone, were also examined. The number and position of the speakers and microphones can be referred to Fig. 2 (a).

palm yield the best performance. To maximize the information we could get with an additional speaker, we placed a second speaker in its opposite position above the back of the hand to obtain more diverse information from both sides of the hand.

In these experiments, we also realized that these transducers had strong directionality, meaning that the emitted sound waves were mostly directed to the fingers directly facing the transducer. These fingers performed well, while other fingers did not work well. This inspired us to choose omnidirectional speakers so that sound waves could travel in all directions, allowing us to place speakers in less-obtrusive positions while still covering most fingers. We ended up using a commodity speaker, which came in a smaller size as well (Fig. 2 (c)).

3.2 Sensor Height

Next, we conducted a quantitative study on the sensor height. We chose three heights, placing the sensors 13 mm, 9 mm, and 5 mm from the skin (measured from the skin surface to the outermost point on the sensor). With the last configuration, the lower edge of the speaker almost touched the skin since the speaker's diameter was 5 mm. Two researchers conducted the experiments, doing a series of pre-defined single-finger movements and complex finger gestures, respectively. The average performance of the three speaker heights was very close (Fig. 3 (b)). Since no strong decrease in performance was observed, we chose a 5 mm sensor height for the final prototype. This allows EchoWrist to adopt a low-profile form factor that can be easily integrated into commercial smartwatches or wristbands.

3.3 Number of Sensors

We then moved to optimize the number of sensors. We separated the emitted signals from the two speakers with different frequency ranges (18-21 kHz and 21.5-24.5 kHz). This way, we could easily separate paths from the two speakers on each microphone. The MJEDE was calculated using data collected from a single pair of a speaker and a microphone. The layout of 2 speakers and 8 microphones is specified in Fig. 2 (a), where S1 was placed on the back of the hand, and S2 was under the palm. Two researchers experimented with using different channels and their combinations. When only using one of the 16 channels, results (Fig. 3 (d)) indicated that microphones close to the speaker yielded slightly better results, but other positions also worked decently. We attributed this to the use of omnidirectional speakers combined with the use of echo profiles to preserve information maximally. Specifically, having the back speakers improved performance when the hand was bent outwards. We placed speakers at both the hand's back and palm sides to maintain reliable performance when the wrist was at different angles. To maximize the benefit of using 2 speakers, we decided to use 2 microphones that are close to the speakers. Compared with having more sensors, the combination of 2 speakers and 2 microphones can be easily achieved as digital audio interfaces such as Inter-IC Sound (I^2S) usually come with stereo audio.

We noticed that the cross-path signals (signal traveling from the palm speaker to the back microphone and vice versa) were ignorable compared with direct-path signals. Another experiment on using the same frequency on the two speakers versus using different frequencies (Fig. 3 (c)) confirmed that no significant performance drop could be observed. Allowing the two speakers to send the same signal can save half of the bandwidth, which can save hardware cost and size by using mono-channel audio amplifiers instead of stereo ones. Therefore, we decided to emit the same signal on the two speakers in our final prototype.

Overall, the final prototype of EchoWrist features two pairs of speakers and microphones strategically positioned on the top and bottom sides of the wrist, respectively. The inclusion of these two pairs of sensors allows for the comprehensive capture of echoes from both sides, thereby providing a wealth of information about hand gestures and interactions.

Figure 4: Hardware of EchoWrist. (a) (b) Wearing EchoWrist at the wrist. All components are mounted on a silicone wristband. (c) Customized PCBs for the microcontroller module and the sensing module. (1) US Quater Dollar coin. (2) Customized PCB with SGW1110 module (front and back views). (3) 3.7 V 70 mAh LiPo battery. (4) Sensor module with speaker and microphone.

4 IMPLEMENTATION

4.1 Hardware Implementation

EchoWrist uses two pairs of speakers (OWR-05049T-38D) and microphones (ICS-43434) mounted on a low-profile silicone band (Fig. 4 (a), (b)). We designed customized printed circuit boards (PCBs) for the sensing module (Fig. 4 (c)). The two sensing modules are connected via a flexible printed circuits (FPC) cable and then connected to the customized microcontroller module. The microcontroller board includes an SGW1110 module (with nRF52840 microcontroller), two MAX98357A audio amplifiers (in the study, only one was used), plus a power management module with a TPS62743 buck regulator. The entire system is powered by a LiPo battery. The sensing and microcontroller modules are attached to 3D-printed cases that can slide along the silicone band to fit different hand sizes. The cases are printed with thermoplastic polyurethane (TPU) so that they are soft and comfortable. The weights of the sensing and microcontroller modules are 0.7 g and 1.2 g, respectively. The weight of the entire system, including the battery, is 16.8 g. While collecting data, the microcontroller drives the speakers to emit sound waves and collects echoes from the microphones. The collected data can be saved on the microSD card with an extended socket. To support real-time data collection, the collected data can also be transmitted to a smartphone (Xiaomi Redmi Note 10 Pro) via Bluetooth low-energy (BLE) operating at 800 kbps. The captured data are truncated to 8 bits to save bandwidth in this case. No performance degradation was observed between using full 16 bits and truncated 8 bits.

4.2 Power Signature

Designed for compact wearable devices such as smartwatches and wristbands, EchoWrist aims to provide a low-power solution. We examine the power signature of EchoWrist using a CurrentRanger [2]. Results show that when the system is on with BLE transmitting data at 800 kbps, the power consumption is 57.9 mW (3.86 V, 15.0 mA).

At the 57.9 mW operation power, EchoWrist can easily last a full day with a common battery size of smartwatches (e.g., Apple Watch Series 8 has around 300 mAh battery size [3], which should last 19 hours). If EchoWrist is integrated into the existing hardware of the smartwatch/wristband, the microcontroller's base power consumption could be saved since the sensors only operate at 10 mW, leading to an even longer battery life. Note that this calculation do not consider the power consumption of the operating system on a smartwatch.

4.3 3D Hand Pose Ground Truth Acquisition

We used MediaPipe [41], which has been widely used in prior projects [9, 24], to acquire the ground truth of 3D hand shapes. With MediaPipe, the shapes of the hand are represented by 21 joints, including the wrist (Fig. 5 (a)). Each joint is represented by 3D coordinates. While recovering the hand shapes, we predict the coordinates of the 20 joints, excluding the wrist, which is set as the origin. While recovering the wrist rotation, we predict the wrist-to-palm vector as illustrated in Fig. 5 (b).

4.3.1 Ground Truth Normalization. For the data collection of 3D hand pose tracking, we asked participants to remount the device between sessions and allowed them to relax and move around. This caused relative position and orientation variances between the hand and the camera. In addition, MediaPipe may not capture the size of the hands reliably due to a lack of objects to compare. To fix these issues, we developed an algorithm to normalize the ground truth. For each frame, we first used MediaPipe to extract the positions of all hand joints. We then calculated the surface of the palm represented by vectors $(\vec{v_5}, \vec{v_{17}})$, where $\vec{v_5}$ is the vector pointing from the wrist to the metacarpophalangeal joint of the index finger (joint 5 in Fig. 5 (b)) and $\vec{v_{17}}$ is the vector pointing from

[2] https://lowpowerlab.com/guide/currentranger/

[3] https://www.xda-developers.com/apple-watch-series-8-and-ultra-battery-size/

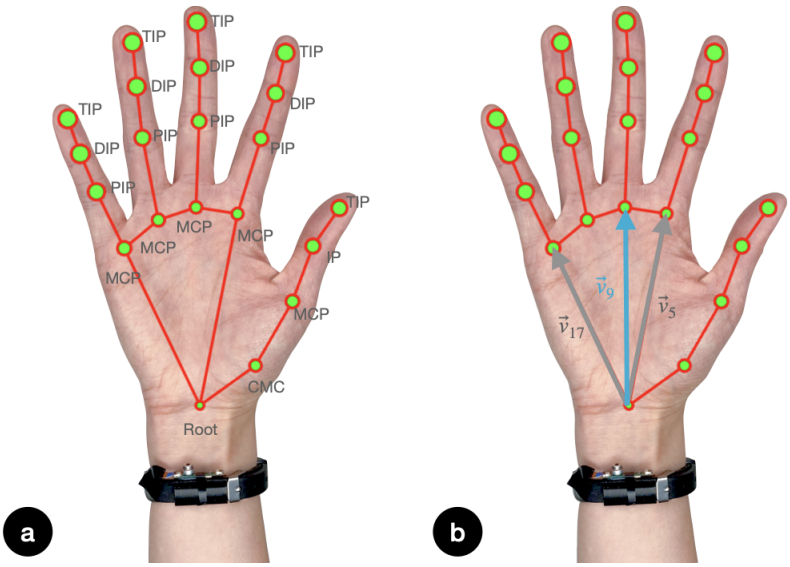

Figure 5: 3D hand pose ground truth annotation. (a) 21 joints detected by MediaPipe. (b) Important vectors used during ground truth normalization. $\vec{v_9}$ is the wrist-to-palm vector used to represent the orientation of the hand. The plane defined by $\vec{v_5}$ and $\vec{v_{17}}$ is used to align the detected hand with the reference hand. The actual length of $\vec{v_{17}}$ is measured in the image against the size of the sensing module to uniform the hand size in each frame.

the wrist to the metacarpophalangeal joint of the little finger (joint 17 in Fig. 5 (b)). We calculated the rotation matrix that rotated this plane to the reference plane $(\hat{\vec{v_5}}, \hat{\vec{v_{17}}})$ extracted from the reference posture (Fig. 5(b)). We chose this plane because it remained largely static even when the user was performing complex gestures. For each participant, we also measured the length of $\vec{v_{17}}$ as against the size of the sensing module. We normalized each frame so that the length of $\vec{v_{17}}$ was equal to the measured value (Fig. 5 (b)). While recovering the wrist rotation, ground truth normalization was not applied so that the wrist rotations were faithfully recorded.

4.4 Data Processing and Deep Learning Pipelines

4.4.1 Echo Profile Analysis. We employed the Frequency Modulated Continuous Wave (FMCW) echo profile analysis as the sensing method, which has demonstrated promise in previous works [35, 77, 84]. With a sampling rate of 50 kHz, we used a frequency range of 18-21 kHz, which was chosen to be inaudible for humans and later raised to 20-24 kHz to further reduce audibleness. The signals of one frequency sweep are denoted as one *FMCW frame*. To eliminate other frequencies, we applied an 18-21 kHz (or 20-24 kHz) bandpass filter. Subsequently, cross-correlation between transmitted and received signals was performed to obtain *echo profiles*, which are formed by temporally stacking *echo frames* and represent the reflection strength of signals traveling from paths with certain distances. Using the current echo profile to subtract the previous one produces *differential echo profiles*, which eliminate static reflections and focus on moving objects. Fig 7 illustrates various hand gestures and their corresponding echo profiles. Both original and differential echo profiles were employed to capture both movements and the static status of the hand and surroundings. The echo profiles were cropped to concentrate on distances of interest. For hand tracking, we used 72 pixels (24.6 cm) to focus on the hand, and for hand-object interactions, we used 88 pixels (30.2 cm) to encompass the hand's surroundings.

4.4.2 Deep Learning Inference. After obtaining the echo profiles, the information related to the hand postures is represented by a 2D image-like array. Due to its wide application and success in image processing, we used a customized deep Convolutional Neural Network (CNN) model to infer the 3D hand shapes or hand-object interactions. We stacked the original and differential echo profiles in channels. For hand pose tracking, we employed a shorter window length of 72 (0.864s) to predict the hand shapes at the last moment of the window. For hand-object interactions, we incorporated a greater temporal context by employing a longer window length of 1050 (12.6s) since the activities are more intricate. In our pilot study with the researchers, we observed significant variability in the time taken to complete each hand-object interaction. To accommodate this variability and prevent information loss, a larger window was utilized. This resulted in input sizes of $72 \times 72 \times 4$ and $1050 \times 88 \times 4$, respectively.

