NEARTRACKER: ACOUSTIC 2-D TARGET TRACKING WITH NEARBY REFLECTOR IN SISO SYSTEM

Chao Liu[†]*, Linlin Gao*[†]*, Ruobing Jiang**

Ocean University of China, Qingdao, 266100, China
liuchao@ouc.edu.cn, gaolinlin@stu.ouc.edu.cn, jrb@ouc.edu.cn

ABSTRACT

Acoustic target tracking has shown significant potential for contactless human-computer interaction. However, most existing acoustic 2-D tracking approaches for portable devices require at least one speaker and two microphones, incapable for universal devices. In this paper, we propose *NearTracker*, a contactless acoustic tracking system, achieves 2-D target tracking with only one speaker and one microphone (i.e., Single Input Single Output, SISO). With the help of a nearby reflector, the additional valuable echoes from target are combined for positioning. Actually, the dynamic interferences from non-target echoes pose huge challenges for target echo extraction. *NearTracker* extracts and enhances these faint target echoes with novel signal processing methods and estimates the target's location accurately via a designed particle filter algorithm. Extensive experiments show that our system achieves on average 1.36 cm error for 2-D target tracking, which can satisfy most devices and application scenarios.

Index Terms— 2-D Target tracking, Acoustic sensing, HCI, Contactless interaction, Wireless sensing

1. INTRODUCTION

Motivation: Acoustic-based interactions allow controlling the devices, e.g., cell phones, computers, TVs, without touching the device or using a controller. However, simply relying on one-dimensional distance cannot provide sufficient gesture information for interaction. Beyond the simple one-dimensional swipe, more functional two-dimensional interactions (e.g., writing and clicking) are desired. A general option for acoustic 2-D interaction is to combine multiple speaker-microphone pairs, but significantly decreasing the usability due to non-universal hardware equipment.

Prior works and limitation: Most existing works cannot operate 2-D interaction with SISO system [1–9]. LLAP [1], Strata [3] and VSkin [4] measures the distance of movement via the phase change caused by hand/finger. However, for 2-D target tracking, their approaches still rely on the combination of multiple microphone-speaker pairs. FingerIO [2] and AMT [5] extract the echoes corresponding to finger from CIR profile and calculates the absolute distance between the finger and device based on its Time of Flight (ToF), which also requires multiple microphone-speaker pairs. Thus these solutions are poorly adapted to the devices with SISO system. Moreover, all the prior works have ignored the additional valuable information of multiple dynamic reflections from environment.

Our approach: In this paper, we propose a novel 2-D target tracking system called *NearTracker*, which requires only one speaker-microphone pair without additional hardware to track moving targets (e.g., hand/finger) in 2-D plane. Two significant challenges should be tackled. Firstly, our approach requires fewer speaker-microphone pairs to fit more devices, which also leads to limited data for 2-D localization. *NearTracker* treats the screen as a reflector, which is placed on the side of target and phone, to provide additional information for positioning. We further track the target's movement by establishing a state-space model and designing a Particle Filter (PF) algorithm. Secondly, sensing environment must contain the interference from the static objects (e.g., wall, desk) and dynamic objects (e.g., arm, body, respiration), which poses new obstacles for the extraction of faint target echoes that undergo multiple reflections. These target echoes are faint but informative. *NearTracker* extracts these echoes with Moving Target Indicator (MTI) filter and enhances them by path combination.

We have implemented *NearTracker* on Samsung Note10 Plus and Xiaomi 10 Ultra with SISO system for 2-D target tracking. The main contributions are summarized as follows.

- We propose a novel 2-D target tracking approach to track moving targets in a 2-D plane, which requires the SISO system without any additional hardware.
- *NearTracker* extracts the target echoes by eliminating dynamic interference from other moving objects with MTI and enhances them by path combination.
- We have established the state-space model and PF algorithm for 2-D tracking. Experimental results show that *NearTracker* achieves 2-D positioning accuracy of 1.36 cm, which is robust in various environments.

[†]Equal contribution, *Corresponding author

This research was supported by the National Key R&D Program No.2020YFB1707700, the National Natural Science Foundation of China Grant No.61902367, 62172377 and 41976185.

