Enabling Voice-Accompanying Hand-to-Face Gesture Recognition with Cross-Device Sensing

Zisu Li[*,†]
zlihe@connect.ust.hk
Tsinghua University
Beijing, China
The Hong Kong University of Science and Technology
Hong Kong SAR, China

Chen Liang[*]
liang-c19@mails.tsinghua.edu.cn
Tsinghua University
Beijing, China

Yuntao Wang[‡]
yuntaowang@tsinghua.edu.cn
Tsinghua University
Beijing, China

Yue Qin
747929791@qq.com
Tsinghua University
Beijing, China

Chun Yu
chunyu@tsinghua.edu.cn
Tsinghua University
Beijing, China

Yukang Yan
yukangy@andrew.cmu.edu
Tsinghua University
Beijing, China

Mingming Fan
mingmingfan@ust.hk
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
The Hong Kong University of Science and Technology
Hong Kong SAR, China

Yuanchun Shi
shiyc@tsinghua.edu.cn
Tsinghua University
Beijing, China
Qinghai University
Xining, China

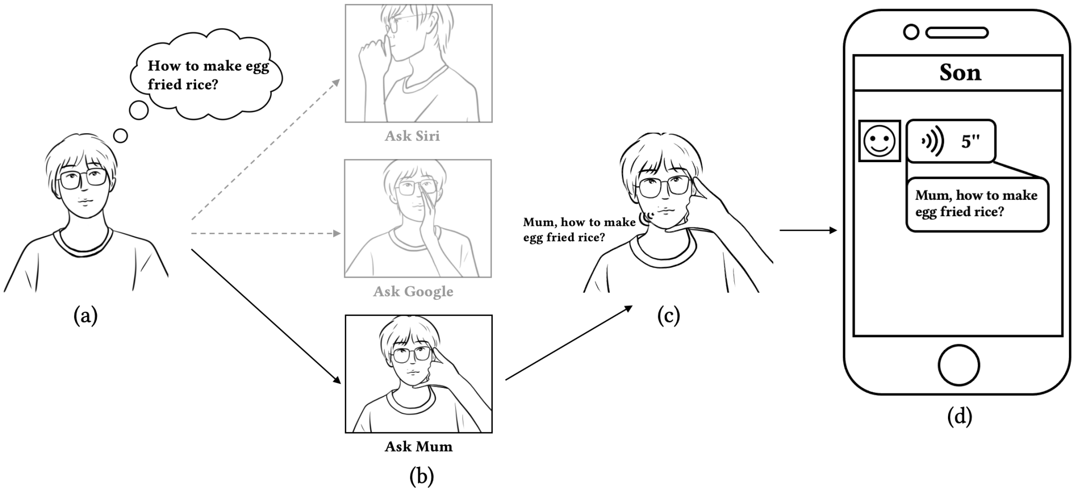

How to make egg fried rice?
(a)
Ask Siri
Ask Google
Ask Mum
(b)
Mum, how to make egg fried rice?
(c)
Son
5''
Mum, how to make egg fried rice?
(d)

Figure 1: A typical usage scenario enabled by voice-accompanying hand-to-face (VAHF) gestures. (a) The user wants to know how to make egg fried rice. (b) The user can perform different VAHF gestures to redirect their voice input to different targets (e.g., asking Siri, searching on Google with the transcribed text, and sending a voice message to mum). (c) The user performs a "phone call" gesture and speaks simultaneously. (d) The smart devices recognize the user's intention through the performed VAHF gesture and simulate sending a voice message to the user's mum.

ABSTRACT

Gestures performed accompanying the voice are essential for voice interaction to convey complementary semantics for interaction purposes such as wake-up state and input modality. In this paper, we investigated voice-accompanying hand-to-face (VAHF) gestures for voice interaction. We targeted on hand-to-face gestures because such gestures relate closely to speech and yield significant acoustic features (e.g., impeding voice propagation). We conducted a user study to explore the design space of VAHF gestures, where we first gathered candidate gestures and then applied a structural analysis to them in different dimensions (e.g., contact position and type), outputting a total of 8 VAHF gestures with good usability and least confusion. To facilitate VAHF gesture recognition, we proposed a novel cross-device sensing method that leverages heterogeneous channels (vocal, ultrasound, and IMU) of data from commodity devices (earbuds, watches, and rings). Our recognition model achieved an accuracy of 97.3% for recognizing 3 gestures and 91.5% for recognizing 8 gestures (excluding the "empty" gesture), proving the high applicability. Quantitative analysis also shed light on the recognition capability of each sensor channel and their different combinations. In the end, we illustrated the feasible use cases and their design principles to demonstrate the applicability of our system in various scenarios.

CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile devices**; **Gestural input**.

KEYWORDS

hand gestures, acoustic sensing, sensor fusion

ACM Reference Format:
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 (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3544548.3581008

1 INTRODUCTION

Voice input has become a natural and always-available interaction modality for wearable devices such as earphones and smartwatches. However, the modality control (e.g., the wake-up state) in voice interaction is still a challenging problem due to the implicitness of modality information in speech and the restricted NLP techniques. People have to repeat the hotword to switch to the modality or the target device actively, which introduce extra burdens for the interaction. Thus, researchers have been seeking supplementary input methods as parallel input channels to assist voice interaction [52, 76, 78].

People tend to perform body gestures accompanying their voice for better expression of certain emotions or intentions during the conversation [61]. In such a case, we defined the gesture, which is performed simultaneously with the speech, as a voice-accompanying gesture. Analogously, in the voice interaction with smart devices, the use of voice-accompanying gestures could provide parallel information that expands the input channel bandwidth [34, 52, 62], simplifies the voice interface flow [59, 76, 78], and make voice interaction more convenient [52, 59, 76]. The underlying logic why voice-accompanying gestures have been prevalent and widely researched is due to the physiological nature that the voice channel and the gesture channel are highly independent and complementary [52] (e.g., performing a gesture as parallel input information while not repressing the voice expressivity). For instance, the user can define a specific voice-accompanying gesture (e.g., covering the mouth) instead of repetitively saying the wake-up keyword to keep the voice interface active. The user can also define multiple gestures to represent the redirection of voice input to different input modalities (e.g., ignored, interrupted, transcribed, or raw audio input), target devices (e.g., whether should the TV or the smartphone accept the input), and shortcuts (e.g., binding certain UI operation flows with the gesture).

In this paper, we investigated the feasibility of using voice-accompanying hand-to-face (VAHF) gestures as parallel channels to improve the traditional voice interaction flow. Specifically, we aim at designing VAHF gestures and recognizing them with an acoustic-based cross-device sensing method. We targeted hand-to-face gestures as the instantiation of voice-accompanying gestures because they have been proven to be natural, expressive (e.g., various landmarks on the face to yield a large gesture space), and more related to the speech by existing researches [71, 74, 76]. Moreover, hand-to-face gestures yielded significant features in voice propagation, which is beneficial for acoustic sensing.

To understand the design space of VAHF gestures, we conducted a user-centric gesture elicitation study with a total proposal of 15 gestures from end-users. Then we narrowed down the gesture set from 15 to 8 gestures with better usability and the least ambiguity.

As for the sensing schemes, the underlying principle is that when the user speaks, each VAHF gesture creates a unique acoustic propagation path from the mouth to the set of microphones on wearable devices. Therefore, our method can recognize the hand-to-face gesture using the acoustic features of each unique propagation path. Further, we incorporated an ultrasound channel and an IMU channel to provide supplementary sensing information and enhance gesture recognition. For the ultrasound channel, the smartwatch served as an ultrasound source and all microphones captured such signals from different positions to indicate position-aware features. For the IMU channel, the IMU on the ring can convey the attitude and motion of the user's hand. We also investigated the fusion mechanism among different devices and channels.

To evaluate our technique, we first built a cross-device VAHF dataset consisting of 1800 samples of 8 VAHF gestures along with one "empty" gesture (meaning not performing a VAHF gesture).

[*] indicates equal contribution.
[†] This work was conducted when Zisu Li was a research intern at Tsinghua University.
[‡] indicates the corresponding author.

CHI '23, April 23–28, 2023, Hamburg, Germany
© 2023 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3581008

Then we conducted 1) a two-factorial evaluation regarding sensor combination and model selection, 2) an extensive evaluation of a reduced gesture set, and 3) an ablation study for the optimal model to validate the computation feasibility and the applicability of our technique. Results showed our model achieved high recognition accuracy of 97.3% for 3 + 1(empty) gestures and 91.5% for 8+1(empty) gestures recognition on our cross-device VAHF dataset. Quantitative analysis also sheds light on the recognition capability of each sensor channel and its different combinations.

At the end of the paper, we discuss real-life application scenarios to demonstrate the applicability of VAHF gestures as well as provide general design implications for VAHF gesture-enhanced voice interaction.

In summary, the contributions of this paper are as follows:

- We conducted a comprehensive study to elicit the gesture space of VAHF gestures and proposed a gesture set with better usability, better social acceptance, less fatigue, and less ambiguity.
- We propose a novel sensor-fusion technique for VAHF gesture recognition which is supported by cross-device sensors. Our quantitative analysis sheds light on the recognition capability of the different sensor combinations over VAHF gestures with different characteristics.
- We demonstrate a set of use cases of our gesture recognition technique that outline new opportunities for VAHF gestures to benefit voice interaction.

2 RELATED WORK

In this section, we presented related work in three aspects: enhancing voice interaction with parallel gestures, hand-to-face interaction, and cross-device sensing for hand gestures.

2.1 Enhancing Voice Interaction with Parallel Gestures

Performing body gestures parallel to voice commands has been a prevalent method to convey certain intentions or information during the voice interaction. People tended to use gestures of different body segments, such as head gestures[17, 52, 56, 65], gaze[18, 49], hand gestures [76], and facial expressions [77], to provide supplementary information obliged for voice interaction. The purpose of introducing of certain parallel body gestures to the voice interface typically included: indicating the wakeup state [59, 76, 78], serving as the control (or trigger) signal [74, 77], and passing scene-related context information [1, 52]. For example, Yan et al. [77] proposed frowning, a facial expression of para-language, to implement interrupting the responses during voice interactions between human and smart devices. Qin et al. [59] leveraged the speech features when user raise the microphone embeded in devices to close to mouth to facilitate wake-up free techniques. WorldGaze [52] used commodity a smartphone to recognize the real-world head-gaze location (e.g., certain buildings or objects) of a user to provide the voice agent with supplementary scene-related information for better comprehension.

Our work was also within the framework of using parallel gestures as active input to enhance voice interaction, which was most related to but achieved a leap over PrivateTalk [76], which allowed users to activate voice input by performing a hand-on-mouth gesture during speaking. Compared with existing work [59, 76, 77] where a specific gesture (e.g., bringing the phone to the mouth[59]) was designed and recognized for specific functionality (e.g., interrupting the conversation or activating the voice assistance), our multiple VAHF gesture recognition on multi-modal wearable devices could effectively broaden the input channel of actions as parallel information with the potential of supporting a larger interaction space, such as defining multiple shortcuts.

2.2 Hand-to-Face Interaction

Gestures involving hand and face have been demonstrated as a natural and easy-to-use way to input commands. Prior research has proposed the validation of the inherent unobtrusiveness, subtlety and social acceptability[63] of hand-to-face interaction. Design space of hand-to-face gestures has been explored by prior research. For example, different face regions such as the ear [35], cheek [63, 75] or nose [40] were demonstrated to have the viability of hand-to-face input. Mahmoud et al. [50] proves that people prefer lower face regions to upper regions, especially chin, mouth and lower cheeks, for naturalistic interaction. Weng et al. [71] proposed recognition of hand-to-face gestures for AR glasses. Serrano et al. [63] provided a set of guidelines for developing effective Hand-to-Face interaction techniques and found that the cheek is the most promising area on the face. Miniaturizing obfuscating, screening, camouflaging and re-purposing have been purposed by Lee et al.[39] as the strategies of the design of socially acceptable hand-to-face gestures. We refer to some principles (e.g. lower face region, social acceptance, etc.) from previous work mentioned above to design our subjective evaluation process. Different from previous work, our design space of hand-to-face gestures are generated from real-life conversations, which has more para-linguistic features resulting in performing more easily and naturally.