The model architecture incorporates a ResNet-18 [17] backbone, followed by an average pooling layer, a dropout layer (with a dropout rate of 0.8), and a fully connected layer. In the task of recovering the 3D hand shapes, the output shape is 60, corresponding to 20 3D coordinates. The wrist is not included in this output as it serves as the origin. In the case of wrist rotation recovery, the output shape is 3, representing the wrist-to-palm vector. To prioritize the joints with larger errors, the model employs Mean Squared Error (MSE) loss. During the training process, Adam optimizer is

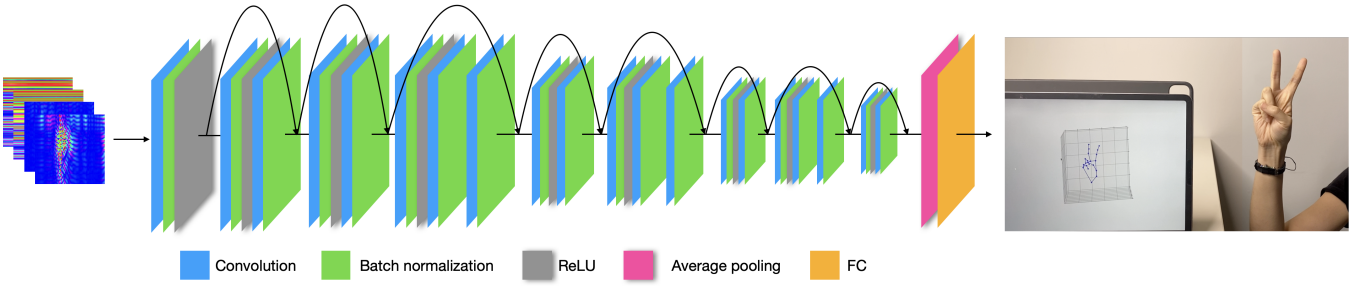

Figure 6: The model architecture of EchoWrist.

utilized, with an initial learning rate set to 0.0002. The batch size for training is set to 30.

In the classification task of hand-object interactions, a similar architecture is employed, although the final layer of the CNN is replaced with a linear classifier. For this task, the output shape is 12, signifying the number of classes, and Cross-Entropy (CE) loss is employed to facilitate accurate classification.

4.4.3 Data Augmentation. We applied data augmentation to improve the robustness of the model. During training, the echo profiles were randomly vertically shifted by ±11 pixels. This was applied to compensate for the variance caused after remounting, where the position of the wristband may shift vertically. In 80% of cases during training, each pixel in the echo profiles was multiplied by a random factor between 0.95 and 1.05. This is applied to avoid overfitting to a fixed set of values after multiple epochs.

4.5 Evaluation Metrics

While the assessment of Simple Gestures and Complex Gestures (Sec. 6) primarily centered on joint movements, the evaluation of Wrist Orientations (Sec. 6) specifically focused on wrist rotation. As a result, we employed different metrics for the two tasks:

4.5.1 3D Hand Pose Estimation. To gauge the precision of our 3D hand pose estimation, we employed two established metrics: mean joint Euclidean distance error (MJEDE) and mean joint angular error (MJAE). These metrics have been utilized in prior works [18, 24, 25, 39, 88].

The MJEDE of each frame is calculated by averaging the Euclidean distance of 20 joints (excluding the wrist). The MJAE is calculated by averaging the angular error of the 15 joint angles (excluding five fingertips). Each joint angle is the angle between two consecutive bone segments.

In addition to MJEDE and MJAE, we also present the error distribution and the error of each joint or joint segment.

4.5.2 Wrist Rotation Estimation. To evaluate the performance of EchoWrist on wrist rotation estimation, we use the mean wrist angular error (MWAE) as the evaluation metric, which is calculated by the angular error of the wrist-to-palm vector. In this context, the wrist-to-palm vector v_9 is defined as the vector pointing from the wrist to the metacarpophalangeal joint of the middle finger (Fig. 5 (b)). Similarly, we reported the error distribution.

5 EVALUATION OVERVIEW

We intend to build a wristband that can not only understand the hands themselves but also the objects that they interact with. For this purpose, we designed two studies to assess the feasibility of using EchoWrist in continuous 3D hand pose tracking and hand-object interaction recognition, respectively. The studies were approved by the Institutional Review Board (IRB) of Cornell University.

With the first study, we designed three sets of hand gestures to demonstrate that EchoWrist is able to continuously recover the 3D hand poses when there is no object in the hand. With the second study, we incorporated 12 activities to demonstrate that beyond the tracking of free-hand poses, EchoWrist possesses the ability to comprehend hand-object interactions. We introduce the details of study design, procedures, and performance in the following sections.

6 USER STUDY 1 - HAND POSE TRACKING

The objective of this first study was to evaluate the performance of EchoWrist in tracking 3D hand poses under different conditions. First, the study analyzed the device's accuracy in tracking different finger gestures across multiple sessions after the device was remounted. Second, the study assessed the effectiveness of the results obtained from training on a different number of sessions of data from each participant.

To explore the efficacy of hand pose tracking with different levels of complexity, three gesture sets were used in the study. The first two sets were designed to confirm the accuracy of 3D hand pose tracking, while the third set was specifically designed to validate the accuracy of wrist rotation estimation:

(1) Simple Gestures: This set of gestures consisted of five single-finger movements, wherein only one finger would bend at a time (Fig. 7 (a)). The purpose of these gestures was to assess the precision of EchoWrist in tracking and distinguishing between distinct finger postures.

(2) Complex Gestures: This set of gestures included ten complex finger gestures, modeled after the American Sign Language (ASL) finger gestures representing the digits 0-9 (Fig. 7 (b)). Unlike Simple Gestures, these gestures involve fingers occluding each other, introducing a higher level of complexity for tracking and recognition. Notably, this particular set of gestures was utilized [24], explicitly selected to challenge EchoWrist's capability to accurately track more complicated motions with multiple finger movements and occlusions.

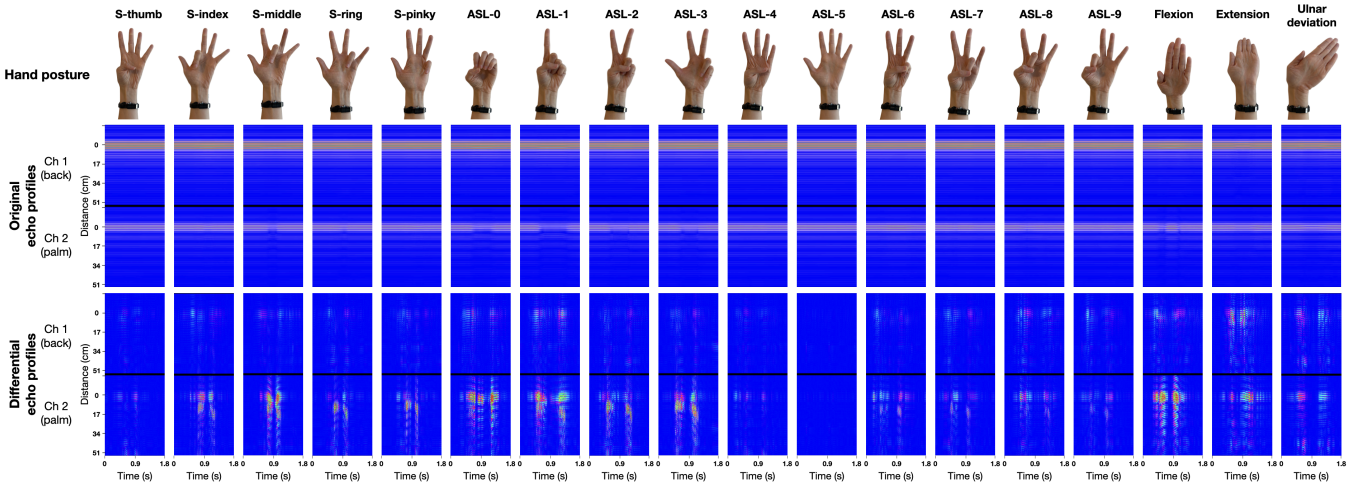

Figure 7: Illustration of gestures used in hand pose tracking user study and their echo profiles. 18 gestures in 3 categories: simple gestures (gestures starting with S-), complex gestures (10 American Sign Language gestures, starting with ASL-), and 3 wrist rotations (the last three columns).

(3) Wrist Orientations: This set of gestures comprised three distinct wrist orientations: flexion, extension, and ulnar deviation (Fig. 7 (c)). It is noteworthy that radial deviation was excluded due to its reported difficulty in the pilot study. These wrist motions are essential in everyday hand gestures. As EchoWrist is mounted on the wrist, it has the unique advantage of observing the hand from that vantage point, making it possible to track wrist orientations accurately.

6.1 Participants

We recruited 12 participants (4 self-reported males, 8 females) aged 20-26 (M = 22.0, SD = 1.5) with snowball sampling at a local university. Ten participants self-identified as right-handed, two left-handed. However, all participants reported wearing or intending to wear watches or wristbands on their left wrists. Consequently, all participants wore the device on their left wrist during the study. After the study, the participants were compensated 15 USD.

6.2 Apparatus

The study was conducted in a quiet experiment room. Participants were seated at a table with a laptop positioned in front of them, serving as a platform for presenting instructions. For capturing the ground truth data, a webcam [4] was employed to record the movements of the participant's hand. These recorded videos of hand postures were processed by MediaPipe [41] to extract the ground truth represented by the 3D positions of 21 finger joints. As previously mentioned, all participants wore the device on their left wrist. A sticker was used to mark the precise position for ease of remounting. As the wrist thickness varied among participants, the device was customized by adjusting the distance between the two pairs of speakers and microphones.

[4]Logitech C615 https://www.logitech.com/en-us/products/webcams/c615-webcam.960-000733.html

During the user study, we streamed the acoustic data from the two microphones to a smartphone (Redmi Note 10 Pro) via BLE at 800 kbps. However, due to heavy traffic, we occasionally experienced packet loss during transmission. In such cases, the lost packets were replaced with zeros, and the data containing the lost periods was removed from the dataset. Throughout the entire study, we experienced only a 0.35% packet loss in BLE.

6.3 Procedure

The procedure followed in the user study was as follows:

(1) Introduction: The study started with participants signing a consent form and receiving an introduction to the study's procedure.

(2) Practice Session: The study was divided into 21 data collection sessions, each following the same process. The initial session served as a practice session, allowing the participants to get familiarized with the testing system and the act of performing gestures. Note that the data from the practice session was used for neither training nor testing later. The participants were not informed of this distinction and treated the practice session as any formal session.

(3) Data Synchronization: During each session, the researcher would snap their fingers in front of the camera to mark the beginning and end of the session. The sound of the snap was captured by the microphones on the device, while the snapping gesture was recorded by the camera. This allowed for the synchronization of the acoustic data and the ground truth video footage.

(4) Data Collection: In each session, the aforementioned three gesture sets were presented in the following order: (1) Simple Gestures, (2) Complex Gestures, and (3) Wrist Orientations. Within each gesture set, each gesture, lasting 2 seconds, was repeated four times. The sequence of gestures was randomized to mitigate potential learning effects.

We created a user-friendly graphical interface to present instructions and streamline the user study process. The interface was

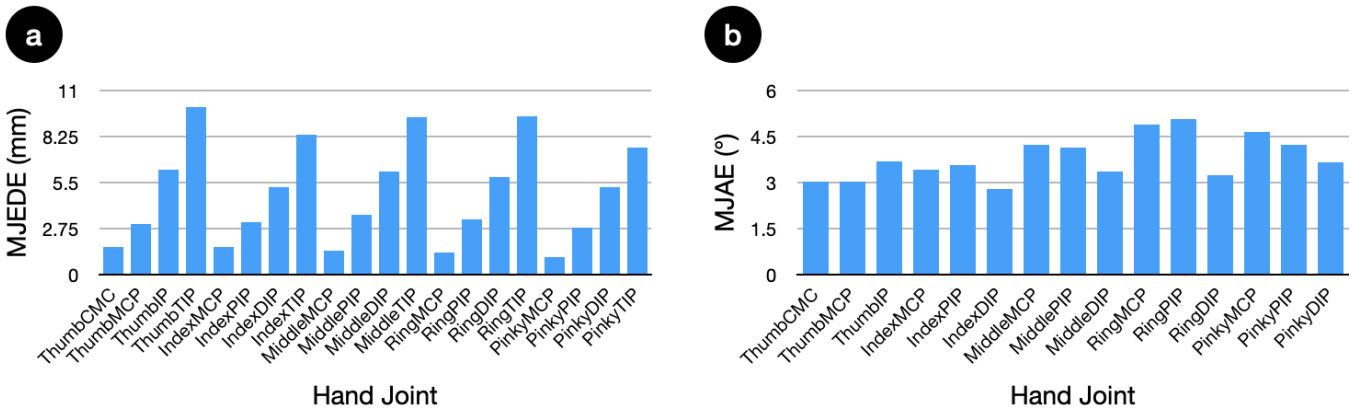

Figure 8: 3D hand shape reconstruction performance on all hand joints. (a) Using MJEDE metric. (b) Using MJAE metric.

displayed on the laptop screen. At the beginning of each trial, an image representing the target gesture appeared, and the participant had 2 seconds to replicate the gesture. The interface also featured a countdown timer for the ongoing gesture and displayed the number of remaining trials. The participants were explicitly instructed to mimic the gesture displayed on the screen and return to a neutral position upon completing each gesture.

