applied sciences

MDPI

Article

Training-Free Acoustic-Based Hand Gesture Tracking on Smart Speakers

Xiao Xu, Xuehan Zhang, Zhongxu Bao, Xiaojie Yu, Yuqing Yin, Xu Yang and Qiang Niu *

School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China; xu_xiao@cumt.edu.cn (X.X.); xuehanzhang@cumt.edu.cn (X.Z.); Baozx@cumt.edu.cn (Z.B.); yuxiaojie@cumt.edu.cn (X.Y.); yinyuqing@cumt.edu.cn (Y.Y.); yang_xu@cumt.edu.cn (X.Y.)
* Correspondence: niuq@cumt.edu.cn

Abstract: Hand gesture recognition is an essential Human–Computer Interaction (HCI) mechanism for users to control smart devices. While traditional device-based methods support acceptable recognition performance, the recent advance in wireless sensing could enable device-free hand gesture recognition. However, two severe limitations are serious environmental interference and high-cost hardware, which hamper wide deployment. This paper proposes the novel system TaGesture, which employs an inaudible acoustic signal to realize device-free and training-free hand gesture recognition with a commercial speaker and microphone array. We address unique technical challenges, such as proposing a novel acoustic hand-tracking-smoothing algorithm with an Interaction Multiple Model (IMM) Kalman Filter to address the issue of localization angle ambiguity, and designing a classification algorithm to realize acoustic-based hand gesture recognition without training. Comprehensive experiments are conducted to evaluate TaGesture. Results show that it can achieve a total accuracy of 97.5% for acoustic-based hand gesture recognition, and support the furthest sensing range of up to 3 m.

Keywords: hand gesture recognition; inaudible acoustic sensing; training-free sensing; device-free sensing

Citation: Xu, X.; Zhang, X.; Bao, Z.; Yu, X.; Yin, Y.; Yang, X.; Niu, Q. Training-Free Acoustic-Based Hand Gesture Tracking on Smart Speakers. *Appl. Sci.* **2023**, *13*, 11954. https://doi.org/10.3390/app132111954

Academic Editor: Antonio Fernández-Caballero

Received: 8 October 2023
Revised: 27 October 2023
Accepted: 30 October 2023
Published: 1 November 2023

1. Introduction

Hand gesture recognition is a convenient and natural Human–Computer Interaction (HCI) technology for users to control smart devices. The report has shown that the market of hand gesture recognition will reach CNY 48 billion by 2024 [1], applying to various fields such as virtual reality, smart home, and health care. Therefore, effective hand gesture recognition will be an important way for humans to interact with machines in the future.

In recent years, diverse hand gesture recognition systems have been developed to serve HCI applications. They can be classified as device-based systems and device-free systems. Among device-based systems, a range of solutions has been proposed based on accelerometers [2], gyroscopes [3], and magnetometers [4]. While promising in recognition accuracy, this series of solutions requires the user to wear or carry devices, which brings inconvenience to their daily life. Among device-free systems, camera-based solutions [5] have issues in regard to occlusion and privacy. Wireless sensing technologies have been successfully employed to sense hand gestures, including Wi-Fi [6], LoRa [7], visible light [8], and radar [9]. Wireless sensing relies on analyzing the wireless signals reflected from the target to obtain information of hand gestures. However, there are several limitations restricting the wide deployment of these solutions. Wi-Fi-based systems are severely affected by environmental interference. Radar-based systems are promising, but they require dedicated high-cost hardware, as do LoRa and visible light.

Acoustic signals [10] have attracted extensive attention recently. Compared with hand gesture recognition systems that employ other wireless signals, acoustic-based systems have two unique advantages. On one hand, speakers and microphones are widely applied

on electronic devices in our daily life, such as smart TVs and smartphones. On the other hand, due to the low propagation speed in the air (340 m/s), acoustic signals have higher sensing precision. Most acoustic-based hand gesture recognition systems adopt machine learning, which requires many data for training. However, there are no publicly available datasets, and it is difficult for us to collect data from diverse users. Inspired by a prior study [11], we adopt acoustic-based hand tracking to intuitively recognize gestures. While promising in many aspects, the microphones of present commodity microphone arrays are usually spaced at several centimeters to optimize speech recognition. Since the wavelength of our using inaudible acoustic signal is much shorter than the spacing of these microphone arrays, the spatial sampling rate is inadequate, which leads to localization angle ambiguity of hand tracking. Therefore, it is particularly challenging for acoustic-based hand gesture recognition with a commodity microphone array.

In this paper, we propose a novel system, TaGesture, which is designed for device-free and training-free hand gesture recognition based on the inaudible acoustic signal at the frequency band of 17 kHz∼23 kHz (audible sound is in 20 Hz∼16 kHz [12]). Specifically, TaGesture utilizes a commercial speaker and microphone array to track hand trajectories for gesture recognition. To achieve TaGesture, we need to address the following challenges. (1) Interference caused by multipath: The interference can cover up the hand information and substantially reduce the sensing precision. (2) Localization angle ambiguity of hand tracking: Due to the large spacing of the microphone array, localization has the angle ambiguity, which leads to hand tracking with a significant error. (3) Training-free gesture recognition: Considering the user diversity, it is challenging to achieve high accuracy and robust gesture recognition without training.

To address the first challenge, we apply background subtraction to obtain clean signals with highlighted target information. To address the second challenge, we propose an acoustic hand-tracking-smoothing algorithm with an Interaction Multiple Model (IMM) Kalman Filter. The smoothing algorithm utilizes nonlinear tracking state equations to perform the optimal estimation on the hand trajectory. Therefore, we can obtain low-error hand trajectories to improve the accuracy of acoustic-based hand gesture recognition with a commercial microphone array. To address the third challenge, we propose a novel classification algorithm without training. We test five types of hand gestures for TaGesture, including swipe left, swipe right, push forward, pull back, and draw a circle. The classification algorithm relies on the unique characteristics of each gesture to realize acoustic-based hand gesture recognition without training.

The main contributions of this paper are listed as follows.

- TaGesture employs an inaudible acoustic signal to realize device-free and training-free hand gesture recognition, with a commercial speaker and microphone array. We believe TaGesture can be widely deployed on smart devices in the real world.
- We propose a novel acoustic hand-tracking-smoothing algorithm with IMM Kalman Filter, which can eliminate localization angle ambiguity of hand tracking. Furthermore, we propose a classification algorithm to realize acoustic-based hand gesture recognition without training.
- We conduct comprehensive experiments to evaluate the performance of TaGesture. Results show that the total accuracy of acoustic-based hand gesture recognition is 97.5%, and the furthest sensing distance is 3 m.