2. SIGNAL SYSTEM

2.1. Acoustic Signal Design

Due to the strong autocorrelation of ZC sequence [10, 11], we choose ZC as the baseband signal for *NearTracker*. The complex value at each position n of each root ZC sequence is given by:

$$z[n] = e^{-j\frac{\pi u n(n+c_f+2q)}{N_{zc}}}, 0 \leq n < N_{zc}, \tag{1}$$

where N_{zc} is the length of sequence, c_f is the result of N_{zc} mod 2, q and u are the parameters of ZC determining by the actual performance. In this paper, we set N_{zc} to 101.

We modulate the baseband signal to the bandwidth from 18 kHz to 24 kHz, thus it can be broadcast in an inaudible way. Considering the sampling rate of device (i.e., 48 kHz) and the audible frequency of human ear, the available bandwidth is insufficient. In this paper, we use frequency-domain interpolation to reduce the bandwidth of ZC sequence [12], the final length of ZC sequence $N'_{zc} = 384$. Furthermore, we use IQ modulation to obtain the transmittable signal via spectrum shift. Correspondingly, we also perform IQ demodulation at receiver to recover the signal [13, 14].

2.2. Target Echo Extraction

Microphone will receive multiple echoes with diverse delays τ due to environmental complexity. We utilize multipath separation to locate the echoes corresponding to hand. The receipt signal can be represented as:

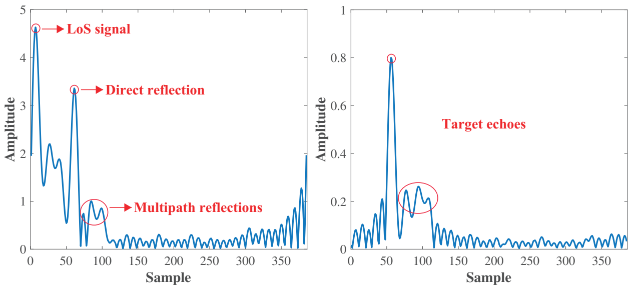
$$Z_r(t) = \sum_{i=0}^{L} A_i e^{-j\phi_i(t)} Z_t(t - \tau_i), \tag{2}$$

where $Z_t(t)$ represents transmitted baseband signal, L is the number of paths, A_i, ϕ_i and τ_i are the amplitude, phase and delay of path i, respectively. Actually, the channel between a microphone and a speaker can be considered as a linear time-variant channel with CIR, which can be denoted as $h(t)$,

$$h(t) = Z_r^*(-t) * Z_t(t) = \sum_{i=0}^{L} A_i e^{-j\phi_i(t)} \delta(t - \tau_i), \tag{3}$$

where $\delta(t)$ is the unit-impulse function. We plot the sampled version $h(n)$ in Fig.1(a), where each peak corresponds to one echo path or an overlapping path involving multiple echo paths with close delay.

We extract the target echoes by eliminating not only static echoes (i.e., LoS signal and static reflection) but also interference from other dynamic objects (e.g., arm, body, respiration). Existing sliding-window subtraction methods [15–17] cannot stably eliminate these echoes. So we choose MTI filter [18], a frequency-domain combo band-pass filter, to better eliminate these interferences, i.e.,



(a) One frame of raw CIR (b) One frame of processed CIR

Fig. 1. Multipath in a frame of $h[n]$ (a) and $D[n]$ (b).

$$D[n] = |h[n] - 3h[n + N'_{zc}] + 3h[n + 2N'_{zc}] - h[n + 3N'_{zc}]|. \tag{4}$$

As shown in Fig.1(b), MTI shows excellent elimination performance in target echo extraction. In this way, the remaining peaks in $D[n]$ indicate the reflected echoes from hand (i.e., target echoes). Besides, there are still some random jitters with tiny amplitudes, which can be filtered out by using an empirical threshold.

2.3. Path Combination

The target echoes in Fig.1(b) consist of direct reflection and multipath reflections. Specifically, direct reflection represents the emitted signal only reflected by the hand. The multipath reflections represent two reflections from the hand and other obstacles (e.g., screen), which are usually disregarded in prior works.