(5) Device Remounting: After each session, the participant was asked to take a short break and then remount the device before the start of the next session. Participants had the freedom to relax their arms or move around during these breaks. This remounting procedure was designed to evaluate the performance of our system in real-world scenarios where users frequently remove and reattach the device.

Each session lasted approximately 2.5 minutes. Accounting for the intervals between sessions, the entire study extended for a duration of 75 to 90 minutes for each participant. In total, each participant contributed 1440 (= (5 (Simple Gestures) + 10 (Complex Gestures) + 3 (Wrist Orientations)) × 4 (repetitions) × 20 (sessions excluding the practice session)) gestures to the dataset, which means that 17280 gestures were collected.

It is worth mentioning that some participants occasionally performed incorrect gestures that differed from the target gestures. However, we decided to include this data in our dataset since the video ground truth faithfully captured the actual gestures performed by the participants. Additionally, almost all participants experienced different physical limitations that made it challenging for them to execute certain hand gestures, such as being unable to bend certain fingers without bending others. In such cases, the participants were instructed to mimic the gestures in a way that was comfortable to them. As a result, our dataset contains many non-standard instances of gestures.

6.4 Training Scheme

To minimize training effort from a new user, we seek to maximize the utilization of data collected from other users. Therefore, we developed a two-step training-fine-tuning scheme. In the first step, we trained a model using data collected from other users. For each new user, we fine-tune the pre-trained model with only a small set of training data collected on the new user. Compared with training a user-dependent model from scratch directly on data provided by this new user, the model not only converges faster but also yields better performance.

For the evaluation, we first trained a leave-one-participant-out (LOPO) model for each participant. Please note that this is a user-independent (UI) model. The UI model was trained for 10 epochs. We then fine-tune this UI model with different amounts of data collected from the same participant for another 5 epochs. The same learning rate (0.0002) was used in both steps.

6.5 Results - 3D Hand Pose Estimation

Following the two-step training scheme, for each participant, we tested the fine-tuned model on the last two sessions of data. When we used the first 18 sessions (i.e., all sessions except the testing ones) as training data to fine-tune the LOPO model, our results yielded an MJEDE of 4.81 mm (SD = 0.99 mm) and MJAE of 3.79°(SD = 0.68°) across all participants.

In general, fingertips exhibited the largest error, with an average of 9.0 mm, across all finger joints. Among the five fingertips, the thumb's fingertip had the highest error, measuring 10.0 mm. However, this discrepancy was not significantly larger than that of other fingertips. The joint angle at the proximal joint of the ring finger has the largest angular error of 5.1°(Fig. 8). Specifically, although fingertips tend to have larger Euclidean distance errors, the angular errors of all joints are similar. This indicates that the larger error on the tips of the fingers largely comes from accumulated error from connected joints. The error distribution of MJEDE of the finger TIPs, DIPs, PIPs, and MCPs are illustrated in Fig. 9 (a). Performance on Simple Gestures and Complex Gestures are similar: the MJEDE is 4.69 mm (SD = 1.17 mm) and 4.87 mm (SD = 0.94 mm), while the MJAE is 3.64°(SD = 0.75°) and 3.87°(SD = 0.67°), respectively. This shows that EchoWrist works consistently in estimating the hand poses while performing hand poses with different complexities.

The results from the above experiments showed promising tracking performance. However, it required 18 sessions of training data from each participant, which took 36 minutes to collect. This long

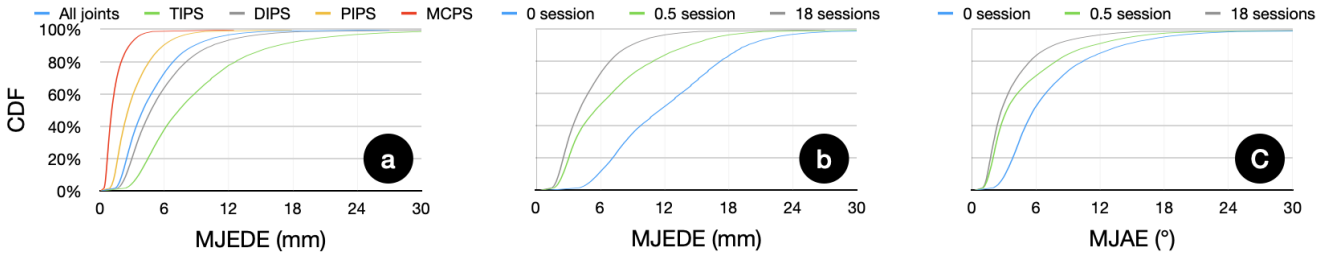

Figure 9: Error distribution in 3D hand shape reconstruction. (a) The error distribution of MJEDE of different joint types. (b) The error distribution of MJEDE of all joints when 0, 0.5, 18 sessions of data was used to fine-tune the UI model. (c) The error distribution of MJAE of all joints when 0, 0.5, 18 sessions of data was used to fine-tune the UI model.

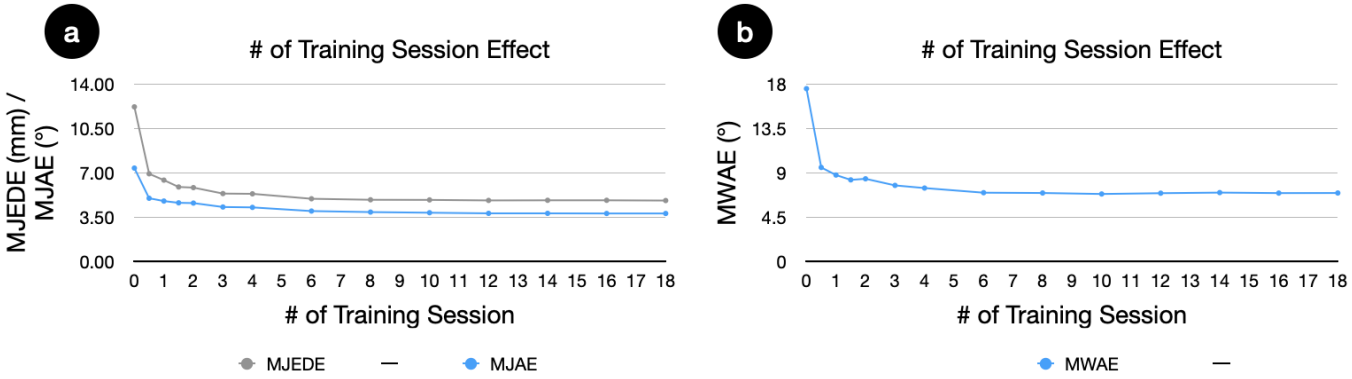

Figure 10: Adjusting the number of sessions used during the fine-tuning step. (a) Performance in 3D hand shape reconstruction. (b) Performance in wrist rotation recovery.

training period might hold back users' acceptance. The high demand for training data has been a long-lasting problem for the data-driven sensing approach.

Therefore, in the follow-up experiment, we examined how much training data is actually needed for a new participant without significantly compromising the performance. To do this, we manipulated the number of sessions used during the fine-tuning stage. When no data from the same participant is used, the MJEDE and MJAE across all participants are 12.2 mm (SD = 3.73mm) and 7.37°(SD = 1.73°), respectively. Note that this is the performance in the user-independent experiment of EchoWrist. When only using 0.5 sessions (about 1 minute) of data, the performance improved to MJEDE 6.92 mm (SD = 2.42 mm) and MJAE 4.99°(SD = 1.28°)(Fig. 10 (a)). The error distribution when 0, 0.5, and 18 sessions were used for fine-tuning is presented in Fig. 9 (b, c). As the figure shows, the performance improves with more training data. However, we noticed the performance flattened at 8 sessions of training data (20 minutes). This indicates that a new user only needs to provide 20 minutes of training data to obtain optimized performance. Even with only 1 minute of training data, EchoWrist can still achieve decent performance. The results from this study were very encouraging, which shows the potential of using the proposed sensing technology for real-world users with minimal calibration from each new user.

6.6 Results - Wrist Rotation Estimation

We follow the same approach as described in Sec 6.5 to evaluate EchoWrist's performance in estimating wrist rotation. When using 18 sessions for fine-tuning, the mean wrist angular error (MWAE) across all participants is 6.95°(SD = 2.23°). The error distribution of MWAE is demonstrated in Fig. 11 (b). By adjusting the number of sessions used during fine-tuning, we obtained the curve shown in Fig. 10 (b). In the User-Independent (UI) setting, where no training data from the same participant was used, the performance is 17.5°(SD = 5.75°). With only 0.5 sessions (72 seconds) of training data, EchoWrist still achieves a performance of 9.54°. As a special note on P02, whose wrist rotation estimation performance was significantly worse than others (13.4°compared to an average of 6.95°), we analyzed the recorded video and found that the ground truth of P02's hand pose was neither stable nor accurate. P02 performed the flexion gesture in close proximity to the camera with a large angle, which caused the camera to easily lose focus and MediaPipe to struggle with capturing the hand reliably due to the blurry images and skewed view angles. In later studies, the camera was placed further away to enable participants to perform gestures without significantly impacting image quality.

6.7 Results - Noise Injection

EchoWrist uses active acoustic sensing as the sensing principle. The frequency range of the signals used during the study is 18-21

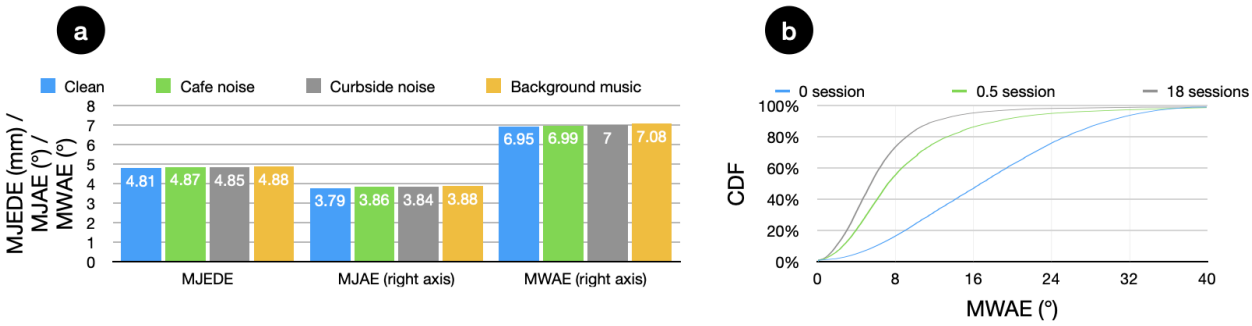

Figure 11: (a) Performance of EchoWrist with different injected noises. (b) Error distribution when 0, 0.5, 18 sessions from the same user was used to fine-tune the UI model.

Table 1: Noises recorded in different scenarios.

Scenario	Cafe	Curbside	Background Music Playing
Noise Level (dB (A)) [5]	61.8	71.5	70.4

kHz, which is well beyond the range of human conversations and most noises. However, in order to examine EchoWrist's robustness against various acoustic noises existing in the real world, we recorded noises in three different scenarios as specified in Table 1 and injected the noises into the testing data. Please note that the model was trained using the training data without noise injection. In this setting, we can evaluate how our system would react to different noises without the need to collect training data for each noise.

