# FaceOri: Tracking Head Position and Orientation Using Ultrasonic Ranging on Earphones


Yuntao Wang*
yuntaowang@tsinghua.edu.cn
Key Laboratory of Pervasive Computing, Ministry of Education, Department of Computer Science and Technology
Tsinghua University
Beijing, China

Jiexin Ding*
djx031906@163.com
Global Innovation Exchange (GIX) Institute
Tsinghua University
Beijing, China

Ishan Chatterjee
ishan.chatterjee1@gmail.com
Paul G. Allen School of Computer Science and Engineering
University of Washington
Seattle, WA, USA

Farshid Salemi Parizi
farshid@uw.edu
Electrical and Computer Engineering
University of Washington
Seattle, WA, USA

Yuzhou Zhuang
zhuangyz19@mails.tsinghua.edu.cn
Global Innovation Exchange (GIX) Institute
Tsinghua University
Beijing, China

Yukang Yan†
yyk@mail.tsinghua.edu.cn
Department of Computer Science and Technology
Tsinghua University
Beijing, China

Shwetak Patel
shwetak@cs.washington.edu
Paul G. Allen School of Computer Science and Engineering
University of Washington
Seattle, WA, USA

Yuanchun Shi
shiyc@tsinghua.edu.cn
Department of Computer Science and Technology
Tsinghua University
Beijing, China



## ABSTRACT

Face orientation can often indicate users' intended interaction target. In this paper, we propose FaceOri, a novel face tracking technique based on acoustic ranging using earphones. FaceOri can leverage the speaker on a commodity device to emit an ultrasonic chirp, which is picked up by the set of microphones on the user's earphone, and then processed to calculate the distance from each microphone to the device. These measurements are used to derive the user's face orientation and distance with respect to the device. We conduct a ground truth comparison and user study to evaluate FaceOri's performance. The results show that the system can determine whether the user orients to the device at a 93.5% accuracy within a 1.5 meters range. Furthermore, FaceOri can continuously track user's head orientation with a median absolute error of 10.9 mm in the distance, 3.7° in yaw, and 5.8° in pitch. FaceOri can allow for convenient hands-free control of devices and produce more intelligent context-aware interactions.



## CCS CONCEPTS

• **Human-centered computing** → **Interaction techniques**; **Sound-based input / output**.

## KEYWORDS

Acoustic ranging, head orientation, earphone, head pose estimation.

**ACM Reference Format:**
Yuntao Wang, Jiexin Ding, Ishan Chatterjee, Farshid Salemi Parizi, Yuzhou Zhuang, Yukang Yan, Shwetak Patel, and Yuanchun Shi. 2022. FaceOri: Tracking Head Position and Orientation Using Ultrasonic Ranging on Earphones. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3491102.3517698


## 1 INTRODUCTION

Earphones are one of the most ubiquitous wireless accessories. As a greater number of smartphones continue to drop the earphone jack, the popularity of these mobile audio devices continues to grow. With the earphones' cord getting cut, the input microphone has now migrated from a placement inline with the cable to a position at each of the user's ears. While most headsets leverage the microphone to take calls and, more recently, to enable the active noise cancellation (ANC) functionality, we find that this unique placement of these sensors can be used to unlock a broader range

*The authors contribute equally to this paper.
†denotes as the corresponding author.



*CHI '22, April 29-May 5, 2022, New Orleans, LA, USA*



ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3517698

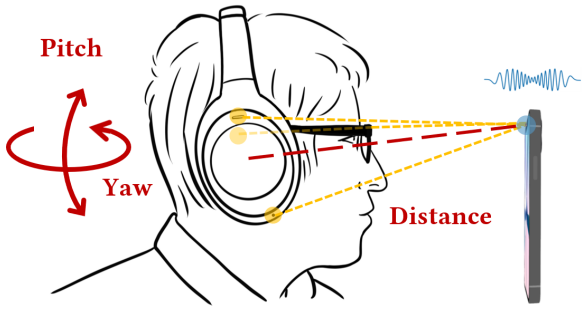


**Figure 1: FaceOri tracks user's face orientation towards the device with acoustic ranging using microphones in an earphone.**

of context-aware interactions when the earphone is transformed into a *spatial input device* via ultrasonic ranging.

In particular, users tend to orient their heads toward their intended interaction targets [49]. To recognize user interaction intention, researchers have leveraged eye-gaze or head tracking by using dedicated devices [12, 24] or by using an external camera [8, 10, 34]. These methods apply vision-based gaze or head tracking, which carry privacy concerns [22], and require the user to be in the field of view of the front-facing camera.

Enabling a device to detect this intention precisely and naturally can simplify user interface flow, enable hands-free interaction, and adapt interfaces to the context of use. For example, smartphones can detect users' proximity and face orientation toward the device to turn on the screen and allow them to read their notifications in a hand-free manner. Additionally, a system can accurately classify whether a user's head is oriented toward a specific device. This device-specific *binary attention detector* can be used to drive context-aware experiences. In this way, devices can adapt the layout and format of visual content based on whether the user is looking at them. Lastly, with *continuous tracking* of user proximity and face orientation, an additional set of applications including activity tracking and head gesture recognition can be realized.

We propose FaceOri, a novel interaction technique that leverages the built-in set of microphones found in almost all modern active noise canceling (ANC) earphones to infer the user's spatial location and head orientation with respect to a smartphone, laptop, smart speaker, or other devices with a built-in speaker. FaceOri works as follows: the computing device (e.g., a smartphone) emits an inaudible, ultrasonic sound from its speaker, and the embedded microphones on the earphone receive the sound. FaceOri calculates the time-of-arrival to estimate the distance from each microphone to the device even when the head occludes direct line-of-sight or introduces the Doppler effect. FaceOri uses these measurements to estimate 2-degree-of-freedom (DoF) face orientation — pitch and yaw — and 1-DoF distance measurement with respect to the device (Fig. 1). The user evaluation demonstrates 93.5% accuracy in *binarized attention detection* and dynamic *continuous tracking* performance to be 10.9 mm in the distance, 3.7° in yaw, and 5.8° in pitch, which significantly outperforms the baseline acoustic ranging method (CAT [32]: 42.0 mm, 11.0°, and 11.6°) on our collected dataset. These outputs enable many hands-free device interactions, including convenient wake-up of the devices, attentive user interfaces, and fitness tracking. To our best knowledge, we are the first to benchmark the head orientation tracking performance with acoustic ranging methods using build-in microphones in commodity ANC earphones. In this paper, we offer three main contributions:

(1) A spatial input technique that applies ultrasonic ranging to enable continuous head orientation and distance tracking with respect to a device with a speaker using a built-in set of microphones in the commodity ANC earphone.
(2) An end-to-end system characterization and user evaluation demonstrate FaceOri's high dynamic performance in continuous tracking and binarized attention detection.
(3) An exploration of the application space afforded by FaceOri with prototypes of selected demonstrative experiences, showcasing the applicability of the proposed approach.

## 2 RELATED WORK

FaceOri employs acoustic ranging to track the earphone's position relative to a device with a speaker (e.g., phone), enabling natural and precise face orientation based interactions. In this section, we first position this paper with respect to the attentive user interface literature. We then review the related works on acoustic ranging with a focus on mobile systems.

### 2.1 Attentive User Interfaces

Attentive user interfaces have been proposed as a natural user interface concept, sensitive to the user's focus of attention [46, 53]. Gaze pointing, as one of the important input modalities, has traditionally used dedicated camera-based eye tracking technology to identify which object a person is looking at [12, 13, 24, 55]. Researchers have explored gaze-aware solutions that enable users to start conversations with software agents [39, 47], select applications on computers [19, 43, 54], and control home appliances [31] by looking to the targets. However, these techniques require users to wear intrusive gaze trackers or environments to be instrumented with dedicated cameras, limiting these methods' ubiquity.

Face orientation can also be used as a proxy for the user's attention [4, 16, 23, 39, 49, 50, 61, 62]. Therefore, prior research has explored tracking users' face orientation to infer their focus toward targets within graphical user interfaces [29, 44], user authentication [27], smart home appliances [18, 20, 45], VR and AR targets [6, 11, 25, 62], wearable computing [7] and assistive interfaces [30]. In industry, several different smartphone applications have been released that incorporate face tracking via the front-facing camera for experiences like Animoji, Memoji, and face filters [3, 14, 48]. Recent works have adopted RGB [1, 8, 17] or depth camera [5, 34] to accurately track the user's head pose or gaze. These methods apply vision-based gaze or head tracking, which carry privacy concerns [22], and require the user to be in the field of view of the front-facing camera. As a result, they would be incompatible for devices without a camera, such as smartwatches [51] or smart speakers.

There have also been related works on face orientation detection that use microphone arrays distributed around the room to predict the direction of the user's voice [2, 36, 37, 63]. Although these

voice-based face orientation detection methods are wearable-free, they require users to speak to the targets. Instead, FaceOri can continuously track the user's face orientation and relative distance without the requirement of speaking. Thus, FaceOri can benefit a wider range of interaction scenarios (e.g., working environment). Further, FaceOri has the potential to achieve higher degrees of freedom and face orientation detection performance.

