UltraHapSense: Simultaneous Ultrasonic Mid-Air Haptics and Hand Localization with Ultrasonic Transducer Array

Changhyeon Park
sac7160@kaist.ac.kr
KAIST
Daejeon, Republic of Korea

Jiwan Kim
kjwan4435@gmail.com
KAIST
Daejeon, Republic of Korea

Ian Oakley
ianoakley@kaist.ac.kr
KAIST
Daejeon, Republic of Korea

Cheng Zhang
chengzhang@cornell.edu
Cornell
New York, USA

Sang Ho Yoon
sangho@kaist.ac.kr
KAIST
Daejeon, Republic of Korea

Abstract

Ultrasonic Mid-Air Haptics (UMH) provide non-contact tactile sensations, enabling immersive interactions without restricting user movement. Still, conventional UMH systems rely on additional external optical sensing devices for hand localization, which introduces system complexity and raises privacy concerns. We propose UltraHapSense, a novel system that enables simultaneous mid-air haptic feedback and hand localization using a single ultrasonic transducer array and four microphones. To achieve this, we combine haptic feedback and acoustic sensing signals into a single envelope via amplitude modulation, enabling simultaneous haptic rendering and hand localization while suppressing harmonic distortion. Additionally, we propose Haptic-Aware Data Augmentation, which introduces structured noise to the input data based on physical interference patterns to improve robustness to haptic signal-induced acoustic interference. We employ a FiLM-conditioned CNN-BiLSTM model that incorporates haptic focal point information, along with a 2-stage estimation strategy that first estimates wrist position and then joint positions. We conduct experiments to evaluate generalization and robustness under diverse conditions. The results show that UltraHapSense achieves accurate hand localization with an average error of 14.73*mm*, while generating haptic feedback at 133.93 dB SPL, demonstrating the feasibility of re-purposing a UMH transducer array for both haptic rendering and acoustic sensing without relying on external optical sensing devices.

CCS Concepts

• **Human-centered computing** → **Interaction techniques**; • **Hardware** → **Tactile and hand-based interfaces**.

Keywords

Acoustic Sensing, Hand Localization, Mid-air Haptics, I/O Device,

ACM Reference Format:
Changhyeon Park, Jiwan Kim, Ian Oakley, Cheng Zhang, and Sang Ho Yoon. 2027. UltraHapSense: Simultaneous Ultrasonic Mid-Air Haptics and Hand Localization with Ultrasonic Transducer Array. In *ACM/IEEE International Conference on Embedded Artificial Intelligence and Sensing Systems (SenSys '27), May 17-20, 2027, Boulder, CO, USA.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3862575.3862822



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.
SenSys '27, Boulder, CO, USA
© 2027 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-3124-2/2027/05
https://doi.org/10.1145/3862575.3862822

1 Introduction

Ultrasonic Mid-Air Haptics (UMH) creates tactile sensation without physical contact by generating acoustic radiation pressure at a focal point in mid-air using ultrasound. As it does not require direct contact or wearable devices, it is beneficial for public environments that require hygienic interaction. By controlling the phase delay of the ultrasonic transducers, tactile stimuli with a peak force of approximately 16mN and a focal diameter of around 20mm can be generated [11]. Since ultrasound can be modulated without mechanical actuation, the resulting haptic feedback operates with low latency, making it suitable for real-time interaction. Based on these features, UMH can enhance the user experience by providing responsive, touch-less haptic feedback across various interactive computing environments, including VR/AR [33, 34], and holographic displays [29, 40]. It can also be applied to public interfaces such as kiosks, elevator panels, and in-vehicle controls, where users interact without wearing any device.

However, existing UMH systems require hand tracking to deliver haptic feedback at appropriate locations during interaction, thereby requiring additional optical sensing devices. Therefore, conventional optical tracking systems have been widely used to support hand tracking. Specifically, the Leap Motion Controller has been widely adopted in commercial UMH devices [1] and prior work [27]. However, previous optical tracking systems require continuous processing of large amounts of visual data and raise privacy concerns. Prior work has explored various approaches to minimize UMH devices, such as reducing the number of ultrasonic transducers [31] or employing smaller transducers [37]. However, these efforts primarily focus on optimizing the ultrasonic actuation. In contrast, little work has explored reusing the haptic transducer array for hand localization. This enables the same array to support both hand localization and haptic feedback within a unified system.

We focus on a lightweight, privacy-preserving hand-localization system for UMH. Specifically, we employ acoustic sensing to estimate the absolute 3D hand position for haptic feedback. We utilized the UMH transducer array to simultaneously emit signals for both haptic feedback and acoustic sensing. Specifically, we employ

amplitude modulation to integrate the two signals onto a single 40 kHz carrier and transmit them simultaneously. Specifically, a low-frequency haptic modulation (200 Hz) and an FMCW sensing signal (16-19 kHz) are combined into a single envelope, enabling concurrent operation without temporal or spatial multiplexing. This frequency-domain separation minimizes mutual interference and enables independent processing, while the near-ultrasonic sensing frequencies remain largely inaudible during operation. We used four small microphones ($3.5 \times 2.65 \times 0.98$) mm to capture the reflected acoustic signals, enabling a compact UMH system. Thus, we hypothesize that *a UMH transducer array originally limited to haptic feedback can be repurposed to support both haptic feedback and hand localization, enabling a compact and privacy-preserving integrated UMH system using only compact microphones instead of bulky optical sensors.*

However, even though the haptic and sensing signals occupy different frequency bands, strong haptic signals interfere with sensing signals due to non-linear effects and produce structured interference patterns in the input data. To address this challenge, we propose a haptic-aware data augmentation (HADA) and deep learning model. HADA improves robustness by injecting structured noise based on the typical interference introduced by generating haptic cues at different locations. Also, a FiLM-CNN-BiLSTM architecture explicitly incorporates haptic focal point information as an external parameter, enabling adaptive feature representations for robust hand localization. Experimental results show that the proposed method achieves a mean absolute error of 14.73 mm and a mean per-joint position error of 30.62 mm with concurrent haptic feedback generation. Simultaneously, our approach maintains a perceptible haptic feedback intensity of 133.93 dB SPL while enabling simultaneous hand localization.

The main contributions are as follows :

- Simultaneous ultrasonic mid-air haptics and hand localization using a single ultrasonic transducer array and 4 microphones;
- Amplitude-modulated integration of ultrasonic mid-air haptics and active acoustic sensing signals into a single transmitted signal;
- Data augmentation leveraging haptic-induced interferences for robust hand localization under concurrent haptic signals; and
- Haptic-Aware deep learning model incorporating haptic focal point information for robust hand localization under concurrent haptic feedback.

2 Related Work

2.1 Haptic-Sensing Integrated Interfaces

In interactive computing environments such as AR/VR and holographic interfaces, relying solely on visual feedback can increase cognitive load and reduce interaction accuracy. Therefore, integrating haptic feedback is essential for enabling effective and immersive user interactions [43]. To enrich user experience, prior work has explored various haptic rendering techniques [30, 41] and haptic modalities beyond vibrotactile feedback, including EMS [6], thermal feedback [51], and electrostatic force [44]. However, effective haptic feedback cannot be sufficiently achieved through simple and predefined haptic mechanisms. To enrich user experience and provide contextually appropriate haptic feedback, the system should understand user inputs and interaction context.

Realizing such context-aware haptics requires additional sensing capabilities to infer user interactions in real time. However, relying on external hardware increases system complexity and introduces form-factor overhead. To address this, interfaces that combine haptic actuation and sensing within a single device have been explored, eliminating the need for external sensing hardware. HapSense [52] demonstrates that a soft deformable material can integrate force sensing and vibrotactile actuation without functional degradation. Similarly, Back-EMF-based approaches [7] repurpose LRAs to enable touch and pressure sensing while maintaining haptic actuation. Also, MagetIO [25] enables sensing and haptic feedback through a single magnetic interaction, detecting contact with soft patches and simultaneously converting it into vibrotactile feedback. In addition, HapTag [5]and HapCoil [53] embed pressure-sensing layers within soft actuators, allowing user input to directly trigger contextually responsive haptic feedback. However, these interfaces fundamentally rely on physical contact between the user and the device. This inherently limits applicability in scenarios where contact-free interaction is required, such as mid-air interaction in AR and VR environments. We extend the haptic-sensing integration from contact-based systems to non-contact systems by combining ultrasonic mid-air haptics with acoustic sensing. This enables simultaneous tactile feedback generation and hand localization without requiring any physical contact with the device.

2.2 Ultrasonic Mid-Air Haptics

UMH controls the phase delay of acoustic waves emitted from multiple transducers to focus acoustic energy at a specific point. The resulting acoustic radiation pressure stimulates mechanoreceptors on the skin to produce tactile sensations [3, 10, 24]. Therefore, UMH can generate precise tactile stimuli without direct physical contact with the user's body and has been used in interactive systems that deliver haptic feedback to various body parts, including the lips [37], face [8, 16], and hands [26]. In addition, prior work has explored generating diverse tactile patterns and richer tactile experiences by adjusting UMH parameters such as modulation methods and frequency [38, 39, 42].

Still, existing UMH systems rely on optical sensing devices such as Leap Motion Controller [16] and OptiTrack [31] to track hand position for delivering haptic feedback at desired locations. Such optical sensing systems raise privacy concerns. To address this limitation, prior works have explored actively repositioning the transducer array using robotic arms [45] or servo motors [14, 31]. However, while they have focused on reducing the size of UMH systems by minimizing the transducer array, minimizing the use of external sensing devices remains unexplored. In this work, we replace optical sensing devices with four microphones, enabling a compact UMH system free from privacy concerns.

2.3 Acoustic Sensing

Acoustic sensing estimates user gestures and motions by analyzing acoustic signals reflected from the user's body or surrounding objects after a speaker emits acoustic waves. Early work leveraged the

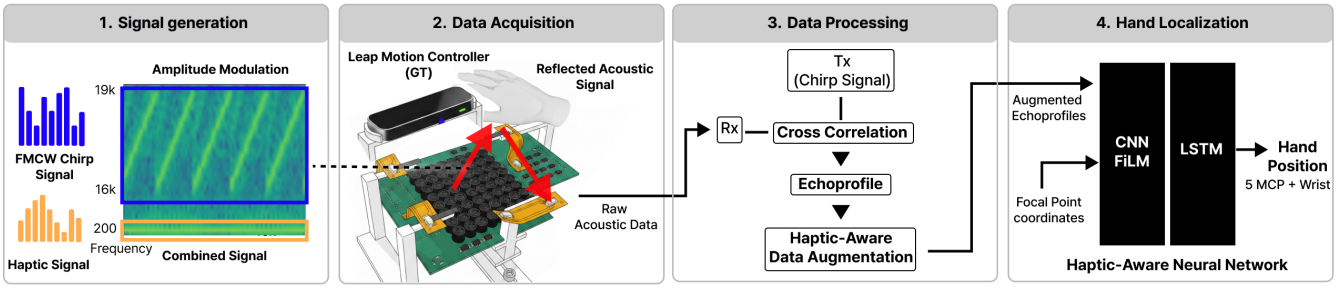

Figure 1: System Overview. The hand localization system consists of 4 main components: (1) sensing and haptic signals are simultaneously transmitted using amplitude modulation, (2) acoustic reflections from the hand are captured by 4 microphones, (3) cross-correlation is applied to generate *echoprofile*, and (4) a CNN-FiLM-BiLSTM model predicts palm position. The estimated hand coordinates are used to generate ultrasonic mid-air haptic feedback.

