MAF: Exploring Mobile Acoustic Field for Hand-to-Face Gesture Interactions

Yongjie Yang
University of Pittsburgh
Pittsburgh, PA, USA
yoy28@pitt.edu

Tao Chen
University of Pittsburgh
Pittsburgh, PA, USA
tac194@pitt.edu

Yujing Huang
University of Pittsburgh
Pittsburgh, PA, USA
yuh115@pitt.edu

Xiuzhen Guo
Zhejiang University
Hangzhou, Zhejiang, China
guoxz@zju.edu.cn

Longfei Shangguan
University of Pittsburgh
Pittsburgh, PA, USA
longfei@pitt.edu

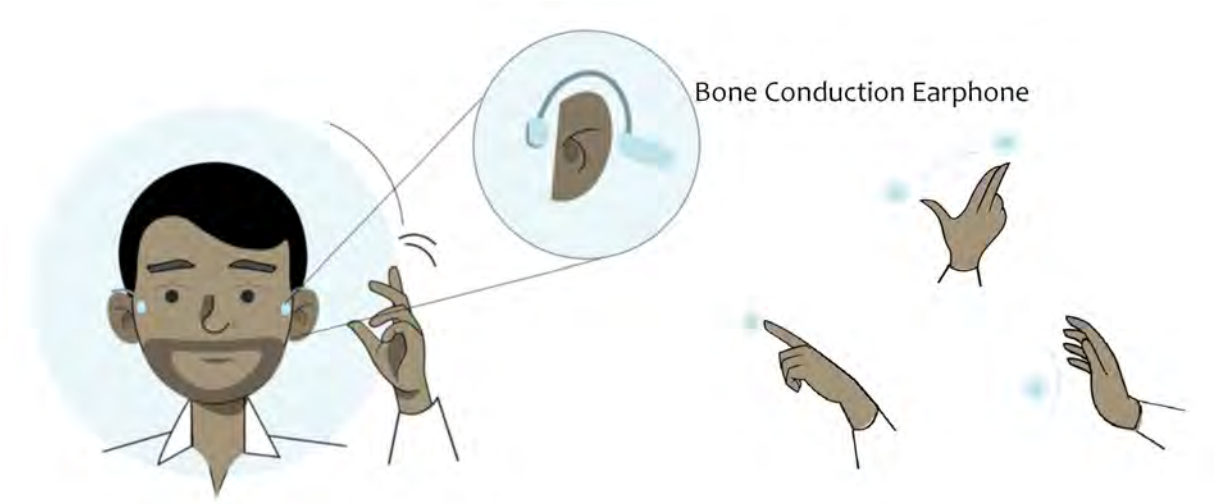

Figure 1: An illustration of Mobile Acoustic Field (MAF). MAF is based on a key observation – when audio is transmitted from the bone conduction earphone, it will not only propagate along the surface of the human face but also dissipate into the air, creating an acoustic field that envelops the individual's head. This acoustic field empowers the mobile user to define their own on-face and over-the-face hand gestures for human-computer interactions.

ABSTRACT

We present MAF, a novel acoustic sensing approach that leverages the commodity hardware in bone conduction earphones for hand-to-face gesture interactions. Briefly, by shining audio signals with bone conduction earphones, we observe that these signals not only propagate along the surface of the human face but also dissipate into the air, creating an acoustic field that envelops the individual's head. We conduct benchmark studies to understand how various hand-to-face gestures and human factors influence this acoustic field. Building on the insights gained from these initial studies, we then propose a deep neural network combined with signal preprocessing techniques. This combination empowers MAF to effectively detect, segment, and subsequently recognize a variety of hand-to-face gestures, whether in close contact with the face or above it. Our comprehensive evaluation based on 22 participants demonstrates that MAF achieves an average gesture recognition accuracy of 92% across ten different gestures tailored to users' preferences.

CCS CONCEPTS

• **Human-centered computing** → **Gestural input**.

KEYWORDS

Wearable Computing, Gesture Detection, Acoustic Sensing

ACM Reference Format:
Yongjie Yang, Tao Chen, Yujing Huang, Xiuzhen Guo, and Longfei Shangguan. 2024. MAF: Exploring Mobile Acoustic Field for Hand-to-Face Gesture Interactions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3613904.3642437

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

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642437

1 INTRODUCTION

Hand-to-face gestures are a natural and intuitive way to control devices or interfaces [89]. It improves the user experience across a wide spectrum of applications from virtual reality to smart home devices. At present, most hand-to-face gesture detection systems rely on dedicated sensing modalities such as IMUs [1, 39, 47, 48, 60, 72] and capacitive sensors [23, 43, 65, 76] to detect gestures having direct contact with the user's face. However, these systems face limitations in capturing contactless gestures (*i.e.*, those gestures performed over the face) because such gestures do not generate signals that are detectable by the aforementioned sensors.

While camera-based solutions [6, 26, 27, 38, 58, 61, 82, 90] can facilitate contactless hand gesture detection, their effectiveness is vulnerable to varying lighting conditions, potential obstructions, and often raises privacy concerns. Similarly, radar-based solutions [24, 49, 79, 80] can be influenced by distance variations and occasional obstructions due to arm movement. In addition, they often come with significant costs and power consumption [74], which limits their wide adoption.

In view of the shortcomings and limits of the existing approaches, in this paper, we ask a simple question – *can we design a system that is able to detect hand-to-face gestures, whether in contact with the face or above it, using widely accessible mobile devices*? A positive response to this question would enable mobile users to experience the advantages of gesture-based interactions in their everyday activities, moving this exciting technology one significant stride closer to widespread adoption. Furthermore, we anticipate that such pervasive interactions, when integrated with emerging Extended Reality (XR) technologies, could offer users unprecedented experiences.

We give an affirmative answer by presenting ***mobile acoustic field (MAF)***, a novel acoustic sensing approach that leverages the commodity hardware in bone conduction earphones for hand-to-face gesture interactions. MAF draws inspiration from the principles of surface acoustic waves (SAW) and leaky surface acoustic waves (LSAW), which are well-studied in seismology. In particular, when the speaker of the bone conduction earphones is in contact with the user's skin, the emitted acoustic signal will generate acoustic radiation around the user's facial structure in the form of surface acoustic waves. In the meantime, part of the sound waves will dissipate into the surrounding air, forming leaky surface acoustic waves. The combination of these two signals effectively creates an acoustic field surrounding the user's head, which we refer to as the Mobile Acoustic Field, as it is generated by widely available earphones and moves with the user. User gestures performed on or in the vicinity of the face can perturb the channel of SAW or LSAW signals, as shown in Figure 2. By observing how the received SAW and LSAW signals change over time, it is possible to detect and further distinguish these gestures.

Compared with existing hand-to-face gesture recognition systems [12, 43, 47, 57, 71, 75, 82, 86, 87, 89], MAF offers several distinct advantages. Firstly, its wearable nature guarantees that users can move freely without any inconvenience or hindrance, always interacting with the device seamlessly. Secondly, MAF does not depend on specialized sensors or require any modifications to standard bone conduction earphones. This means that mobile users can effortlessly enjoy hand-to-gesture interactions without any additional equipment or alterations.

To harvest the aforementioned benefits, we first design a series of benchmark studies to understand the capacity and capability of the mobile acoustic field, answering a plethora of fundamental questions such as "How does the sound volume affect the gesture detection accuracy"? "Would the spacing between the hand and the face matter"? and "What is the effective size of the mobile acoustic field"? We then craft a plethora of user studies to *i*) examine the impact of various human factors on the mobile acoustic field; and *ii*) assess the social acceptance of this mobile acoustic field-based gesture interaction by interviewing 22 participants. The preliminary results are promising and the user feedback is generally positive.

Based on these promising preliminary results, we then build a signal-processing pipeline for detecting and recognizing hand-to-face gestures to showcase the potential of the mobile acoustic field. To begin, we first create a set of 12 hand gestures, comprising six performed on the face and six over the face. From this set, we curate a selection of 10 gestures, consisting of four on-face gestures and six over-the-face gestures, based on the preference of 22 participants. Given that over-the-face gestures tend to produce relatively weak channel distortions, we propose a series of signal-processing techniques combined with a Convolutional Recurrent Neural Network (CRNN) model to segment the signal, enhance its SNR, and subsequently recognize each of them accurately.

We follow the IRB protocol to conduct comprehensive field studies based on 22 volunteers. The experiment results show that MAF achieves an average gesture recognition accuracy of above 92% for all ten types of testing gestures. We further conduct benchmark experiments in various environmental settings to scrutinize the influences of environmental noise, earphone remounting, human speech, body movement, skin moisture level, and music playback on MAF's performance. We observe that despite fluctuations amongst these variables, MAF could adequately accommodate the demands of most daily life settings.

Our contributions are summarized as follows:

- We identify a new opportunity for hand-to-face gesture interaction based on commodity bone conduction earphones. We study the capacity, robustness, and user acceptance of this new interaction opportunity by designing a series of user studies.
- We demonstrate the potential of such a mobile acoustic field for hand-to-face gesture interaction by building an end-to-end, data-driven signal processing pipeline. The proposed approach can effectively detect and further recognize ten user-selected gestures at high accuracy ($\geq$92%).
- We evaluate the performance of our prototyping system with 22 participants. In addition, we also conduct a comprehensive UX study to gauge the users' attitudes toward this new technology.

The remaining of this paper is organized as follows: Section 2 discusses related works in this domain. Section 3 starts with a detailed introduction of SAW and LSAW, briefly touches on the feasibility of the MAF systems, and then delves into their real-world application scenarios. Section 4 examines the practical possibility of integrating SAW and LSAW signals, evaluating them from three different perspectives. Section 5 describes the development and validation of signal processing and machine learning frameworks

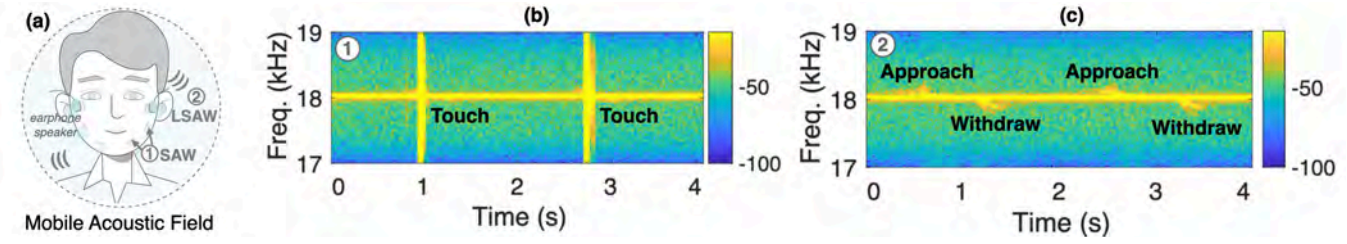

Figure 2: (a): Illustration of surface acoustic wave (SAW) and leaky surface acoustic wave (LSAW); (b): Touching the check twice results in two peaks in the SAW signals; (c): Approaching the face leads to weak yet observable variations in LSAW signals.

for MAF. Section 6 assesses user perceptions of the MAF system and examines the performance benchmarks of the MAF system in a variety of challenging situations. Section 7 discusses the system limitations and potential improvement. Section 8 concludes.

2 RELATED WORK

In this section, we review sensor-based solutions for hand-to-face gesture interactions, with a particular focus on acoustic-based and earable-based solutions that are closely related to our design.

2.1 Vision-, RADAR-, and IMU-based Solutions

There exists a wide range of designs that employ various sensors to detect facial gestures. The most extensively researched approach is using computer vision [6, 26, 27, 38, 58, 61, 82, 90] for non-contact detection. However, such an approach usually consumes significant computational resources, which becomes a critical issue for mobile and wearable platforms. Moreover, this approach also suffers from low-lighting conditions and occlusions, raising privacy concerns. mmWave radar [24, 49, 79, 80], on the other hand, has also been employed to detect face contact. For instance, Li et al. [44] demonstrated the use of microwave radar systems for hand gesture recognition. Rojas et al. [68] employed sonar-inspired techniques to measure the distance between the hand and the face and trigger an alarm if the user approaches too closely. Besides, past works also explored various types of wearable sensors for hand-to-face gesture recognition. FaceTouch [52] utilized a vibration sensor placed on the wrist or finger to track hand movements towards the facial region, and the followups [1, 11, 53, 57, 60] had also leveraged accelerometer and inertial measurement unit (IMU) sensors on commodity smartphones to detect facial touch.

Unlike these previous studies, we utilize a speaker-microphone pair commonly found in everyday earphones for facial gesture interaction. Our approach overcomes issues related to lighting conditions and ensures privacy protection. In contrast to vibration or IMU sensing, which operates at low frequencies, our system achieves greater precision in gesture recognition thanks to its capability to achieve a high sampling rate of up to 48kHz. When compared to mmWave sensing, our solution is both more cost-effective and power-efficient.

2.2 Acoustic-based Solutions

There is also plenty of research on using speaker-microphone pairs for acoustic sensing. For instance, FaceOri [81] leveraged an ultrasonic chirp to track head position and orientation on earphones. EchoSpeech [92] leveraged the inaudible sound emitted from an eyewear device to detect and further recognize silent speech. SoundWave [25] employed inaudible band tones generated by the PC speaker to detect gestures in the surrounding space, exploiting the concept that various gestures induce distinct frequency shifts due to Doppler effects. Sonicoperator [46] devised a recursive neural network and implemented it on mobile devices to recognize mid-air human gestures. Additionally, Dolphin [63], Strata [91], and AudioGest [69] also explored similar techniques for recognizing human gestures. In another approach, CAT [56] employed frequency-modulated continuous wave (FMCW) signals to estimate the relative displacement between smartphone speakers and microphones, subsequently integrating Doppler shift data obtained through FMCW with IMU measurements to enhance gesture tracking precision. Furthermore, fingerIO [64] had embraced orthogonal frequency-division multiplexing (OFDM) modulation to monitor subtle finger movements in the proximity of the phone. FingerPing [93] further took advantage of multiple microphones mounted on the wrist and thumb to recognize different hand gestures by analyzing acoustic resonance features.

While these studies show promise, they often require explicit user involvement or are primarily effective in static settings. For instance, Sonicoperator [46] mandated that users hold the phone and aligned the microphone to face forward for gesture detection, restricting its usability to relatively stationary scenarios. Similarly, SoundWave [25] functioned effectively when the user was seated in front of a computer. In contrast, our system capitalizes on the distinctive potential of the mobile acoustic field generated by readily available bone conduction earphones. This enables us to detect hand-to-face gestures in both stationary and mobile environments without necessitating explicit user intervention.

