WatchHand: Enabling Continuous Hand Pose Tracking On Off-the-Shelf Smartwatches

Jiwan Kim*
School of Electrical Engineering
KAIST
Daejeon, Republic of Korea
jiwankim@kaist.ac.kr

Chi-Jung Lee*
Cornell University
Ithaca, New York, USA
cl2358@cornell.edu

Hohurn Jung
School of Electrical Engineering
KAIST
Deajon, Republic of Korea
cllocker@kaist.ac.kr

Tianhong Catherine Yu
Cornell University
Ithaca, New York, USA
ty274@cornell.edu

Ruidong Zhang
Information Science
Cornell University
Ithaca, New York, USA
rz379@cornell.edu

Ian Oakley†
School of Electrical Engineering
KAIST
Daejeon, Republic of Korea
ian.r.oakley@gmail.com

Cheng Zhang†
Computing and Information Science
Cornell University
ITHACA, New York, USA
chengzhang@cornell.edu

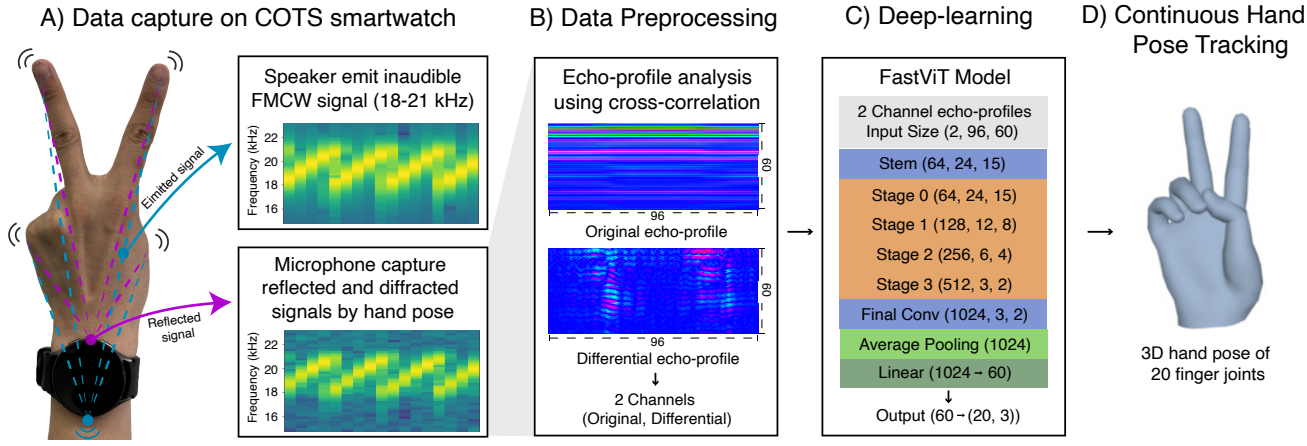


Figure 1: Overview of WatchHand: (A) Commercial off-the-shelf smartwatch emits inaudible frequency-modulated continuous waves (FMCW) via its speaker while capturing reflected and diffracted signals from hand motion through its microphone. (B) The captured acoustic signals are preprocessed using cross-correlation to generate echo profiles—a spatiotemporal representation of signal reflections. (C) A deep-learning model then processes echo profiles to (D) continuously track 3D hand poses.

Abstract

Tracking hand poses on wrist-wearables enables rich, expressive interactions, yet remains unavailable on commercial smartwatches, as prior implementations rely on external sensors or custom hardware, limiting their real-world applicability. To address this, we present WatchHand, the first continuous 3D hand pose tracking system implemented on off-the-shelf smartwatches using only their built-in speaker and microphone. WatchHand emits inaudible frequency-modulated continuous waves and captures their reflections from the hand. These acoustic signals are processed by a deep-learning model that estimates 3D hand poses for 20 finger joints. We evaluate WatchHand across diverse real-world conditions—multiple smartwatch models, wearing-hands, body postures, noise conditions, pose-variation protocols—and achieve a mean per-joint position

*Contributed equally to this paper
†Co-corresponding authors

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

CHI '26, Barcelona, Spain

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3790932

error of 7.87 mm in cross-session tests with device remounting. Although performance drops for unseen users or gestures, the model adapts effectively with lightweight fine-tuning on small amounts of data. Overall, WatchHand lowers the barrier to smartwatch-based hand tracking by eliminating additional hardware while enabling robust, always-available interactions on millions of existing devices.

CCS Concepts

• **Human-centered computing → Gestural input**; *Sound-based input / output*; *Mobile devices.*

Keywords

Hand pose, Smartwatch, Wearable, Acoustic sensing

ACM Reference Format:

Jiwan Kim, Chi-Jung Lee, Hohurn Jung, Tianhong Catherine Yu, Ruidong Zhang, Ian Oakley, and Cheng Zhang. 2026. WatchHand: Enabling Continuous Hand Pose Tracking On Off-the-Shelf Smartwatches. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3772318.3790932

1 Introduction

Hands are our fundamental tools for interacting with the surrounding world, expressing our physical intentions, and performing activities. However, despite millions of smartwatches on the market, none of these commercial devices can continuously track hand poses, a key technique to understanding user intentions and behaviors, such as supporting context-aware applications or behavioral communication and expression (e.g., finger-spelling [11, 21, 30]). Recognizing the importance of tracking hand poses for interaction, researchers have explored various sensing methods using a wrist-mounted form factor, including cameras [3, 13, 28], radio frequency (RF) antennas [12], electromyography (EMG) [22], radar [7], active sonar [15, 19], and motion sensing [44, 46]. However, all these prior works either focus solely on recognizing discrete hand gestures (rather than continuously tracking hand poses) [15, 44, 46] or require additional hardware [3, 9, 12, 13, 17, 19, 22] that is not available on millions of existing commercial off-the-shelf (COTS) smartwatches. As a result, continuous hand pose tracking remains only available on customized hardware or in research prototypes. We believe that enabling continuous hand pose tracking on COTS smartwatches without requiring additional hardware would significantly advance research in this area by immediately unlocking a wide range of downstream applications that could substantially improve end-user interaction experience.

The primary challenge of continuous hand pose tracking lies in capturing detailed information about all finger joint movements and hand dynamics. Recent work by Lee et al. [19] demonstrated the feasibility of using two speaker-microphone pairs to form micro-sonars on a customized wristband, enabling ultrasound-based tracking of finger movements. Inspired by this, we observed that modern smartwatches typically contain at least one speaker and one microphone, which could form a single micro-sonar. We thus hypothesize that *repurposing the built-in sensors can provide sufficient information to continuously track finger joint movements and hand poses on COTS smartwatches.* In this paper, we aim to evaluate this hypothesis by addressing several key challenges that must be tackled to adapt a prior proof-of-concept continuous hand pose tracking sonar implementation on bespoke hardware [19] into a practical and deployable solution for existing COTS smartwatches.

- **Hardware configuration**: Commercial smartwatches typically include only a single speaker-microphone pair, often placed on opposite sides of the case (see Figure 2). It remains an open question whether a single speaker-microphone pair on COTS smartwatches can capture sufficient reflections for fine-grained hand-pose estimation. Moreover, a practical sensing system needs to support different smartwatch models, varying in microphone and speaker quality, placement, and design.
- **Robustness to device re-wearing**: When users remove and re-wear the device, small shifts in placement or orientation can alter sensor data patterns, leading to variability in tracking performance [3, 9, 17]. A practical sensing system should provide reliable hand pose tracking performance even after the watch is remounted.
- **Real-world variability**: A practical sensing system must also remain robust under different wearing habits, postures, noises, and motion patterns.
  - Accounting for different watch-wearing hand preferences (e.g., left/right hand), which reverses the spatial orientation of the speaker and microphone.
  - Supporting varied body postures, such as arm-raised to view the screen or arm-resting naturally beside the body, which may affect signal reflections and thus hand pose tracking performance.
  - Maintaining robustness under challenging conditions such as environmental (e.g., loud music or nearby human movement) or user-induced noises (e.g., walking conditions or altered hand poses).
  - Covering dynamic hand pose variations, including both pose-neutral-pose (i.e., returning to a fixed neutral pose between pose changes) and direct pose-to-pose transitions, as well as varying finger movement speeds.

Despite many pioneering efforts, these real-world variations have not been fully addressed in prior work, yet they represent critical challenges towards building robust, generalizable, and scalable continuous 3D hand pose tracking systems that can make an immediate impact on millions of COTS smartwatches.

In this paper, we aim to tackle the above challenges and lower the barrier of continuous hand pose tracking on wrist-mounted devices by presenting WatchHand, the first continuous hand-pose tracking system for COTS smartwatches that relies solely on the built-in speaker and microphone. Our approach leverages the cross-correlation-based frequency-modulated continuous waves (C-FMCW) method [42]. The system emits inaudible FMCW signals (18–21 kHz) through its speaker into the air and captures reflections caused by hand poses via its microphone. The signals are then analyzed with a cross-correlation-based method to generate spatiotemporal echo profiles. These profiles are fed into a deep learning model to continuously estimate 3D hand poses. To evaluate WatchHand, we adopted established study protocols from prior work that employed custom arrays of speakers and microphones [19, 48]. To demonstrate WatchHand's real-world applicability, we ran a first

study (N = 24) that tested it on three recent COTS smartwatches from three different manufacturers (Samsung, Xiaomi, and Google), which were worn on either hand. We then conducted a follow-up second study by re-inviting 6 participants and investigating system performance across varying body postures. In the third study (N = 8), we assessed robustness under various noise conditions, including loud music, nearby human motion, walking, and altered hand poses. Finally, we conducted a fourth study (N = 8) to evaluate the system's capability in tracking more dynamic hand pose variations. These studies collectively sought to establish performance in diverse real-world conditions.

The results show WatchHand reliably tracks 3D hand poses, achieving a mean per-joint position error (MPJPE) of 7.87 mm in a cross-session model and 14.88 mm in a cross-user model across 20 finger joints spanning different smartwatch models and watch-wearing hands. The system also remains robust under variations in arm and body postures and in noisy conditions with only a single session (2 minutes of data) of fine-tuning. Furthermore, WatchHand covers dynamic hand pose variation (mean MPJPE of 15.48 mm) in cross-session testing. To provide a comprehensive view of system performance, we further extended our analysis to within-session and cross-user protocols, enabling a direct comparison with prior works and highlighting WatchHand's robustness across diverse evaluation settings. These findings confirm that WatchHand provides promising performance even after the device is remounted, while revealing expected drops in cross-user performance or when exposed to dynamic gesture variations. Nevertheless, the system adapts effectively to different body postures and noise conditions, highlighting its practicality and suitability for real-world hand-tracking applications. The dataset collected across four studies with 40 participants is publicly available at https://github.com/witlab-kaist/WatchHand. Consequently, by leveraging only pre-existing built-in hardware, WatchHand demonstrates the feasibility of enabling continuous hand-pose tracking on COTS smartwatches. Our contributions include:

- Significantly lowering the barrier for continuous hand pose tracking on wrist-mounted devices by being the first to enable and demonstrate a continuous 3D hand pose tracking system developed exclusively using the built-in speaker and microphone on off-the-shelf smartwatches, requiring no external sensors or hardware modifications.
- The design and evaluation of deep learning models for hand pose estimation with multiple training strategies, including within-session, cross-session, and cross-user approaches to support broad comparisons with prior work.
- Extensive empirical evaluations using three widely used smartwatch brands (Samsung, Xiaomi, and Google). The promising performance exhibited across diverse scenarios (e.g., different watch models, wearing hands, body postures, noise conditions, and hand pose variation protocols) demonstrates WatchHand's real-world feasibility. We released the dataset, comprising a total of 35.6 hours of hand pose variations captured from 40 participants across all studies, to support further research in this area.
- Discussion of key insights toward generalization across diverse hand poses, unseen users, and use contexts, along with practical considerations, such as power consumption, privacy, system deployment, and application designs that can support real-world adoption.

2 Related Work

2.1 Active Acoustic Sensing

Active acoustic sensing has emerged as a promising technique for around-device interaction by analyzing reflected or diffracted inaudible sound waves emitted from speakers to detect user gestures or movements. Early work such as SoundWave [6] leveraged Doppler shifts and a laptop's built-in speaker and microphone to recognize mid-air gestures. Similarly, FingerIO [28] and LLAP [43] explored orthogonal frequency-division multiplexing (OFDM) or multiple continuous wave signals on customized smartwatches or commercial smartphones equipped with two pairs of microphones for precise near-device finger tracking. Building on these initial demonstrations, researchers have implemented richer acoustic sensing paradigms on various wearable devices, such as wristbands [18, 19, 23], smartwatches [15, 16, 50], rings [48], smartglasses [24, 25, 51], and earbuds [37]. This research has sought to enable more expressive and context-aware interaction paradigms, such as discrete gesture recognition [15, 18, 37], continuous hand pose tracking [19, 48], silent speech recognition [50, 51], and activity recognition [25].

Collectively, these studies highlight the potential of active acoustic sensing with a simple hardware configuration (using only speakers and microphones) to realize diverse and expressive interaction techniques. However, while their achievements are impressive, their reliance on specialized, bespoke hardware designs—often requiring multiple carefully positioned speaker and microphone pairs [19, 25, 28, 51]—limits practical integration into commercial devices. We argue that work in this area needs to move beyond proof-of-concept prototypes to tackle the real-world deployment challenges, such as relying on single, suboptimally positioned speakers and microphones (e.g., distant, opposite-facing). WatchHand directly addresses these limitations, utilizing only built-in smartwatch hardware to tackle the challenging task of 3D hand pose tracking. This advances the state of the art by demonstrating how active sonar systems can be implemented on current commercial wearable devices.