## 2.2 Background on Acoustic Ranging

Acoustic signals have been studied extensively for various tracking applications. Traditionally, acoustic tracking systems are based on the Doppler effect, which calculates the frequency shift to infer a moving object's speed, and thus distance [15, 21]. AAMouse [64] uses the frequency shifts of transmitted signals to enable accurate device tracking and achieves a median error of 1.4 cm. Another set of acoustic tracking systems performs auto-correlation to determine the travel time (and thus distance) between the speaker and the microphone [38, 40], achieving centimeter-level accuracy. Phase-based methods treat received signals as phase-modulated signals and analyze phase changes to obtain fine-grained distance information [57, 65], achieving a mean distance error of 1.3 cm in 3D space. EarphoneTrack [9] adopts the speaker in the earphone as the transmitter for acoustic ranging. It utilizes the leakage signal from the earphone's speaker to the microphone as a reference signal to calculate the distance from the earphone to the connected device.

Most similar to our work, acoustic ranging via Frequency Modulated Continuous Wave (FMCW) was proposed for high-precision distance estimation. CAT [32] proposed a distributed FMCW technique to accurately estimate the absolute distance between a transmitter and a receiver. It further combines IMU measurements to achieve mm-level tracking performance. Based on CAT, MilliSonic [56] utilizes the phase information in the demodulated FMCW signal to compute distances and further refine the tracking accuracy. The paper prototypes a 4-microphone array setup and achieves 2.6 mm median 3D tracking accuracy for smartphones. DroneTrack [33] applies MUltiple SIgnal Classification (MUSIC) for solving the multipath and strong noise issues, achieving 2-3 cm distance median error and 1°-3° orientation median error.

However, these acoustic ranging methods rely on a direct line-of-sight (LOS) and low moving speed between the speaker and the microphone, limiting the applicability of the approaches to our face orientation application. In our scenario, the head occludes the direct path from the microphone to the device speaker, resulting in a severe non-LOS issue. Further, the quick head movement introduces a significant Doppler effect. These issues significantly damage the tracking performance. Inspired by advanced techniques in the FMCW radar research field [28, 35, 58], we extended CAT [32] with optimization approaches, including adopting the triangular modulated chirp signal to reduce the Doppler effect [35], and applying an advanced filtering method to increase the signal-to-noise-ratio (SNR) [28, 41, 58]. To our best knowledge, we are the first to introduce acoustic ranging for head orientation to a practical usage setup – commodity ANC earphone and a smart device and benchmark its performance.

# 3 METHOD

FaceOri tracks head *distance* and head *orientation* in relation to a device to feed smarter device interactions, using the following two-step process. As Fig. 2A shows, a set of distance measurements from the device speaker to the earphone microphones are produced via FMCW acoustic ranging, requiring a low-effort calibration procedure (described below). Second, these distances are fed to a geometric model to continuously calculate the face orientation (both yaw and pitch) with respect to the speaker. Separately, to enable context-aware applications that only require information on whether the user's face orients to the device or not (*binarized attention detection*), we employ a binary classifier on a set of acoustic features (Fig. 2B). Notably, *binarized attention detection* is calibration-free. We describe the methods and algorithms below.

## 3.1 Acoustic Ranging Using FMCW

*3.1.1 CAT Acoustic Ranging.* This section provides a brief review of the fundamental aspects of CAT [32] for our method. The speaker emits a chirp signal whose frequency changes linearly with time, $f(t) = f_0 + \frac{B}{T}t$, where $B$ is the frequency bandwidth and $T$ is the sweep time. By integrating the frequency, we can express the transmitted signal in time domain as $y_t(t) = A_0 cos(2\pi f_0 t + \pi\frac{B}{T}t^2)$. After some time delay $t_d$, the microphone receives the signal as $y_r(t) = A_1 cos(2\pi f_0(t - t_d) \pm \pi\frac{B}{T}(t - t_d)^2)$. By mixing the received signal with the transmitted signal and applying a low pass filter, we obtain following signal:

$$y_m = \frac{A_0 A_1}{2} cos(2\pi\frac{B}{T}t_d t + 2\pi f_0 t_d - \pi\frac{B}{T}t_d^2) \quad (1)$$

The time delay $t_d$ can be calculated with the frequency and phase from the mixed signal $y_m$. FMCW-based ranging methods with shared transmitter and receiver clock can directly calculate the peak at $f_p^d = \frac{B}{T}t_d$ in the frequency domain. From the peak frequency $f_p^d$, one can estimate the delay time and thus the distance. However, with a common problem of distributed FMCW systems [32, 33, 56], FaceOri has a separate transmitter (device speaker) and receiver (earphone microphone) with unsynchronized clocks. Therefore, FaceOri requires a calibration procedure to establish a reference position with the peak frequency of $f_p^0$, and the detail is described in Sec.3.1.2. The distance between the speaker and microphone can be calculated with the following equation, where $c$ is the speed of sound.

$$D = c\frac{(f_p^d - f_p^0)T}{B} \quad (2)$$

*3.1.2 Calibration.* The calibration procedure is required for *head tracking* but not mandatory for *binarized attention detection*. Referring to MilliSonic [56], we require the user to place the left ANC microphone against the speaker with around a 2 mm gap for a couple of seconds (4 seconds are sufficient based on our evaluation in Sec. 5.3.4). Therefore, FaceOri can 1) establish a reference position with the peak frequency of $f_p^0$; 2) perform an approximation synchronization by correlating the received signals with the original one; 3) handle the continuous clock time drift between the transmitter and the receiver by applying the linear curve fitting solution [32, 56]. We acknowledge that this procedure limits

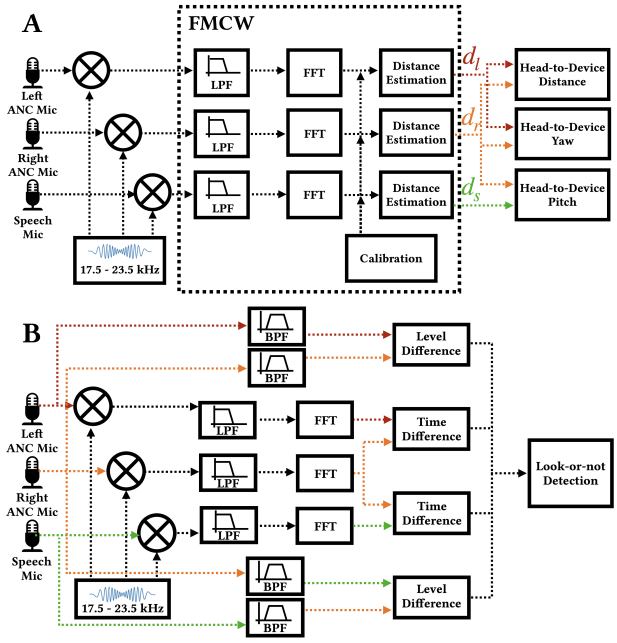


**Figure 2: FaceOri can enable continuous head position and orientation tracking (A) with FMCW-based acoustic ranging or binarized attention classification without calibration (B).**

the convenience of using FaceOri in the real-world application. Alternative calibration methods are discussed in Sec. 7.1.

*3.1.3 Optimizations for missing LOS and the Doppler effect.* Related acoustic ranging works [9, 32, 56] assume that there is a direct propagation path from the speaker to the microphone. However, since the microphones are located on the side of the head, minimal head shifts can cause the microphones to be occluded from the speaker. This loss of line-of-sight (LOS) significantly degrades the signal-to-noise ratio (SNR). It distorts the peaks ($f_p^d$) in the frequency domain, resulting in multiple peaks or the direct path peak merging with anomalous nearby peaks. Further, the quick head motion can introduce a significant Doppler effect. These issues can significantly degrade the tracking performance. Therefore, we developed and applied optimizations to the existing method to better support our application scenarios. Inspired by advanced techniques in the FMCW radar research field [28, 35, 58], we extended CAT [32] with optimization approaches to solve the non-LOS and Doppler effect issues. FaceOri adopts an inaudible triangular modulated chirp signal to reduce the Doppler effect. Our implementation adopted an up-chirp from 17.5 kHz ($f_0$) to 23.5 kHz ($f_1$) followed by a down-chirp to 17.5 kHz with a total sweep time of 42.7 ms (2048/48000). FaceOri further averages the two parts of measurements at different edges of the triangular pattern [41]. Therefore, FaceOri can achieve a more accurate distance estimation despite the frequency shift caused by the Doppler frequency, as prior works [35, 66] explained. To further increase the SNR, FaceOri adopts a non-coherent integration method [28] by averaging the intermediate FFTs from a small set of recent frames (2 frames in our implementation).

To get the correct peak corresponding to the direct path, we used the Constant False Alarm Rate (CFAR) adaptive thresholding algorithm [41, 59] on the FFT values of the mixed signal — $y_m$. The algorithm combines the following heuristics: 1) an early and high peak that is closest to the previous peak is selected corresponding to the direct peak due to the continuous change in the distance; 2) when a sudden peak shift appear in one microphone channel but not the others, indicating a loss-track event, a fallback algorithm is utilized to predict the peak frequency from recent historical frames (5 frames in our implementation) through interpolation.