2.3 Earable-based Solutions

Earable computing [67] is a rapidly growing research field, with an increasing amount of attention given to technologies surrounding ear-based or headset applications for acoustic sensing. Lissermann et al. [50] defined the possible ways of interacting around the ear. Then there was a groundbreaking study by Chen et al. [12] discovered that the majority of desired ear-based interactive gestures involve mid-air hand interactions. HeadFi [18] transformed everyday headphones into smart devices, making earable sensing easily accessible. However, HeadFi required additional hardware support and only allowed for conventional interaction with the existing headset. FreeDigiter [59] integrated proximity sensors into earbuds, enabling near-ear non-contact input from finger gestures.

FaceSense [37] designed an earbud with impedance sensing and thermal sensing for gesture recognition. Earbuddy [86] leveraged feed-forward microphones on ANC earphones to detect the sound of touching gestures in the facial and ear areas for gesture recognition. However, it cannot detect non-contact, over-face gestures.

In a different approach, SonicASL [36] leveraged an external speaker and a microphone facing toward the deaf individual to recognize his/her sign language. Further, EarEcho [20] leveraged in-ear microphones to identify different users based on unique ear canal structures. Amesaka et al. [4] used facial muscle movements to alter the transfer function of the user's ear canal for facial expression recognition. Meanwhile, BrianyHand [75] employed a mini-projector and color camera within an earbud to relay input feedback to the user's ear. PrivateTalk [87] employed audio signals reaching the left and right ears to interpret the user's intent to interact. Similarly, Li et al. [47] designed voice-accompanying hand-to-face (VAHF) gestures for voice interaction.

Different from the above approaches, we propose an active probing-based approach that explores the surface acoustic wave and leaky surface acoustic waves produced by bone conduction earphones for both on-face and over-the-face gesture recognition. This approach leverages the natural properties of bone conduction to provide a more immersive and interactive user experience.

2.4 Surface Acoustic Waves and Leaky Surface Acoustic Waves

We are not the first to explore surface acoustic waves for human-centric sensing. There are a bunch of works [22, 30, 51, 62, 73, 83] have already explored various types of sensors to generate surface acoustic waves. These sensors include microphones [5, 28, 29, 34, 40, 54], IMU sensors [21, 35, 42], geophones [31, 66], and piezoelectric devices [19]. The most recent work, SAWSense [33], explored a newly emerging sensor known as a voice pick-up unit (VPU) for on-desk gesturing. In a different vein, Leaky Acoustic Surface Waves (LSAW) have more recently found utility in collision avoidance for robotics [17, 19]. This concept was realized by deploying a pair of piezoelectric sensors on a robotic arm to generate LSAW, enabling the monitoring of obstacles encountered by the robotic arm.

Our work draws inspiration from these pioneering efforts but distinguishes itself in two key ways. Firstly, MAF relies solely on a pair of off-the-shelf bone conduction earphones, eliminating the need for specialized sensors, such as piezoelectric sensors. Secondly, MAF possesses the capability to monitor both on-face and over-the-face gestures without any hardware modifications to the earphones. Consequently, it holds significant potential to enhance a wide range of facial gesture applications.

3 MOBILE ACOUSTIC FIELD: PRELIMINARY, FEASIBILITY, AND APPLICATIONS

In this section, we first introduce the concept of the mobile acoustic field (§3.1), highlighting its potential for hand-to-face gesture interaction through a feasibility study (§3.2). Subsequently, we describe three representative mobile applications that can directly benefit from the mobile acoustic field (§3.3).

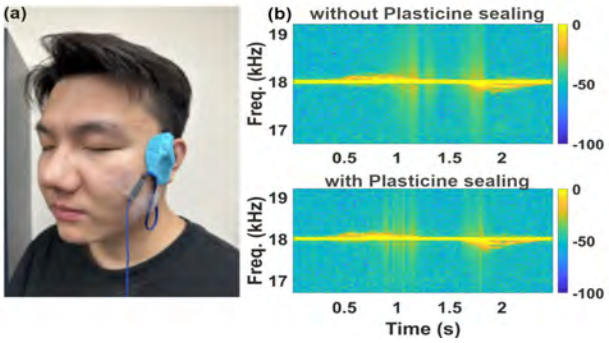

Figure 3: (a) Sealing the earphone with a plasticine. (b) Top: The signal wave produced by an approaching gesture without the plasticine sealing. Bottom: The signal wave produced by an approaching gesture with the plasticine sealing.

3.1 MAF Primer

When a mobile user uses bone conduction earphones to listen to an audio clip, the electrical audio signals get transformed into mechanical waves by the diaphragm of the earphone's speaker. Because the earphone is in direct contact with the user's head, these mechanical waves transfer their energy into the tissues of the human head, forming *Surface Acoustic Waves (SAW)* [7]. SAWs are a type of mechanical waves that propagate along the interface between a solid material and its adjacent medium [7], exhibiting a longitudinal and vertical shear component along the surface. Furthermore, these surface acoustic waves also disperse into the air as they travel through the user's facial region, creating another type of signal known as *Leaky Surface Acoustic waves (LSAW)* [77].

Both SAW and LSAW waves persist as long as the mobile user continues to play audio, offering opportunities for interactions in close proximity to the user's face. Essentially, the combination of these two signals generates an acoustic field that envelops the user's head, as shown in Figure 2(a). We term it as *Mobile Acoustic Field* as it is produced by the headphones and moves with the user.

3.2 MAF for Hand-to-Face Gesture Interaction

We envision the mobile acoustic field can be leveraged to detect and recognize different types of gestures that are performed both *in contact with the human face* and *in the vicinity of the human face (i.e., over the face)*, without the instrument of any dedicated sensors. To validate this feasibility, we invite a volunteer to wear a pair of bone conduction earphone. The earphone emits a single tone at the ultrasound frequency band. As the bone conduction earphone is in close contact with the human face, this single-tone probing signal produces surface acoustic waves and leaky surface acoustic waves that propagate along the human head and get picked up by a microphone attached to the bottom part of the left face.

We first ask the volunteer to gently touch her cheek. This physical contact not only generates a new signal but also has an impact on how surface acoustic waves propagate. Consequently, in Figure 2(b), we can clearly observe significant deviations from the original signal when two cheek-touching gestures are performed. Leaky Surface Acoustic Waves (LSAWs) create an acoustic field above the face due to the energy leakage from the surface acoustic waves. As illustrated in Figure 2(c), when the volunteer moves her

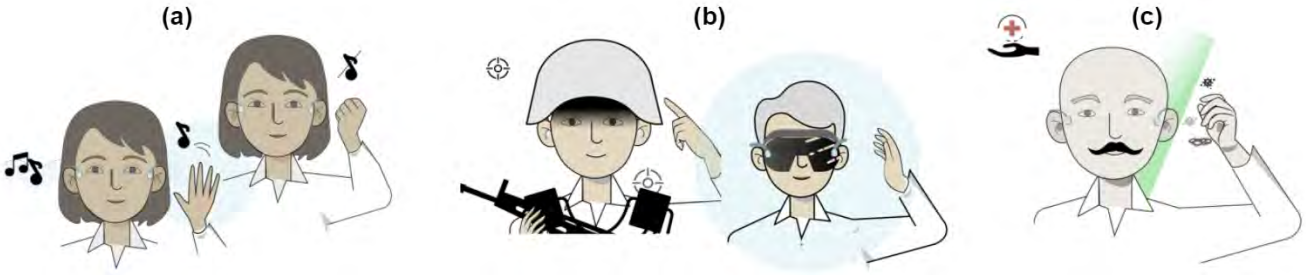

Figure 4: Three potential applications of MAF, out of many. (a): software-defined headphones that allow users to play and stop music with in-air gestures. (b): enhanced gaming experience – the military game can automatically recognize user gestures in the air with MAF. (c): face-touching awareness and prevention.

hand closer to or further away from her face (around 2 cm apart), we also detect a subtle change in the received signals. This occurs because the approaching hand gesture alters the way the leaky waves propagate through the air.

One might wonder about the origin of these LSAW signals, questioning whether they stem from the movement of the rear section of the earphone or the mobile acoustic field. To answer this question, we seal the earbud with plasticine, as shown in Figure 3(a). Figure 3(b) shows the signal wave produced by an approaching gesture in the presence and absence of plasticine sealing, respectively. We observed a clear signal pattern when the earphone was sealed, which demonstrates that the LSAW signal is due to the mobile acoustic field, not the motion of the rear part of the earpiece.

3.3 Applications of MAF

With the all-encompassing mobile acoustic field, the human head becomes an independent space for interaction. Below we list a few potential applications of MAF, out of many.

• **Software-defined intelligent headphones**. The current market offers intelligent headphones like Bose QC35 and Microsoft Surface headphones, which utilize built-in accelerometers to detect touch gestures, enabling gesture control. However, these existing intelligent headphones are expensive and bulky, limiting their widespread adoption. As Figure 4(a) shows, we envision that mobile acoustic sound can enable mobile users to define personalized gestures for controlling volume, playback, and muting without the need for dedicated sensors. This approach eliminates the cost and bulk associated with current intelligent headphones, offering a more accessible and customizable gesture control solution for mobile users worldwide.

• **Enhanced VR gaming experience**. The on-face and over-the-face human gesture interaction can be leveraged to enhance the virtual reality (VR) experience. By accurately detecting and interpreting the gestures of the user's hands, the VR applications allow users to manipulate and control virtual objects and perform various actions, without the need for physical controllers. Moreover, it opens up possibilities for enhanced social interactions, particularly in scenarios depicted in Figure 4(b), like VR team-based shooter games Larcenauts [3]. Users can communicate non-verbally through their over-the-face hand gestures, fostering a more engaging shooting game experience.

• **Face-touching awareness and prevention**. The identification and monitoring of face-touching behavior are crucial in preventing virus transmission and promoting hygienic practices, particularly during the COVID-19 pandemic when the virus can spread through contaminated surfaces and close contact. However, existing solutions, such as those based on earables and smart rings [37, 52], can only detect face-touching after it has already occurred, lacking the ability to proactively prevent virus transmission. In contrast, the implementation of over-the-face detection utilizing the mobile acoustic field allows for the detection of face-touching intentions by capturing hand movements approaching the face, such as in Figure 4(c) shows. This capability enables a proactive approach to prevent face-touching behavior and mitigate the risk of virus transmission.

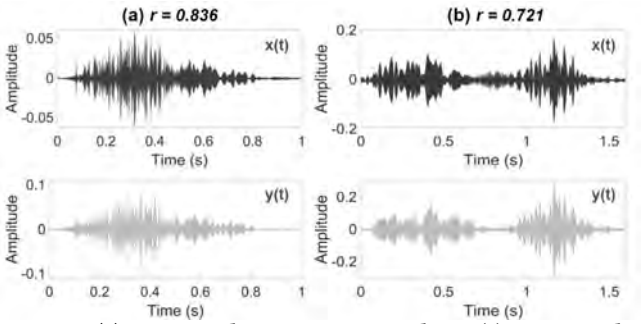

Figure 5: (a) Top: Touching gesture template $x(t)$; Bottom: the received touching gesture $y(t)$. (b) Top: Approaching gesture template $x(t)$; Bottom: the received approaching gesture $y(t)$.

4 FEASIBILITY STUDIES

To gain a comprehensive understanding of the mobile acoustic field (MAF) for gesture-based human-computer interaction, we design a plethora of benchmarks and user studies to effectively assess the capacity (§4.1), robustness (§4.2), and user acceptance (§4.3) of MAF in real-world scenarios. All human evaluations in this section are conducted in full accordance with the internal Institutional Review Board (IRB) protocol. The maximum signal transmission power is set to 60 dB SPL, 10 dB lower than the CDC's regulation [70].

We first introduce a gesture detection algorithm for feasibility studies. Specifically, we apply cross-correlation to detect the presence of a human gesture in the time domain. Consider a template signal $x(t)$ and a received data segment $y(t)$ where t=0,1,2...N-1, the cross-correlation r of these two signals is defined as follows:

$$r = \frac{\sum_{t=1}^{N}[(x(t)-\bar{x})(y(t)-\bar{y})]}{\sqrt{\sum_{t=1}^{N}(x(t)-\bar{x})^2 \sum_{t=1}^{N}(y(t)-\bar{y})^2}} \quad (1)$$

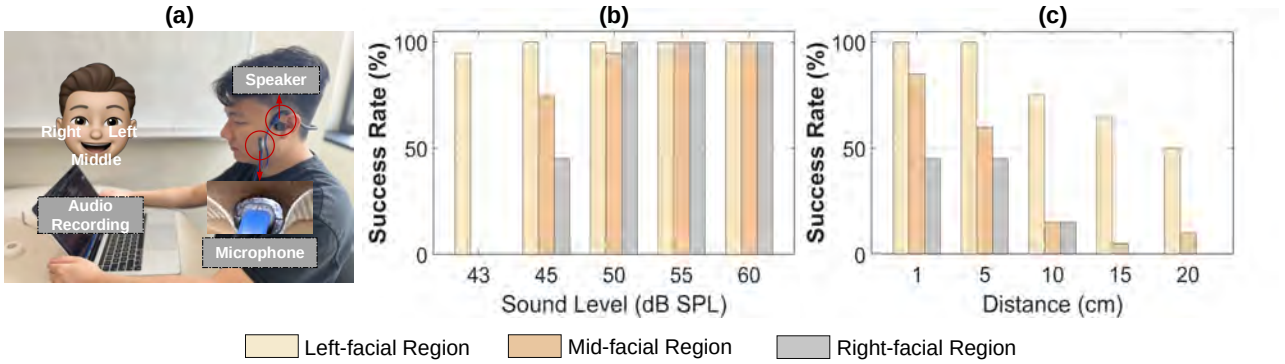

Figure 6: Detecting approaching gestures in different volume levels or distance settings. (a): experiment setups; (b): the gesture detection success rate in different volume settings (dB SPL); (c): the gesture detection success rate in different distance settings.

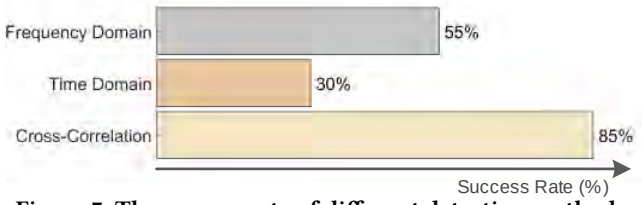

Figure 7: The success rate of different detection methods.

where $\bar{x}$ and $\bar{y}$ are the mean of the template and received data, respectively. A gesture is considered to be successfully detected as long as the cross-correlation coefficient r is higher than a pre-defined threshold. A correlation coefficient close to 1 suggests a strong positive relationship. In many cases, a value above 0.6 or 0.7 is often considered indicative of a strong positive correlation. So we set the threshold to 0.7 in our experiment. The algorithm runs by sliding the window by one sample each time. As a feasibility study, we did not prioritize the optimality of the gesture detection algorithm chosen. Instead, we opted for cross-correlation due to its resilience to varying gesture intensities. It is worth noting that many other advanced gesture detection algorithms could be applied for improved performance.