The direct reflection is the first to arrive at microphone with relatively higher amplitude, which normally occupies one or two peaks. We can also see some lower peaks behind the direct reflection (i.e., multipath reflections), which are caused by the hand and screen. Compared to other reflective planes (e.g.,wall), glass is actually a great reflector of sound waves. Specifically, the emitted signal is reflected by the screen first and then by the hand, or in reverse order. Different from static reflections, multipath reflections show lower amplitudes and variation due to the multiple attenuations from moving hand and stationary screen.

Actually, there are multiple reflection paths that pass the screen and reach the microphone with different propagation paths, as shown in Fig.2. We call this phenomenon acoustic dispersion, which is caused by the material and structure of the screen. Most glossy screens decrease the screen glare by covering anti-glare coating or plastic polymer, which will aggravate the acoustic dispersion.

Intuitively, we combine those multiple paths related to acoustic dispersion to uniquely represent multipath reflections. Specifically, we first choose the peak with largest

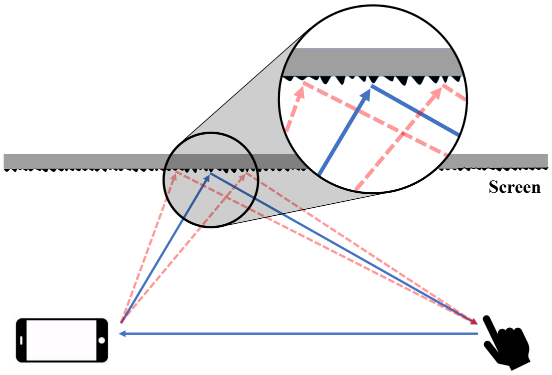


Fig. 2. Multipath effects of screen surface.

amplitude among them as central echo N_c. Then we calculate the weights w_i for each surrounding echo N_i around the central echo via cross-correlation. Finally, we combine surrounding echoes to the center echo with weighted-sum,

$$N_{mr}(n) = N_c(n) + \sum_{i=1}^{L} N_i(n)w_i. \tag{5}$$

In this way, we combine these multiple paths to uniquely represent multipath reflections with N_{mr}. We further extract the estimated arrival time of N_{mr} denoted as $\hat{t}_{mr}$ for the following 2-D target tracking.

3. MOVEMENT TRACKING

3.1. Direct Reflection Measurement

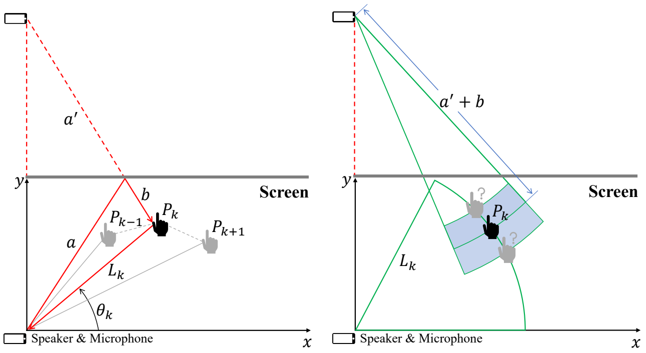
The distance between the hand and microphone can be easily obtained via the direct reflection. Firstly, we obtain the precise arrival time t_d of LoS signal by locating the maximum peak sample subscript $n_{_{LoS}}$. The ToF of LoS signal, denoted by $T_{_{LoS}}$, can be determined by the fixed distance from speaker to microphone. Thus, the initial time of signal t_s from the speaker can be represented as:

$$t_s = t_d - T_{_{LoS}}. \tag{6}$$

Similarly, we can get the arrival time t_r of direct reflection signal by detecting the maximum peak subscript n_{dr} among all the multipath echoes. Then we calculate the ToF T_{dr} along the direct reflection echo to get the path length l_{dr},

$$l_{dr} = \frac{cT_{dr}}{2} = \frac{c(t_r - t_d)}{2}, \tag{7}$$

where c is the sound velocity, l_{dr} is the distance between phone and hand. Finally, a ranging ellipse can be constructed using two focal points (i.e., speaker and microphone) and $2l_{dr}$ as the major Axis. In this way, the tracked target is located at an arbitrary point on the ellipse.