2. Preliminaries

Since the circular microphone array is employed in most commodity smart speakers, we first introduce the hand localization model based on received chirp signals at the circular microphone array. The chirp signal is widely employed in acoustic sensing [13], which can separate reflections from different positions. During the process of hand tracking, the smart speaker continuously transmits a sequence of chirp signals. The chirp signal is a sine wave whose frequency varies linearly over time. The transmitted signal can be represented as

$$S_T(t) = cos\left(2\pi f_c t + \pi k t^2\right), \tag{1}$$

where f_c is the start frequency, $k = \frac{B}{T}$ is the sweep rate, B is the frequency bandwidth, and T is the chirp duration. After the transmitted signal is reflected from the hand, the received signal with delay and attenuation can be obtained, which can be represented as

$$S_R(t) = acos\left(2\pi f_c(t-\tau) + \pi k(t-\tau)^2\right), \tag{2}$$

where a is the attenuation factor, and τ is the Time-of-Flight (ToF) of the signal reflected by the hand. Then, the transmitted and received signals are processed to generate the mixed signal, which is a complex signal with the In-Phase component $I(t) = S_R(t) \cdot S_T(t)$ and Quadrature component $Q(t) = S_R(t) \cdot S_T^*(t)$. $S_T^*(t) = sin\left(2\pi f_c t + \pi k t^2\right)$ is the 90° phase-shifted transmitted signal. By applying the formula $cos\alpha \cdot cos\beta = \frac{1}{2}(\cos(\alpha+\beta) + \cos(\alpha-\beta))$ and a low-pass filter, the mixed signal can be combined as

$$S_M(t) = I(t) + j \cdot Q(t) = \frac{1}{2} a e^{j2\pi\tau(kt+f_c)}. \tag{3}$$

We can analyze the TOF τ to obtain the hand position information, i.e., the distance and angle. If the distance from the hand to the center of the circular array is d, the TOF of the signal at this path is $\frac{2d}{c}$, where $2d$ is the round-trip distance, and c is the speed of sound. Suppose that there are M microphones distributing equally at the circumference as shown in Figure 1. The angle between the m^{th} microphone and the first microphone can be calculated by $\varphi(m) = 2\pi \cdot \frac{m-1}{M}$. Therefore, the TOF $\tau(m)$ of the received signal at the m^{th} microphone can be calculated by

$$\tau(m) = \frac{2d - rcos(\varphi(m) - \theta)}{c}, \tag{4}$$

where r is the radius of the circle, and θ is the incidence angle of the reflected signal. Then, we substitute $\tau(m)$ into (3), and the mixed signal can be rewritten as

$$\begin{aligned} S_M(t_n,\ m) &= \frac{1}{2} a e^{j2\pi(2d - rcos(\varphi(m)-\theta))(kt_n + f_c)/c} \\ &= \frac{1}{2} a e^{j\phi(t_n,\ m)} \end{aligned}, \tag{5}$$

where t_n is the n^{th} sampling timestamp, and $\phi(t_n,\ m)$ is the phase of the mixed signal. Therefore, we can obtain distance d and angle θ from the mixed signal to represent the hand position information.

In practice, the signal is not only reflected by the hand but also by ambient objects, including static objects (e.g., walls and furniture) and dynamic objects (e.g., other humans). Thus, the mixed signal can be viewed as a superposition of reflections from L paths

$$\begin{aligned} S(t_n,\ m) &= \frac{1}{2} \sum_{l=1}^{L} a_l e^{j2\pi(2d_l - rcos(\varphi(m)-\theta_l))(kt_n + f_c)/c} \\ &= \frac{1}{2} \sum_{l=1}^{L} a_l e^{j\phi_l(t_n,\ m)} \end{aligned}, \tag{6}$$

where a_l, d_l, θ_l and $\phi_l(t_n,\ m)$ are the attenuation factor, the distance of the object, the angle of the object, and the phase at the l^{th} path, respectively. In Section 3.3, we describe how to extract the required hand position information from the superposed mixed signal.

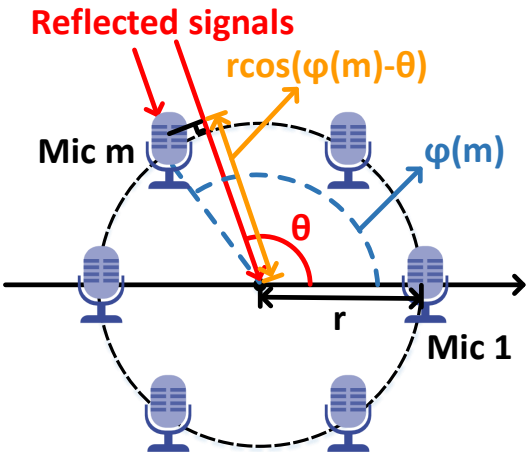

Figure 1. The received signal modal at the circular microphone array.

3. Materials and Methods

Figure 2 shows the overall architecture of our proposed TaGesture, including five main modules: interference cancellation, signal enhancement, position estimation, hand tracking, and hand gesture recognition. In this section, we introduce each module of TaGesture in detail.

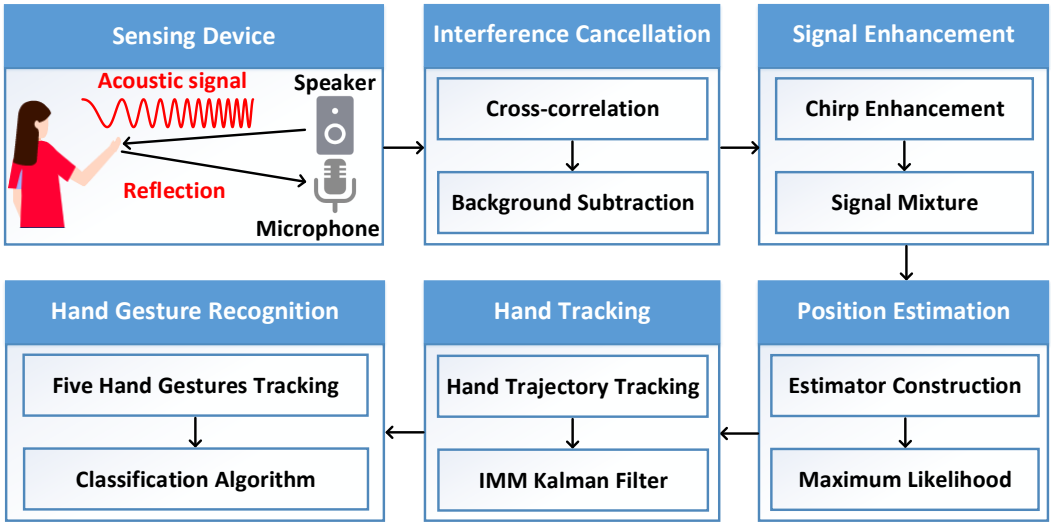

Figure 2. The overall architecture of TaGesture.

3.1. Interference Cancellation

As mentioned in Section 2, we use the TOF of the signal to obtain the hand position information. However, because of the non-synchronization between the speaker and microphone, there is a random time delay for the signals to be transmitted by the speaker. Since the speaker and microphone are co-located, the direct transmission time can be ignored, and the direct received signal is much stronger than the reflected received signal. Therefore, we first adopt a bandpass filter on the received signal to remove the low-frequency noise. Then, we perform cross correlation on the transmitted and received signals, and search for the lag point corresponding to the largest peak. This lag point is considered the start time point of the signal transmission.

Now, we obtain the received signal with the accurate TOFs of multipath, which contains not only the reflection path from the hand but also the direct path and the environment reflection paths. As shown in Figure 3a, the reflected signal from the hand is drowned out by these background signals. Therefore, we apply background subtraction to solve this problem. Specifically, each chirp of the received signal subtracts its previous chirp, and the

first chirp is removed. In this way, the amplitudes of the interference peaks become smaller as shown in Figure 3b.