Combining voice input with hand gestures can provide rich information to enable convenient and expressive multi-modal interaction [34]. "Put that there" was the first multi-model interaction method introducing the pointing gesture to indicate the location mentioned in the voice command [5]. Bourguet et al. studied the temporal synchronization between speech (Japanese) and hand pointing gestures during multi-modal interaction [6]. Sauras-Perez et al. [62] proposed a human vehicle interaction system based on voice and pointing gestures that enables the user making spontaneous decisions over the route and communicate them to the car. Closest to our work, Yan et al. [76] proposed to enable hand-to-mouth gesture interaction as a wake-up action for voice interfaces. However, the combination of the hand-to-face gestures and voice interaction in previous work is limited to few types of gestures and the fixed usage patterns. In our work, we discussed more types of gestures which can be used in versatile scenarios of voice interaction.

2.3 Cross-Device Sensing for Hand Gestures

Free-form hand gesture sensing is key to enabling rich interaction space taking advantage of the expressiveness of human hand. Previous literature has investigated sensing solutions for free-form hand gestures with different sensors including cameras (RGB [26, 33, 57, 79], IR [8, 71], and depth [27, 45]), EMG sensors [47, 53], capacitive sensors [12], millimeter-wave radar [22, 44, 67], acoustic sensors [25, 46, 58, 68], and inertial sensors [7, 38, 66, 69]. Among these sensors, the camera is most widely researched since it has the strongest sensing capability to capture pixel-wise image data, based on which many computer vision models have been developed for fine-grained hand sensing such as detecting hand keypoints and recovering the hand pose. However, vision-based hand gesture sensing often requires externally-mounted camera and heavy computation (e.g., using a GPU), which prevents the practical use in mobile and pervasive scenarios. Similarly, sensing methods based on EMG sensors [47, 53], capacitive sensors [12], and millimeter-wave radar [22, 44, 67] require additional wearing on the human body, making them far from practical deployment. In our work, we focused on the latter two - microphones and inertial sensors - because they are computational efficient and largely equipped on commodity devices such as smartphones, earbuds, smartwatches, and smart rings. Below we presented related work on acoustic- and inertial-based hand gesture sensing.

2.3.1 *Acoustic sensing.*

The principle of acoustic sensing for hand gestures is to measure how specific hand gesture influences the propagation of active (e.g., active ultrasound) or passive (e.g., human voice) sound sources or makes a sound. Based on the presence of an active sound source, it can be categorized into active acoustic sensing and passive acoustic sensing.

Active acoustic sensing methods [68] has been widely explored for gesture recognition including in-air hand gesture [25], ambient activity [54], touch gesture on everyday objects or surfaces [46, 58], finger tracking [80], silent speech interface [19, 82]. For example, Touch & Activate [58] enabled touch interface on everyday objects by measuring the acoustic frequency response of the object and the touch gesture. Strata [80] enabled 2-dimensional finger position tracking using the reflective impulse audio signal on commodity smartphones. Ando et al. [2] modeled the transfer function, or the propagation path of the sound, and used it in gesture recognition. These methods recognized the acoustic echo or propagation features caused by the gestures using a speaker and one or more microphones. Doppler effect was also frequently used for sensing subtle hand gestures involving relative movements [25, 46]. EchoWhisper [19] also leveraged the Doppler shift of reflection for near-ultrasound sound waves caused by the mouth and tongue movements to interpret the speech and build a silent speech interface.

Passive acoustic sensing recognizes sound activities using merely microphones [28, 29, 37, 54, 72, 73]. For instance, Toffee [73] enabled an ad-hoc touch interface on a table around the device using time of arrival correction. TapSense [28] enhanced finger interaction on touch screens by detecting the unique sound features of fingertip, pad, knuckle, and nail. Acoustic Barcodes [29] was an identifying tag that used notches to produce sound when dragged across, which can be recognized by a microphone for information retrieval or triggering interactive functions. Daily activities or ambient environment (e.g., taking a bus) can be detected based on features of the sound collected by a single microphone [54] or a microphone array. Ubicoustics [37] proposed a plug-and-use sound recognition pipeline for general-purpose activity recognition. Wu et al. [72] further extended the environment acoustic event detection using an end-to-end system for self-supervised learning of events labeled through one-shot interaction.

In our work, we took advantage of both passive and active acoustic sensing. To be more specific, passive acoustic sensing recognized the hand's influence on the features of the accompanying speech, including frequency response, amplitude, etc. Active acoustic sensing helped to determine relative position-based features among different devices. The two channels can provide supplementary capability in VAHF gesture sensing.

2.3.2 *Inertial Sensing.*

Inertial sensors, commonly integrated into commodity devices, are efficient in detecting motion- and attitude-related hand/finger gestures. For instance, a number of previous works [7, 38, 42, 43, 66, 69] used acceleration and rotation with wrist-worn inertial sensors to recognize hand gesture. Serendipity [70] recognized five fine-grained gestures based on the IMU in off-the-shelf smartwatches. Mo-Bi [36] used a smartphone and two accelerometer-embedded wrist-worn devices for each hand to collect the hand-posture data and developed the implicit hand-posture recognition software. Leveraging inertial sensors integrated into smartwatches, Float [64] recognized wrist-to-finger gestures to enhance one-hand and smartwatch interaction. Gu et al. [24] enabled one-finger typing with an index-finger-worn IMU ring by detecting hand-to-surface touching events and rotation angles. Lu et al. [48] studied the sensing capability of dual wrist-worn devices and analyzed cross-device features for more accurate gesture inference.

Inspired by these works, we incorporated an IMU ring into our sensing system to capture the motion features of VAHF gestures.

2.3.3 *Sensor Fusion Methods for Hand Gestures.*

Previous research has explored cross-device sensor fusion methods to enhance the recognition capability of different types of hand gestures, especially when the sensing capability of different sensors are complementary for recognizing different features. Sensor fusion methods included homogeneous fusion and heterogeneous fusion. Homogeneous fusion aimed to add more homogeneous sensor nodes into the sensing system to capture fine-grained information. For example, fusing camera captures from different views [55] is a classical and effective solution to reduce 3D reconstruction or detection error which is also widely used in generating high-quality machine annotations. For acoustic sensing, adding more microphones to the scene achieved more fine-grained acoustic measurements, which is beneficial to various sensing purposes such as 2D localization [80] and gesture classification [68].

The other type is heterogeneous fusion, where different types of sensors are combined to merge their strengths [9, 14, 20, 31, 41]. For example, Li et al. [41] presented a hierarchical sensor fusion approach for human micro-gesture recognition by combining an Ultra Wide Band (UWB) Doppler radar and wearable pressure sensors. Ceolini et al. [9] presented a sensor fusion framework that integrates complementary systems: the electromyography (EMG) signal from muscles and visual information. Ceolini et al. [10] also investigated the fusion of EMG and a camera on limited computational resources of mobile phones to detect gestures. A more typical scene is fusing an IMU with a camera [23], where the IMU detects the subtle contact signal and the camera senses the global hand state. Acoustico [21] fused acoustic and IMU signal for 2D

Table 1: Text description and empirical categorization for all the 15 gestures in our gesture set. Each gesture was empirically categorized in three dimensions: contact position, contact type, and occlusion state. Contact position is represented in 5 levels: ear (E), mouth (M), chin(CN), cheek(CK), and none(N). Contact type is represented in 3 levels: finger(F), palm(P), and hand segments(HS). The occlusion state on the sound propagation path for the human voice to ears is represented in 3 levels: hardly(H), partially(P), and completely(C).

Index	Gesture	Contact Position	Contact Type	Occlusion State	Semantics
1	pinch ear rim	E	F	H	Earphone Manipulation
2	thinking face gesture	M, CN	HS	H	Thinking, Querying
3	support cheek with fist	M, CK	P	P	Thinking, Resting
4	non-contact cover mouth with palm	N	P	P	Directional Speech, Whisper
5	support cheek with palm	M, CK	P	P	Thinking, Concentrating
6	cover mouth with fist	M	HS	C	Interphone, Messaging
7	cover ear with arched palm	E	P	C	Hearing, Phone Call
8	hold up palm beside nose and mouth	M, CK	P	C	Directional Speech, Block
9	touch earphone with index finger	E	F	H	Earphone Manipulation
10	touch top ear rim	E	F	H	Earphone Manipulation
11	touch vocal cord	N	F	H	Voice Distortion
12	cover mouth with palm	M, CK, CN	P	C	Silence, Whisper
13	shushing gesture	M	F	H	Silence, Interruption
14	touch the back of ear rim	E	F	H	Hearing, Attention
15	calling gesture	M, CK, CN, E	HS	P	Communication, Phone Call

tap position localization based on the TDOA of the tap sound's two propagation paths.

In our work, we adopted both homogeneous fusion and heterogeneous fusion strategies. The former aimed to probe more measurement nodes into the sensing space, while the latter aimed to capture different types of features from different channels.

3 DESIGNING VOICE-ACCOMPANYING HAND-TO-FACE GESTURES

To thoroughly understand and explore the gesture space of voice-accompanying gestures, we conducted a user-centric gesture elicitation study to elicit gesture design from end users. Subsequently, we conducted a hierarchical analysis process to narrow down the gesture space from 15 gestures to 8 gestures which are easy to perform, easy to memorize, and with less ambiguity. The design of this study was in line with the previous gesture elicitation work [13, 48, 74] consisting the typical phases of gesture proposal, gesture evaluation, and gesture set refinement.

3.1 Voice-accompanying Hand-to-Face Gesture Proposal

We conducted a brainstorming gesture proposal study to understand users' preference on VAHF gestures and derive a gesture set with general agreements.

3.1.1 Participants, Brainstorming Design, and Procedure. We recruited 10 participants (4 female, all right-handed) from the local campus, with an average age of 21.3 (from 18 to 27, SD=2.4). Their familiarity score of wearable devices and voice interaction was 3.35 (SD=1.2). The whole study took about 1 hour and each participant received 15$ for compensation.

The purpose of this study is to encourage participants to brainstorm as many voice-accompanying hand-to-face (VAHF) gestures as they could without considering the sensing feasibility and together work out a usable VAHF gesture set with common agreements. To achieve this point, we do not restrict the gestures to specific application scenarios or tasks, which maintained their focus on the physical nature of performing different VAHF gestures. The only constraint we imposed on the design was that the gestures should be static and durable to meet the nature of "voice-accompanying". The whole study consisted of 4 stages: 1) icebreaking and introduction, 2) individual thinking, 3) individual proposal, and 4) group discussion.

After a short icebreaking procedure where all the participants introduced themselves and familiarized themselves with each other, the experimenter acknowledged the participants of the purpose and the procedure for the study as well as the definition of VAHF gestures to the participants. Then the participants went through an individual thinking process for 10 minutes where participants worked separately to come up with as many gestures as possible and wrote them down on a notebook as detailed as possible (e.g., encouraging them to write down the motivations, semantics, and potential usages of the gestures other than simply the descriptions). After that, each participant was asked to verbalize their proposal (including the gesture descriptions, motivations, semantics, and potential usages) and perform the proposed gestures by hand orderly. They could also sketch and show their ideas on a public whiteboard. Participants then came up with a group discussion where one could either show the pros and cons of the others' proposal or generate new gestures from the others' inspiration. The discussion ended until all participants worked out a final gesture set with the consistent agreement. The whole brainstorming process was hosted by two experimenters - one guiding the experiment while the other taking notes of the key points presented by the participants.