(a) The coordinate system for tracking the location of target.

(b) Narrowing the target scope.

Fig. 3. Tracking the movement of target.

3.2. 2-D Tracking

We first build a dynamic state-space model that incorporates multipath reflections from the screen. Specifically, we suppose that the tracked target's movement obeys First-order Markov Model [19], which can be shown as,

$$x_k = f_k(x_{k-1}, v_{k-1}) \tag{8}$$

$$y_k = h_k(x_k, n_k), \tag{9}$$

where x represents the target's state, y is the observation function, f, h are the state transition function and measurement function, v, n are process noise and measurement noise.

We construct a Cartesian coordinate for the detection area in Fig.3(a), where the speaker and microphone are regarded as the origin. The tracked target's position is represented as $P_k(L_k, \theta_k)$, where L_k is the 1-D distance (i.e., l_{dr}) between the phone and target obtained in section 3.1. The critical point for 2-D tracking is to determine the angle of θ, i.e.,

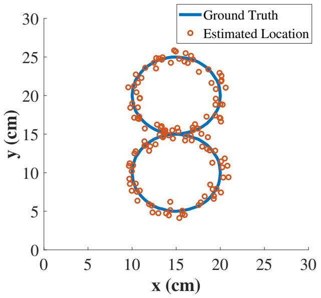
$$\tan\theta_k = \frac{L_{k-1}\sin\theta_{k-1} + v_k \sin\theta_k \Delta t}{L_{k-1}\cos\theta_{k-1} + v_k \cos\theta_k \Delta t} + v_{k-1} \tag{10}$$

$$L_k = L_{k-1} + v_k \sin\theta_k \Delta t \sin\theta_{k-1} + v_k \cos\theta_k \Delta t \cos\theta_{k-1}, \tag{11}$$

where Δt is the fixed time interval, v_k is the target's velocity indicated from the magnitude of the peak in $D[n]$. According to the Equ.9, the observation function is related to target's current state (distance, angle). Specifically, different states of the target, i.e., different positions, will result in different received signal power at the microphone. Therefore, we treat the power of received signal from target as the observation function, which can be represented as,

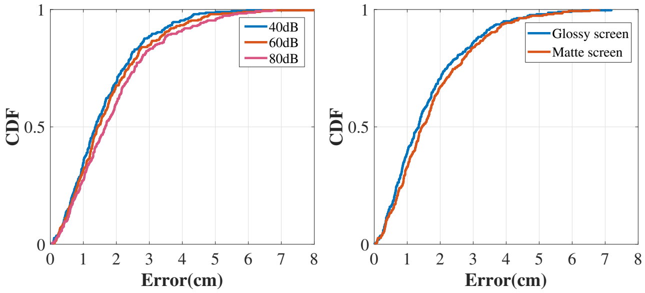
$$y_k = AP_{dr} + AP_{mr} + n_k, \tag{12}$$

where A is the amplitude of the reflected echo from the target, P_{dr} denotes the effect that only the target's reflection exerts on the amplitude, P_{mr} is the effect that target's reflection and



(a) The pattern ‘8’ (b) The pattern ‘star’

Fig. 4. The ground truth trace (blue lines) and *NearTracker*’s estimated location (orange circles) for two patterns.



(a) Different noise levels (b) Different screen types

Fig. 5. CDF of the 2-D positioning error with different noise levels (a) and different screen types (b).

screen’s reflection on the amplitude. Thus we can estimate the target’s state via state equation and observation equation.

For tracking the target’s position $P_k(L_k, \theta_k)$ at time k, we first significantly narrow the target scope using the arrival time $\hat{t}_{mr}$ mentioned in section 2.3. Specifically, as shown in Fig.3(b), we treat the screen as a symmetry axis to get the phone’s projection. Thus the multipath reflection $a + b + L_k$ is replaced by $a' + b + L_k$. Two positioning circles are constructed with direct reflection L_k and multipath reflection $a' + b$, the joint part is the target scope. Note that the joint part is a scope instead of an intersection, because the arrival time $\hat{t}_{mr}$ is an estimated version, which leads to the length of multipath reflection being in a range. We further narrow the scope in the following until obtaining the position of target.