Doppler effect to recognize hand gestures from reflected acoustic signals [9, 35]. To improve the accuracy of motion estimation, other works have advanced both signal design and processing techniques. For example, FingerIO [28] used orthogonal frequency division multiplexing (OFDM) to track fine-grained 2D finger movements, and MilliSonic [46] leveraged frequency-modulated continuous wave (FMCW) signals to enable distance-aware tracking and localization. In particular, C-FMCW based approaches achieved a ranging resolution of approximately $0.4cm$ at 48 kHz sampling rate, enabling the detection of subtle motions. This is because spatial resolution is fundamentally constrained by the sampling rate [49]. Beyond these advances, prior work has progressed from point-based hand localization [19, 47, 50, 55] to continuous full hand reconstruction by estimating the 3D positions of individual joints. With this improved resolution, acoustic sensing can capture more fine-grained human movements, making it possible to extend beyond hand gesture recognition to tasks such as facial expression tracking [21] and upper-body pose estimation [23].

Ring-a-Pose [54] enables high-resolution hand pose estimation with a compact ring form factor using a microphone-speaker pair. WatchHand [15] extended the previous technique to commodity devices, enabling continuous finger tracking with limited hardware. In addition to relative pose reconstruction, recent work estimates absolute hand positions in 3D space and enables multi-hand tracking [22, 47, 48]. These advances extend acoustic sensing from basic input recognition to precise spatial localization. On the other hand, UMH systems require hand position information to determine where to generate haptic feedback. In this context, the same ultrasonic transducers used for haptic feedback can be repurposed for sensing, enabling hand localization without an optical sensing device and reducing system complexity. With the proposed haptic-sensing integrated system, we develop a camera-free approach that can mitigate privacy concerns and enable a compact system.

3 System

3.1 System Overview

In this paper, we propose *UltraHapSense*, a novel system that enables simultaneous ultrasonic mid-air haptic feedback and hand localization within a unified framework. *UltraHapSense* simultaneously transmits haptic and sensing signals by amplitude modulation. It then estimates hand locations by processing the reflected signals, ensuring robust localization in the presence of haptic interference. The ultimate goal of the system is to precisely localize the hand and deliver haptic feedback to the user's palm. To achieve this, the system should provide a perceptible level of haptic feedback intensity despite the simultaneous emission of sensing and haptic driving signals and maintain continuous hand tracking capabilities. As a proof-of-concept for our integrated framework, we focus on palm localization to demonstrate the fundamental feasibility of simultaneous sensing and haptic feedback. This approach aligns with practical interaction as the palm remains the primary target for mid-air haptics due to its large surface and high sensitivity. In addition, *UltraHapSense* achieves a mean hand localization latency of 2.27 ms per inference. With a 500 ms input data window and 90 % overlap, the system produces localization outputs at 20 fps. As shown in Figure 1, *UltraHapSense* consists of 4 main components: (1) signal generation, (2) data acquisition, (3) data preprocessing, and (4) hand localization.

3.2 Signal Generation

3.2.1 Ultrasonic Mid-Air Haptics. Ultrasonic mid-air haptics (UMH) is a technology that generates tactile sensations in mid-air using an array of ultrasonic transducers. By controlling the phase delay of each transducer, the emitted acoustic waves constructively interfere at a designated focal point, concentrating acoustic radiation pressure at that location. When this concentrated acoustic radiation pressure is applied to the skin, it induces subtle mechanical vibrations, which are perceived as tactile sensations. Each transducer emits a spherical wave, and according to Huygens' principle, the acoustic pressure at a position (x, y, z) is given by the superposition of the waves emitted by all transducers. Under the far-field approximation, the acoustic pressure field generated by a single ultrasonic transducer can be represented as:

$$P(r) = P_0 A \frac{D_f(\theta)}{d} e^{i(\varphi + kd)} \tag{1}$$

where P_0 denotes the reference pressure amplitude of the transducer, A represents the peak-to-peak amplitude of the driving signal, $D_f(\theta)$ is the far-field directivity function, ϕ is the initial phase, k is the wave number, and d is the distance between the transducer and the focal point.

In this way, controlling the phase of each transducer enables constructive interference at the desired focal point, thereby generating concentrated acoustic radiation pressure. Additionally, humans cannot perceive vibrations at 40 kHz (the resonant frequency of the ultrasonic transducers, MA40S4S, Murata). Therefore, the ultrasonic carrier signal must be modulated to generate low-frequency vibrations perceivable by human mechanoreceptors. Three modulation methods are commonly used in UMH, including Amplitude Modulation (AM), Lateral Modulation (LM), and Spatiotemporal Modulation (STM) [38]. First, AM generates vibrations by periodically varying the acoustic pressure of a fixed focal point. Second, LM induces vibrations by rapidly oscillating the focal point laterally while maintaining constant pressure intensity. Lastly, STM provides haptic feedback by moving a single focal point along a trajectory. Among the three modulation methods, we selected AM because it enables straightforward integration of haptic and sensing signals through a shared carrier signal. In addition, we employed a modulation frequency of 200 Hz, as humans are sensitive to UMH stimuli in this frequency range [4].

3.2.2 Ultrasonic Sensing via C-FMCW. A naive approach to active acoustic sensing is the pulse-echo method, which emits a pulse signal and directly measures the time-of-flight (ToF) of the reflected signal. However, this method is highly sensitive to ambient noise and struggles to distinguish multi-path reflections. In addition, it provides limited resolution, expressed as $\Delta d = \frac{c \cdot \tau}{2}$, where c is the speed of sound and τ denotes the pulse duration. In contrast, Frequency-Modulated Continuous Wave (FMCW) sensing continuously transmits a chirp signal whose frequency linearly sweeps over a bandwidth B during a time period T. The distance is then estimated by analyzing the frequency difference between the transmitted and received signals. The range resolution of FMCW is determined by the signal bandwidth, expressed as $R_{L\text{-}FMCW} = \frac{c}{2B}$. With a bandwidth of 3 kHz, the resolution is approximately 57.2mm, which is insufficient for the precise hand localization required for mid-air haptic feedback. Therefore, we adopt a cross-correlation-based FMCW (C-FMCW). Instead of analyzing individual frequency components, this approach identifies the segment in the received signal that matches the transmitted chirp signal (16-19 kHz). Since the received signal is a delayed version of the transmitted signal, computing the cross-correlation between them produces a peak corresponding to the propagation delay, which can be used to estimate the distance. The resolution is determined by the sampling rate and is expressed as:

$$R_{C\text{-}FMCW} = \frac{c}{2f_s} \tag{2}$$

c denotes the speed of sound in air, and f_s represents the sampling rate. In our system, the microphone samples at 40 kHz. Assuming a sound speed of $343m/s$, the resolution is approximately $4.29mm$, which is sufficient for accurate hand localization for haptic feedback.

To extract spatial features from the received signals, we perform the following signal processing steps. We first apply a band-pass filter to remove noise in the FMCW chirp frequency range (16-19 kHz), and compute the cross-correlation between the received signal and the transmitted chirp signal. The direct path signal between the ultrasound transducer and the microphone corresponds to the shortest propagation path and therefore has the minimum propagation delay. Using this signal as a reference, we temporally align the received signal and analyze the subsequent echoes reflected from the hand. This alignment compensates for residual timing offsets that remain after synchronization to the shared clock. The aligned received signal is segmented into windows corresponding to the sweep duration ($10ms$). For each segment, time delays are estimated from the cross-correlation and mapped to distances to construct a vector representing reflection intensity across the spatial range. Consistent with prior works using the C-FMCW method [20, 21, 54], we refer to this distance distribution vector as an *echo frame*. *Echo frames* computed sequentially over time are stacked along the temporal axis to form a two-dimensional *echo profile*. In the *echo profile*, the horizontal axis corresponds to time, while the vertical axis represents distance bins. Each pixel represents the reflection strength at a specific time and distance. Additionally, we compute a *differential echo profile* to suppress reflections from the static environment and emphasize changes caused by hand movement. This representation is obtained by calculating the difference between two consecutive *echo frames*. In our system, both the *echo profile* and the *differential echo profile* are used as input to the hand localization model.

3.2.3 AM-based Signal Integration. *UltraHapSense* generates a 40 kHz ultrasound drive signal that incorporates both haptic feedback and sensing signals by modulating the carrier's amplitude. As described in the section 3.2.1, a 200 Hz sinusoidal amplitude modulation is employed to enable human tactile perception. Simultaneously, as described in the Section 3.2.2, a 16-19 kHz FMCW chirp signal is utilized for hand localization. These two signals are combined into a single envelope through linear superposition, expressed as $g(t) = g_{\text{haptic}}(t) + g_{\text{fmcw}}(t)$. Then, to suppress undesired harmonic distortion during the amplitude modulation process, this envelope is processed according to Single-Sideband (SSB) modulation principles. The final modulated amplitude $A(t)$ is calculated

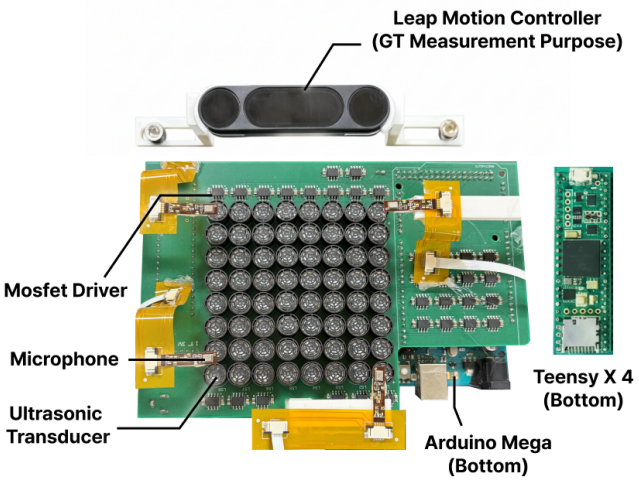

Figure 2: Top view of the fabricated hardware prototype, including the ultrasonic transducer array and microphones.

as follows:

$$A(t) = \sqrt{(1 + ag(t))^2 + (a\hat{g}(t))^2} \tag{3}$$

where a is the modulation index and $\hat{g}(t)$ denotes the Hilbert transform of $g(t)$. The resulting amplitude $A(t)$ is mapped to the duty cycle of a 40 kHz PWM signal to drive the transducers. According to the Fourier series expansion of a PWM signal, the amplitude of the fundamental frequency component is determined by the duty cycle. Since ultrasound transducers(MA40S4S) are characterized by a narrow bandwidth, they naturally filter out higher-order harmonics and emit only the fundamental frequency component. Consequently, the relationship between the amplitude and the duty cycle is defined by the following mapping equation:

$$d(t) = \frac{100}{\pi} \arcsin(A(t)) \tag{4}$$

The calculated duty cycle $d(t)$ is quantized into discrete levels of 10 %, 20 %, 30 %, 40 %, and 50 % to accommodate hardware constraints before driving the ultrasonic transducer array. Through this process, *UltraHapSense* successfully integrates the FMCW sensing signal and low-frequency haptic signal into a single carrier, thereby achieving simultaneous mid-air haptic feedback and reflection-based hand localization.

