V-Pen: An Acoustic-Based Fine-Grained Virtual Pen Input System Using Hand Tracking

Wei Han, *Student Member, IEEE*, Yinghao Li, *Student Member, IEEE*, Hao Yu, *Student Member, IEEE*, and Jiabin Jia, *Senior Member, IEEE*

***Abstract*— Acoustic-based hand-tracking technology leads to the next generation of the Human-Computer interaction (HCI) mechanism. This approach uses embedded speakers and microphones on commercial devices to send and receive acoustic signals simultaneously, then the echo can be processed to obtain the hand's position. However, existing tracking approaches do not support multistroke input, as a result, the trajectory is incapable of character recognition with models trained by simple character images from databases such as MNIST and EMNIST. In this article, V-Pen is proposed to estimate the status of the hand with the energy information acquired from the echo. Subsequently, Zadoff-Chu (ZC) sequences are used to obtain the initial position of the hand and track the hand continuously with the change of phase for a smooth trajectory. While inputting the characters, V-Pen allows the user to input multiple strokes to get rid of redundant trajectory which affects the recognition. The results show that V-Pen achieves an average error of 4.3 mm for tracking and 94.8% recognition accuracy for 52 English letters, ten numbers, and 20 Chinese characters.**

***Index Terms*— Acoustic tracking, hand trajectory recognition, time-of-flight (ToF) estimation.**

I. Introduction

HAND tracking and recognition have emerged as a compelling Human-Computer interaction (HCI) mechanism, garnering significant attention in various HCI-related domains [1] such as mobile phones [2], virtual reality (VR), and augmented reality (AR) devices [3]. In contrast to traditional HCI systems that rely on physical contact with keyboards or touch pads [4], hand tracking and gesture recognition present a novel communication approach, offering unique possibilities for gaming and interactive experiences [5].

Hand tracking methodologies can be broadly categorized into four types, as outlined in Table I: wearable-sensor-based [6], [7], radio frequency (RF)-based [8], [9], camera-based [10], [11], and acoustic-based solutions [12]. Wearable-sensor-based solutions leverage precise hand motion data acquisition, achieving optimal accuracy for gesture recognition. RF-based approaches find extensive application in localization and tracking, yet their resolution is constrained by the speed of light. Unlike the first two types, which necessitate external devices, the ubiquity of built-in cameras in most smart devices facilitates the implementation of vision-based hand sensing. However, this approach contends with privacy concerns, susceptibility to lighting conditions, limited viewing range, and high power consumption. Moreover, obtaining accurate distance information from visual data poses a significant challenge. To address these limitations, acoustic-based approaches have been introduced. The compatibility of these approaches is similar to vision-based ones as smart devices also have embedded speakers and microphones [13].

Acoustic-based hand-tracking solutions exploit physical phenomena named Doppler shift (DS) [20], coupled with advancements in signal modulation techniques. These solutions often incorporate either active or passive acoustic sensing through embedded transceivers [12]. Active sensing methods, represented by AAMouse [16], high-precision acoustic tracker (CAT) [21], and DopLink [22], provide accurate tracking of hand movements down to centimeter-level precision. However, these methods require the user to hold the tracking device physically, introducing limitations in terms of implementation flexibility.

Conversely, passive acoustic sensing methods, or device-free tracking solutions, are becoming increasingly popular due to their innate benefits [12]. The operation of a device-free passive sensing system typically consists of three stages. In the first stage, the system determines a coarse-grained estimation of the hand's position by measuring the absolute distance of the signal transmission path. Yet, due to bandwidth limitations, the resolution of these measurements is relatively low. For instance, low-latency acoustic phase (LLAP) [17] calculates absolute distance using a delay profile, yielding a resolution of only 6.16 cm. Acoustic multitarget tracking (AMT) [19] utilizes modulated Zad-off Chu (ZC) sequences [23] to compute time-of-flight (ToF) [24], capitalizing on their high auto-correlation and weak cross correlation properties. However, this approach may produce redundant results even in the absence of a hand. The second stage involves continuous hand tracking by applying relative distance changes to the initial known position. Technologies such as FingerIO [25], LLAP [17], and Strata [18] leverage phase changes to detect fine-grained distance variations. However, the error accumulates over time. The last stage is to recognize the tracking results. Current research lacks support for multiple stroke inputs for single characters. For example, LLAP [17] overlooks characters like the lowercase "i." Moreover, LLAP [17] uses MyScript [26] instead of the readily available MNIST [27] and EMNIST [28] databases, thereby increasing the costs associated with commercial applications.

Manuscript received 29 August 2023; revised 15 October 2023; accepted 30 October 2023. Date of publication 28 November 2023; date of current version 28 December 2023. The Associate Editor coordinating the review process was Dr. Huang-Chen Lee. *(Corresponding author: Jiabin Jia.)*

The authors are with the Agile Tomography Group, School of Engineering, Institute for Imaging, Data and Communications, The University of Edinburgh, EH8 9YL Edinburgh, U.K. (e-mail: jiabin.jia@ed.ac.uk).

Digital Object Identifier 10.1109/TIM.2023.3335525

TABLE I

HAND TRACKING SOLUTIONS AND COMPARISON

Sensing system		Hardware	Methodology	Tracking accuracy	Recognition accuracy	Platform /Datasets	Character numbers	Multi-stroke support
Wearable		Electromyogram [14]	Support vector machine	N/A	89% (gesture)	N/A	N/A	N/A
		Motion sensors [7]	Artificial neural network	N/A	95.07% (gesture)	N/A	N/A	N/A
RF		RFID devices [15]	Phase change	Median position error: 12.8 cm	98% (character)	Sogou	Not given	No
Vision		Camera [11]	Circular Fuzzy Neural Network	N/A	96% (gesture)	N/A	N/A	N/A
Acoustic	Active	Mobile phone [16]–[19]	Doppler shift	Median error: 1.4 cm	Not given	N/A	N/A	N/A
	Passive		Phase change Delay profile	Average error: 4.6mm	92.3% (character)	Myscript	26 lower case letters	No
			Phase change Channel information	Median error: 5.7mm	Not given	N/A	N/A	N/A
			ZC sequence	Average error: 1.13cm	97% (gesture)	N/A	N/A	N/A
			V-Pen: Phase change ZC sequence	Average error: 4.3mm Median error: 4.0mm	94.8% (character)	MNIST EMNIST QPen	82	Yes

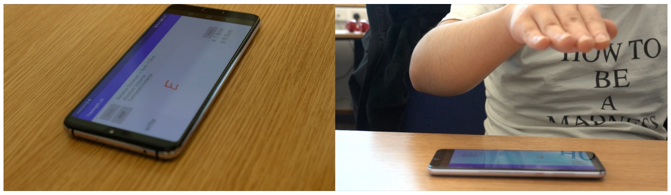


Fig. 1. Input a character in the air using V-Pen.

To address the aforementioned limitations, this article presents V-Pen, a novel hand tracking and recognition system. V-Pen accesses the energy of the echo to ascertain hand movements, then filters out superfluous absolute distance data gathered when the hand remains stationary. This process allows for more precise estimation of absolute distances. Furthermore, the system takes advantage of the period when the user moves from the end of the current character to the start of the next character to reevaluate the absolute distance, mitigating the accumulation of phase errors. A notable enhancement offered by V-Pen is its ability to accommodate multiple stroke inputs for a single character as Fig. 1 shows, which significantly improves recognition performance.

The key contributions of this article are threefold.

1) To the best of our knowledge, V-Pen is the first work that allows the user to input characters with multiple strokes using acoustic hand tracking. Thus, the accuracy of the recognition is improved accordingly.
2) V-Pen proposed a hand status estimation method based on the echo's energy to remove redundant distance estimation results, which improves the tracking accuracy.
3) V-Pen has been successfully integrated as an input method on commercial mobile phones. The system's performance boasts an average tracking error of merely 4.3 mm and maintains a remarkable recognition accuracy rate of 94.8% for 52 English letters, ten numbers and 20 Chinese characters.