3.1.2 Results and Discussion. Fig. 2 and the "Gesture" column of Table 1 illustrated the 15 VAHF gestures and their text descriptions proposed by users in the brainstorming study. The "Semantics" column summarized some of the typical semantics of each gesture from participants' quotes. An interesting finding is that participants tended to design the gestures in a mimetic and semantic-based manner, borrowing the inspirations from their daily activities and usages of smart devices. For example, touching gestures on the ear (gestures 1, 9, and 10 in Table 1, same as below) was regarded as the metaphor of earphone manipulations related to voice interaction (N=9), which came from their experiences of using wireless earbuds (e.g., triggering the voice assistant and controlling the volume and the progress). Participants also presented their potential usages, such as waking up Siri (P2), making a voice memo (P3), and sending a voice message to a specific person (P8). Similarly, participants described gestures 6, 7, and 15 as "the imitation of using certain devices" (N=10). *"Holding up the fist in front of the mouth is a cool gesture. It is just like sending an instant command with an interphone"* (P4). *"Covering the ear with the arched palm is like you are holding the phone while the 'calling' gesture is like you are imitating an old-fashioned telephone. I would prefer the former one because it is easy to perform and seems more natural to others"* (P10).

In addition to the above gestures related to device usage, some gestures were proposed for their prevalence in daily communication and social expression. Participants (N=10) showed their will to transfer these gestures to the interaction with voice assistance. For example, the "shushing" gesture (gesture 13) and the "flaring ear" gesture (gesture 14) were proposed because they were frequently used in daily dialog. *"Shushing has the meaning of silence and interruption. We can also use it to interrupt the conversation with the voice assistant" (P3). "The 'flaring ear' gesture means 'pardon' or shows attention to the speaker. I guess it would be nice to assign this gesture to functions with similar meanings" (P1).* Gestures 4, 6, 7, and 12 were mouth-related gestures proposed by the participants, with the general meanings of special speech, lowered volume, whisper, and silence. The gestures were distinguished by different ways of covering the mouth. *"When I hold up the palm on one side of the mouth, I probably want to speak to the one on the other side directionally. However, when I cover my mouth, the meaning could be totally different" (P3).* Similarly, participants designed three face-related gestures (gestures 2, 3, and 5), indicating thinking, querying, resting, or concentrating, yet with slightly different implications. *"It would be wonderful the voice assistant could respond to my 'thinking face' gesture by querying my words on the searching engine"* (P2). Exceptionally, P5 proposed a "touch vocal cord" gesture (gesture 11) with a unique position. "The vocal cord affects the timbre, meaning to 'make a different sound' (P5)."

Although the semantics of each gesture seemed clear to individuals, we found some conflicts in the group discussion stage. For example, regarding the "cover mouth with fist" gesture (gesture 6), most participants showed approval of the "interphone" metaphor while some participants (P2, P5) thought it should be with the semantics of "silence" and "secrete". Some participants also mentioned the meanings and preferences of certain gestures might vary under different cultural backgrounds, especially for the gestures with social functionalities.

3.2 Optimizing VAHF Gesture Set

To derive the final user-defined gesture set from all gestures proposed by all participants, we collated the gestures and asked participants to perform all the gestures, and conducted subjective ratings from 4 dimensions. We resolved repeatability between gestures by empirical categories, which intuitively characterized the similarity between gestures from 3 dimensions. We chose one gesture from each category to a subset of the most preferable gestures.

3.2.1 Subjective Evaluation. After deriving a gesture set with 15 gestures, we sought to find out which gesture is most suited for voice interactions, especially in social acceptance and using fatigue. We recruited 25 participants (10 male and 15 female) for our subjective evaluation, with an average age of 21(from 19 to 32, SD = 2.1). All of the participants were right-handed. Each participant performed all gestures three times using their right hand. The order of the gestures was pre-determined to counterbalance ordering effects. For each gesture, the experimenter would show an example video of this gesture to ensure the participant could perform the gesture correctly. The participant then followed the instructions provided on a laptop screen to perform gestures. After performing the gesture three times, the participant was asked to rate the gesture according to the following four criteria along a 7-point Likert scale (1: strongly disagree to 7: strongly agree), and the results are shown in Fig. 2:

- **Usability** measured ergonomics to reflect the comfort of the gesture. The participants are required to consider the gesture not only in stationary conditions (e.g., sitting) but also under moving conditions (e.g., running). The higher the score, the easier the gesture is to perform.
- **Social Acceptance** measures if the user will feel uncomfortable or embarrassed, or if performing the gesture will disturb others in public settings. The higher the score, the more acceptable the gesture is in social environments.
- **Disambiguity** measures the difficulty of confusing the gesture with daily hand movements or with other gestures. The higher the score, the less ambiguous the gesture is.
- **Fatigue** measures the physiological burden to perform the gesture. The higher the score, the less fatigue the gesture is to perform.

3.2.2 Design Principles and Finalized Gesture Set. In order to eliminate the design consistency and gestures with signal similarity to derive gestures that can be naturally performed and quickly remembered, we categorized all the gestures to propose the most representative one in each category. Considering the propagation path of the human voice around the head, we identify three structural properties to represent the proposed gesture set, which are illustrated in Table 1:

- **Contact Position:** Due to the different contact positions of the fingers, the microphone mounted on the ring can receive different sounds. Because the mouth is the source of the sound, the closer the finger touches the sound source, the louder the sound will be picked up by the microphone on the ring. In all gestures, the contact positions of the fingers are the mouth(M), the cheek(CK), the chin(CN), and the ear(E).
- **Contact Type:** Different contact types, specifically divided into fingers(F), palms(P), and hand segments(HS) by which hands used to contact the face region, have clear distinctions in morphology which can be distinguished easily by users without ambiguity. Furthermore, different contact types will form a unique structure on the face to affect the collection of the earphones' feed-forward microphones.
- **Occlusion State:** The occlusion state, which is separated in 3 levels(hardly(H), partially(P), and Completely(C), will produce different sounds by affecting the propagation path from the human voice to the ears. For example, gesture 7 (cover ear with arched palm) and gesture 12 (cover mouth with palm) shown in Fig. 2, which 'completely' occlude the receiver (ears) and the transmitter (mouth) of the sound the propagation path in the air, will cause a loss of high-frequency sound.

We combined the subjective evaluation results shown on Fig. 2 (in 4 dimensions: usability, social acceptance, disambiguity, and fatigue) with the gesture set optimization process. The process was based on first grouping gestures with the same structural property combinations from the above three dimensions and chose one gesture according to the subjective scores to represent each category. From the categorization results, we found that gestures belonging to the E-F-H category include gestures 1, 9, 10, and 14. This type of gesture with fingers to contact the ear region and have similarity in signal. Moreover, E-F-H gestures are also commonly used to

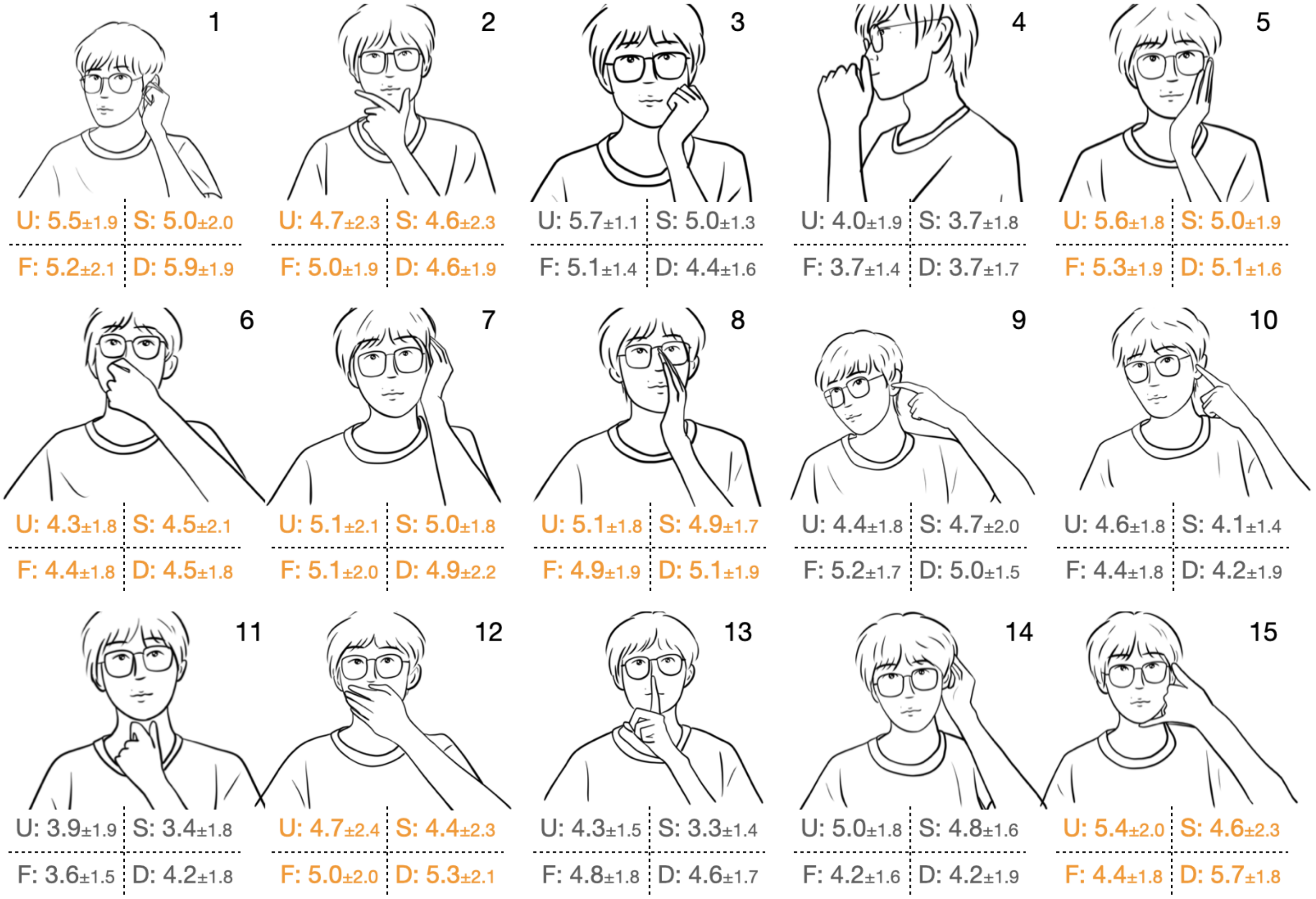


Figure 2: Drafts illustrating each gesture in the gesture set: 1) pinch the ear rim, 2) thinking face gesture, 3) support cheek with fist, 4)non-contact cover mouth with palm, 5)support cheek with palm, 6) cover mouth with fist 7) cover ear with arched palm, 8) hold up palm beside nose and mouth, 9) touch the earphone with index finger, 10) touch the top ear rim, 11) touch the vocal cord, 12) cover mouth with palm, 13) shushing gesture, 14) touch back of ear rim with fingers, 15) calling gesture. The mean±s.d. of users' subjective scores (1-7, the higher the better) on Usability (U), Social Acceptance (S), Fatigue (F) and Disambiguity (D) is shown on the bottom of each draft. Scores of gestures in our final gesture set are highlighted in orange.

interact with earphones. We chose gesture 1 to represent E-F-H gestures in our final subset of VAHF gestures according to subjective ratings; Gestures 3,5 belong to M,CK-P-P gestures. Considering the operating region of gesture 4 is also near mouth and cheek(M, CK) although it does not actually contact the face and the gesture 4's other two dimensions are the same P(Contact Type)-P(Occulasion State) with gesture 3, 5, we grouped gestures 3,4,5 into one category and chose gesture 5 to represent this category. Gestures 11 and 13 are omitted due to their lower social acceptance (e.g., lower than 3.5). And the remaining gestures (2,6,7,8,12,15) are kept in the gesture set due to their specificity in the three dimensions and higher subjective scores. The gesture selection process resulted in 8 gestures. We checked the subjective scores of each dimension of the eight gestures selected again and found they are all above 4.4 and have a comprehensively higher score over others, which proves that our gesture selection is subjectively reasonable and practical for users.

The above gesture selection process filtered out the following 8 gestures: gesture 1 (E-F-H), gesture 2 (M,CN-HS-H), gesture 5 (M,CK-P-P), gesture 6 (M-HS-C), gesture 7 (E-P-C), gesture 8 (M,CK-P-C), gesture 12 (M,CK,CN-P-C), gesture 15 (M,CK,CN,E-HS-P) where the indexes were consistent with Fig. 2. These gestures constituted our final gesture set.