3.3 Hardware

The hardware architecture of *UltraHapSense* is primarily divided into two functional modules (Figure 2): a transmission unit for ultrasonic signal emission and an acquisition unit for capturing reflected acoustic signals from the hand. To ensure system compactness and high integration, both modules are integrated onto a single PCB. The transmission unit comprises an Arduino Mega for signal generation, a MOSFET gate driver (IXDN602SIA, Littelfuse Inc.) for the driving signal amplification, a DC power jack for external power, and an 8 $\times$ 8 array of 64 ultrasonic transducers (MA40S4S, Murata). The ultrasonic driving signals generated by the Arduino Mega are delivered to the transducers via a MOSFET gate driver. To ensure a stable power supply, the gate driver circuit incorporates two decoupling capacitors (0.1 μF and 4.7 μF) configured in parallel. This driver amplifies the input signal voltage and delivers it to the transducers, supporting a maximum driving voltage of 35 V. The array is driven at 25 V and consumes 11.8 W during simultaneous haptic and sensing signal emission. Consequently, it generates sufficient acoustic pressure to provide users with perceptible haptic feedback. The acquisition unit includes four Teensy 4.1 microcontrollers for data processing and 4 digital I^2S MEMS microphones (T5848, TDK InvenSense). We selected four microphones based on a comparison of localization accuracy across different microphone counts. (Appendix A). In the current prototype, we use a separate Teensy board for each microphone to facilitate development and debugging. The acquisition stage can also be implemented on a single MCU. The T5848 provides a high signal-to-noise ratio (SNR) of 68 dBA and a superior acoustic overload point (AOP) of 133 dB SPL. The high AOP ensures that the microphone does not saturate under the intense acoustic pressure generated by the UMH transducer array. The four microphones are arranged at the corners of a 60 mm square centered on the transducer array, elevated 20 mm above the transducer array surface. Additionally, a Leap Motion Controller is employed specifically during the data collection phase to acquire ground-truth hand position data. The Leap Motion Controller is used solely for labeling and is excluded from the final *UltraHapSense* hand localization pipeline.

3.4 Software

3.4.1 Ultrasonic Signal Emission Control. *UltraHapSense* adopts amplitude modulation based on its ability to directly control the amplitude by adjusting the duty cycle in the system. This approach minimizes system latency by reducing the computational load during signal generation and transmission. Furthermore, it maintains high signal integrity by eliminating the need for additional signal conversion, ensuring stable and reliable transmission. As described in section 3.2.3, the PC combines the 200 Hz haptic signal and the 16-19 kHz FMCW sensing signal into a single envelope, converts the resulting amplitude into duty cycle sequences. Then the PC transmits control commands to the Arduino Mega via serial communication, utilizing a simplified version of the *Ultraino* protocol [24]. Unlike the original haptic-specific protocol, our protocol transmits only pattern sequences to accommodate the continuously varying duty cycle required for the FMCW sensing signal.

The phase delay of each transducer is individually computed from its distance to the target coordinate. In the main loop of the Arduino Mega, each 40 kHz carrier period is divided into 10 discrete steps, providing a phase resolution of $\pi/5$. At each step, a 64 bit binary stream is generated in synchronization with the sync signal period. The i th bit of this word specifies the on/off state of the i th transducer at that step. Thus, each transducer is individually driven by a unique switching pattern that encodes its phase and the duty cycle. The resulting binary stream is amplified through MOSFET gate drivers and emitted from the array as 40 kHz ultrasound. To prevent clock drift between transmission and reception, all boards are driven by a shared timing source. We configure the Teensy Audio Library to use a 40 kHz sampling rate, making the I^2S frame clock (LRCLK) of the master Teensy a 40 kHz square wave. This frame clock is directly wired to the three slave Teensy boards, which sample their microphone signals according to the shared clock. The same frame clock is also connected to a GPIO pin on the Arduino Mega. This shared timing ensures consistent temporal alignment throughout the echo profile computation described in Section 3.2.2.

To maintain continuous ultrasonic transmission while simultaneously receiving new signal data from the PC, the Arduino Mega employs a double-buffering architecture. This system consists of an RX buffer for storing incoming emission patterns from the PC and an active TX buffer for the current transducer output. When a pattern-change command (0b00000000) from PC is detected, the system swaps the two buffers. This double-buffering approach ensures that the system can receive new data while maintaining a continuous ultrasonic output with minimal interference.

3.4.2 Ultrasonic Signal Acquisition. Each Teensy 4.1 samples ultrasonic data acquired from its corresponding microphone at a synchronized rate of 40 kHz via the I^2S interface, corresponding to the operating frequency of the ultrasonic transducers used for signal generation. The four Teensy boards operate in parallel, independently acquiring and transmitting data to the PC. The sampled data is buffered in an audio record queue and sent without additional on-device processing to reduce latency and improve data

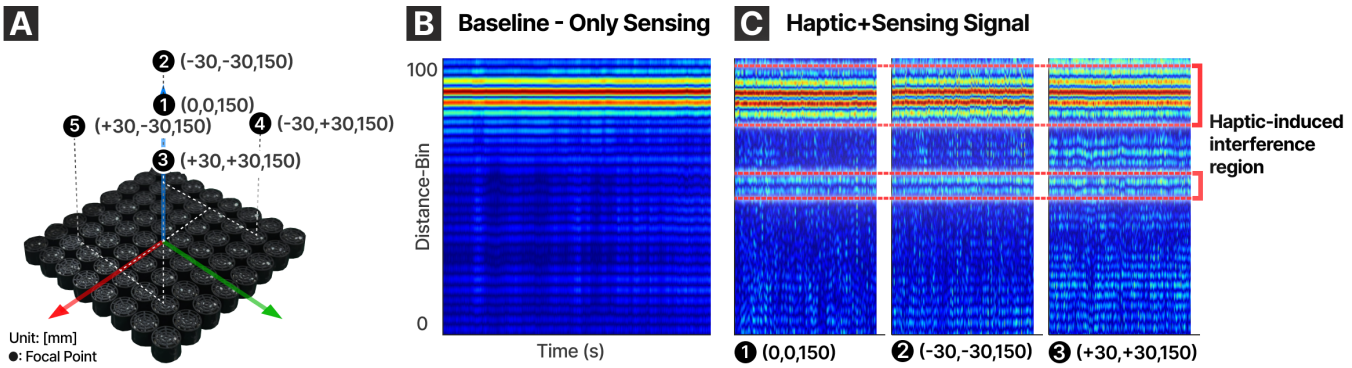

Figure 3: Overview of Haptic-Aware Data Augmentation. (A) Haptic focal point configurations used for Haptic-Aware Data Augmentation. (B) *Echo profile* when only sensing signal is transmitted (C) *Echo profile* when both sensing and haptic signals are simultaneously transmitted, with the focal points at (0, 0, 150), (-30,-30,150), (30,30,150). For each haptic condition, *echo profiles* are generated and compared with the baseline (only sensing) *echo profile*. Common interference regions by haptic signals are identified as haptic-induced inference regions that consistently show significant deviations from the only-sensing baseline across haptic conditions.

throughput for real-time processing. For efficient transmission, the collected signals are packetized into 256 byte binary blocks and transmitted to a PC through serial communication.

3.5 Data Preprocessing

The raw acoustic signals acquired using four microphones are transformed into a feature space optimized for training the deep learning model. As described in Section 3.2.2, the raw signals are first band-pass filtered to attenuate out-of-band noise, and subsequently processed via cross-correlation with the transmitted FMCW chirp, resulting in an *echo profile* that encodes distance information [15, 54]. Here, we apply frame-level normalization to reduce the effects of environmental noise and inter-user variability in hand reflection intensity. This step encourages the system to focus on reflection patterns rather than absolute signal magnitude, allowing it to capture subtle variations. We then crop 100 distance bins relative to the direct path from the transducer to improve computational efficiency and exclude irrelevant distant reflections. Considering the speed of sound and the sampling rate, the cropped *echo profile* covers approximately $0.4m$ in range, which aligns with the typical UMH interaction space of approximately $0.4m$ above the transducer array [10, 22]. Also, to capture dynamic motion characteristics beyond static information, we compute the intensity difference between consecutive frames to construct a *differential echo profile*. This process suppresses signals from stationary objects and highlights motion-induced signals. However, since the *differential echo profile* retains only the changes between frames, it loses the absolute distance information to the hand and produces little signal when the hand remains stationary [17]. Accordingly, we jointly use the *echo profile*, which preserves absolute distance information and the *differential echo profile* which captures motion information as inputs to the system. As a final step, we concatenate the *echo profiles* from 4 Rx channels with their corresponding *differential echo profiles* to form an 8-channel input tensor of size $T \times 100 \times 8$. Each channel contains distance information measured by a microphone at a different spatial location. This information is preserved along the channel dimension without explicit aggregation, allowing the model to learn spatial relationships across channels. We segment the data using a sliding window of 50 frames (500 *ms*) with a 90 % overlap to effectively capture hand movement over time.

4 UltraHapSense Operating Principle

4.1 Haptic-Aware Data Augmentation (HADA)

Achieving generalization in deep learning-based hand localization requires a large and diverse training dataset that spans a wide range of environments and conditions. However, ultrasonic sensing data is often collected under limited conditions, restricting coverage of in-deployment signal variations. Prior work has addressed the limited training data problem through various data augmentation techniques to improve generalization. Specifically, Beyond-Voice [22] applied slight shifts to both the feature map and ground truth, and Ram-Hand [48] added random noise to the feature map. However, *UltraHapSense* simultaneously generates haptic signals and sensing signals, introducing interference patterns in the *echo profile* that are absent in conventional sensing systems. If such interference patterns are not sufficiently reflected in the training data, the model fails to handle changes in the input signal distribution under real-world conditions. To address this, we propose Haptic-Aware Data Augmentation (HADA).