Results (Fig. 11) indicate that there is nearly no performance degradation when different types of noises are injected. The maximum performance drop is on background music playing, where MJEDE increases from 4.81 mm to 4.88 mm. We performed one-way ANOVA tests on all three noises and did not find a significant difference between performance in the clean environment and with different noise injections: for 3D hand pose estimation, $F(3, 44) = 0.010$, $p = 1.00 > 0.05$; for wrist rotation estimation, $F(3, 44) = 0.0066$, $p = 1.00 > 0.05$.

6.8 Follow-Up Study - Hand Motion Speed

As EchoWrist relies on both original and differential echo profiles, we want to investigate whether the hand motion speed affects the tracking performance. To address this, we conducted a follow-up study, collecting data from participants executing the same gestures at different speeds.

6.8.1 Participants. We recruited another 12 participants (5 self-reported males, 7 females) aged 18-31 (M = 24.1, SD = 4.8) using snowball sampling at a local university. Eleven participants self-identified as right-handed, one left-handed. To align with the initial study, we asked all the participants to wear the device on their left wrist during the study. After the study, the participants were compensated 20 USD.

6.8.2 Apparatus. The apparatus is closely identical to the initial study, with two differences. Firstly, a laptop's built-in camera [6] was employed instead of the webcam. Secondly, the data was stored on a microSD card to prevent packet loss. This change in data storage is to avoid data loss since we collect less data for each condition in this study.

6.8.3 Procedure. The procedure closely mirrors that of the initial study, involving 24 data collection sessions, each adhering to the same protocol. The participants were instructed to perform the gestures at varying speeds during each session. Three different speeds were implemented: (1) Fast (1.5 s/gesture), (2) Medium (2.0 s/gesture), and (3) Slow (2.5 s/gesture). Instead of images, videos showing a hand executing the gestures at the specified speed were shown to guide the participants, who were instructed to try their best to follow the movement/speed of the videos. The initial three sessions, one for each speed, were designated as practice sessions, and the data collected during this period was excluded from both training and testing later. The sequence of speeds was randomized for the remaining 21 sessions.

The entire study took 90 minutes for each participant. In total, each participant contributed 504 (= 18 (gestures) × 4 (repetitions) × 7 (sessions excluding the practice session)) gestures to the dataset for each of the three speeds, resulting in 3 datasets, each containing 6048 gestures. Note that one of the participants had an emergency during the study, so they completed the last five sessions the next day. In addition, due to a hardware issue, some data were lost for three participants. To address this, two participants were invited to redo one session on the other day, while the other one redid two. Each participant was compensated an additional $15 in local currency.

6.8.4 Training Scheme. The training scheme is the same as the initial study. A model was first trained for each participant at the Medium speed, consistent with our initial study, where each gesture was performed in a two-second interval. The User-independent (UI)

[6] MacBook Pro 14-inch, 2021 https://www.apple.com/macbook-pro/

model underwent training for 20 epochs. Subsequently, we fine-tuned the UI model using the data from a specific participant at the Medium speed for an additional 10 epochs, with a consistent learning rate of 0.0002. 6 sessions were utilized as training data for fine-tuning, while the remaining session served as the testing data. Following the fine-tuning on the Medium speed model, we tested the model using the 7 sessions of data at both the Fast and Slow speeds from the same participant, respectively.

6.8.5 Results. When we tested the model with the same, i.e., the Medium, speed, the average MJEDE is 11.01 mm (SD = 4.99 mm), and MJAE is 7.38°(SD = 1.89°). For testing the model with the Fast speed data, the average MJEDE is 12.26 mm (SD = 7.80 mm), and MJAE is 7.80°(SD = 2.09°). In the case of the Slow speed data, the average MJEDE is 13.47 mm (SD = 8.27 mm), and MJAE is 8.54°(SD = 2.50°). Note that the dataset used in this study was only one-third of the initial dataset, explaining the difference in performance compared to the initial study. In addition, the use of video instructions, as opposed to image instructions, led to shorter reaction times and posed challenges for some participants in following the instructions accurately. Despite instructing participants to complete the incorrect gestures, many attempted corrections, introducing more unseen poses and reducing the number of target poses in the dataset. Notably, one participant had a significantly larger hand, with the longest part measuring 22.8 cm, compared to the average of 18.2 cm for other participants. This variation contributed to unexpected outcomes. Nevertheless, the results of this study remain comparable to prior work [9, 24].

In this study, we aimed to investigate the impact of motion speed on the performance of our hand pose tracking system. We initiated our analysis with a one-way ANOVA. In the case of MJEDE, the results did not indicate a significant difference ($F_{2,33} = 0.37, p = 0.69$) when testing the Medium-speed model on datasets with Fast, Medium, or Slow speeds. This suggests that the performance of our system, as measured by MJEDE, remains consistent across different motion speeds. Similarly, when considering MJAE, the results did not reveal a significant difference ($F_{2,33} = 0.86, p = 0.43$). This finding suggests that there is consistency in both MJEDE and MJAE across different motion speeds. In summary, our comprehensive statistical analyses reinforce the overall robustness of our system, highlighting its ability to maintain consistent performance across different motion speeds.

7 USER STUDY 2 - HAND-OBJECT INTERACTION RECOGNITION

In the second user study, our primary objective was to evaluate EchoWrist's capability to recognize everyday hand activities within a naturalistic environment. Owing to the active acoustic sensing techniques, EchoWrist can effectively recognize both *Static Interactions*, such as holding objects in hand, and *Dynamic Interactions*, such as moving objects. As a result, an interaction set, with half of them being static interactions and the remaining half being dynamic interactions, was used to comprehensively assess the device's performance in capturing these real-world activities.

On the other hand, although humans engage in extensive hand-object interactions across diverse contexts in our daily lives, we have confined our study's context to the kitchen for a proof of concept. This decision stems from the fact that the kitchen setting is a frequently chosen dataset scenario in prior research [4–6, 62, 89]. It can be applied to support a wide range of applications, including but not limited to elder care, smart home technology, and accessibility solutions.

As a result, an interaction set consisting of 12 hand-object interactions specifically tailored to the kitchen environment was proposed. This included 6 *Static Interactions*: **holding a paper cup**, **a pair of chopsticks**, **a glass water bottle**, **a pot**, **a pan**, and **a kettle**, and 6 *dynamic interactions*: **drinking**, **stirring**, **peeling**, **twisting**, **chopping**, and **pouring**.

7.1 Participants

We recruited 12 participants (9 females, all right-handed) aged 21-33 (M = 27.0, SD = 3.7, one preferred not to state their age) with snowball sampling at a local university. The participants were compensated 20 USD.

7.2 Apparatus

Different from the first study, we conducted this study mimicking real-life settings. Therefore, the participants conducted this study at one researcher's home (a 2 bedroom 2.5 bathroom townhouse), where their roommate continued with their usual activities, including generating background noise. The study took place in the kitchen area, which adjoined the living room. Participants moved around the kitchen to complete various tasks while the researcher oversaw the study in the living room. All objects, except for paper cups, were what the researcher regularly used at home and were in their usual place. Disposable paper cups were used for drinking to maintain hygiene standards. In addition, the participants had the freedom to engage in conversations with the researcher during the study.

In this study, the participants had to interact with different objects. As this interaction was mostly done through the dominant hand, they were requested to wear EchoWrist on their dominant arm (all the participants were right-handed) at the relative position where they usually wear a watch. To facilitate the remounting process, a sticker was used to mark the position. The device was customized for each participant based on the thickness of their wrist by adjusting the length between the two pairs of speakers and microphones. Additionally, the participants were asked to wear earbuds to allow for the use of voice commands throughout the study. We saved the data into a microSD card to avoid unstable BLE data package loss in a more complex electromagnetic environment.

7.3 Procedure

The study shared similar procedures with the following differences:

(1) Introduction: A demonstration of the 12 interactions and the environment configuration was included. Each participant was asked to go through the 12 interactions independently to confirm their full comprehension of the procedure and the setup.

(2) Data Collection: The study consisted of 5 data collection sessions, all of which followed the same process. Given that all the interactions featured in our interaction set were everyday activities

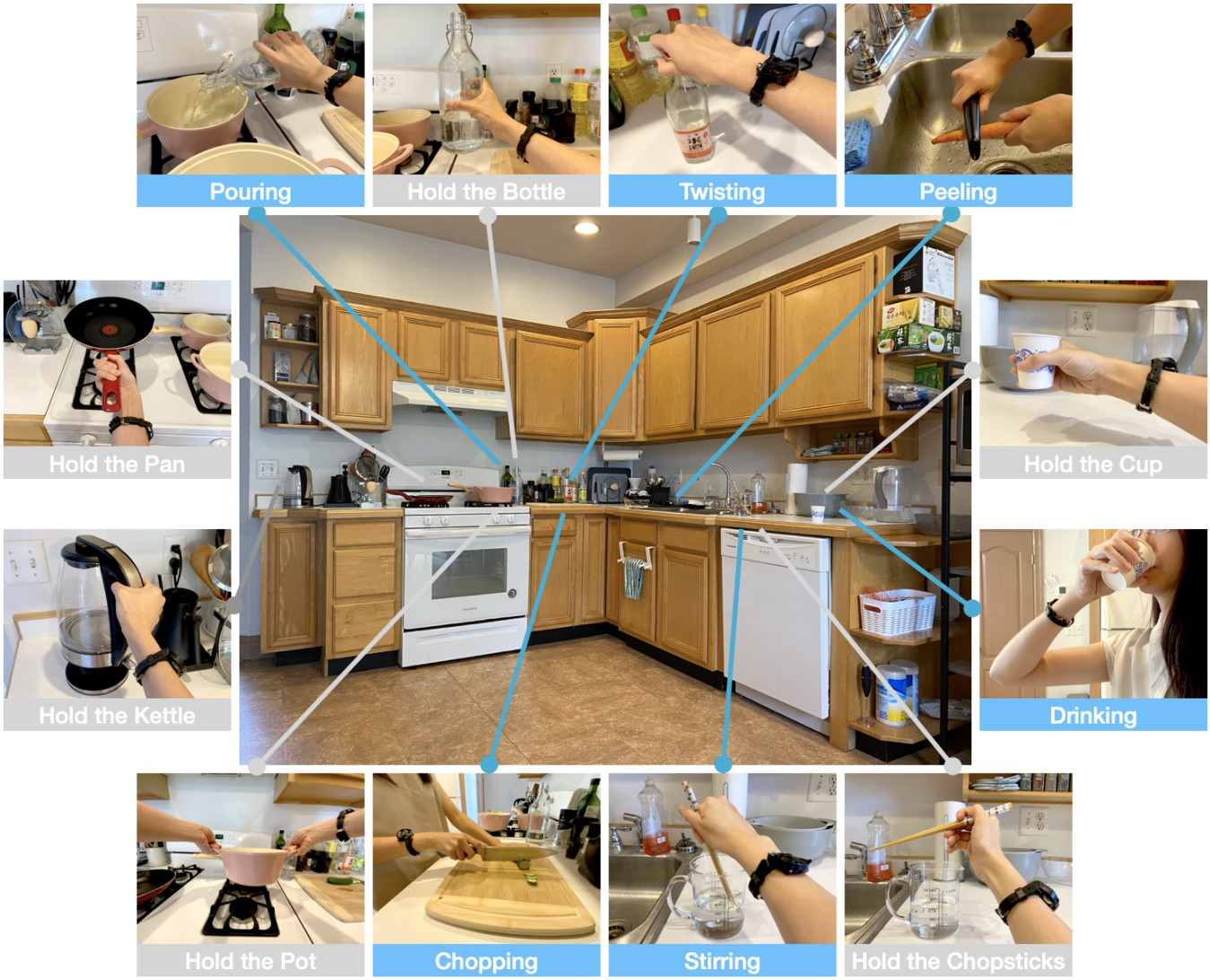

Figure 12: The kitchen and the activities used in the user study. The blue background indicates Dynamic Interactions, while the gray background indicates Static Interactions.

that should be inherently familiar to all participants, there was no need for practice sessions, unlike in the first study.

During each session, participants were engaged in a series of interactions, each lasting 10 seconds and repeated four times. To minimize any learning effects, the order of these interactions was randomized.

Voice commands were used in this study to present instructions and streamline the user study process. At the beginning of each trial, a voice command regarding the target interaction was issued to the participant. Subsequently, participants were allotted a 10-second timeframe to execute the interaction as instructed.

