WSTrack: A Wi-Fi and Sound Fusion System for Device-free Human Tracking

Yichen Tian *†, Yunliang Wang *†, Ruikai Zheng *†, Xiulong Liu *†, Xinyu Tong*†, Keqiu Li *†
*College of Intelligence and Computing, Tianjin University, China
†Tianjin Key Laboratory of Advanced Networking (TANK)
Email:{tianyichentyc, wangyunliang_1016, zzzrk, xiulong liu, xytong, keqiu}@tju.edu.cn

***Abstract*—The voice assistants benefit from the ability to localize users, especially, we can analyze the user's habits from the historical trajectory to provide better services. However, current voice localization method requires the user to actively issue voice commands, which makes voice assistants unable to track silent users most of the time. This paper presents *WSTrack*, a Wi-Fi and Sound fusion tracking system for device-free human. In particular, current voice assistants naturally support both Wi-Fi and acoustic functions. Accordingly, we are able to build up the multi-modal prototype with just voice assistants and Wi-Fi routers. To track the movement of silent users, our insights are as follows: (1) the voice assistants can hear the sound of the user's pace, and then extract which direction the user is in; (2) we can extract the user's velocity from the Wi-Fi signal. By fusing multi-modal information, we are able to track users with a single voice assistant and Wi-Fi router. Our implementation and evaluation on commodity devices demonstrate that *WSTrack* achieves better performance than current systems, where the median tracking error is $0.37m$.**

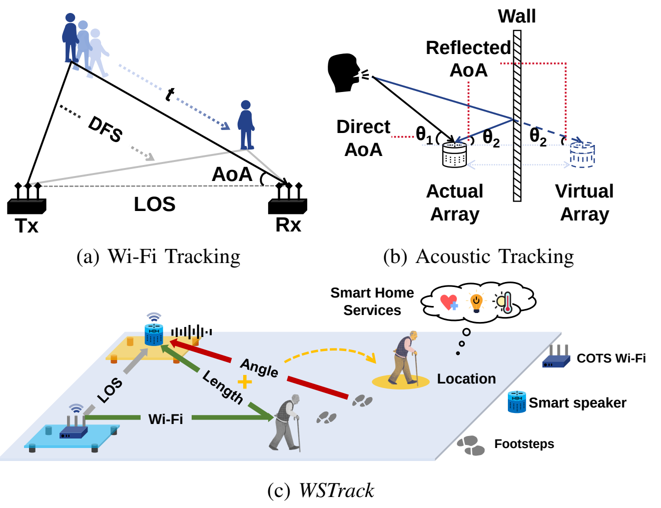
***Index Terms*—Multi-modal Information, Device-free Tracking, Wi-Fi, Acoustic**

I. Introduction

A. *Background and Motivation*

Currently, voice assistants become more and more popular in our daily life. For example, smart speakers including Amazon Alexa [1] and Google Home [2], [3] support various applications such as appliances control and human-machine conversation. The voice assistants benefit from the ability to localize users, especially, we can analyze the user's habits from the historical trajectory to provide better services. To this end, the arts [4]–[6] propose to utilize the microphone to localize the user's voice. However, these methods require the user to actively issue voice commands, and therefore cannot continuously track the silent user.

To realize ubiquitous localization services, we are inspired by the integration of Wi-Fi and acoustic tracking technologies. Specifically, voice assistants support both Wi-Fi communication and sound collection functions, hence, it is feasible to build up the WiFi-sound tracking prototype with the smart speaker and the Wi-Fi router. It is promising to deploy such a prototype in some important applications including smart home [7]–[11] and elderly care [12], [13], due to the following reasons: (1) **Convenient.** This device-free tracking technology does not require the user to carry any device; (2) **Ubiquitous.** The smart speaker and the Wi-Fi router have been widely deployed in our daily life; (3) **Accurate.** Tracking performance benefits from multi-modal wireless features. Despite these advantages, we need to address a key issue: *There is usually only one smart speaker and Wi-Fi router at home, how to realize the goal with limited devices?*



(a) Wi-Fi Tracking (b) Acoustic Tracking

(c) *WSTrack*

Fig. 1: The schematic diagram of *WSTrack*.

B. *Limitations of Prior Art*

The state-of-the-art research in the field can be divided into Wi-Fi and acoustic tracking.