This haptic-induced interference can be derived from the signal integration process. Amplitude modulated ultrasound undergoes self-demodulation in air which squares the envelope. [2] Consequently, the demodulated signal available at the receiver is $A^2(t)$ rather than $A(t)$. Squaring the envelope in equation 3 leaves a cross term between the two signals in addition to the DC and self terms:

$$g_{\text{cross}}(t) = 2a^2 \cos\big(\theta(t) - \omega_h t\big) \tag{5}$$

where ω_h is the haptic modulation frequency and $\theta(t)$ is the phase of the chirp. Single-sideband modulation eliminates this term only when the envelope contains a single tone. Since our system embeds both signals into a single envelope, this term remains. [36] And it corresponds to a replica of the chirp shifted by the modulation

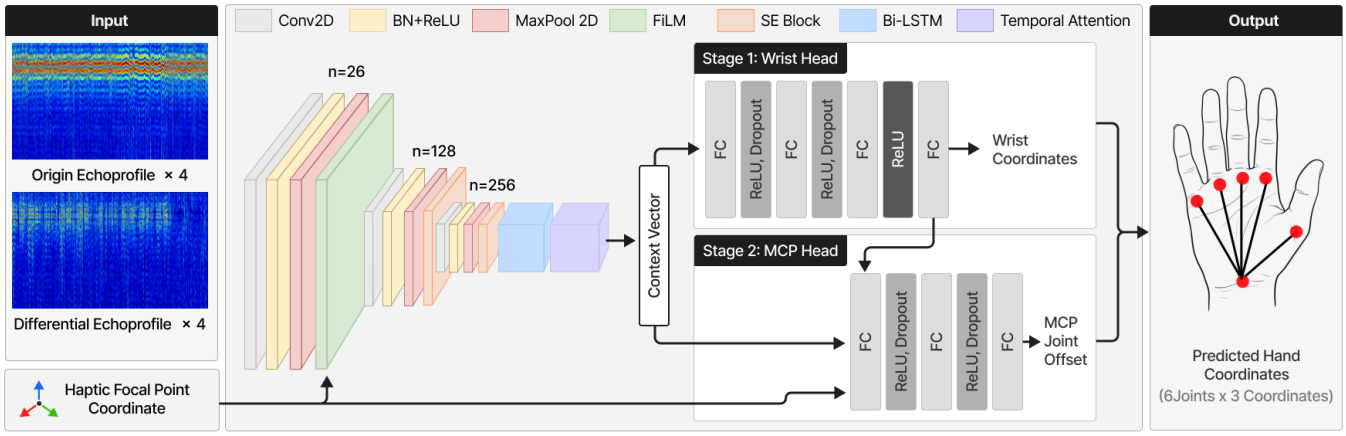

Figure 4: Overview of the haptic-aware CNN-FiLM-BiLSTM model. 4 *echo profiles* and 4 *differential echo profiles* are temporally folded and processed by a CNN backbone, where FiLM layers modulate feature maps based on the haptic focal point. The extracted features are then fed into a BiLSTM to capture temporal dependencies, enabling robust hand localization under haptic-induced interference.

frequency within the sensing band. The cross-correlation with a linearly swept chirp maps a frequency offset onto a delay offset. [18] Consequently, this replica appears in the echo profile at a fixed distance offset from the original reflection peak:

$$\Delta_{\text{bin}} = \frac{m \cdot f_{\text{mod}}}{k} \cdot f_s, \qquad k = \frac{B}{T} \tag{6}$$

Δ_{bin} denotes the distance-bin offset between the original reflection peak and the peak generated by the cross term, k is the chirp slope, f_{mod} is the haptic modulation frequency, and m denotes the harmonic order of the haptic component in the envelope. Due to the quantization of the duty cycle, the envelope also contains harmonics of the modulation frequency.

HADA selectively injects structured noise into the interference regions of the *echo profile* predicted by the model. To validate this prediction, we compute the mean amplitude at each distance bin across the *echo profile* collected under five haptic focal point locations (Figure 3(A) 1-5. center (0, 0, 150), and edge (30, 30, 150) mm). We then compute the difference between each haptic condition and the baseline (only sensing). As a result, consistent amplitude changes are concentrated in specific distance bin ranges across conditions. Two distinct interference regions (distance-bin 50-58, 75-97) are identified, which match the predicted offsets. Based on this observation, we inject Gaussian noise $\mathcal{N}(0, \sigma^2)$ into the identified interference regions when the haptic signal is emitted. We also apply uniform noise across all bins. By incorporating structured noise that reflects the physical interference patterns induced by haptic signals, the model learns to maintain stable hand localization performance regardless of whether haptic feedback is generated.

4.2 Haptic-Aware Neural Network for Continuous Hand Localization

The proposed haptic-aware model adopts a CNN-FiLM-BiLSTM architecture and estimates the coordinates of the wrist and MCP joints in 2 stages, as shown in Figure 4. In the following, we describe the design of each component.

4.2.1 Haptics-Aware Deep Learning Model. The *echo profile* obtained by computing the cross-correlation of the acquired acoustic data contains temporal dependencies. In addition, haptic signals generate acoustic interference when haptic and sensing signals are emitted simultaneously. This interference distorts the *echo profile*. To address this issue, we integrate a FiLM (Feature-wise Linear Modulation) [32] layer into a CNN-BiLSTM architecture to incorporate the haptic focal point coordinates to reflect the spatial information of the haptic feedback into hand localization. FiLM dynamically modulates the scale and shift of intermediate feature maps based on an external conditioning signal that explicitly provides the model with spatial information from the haptic feedback. In our approach, the haptic focal point $f \in \mathbb{R}^3$ serves as the external conditioning signal. Since haptic-induced interference in the *echo profile* depends on the focal point location, we incorporate f into the feature extraction stage as a conditioning signal to enable adaptive modulation of feature maps. This allows the model to learn end-to-end feature representations that are robust to interference.

Temporal Folding. The input *echo profile* sequence of length T is divided into $N_{\text{fold}} = 10$ folds. Each fold is represented as a 3D tensor of shape $(T_{\text{bin}}, T_{\text{fold}}, C)$, where $C = 8$ corresponds to four-channel *echo profiles* and their corresponding *differential echo profiles*, $T_{\textbf{bin}} = 100$ is the number of distance-bins per *echo frame*, and $T_{\textbf{fold}}$ is the number of frames per fold. This temporal folding allows the network to capture spatiotemporal patterns of the echo through convolution.

CNN-FiLM. The CNN backbone consists of three convolutional blocks, followed by a FiLM layer after the first block. The FiLM layer takes haptic focal point f as input and generates channel-wise scale γ and shift β, which are then used to modulate the feature map as follows:

$$\tilde{x} = x \odot \big(1 + \gamma(f)\big) + \beta(f) \tag{5}$$

Squeeze-and-Excitation (SE) modules are then added to help the network focus on channels relevant to hand localization.

BiLSTM and Temporal Attention. The features extracted by the CNN for the N_{fold} folds are organized into a sequence and fed into a bidirectional LSTM, which captures temporal dependencies in both directions. A subsequent temporal attention layer assigns weights to each fold based on its relevance to hand localization, producing a context vector z.

4.2.2 Hierarchical 2-Stage Estimation. As mentioned in Section 3.1, the goal of *UltraHapSense* is to localize the hand for delivering haptic feedback to the palm. Therefore, we focus on localizing the palm (specifically, the wrist joint and 5 MCP joints), which is the primary target for haptic rendering, and estimating all joints (including the fingertips) is not essential for our purpose. However, acoustic sensing is more susceptible to ambiguity in reflected signals and limited spatial resolution. To address these limitations, we adopt a 2-stage approach. The first stage estimates the absolute wrist position, while the second stage predicts the relative positions of the five MCP joints with respect to the wrist, enabling more stable and accurate localization. Together, the two stages form a palm localization pipeline that estimates the six joint positions spanning the palm region.

Stage 1. Wrist Localization. The context vector z is fed into a fully connected layer to predict the 3D global coordinates of the wrist $p^{\text{wrist}} \in \mathbb{R}^3$. Compared to fingertips, the wrist has a more constrained range of motion and simpler structure, which makes it a stable spatial anchor. Estimating the wrist first reduces the search space for subsequent joint predictions.

Stage 2. MCP Joints Estimation. In stage 2, the context vector z, the wrist position p^{wrist}, and the haptic focal point coordinates $\mathbf{f} \in \mathbb{R}^3$ are concatenated and fed into a fully connected layer. The network predicts the relative offsets of the five MCP joints with respect to the wrist $\Delta p^{\text{mcp}} \in \mathbb{R}^{5\times 3}$. The final joint coordinates are reconstructed as $p^{\text{mcp}} = p^{\text{wrist}} + \Delta p^{\text{mcp}}$. This constrains the predictions to anatomically plausible configurations, improving stability and consistency. In this stage, the haptic focal point f is incorporated as an additional input to further mitigate haptic-induced interference during joint estimation.

4.2.3 Loss Function. The two stages are trained end-to-end without blocking gradient flow, using the following loss function that jointly considers both stages:

$$\mathcal{L} = \lambda_1 \sum_i \left\| \hat{\mathbf{p}}^{(i)}_{\text{wrist}} - \mathbf{p}^{(i)}_{\text{wrist}} \right\|_2^2 + \lambda_2 \sum_i \left\| \hat{\mathbf{p}}^{(i)}_{\text{mcp}} - \mathbf{p}^{(i)}_{\text{mcp}} \right\|_2^2 \tag{6}$$

p denotes the ground truth coordinates captured using the Leap Motion Controller, and λ_1 and λ_2 are weighting factors. This joint loss enables end-to-end optimization of both stages. The model is evaluated using the Mean Absolute Error (MAE), which measures the average coordinate-wise deviation and provides an interpretable measure of overall model performance.

4.3 Haptic Feedback Intensity Validation

To validate the feasibility of concurrent modulation of sensing (16-19 kHz) and haptic (200 Hz) signals, we verified whether perceptible haptic feedback is produced. The transducer array of *UltraHapSense* has a square and symmetric structure. Thus, we selected two representative measurement locations: an edge point (-30, -30, 150) where the acoustic pressure is relatively weak and a center point (0, 0, 150) where the acoustic pressure is relatively strong (Figure 3). We defined a coordinate system with the center of the transducer array as the origin, the signal propagation direction as the z-axis, and units in millimeters (Figure 3 (A)). An electret condenser microphone is mounted on a 3D stage to measure acoustic pressure, while the ultrasonic transducer array is positioned on the ground facing upward. The microphone output is amplified using an operational amplifier and then analyzed using FFT to extract the magnitude of the modulated frequency component. At each location, signals are recorded for 30 seconds, and the averaged value is used as the final acoustic pressure. The results show 129.54 dBSPL at the edge and 133.93 dBSPL at the center, both exceeding the perceptual threshold of 118 dBSPL for human hand haptic perception [31]. This indicates that *UltraHapSense* can generate perceptible haptic feedback while simultaneously emitting sensing signals.

Additionally, we measured the acoustic radiation pressure around the focal point at 2 mm intervals and compared the spatial distributions between the haptic-only signal and the simultaneous haptic and sensing signal under the *Center Haptic* condition. The maximum sound pressure level (SPL) of the main lobe was 134.50 dB SPL for the haptic-only signal and 133.93 dB SPL for the simultaneous signal, corresponding to a difference of 0.57 dB. The main-lobe widths at -3 dB were 17.3 mm and 17.2 mm, respectively. Also, the peak-to-sidelobe ratios were 11.2 dB and 11.5 dB, with a difference of 0.3 dB. This confirms that integrating the sensing signal into the haptic signal does not substantially degrade haptic feedback quality, either in terms of the peak sound pressure or the spatial pressure distribution around the focal point.