The structure of the article is organized as follows: Section II starts with an overview of the V-Pen and then illustrates algorithms in detail. Section III demonstrates the implementation of the V-Pen on a commercial mobile phone. Section IV introduces the system performance. Finally, Section V illustrates the concluding remarks and discusses future works.

II. METHODOLOGY

A. Overview

The workflow of the V-Pen signal processing is shown in Fig. 2. The V-Pen system architecture has two components: the transmitter and the receiver. On the transmitter side, the signal designed was modulated and then transmitted by the transmitter on the mobile phone as shown in the upper part of Fig. 2. For the receiver, an initialization step was required, which contains the demodulation of the received signal and transceiver synchronization. This step was ensured to acquire precise ToF measurements. Subsequently, the algorithm would be waiting for the determination of the initial hand's position. During this period, the hand status estimation algorithm was used to detect the start and the end of hand movement. Upon the hand arrived at the desired start position and stops, the hand status estimation algorithm would catch this moment and allow the system to estimate the initial position of the hand. With the initial hand position determined, V-Pen ensured continuous and smooth hand trajectories by tracking the relative distance by monitoring the phase change. It then calculated the hand's coordinates utilizing the distance data procured from the phone's dual microphones. In the concluding stage, V-Pen activated its multistroke writing functionality, which comprises two modes: normal writing and stroke change. The transition between these modes was easily instigated by detecting short-term distance variations. When these changes fall beneath a specified threshold, it indicates the completion of the current stroke or character. The recorded trajectory will next be saved for recognition. After finishing the writing of the current character, the system would reset the initial position using the period that the hand moves to the desired start position for the next character. Sections II-B–II-K will discuss the algorithms in detail following the order of the workflow.

B. Signal Design

Two distinct signals were designed for different algorithm modes. When obtaining the initial position or re-locating

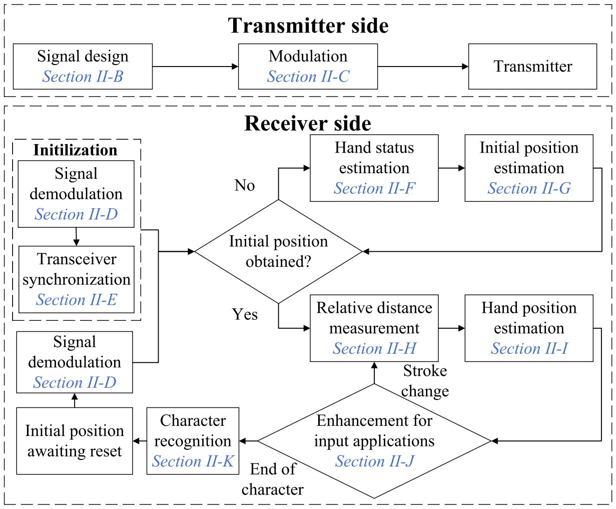


Fig. 2. V-Pen signal processing flowchart.

the hand following the completion of writing one character, a single frequency sinusoidal wave at 19.5 kHz was utilized in conjunction with a ZC signal occupying a bandwidth of 20–24 kHz. In order to facilitate implementation on commercial devices, the sampling rate was set to 48 kHz. The equation used for generating the ZC sequence is provided below

$$z[n] = \begin{cases} e^{-j\pi \frac{u}{N_{z\,c}} n(n-1)}, & \text{for odd } N_{z\,c} \\ e^{-j\pi \frac{u}{N_{z\,c}} n^2}, & \text{for even } N_{z\,c} \end{cases} \tag{1}$$

where N_{zc} is the length of the sequence, u is a parameter that specifies the ZC sequences, which is also called the root index, and n is the time index. In this article, the ZC sequence was generated with a length of 63 and a root index of 1. A length of 63 strikes a balance between resolution and bandwidth, allowing V-Pen to achieve precise tracking accuracy with around 4 kHz bandwidth. Regarding the choice of the root index, one of the valuable characteristics of the ZC sequence is that two sequences that were generated with co-prime root indexes have nearly zero cross correlation [19]. V-Pen used 1 for convenience as there is only one transmitter in the system.

After determining the initial position, the ZC signal was replaced with an additional set of seven sinusoidal waves. These sinusoidal waves were generated with frequencies starting from 21 kHz, then, a constant gap of 350 Hz between each consecutive frequency until 23.1 kHz. It is important to note that the generation of the ZC signal continues during this stage, although the amplitude of the ZC signal was set to zero. It aims to prevent the loss of synchronization between the transceivers when the ZC signal is needed for resetting the absolute distance. This modification ensures a seamless and uninterrupted tracking process throughout the handwriting input session.

C. Modulation

One notable characteristic of ZC sequences is that their correlation properties remain unchanged regardless of whether fast Fourier transform (FFT) [29] or inverse fast Fourier transform (IFFT) operations are applied to the sequence. Exploiting this property, the real part of the FFT results of the generated sequence could be inserted into the central portion of a 384-bit blank array. Subsequently, an IFFT operation was performed on the array. The real part of the resulting signal was extracted to obtain the modulated sequence. In this configuration, the modulated signal occupied the frequency band of 20–24 kHz, with a period of 8 ms. The frequency spectrum of the modulated ZC sequence is shown in Fig. 3.

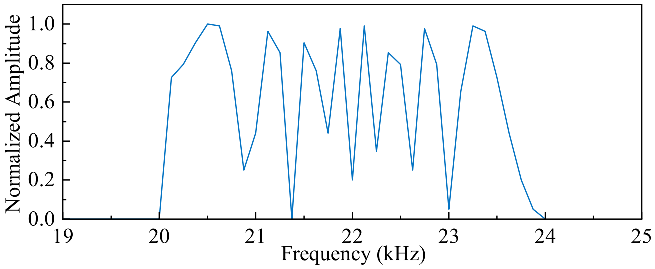


Fig. 3. Frequency spectrum of the modulated ZC signal.

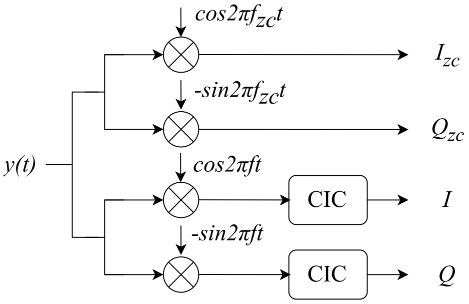


Fig. 4. Demodulation procedure.

The sinusoidal waves can be conveniently generated as follows:

$$\cos(2 \times \pi \times f_c \times t) \tag{2}$$

where f_c is the carrier frequency and t represents time. The modulated ZC sequence and sinusoidal waves were added together for transmission.

D. Demodulation

The demodulation process can be separated into two distinct paths as shown in Fig. 4. To process the ZC signal and neutralize potential interference caused by the sinusoidal waves, the carrier frequency $f_{\text{zc}} = 19.5$ kHz of the wave sent together with the ZC signal was employed for demodulating the received signal. The conventional procedure of IQ demodulation [30] was applied, retaining only the real portion of the demodulated results I_{zc}. Despite the demodulation process generating an additional sinusoidal wave at a higher frequency, it was unnecessary to filter it out due to the weak cross correlation between the ZC sequence and other signals. Furthermore, for the ensuing correlation computation, which requires a reference signal, the previously acquired modulated ZC signal is also demodulated at the frequency of 19.5 kHz. This demodulation result was referred to as the *reference signal* in Section II-E. Another demodulation path caters to the sinusoidal waves, employing their inherent frequencies f (e.g., 21 kHz) for demodulation.

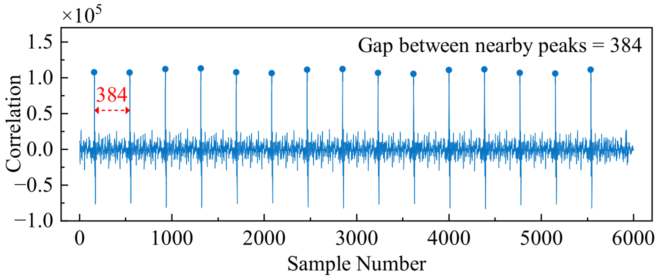


Fig. 5. Correlation graph for synchronization.