Figure 5 shows two time series templates $x(t)$ and the received gestures $y(t)$, respectively. To obtain a representative template signal $x(t)$, we collected the touching and approaching gestures 100 times, respectively. We then manually segment these 100 gestures and pass each gesture through Wiener filter to remove background interference. Since these 100 gestures last for different time periods, we resize them into equal length (*i.e.*, with the same length of signal samples) through interpolation. The final template signal $x(t)$ is derived by calculating the mean at each signal sample across these 100 aligned and normalized gestures, effectively capturing the average gesture pattern and its variability.

Figure 7 shows the gesture detection rate of the cross-correlation-based solution and two other baselines that detect gestures in the time and frequency domain, respectively. The time-domain baseline compares the signal energy with a threshold to detect the gesture, while the frequency-domain baseline compares the energy within the Doppler frequency shift with a pre-defined threshold. To be fair, the two energy thresholds[1] are obtained by computing the average energy of the same 100 gestures being used for determining the threshold for our proposed gesture detection solution. The testing set contains approaching gestures under different distances, angles with respect to the face, and speed. As the result shows, the time-domain baseline achieves merely 30% accuracy. This is expected since different gesture speeds result in different signal intensities while a fixed threshold fails to factor in such impact. Moreover, body motions may also lead to a signal energy surge, leading to false positives. On the other hand, the frequency-domain method achieves slightly better performance (55% accuracy) since it is resilient to body motions. However, its performance is also sensitive to the variations of gesture intensities. In contrast, the cross-correlation approach achieves 85%[2] accuracy on average.

4.1 Understanding MAF's Capacity

Q1: How does the sound volume affect the gesture detection accuracy? Our initial investigation focuses on investigating whether the intensity of the acoustic field generated by leaky surface acoustic waves is sufficient for detecting gestures. We invite two participants, named *A* and *B*, to conduct this experiment in a controlled laboratory environment. As shown in Figure 6(a), participant *A* wears a pair of GZCRDZ bone conduction wired headphones [2] and sits on a chair. The headphone speakers are in close contact with the face so that the acoustic signal will not leak into the air. The microphone on the bone conduction earphone is facing toward the participant *A*'s face but has no direct contact with the skin. This allows the microphone to detect both SAW and LSAW signals. The left speaker of the earphone emits a single-tone probing signal at the ultrasound frequency of 18 kHz. At the beginning of the experiment, participant *B* proceeds to perform approaching gestures by moving her palm closer to participant *A*'s face, with 1 cm spacing in-between. The involvement of two participants ensures the absence of body motion artifacts from participant *A* that could interfere with gesture detection. It also guarantees consistent alignment of the palm each time they approach the face.

[1]For the frequency-domain baseline, we first identify the Doppler frequency shift due to gestures and then calculate the average energy within the Doppler frequency band.
[2]We evaluate these three methods in challenging scenarios that require more intricate hand or finger movements than those performed in Section 4.1 and Section 4.2.

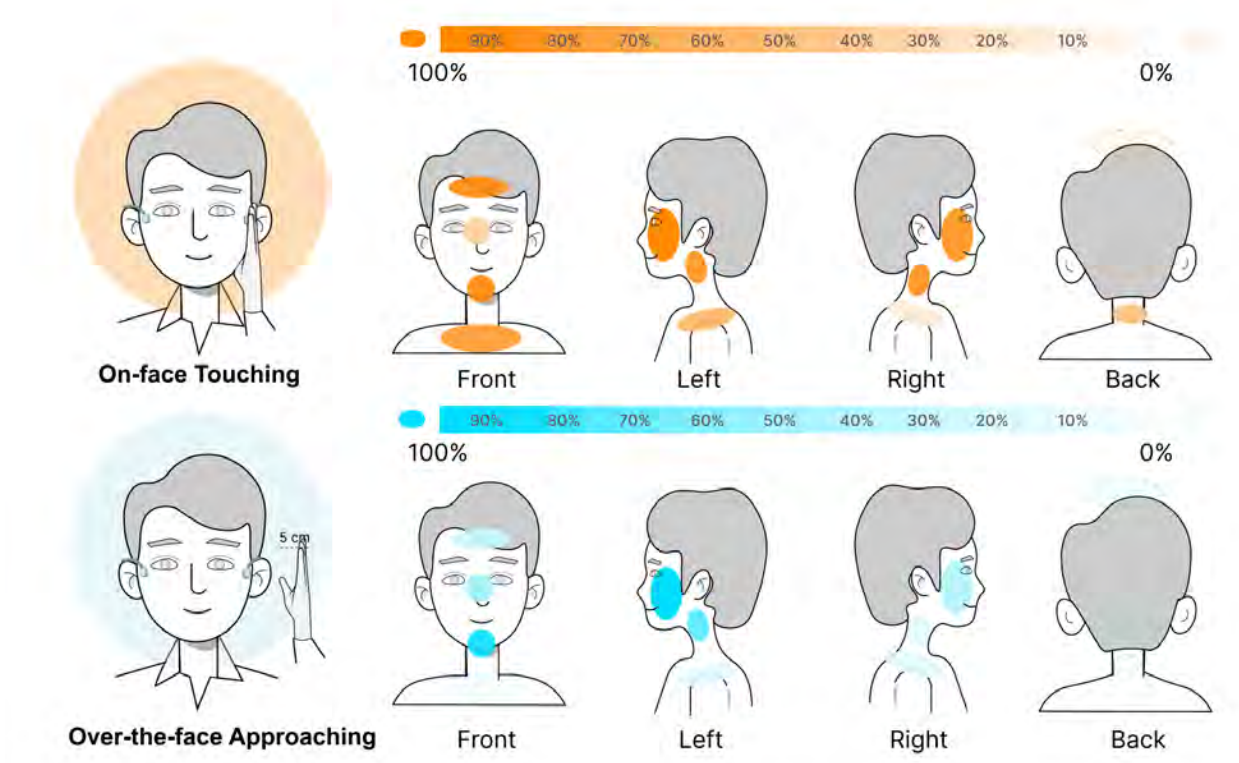

Figure 8: The effective coverage area of SAW signals and LSAW signals. The detection success rate of touching (top) and approaching (bottom) gestures grows with the darkness of the heatmap.

▷ **Experiment Results**. We divide the human face into the left-facial region, middle-facial region, and right-facial region. As depicted in Figure 6(a), within each region, we vary the sound volume between the minimum 43 dB SPL to 60 dB SPL to assess the success rate of gesture detection at different volume levels. We repeat the experiment 20 times for each volume setting. Figure 6(b) shows the success rate of gesture detection (referred to as the *success rate* in the figure) across the left-, middle-, and right-facial regions of the human face. We observe consistently high success rates (>95%) in the left-facial region across all five volume settings. However, in the middle-facial and right-facial regions, we notice that approaching gestures are difficult to detect at low sound volumes (43 dB SPL). This is expected because the sound source (the speaker) is positioned on the left side of the human head. Consequently, the sound signals (LSAW) attenuate significantly before reaching the middle-facial and right-facial regions, resulting in a low success rate for approaching gesture detection.

As we increase the sound volume to 45 dB SPL and further to 50 dB SPL, we observe a substantial improvement in the success rate for the middle-facial region, reaching over 60% and eventually surpassing 95%. Similarly, in the right-facial region, the success rate increases to over 45% at 45 dB SPL and then reaches 95% at 50 dB SPL. These preliminary experiments demonstrate the sensitivity of gesture detection success rates to sound volume. However, a sound volume of 45 dB SPL proved to be sufficiently strong for successful gesture detection in the left-facial region. To ensure coverage of all three regions, a slight increase in sound volume to 50 dB SPL can be implemented without compromising safety requirements.

Q2: How does the gesture detection success rate change with the spacing between hand and face? Next, we set the sound volume to 45 dB SPL and examine how the gesture detection success rate changes with the spacing between the hand and the face. We anticipate the benchmark results could reveal the smallest spacing in which the bone conduction earphones can generate detectable waves around the user's face. The experiment setup is similar to the previous experiment.

▷ **Experiment Results.** Figure 6(c) illustrates that the approaching gesture can consistently be detected when the hand is within 5 cm of participant *A*'s left face. However, the success rate declines to below 80% as the spacing between the hand and the left face increases. Furthermore, when the hand is 20 cm away from volunteer A's left face, the success rate drops below 50%. Notably, the success rate decreases significantly in both the middle- and right-facial regions compared to the left-facial region, primarily due to the greater distance from the signal source (headphone speaker). For instance, when the hand-to-face spacing is reduced to 10 cm, the success rate in the middle- and right-facial regions decreases to 15%,

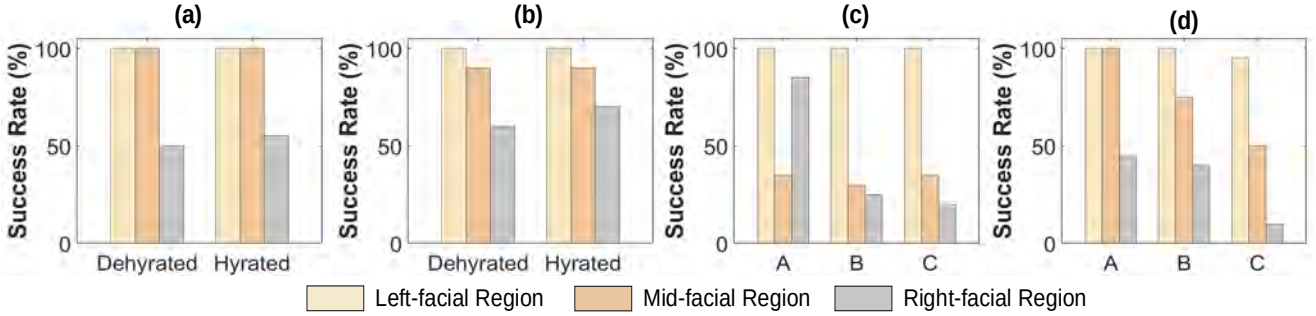

Figure 9: The set of histograms illustrates the success rates of gesture detection across different facial regions (left, middle, and right) under various conditions. Histogram (a) focuses on the face-touching gesture under different hydration conditions, while histogram (b) showcases the face-approaching gesture under the same hydration conditions. Evaluating (c) the detection success rate of the face-touching gesture and (d) the face-approaching gesture under varied motion conditions. The symbols *A*, *B*, and *C* denote the stationary state, walking state, and jogging state, respectively.

representing a 60% reduction compared to the left-facial region. In practice, a longer detection distance may cause false alarms as a nearby user may unintentionally trigger a gesture. Hence, we set a targeting distance of 5 cm in order to minimize the false positives.

Q3: How does the mobile acoustic field look like? Finally, we measure the effective coverage area of the mobile acoustic field. In this experiment, we divide the region of interest into 12 sub-regions: the left and right cheeks, forehead, chin, nose, top of the head, back of the head, left and right neck, back neck, and left and right shoulders. We detect the coverage area of both SAW signals and LSAW signals by performing touching and approaching gestures at these 12 sub-regions. The approaching gesture maintains an effective spacing of roughly 5 cm ±1 cm between the hand and the face. The sound volume is set to 45 dB SPL.

▷ **Experiment Results.** We use the *heatmap* to represent the gesture detection success rate. As shown in Figure 8, we observe that touching gestures generally achieve a broader coverage area, with a significantly higher success rate due to the following two reasons. Firstly, the touching gestures not only impact the surface acoustic wave channels but also generate new signals that are detectable by the microphone. This dual effect of touching face gestures substantially modifies the received signal detected by the microphone, leading to a notable increase in amplitude. Secondly, the touching gestures occur at a closer distance to the microphone sensor than approaching gestures.

Taking further scrutiny of the touching gestures (top figure in Figure 8), the predominantly deep orange coloration around the entire head area (>80%) indicates a high gesture detection success rate. Interestingly, there is a notable exception in the user's nose. This is due to the relatively small contact area between hand and nose, which results in a less pronounced signal path alteration. In addition, other positions away from the head, *e.g.*, the neck and the front chest can still achieve around a 70% success rate. The success rate declines significantly in areas such as the shoulders and the back of the head, dropping to less than 30%. This decline is due to the significant SAW signal attenuation over distance. Thus, from a practical perspective regarding touching face gestures, we focused on the head area due to its exceptional stability and optimal recognition success rate.

Lastly, we shift our focus to approaching gestures, as shown in the bottom figure in Figure 8. the left-side regions of the head, including the left face, left neck, and left chin, represented as triangular areas on the heatmap, exhibit a distinctly higher success rate (>75%) compared to the rest. This striking difference underscores the fact the LSAW attenuates more severely than SAW due to its air propagation path: it emanates from the left speaker on the earphone, permeating the left face, and extending to the head and shoulder areas. Hence the users could be instructed to focus their gestures in the areas with the strongest signal transmission for the highest recognition success rates. This would enhance the overall user experience by ensuring the system responds accurately and reliably to user inputs.

4.2 Understanding MAF's Robustness

Next, we examine the robustness of the mobile acoustic field under diverse conditions. Ideally, MAF should maintain its performance within an acceptable range in the presence of both body movement and changes in skin conditions.

Q1: Will facial hydration condition affect the success rate of gesture detection? Firstly, we evaluate the potential effects of daily facial skin condition changes on system performance. This is crucial since the circadian rhythm can significantly influence various skin conditions, including its hydration levels [55]. We invite a volunteer (participant *A*) to perform on-face touching and over-the-face approaching gestures, once after applying a moisturizing mask in the morning (hydrated facial state) and once following a typical workday in the evening (dehydrated facial state). We label the microphone's position on the user's face with a marker to ensure precise remounting at the same location. This experiment is designed to simulate common skin conditions and monitor their potential impact on system performance. The experimental design follows the setup used in the facial hydration state experiment, wherein the left channel of a bone conduction earphone transmits an 18kHz ultrasonic wave with an energy level of 45 dB SPL. The effective detection distance for approaching gestures is also maintained at 5 cm ±1 cm.