## 3.2 Yaw and Pitch Estimation

The FMCW-based acoustic ranging technique provides three distances between the speaker and the earphones' three microphones. Two microphones used for active noise cancellation (ANC) sit at a similar elevation at the top of the earcup (see Fig. 4). A single speech microphone sits at a lower elevation on the right earcup. By comparing distances between the left and right ANC microphones from the speaker, yaw can be calculated. By comparing distances between the right ANC microphone and the speech microphone, the pitch can be calculated.

The yaw and pitch angles are calculated as angles between the face orientation vector and the vector from the center of the head to the speaker location. The mic-speaker distances form triangles in the transverse (top view, see Fig. 3A) and sagittal (side view, see Fig. 3B) plane of the head. In each, the triangle's altitude is aligned with the face orientation vector, and the triangle's median is aligned with the vector between the head center and speaker. We refer to the distance from the speaker to the left ANC microphone as $d_l$, the right ANC microphone as $d_r$, and the right speech microphone as $d_s$. The distance between the left and right ANC microphones is $d_e$, which can be measured manually or set by an average value across users. The distance between the right speech microphone and the right ANC microphone is $d_b$, a known quantity. To explain our method, we will detail how the yaw angle is calculated. A similar approach is employed for pitch estimation. The length of the median line ($d_m$) is calculated as:

$$d_m = \frac{\sqrt{2{d_l}^2 + 2{d_r}^2 - {d_e}^2}}{2} \tag{3}$$

and the angle between the median line and the base line of the triangle ($\alpha$) defines yaw ($\varphi$) as follows:

$$\alpha = \arccos\left(\frac{{d_m}^2 + \frac{1}{2}{d_e}^2 - {d_r}^2}{d_m d_e}\right) \tag{4}$$

$$\varphi = \alpha - 90^\circ \tag{5}$$

The same approach can be used to calculate pitch ($\theta$), by replacing the triangle formed by $d_r$,$d_s$, and $d_b$ with the one formed by $d_l$,$d_r$, and $d_e$. Therefore, we can obtain both the yaw ($\varphi$) and pitch ($\theta$) angles of the user's face towards the speaker. This is achieved by subtracting the current angle with the known initial angle ($\theta_0$), since the speech microphone is slightly skewed off in vertical from the right ANC microphone in the earphone design (Sec. 4).

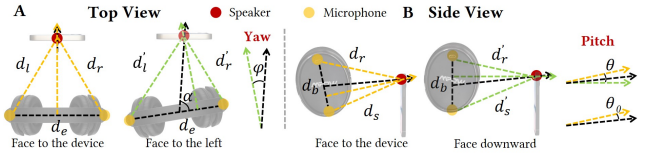


**Figure 3: FaceOri estimates face orientation towards the sound source by calculating the angle between the median line and the altitude line.**

## 3.3 Binarized Attention Detection

FaceOri requires a calibration procedure for continuous face tracking. However, we present an alternative binarized attention classification method that detects whether the user is looking at the device without any calibration. This binarized detection can still be useful in various scenarios, including attentive user interfaces [31, 39, 47]. As Fig.2B shows, we first applied a bandpass filter with a frequency range of 17.5 kHz to 23.5 kHz to the audio signals from the three microphones. We evenly divided the frequency range into 20 bands. In each band, we can obtain the level difference (LD), which is the amplitude level ratio of audio signals from two microphones. We obtained $20 \times 3 = 60$ LD features among three microphones. We also adopted 3 time difference features between the microphones. Each time difference feature is the frequency gap between the peaks ($f_p^d$) in the frequency domain of the mixed signals from two microphones. Using the 63 features (Fig. 2B), FaceOri can detect whether the user is looking at the device by training a binary classifier. In our implementation, we adopted the supported vector machine (SVM, RBF kernel, C = 1.0) as the binary classifier.

# 4 IMPLEMENTATION

## 4.1 FaceOri Hardware

*4.1.1 Headphone Prototype.* Modern ANC earphones share a similar design in the microphone placement [42] as Fig. 4 shows. Two microphones are located at the top of the earcups for collecting environmental noises. A speech microphone or microphone array sits at a lower elevation on one earcup. We adopted MPOW H19 [1] for evaluation. Further, we demonstrated FaceOri's applications using Hush earphone by 233621 [2], Sony WF-1000XM3 [3], and ANC earbud — Sony WH-1000XM3 [4] without an extra speech microphone. To obtain the high-resolution raw acoustic stream, we wired out two feed-forward ANC microphones and the speech microphone with 3.5mm TRS plugins. The plugins were connected to a Zoom H6 [5] audio interface via VXLR to a 3.5 mm audio adapter. Zoom H6 supports up to 6 synchronized channels of real-time audio streaming through USB. Therefore, the audio signals from the three microphones on the earphone were streamed by the Zoom H6 to a Thinkpad X1 carbon laptop (CPU: i7-10710U, 6 cores, 1.1 GHz, RAM: 16GB, Storage: 512GB), which ran the audio signal processing algorithms in real-time. The sampling rate and the bit depth were set to 48 kHz and 16 bits. To further compare FaceOri's performance to the inertial measurement unit (IMU) based solution, we adopted the MPU-9250 [6] IMU module, which has a 3-axis accelerometer, a 3-axis gyroscope, and a 3-axis magnetometer. The data was streamed to the laptop with an Arduino Uno using the I2C protocol. The laptop read the IMU data with the same sampling rate — 23.4 frames per second (fps). To avoid the effect of the speaker magnet, we mounted the IMU module to the top of the earphone. Before each measurement, we calibrated the magnetometer inside the IMU by drawing the ∞ shape in the air.

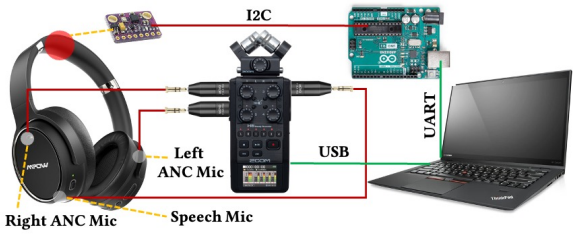


**Figure 4: FaceOri's earphone hardware has a commodity earphone hardware (MPOW H19 for demonstration), an MPU-9250 IMU, an audio interface, and a laptop to process the audio signal.**

*4.1.2 Audio Transmitter.* A common device with a speaker capable of generating inaudible ultrasonic sound (e.g., above 17 kHz) can be adopted as an audio transmitter. During our evaluation, a Samsung Galaxy S21 Ultra smartphone (256GB storage, 12GB RAM) with stereo speakers was adopted as the audio transmitter. Further, we demonstrated FaceOri's applications using Thinkpad X1 Carbon laptop ( Intel i7-10710U CPU, 16G RAM, 512G storage), Mi Watch (8GB storage, 1GB RAM), and Huawei Matepad PRO (10.8 inches, 256 GB storage). We generated a one-hour mono-channel audio file with continuous triangular chirp signals modulated signals (see Sec. 3.1.3). Then, the transmitter played the audio file from only one speaker using the HibyMusic [7] application, which supports the sampling rate and bit depth at 48 kHz and 16 bits, respectively.

## 4.2 FaceOri Software

We implemented FaceOri (see Sec. 3.1) using Python on the Thinkpad X1 Carbon laptop. As Fig. 5 shows, we used PyAudio [8] to read the triple-channel raw audio signal from the Zoom H6 audio interface. All the raw audio data was stored for further offline analysis. The calibration module was launched when the user clicked the calibration button on the user interface. FaceOri requires two parameters to be calibrated that are (1) the distance between the two ANC microphones ($d_e$) when the user wears the earphone (see Sec. 3.2), and (2) the reference origin for precise acoustic ranging (see Sec. 3.1.2). Then we pressed the *Calibrate* button on the launch user interface and kept the two devices still for a 10-seconds duration. Notably,

[1] https://www.xmpow.com/products/mpow-h19-hybrid-noise-cancelling-headphones
[2] https://www.233621.com/
[3] https://www.sony.com.sg/electronics/truly-wireless/wf-1000xm3
[4] https://www.sony.com.sg/electronics/headband-headphones/wh-1000xm3
[5] https://zoomcorp.com/en/us/handheld-recorders/handheld-recorders/h6-audio-recorder/
[6] https://invensense.tdk.com/products/motion-tracking/9-axis/mpu-9250/
[7] https://store.hiby.com/
[8] https://pypi.org/project/PyAudio/

10 seconds are redundant for later evaluation (Sec. 5.3.4). Once the "Start" button is pressed, FaceOri software displays the distance, yaw, and pitch values onto the launch user interface in real-time. Further, another interface popped up showing the measured distance curves with the three channels of the audio signal as Fig. 10C shows.