4 RECOGNIZING VOICE-ACCOMPANYING HAND-TO-FACE GESTURES WITH CROSS-DEVICE SENSING

In this section, we introduce the design considerations and the technical details of our cross-device sensing method to recognize VAHF gestures. We explain the implementation considerations regarding device and channel selection. Then we present individual sensing models for vocal, ultrasonic, and IMU channels. Finally, we clarify the sensor combination and fusion strategies for real-world deployment.

4.1 Considerations and Technical Overview

We first clarify the considerations of our implementation before going into the technical details. To recognize VAHF gestures, we chose 3 types of commercial wearable devices - wireless ANC earbuds, smartwatches, and smart rings - as the sensor nodes in consideration of real-life deployment. Each wireless ANC earbud consists of an inner microphone and an outer microphone for noise canceling. The smartwatch is equipped with a microphone and a speaker which is capable to play sounds at 22.5 KHz and the ring is equipped with a microphone and an IMU. We chose the microphones and the IMUs as the sensor candidates in consideration of the computation efficiency for the always-availability (e.g., raise-to-speak technique on an Apple watch). These sensors are widely equipped on the aforementioned commercial wearable devices (earbuds, watch, and ring).

An illustration of the entire system is shown in Figure 3. The sensing system consists of three independent models: vocal model, ultrasonic model, and IMU model. Each channel takes the corresponding preprocessed signal from the devices and outputs the feature vectors, which are fused and fed to the classifier layers to output the prediction logits.

4.2 Recognizing VAHF Gestures with Single-Modality Solutions

To facilitate efficient recognition of VAHF gestures, we first build three individual sensing models involving three independent channels of features - vocal features, ultrasonic features, and IMU features. Each channel of features serves as individual input of recognition in different dimensions.

4.2.1 Vocal Model. Performing hand-to-face gestures while speaking leads to changes in the acoustic property including amplitude, frequency response, and reverberation for the received signal. For example, an "hold up the palm beside nose and mouth" gesture may impede the direct transmission of sound to the left earbud's microphone, resulting in a lower amplitude and decay in high frequency in the corresponding channel. Since we focus on the difference in the acoustic property among the distributed microphones, we first figured out the reference channel. Typically, we chose the inner microphone as the reference channel when inner and outer microphones were simultaneously used and the outer microphone of the right earbud when inner channels were disabled.

Similar to prior work, we processes the audio data for classification using mel spectrum [30]. Given a set of audio segments from all channels $[a_1, \cdots, a_n; a_{ref}]$ with the sample rate of 16 KHz, we convert each segment into the frequency domain by first applying the short-time Fourier transform then adopting a mel-scale transform with 128 mel filterbanks, after which we pad or trunk each spectrum in the temporal axis with zeros into 128×250 ($\approx$ 3 seconds) and acquire n+1 maps $[m_1, \cdots, m_n; m_{ref}]$. Then we subtract the reference map m_{ref} from all the monitored map m_i to acquire the channel-wise difference in mel spectrum $[m_1 - m_{ref}, \cdots, m_n - m_{ref}; m_{ref}]$. Finally, we concatenate all the maps in the first axis into a single input frame that can be fed into a deep-learning classification model.

A MobileNet V3 Large [32] model pretrained on ImageNet [60] is used as the backbone network for feature extraction. The input frame is fed to the feature extractor layers of the pretrained model to generate a 1-D feature series f_{spec}. Such a model is chosen in consideration of the balance between computational complexity and performance [32]. Benefitting from the well-designed network structure and the large parameter space with good initialization, MobileNet V3 Large has the potential in capturing fine-grained textural and geometry features from the concatenated spectrum map.

Despite the direct use of the neural network on the mel frequency map, we extracted two additional sets of statistics features - transient signal amplitude and pair-wise similarity among Mel-frequency spectrum coefficients (MFCC) series - as classifier input, which is inspired by PrivateTalk's [76] solution in dealing with channel difference and delay between audio segments. For the transient amplitude feature, we use a sliding window with the size and stride of 200 to compute the amplitude series for each segment, after which we pad or trunk each series to a fixed length of 250. Then we concatenate all the amplitude series into a 1-D feature series f_{amp}. For pair-wise MFCC similarity, we first compute the MFCC series for each audio channel, then resample each MFCC map in the temporal domain into 20-frame segments with a stride of 10. For each pair of segment series, we compute their similarity using dynamic time warping (DTW) [3]. We acquired the pair-wise similarity feature vector f_{MFCC} by concatenating all the above $\frac{1}{2}n(n+1)$ simularity values.

After getting f_{spec}, f_{amp}, and f_{MFCC}, we concatenate them into a 1-dimensional vocal feature f_{vol}, which can either be used in an individual recognition model or be combined with other features. For an individual recognition model, f_{vol} is fed into a multi-layer perceptron (MLP) classifier to predict the performed gesture.

4.2.2 Ultrasonic Model. When the user performs a hand-to-face gesture, with his hand reaching different position on the face, the relative positions among the wrist, the finger, and the ears are temporally changing, thus yielding salient positional features. To facilitate such features, we devised an embedded ultrasonic sensing component, where the speaker on the smart watch works as an active source transmitting a 17.5 KHz - 22.5 KHz linear chirp signal which is captured by the microphones on the target devices. Such a design is inspired by the theory of Frequency Modulated Continuous Wave (FMCW) [51], which is widely used in radar and indoor positioning systems to acquire positional tracking information. The sensing principals can be formalized as a typical linear chirp based FMCW. Let $x_0(t) = A_0cos(2\pi f_0t + \pi\frac{B}{T}t^2)$ be the source signal and $x_i(t) = A_1cos(2\pi f_0(t - t_0) + \pi\frac{B}{T}(t - t_0)^2)$ be the signal received by the i^{th} device, where $B = f_1 - f_0$ is the bandwidth and T is the period of the chirp. We first compute the correlation $x_0(t)x_i(t)$ and then pass the result to a low-pass filter to acquire the low-frequency component:

$$LPF(x_0(t)x_i(t)) = \frac{1}{2}A_0A_1cos(2\pi f_0t_0 + \pi\frac{B}{T}(2t_0t - t_0^2)) \quad (1)$$

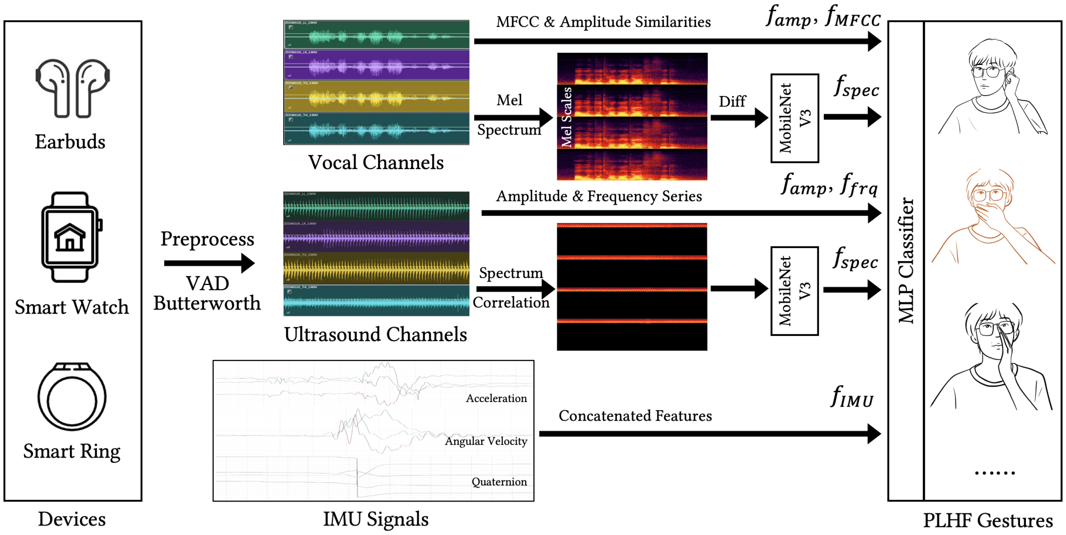


Figure 3: The sensing algorithm pipeline.

Note that $LPF(x_0(t)x_i(t))$ is a cosine function form of t with an amplitude of $\frac{1}{2}A_0A_1$ and a frequency of $2\frac{B}{T}t_0$, where the amplitude indicates the decay of the signal transmission and the frequency is proportional to the delay t_0 of the received signal. So we first compute the spectrum of $LPF(x_0(t)x_i(t))$ using short time Fourier transform (STFT). We extracted the following two features based on the spectrum: 1) the image feature f_{spec} of the spectrum using a pre-trained MobileNet V3 network and 2) the amplitude and frequency series f_{stats} (which is flattened into a 1-D vector) derived from the spectrum. Finally, we concatenate f_{spec} and f_{stats} into f_{ultra}, which can be either fed into a downstream classifier or combined with other features as mentioned above.

4.2.3 IMU Model. Devices worn on the user's hand, such as a watch and ring, help to capture the movement and attitude of the user's moving hand, thus beneficial for recognizing hand-to-mouth gestures. In our setting, we choose to mount a 9-axis wireless IMU on the ring as previous work [43] did. The IMU reports 3-axis acceleration, 3-axis angular velocity, and the quaternion at 200 Hz. For gesture recognition, we adopted a fixed window of 400 frames (or 2 seconds), concatenating the acceleration, angular velocity, and quaternion series into a 4000-length vector. Then we used a 3-layer MLP with the structure of $Dropout(0.5) \longrightarrow Linear(4000, 512) \longrightarrow ReLU \longrightarrow Linear(512, 512) \longrightarrow ReLU \longrightarrow Linear(512, 9)$ for classification.

4.3 Sensor Combination and Fusion Strategies

In consideration of the real-world deployment, we first figure out the reasonable device and sensor combinations. The devices include 1) left earbud (LE) with inner and outer microphones ($m_{l,i}$, $m_{l,o}$), 2) right earbud (RE) with inner and outer microphones ($m_{r,i}$, $m_{r,o}$), 3) watch with a microphone (m_w), and 4) ring with a microphone (m_r) and an IMU. Considering earbuds are most commonly used, we chose them as the primary device, which would work in different forms including two-side, one-side (wearing one earbud), and outer-only (for ones without active noise canceling). The introduction of the watch could be beneficial in providing an active ultrasound source as well as a hand-mounted microphone. Last, a ring device with an IMU and a microphone could track the movement of the hand and finger as well as provide a finger-mounted microphone. Based on the above observation, we devised four typical settings as follows for investigation: 1) single earbud (RE); 2) two earbuds (LE+RE); 3) two earbuds + watch (LE+RE+W); and 4) all devices (LE+RE+W+R).

For settings 3) and 4), since the active ultrasound source and the IMU enable the ultrasonic model and the IMU model, a fusion method is required to fuse different recognition models from different channels. We investigated two fusion strategies: 1) logit-level fusion and 2) feature-level fusion.

Let F_v, F_u, and F_i be the feature extractor network of vocal, ultrasonic, and IMU models respectively and C_v, C_u, and C_i be the corresponding multilayer classifier that outputs the logits. For logit-level fusion, the output logits are computed as

$$logits = a \cdot C_v(F_v(x_v)) + b \cdot C_u(F_u(x_u)) + c \cdot C_i(F_i(x_i)) \quad (2)$$

where a, b, and c are learnable weight parameters (x_v, x_u, and x_i are the corresponding channels of input). For feature-level fusion, the output logits are computed as

$$logits = C_{fuse}([F_v(x_v), F_u(x_u), F_i(x_i)]) \quad (3)$$

, where [*, *, *] refers to concatenation and C_{fuse} is another MLP classifier that takes the concatenated features as input and outputs the logits.

5 EVALUATION

In this section, we conducted a systematic evaluation on the cross-device sensing method illustrated in the previous section. We first built a cross-device VAHF dataset consisting of 10 users × 20 sentences × (8+1) gestures = 1800 samples. Then we evaluated our cross-device sensing method on the dataset in the dimensions of sensor combination, model selection, gesture reduction, and model ablation.