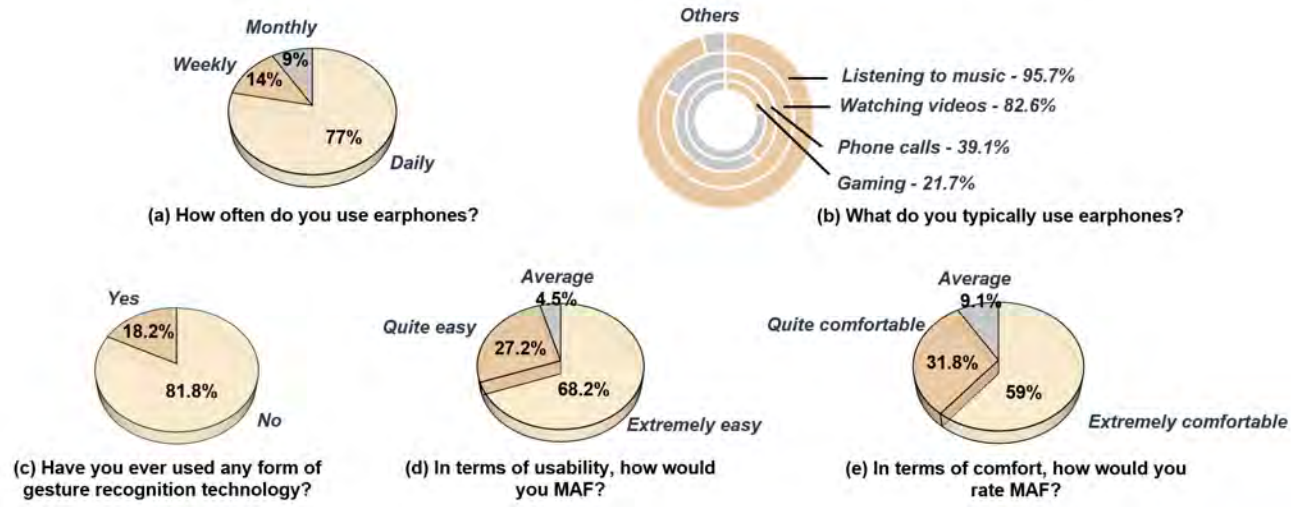

Figure 10: Gauging the initial feedback from 22 participants on the MAF system.

▷ **Experiment Results.** Figure 9(a) and Figure 9(b) show the success rate of on-face touching gestures and over-the-face approaching gestures in dehydrated and hydrated facial states, respectively. The state of dehydration is characterized by the participant's facial skin appearing oily and taut. Conversely, with hydration, the participant's skin condition is smooth and delicate. We observe the success rate of both touching gestures and approaching gestures remains consistently high in the two facial states, with slight differences when the user performs gestures in the right-facial regions. This shows that facial skin conditions do not remarkably impact the sensitivity of the microphone or the headset's audio output, thus not leading to a significant alteration in the signal path.

Q2: Will motion artifact affect the success rate of gesture detection? Next, we experimentally validate whether different body motions could disrupt the efficacy of the mobile acoustic field for on-face (touching) and over-the-face (approaching) gesture detection. In this controlled lab experiment, participants carry out touching and approaching gestures in three movements: stationary, walking (≈ 1.4m/s), and joggling (≈ 3m/s). Similar to the previous experiment, the user plays a single-tone signal at 18kHz through the left speaker on her bone conduction earphone. The sound volume is fixed to 45 dB SPL. However, compared with the previous experimental settings, participant *A* has been asked to complete all experiments independently in this study, while participant *B* only played a supervisory role to ensure that the proximity gesture was in line with 5 cm ±1 cm each time.

▷ **Experiment Results.** Figure 9(c) and (d) show the success rate of on-face and over-the-face gesture detection at different regions and in different motion state settings, respectively. We observe that in the stationary state, the success rate of both on-face and over-the-face gestures aligns closely with the heatmap results shown in Figure 8. This essentially reveals that the resting participant using their own hands did not significantly affect the MAF's performance. In the context of the two motion states under different speeds, we observe a decrease in the success rate compared to the stationary state. This is particularly apparent with the success rate of touching face gestures. For instance, the success rate of the right-facial gesture drops from an original 85% to 25% during walking and 20% while jogging as shown in Figure 9(c). This decrease can be attributed to a shift in the headphone's speaker caused by touch and body movement, leading to a corresponding shift in the transmission ultrasound signal. Similarly, the approaching face gesture in Figure 9(d) also shows a downward trend. However, despite a noticeable decrease in the success rate from the static state, the left-facial gestures consistently exhibit a robust success rate (>95%), thereby assuring good gesture detection under any motion state. This result manifests that the left-facial region can be fully relied upon for over-the-face interactions under any tested motion states.

4.3 Understanding MAF's social acceptance

Q3: Is MAF acceptable to mobile users? We create a Likert Scale questionnaire and a simple gesture test to evaluate the user acceptance of the MAF system.

Participants & Apparatus. We have conducted tests involving *N* = 22 participants, encompassing diverse skin conditions, ranging from 16 to 54 years old, including both males and females. The experimental setup aligns with the methodology employed in prior studies [78], where all participants are directed to interact with their only left facial regions—specifically by touching or positioning their hands in close proximity. We execute each 20 sets of trials of these two gestures under stable acoustic intensity and effective distance. It only takes one person five minutes to complete this set of experiments. To uphold the experimental rigor, the procedure is conducted under the watchful guidance of a designated supervisor. When each person completes these experiments, we show them their gesture signals diagrams which are similar in Figure 2(b) and (c), and explain the results to them.

▷ **Experiment Results.** The users' feedback is shown in Figure 10. First of all, we clearly see that users often use headphones in most of their daily lives to carry out various entertainment activities, such as

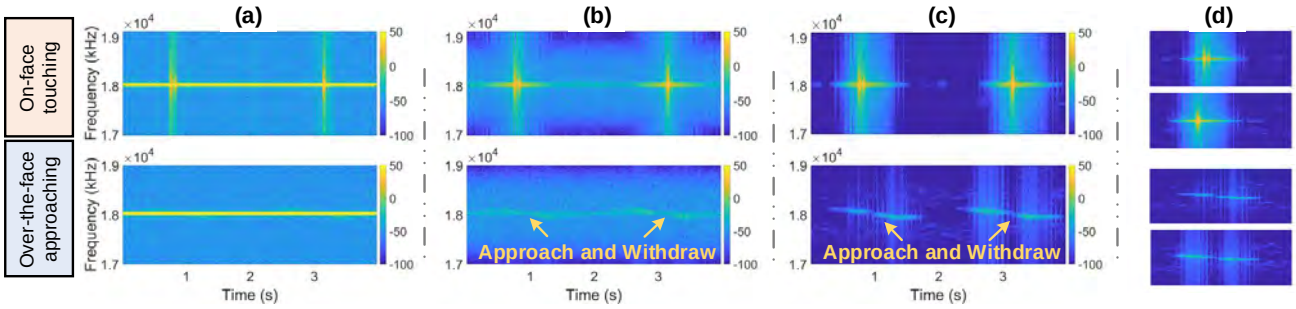

Figure 11: (a): the raw spectrogram of on-face and over-the-face gestures. (b): the spectrogram after applying narrow bandpass and bandstop filters. (c): the spectrogram after the signal enhancement. (d): the two signal segments after applying KL divergence-based signal segmentation.

listening to music and watching videos, which indicates that headphones occupy an important position in our daily lives. Secondly, in the initial stage of our straightforward experiment, we query participants about their prior experience with gesture recognition technology. A substantial 81.8% reveal they have never interacted with such technology, whereas a minor portion, 18.2%, have some experience, mainly through VR or AR platforms. Following this, we enable users to explore face-touching and face-approaching gesture recognition activities, later gathering their responses through a final Likert scale questionnaire, focusing particularly on questions (d) and (e). To our delight, an impressive 95% of participants describe the gesture recognition method as incredibly intuitive and simple to use. Furthermore, 91% find it is still comfortable to wear with only minimal attaching to the microphone's mouthpiece. Overall, users' recognition and expectations of the MAF system are positive.

5 SYSTEM DESIGN

Our system leverages the acoustic signal emitted from the bone conduction earphones to generate the mobile acoustic field. Although music signals can produce both surface acoustic waves and leaky surface acoustic waves, their frequency and amplitude both change abruptly over time, introducing variations to both the SAW and LSAW. It's thus challenging to disentangle the signal variation caused by human gestures from the raw signal receptions.

In MAF, we proactively send out a probing signal on the ultrasound band to produce stable surface acoustic waves and leaky surface acoustic waves. The probing signal works on the ultrasound band for three key reasons. Firstly, it allows mobile users to perform gesture control while listening to music without interfering with each other. Secondly, it is imperceptible to our human beings and thus will not negatively affect the user experience. Thirdly, compared to audible band signals, ultrasound at a higher frequency band attenuates more rapidly and thus is less prone to false alarms triggered by other users nearby. Moreover, it suffers less from ambient noises since most environmental noises are below 18kHz [9, 16]. We have measured the frequency response of three different pairs of bone conduction earphones and empirically set the central frequency of the probing signal to 18kHz. The user is free to use a higher frequency within the range of 18kHz to 22kHz to transmit probing signals if they can hear the lower frequency probing signal.

Single Tone vs. FMCW. We choose a single tone instead of the chirp signal (FMCW) for two reasons. Firstly, we find that the frequency response of most earphone speaker transducers in the ultrasound band varies significantly. This implies that the power of a chirp signal is not uniform across the frequency band. Given that the power fluctuation of the received signal is a crucial feature of our gesture recognition model, the inconsistency in chirp signal power has the potential to impact the performance of our model. Secondly, we find that sending continuous chirps will lead to hearable noises. This is because continuous chirp signals will trigger sharp impulse responses in the system, leading to the generation of transient signals that manifest as audible noise.

5.1 Signal Pre-processing

Figure 11 shows our proposed signal pre-processing pipeline. The raw signal received by the microphone first passes through a series of filters to extract the gesture-induced signals from the noise receptions. These processed signals are then fed into a signal enhancement module to improve their SNR.

Step One: Filtering. The received signal is first fed into a Butterworth bandpass filter with a cutoff frequency of f_{prob}±50Hz in order to remove the out-of-band noises. f_{prob} is the frequency of our probing signal on the ultrasound band. Subsequently, we pass the filtered signal through a Butterworth band-stop filter with the central frequency of f_{prob}. This allows us to remove the probing signal from the receptions while preserving the frequency variation caused by hand-to-face gestures, thereby enhancing its SNR. Figure 11(b) shows the received signal after passing the filters. Evidently, the signal variation due to facial gestures becomes more prominent after the filtering step.

Step Two: Signal Enhancement. To attain precise segmentations in MAF, it is essential to mitigate the effects of the in-band noise artifacts as well. One significant contributor to these artifacts is the probe signal, which generates multipath components [9] as it traverses different channels on the face, such as bones and fats, subsequently affecting the accurate detection of SAW and LSAW when the gesture commences (as illustrated in Figure 11(b)). Given that these multipath artifacts and the probing signal exhibit overlapping frequency components, prior band-pass filter (BPF) strategies are incapable of isolating the noise effectively. Thus, we leverage Wiener filtering [15] to manage such frequency-overlapped noise.

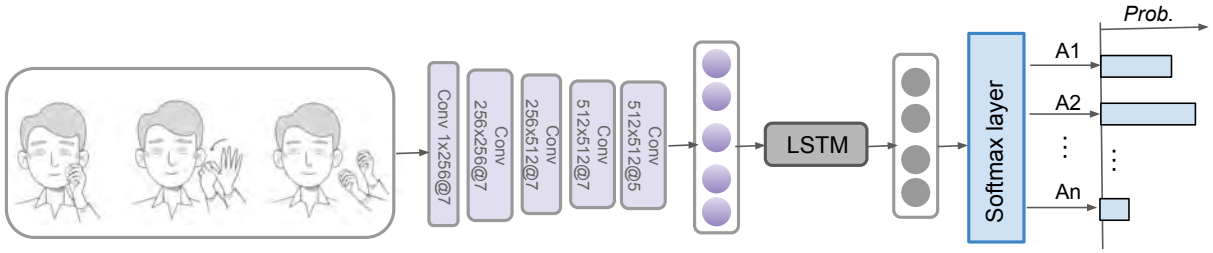

Figure 12: Model structure of MAF, it consists of five Convolution layers, a bi-directional LSTM layer, and an MLP.

Threshold	Precision	Recall	F1 score	Threshold	Precision	Recall	F1 score
0	0.550	1	0.762	0.55	0.993	0.700	0.821
0.05	0.870	1	0.930	0.6	1	0.615	0.762
0.1	0.948	1	0.973	0.65	1	0.560	0.718
0.15	**0.971**	**1**	**0.985**	0.7	1	0.550	0.710
0.2	0.969	0.950	0.960	0.75	1	0.550	0.710
0.25	0.978	0.905	0.940	0.8	1	0.520	0.684
0.3	0.978	0.880	0.926	0.85	1	0.495	0.662
0.35	0.982	0.820	0.894	0.9	1	0.455	0.625
0.4	0.981	0.780	0.869	0.95	1	0.435	0.606
0.45	0.980	0.735	0.840	1	1	0.435	0.606
0.5	0.993	0.730	0.842				

Table 1: Find the threshold for segmentation algorithm.

Specifically, we first collect a short segment of noise samples lasting 0.3 seconds. This step facilitates the analysis of the noise's frequency characteristics, thereby assisting in the accurate determination of the Wiener filter's parameters. Subsequently, these parameters are employed to filter out successive time frames sharing identical frequency characteristics. In MAF, the Wiener filter predominantly gathers the time frames that encompass the multipath occurring between the speaker and microphone pair when the probing signal initiates. As depicted in Figure 11(c), the application of the Wiener filter substantially reduces these multipaths when a gesture commences, thereby enhancing the discernibility of each gesture's initiation point and duration.

5.2 Segmentation

Next, we divide the received audio wave into a sequence of audio segments and feed those segments containing human gestures into the classifier for gesture recognition. To detect the presence of a gesture in a segment, an intuitive solution is to apply a pre-defined threshold to the audio wave to detect the energy variations caused by human gestures. However, this method is not scalable as it does not consider the fluctuations in signal energies resulting from diverse human behaviors, such as varying user strengths.

To tackle this challenge, we employ a Kullback-Leibler (KL) divergence-based method [13] to detect the presence of a gesture within each segment. Specifically, for any two consecutive audio segments, we compute the energy probability distribution of these two segments, denoted as P and Q, respectively. The KL divergence quantifies the information loss when Q approximates P:

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_i P(i) \log \left(\frac{P(i)}{Q(i)} \right)$$

When there is no gesture shows up, the two audio segments are full of ambient noises. Accordingly, their energy distribution would be similar. Accordingly, the KL divergence value would be close to 0. Conversely, when a gesture shows up, its energy distribution would differ from the audio segment containing purely ambient noise. Hence, we are expected to see a large KL divergence value.