To preserve the DS information caused by the moving hand, a high-efficiency low-pass filter known as the cascaded integrator-comb (CIC) filter [31] was utilized. Additionally, the CIC filter played a crucial role in down-sampling, thereby reducing the volume of data to be processed in subsequent steps. The CIC filter parameters were set as follows: a decimation ratio (R) of 16, a delay (M) of 17, and the number of stages (N) set to 3. These parameter settings ensured efficient filtering while maintaining the integrity of the DS information. For this particular pathway, both the I and Q components were retained for future phase change measurement.

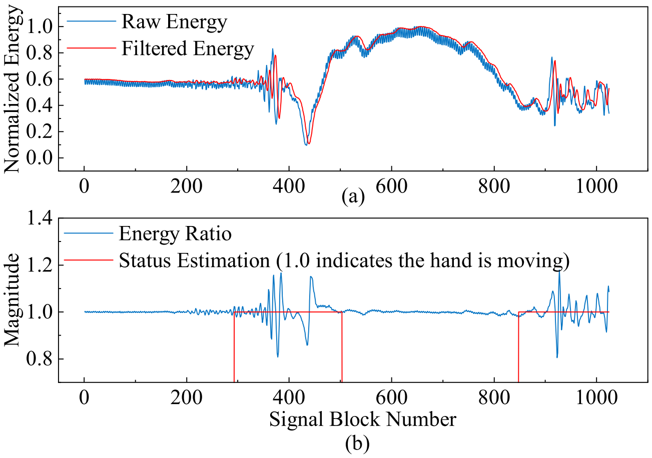
E. Transceiver Synchronization

Correlation peaks can arise from both the direct signal path between the transceivers and the echo originating from nearby objects. However, the direct path produced the strongest peak due to its relatively higher energy compared to the echoes. At the initiation of the algorithm, considering that the transceivers take some time to be stable, the initial 80 000 samples were ignored. The correlation between the next demodulated 6000 samples and the *reference signal* was assessed. As anticipated, correlation peaks are observed at consistent intervals equal to the length of the ZC signal ($N = 384$) in Fig. 5. These peaks signify the moments when the receiver captures the signal directly transmitted from the transmitter. The synchronization process only needs to be performed once at the algorithm's start, as the subsequent correlation peaks can be reliably predicted.

F. Hand Status Estimation

Prior to obtaining energy information, the initial step involves filtering out environmental noises. To achieve this, a CIC filter was already implemented during the demodulation process. By computing the total energy of the IQ signal within each signal block, it becomes evident that the energy levels vary significantly based on the hand's status. However, due to the presence of noise, it is not feasible to determine the hand's status solely by setting a threshold. Therefore, an additional three-stage CIC filter was employed to further mitigate the noise as shown in Fig. 6(a), with the delay and decimation factor both set to be 1.

Ultimately, by leveraging the energy change ratio between consecutive signal blocks and appropriate threshold values, it becomes straightforward to ascertain the status of the hand as shown in Fig. 6(b). When initializing the V-Pen, the algorithm captured the highest and lowest energy ratios while assuming a static environment. Subsequently, the thresholds th_1 and th_2 were established by adding 0.02 to the maximum ratio value and subtracting 0.02 from the minimum ratio value obtained. These values were chosen to establish a tolerance range around the extreme energy ratios observed during initialization. This choice enables the algorithm to gracefully accommodate variations in real-world scenarios while retaining sensitivity to significant changes in the hand's status. The counter variable plays an essential role in filtering transient fluctuations, it increases when energy ratios deviate from the threshold range. Upon reaching a predefined threshold of 5, the counter triggers a reset of the status estimation to 0, indicating the cessation of hand movement. During algorithm development, it was observed that the energy ratio might not change rapidly enough or fail to reach thresholds during slow hand movement or near cessation due to directional changes. Consequently, additional thresholds E_1 and E_2 were introduced to provide extra tolerance for these phenomena. The quantity of data included in E_1 and E_2 was fine-tuned based on hand status estimation performance to strike a balance between sensitivity and specificity. In order to achieve a precise estimation of the hand condition, specific conditions based on these thresholds were devised in Algorithm 1.



(a)

(b)

Fig. 6. Energy information and hand status estimation. (a) Energy filtering. (b) Hand status estimation based on energy ratio.

G. Initial Position Estimation

Following the application of demodulation to the received signal, a correlation is established between this signal and the reference signal as shown in Fig. 7(a). In order to minimize the impact of environmental noise under 17 kHz [16] and direct signal interference, the disparity between two successive correlation results was computed.

Subsequently, the ToF could be determined by locating the correlation peak resulting from the movement of the hand [19]. For example, in Fig. 7(b), the absolute distance could be obtained by utilizing ToF and the speed of sound

$$\begin{aligned}\text{Absolute distance}\\ &= \left(\frac{\text{Speed of sound (cm/s)}}{\text{Sampling rate}}\right) \times \text{(Delay in sample)}\\ &= \left(\frac{34\,300}{48\,000}\right) \times (384 - 351) = 23.58 \text{ cm.} \qquad (3)\end{aligned}$$

Algorithm 1 Hand Status Estimation

```
Input: Energy Ratio
Output: Status Estimation
Initialize:
Status estimation = 0 (0 indicates the hand is NOT moving and 1 indicates the hand is moving)
Counter = 0
$E_1$ ← last three consecutive energy ratio values
$E_2$ ← last five consecutive energy ratio values

if status estimation == 0 then
    if energy ratio>= $th_1$ or energy ratio <= $th_2$ then
        status estimation ← 1
    else
        status estimation ← 0
    end if
else
    if energy ratio< $th_1$ or energy ratio > $th_2$ then
        counter += 1
    end if
    if energy ratio>= $th_1$ or energy ratio <= $th_2$ or all of $E_1 > 1$ or all of $E_1 < 1$ or $(max(E_2) - min(E_2)) > (th_1 - 0.02 - (th_2 + 0.02) + 0.002)$ then
        counter = 0
    end if
    if counter = 5 then
        status estimation ← 0
        counter ← 0
    end if
end if
```

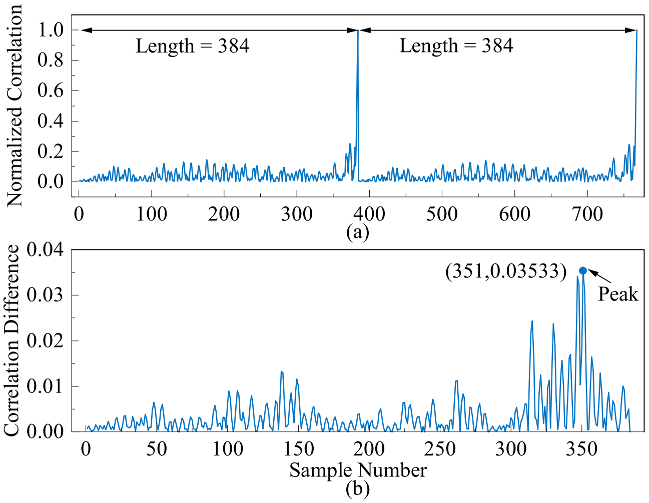


Fig. 7. Correlation results. (a) Correlation of two consecutive signal blocks. (b) Correlation difference of two consecutive signal blocks.

