HeadTracker: Fine-Grained Head Orientation Tracking System Based on Headphones

Jinpeng Song[1], Haipeng Dai[1(✉)], Shuyu Shi[1], Lei Wang[2], Haoran Wan[1], Zhizheng Yang[1], Fu Xiao[3(✉)], and Guihai Chen[1(✉)]

[1] Department of Computer Science and Technology, Nanjing University, Nanjing, China
{jinpengsong,wanhr,yzz}@smail.nju.edu.cn, {haipengdai,ssy,gchen}@nju.edu.cn

[2] Department of Computer Science and Technology, Peking University, Bejing, China
wang_l@pku.edu.can

[3] School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing, China
xiaof@njupt.edu.cn

Abstract. Head orientation tracking has many potential applications in various fields, *e.g.*, online courses, online meetings, and somatosensory games. Undoubtedly, with the information of the user's head orientation, these applications will have more opportunities to enhance performance and provide better user experience. However, reviewing existing works regarding head tracking, the CV-based solutions have limited tracking angle range and privacy issues and the IMU-based solutions have accumulated errors. None of these methods provide accurate and stable user head orientation. In this paper, we propose HeadTracker, a fine-grained 3D head orientation tracking system based on a single headphone. Headtracker achieves high-precision head orientation tracking by installing ultrasonic transmitters on an ordinary headphone and deploying ultrasonic receivers in the environment. We conducted experiments to evaluate the performance of HeadTracker in the real use environment, and the experimental results show that the system can achieve an average error of 6° in the 3D head orientation tracking. To the best of our knowledge, HeadTracker is the first system to use head-mounted ultrasound device to achieve 3D head orientation tracking and achieves the state-of-the-art in this category.

Keywords: Head orientation · Wireless sensing · Ultrasonic signal · Wearable devices

This work was supported in part by the National Natural Science Foundation of China under Grant 61872178, 61832005, 62102006, in part by the Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing University, and in part by the Jiangsu High-level Innovation and Entrepreneurship (Shuangchuang) Program.

L. Wang et al. (Eds.): WASA 2022, LNCS 13472, pp. 317–329, 2022.
https://doi.org/10.1007/978-3-031-19214-2_26

1 Introduction

User tracking, which refers to locating users in real time, has become the focus of many research work in recent years [2,5,6,9,12]. Nevertheless, most user tracking systems only focus on the user's location but ignore the user's head orientation, which can reveal important and valuable information, such as the user's attention and intent. If the user's accurate 3D head orientation can be obtained in real time, we can envision and expect its wide usage in many scenarios. For example, in online courses scenarios, we can know where the students' attention is through their head orientation. In addition, it also has promising usage in motion-sensing games as an alternative to mouse and keyboard. Besides, in driving scenarios, we can implement many intelligent driving applications such as estimating the driver's intention based on his head orientation.

According to our survey, most of the existing head orientation tracking work is based on computer vision [1,10,13]. These CV-based solutions can only achieve a small range of head tracking due to the narrow angle of view of camera, and they are severely affected by environmental factors such as light. Moreover, the use of cameras will bring certain privacy risks. There are also some IMU-based head tracking solutions [6], but such solutions are limited by the cumulative error of the six-axis IMU and need to be continuously calibrated in use. Although the nine-axis IMU addresses the cumulative error problem to a certain extent [4,8], it is seriously affected by the external magnetic field [3]. Most importantly, both of the CV-based solutions and the IMU-based solutions obtain head orientation in their own internal coordinate system, which is difficult to be converted to the world coordinate system for interaction with other devices. Besides, there are some solutions based on microphone arrays [15,16], but the accuracy of these solutions are relatively low.