5 User Study 1: Continuous Hand Localization

We conducted a user study to collect acoustic signals reflected from the hands during continuous hand movement, which serves as training data for our localization model. We recruited 11 participants (aged 20-30 years, mean = 24, SD = 3.2; 5 males) with varying hand sizes. Participants' hand length (from wrist to tip of the middle finger) ranged from 16.5 to 21 *cm* (mean = 17.8, SD = 1.3) and palm width ranged from 7 to 9 *cm* (mean = 8.09, SD = 0.73), covering the central range of the adult hand size distribution. The study took two hours to complete. This study was approved by the local Institutional Review Board (IRB).

5.1 Study Design and Procedure

We collected data in a small conference room (Figure 5 (A)) for the open palm gesture, which is a naturally extended hand posture and one of the most commonly used gestures for system control in mid-air gesture sets [12, 56]. During the sessions, the target position and the participant's hand position were simultaneously visualized in real time on a monitor (figure 5 (D)). The target position was randomly updated within a 20 cm $\times$ 20 cm $\times$ 20 cm workspace that supports haptic feedback generation. We prevented participants from becoming biased toward hand trajectories by randomizing the target positions rather than following predefined paths. By instructing participants to move across the entire workspace, this

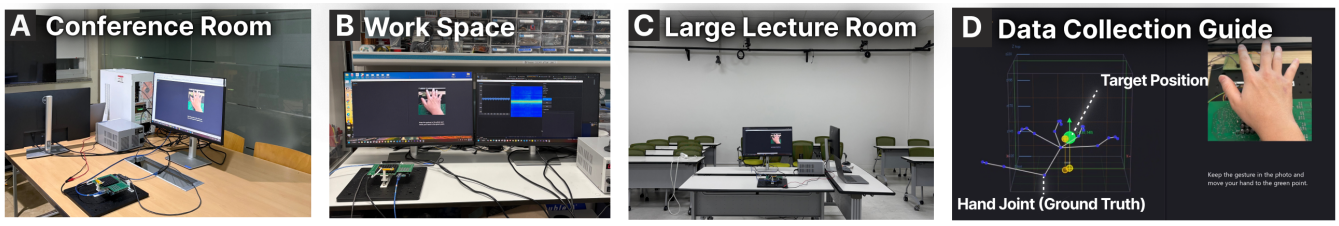

Figure 5: Three different environmental conditions and data collection interface. (A) small conference room (baseline), (B) workspace, and (C) large lecture room. (D) User guidance interface displaying the target position and hand joints ground truth using Leap Motion Controller, instructing participants to move their hand to the designated location during data collection.

design minimized bias toward specific spatial locations in the data. Acoustic signals were sampled at 25 μs intervals, while the ground truth hand joint coordinates were simultaneously recorded using a Leap Motion Controller.

After signing the informed consent form, participants received a brief explanation of the experimental procedure. Then, the experimenter measured the participant's hand size and conducted a training session to help them become familiar with the task. We acquired data under 3 signal conditions: FMCW signal only (*Only Sensing*), FMCW signal with haptic signal (*Center Haptic*), and FMCW signal with haptic feedback generated at the microphone location (*Edge Haptic*). By collecting and training on data from three signal conditions, we aim to evaluate not only the hand localization accuracy when haptic signals are transmitted, but also the impact on accuracy under strong interference when haptic feedback is generated toward the microphone. Each condition included 10 sessions, for a total of 30, and each lasted 2 minutes. Participants were instructed to move their hand to the target position while maintaining an open palm gesture. A 30 second rest was provided between sessions, with additional breaks permitted as needed.

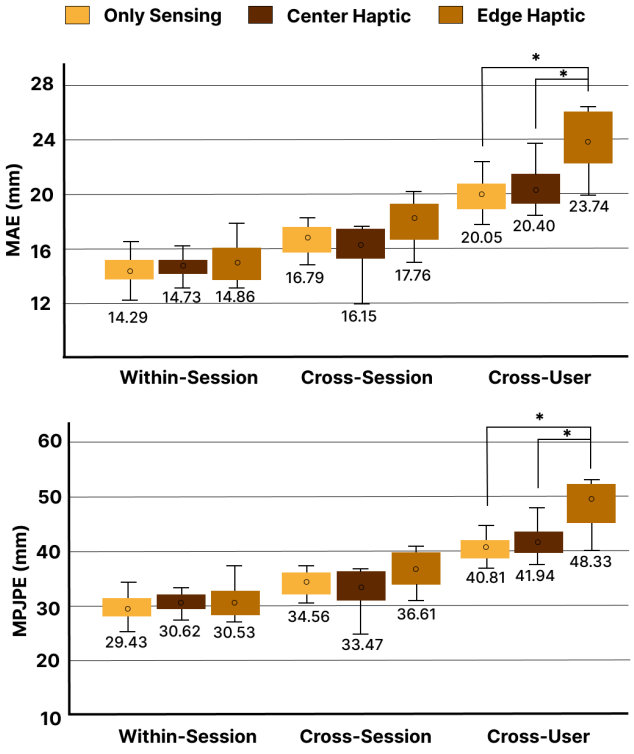

Figure 6: Hand localization performance results. MAE (up) and MPJPE (down) across three protocols (within-session, cross-session, and cross-user) under three signal conditions (*Only Sensing, Center Haptic, and Edge Haptic*). Circles indicate the mean, numerical labels indicate the corresponding exact values, and * denotes the significance level of $p < .05$.

5.2 Result

We collected a total of 660 minutes of data from 11 participants, with 220 minutes per signal type. During data collection, some participants occasionally moved their fingers due to reduced concentration and fatigue. However, we included all recorded data in the dataset without exclusion. This variability increases dataset diversity and enhances model robustness. Following prior work on hand pose estimation via acoustic sensing [22], we use Mean Absolute Error (MAE) as the primary metric for its intuitive interpretability and analyze errors along the x,y, and z axes. We also report the Mean Per-Joint Position Error (MPJPE) to enable direct comparison with previous studies [48].

To evaluate the hand localization performance of *UltraHapSense*, we train and evaluate models under three protocols. First, the cross-user model serves as a foundation model trained on data from multiple users, providing a generalized initialization for fine-tuning. We train the model on data from all participants except one and test it on the held-out participant. This reflects a practical deployment scenario where a pre-trained model is directly applied to a new user. Second, the cross-session model fine-tunes the cross-user foundation model on nine sessions from the target user and evaluates it on the held-out session. This reflects a practical scenario in which only a limited amount of data from the target user is available for fine-tuning, while the model is expected to generalize to unseen sessions. Lastly, the within-session model fine-tunes the cross-user foundation model using the first 8 sessions and half of the last 2 sessions from the target user, and evaluates it on the remaining data. In other words, the model first learns general hand motion patterns from data across multiple users and is then adapted to each target user via fine-tuning.

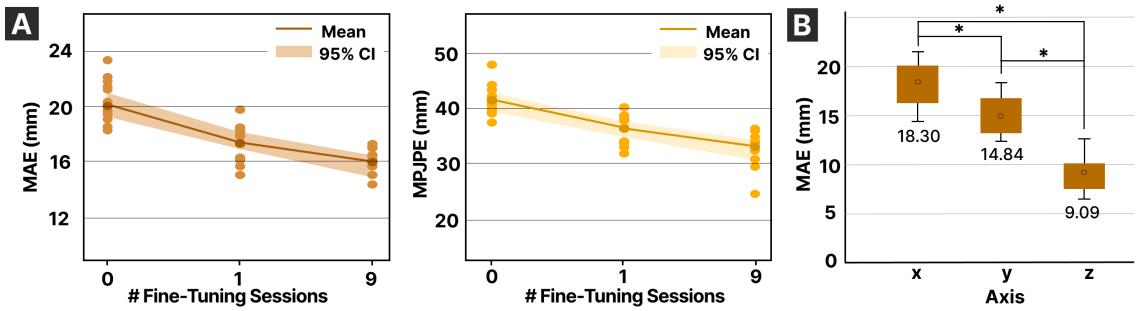

Figure 7: (A) Effect of fine-tuning data size on hand localization performance in the cross-session protocol: MAE(left) and MPJPE (right). Lines indicate the mean, and shaded regions represent the 95% confidence interval across participants. (B) Per-axis MAE under the cross-session model. Circles indicate the mean, numerical labels indicate the corresponding exact values, and * denotes the significance level of $p < .05$.

5.2.1 Cross-User Model. We evaluated the cross-user model using a leave-one-participant-out protocol to assess performance on unseen users without user-specific calibration. As shown in Figure 6, signal condition significantly affected localization performance for both MAE ($F(2, 20) = 25.09$, $p < .001$) and MPJPE ($F(2, 20) = 25.99$, $p < .001$). Post-hoc analysis showed no significant difference between the *Only Sensing* and *Center Haptic* conditions, whereas the *Edge Haptic* condition showed significantly larger localization errors than both other conditions.

These results indicate that strong haptic inference significantly degrades localization performance in the cross-user model, especially when the haptic focal point is located near the microphones. While no significant differences across signal conditions were observed in the within-session and cross-session models, the effect of the edge haptic condition was statistically significant only in the cross-user model. The *Center Haptic* condition introduces relatively weak interference, which the proposed haptic-aware neural network and data augmentation approach can effectively mitigate interference without user-specific data. In contrast, the *Edge Haptic* condition represents a worst-case scenario in which a strong haptic focal point is generated near the microphones, causing severe distortion in the *echoprofile*. However, based on the results in section 5.2.3 and section 5.2.2, even such extreme interference can be effectively mitigated through fine-tuning with a small amount of user-specific data.

5.2.2 Cross-Session Model. *UltraHapSense* achieved stable hand localization performance across all signal conditions, including *Only Sensing*, *Center Haptic*, and *Edge Haptic*, as shown in Figure 6. Since normality assumptions were violated, we conducted the Friedman test, which revealed no significant effect of signal condition on either MAE ($\chi^2 = 5.64$, $p = .060$) or MPJPE ($\chi^2 = 3.82$, $p = .148$). These results indicate that *UltraHapSense* maintains stable localization performance even during simultaneous mid-air haptic feedback.

Ablation study Next, we conducted an ablation study by varying the number of fine-tuning sessions (0, 1, and 9 sessions) to evaluate the effect of fine-tuning data size on localization performance. As shown in Figure 7(A), MAE progressively decreased from 20.40 mm (SD = 1.46) without fine-tuning to 17.57 mm (SD = 1.45) with a single fine-tuning session and further to 16.15 mm (SD = 1.70) with nine sessions. Similar trends were observed for MPJPE. Since the data were not normally distributed, we conducted the Friedman test, which showed a significant effect of the number of fine-tuning sessions on both metrics ($p < .001$).

These results indicate that increasing the amount of fine-tuning data consistently improves localization performance. Notably, even a single fine-tuning session substantially improved performance compared to no fine-tuning, suggesting that effective user adaptation can be achieved with minimal calibration data. Moreover, the performance gain from 1 to 9 sessions was smaller than that from 0 to 1 session, indicating that a single fine-tuning session can be sufficient for practical deployment.