For the *Static Interactions*, the participants were instructed to hold the object until the next interaction was presented. For the *Dynamic Interactions*, the participants were granted the freedom to perform the interaction as many times as they wished within the 10-second timeframe. If the interaction was completed before the 10 seconds elapsed, participants could take a brief rest while placing their hands in a comfortable position.

Each session lasted about 8 minutes. When accounting for the intervals between sessions, the entire study spanned a duration of 75 to 90 minutes for each participant. In total, 2880 (= 12 (participants) × 12 (interactions) × 4 (repetitions) × 5 (sessions)) interactions were collected.

7.4 Results

Following a similar two-step training scheme as described in Sec 6.4 used in the first study, we evaluated EchoWrist's performance in recognizing hand-object interactions in both UI and UD ways. To do so, we first trained a LOPO model for each participant and then fine-tuned the model using only the specific participant's

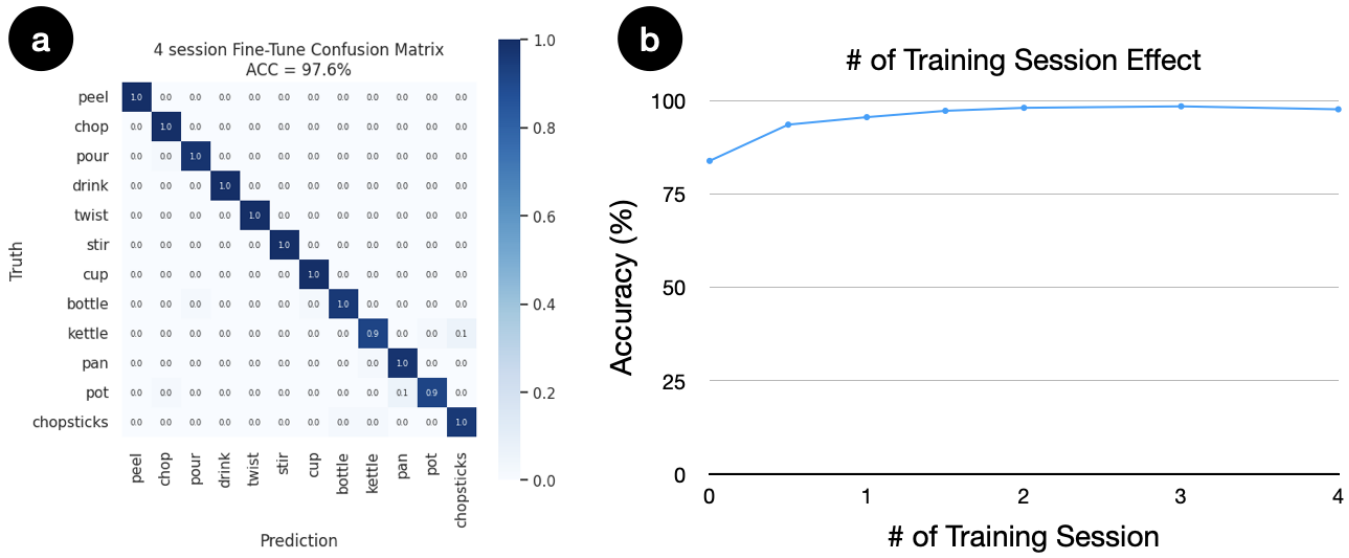

Figure 13: (a) The confusion matrix of 4-session fine-tune model. (b) The performance trend of adjusting the number of sessions used during the fine-tuning step.

data. After fine-tuning the LOPO model with four sessions (i.e., all sessions except the testing ones) of data from the participant, the overall accuracy by averaging the results from all the participants was 97.6% (SD = 0.82%), and all the interactions have at least 80% accuracy (Fig. 13 (a)). More specifically, all the Dynamic Interactions achieved 100% accuracy of prediction, while holding bottle, pot, and chopsticks was still a little bit confusing to the model.

7.4.1 Performance with Different Sizes of Training Data. By manipulating the number of sessions used during the fine-tuning process, we generated the curve shown in Fig. 13 (b). In the UI setting, where no training data from the same participant was employed, the achieved accuracy reached 83.8% (SD = 15.06%). It is important to note that the participants exhibited variations in wrist thickness, hand size, and preferred position for wearing the device. Specifically, some participants favored wearing the device in front of the ulnocarpal joint, while others preferred it directly on the ulnocarpal joint, and still others chose to wear it behind the ulnocarpal joint. Additionally, participants exhibited diverse approaches to interacting with objects. For example, due to the bottle's thickness, some participants with smaller hands could not grab its bottom portion and instead opted to grip the bottleneck, which was thinner and more accessible for them. Furthermore, the manner in which participants grasped the pan also displayed variations. Despite these individual discrepancies, EchoWrist consistently delivered promising results in recognizing and interpreting hand-object interactions.

Our observations from the fine-tuned results revealed an interesting trend. It became evident that even when the base model was trained using data from different individuals, incorporating a small amount of data from a new user resulted in significant performance improvements. For instance, when just 0.5 sessions (4 minutes) of data from the new user were added to the training set, the accuracy notably increased to 93.5% (SD = 5.88%) (Fig. 13 (b)). Furthermore, the variation among users' results decreased with fine-tuning. This serves as a strong validation of the fine-tuning process's effectiveness. It is reasonable to expect that with a more extensive dataset, we could achieve even better performance.

7.4.2 User Experience Survey. Based on the post-study survey results, the participants found EchoWrist comfortable to wear in a real-world setting (M = 6.0, SD = 0.77 on the Likert scale; 1 = extremely uncomfortable, 7 = extremely comfortable). Some participants (P05, P04, P06, P10) also commented that *"it is just a general watch which I usually wear."* Additionally, EchoWrist is *"lightweight"* (P03, P07, P08), and some participants reported that *"I forgot that I was wearing the device during the experiment."* (P09, P11, P12). Overall, EchoWrist provided a good wearing experience. Furthermore, in our post-study survey, all participants reported that they did not perceive any audible sound emitted from the device.

8 DISCUSSION

8.1 Comparison with Other Sensing Methods on Hand Pose Recognition/Tracking

Our study results showed that EchoWrist has a promising performance in continuously tracking hand poses and recognizing hand-object interactions. To help readers better situate the performance of EchoWrist with prior work, we highlight the key characteristics of prior work and EchoWrist in Table 2.

As the table shows, previous work on wristbands that were able to track hand poses continuously either require multiple form factors [75, 76] or consumes significant energy ranging from 0.4 W [25] to 4.5 W [24], which is 10 to 100 times higher than EchoWrist. Furthermore, many of these works require training data from a new user [18, 24, 64, 75, 76] or even a new session [18, 24].

In comparison, EchoWrist can track hand poses continuously using a single wristband, consuming at most 1/10 of the prior works' energy (0.058 W) and providing promising tracking accuracy even

Table 2: Comparison with other sensing methods.

	Technique	Form Factor	Hand Pose Tracking	Hand-Object Interaction Recognition	Thickness	Power	SI	UI
Digits [25]	IR Camera + IMU	Wristband	Continuous Hand Pose (Mean Errors All < 9°)	✗	-	< 0.4 W	-	-
DiscoBand [9]	Depth Sensor	Wristband (16 Depth Sensors)	Continuous Hand Pose (MJEPE 11.69 mm)	Preliminary Exploration	< 1 cm	3.6 W	MPJPE 17.87 mm	MPJPE 19.98 mm
FingerTrak [18]	Thermal Camera	Wristband (4 Thermal Cameras)	Continuous Hand Pose (Average Angular Error 6.46°)	✗	1.19 cm	0.44 W	✗	✗
Z-Ring [75, 76]	Impedance	Ring + Armband (1 VNA)	Continuous Hand Pose (Average Euclidean Error 7.2 mm)	6 Objects (94.5% Accuracy)	-	2.4 W	-	✗
EtherPose [24]	Impedance	Wristband (1 VNA)	Continuous Hand Pose (MPJPE 11.57 mm)	✗	-	4.5 W	✗	✗
Rudolph et al. [64]	Capacitance	Wristband	✗	6 Interactions (99% Accuracy)	0.7 cm	-	-	✗
AudioGest [63]	Acoustic Signals	Commercial Laptop or Smartphone	Discrete Hand Gesture (6 Gestures, 96% Accuracy)	✗	-	-	✓	✓
FingerIO [49]	Acoustic Signals	Smartwatch or Smartphone	Continuous 2D Tracking (Accuracy 8 mm)	✗	-	last 4 hr	✓	✓
BeamBand [22]	Acoustic Signals	Wristband	Discrete Hand Gesture (6 Gestures, 94.6% Accuracy)	✗	1 cm	5 v, 400 mA	89.4% Accuracy	51.7% Accuracy
EchoWrist	Acoustic Signals	Continuous Hand Pose	Wristband (MJEDE 4.81 mm)	12 Interactions (97.6% Accuracy)	0.6 cm	0.0579 W	MJEDE 4.81 mm 97.6% Accuracy	MJEDE 12.2 mm 83.8% Accuracy

without the training data from a new user. The closest prior work is DiscoBand [9], which also tracks hand poses continuously and reports user-dependent and independent performance. In comparison, EchoWrist presents a better tracking performance, as shown in the table, with only 1/63 of their power signature. However, given the pose set is different, the comparison of performance may not be completely fair. Therefore, we intend to present it as a reference for future directions.

8.2 User Dependency of the Deep Learning Model

A data-driven wearable sensing system usually requires a significant amount of training data from a user before using the system. To improve the user experience, we strive to minimize user dependency to allow new users to access the system easily.

For hand pose tracking (Fig. 10), when the new user does not provide any training data, EchoWrist still achieves 12.2 mm MJEDE or 7.37°MJAE. With only 0.5 sessions of training data, EchoWrist obtains a significant performance improvement to 6.92 mm MJEDE or 4.99°MJAE. The duration of 0.5 sessions is roughly one minute. In practice, one minute is close to the time needed to set up a fingerprint sensor or FaceID.

Hand-object interaction recognition achieved an average accuracy of 82.9% for a new user without any training data. Similar to hand pose tracking, with only 0.5 sessions of training data, the accuracy significantly improved, reaching an average of 93.5%. Although the data collection time is longer, taking four minutes, it remains comparable to the setup time of commonly used smart devices.

EchoWrist uses a pre-train-fine-tune scheme, which allows flexibility in incorporating new data and using a larger base dataset to improve performance. In the study, the base dataset was collected from 11 other users. In the future, the base dataset can be expanded at scale to improve EchoWrist's performance further and eventually push EchoWrist towards a real user-independent system with strong performance.

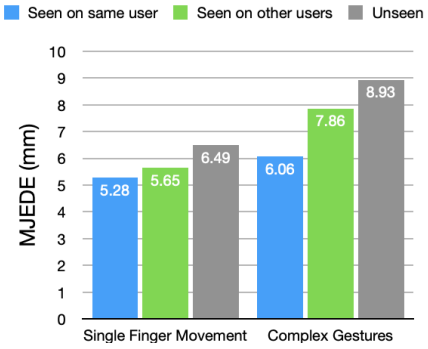

Figure 14: Performance of EchoWrist on unseen gestures. "Seen on same user": gestures in the testing set were also present in the training set from the same participant. "Seen on other users": gestures in the testing set were present in the training set but collected by different participants. "Unseen": gestures in the testing set were not present in the training set.

8.3 Hand Pose Tracking of Unseen Gestures

EchoWrist includes two sets of hand gestures: 5 Simple Gestures and 10 Complex Gestures. Ideally, a hand-tracking technology would work on all hand poses. While it's difficult to evaluate all possible hand poses, testing the system's ability to estimate postures with unseen hand gestures (without training data) is a practical approach. It's worth noting that generalizing to unseen gestures is an extremely challenging task in the wearable community, and we have not seen any prior wearable hand-tracking system that conducted similar experiments. In this section, we present a preliminary analysis to provide insights into the promising potential of deploying EchoWrist in such real-world hand-tracking applications.

In this experiment, we used the two gesture sets collected in the user study as the training and testing sets, respectively.

We first established a baseline by evaluating the model's performance on data from the same gesture set for training and testing. In contrast to the previous experiment, we used a one-step training approach, where training data from all participants were combined

Table 3: Delay breakdown in the two use cases.