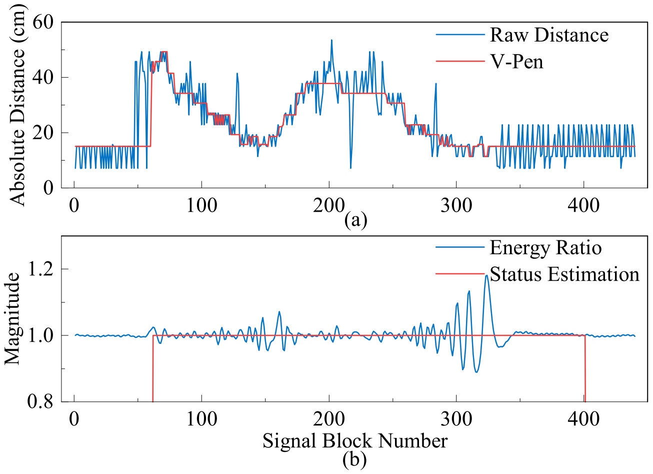


Fig. 8. Initial position estimation from ToF. (a) Raw distance results from ToF and smoothed results of V-Pen. (b) Hand status estimation based on energy ratio.

However, the raw ToF results still exhibit considerable levels of noise as Fig. 8(a) shows, especially when the hand remains stationary. Recognizing that ToF should remain relatively stable over short time intervals, only ToF measurements with a maximum difference among five consecutive readings below a predefined threshold (i.e., 6 cm in V-Pen) were recorded. Another limitation of the ToF approach is that the correlation peak persists regardless of the presence of a moving hand. By utilizing the hand status information, valid initial position estimation results could be derived based on the hand status shown in Fig. 8(b). When the hand status estimation indicates that movement has stopped, the average of the last ten distance values was computed as the final result.

H. Relative Distance Measurements

The variation in phase reflects the alteration in the relative distance of the signal transmission path. A phase change of 2π corresponds to a wavelength change in distance [32]. To achieve precise measurement of the phase change in the signal, the local extreme values detection (LEVD) algorithm was implemented by LLAP [17], which was inspired by the well-known empirical mode decomposition (EMD) algorithm [33]. This algorithm compares the local extreme values with a threshold (i.e., five times the standard deviation of I/Q measured under a static environment) and extracts the signal that represents the echo from the hand, thus, enabling accurate determination of the phase change. However, while implementing LEVD in V-Pen, it was observed that the algorithm could be wrongly triggered under noise. For example, LEVD was activated by local extreme values caused by the noise as Fig. 9 shows, and then output the wrong phase change as a result. Once LEVD has been initialized (i.e., found two satisfied extreme values with a difference larger than the threshold), the modified LEVD in V-Pen and the original version show similar performance as indicated by Fig. 9. The pseudocode of the modified LEVD algorithm is given in Algorithm 2. Additionally, as there were several sinusoidal waves, each of them can extract independent distance change information. Theoretically, the total distance change measured by each sinusoidal wave should be the same, however, it can be affected by the noise. As the speed of the hand remains constant during a very short period (i.e., 8 ms), the distance change should be linear during this time. V-Pen measured the linearity of the distance change detected and ignored the measurements from the frequency with the most nonlinear

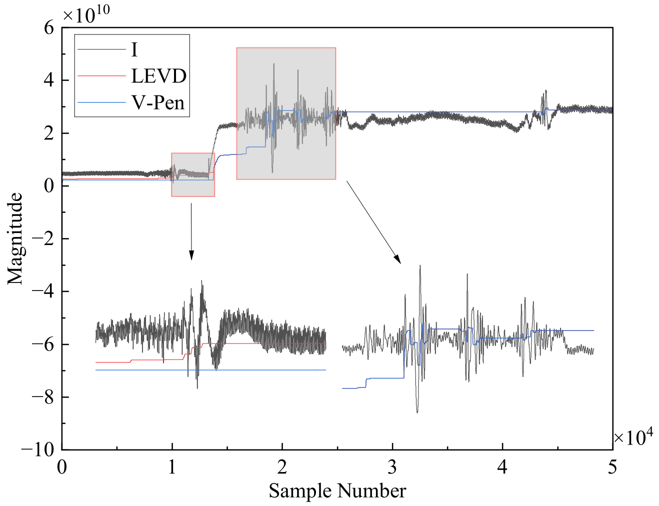


Fig. 9. Results of the original and improved LEVD algorithms.

distance changes. The rest of the results were averaged to reduce the noise.

I. Hand Position Estimation

With the original initial position and continuous relative distance change information, the length of the signal transmitting path was monitored. As there were two microphones on the mobile phone, each of them can provide distance information. Establishing one microphone and the speaker to be the focus of the ellipse as demonstrated in Fig. 10, two equations can be formulated. Equation (4) can be simplified to be a standard equation of a circle as the transceivers share the same position

$$x^2 + y^2 = \frac{p_1^2}{4} \tag{4}$$

$$\frac{4(x - L/2)^2}{p_2^2} + \frac{4y^2}{p_2^2 - L^2} = 1. \tag{5}$$

Then the coordinates of the hand can be obtained by solving the above equations

$$x = \frac{L^2 - p_2^2 + p_1 p_2}{2L} \tag{6}$$

$$y = \frac{\left(-\left(L^2 - p_2^2\right)(L + p_1 - p_2)(L - p_1 + p_2)\right)^{1/2}}{2L}. \tag{7}$$

J. Enhancement for Input Applications

For the purpose of recognition, not all the tracking results are needed. For example, the number “4” is usually written in two separate strokes. If only one stroke is allowed, the trajectory will be mixed with redundant contents. To solve this problem, the algorithm continuously detects the total moving distance in the last 0.8 s. A threshold (i.e., 1.5 cm in V-Pen) was set to detect whether the hand stops moving, then the system goes into the stroke change mode accordingly. There would be a black spot on the screen displaying the current position of the hand. Once the hand stops moving, which indicates that it has been moved to the start position of the next stroke, the system will be back to normal writing mode. If the hand does not move in the stroke change mode, the system will consider it as the end of the current character, so that the number of strokes needed can be fully controlled by the user, instead of setting it as a fixed number.

Algorithm 2 V-Pen LEVD Algorithm

```
Input: $I/Q$ Vectors
Output: Echo Reflected by Hand
Initialize:
Echo from hand = $D(t)$, $t = 0, \dots, T$
Echo from where else = $S(t)$, $t = 0, \dots, T$
Extreme list = $E(n)$, $n = 0$
Extreme update value = $EV$, $EV = 0$
$I/Q$ mean value in a static environment = $S_{end}$
All extreme values in the current signal block = $peak_{local}$
All extreme values found = $peak_{max}$ and $peak_{min}$
Initialization ($ini$) = $false$

$S(0) = S_{end}$
if $ini == false$ and $peak_{max}$ is not empty and $peak_{min}$ is not empty then
  compare location of latest $peak_{max}$ and $peak_{min}$
  update: latest extreme is a max/min
  $n \leftarrow n + 1$, $E(n) \leftarrow$ latest $peak_{min}/peak_{max}$
  $n \leftarrow n + 1$, $E(n) \leftarrow$ latest $peak_{max}/peak_{min}$
  $ini \leftarrow true$
end if
if $ini == false$ then
  $S(t) \leftarrow S(0), t = 1, \dots, T$
end if
if $ini == true$ then
  if $peak_{local}$ is empty then
    if $EV == 0$ then
      $S(t) \leftarrow S(0), t = 1, \dots, T$
    else
      $S(t) \leftarrow S(t-1) * 0.9 + EV, t = 1, \dots, T$
    end if
  else
    for $t \leftarrow 1$ to $T$ do
      if both latest extreme and $S(t)$ is max/min and $S(t)$ is larger/smaller than $E(n)$ then
        $E(n) \leftarrow S(t)$
      end if
      if one of latest extreme and $S(t)$ is max and the other is min then
        $n \leftarrow n + 1$, $E(n) \leftarrow S(t)$
        update: latest extreme is a max/min
      end if
      if $E(n) - E(n-1) > threshold$ then
        $EV \leftarrow 0.1 * (E(n) + E(n-1))/2$
        $S(t) \leftarrow S(t-1) * 0.9 + EV$
      else
        $S(t) \leftarrow S(t-1) * 0.9 + EV$
      end if
    end for
    $S_{end} \leftarrow S(T) * 0.9 + EV$
  end if
end if
$D(t) \leftarrow I/Q - S(t), t = 0, \dots, T$
```