The basic principle of Wi-Fi tracking is to extract location-related features including angle, distance and movement velocity from the raw Wi-Fi signals as shown in Fig. 1a. Recent studies [14]–[17] realize device-free human tracking with a single Wi-Fi link. However, they suffer from one of the following limitations: (1) *Time-consuming Manual Calibration.* Only when we manually calibrate the phase offset across received antennas, these systems [14], [15], [17] provide satisfactory accuracy. This calibration requires us to utilize the power splitter/combiner [18] and cable to physically connect the transceiver antennas, but the power splitter/combiner is more expensive than the Wi-Fi router; (2) *Massive Antennas.* The art [16] requires massive antennas, *e.g.*, 6 antennas, to maintain the tracking accuracy. However, there are usually only 2 $\sim$ 3 antennas for commercial Wi-Fi devices.

The basic principle of acoustic tracking is illustrated in Fig. 1b, where we first extract the angle of arrival (AoA) of the acoustic source, and then leverage wall reflections to localize targets. Such methods require users to actively issue

Corresponding author: Xinyu Tong.

voice commands. Moreover, there should be a suitable distance between the wall and the voice assistant, so that the signal reflected off the wall can be correctly separated.

C. *WSTrack in a Nutshell*

The Wi-Fi or acoustic tracking system suffers from limitations, which motivates us to develop a multi-modal tracking system, *i.e.*, *WSTrack*. As illustrated in Fig. 1c, as current voice assistants support both Wi-Fi communication and sound collection, we can regard the voice assistant as a receiver for both Wi-Fi and sound signals. When the user moves, the voice assistant can collect the Wi-Fi signal reflected off the human body, as well as the sound of human pace. Thereafter, we extract location-related features from multi-modal information and finally realize silent user tracking.

D. *Challenges and Solutions*

It is not trivial to design such a system, and we need to address the following challenges:

The first challenge is how to extract location-related features from the sound of human pace. In contrast to human speech, the human pace is weaker and only lasts for a very short time. Meanwhile, between the user's adjacent paces, there is usually a silent state that interferes with AoA measurements. To address this issue, we design the pace detection, matrix-based AoA estimation, and AoA sequence recovery methods. The pace detection can recognize the human pace, remove irrelevant sounds and thus reduce the execution time. The matrix-based AoA estimation fuses the sample shifts of multiple microphones to improve the accuracy. The AoA sequence recovery algorithm is responsible for inferring missing AoAs caused by the silent state between adjacent paces.

The second challenge is how to integrate the features of Wi-Fi and acoustic signals. Specifically, we are able to extract the change rate of the reflection path from the Wi-Fi signal, while being able to extract the AoA from the acoustic signal. In order to integrate these information, IndoTrack [19] proposes a probabilistic model, but this probability relies on subjective judgment. In contrast, we establish a physical model of the proposed problem. First, based on the Fresnel Zone, the change rate of the reflection path is modeled as a series of ellipses, while the AoA is modeled as the ray. The user can then be tracked by solving the intersection location of the ellipse and the ray.

The third challenge is how to determine the initial location of the user. In order to determine the initial location, previous work [19], [20] usually iterates over all possible positions and finally determines the desired one. Such process is time-consuming, for example, it takes several minutes to find the initial location for the 10-second trajectory [20]. To reduce the execution time, our iteration is to determine the optimal Fresnel zone. Compared to thousands of iterations in previous work, we only need hundreds of iterations to cover the same area with the same granularity. For each candidate Fresnel ellipse, based on acoustic pace detection, we can obtain the corresponding gait features between adjacent paces. Since users tend to maintain a steady gait, we are able to select the optimal Fresnel ellipse. Finally, it is feasible to calculate initial location based on the Fresnel ellipse and the AoA.

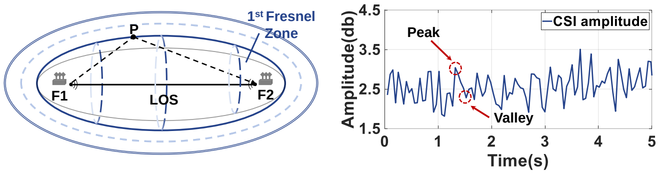


Fig. 2: The schematic diagram of Fresnel zone.

E. *Contributions and Advantages over Prior arts*

This paper for the first time integrates Wi-Fi and acoustic signals to realize device-free human tracking. The technical novelty of *WSTrack* lies in tackling three challenges: parsing location-related features from footsteps; integrating the features of Wi-Fi and acoustic signals; inferring initial location of the user. We implement the prototype with commercial off-the-shelf (COTS) Wi-Fi and acoustic devices. The experimental results demonstrate that the median tracking error of *WSTrack* is less than $0.37m$, which is less than most of the existing tracking methods with single-modal information.