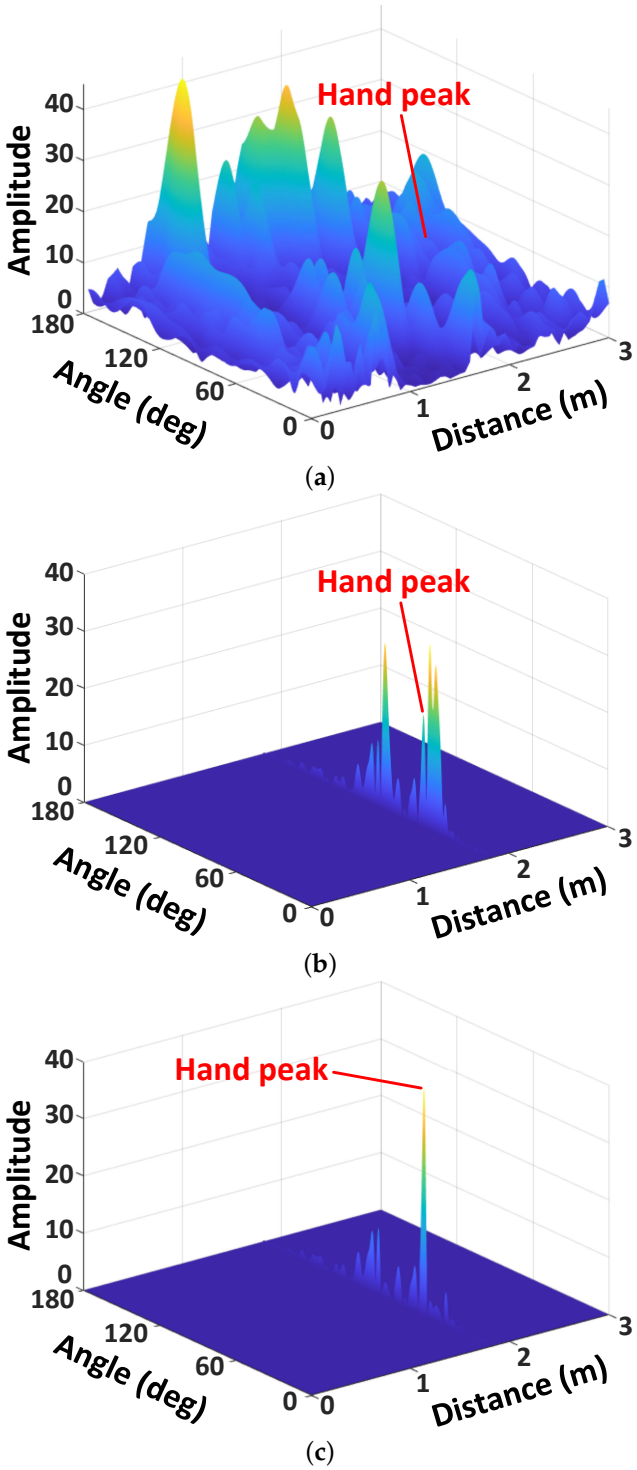

Figure 3. The distance–angle profiles for the hand at 1.8 m and 50°. (**a**) Original. (**b**) After interference cancellation. (**c**) After signal enhancement.

3.2. Signal Enhancement

To preliminarily alleviate the localization angle ambiguity, we adopt wideband chirp signals with the frequency band from 17 kHz to 23 kHz [14], and a series of signal processing methods to enhance the mixed signal.

First, we enhance the transmitted signal purely in software. The sweep time, bandwidth, and sampling rate of the actual signal transmitted by the speaker are 0.08 s, 6 kHz, and 48,000 Hz, respectively. Before mixing the transmitted and received signals, we change the parameters of the transmitted signal in the software. Specifically, the sweep time, frequency band, and sampling rate are set as 0.12 s, 17 kHz $\sim$ 26 kHz, and 96,000 Hz. In this way, the enhanced transmitted chirp can cover the entire received chirp to increase the effective information of the mixed signal. Then, the received signal is upsampled by a factor of two and mixed with the enhanced transmitted signal as introduced in Section 2. Finally, we downsample the In-Phase and Quadrature components by a factor of two and obtain the enhanced mixed signal. As shown in Figure 3c, the signal reflected from the hand stands out in the enhanced mixed signal.

3.3. Position Estimation

For hand tracking, we apply a distance–angle joint estimation algorithm. It is known from Equation (6) that the measured mixed signal $S(t_n,\ m)$ of one chirp contains $N \times M$ samples, where N is the sample size of a chirp, and M is the number of microphones. And for each possible distance d and angle θ, we construct the theoretical mixed signal sample for the m^{th} microphone at the n^{th} sampling timestamp by (5), which can be calculated by

$$\begin{aligned} S^t(t_n,\ m,\ d,\ \theta) &= e^{j2\pi(2d - rcos(\varphi(m)-\theta))(kt_n+f_c)/c} \\ &= e^{j\phi^t(t_n,\ m,\ d,\ \theta)} \end{aligned}, \tag{7}$$

where $\phi^t(t_n,\ m,\ d,\ \theta)$ is the theoretical phase caused by d and θ. By the ratio of the measured and theoretical mixed signal samples, we can obtain the distance–angle joint estimation sample, which can be calculated by

$$\begin{aligned} E_s(t_n,\ m,\ d,\ \theta) &= \frac{S(t_n,\ m)}{S^t(t_n,\ m,\ d,\ \theta)} \\ &= \frac{1}{2}\sum_{l=1}^{L} a_l e^{j(\phi_l(t_n,\ m) - \phi^t(t_n,\ m,\ d,\ \theta))} \end{aligned}. \tag{8}$$

Due to the interference cancellation and signal enhancement, the attenuation factor at the path of the hand a_h is much larger than the attenuation factors at other paths. The distance–angle joint estimation sample can be rewritten as

$$E_s(t_n,\ m,\ d,\ \theta) \approx \frac{1}{2} a_h e^{j(\phi_h(t_n,\ m) - \phi^t(t_n,\ m,\ d,\ \theta))}, \tag{9}$$

where $\phi_h(t_n,\ m)$ is the phase caused by the hand. By summing all distance–angle joint estimation samples for the M microphones at the N sampling timestamps, we can construct the distance–angle joint estimation as

$$E(d,\ \theta) = \sum_{n=1}^{N}\sum_{m=1}^{M} E_s(t_n,\ m,\ d,\ \theta). \tag{10}$$

If distance d and angle θ are the actual position of the hand, theoretical phase $\phi^t(t_n,\ m,\ d,\ \theta)$ is approximately equal to measured phase $\phi_h(t_n,\ m)$, i.e., $E_s(t_n,\ m,\ d,\ \theta)$ are in-phase and closer to the real component. And thus, the amplitude of the summed vector $E(d,\ \theta)$ is maximal. Therefore, we search for the distance $\hat{d}$ and angle $\hat{\theta}$, which make $|E(d,\ \theta)|$ maximized:

$$\left(\hat{d},\ \hat{\theta}\right) = \underset{d,\ \theta}{\arg max} |E(d,\ \theta)|. \tag{11}$$

where $\hat{d}$ and $\hat{\theta}$ are the estimation results of the hand's position information for one chirp.

The search range of the distance and angle can be set based on the application. For hand tracking, we apply a two-step estimation to reduce the computation time. Due to the high accuracy of the acoustic distance estimation, we first narrow the search range of the distance and then estimate the fine-grained position information. For the first estimation, the search ranges of the distance and angle are set as [0, 4 m] and [30°, 150°] at the step sizes of 0.1 m and 40°. Suppose that the estimated distance is d. For the second fine-grained estimation, the search ranges of the distance and angle are set as $[d - 0.1,\ d + 0.1]$ and [30°, 150°] at the step sizes of 0.01 m and 1°.

3.4. Hand Tracking

By the distance–angle joint estimation algorithm, we can obtain the position of the hand for one chirp. Through continuously transmitting chirps from the speaker, we can track the moving trajectory of the hand. However, although we alleviated the angle ambiguity of localization, there are still several outliers in the trajectory due to the low SNR of the acoustic signal. These outliers will reduce the gesture recognition accuracy. Therefore, we perform the Interaction Multiple Model (IMM) Kalman Filter on the process of hand tracking. The detailed process of the IMM Kalman Filter for the k^{th} chirp proceeds as the following steps.