To find a proper length of the audio segment, we assessed the duration of the gestures collected from 22 users across different ages (details can be found in §6.2.1) and found the longest gesture lasting around 2s. So we adopt a slightly larger segment size of 2.5s with a 50% overlap to ensure the completeness of the gesture within each audio segment. Figure 11(d) shows the on-face and over-the-face gestures after segmentation. Additionally, to find a proper threshold for KL divergence that effectively distinguishes significant probability distribution differences between P and Q in gesture detection, we collected 100 touch gestures and 100 proximity gestures and analyzed the detection precision and recall in different threshold settings. As shown in Table 1, a lower threshold increases detection sensitivity (a higher Recall). But the precision suffers as small signal fluctuations would be taken as a gesture. Conversely, a higher threshold risks missing subtle but important signal variations, leading to a lower recall but a higher precision. In MAF, we adopt 0.15 as our threshold, which strives for a balance between precision and recall, as shown in Table 1.

5.3 Gesture Classification

In the final stage, we aim to differentiate the specified gestures within MAF. Inspired by the success of the Deep Neural Network in the applications of image and audio fingerprinting classifications, we introduce a data-driven framework to identify these on-face and over-the-face gestures in MAF. The overall frame consists of two parts: feature extraction, and model training.

Feature extraction. MAF processes audio data for gesture classification using a short-time Fourier transform (STFT) spectrogram directly. Compared with the 1D time series signal, the 2D STFT spectrogram is considered to generally provide richer information on the feature representations [8] and has better temporal and frequency localization properties than a one-dimension waveform in the time-domain [14], making it a unique fit to the classify the nonstationary human gestures. Different from the prior works [86] applying a Mel spectrogram for classification, we apply the STFT spectrogram directly in MAF. The reason is that the non-linear Mel-scale emphasizes the fine-grained spectral structure in the lower frequency range, which is more important for speech recognition, but less critical for gesture detection in the high-frequency band

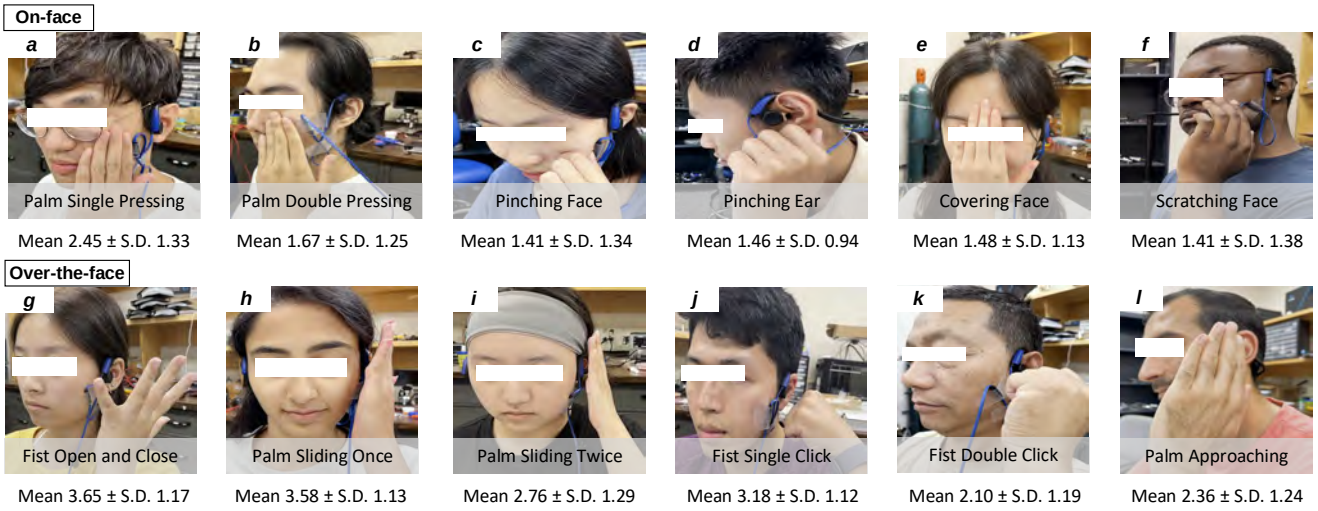

Figure 13: 12 gesture candidates for user evaluation. Among them, the first six (*a*, *b*, *c*, *d*, *e*, *f*) are on-face gestures whereas the last six (*g*, *h*, *i*, *j*, *k*, *l*) are over-face gestures. At the bottom of each gesture, we put the average score indicating user preference, with a higher score indicating greater user satisfaction with that particular gesture.

(>18kHz). Due to the acoustic signal being quasi-stationary within a short time (*e.g.*, 2–50ms) [10], we select the frame length of our spectrogram input to 2048, corresponding to 20ms within the sampling rate of 48,000 Hz. The hop length is set to 1024. Accordingly, the frequency resolution is around 23Hz within each sample point.

Model structure. MAF adopts a hybrid neural network architecture to enhance the classification performance, as depicted in Figure 12. Specifically, we employ a combination of Convolutional Neural Networks (CNNs) and a Recurrent Neural Network (RNN) layer to facilitate superior feature extraction before inputting the spectrum feature representations into the multilayer perceptron (MLP) for classification [52]. The architecture encompasses five CNN encoder layers, a bi-directional LSTM layer, and a classic multilayer perception (MLP) structure. Each CNN layer is configured with a 2D convolution, a batch norm, a ReLU function, and a dropout regularization. The stride is set to 2.

Given that the gesture representation usually spans across a frequency band of 150Hz, we have designed the kernel size of the initial two convolution layers to be 7×7. This decision ensures that the receptive field is adequately sized to encapsulate a complete gesture component within the spectrogram, thus enhancing the feature extraction efficacy. Subsequently, the high-dimensional features extracted are forwarded to the LSTM layer, which enhances the temporal connections between individual time frames. This LSTM layer acts as a bridge, conveying the refined feature set to the MLP. The MLP processes the features received from the LSTM and outputs the prediction results. To compute the loss, we utilize the cross-entropy loss function.

6 EVALUATION

6.1 Study One: Gesture Selection

Building upon prior research on hand-to-face interactions [45, 85, 86, 88], we devise a set of 12 distinct gestures. Among these, six are performed directly on the face, while the remaining six are performed in proximity to the face (*a.k.a.*, over-the-face gestures). Figure 13 illustrates these 12 gestures. Subsequently, we invite participants to rate each gesture, indicating their personal preferences. Our objective is to assess whether the gesture set we are crafting is aligned with user preferences and intuitive behavior.

Participants & Apparatus. The identical group of 22 participants who participate in our previous user study Section (§4.3) are enlisted for this particular experiment. To be precise, each participant is asked to wear the bone conduction earphones and execute the set of 12 gestures illustrated in Figure 13. Upon the completion of the gesture performance, each participant is requested to complete a Likert scale questionnaire. In this questionnaire, they are prompted to rank these 12 gestures (on a scale from 1 to 5, the higher the better) based on their individual preferences. Notice that the volunteers #6, #10, and #20 have noticeable facial hair. So we tape the microphone on top of their facial hair.

Results. In Figure 13 representing each gesture, we mark the average score associated with each gesture at the bottom. From the results, we observe that the participants generally prefer over-the-face gestures to on-face gestures. On the other hand, comparing gestures *h* and *i*, as well as *j* and *k*, we can also see that users prefer to perform gestures only once, indicating that users prefer simple and efficient gestures. Notably, gestures *c* and *f* receive the lowest scores, averaging around 1.41. Participants' feedback highlights that they find these two types of gestures uncomfortable as they involve pinching the face, potentially causing discomfort or even pain. Furthermore, they indicate that performing these gestures in public could be viewed as inappropriate or uncouth. Based on these user study results, we finally decide to keep the palm single and double pressing, pinching ear and covering the face for on-face gestures (*a*, *b*, *d*, *e*) and including all of the over-face gestures: fist open and close, the palm sliding once or twice, the fist single and double click, and one palm approaching (*g*, *h*, *i*, *j*, *k*, *l*).

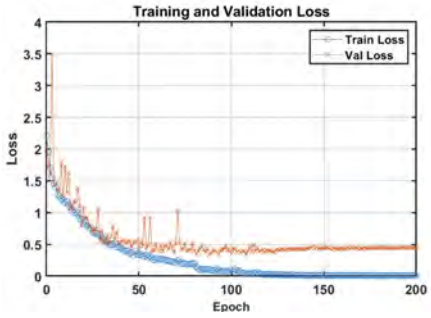

Figure 14: Training loss and validation loss curve over different training epochs.

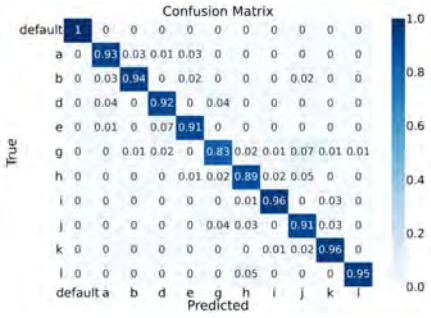

Figure 15: The confusion matrix of gesture recognition.

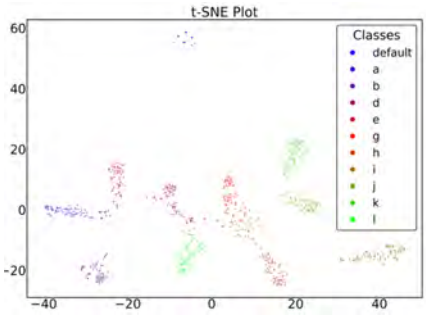

Figure 16: The feature distribution of ten different gestures.

6.2 Study Two: Gesture Recognition

In this section, we first describe the experiment setups (§6.2.1) and then discuss the experiment results (§6.2.2).

6.2.1 ***Experiment Setups.*** The same group of 22 participants was recruited for our field study. Within this group, there were 12 male participants and 10 female participants, with an average age of 26.8 years and a standard deviation of 9.0.

Data collection. All participants used the same pair of bone conduction earphones during the data collection process. To ensure consistency, the microphone's mouthpiece was taped approximately three fingers' width away from the participant's earlobe, on their cheek.[3] Following this, a supervisor connected the earphones to a laptop and emitted an 18kHz frequency signal at a consistent volume (45 dB SPL), ensuring that the ultrasound signal remained inaudible to the participants. The microphone recordings from the earphones were captured using a Matlab program. Simultaneously, the supervisor recorded video footage of the participants performing gestures to establish the ground truth. Each participant was instructed to execute each type of gesture 20 times, resulting in a total time commitment of approximately 15 minutes. In total, we collected 4,400 gesture recordings.

Model Training and Evaluation. Our CRNN model is implemented in Pytorch and trained on an NVIDIA A100 GPU for 150 epochs, using a batch size of 32. We employ the Adam optimizer with a learning rate set to 0.001. Then we adopt a Leave-One-Out approach, specifically a 5-fold cross-validation approach, to evaluate our CRNN model. In this setup, we divide the data from 22 participants into 5 groups, with each group containing data from 4 or 5 participants. During each iteration of cross-validation, one group is left out as the test set, while the combined data from the remaining groups are used as the training set. This ensures that each group has the opportunity to serve as an independent test set, and also accounts for potential correlations between participants' data, providing a more comprehensive assessment of the model's generalization capabilities.

Evaluation Metrics. We adopt gesture recognition accuracy as the metric to evaluate the performance of our proposed solution.

[3]The microphone is taped on the participant's face because the bone conduction earphone being used in our experiment adopts an inline microphone.

The recognition accuracy is formally defined as:

$$\text{Recognition Accuracy} = \frac{\text{\# of correctly recognized gestures}}{\text{\# of gestures being tested}} \quad (2)$$

Additionally, we also use precision, recall, F1 score, and accuracy to assess the model's performance.

Prevent Overfitting. We take the following actions to prevent model overfitting. Firstly, we adopt leave-one-out cross-validation to ensure the testing is performed on unseen data (i.e., collected from other users). Secondly, our model adopts L2 regularization to penalize large weights, which helps prevent the model from fitting the training data too closely. Thirdly, to mitigate overfitting, we have implemented early stopping and dropout layers. Figure 14 shows the training and validation loss curves both trend downwards consistently. The training loss gradually decreases upon adding training examples and flattens gradually. The validation loss decreases upon adding training examples and flattens gradually. We notice that there is a gap between the training loss and validation loss after 50 epochs, indicating addition of more training examples doesn't improve the model performance on unseen data. Hence our model did not overfit.

6.2.2 ***Experiment Results.*** In this section, we report the evaluation results based on the data collected.

① **Gesture Recognition Accuracy Across Ten Types of Gestures.** Figure 15 shows the confusion matrix illustrating the results for the ten tested gestures. Among them, gestures *a*, *b*, *d*, and *e* are detected using SAW signals, while the remaining six gestures (*g*, *h*, *i*, *j*, *k*, *l*) are recognized using LSAW signals. Additionally, we have included a *default* gesture to serve as a reference, indicating no specific gesture being performed.

We have four observations. Firstly, the model can successfully differentiate between signals that contain gestures and those contain no gestures. Secondly, the recognition accuracy for most of these gestures surpasses 92%, demonstrating the efficacy of our proposed signal processing algorithms. Thirdly, LSAW signals outperform SAW signals in recognizing facial gestures. This observation suggests that the MAF system demonstrates increased sensitivity when detecting expansive movements or gestures that span a wider spatial area. Fourthly, we observe a notable challenge in accurately classifying the "fist open and close" gesture (*g*), with a recognition accuracy of merely 83%. To understand the reason for this inferior performance, we further visualized the feature distribution of these

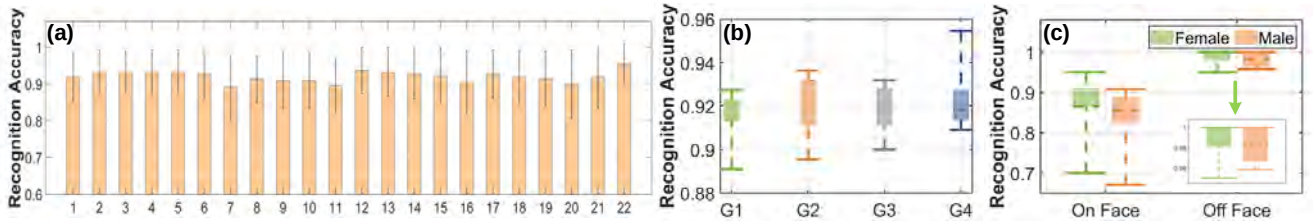

Figure 17: Examine the gesture recognition accuracy. (a): the gesture recognition accuracy across 22 participants. (b): the gesture recognition accuracy across four age groups. (c): the gesture recognition accuracy across different genders.