5.1 Participants and Apparatus

We recruited 10 participants (6 female and 4 male) with an average age of 21.2 (from 19 to 28, SD=2.64) and all participants were right-handed. All participants were recruited via emails and websites in our organization. We used a pair of earbuds with microphones, a smart watch with a motion sensor and a microphone, and a smart ring with a motion sensor and a microphone. The data of the microphones and motion sensors were fetched synchronously by a data collection thread.

5.1.1 *Microphones.*

We used the Sony WF-1000XM3 wireless noise-canceling headphone [1] and ZOOM H6 Handy recorder [2] in this paper. We used four one-channel TRS audio cables to connect to the feed-forward and feed-back microphones with headphones and a two-channel TRS audio cable to connect to the watch and the ring respectively. The ZOOM H6 audio recorder can record these six channels of timely synchronized audio data to a TF storage card (32 GB). The audio sampling rate was set to 48 KHz. To remain the same acoustic characteristic, we kept all the hardware in its original position in the earphone. The battery was run out of power to disable the on-chip software including the active noise canceling.

We use the MI Watch [3] with 3-axis accelerometer and 3-axis gyroscope at 100Hz. The data is kept locally on watch and would be pulled after each round of the experiment.

5.1.2 *Inertial Measurement Unit.*

We used a ring embedded with a wireless BMI-055 9-axis Inertial Measurement Unit (IMU) module, as shown in Figure 4. The IMU data (3-axis acceleration, 3-axis gyroscope data, 3-axis geomagnetic data, and 3-axis Euler angle, current system time) is transmitted to a PC with a Bluetooth module at 200Hz (460800 baud rate).

5.2 Data Collection

We collected gesture samples from 10 participants. The data collection entailed recording voice, ultrasound, and motion data while participants performed VAHF gestures corresponding to Section 3 and speak with daily voice commands. Each data collection study lasted about 60 minutes. Initially, participants were asked to read and sign consent forms. They were then shown instruction slides explaining the overall procedure of the data collection session and videos of VAHF gesture set from Section 3. Then we instructed participants to put on the earbuds, the smartwatch, and the ring properly, helping them to adjust the wearing until they felt comfortable with the devices.

Each participant was required to perform 15 gestures and record 10 voice commands for each gesture (150 gesture samples in total). For each gesture, the participant was shown a slide with the gesture's name, the 10 voice commands, and the posture (sitting or standing). The order of the gestures and the posture condition were randomly picked to remove the order effect. The 10 voice commands were randomly picked from the daily Siri voice commands[4], as shown in Table 2.

Table 2: The English translations of the voice commands used in data collection. The participants read the commands in Chinese. These voice commands were picked from Apple Siri's tutorial.

Index	Voice Command	Index	Voice Command
1	Text Mom.	11	Turn the temperature up to 24 degrees.
2	Read my messages.	12	Show the photos taken today.
3	Who is calling?	13	Find the popular restaurants nearby.
4	Set an alarm for eight o'clock.	14	What is the latest movie?
5	Pay with Apple Pay.	15	How to take a holiday on National Day?
6	Transfer 20 yuan to Amy.	16	Buy train tickets to Beijing.
7	Remind me to pick up the clothes.	17	How is the weather today?
8	What is my plan today?	18	Open Voice Memos.
9	Play my favorite song.	19	How to go to the nearest metro station?
10	Turn on the living room lights.	20	Countdown 20 minutes.

During the recording of the 10 gesture samples of each gesture, the experimenter first turned on the recording of the IMU ring, the watch's ultrasound, and the recorder. Then the participant clapped his or her hands to provide a synchronous signal used for the synchronization of different sensors. For each gesture sample, the experimenter first pressed a key on the PC to label a tick and record the system time, which was used for gesture sample segmentation, and then signaled the participant to perform the gesture and read the corresponding voice command while keeping the gesture. This process was repeated 10 times until the participant finished all 10 gesture samples. After that, the experimenter turned off the recording.

5.3 Data Preprocessing

Our data preprocessing process consisted of four steps: channel synchronization, segmentation, voice activation detection (VAD), and vocal-ultra sound separation. Below we illustrate the implementation details.

5.3.1 *Channel Synchronization.*

The synchronization between audio channels was achieved by the audio card in ZOOM H6. To synchronize the audio channels and the IMU channel, we required the user to clap their hand to provide a signal for alignment before starting data collection for each session. Then we located such a clapping peak in both the audio channels and the IMU channel to acquire the relative time shift. The peak in the audio channels and the IMU channel was detected by finding the first local maximum in the amplitude spectrogram and the acceleration spectrogram, respectively.

5.3.2 *Audio Segmentation and VAD.*

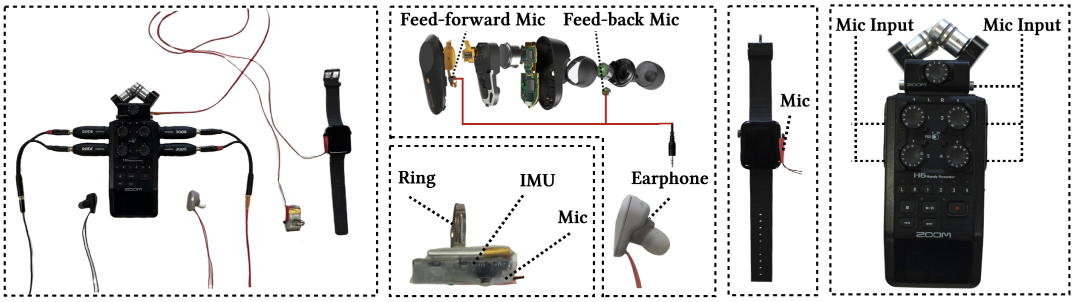
After aligning all the channels, we segment the audio data which includes 10 voice commands for each. This was achieved by simply separating a piece of audio using the keystroke points annotated by the participants during recording. After getting the coarse segmentation, we further ran a VAD algorithm [11] to remove the silent period at the two ends in each segment.

[1] https://www.sony.com/electronics/truly-wireless/wf-1000xm3

[2] https://zoomcorp.com/en/us/handheld-recorders/handheld-recorders/h6-audio-recorder/

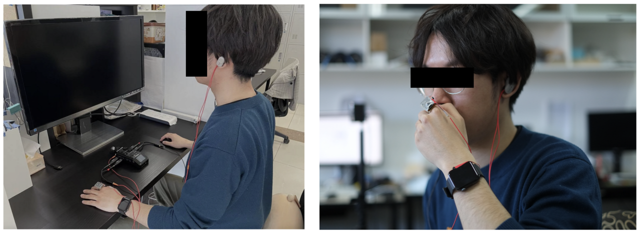
[3] https://www.mi.com/global/mi-watch-lite

[4] https://support.apple.com/siri



(a) (b) (c) (d) (e)

Figure 4: The apparatus of data collection. (a) Hardware overview. (b) A ring with an IMU and a microphone. (c) An earphone with a feed-forward microphone and a feed-back microphone. (d) A smartwatch with a microphone and a speaker. (e) A Zoom H6 recorder with six audio input channels.



(a) (b)

Figure 5: (a) Experiment setup. (b) A gesture example.

5.3.3 Separation of Ultrasound and Vocals. We used a Butterworth [5] highpass filter with 17500Hz cutoff frequency to separate the ultrasound and vocal from the audio data.

5.4 Evaluation Design

The evaluation consists of three sessions. In the first session, we conducted a two-factorial evaluation to analyze the recognition performance with regard to sensor combination and model selection. For sensor combination, corresponding to Section 4.3, we investigated five settings: 1) single (right) earbud with inner and outer microphones (RE, 2 audio channels), 2) two earbuds with inner and outer microphones (LE+RE, 4 audio channels), 3) two earbuds with outer microphones + watch (LE+RE+W, 3 audio channels), 4) all devices without the earbuds' inner channels (ALL-4ch, 4 audio channels), and 5) all devices with all channels (ALL-6ch, 6 audio channels). For model selection, we investigated the following six models: 1) vocal only (V), 2) ultrasound only (U), 3) IMU only (I), 4) vocal + ultrasound(V+U), 5) vocal + ultrasound + IMU with logit-level fusion (ALL-L), and 6) vocal + ultrasound + IMU with feature-level fusion (ALL-F). It is worth mentioning that the above two factors are correlated. The ultrasound channel would be activated unless the watch is used. Similarly, the IMU channel would be activated when the ring is used. Other factors, including the network structure, hyperparameters (max training epoch=100, dropout=0.5), and optimization strategies, are strictly controlled. We adopt three optimization strategies - pretraining, dropout, and warm-up to improve the performance and training robustness of our model. For pretraining, we initialized the MobileNet V3's parameters with the one pretrained on ImageNet [60]. For dropout, we added a dropout layer with a probability of 0.5 after the input layer to alleviate overfitting during training. For warm-up, we adopted a warm-up and weight decay strategy on the learning rate using the following piecewise function: if $n \leq 10$, then $lr(n) = 0.1 \times n \times lr(0)$, else $lr(n) = 0.97^{n-10} \times lr(0)$, where n is the training epoch and $lr(n)$ is the learning rate of the n^{th} epoch.

In the second session, we conducted an extensive evaluation on a reduced gesture set to analyze the optimal performance and usability of each sensor combination for practical deployment. The reduced gesture set contains three signature gestures - cover mouth with palm (G1), cover ear with arched palm (G2), and hold up the palm beside nose and mouth (G3) - which received high preference scores from the previous user study and intuitively had significant effects on the acoustic propagation. For each sensor combination, we chose the optimal model, as acquired above, to compute the classification accuracies of each gesture and all three gestures ({G1, E}, {G2, E}, {G3, E}, and {G1, G2, G3, E}, where E refers to the empty gesture). All the evaluation settings were consistent with the first session. Such an evaluation helps to ground the applicability value of the minimal functionality under different hardware settings.

In the last session, we conducted an ablation study for the optimal model to analyze the effects of the optimization strategies in the model design: 1) pretraining, 2) dropout, and 3) warm-up. After getting the optimal model above, we ran the model in the same setting except disabling 1) all the three optimizations, 2) pretraining, 3) dropout, and 4) warm-up to acquire the recognition accuracy in these 4 ablation settings. Such a study helped to validate the effectiveness of our model design.

All the above evaluations were conducted with leave-one-user-out cross-validation. For all the numerical comparisons, we reported the results along with the Wilcoxon Signed-Rank test to indicate the significance.

5.5 Results

Table 3 showed the results of the recognition performance regarding sensor combination and model selection.

[5] https://en.wikipedia.org/wiki/Butterworth_filter

Table 3: Evaluation results regarding different sensor combinations and model selection. The notions in the table are consistent with Section 5.4. The numbers in the table indicate recognition accuracies in % with standard deviations.

	V	U	I	V+U	ALL-L	ALL-F
RE	39.5(6.3)	-	-	-	-	-
LE+RE	70.3(10.0)	-	-	-	-	-
LE+RE+W	84.2(12.0)	52.4(11.5)	-	85.8(13.2)	-	-
ALL-4ch	89.9(10.5)	66.7(13.6)	49.0(5.3)	89.8(10.3)	90.8(10.2)	**91.5(8.9)**
ALL-6ch	90.0(10.9)	70.9(14.5)	49.0(5.3)	89.2(11.4)	90.7(9.9)	90.9(9.4)

For vocal-only models, we observed a constant increase in recognition accuracy as more sensor nodes were introduced (e.g., from 39.5% with a single earbud to 90.0% with all the sensors, $Z = -2.81, p < 0.05$). However, the difference between ALL-4ch and ALL-6ch was not significant, meaning when multiple devices were used, the introduction of the earbuds' inner channels brings limited information for the vocal channel. For ultrasonic-only models, the performance increased from 52.5% to 70.9% ($Z = -2.81, p < 0.05$) as the ring microphone and the earbuds' inner microphones were added. Notably, the independent use of the ultrasonic channels has its unique advantage of not relying on the vocal feature so that the model can still work well in scenarios such as noisy environments and whispering. The IMU model achieved an accuracy of 49.0%, meaning the IMU could provide complementary information on hand and finger movement, though far from practical as an individual model.