2.2 Hand Pose Input on Commercial Smartwatches

Smartwatch hand pose input involves using actions of the watch-wearing hand to issue commands or respond to events. It is a compelling paradigm as it inherently enables single-handed use [5, 11, 44, 46], and rich expressivity through the high dexterity of our hands [3, 7, 12, 19, 22]. However, sensing these fine-grained actions remains challenging. Prior work has predominantly focused on leveraging motion signals captured by a watch's IMU. For example, Serendipity [44] proposed a discrete five-class hand-pose recognition system using watch motion data with an average 87% F1-score. Xu et.al. [46] further introduced a user-adaptive method for personalizing smartwatch-based pose input that enabled detection of a more substantial set of discrete poses (e.g., 12 to 16) with

Table 1: An overview of existing wrist-worn continuous hand pose tracking systems. The Pose-N-Pose protocol denotes returning to a neutral (default) pose between variations, while Pose to Pose signifies direct transitions between target hand poses. Cross-session evaluation refers to session-independent testing with device re-mounting [9, 19] or across different days [17]. Errors are reported in MPJPE (mm), except for * denoting MPJAE (°).

Related Work	Sening Modality (# of sensors)	Study Protocol (# of Ref Poses)	Within Session Error	Cross Session Error	Cross User Error
FingerTrak [9]	Thermal cameras (4)	Pose-N-Pose (19)	12 mm	27.2mm	87 mm
WR-Hand [22]	EMG sensors (8), gyro (1)	Pose-N-Pose (12)	25.7 mm	-	31.3 mm
EchoWrist [19]	Speakers (2), mics (2)	Pose-N-Pose (15)	-	4.81 mm	12.2 mm
WatchHand	COTS smartwatch's built-in speaker (1), mic (1)	Pose-N-Pose (15)	6.06 mm	7.87 mm	14.88 mm
		Pose to Pose (15)	12.68 mm	15.48 mm	22.60 mm
Digits [13]	IR camera (1), IR laser (1)	Pose to Pose (10)	2° - 9°*	-	-
DiscoBand [3]	Depth cameras (8)	Pose to Pose (10)	11.69 mm	17.87 mm	19.98 mm
EtherPose [12]	RF antennas (2)	Pose to Pose (11)	11.57 mm	-	-
EITPose [17]	EIT electrodes (8)	Pose to Pose (12)	11.06 mm	17.81 mm	18.91 mm

up to 92.1% F1 score. Tsunoda et al.[39] detected 11 thumb-to-finger gestures with 90.2% accuracy using the watch's accelerometer.

While these results are promising, their reliability in real-world applications remains in doubt, and commercial realizations of these techniques have tended to focus on highly limited gesture sets composed of simple poses such as thumb-index pinches [1] or making a fist [29] or rely on integrating additional sensors (e.g., such as PPG [52]). We argue that the full potential of hand pose input as an expressive modality is inadequately realized by current motion sensing implementations. To fully unlock richer and expressive interactions, we need new systems that explore alternative sensing channels available on commercial devices, such as the sonar used in WatchHand.

2.3 Continuous Hand Pose Tracking on Wrist Wearables

Continuous hand pose tracking enables expressive, seamless, and context-aware interactions, making it a valuable modality for wearable computing. Existing approaches have explored various around-device sensing techniques integrated into wrist-worn wearables, such as camera [3, 9, 13, 45], radar [7], sound [19], RF antenna [12], proximity [5], or electromyography (EMG) [22] to track our hand movements. Camera-based methods provide high precision but face significant practical constraints. They generally require dedicated space for camera placements, restrict interaction area to the camera's fields of view, are sensitive to lighting changes, and require a high computational cost for real-time image processing. Moreover, camera usage in wearables raises privacy concerns, particularly in public settings [20, 31]. Alternative sensing modalities have thus been proposed to mitigate these issues. EtherPose [12] tracks hand poses using wrist-worn RF antennas, while WR-Hand [22] analyzes EMG signals captured from forearm muscles.

However, these solutions require external sensors or specialized hardware, which are cumbersome and impractical for daily use. For instance, EMG-based systems require electrodes in constant skin contact with tight-fitting wristbands, impacting user comfort [22]. Similarly, antenna-based [12] and radar-based methods [7] involve additional bulky components that preclude seamless integration. Moreover, these systems often require careful sensor positioning, as slight shifts in placement can distort sensor signals or substantially degrade tracking reliability [3, 9, 17], which is unrealistic in everyday use, where users frequently remove and re-wear devices. Recently, EchoWrist [19] introduced a wristband-based active-sonar system using two carefully positioned pairs of speakers and microphones, achieving accurate hand pose estimation (MPJPE of 4.81 mm) and enabling hand-object interaction, even across device re-mounting sessions. However, its reliance on bespoke hardware configurations—requiring spatial separation across both sides of the wrists—makes it impractical for the existing commercial smartwatches.

Table 1 provides an overview of wrist-worn continuous 3D hand pose tracking techniques explored in related works. Based on this analysis, to the best of our knowledge, there is no existing continuous hand pose estimation system that is a viable deployment candidate for current (and likely future) smartwatch form factors. We propose WatchHand as the first solution that achieves continuous hand tracking on unmodified commercial smartwatches.

3 System Design and Implementation

3.1 Design Consideration and Principle

The main research question this paper aims to answer is: *Can we enable an off-the-shelf smartwatch to continuously track hand poses?* This is a highly challenging task, as it requires accurately estimating the 3D positions of 20 finger joints using only the limited sensing utility of COTS smartwatches. To address this problem, we examine the most common sensors embedded into modern commercial smartwatches, including the microphone, speaker, and IMU. To track hand poses, the key to sensing is capturing high-quality information on the detailed motion and shape of the hand and fingers. The most obvious sensor to track motion is the IMU, which has been explored extensively and used by prior work to recognize discrete hand gestures by learning the fingerprints of the motion on the wrist resulting from performing different hand gestures [39, 44, 46]. However, the motion of the wrist does not contain enough information to infer the postures of all 20 finger joints. A more capable sensing modality is needed, one that can capture richer information on finger articulation. Recent work [19] has shown that two

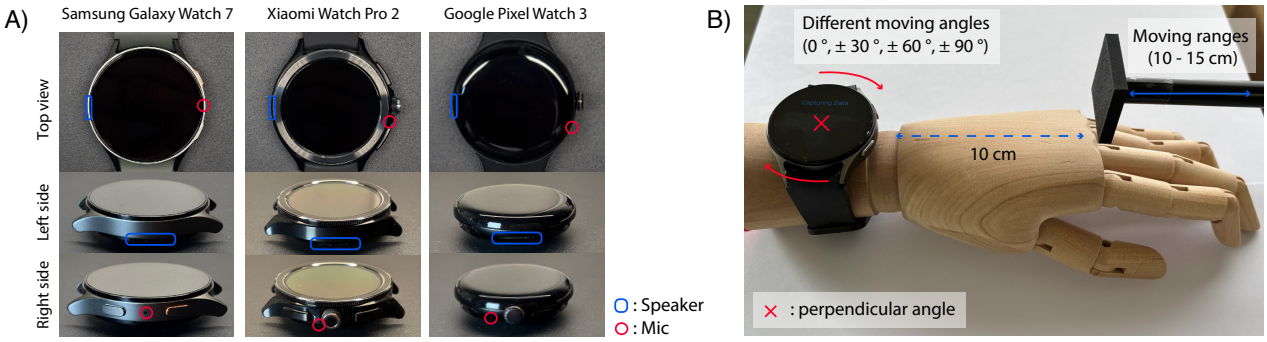


Figure 2: (A) Three different COTS smartwatches we evaluated and the physical locations of their built-in speaker and microphone, and (B) system evaluation setup: A smartwatch is worn on a prop hand while a stepper motor–driven linear stage moves a flat plate back and forth within typical finger movement ranges (10–15 cm). To evaluate different angles, the prop hand was rotated to various orientations (e.g., 0°, ±30°, ±60°, ±90°, and perpendicular) relative to the direction of plate motion.

pairs of microphones and speakers with specific configurations on a wristband can be used as micro-sonars to capture and infer accurate hand poses. Building on this insight, we hypothesize that *repurposing the built-in microphone and speaker as a micro sonar can provide sufficient information to continuously track finger joint movements and hand poses on commercial smartwatches.*

To explore this idea, we implemented a system that leverages active acoustic sensing with a built-in speaker and microphone on COTS smartwatches to capture fine-grained finger movements. Our system emits encoded acoustic waves, and the microphone captures the reflected and diffracted signals, essentially functioning as an airborne sonar system. The emitted acoustic waves propagate through the air toward the hand; when they encounter the hand, a portion of the energy is reflected back toward the source, specifically the microphone, as an echo. The system then measures the time elapsed between the emission and the reception of the echo. To enhance the ranging resolution and continuously capture the complex spatial and temporal reflection strengths around the hand, we employed the C-FMCW technique [42], as demonstrated in prior work [19, 48]. This technique estimates the round-trip time of the acoustic signal by identifying the correlation between the transmitted and received signals. The resulting spatiotemporal reflection patterns, encoding hand geometries, are subsequently used to infer hand poses. More detailed signal processing is described in Section 3.3.

3.2 Data Collection System

We implemented a sensor data capture system on three different COTS smartwatches running on WearOS: Samsung Galaxy Watch 7, Xiaomi Watch 2 Pro, and Google Pixel Watch 3. These devices feature 425 mAh, 495 mAh, and 420 mAh Li-ion batteries, respectively. All three smartwatches feature a left-side speaker and a right-side microphone when worn on the right dorsal wrist and support simultaneous 48 kHz, 16-bit PCM audio playback and recording [14]. While the speaker is consistently located at the center-bottom of the left side across all three models, the microphone placement varies slightly—centered on the right side in the Galaxy Watch, and shifted slightly toward the left and bottom on the Xiaomi and Pixel watches (see Figure 2). The position will be opposite if the watch is worn on the left wrist. Different sensor positions will significantly impact the paths along which acoustic signals are reflected, a factor we explicitly explore in this paper.

Our system emits FMCW signals in the inaudible range (18–21 kHz) and simultaneously captures reflected and diffracted signals from hand movements. To characterize the output levels across devices, we measured the playback sound levels using a sound level meter positioned 10 cm from the speaker. The Galaxy Watch 7 emitted 61.6 dBA, the Xiaomi Watch 2 Pro produced 66.9 dBA, and the Pixel Watch 3 emitted the highest amplitude at 80.8 dBA. To ensure consistently characterized signals across different hardware, our system records *UNPROCESSED* signals using the Android MediaRecorder API [1]. Additionally, we recorded 6-axis IMU data at the maximum available sampling rate on each device (Galaxy Watch: 100 Hz; Xiaomi Watch: 50 Hz; Pixel Watch: 200 Hz) to explore its potential contribution to hand pose tracking.

3.3 Acoustic Data Preprocessing

3.3.1 C-FMCW Based Echo Profiles. We employ airborne active acoustic sensing based on the C-FMCW approach [42]. In the conventional Linear FMCW method, the theoretical distance estimation, or range-finding, resolution (R_{L-FMCW}) is restricted by the sweep bandwidth (B) under the speed of sound in air ($C = 343m/s$). Using our operational bandwidth, the resolution would be 57.2 mm ($= \frac{C}{2B} = \frac{343m/s}{2\times(21000-18000)} = 0.0572m$) [36]. However, by leveraging the C-FMCW technique, the effective theoretical distance estimation resolution (R_{C-FMCW}) is instead restricted by the system's sampling rate (F_s), allowing for a substantial increase in spatial precision:

$$R_{C-FMCW} = \frac{C}{2F_s} = \frac{343m/s}{2\times(48000)} = 0.00357m = 3.57mm \quad (1)$$

[1] https://developer.android.com/media/platform/mediarecorder

Algorithm 1 Echo Profile Generation

Require: Rx: Raw received audio signal in the full recording $(N_{samples},)$; $N_{samples}$: The length of Rx; Tx: Reference transmitted FMCW signal of one chirp, $(L,)$; L: Frame length, the samples per chirp, which is (600); F_s: Sampling rate, which is (48000); $[f_{min}, f_{max}]$: Bandpass frequency range, the frequency range of the transmitted signal, which is $(18000, 21000)$; N_{order}: Butterworth bandpass filter order, which is (5) N_{frames}: The number of total echo frames.