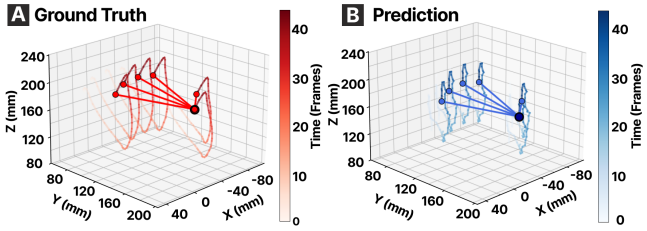

Figure 8: Comparison of (A) ground truth and (B) predicted hand joint trajectories over 45 frames

Per-Axis Error Analysis We then evaluated localization accuracy along each axis to analyze axis-wise error distribution. Since the data satisfied the assumptions of normality and sphericity, we conducted a one-way repeated-measures ANOVA, which showed a significant effect of axis on localization error ($F(2, 20) = 80.74$, $p < .001$). Post-hoc analysis showed significant differences across all axis pairs. As shown in Figure 7 B, the localization error was highest along the x-axis and lowest along the z-axis.

These differences across axes reflect the system's physical characteristics. The *echo profile* encodes distance information based on ultrasonic signals reflected from the hand to the microphones. As the reflection is oriented primarily along the z-axis, hand motion along the z-axis is most directly captured in the *echo profile*, resulting in the lowest error among the axes. Moreover, the higher error

Table 1: Overview of related work on hand localization using acoustic sensing. The reported accuracies were obtained under different hardware configurations, datasets, and evaluation protocols. This table is intended to highlight the design differences from prior work, rather than to claim comparable or superior accuracy. (RAM-Hand [48] primarily targets full hand pose estimation for both hands; its wrist localization accuracy is listed here for reference only.)

System	Model	Joints	Hardware Platform	Haptic	Protocol	Accuracy	
						MAE	MPJPE
Beyond Voice [22]	CNN+LSTM	21 joints	1 spk + 7 mic array	×	Cross-user, 11 users	16.47 mm	21.70 mm
RAM-Hand [48]	Transformer+Attention	Wrist only	1 spk + L-shaped 16 mic array	×	Within-session, 8 users	—	29.3 mm
UltraHapSense (Ours)	CNN+FiLM+BiLSTM	6 joints	64 transducer + 4 mics	✓	Within-session, 11 users	14.73 mm	30.62 mm

Table 2: Ablation study of each core component. Baseline denotes a simple CNN+LSTM model without FiLM layer, HADA, and the two-stage structure.

Model	MAE (mm)	MPJPE (mm)
Baseline (CNN+LSTM)	18.80	36.59
+ Two-Stage	17.23	35.69
+ FiLM	15.51	32.34
+ HADA	15.07	31.23
UltraHapSense (All)	14.73	30.62

along the x-axis compared to the y-axis reflects the symmetry of the hand. The symmetric structure of the hand produces similar echo patterns, introducing ambiguity in localization along the x-axis. This finding is consistent with our exploratory pilot study that varied the number (1-4) and placement (center and corners) of microphones, confirming that microphone configuration directly affects axis-wise localization performance.

5.2.3 Within-Session Model. Lastly, we evaluated the within-session model. Consistent with the cross-session results, no statistically significant differences were observed across the three signal conditions for either MAE ($p = .356$) or MPJPE ($p = .913$), indicating that simultaneous haptic feedback does not significantly affect localization performance.

For direct comparison with prior work, we also report the within-session results under the *Center Haptic* condition, in which haptic and sensing signals are transmitted simultaneously (Table 1). Under this condition, *UltraHapSense* achieved an MAE of 14.73 mm (SD = 0.79) and an MPJPE of 30.62 mm (SD = 1.63). The qualitative results are shown in Figure 8. Additionally, we conducted an ablation study to isolate the contribution of each component. Under the within-session protocol with the *Center Haptic* condition, we added the two-stage structure, FiLM layer, and HADA to a baseline CNN-LSTM model and compared performance (Table 2). With our final model, the MAE decreases by 21.6% relative to the baseline. Each component independently improved both MAE and MPJPE. Especially, HADA achieved the greatest performance improvement among the three components. This result suggests that interference induced by haptic signals is the primary limitation on hand localization in our system, and HADA directly mitigates this interference. Overall, these results indicate that our system maintains localization accuracy sufficient for providing haptic feedback on the palm, even while the sensing and haptic signals are emitted simultaneously.

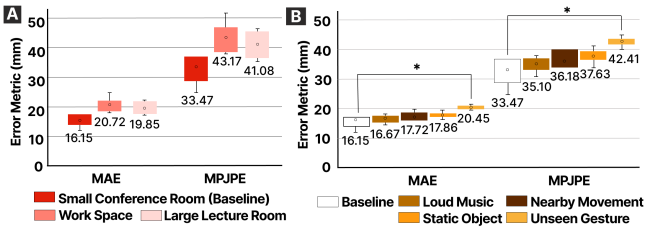

Figure 9: Robustness evaluation under various real-world conditions. (A) Hand localization performance (MAE and MPJPE) across three different environments. (B) Hand localization performance under various noise conditions, including loud music, nearby human movement, static objects, and unseen gestures. Circles indicate the mean, numerical labels indicate the corresponding exact values, and * denotes the significance level of $p < .05$.

5.3 Robustness Evaluation

We then evaluated the robustness of *UltraHapSense* under various real-world environments and noise conditions. The same 11 participants from the previous user study were recruited, and additional data were collected on a different day to naturally incorporate variability in user posture and movement. This setup enables a more realistic assessment of the model's generalization performance in practical scenarios. Following the same methodology in the user study, participants were instructed to maintain a hand gesture while moving their hand to follow the target point displayed on a screen, and each session lasted 2 minutes. To evaluate robustness under realistic noise conditions, we collected six additional sessions per participant, each corresponding to a different noise condition. The

conditions are as follows: (1) Different environments: a large lecture room and a workspace, with the small conference room as a baseline (2) Nearby human movement: a person walked at a distance of 1 m (3) Nearby static object: a metal bottle was placed 5 cm from the ultrasonic transducer array (4) Loud music: a random YouTube video was played at approximately 85 dB from a laptop located 1 m away from the participant (5) Unseen gesture: a pointing gesture (index finger extended) instead of a open palm gesture. All robustness evaluations were conducted without retraining on user-specific data.

As shown in Figure 9, *UltraHapSense* maintained stable localization performance under various real-world noise conditions, including different environments, loud music, static objects, and nearby human movement, with no significant performance degradation observed (p = .156–.313). In contrast, performance significantly degraded under the *Unseen Gesture* condition, where the model was evaluated on hand shapes not observed during training (p = .031). This environmental robustness stems from the use of ultrasonic frequency bands and *differential echo profiles*, which reduce the impact of static environmental noise. However, the performance degradation under the *Unseen Gesture* condition suggests limited generalization to unseen hand geometries, as different hand shapes produce substantially different echo patterns. Additionally, to evaluate robustness to changes in hand orientation, we evaluated the model on data collected with the hand rotated by 45° in the horizontal plane while keeping the palm parallel to the array. The model's performance degraded substantially, with an MAE of 21.35 mm and an MPJPE of 44.23 mm. As in the *Unseen Gesture* condition, this result suggests that the learned representations rely on reflection patterns observed during training and do not generalize well to hand configurations outside the training distribution. Lastly, we trained and evaluated the model using all 21 hand joints to evaluate localization performance across the entire hand. The model achieved an MAE of 16.41 mm and an MPJPE of 34.28 mm (*Center Haptic* condition, within-session protocol). This suggests that *echo profiles* capture information about the overall hand pose, rather than only the palm position.

6 User Study 2: Real-Time Haptic Subjective Evaluation

We conducted an additional user study to evaluate whether users could perceive haptic feedback with appropriate intensity and timing when the haptic focal point was generated in real time based on hand positions estimated by our system. To enable comparison under conditions consistent with existing UMH systems, we compared our system with a baseline that used a Leap Motion Controller for hand localization.

8 new participants were recruited and provided informed consent before the experiment. They first completed a practice session to familiarize themselves with ultrasonic haptic feedback and then proceeded to the main experiment. Participants moved their hands in specified directions within the 3D space above the transducer array. At each moment, haptic feedback was generated to the hand position estimated by the system or Leap Motion Controller. After perceiving the feedback, participants rated two aspects on a 7-point Likert scale. *(1) "How strong did the vibration feel?" (2) "Was the timing of the feedback appropriate?"* Each participant completed 15 trials per condition, for a total of 30 trials. The order of the two conditions was counterbalanced across participants.

The Shapiro-Wilk test showed that both conditions satisfied the normality assumption ($p > .05$). Therefore, we performed paired t-tests. Perceived intensity was rated 5.52 (SD = 0.84) for the Leap Motion condition and 5.40 (SD = 0.71) for our system, with no significant difference between the conditions ($t(7) = 1.83$, $p = .111$). In contrast, timing appropriateness was rated 4.92 (SD = 1.19) for the Leap Motion condition and 3.54 (SD = 0.56) for our system, showing a significant difference ($t(7) = 3.68$, $p = .008$). This difference is attributable to the additional latency introduced by our system's 500 ms input window for localization, which is required for stable tracking. Nevertheless, participants consistently perceived the haptic feedback in both conditions, and perceived intensity did not differ significantly between them.

7 Discussion

In this section, we discuss the generalization capability and robustness as an integrated haptic-sensing system of *UltraHapSense* and outline the directions for future work to address the system's current limitations.

Consistent with previous hand tracking systems based on active acoustic sensing [15, 22, 48], localization performance remained robust under diverse environmental noises but degraded for unseen gestures. This suggests that the primary limitation of the proposed system is the ability to generalize to hand geometries not represented in the training data. Furthermore, the challenge of generalizing to unseen gestures is closely related to the problem of reconstructing the full hand pose. Just as unseen gestures introduce distributional shifts in reflection patterns, full hand reconstruction requires capturing diverse finger movements and poses that lead to substantial variations in the *echo profile*. To further investigate the feasibility of extending the system to full hand pose estimation with 21 joints, we additionally analyze joint-wise localization accuracy.

The joint-wise accuracy analysis in Table 4 shows two key patterns that reflect characteristics of hand localization using acoustic sensing. First, across all models, the wrist joint consistently shows lower MAE and MPJPE than the five MCP joints (average MAE across all conditions: wrist 14.99 mm, MCP joints 17.51 mm). This is because the wrist has a relatively constrained range of motion and serves as a stable spatial anchor for the hand localization. Joints closer to the fingertips have higher degrees of freedom, leading to higher variability in acoustic reflection patterns even under similar hand positions. This aligns with the proposed two-stage localization strategy, where the wrist is first localized, and the MCP joints are then estimated. Second, the thumb MCP joint (SD 9.24-12.47) shows a notably larger standard deviation in MPJPE compared to other joints (average of 2.63 for other MCP joints). This reflects the thumb's unique anatomical structure and motion characteristics. Unlike the other fingers, the thumb has a higher degree of independent movement and is positioned laterally relative to the hand, resulting in higher variability in acoustic reflection patterns. These findings indicate that localization error increases with joint mobility. Therefore, extending to full hand pose estimation poses

additional challenges, as distal joints have greater motion variability and shorter inter-joint distances, making reflection patterns harder to distinguish. Incorporating anatomical constraints of finger joints into the model or improving spatial resolution through additional microphones are promising directions for enabling full hand reconstruction.