Model	Precision	Recall	F1 score	Accuracy
SVM	0.550	0.547	0.550	0.547
Random Decision Tree	0.555	0.541	0.543	0.541
K-Nearest Neighbors	0.565	0.559	0.545	0.559
RNN only	0.613	0.613	0.613	0.565
CNN only	0.823	0.793	0.808	0.817
CRNN (ours)	**0.949**	**0.949**	**0.949**	**0.928**

Table 2: Test results of different models, and the final weighted values of Precision, Recall, F1 score, and Accuracy.

ten gestures in Figure 16. As shown, the features of the gesture "fist open and close" (g) overlap with the feature of the gesture "fist single click" (j). This confusion stems from the similarity in the hand movement involved in both gestures.

② **Overall Performance Across Different Users.** We first examine the gesture recognition accuracy across all 22 participants. To do this, we have calculated the mean gesture recognition accuracy for each individual based on their performance across 10 different gestures and repeated each gesture twenty times to ensure the reliability of the results. The results are shown in Figure 17(a). We observe that MAF consistently demonstrates strong performance, consistently exceeding 91% accuracy across all 22 participants. Notably, only participants #7 and #11 achieve a slightly lower accuracy of approximately 89%. On the whole, these results reaffirm the overall effectiveness of MAF.

③ **The Impact of Different Ages.** Prior works have shown that people of different ages tend to perform gestures differently [94, 95]. For instance, a senior may perform a gesture slightly slower than a junior. Moreover, the difference in palm size and finger length across different age groups can lead to distinct effects on the received signal. Furthermore, people of different age brackets may exhibit distinct levels of facial skin moisture, which could impact the propagation of both SAW and LSAW signals. Hence we conduct experiments to examine the impact of different ages on our system. In this experimental, we classify the 22 participants into four age groups: Group 1 (under 22 years old), Group 2 (between 22 and 25 years old), Group 3 (between 26 and 30 years old), and Group 4 (above 30 years old), as illustrated in Figure 17(b). The results reveal a relatively consistent gesture recognition accuracy across these four groups, averaging 92%. However, it's worth noting that the variance in gesture recognition accuracy is slightly higher in Groups 1, 2, and 4 compared to Group 3. The superior accuracy observed in Group 3 may be attributed to the stability and consistency demonstrated by participants in this age bracket in terms of ease of gesture manipulation performance, physical coordination, and even skin moisture conditions.

④ **The Impact of Different Genders.** Likewise, there will be inherent differences in how men and women naturally execute gestures as people in different genders differ in their palm size, finger length, and the strength as well as the speed when performing a gesture [95]. This could lead to variations in the received SAW and LSAW signals. Consequently, we conduct an analysis of gesture recognition performance across different genders. Figure 17(c) shows the accuracy of gesture recognition for both on-face and over-the-face gestures. Across both genders, we observe consistently high performance for gestures made away from the face (over-the-face), averaging 98%. However, regarding the recognition of on-face gestures, there is a noticeable disparity: females exhibit an accuracy rate of 88%, compared to males who achieve an accuracy of 84%. Moreover, there is a notably higher variance in the accuracy of on-face gesture recognition across both genders. While differences in gesture intensity might contribute to this variation, we also suspect that the smaller amount of training data available for on-face gestures (in comparison to eight types of over-face gestures) could be another contributing factor.

⑤ **Comparison Between Different Models**. We also compare our CRNN model against a few other classification methods such as random decision trees, K Nearest Neighbor (kNN), Support Vector Machine (SVM), and basic deep learning models consisting solely of CNN or RNN components. All methods have undergone identical signal pre-processing procedures, including filtering, enhancement, and segmentation. Furthermore, we employ identical training and testing datasets for all of these models to ensure a fair comparison.

The results are shown in Table 2. Our CRNN model exhibits remarkable superiority over traditional classifiers, achieving an accuracy of 92.8% compared to 54.7% for SVM, 54.1% for Random Decision Trees, and 55.9% for K-Nearest Neighbors. The gesture recognition accuracy of the RNN model experiences only a slight increase to 56.5%. This limited improvement may be attributed to the RNN's predominant focus on temporal data, potentially overlooking crucial spatial characteristics inherent in gestures. The utilization of the CNN model results in a significant leap in gesture recognition accuracy to 82%. This improvement can be attributed to CNN's robust capability to extract essential spatial features from the spectrogram data. When we combine the strengths of both CNN and RNN in our hybrid RCNN model, the gesture recognition accuracy peaks at 92.8%, surpassing all baseline models significantly.

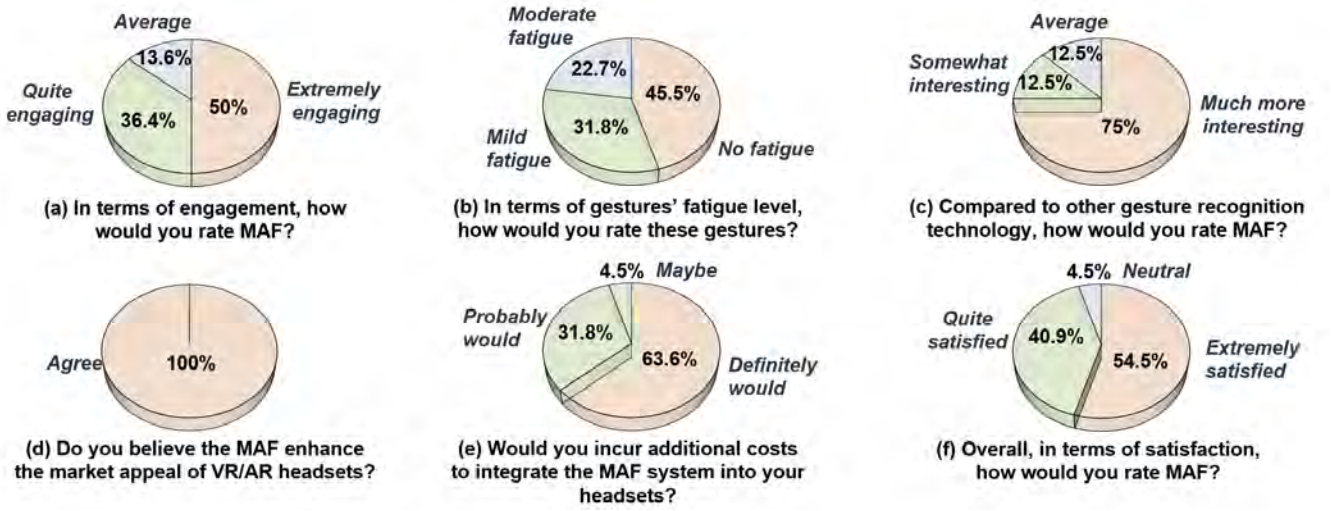

Figure 18: Feedback from 22 users was gathered through a Likert scale questionnaire after using MAF.

⑥ **Subjective Evaluation**. We have administered a second Likert scale survey to each individual after using MAF. The goal is to evaluate their perspectives on the usability of MAF. The results are shown in Figure 18. Encouragingly, a substantial 86.4% finds the MAF approach engaging, with no one stating that it wasn't. Furthermore, we find that 17 participants considered the difficulty of executing the 10 gestures as acceptable. As a follow-up step, from the results in the first survey of Figure 10(c), only 8 users have previously encountered the conveniences afforded by gesture recognition technology, and 6 of them found our MAF technology to be more appealing. Moreover, all 22 interviewees believe that MAF could enhance the market allure of VR/AR headsets, with 21 indicating a willingness to purchase the service should it become available on the market. Ultimately, a positive sentiment towards the overall satisfaction with MAF is expressed by 21/22 participants.

6.3 Study Four: Micro-Benchmarks

We also conduct micro-benchmarks to understand the impact of various factors on system performance.

① **The Impact of Body Motions on Segmentation.** We next evaluate whether body motions can fool the segmentation algorithm. A volunteer was asked to follow the instructions to perform gestures: starting with the user sitting down, followed by touching his face, eating, approaching the face, and speaking, each for 2.5 seconds. The volunteer repeated the above process 100 times. Table 3 shows the result. The touching gestures and approaching gestures can be successfully detected at a rate of 100% and 96%, respectively. For body motions, we found sitting down and eating are rarely recognized as the presence of gestures. However, we noticed that human talking are easily recognized as the presence of gestures, with a rate of 82%. These false positives will be inherently sent to the classification model for gesture recognition. Recall that our classification model was trained with both gestures and non-gestures. So it will filter out these body motions during classification (*i.e.*, putting them into the default non-gesture class).

Action	Identified as a Gesture
Sitting	4%
Touching	100%
Eating	11%
Approaching	96%
Talking	82%

Table 3: A volunteer is asked to execute two gestures (touching and approaching) and three distinct body motions (sitting down, eating, and talking), each repeated 100 times. We feed the corresponding signals into the segmentation algorithm.

② **The Impact of Environmental Noise**. To evaluate the robustness of our system in environments with varying noise levels, we simulate different noise levels typically encountered in daily human activities as referenced in [32]. During these experiments, a participant is instructed to perform each of the 10 gestures 20 times, amidst white noise emitted from a speaker at different volumes: 40 dB SPL, 60 dB SPL, and 80 dB SPL, as illustrated in Figure 19(a). Utilizing our pre-trained model, we then assess the impact of these noise levels on gesture recognition accuracy. Our analysis indicates that increasing noise levels could compromise the system's ability to accurately recognize gestures. This could be attributed to the interference created by noise reflections over the face, affecting the formation of the MAF that could obstruct precise gesture recognition. However, it is encouraging to note that our system maintains stability at a noise level of 40 dB SPL, proving to be highly reliable for indoor activities such as VR gaming. This underscores that MAF can accommodate the majority of daily activities.

③ **The Impact of Earphone Remounting.** We conduct experiments to evaluate the impact of earphone remounting on system performance. Initially, under the supervision of a researcher, a participant places the earphone's microphone at a designated position

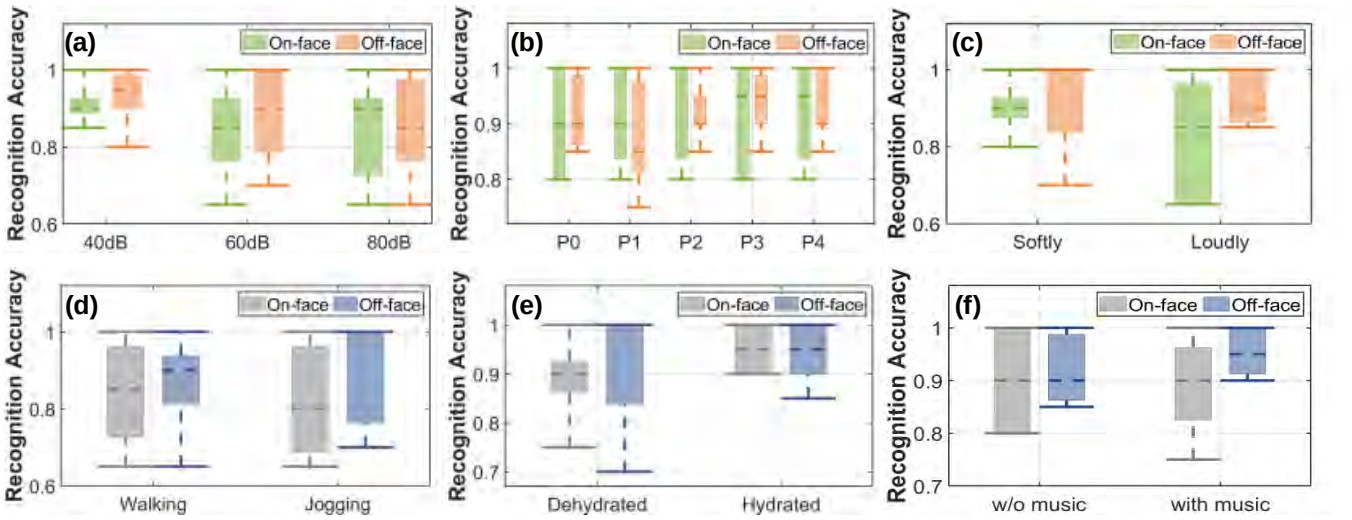

Figure 19: Examine the gesture recognition accuracy in different environments and human factor settings. (a): the gesture recognition accuracy in different ambient noise environment settings.(b): the gesture recognition accuracy at different remounting positions. (c): the gesture recognition accuracy in the presence of human speech. (d): the gesture recognition accuracy in different levels of human activities. (e): the gesture recognition accuracy in different skin hydration settings. (f): the gesture recognition accuracy in the absence and presence of music playback.

on her left cheek, labeled as position P0. Subsequently, the participant independently remounts the microphone and performs each gesture 20 times, respectively. The participant repeats the above process four times. We denote these remounted positions as P1, P2, P3, and P4, respectively. Figure 19(b) shows the results. We observe that at the first repositioning (P1), there is a slight decrease in segmentation accuracy. This performance drop is attributed to the participant's unfamiliarity with the microphone-taping process, which results in suboptimal microphone placement affecting the detection of off-face gestures. When the participant gets familiar with the microphone mounting process, there is a noticeable improvement in segmentation accuracy in the subsequent three repositioning attempts (*e.g.*, P2, P3, and P4). This improvement indicates that the participant becomes increasingly adept at using the MAF system, leading to more effective microphone placements.

④ **The Impact of Human Speech**. To assess the impact of verbal communication on our system's performance, we have devised an experiment where participants are required to perform gestures while concurrently having a conversation. In this trial, each participant is tasked with repeating each gesture 20 times, and the collected data is classified using the same pre-trained model. As illustrated in Figure 19(c), a noticeable discrepancy in accuracy is observable when comparing on-face gesture recognition of loud speech to soft speech. This phenomenon can be attributed to the significant movement in the cheek area that occurs during loud speech, which interferes with the signal propagation of gesture transmission. Nevertheless, the median results indicate that the recognition accuracy remains approximately 90%, regardless of the volume of speech. This stability is largely due to our signal preprocessing filtering phase, where the frequencies commonly associated with verbal communication and their harmonics are effectively eliminated, ensuring the integrity of gesture recognition.

⑤ **The Impact of Body Motions**. We evaluate the impact of body motions on gesture recognition accuracy. Specifically, we invite a participant to perform 10 gestures under the walking state and jogging state. Each gesture is repeated 20 times. The same pre-trained model is used to recognize gestures collected in these two states. Figure 19(d) shows the results. We find that under different body movements, we can still recognize 10 different gesture signals, but the median recognition accuracy drops to below 90% for walking and performs worse in jogging. This also reaffirms the results of Section (§4.2), indicating that the motion state affects our posture recognition rate. The reason is that the recognition can be influenced by the movement of the head and the tightness of the headphone wear, affecting the formation of the MAF. However, the above 80% recognition accuracy is still acceptable, showcasing significant accuracy despite the challenges presented by physical movements. Future improvement of MAF might focus on implementing advanced algorithms capable of compensating for the disturbances generated by these physical activities.