3.4.1. Tracking Model

After position estimation, we convert the distance d and angle θ to a point $(dcos\theta,\ dsin\theta)$ on the two-dimensional x-y plane. For the k^{th} chirp, the hand state vectors of the constant velocity (CV) and constant acceleration (CA) models can be represented as

$$\mathbf{x}_k^{CV} = \left[p_x, p_y, \dot{p}_x, \dot{p}_y\right]^T \tag{12}$$

$$\mathbf{x}_k^{CA} = \left[p_x, p_y, \dot{p}_x, \dot{p}_y, \ddot{p}_x, \ddot{p}_y\right]^T \tag{13}$$

where p_x and p_y are the coordinates of the hand's position, $\dot{p}_x$ and $\dot{p}_y$ are the hand's velocities along the x and the y directions, and $\ddot{p}_x$ and $\ddot{p}_y$ are the hand's accelerations along the x and the y directions. The hand trajectory state equations can be represented as

$$\mathbf{x}_{k+1}^{CV} = F^{CV}\mathbf{x}_k^{CV} + w_k, \tag{14}$$

$$\mathbf{x}_{k+1}^{CA} = F^{CA}\mathbf{x}_k^{CA} + w_k, \tag{15}$$

where w_k is the process noise. F^{CV} and F^{CA} are the transition matrices, which can be represented as

$$F^{CV} = \begin{bmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{16}$$

$$F^{CA} = \begin{bmatrix} 1 & 0 & \Delta t & 0 & \frac{1}{2}\Delta t^2 & 0 \\ 0 & 1 & 0 & \Delta t & 0 & \frac{1}{2}\Delta t^2 \\ 0 & 0 & 1 & 0 & \Delta t & 0 \\ 0 & 0 & 0 & 1 & 0 & \Delta t \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \tag{17}$$

where Δt is the time increment. The measurement equations can be represented as

$$\mathbf{z}_k^{CV} = H^{CV}\mathbf{x}_k + v_k, \tag{18}$$

$$\mathbf{z}_k^{CA} = H^{CA}\mathbf{x}_k + v_k, \tag{19}$$

where v_k is the measurement noise. H^{CV} and H^{CA} are the measurement matrices, which can be represented as

$$H^{CV} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \tag{20}$$

$$H^{CA} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{21}$$

3.4.2. Input Interaction

Suppose that there are R models for the k^{th} chirp, and the mixing probabilities for model i and model j can be represented as

$$\mu^{j|i}_{k-1|k-1} = \frac{p_{ij}\mu^{i}_{k-1}}{\bar{c}_j}, \tag{22}$$

where μ^{i}_{k-1} is the probability of model i, p_{ij} is the transition probability from model i to model j, and $\bar{c}_j = \sum_{i=1}^{R} p_{ij}\mu^{i}_{k-1}$ is the normalization factor. Since the measured value $\mathbf{z}_k$ can be predicted by the hand trajectory state and covariance, we perform the Markovian operation on different models to obtain the new initial input of model i for the k^{th} chirp, which can be calculated by

$$\overline{\mathbf{x}}^{i}_{k-1|k-1} = \sum_{j} \hat{\mathbf{x}}^{j}_{k-1|k-1}\mu^{i|j}_{k-1|k-1}, \tag{23}$$

$$\begin{aligned} \overline{P}^{i}_{k-1|k-1} = \sum_{j} \mu^{i|j}_{k-1|k-1}(P^{j}_{k-1|k-1} + (\overline{\mathbf{x}}^{i}_{k-1|k-1} \\ -\hat{\mathbf{x}}^{j}_{k-1|k-1}) \times (\overline{\mathbf{x}}^{i}_{k-1|k-1} - \hat{\mathbf{x}}^{j}_{k-1|k-1})^{T}), \end{aligned} \tag{24}$$

where $\hat{\mathbf{x}}^{j}_{k-1|k-1}$ and $P^{j}_{k-1|k-1}$ are the state estimation and covariance matrix of model j for the $(k-1)^{th}$ chirp, respectively.

3.4.3. Kalman Filter

For model i of the k^{th} chirp, the predicted state $\hat{\mathbf{x}}_{k|k-1}$ can be calculated by

$$\hat{\mathbf{x}}^{i}_{k|k-1} = F^{i}\hat{\mathbf{x}}^{i}_{k-1|k-1}, \tag{25}$$

where F^i is the transition matrix. And the predicted error covariance matrix can be calculated by

$$P^{i}_{k|k-1} = Q^{i}_{k-1} + F^{i}P^{i}_{k-1|k-1}(F^{i})^{T}, \tag{26}$$

where Q^{i}_{k-1} is the covariance of the process noise. Therefore, the state estimation and covariance matrix after the Kalman Filter can be calculated by

$$\hat{\mathbf{x}}^{i}_{k|k} = \hat{\mathbf{x}}^{i}_{k|k-1} + K^{i}_{k}\left(\mathbf{z}^{i}_{k} - H^{i}\hat{\mathbf{x}}^{i}_{k|k-1}\right), \tag{27}$$

$$P^{i}_{k|k} = P^{i}_{k|k-1} - K^{i}_{k}H^{i}P^{i}_{k|k-1}, \tag{28}$$

where H^i is the measurement matrix, and $K^{i}_{k} = P^{i}_{k|k-1}\left(H^{i}\right)^{T}\left(H^{i}P^{i}_{k|k-1}\left(H^{i}\right)^{T} + Q^{i}_{k}\right)^{-1}$ is the Kalman gain.

3.4.4. Model Probability Update

For model i of the k^{th} chirp, we adopt maximum likelihood function $\Lambda_k^i = \mathbb{N}(u_k^i; 0, S_k^i)$ to update the probability, where $\mathbb{N}$ is the Gaussian distribution, and u_k^i and S_k^i are the measurement residual and its covariance, respectively. Therefore, the updated probability can be calculated by

$$\mu_{k|k}^i = \frac{1}{c}\Lambda_k^i \bar{c}_i, \tag{29}$$

where $\bar{c}_i = \sum_{j=1}^{R} p_{ij}\mu_{k-1}^j$ is the normalization factor, and $c = \sum_{i=1}^{R} \Lambda_k^i \bar{c}_i$ is the normalization constant.

3.4.5. Combination

The combined state estimation and covariance matrix are generated by

$$\hat{\mathbf{x}}_{k|k} = \sum_i \hat{\mathbf{x}}_{k|k}^i \mu_{k|k}^i, \tag{30}$$

$$\begin{aligned} P_{k|k} = \sum_i \mu_{k|k}^i (P_{k|k}^i + (\hat{\mathbf{x}}_{k|k} - \hat{\mathbf{x}}_{k-1|k-1}^i) \\ \times (\hat{\mathbf{x}}_{k|k}^i - \hat{\mathbf{x}}_{k|k}^i)^T). \end{aligned} \tag{31}$$

$\hat{\mathbf{x}}_{k|k}$ is regarded as the filtered position of the hand for the k^{th} chirp, and these two outputs are used as the next interactive input to complete the cycle of the whole hand tracking.

3.5. Hand Gesture Recognition

We test five types of hand gestures for TaGesture, including swipe left, swipe right, push forward, pull back, and drawing a circle. Figure 4 shows the estimated trajectories of the five hand gestures. The transceiver is at the zero point, and the yellow and red dots represent the starting and ending points, respectively. For hand gesture recognition, we design a brief classification algorithm without machine learning as the following steps.

(a) We extract the first, middle, and last points from the trajectory. Then, we take the middle point as the vertex of the angle β, which can be calculated by the cosine law. If $\beta < 90°$, the trajectory is draw a circle.

(b) If $\beta > 90°$, we define the moving displacement along the x and y axis as d_x and d_y, respectively. If $d_x - d_y > 0$, the trajectory is swipe left or swipe right. If $d_x - d_y < 0$, the trajectory is push forward or pull back.