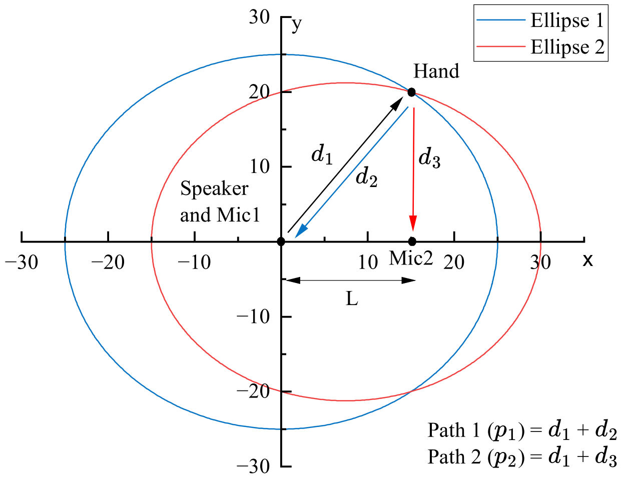


Fig. 10. Intersection of the ellipses.

K. Character Recognition

To recognize the trajectory, the MNIST [27] and EMNIST [28] datasets were used for English letters and numbers. The recorded coordinates of the trajectory were used to draw images. These images then will be used as input for the model trained by the mentioned database. Additionally, Qpen [34] was used for Chinese characters.

III. Implementation

V-Pen has been successfully implemented on a commercial mobile phone Huawei P20 Pro, using Java for all processing algorithms. The application consists of four main threads. The first thread is responsible for playing acoustic signals, while the second thread writes received signals to a file. The remaining two threads handle signal processing and drawing the results, respectively.

Similar to CTrack [35], the implementation of the V-Pen system is primarily constrained by the design and performance of the embedded transceivers. To achieve optimal echo strength, it is necessary for the hand to be positioned facing the transmitter, and the transmitter's power should ideally be high. However, due to the hardware configuration involving two microphones facing opposite directions and the main speaker positioned opposite one of the microphones, using the main speaker as the transmitter would compromise one of the microphones' ability to capture a significant amount of echo. Consequently, the signal is routed through the earpiece while the user writes in the air above the screen. As a result, the effective range of the V-Pen system is affected by the power of the earpiece, as it is only a tenth of the main speaker's output. The decision to utilize Samsung S5 for algorithm implementation in previous research was influenced by the high signal-to-noise ratio (SNR) of its transceivers, a characteristic that may not be present in newer phone models. The noise level of the received signal is discernibly higher in comparison to earlier studies, as evidenced by comparing the I/Q signal from V-Pen and the graph given in LLAP [17]. This suboptimal SNR can affect the accuracy of distance estimation. Furthermore, the performance of distance estimation for signals from Mic2 is impacted by its physical distance from the signal source. Consequently, the absolute distance estimation process exclusively relies on Mic1, necessitating the hand to stop right above the earpiece at the end of the initial position estimation or absolute distance reset phase, the height of the hand is then determined through absolute distance estimation. Therefore, the position of the hand can be confirmed as $(0, \text{height})$. Regarding the latency of the algorithm, no issues were observed with the Kirin 970 processor, which was developed in 2017. The performance of the V-Pen system might be improved with the adoption of newer, more advanced devices.

IV. Performance Evaluation

Three assessments of V-Pen's performance were evaluated. Firstly, its 1-D-ranging capabilities were evaluated by measuring both 1-D relative and absolute distance errors. The effects of the hand configuration were also discussed. Secondly, a square template was employed to evaluate its 2-D tracking performance. Finally, the accuracy of trajectory recognition was evaluated.

A. One-Dimensional Ranging

1) Relative Distance Ranging Error: To assess the relative distance ranging performance, a ruler was placed near the mobile phone to measure the ground truth distance, and the hand was moved from a position 20 cm away from the mobile phone and brought to a stop at a 15 cm position. The ranging error is defined to be the absolute value of the difference between the algorithm's results and the ground truth values measured by the ruler. Fig. 11(a) illustrates the cumulative distribution function (cdf) of the measurement error under different noise levels. Under a normal environment (i.e., 45 dB), the average error over 50 movements was found to be 2.9 mm, which is comparable to LLAP [17] and Strata [18] as they all rely on detecting phase changes. The 90th percentile measurement error was determined to be 5.2 mm. While there are noises from speech and music, the ranging error also increases accordingly. The average error in environments with noise levels of 55 and 65 dB was measured to be 6.8 and 9.1 mm, and the 90th percentile measurement error was around 11.2 and 13.5 mm, respectively. The sub-centimeter level of average measurement error demonstrates the high robustness of V-Pen. Additionally, the error was evaluated while the hand moved at different distances under a normal environment. As shown in Fig. 11(b), when the hand position is within the range of 50–25 cm, the average error is 4.4 mm, and the standard deviation varies from 1.8 to 7.7 mm. The results indicate that V-Pen slightly outperforms LLAP [17] due to improvements made in the LEVD algorithm. However, errors of over 10 mm occur when the hand is more than 30 cm away from the device, which is attributed to the limited power of the transmitter. Moreover, when the hand is too close to the device (e.g., 5 cm), the performance also slightly deteriorates due to the multipath effects.

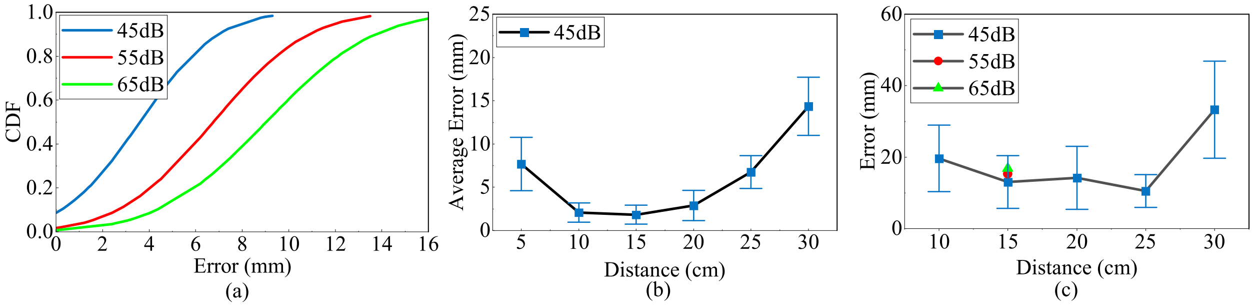


Fig. 11. One-dimensional ranging performance evaluation. (a) Movement error under different noise levels. (b) Relative distance measurement error at difference distance. (c) Absolute distance measurement error at different distances.

2) Absolute Distance Ranging Error: The absolute distance ranging performance was demonstrated in Fig. 11(c), the error was measured by moving the hand from any position and stopping at a distance of 15 cm away from the device. In a normal environment, the average and median errors were 1.30 and 1.07 cm, which outperforms LLAP [17] (i.e., average 1.8 cm) and is similar to Strata [18] (i.e., median 1.0 cm). However, the performance under noise was not evaluated in Strata [18] while the average error of V-Pen only increased to 1.53 and 1.68 cm when the noise levels reached 55 and 65 dB, respectively. The minor impact on accuracy caused by noise underscores the robustness of the ToF-based method. The absolute distance estimation shows a similar performance when the hand stops between 15 and 25 cm as shown in Fig. 11(c), the error is averaged as 1.26 cm. Similar to the relative distance ranging performance, the accuracy decreases when the hand is 30 cm away from the device, as the power of the transmitter and the strength of the received echo can also affect the absolute distance ranging performance. Also, when the hand is too close to the device, the accuracy drops as well.