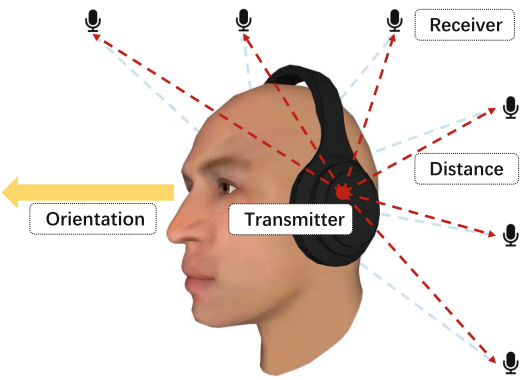


Fig. 1. HeadTracker

In this paper, we propose HeadTracker (Fig. 1), a fine-grained 3D head orientation tracking system based on headphones. Compared with other existing work, HeadTracker significantly improves the accuracy of head orientation tracking. On the hardware side, we add two ultrasonic transmitters to both sides of

an ordinary headphone and deploy some ultrasonic receivers in the environment to complete the positioning of the headphone. On the software side, we design several algorithms to calculate the head orientation and keep the system running smoothly. Specifically, the contributions of our paper are as follows:

1) We use the Zadoff-Chu sequence as the baseband signal and modulate it to the ultrasound band as our transmitting signal. We demodulate it on the receiving side, and decompose different paths from the accurate CIR. On this basis, we design a frequency division multiplexing method to realize the simultaneous positioning of two transmitters.
2) To solve the problem that signal direct path is easily blocked, we borrow the idea of GPS satellite positioning systems [7]. That is, we deploy multiple receivers in the environment and propose a receiver selection algorithm based on signal quality to accomplish positioning.
3) We use neural network to design a special head orientation tracking algorithm based on head movement recognition, which enables approximately 6-DoF head orientation tracking using only two coordinates on the head.

The remaining of this paper is organized as follows. In Sect. 2, we describe the system design and processing flow of our proposed approach HeadTracker in detail. Then, we introduce the deployment of the system and conduct a large number of experiments to evaluate the effectiveness of it in Sect. 3. Finally, we conclude this paper in Sect. 4.

2 System Design

In this section, we introduce the technical details of the HeadTracker. The system mainly consists of four modules: *signal process, headphone positioning, movement recognition, and orientation calculation*, as depicted in Fig. 2.

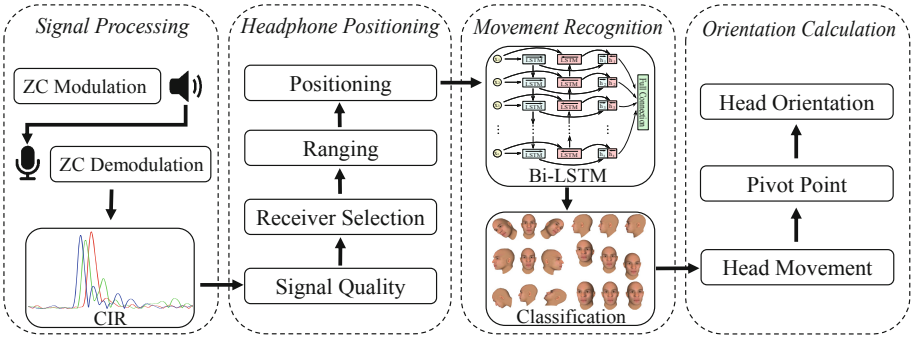


Fig. 2. Overview of HeadTracker

2.1 Signal Progressing

Signal Design. We use ZC sequences as baseband signal as ZC is a kind of CAZAC (Constant Amplitude Zero AutoCorrelation waveform), which means ZC sequences have ideal auto-correlation properties [11,14]. Compared with the common CW (Continuous Wave) signal, ZC signal can separate paths at different distances and reduce the influence of multipath. And compared with the FMCW (Frequency Modulated Continuous Wave) signal, ZC signal has better range resolution. We modulate the ZC sequence by a sinusoid carrier at the transmitter, and the mathematical form of the ZC sequence is

$$ZC[n] = e^{-j\frac{\pi un\left(n+c_f+2q\right)}{N_{ZC}}}, \tag{1}$$

where N_{ZC} is the length of ZC sequence, the value range of n is $0 \leq n < N_{ZC}$, and c_f takes 0 or 1 as the remainder of N_{ZC} modulo 2. The ZC sequence contains two integer parameters q and u. Generally, q is set to 0, and the ZC sequence degenerates into Chu sequence. Moreover, u is in the range $[0, N_{ZC}]$, and it is relatively prime to N_{ZC}.