Ensure: P: Original Echo Profiles (L, N_{frames}); P_{diff}: Differential Echo Profiles $(L, N_{frames} - 1)$

Step 1: Data Formatting and Filtering

$Rx \leftarrow \text{ButterworthBandpassFilter}(Rx, f_{min}, f_{max}, F_s, N_{order})$ ▷ Filter to match transmitted signal bandwidth

Step 2: Start Position Detection

$Corr_{sync} \leftarrow Rx \star Tx$ ▷ $\star$ denotes Cross-Correlation

$p_{\text{start}} \leftarrow \arg\max(|Corr_{sync}|)$ ▷ Find index of strongest correlation

$p_{\text{start}} \leftarrow (p_{\text{start}} + L - \lfloor L/2 \rfloor) \mod L$ ▷ Adjust to frame center

$Rx \leftarrow Rx[p_{start}:,:]$ ▷ Align to start position, Rx shape: $(N_{samples} - p_{start},)$

Step 3: Original Echo Profile Computation

$y \leftarrow Rx \star Tx$ ▷ $\star$ denotes Cross-Correlation, y shape: $(\text{Length}(Rx) + L - 1,)$

$N_{frames} \leftarrow \lfloor \text{Length}(y)/L \rfloor$ ▷ Calculate N_{frames} with L

$P \leftarrow \text{Reshape}(y, [N_{frames}, L])$ ▷ Reshape into frames

$P \leftarrow (P)^T$ ▷ Transpose to (Distance, Time), P shape: (L, N_{frames})

Step 4: Differential Echo Profile Computation

for frame $f = 1$ to N_{frames} **do**

  $P_{diff}[f] \leftarrow |P[f]| - |P[f-1]|$ ▷ Calculate frame-to-frame magnitude difference

end for ▷ P_{diff} shape: $(L, N_{frames} - 1)$

This marked improvement in resolution is critical for the fine-grained hand pose tracking required by our system. Prior work has shown that C-FMCW can support continuous hand pose tracking with a single speaker-microphone pair on a custom ring [48]. We apply this technique in our system and detail the signal processing pipeline in Algorithm 1. Once activated, our system continuously emits and records FMCW signals in the 18-21 kHz band. The received signals are first passed through a 5th-order Butterworth bandpass filter with cutoff frequencies that match our emission band to eliminate out-of-band noise. Since the direct path between the speaker and microphone (without reflection) represents the shortest propagation path, and thus the earliest arrival, we use it as a reference for absolute distance. To locate this direct-path arrival, we calculate the cross-correlation between one transmitted FMCW sweep and the filtered received signals. The timestamp of the maximum correlation peak identifies the start of the received sweep, corresponding to the direct path, and serves as a reference for subsequent signal alignment. We then process the received signals in segments corresponding to each FMCW sweep, where each sweep consists of 600 samples (L). For every sweep, we calculate the cross-correlation between the transmitted and received signals. The resulting values encode reflection strength across different time lags, which we map to the distances using the speed of sound. This distance-mapped vector forms a slice of our echo profile, which we call an *echo frame*. Temporally stacking these slices forms the complete *original echo profile*. In the original echo profiles, each pixel along the x-axis represents the duration of one frequency sweep, which is 12.5 ms ($= \frac{L}{F_s} = \frac{600}{48000} = 0.0125s$). Along the y-axis, which consists of 600 pixels representing the sensing range, each pixel corresponds to a distance of approximately 3.57 mm [42, 48]. The intensity of each pixel represents the strength of the correlation at that specific time/distance.

3.3.2 Differential Echo Profiles. The original echo profiles preserve the hand shape information through the reflection strengths over time and distance, but they also include static reflections from the nearby surfaces and environments. To suppress these static components and highlight motion, we introduce *differential echo profiles* by computing the frame-to-frame differences in the original echo profiles along the time axis, thereby capturing the temporal variations in acoustic reflections. Specifically, we subtract each preceding echo frame from the current one. This process helps attenuate static reflections while emphasizing moving elements, enhancing the system's ability to track finger movements.

To focus on hand reflections and reduce environmental interference, we only retain the 60 closest pixels along the vertical axis of the echo profiles, corresponding to a sensing range of 21.42 cm. This range covers the typical wrist-to-middle-fingertip distance (average male: 19.05 cm, female: 18.29 cm [38]). Along the horizontal axis, we use a 1.2-second window, which corresponds to 96 pixels. Altogether, combining the original and differential echo profiles, this results in two channels of 60×96 acoustic input to our models (see Figure 1-B). The echo profiles within each window are normalized prior to being fed into the model to ensure consistency across different conditions.

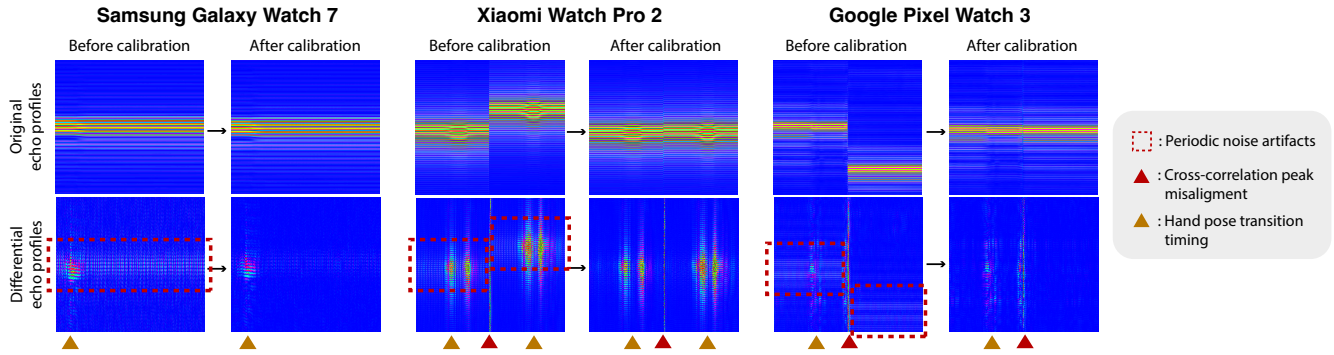


Figure 3: Visual examples of echo profile calibration across three commercial smartwatches. Sliding-window cross-correlation peak correction removes peak misalignment (red triangles), while periodic drift calibration mitigates repeating noise artifacts (red dashed boxes). Hand pose transition timings are marked (yellow triangles). Together, these calibration steps produce cleaner and more temporally stable echo profiles during dynamic hand pose transitions.

To enhance signal clarity and highlight salient patterns, we clip echo profiles to a threshold of $\pm 10^{10}$ for visualization. This threshold is determined empirically to best illustrate the finger motion patterns. All visualized echo profile figures in this paper use these clipped versions, offering a clearer view of the extracted features. However, to preserve maximum information for the deep-learning model, we use the raw unclipped echo profiles during training.

3.3.3 Echo Profile Calibration for Different Watches. Ideally, the strongest cross-correlation peak, representing the fixed direct path between the smartwatch's microphone and speaker, should appear at a consistent position in each echo frame in the same recording. Reflections, inherently weaker, would then be analyzed relative to this point. However, our pilot study revealed a noticeable drift in the position of this dominant peak across consecutive echo frames in some recordings, especially for the Xiaomi and Pixel Watch (see Figure 3). We attribute this to device-specific timing jitter in the audio stack: the intermittent delay between chirp repetitions. This misalignment can shift echo profiles to incorrect ranges, disrupting the pattern, and leading the model to make entirely wrong predictions. To resolve this problem, we adopted a simple yet practical approach, a sliding-window-based correction method. This method dynamically estimates the peak location within overlapping windows and realigns segments exhibiting substantial shifts.

A further challenge was a periodic drift of echo frames within the original echo profiles. This drift appeared as periodic artifacts in the differential echo profiles. Several factors may contribute to this issue, including subtle periodic variations in the effective sampling rate and modulated aliasing effects. To mitigate these periodic artifacts, we implemented a median filter on the original echo profiles. The second row of Figure 3 shows that this filtering approach effectively eliminates the subtle drifts while preserving temporal consistency in the original data, consequently removing the artifacts from the differential echo profiles and ensuring cleaner signal representation.

We found that the Xiaomi and Pixel Watch exhibit more misalignment issues, while the Galaxy Watch shows more drifting issues. These device-specific inconsistencies in echo profiles can significantly degrade performance, especially when training on data pooled across different brands. Our calibration process can effectively address these issues with a universal, hardware-agnostic algorithm that corrects these misalignment and drifting issues. Importantly, all echo profiles from all devices are processed using the identical calibration process, with no per-watch or per-model specialization.

3.4 IMU Motion Data Preprocessing

To enhance feature extraction from IMU data while reducing noise, we adopted a data preprocessing method from prior work [46]. The raw 6-axis IMU data, comprising 3 axes of accelerometer data and 3 axes of gyroscope data, yields 6 initial data channels. Each channel undergoes processing through a 2nd-order Butterworth bandpass filter with three cutoff frequency ranges: 0.22-8 Hz, 8-32 Hz, and above 32 Hz. This filtering approach separates different frequency components of the signal. By preserving the original raw data and adding these three filtered versions, each original channel expands into 4 channels. Consistent with the acoustic data processing, we employ a window of 1.2 seconds for the IMU data, and the data within each window is normalized prior to being fed into the model.

3.5 Preliminary Viability Testing on COTS Watches

To demonstrate the viability of our system for 3D hand pose tracking on COTS smartwatches, we visually examined processed data to assess how hardware configurations and watch-wearing hands affect echo profiles, whether a single built-in speaker–microphone pair can reliably capture 3D around-device movement, and how echo profiles can support reliable 3D hand pose tracking.

3.5.1 Impact of hardware and watch-wearing hand. We first investigated how differences in smartwatch hardware and the wearing hand affect captured echo profiles. To simulate consistent around-device movement, we used a stepper motor-driven linear stage to move a flat plate back-and-forth at 6 cm/s over a prop hand wearing

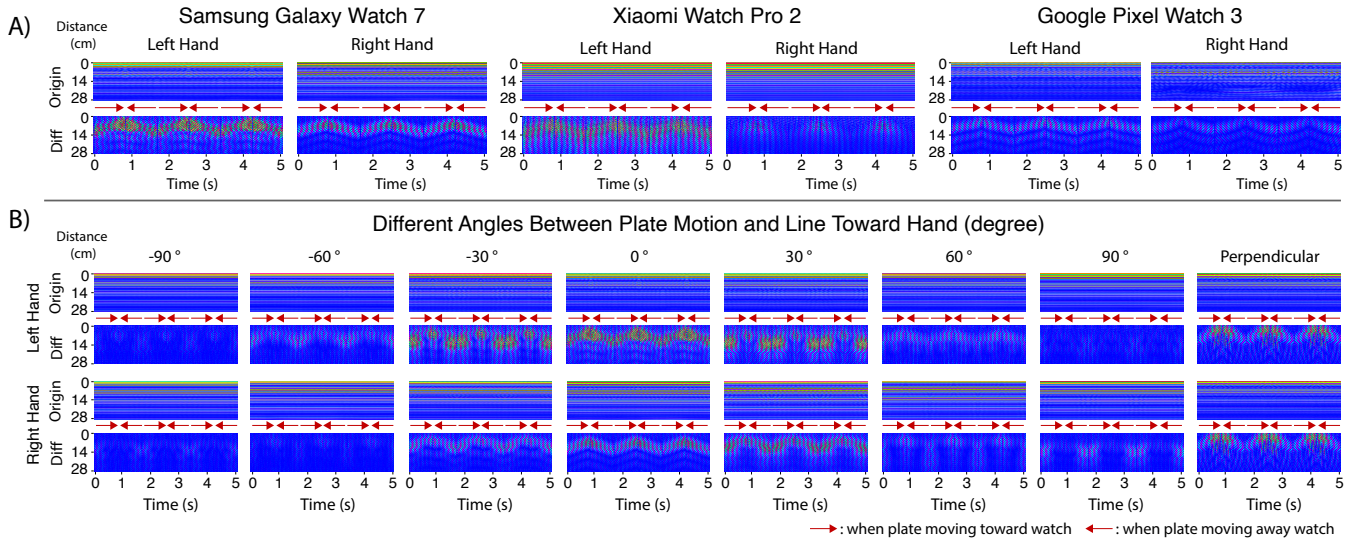


Figure 4: Captured original and differential echo profiles across different COTS smartwatches (Galaxy, Xiaomi, and Pixel) (A) and varying angles between the plate's motion and line toward the hand (B) during repetitive back-and-forth movements toward and away from the smartwatch.

a smartwatch, within a 10–15 cm range—an area corresponding to typical finger movement (see Figure 2-B). Visual inspection of the resulting differential echo profiles (Figure 4-A) shows that while all watches reliably capture distinct motion patterns, the signal characteristics vary across devices and hand orientations. For instance, the Galaxy Watch exhibits clear zigzag patterns, the Xiaomi Watch produces broader profiles, and the Pixel Watch shows narrower responses. Differences in signal intensity were also observed between left- and right-hand configurations (which flip the watch, and thus the orientations of the microphone and speaker). This highlights the importance of accounting for real-world variability in both hardware and watch-wearing hands.

3.5.2 3D Around-Device Movement Tracking Viability. Next, we investigated the capability of active acoustic sensing on a COTS smartwatch to support 3D around-device movement tracking. While SonarSelect [16] demonstrated robust sensing of one-dimensional motion using sonar on a COTS smartwatch, other recent studies [15, 50] have shown that more sophisticated around-device movement (e.g., lip motions for silent speech recognition) can be detected using only a single built-in speaker and microphone. To evaluate this capability in the context of 3D hand pose tracking, we visually examined how changes in movement direction affect the captured echo profiles. Specifically, we varied the angle between the motion path of a moving plate and the long axis of the hand/wrist from −90° to 90° in 30° increments and additionally tested a perpendicular movement angle involving motions towards and away from the palm (see Figure 2-B). Figure 4-B illustrates the results. It shows distinct patterns corresponding to different around-movement angles. For example, while a clear straight zigzag pattern appears when the movement is directed toward the hand (0°), as the angle changes, the echo profiles vary, showing discontinuities at ±30° on the left-hand, curved signals at ±30° on the right-hand, signal shifts when the movement is perpendicular to the watch screen, and weaker signals when the angle exceeds 60° for both hands. These observations highlight that the captured echo profiles contain rich 3D around-device movement information, which can potentially be leveraged to enable continuous 3D hand pose tracking.

3.5.3 3D Hand Pose Tracking Viability. Given these insights, we further explored the viability of continuous hand pose tracking across different hand configurations. Figure 5 presents samples of processed acoustic and motion data recorded by our system for various hand poses performed by a single user. Each pose was executed by transitioning from an open-hand position to the target pose and then returning to the open-hand position. A visual examination of the sensor data indicates that different hand poses generate sufficiently distinct patterns, supporting the potential for continuous hand pose tracking. Most poses exhibit noticeable variations, except for ASL-5, where the hand remains open with minimal movement. Additionally, wrist rotation poses produce clearer patterns in the differential echo profiles. Collectively, these observations confirm the viability of 3D hand pose tracking on COTS smartwatches using only their built-in sensors.