The *WSTrack* system has three main advantages over previous works: (1) Compared with Wi-Fi tracking systems [14]–[17], we remove the extra cost for phase calibration and improve tracking accuracy; (2) Compared with acoustic tracking systems [4]–[6], we can track the users even if they do not actively issue voice commands; (3) In terms of multi-modal fusion, such devices indeed exist in reality that naturally have Wi-Fi and acoustic communication capabilities. *WSTrack* makes voice assistants more powerful.

The remainder of this paper is organized as follows. We present the preliminary knowledge in §Section. II, and system overview in §Section. III. We introduce the proposed system in detail in §Section. IV. §Section. V presents the implementation and the experimental results. We discuss the related work in §Section. VI and conclude the paper in §Section. VII.

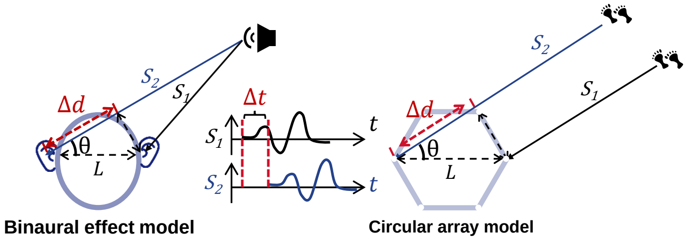
II. Preliminary

This section presents a relevant preliminary for this paper, centered around the Wi-Fi information processing and acoustic geometric property. The preliminary will lay the foundation for multi-modal fusion and tracking.

A. *Wi-Fi CSI and Fresnel Zone*

Current Wi-Fi device-free tracking systems typically extract location features from raw channel state information (CSI) readings. CSI reflects the influence of the wireless channel on the transmitting signal's properties including amplitude and phase. As shown in Fig. 2, the user movement will change the propagation distance of the signal reflected off the human body, which results in a change in the features of the received signals. Theoretically, we have

$$H(f,t) = H_s(f,t) + H_d(f,t) = H_s(f,t) + A(f,t)e^{-j2\pi \frac{d(t)}{\lambda}}, \tag{1}$$



Binaural effect model

Circular array model

Fig. 3: The binaural model and acoustic AoA estimation.

where $H(f,t)$ denotes the CSI, $H_s(f,t)$ and $H_d(f,t)$ respectively represent the static and dynamic components, $A(f,t)$, $2\pi\frac{d(t)}{\lambda}$ are the amplitude and phase of dynamic signal, and $d(t)$ is path length of dynamic signal. The user's movement will cause a change in the dynamic path length $L_d(t)$, so that the dynamic signal and the static signal are superimposed with different phases. When the dynamic signal and static signal are in-phase, we are able to observe stronger energy. In turn, when the dynamic signal and static signal are anti-phase, we are able to observe weaker energy.

According to the proportional relationship between the phase and the propagation distance, the above-mentioned phase superposition method can be reflected in the physical space as the Fresnel zone [21]. We know that it takes half a wavelength to go from in-phase to anti-phase, hence, we divide the space by every half wavelength as follows:

$$|PF_1| + |PF_2| - |F_1F_2| = n\lambda/2, \tag{2}$$

where P represents the location of the user, F_1 and F_2 are the locations of the transmitter and the receiver respectively. The reflected propagation distance is $|PF_1|+|PF_2|$, while the direct propagation distance is $|F_1F_2|$ in Fig. 2.

Obviously, for each given n, P satisfies the elliptic equation. When extended to all n, Fresnel zones refer to the concentric ellipses with foci in a pair of transceivers [22]. Based on the Fresnel zones, We have the following observations: First, when the user walks along a given ellipse, in the ideal situation, the reflected path length does not change, and the received signal does not change; second, when the user walks across one Fresnel ellipse, we will observe one peak (valley) in the amplitude of the received signal as shown in the right part of Fig. 2. To summarize, by counting the peaks and valleys of the CSI readings, we can determine the number of Fresnel zones that the user crosses.