⑥ **The Impact of Skin Hydration**. The circadian rhythm notably influences the skin's water permeability, the hydration level of facial muscles, and the concentration and dispersion of oils in the stratum corneum in individuals [55]. Since these factors can vary at different points in the day, we design an experiment to assess whether these changes affect the accuracy of gesture recognition signals. We experiment 20 times with 10 gestures in the morning and evening with the same participant. We label the microphone's position on the user's face with a marker to ensure precise remounting at the same location. From the results of Figure 19(e), the pre-trained model struggles to perform well when participants have oily skin after a day's activities. This is largely due to the natural oil buildup on the facial skin over the course of the day, which can disrupt the accuracy of the gesture recognition process.

⑦ **The Impact of Music Playback**. We then conduct an experiment to assess if music playback affects gesture recognition. The participant wears the headphones while a computer played a song mixed with our 18kHz probing signal. We evaluate the recognition accuracy of ten gestures performed by the participant. We also plot the gesture recognition accuracy in the absence of music playback for comparison. As shown in Figure 19(f), our system achieves above 90% accuracy in the presence of music playback, which is consistent with that achieved in the absence of music playback. This is because the frequency of music signals is usually below 15kHz, and thus will not interfere with our probing signals. These music signals will be further filtered out during our signal processing. These results affirm that our system supports gesture recognition without compromising the headset's music playback function.

7 LIMITATIONS

Restricted to Bone Conduction Headphones. We experiment with different types of headphones, including on-ear, over-ear, in-ear, and bone-conduction. However, the SAW and LSAW waves only show up when the user wears a pair of bone conduction headphones. We suspect that insufficient skin contact or a small contact surface of on-ear, over-ear, and in-ear headphones are the primary reason. Additionally, the on-ear and over-ear headphones are usually equipped with soft earcups or ear pads that can absorb acoustic energy, further diminishing the generation of these SAW and LSAW waves.

Taping to the user's head. Our current prototype requires the user to tape the microphone to her head for signal reception, which may not be practical for everyday use. However, we noticed numerous bone conduction earphones have built-in microphones. These microphones have natural contact with human face, making it possible to receive SAW and LSAW signals without taping. As these built-in microphones can come into contact with various positions on the face, the impact of microphone placement is worth further exploration.

Extend to Micro-Gestures. The proposed CRNN model can effectively identify and further categorize ten distinct gestures performed on and above the face. Nevertheless, it encounters challenges in accurately recognizing micro-gestures involving fine-grained finger motions, such as scratching the face, pinching the ear, and sliding the face with two fingers, owing to its constrained model capacity. One potential remedy is to add more layers to the current CRNN model to augment its capacity. However, this may increase the model inference latency, adversely impacting the user experience. The tradeoff between model capacity and latency is worth further exploration.

8 CONCLUSION

We have presented the design, implementation, and evaluation of mobile acoustic field (MAF), a novel acoustic sensing approach that leverages commodity hardware in bone conduction earphones for hand-to-face gesture interactions. This new approach hinges on the principles of Surface Acoustic Waves (SAW) and Leaky Surface Acoustic Waves (LSAW) to create signals that not only traverse the user's facial surface but also radiate into the surrounding air, forming an encompassing acoustic field surrounding the individual's head. Our evaluation involving 22 participants demonstrated MAF can accurately recognize ten distinct gestures performed both on face and above face with high precision. The user feedback is also promising: an overwhelming majority of our 22 interviewees showcased enthusiasm toward adopting the MAF, anticipating its seamless integration into contemporary scenarios. We envision that MAF stands as a significant approach in the field of hand-to-face gesture recognition technology.

ACKNOWLEDGMENTS

This work is supported by the National Science Foundation under Grant No.2337537. We also thank University of Pittsburgh Center for Research Computing (CRC) for providing us the GPU resource.

REFERENCES

[1] [n. d.]. https://store.immutouch.com/collections/all
[2] [n. d.]. GZCRDZ bone conduction headphone. https://www.amazon.in/GZCRDZ-Conduction-Headphones-Microphone-Headphone/dp/B07QVPGPZ5
[3] [n. d.]. Larcenauts. https://www.larcenauts.com/
[4] Takashi Amesaka, Hiroki Watanabe, and Masanori Sugimoto. 2019. Facial Expression Recognition Using Ear Canal Transfer Function. In *Proceedings of the 2019 ACM International Symposium on Wearable Computers* (London, United Kingdom) *(ISWC '19)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3341163.3347747
[5] Andreas Braun, Stefan Krepp, and Arjan Kuijper. 2015. Acoustic Tracking of Hand Activities on Surfaces. In *Proceedings of the 2nd International Workshop on Sensor-Based Activity Recognition and Interaction* (Rostock, Germany) *(iWOAR '15)*. Association for Computing Machinery, New York, NY, USA, Article 9, 5 pages. https://doi.org/10.1145/2790044.2790052
[6] Alex Butler, Shahram Izadi, and Steve Hodges. 2008. SideSight: Multi-"touch" Interaction around Small Devices. In *Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology* (Monterey, CA, USA) *(UIST '08)*. Association for Computing Machinery, New York, NY, USA, 201–204. https://doi.org/10.1145/1449715.1449746
[7] Colin Campbell. 1989. 1 - Introduction. In *Surface Acoustic Wave Devices and their Signal Processing Applications*, Colin Campbell (Ed.). Academic Press, 1–7. https://doi.org/10.1016/B978-0-12-157345-4.50005-9
[8] Qun Chao, HaoHan Gao, JianFeng Tao, YuanHang Wang, Jian Zhou, and ChengLiang Liu. 2022. Adaptive decision-level fusion strategy for the fault diagnosis of axial piston pumps using multiple channels of vibration signals. *Science China Technological Sciences* 65, 2 (2022), 470–480.
[9] Tao Chen, Longfei Shangguan, Zhenjiang Li, and Kyle Jamieson. 2020. Metamorph: Injecting inaudible commands into over-the-air voice controlled systems. In *Network and Distributed Systems Security (NDSS) Symposium*.
[10] Tao Chen, Longfei Shangguan, Zhenjiang Li, and Kyle Jamieson. 2023. The Design and Implementation of a Steganographic Communication System over In-Band Acoustical Channels. *ACM Transactions on Sensor Networks* (2023).
[11] Xiang 'Anthony' Chen. 2020. FaceOff: Detecting Face Touching with a Wrist-Worn Accelerometer. *ArXiv* abs/2008.01769 (2020).
[12] Yu-Chun Chen, Chia-Ying Liao, Shuo-wen Hsu, Da-Yuan Huang, and Bing-Yu Chen. 2020. Exploring User Defined Gestures for Ear-Based Interactions. *Proc. ACM Hum.-Comput. Interact.* 4, ISS, Article 186 (nov 2020), 20 pages. https://doi.org/10.1145/3427314
[13] Han Ding, Longfei Shangguan, Zheng Yang, Jinsong Han, Zimu Zhou, Panlong Yang, Wei Xi, and Jizhong Zhao. 2015. FEMO: A Platform for Free-Weight Exercise Monitoring with RFIDs. In *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems* (Seoul, South Korea) *(SenSys '15)*. Association for Computing Machinery, New York, NY, USA, 141–154. https://doi.org/10.1145/2809695.2809708
[14] A. Djebbari and F. Bereksi Reguig. 2000. Short-time Fourier transform analysis of the phonocardiogram signal. In *ICECS 2000. 7th IEEE International Conference on Electronics, Circuits and Systems (Cat. No.00EX445)*, Vol. 2. 844–847 vol.2. https://doi.org/10.1109/ICECS.2000.913008
[15] Matthias Doerbecker and Stefan Ernst. 1996. Combination of two-channel spectral subtraction and adaptive Wiener post-filtering for noise reduction and dereverberation. In *1996 8th European Signal Processing Conference (EUSIPCO 1996)*. IEEE.
[16] Tingchao Fan, Huangwei Wu, Meng Jin, Tao Chen, Longfei Shangguan, Xinbing Wang, and Chenghu Zhou. 2023. Towards Spatial Selection Transmission for Low-end IoT devices with SpotSound. (2023).

[17] Xiaoran Fan, Daewon Lee, Larry Jackel, Richard Howard, Daniel Lee, and Volkan Isler. 2022. Enabling Low-Cost Full Surface Tactile Skin for Human Robot Interaction. *IEEE Robotics and Automation Letters* 7, 2 (2022), 1800–1807. https://doi.org/10.1109/LRA.2022.3142433

[18] Xiaoran Fan, Longfei Shangguan, Siddharth Rupavatharam, Yanyong Zhang, Jie Xiong, Yunfei Ma, and Richard Howard. 2021. HeadFi: Bringing Intelligence to All Headphones. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking* (New Orleans, Louisiana) *(MobiCom '21)*. Association for Computing Machinery, New York, NY, USA, 147–159. https://doi.org/10.1145/3447993.3448624

[19] Xiaoran Fan, Riley Simmons-Edler, Daewon Lee, Larry Jackel, Richard Howard, and Daniel Lee. 2021. AuraSense: Robot Collision Avoidance by Full Surface Proximity Detection. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 1763–1770. https://doi.org/10.1109/IROS51168.2021.9635919

[20] Yang Gao, Wei Wang, Vir V. Phoha, Wei Sun, and Zhanpeng Jin. 2019. EarEcho: Using Ear Canal Echo for Wearable Authentication. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3, Article 81 (sep 2019), 24 pages. https://doi.org/10.1145/3351239

[21] Mayank Goel, Brendan Lee, Md. Tanvir Islam Aumi, Shwetak Patel, Gaetano Borriello, Stacie Hibino, and Bo Begole. 2014. SurfaceLink: Using Inertial and Acoustic Sensing to Enable Multi-Device Interaction on a Surface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 1387–1396. https://doi.org/10.1145/2556288.2557120

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

[23] Tobias Grosse-Puppendahl, Christian Holz, Gabe Cohn, Raphael Wimmer, Oskar Bechtold, Steve Hodges, Matthew S. Reynolds, and Joshua R. Smith. 2017. Finding Common Ground: A Survey of Capacitive Sensing in Human-Computer Interaction. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3293–3315. https://doi.org/10.1145/3025453.3025808

[24] Changzhan Gu and Jaime Lien. 2017. A Two-Tone Radar Sensor for Concurrent Detection of Absolute Distance and Relative Movement for Gesture Sensing. *IEEE Sensors Letters* 1, 3 (2017), 1–4. https://doi.org/10.1109/LSENS.2017.2696520

[25] 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. https://doi.org/10.1145/2207676.2208331

[26] Sean Gustafson, Daniel Bierwirth, and Patrick Baudisch. 2010. Imaginary Interfaces: Spatial Interaction with Empty Hands and without Visual Feedback. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology (UIST '10)*. Association for Computing Machinery, New York, NY, USA, 3–12. https://doi.org/10.1145/1866029.1866033

[27] Sean Gustafson, Christian Holz, and Patrick Baudisch. 2011. Imaginary Phone: Learning Imaginary Interfaces by Transferring Spatial Memory from a Familiar Device. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology* (Santa Barbara, California, USA) *(UIST '11)*. Association for Computing Machinery, New York, NY, USA, 283–292. https://doi.org/10.1145/2047196.2047233

[28] Chris Harrison and Scott E. Hudson. 2008. Scratch Input: Creating Large, Inexpensive, Unpowered and Mobile Finger Input Surfaces. In *Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology* (Monterey, CA, USA) *(UIST '08)*. Association for Computing Machinery, New York, NY, USA, 205–208. https://doi.org/10.1145/1449715.1449747

[29] Chris Harrison, Julia Schwarz, and Scott E. Hudson. 2011. TapSense: Enhancing Finger Interaction on Touch Surfaces. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology* (Santa Barbara, California, USA) *(UIST '11)*. Association for Computing Machinery, New York, NY, USA, 627–636. https://doi.org/10.1145/2047196.2047279

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

[31] Yan He, Hanyan Zhang, Edwin Yang, and Song Fang. 2020. Virtual Step PIN Pad: Towards Foot-input Authentication Using Geophones. In *2020 IEEE 17th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*. 649–657. https://doi.org/10.1109/MASS50613.2020.00084

[32] Institute for Quality and Efficiency in Health Care (IQWiG). 2008. Hearing loss and deafness: Normal hearing and impaired hearing. https://www.ncbi.nlm.nih.gov/books/NBK390300/

[33] Yasha Iravantchi, Yi Zhao, Kenrick Kin, and Alanson P. Sample. 2023. SAWSense: Using Surface Acoustic Waves for Surface-Bound Event Recognition. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 422, 18 pages. https://doi.org/10.1145/3544548.3580991

[34] Hiroshi Ishii, Craig Wisneski, Julian Orbanes, Ben Chun, and Joe Paradiso. 1999. PingPongPlus: Design of an Athletic-Tangible Interface for Computer-Supported Cooperative Play. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania, USA) *(CHI '99)*. Association for Computing Machinery, New York, NY, USA, 394–401. https://doi.org/10.1145/302979.303115

[35] Naoya Isoyama, Tsutomu Terada, and Masahiko Tsukamoto. 2014. An Interactive System for Recognizing User Actions on a Surface Using Accelerometers. In *Proceedings of the 5th Augmented Human International Conference* (Kobe, Japan) *(AH '14)*. Association for Computing Machinery, New York, NY, USA, Article 57, 2 pages. https://doi.org/10.1145/2582051.2582108

[36] Yincheng Jin, Yang Gao, Yanjun Zhu, Wei Wang, Jiyang Li, Seokmin Choi, Zhangyu Li, Jagmohan Chauhan, Anind K. Dey, and Zhanpeng Jin. 2021. SonicASL: An Acoustic-Based Sign Language Gesture Recognizer Using Earphones. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 2, Article 67 (jun 2021), 30 pages. https://doi.org/10.1145/3463519

[37] Vimal Kakaraparthi, Qijia Shao, Charles J Carver, Tien Pham, Nam Bui, Phuc Nguyen, Xia Zhou, and Tam Vu. 2021. FaceSense: sensing face touch with an ear-worn system. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–27.