3.6 Deep-Learning Pipelines

3.6.1 Ground Truth Acquisition & Normalization. Since our sensing technique is sensitive to the physical geometry of the hand, relying on marker-based systems potentially distorts the signals. To establish reliable ground truth for continuous 3D hand pose information without altering the hand's acoustic reflective properties, we employ MediaPipe Hands [49]. Although monocular RGB cameras can suffer from reduced accuracy when finger joints are occluded, we minimized such occlusions by positioning the camera

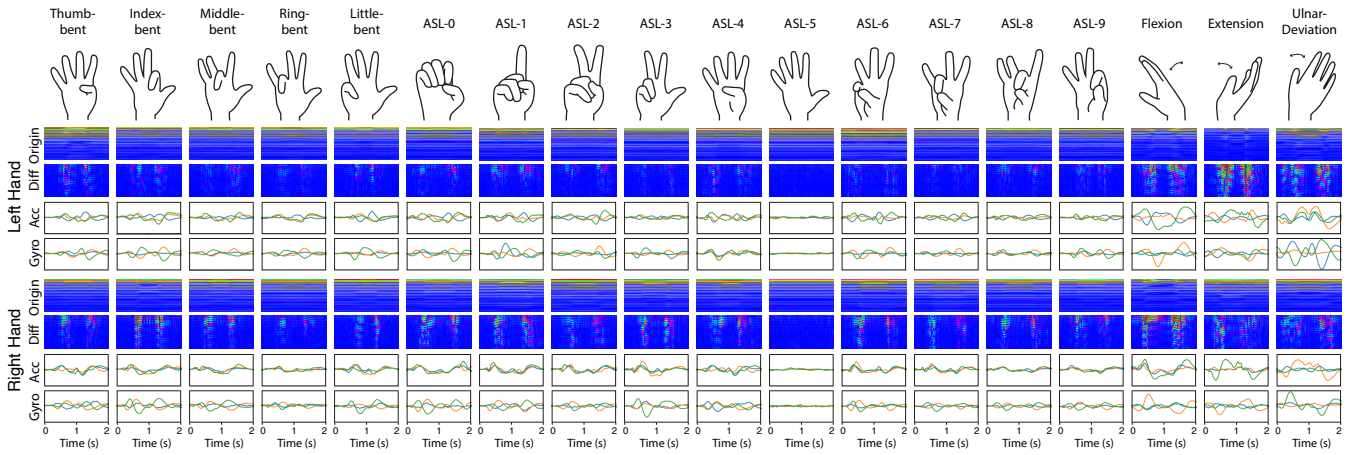


Figure 5: Processed original and differential echo profiles (see Section 3.3) and bandpass-filtered IMU sensor data in the 32-100 Hz range (see Section 3.4) captured from a single user using a COTS smartwatch (Galaxy Watch 7) during different hand postures.

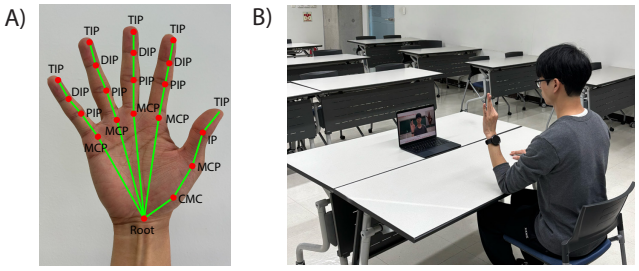


Figure 6: (A) MediapPipe-based 3D hand joint annotation used as ground truth, showing 20 labeled landmarks from the root to each fingertip. (B) Study setup in which the laptop's front camera faces the participant's palm to capture ground truth without optical occlusion during hand pose demonstrations.

to face the user's palm directly (see Figure 6). This RGB camera-based approach has been widely adopted by prior work in wearable hand tracking for generating ground truth [3, 12, 19, 41, 48]. Hand poses are captured at 30 frames per second (FPS) using a built-in webcam on a MacBook Air. Processing webcam-captured video footage with MediaPipe Hands produces 3D hand representations consisting of 21 hand joint coordinates in 3D space. We designate the wrist coordinate as the reference point (origin) of our coordinate system and calculate the relative positions of the remaining 20 landmarks by subtracting the wrist coordinate from each point. This approach provides a consistent frame of reference centered on the wrist position. We normalize the 3D hand pose ground truth to address session-to-session and user-to-user differences in hand position, orientation, and size. For each generated frame, we first define the palm plane using vectors originating from the wrist and extending to the metacarpophalangeal (MCP) joints of the index and little fingers. We then calculate a rotation matrix to align this palm plane with a reference palm plane derived from a baseline posture. This rotation correction compensates for the inevitable variations in hand orientation that occur between different recording sessions and when users remount the device. Additionally, for each user, we normalize the hand size by scaling each frame so that the distance between the wrist and the little finger MCP joint matches a measured physical length. Note that the wrist rotation is not normalized to preserve its natural variation. This normalization method ensures that generated hand pose data remains comparable across different recording sessions and users.

3.6.2 Deep-learning Model Framework. We construct a regression model that learns the mapping from acoustic data to 3D hand poses. Each ground truth video frame serves as an individual sample at 30 FPS, with echo profiles mapped to ground truth using timestamp alignment. To support continuous hand pose tracking, all the frames are used, including the transition segments. We employ the FastViT-T12 [40], a state-of-the-art hybrid model that combines the efficiency of convolutional neural networks (CNNs) with the powerful feature mixing capabilities of vision transformers. In this process, we also explored ResNet-18 [8] and CNN-LSTM that have been explored in closely related prior works [18, 19, 48], and we chose the best-performing and most lightweight model. The model takes a two-channel tensor as input, comprising original and differential echo profiles. The architecture processes this input through a convolutional stem for initial $4x$ downsampling, followed by four hierarchical stages that progressively refine features using RepMixerBlocks. Between stages, a PatchEmbed layer further downsamples the feature maps. A final regression head uses global average pooling and a linear layer to map the learned features to the 60 output parameters representing the 3D coordinates of 20 hand landmarks (see Figure 1). These landmarks can be used directly to reconstruct the wireframe of the hand as shown in Figure 6. We

subsequently generate all the visualizations using MANO [32]. The model is trained end-to-end using a composite loss function that promotes both positional accuracy and temporal smoothness. This loss is a weighted sum of the standard Mean Squared Error (MSE) between the predicted and ground truth coordinates, and a velocity loss. The velocity component, weighted by a factor of 0.1, penalizes abrupt changes between consecutive frames, encouraging the model to produce more physically plausible motion. Overall, the model has 6.62M parameters.

To explore whether IMU signals could complement acoustic sensing, we extended our pipeline with a lightweight IMU encoder. Inspired by prior work on IMU-based full-body pose tracking [27], filtered motion signals (see Section 3.4) were first mapped to a 512-dimensional embedding via a linear layer, then processed using a bidirectional LSTM. The resulting features were concatenated with the acoustic features and passed through a simple fusion decoder to predict the 3D hand landmarks.

3.6.3 Training Pipeline. We use three protocols to train and evaluate our models:

- Cross-user models: We train cross-user foundation models using data from all participants except one, whose data was reserved exclusively for the test set. This protocol simulates the deployment scenario where a system is pre-trained on large-scale datasets without access to the end user's data.
- Cross-session models: User-specific models created by fine-tuning cross-user foundation models with target user data from all wearing sessions except the final session, which is reserved for testing. This protocol mimics realistic usage where users provide personalization data before regular system deployment.
- Within-session models: User-specific models that fine-tune cross-user foundation models using data from sessions up to the antepenultimate session plus a 50% subset from the final two sessions. The remaining 50% of data from the final two sessions serves as the test set. This protocol evaluates system performance when users collect calibration data after watch remounting, simulating scenarios where brief personalization occurs with every device repositioning.

We employed a consistent two-phase fine-tuning training strategy for the acoustic model to maximize the utility of our collected dataset. In the first phase, we develop a cross-user foundation model that captures generalizable hand pose tracking patterns across participants. This cross-user base model is trained using the Adam optimizer with an initial learning rate of 0.001, incorporating a learning rate scheduler to dynamically adjust optimization during training. Training continues for 100 epochs with early stopping if the validation metric does not improve for 9 epochs, utilizing a relatively large batch size of 256 to promote stable convergence. The second phase is user-specific cross-session fine-tuning. Starting with the pre-trained cross-user base model, we adapt the network to individual hand characteristics using user-specific data. This fine-tuning process employs the Adam optimizer with a reduced initial learning rate of 0.0003, allowing for more subtle parameter adjustments. We train for 100 epochs with a smaller batch size of 64, enabling the model to better capture user-specific nuances. This two-phase approach leverages our complete dataset efficiently, establishing broad hand-tracking capabilities in the base model while accommodating individual variations through targeted fine-tuning.

3.6.4 Data Augmentation. To enhance model robustness towards interference noise and mitigate overfitting, we implement data augmentation during training. While the original data is preserved, we introduce controlled randomness into copied data samples. For acoustic data, we apply three augmentation methods targeting different aspects of signal variability. First, we introduce random vertical shifts to the echo profiles, simulating variations in watch-wearing positions across different users and sessions. Specifically, we apply random shifts of ±5 pixels, corresponding to around 1.8 cm ranges to account for these natural positioning inconsistencies across users and watch-wearing sessions and enhance generalizability. Second, we incorporate amplitude variations to increase resilience to signal strength fluctuations. With 80% probability, we multiply echo profile values by a random factor between 0.95 and 1.05 (±0.5%). Third, to further challenge the model's ability to extract robust features, we applied time and frequency/distance masking. With 80% probability, we mask random sections: 5 to 15 pixels on the x-axis (time) and 5 to 10 pixels on the y-axis (distance). This introduces controlled randomness into the training process, challenging the model to maintain performance despite minor signal variations rather than memorizing exact patterns.

4 Study 1: Continuous Hand Pose Tracking

We conducted a study to evaluate the continuous 3D hand pose tracking performance of WatchHand across three different COTS smartwatches and different watch-wearing hands. This study was approved by the local Institutional Review Board (IRB).

4.1 Participants

We recruited 24 participants (12 males, a mean age of 22.5 (SD 2.3)) from the local university. 23 participants were right-handed. While most participants reported wearing a smartwatch on their non-dominant hand, two right-handed participants preferred wearing it on their dominant hand for reasons such as easier access to notifications or increased comfort. Participants' hand sizes averaged 17.95 cm (SD 0.95) in length (from the base of the hand to the tip of the middle finger) and 7.88 cm (SD 0.53) in palm width. The study took approximately two hours to complete, and participants were compensated with 30 USD in local currency.

4.2 Study Design

We followed an established methodology previously tested on customized hardware [19] to ensure the comparability of our results. Our study evaluated a total of 18 gestures (see Figure 5), categorized into three groups: (1) simple gestures (five single-finger bending movements), (2) complex gestures (ten complex finger gestures representing American Sign Language (ASL) digits 0–9), and (3) wrist orientations (three distinct wrist movements: flexion, extension, and ulnar deviation). Although the thumb-bent and ASL-4 poses represent similar poses and appear in both the simple and complex gesture sets, we retained them to enable direct comparison with prior work [19]. Since our system performs continuous

per-frame hand pose tracking rather than discrete gesture classification, this duplication has minimal impact on the validity of our results. Each session consisted of four repetitions per gesture, with each gesture lasting two seconds. We randomized the gesture order within each group to minimize potential order effects, but kept the group order (simple → complex → wrist rotation) fixed to reduce participant confusion. This design choice follows prior work [19], enabling direct comparison with their results. While prior work conducted 20 sessions per participant using only a single hand, our study introduced two independent variables: hardware (Samsung Galaxy Watch 7, Xiaomi Watch 2 Pro, Google Pixel Watch 3) and hand orientation (left, right). The hardware variable was between groups, while the hand orientation variable was within groups. We evenly distributed the 24 participants across the three smartwatch groups, with eight different participants performing all study tasks on each smartwatch. Each participant completed 20 sessions, with 10 sessions per hand, to ensure comprehensive data collection across different conditions. The order in which participants completed trials with the watch on their right and left hands was fully counter-balanced within each watch group.

4.3 Study Procedure

The study was conducted in an empty room with participants seated at a desk. To prevent physical fatigue, participants rested their elbows on the desk. Participants first signed an informed consent form, had their hand size measured, and read the experimental instructions. Before starting the experiment, they practiced all 18 different hand poses for five minutes. If some poses could not be performed due to ergonomic constraints, participants were instructed to approximate the pose. For example, 13 participants were unable to bend their little finger alone, so they bent the ring finger together (8) or rotated it toward the palm (5). Next, participants donned the smartwatch on the instructed wrist, and the study began. For each session, the target pose appeared on a screen in front of the participant for 2 seconds (see Figure 6-B), and they were instructed to perform the pose and then return to the default pose (ASL-5). The next target pose would then appear. Each session lasted approximately 3 minutes. Participants were instructed to remove and re-wear the smartwatch between sessions and were able to rest between sessions. This re-wearing process simulates real-world use, as users often remove and re-wear smartwatches. After completing 10 sessions, they switched the smartwatch to the opposite wrist and conducted the remaining 10 sessions. At the end of the study, participants completed a demographic questionnaire.

5 Results

We recorded a total of 1152 minutes of gesture data from 24 participants, consisting of 192 minutes for each of the six combinations of two independent factors: three smartwatch models and two wearing hands. During the study, while some participants at times performed incorrect gestures due to physical constraints or loss of focus, we included all recorded data, as our ground truth capture system reliably tracks the actual gestures performed, and our interest lies in continuous hand pose tracking. These natural variations enrich the dataset's diversity and contribute to the development of models that are more robust to continuous and unconstrained hand pose tracking scenarios.