B. AoA of Acoustic Signals

Besides the path length, the AoA is also an important feature to depict user's location. In particular, the AoA refers to which direction the acoustic signal arrives at the received devices or ears. An intuitive example known as the binaural effect is shown in Fig. 3. Specifically, the acoustic signal reaching both ears experiences different propagation times, which helps the human localize the acoustic source. In practice, the microphone arrays are the common devices to collect and capture ambient acoustic signals. The principle of AoA measurement in the microphone arrays is similar to human ears. In particular, as the inter-distance between the microphones is only a few centimeters, which is small relative to the sensing area, so the propagation paths of the sound signal can be regarded as a series of parallel lines [6]. Assuming that the signal arrives at the microphone array with the AoA denoted by θ, we can observe an extra flight time denoted by

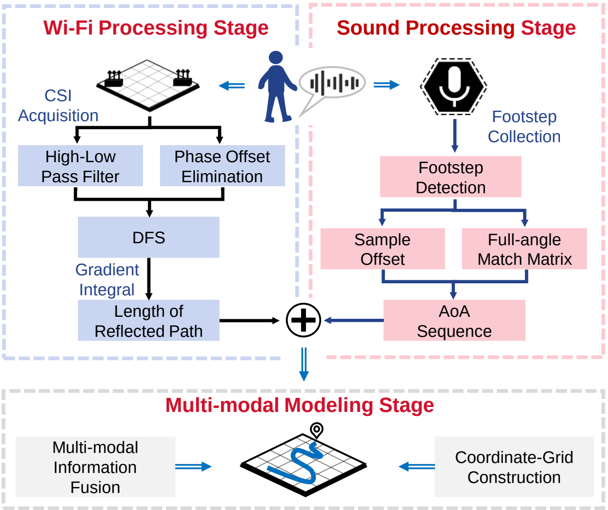


Fig. 4: System overview.

$$\Delta t = \frac{Lcos\theta}{c_{sound}}, \tag{3}$$

where L denotes the length between a microphone pair, and c_{sound} denotes the speed of the acoustic signals.

Since microphones characterize the speech signal in a discrete manner, the time shift described above appears as a discrete sample shift. For acoustic sampling rate f_s, the time interval between adjacent samples is $1/f_s$. Hence, the sample shift n_s corresponding to the above time shift should be

$$n_s = round(\Delta t f_s) = round(\frac{Lf_s cos\theta}{c_{sound}}), \tag{4}$$

where $round(\cdot)$ represents the rounding operation.

Theoretically, AoA measurement is to solve θ from the observed sample shift n between different microphone pairs. In this paper, our insight is when the person is walking, the microphone will capture the sound of the footsteps, which helps us determine the AoA of the user.

In a nutshell, our insight is to fuse the multi-modal mobility information from the Wi-Fi and acoustic signals. Mathematically, we need to solve the intersection location of the Fresnel ellipse and the ray corresponding to AoA. Next, we will present our system overview and designs.

III. System Overview

As shown in Fig. 4, *WSTrack* consists of Wi-Fi processing, sound processing, and multi-modal modeling stages.

Wi-Fi Processing Stage. The purpose is to build up the ellipse model based on the Wi-Fi signals. This stage first analyzes the raw CSI readings, and then extracts location-related features including the path length change rate (PLCR) and timestamp. Finally, we convert the PLCR and timestamp

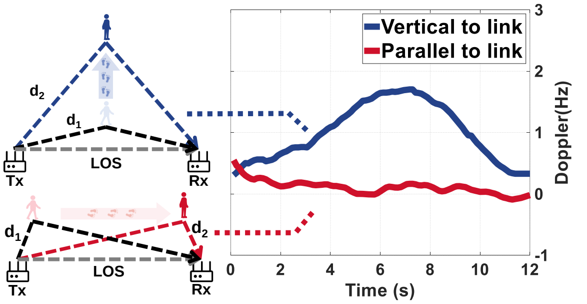


Fig. 5: The DFS of different trajectories.

into reflected path length, which depicts how many Fresnel ellipses the user has passed through.

Sound Processing Stage. The aim is to estimate the AoA with respect to the human pace, and set up the ray model. First, in order to remove silent acoustic signals between adjacent paces, we design the pace detection method. Then, we exploit the sample offsets between all pairs of microphones to realize matrix-based estimation, which can improve the accuracy of AoA. Finally, we recover the AoA for those silent acoustic signals and obtain the AoA sequence for tracking.

Multi-modal Modeling Stage. We realize multi-modal fusion through modeling the ellipse and the ray. The crux of this stage is to determine the initial location of the user. To this end, the previous work [20] traverses all tracking areas and performs tracking for each candidate initial position. To reduce the execution time, we model the tracking process as the intersection positions of a set of rays with a set of ellipses. Moreover, calculating initial position becomes the search problem of the initial Fresnel ellipse. Since users tend to maintain a steady gait, we are able to select the initial Fresnel ellipse via the pace analysis.