(c) To further distinguish between swipe left and swipe right, if $d_x > 0$, the trajectory is swipe left. If $d_x < 0$, the trajectory is swipe right. To further distinguish between push forward or pull back, if $d_y > 0$, the trajectory is pull back. If $d_y < 0$, the trajectory is push forward.

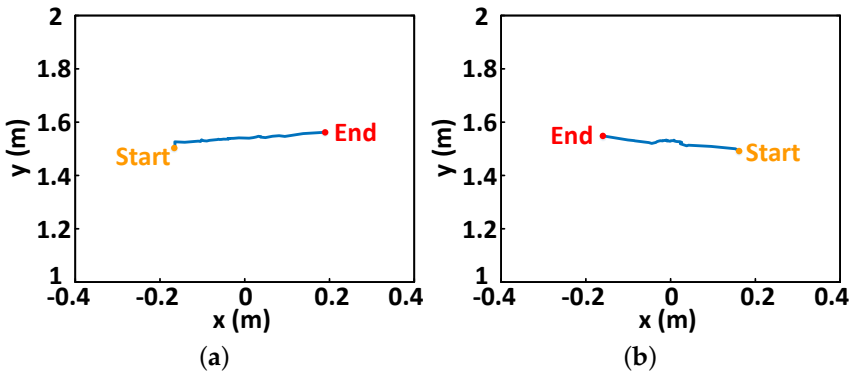

Figure 4. *Cont.*

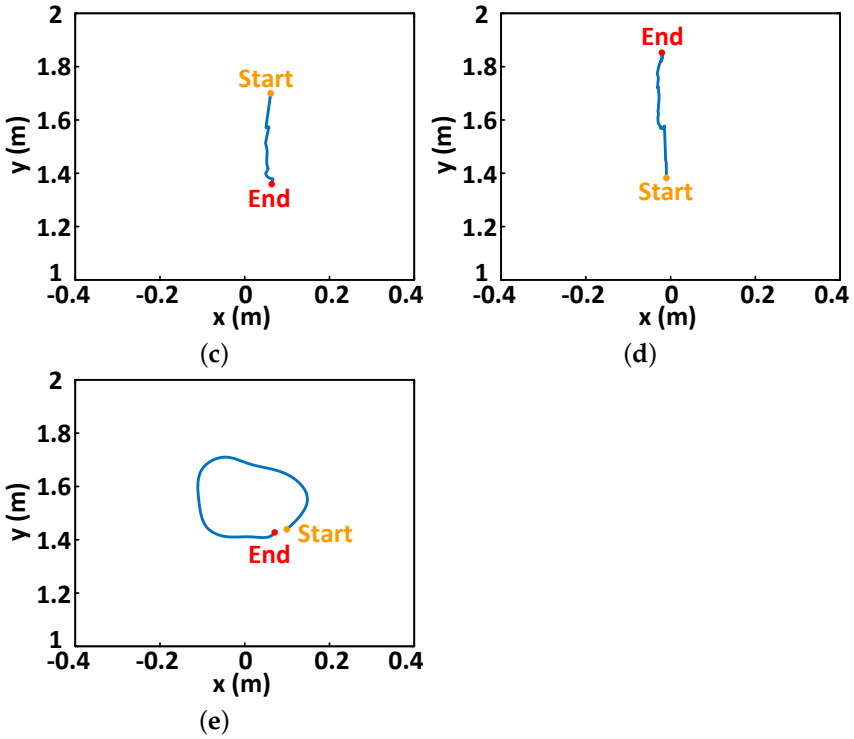

Figure 4. The trajectory of five hand gestures after the IMM Kalman Filter. (**a**) Swipe left. (**b**) Swipe right. (**c**) Push forward. (**d**) Pull back. (**e**) Draw a circle.

4. Results

4.1. Implementation

Experiment scenario and setup. We recruited five volunteers to participate in the study, including four females and one male. We conducted the experiments in the laboratory, and informed the volunteers of the experimental details. For each volunteer, we collected 300 hand gestures, where there were 60 trials for each of the five types of hand gestures, including swipe left, swipe right, push forward, pull back, and draw a circle. For each trial, we told the volunteers the type of hand gesture, and they performed the hand gesture after the speaker started transmitting acoustic signals. In the experiments, the volunteers were asked to sit in front of the transceiver in the range from 0.5 m to 3 m, and the transceiver was placed at the same height as the hand.

Hardware. As shown in Figure 5, we adopt a commercial speaker (JBL Jembe, London, UK, 6 Watt, 80 dB) to transmit acoustic chirp signals and a commercial microphone array (ReSpeaker 6-Mic Circular Array Seeed Studio, Inc., Shenzhen, China) as the receiver. The speaker and microphone array are connected to Raspberry Pi 3B, which is connected with the laptop (Dell Inspiron 7566 Dell Technologies, Round Rock, YX, USA) through the network. We use the laptop to control transmitting and receiving acoustic signals, and process the signals with MATLAB R2021a . For the acoustic chirp signal transmitted by the speaker, the parameters are the starting frequency f_c = 17 kHz, the frequency bandwidth B = 6 kHz, and the chirp period T = 0.08 s. The sampling rate of the microphone array is 48 kHz.

4.2. Overall Performance

We first evaluate the overall performance of TaGesture. As described in the implementation, we collected 1500 hand gestures in total, where there were 300 trials for each of the five types of hand gestures. Figure 6 shows the confusion matrix of the five types of hand gestures. TaGesture can achieve excellent overall performance for hand gesture recognition with an accuracy of 97.5%. The results demonstrate TaGesture's effectiveness and robustness for hand gesture recognition.

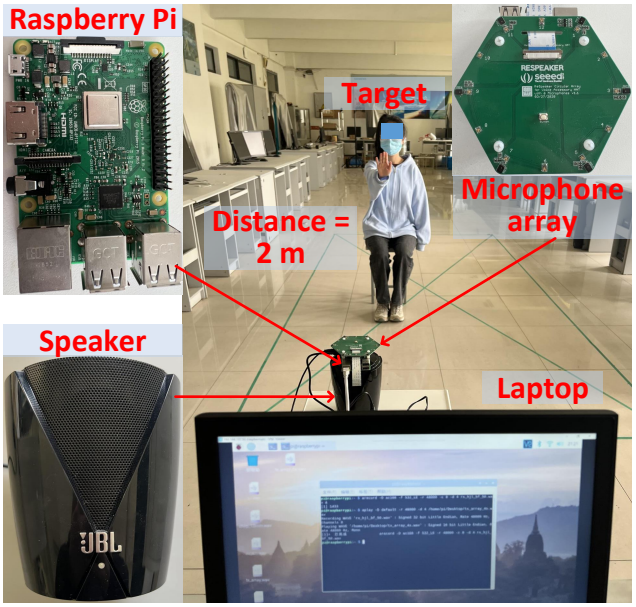

Figure 5. Implementation.

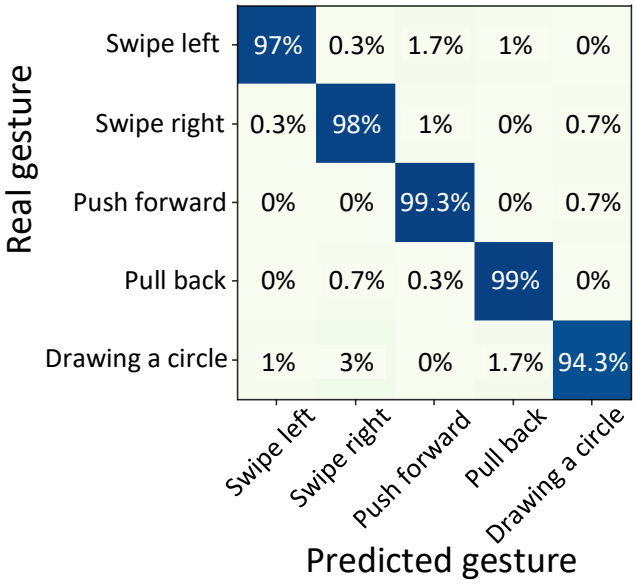

Figure 6. The overall confusion matrix.