To evaluate the performance of WatchHand for continuous hand pose tracking, we first segmented the collected gesture data into two distinct subsets, following prior work [12, 19]: (1) fine-grained finger movements, consisting of five simple gestures and ten complex gestures, and (2) gross wrist rotations, including three wrist rotation gestures. Then, we explored the performance result with respect to different smartwatch models, variations in watch orientation, and evaluation conditions (e.g., cross-session, within-session, and cross-user models). We processed all the acoustic and motion data following the procedures outlined in Sections 3.3 and 3.4, then trained the models in six conditions (three watches by two watch-wearing hands). In this result, for the acoustic model, the input consists of two channels of echo profiles (original and differential), each of size 96×60, as illustrated in Figure 1-C. The multimodal model additionally includes 24 IMU channels (six axes × four bands: raw plus three filtered bands). Importantly, the model is trained and evaluated on all frames in each recording, including both the instructed poses and the transitions into and out of those poses. We use a 1.2-second sliding window with a stride of 1/30 second for each step, so that we obtain one input window per video frame (30 FPS). This design explicitly targets continuous hand pose tracking rather than static pose classification.

As detailed in Section 3.6.3, we implemented and evaluated cross-session, within-session, and cross-user models, following a two-phase training pipeline. In this analysis, as all the results were normally distributed and met the upheld-sphericity assumptions, we conducted two-way ANOVAs on both metrics with different watch hardware and watch-wearing hand as independent variables.

5.1 Cross-Session Model

5.1.1 Continuous Fine-Grained Finger Movements. We evaluated system performance using two standard metrics: mean per-joint position error (MPJPE) and mean per-joint angle error (MPJAE). Across all conditions, the acoustic model achieved an average MPJPE of 7.87 mm (SD 1.67) and an MPJAE of 6.06° (SD 1.52). For both metrics, no significant main effects or interactions were found (F-values 0.005-0.42, p-values 0.66-0.95). These results confirm that WatchHand can reliably track continuous 3D hand poses for fine-grained finger movements across varying hardware and watch-wearing hands, using only acoustic data. By breaking down MPJPE and MPJAE metrics, Figure 7-A illustrates the average 3D hand reconstruction errors for each finger. While MPJPE errors generally correlate with finger length, MPJAE more clearly highlights the middle, ring, and little fingers as primary sources of angular error, corresponding to regions where unintended or co-activated movements most commonly occur (e.g., involving ring finger motion when instructed to bend the little finger solely, a physical constraint that varies across participants).

5.1.2 Continuous Wrist Rotation. Next, we evaluated our system's wrist-rotation tracking performance using the mean wrist angle error (MWAE) metric. The acoustic model demonstrated high accuracy in wrist-rotation tracking, achieving an average MWAE of

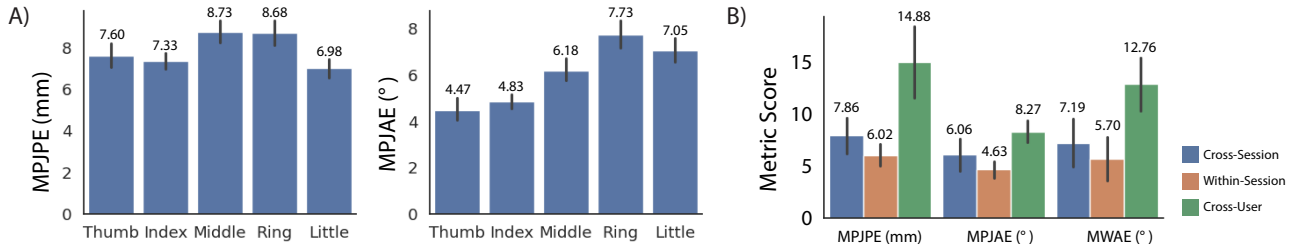


Figure 7: Hand pose tracking performance in Study 1: (A) average 3D hand reconstruction errors in cross-session evaluation for each finger in both MPJPE and MPJAE metrics, and (B) average hand pose tracking performance across all evaluation models and metrics.

7.19° (SD 2.26). Similar to the results for fine-grained finger movement tracking, no significant main effects or interactions were observed (F-values 0.12-2.37, p-values 0.11-0.73).

5.1.3 Multimodal Model Result. We then evaluated the multimodal model combining acoustic and IMU data, which achieved an average MPJPE of 7.96 mm (SD 1.69), MPJAE of 6.03° (SD 1.49), and MWAE of 7.48° (SD 2.23). Similar to the acoustic model, performance showed no significant effects of hardware or watch-wearing hand (F-values 0.02-1.83, p-values 0.17-0.96). This indicates performance is stable and similarly robust across all three watches and on both left and right wrists. A one-way ANOVA comparing acoustic and multimodal models revealed a non-significant effect with IMU addition in all metrics (F-values 0.01-0.41, p-values 0.52-0.92), indicating inclusion of IMU data did not impact model performance. This may be attributed to redundancy between the acoustic and motion signals. IMU data primarily captures wrist motion and its dynamics, and thus lacks the granularity needed to represent detailed finger articulation. In contrast, active acoustic sensing can detect fine-grained finger movements while simultaneously encoding wrist motion through reflected signals [19]. As a result, the acoustic channel alone already provides both local (finger) and global (wrist) information, limiting the additional benefit that IMU data can offer, rather than indicating overfitting of the acoustic model. Based on this observation, we decided to omit further analysis of the IMU data in subsequent evaluations to avoid redundant analysis.

5.2 Within-Session Model

To enable full comparison with prior work, we further evaluated WatchHand using a within-session model. It achieved an average MPJPE of 6.02 mm (SD 1.04), MPJAE of 4.63° (SD 0.80), and MWAE of 5.70° (SD 2.07). Similar to the cross-session model, performance showed no significant effects of hardware or watch-wearing hand (F-values 0.004-1.51, p-values 0.27-0.95). The hand pose tracking performances are notably better than those from the cross-session model (see Figure 7-B). These results align with prior works showing that within-session testing benefits from stable device placement and minimized variability [3, 9, 17]. While this protocol represents a somewhat unrealistic best-case scenario with highly stable device placement, it does provide a useful benchmark against prior systems that have followed similar procedures and serves to underscore the effectiveness of WatchHand's acoustic pipeline.

5.3 Cross-User Model

We next further validate WatchHand's performance using a cross-user model with a leave-one-participant-out scheme, where each participant's data was excluded from training and used solely for testing. The acoustic model shows an average MPJPE of 14.88 mm (SD 3.37), MPJAE of 8.27° (SD 1.04), and MWAE of 12.76° (SD 2.56). Similarly, no significant main or interaction effects were observed for MPJPE and MWAE (F-values 0.09-1.78, p-values 0.18-0.77). However, hardware effects emerged for MPJAE ($F(2, 42) = 4.76$, $p = 0.01$, $\hat{\eta}^2_G$=0.18). Post hoc t-tests revealed that the Galaxy watch shows weaker prediction performance in MPJAE, with a mean of 8.84° (SD 0.80), compared to the Xiaomi watch (7.77° (SD 1.10), $p = 0.003$). While we observed significant hardware effects in cross-user evaluation, these hardware differences were not evident in cross-session or within-session evaluations. This suggests that hardware-induced variability primarily emerges when models are generalized across users, where differences in acoustic signal amplitude and articulation patterns interact with device-specific and user-specific characteristics. We note that performance markedly decreased in the cross-user model compared to the cross-session or within-session model (see Figure 7-B). This pattern is consistent with prior wrist-worn continuous hand pose tracking systems [3, 9, 13, 17, 19], and likely due to the high inter-user variability in hand shape, finger articulation, and motion styles rather than model overfitting. These results suggest the potential value of incorporating lightweight personalization techniques, such as few-shot fine-tuning, to bridge the performance gap between generalized and personalized models.

6 Study 2: Adaptation to Diverse Body Postures

Given the positive hand pose tracking accuracy observed in sitting posture during Study 1, we further conducted a follow-up study to evaluate performance across more diverse postures. Specifically, we examined two common standing postures, including a *watch-raised* pose and an *arm-resting* pose (Figure 8-A). This study was approved by the local IRB.

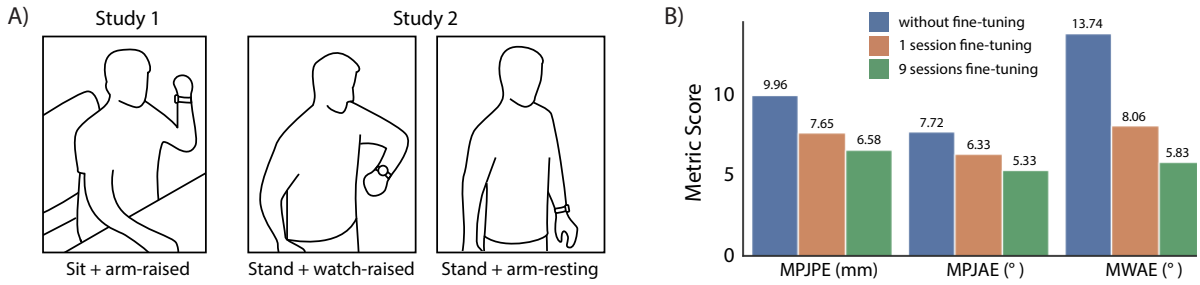


Figure 8: (A) Three key postures covered in this work: sitting and arm-raised pose, and watch-raised pose, where the arm is naturally positioned horizontally in front of the body to view the watch screen, and arm-resting, with the arm lowered in a relaxed position adjacent to the thigh. (B) Average hand pose tracking performances from standing postures (including both watch-raised and arm-resting poses) for each fine-tuning condition—without (trained with data captured from the sitting poses), 1-session, and 9-session, reported across all metrics.

6.1 Participants

We re-invited six participants (three males, a mean age of 22.17 (SD 2.86), all right-handed) from the first study, sampled opportunistically according to their availability. Each participant wore the same smartwatch model they had previously used (3 Xiaomi Watch, 2 Google Pixel Watch, and 1 Samsung Galaxy Watch) on their non-dominant hand. This study lasted approximately two hours, and each participant was compensated with 30 USD in local currency.

6.2 Study Design and Procedure

We conducted 20 standing sessions—including 10 in the *watch-raised* pose and 10 in the *arm-resting* pose, following the same protocol as in Study 1, which included 18 gestures with four repetitions for each session—to evaluate robustness to variations in posture, addressing: *Can our system track hand poses when users are in different body postures?* For the collection of ground-truth data, we used an external webcam positioned below for the watch-raised pose and to the side for the arm-resting pose, with camera angles facing the palm to get the full finger joint ground truth.

6.3 Results

We collected a total of 288 minutes of data from 6 participants, including 144 minutes from the two postures: *watch-raised* and *arm-resting*. Following the first study, we evaluated the system performance on two subsets: (1) continuous fine-grained finger movement tracking using MPJPE and MPJAE metrics, and (2) gross wrist rotation using the MWAE metric. To evaluate WatchHand's robustness across varying body postures, we conducted a targeted analysis across three distinct conditions: (1) posture-independent model, (2) single-session fine-tuned model, and (3) nine-session fine-tuned model. The data collected for the two new postures—watch-raised and arm-resting—were analyzed separately throughout this evaluation. Given the small sample size, we focused on comparing the numerical values rather than formal statistical analysis.

We first evaluated generalization through a posture-independent model trained only on arm-raised sitting data (cross-session protocol, Section 5.1) and tested on unseen postures without retraining (zero-shot evaluation), achieving 9.96 mm MPJPE (SD 1.67), 7.72° MPJAE (SD 1.24), and 13.74° MWAE (SD 6.11). Next, with limited adaptation, fine-tuning on one new session (2 minutes of data for finger movements; 0.4 minutes of data for wrist rotation) improved performance to 7.65 mm MPJPE (SD 1.39), 6.33° MPJAE (SD 1.43), and 8.06° MWAE (SD 2.57). Lastly, more extensive adaptation with nine new sessions (18 minutes of data for finger movements; 3.6 minutes of data for wrist rotation, matching the original training volume) further reduced errors to 6.58 mm MPJPE (SD 1.12), 5.33° MPJAE (SD 1.05), and 5.83° MWAE (SD 1.72). Overall, while the posture-independent model consistently demonstrated centimeter-level tracking accuracy, fine-tuning approaches revealed progressively improved performance as more data was used for training (see Figure 8-B). This highlights the importance of data diversity and quantity for improving model generalization across body postures.

7 Study 3: Robustness to Noises

We further evaluated WatchHand under various noise conditions, reflecting common real-world usage where acoustic signals may be disturbed by external sounds and motion. This study was approved by the local IRB.

7.1 Participants

We recruited 8 participants (5 male, mean age 23.88, SD 4.39) from a local university, matching the number of participants who used a single smartwatch in Study 1. Their mean hand length was 18.18 cm (SD 1.21), and their mean palm width was 8.04 cm (SD 0.62). Each session lasted two hours, and participants received 30 USD in local currency for compensation.