ZC Modulation and Demodulation. In the process of signal modulation and demodulation, we use an OFDM-based interval interpolation method, which makes it possible to modulate two different ZC sequences to the same center frequency. Similarly, in the demodulation process of the received signal, we use the frequency domain interval sampling method to separate the two ZC sequences from the same received signal.

We know that according to the characteristics of the ZC sequence, the auto-correlation result of the ZC sequence is non-zero only at $t = 0$, which ideally will be a Dirac impulse function $\delta(t)$ and is a *sinc* function practically due to limited bandwidth. Because the received signal is composed of multiple transmitted signals with different time delay versions through multiple different paths, the result of cross-correlation between the transmitted signal and the received signal is $h(t)$, which is a combination of $\delta(t - \tau_i)$ signals with different time delays τ_i:

$$h(t) = \sum_{i=1}^{P} A_i e^{-j\phi_i(t)} \delta(t - \tau_i), \tag{2}$$

where P is the number of paths, A_i is the signal strength of signal path i, and ϕ_i is the phase offset of the signal on path i. And we use Dirac function here for convenience. We can express the channel impulse response (CIR) as $h(t)$, as shown in Fig. 3.

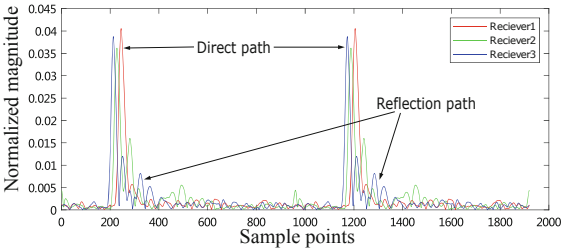


Fig. 3. CIR

Ranging. The abscissa of the CIR corresponds to the delay, while the ordinate corresponds to the cross-correlation value between transmitted signal $ZC\,[t]$ and $ZC\,[t-\tau_i]$, which is transmitted signal after a certain delay τ_i. The larger the cross-correlation value, the stronger the delayed signal. Generally, the path corresponding to the highest peak of the CIR is the direct path in the case that it is not blocked. So we can calculate the direct path using the following equation:

$$d = \underset{1<=i<=\frac{L}{2}}{\arg\max}\ CIR\,[i]\,\frac{c}{f_s}, \tag{3}$$

where L is the length of CIR. Therefore, after ranging, we can obtain the straight-line distance between each transmitter and each receiver, which makes preparations for our subsequent positioning work.

2.2 Headphone Positioning

Receiver Selection. As is well known, a major challenge for ultrasonic positioning in practice is that the direct path of sound waves between transmitter and receiver can be blocked frequently. To solve this challenge, we refer to the idea of satellite positioning systems like GPS, which is to deploy multiple satellites in orbit to achieve full coverage of the ground. Similarly, we can deploy multiple ultrasonic receivers in the environment so that no matter how the user's head rotates and moves, the direct path between each transmitter and at least three receivers is not blocked. To this end, we propose a receiver selection algorithm by which the system will select the most suitable three receivers to positioning the transmitter each time. Firstly, we propose an indicator named SNR_{los}, which is used to evaluate the signal quality between receivers and transmitters. Formally, SNR_{los} is defined as the ratio of the amplitude of the highest peak to the average of all other peaks' amplitude in the CIR.

$$SNR_{los} = \frac{\max CIR\,[i]}{\sum_{i=1}^{\frac{L}{2}} CIR\,[i] - \max CIR\,[i]}. \tag{4}$$