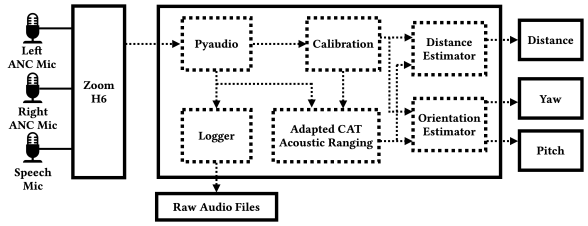


**Figure 5: FaceOri software's key components.**

## 5 EVALUATION

In this section, we describe the in-lab user evaluation study to benchmark FaceOri's performance of distance, orientation tracking, and binarized attention detection when the user is at different positions in relation to the device.

### 5.1 Participant and Apparatus

We recruited 12 participants (7 females, 5 males) with an average age of 22.5 (SD = 3.4). Each had previously used earphones and smartphones. The experiment was conducted in a room with the size of 4 by 3 meters. To obtain the ground truth positions, we utilized the OptiTrack motion capture system (10 Prime 17 cameras) with its coordinate calibrated. The tracking markers were located at the phone's front speaker, which was located next to the front camera, and each microphone location on the earphone. In this evaluation study, we used a Samsung Galaxy S21 Ultra smartphone as our sound transmitter and MPOW H19 earphone as our receiver. A tripod with adjustable height was used to hold the phone at different heights.

### 5.2 Experiment Design and Procedure

Each participant was informed about the purpose and the procedure of the experiment. An experimenter assisted each participant in putting on the earphone and then measured the approximate distance between the left and right ANC microphones with a ruler. The experimenter conducted the calibration by placing the smartphone's speaker to the left ANC microphone of the earphone for 10 seconds. To understand the effect of relative position on FaceOri's distance and orientation tracking accuracy, we created a 3D grid of test positions in front of the smartphone. With the smartphone located at the origin (0,0) in the top view, three rows of grids were located at 50 cm, 100 cm, and 150 cm away from (0,0) in the $y$ direction. The three columns of grids were located at -50 cm, 0 cm, and 50 cm away from (0,0) in $x$. We chose the maximum tracking distance to be around 160 cm — (150, -50) or (150, 50) — because we targeted FaceOri's usage scenarios within the range of a personal workspace. We tested three speaker heights: 80 cm, 120 cm, and 160 cm away from the floor. The participant adjusted the seated chair's height to a comfortable position. Therefore, the relative heights of the smartphone with respect to the earphone are different across the participants. During the user study, the lab noise levels ranged between 54.3 dBA to 62.7 dBA with a server running and people talking.

Each participant finished three head movement sessions at each 3D grid point. Each head movement session consisted of 6 subtasks: 1) look at the smartphone's speaker for 5 seconds, called the neutral state; 2) move forward and backward for 3 times; 3) rotate the head in the yaw direction for 3 times to the maximum range and return to the neutral state; 4) tilt the head in the pitch direction for 3 times to the maximum range and then return to the neutral state; 5) draw the zigzag shape from top left to the bottom right with 2 folds; and 6) randomly move the head for 3 seconds. The order of the 2D grids was randomized under each height condition. We re-calibrated FaceOri when we collected the data at a different height. Therefore, we conducted three calibrations in total. Each participant received a 20 USD gift card for their effort and time (40 minutes).

### 5.3 Results

Same as CAT [32] and MilliSonic [56], the deviation of FaceOri's measurements (distance, yaw, and pitch) follow non-Gaussian distributions, the median absolute error (MedAE) is a better measure compared to the mean absolute error (MAE). Therefore, we report FaceOri's tracking performance using the MedAE and the interquartile range (IQR). Nonetheless, we also derive MAE in the

**Table 1: The tracking performance within 9 grids. Each value group A/B/C indicates performance when the smartphone is placed on the height of 80/120/160 cm. Distance (mm), Yaw (°), Pitch (°).**

| Y/X | | -50 cm | | 0 cm | | 50 cm | |
|---|---|---|---|---|---|---|---|
| | | MedAE | IQR | MedAE | IQR | MedAE | IQR |
| **50 cm** | Distance | 12.9/9.5/5.4 | 19.2/14.7/7.8 | 7.9/4.1/3.4 | 13.9/8.1/6.3 | 10.1/6.6/4.6 | 27.3/13.2/8.5 |
| | Yaw | 4.7/4.9/2.0 | 7.8/8.2/2.6 | 3.7/1.7/1.5 | 5.9/3.2/2.6 | 5.7/3.9/2.0 | 11.9/8.0/3.5 |
| | Pitch | 6.2/4.2/3.1 | 10.0/5.4/3.8 | 8.9/3.4/2.8 | 15.9/6.0/4.5 | 9.7/5.7/4.1 | 18.8/8.8/7.3 |
| **100 cm** | Distance | 16.5/11.0/7.4 | 21.1/12.7/10.7 | 14.6/12.4/9.2 | 22.8/14.6/12.5 | 13.4/10.4/10.8 | 23.0/16.1/20.7 |
| | Yaw | 3.7/3.0/2.3 | 6.4/4.0/3.6 | 5.8/4.1/3.1 | 9.5/5.7/3.8 | 4.8/4.9/3.9 | 11.7/7.0/8.5 |
| | Pitch | 6.3/4.5/3.6 | 9.5/6.0/5.4 | 6.7/6.6/4.9 | 12.5/9.1/6.7 | 8.4/6.5/7.4 | 13.5/9.8/13.9 |
| **150 cm** | Distance | 24.5/17.4/14.6 | 29.8/19.6/25.0 | 19.4/11.2/10.6 | 55.1/17.7/22.3 | 19.6/15.5/13.1 | 41.5/29.5/23.8 |
| | Yaw | 4.7/3.7/4.0 | 10.8/6.4/6.7 | 6.6/5.0/3.9 | 14.9/8.9/6.9 | 5.5/6.6/4.5 | 12.0/13.4/11.5 |
| | Pitch | 9.3/7.3/6.7 | 15.0/11.9/10.7 | 9.8/8.3/6.3 | 15.1/12.2/10.3 | 8.3/7.5/8.2 | 13.6/11.9/14.4 |

discussion to compare against camera-based methods [1, 52] in literature. The ground truth distance and orientation of a user's face towards the device were calculated as described in Sec. 3.2 using coordinates of the tracker attached to the microphones and the speaker in the OptiTrack system. We utilized Aligned Rank Transform Factorial ANOVA for within-subject non-parametric statistical analysis ($p < 0.05$) with Wilcoxon signed-rank test for post-hoc analysis ($p < 0.05$).

We summarize results in Table 1. The table shows the MedAE and IQR of distance in millimeters, yaw in degrees, and pitch in degrees when the smartphone's speaker was placed at the height of 80, 120, and 160 cm. Each cell of the $3 \times 3$ cell represents one grid during the experiment (see Sec. 5.2).

*5.3.1 Distance Tracking Accuracy.* Results show that FaceOri can continuously track the distance from the user's head to the smartphone with a MedAE of 10.9 mm and an IQR of 18.8 mm. Statistical analysis shows that there are significant effects of 2D location (grid) ($F_{(8,253100)} = 1260, p < 0.001$) and height ($F_{(2,253106)} = 3115, p < 0.001$) on the distance tracking performance. Fig. 6 and post-hoc pairwise tests show that FaceOri can achieve a better distance tracking performance when the user is closer to the smartphone ($p < 0.01$) and in the center column of grids ($p < 0.01$). Results show that FaceOri can achieve the best performance when the smartphone was placed at height of 160 cm ($p < 0.001$).

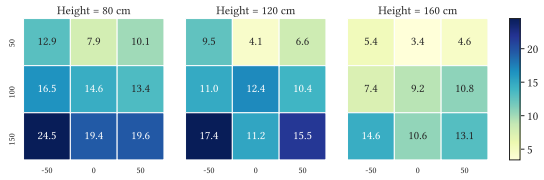


**Figure 6: The median absolute error (mm) of the distance at different relative positions (cm) between the smartphone and the earphone.**

*5.3.2 Head Orientation Tracking Accuracy.* Results show that FaceOri can continuously track the user's head orientation with respect to the smartphone. The MedAE and IQR of the yaw angle were 3.7° and 6.8° and those of the pitch angle were 5.8° and 10.0°. Results also show that the yaw range among all participants was from -81.1° (s.d. = 10.1) to 76.9° (s.d. = 9.1), the pitch range was from -87.8° (s.d. = 6.3) to 84.3° (s.d. = 9.9). The maximum, MedAE, and IQR of head orientation speed of the yaw angle were 250.9 °/s, 185.6 °/s, and 33.2 °/s and those of the pitch angle were 185.6 °/s, 13.6 °/s, and 44.1 °/s. These results can be helpful references for experience development.