IV. System Design

A. *Wi-Fi Processing Stage*

This section discusses how to determine the reflected path length $L_d(t)$ based on the raw CSI readings. As discussed in the §Section. II-A, when the target moves relative to the Wi-Fi link, the propagation path of the signal reflected off the target will change, which results in the Doppler Frequency Shift (DFS) and the Path Length Change Rate (PLCR). Specifically, when the object moves away from the link vertically (*e.g.*, from d_1 to d_2 in the top left part of Fig. 5), the DFS shows a bulge (labeled by the blue line in Fig. 5). In contrast, when the target moves parallel to the WiFi link, the signal path only presents imperceptible changes (labeled by the red line in Fig. 5). To depict the above tracking process, the path length change rate (PLCR) has been proposed in Widar [20] and Indotrack [19]. In particular, the PLCR can be defined as

$$\mathbb{R}(t) = L'_d(t), \tag{5}$$

where $L_d(t)$ is the reflected path length at time t, and $(\cdot)'$ denotes the derivation operation. Since the DFS and PLCR respectively describe the change of frequency and path length, the relationship between them can be written as follows according to the Doppler equation [20]

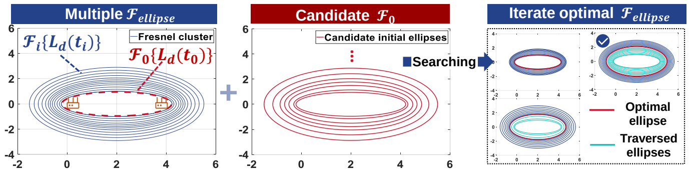

Multiple $\mathcal{F}_{ellipse}$
$\mathcal{F}_i\{L_d(t_i)\}$
Fresnel cluster
$\mathcal{F}_0\{L_d(t_0)\}$
Candidate $\mathcal{F}_0$
Candidate initial ellipses
Searching
Iterate optimal $\mathcal{F}_{ellipse}$
Optimal ellipse
Traversed ellipses
6
4
2
0
-2
-4
-2
0
2
4
6

Fig. 6: The Fresnel ellipse cluster model.

$$f_D(t) = -\frac{\mathbb{R}(t) f}{c_{light}} = -\frac{\mathbb{R}(t)}{\lambda}, \tag{6}$$

where f denotes the communication frequency, and c_{light} denotes the speed of the light. According to Eq. 5 and Eq. 6, we can obtain $L_d(t)$ based on the DFS f_D as follows

$$\begin{cases} L_d(t_0) = \left\|\vec{l}_u(t_0) - \vec{l}_t\right\| + \left\|\vec{l}_u(t_0) - \vec{l}_r\right\|, \\ L_d(t) = L_d(t_0) + \int_{t_0}^{t} -\lambda f_D(t)dt, \end{cases} \tag{7}$$

where $L_d(t_0)$ denotes the reflected path length at time t_0, $\vec{l}_u(t_0)$ denotes user's location at time t_0, $\vec{l}_t$ and $\vec{l}_t$ respectively denote the locations of the transmitter and the receiver. According to Eq. 7, the crux of $L_d(t)$ calculation is to determine $L_d(t_0)$ and $f_D(t)$.

First, we can extract DFS from the raw CSI readings via the conjugate multiplication and STFT. Since the method to determine DFS has been well studied in previous work [14], [20], we do not discuss the details here.

Second, for any $L_d(t_0)$, according to the Fresnel zone in §Section. II-A, each $L_d(t)$ corresponds to one Fresnel ellipse. Hence, the path length sequence $[L_d(t_1), L_d(t_2), ...]$ corresponds to a series of ellipses, *i.e.*, our proposed ellipse cluster. Supposing that the ellipse cluster is defined as the set $\mathcal{F}_{ellipse}$, we have

$$\mathcal{F}_{ellipse} = \{L_d(t) | L_d(t_0)\}. \tag{8}$$

When $L_d(t_0)$ is unknown, $\mathcal{F}_{ellipse}$ refers to multiple candidate ellipse clusters as illustrated in Fig. 6. Once the $L_d(t_0)$ is given, we can uniquely determine the ellipse cluster $\mathcal{F}_{ellipse}$ in Fig. 6. Until now, we have established the ellipse cluster model. Further, we will introduce the ray model, followed by how to fuse the ellipse and ray models. Finally, we discuss how to determine the initial path length $L_d(t_0)$.

B. *Sound Processing Stage*

This section will focus on calculating the sample shift and estimating AoA via footsteps.