7.2 Study Design and Procedures

We generally followed the methodology outlined in Study 1. First, we collected 10 baseline sessions per participant in a seated posture using the same design as Study 1. To evaluate and compare robustness under noise conditions, we then collected 8 additional sessions

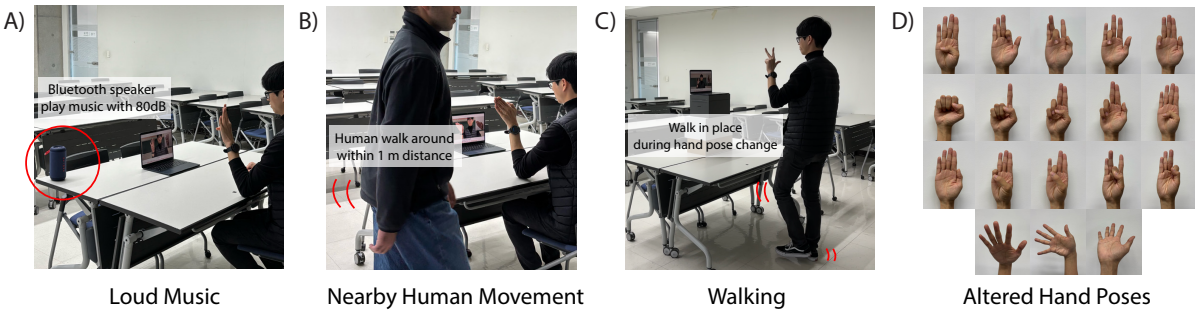


Figure 9: Four noise and interference conditions used in Study 3: (A) high-volume audio played near the participant, (B) another person walking around the participant within 1 m, (C) participant walking in place while executing pose transitions, and (D) participants intentionally performing modified gestures (e.g., reduced or increased finger spacing) to simulate natural inconsistencies.

per participant (2 for each of four noise scenarios), enabling both cross-condition and one-session fine-tuned evaluation, based on the promising fine-tuning results from Study 2. In total, each participant completed 18 sessions, following the same procedure outlined in Study 1, including consent, equipment setup, instructions, practice, main task, and questionnaire. The four noise/interference conditions were instructed as follows: (1) Loud music: high-volume audio (around 80 dB) was played through a Bluetooth speaker positioned one meter from the participant. (2) Nearby human movement: another person walked around the participant at around a one-meter distance. (3) Walking scenario: participants walked continuously while executing pose variations. They were instructed to walk in place at a natural walking pace to introduce motion-induced noise. (4) Altered hand poses: participants intentionally performed modified gestures to simulate natural inconsistencies. Specifically, they were instructed to eliminate finger spacing during hand-pose variations and to widen finger spacing during wrist-rotation poses, deviating from the poses instructed in previous studies.

7.3 Results

We collected a total of 345.6 minutes of data from 8 participants, consisting of 192 minutes from the sitting condition and 38.4 minutes for each of the four noise/interference scenarios. In this study, we report three sets of results: 1. Baseline: training on 9 sessions and testing on the remaining session under the sitting condition; 2. Cross-condition: evaluating the same trained model directly on each of the four noise conditions; 3. One-session fine-tuning: adapting the model with one session (2 minutes of data for finger movements; 0.4 minutes of data for wrist rotation, following Study 2) from each noise condition and testing on the other session. All evaluations were conducted under a cross-session protocol to reflect realistic real-world usage.

The baseline model achieved an average MPJPE of 10.26 mm (SD 1.37), MPJAE of 7.28° (SD 0.76), and MWAE of 9.6° (SD 2.11). Although collected under the same protocol as Study 1, this baseline performance exhibits a degradation, which we attribute to participant-dependent variability, implying that inter-participant articulation styles contribute to model performance. We then examined how this baseline model generalizes to the four noise scenarios (summarized in Figure 10-A). Performance under loud music and nearby human movement remained statistically comparable to the baseline, with no significant differences observed across all metrics (p-values 0.18-0.76). This suggests that our system is robust to audible audio interference and that most nearby human motion occurring beyond the 21.42 cm sensing range is effectively filtered out during preprocessing. However, self-induced physical noise conditions led to significant degradations in the walking condition (p-values 0.003-0.028) and altered hand poses (all $p < 0.001$), indicating that physical interference within the sensing range (e.g., arm motion while walking or unseen variation in altered hand poses) can meaningfully disrupt the echo profiles. Finally, fine-tuning with a single session per noise condition substantially restored performance (See Figure 10-B). Notably, the walking condition shows no significant difference from baseline across all metrics (p-values 0.13-0.54). These results demonstrate that WatchHand can accommodate real-world variability with minimal additional data, and that brief scenario-specific calibration can greatly improve accurate hand pose tracking.

8 Study 4: Evaluation on Dynamic Hand Pose Variations

Our Studies 1-3 demonstrated promising results in continuous hand pose tracking using only the built-in microphone and speaker, following closely related protocols [19]. However, the evaluated pose variations were constrained, as each of the 15 gestures (10 complex, 5 simple) returned to a neutral pose between transitions. To address this, our fourth study captured direct pose-to-pose transitions—100 ordered transitions between the 10 complex gestures and 25 between the 5 simple gestures—for 125 total. This design exposes the model to richer, more realistic hand-motion dynamics. Building on the robustness to different watches and wearing hands established earlier, we selected a single representative device—Xiaomi Watch Pro 2 (chosen for its average tracking performance)—worn on the left hand. In addition, we evaluated the model's robustness under

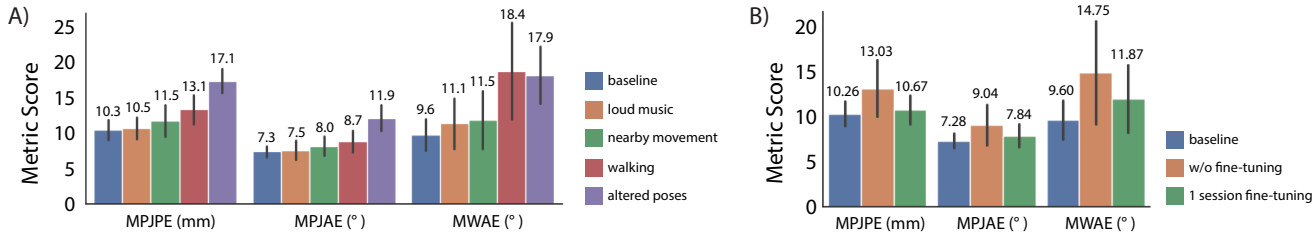


Figure 10: Average hand pose tracking performances across all metrics, comparing baseline (no noise) to (A) four noisy conditions—loud music, nearby movement, walking, and altered poses, and (B) average performance across the noisy conditions with and without one-session fine-tuning.

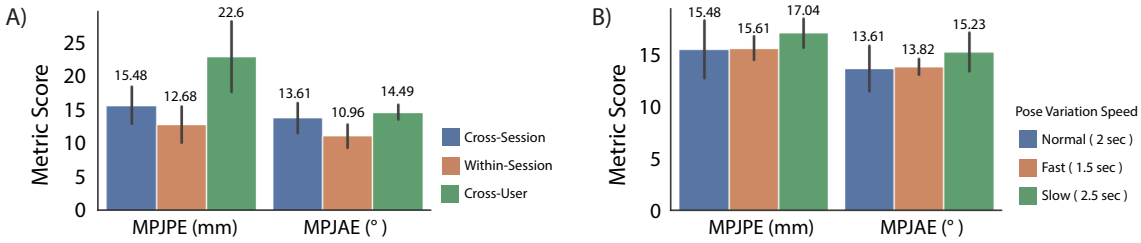


Figure 11: Average hand pose tracking performance in Study 4 across (A) evaluation protocols and (B) hand pose transition speeds.

different hand movement speeds, considering real-world variation. This study was approved by the local IRB.

8.1 Participants

We recruited 8 participants (4 male, all right-handed, mean age 23.5, SD 4.17) from a local university. Their mean hand length was 18.46 cm (SD 1.94), and their mean palm width was 7.99 cm (SD 0.94). Each session lasted approximately 90 minutes, and participants received a 25 USD voucher for their participation.

8.2 Study Design

We evaluated the same set of 15 hand gestures as in Study 1, consisting of 5 simple and 10 complex gestures. Wrist orientation gestures were excluded due to their limited pose-to-pose variation (only 6 combinations). Each session included diverse pose-to-pose transitions, including 4 repetitions of both simple and complex gestures, each lasting two seconds, presented in a randomized order, consistent with the Study 1 protocol. Participants completed 20 sessions at a normal transition speed (two seconds per pose change). To further assess robustness, each participant completed an additional session in both fast (1.5 s transition) and slow (2.5 s transition) conditions.

8.3 Study Procedure

This study followed a procedure similar to that of Study 1, including obtaining consent, donning equipment, delivering instructions, conducting practice trials, executing the main task, and completing a demographic questionnaire. Before beginning the experimental sessions, participants completed a practice trial. They then performed 20 normal-speed sessions, followed by one session each under fast and slow conditions. In the fast and slow conditions, participants were instructed to adjust their hand pose transition times to 1.5 s (fast) or 2.5 s (slow), following protocols in prior work [19]. Considering practical and realistic variability, participants removed and re-wore the smartwatch between sessions.

8.4 Results

We collected a total of 352 minutes of data from 8 participants. This dataset includes 320 minutes of normal-speed sessions and additional test sessions at varying speeds and styles: 12 minutes of fast, 20 minutes of slow conditions. One participant (P28) exhibited a noticeable hand tremor when attempting to bend only the middle finger. To prevent muscle strain and ensure consistent data, we instructed him to stabilize the middle finger with his thumb (e.g., in ASL-8). We processed the captured data and implemented the acoustic model using the same procedures outlined in Study 1. Below, we report results under three evaluation protocols to provide a comprehensive view of WatchHand's continuous 3D hand pose performances in pose-to-pose variation.

We first evaluated WatchHand under a *cross-session* protocol, where the participants re-wore the device across different sessions, reflecting realistic everyday use. In this result, the model achieved an average MPJPE of 15.48 mm (SD 2.75) and an MPJAE of 13.61° (SD 2.18). Next, we evaluated WatchHand under a *within-session* protocol—as an upper-bound performance benchmark—since device placement and environmental factors remain constant. In this

result, the model achieved an average MPJPE of 12.68 mm (SD 2.63) and an MPJAE of 10.96° (SD 1.72). Finally, we performed a *cross-user* evaluation, where models were trained on data from all other participants and tested on unseen participants, providing the system's generalizability to new users. In this result, the model achieved an average MPJPE of 22.60 mm (SD 5.18) and an MPJAE of 14.49° (SD 1.12). Overall, these results show consistent tendencies with Study 1 across protocols: strongest performance in within-session settings, moderate error increases in cross-session evaluation, and the largest drop in cross-user testing (see Figure 11-A)—patterns consistent with prior work [3, 9, 17].

We then analyzed the robustness of WatchHand under different hand pose variation speeds. We used a cross-session model for this testing. For different speeds, the model achieved an average MPJPE of 15.61 mm (SD 1.13) for the fast condition, 17.04 mm (SD 1.35) for the slow condition. The corresponding MPJAE results were 13.82° (SD 0.73) and 15.23° (SD 1.82), respectively. These results show that WatchHand maintains tracking during fast pose changes, with modest performance drops under slower transitions (see Figure 11-B), highlighting both the system's robustness and the need for broader training data to capture diverse real-world variability. As expected, the more dynamic pose-to-pose transitions in Study 4 introduced additional variability compared to the pose-neutral-pose protocol in Studies 1-3. Unlike Study 1, where each gesture transition was anchored at a neutral open-hand pose, Study 4 involved direct transitions between arbitrary gesture pairs (125 possible combinations), which led to noisier intermediate states and a broader range of temporal dynamics. Overall, these results confirm that pose-to-pose transitions are indeed more challenging, but WatchHand still achieves sub-centimeter accuracy in within-session tests and competitive cross-session/user performances.

We further examined WatchHand's behavior by qualitatively inspecting prediction frames. Figure 12 illustrates example reconstructions from a single participant in the pose-neutral-pose cross-session model (Study 1) and the pose-to-pose cross-session model (Study 4), including failure cases, along with the corresponding instruction images. The participants (P11 and P34) were chosen because their overall tracking performance was close to the average score for each study. Overall, our system is particularly effective at predicting thumb movement, likely because the thumb is closer to the smartwatch and follows a distinct bending direction compared to the other fingers, and appears more frequently during training (both in thumb-bent and ASL-4 poses). However, several key error patterns were observed. First, neighboring-finger effects: the model occasionally activated adjacent fingers incorrectly—for example, predicting motion in both the index and middle fingers when only the index should move, likely reflecting biomechanical coupling between fingers and the limited separability of their acoustic signatures, and is especially common around the middle finger, where inter-finger spacing is small [10]. Second, ergonomic constraints: since many participants could not bend the little or ring finger in isolation, the model sometimes reflected these real-world limitations, leading to misrepresentation in those poses. Third, anomalies: we observed occasional random errors (e.g., a single-finger bend predicted during a complex gesture), which we suspect are caused by transient noise, such as subtle posture adjustments or arm movements. These anomalies appear more frequently in the pose-to-pose model, where dynamic transitions are richer, and training data per transition is sparser compared to Study 1. In contrast, in the pose-neutral-pose model, the default open-palm pose (ASL-5) is sometimes over-predicted when the reflected signals are weak (e.g., for isolated little-finger bends). Overall, these qualitative observations are consistent with inter-participant differences in behavior and finger articulation.

9 Discussion and Limitation

Our studies confirm the encouraging performance of WatchHand for inferring continuous 3D hand poses across three common COTS smartwatches. Because it operates entirely with built-in hardware and does not require bespoke sensors, WatchHand has strong potential to immediately expand the interaction capabilities of COTS smartwatches. The results also suggest its applicability across diverse scenarios, including more complex hand poses, different smartwatch models, watch-wearing hands, body postures, and noise conditions. In this section, we further discuss the opportunities and challenges of deploying WatchHand more broadly across diverse hand poses and use contexts to support everyday use in the future.

9.1 Generalizing to Diverse Hand Poses, Users, and Contexts

9.1.1 Dynamic Hand Pose Variation. We evaluated WatchHand under two hand pose variation protocols: pose-neutral-pose (returning to a neutral hand pose between pose variations) in Studies 1-3 and pose-to-pose (direct transitions between poses) in Study 4. The latter introduces richer intermediate states and greater temporal variability, which, as expected, led to a modest accuracy drop (96.7% increase in MPJPE in cross-session evaluation) relative to pose-neutral-pose. However, this drop must be understood in the context of the greatly increased complexity of the task: the number of possible pose transitions increased by 8.3 times (from 15 to 125). Despite this more challenging setting, WatchHand outperforms prior pose-to-pose systems (e.g., DiscoBand [3] (17.87 mm) and EIT-Pose [17] (17.81 mm)) in cross-session evaluation, demonstrating its robustness and the feasibility of scaling continuous hand pose tracking to real-world smartwatch use.