3) Effects of Hand Configurations: The hand-tracking system's performance was also evaluated across diverse hand configurations, encompassing variations in the number of fingers and hand poses. Specifically, the evaluation was performed in scenarios where the user held out 4–5 fingers, 2–3 fingers, and 1 finger. When the palm faced down, the relative distance error ranged from 2.9 to 8.6 mm, the performance dropped as the number of fingers held out decreased, and the absolute distance error varied from 1.3 to 1.6 cm. The change in number of the fingers adjusts the size of the area facing the transmitter on the bottom, which results in a difference in the strength of the echo received. Based on the results of the experiment, the area facing down is directly linked to the relative distance measurement accuracy. However, the absolute distance estimation appeared to be less affected by variations in hand configuration. Another group of experiments was repeated by letting the palm face left. When 4–5 fingers were held out, the area facing downward decreased significantly compared to the palm facing down scenario, resulting in an expected drop in performance. Notably, in scenarios where users held out 2–3 fingers, the bending of the remaining fingers increased the area facing downward, resulting in enhanced tracking accuracy compared with the 4–5 fingers case. The accuracy measurements when palms faced left included a relative distance error of 7.3 to 11.7 mm and an absolute distance error of 1.6 to 1.8 cm. Remarkably, the absolute distance error exhibited resilience to variations, similar to the observations when the palm faced down. Additionally, the system exhibited consistent performance when tilting the palms at a 45° angle, showcasing comparable results to configurations with no tilt. This resilience to variations in hand poses underscores the system's adaptability in real-world scenarios, further emphasizing its robust tracking capabilities. In the subsequent performance evaluations, the default setting will be palm facing down with five fingers held out. Additionally, the accuracy when 2–3 fingers were held out will be discussed to assess system performance under slightly nonideal hand configurations.

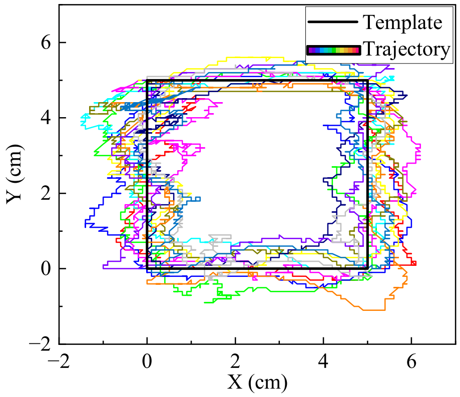


Fig. 12. Template and multiple tracking results.

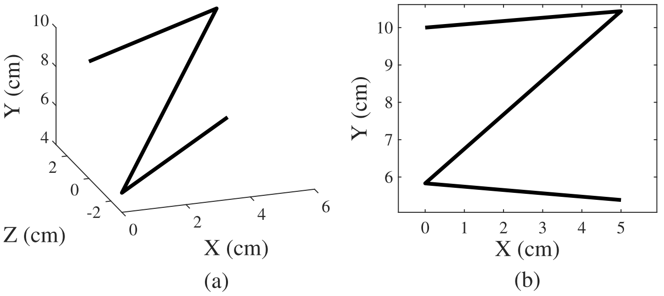


Fig. 13. Effects of 3-D variations. (a) Ground truth 3-D movement. (b) 2-D trajectory obtained by V-Pen.

B. Two-Dimensional Tracking

In order to assess the 2-D tracking performance, a 5 × 5 cm square template was employed. Once the initial position of

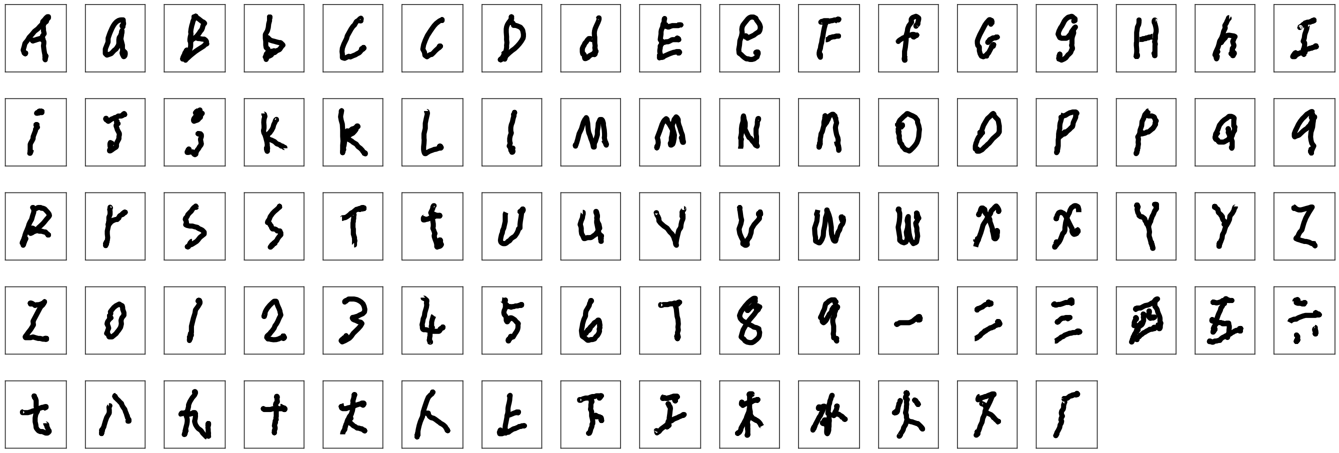


Fig. 14. Tracking results for 52 English letters, ten numbers, and 20 Chinese characters.

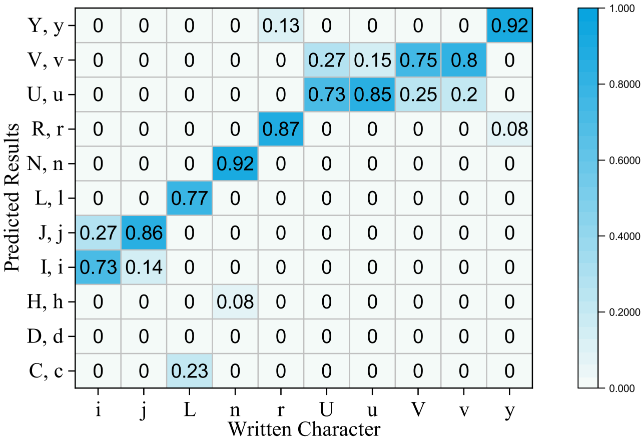


Fig. 15. Normalized confusion matrix.

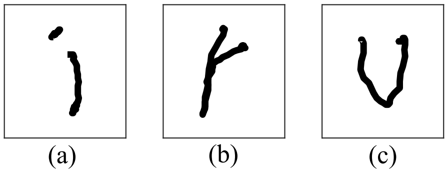


Fig. 16. Examples of misclassification results. (a) Classify “i” as “j.” (b) Classify “r” as “y.” (c) Classify “V” as “U.”

the hand is determined, the template will be displayed on the screen for the user to follow its outline while moving their hand. The user interface provides real-time visualization of the tracking results, allowing the user to make small adjustments to compensate for any errors during movement. An example of the template and some tracking results are given in Fig. 12, which shows the fine accuracy and repeatability of the V-Pen. The definition of 2-D error is defined as the distance between the points on the tracking results to the nearest point on the template, which is the same as LLAP [17] and Strata [18]. The system achieves an average and median tracking error of 4.3 and 4.0 mm for 2-D tracking, which outperforms LLAP [17] (i.e., average 4.6 mm) and Strata [18] (i.e., median 5.7 mm). While using the 2–3 fingers hand configuration, the average tracking error dropped to 4.7 mm. The superior performance on absolute distance estimation significantly contributes to the 2-D tracking accuracy.