Delay (s)	BLE	Prediction	Echo Profile Calc	WiFi	Rendering	Total
Hand Pose Tracking	0.2567	0.0578	0.0018	0.0550	0.0716	0.4429
Hand-Object Interaction Recognition	0.2300	0.2703	0.0054	0.0350	0.0001	0.5408

to form the training set, and testing data from all participants were combined to form the testing set. In this experiment, the model saw the same gesture (different instances) from the same participant in both the training and testing datasets. There was no overlap between the training and testing datasets. The results (Fig. 14) reveal that the MJEDE for the Simple Gestures and the Complex Gestures was 5.28mm and 6.06mm, respectively (bar "Seen on Same User"). Please note that this performance is worse than that reported in Sec 6.5 since the 2-step training was not applied.

We then evaluated the model performance on one gesture set that was trained on a different gesture set. In this case, gestures in the testing set were not present in the training set. The performance of the system decreased to 6.49mm and 8.93mm on Simple Gestures (model trained using Complex Gestures) and Complex Gestures (model trained using Simple Gestures), respectively.

To improve the performance, we conducted experiments by including the same gesture collected from other users in the training set. In this way, the model still did not see any training data from the testing user on the target gestures. Results are shown in Fig 14, and the performance improved over the "Unseen" case, but there is still a gap between seen and unseen gestures. This indicates that incorporating other users' data on new gestures can improve performance.

The performance of EchoWrist decreased when estimating the poses of unseen hand gestures. However, to the best of our knowledge, this is the first experiment to predict unseen gestures in similar data-driven wearable hand posture tracking technologies. The performance was very encouraging (under 10 mm) compared to other prior work. This again confirms the promising potential of this novel hand pose tracking technology for future deployment in real-world applications where hand poses vary significantly.

8.4 Real-Time Inference

We developed a real-time inference system on smartphones to further support various real-life applications. With the BLE module, EchoWrist enables real-time data transmission to a smartphone. We implemented the data processing and deep learning pipeline on the smartphone using PyTorch Mobile [7]. The inference results can subsequently be transmitted to a laptop via WiFi for visualization or additional applications.

The delay of our real-time inference system spans from 0.3 seconds to 1 second for both hand pose tracking and hand-object interaction recognition. We logged the timestamp for each step during real-time inference testing, and ten random frames were chosen to be recorded. The average delay of each step is detailed in Table 3. The prediction delay of hand-object interaction recognition is larger since the window length and pixel of interest are larger, which results in a larger input size. The current bottleneck is in BLE transmission, stemming from a tradeoff between data throughput and latency. This challenge could be addressed by implementing data compression before transmission or relocating the computation process from the smartphone to the microcontroller. The fluctuating delay is primarily attributed to the performance of the smartphone and laptop. In instances of lower performance, e.g., when numerous other applications are running in the background of the smartphone or laptop, the prediction and rendering delay will noticeably increase. The delay can be mitigated by utilizing high-performance devices and further optimizing our algorithms on the operating system of the smartphone.

[7]PyTorch Mobile https://pytorch.org/mobile/home/

8.5 Integrating Hand Tracking and Hand-Object Interaction Recognition

EchoWrist is able to track 3D hand poses as well as recognizing hand-object interactions. Both modules take in the exact same input and go through slightly different processes. While we did not evaluate how 3D hand tracking works when there are objects in hand, we demonstrate that it is possible to directly recognize the hand's interaction intentions with an end-to-end approach. In real use cases, it is possible to integrate the two modules together to allow for a more comprehensive understanding of the hand's activities. For instance, it is possible to use the hand-object interaction module to detect what object is interacting with the hand first. If no object is in the hand, the hand tracking module can be activated to understand the hand shape.

8.6 Smartwatch Integration

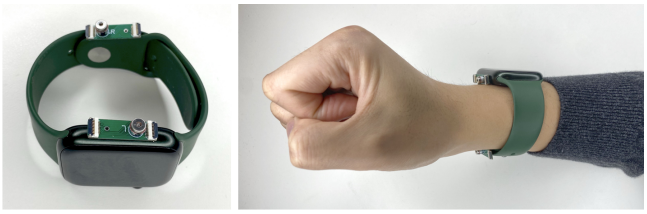

Figure 15: The mockup prototype integrating with an off-the-shelf smartwatch.

Our proposed technology is low-power and minimally obtrusive, demonstrating the potential for integration into commodity smartwatches. However, several questions need to be addressed before it can be fully deployed on commodity devices.

8.6.1 Form Factor and Hardware. Unlike previous wrist-mounted sensing technologies that require cameras or sensors placed high above the skin, EchoWrist only requires two MEMS microphones and speakers placed 5 mm above the skin. Therefore, integrating this sensing technology into future smartwatches is much easier.

Fig 15 illstrate how EchoWrist can be possibly integrated with Apple Watch Series 8 45mm [8]. EchoWrist takes advantage of the thickness of the watch's body and the straps to keep the sensors minimally visible.

It is worth noting that many smartwatches on the market today, including the Apple Watch, already come equipped with speakers and microphones. This means that it is possible to adjust the position and orientation of these components in order to match our requirements. In our previous study, we found that this approach produced promising results.

Furthermore, the integration of the two pairs of microphones and speakers into the watch and the band, respectively, serves as a practical solution. Given the minimal cost, low energy consumption, and compact size of microphones and speakers, incorporating an extra pair into a smartwatch is well within the capabilities of watch manufacturers with the appropriate resources. Moreover, integrating these sensing modules into existing hardware can lead to additional power savings, as our sensing solution can leverage the microprocessor and BLE module on the smartwatch. If we focus solely on the power consumption of the two pairs of microphones and speakers, the estimated power signature is as low as 10.0 mW.

8.6.2 Privacy Protection. EchoWrist uses 18-21 kHz or 20-24 kHz acoustic sounds, which are generally inaudible to most people. Human conversations and most environmental sounds are usually distributed in low-frequency ranges that can be easily filtered out. However, in the current implementation of EchoWrist, all sounds under 25 KHz are recorded and transmitted to smartphones, which may expose potential privacy risks to users if the BLE transmission is hacked and raw audio is leaked. To mitigate this risk, several solutions can be adopted. Firstly, an analog band-pass filter can be implemented in the hardware to remove low-frequency sounds before they are converted into digital signals. Secondly, raw audio data transmission can be avoided in the processing system. This can be achieved by implementing a digital filter to remove audible frequencies in the data or by extracting features on the smartwatch and only transmitting processed features. Lastly, a complete local data processing pipeline can be employed on the user's personal devices, such as a smartphone or even on embedded microcontrollers, to avoid transmitting the data into the cloud.

9 LIMITATIONS AND FUTURE WORK

9.1 Cloth Coverage

One limitation shared by EchoWrist and many other wrist-mounted sensing methods based on cameras [18] is the potential degradation of system performance when the sensing unit on the wristband is obscured by clothing. In such cases, the acoustic signals may be obstructed by the fabric and fail to reach the hands, particularly when users are wearing long-sleeved clothing. This limitation imposes a requirement for users of our system to wear clothing that does not cover the wrist. In some cases, clothing might obstruct some sensing units partially while leaving others unaffected. In future research, this feature could be leveraged to enhance the algorithm's robustness or integrated with other sensing methods to mitigate occlusion issues arising from clothing.

[8] https://www.apple.com/apple-watch-series-8/

9.2 Performance under Intense Movements and Holding Objects

While EchoWrist successfully tracks both hand postures and hand-object interactions, its performance in tracking hand postures while holding objects or during intense movements remains unexplored. This is a challenging issue encountered by many hand activity tracking systems. We believe that performing intense movement and holding objects could introduce distinctive acoustic echo profiles. To address this, further investigation is needed, potentially involving the collection of additional training data specific to these scenarios. Future research may delve into this aspect to enhance the system's capabilities.

9.3 Hand-Object Interaction Contexts

In study 2 (Section 7, we evaluated EchoWrist's performance in recognizing hand-object interactions within a kitchen setting. However, this does not mean that EchoWrist can only be used in the kitchen. Pilot studies involving one researcher testing the system in different contexts were conducted. The interactions that have been tested included writing, typing, scrolling, cutting, and more. Most of the tested interactions demonstrated impressive results, achieving accuracy rates exceeding 90% when training and testing with the data from the same user. However, it is challenging for EchoWrist to distinguish holding objects sharing similar shapes and grabbing poses, e.g., fork and spoon, pencil and marker, and hot glue gun and drill. In addition, objects that are fully in hand can not be recognized as well. These include AirPods, erasers, and candies. Future research could encompass more expansive contexts to further gauge EchoWrist's adaptability and robustness across a broader spectrum of everyday scenarios. Also, multimodel approaches could be deployed to extend the capability further.

10 CONCLUSION

Through this paper, we present the design, implementation, and evaluation of EchoWrist, the first wristband that can both track 3D hand poses continuously and recognize hand-object interactions. Two user studies with 24 participants in total demonstrate EchoWrist's capability and robustness in these two tasks. EchoWrist operates at 56.9mW while maintaining a low-profile minimally-obtrusive form factor. With further optimization, we believe that it is promising to deploy EchoWrist at scale.

ACKNOWLEDGMENTS

We extend our gratitude to all our colleagues at SciFi Lab for their invaluable support. Additionally, we express our heartfelt appreciation to all the participants who generously contributed to the user study. This project was supported by the National Science Foundation Grant No. 2239569, the National Science Foundation Award No. IIS-1925100, the National Science Foundation's I-Corps Award No. 2346817, the Cornell University IGNITE Innovation Acceleration Program, and the Nakajima Foundation.

REFERENCES

[1] Brian Amento, Will Hill, and Loren Terveen. 2002. The Sound of One Hand: A Wrist-Mounted Bio-Acoustic Fingertip Gesture Interface. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems* (Minneapolis, Minnesota, USA)

(CHI EA '02). Association for Computing Machinery, New York, NY, USA, 724–725. https://doi.org/10.1145/506443.506566

[2] Sarnab Bhattacharya, Rebecca Adaimi, and Edison Thomaz. 2022. Leveraging Sound and Wrist Motion to Detect Activities of Daily Living with Commodity Smartwatches. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 42 (jul 2022), 28 pages. https://doi.org/10.1145/3534582

[3] Wenqiang Chen, Lin Chen, Meiyi Ma, Farshid Salemi Parizi, Shwetak Patel, and John Stankovic. 2021. ViFin: Harness Passive Vibration to Continuous Micro Finger Writing with a Commodity Smartwatch. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1, Article 45 (mar 2021), 25 pages. https://doi.org/10.1145/3448119

[4] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2018. Scaling Egocentric Vision: The EPIC-KITCHENS Dataset. In *European Conference on Computer Vision (ECCV)*.

[5] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2022. Rescaling Egocentric Vision: Collection, Pipeline and Challenges for EPIC-KITCHENS-100. *International Journal of Computer Vision (IJCV)* 130 (2022), 33–55. https://doi.org/10.1007/s11263-021-01531-2

[6] Pradipto Das, Chenliang Xu, Richard F. Doell, and Jason J. Corso. 2013. A Thousand Frames in Just a Few Words: Lingual Description of Videos through Latent Topics and Sparse Object Stitching. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*. 2634–2641. https://doi.org/10.1109/CVPR.2013.340

[7] Artem Dementyev and Joseph A. Paradiso. 2014. WristFlex: Low-Power Gesture Input with Wrist-Worn Pressure Sensors. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology* (Honolulu, Hawaii, USA) *(UIST '14)*. Association for Computing Machinery, New York, NY, USA, 161–166. https://doi.org/10.1145/2642918.2647396

[8] Tamara Denning, Zakariya Dehlawi, and Tadayoshi Kohno. 2014. In Situ with Bystanders of Augmented Reality Glasses: Perspectives on Recording and Privacy-Mediating Technologies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 2377–2386. https://doi.org/10.1145/2556288.2557352

[9] Nathan Devrio and Chris Harrison. 2022. DiscoBand: Multiview Depth-Sensing Smartwatch Strap for Hand, Body and Environment Tracking. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 56, 13 pages. https://doi.org/10.1145/3526113.3545634

[10] Travis Deyle, Szabolcs Palinko, Erika Poole, and Thad Starner. 2007. Hambone: A Bio-Acoustic Gesture Interface. 3–10. https://doi.org/10.1109/ISWC.2007.4373768

[11] Yu Du, Yongkang Wong, Wenguang Jin, Wentao Wei, Yu Hu, Mohan Kankanhalli, and Weidong Geng. 2017. Semi-Supervised Learning for Surface EMG-based Gesture Recognition. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2017/225

[12] Junjun Fan, Xiangmin Fan, Feng Tian, Yang Li, Zitao Liu, Wei Sun, and Hongan Wang. 2018. What is That in Your Hand? Recognizing Grasped Objects via Forearm Electromyography Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 4, Article 161 (dec 2018), 24 pages. https://doi.org/10.1145/3287039

[13] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 2019. 3d hand shape and pose estimation from a single rgb image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10833–10842.