Positioning. After the receiver selection, each transmitter has found the three most suitable receivers. According to the triangulation method, knowing the distance between a certain point and three known anchor points, a ternary quadratic equation can be established to calculate this point's coordinates:

$$f = \begin{cases} (x - x_1)^2 + (y - y_1)^2 + (z - z_1)^2 - d_1^2 \\ (x - x_2)^2 + (y - y_2)^2 + (z - z_2)^2 - d_2^2 \\ (x - x_3)^2 + (y - y_3)^2 + (z - z_3)^2 - d_3^2, \end{cases} \tag{5}$$

where (x, y, z) is the position of the transmitter, (x_i, y_i, z_i) is the position of selected receiver, and d_i is the distance between the transmitter and receiver measured by ultrasonic ranging. We use Newton's iterative method to solve this ternary quadratic equation system. To improve the calculation speed, we set the initial iteration value of each positioning as the result of the last positioning, which can greatly reduce the number of iterations. Generally, each positioning can be completed only after three or four iterations in this way. After positioning, we can obtain the trajectory data of the headphone, which can be used to identify the current movement of the head.

2.3 Movement Recognition

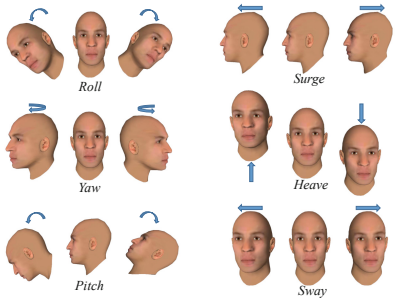


Fig. 4. Head movements

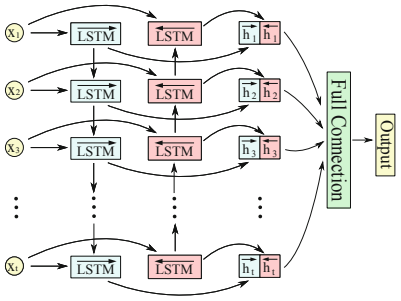


Fig. 5. Bi-LSTM

Movement Definition. We know that a rigid body has six degrees of freedom in three-dimensional space. We use the definition in the field of navigation to describe these movements, which are the three translational movements (surge, sway, and heave) and the three rotational movements (roll, pitch, and yaw), as shown in Fig. 4. Specifically, surge, heave, and sway are the translation movements along the x-axis, z-axis, and y-axis, respectively; roll, yaw, and pitch are the rotation movements around the x-axis, z-axis, and y-axis, respectively.

We ignore overly complicated head movements here as we believe that the six basic movements account for the vast majority in our daily life, while other

complex movements are relatively rare. Besides, adding other uncommon movements will increase the complexity of the classification model and reduce the overall classification performance, which we think is not worth the gain.

Classification Model. Head movements recognition is a classification task with the data of headphone trajectory. Since trajectory is a kind of time series data, we adopt Bi-LSTM as the classification model to complete this classification task. Bi-LSTM is a special kind of recurrent neural network that has a good representation ability for the time series data. Its excellent performance has been proven in many fields such as speech recognition. Bi-LSTM combines the forward LSTM with the backward LSTM as shown in Fig. 5. Therefore, Bi-LSTM can make better use from the information of the subsequent data compared with traditional LSTM.

In this step, we use the headphone trajectory as training data to train a classification model, which can be used to identify the ongoing head movement. After obtaining the head movement, we can calculate the head orientation according to some head movement rules, which is the next step of our system.