4.3. Evaluation of IMM Kalman Filter

We verify the effectiveness of IMM Kalman Filter through benchmark experiments. Specifically, we evaluate the performances of hand gesture recognition with and without the IMM Kalman Filter for five types of hand gestures. As shown in Figure 7, the accuracy of hand gesture recognition with the IMM Kalman Filter are all significantly higher for the five types of hand gestures. Note that the IMM Kalman Filter works better for complex gestures, such as drawing a circle. The results demonstrate that the IMM Kalman Filter can effectively smooth the trajectories of hand tracking and improve the performance of hand gesture recognition.

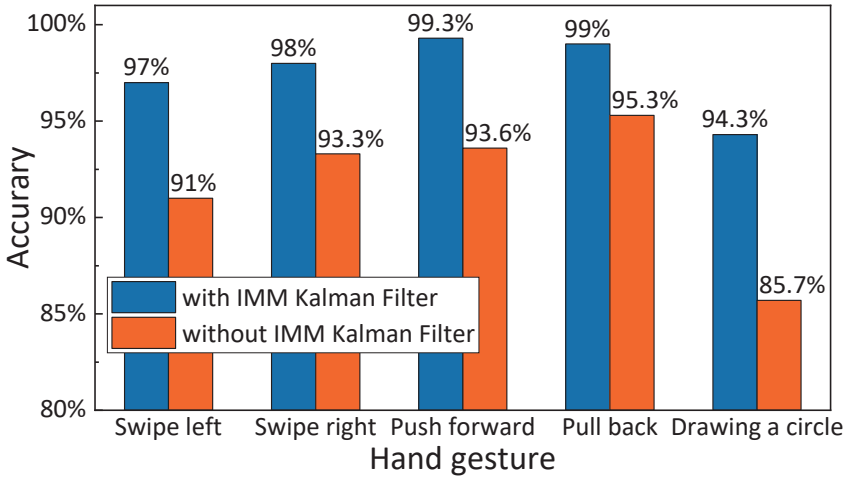

Figure 7. Evaluation of IMM Kalman Filter.

4.4. Impact of Different Distances

To evaluate the performance of hand gesture recognition at different distances, we conducted the experiment at the varied distance between the volunteer and transceiver from 0.5 m to 3 m at a step size of 1 m. For each distance, there were 50 trials of each hand gesture. As shown in Figure 8, at 0.5 m, TaGesture achieved the highest accuracy of 99.2%. As the distance increased, the accuracy decreased due to the lower SNR caused by attenuation. But even at 3 m, TaGesture still achieved an accuracy of 94.8%. The results demonstrate that TaGesture can still work effectively at a distance of 3 m.

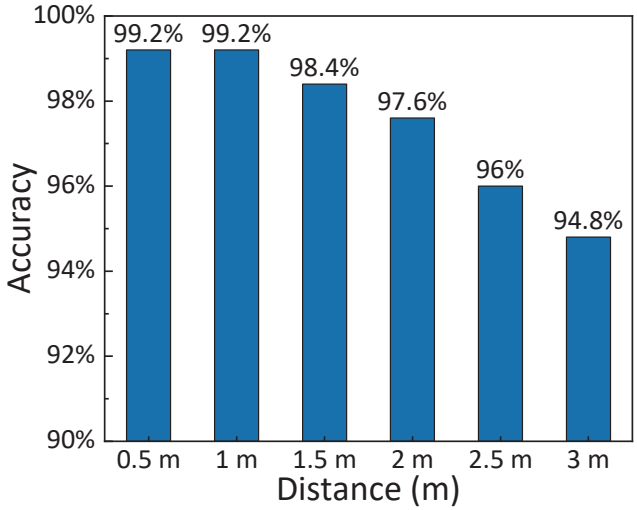

Figure 8. Impact of different distance.

4.5. Impact of Ambient Noise

Since our TaGesture relies on acoustic signals, we evaluated the impact of ambient sound on it. The volunteers were at 2 m in front of the transceiver and we set two types of noises at 0.5 m away from the transceiver. The first type of noise is human voice. We asked another volunteer to read an article with a normal speech volume. The second type of noise is external music played by an external mobile phone, which played music at 40% and 80% of its maximum volume. We used the Decibel X app on iPhone 13 Pro (Apple Inc., Cupertino, CA, USA) to measure the sound pressure level at the position of the transceiver. The sound pressure level of four cases are quiet (37.9 dB), speak (55.1 dB), music at 40% maximum volume (63.1 dB), and music at 80% maximum volume (69.2 dB), respectively. As shown in Figure 9, the performance of hand gesture recognition under different noises

are similar. The reason is that the frequency of the ambient noise is below 14 kHz, and thus they cannot affect our TaGesture with the signal in the frequency range of 17–23 kHz.

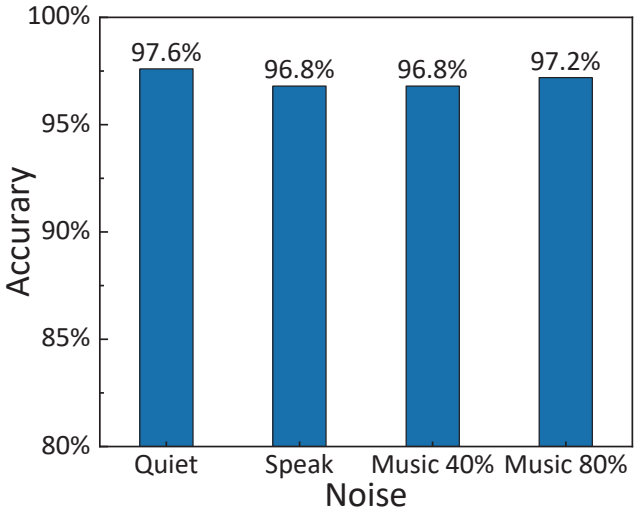

Figure 9. Impact of ambient noise.

4.6. Impact of User Diversity

To evaluate the impact of user diversity, we display the accuracies of all five volunteers in Figure 10. From the results, we observe that the performance of gesture recognition is dependent on the hand size. Since the larger hand size can induce more obvious signal variations, we can extract the precise hand position by the outstanding reflection peak. But overall, TaGesture can achieve stable recognition accuracy for different users.

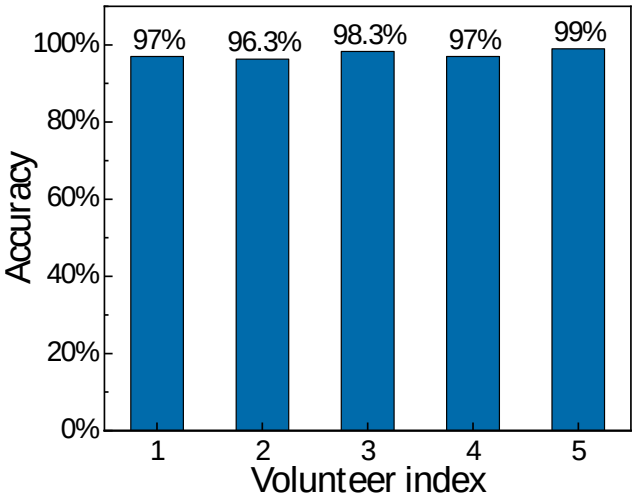

Figure 10. Impact of user diversity.

4.7. Impact of Different Environments

To evaluate the system robustness, we conducted experiments in three different environments, including a laboratory, meeting room, and corridor. The experiment setup is the same as the implementation in Section 4.1, and the distance between the volunteer and the transceiver is 2 m. The experimental results are shown in Figure 11. Due to a series of anti-interference algorithms, TaGesture can perform well in all three environments.