As for the sensor fusion models, we notice the vocal+ultra model had a performance increase over the vocal-only model with fewer input channels (LE+RE+W, 84.2% V.S. 85.8%, $Z = -1.64, p = 0.1$), while it had no increase for ALL-4ch (89.9% V.S. 89.8%, $Z = -0.18, p = 0.86$) and had a decrease for ALL-6ch (90.0% V.S. 89.2%, $Z = -0.98, p = 0.33$). This is probably because the vocal-only model with multiple channels (e.g., 6 channels) is a strong baseline, and combining it with an inferior model would introduce additional noise. Regarding all-channel fusion, we found feature-level fusion slightly outperformed logit-level fusion in accuracy (91.5% V.S. 90.8%, $Z = -0.98, p = 0.33$), probably due to the larger parameter space. We also observed a slight performance decrease for fusion models when adding the inner channels of the earbuds to ALL-4ch (91.5% V.S. 90.9%, $Z = -0.36, p = 0.72$), although the difference was not significant. The optimal model (all-channel feature-level fusion for ALL-4ch) achieved a 9-class recognition accuracy of 91.5%, which significantly outperformed the vocal-only model ($Z = -1.96, p < 0.05$) with the same channels.

To ground a better understanding of how each channel (vocal, ultrasound, and IMU) contributed to the recognition, we analyzed the confusion matrix of four models (vocal-only, ultra-only, IMU-only, and feature-level fusion) under ALL-4ch, as shown in Figure 6. This result was understandable because for the gestures with larger confusion, we could easily figure out their similarity based on semantics. For example, gesture pairs $(0, 4)$ and $(3, 7)$ yield larger confusion for vocal and ultrasound models, where we observed similar touch positions for each pair of gestures (ear for $(0, 4)$ and mouth for $(3, 7)$). Gesture 1 confuses with gestures 3 and 7 in the ultrasound model probably due to a similar hand position, though it yields less confusion for the vocal model probably due to different occlusion levels (gestures 3 and 7 yield greater occlusion) that may influence the frequency response of the human voice.

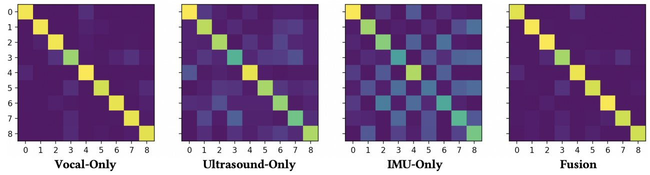


Figure 6: The confusion matrix of different models: vocal-only, ultra-only, IMU-only, and feature-level fusion. 0-9 represent the following gestures respectively: 0 - pinch the ear rim, 1 - calling gesture, 2 - support cheek with palm, 3 - cover mouth with palm, 4 - cover ear with arched palm, 5 - thinking face gesture, 6 - hold up the palm beside nose and mouth, 7 - cover mouth with fist, and 8 - empty gesture.

Table 4: Recognition accuracy on the reduced gesture set. G1: cover mouth with palm, G2: cover ear with arched palm, and G3: hold up the palm beside nose and mouth.

	RE	LE+RE	LE+RE+W	ALL-4ch
G1	82.3(11.4)	92.1(12.8)	97.8(6.7)	95.7(7.3)
G2	83.0(13.0)	98.8(1.9)	97.9(6.3)	100.0(0.0)
G3	75.5(26.7)	90.6(9.0)	94.2(8.4)	100.0(0.0)
G1+G2+G3	64.4(10.6)	87.3(8.9)	91.3(11.3)	97.3(4.5)

Results on the reduced gesture set were shown in Table 4. Since ALL-4ch achieved higher recognition accuracy than ALL-6ch in the fusion model (e.g., 91.5% V.S. 90.9%), we dropped ALL-6ch in this table. We had the following observations: 1) Using one earbud with inner and outer microphones (RE), which is a severely restricted setting, could achieve a narrowly applicable accuracy of over 80% for recognizing a specific single gesture (82.3% for G1 and 83.0% for G2) while it performed worse in recognizing other gesture (e.g., G3) or multiple gestures, which is understandable due to limited sensing information. 2) Using a pair of earbuds (LE+RE) could significantly boost the performance, with promising accuracies of 87.3% for recognizing all three gestures and 98.8% for recognizing G2, indicating the high applicability of such compact hardware form. 3) Additional hardware including a watch and ring brought the feasibility of fusing more input channels (e.g., ultrasound), which constantly improved the performance to a highly robust one (e.g., 97.3% for recognizing all three gestures and 100% for recognizing G2 and G3) and meanwhile lifting the distinguishable gesture space (e.g., from 3 gestures to 8 gestures, see Table 3) with high applicability (e.g., 91.5% for simultaneously recognizing 8 gestures). The above results showed a leap over previous work with similar interaction modality (e.g., PrivateTalk [76]), revealing the feasibility of broadened gesture space (e.g., recognizing 8 gestures simultaneously) and the effectiveness of multi-device sensing.

The results of the ablation study are shown in Table 5. We found disabling pretraining, dropout, and warm-up caused different levels of performance degradation. Disabling pretraining caused the most significant decrease in performance (-14.3%, $Z = -2.81, p < 0.05$), which is probably because the feature extractor network (MobileNet

Table 5: Results of the ablation study. The numbers in the table indicate recognition accuracies in % with standard deviations.

Techniques	Accuracy
No Optimization	75.8(12.2)
No Pretraining	77.2(12.8)
No Dropout	91.2(7.8)
No warm-up	87.8(11.1)
Full Model	91.5(8.9)

V3) with pretraining on large-scale datasets could better extract different levels of image features. Meanwhile, disabling dropout caused slight decrease of 0.3% ($Z = -0.18, p = 0.86$), which was not significant, and disabling warm-up caused a decrease of 3.7% ($Z = -2.67, p < 0.05$). The introduction of warm-up and dropout aims to optimize the training procedure (e.g., alleviating overfitting) and improve the robustness of the model. Compared with the raw model with no optimization, our model achieved a significant increase of 15.7% ($Z = -2.81, p < 0.05$), showing the superiority of all the optimization techniques.

6 APPLICATION SCENARIOS

To demonstrate the applicability of VAHF gestures in voice interaction, we first presented the interaction space created by VAHF gestures along with example real-life scenarios. Then we discussed the design considerations and implications regarding the deployment of VAHF gestures in real practice.

6.1 Interaction Space and Scenario Description

The introduction of VAHF gestures achieves the unique benefit of assigning a multi-class label to speech segments, which brings great potential to broaden the traditional voice interaction space in the following aspects.

6.1.1 VAHF Gestures as Modality Control Signals. **Wakeup-free interaction.** The most intuitive function for modality control in voice interface is to use hand-to-face gestures (e.g., covering the mouth) to indicate whether the current speech is with interaction intention that should be processed by the voice assistant, which has been achieved and widely researched by previous work [59, 76, 78]. In our work, VAHF gestures have the inherited capability to support wakeup-free interaction simply by assigning one of the gestures for the wakeup state control.

Dynamic modality control in multi-round interaction with voice assistant. We demonstrate an example scenario using VAHF gestures for dynamic modality control in the multi-round dialog that has never been achieved before. When the user is enrolled in a multi-round dialog with the voice assistant, the complexity of the interaction behavior increases significantly. For example, in a specific dialog round, the user has different options to proceed with the dialog: 1) appending - the user appends a voice command and expects the voice assistant to process the command based on the dialog context in the regular order; 2) interrupting - the user wants to interrupt the current dialog (e.g., the voice assistant stops immediately and waits for new voice commands) and start a new dialog (abandoning the dialog context) with the new commands; and 3) editing - the user wants the voice assistant to edit the commands that they previously asked based on the dialog context and the brief editing command (e.g., the user says "How is the weather today?" When the assistant is answering, the user adds an editing command "No, I mean tomorrow."). Since our technique enables a channel width of up to 9 gestures (including the empty gesture) as modality input, we can assign different VAHF gestures to the three modalities of voice input - appending (e.g., covering the mouth), interrupting (e.g., covering the earphone), and editing (e.g., holding up the palm beside the mouth) - in the multi-round dialog scenario to enable more flexible and intelligent voice interaction flow.

6.1.2 Binding Shortcuts to VAHF Gestures. **VAHF Gestures as UI shortcuts.** Simulating the execution of certain interaction paths through voice commands is a prevalent form of voice interaction on smartphones and wearable devices. When an interaction path takes a text entry slot or a period of raw speech as the input, it can be replaced with certain VAHF gestures. For example, the user can define the "phone call" VAHF gesture as opening WeChat and sending a voice message of the user's raw speech to Alice. Another example is to define the "thinking face" gesture as opening the Google website and searching for the text transcribed from the raw speech input. Such replacements of complex UI shortcuts with VAHF gestures could potentially reduce the repetition of the interaction path in speech, especially in a multi-round interaction.

Registration and reservation of the VAHF-gesture-enabled shortcut session. Regarding the binding of shortcuts with VAHF gestures, a more exciting design question is how the VAHF gestures are binded in real-world practice. Normally, the binding is fixed and can be set by the GUI (e.g., on a smartphone). On the contrary, we here present a dynamic registration and reservation mechanism for VAHF-gesture-enabled shortcut sessions, which are worthy of extensive exploration. In such a mechanism, for an unbinded VAHF gesture, when the user performs the gesture while narrating the full voice command, the voice assistant would automatically extract the UI shortcut from the command and bind it with the performed gesture. Later when the user wants to access the shortcut for a second time, they could perform the binded gesture while saying the input slot instead of the full command. The session and the dialog context are fully preserved for the gesture until a new command with UI shortcut semantics is input. The voice assistant would ask the user whether to update the binding of the gesture to a new shortcut. We believed such a design of a dynamic registration mechanism for VAHF gestures would benefit memorability, flexibility, and lower cold-start cost.

6.1.3 VAHF Gestures as Spatial Indicators. VAHF gestures in voice interaction are also capable of indicating the target to interact with from the multiple interactable devices or elements. For example, in an IoT scenario where multiple voice-interactable devices (e.g., a smartphone, a TV, and a smart speaker) are in the same room, the user could perform different VAHF gestures with voice commands to trigger voice interaction with different devices. Similarly, in a complex UI control scenario (e.g., filling in a form with multiple text boxes), a VAHF gesture is displayed beside each text box, and the user could perform the corresponding VAHF gesture to input a particular text box.

6.2 Design Considerations for VAHF Gestures to Enhance Voice Interaction

The VAHF gestures proposed in our paper open the opportunity to design novel voice interactions for mobile, wearable, and HMD devices that allow users to quickly switch among modalities, accelerate common tasks, and manage multi-device interaction in different scenarios. We discussed two issues regarding the real-world deployment of VAHF gestures. **1) Combination strategy for better performance.** Although VAHF gestures have shown great potential in applicability, simply adding on all the functions described in the previous section is not feasible due to the channel capacity and the recognition accuracy. For example, as shown in Tables 3 and 4, an accuracy of 91.5% for 9 classes is not yet highly usable, but a 4-class sub-gesture set achieved an accuracy of 97.3%, which is considered highly usable. Therefore, a fine-grained design on the selection of gestures (e.g., alleviating using two gestures with higher confusion at the same time) and the switch of gesture sets in different scenarios is key to implementing a highly usable VAHF-gesture-enhanced voice interaction system. **2) Scalability and extensibility.** Although we only investigated an optimized VAHF gesture set with 8 gestures in our work, our sensing method was open to absorbing other extensive VAHF gestures. Our analysis method in Sections 3.1 and 3.2 provide a practical design guideline to elicit new gestures and analyze their feasibility. Further, our framework of recognizing VAHF gestures by multiple wearable devices has the advantage of appending or cutting down certain sensing channels easily, so the gesture set should be scalable and convertible for the system's flexibility.