2.4 Orientation Calculation

Clearly, to determine the posture of an object in three dimensions, at least the coordinates of three different points need to be known. The posture of the rigid body is not unique with only two coordinates, because it can rotate around the axis formed by the two points. But after headphone positioning we can only obtain two points on the head, now the problem is ***how to estimate the head posture based on the positions of only two points?***

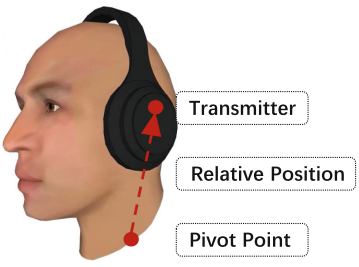


Fig. 6. Pivot point

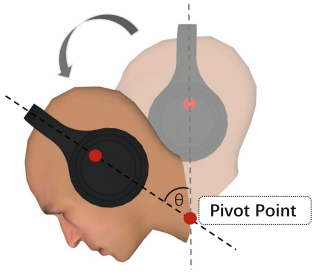


Fig. 7. Rotation

In fact, the head is not an object that can move freely in three-dimensional space, which is limited by its connection to the body. By observing and analyzing, we find some rules of head movement. That is, the movement of the human head are carried out around a point in the neck, we call it the **pivot point** (Fig. 6). The position change of the pivot point is closely related to the head's movements.

When the head is only rotating without moving, the absolute position of the pivot point is almost unchanged (Fig. 7). When the head is only moving without rotating, the relative position of the pivot point and the headphone remains almost unchanged. Therefore, these rules give us the possibility to determine the position of the pivot point by the head's movement. The relative position between the pivot point and the headphone is unchanged for a person, which is determined by the bones of the head and neck. So we only need to initialize the pivot point once at the beginning and then we can update it in real time according to the movement of the head.

The pivot point position updating formula is expressed as follows:

$$P_{pivot} = \begin{cases} \frac{P_{left}+P_{right}}{2} - V_{relative} & M \in \{surge, sway, heave\} \\ P_{pre} & M \in \{roll, pitch, yaw, static\} \end{cases}, \tag{6}$$

where P_{pivot} is the position of the current required pivot point, P_{left} and P_{right} are the positions of two transmitters, $V_{relative}$ is the vector between the midpoint of the two transmitters and the pivot point, P_{pre} is the position of the pivot point in the previous frame, and M is the ongoing movement of the head.

We now have the coordinates of the three points on the head in total, *i.e.*, the pivot point and two transmitters. Sequentially, we can calculate the head orientation according to the following formula:

$$V_{orientation} = (P_{right} - P_{pivot}) \times (P_{left} - P_{pivot}). \tag{7}$$

3 Implementation and Evaluation

3.1 Implementation

(a) Headphone

(b) Piezoceramics

(c) Coaxial line

(d) NI I/O device

Fig. 8. Hardware

Fig. 9. Experimental scene

Figure 8 shows the devices used in our experiment. We choose piezoelectric ceramics as the transmitter and receiver of ultrasonic waves. We install the two receivers on both sides of the headphone and install the receivers in the environment. We use Murata MA40H1 piezoelectric ceramics as sound sources for

transmitting and receiving ultrasonic waves. The I/O device we use is USB-6356 produced by National Instruments, which can support up to 2 analog signal outputs and 8 analog signal inputs. These piezoelectric ceramics are connected to I/O device through coaxial cables and the experimental scene is shown in Fig. 9. In the part of software, we use MATLAB to drive the device for signal acquisition and data processing. In the experiment, we set the center frequency of ZC signal to 40 KHz with 96 KHz sampling rate, which is far beyond the hearing range of human ears. As for the cost of this system, we admit that the price will be higher than other solutions, such as CV and Bluetooth. We are studying how to complete this task with the help of loudspeakers and microphones commonly used in life to reduce costs.

3.2 Performance of Head Movement Recognition

We collect a data set of more than 7000 trajectories information to evaluate the performance of the head movement recognition module. There are about 1000 trajectories for each type of movements, each of which is a two-second coordinate sequence of two transmitters. Specifically, 80% of the data in the dataset is used to train the model while the remaining 20% is used for testing. According to the confusion matrix in Fig. 10, the average classification accuracy on the dataset is more than 99%. For a single head movement, the one with the lowest accuracy is pitch, which achieves the accuracy of 98.64%, and the one with the highest accuracy is sway, roll, yaw and static movements, which reach 100%. It can be seen that the classification accuracy of the two movements (*i.e.*, surge and pitch) are relatively low compared with others. This is also in line with our intuition, because surge is the forward and backward translation of the head, and pitch is the forward and backward rotation of the head. The two movements are very similar when the movement range is not large, so they are easy to be confused.