Additionally, a noteworthy difference in the error magnitude between the x-axis and y-axis was observed, because the hand position estimation algorithm has different tolerance to the distance measurements error in different moving directions. Specifically, during the transition from coordinates (0, 10) to (1, 10), inducing a 0.5 cm deviation in the y-axis necessitates distance measurement errors of approximately 1.2 and 1.0 cm for each microphone, respectively. In contrast, achieving a 0.5 cm error along the x-axis during the progression from (0, 10) to (0, 9) requires significantly smaller distance measurement errors of approximately 0.42 and 0.02 cm for each microphone, respectively.

Another factor that could affect the accuracy of the tracking and recognition is the 3-D variation during the writing period. Considering that the accurate 3-D variation of the hand is hard to control, a simulation was performed. Assuming writing a “Z” in the air and including some involuntary movements in the 3-D, the moving trajectory can be drawn by the following four positions as Fig. 13(a) shows: $(0, 10, 0)$, $(5, 10, 3)$, $(0, 5, -3)$, and $(5, 5, 2)$. As Section II-F stated, the algorithm assumed that the hand is always in the 2-D plane. When there is a movement in the 3-D, the algorithm is actually calculating the rotated position of the ground truth along with the x-axis. The estimated hand position is given in Fig. 13(b), the maximum error of the trajectory is still under 1 cm for a 3 cm 3-D variation in this case. In real application, the variations would be expected to be much less than 3 cm.

C. Trajectory Recognition

To assess the recognizability of the trajectory captured by V-Pen, Qpen [34] was used for Chinese characters, and the models trained with MNIST [27] and EMNIST [28] datasets were utilized [36] for numbers and English letters. V-Pen allows the user to input multiple strokes and control the number of strokes required (e.g., two strokes for number “4” and three strokes for letter “E”). For the purpose of evaluation, three participants were engaged in the experiments, with each individual performing multiple repetitions of writing every character, ranging from 3 to 5 times. The volume of the characters is constrained to a maximum of 10 × 10 cm, and the average dimension of the characters written was 3.2 × 5.9 cm. A demonstration was given to the volunteers regarding how the system works, such as how to write the next stroke or finish the current character. And they were given 10 min to get familiar with the system before experiments. Examples of English letters and digit numbers tracked by V-Pen are illustrated in Fig. 14. The tracking results in the figure benefit from the multistroke input function and have greater similarity to the real-world handwritten trajectory compared with the previous research. V-Pen achieves an impressive recognition accuracy of 94.8% for 52 English letters, ten numbers, and 20 Chinese characters. Regarding the recognition accuracy while using the 2–3 fingers hand configuration, as there is only a 0.4 mm drop in the 2-D tracking performance, the recognition performance remains similar. For 3-D variations during the movement, the sub-centimeter error incurred in the tracking results also does not affect the recognition accuracy as the shape of the character remains similar. In comparison, LLAP [17] achieved over 90% recognition accuracy by sacrificing some characters and relying on MyScript [26]. In contrast, V-Pen offers enhanced flexibility and a cost-effective approach for future implementations, enabling precise recognition of English letters and numbers without the need for a third-party recognition platform.

The confusion matrix delineating misrecognized English characters is presented in Fig. 15. Characters sharing similar shapes or structures as given in Fig. 16, such as "V" and "U," are particularly susceptible to confusion, a common phenomenon in prior literature [28]. Remarkably, for numeric characters, the model, trained on the MNIST dataset, exhibited commendable performance, achieving a flawless 100% recognition accuracy for the ten numbers. In terms of Chinese characters, V-Pen attains an accuracy rate of 86%. Notably, the recognition dataset contains the entirety of Chinese characters, diverging from V-Pen's focus on a set of 20 commonly used characters. Consequently, instances of misclassification often result in the recognition of characters beyond the selected 20, precluding the generation of a meaningful confusion matrix in this context.

V. Conclusion

This research presents V-Pen, an innovative acoustic-based hand-tracking solution designed for input systems. V-Pen achieves remarkable tracking accuracy at the sub-centimeter level, ensuring precise and reliable tracking results. Moreover, the algorithm's efficient processing capabilities enable real-time applications, enhancing user experience and system responsiveness. In addition, a novel hand status estimation algorithm is introduced to tackle issues caused by redundant data, resulting in further improvements in tracking performance. Furthermore, to overcome limitations commonly encountered in previous approaches, V-Pen successfully addresses the challenge of substandard recognition accuracy for characters with multiple strokes. These advancements collectively contribute to the effectiveness and usability of V-Pen as an input system solution.

Future work will study further improvements on the absolute distance estimated with limited bandwidth, and investigate the solution to overcome the multipath effects in order to use a single sinusoidal wave to measure accurate phase change instead of the averaging results from multiple waves.

References

[1] K. Terajima, T. Komuro, and M. Ishikawa, "Fast finger tracking system for in-air typing interface," in *Proc. CHI Extended Abstr. Human Factors Comput. Syst.* New York, NY, USA: Association for Computing Machinery, Apr. 2009, pp. 3739–3744, doi: 10.1145/1520340.1520564.

[2] J. Lien et al., "Soli: Ubiquitous gesture sensing with millimeter wave radar," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 1–19, Jul. 2016, doi: 10.1145/2897824.2925953.

[3] K. M. Sagayam and D. J. Hemanth, "Hand posture and gesture recognition techniques for virtual reality applications: A survey," *Virtual Reality*, vol. 21, no. 2, pp. 91–107, Jun. 2017, doi: 10.1007/s10055-016-0301-0.

[4] S. Jiang et al., "Feasibility of wrist-worn, real-time hand, and surface gesture recognition via sEMG and IMU sensing," *IEEE Trans. Ind. Informat.*, vol. 14, no. 8, pp. 3376–3385, Aug. 2018.

[5] V. Tam and L.-S. Li, "Integrating the Kinect camera, gesture recognition and mobile devices for interactive discussion," in *Proc. IEEE Int. Conf. Teaching, Assessment, Learn. Eng. (TALE)*, Aug. 2012, pp. H4C-11–H4C-13.

[6] Q. Li, Z. Luo, and J. Zheng, "A new deep anomaly detection-based method for user authentication using multichannel surface EMG signals of hand gestures," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–11, 2022.

[7] A. Calado, V. Errico, and G. Saggio, "Toward the minimum number of wearables to recognize signer-independent Italian sign language with machine-learning algorithms," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–9, 2021.

[8] B. Li, J. Yang, Y. Yang, C. Li, and Y. Zhang, "Sign language/gesture recognition based on cumulative distribution density features using UWB radar," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.

[9] Z. Chen, G. Li, F. Fioranelli, and H. Griffiths, "Dynamic hand gesture classification based on multistatic radar micro-Doppler signatures using convolutional neural network," in *Proc. IEEE Radar Conf. (RadarConf)*, Apr. 2019, pp. 1–5.

[10] G. Plouffe and A.-M. Cretu, "Static and dynamic hand gesture recognition in depth data using dynamic time warping," *IEEE Trans. Instrum. Meas.*, vol. 65, no. 2, pp. 305–316, Feb. 2016.

[11] A. R. Varkonyi-Koczy and B. Tusor, "Human–computer interaction for smart environment applications using fuzzy hand posture and gesture models," *IEEE Trans. Instrum. Meas.*, vol. 60, no. 5, pp. 1505–1514, May 2011.

[12] Z. Wang et al., "Hand gesture recognition based on active ultrasonic sensing of smartphone: A survey," *IEEE Access*, vol. 7, pp. 111897–111922, 2019.

[13] X. Xu, J. Yu, Y. Chen, Y. Zhu, and M. Li, "SteerTrack: Acoustic-based device-free steering tracking leveraging smartphones," in *Proc. 15th Annu. IEEE Int. Conf. Sens., Commun., Netw. (SECON)*, Jun. 2018, pp. 1–9.

[14] W. Chen and Z. Zhang, "Hand gesture recognition using sEMG signals based on support vector machine," in *Proc. IEEE 8th Joint Int. Inf. Technol. Artif. Intell. Conf. (ITAIC)*, May 2019, pp. 230–234.