Based on the state-space model, we found PF [20–22] has excellent performance in target tracking. Specifically, we first generate N particles and distribute them on the target scope via the Gaussian distribution, each particle being given the same weight $1/N$. Then we sample N' particles to obtain the paticles set $\{\widetilde{x}_k^{(i)}\}_{i=1}^{N'}$ and calculate the weight of them $\widetilde{w}_k^{(i)}$. To be specific, we calculate the amplitude of target echo at each particle’s position as the observation function y and combine it with state function x to calculate the weights and normalize them. Furthermore, we resample the particles set $\{\widetilde{x}_k^{(i)}, \widetilde{w}_k^{(i)}\}$ according to the calculated weights, placing more particles around the particles with high weights $\widetilde{w}_k^{(i)}$, which further narrows the location range of target. Finally, the weighted sum of state functions at moment k is the obtained position of target.

4. PERFORMANCE EVALUATION

We run the experiments on Samsung Note10 Plus and Xiaomi 10 Ultra with only one speaker and one microphone open. Ten volunteers (5 males and 5 females) between the ages of 25-40 are recruited to perform 2-D gestures within the sensible area. Specifically, we first place the screen on the side of hand and phone. Then volunteers are asked to draw two pre-defined patterns with their finger, each pattern is repeated 10 times. We finally evaluate the average 2-D tracking accuracy by comparing ground truth with our collected data.

The general 2-D tracking performance of *NearTracker* is shown in Fig.4. Blue lines (ground truth) are the pre-drawn patterns on paper and orange circles are the estimated location computed by *NearTracker*. Experimental results show that the estimated position is close to the ground truth trace, even for the complex pattern ‘star’. We further evaluate the 2-D tracking error under different conditions.

Different noise level: We evaluate the 2-D tracking performance under these three different noise levels: (1) quiet level with 40 dB noise pressure; (2) medium level with 60 dB noise pressure; (3) noisy level with 80 dB noise pressure. We maintain a constant noise level by playing music. As shown in Fig.5(a), the average positioning error is 1.37 cm, 1.43 cm, 1.67 cm respectively under the different noise level of 40 dB, 60 dB, 80 dB. The results demonstrate *NearTracker*’s strong robustness to noise.

Different screen type: The mainstream screen types are matte screen and glossy screen. Fig.5(b) shows the CDF of 2-D tracking error with these two screen types. The average error of glossy screen across all participants is around 1.38 cm, 4.82% lower than the 1.45 cm achieved by matte screen. The result proves that our approach can reduce the interference caused by acoustic dispersion on different screen types.

5. CONCLUSION

In this paper, we propose an acoustic-based 2-D tracking system with only one speaker and one microphone, which cooperates with a nearby screen for positioning. We address the significant challenge of target echo extraction with MTI Filter and path combination. *NearTracker* also uses the nearby reflector and PF filter to tackle the challenge of insufficient data. Extensive experiment results exhibit superior performance of our approach in 2-D target tracking, which has significant potential for diverse devices and application scenarios.

6. REFERENCES

[1] Wei Wang, Alex X. Liu, and Ke Sun, "Device-free gesture tracking using acoustic signals," in *MobiCom*. 2016, pp. 82–94, ACM.

[2] Rajalakshmi Nandakumar, Vikram Iyer, Desney S. Tan, and Shyamnath Gollakota, "Fingerio: Using active sonar for fine-grained finger tracking," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing*. 2016, pp. 1515–1525, ACM.

[3] Sangki Yun, Yi-Chao Chen, Huihuang Zheng, Lili Qiu, and Wenguang Mao, "Strata: Fine-grained acoustic-based device-free tracking," in *MobiSys*. 2017, pp. 15–28, ACM.

[4] Ke Sun, Ting Zhao, Wei Wang, and Lei Xie, "Vskin: Sensing touch gestures on surfaces of mobile devices using acoustic signals," in *MobiCom*. 2018, pp. 591–605, ACM.