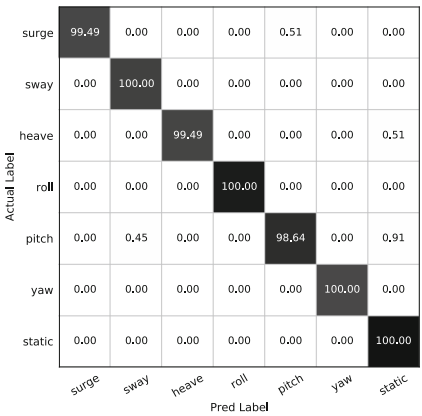


Fig. 10. Confusion matrix of head movement recognition

As for the reason why the classification result is so accurate, we believe that it can be attributed to the strong representation ability of Bi-LSTM for time series data, the sufficient training data and the simple classification task.

3.3 Performance of Head Orientation Tracking

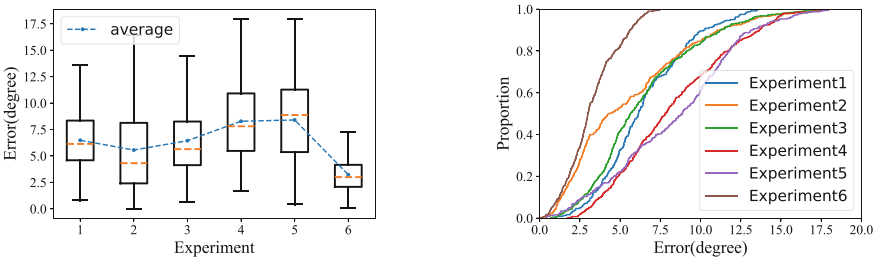


Fig. 11. Results of head orientation tracking

To get the groundtruth the head is facing, we use a nine-axis IMU (including accelerometer, gyroscope, and magnetometer). We attach this IMU to the same headphone together with the HeadTracker system. This IMU can feed back the three-axis angle changes of the head to us in real time. Based on this, we can calculate the orientation of the head as the groundtruth. We then compare the groundtruth with the HeadTracker measurements to evaluate the performance of the system. However, as we mentioned earlier, the IMU has the problem of cumulative error, which can adversely affect the experimental results. To reduce the influence of the cumulative error of the IMU, we try to shorten the duration of each experiment, which is about 20 s to 60 s. During each experiment, the participants are first told what to do and then put on the equipment with the help of the experimenter. The participants will repeat the following actions during the experiment: surge, sway, heave, roll, pitch, and yaw. During the experiment, we record the groundtruth of the IMU and the measurements of the HeadTracker in real time at ten frames per second.

We conduct a total of 6 groups of experiments. The system samples the head's orientation at a frequency 10 Hz during the volunteer's rotation. Figure 11(a) shows the error of the 6 groups. From the figure, it can be seen that the median error of each experiment is between 3° and 7°, and the maximum error is about 17.5°. According to the calculation, the average error of these 6 groups is about 6°. Figure 11(b) shows the CDF of the errors of all groups, where the 50% error of data is less than 7° and the 90% error of data is less than 12°.