9.1.2 Unseen Hand Poses. While our study includes a diverse set of hand poses, comprehensively covering the entire space of possible hand configurations remains challenging. In Study 3, for example, altered hand poses (e.g., reducing or enlarging gaps between fingers) led to a 66.19% increase in MPJPE in cross-condition testing, yet a single-session (2 minutes of data) fine-tuning reduced this to only a 5.3% drop relative to the baseline. In Study 4, we observed only a modest drop in tracking performance under different hand pose variation speeds—mean MPJPE increased by 2.6% in the fast conditions and 11.5% in the slow conditions—reflecting the added ambiguity of intermediate states and trajectory diversity. Importantly, these effects were bounded rather than catastrophic, suggesting the sensing pipeline remains informative even when encountering out-of-distribution poses. These results demonstrate strong generalization to unseen poses, an important step toward real-world usability with larger, more diverse datasets.

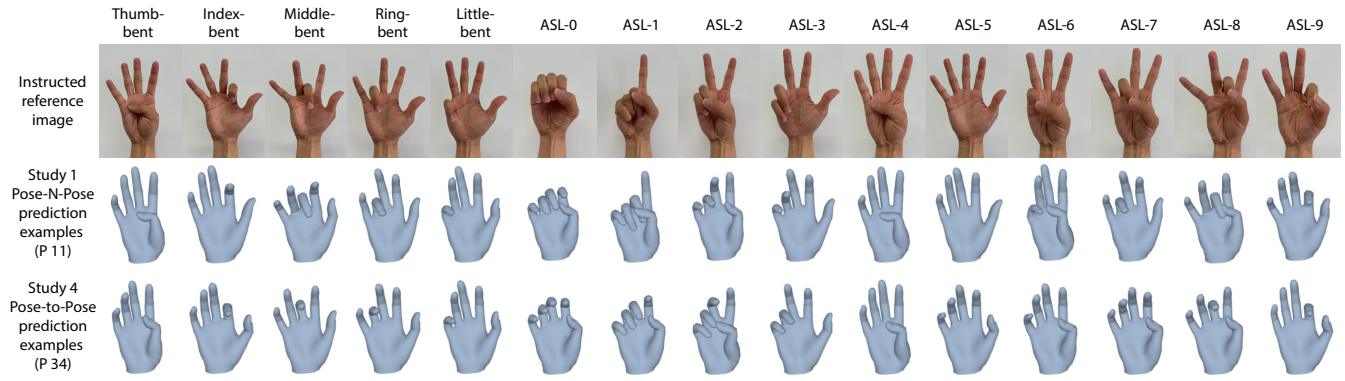


Figure 12: Hand poses instructed in the studies (top row) and example predictions (visualized with MANO [32]) from the pose-neutral-pose cross-session model in Study 1 (middle row) and pose-to-pose cross-session model in Study 4 (bottom row).

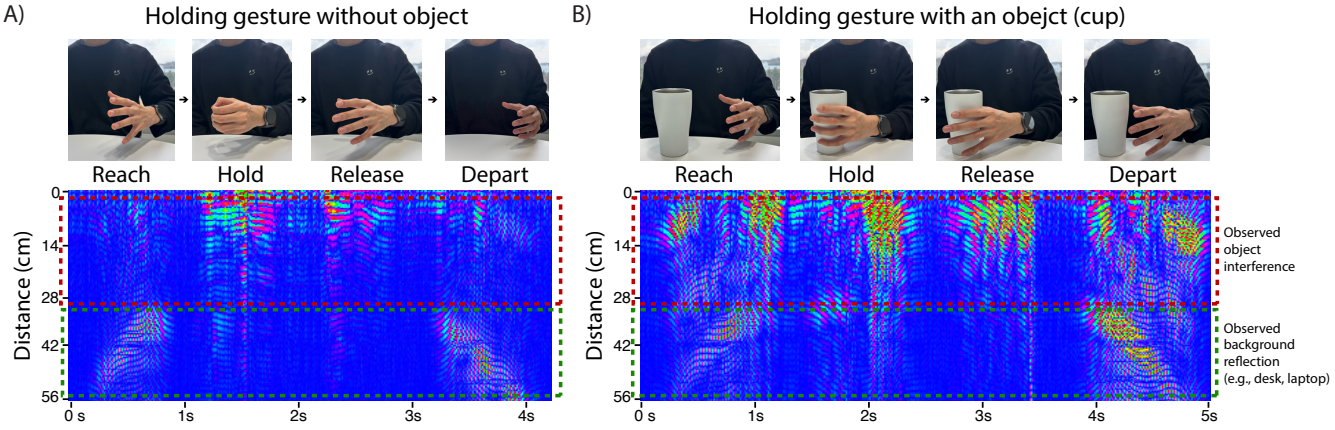


Figure 13: Impact of hand–object interaction on acoustic hand pose tracking. (A) Echo profiles of a holding gesture performed without an object and (B) with a cup. Each sequence consists of four phases: Reach, Hold, Release, and Depart. Compared to the object-free condition, performing the gesture while holding an object produces noticeable acoustic interference near the smartwatches.

9.1.3 Unseen Users. Cross-user evaluation yielded encouraging results, with MPJPE ranging from 15 to 23 mm across both pose-neutral-pose and pose-to-pose protocols. Performance degradation was primarily due to inter-individual differences in hand size and articulation, which introduced greater variability in pose execution. This pattern is consistent with prior wrist-worn hand-pose tracking systems, where cross-user performance is substantially weaker than within-session [3, 17] or cross-session results [9, 19]. These findings underscore a key challenge of cross-user generalization: while the system adapts well within a user, accommodating inter-individual differences remains challenging. Addressing this will require larger and more heterogeneous datasets spanning a wider range of users and hand characteristics, and meta-learning–style adaptation could further improve cross-user robustness.

9.1.4 Robustness to Noise. Our noise study (Study 3) suggests that the primary vulnerability of smartwatch-based acoustic hand tracking is not ambient sound or distant movement, but physical changes near the sensing region. Loud music and nearby human movement had a negligible impact, indicating that WatchHand, operating in the inaudible band and with a 21.42 cm sensing range, is effectively decoupled from most everyday audio activity and background motion. In contrast, mobile scenarios (e.g., walking) altered the propagation paths around the wrist and hand, leading to measurable degradations and confirming that nearby physical changes around the wrist and hand disrupt the echo signals. Nevertheless, a single short fine-tuning session per condition largely restored performance, showing that WatchHand can adapt to real-world noise and motion scenarios with minimal calibration, reinforcing its practicality for everyday smartwatch use.

9.1.5 Holding Object Interferences. While WatchHand reliably tracks 3D hand poses in free-hand scenarios, its performance will likely be substantially impacted when users hold objects. An exploratory analysis (see Fig 13) supports this observation: when participants performed the same reach–hold–release–depart sequence [34] with and without a cup, the object introduced localized interference in the echo profiles. These reflections stemmed from object surfaces rather than the hand, masking critical pose-dependent patterns and confirming that object interaction is a meaningful disruptor to single-channel acoustic sensing. Quantifying the resulting degradation is challenging, as object manipulation fundamentally changes hand geometry and invalidates the original pose protocol, making direct comparison with free-hand conditions infeasible. However, this limitation is not unique to our approach. Commercial AR/VR systems typically suspend hand tracking under occlusion rather than produce unreliable poses [26]. A smartwatch-based acoustic system could adopt a similar strategy—detecting object presence and temporarily disabling pose tracking. This limitation highlights an exciting direction for future research. Integrating object-awareness—such as recognizing grasp states or supporting micro-gestures during holding, as explored in bespoke systems [18], could enable WatchHand to transition seamlessly between free-hand and object-mediated interactions in real-world use.

9.1.6 Strategies for Improving and Enriching Datasets. Despite our extensive evaluations, the dataset size remains a key limitation due to resource constraints. As prior work [3, 9, 12, 17, 19] has noted, larger-scale datasets are essential for training more sophisticated models to handle unseen users, poses, and contexts. Given that our system is implemented with COTS smartwatches and optical tracking systems, we can leverage this accessibility to implement a distributed data-collection approach, potentially through an opt-in framework that allows users to contribute anonymized hand-pose data during everyday use. Beyond collecting additional real-world data, developing a physics-based or data-driven acoustic simulator could generate synthetic echo profiles for underrepresented hand poses, helping to fill gaps in pose coverage. Furthermore, lab studies using a research-grade optical tracking system (e.g., Optitrack or similar) could contribute more accurate and reliable baseline data. In parallel, self-supervised learning could enable the model to learn meaningful representations from unlabeled acoustic data, reducing reliance on dense ground-truth annotations. We further acknowledge that unsupervised domain adaptation methods, which encourage invariance across users, sessions, or poses, are a promising direction to narrow the remaining generalization gap, especially in cross-user settings. Together, these strategies would strengthen generalization and pave the way toward broader real-world deployment of WatchHand.

9.2 Toward Real-World Implementation

9.2.1 Bespoke Hardware vs Off-the-shelf Smartwatches. Recent acoustic hand-tracking systems such as EchoWrist [19] achieve high-quality pose estimation by using two speaker–microphone pairs—one placed on the watch body and another embedded in the wrist strap to directly face the palm. Although this configuration produces stronger acoustic reflections and delivers improved tracking performance compared to WatchHand (4.81 vs 7.87 mm MPJPE in cross-session evaluation), it relies on non-commercial hardware modifications. Embedding electronics into straps is uncommon in today's smartwatches, as straps are intentionally user-replaceable, vary across brands, and are not currently designed or optimized to host powered sensing components. In contrast, WatchHand demonstrates that continuous 3D hand pose tracking can be achieved using only a single speaker and microphone already embedded in commercial smartwatches. This distinction is critical. Bespoke systems illustrate what enhanced sensing could achieve if hardware were redesigned; WatchHand shows what is possible today, on devices that already ship at scale. By enabling fine-grained 3D hand pose tracking on millions of existing smartwatches, WatchHand represents the first practical pathway for transitioning continuous hand tracking from bespoke academic prototypes to real-world smartwatch platforms.

9.2.2 Hardware Configuration and Watch-wearing Hands. WatchHand demonstrates robustness to variations in hardware platforms and watch orientations. Our results show no significant impact from differences in hardware configurations across manufacturers or from flipped orientations due to the watch-wearing hand (left vs. right). This highlights WatchHand's potential for universal applicability. In our main experiment, we train separate models for each watch–hand configuration, so each model learns a stable mapping for its specific setup. To further examine generalizability, we trained a general model with training data aggregated across all smartwatch models and watch-wearing hands in cross-session testing. The results are promising, demonstrating a comparable performance to the per-configuration cross-session models, with an average MPJPE of 7.88 mm (SD 1.49), MPJAE of 6.22° (SD 1.45), and MWAE of 7.68° (SD 1.35). Since all tested devices share a similar acoustic layout (speaker and microphone on opposite edges), these results suggest that a single model can generalize across brands and orientations, supporting scalable deployment on COTS smartwatches.

9.2.3 System Deployment. In this study, as in many prior works [3, 9, 12, 17, 19, 22, 48], our system initially streamed the raw sensor data from the smartwatch to external servers (e.g., laptop or smartphone), with signal processing and model inference performed on a remote server. However, to be deployed in real-world settings, the system needs to run entirely on-device without relying on external infrastructure. To demonstrate the feasibility of this, we deployed the entire pipeline model—including signal processing and model inference—on commercial smartwatches. The model (26.7 MB) executes fully on-device (tested on Samsung Galaxy Watch 7), enabling real-time, continuous hand pose tracking without transmitting data to external systems. The average latency per prediction is 0.115s (SD 0.006), comprising 0.035s (SD 0.004) for preprocessing and 0.08s (SD 0.005) for inference, supporting responsive interaction in practical use.

9.3 Practical Consideration

9.3.1 Power Consumption. Running continuous playback and recording for active acoustic sensing on a smartwatch introduces non-trivial power demands. To estimate this cost, we monitored the battery usage over a 3-minute session—including key operations such as recording, playback, and data transmission, along with

ambient power consumption from screen lighting and background processes. The Galaxy Watch 7 consumed approximately 1% of its battery (4.25 mAh of 425 mAh), the Xiaomi Watch 2 Pro about 0.87% (4.31 mAh of 495 mAh), and the Pixel Watch 3 about 1.52% (6.36 mAh of 420 mAh). These differences in power consumption may be attributed to variations in hardware configurations and sensor characteristics—for instance, the Pixel Watch 3 exhibited a higher speaker output level, which may contribute to its increased energy use. It is important to note that these numbers reflect the entire system on the COTS watch (e.g., OS, display, and other functions), and our system is not yet optimized for this sensing task, as typical wearable microphones and speakers are reported to have power signatures [19, 48] under 1 mW. This indicates that further optimization could significantly extend the usage time for a smartwatch-based acoustic hand pose tracking system. In practice, to conserve power, hand pose tracking with acoustic sensors could be activated only when low-power IMU-based gestures [1, 29] are detected.

9.3.2 Privacy. Operating in the 18-21 kHz frequency range offers WatchHand privacy advantages, as this ultrasonic range lies significantly above the typical spectrum of human speech, which spans approximately 0.1 to 8 kHz. Our system implements frequency filtering to isolate only signals within our designated operating range, effectively eliminating speech-related audio content from processing. Although the current WatchHand system initially records broadband audio, including speech frequencies, our processing pipeline immediately discards this lower-frequency information. In future implementations, when the entire system can be deployed directly on the smartwatch without requiring external processing, privacy protection can be further enhanced through on-device computation, ensuring that potentially sensitive audio data never leaves the user's personal device.