To extend *UltraHapSense* toward full hand reconstruction, there will be two potential directions in terms of hardware and signal design. From a hardware perspective, increasing the number of microphones will improve spatial resolution and help distinguish reflection patterns across individual finger joints, as demonstrated in Ram-Hand [48]. Furthermore, the per-axis error suggests that the x-axis is most susceptible due to the symmetric reflection patterns of the hand, indicating that asymmetric microphone placements could improve localization accuracy along this axis. Additionally, Ring-a-Pose [54] achieved continuous full hand reconstruction using a single microphone-speaker pair. This suggests that full hand reconstruction is a feasible goal for *UltraHapSense*, which employs 64 ultrasonic transducers and multiple microphones. Second, signal design can be further improved. Specifically, the current system employs a single FMCW signal in the 16-19 kHz band. Using multiple FMCW signals across different frequency bands enables the acquisition of more diverse and higher-resolution reflection signals. In particular, *UltraHapSense* employs 64 ultrasonic transducers, which can be divided into multiple groups to simultaneously emit multiple FMCW signals. Through these improvements in both hardware and signal design, *UltraHapSense* can serve as an integrated haptic-sensing platform capable of fully replacing the optical sensors.

From a practical perspective, we evaluated whether integrating the sensing signal introduces audible noise. The 16–19 kHz band lies near the upper limit of human hearing, which varies across individuals, and hardware nonlinearities may cause part of the signal energy to leak into the audible range. To assess this effect, we placed a microphone approximately 50 cm from the transducer array to approximate a practical usage scenario and measured a sound level under two conditions: haptic-only emission and simultaneous haptic and sensing emission. Adding the sensing signal increased the sound level by approximately 11 dB. This increase represents a practical cost of performing haptic feedback and sensing simultaneously with a single array, and reducing the resulting audible noise is left for future work. Also, our system uses amplitude modulation (AM) to generate ultrasonic haptic feedback. However, prior work has shown that modulation techniques such as lateral modulation (LM) and spatiotemporal modulation (STM) can provide richer haptic experiences. [38, 39] These techniques require rapid focal point movement, making the update rate of the emission patterns an important factor. Therefore, we measured the time required to generate and transmit the emission pattern for a new focal point. Using an Intel Core i9-14900K CPU and an NVIDIA RTX 4070 GPU, pattern generation took 2.9 ms and serial transmission took 87.0 ms, for a total of 89.9 ms. Since most of this latency comes from the limited bandwidth of serial transmission, it could be further reduced by increasing the transmission rate or preloading the modulation patterns onto the microcontroller.

Furthermore, to integrate our system into commercial platforms such as Ultraleap [1], further validation is needed to determine whether the proposed approach can be applied without modification to transducer arrays with different geometries and configurations. Changes in array geometry alter both the reflection paths from the focal point and the relative positions of the microphones. Evaluating the effects of these changes remains future work.

In addition, the current system is evaluated within a constrained $(20 \times 20 \times 20)$ cm workspace, whereas real-world interaction scenarios require hand tracking over a larger spatial range [13, 45]. The C-FMCW-based sensing employed in this system uses a 10 ms chirp length and 100 distance bins, enabling an effective sensing range of approximately 0.4 m, which exceeds the current workspace. This indicates that hand tracking can be extended beyond the current spatial constraints. However, increasing the workspace also increases the distance between the hand and microphones, which may reduce the strength of reflected signals. As a result, experimental validation of estimation performance under extended range conditions remains an open research question for future work.

8 Conclusion

Conventional UMH systems rely on external optical sensing devices such as the Leap Motion Controller or OptiTrack for hand localization, which introduces limitations including increased system complexity and potential privacy concerns. Therefore, we present *UltraHapSense*, a system that simultaneously enables mid-air haptic feedback generation and hand localization using only a UMH ultrasonic transducer array and four small microphones. Through amplitude modulation-based signal integration of haptic signal and sensing signals, the proposed system maintains a sufficient haptic intensity of 133.93 dBSPL to perceive while concurrently supporting accurate hand localization. To achieve robust localization under acoustic interference induced by haptic signals, we introduce a haptic-aware data augmentation (HADA) and a neural network model that leverages haptic focal point coordinates as a conditioning signal. Experimental results showed that the proposed system achieves localization accuracies of 14.73 mm MAE and 30.62 mm MPJPE while simultaneously providing haptic feedback, and shows no significant performance degradation under diverse noise conditions. Furthermore, the system effectively adapts to new users through fine-tuning on a single session's data (2 minutes), demonstrating its practical applicability in real-world scenarios. These results indicate that a single ultrasonic transducer array with four microphones can serve as an integrated platform for both haptic feedback and acoustic sensing, while mitigating the system size and privacy limitations of conventional UMH systems.

Acknowledgments

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2026-RS-2024-00436398) supervised by the IITP(Institute for Information Communications Technology Planning Evaluation)(50%). And this work was supported by the Institute of Information Communications Technology Planning Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2026-25507551, Development of Egocentric Data Sensing and Spatial Immersive Experience Technology)(50%).

References

[1] 2025. *Ultraleap.* https://docs.ultraleap.com/

[2] HO Berktay. 1965. Possible exploitation of non-linear acoustics in underwater transmitting applications. *Journal of Sound and Vibration* 2, 4 (1965), 435–461.

[3] Tom Carter, Sue Ann Seah, Benjamin Long, Bruce Drinkwater, and Sriram Subramanian. 2013. UltraHaptics: multi-point mid-air haptic feedback for touch surfaces. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology* (St. Andrews, Scotland, United Kingdom) *(UIST '13)*. Association for Computing Machinery, New York, NY, USA, 505–514. doi:10.1145/2501988.2502018

[4] Antonio Cataldo, William Frier, and Patrick Haggard. 2023. Quantifying spatial acuity of frequency resolved midair ultrasound vibrotactile stimuli. *Scientific Reports* 13, 1 (2023), 21149.

[5] Yanjun Chen, Xuewei Liang, Si Chen, Yuwen Chen, Hongnan Lin, Hechuan Zhang, Chutian Jiang, Feng Tian, Yu Zhang, Shanshan Yao, and Teng Han. 2022. HapTag: A Compact Actuator for Rendering Push-Button Tactility on Soft Surfaces. 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 70, 11 pages. doi:10.1145/3526113.3545644

[6] Siya Choudhary, Romain Nith, Yun Ho, Jas Brooks, Mithil Guruvugari, and Pedro Lopes. 2025. Adaptive Electrical Muscle Stimulation Improves Muscle Memory. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 361, 11 pages. doi:10.1145/3706598.3713676

[7] Artem Dementyev, Alex Olwal, and Richard F. Lyon. 2020. Haptics with Input: Back-EMF in Linear Resonant Actuators to Enable Touch, Pressure and Environmental Awareness. 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, 420–429. doi:10.1145/3379337.3415823

[8] Hyunjae Gil, Hyungki Son, Jin Ryong Kim, and Ian Oakley. 2018. Whiskers: Exploring the Use of Ultrasonic Haptic Cues on the Face. 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, 1–13. doi:10.1145/3173574.3174232

[9] 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

[10] Keisuke Hasegawa and Hiroyuki Shinoda. 2018. Aerial Vibrotactile Display Based on Multiunit Ultrasound Phased Array. *IEEE Transactions on Haptics* 11, 3 (2018), 367–377. doi:10.1109/TOH.2018.2799220

[11] Takayuki Hoshi, Masafumi Takahashi, Takayuki Iwamoto, and Hiroyuki Shinoda. 2010. Noncontact Tactile Display Based on Radiation Pressure of Airborne Ultrasound. *IEEE Transactions on Haptics* 3, 3 (2010), 155–165. doi:10.1109/TOH.2010.4

[12] Masoumehsadat Hosseini, Tjado Ihmels, Ziqian Chen, Marion Koelle, Heiko Müller, and Susanne Boll. 2023. Towards a Consensus Gesture Set: A Survey of Mid-Air Gestures in HCI for Maximized Agreement Across Domains. 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 311, 24 pages. doi:10.1145/3544548.3581420

[13] Thomas Howard, Maud Marchal, Anatole Lécuyer, and Claudio Pacchierotti. 2019. PUMAH: pan-tilt ultrasound mid-air haptics for larger interaction workspace in virtual reality. *IEEE transactions on haptics* 13, 1 (2019), 38–44.

[14] Thomas Howard, Maud Marchal, Anatole Lécuyer, and Claudio Pacchierotti. 2020. PUMAH: Pan-Tilt Ultrasound Mid-Air Haptics for Larger Interaction Workspace in Virtual Reality. *IEEE Transactions on Haptics* 13, 1 (2020), 38–44. doi:10.1109/TOH.2019.2963028

[15] 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. arXiv:2602.21610 [cs.HC] https://arxiv.org/abs/2602.21610

[16] Ruiheng Lan, Xu Sun, Qingfeng Wang, and Bingjian Liu. 2024. Ultrasonic Mid-Air Haptics on the Face: Effects of Lateral Modulation Frequency and Amplitude on Users' Responses. 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 712, 12 pages. doi:10.1145/3613904.3642417

[17] 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. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 3, Article 99 (Sept. 2025), 27 pages. doi:10.1145/3749469

[18] Nadav Levanon and Eli Mozeson. 2004. Radar signals. *New Jersey* (2004), 426.

[19] Dong Li, Jialin Liu, Sunghoon Ivan Lee, and Jie Xiong. 2020. FM-track: pushing the limits of contactless multi-target tracking using acoustic signals. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems* (Virtual Event, Japan) *(SenSys '20)*. Association for Computing Machinery, New York, NY, USA, 150–163. doi:10.1145/3384419.3430780

[20] Ke Li, Ruidong Zhang, Siyuan Chen, Boao Chen, Mose Sakashita, Francois Guimbretiere, and Cheng Zhang. 2024. EyeEcho: Continuous and Low-power Facial Expression Tracking on Glasses. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 319, 24 pages. doi:10.1145/3613904.3642613

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

[22] Yin Li, Rohan Reddy, Cheng Zhang, and Rajalakshmi Nandakumar. 2024. BeyondVoice: Towards Continuous 3D Hand Pose Tracking on Commercial Home Assistant Devices. In *2024 23rd ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. 151–162. doi:10.1109/IPSN61024.2024.00017

[23] 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

[24] Asier Marzo, Tom Corkett, and Bruce W. Drinkwater. 2018. Ultraino: An Open Phased-Array System for Narrowband Airborne Ultrasound Transmission. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* 65, 1 (2018), 102–111. doi:10.1109/TUFFC.2017.2769399

[25] Alex Mazursky, Shan-Yuan Teng, Romain Nith, and Pedro Lopes. 2021. MagnetIO: Passive yet Interactive Soft Haptic Patches Anywhere. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 213, 15 pages. doi:10.1145/3411764.3445543

[26] Yasuaki Monnai, Keisuke Hasegawa, Masahiro Fujiwara, Kazuma Yoshino, Seki Inoue, and Hiroyuki Shinoda. 2014. HaptoMime: mid-air haptic interaction with a floating virtual screen. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology* (Honolulu, Hawaii, USA) *(UIST '14)*. Association for Computing Machinery, New York, NY, USA, 663–667. doi:10.1145/2642918.2647407

[27] Lendy Mulot, Thomas Howard, Guillaume Gicquel, Claudio Pacchierotti, and Maud Marchal. 2024. Bimanual ultrasound mid-air haptics for virtual reality manipulation. *IEEE Transactions on Visualization and Computer Graphics* (2024).