3.4 Impact of Speed

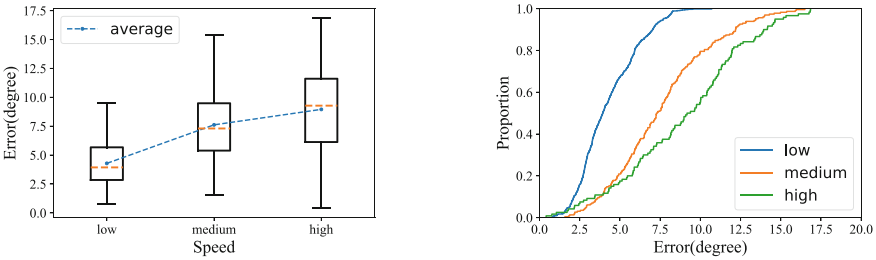


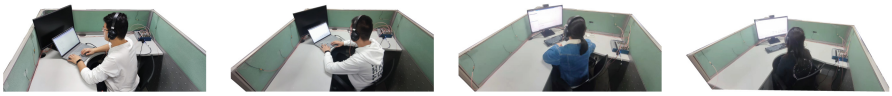
Fig. 12. Results for different rotation speeds

We also conduct experiments at different rotation speeds. First of all, according to the habit of human head rotation, we divide head's rotation speed into low speed, medium speed, and high speed. Low speed means that the rotation speed is about 1.5 degrees per second, medium speed is about 3 degrees per second, and high speed is about 9 degrees per second. We let the volunteer rotate the head at different speeds, and then evaluate the head orientation tracking performance. Figure 12 shows the experimental results at different rotation speeds.

It can be seen from Fig. 12(a) that the error of low-speed rotation is smaller than that of medium-speed rotation, and the error of medium-speed rotation is smaller than that of high-speed rotation. Regardless of the average value, median, maximum value, and other indicators for comparison, the result of low-speed rotation is almost always the best. Figure 12(b) also shows that the error of low-speed rotation is the smallest, achieving a result that the 50% error of the data is less than 5° and the 90% error of the data is less than 12°. The results are in line with our intuition, because the lower the rotation speed, the more stable the head is, the easier it is to control the head orientation.

3.5 Impact of Participants

Considering that the performance of our HeadTracker system is closely related to the user's physiological characteristics, especially the size and shape of the bones in the head and neck, different users may bring different experimental results. To evaluate the robustness of our system to users, we invite 10 participants (7 males, 3 females) to conduct the experiment (Fig. 13). These volunteers range in age from 20 to 25. We let each participant wear the equipment to conduct the same experiment and evaluate the results of these experiments. The results of all experiments are shown in Fig. 14. Because of the different physiological structure and head movement habits among people, there are some differences between the results from different participants as shown in Fig. 14(a). Among them, participant M5 has the largest average error (about 5°), while participant



(a) Male1 (b) Male2 (c) Female1 (d) Female2

Fig. 13. Different participants

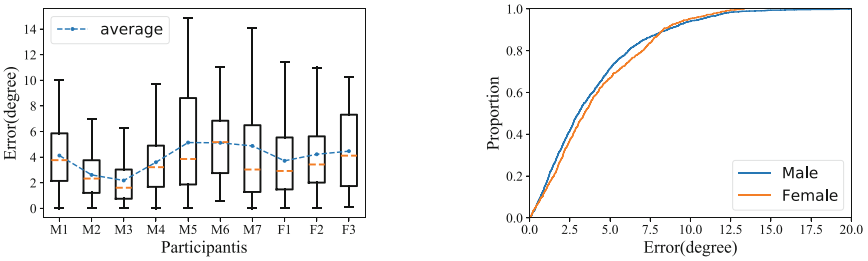


Fig. 14. Results for different participants

M3 has the smallest average error (about 2.5°). Moreover, the average error of all the participants' data is about 4°, which is basically the same as the error we measured above. These experiments show that HeadTracker is robust to different participants. In addition, we count the head orientation errors of male and female participants and draw the CDF of them in Fig. 14(b). We can see that the two curves basically overlap, which prove that the results are almost not affected by gender.