[14] Jun Gong, Aakar Gupta, and Hrvoje Benko. 2020. Acustico: Surface Tap Detection and Localization Using Wrist-Based Acoustic TDOA Sensing. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 406–419. https://doi.org/10.1145/3379337.3415901

[15] Jun Gong, Xing-Dong Yang, and Pourang Irani. 2016. WristWhirl: One-Handed Continuous Smartwatch Input Using Wrist Gestures. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (Tokyo, Japan) *(UIST '16)*. Association for Computing Machinery, New York, NY, USA, 861–872. https://doi.org/10.1145/2984511.2984563

[16] Chris Harrison, Desney Tan, and Dan Morris. 2010. Skinput: Appropriating the Body as an Input Surface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 453–462. https://doi.org/10.1145/1753326.1753394

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778. https://doi.org/10.1109/CVPR.2016.90

[18] Fang Hu, Peng He, Songlin Xu, Yin Li, and Cheng Zhang. 2020. FingerTrak: Continuous 3D Hand Pose Tracking by Deep Learning Hand Silhouettes Captured by Miniature Thermal Cameras on Wrist. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 2, Article 71 (jun 2020), 24 pages. https://doi.org/10.1145/3397306

[19] Jun Ho Huh, Hyejin Shin, HongMin Kim, Eunyong Cheon, Youngeun Song, Choong-Hoon Lee, and Ian Oakley. 2023. WristAcoustic: Through-Wrist Acoustic Response Based Authentication for Smartwatches. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 4, Article 167 (jan 2023), 34 pages. https://doi.org/10.1145/3569473

[20] Tap Systems Inc. 2022. *Tap.* Retrieved Sep 14, 2023 from https://www.tapwithus.com/product/tap-strap-2/

[21] Intel. 2022. *Intel RealSense Technology.* Retrieved Feb 12, 2023 from https://www.intel.com/content/www/us/en/architecture-and-technology/realsense-overview.html

[22] Yasha Iravantchi, Mayank Goel, and Chris Harrison. 2019. BeamBand: Hand Gesture Sensing with Ultrasonic Beamforming. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3290605.3300245

[23] Frederic Kerber, Michael Puhl, and Antonio Krüger. 2017. User-Independent Real-Time Hand Gesture Recognition Based on Surface Electromyography. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Vienna, Austria) *(MobileHCI '17)*. Association for Computing Machinery, New York, NY, USA, Article 36, 7 pages. https://doi.org/10.1145/3098279.3098553

[24] Daehwa Kim and Chris Harrison. 2022. EtherPose: Continuous Hand Pose Tracking with Wrist-Worn Antenna Impedance Characteristic Sensing. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 58, 12 pages. https://doi.org/10.1145/3526113.3545665

[25] David Kim, Otmar Hilliges, Shahram Izadi, Alex D. Butler, Jiawen Chen, Iason Oikonomidis, and Patrick Olivier. 2012. Digits: Freehand 3D Interactions Anywhere Using a Wrist-Worn Gloveless Sensor. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology* (Cambridge, Massachusetts, USA) *(UIST '12)*. Association for Computing Machinery, New York, NY, USA, 167–176. https://doi.org/10.1145/2380116.2380139

[26] Jungsoo Kim, Jiasheng He, Kent Lyons, and Thad Starner. 2007. The Gesture Watch: A Wireless Contact-free Gesture based Wrist Interface. 15–22. https://doi.org/10.1109/ISWC.2007.4373770

[27] Jiwan Kim and Ian Oakley. 2022. SonarID: Using Sonar to Identify Fingers on a Smartwatch. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>New Orleans</city>, <state>LA</state>, <country>USA</country>, </conf-loc>) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 287, 10 pages. https://doi.org/10.1145/3491102.3501935

[28] Olya Kudina and Peter-Paul Verbeek. 2019. Ethics from within: Google Glass, the Collingridge dilemma, and the mediated value of privacy. *Science, Technology, & Human Values* 44, 2 (2019), 291–314.

[29] Gierad Laput, Eric Brockmeyer, Scott E. Hudson, and Chris Harrison. 2015. Acoustruments: Passive, Acoustically-Driven, Interactive Controls for Handheld Devices. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 2161–2170. https://doi.org/10.1145/2702123.2702414

[30] Gierad Laput and Chris Harrison. 2019. Sensing Fine-Grained Hand Activity with Smartwatches. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300568

[31] Gierad Laput, Robert Xiao, and Chris Harrison. 2016. ViBand: High-Fidelity Bio-Acoustic Sensing Using Commodity Smartwatch Accelerometers. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (Tokyo, Japan) *(UIST '16)*. Association for Computing Machinery, New York, NY, USA, 321–333. https://doi.org/10.1145/2984511.2984582

[32] Dong Li, Jialin Liu, Sunghoon Ivan Lee, and Jie Xiong. 2023. Room-Scale Hand Gesture Recognition Using Smart Speakers. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems* (Boston, Massachusetts) *(SenSys '22)*. Association for Computing Machinery, New York, NY, USA, 462–475. https://doi.org/10.1145/3560905.3568528

[33] Ke Li, Ruidong Zhang, Boao Chen, Siyuan Chen, Sicheng Yin, Saif Mahmud, Qikang Liang, François Guimbretière, and Cheng Zhang. 2024. GazeTrak: Exploring Acoustic-based Eye Tracking on a Glass Frame. In *Proceedings of the Annual International Conference on Mobile Computing and Networking* (Washington D.C., DC, USA) *(MobiCom '24)*. Association for Computing Machinery, New York, NY, USA, 16 pages. https://doi.org/10.1145/3636534.3649376

[34] Ke Li, Ruidong Zhang, Siyuan Chen, Boao Chen, Mose Sakashita, François Guimbretière, and Cheng Zhang. 2024. EyeEcho: Continuous and Low-power Facial Expression Tracking on Glasses. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, 24 pages. https://doi.org/10.1145/3613904.3642613

[35] 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

[36] Hyunchul Lim, Yaxuan Li, Matthew Dressa, Fang Hu, Jae Hoon Kim, Ruidong Zhang, and Cheng Zhang. 2022. BodyTrak: Inferring Full-body Poses from Body Silhouettes Using a Miniature Camera on a Wristband. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 3, Article 154 (sep 2022), 21 pages. https://doi.org/10.1145/3552312

[37] Hyunchul Lim, Ruidong Zhang, Samhita Pendyal, Jeyeon Jo, and Cheng Zhang. 2023. D-Touch: Recognizing and Predicting Fine-Grained Hand-Face Touching Activities Using a Neck-Mounted Wearable. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 569–583. https://doi.org/10.1145/3581641.3584063

[38] Jhe-Wei Lin, Chiuan Wang, Yi Yao Huang, Kuan-Ting Chou, Hsuan-Yu Chen, Wei-Luan Tseng, and Mike Y. Chen. 2015. BackHand: Sensing Hand Gestures via Back of the Hand. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology* (Charlotte, NC, USA) *(UIST '15)*. Association for Computing Machinery, New York, NY, USA, 557–564. https://doi.org/10.1145/2807442.2807462

[39] Yang Liu, Chengdong Lin, and Zhenjiang Li. 2021. WR-Hand: Wearable Armband Can Track User's Hand. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 3, Article 118 (sep 2021), 27 pages. https://doi.org/10.1145/3478112

[40] Yilin Liu, Shijia Zhang, and Mahanth Gowda. 2021. NeuroPose: 3D Hand Pose Tracking Using EMG Wearables. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 1471–1482. https://doi.org/10.1145/3442381.3449890

[41] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172* (2019).

[42] Takuya Maekawa, Yasue Kishino, Yutaka Yanagisawa, and Yasushi Sakurai. 2012. WristSense: Wrist-worn sensor device with camera for daily activity recognition. In *2012 IEEE International Conference on Pervasive Computing and Communications Workshops*. 510–512. https://doi.org/10.1109/PerComW.2012.6197551

[43] Saif Mahmud, Ke Li, Guilin Hu, Hao Chen, Richard Jin, Ruidong Zhang, François Guimbretière, and Cheng Zhang. 2023. PoseSonic: 3D Upper Body Pose Estimation Through Egocentric Acoustic Sensing on Smartglasses. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 3, Article 111 (sep 2023), 28 pages. https://doi.org/10.1145/3610895

[44] Jess McIntosh, Asier Marzo, and Mike Fraser. 2017. SensIR: Detecting Hand Gestures with a Wearable Bracelet Using Infrared Transmission and Reflection. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (Québec City, QC, Canada) *(UIST '17)*. Association for Computing Machinery, New York, NY, USA, 593–597. https://doi.org/10.1145/3126594.3126604

[45] Jess McIntosh, Asier Marzo, Mike Fraser, and Carol Phillips. 2017. EchoFlex: Hand Gesture Recognition Using Ultrasound Imaging. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 1923–1934. https://doi.org/10.1145/3025453.3025807

[46] Microsoft. 2022. *Kinect for Windows*. Retrieved Feb 12, 2023 from https://learn.microsoft.com/en-us/windows/apps/design/devices/kinect-for-windows

[47] Vimal Mollyn, Karan Ahuja, Dhruv Verma, Chris Harrison, and Mayank Goel. 2022. SAMoSA: Sensing Activities with Motion and Subsampled Audio. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 3, Article 132 (sep 2022), 19 pages. https://doi.org/10.1145/3550284

[48] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. 2018. GANerated Hands for Real-Time 3D Hand Tracking from Monocular RGB. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 49–59. https://doi.org/10.1109/CVPR.2018.00013

[49] Rajalakshmi Nandakumar, Vikram Iyer, Desney Tan, and Shyamnath Gollakota. 2016. FingerIO: Using Active Sonar for Fine-Grained Finger Tracking. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 1515–1525. https://doi.org/10.1145/2858036.2858580

[50] Seungjae Oh, Gyeore Yun, Chaeyong Park, Jinsoo Kim, and Seungmoon Choi. 2019. VibEye: Vibration-Mediated Object Recognition for Tangible Interactive Applications. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300906

[51] Joseph O'Hagan, Pejman Saeghe, Jan Gugenheimer, Daniel Medeiros, Karola Marky, Mohamed Khamis, and Mark McGill. 2023. Privacy-Enhancing Technology and Everyday Augmented Reality: Understanding Bystanders' Varying Needs for Awareness and Consent. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 4, Article 177 (jan 2023), 35 pages. https://doi.org/10.1145/3569501

[52] Katsunori Ohnishi, Atsushi Kanehira, Asako Kanezaki, and Tatsuya Harada. 2016. Recognizing Activities of Daily Living with a Wrist-Mounted Camera. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3103–3111. https://doi.org/10.1109/CVPR.2016.338

[53] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A. Argyros. 2011. Full DOF tracking of a hand interacting with an object by modeling occlusions and physical constraints. In *2011 International Conference on Computer Vision*. 2088–2095. https://doi.org/10.1109/ICCV.2011.6126483

[54] Makoto Ono, Buntarou Shizuki, and Jiro Tanaka. 2013. Touch & Activate: Adding Interactivity to Existing Objects Using Active Acoustic Sensing. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology* (St. Andrews, Scotland, United Kingdom) *(UIST '13)*. Association for Computing Machinery, New York, NY, USA, 31–40. https://doi.org/10.1145/2501988.2501989

[55] Farshid Salemi Parizi, Eric Whitmire, and Shwetak Patel. 2020. AuraRing: Precise Electromagnetic Finger Tracking. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 4, Article 150 (sep 2020), 28 pages. https://doi.org/10.1145/3369831