We further evaluated the effect of the relative position of the user's head with respect to the smartphone on the orientation tracking performance. Statistical analysis shows that there are significant effects of 2D location (grid) on the yaw ($F_{(8,234813)} = 745, p < 0.001$) and the pitch ($F_{(8,226907)} = 730, p < 0.001$) tracking performance. Fig. 7 and post-hoc pairwise tests show that FaceOri can achieve a better tracking performance when the user is closer to the smartphone and in the center column of grids ($p < 0.01$) in general. Further, there are significant effects of height on the yaw ($F_{(2,234819)} = 1056, p < 0.001$) and the pitch tracking performance ($F_{(2,226913)} = 390, p < 0.001$). Again, FaceOri can achieve a better performance when the smartphone was placed at height of 160 cm ($p < 0.05$) as compared to the other heights. Further, we observed a significant effect of the relative height ($p < .01$) of the earphone to the smartphone's speaker but not the absolute sitting height of the participant ($p = 0.07$) on the tracking performance.

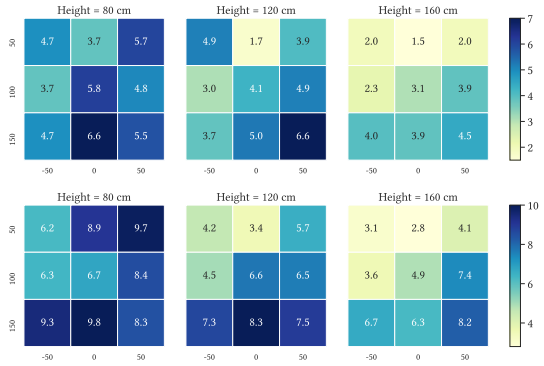


**Figure 7: The median absolute errors of the yaw (left) and pitch (right) angles with different relative position (cm) between the smartphone and the earphone.**

*5.3.3 Binarized Attention Classification Accuracy.* Using the trackers on the smartphone, we labeled the smartphone's margin as a rectangle in the OptiTrack coordinate space. We regarded the ground truth of orienting to the device as the vector of the user's head intersecting with this rectangle. When evaluated on the performance through *leave-one-out cross-user validation* for the look-or-not binary classification, FaceOri can achieve an average classification accuracy of 93.5% across users ( s.d.= 2.5%). The calculation latency is within 42 ms.

*5.3.4 Effect of Setup Configuration.* During the user study, the experimenter measured the distance between two ANC microphones ($d_e$ in Fig. 3) manually. Results show that the mean absolute error of the measurement with a ruler is 3.4 mm against OptiTrack. But choosing which one from these two measuring methods has no significant effect (p = 0.33) on the yaw tracking performance. Further, when setting $d_e$ to an fixed distance of 235 mm (considering human's head breadth 155 mm [9] + extra earcup depth 40 mm × 2), FaceOri was still able to achieve a satisfying performance that the MedAE in yaw direction increases by only 1% without a significant difference ($p = 0.1$). Therefore, there is no evidence that it is necessary to manually measure the distance between two ANC microphones.

We applied a redundant clock-sync calibration duration — 10 seconds. Since we recorded all the data from the study, we reran our method with a 4-second calibration duration, resulting in only a 3% increment in the median absolute error across all angles and distances, adequate for a whole experiment session of 15 minutes.

[9] https://en.wikipedia.org/wiki/Human_head

Our evaluation environment had a background noise from 54.3 dBA to 62.7 dBA with a server running and people talking, indicating the robustness of ultrasonic ranging against noise, which aligns with the result from MilliSonic [56].

*5.3.5 Comparison with CAT Baseline Method.* We evaluated the effectiveness of our optimization method mentioned in Sec. 3.1.3 to overcome the issues introduced by the head motion. We compared our method to the baseline CAT acoustic ranging method [32]. Results show that our method (10.9 mm in the distance, 3.7° in yaw, and 5.8° in pitch) significantly outperforms CAT (42.0 mm, 11.0° and 11.6°) on our collected dataset ($p < 0.001$ for all cases). Here, we define a dropped frame with the feature of a large distance offset away from the ground truth (37.6 mm as our threshold - $2.0 \times IQR$). Fig. 8 shows the dropped frame rate along with the ground-truth yaw (left figure) and the pitch (right figure) angle. Results show that our method can effectively decrease the dropped frame rate with an average rate of 14.8% versus 52.8% (CAT method). Head rotation of a larger amplitude in yaw or pitch angles results in more dropped frames. This demonstrates that our method is more robust against the non-LOS and Doppler effect introduced by the head motion. Further, the large spread of the head orientation angle and speed also indicate FaceOri's robustness against noises introduced by the head motions.

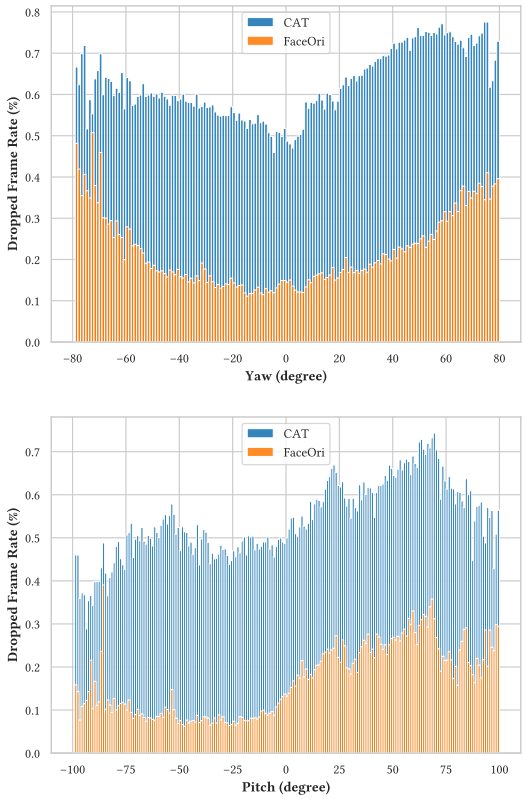


**Figure 8: FaceOri has less dropped frames compared with CAT acoustic tracking method [32].**

*5.3.6 Comparison with IMU-Based Head Orientation Tracking Solutions.* To compare the performance of FaceOri and the IMU-based solution (MPU-9250 with Arduino code [10]), we performed calibration to the IMU by aligning its initial yaw and pitch angle with the OptiTrack coordinate system. Results show that IMU-based solution can continuously track the user's head orientation with a MedAE of 17.2° (IQR = 330.4) in yaw, and 4.9° (IQR = 10.8) in pitch. The pitch tracking performance is significantly better than FaceOri ($p < 0.001$). However, we observed significant yaw drift, a well-studied problem in IMU tracking [26], even with the reference calibration provided by the magnetometer. Therefore, the IMU-based solution is insufficient for our applications, which require more accurate yaw estimation.

There are additional functionality limitations regarding to the IMU-based solution. The IMU tracks it orientation relative to the inertial world reference frame rather than the mobile device reference frame as FaceOri does. Modern earphones with IMUs do not contain the magnetometer due to the speaker's strong magnet. Further, IMU cannot provide accurate absolute distance to the mobile device (>20 cm error with calibration [60]).

*5.3.7 Comparison with Camera-Based Head Orientation Tracking Solutions.* To compare with camera-based solutions in the literature, we measure FaceOri's performance with a mean absolute error of 8.3° in yaw and 9.6° in pitch. This is comparable with cutting-edge RGB camera-based technique, which can track head orientation with MAE of 7.6° in both yaw and pitch [1]. Further, RGBD-based methods (ARKit) achieved higher performance — 1.8 mm in the distance, 0.9° in yaw, and 0.7° in pitch [52]. However, FaceOri has advantages in a wider field of view, preserving visual privacy, and has the potential to support the interaction with devices without cameras (e.g., smartwatch). FaceOri can be complementary to the vision-based method regarding usage scenario, power consumption, privacy, etc.

*5.3.8 Sensor Fusion of FaceOri and IMU.* To further reduce the power consumption during real-world deployment, we can adopt a sensor fusion method by combining FaceOri and the IMU if available. FaceOri can calibrate the IMU every certain amount of time $T_{cal}$. Therefore, we can track the yaw and pitch angles with higher accuracy than the IMU-based solution but consume less power than the sole ultrasonic-ranging-based solution. Since the IMU already achieved a better tracking performance in pitch than FaceOri, we evaluated the sensor fusion method on the yaw angle. We ran the FaceOri for 0.5 seconds to establish an accurate yaw angle and distance. Then we tested the effect of $T_{cal}$ (in second) on the tracking performance in yaw. Results show that the sensor fusion method can achieve a MedAE of 5.1° (IQR = 14.2°), 7.0° (IQR = 20.0°), and 9.7° (IQR = 21.5°) when $T_{cal} = 1, 3, 5$ seconds.

## 6 APPLICATIONS

FaceOri provides three outputs as a user interacts with a particular computing device: *distance measurement*, *binarized attention detection*, and *continuous orientation tracking*. These metrics can be used individually or in conjunction to enable and enhance applications.

[10] https://github.com/hideakitai/MPU9250

## 6.1 Activity and Gesture Sensing

The head distance and orientation measurements can be used to analyze user activity or capture explicit user input. We prototyped an exercise smartwatch application that can count the number of exercise repetitions that a user performs by analyzing the periodicity of the *distance measurement* provided by FaceOri (Fig. 9). We used the Sony WH-1000XM3 earbud as the receiver and the Mi Watch as the transmitter in this application.