*1) **Footstep Detection**:* According to previous studies [4]–[6], the accuracy of acoustic localization is dependent on the type and quality of the sound source. Currently, VoLoc [5] and Symphony [6] utilize several kinds of daily objects' sounds like human speech, electronic rings, etc. These systems realize decimeter-level accuracy, because their acoustic signals have strong sustainability and Signal to Noise Ratio (SNR). However, sound localization using footsteps faces the following issues: 1) *Higher dynamics.* The location of household

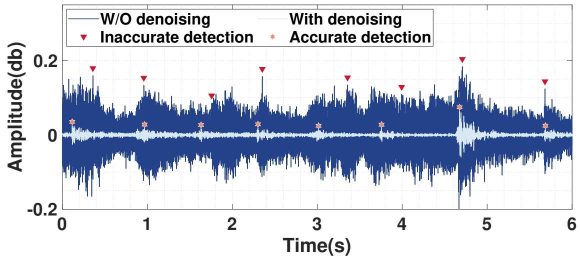


Fig. 7: The footstep detection schematic diagram.

appliances and the human voice is often static in previous work [4]–[6]. However, the sound of footsteps usually accompanies the user's movement, which causes a dynamic sound source; 2) *Shorter duration.* A sentence of speech usually lasts for several seconds, which provides abundant information for localization. In contrast, the sound of footsteps has a shorter duration and is usually only effective for 0.1 seconds. To summarize, we need to track dynamic targets with shorter effective sounds, which is thus challenging.

In order to address the above issues, the method should satisfy the following requirements: 1) We should segment the audio sequence by step recognition, because footsteps correspond to different locations; 2) we should identify valid data segments because valid data segment corresponds to user's footstep instead of environmental noise [23]. Accordingly, an intuitive solution is to apply the sliding windows to segment the overall audio sequence into small pieces, and calculate AoA for each piece. In this case, every signal piece could be considered as static in a short window. Nevertheless, this traverse method causes a huge execution time in practice. Our experimental results in Fig. 10a indicate that it takes more than 1.1 seconds to deal with the 1-second audio sequence.

To save the execution time, our insight is to identify the footstep, and only estimate AoA for those valid acoustic data. First, we set the length of the sliding window according to the sampling rate of the microphone and gait duration. We then leverage the threshold to filter the raw footstep sequence. Finally, we determine the exact time of K footsteps.

Considering that talking, music and other common types of household noise have different frequencies with footstep, we use a band-pass filter to remove the background noise. Fig. 7 shows the audio signals with and without denoising and our footstep detection results (the background noise is loud enough to cover footsteps), which demonstrates the possibility of working in noisy environments.

*2) **AoA Estimation**:* The basic principle of AoA estimation is that the sound from the target arrives at different microphones with different sample shifts in Eq. 4. Consequently, we should first determine the sample shift. Generalized cross-correlation is a well-known method to localize the sound source based on time-delay estimation (TDE). We assume that there are M microphones in the voice assistant. For two microphones denoted by m_i and m_j, there exists a sample offset between them. Based on the cross-correlation method, we have

$$\mathcal{C}[n_s] = \sum (m_i[n] \cdot m_j[n+n_s]), \tag{9}$$

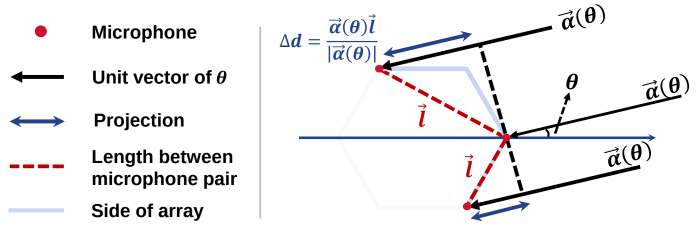


Fig. 8: Projection principle of AoA.

where n_s is the sample shifts of m_i and m_j. The sequence $\mathcal{C}[n_s]$ has the correlation peak, where the sample shift $\hat{n}_s = \arg\max_{n_s} \mathcal{C}[n_s]$ is the measured sample shift.

In practice, the calculated sample shift is not the same as the actual situation. As discussed in Eq. 4, we can only get an integer sample shift via the rounding operation. For example, if the actual sample shift is 3.4, estimated $\hat{n}_s$ is 3. If the actual sample shift is 3.6, estimated $\hat{n}_s$ is 4. A natural question is how to achieve accurate AoA measurements with approximate sample shifts.