[56] Farshid Salemi Parizi, Eric Whitmire, and Shwetak Patel. 2020. AuraRing: Precise Electromagnetic Finger Tracking. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 4, Article 150 (sep 2020), 28 pages. https://doi.org/10.1145/3369831

[57] JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. 2022. HandOccNet: Occlusion-Robust 3D Hand Mesh Estimation Network. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1486–1495. https://doi.org/10.1109/CVPR52688.2022.00155

[58] John Perng, B. Fisher, Seth Hollar, and Kristofer Pister. 1999. Acceleration sensing glove (ASG). 178 – 180. https://doi.org/10.1109/ISWC.1999.806717

[59] Fernando Quivira, Toshiaki Koike-Akino, Ye Wang, and Deniz Erdogmus. 2018. Translating sEMG signals to continuous hand poses using recurrent neural networks. In *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. 166–169. https://doi.org/10.1109/BHI.2018.8333395

[60] Sumit Raurale, John McAllister, and Jesus Martinez del Rincon. 2018. Emg Acquisition and Hand Pose Classification for Bionic Hands from Randomly-Placed Sensors. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Calgary, AB, Canada). IEEE Press, 1105–1109. https://doi.org/10.1109/ICASSP.2018.8462409

[61] Jun Rekimoto. 2001. GestureWrist and GesturePad: unobtrusive wearable interactiondevices. *International Symposium on Wearable Computers, Digest of Papers*, 21–27. https://doi.org/10.1109/ISWC.2001.962092

[62] Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. 2012. A database for fine grained activity detection of cooking activities. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 1194–1201. https://doi.org/10.1109/CVPR.2012.6247801

[63] 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 *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Heidelberg, Germany) *(UbiComp '16)*. Association for Computing Machinery, New York, NY, USA, 474–485. https://doi.org/10.1145/2971648.2971736

[64] Julius Cosmo Romeo Rudolph, David Holman, Bruno De Araujo, Ricardo Jota, Daniel Wigdor, and Valkyrie Savage. 2022. Sensing Hand Interactions with Everyday Objects by Profiling Wrist Topography. In *Sixteenth International Conference on Tangible, Embedded, and Embodied Interaction* (Daejeon, Republic of Korea) *(TEI '22)*. Association for Computing Machinery, New York, NY, USA, Article 14, 14 pages. https://doi.org/10.1145/3490149.3501320

[65] Ashwin De Silva, Malsha V. Perera, Kithmin Wickramasinghe, Asma M. Naim, Thilina Dulantha Lalitharatne, and Simon L. Kappel. 2020. Real-Time Hand Gesture Recognition Using Temporal Muscle Activation Maps of Multi-Channel Semg Signals. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1299–1303. https://doi.org/10.1109/ICASSP40776.2020.9054227

[66] Ivan Sosin, Daniel Kudenko, and Aleksei Shpilman. 2018. Continuous Gesture Recognition from sEMG Sensor Data with Recurrent Neural Networks and Adversarial Domain Adaptation. In *2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV)*. IEEE. https://doi.org/10.1109/icarcv.2018.8581206

[67] Ke Sun, Ting Zhao, Wei Wang, and Lei Xie. 2018. VSkin: Sensing Touch Gestures on Surfaces of Mobile Devices Using Acoustic Signals. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking* (New Delhi, India) *(MobiCom '18)*. Association for Computing Machinery, New York, NY, USA, 591–605. https://doi.org/10.1145/3241539.3241568

[68] Rujia Sun, Xiaohe Zhou, Benjamin Steeper, Ruidong Zhang, Sicheng Yin, Ke Li, Shengzhang Wu, Sam Tilsen, Francois Guimbretiere, and Cheng Zhang. 2023. EchoNose: Sensing Mouth, Breathing and Tongue Gestures inside Oral Cavity using a Non-contact Nose Interface. In *Proceedings of the 2023 ACM International Symposium on Wearable Computers* (Cancun, Quintana Roo, Mexico) *(ISWC '23)*. Association for Computing Machinery, New York, NY, USA, 22–26. https://doi.org/10.1145/3594738.3611358

[69] Wei Sun, Franklin Mingzhe Li, Congshu Huang, Zhenyu Lei, Benjamin Steeper, Songyun Tao, Feng Tian, and Cheng Zhang. 2021. ThumbTrak: Recognizing Micro-finger Poses Using a Ring with Proximity Sensing. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction*. ACM. https://doi.org/10.1145/3447526.3472060

[70] Hoang Truong, Phuc Nguyen, Nam Bui, Anh Nguyen, and Tam Vu. 2017. Demo: Low-Power Capacitive Sensing Wristband for Hand Gesture Recognition. In *Proceedings of the 9th ACM Workshop on Wireless of the Students, by the Students, and for the Students* (Snowbird, Utah, USA) *(S3 '17)*. Association for Computing Machinery, New York, NY, USA, 21. https://doi.org/10.1145/3131348.3131358

[71] Hoang Truong, Phuc Nguyen, Anh Nguyen, Nam Bui, and Tam Vu. 2017. Capacitive Sensing 3D-Printed Wristband for Enriched Hand Gesture Recognition. In *Proceedings of the 2017 Workshop on Wearable Systems and Applications* (Niagara Falls, New York, USA) *(WearSys '17)*. Association for Computing Machinery, New York, NY, USA, 11–15. https://doi.org/10.1145/3089351.3089359

[72] Hoang Truong, Shuo Zhang, Ufuk Muncuk, Phuc Nguyen, Nam Bui, Anh Nguyen, Qin Lv, Kaushik Chowdhury, Thang Dinh, and Tam Vu. 2018. CapBand: Battery-Free Successive Capacitance Sensing Wristband for Hand Gesture Recognition. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems* (Shenzhen, China) *(SenSys '18)*. Association for Computing Machinery, New York, NY, USA, 54–67. https://doi.org/10.1145/3274783.3274854

[73] Hsin-Ruey Tsai, Cheng-Yuan Wu, Lee-Ting Huang, and Yi-Ping Hung. 2016. ThumbRing: Private Interactions Using One-Handed Thumb Motion Input on Finger Segments. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct* (Florence, Italy) *(MobileHCI '16)*. Association for Computing Machinery, New York, NY, USA, 791–798. https://doi.org/10.1145/2957265.2961859

[74] UltraLeap. 2022. *World-leading Hand Tracking Products: Small. Fast. Accurate. | Ultraleap.* Retrieved Feb 12, 2023 from https://www.ultraleap.com/product/

[75] Anandghan Waghmare, Youssef Ben Taleb, Ishan Chatterjee, Arjun Narendra, and Shwetak Patel. 2023. Z-Ring: Single-Point Bio-Impedance Sensing for Gesture, Touch, Object and User Recognition. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 150, 18 pages. https://doi.org/10.1145/3544548.3581422

[76] Anandghan Waghmare, Ishan Chatterjee, and Shwetak Patel. 2023. Z-Pose: Continuous 3D Hand Pose Tracking Using Single-Point Bio-Impedance Sensing on a Ring. In *Proceedings of the 2nd Workshop on Smart Wearable Systems and Applications* (Madrid, Spain) *(SmartWear '23)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3615592.3616851

[77] Tianben Wang, Daqing Zhang, Yuanqing Zheng, Tao Gu, Xingshe Zhou, and Bernadette Dorizzi. 2018. C-FMCW Based Contactless Respiration Detection Using Acoustic Signal. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 4, Article 170 (jan 2018), 20 pages. https://doi.org/10.1145/3161188

[78] Erwin Wu, Ye Yuan, Hui-Shyong Yeo, Aaron Quigley, Hideki Koike, and Kris M. Kitani. 2020. Back-Hand-Pose: 3D Hand Pose Estimation for a Wrist-Worn Camera via Dorsum Deformation Network. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 1147–1160. https://doi.org/10.1145/3379337.3415897

[79] Chao Xu, Parth H. Pathak, and Prasant Mohapatra. 2015. Finger-Writing with Smartwatch: A Case for Finger and Hand Gesture Recognition Using Smartwatch. In *Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications* (Santa Fe, New Mexico, USA) *(HotMobile '15)*. Association for Computing Machinery, New York, NY, USA, 9–14. https://doi.org/10.1145/2699343.2699350

[80] Xuhai Xu, Jun Gong, Carolina Brum, Lilian Liang, Bongsoo Suh, Shivam Kumar Gupta, Yash Agarwal, Laurence Lindsey, Runchang Kang, Behrooz Shahsavari, Tu Nguyen, Heriberto Nieto, Scott E Hudson, Charlie Maalouf, Jax Seyed Mousavi, and Gierad Laput. 2022. Enabling Hand Gesture Customization on Wrist-Worn Devices. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 496, 19 pages. https://doi.org/10.1145/3491102.3501904

[81] Hui-Shyong Yeo, Erwin Wu, Juyoung Lee, Aaron Quigley, and Hideki Koike. 2019. Opisthenar: Hand Poses and Finger Tapping Recognition by Observing Back of Hand Using Embedded Wrist Camera. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 963–971. https://doi.org/10.1145/3332165.3347867

[82] Cheng Zhang, Qiuyue Xue, Anandghan Waghmare, Ruichen Meng, Sumeet Jain, Yizeng Han, Xinyu Li, Kenneth Cunefare, Thomas Ploetz, Thad Starner, Omer Inan, and Gregory D. Abowd. 2018. FingerPing: Recognizing Fine-grained Hand Poses using Active Acoustic On-body Sensing. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Montreal QC</city>, <country>Canada</country>, </conf-loc>) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3173574.3174011

[83] Ruidong Zhang, Hao Chen, Devansh Agarwal, Richard Jin, Ke Li, François Guimbretière, and Cheng Zhang. 2023. HPSpeech: Silent Speech Interface for Commodity Headphones. In *Proceedings of the 2023 ACM International Symposium on Wearable Computers* (Cancun, Quintana Roo, Mexico) *(ISWC '23)*. Association for Computing Machinery, New York, NY, USA, 60–65. https://doi.org/10.1145/3594738.3611365

[84] Ruidong Zhang, Ke Li, Yihong Hao, Yufan Wang, Zhengnan Lai, François Guimbretière, and Cheng Zhang. 2023. EchoSpeech: Continuous Silent Speech Recognition on Minimally-Obtrusive Eyewear Powered by Acoustic Sensing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 852, 18 pages. https://doi.org/10.1145/3544548.3580801

[85] Ruidong Zhang, Jihai Zhang, Nitish Gade, Peng Cao, Seyun Kim, Junchi Yan, and Cheng Zhang. 2022. EatingTrak: Detecting Fine-Grained Eating Moments in the Wild Using a Wrist-Mounted IMU. *Proc. ACM Hum.-Comput. Interact.* 6, MHCI, Article 214 (sep 2022), 22 pages. https://doi.org/10.1145/3546749

[86] Yang Zhang and Chris Harrison. 2015. Tomo: Wearable, Low-Cost Electrical Impedance Tomography for Hand Gesture Recognition. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology* (Charlotte, NC, USA) *(UIST '15)*. Association for Computing Machinery, New York, NY, USA, 167–173. https://doi.org/10.1145/2807442.2807480

[87] Yang Zhang, Robert Xiao, and Chris Harrison. 2016. Advancing Hand Gesture Recognition with High Resolution Electrical Impedance Tomography. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (Tokyo, Japan) *(UIST '16)*. Association for Computing Machinery, New York, NY, USA, 843–850. https://doi.org/10.1145/2984511.2984574

[88] Hao Zhou, Taiting Lu, Yilin Liu, Shijia Zhang, and Mahanth Gowda. 2022. Learning on the Rings: Self-Supervised 3D Finger Motion Tracking Using Wearable Sensors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 90 (jul 2022), 31 pages. https://doi.org/10.1145/3534587

[89] Luowei Zhou, Chenliang Xu, and Jason Corso. 2018. Towards Automatic Learning of Procedures From Web Instructional Videos. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (April 2018). https://doi.org/10.1609/aaai.v32i1.12342

[90] Christian Zimmermann and Thomas Brox. 2017. Learning to Estimate 3D Hand Pose from Single RGB Images. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 4913–4921. https://doi.org/10.1109/ICCV.2017.525