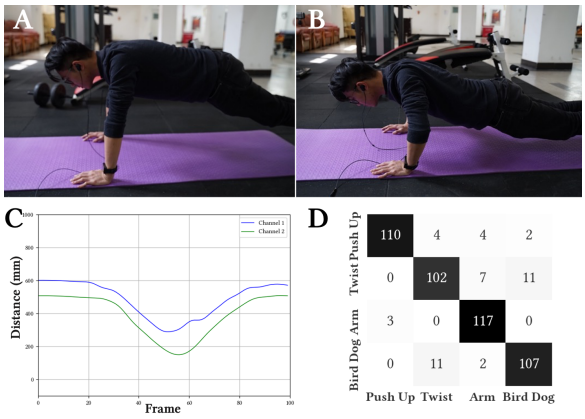


**Figure 9: FaceOri can track and count the exercise activities with a smartwatch as the audio transmitter and an ANC earbud as the receiver. (A) (B) User does push-ups. (C) The distances between two ears and the speaker when user does a push-up. (D) Confusion matrix of the four classifications (push-up, body twist, touching shoulder, and bird dog).**

To evaluate the feasibility of FaceOri in activity recognition, we conducted a recognition evaluation for this application. Participants were asked to perform four activities: push-up, body twist, touching shoulder with contralateral hand (arm for abbreviation), and bird dog for 3 rounds with 5 repetitions per round. We manually segmented the dual-channel raw audio signal and aligned data to the length of 160 by interpolation. We chose 7 features for each frame: the distance between two microphones and smartwatch, the first derivative of two distances, the level difference between two microphones, and whether the signals of two channels lose track. Using SVM (RBF kernel, C = 1.0) to classify each exercise activity, FaceOri can achieve an average accuracy of 90.9% using leave-one-out cross-user validation.

The *continuous orientation tracking* metric could also be used to enable gesture input. By analyzing oscillations in pitch and yaw, "yes" and "no" head shake gestures can be recognized. Finally, *continuous orientation tracking* could drive a selector or pointer for accessible interfaces where a user may lack muscle control below the neck.

## 6.2 Context-aware and Attentive User Interfaces

Real-time data on user position can be used to drive smarter, more context-aware interfaces. *Distance measurement* can be used to lock a phone or laptop or dim the screen when the user moves beyond a certain distance threshold away from the device (Fig. 10). We implemented this example on a Thinkpad X1 Carbon laptop. The distance threshold was set to 1.5 meters.

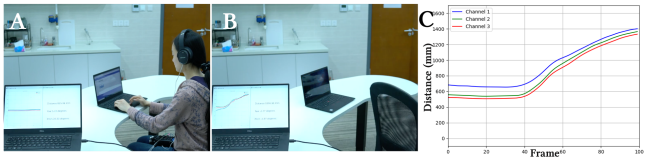


**Figure 10: FaceOri dims the device when the user moves beyond a certain distance threshold away from the device. (A)(B) The User walks away from the laptop and the screen is dimmed. (C) FaceOri measures distances from the laptop speaker to the three microphones in the earphone.**

*Binarized attention detection* can help ease switching between multiple tasks or points of interest. For example, as a user follows a recipe video on their laptop, the video can automatically pause as the user turns to the stove or cutting board and resume when they return their attention to the screen. In Fig. 11, a user can provide input to their smartphone device even if they are otherwise preoccupied and unable to easily perform touch input.

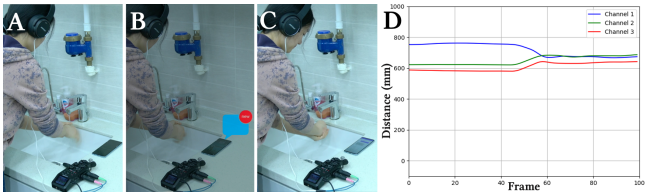


**Figure 11: FaceOri can light up the smartphone and open to the message application when the user is unable to easily perform touch input. (A) The user is washing her hands. (B) When a message comes, (C) the user turns her head towards the phone to wake it up, and then the detailed message is displayed. (D) FaceOri measures distances from the smartphone speaker to the three microphones in the earphone.**

## 6.3 Attentive Detection from Multiple Devices

FaceOri can be used to detect when users direct their intention toward a specific device. By orienting toward a smart speaker, a user can issue a command without requiring a keyword. Finally, we prototype a demonstrative application that applies FaceOri's *binarized attention detection* in a simple multi-device scenario. As a user looks between two laptops, the keyboard and the mouse pair automatically to the laptop that the user watches (Fig. 12). To implement this application, the devices share a central server and time-multiplex their transmitted chirps.

We implemented the multi-device application on two Thinkpad X1 Carbon laptops. We used a third laptop as the proxy to 1) transfer the mouse and keyboard inputs to the two Thinkpad laptops; 2) run the FaceOri algorithm to recognize which device the user orients to. These three laptops were connected through WiFi. The proxy coordinated the two Thinkpad laptops to let them emit the

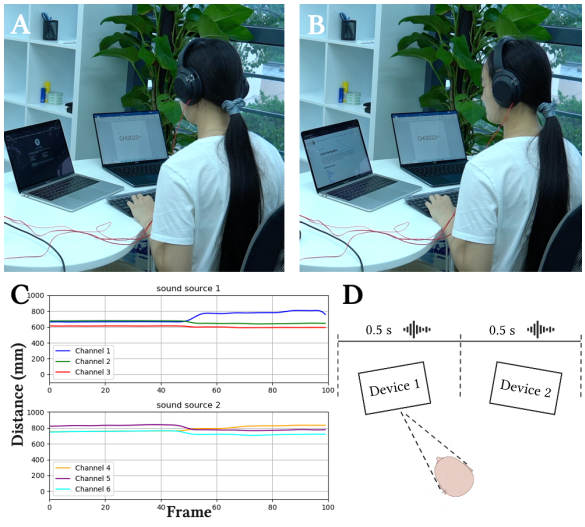


**Figure 12: The mouse and keyboard can pair to the device that the user faces towards automatically. (A)(B) The user switched her attention between two laptops. (C) FaceOri measures distances from each laptop speaker to the three microphones in the earphone. (D) The time-multiplex approach for the multi-device application.**

ultrasonic chirp signal alternately, 0.5 seconds for each. Therefore, the proxy knew which device was emitting the sound and then performed the *binarized attention detection* to detect which device the user orients to. The first two seconds are skipped due to the delay of speakers and echoes. The time-multiplex approach caused a delay during recognition, but there was no evidence that it would influence the tracking performance.

# 7 DISCUSSION AND FUTURE WORK

This paper proposes FaceOri, a novel end-to-end head position and orientation tracking system based on acoustic ranging using existing microphones in commodity earphones. Due to its high tracking performance, FaceOri can support a wide range of novel interaction applications. In this section, we discuss the findings, limitations, and avenues for future work.

## 7.1 Alternative Calibration Methods

The biggest limitation of FaceOri is the requirement of calibration (Sec. 3.1.2). During our evaluation, to enable accurate continuous acoustic ranging, FaceOri synchronizes the transmitter and receivers by holding one of the microphones to the speaker every session. However, existing work found that the calibration is only required once in each battery circle [9]. Therefore, the per-session calibration is not necessary. We would expect future work to validate a per-battery-circle calibration method.

To make the calibration procedure more user friendly, future work could explore using the front-facing RGB or RGBD camera on a device (if available) to establish the reference point and synchronize the clocks of the device and the wireless earphone. Further, calibration can be completely side-stepped if the earphone is connected to the transmitter via Bluetooth 5.0 or other wireless channels method with time synchronization protocol; thus, a sufficiently synchronized clock can be established. Future work can explore combining ultrasonic ranging with synchronization provided over these channels, avoiding calibration and additional complexity for drift compensation. A heuristic can be applied for a quick calibration procedure for applications requiring absolute distance but not requiring high accuracy. For example, the user can be instructed to hold their phone out at arm's length, and the origin can be set by substituting the average human arm length.

Notably, binarized attention detection requires no calibration and can be useful in various applications. Further, applications that use only a relative distance (such as the exercise application in Fig. 9) do not necessarily require calibration.

## 7.2 Deployment and Generalizability

We developed and evaluated FaceOri using the existing hardware in commodity earphones and mobilephones. To clarify, our test hardware is a proof-of-concept to evaluate our end-to-end face orientation and distance tracking system. In our implementation, we wired the built-in ANC microphones to an off-board laptop to host the signal processing program. However, modern ANC earphones have on-device processors. For instance, Sony 1000XM3 has a CSR8675 chip with DSP (120 MHz, 48kHz audio sampling rate) and MCU (80 MHz). We believe the proposed algorithms can be deployed on the micro-controller in the future for real-world applications.

We evaluated FaceOri's performance with only one type of earphone and one type of smartphone. However, we observed that ANC earphones adopt a common design in microphone placement as the tested one. Further, many newer models of earphones possess an extensive array of distributed microphones (e.g., Apple AirPods Max and the Bose 700). These characteristics can further improve the performance and increase the degrees of freedom. We expect future work to investigate FaceOri's generalizability across earphone models.