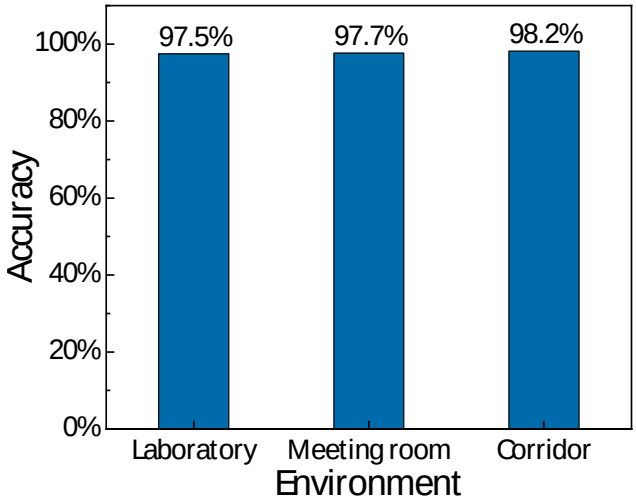

Figure 11. Impact of different environments.

5. Related Work

5.1. Gesture Recognition Based on Wireless Signal

Wireless sensing technologies have been widely employed for gesture recognition, including Wi-Fi, radar, UWB, visible light, LoRa, and RFID. Wireless sensing relies on analyzing the wireless signals reflected from the target to obtain gesture information. DPSense [15] leverages the proposed EDP metric to effectively classify signals into segments according to their sensing qualities and process them differently. This system enhances the sensing quality and avoids distortion from low-quality signals to assure the robustness of gesture recognition. Ahmed et al. [9] presented a multistream convolutional neural network (MS-CNN)-based in-air digits recognition method using a frequency-modulated continuous-wave (FMCW) radar. With one FMCW radar comprising two receiving channels, a novel three-stream CNN network with range–time, Doppler–time, and angle–time spectrograms as inputs was constructed, and the features were fused together in the later stage before making a final recognition. Li et al. [16] proposed a sign language (SL)/hand gesture recognition method based on a novel discriminative feature, built a measurement system of hand movements using an UWB radar, measured 10-type, 15-type SL actions, and 10-type hand gesture actions, and completed the tasks of 10-type SL/hand gesture recognition based on the extracted new features. Webber et al. [17] presented a method for recognizing gestures via the analysis of interrupted light patterns using visible light. MineSOS [18] is a distress gesture sensing system utilizing LoRa technology, in response to rescues in long-distance coal mine tunnels full of dust and dangerous gases. Merenda et al. [19] proposed a device-free gesture identification system that recognizes different hand movements by processing through Edge Machine Learning (EML) algorithms, received signal strength indication (RSSI) and phase values from backscattered signals of a collection of RFID tags mounted on a plastic plate. Among these device-based systems, Wi-Fi-based systems have low sensing precision because of environmental interferences, and systems based on other signals usually require expensive dedicated devices.

5.2. Wireless Sensing Based on Acoustic Signal

In previous work, sensing applications based on acoustic signals mainly utilize their high sensing precision and low cost. A large variety of applications have been enabled with acoustic sensing, including localization, gesture recognition, daily activity identification, gait estimation, elderly fall detection, lip reading, and respiration monitoring. EchoSpot [20] is a novel device-free localization system that leverages only one speaker and one microphone for precisely locating a human. RobuCIR [21] adopts a frequency-hopping mechanism to achieve a robust contact-free gesture recognition system that can work under different practical impact factors with high accuracy and robustness. Aitness [22] enables

non-intrusive, passive, and high-precision fitness detection based on acoustic sensing. EchoSensor [23] leverages speakers and microphones in smart home devices to capture human gait patterns for individual identification. Lian [24] developed a novel and lightweight fall detection system by relying solely on a home audio device via inaudible acoustic sensing to recognize fall occurrences for wide home deployment. SVoice [25] supports accurate audible speech reconstruction by analyzing the disturbance of tiny articulatory gestures on the reflected ultrasound signal. The design introduces a new model that provides the unique mapping relationship between ultrasound and speech signals so that audible speech can be successfully reconstructed from silent speech. SymListener [26] detects respiratory symptoms in a driving environment. By continuously recording acoustic data through a built-in microphone, SymListener can also detect the sounds of coughs, sneezes and sniffles. These acoustic-based works provide good basics for future research.

6. Discussion

6.1. Traditional Tracking Method

The multiple signal classification (MUSIC) algorithm is a method based on matrix feature space decomposition. First, we obtain the covariance matrix estimate $\hat{R}_x$ of the output signal based on N received signal vectors:

$$\hat{R}_x = \frac{1}{K}\sum_{1}^{K} x_k x_k^H, \tag{32}$$

where K is the signal sample size, x_k is the signal sample, and H computes the transpose of the matrix. Then, eigenvalue decomposition is performed on the covariance matrix obtained above:

$$\begin{aligned} \hat{R}_x &= \hat{U}\Sigma\hat{U}^H \\ &= \hat{U}_S\Sigma_S\hat{U}_S^H + \hat{U}_N\Sigma_N\hat{U}_N^H \end{aligned} \tag{33}$$

where $\hat{U}_S$ is the subspace spanned by the eigenvectors corresponding to the large eigenvalues (i.e., the signal subspace), $\hat{U}_N$ is the subspace spanned by the eigenvectors corresponding to the small eigenvalues (i.e., the noise subspace), and $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_N^2)$ is the eigenvalue matrix. According to the order of the eigenvalues, the eigenvalues equal to the number of signals M and the corresponding eigenvectors are regarded as the signal part space, and the remaining eigenvalues and eigenvectors are regarded as the noise part space, and the noise matrix is obtained. Because the array direction vector is orthogonal to the noise subspace, the orientation of the target will appear minimal in the signal incident direction. Finally, the spatial spectral function $P_{MUSIC}(\theta)$ is sorted out, and the maximum spectral peak is searched:

$$P_{MUSIC}(\theta) = \frac{1}{\mathbf{a}^H(\theta)\hat{\mathbf{U}}_N\hat{\mathbf{U}}_N^H\mathbf{a}(\theta)}, \tag{34}$$

where $\mathbf{a}(\theta)$ is the guiding value of the spatial array.

The MUSIC algorithm is calculated under a number of known signal sources, which is impossible in practical application, and the number of sources can only be estimated based on the observed data. This method of source estimation based on the distribution of eigenvalues of array covariance matrix is perfect in theory, at least for independent sources and partially related sources, but in fact, due to the limited data length, the source number can only be determined by subjective judgment to a large extent. This is also the biggest drawback of the MUSIC algorithm.

6.2. Limitation and Future Work

First, the hand gestures recognized by our system are simple. Since our system is training free, we must design a more complex classification algorithm for complex gestures. Second, the sensing distance of the system has limitations. It is mainly due to the severe

attenuation of acoustic signals in the air. In this paper, we designed and used some algorithms to increase the sensing distance to 3 m. We will design the microphone array development board ourselves to adapt to more tasks. This can solve the problem of angle ambiguity and further increase the sensing range. Finally, the system in this paper is only suitable for a single target. Multi-target gesture recognition is also a future direction. We plan to employ beamforming with a microphone array to separate multiple targets and enhance the signal amplitude.