[38] Naoaki Kashiwagi, Yuta Sugiura, Natsuki Miyata, Mitsunori Tada, Maki Sugimoto, and Hideo Saito. 2017. Measuring Grasp Posture Using an Embedded Camera. In *2017 IEEE Winter Applications of Computer Vision Workshops (WACVW)*. 42–47. https://doi.org/10.1109/WACVW.2017.14

[39] Wolf Kienzle, Eric Whitmire, Chris Rittaler, and Hrvoje Benko. 2021. ElectroRing: Subtle Pinch and Touch Detection with a Ring. 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 3, 12 pages. https://doi.org/10.1145/3411764.3445094

[40] Hyosu Kim, Anish Byanjankar, Yunxin Liu, Yuanchao Shu, and Insik Shin. 2018. UbiTap: Leveraging Acoustic Dispersion for Ubiquitous Touch Interface on Solid Surfaces. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems* (Shenzhen, China) *(SenSys '18)*. Association for Computing Machinery, New York, NY, USA, 211–223. https://doi.org/10.1145/3274783.3274848

[41] Fan-Jie Kung and Mingsian R Bai. 2020. Estimation of the Noise and Reverberation Covariance Matrices with Application in Speech Enhancement using the Multichannel Wiener Filter. In *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*. Institute of Noise Control Engineering.

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

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

[44] Changzhi Li, Zhengyu Peng, Tien-Yu Huang, Tenglong Fan, Fu-Kang Wang, Tzyy-Sheng Horng, José-María Muñoz-Ferreras, Roberto Gómez-García, Lixin Ran, and Jenshan Lin. 2017. A Review on Recent Progress of Portable Short-Range Noncontact Microwave Radar Systems. *IEEE Transactions on Microwave Theory and Techniques* 65, 5 (2017), 1692–1706. https://doi.org/10.1109/TMTT.2017.2650911

[45] Wei Li, Ping Shi, and Hongliu Yu. 2021. Gesture recognition using surface electromyography and deep learning for prostheses hand: State-of-the-art, Challenges, and future. https://www.frontiersin.org/articles/10.3389/fnins.2021.621885/full

[46] Xingyu Li, Hongjun Dai, Lizhen Cui, and Ya Wang. 2017. SonicOperator: Ultrasonic gesture recognition with deep neural network on mobiles. In *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*. 1–7. https://doi.org/10.1109/UIC-ATC.2017.8397483

[47] Zisu Li, Chen Liang, Yuntao Wang, Yue Qin, Chun Yu, Yukang Yan, Mingming Fan, and Yuanchun Shi. 2023. Enabling Voice-Accompanying Hand-to-Face Gesture Recognition with Cross-Device Sensing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 313, 17 pages. https://doi.org/10.1145/3544548.3581008

[48] Chen Liang, Chun Yu, Yue Qin, Yuntao Wang, and Yuanchun Shi. 2021. DualRing: Enabling Subtle and Expressive Hand Interaction with Dual IMU Rings. *Proc.*

ACM Interact. Mob. Wearable Ubiquitous Technol. 5, 3, Article 115 (sep 2021), 27 pages. https://doi.org/10.1145/3478114

[49] Jaime Lien, Nicholas Gillian, M. Emre Karagozler, Patrick Amihood, Carsten Schwesig, Erik Olson, Hakim Raja, and Ivan Poupyrev. 2016. Soli: Ubiquitous Gesture Sensing with Millimeter Wave Radar. *ACM Trans. Graph.* 35, 4, Article 142 (jul 2016), 19 pages. https://doi.org/10.1145/2897824.2925953

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

[51] Jian Liu, Chen Wang, Yingying Chen, and Nitesh Saxena. 2017. VibWrite: Towards Finger-Input Authentication on Ubiquitous Surfaces via Physical Vibration. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (Dallas, Texas, USA) *(CCS '17)*. Association for Computing Machinery, New York, NY, USA, 73–87. https://doi.org/10.1145/3133956.3133964

[52] Li Liu, Zhichao Cao, and Tianxing Li. 2023. FaceTouch: Practical Face Touch Detection with a Multimodal Wearable System for Epidemiological Surveillance. In *Proceedings of the 8th ACM/IEEE Conference on Internet of Things Design and Implementation*. 13–26.

[53] Vaughn & amp;amp;Davenport Development LLC. 2020. Face touch aware. https://apps.apple.com/us/app/face-touch-aware/id1502864331

[54] Pedro Lopes, Ricardo Jota, and Joaquim A. Jorge. 2011. Augmenting Touch Interaction through Acoustic Sensing. In *Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces* (Kobe, Japan) *(ITS '11)*. Association for Computing Machinery, New York, NY, USA, 53–56. https://doi.org/10.1145/2076354.2076364

[55] Alexis B Lyons, Lauren Moy, Ronald Moy, and Rebecca Tung. 2019. Circadian rhythm and the skin: A review of the literature. *J. Clin. Aesthet. Dermatol.* 12, 9 (Sept. 2019), 42–45.

[56] Wenguang Mao, Jian He, and Lili Qiu. 2016. CAT: High-Precision Acoustic Motion Tracking *(MobiCom '16)*. Association for Computing Machinery, New York, NY, USA, 69–81. https://doi.org/10.1145/2973750.2973755

[57] S. Marullo, T. Lisini Baldi, G. Paolocci, N. D'Aurizio, and D. Prattichizzo. 2021. No Face-Touch: Exploiting Wearable Devices and Machine Learning for Gesture Detection. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*. 4187–4193. https://doi.org/10.1109/ICRA48506.2021.9561178

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

[59] C. Metzger, M. Anderson, and T. Starner. 2004. FreeDigiter: a contact-free device for gesture control. In *Eighth International Symposium on Wearable Computers*, Vol. 1. 18–21. https://doi.org/10.1109/ISWC.2004.23

[60] Allan Michael Michelin, Georgios Korres, Sara Ba'ara, Hadi Assadi, Haneen Alsuradi, Rony R Sayegh, Antonis Argyros, and Mohamad Eid. 2021. FaceGuard: A wearable system to avoid face touching. *Front. Robot. AI* 8 (April 2021), 612392.

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

[62] Adiyan Mujibiya, Xiang Cao, Desney S. Tan, Dan Morris, Shwetak N. Patel, and Jun Rekimoto. 2013. The Sound of Touch: On-Body Touch and Gesture Sensing Based on Transdermal Ultrasound Propagation. In *Proceedings of the 2013 ACM International Conference on Interactive Tabletops and Surfaces* (St. Andrews, Scotland, United Kingdom) *(ITS '13)*. Association for Computing Machinery, New York, NY, USA, 189–198. https://doi.org/10.1145/2512349.2512821

[63] Lendy Mulot, Guillaume Gicquel, Quentin Zanini, William Frier, Maud Marchal, Claudio Pacchierotti, and Thomas Howard. 2021. DOLPHIN: A Framework for the Design and Perceptual Evaluation of Ultrasound Mid-Air Haptic Stimuli. In *ACM Symposium on Applied Perception 2021* (Virtual Event, France) *(SAP '21)*. Association for Computing Machinery, New York, NY, USA, Article 2, 10 pages. https://doi.org/10.1145/3474451.3476232

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

[65] Frazer Noble. 2023. Static Hand Gesture Recognition Using Capacitive Sensing and Machine Learning. " Sensors. *Sensors* 23 (2023).

[66] Shijia Pan, Ceferino Gabriel Ramirez, Mostafa Mirshekari, Jonathon Fagert, Albert Jin Chung, Chih Chi Hu, John Paul Shen, Hae Young Noh, and Pei Zhang. 2017. SurfaceVibe: Vibration-Based Tap & Swipe Tracking on Ubiquitous Surfaces. In *2017 16th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. 197–208.

[67] Tobias Röddiger, Christopher Clarke, Paula Breitling, Tim Schneegans, Haibin Zhao, Hans Gellersen, and Michael Beigl. 2022. Sensing with Earables: A Systematic Literature Review and Taxonomy of Phenomena. 6, 3, Article 135 (sep 2022), 57 pages. https://doi.org/10.1145/3550314

[68] Camilo Rojas, Niels Poulsen, Mileva Van Tuyl, Daniel Vargas, Zipporah Cohen, Joe Paradiso, Pattie Maes, Kevin Esvelt, and Fadel Adib. 2021. A Scalable Solution for Signaling Face Touches to Reduce the Spread of Surface-Based Pathogens. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1, Article 31 (mar 2021), 22 pages. https://doi.org/10.1145/3448121

[69] Wenjie Ruan, Quan Z. Sheng, Lei Yang, Tao Gu, Peipei Xu, and Longfei Shangguan. 2016. AudioGest: Enabling Fine-Grained Hand Gesture Detection by Decoding Echo Signal. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Heidelberg, Germany) *(UbiComp '16)*. Association for Computing Machinery, New York, NY, USA, 474–485. https://doi.org/10.1145/2971648.2971736

[70] [n. d.]. Safety sound level recommended by EPA and WHO. https://www.cdc.gov/nceh/hearing_loss/what_noises_cause_hearing_loss.html.

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

[72] Sheng Shen, He Wang, and Romit Roy Choudhury. 2016. I Am a Smartwatch and I Can Track My User's Arm. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services* (Singapore, Singapore) *(MobiSys '16)*. Association for Computing Machinery, New York, NY, USA, 85–96. https://doi.org/10.1145/2906388.2906407

[73] Yilei Shi, Haimo Zhang, Jiashuo Cao, and Suranga Nanayakkara. 2020. VersaTouch: A Versatile Plug-and-Play System That Enables Touch Interactions on Everyday Passive Surfaces. In *Proceedings of the Augmented Humans International Conference* (Kaiserslautern, Germany) *(AHs '20)*. Association for Computing Machinery, New York, NY, USA, Article 26, 12 pages. https://doi.org/10.1145/3384657.3384778

[74] Merrill I. Skolnik. 2008. *DESCRIPTION OF RADAR*. McGraw-Hill, 1.3–1.5.

[75] Emi Tamaki, Takashi Miyaki, and Jun Rekimoto. 2009. Brainy Hand: An Ear-Worn Hand Gesture Interaction Device. In *CHI '09 Extended Abstracts on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI EA '09)*. Association for Computing Machinery, New York, NY, USA, 4255–4260. https://doi.org/10.1145/1520340.1520649

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

[77] By Vanneste and O Bühler. 2011. Streaming by leaky surface acoustic waves. *Proceedings of The Royal Society A Mathematical Physical and Engineering Sciences* 467 (06 2011), 1779–1800. https://doi.org/10.1098/rspa.2010.0457

[78] Santiago Villarreal-Narvaez, Jean Vanderdonckt, Radu-Daniel Vatavu, and Jacob O. Wobbrock. 2020. A Systematic Review of Gesture Elicitation Studies: What Can We Learn from 216 Studies?. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference* (Eindhoven, Netherlands) *(DIS '20)*. Association for Computing Machinery, New York, NY, USA, 855–872. https://doi.org/10.1145/3357236.3395511

[79] Chuyu Wang, Jian Liu, Yingying Chen, Hongbo Liu, Lei Xie, Wei Wang, Bingbing He, and Sanglu Lu. 2018. Multi - Touch in the Air: Device-Free Finger Tracking and Gesture Recognition via COTS RFID. In *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*. 1691–1699. https://doi.org/10.1109/INFOCOM.2018.8486346

[80] Saiwen Wang, Jie Song, Jaime Lien, Ivan Poupyrev, and Otmar Hilliges. 2016. Interacting with Soli: Exploring Fine-Grained Dynamic Gesture Recognition in the Radio-Frequency Spectrum. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (Tokyo, Japan) *(UIST '16)*. Association for Computing Machinery, New York, NY, USA, 851–860. https://doi.org/10.1145/2984511.2984565

[81] 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 *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 290, 12 pages. https://doi.org/10.1145/3491102.3517698

[82] Yueting Weng, Chun Yu, Yingtian Shi, Yuhang Zhao, Yukang Yan, and Yuanchun Shi. 2021. FaceSight: Enabling Hand-to-Face Gesture Interaction on AR Glasses with a Downward-Facing Camera Vision. 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 10, 14 pages. https://doi.org/10.1145/3411764.3445484

[83] Robert Xiao, Greg Lew, James Marsanico, Divya Hariharan, Scott Hudson, and Chris Harrison. 2014. Toffee: Enabling Ad Hoc, around-Device Interaction with

Acoustic Time-of-Arrival Correlation. In *Proceedings of the 16th International Conference on Human-Computer Interaction with Mobile Devices & Services* (Toronto, ON, Canada) *(MobileHCI '14)*. Association for Computing Machinery, New York, NY, USA, 67–76. https://doi.org/10.1145/2628363.2628383

[84] Zhen Xiao, Tao Chen, Yang Liu, and Zhenjiang Li. 2020. Mobile phones know your keystrokes through the sounds from finger's tapping on the screen. In *2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS)*. IEEE.

[85] Chenhan Xu, Bing Zhou, Gurunandan Krishnan, and Shree Nayar. 2023. AO-Finger: Hands-Free Fine-Grained Finger Gesture Recognition via Acoustic-Optic Sensor Fusing. 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 306, 14 pages. https://doi.org/10.1145/3544548.3581264

[86] Xuhai Xu, Haitian Shi, Xin Yi, WenJia Liu, Yukang Yan, Yuanchun Shi, Alex Mariakakis, Jennifer Mankoff, and Anind K. Dey. 2020. EarBuddy: Enabling On-Face Interaction via Wireless Earbuds. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376836

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

[88] Linchu Yang, Ji'an Chen, and Weihang Zhu. 2020. Dynamic Hand Gesture Recognition Based on a Leap Motion Controller and Two-Layer Bidirectional Recurrent Neural Network. *Sensors* 20, 7 (2020). https://www.mdpi.com/1424-8220/20/7/2106

[89] Mais Yasen and Shaidah Jusoh. 2019. A systematic review on hand gesture recognition techniques, challenges and applications. *PeerJ Comput. Sci.* 5 (Sept. 2019), e218.

[90] Chun Yu, Xiaoying Wei, Shubh Vachher, Yue Qin, Chen Liang, Yueting Weng, Yizheng Gu, and Yuanchun Shi. 2019. HandSee: Enabling Full Hand Interaction on Smartphone with Front Camera-Based Stereo Vision. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300935

[91] 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. https://doi.org/10.1145/3081333.3081356

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

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

[94] Caroline Landelle, Jean-Luc Anton, Bruno Nazarian, Julien Sein, Ali Gharbi, Olivier Felician, and Anne Kavounoudias. 2020. Functional brain changes in the elderly for the perception of hand movements: A greater impairment occurs in proprioception than touch. In *NeuroImage (220)*. 117056. https://www.sciencedirect.com/science/article/pii/S1053811920305425

[95] Wing Lok Au, Irene Soo Hoon Seah, Wei Li, and Louis Chew Seng Tan. 2015. Effects of Age and Gender on Hand Motion Tasks. *Parkinson's Disease* 2015, 862427. ISSN: 2090-8083 (Print), 2042-0080 (Electronic) https://doi.org/10.1155/2015/862427