We find the cause of estimation error is that previous work just estimates the sample shift with respect to one certain reference microphone. To address these issues, we design a matrix-based representation method, and the insight is to estimate the sample shift between all pairs of microphones. Supposing that there are three microphones, the corresponding sample shifts should satisfy: $n_s(1,2) + n_s(2,3) \equiv n_s(1,3)$. For example, the actual sample shifts satisfy: $n_s(1,2) = 3.4$, $n_s(2,3) = 4.3$, $n_s(1,3) = 7.7$, where $3.4 + 4.3 = 7.7$. Under this situation, the estimated sample shifts satisfy: $\hat{n}_s(1,2) = 3$, $\hat{n}_s(2,3) = 4$, $\hat{n}_s(1,3) = 8$. However, after the rounding operation, we find that $\hat{n}_s(1,2) + \hat{n}_s(2,3) \neq \hat{n}_s(1,3)$. Compared to only calculating $\hat{n}_s(1,2)$ and $\hat{n}_s(1,3)$ in previous work, this extra $\hat{n}_s(2,3)$ can help us further determine the range of sample shifts. That is, $3 \leq \hat{n}_s(1,2) < 3.5$, $4 \leq \hat{n}_s(2,3) < 4.5$, and $7.5 \leq \hat{n}_s(1,3) \leq 8$. This accounts for why the proposed method can realize accurate AoA estimation.

In our system, for the k^{th} footstep, the sample shift between m_i and m_j is defined as

$$O_{act}^{(k)}(m_i,m_j) = \arg\max_{n_s} \mathcal{C}[n_s]. \tag{10}$$

Obviously, $O_{act}^{(k)}(m_i,m_j)$ is a $M \times M$ matrix for a certain k. where M is the number of microphones.

*3) **Sample Shift Calculation**:* In order to calculate the AoA, we first establish a reference matrix to depict the theoretical sample shifts with respect to different AoAs. Then, we determine which candidate situation is closest to the current measurement. The workflow is as follows: As illustrated in §Section. II, the sample shift is caused by the extra flight distance of the sound.

It can be expressed as the projection of $\vec{l}_{m_i,m_j}$ on the unit vector $\vec{\alpha}(\theta)$, which is corresponding to the AoA θ [24]. For a given AoA, the corresponding sample shift should be

$$O_{sim}^{(\theta)}(m_i,m_j) = \frac{\Delta d}{c_{sound}} f_s = \frac{\vec{\alpha}(\theta)\vec{l}_{m_i,m_j}}{|\vec{\alpha}(\theta)| c_{sound}} f_s, \tag{11}$$

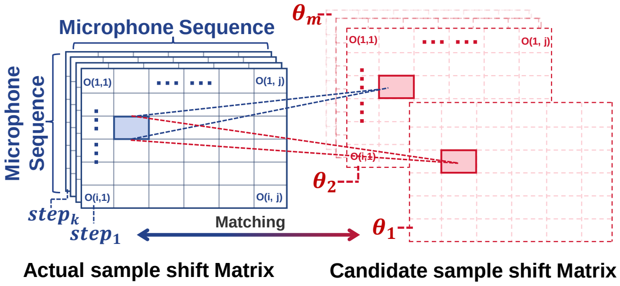


Fig. 9: The AoA estimation via matrix matching.

where $\frac{\vec{\alpha}(\theta)\vec{l}_{m_i,m_j}}{|\vec{\alpha}(\theta)|}$ denotes the projection distance. For k^{th} footstep, we can obtain AoA through comparing the actual sample shift $O_{act}^{(k)}(m_i,m_j)$ with candidate sample shift of different AoAs, and find the optimal one. It is equivalent to solving the following optimization equation:

$$\hat{\theta}_k = \arg\min_{\theta} \sum_{i,j} |O_{sim}^{(\theta)}(m_i,m_j) - O_{act}^{(k)}(m_i,m_j)|. \tag{12}$$

Based on the proposed method, we can obtain the AoAs with respect to all footsteps.

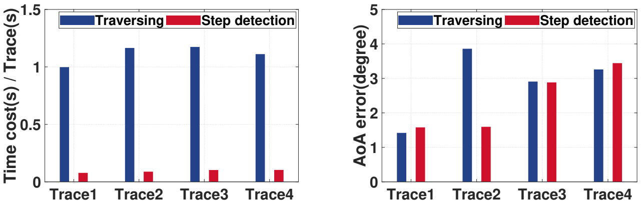
Until now, we can only obtain the sample shift and AoA with respect to the moment the footsteps are generated. Next, we need to recover the AoA between adjacent footsteps, where we can just collect environmental noise. Our insight is that the user tends to maintain a steady velocity between adjacent footsteps, hence, we can recover these missing AoAs via interpolation. In contrast to calculating sample shift and AoA, it takes less time to perform the interpolation. We conduct experiments to compare the performance of our footstep-based method and traversal method, and the results are shown in Fig. 10. In terms of the execution time, for each one-second sound sequence, the average processing time is reduced from $1.11s$ to $0.09s$, hence, the proposed method is $11\times$ faster than the traversal method. Moreover, the average error of AoA estimation has been reduced by $0.49°$.