9.3.3 Potential Applications. Building on the strong and robust hand pose tracking performance results we report, we suggest potential practical applications of WatchHand in daily wearable contexts. One key application area is gesture-based interaction. WatchHand allows users to interact with devices through natural hand movements without requiring physical touch, which is particularly valuable for hands-busy or touch-free scenarios, such as controlling music playback while cooking, navigating slides during a presentation, or issuing commands [2, 4, 35]. Continuous tracking also enables more expressive gestures beyond discrete taps or swipes, allowing for nuanced and expressive controls such as pinch-to-zoom, drag, or swipe variations based on finger articulation [47]. In addition, WatchHand can support assistive technologies, particularly for users with limited mobility or speech. By detecting subtle and continuous finger movements, the system could facilitate silent communication (e.g., fingerspelling [21, 30]) and customizable input for accessibility settings. Finally, WatchHand has the potential to extend cross-device interaction. When paired with smartphones, tablets, or smart glasses, smartwatches running WatchHand could serve as lightweight and low-cost, wrist-mounted input devices for spatial interfaces [33]. These applications outline how continuous hand pose tracking—achieved solely by the sensors within existing COTS smartwatches—can empower more natural, expressive, and context-aware interaction in a range of real-world scenarios.

10 Conclusion

We presented WatchHand, the first continuous 3D hand pose tracking system running entirely on commercial smartwatches using only the built-in speaker and microphone. Leveraging active acoustic sensing, WatchHand reliably tracks finger movements and wrist rotations without requiring external hardware. Our evaluations across different smartwatch models, watch-wearing hands, body postures, noise conditions, and hand pose variation protocols confirm the system's robustness, adaptability, and generalizability. By utilizing existing smartwatch hardware, WatchHand offers a practical, privacy-preserving, and scalable solution that is capable of immediate deployment. This breakthrough substantially lowers the barrier to smartwatch-based hand tracking and unlocks new opportunities for always-available interactions on millions of devices already in use today.

Acknowledgments

We thank all the participants who took part in our user studies. This work was supported by the National Science Foundation under Grant No.2239569 (NSF CAREER Award) and the IITP(Institute of Information & Communications Technology Planning & Evaluation)-ITRC(Information Technology Research Center) grant funded by the Korea government(Ministry of Science and ICT)(IITP-2026-RS-2024-00436398).

References

[1] Apple. 2024. Use AssistiveTouch on your Apple Watch. https://support.apple.com/guide/watch/assistivetouch-apdec70bfd2d/watchos.

[2] Riku Arakawa, Hiromu Yakura, and Mayank Goel. 2024. PrISM-Observer: Intervention Agent to Help Users Perform Everyday Procedures Sensed using a Smartwatch. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 15, 16 pages. doi:10.1145/3654777.3676350

[3] 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. doi:10.1145/3526113.3545634

[4] Nathan DeVrio, Vimal Mollyn, and Chris Harrison. 2023. SmartPoser: Arm Pose Estimation with a Smartphone and Smartwatch Using UWB and IMU Data. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 79, 11 pages. doi:10.1145/3586183.3606821

[5] Jun Gong, Zheer Xu, Qifan Guo, Teddy Seyed, Xiang 'Anthony' Chen, Xiaojun Bi, and Xing-Dong Yang. 2018. WrisText: One-handed Text Entry on Smartwatch using Wrist Gestures. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3173574.3173755

[6] Sidhant Gupta, Daniel Morris, Shwetak Patel, and Desney Tan. 2012. SoundWave: using the doppler effect to sense gestures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 1911–1914. doi:10.1145/2207676.2208331

[7] Ryo Hajika, Tamil Selvan Gunasekaran, Chloe Dolma Si Ying Haigh, Yun Suen Pai, Eiji Hayashi, Jaime Lien, Danielle Lottridge, and Mark Billinghurst. 2024. RadarHand: A Wrist-Worn Radar for On-Skin Touch-Based Proprioceptive Gestures. *ACM Trans. Comput.-Hum. Interact.* 31, 2, Article 17 (jan 2024), 36 pages. doi:10.1145/3617365

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[9] 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 (June 2020), 24 pages. doi:10.1145/3397306

[10] Jose M Jimenez-Olmedo, Alfonso Penichet-Tomás, Manuel Ortega Becerra, Basilio Pueo, and Jose J Espina-Agullo. 2019. Relationships between anthropometric

parameters and overarm throw in elite beach handball. *Human Movement* 20, 2 (2019), 16–24.

[11] Patrick Kaifosh and Thomas R Reardon. 2025. A generic non-invasive neuromotor interface for human-computer interaction. *Nature* (2025), 1–10.

[12] 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. doi:10.1145/3526113.3545665

[13] 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. doi:10.1145/2380116.2380139

[14] Jiwan Kim, Hohurn Jung, and Ian Oakley. 2026. OpenAcoustics: An Open-Source Framework for Acoustic Data Capture on Smart Devices and Microcontrollers. In *Companion of the 2025 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Finland) *(UbiComp Companion '25)*. Association for Computing Machinery, New York, NY, USA, 662–664. doi:10.1145/3714394.3756156

[15] 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* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 287, 10 pages. doi:10.1145/3491102.3501935

[16] Jiwan Kim, Jiwan Son, and Ian Oakley. 2025. Cross, Dwell, or Pinch: Designing and Evaluating Around-Device Selection Methods for Unmodified Smartwatches. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 604, 11 pages. doi:10.1145/3706598.3714308

[17] Alexander Kyu, Hongyu Mao, Junyi Zhu, Mayank Goel, and Karan Ahuja. 2024. EITPose: Wearable and Practical Electrical Impedance Tomography for Continuous Hand Pose Estimation. 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, Article 402, 10 pages. doi:10.1145/3613904.3642663

[18] Chi-Jung Lee, Jiaxin Li, Tianhong Catherine Yu, Ruidong Zhang, Vipin Gunda, François Guimbretière, and Cheng Zhang. 2025. Grab-n-Go: On-the-Go Microgesture Recognition with Objects in Hand. *arXiv preprint arXiv:2508.11620* (2025).

[19] Chi-Jung Lee, Ruidong Zhang, Devansh Agarwal, Tianhong Catherine Yu, Vipin Gunda, Oliver Lopez, James Kim, Sicheng Yin, Boao Dong, Ke Li, Mose Sakashita, Francois Guimbretiere, 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 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 403, 21 pages. doi:10.1145/3613904.3642910

[20] Sarah M. Lehman, Abrar S. Alrumayh, Kunal Kolhe, Haibin Ling, and Chiu C. Tan. 2022. Hidden in Plain Sight: Exploring Privacy Risks of Mobile Augmented Reality Applications. *ACM Trans. Priv. Secur.* 25, 4, Article 26 (July 2022), 35 pages. doi:10.1145/3524020

[21] Hyunchul Lim, Nam Anh Dang, Dylan Lee, Tianhong Catherine Yu, Jane Lu, Franklin Mingzhe Li, Yiqi Jin, Yan Ma, Xiaojun Bi, François Guimbretière, et al. 2025. SpellRing: Recognizing Continuous Fingerspelling in American Sign Language using a Ring. *arXiv preprint arXiv:2502.10830* (2025).

[22] 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 (Sept. 2021), 27 pages. doi:10.1145/3478112

[23] Kian Mahmoodi, Yudong Xie, Tan Gemicioglu, Chi-Jung Lee, Jiwan Kim, and Cheng Zhang. 2025. EchoForce: Continuous Grip Force Estimation from Skin Deformation Using Active Acoustic Sensing on a Wristband. In *Proceedings of the 2025 ACM International Symposium on Wearable Computers* (Espoo, Finland) *(ISWC '25)*. Association for Computing Machinery, New York, NY, USA, 53–60. doi:10.1145/3715071.3750405

[24] 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 (Sept. 2023), 28 pages. doi:10.1145/3610895

[25] Saif Mahmud, Vineet Parikh, Qikang Liang, Ke Li, Ruidong Zhang, Ashwin Ajit, Vipin Gunda, Devansh Agarwal, Francois Guimbretiere, and Cheng Zhang. 2024. ActSonic: Recognizing Everyday Activities from Inaudible Acoustic Wave Around the Body. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 4, Article 183 (Nov. 2024), 32 pages. doi:10.1145/3699752

[26] Meta. 2025. Hands best practices. https://developers.meta.com/horizon/design/hands-best-practices/

[27] Vimal Mollyn, Riku Arakawa, Mayank Goel, Chris Harrison, and Karan Ahuja. 2023. IMUPoser: Full-Body Pose Estimation using IMUs in Phones, Watches, and Earbuds. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 529, 12 pages. doi:10.1145/3544548.3581392

[28] 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. doi:10.1145/2858036.2858580

[29] Samsung Newsroom. 2023. More Than a Gesture: How Galaxy Watch's Universal Gestures Feature Enhances Accessibility. https://news.samsung.com/global/more-than-a-gesture-how-galaxy-watchs-universal-gestures-feature-enhances-accessibility.

[30] Nicolas Pugeault and Richard Bowden. 2011. Spelling it out: Real-time ASL fingerspelling recognition. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. 1114–1119. doi:10.1109/ICCVW.2011.6130290

[31] Franziska Roesner, Tadayoshi Kohno, and David Molnar. 2014. Security and privacy for augmented reality systems. *Commun. ACM* 57, 4 (April 2014), 88–96. doi:10.1145/2580723.2580730

[32] Javier Romero, Dimitrios Tzionas, and Michael J. Black. 2017. Embodied hands: modeling and capturing hands and bodies together. *ACM Trans. Graph.* 36, 6, Article 245 (Nov. 2017), 17 pages. doi:10.1145/3130800.3130883

[33] Sasha Salter, Richard Warren, Collin Schlager, Adrian Spurr, Shangchen Han, Rohin Bhasin, Yujun Cai, Peter Walkington, Anuoluwapo Bolarinwa, Robert J Wang, et al. 2024. emg2pose: A large and diverse benchmark for surface electromyographic hand pose estimation. *Advances in Neural Information Processing Systems* 37 (2024), 55703–55728.

[34] Adwait Sharma, Alexander Ivanov, Frances Lai, Tovi Grossman, and Stephanie Santosa. 2024. GraspUI: Seamlessly Integrating Object-Centric Gestures within the Seven Phases of Grasping. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (Copenhagen, Denmark) *(DIS '24)*. Association for Computing Machinery, New York, NY, USA, 1275–1289. doi:10.1145/3643834.3661551

[35] Adwait Sharma, Christina Salchow-Hömmen, Vimal Suresh Mollyn, Aditya Shekhar Nittala, Michael A. Hedderich, Marion Koelle, Thomas Seel, and Jürgen Steimle. 2023. SparseIMU: Computational Design of Sparse IMU Layouts for Sensing Fine-grained Finger Microgestures. *ACM Trans. Comput.-Hum. Interact.* 30, 3, Article 39 (June 2023), 40 pages. doi:10.1145/3569894

[36] Andrew G Stove. 1992. Linear FMCW radar techniques. In *IEE Proceedings F (Radar and Signal Processing)*, Vol. 139. IET, 343–350.

[37] Shunta Suzuki, Takashi Amesaka, Hiroki Watanabe, Buntarou Shizuki, and Yuta Sugiura. 2024. EarHover: Mid-Air Gesture Recognition for Hearables Using Sound Leakage Signals. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 129, 13 pages. doi:10.1145/3654777.3676367

[38] Alvin R Tilley. 2001. *The measure of man and woman: human factors in design*. John Wiley & Sons.

[39] Riku Tsunoda, Myungguen Choi, and Buntarou Shizuki. 2024. Thumb-to-Finger Gesture Recognition Using COTS Smartwatch Accelerometers. In *Proceedings of the International Conference on Mobile and Ubiquitous Multimedia (MUM '24)*. Association for Computing Machinery, New York, NY, USA, 184–195. doi:10.1145/3701571.3701600

[40] Pavan Kumar Anasosalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, and Anurag Ranjan. 2023. Fastvit: A fast hybrid vision transformer using structural reparameterization. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5785–5795.

[41] 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. doi:10.1145/3615592.3616851

[42] 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. doi:10.1145/3161188

[43] Wei Wang, Alex X. Liu, and Ke Sun. 2016. Device-free gesture tracking using acoustic signals. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking* (New York City, New York) *(MobiCom '16)*. Association for Computing Machinery, New York, NY, USA, 82–94. doi:10.1145/2973750.2973764

[44] Hongyi Wen, Julian Ramos Rojas, and Anind K. Dey. 2016. Serendipity: Finger Gesture Recognition using an Off-the-Shelf Smartwatch. 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, 3847–3851. doi:10.1145/2858036.2858466

[45] 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. doi:10.1145/3379337.3415897

[46] 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. doi:10.1145/3491102.3501904

[47] Chun Yu, Yizheng Gu, Zhican Yang, Xin Yi, Hengliang Luo, and Yuanchun Shi. 2017. Tap, Dwell or Gesture? Exploring Head-Based Text Entry Techniques for HMDs. 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, 4479–4488. doi:10.1145/3025453.3025964

[48] Tianhong Catherine Yu, Guilin Hu, Ruidong Zhang, Hyunchul Lim, Saif Mahmud, Chi-Jung Lee, Ke Li, Devansh Agarwal, Shuyang Nie, Jinseok Oh, François Guimbretière, and Cheng Zhang. 2024. Ring-a-Pose: A Ring for Continuous Hand Pose Tracking. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 4, Article 189 (Nov. 2024), 30 pages. doi:10.1145/3699741

[49] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. 2020. Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214* (2020).

[50] Qian Zhang, Yubin Lan, Kaiyi Guo, and Dong Wang. 2024. Lipwatch: Enabling Silent Speech Recognition on Smartwatches using Acoustic Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 2, Article 84 (May 2024), 29 pages. doi:10.1145/3659614

[51] 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. doi:10.1145/3544548.3580801

[52] Yu Zhang, Tao Gu, Chu Luo, Vassilis Kostakos, and Aruna Seneviratne. 2018. FinDroidHR: Smartwatch Gesture Input with Optical Heartrate Monitor. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 1, Article 56 (March 2018), 42 pages. doi:10.1145/3191788