To further improve robustness and performance for real-world deployment, we would expect future work to further evaluate the effect of ambient noise on FaceOri's performance in various mobile scenarios. Meanwhile, future work can explore a more comprehensive sensor fusion method using the absolute (in device-frame) orientation provided by FaceOri with the relative (with respect to an inertial reference frame) information supplied by the IMU.

## 7.3 Supporting Multiple Devices

In section 6.2, we briefly demonstrated a possible time-multiplexed solution to support multiple audio transmission devices, allowing FaceOri to enable richer multi-device applications. However, this method assumes that all transmitters and the receivers can communicate via an additional channel (e.g., WiFi). We expect future work to explore other audio-only solutions to enable multiple device applications, such as using a frequency or phase-modulated chirp signal to provide unique device identification. Further, future work can adopt existing wireless multi-transmitter communication methods such as frequency hopping or code-multiplexing.

## 8 CONCLUSION

In this work, we have presented FaceOri, a novel spatial input technique using ultrasonic ranging. FaceOri leverages the microphones found in typical active noise cancellation (ANC) earphones to glean user head proximity and orientation with respect to a computing device that emit an inaudible chirp from its speaker. Through a user study, we evaluated FaceOri's performance for continuous head position and orientation tracking, and binarized attention detection. We explored and demonstrated how FaceOri can be used to capture user activity and gestural input, and enable more context-aware interactions. As the number and type of computing devices continue to proliferate, techniques like FaceOri can help to make our interaction experiences more human-centered.

## ACKNOWLEDGMENTS

This work is supported by the Natural Science Foundation of China (NSFC) under Grant No.6213000120, 62002198, Tsinghua University Initiative Scientific Research Program, the China Postdoctoral Science Foundation under Grant No.2021M691788, Beijing Key Lab of Networked Multimedia, and the Institute for Guo Qiang and Institute for Artifcial Intelligence, Tsinghua University.

## REFERENCES

[1] A. F. Abate, P. Barra, C. Bisogni, M. Nappi, and S. Ricciardi. 2019. Near Real-Time Three Axis Head Pose Estimation Without Training. *IEEE Access* 7 (2019), 64256–64265. https://doi.org/10.1109/ACCESS.2019.2917451

[2] Karan Ahuja, Andy Kong, Mayank Goel, and Chris Harrison. 2020. Direction-of-Voice (DoV) Estimation for Intuitive Speech Interaction with Smart Devices Ecosystems. 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, 1121–1131. https://doi.org/10.1145/3379337.3415588

[3] Apple. 2018. ARKit. https://developer.apple.com/documentation/arkit Accessed: 2018-03-29.

[4] Ilse Bakx, Koen Van Turnhout, and Jacques Terken. 2003. Facial orientation during multi-party interaction with information kiosks. *Proceedings of INTERACT 2003 Zurich, Switzerland* (2003), 163–170.

[5] G. Borghi, M. Fabbri, R. Vezzani, S. Calderara, and R. Cucchiara. 2020. Face-from-Depth for Head Pose Estimation on Depth Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 3 (2020), 596–609. https://doi.org/10.1109/TPAMI.2018.2885472

[6] Doug A. Bowman, Ernst Kruijff, Joseph J. LaViola, and Ivan Poupyrev. 2004. *3D User Interfaces: Theory and Practice*. Addison Wesley Longman Publishing Co., Inc., USA.

[7] Stephen Brewster, Joanna Lumsden, Marek Bell, Malcolm Hall, and Stuart Tasker. 2003. Multimodal 'eyes-Free' Interaction Techniques for Wearable Devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. Association for Computing Machinery, New York, NY, USA, 473–480. https://doi.org/10.1145/642611.642694

[8] Mihai Bâce, Sander Staal, and Andreas Bulling. 2019. Accurate and Robust Eye Contact Detection During Everyday Mobile Device Interactions. arXiv:1907.11115 [cs.HC]

[9] Gaoshuai Cao, Kuang Yuan, Jie Xiong, Panlong Yang, Yubo Yan, Hao Zhou, and Xiang-Yang Li. 2020. *EarphoneTrack: Involving Earphones into the Ecosystem of Acoustic Motion Tracking*. Association for Computing Machinery, New York, NY, USA, 95–108. https://doi.org/10.1145/3384419.3430730

[10] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh. 2021. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 1 (2021), 172–186. https://doi.org/10.1109/TPAMI.2019.2929257

[11] R. M. S. Clifford, N. M. B. Tuanquin, and R. W. Lindeman. 2017. Jedi Force-Extension: Telekinesis as a Virtual Reality interaction metaphor. In *2017 IEEE Symposium on 3D User Interfaces (3DUI)*. 239–240. https://doi.org/10.1109/3DUI.2017.7893360

[12] Heiko Drewes, Alexander De Luca, and Albrecht Schmidt. 2007. Eye-gaze interaction for mobile phones. In *Proceedings of the 4th international conference on mobile technology, applications, and systems and the 1st international symposium on Computer human interaction in mobile technology*. 364–371.

[13] Augusto Esteves, Eduardo Velloso, Andreas Bulling, and Hans Gellersen. 2015. Orbits: Gaze interaction for smart watches using smooth pursuit eye movements. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. 457–466.

[14] Google. 2020. ARCore API. https://developers.google.com/ar/reference Accessed: 2020-12-23.

[15] 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*. 1911–1914.

[16] Sebastian Hueber, Christian Cherek, Philipp Wacker, Jan Borchers, and Simon Voelker. 2020. Headbang: Using Head Gestures to Trigger Discrete Actions on Mobile Devices. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services* (Oldenburg, Germany) *(MobileHCI '20)*. Association for Computing Machinery, New York, NY, USA, Article 17, 10 pages. https://doi.org/10.1145/3379503.3403538

[17] Dries Hulens, Kristof Van Beeck, and Toon Goedemé. 2016. Fast and Accurate Face Orientation Measurement in Low-resolution Images on Embedded Hardware. https://doi.org/10.5220/0005716105380544

[18] Eiichi Itoh. 2001. Multi-modal Interface with Voice and Head Tracking for Multiple Home Appliances.. In *INTERACT*. 727–728.

[19] Robert J. K. Jacob. 1990. What You Look at is What You Get: Eye Movement-Based Interaction Techniques. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Seattle, Washington, USA) *(CHI '90)*. Association for Computing Machinery, New York, NY, USA, 11–18. https://doi.org/10.1145/97243.97246

[20] Vikram Jeet, Hardeep Singh Dhillon, and Sandeep Bhatia. 2015. Radio frequency home appliance control based on head tracking and voice control for disabled person. In *2015 Fifth International Conference on Communication Systems and Network Technologies*. IEEE, 559–563.

[21] Kaustubh Kalgaonkar and Bhiksha Raj. 2009. One-handed gesture recognition using ultrasonic Doppler sonar. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1889–1892.

[22] Christina Katsini, Human Opsis, Yasmeen Abdrabou, George E Raptis, Mohamed Khamis, and Florian Alt. 2020. The Role of Eye Gaze in Security and Privacy Applications: Survey and Future HCI Research Directions. In *Proceedings of the 38th Annual ACM Conference on Human Factors in Computing Systems (Honolulu, Hawaii, USA)(CHI'20). ACM, New York, NY, USA*, Vol. 21.

[23] MARUYAMA KINYA and ENDO MITSUO. 1983. The effect of face orientation upon apparent direction of gaze. *Tohoku Psychologica Folia* 42, 1-4 (1983), 126–138.

[24] Manu Kumar, Andreas Paepcke, and Terry Winograd. 2007. EyePoint: practical pointing and selection using gaze and keyboard. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 421–430.

[25] Mikko Kytö, Barrett Ens, Thammathip Piumsomboon, Gun A. Lee, and Mark Billinghurst. 2018. Pinpointing: Precise Head- and Eye-Based Target Selection for Augmented Reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173655

[26] Steven M. LaValle, Anna Yershova, Max Katsev, and Michael Antonov. 2014. Head tracking for the Oculus Rift. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*. 187–194. https://doi.org/10.1109/ICRA.2014.6906608

[27] Chen Liang, Chun Yu, Xiaoying Wei, Xuhai Xu, Yongquan Hu, Yuntao Wang, and Yuanchun Shi. 2021. Auth+Track: Enabling Authentication Free Interaction on Smartphone by Continuous User Tracking *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 2, 16 pages. https://doi.org/10.1145/3411764.3445624

[28] Yi Liang, Long Zhang, Mengdao Xing, and Zheng Bao. 2009. High-speed ground moving target detection research using triangular modulation FMCW. *Frontiers of Electrical and Electronic Engineering in China* 4, 2 (2009), 127–133. https://doi.org/10.1007/s11460-009-0032-z

[29] Rainer Malkewitz. 1998. Head pointing and speech control as a hands-free interface to desktop computing. In *Proceedings of the third international ACM conference on Assistive technologies*. 182–188.