[15] H. Wang and W. Gong, "RF-pen: Practical real-time RFID tracking in the air," *IEEE Trans. Mobile Comput.*, vol. 20, no. 11, pp. 3227–3238, Nov. 2021.

[16] S. Yun, Y.-C. Chen, and L. Qiu, "Turning a mobile device into a mouse in the air," in *Proc. 13th Annu. Int. Conf. Mobile Syst., Appl., Services*. New York, NY, USA: Association for Computing Machinery, May 2015, pp. 15–29, doi: 10.1145/2742647.2742662.

[17] W. Wang, A. X. Liu, and K. Sun, "Device-free gesture tracking using acoustic signals," in *Proc. 22nd Annu. Int. Conf. Mobile Comput. Netw.* New York, NY, USA: Association for Computing Machinery, Oct. 2016, pp. 82–94, doi: 10.1145/2973750.2973764.

[18] S. Yun, Y.-C. Chen, H. Zheng, L. Qiu, and W. Mao, "Strata: Fine-grained acoustic-based device-free tracking," in *Proc. 15th Annu. Int. Conf. Mobile Syst., Appl., Services*. New York, NY, USA: Association for Computing Machinery, Jun. 2017, pp. 15–28, doi: 10.1145/3081333.3081356.

[19] C. Liu, P. Wang, R. Jiang, and Y. Zhu, "AMT: Acoustic multi-target tracking with smartphone MIMO system," in *Proc. IEEE INFOCOM IEEE Conf. Comput. Commun.*, May 2021, pp. 1–10.

[20] T. Liu, X. Niu, J. Kuang, S. Cao, L. Zhang, and X. Chen, "Doppler shift mitigation in acoustic positioning based on pedestrian dead reckoning for smartphone," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–11, 2021.

[21] W. Mao, J. He, and L. Qiu, "CAT: High-precision acoustic motion tracking," in *Proc. 22nd Annu. Int. Conf. Mobile Comput. Netw.* New York, NY, USA: Association for Computing Machinery, 2016, pp. 69–81, doi: 10.1145/2973750.2973755.

[22] M. T. I. Aumi, S. Gupta, M. Goel, E. Larson, and S. Patel, "DopLink: Using the Doppler effect for multi-device interaction," in *Proc. ACM Int. joint Conf. Pervasive Ubiquitous Comput.* New York, NY, USA: Association for Computing Machinery, Sep. 2013, pp. 583–586, doi: 10.1145/2493432.2493515.

[23] J. C. L. Ng, K. B. Letaief, and R. D. Murch, "Complex optimal sequences with constant magnitude for fast channel estimation initialization," *IEEE Trans. Commun.*, vol. 46, no. 3, pp. 305–308, Mar. 1998.

[24] G. Hackenberg, R. McCall, and W. Broll, "Lightweight palm and finger tracking for real-time 3D gesture control," in *Proc. IEEE Virtual Reality Conf.*, Mar. 2011, pp. 19–26.

[25] R. Nandakumar, V. Iyer, D. Tan, and S. Gollakota, "FingerIO: Using active sonar for fine-grained finger tracking," in *Proc. CHI Conf. Human Factors Comput. Syst.* New York, NY, USA: Association for Computing Machinery, May 2016, pp. 1515–1525, doi: 10.1145/2858036.2858580.

[26] *Myscript*. Accessed: May 10, 2023. [Online]. Available: https://www.myscript.com/

[27] L. Deng, "The MNIST database of handwritten digit images for machine learning research [best of the web]," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 141–142, Nov. 2012.

[28] G. Cohen, S. Afshar, J. Tapson, and A. van Schaik, "EMNIST: An extension of MNIST to handwritten letters," 2017, *arXiv:1702.05373*.

[29] R. N. Bracewell and R. N. Bracewell, *The Fourier Transform and Its Applications*, vol. 31999. New York, NY, USA: McGraw-Hill, 1986.

[30] T. Sakamoto, X. Gao, E. Yavari, A. Rahman, O. Boric-Lubecke, and V. M. Lubecke, "Hand gesture recognition using a radar echo I–Q plot and a convolutional neural network," *IEEE Sensors Lett.*, vol. 2, no. 3, pp. 1–4, Sep. 2018.

[31] G. Jovanovic Dolecek and S. K. Mitra, "On design of CIC decimation filter with improved response," in *Proc. 3rd Int. Symp. Commun., Control Signal Process.*, Mar. 2008, pp. 1072–1076.

[32] F. Gueuning, M. Varlan, C. Eugene, and P. Dupuis, "Accurate distance measurement by an autonomous ultrasonic system combining time-of-flight and phase-shift methods," in *Proc. Quality Meas. Indispensable Bridge Between Theory Reality (No Meas.? No Sci.! Joint Conf. IEEE Instrum. Meas. Technol. Conf. IMEKO TEC*, vol. 1, Jun. 1996, pp. 399–404.

[33] N. E. Huang et al., "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. Roy. Soc. London. Ser. A, Math., Phys. Eng. Sci.*, vol. 454, no. 1971, pp. 903–995, Mar. 1998. [Online]. Available: http://www.jstor.org/stable/53161

[34] *Qpen*. Accessed: May 10, 2023. [Online]. Available: https://hanzi.unihan.com.cn/Qpen

[35] H. Jiang, M. Wang, D. Liu, and S. Zhou, "CTrack: Acoustic device-free and collaborative hands motion tracking on smartphones," *IEEE Internet Things J.*, vol. 8, no. 19, pp. 14658–14671, Oct. 2021.

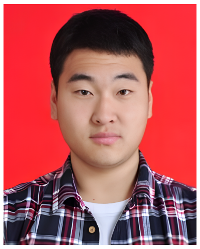
[36] H. M. D. Kabir et al., "SpinalNet: Deep neural network with gradual input," *IEEE Trans. Artif. Intell.*, vol. 4, no. 5, pp. 1165–1177, Oct. 2023.

Wei Han (Student Member, IEEE) received the B.Eng. degree in electronics and communication engineering from The University of Sheffield, Sheffield, U.K., in 2019, and the M.Sc. degree in signal processing and communications from The University of Edinburgh, Edinburgh, U.K., in 2020. He is currently pursuing the Ph.D. degree with the Agile Tomography group, School of Engineering, The University of Edinburgh.

His research interests include Human-Computer interaction, acoustic-based hand tracking, and gesture recognition.

Yinghao Li (Student Member, IEEE) was born in Beijing, China, in 1996. He received the B.S. degree in electrical and electronic engineering from North China Electric Power University (NCEPU), Beijing, in 2016, and the B.S. degree in electronic engineering from The University of Edinburgh, Edinburgh, U.K., in 2018. He is currently pursuing the Ph.D. degree with the Agile Tomography Group, School of Engineering, The University of Edinburgh.

His research interests include ultrasonic remote sensing, Human-Computer-Interaction, and deep learning in the industrial and smart sensing fields.



Hao Yu (Student Member, IEEE) was born in Heilongjiang, China, in 1996. He received the B.S. degree in electrical engineering from North China Electric Power University, Baoding, China, in 2018, and the M.Sc. degree in electrical engineering from the Harbin Institute of Technology, Harbin, China, in 2020. He is currently pursuing the Ph.D. degree with the Agile Tomography Group, School of Engineering, The University of Edinburgh, Edinburgh, U.K.

His research interests include disturbance rejection control, computational imaging, and deep learning for electrical impedance tomography in the industrial and medical fields.

Jiabin Jia (Senior Member, IEEE) received the B.Eng. and M.S. degrees in electrical and electronics engineering from Wuhan University, Wuhan, China, in 2002 and 2005, respectively, and the Ph.D. degree from the University of Leeds, Leeds, U.K., supported by the Overseas Research Students Award Scheme in 2010.

He is a Senior Lecturer with the School of Engineering, The University of Edinburgh, Edinburgh, U.K. He has authored more than 60 peer-reviewed journal articles and has led and contributed to a range of research projects. His current research interests include electrical tomography, multiphase flow measurement, medical imaging, and human machine interface.