7 DISCUSSION AND LIMITATIONS

7.1 Form Factor for Deployment

Currently, our sensing algorithms were run and evaluated in an offline setting with a full-functional prototype. As the instantiation, we also implemented a prototypical realtime VAHF gesture recognition system with the devices shown in Fig. 4, a laptop, and a GPU server. The Zoom H6 recorder served as an audio card that streamed realtime audio data to the laptop. The PC ran a realtime data prepocessing program (written in Python, similar to Section 5.3) on one 2.3GHz Intel CPU core. The PC sent the processed audio segment to the GPU server using a sliding-window strategy while the recognition model processed the audio segment on one Nvidia RTX 3090 GPU and sent the results back to the PC. The whole pipeline ran at 60FPS with a delay of less than 50ms (excluding the network delay). The actual FPS could be controlled by adjusting the stride of the sliding window (typically an FPS higher than 5 could provide a good immediate experience).

Although our work demonstrated the computational feasibility of recognizing VAHF gestures, we should further consider the form factors for real-life deployment regarding synchronization, channel access, computational complexity, etc. We discussed the following three questions:

(1) How to synchronize and transmit the signal from different channels? In our implementation, we used a strong synchronization system, where all the audio channels were wired to the audio card of ZOOM H6. In real deployment, the system could be implemented using Bluetooth low energy (BLE) technique for real-time signal transmission and synchronization (e.g., using broadcast mode or mesh mode for communication[6]), allowing dynamic communication among devices.

(2) How to determine the proper channels or devices to enable in different scenarios? As discussed in Section 6, we suggest the channels and devices should be enabled dynamically based on context information (e.g., the activated devices and the surrounding environment). The system would provide multiple levels of interaction progressively based on the activated devices (e.g., more complex gesture set for more devices) while preserving certain environment constraints (e.g., avoid using ultrasound in quiet scenarios or degrading the interaction capability in a noisy environment). With such context-aware optimizations, our technique could be implemented in a more user- and energy-friendly manner.

(3) How to reduce the computational complexity? In our implementation, we chose MobileNet V3, a light-weight NN model capable for mobile devices, as the backbone model in consideration of the computational efficiency. Further, there are three possible ways to reduce the computational complexity: 1. using dynamic channels (e.g., using the minimal channels in an efficient mode); 2. using more light-weight feed-forward NN models (e.g., ShuffleNet[81]) for recognition; and 3. adopting bottom-level optimization (e.g., parameter quantization [15, 16] or customized hardware such as FPGA [4]).

7.2 Robustness against Environmental Interference

Currently, our data were collected in an indoor environment with no background noise, aiming to validate the feasibility of recognizing VAHF gestures in an ideal setting. For real-world deployment, the recognition model is expected to deal with more complicated data with lower signal-noise ratio (SNR) and more environmental noise. So further research on the effect of environmental interference and how to build a robust recognition model should be conducted. Two strategies - 1) training the model with more diverse data coming from real-world scenarios or synthesization; 2) using advanced preprocessing techniques (e.g., active noise canceling algorithms) to reduce the noise and improve the SNR - may resolve this issue, which are worthy of further investigation.

7.3 Ultrasound Usage

We were well acknowledged the use of ultrasound for sensing could be controversial due to the interference and damage to one's hearing. In our work, we used a chirp signal from 17.5KHz to 22.5KHz and we noticed in our data collection procedure, some of the participants could hear the ultrasound and found it annoying. Further, ultrasound at sufficient sound pressure levels exert underlying danger of hearing damage even if it cannot be heard (though we strictly controlled the ultrasound amplitude in our study). Therefore, the use of ultrasound, including the amplitude, frequency, and duration should be more carefully designed for a gesture recognition system. More research should be conducted to explore the use of ultrasound and alternative sensing methods.

[6] https://www.bluetooth.com/learn-about-bluetooth/tech-overview/

8 CONCLUSION

In this paper, we investigated the design space and the recognition method of voice-accompanying hand-to-face (VAHF) gestures to enhance voice interaction with parallel gesture channels. To design VAHF gestures, we first conducted an elicitation study, resulting in a total proposal of 15 gestures, followed by a hierarchical analysis process to output the most salient 8 gestures with the least ambiguity and physical confusion. Then we proposed a novel cross-device sensing method fusing different sensor channels to recognize para-linguistic hand-to-face gestures, achieving a high recognition accuracy of 97.3% for 3+1(empty) gestures and 91.5% for 8+1(empty) gestures recognition on our cross-device VAHF dataset. The uniqueness of our work is that we explored a broadened and scalable VAHF-gesture-based interaction space, which remains under-researched, to facilitate voice interaction in a more diverse manner (e.g., defining a shortcut or parsing parameters). Compared with prior work [59, 76] where a specific gesture (e.g., bringing the phone to the mouth[59]) was designed and recognized for 1-bit modality control (e.g., activating the voice assistant), our multi-device sensing framework is not only capable for recognizing up to 8 VAHF gestures simultaneously from the hand-off "empty" gesture, but also benefits from the scalability (e.g., adding a device or adding a gesture is easy under our framework). As mobile devices and scenarios are becoming prevalent these years, voice input has become an essential modality of pervasive interaction. We envision our work would further enhance the efficiency and capability of current voice interaction and serve an important role in the future voice interaction of various scenarios like AR and IoT.

ACKNOWLEDGMENTS

This work is supported by the Natural Science Foundation of China (NSFC) under Grant No. 62132010 and No. 62002198, Tsinghua University Initiative Scientific Research Program, Beijing Key Lab of Networked Multimedia, and Institute for Artificial Intelligence, Tsinghua University.

REFERENCES

[1] Karan Ahuja, Andy Kong, Mayank Goel, and Chris Harrison. 2020. Direction-of-Voice (DoV) Estimation for Intuitive Speech Interaction with Smart Devices Ecosystems. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 1121–1131. https://doi.org/10.1145/3379337.3415588

[2] Toshiyuki Ando, Yuki Kubo, Buntarou Shizuki, and Shin Takahashi. 2017. CanalSense: Face-related movement recognition system based on sensing air pressure in ear canals. *UIST 2017 - Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (2017), 679–689. https://doi.org/10.1145/3126594.3126649

[3] Donald J Berndt and James Clifford. 1994. Using dynamic time warping to find patterns in time series.. In *KDD workshop*, Vol. 10. Seattle, WA, USA:, 359–370.

[4] Merwan Birem and François Berry. 2014. DreamCam: A modular FPGA-based smart camera architecture. *Journal of Systems Architecture* 60, 6 (2014), 519–527. https://doi.org/10.1016/j.sysarc.2014.01.006

[5] Richard A Bolt. 1980. "Put-That-There": Voice and Gesture at the Graphics Interface. *SIGGRAPH Comput. Graph.* 14, 3 (1980), 262–270. https://doi.org/10.1145/965105.807503

[6] Marie-Luce Bourguet and Akio Ando. 1998. Synchronization of Speech and Hand Gestures during Multimodal Human-Computer Interaction. In *CHI 98 Conference Summary on Human Factors in Computing Systems (CHI '98)*. Association for Computing Machinery, New York, NY, USA, 241–242. https://doi.org/10.1145/286498.286726

[7] Daniel Buschek, Bianka Roppelt, and Florian Alt. 2018. *Extending Keyboard Shortcuts with Arm and Wrist Rotation Gestures*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173595

[8] 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

[9] Enea Ceolini, Charlotte Frenkel, Sumit Bam Shrestha, Gemma Taverni, Lyes Khacef, Melika Payvand, and Elisa Donati. 2020. Hand-gesture recognition based on EMG and event-based camera sensor fusion: A benchmark in neuromorphic computing. *Frontiers in Neuroscience* 14 (2020), 637.

[10] Enea Ceolini, Gemma Taverni, Lyes Khacef, Melika Payvand, and Elisa Donati. 2019. Sensor fusion using EMG and vision for hand gesture classification in mobile applications. In *2019 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 1–4.

[11] Joon-Hyuk Chang, Nam Soo Kim, and Sanjit K Mitra. 2006. Voice activity detection based on multiple statistical models. *IEEE Transactions on Signal Processing* 54, 6 (2006), 1965–1976.

[12] Wook Chang, Kee Eung Kim, Hyunjeong Lee, Joon Kee Cho, Byung Seok Soh, Jung Hyun Shim, Gyunghye Yang, Sung-Jung Cho, and Joonah Park. 2006. Recognition of grip-patterns by using capacitive touch sensors. In *2006 IEEE International Symposium on Industrial Electronics*, Vol. 4. IEEE, 2936–2941.

[13] Victor Chen, Xuhai Xu, Richard Li, Yuanchun Shi, Shwetak Patel, and Yuntao Wang. 2021. Understanding the Design Space of Mouth Microgestures. In *Designing Interactive Systems Conference 2021* (Virtual Event, USA) *(DIS '21)*. Association for Computing Machinery, New York, NY, USA, 1068–1081. https://doi.org/10.1145/3461778.3462004

[14] J. Guillermo Colli-Alfaro, Anas Ibrahim, and Ana Luisa Trejos. 2019. Design of User-Independent Hand Gesture Recognition Using Multilayer Perceptron Networks and Sensor Fusion Techniques. In *2019 IEEE 16th International Conference on Rehabilitation Robotics (ICORR)*. 1103–1108. https://doi.org/10.1109/ICORR.2019.8779533

[15] Matthieu Courbariaux, Yoshua Bengio, and Jean-Pierre David. 2014. Training deep neural networks with low precision multiplications. arXiv:1412.7024 [cs.LG]

[16] Tim Dettmers. 2015. 8-Bit Approximations for Parallelism in Deep Learning. arXiv:1511.04561 [cs.NE]

[17] S. Fujie, Y. Ejiri, K. Nakajima, Y. Matsusaka, and T. Kobayashi. 2004. A conversation robot using head gesture recognition as para-linguistic information. In *RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No.04TH8759)*. 159–164. https://doi.org/10.1109/ROMAN.2004.1374748

[18] Shinya Fujie, Toshihiko Yamahata, and Tetsunori Kobayashi. 2006. Conversation Robot with the Function of Gaze Recognition. In *2006 6th IEEE-RAS International Conference on Humanoid Robots*. 364–369. https://doi.org/10.1109/ICHR.2006.321298

[19] Yang Gao, Yincheng Jin, Jiyang Li, Seokmin Choi, and Zhanpeng Jin. 2020. EchoWhisper: Exploring an Acoustic-Based Silent Speech Interface for Smartphone Users. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 3 (2020). https://doi.org/10.1145/3411830

[20] Nicholas Gillian, Sara Pfenninger, Spencer Russell, and Joseph A. Paradiso. 2014. Gestures Everywhere: A Multimodal Sensor Fusion and Analysis Framework for Pervasive Displays. In *Proceedings of The International Symposium on Pervasive Displays* (Copenhagen, Denmark) *(PerDis '14)*. Association for Computing Machinery, New York, NY, USA, 98–103. https://doi.org/10.1145/2611009.2611032

[21] Jun Gong, Aakar Gupta, and Hrvoje Benko. 2020. *Acustico: Surface Tap Detection and Localization Using Wrist-Based Acoustic TDOA Sensing*. Association for Computing Machinery, New York, NY, USA, 406–419. https://doi.org/10.1145/3379337.3415901

[22] 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

[23] Yizheng Gu, Chun Yu, Zhipeng Li, Weiqi Li, Shuchang Xu, Xiaoying Wei, and Yuanchun Shi. 2019. Accurate and Low-Latency Sensing of Touch Contact on Any Surface with Finger-Worn IMU Sensor. 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, 1059–1070. https://doi.org/10.1145/3332165.3347947

[24] Yizheng Gu, Chun Yu, Zhipeng Li, Zhaoheng Li, Xiaoying Wei, and Yuanchun Shi. 2020. QwertyRing: Text Entry on Physical Surfaces Using a Ring. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 4, Article 128 (Dec. 2020), 29 pages. https://doi.org/10.1145/3432204

[25] Sidhant Gupta, Dan Morris, Shwetak N. Patel, and Desney Tan. 2012. SoundWave: Using the Doppler effect to sense gestures. In *Conference on Human Factors in Computing Systems - Proceedings*. 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*. 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, 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 (UIST '11)*. Association for Computing Machinery, New York, NY, USA, 627–636. https://doi.org/10.1145/2047196.2047279

[29] Chris Harrison, Robert Xiao, and Scott E. Hudson. 2012. Acoustic barcodes: Passive, durable and inexpensive notched identification tags. *UIST'12 - Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology* (2012), 563–567.