[30] Rainer Malkewitz. 1998. Head Pointing and Speech Control as a Hands-Free Interface to Desktop Computing. In *Proceedings of the Third International ACM Conference on Assistive Technologies* (Marina del Rey, California, USA) *(Assets '98)*. Association for Computing Machinery, New York, NY, USA, 182–188. https://doi.org/10.1145/274497.274531

[31] Aadil Mamuji, Roel Vertegaal, J Shell, Thanh Pham, and Changuk Sohn. 2003. AuraLamp: contextual speech recognition in an eye contact sensing light appliance. In *Extended abstracts of ubicomp*, Vol. 3.

[32] Wenguang Mao, Jian He, and Lili Qiu. 2016. CAT: High-Precision Acoustic Motion Tracking. 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, 69–81. https://doi.org/10.1145/2973750.2973755

[33] Wenguang Mao, Zaiwei Zhang, Lili Qiu, Jian He, Yuchen Cui, and Sangki Yun. 2017. Indoor Follow Me Drone. 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, 345–358. https://doi.org/10.1145/3081333.3081362

[34] Sven Mayer, Gierad Laput, and Chris Harrison. 2020. Enhancing Mobile Voice Assistants with WorldGaze. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3313831.3376479

[35] I Skolnik Merrill et al. 2001. Introduction to radar systems. *Mc Grow-Hill* 7, 10 (2001).

[36] H. Nakajima, K. Kikuchi, T. Daigo, Y. Kaneda, K. Nakadai, and Y. Hasegawa. 2009. Real-time sound source orientation estimation using a 96 channel microphone array. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 676–683. https://doi.org/10.1109/IROS.2009.5354285

[37] Alberto Yoshihiro Nakano, Seiichi Nakagawa, and Kazumasa Yamamoto. 2009. Automatic estimation of position and orientation of an acoustic source by a microphone array network. *The Journal of the Acoustical Society of America* 126, 6 (2009), 3084–3094. https://doi.org/10.1121/1.3257548 arXiv:https://doi.org/10.1121/1.3257548

[38] 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*. 1515–1525.

[39] Alice Oh, Harold Fox, Max Van Kleek, Aaron Adler, Krzysztof Gajos, Louis-Philippe Morency, and Trevor Darrell. 2002. Evaluating Look-to-Talk: A Gaze-Aware Interface in a Collaborative Environment. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems* (Minneapolis, Minnesota, USA) *(CHI EA '02)*. Association for Computing Machinery, New York, NY, USA, 650–651. https://doi.org/10.1145/506443.506528

[40] Chunyi Peng, Guobin Shen, Yongguang Zhang, Yanlin Li, and Kun Tan. 2007. Beepbeep: a high accuracy acoustic ranging system using cots mobile devices. In *Proceedings of the 5th international conference on Embedded networked sensor systems*. 1–14.

[41] Mark A Richards. 2014. *Fundamentals of radar signal processing*. McGraw-Hill Education.

[42] Piero Rivera Benois, Patrick Nowak, and Udo Zoelzer. 2018. Hybrid Active Noise Control Structures: A Short Overview. In *Speech Communication; 13th ITG-Symposium*. 1–5.

[43] Florian Roider, Lars Reisig, and Tom Gross. 2018. Just Look: The Benefits of Gaze-Activated Voice Input in the Car. In *Adjunct Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Toronto, ON, Canada) *(AutomotiveUI '18)*. Association for Computing Machinery, New York, NY, USA, 210–214. https://doi.org/10.1145/3239092.3265968

[44] Andrey Ronzhin and Alexey Karpov. 2005. Assistive multimodal system based on speech recognition and head tracking. In *2005 13th European Signal Processing Conference*. IEEE, 1–4.

[45] Carlos Segura, Cristian Canton-Ferrer, Alberto Abad, Josep R Casas, and Javier Hernando. 2007. Multimodal head orientation towards attention tracking in smartrooms. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, Vol. 2. IEEE, II–681.

[46] Jeffrey S Shell, Ted Selker, and Roel Vertegaal. 2003. Interacting with groups of computers. *Commun. ACM* 46, 3 (2003), 40–46.

[47] Jeffrey S. Shell, Roel Vertegaal, and Alexander W. Skaburskis. 2003. EyePliances: Attention-Seeking Devices That Respond to Visual Attention. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI EA '03)*. Association for Computing Machinery, New York, NY, USA, 770–771. https://doi.org/10.1145/765891.765981

[48] Snapchat. 2020. Lens Studio. https://lensstudio.snapchat.com/api/ Accessed: 2020-12-08.

[49] R. Stiefelhagen. 2002. Tracking focus of attention in meetings. In *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*. 273–280. https://doi.org/10.1109/ICMI.2002.1167006

[50] Rainer Stiefelhagen and Jie Zhu. 2002. Head orientation and gaze direction in meetings. In *CHI'02 Extended Abstracts on Human Factors in Computing Systems*. 858–859.

[51] Ke Sun, Yuntao Wang, Chun Yu, Yukang Yan, Hongyi Wen, and Yuanchun Shi. 2017. *Float: One-Handed and Touch-Free Target Selection on Smartwatches*. Association for Computing Machinery, New York, NY, USA, 692–704. https://doi.org/10.1145/3025453.3026027

[52] David Joseph Tan, Federico Tombari, and Nassir Navab. 2018. Real-Time Accurate 3D Head Tracking and Pose Estimation with Consumer RGB-D Cameras. *International Journal of Computer Vision* 126, 2 (2018), 158–183. https://doi.org/10.1007/s11263-017-0988-8

[53] Roel Vertegaal. 2003. Attentive User Interfaces. *Commun. ACM* 46, 3 (March 2003). https://doi.org/10.1145/3263733

[54] Roel Vertegaal, Aadil Mamuji, Changuk Sohn, and Daniel Cheng. 2005. Media Eyepliances: Using Eye Tracking for Remote Control Focus Selection of Appliances. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems* (Portland, OR, USA) *(CHI EA '05)*. Association for Computing Machinery, New York, NY, USA, 1861–1864. https://doi.org/10.1145/1056808.1057041

[55] Roel Vertegaal, Robert Slagter, Gerrit van der Veer, and Anton Nijholt. 2001. Eye Gaze Patterns in Conversations: There is More to Conversational Agents than Meets the Eyes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Seattle, Washington, USA) *(CHI '01)*. Association for Computing Machinery, New York, NY, USA, 301–308. https://doi.org/10.1145/365024.365119

[56] 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, Article 18, 11 pages. https://doi.org/10.1145/3290605.3300248

[57] 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*. 82–94.

[58] Chin-Der Wann and Chih-Sheng Hsueh. 2007. NLOS mitigation with biased Kalman filters for range estimation in UWB systems. In *TENCON 2007 - 2007 IEEE Region 10 Conference*. 1–4. https://doi.org/10.1109/TENCON.2007.4429031

[59] M. Weiss. 1982. Analysis of Some Modified Cell-Averaging CFAR Processors in Multiple-Target Situations. *IEEE Trans. Aerospace Electron. Systems* AES-18, 1 (1982), 102–114. https://doi.org/10.1109/TAES.1982.309210

[60] Cheng Xu, Jie He, Yuanyuan Li, Xiaotong Zhang, Xinghang Zhou, and Shihong Duan. 2019. Optimal Estimation and Fundamental Limits for Target Localization Using IMU/TOA Fusion Method. *IEEE Access* 7 (2019), 28124–28136. https://doi.org/10.1109/ACCESS.2019.2902127

[61] Xuhai Xu, Chun Yu, Yuntao Wang, and Yuanchun Shi. 2020. Recognizing Unintentional Touch on Interactive Tabletop. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 1, Article 33 (mar 2020), 24 pages. https://doi.org/10.1145/3381011

[62] Yukang Yan, Yingtian Shi, Chun Yu, and Yuanchun Shi. 2020. HeadCross: Exploring Head-Based Crossing Selection on Head-Mounted Displays. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 1, Article 35 (mar 2020), 22 pages. https://doi.org/10.1145/3380983

[63] Jackie (Junrui) Yang, Gaurab Banerjee, Vishesh Gupta, Monica S. Lam, and James A. Landay. 2020. Soundr: Head Position and Orientation Prediction Using a Microphone Array. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376427

[64] Sangki Yun, Yi-Chao Chen, and Lili Qiu. 2015. Turning a mobile device into a mouse in the air. In *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*. 15–29.

[65] Cheng Zhang, Qiuyue Xue, Anandghan Waghmare, Sumeet Jain, Yiming Pu, Sinan Hersek, Kent Lyons, Kenneth A Cunefare, Omer T Inan, and Gregory D Abowd. 2017. Soundtrak: Continuous 3d tracking of a finger using active acoustics. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2 (2017), 1–25.

[66] Yuzhou Zhuang, Yuntao Wang, Yukang Yan, Xuhai Xu, and Yuanchun Shi. 2021. ReflecTrack: Enabling 3D Acoustic Position Tracking Using Commodity Dual-Microphone Smartphones. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21)*. Association for Computing Machinery, New York, NY, USA, 1050–1062. https://doi.org/10.1145/3472749.3474805