[5] Liu Chao, Penghao Wang, Ruobing Jiang, and Yanmin Zhu, "Amt: Acoustic multi-target tracking with smartphone mimo system," in *INFOCOM*, 2021.

[6] Xiangyu Xu, Jiadi Yu, Yingying Chen, Yanmin Zhu, and Minglu Li, "Leveraging acoustic signals for vehicle steering tracking with smartphones," *IEEE Trans. Mob. Comput.*, vol. 19, no. 4, pp. 865–879, 2020.

[7] Peng Cheng, Ibrahim Ethem Bagci, Utz Roedig, and Jeff Yan, "Sonarsnoop: Active acoustic side-channel attacks," *CoRR*, vol. abs/1808.10250, 2018.

[8] Sangki Yun, Yi-Chao Chen, and Lili Qiu, "Turning a mobile device into a mouse in the air," in *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services, MobiSys*. 2015, pp. 15–29, ACM.

[9] Li Lu, Jian Liu, Jiadi Yu, Yingying Chen, Yanmin Zhu, Xiangyu Xu, and Minglu Li, "Vpad: Virtual writing tablet for laptops leveraging acoustic signals," in *2018 IEEE 24th International Conference on Parallel and Distributed Systems (ICPADS)*, 2018, pp. 244–251.

[10] B.M. Popovic, "Generalized chirp-like polyphase sequences with optimum correlation properties," *IEEE Transactions on Information Theory*, vol. 38, no. 4, pp. 1406–1409, 1992.

[11] Jun Tao, Le Yang, and Xiao Han, "Enhanced carrier frequency offset estimation based on zadoffchu sequences," *IEEE Communications Letters*, vol. 23, no. 10, pp. 1862–1865, 2019.

[12] Juanjuan Chen and Zhanchuan Cai, "Cardinal mk-spline signal processing: Spatial interpolation and frequency domain filtering," *Inf. Sci.*, vol. 495, pp. 116–135, 2019.

[13] J. Tuthill and A. Cantoni, "Optimum precompensation filters for iq modulation systems," *IEEE Transactions on Communications*, vol. 47, no. 10, pp. 1466–1468, 1999.

[14] Ivo Bizon Franco de Almeida, Marwa Chafii, Ahmad Nimr, and Gerhard Fettweis, "In-phase and quadrature chirp spread spectrum for iot communications," in *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, 2020, pp. 1–6.

[15] Haiming Cheng and Wei Lou, "Push the limit of device-free acoustic sensing on commercial mobile devices," in *40th IEEE Conference on Computer Communications, INFOCOM 2021, Vancouver, BC, Canada, May 10-13, 2021*. 2021, pp. 1–10, IEEE.

[16] Yanwen Wang, Jiaxing Shen, and Yuanqing Zheng, "Push the limit of acoustic gesture recognition," in *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*, 2020, pp. 566–575.

[17] Li Lu, Jiadi Yu, Yingying Chen, Yanmin Zhu, Xiangyu Xu, Guangtao Xue, and Minglu Li, "Keylistener: Inferring keystrokes on qwerty keyboard of touch screen through acoustic signals," in *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*, 2019, pp. 775–783.

[18] A. Zverev, "Digital mti radar filters," *IEEE Transactions on Audio and Electroacoustics*, vol. 16, no. 3, pp. 422–432, 1968.

[19] Aravind Kailas, Chia-Chin Chong, and Fujio Watanabe, "A first-order markov model for wellness mobile applications," in *2010 IEEE 71st Vehicular Technology Conference*, 2010, pp. 1–6.

[20] M.S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 174–188, 2002.

[21] Chenshu Wu, Feng Zhang, Beibei Wang, and K. J. Ray Liu, "Easitrack: Decimeter-level indoor tracking with graph-based particle filtering," *IEEE Internet of Things Journal*, vol. 7, no. 3, pp. 2397–2411, 2020.

[22] Yuan Jing, Zhifeng Li, and Chang Liu, "Acoustic source tracking based on adaptive distributed particle filter in distributed microphone networks," *Signal Processing*, vol. 154, pp. 375–386, 2019.