[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] Karri Palovuori, Ismo Rakkolainen, and Antti Sand. 2014. Bidirectional touch interaction for immaterial displays. In *Proceedings of the 18th International Academic MindTrek Conference: Media Business, Management, Content & Services* (Tampere, Finland) *(AcademicMindTrek '14)*. Association for Computing Machinery, New York, NY, USA, 74–76. doi:10.1145/2676467.2676503

[30] Changhyeon Park, Seokmin Hong, and Jaeyoung Park. 2024. Effect of Rendering Virtual Vibrotactile Motion on the Perceived Lateral Force. *IEEE Access* 12 (2024), 173792–173799. doi:10.1109/ACCESS.2024.3502903

[31] Changhyeon Park, Yubin Lee, and Sang Ho Yoon. 2025. UltraBoard: Always-available Wearable Ultrasonic Mid-air Haptic Interface for Responsive and Robust VR Inputs. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 2, Article 44 (June 2025), 31 pages. doi:10.1145/3731413

[32] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[33] Ismo Rakkolainen, Euan Freeman, Antti Sand, Roope Raisamo, and Stephen Brewster. 2020. A survey of mid-air ultrasound haptics and its applications. *IEEE Transactions on Haptics* 14, 1 (2020), 2–19.

[34] Ted Romanus, Sam Frish, Mykola Maksymenko, William Frier, Loïc Corenthy, and Orestis Georgiou. 2019. Mid-Air Haptic Bio-Holograms in Mixed Reality. In *2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. 348–352. doi:10.1109/ISMAR-Adjunct.2019.00-14

[35] Wenjie Ruan, Quan Z. Sheng, Lei Yang, Tao Gu, Peipei Xu, and Longfei Shangguan. 2016. AudioGest: enabling fine-grained hand gesture detection by decoding echo signal. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Heidelberg, Germany) *(UbiComp '16)*. Association for Computing Machinery, New York, NY, USA, 474–485. doi:10.1145/2971648.2971736

[36] Ricardo San Martín, Pablo Tello, Ana Valencia, and Asier Marzo. 2020. Experimental evaluation of distortion in amplitude modulation techniques for parametric loudspeakers. *Applied Sciences* 10, 6 (2020), 2070.

[37] Vivian Shen, Craig Shultz, and Chris Harrison. 2022. Mouth Haptics in VR using a Headset Ultrasound Phased Array. 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 275, 14 pages. doi:10.1145/3491102.3501960

[38] Zhouyang Shen, Zak Morgan, Madhan Kumar Vasudevan, Marianna Obrist, and Diego Martinez Plasencia. 2024. Controlled-STM: A Two-stage Model to Predict User's Perceived Intensity for Multi-point Spatiotemporal Modulation in Ultrasonic Mid-air Haptics. 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 709, 12 pages. doi:10.1145/3613904.3642439

[39] Zhouyang Shen, Madhan Kumar Vasudevan, Jan Kučera, Marianna Obrist, and Diego Martinez Plasencia. 2023. Multi-point STM: Effects of Drawing Speed and Number of Focal Points on Users' Responses using Ultrasonic Mid-Air Haptics. 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 83, 11 pages. doi:10.1145/3544548.3580641

[40] Hiroyuki Shinoda and Yasutoshi Makino. 2022. *Superimposing Visual Images on Mid-Air Ultrasonic Haptic Stimulation.* Springer International Publishing, Cham, 281–298. doi:10.1007/978-3-031-04043-6_12

[41] Youjin Sung, Rachel Kim, Kun Woo Song, Yitian Shao, and Sang Ho Yoon. 2024. HapticPilot: Authoring In-situ Hand Posture-Adaptive Vibrotactile Feedback for Virtual Reality. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 4, Article 179 (Jan. 2024), 28 pages. doi:10.1145/3631453

[42] Ryoko Takahashi, Keisuke Hasegawa, and Hiroyuki Shinoda. 2018. Lateral Modulation of Midair Ultrasound Focus for Intensified Vibrotactile Stimuli. In *Haptics: Science, Technology, and Applications*, Domenico Prattichizzo, Hiroyuki Shinoda, Hong Z. Tan, Emanuele Ruffaldi, and Antonio Frisoli (Eds.). Springer International Publishing, Cham, 276–288.

[43] Sam Van Damme, Nicolas Legrand, Joris Heyse, Femke De Backere, Filip De Turck, and Maria Torres Vega. 2022. Effects of Haptic Feedback on User Perception and Performance in Interactive Projected Augmented Reality. In *Proceedings of the 1st Workshop on Interactive EXtended Reality* (Lisboa, Portugal) *(IXR '22)*. Association for Computing Machinery, New York, NY, USA, 11–18. doi:10.1145/3552483.3556456

[44] Nicha Vanichvoranun, Hojeong Lee, Seoyeon Kim, and Sang Ho Yoon. 2024. EStatiG: Wearable Haptic Feedback with Multi-Phalanx Electrostatic Brake for Enhanced Object Perception in VR. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 3, Article 131 (Sept. 2024), 29 pages. doi:10.1145/3678567

[45] Steeven Villa, Sven Mayer, Jess Hartcher-O'Brien, Albrecht Schmidt, and Tonja-Katrin Machulla. 2022. Extended Mid-air Ultrasound Haptics for Virtual Reality. *Proc. ACM Hum.-Comput. Interact.* 6, ISS, Article 578 (Nov. 2022), 25 pages. doi:10.1145/3567731

[46] Anran Wang and Shyamnath Gollakota. 2019. MilliSonic: Pushing the Limits of Acoustic Motion Tracking. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. doi:10.1145/3290605.3300248

[47] Penghao Wang, Ruobing Jiang, Jingyang Hu, Yanmin Zhu, Hongbo Jiang, Minglu Li, and Chao Liu. 2024. AMT$^+$+: Acoustic Multi-Target Tracking With Smartphone MIMO System. *IEEE Transactions on Mobile Computing* 23, 12 (2024), 12650–12665. doi:10.1109/TMC.2024.3417474

[48] Shiyang Wang, Henglin Pu, Qiming Cao, Wenjun Jiang, Xingchen Wang, Tianci Liu, Zhengxin Jiang, Hongfei Xue, and Lu Su. 2025. *RAM-Hand: Robust Acoustic Multi-Hand Pose Reconstruction Using a Microphone Array.* Association for Computing Machinery, New York, NY, USA, 130–143. https://doi.org/10.1145/3715014.3722055

[49] 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

[50] 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

[51] Hyung Ii Yi, Hojeong Lee, and Sang Ho Yoon. 2024. ThermicVib: Enabling Dynamic Thermal Sensation with Multimodal Haptic Glove for Thermal-Responsive Interaction. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 614–623. doi:10.1109/ISMAR62088.2024.00076

[52] Sang Ho Yoon, Siyuan Ma, Woo Suk Lee, Shantanu Thakurdesai, Di Sun, Flávio P. Ribeiro, and James D. Holbery. 2019. HapSense: A Soft Haptic I/O Device with Uninterrupted Dual Functionalities of Force Sensing and Vibrotactile Actuation. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 949–961. doi:10.1145/3332165.3347888

[53] Jung-Hwan Youn, Seung Heon Lee, and Craig Shultz. 2025. HaptiCoil: Soft Programmable Buttons with Hydraulically Coupled Haptic Feedback and Sensing. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 155, 16 pages. doi:10.1145/3706598.3713175

[54] 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

[55] Sangki Yun, Yi-Chao Chen, Huihuang Zheng, Lili Qiu, and Wenguang Mao. 2017. Strata: Fine-Grained Acoustic-based Device-Free Tracking. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services* (Niagara Falls, New York, USA) *(MobiSys '17)*. Association for Computing Machinery, New York, NY, USA, 15–28. doi:10.1145/3081333.3081356

[56] Xiaoyan Zhou, Adam Sinclair Williams, and Francisco Raul Ortega. 2022. Eliciting Multimodal Gesture+Speech Interactions in a Multi-Object Augmented Reality Environment. In *Proceedings of the 28th ACM Symposium on Virtual Reality Software and Technology* (Tsukuba, Japan) *(VRST '22)*. Association for Computing Machinery, New York, NY, USA, Article 17, 10 pages. doi:10.1145/3562939.3565637

Table 3: Localization performance with different numbers of microphones, evaluated under the within-session protocol in the *Center Haptic* condition.

# of Microphones	MAE (mm)	MPJPE (mm)
1	21.89	45.35
2	19.49	37.81
4	14.73	30.62

Table 4: Per-joint hand localization performance (MAE and MPJPE, in mm) across three evaluation protocols under the *Center Haptic* condition.

	Within Session		Cross Session		Cross User	
	MAE	**MPJPE**	**MAE**	**MPJPE**	**MAE**	**MPJPE**
Wrist Joint	12.89 ± 0.96	26.45 ± 1.98	14.06 ± 1.28	28.93 ± 2.59	18.02 ± 1.40	36.70 ± 2.78
Thumb MCP	14.77 ± 1.30	28.27 ± 9.24	16.42 ± 1.84	31.00 ± 10.43	20.42 ± 1.66	38.32 ± 12.47
Index Finger MCP	15.31 ± 1.09	31.93 ± 2.41	17.00 ± 1.91	35.28 ± 4.09	20.95 ± 1.70	42.92 ± 3.34
Middle Finger MCP	14.84 ± 0.91	30.83 ± 1.98	16.31 ± 1.85	33.80 ± 4.05	20.64 ± 1.65	42.44 ± 3.24
Ring Finger MCP	15.10 ± 0.96	31.38 ± 2.10	16.34 ± 1.91	33.99 ± 4.32	20.84 ± 1.54	43.08 ± 3.02
Pinky Finger MCP	15.56 ± 0.97	32.31 ± 2.10	16.74 ± 2.10	34.95 ± 4.87	21.38 ± 1.55	44.48 ± 3.19

Mean ± Standard Deviation.

A Effect of the Number of Microphones

To evaluate the effect of the microphone configuration, we trained and evaluated the model using subsets of the four microphone channels. All results are obtained under the within-session protocol in the *Center Haptic* condition (Table 3).

B Per-Joint Hand Localization Performance

Table 4 reports the localization error for each joint under the *Center Haptic* condition across the three evaluation protocols.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009