7. Conclusions

This paper proposed TaGesture, which is a device-free and training-free hand gesture recognition system using inaudible acoustic signals. Our proposed system is a promising alternative compared to traditional device-based methods and other wireless sensing technologies. We address several challenges, such as an acoustic hand-tracking-smoothing algorithm with IMM Kalman Filter to address the issue of localization angle ambiguity, and a classification algorithm to realize acoustic-based hand gesture recognition without training. And then, we implement TaGesture on a commercial speaker and microphone array. Comprehensive experiments demonstrate the effectiveness and robustness of the proposed system. We believe the proposed TaGesture can trigger a long range of hand gesture recognition applications for users. The unique opportunities and challenges make this an exciting research direction to explore.

Author Contributions: Conceptualization, X.X.; Data curation, X.X.; Formal analysis, X.Z.; Funding acquisition, Q.N.; Investigation, X.Z.; Methodology, X.Y. (Xiaojie Yu); Project administration, Q.N.; Resources, X.Y. (Xiaojie Yu); Software , X.Y. (Xu Yang); Supervision, Y.Y.; Validation, X.Y. (Xu Yang); Visualization, Y.Y.; Writing—original draft, Z.B.; Writing—review and editing, Z.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the China Postdoctoral Science Foundation (grant number 2022TQ0366 and 2023M733770).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the authors.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wang, C.; Liu, J.; Chen, Y.; Liu, H.; Xie, L.; Wang, W.; He, B.; Lu, S. Multi-touch in the air: Device-free finger tracking and gesture recognition via cots rfid. In Proceedings of the IEEE INFOCOM 2018—IEEE Conference on Computer Communications, Honolulu, HI, USA, 15–19 April 2018; pp. 1691–1699.
2. Pan, T.Y.; Tsai, W.L.; Chang, C.Y.; Yeh, C.W.; Hu, M.C. A hierarchical hand gesture recognition framework for sports referee training-based EMG and accelerometer sensors. *IEEE Trans. Cybern.* **2020**, *52*, 3172–3183. [CrossRef] [PubMed]
3. Han, H.; Yoon, S.W. Gyroscope-based continuous human hand gesture recognition for multi-modal wearable input device for human machine interaction. *Sensors* **2019**, *19*, 2562. [CrossRef] [PubMed]
4. Koch, P.; Dreier, M.; Böhme, M.; Maass, M.; Phan, H.; Mertins, A. Inhomogeneously stacked rnn for recognizing hand gestures from magnetometer data. In Proceedings of the 2019 27th European Signal Processing Conference (EUSIPCO), A Coruña, Spain, 2–6 September 2019; pp. 1–5.
5. León, D.G.; Gröli, J.; Yeduri, S.R.; Rossier, D.; Mosqueron, R.; Pandey, O.J.; Cenkeramaddi, L.R. Video hand gestures recognition using depth camera and lightweight cnn. *IEEE Sens. J.* **2022**, *22*, 14610–14619. [CrossRef]
6. Li, C.; Liu, M.; Cao, Z. WiHF: Enable user identified gesture recognition with WiFi. In Proceedings of the IEEE INFOCOM 2020—IEEE Conference on Computer Communications, Toronto, ON, Canada, 6–9 July 2020; pp. 586–595.
7. Xie, B.; Xiong, J. Combating interference for long range LoRa sensing. In Proceedings of the 18th Conference on Embedded Networked Sensor Systems, Virtual, 16–19 November 2020; pp. 69–81.
8. Venkatnarayan, R.H.; Shahzad, M. Gesture recognition using ambient light. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2018**, *2*, 1–28. [CrossRef]
9. Ahmed, S.; Kim, W.; Park, J.; Cho, S.H. Radar-Based Air-Writing Gesture Recognition Using a Novel Multistream CNN Approach. *IEEE Internet Things J.* **2022**, *9*, 23869–23880. [CrossRef]

10. Ling, K.; Dai, H.; Liu, Y.; Liu, A.X.; Wang, W.; Gu, Q. Ultragesture: Fine-grained gesture sensing and recognition. *IEEE Trans. Mob. Comput.* **2020**, *21*, 2620–2636. [CrossRef]
11. Li, D.; Liu, J.; Lee, S.I.; Xiong, J. FM-track: Pushing the limits of contactless multi-target tracking using acoustic signals. In Proceedings of the 18th Conference on Embedded Networked Sensor Systems, Virtual, 16–19 November 2020; pp. 150–163.
12. Møller, H.; Pedersen, C.S. Hearing at low and infrasonic frequencies. *Noise Health* **2004**, *6*, 37–57. [PubMed]
13. Cai, C.; Zheng, R.; Luo, J. Ubiquitous acoustic sensing on commodity iot devices: A survey. *IEEE Commun. Surv. Tutor.* **2022**, *24*, 432–454. [CrossRef]
14. Li, D.; Liu, J.; Lee, S.I.; Xiong, J. Room-Scale Hand Gesture Recognition Using Smart Speakers. In Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems, Boston, MA, USA, 6–9 November 2023; pp. 462–475.
15. Gao, R.; Li, W.; Xie, Y.; Yi, E.; Wang, L.; Wu, D.; Zhang, D. Towards robust gesture recognition by characterizing the sensing quality of WiFi signals. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2022**, *6*, 1–26. [CrossRef]
16. Li, B.; Yang, J.; Yang, Y.; Li, C.; Zhang, Y. Sign language/gesture recognition based on cumulative distribution density features using UWB radar. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–13. [CrossRef]
17. Webber, J.; Mehbodniya, A. Recognition of Hand Gestures using Visible Light and a Probabilistic-based Neural Network. In Proceedings of the 2022 4th IEEE Middle East and North Africa COMMunications Conference (MENACOMM), Amman, Jordan, 6–8 December 2022; pp. 54–58.
18. Yin, Y.; Yu, X.; Gao, S.; Yang, X.; Chen, P.; Niu, Q. MineSOS: Long-Range LoRa-Based Distress Gesture Sensing for Coal Mine Rescue. In Proceedings of the International Conference on Wireless Algorithms, Systems, and Applications, Dalian, China, 24–26 November 2022; pp. 105–116.
19. Merenda, M.; Cimino, G.; Carotenuto, R.; Della Corte, F.G.; Iero, D. Edge machine learning techniques applied to rfid for device-free hand gesture recognition. *IEEE J. Radio Freq. Identif.* **2022**, *6*, 564–572. [CrossRef]
20. Lian, J.; Lou, J.; Chen, L.; Yuan, X. Echospot: Spotting your locations via acoustic sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2021**, *5*, 113. [CrossRef]
21. Wang, Y.; Shen, J.; Zheng, Y. Push the limit of acoustic gesture recognition. *IEEE Trans. Mob. Comput.* **2020**, *21*, 1798–1811. [CrossRef]
22. Wang, P.; Jiang, R.; Guo, Z.; Liu, C. Afitness: Fitness Monitoring on Smart Devices via Acoustic Motion Images. *ACM Trans. Sens. Netw.* **2023**. [CrossRef]
23. Lian, J.; Du, C.; Lou, J.; Chen, L.; Yuan, X. EchoSensor: Fine-Grained Ultrasonic Sensing for Smart Home Intrusion Detection. *ACM Trans. Sens. Netw.* **2023**, *20*, 1–24. [CrossRef]
24. Lian, J.; Yuan, X.; Li, M.; Tzeng, N.F. Fall detection via inaudible acoustic sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2021**, *5*, 114. [CrossRef]
25. Fu, Y.; Wang, S.; Zhong, L.; Chen, L.; Ren, J.; Zhang, Y. SVoice: Enabling Voice Communication in Silence via Acoustic Sensing on Commodity Devices. In Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems, Boston, MA, USA, 6–9 November 2022; pp. 622–636.
26. Wu, Y.; Li, F.; Xie, Y.; Wang, Y.; Yang, Z. SymListener: Detecting Respiratory Symptoms via Acoustic Sensing in Driving Environments. *ACM Trans. Sens. Netw.* **2023**, *19*, 1–21. [CrossRef]