4 Conclusion

Users' head orientation provides valuable information to various fields such as online courses, online conferences, and somatosensory games. To effectively obtain and utilize this information, we propose HeadTracker in this paper, which is a fine-grained 3D head orientation tracking system based on a headphone. To achieve high-precision tracking, we first install the ultrasonic transmitters on the headphone and deploy the ultrasonic receivers in the environment to realize the positioning of the user's headphone. Then, we use the trajectory of the headphone and Bi-LSTM to complete the recognition of the user's head movement. Finally, we calculate the real-time position of the pivot point based on the head movement and then calculate the head orientation. Our experimental results show that the average error of head orientation tracking is about 6° in real environment, which is the best performance known at present and indicates that our system has great development potential and application prospects.

References

1. Cordea, M.D., Petriu, E.M., Georganos, N., Petriu, D.C., Whalen, T.E.: Real-time 2(1/2)-D head pose recovery for model-based video-coding. IEEE Trans. Instrum. Meas. **50**(4), 1007–1013 (2001)
2. Correa, A., Munoz Diaz, E., Bousdar Ahmed, D., Morell, A., Lopez Vicario, J.: Advanced pedestrian positioning system to smartphones and smartwatches. Sensors **16**(11), 1903 (2016)
3. Das, S.S.: Simple, inexpensive, accurate calibration of 9 axis inertial motion unit. In: 2019 28th IEEE International Conference on Robot and Human Interactive Communication, pp. 1–6. IEEE (2019)
4. Euston, M., Coote, P., Mahony, R., Kim, J., Hamel, T.: A complementary filter for attitude estimation of a fixed-wing UAV. In: 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 340–345. IEEE (2008)
5. Jimenez, A.R., Seco, F., Prieto, C., Guevara, J.: A comparison of pedestrian dead-reckoning algorithms using a low-cost mems IMU. In: 2009 IEEE International Symposium on Intelligent Signal Processing, pp. 37–42. IEEE (2009)
6. Kang, W., Han, Y.: SmartPDR: smartphone-based pedestrian dead reckoning for indoor localization. Sensors **15**(5), 2906–2916 (2014)
7. Kaplan, E.D., Hegarty, C.: Understanding GPS/GNSS: Principles and Applications. Artech House (2017)
8. Madgwick, S.O., Harrison, A.J., Vaidyanathan, R.: Estimation of imu and marg orientation using a gradient descent algorithm. In: 2011 IEEE International Conference on Rehabilitation Robotics, pp. 1–7. IEEE (2011)
9. Roy, N., Wang, H., Roy Choudhury, R.: I am a smartphone and i can tell my user's walking direction. In: Proceedings of the 12th Annual International Conference on Mobile Systems, Applications, and Services, pp. 329–342 (2014)
10. Tong, Y., Wang, Y., Zhu, Z., Ji, Q.: Robust facial feature tracking under varying face pose and facial expression. Pattern Recogn. **40**(11), 3195–3208 (2007)
11. Wan, H., Shi, S., Cao, W., Wang, W., Chen, G.: Resptracker: multi-user room-scale respiration tracking with commercial acoustic devices. In: Proceedings of IEEE INFOCOM Conference on Computer Communications, pp. 1–10. IEEE (2021)
12. Wang, H., Sen, S., Elgohary, A., Farid, M., Youssef, M., Choudhury, R.R.: No need to war-drive: unsupervised indoor localization. In: Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services, pp. 197–210 (2012)
13. Wang, J.G., Sung, E.: EM enhancement of 3D head pose estimated by point at infinity. Image Vis. Comput. **25**(12), 1864–1874 (2007)
14. Wang, L., et al.: Watching your phone's back: gesture recognition by sensing acoustical structure-borne propagation. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **5**(2), 1–26 (2021)
15. Yang, J.J., Banerjee, G., Gupta, V., Lam, M.S., Landay, J.A.: Soundr: head position and orientation prediction using a microphone array. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–12. Association for Computing Machinery (2020)
16. Yang, Q., Zheng, Y.: Model-based head orientation estimation for smart devices. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **5**(3), 1–24 (2021)