[30] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. 2017. CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 131–135. https://doi.org/10.1109/ICASSP.2017.7952132

[31] Hossein Mousavi Hondori, Maryam Khademi, and Cristina V Lopes. 2012. Monitoring intake gestures using sensor fusion (microsoft kinect and inertial sensors) for smart home tele-rehab setting. In *2012 1st Annual IEEE Healthcare Innovation Conference*.

[32] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. 2019. Searching for MobileNetV3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[33] 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

[34] Athanasios Katsamanis, Vassilis Pitsikalis, Stavros Theodorakis, and Petros Maragos. 2017. *Multimodal Gesture Recognition*. Association for Computing Machinery and Morgan & Claypool, 449–487. https://doi.org/10.1145/3015783.3015796

[35] Takashi Kikuchi, Yuta Sugiura, Katsutoshi Masai, Maki Sugimoto, and Bruce H. Thomas. 2017. EarTouch: Turning the Ear into an Input Surface. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Vienna, Austria) *(MobileHCI '17)*. Association for Computing Machinery, New York, NY, USA, Article 27, 6 pages. https://doi.org/10.1145/3098279.3098538

[36] Han-Jong Kim, Seijin Cha, Richard C. Park, Tek-Jin Nam, Woohun Lee, and Geehyuk Lee. 2016. Mo-Bi: Contextual Mobile Interfaces through Bimanual Posture Sensing with Wrist-Worn Devices. In *Proceedings of HCI Korea* (Jeongseon, Republic of Korea) *(HCIK '16)*. Hanbit Media, Inc., Seoul, KOR, 94–99. https://doi.org/10.17210/hcik.2016.01.94

[37] Gierad Laput, Karan Ahuja, Mayank Goel, and Chris Harrison. 2018. Ubicoustics: Plug-and-play acoustic activity recognition. *UIST 2018 - Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (2018), 213–224. https://doi.org/10.1145/3242587.3242609

[38] 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

[39] 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

[40] Juyoung Lee, Hui-Shyong Yeo, Murtaza Dhuliawala, Jedidiah Akano, Junichi Shimizu, Thad Starner, Aaron Quigley, Woontack Woo, and Kai Kunze. 2017. Itchy Nose: Discreet Gesture Interaction Using EOG Sensors in Smart Eyewear. In *Proceedings of the 2017 ACM International Symposium on Wearable Computers* (Maui, Hawaii) *(ISWC '17)*. Association for Computing Machinery, New York, NY, USA, 94–97. https://doi.org/10.1145/3123021.3123060

[41] Haobo Li, Xiangpeng Liang, Aman Shrestha, Yuchi Liu, Hadi Heidari, Julien Le Kernec, and Francesco Fioranelli. 2020. Hierarchical Sensor Fusion for Micro-Gesture Recognition With Pressure Sensor Array and Radar. *IEEE Journal of Electromagnetics, RF and Microwaves in Medicine and Biology* 4, 3 (2020), 225–232. https://doi.org/10.1109/JERM.2019.2949456

[42] Chen Liang, Chi Hsia, Chun Yu, Yukang Yan, Yuntao Wang, and Yuanchun Shi. 2022. DRG-Keyboard: Enabling Subtle Gesture Typing on the Fingertip with Dual IMU Rings. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 4, Article 170 (dec 2022), 30 pages. https://doi.org/10.1145/3569463

[43] 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

[44] 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 (July 2016), 19 pages. https://doi.org/10.1145/2897824.2925953

[45] Mingyu Liu, Mathieu Nancel, and Daniel Vogel. 2015. Gunslinger: Subtle arms-down mid-air interaction. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. 63–71.

[46] W Liu, W Shen, B Li, and L Wang. 2019. Toward Device-Free Micro-Gesture Tracking via Accurate Acoustic Doppler-Shift Detection. *IEEE Access* 7 (2019), 1084–1094. https://doi.org/10.1109/ACCESS.2018.2886279

[47] Yilin Liu, Shijia Zhang, and Mahanth Gowda. 2021. NeuroPose: 3D Hand Pose Tracking using EMG Wearables. In *Proceedings of the Web Conference 2021*. 1471–1482.

[48] Yiqin Lu, Bingjian Huang, Chun Yu, Guahong Liu, and Yuanchun Shi. 2020. Designing and Evaluating Hand-to-Hand Gestures with Dual Commodity Wrist-Worn Devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 1, Article 20 (March 2020), 27 pages. https://doi.org/10.1145/3380984

[49] Yanpeng Lv, Shangfei Wang, and Peijia Shen. 2011. A Real-Time Attitude Recognition by Eye-Tracking. In *Proceedings of the Third International Conference on Internet Multimedia Computing and Service* (Chengdu, China) *(ICIMCS '11)*. Association for Computing Machinery, New York, NY, USA, 170–173. https://doi.org/10.1145/2043674.2043723

[50] Marwa Mahmoud and Peter Robinson. 2011. Interpreting Hand-Over-Face Gestures. In *Affective Computing and Intelligent Interaction*, Sidney D'Mello, Arthur Graesser, Björn Schuller, and Jean-Claude Martin (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 248–255.

[51] Wenguang Mao, Jian He, and Lili Qiu. 2016. Cat: high-precision acoustic motion tracking. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*. 69–81.

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

[53] Jess McIntosh, Charlie McNeill, Mike Fraser, Frederic Kerber, Markus Löchtefeld, and Antonio Krüger. 2016. EMPress: practical hand gesture classification with wrist-mounted EMG and pressure sensing. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2332–2342.

[54] Annamaria Mesaros, Toni Heittola, Antti Eronen, and Tuomas Virtanen. 2010. Acoustic event detection in real life recordings. *European Signal Processing Conference* (2010), 1267–1271.

[55] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. 2020. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *European Conference on Computer Vision*. Springer, 548–564.

[56] Louis-Philippe Morency, Candace Sidner, Christopher Lee, and Trevor Darrell. 2005. Contextual Recognition of Head Gestures. In *Proceedings of the 7th International Conference on Multimodal Interfaces* (Torento, Italy) *(ICMI '05)*. Association for Computing Machinery, New York, NY, USA, 18–24. https://doi.org/10.1145/1088463.1088470

[57] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[58] Makoto Ono, Buntarou Shizuki, and Jiro Tanaka. 2013. Touch & activate: adding interactivity to existing objects using active acoustic sensing. *Proceedings of the 26th annual ACM symposium on User interface software and technology* (2013), 31–40. https://doi.org/10.1145/2501988.2501989

[59] Yue Qin, Chun Yu, Zhaoheng Li, Mingyuan Zhong, Yukang Yan, and Yuanchun Shi. 2021. ProxiMic: Convenient Voice Activation via Close-to-Mic Speech Detected by a Single Microphone. 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 8, 12 pages. https://doi.org/10.1145/3411764.3445687

[60] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115, 3 (2015), 211–252.

[61] Tomasz Sapiński, Dorota Kamińska, Adam Pelikant, Cagri Ozcinar, Egils Avots, and Gholamreza Anbarjafari. 2019. Multimodal Database of Emotional Speech, Video and Gestures. In *Pattern Recognition and Information Forensics*, Zhaoxiang Zhang, David Suter, Yingli Tian, Alexandra Branzan Albu, Nicolas Sidère, and Hugo Jair Escalante (Eds.). Springer International Publishing, Cham, 153–163.

[62] Pablo Sauras-Perez, Andrea Gil, Jasprit Singh Gill, Pierluigi Pisu, and Joachim Taiber. 2017. VoGe: A Voice and Gesture System for Interacting with Autonomous Cars. In *WCX™ 17: SAE World Congress Experience*. SAE International. https:

//doi.org/10.4271/2017-01-0068

[63] 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)*. Association for Computing Machinery, New York, NY, USA, 3181–3190. https://doi.org/10.1145/2556288.2556984

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

[65] Sanjeev Kadagathur Vadiraj, Achuth Rao M .V., and Prasanta Kumar Ghosh. 2020. Automatic Identification of Speakers From Head Gestures in a Narration. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6314–6318. https://doi.org/10.1109/ICASSP40776.2020.9053124

[66] Tran Huy Vu, Archan Misra, Quentin Roy, Kenny Choo Tsu Wei, and Youngki Lee. 2018. Smartwatch-Based Early Gesture Detection 8 Trajectory Tracking for Interactive Gesture-Driven Applications. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 1, Article 39 (March 2018), 27 pages. https://doi.org/10.1145/3191771

[67] 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

[68] Z Wang, Y Hou, K Jiang, W Dou, C Zhang, Z Huang, and Y Guo. 2019. Hand Gesture Recognition Based on Active Ultrasonic Sensing of Smartphone: A Survey. *IEEE Access* 7 (2019), 111897–111922. https://doi.org/10.1109/ACCESS.2019.2933987

[69] Jamie A. Ward, Paul Lukowicz, and Gerhard Tröster. 2005. Gesture Spotting Using Wrist Worn Microphone and 3-Axis Accelerometer. In *Proceedings of the 2005 Joint Conference on Smart Objects and Ambient Intelligence: Innovative Context-Aware Services: Usages and Technologies* (Grenoble, France) *(sOc-EUSAI '05)*. Association for Computing Machinery, New York, NY, USA, 99–104. https://doi.org/10.1145/1107548.1107578

[70] Hongyi Wen, Julian Ramos Rojas, and Anind K. Dey. 2016. *Serendipity: Finger Gesture Recognition Using an Off-the-Shelf Smartwatch*. Association for Computing Machinery, New York, NY, USA, 3847–3851. https://doi.org/10.1145/2858036.2858466

[71] 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*. 1–14.

[72] Jason Wu, Chris Harrison, Jeffrey P. Bigham, and Gierad Laput. 2020. Automated Class Discovery and One-Shot Interactions for Acoustic Activity Recognition. *Conference on Human Factors in Computing Systems - Proceedings* (2020). https://doi.org/10.1145/3313831.3376875

[73] 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 &#38; Services (MobileHCI '14)*. ACM, New York, NY, USA, 67–76. https://doi.org/10.1145/2628363.2628383

[74] 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*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376836

[75] Koki Yamashita, Takashi Kikuchi, Katsutoshi Masai, Maki Sugimoto, Bruce H. Thomas, and Yuta Sugiura. 2017. CheekInput: Turning Your Cheek into an Input Surface by Embedded Optical Sensors on a Head-Mounted Display. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology* (Gothenburg, Sweden) *(VRST '17)*. Association for Computing Machinery, New York, NY, USA, Article 19, 8 pages. https://doi.org/10.1145/3139131.3139146

[76] Yukang Yan, Chun Yu, Yingtian Shi, and Minxing Xie. 2019. PrivateTalk: Activating voice input with hand-on-mouth gesture detected by bluetooth earphones. In *UIST 2019 - Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. 1013–1020. https://doi.org/10.1145/3332165.3347950

[77] Yukang Yan, Chun Yu, Wengrui Zheng, Ruining Tang, Xuhai Xu, and Yuanchun Shi. 2020. *FrownOnError: Interrupting Responses from Smart Speakers by Facial Expressions*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376810

[78] Zhican Yang, Chun Yu, Fengshi Zheng, and Yuanchun Shi. 2019. ProxiTalk: Activate Speech Input by Bringing Smartphone to the Mouth. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3, Article 118 (sep 2019), 25 pages. https://doi.org/10.1145/3351276

[79] 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*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300935

[80] Sangki Yun, Yi Chao Chen, Huihunag Zheng, Lili Qiu, and Wenguang Mao. 2017. Strata: Fine-grained acoustic-based device-free tracking. *MobiSys 2017 - Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services* (2017), 15–28. https://doi.org/10.1145/3081333.3081356

[81] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. 2018. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6848–6856.

[82] Yongzhao Zhang, Wei Hsiang Huang, Chih Yun Yang, Wen Ping Wang, Yi Chao Chen, Chuang Wen You, Da Yuan Huang, Guangtao Xue, and Jiadi Yu. 2020. Endophasia: Utilizing acoustic-based imaging for issuing contact-free silent speech commands. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020). https://doi.org/10.1145/3381008