C. *Multi-modal Modeling Stage*

The purpose of this stage is to fuse the Wi-Fi and sound model to track the user, and we realize this target with the following two parts: (1) We first leverage Wi-Fi and acoustic information to establish a physical model; (2) Then, *WSTrack* applies the multi-modal model to joint localize the target and estimate the initial location.

*1) **Tracking Model**:* In previous stage, we have derived the ellipse cluster and the AoA. When there is only one receiver, we need to fuse these information to estimate the trajectories of the user. In particular, the Fresnel ellipse with respect to $L_d(t)$ can be written as follows:

$$\begin{cases} a_t = L_d(t)/2 \\ b_t = \sqrt{a_t^2 - c_t^2} \\ c = \frac{\left\|\vec{l}_r - \vec{l}_t\right\|_2}{2} \end{cases} \Rightarrow \frac{x_t^2}{a_t^2} + \frac{y_t^2}{b_t^2} = 1, \tag{13}$$



(a) Time Cost

(b) AoA Estimation Error

Fig. 10: Performance verification.

where a_t, b_t, c are the parameters of current ellipse, $\boldsymbol{p}_t = (x_t, y_t)$ denotes the location of the user at time t. Meanwhile, we construct the ray equation based on AoA denoted by θ_t:

$$\begin{cases} k_{path}^{(t)} = \tan\theta_t \\ y_t = k_{path}^{(t)} \cdot x_t \end{cases}, \tag{14}$$

where $k_{path}^{(t)}$ denotes the slope of the ray.

In order to determine the intersection locations of the above model in Eq. 13 and Eq. 14, we can solve the following optimal function

$$\begin{cases} E_t = |y_t - k_{path}^{(t)} \cdot x_t| + |\frac{x_t^2}{a_t^2} + \frac{y_t^2}{b_t^2} - 1|, \\ \hat{\mathbf{p}}_t = (\hat{x}_t, \hat{y}_t) = \arg\min\limits_{(x_t,y_t)} |E_t|. \end{cases} \tag{15}$$

When we deal with all values in the sequence, we can obtain the overall trajectory $\hat{\mathcal{T}}$ denoted by

$$\hat{\mathcal{T}} = [\hat{\boldsymbol{p}}_1, \hat{\boldsymbol{p}}_2, \cdots, \hat{\boldsymbol{p}}_t]. \tag{16}$$

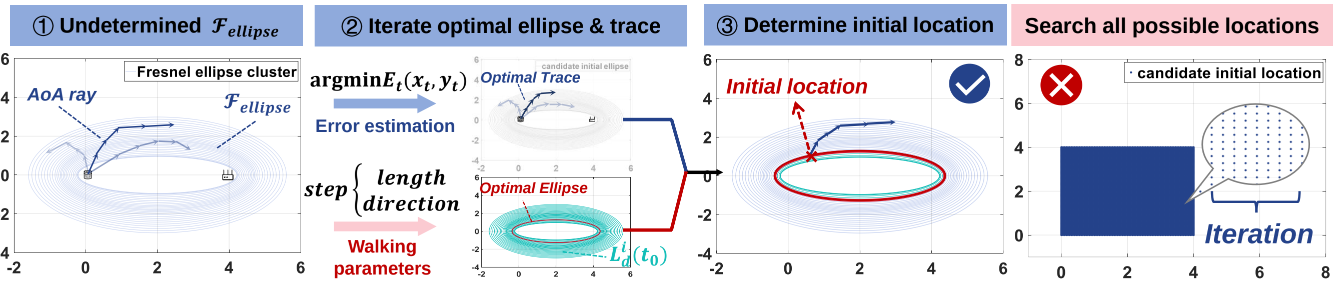
*2) **Initial Position Determination**:* To solve the trajectory, we should master the previous knowledge about the ellipse equation such as a_t and b_t in Eq. 13. However, a_t and b_t are determined by $L_d(t)$, and we can only determine $L_d(t)$ with the exact knowledge about $L_d(t_0)$ in Eq. 7. Hence, we should design a method to determine the initial path length $L_d(t_0)$.

For this purpose, the previous method usually searches the overall space to determine the initial location of the user. For example, for a $4m \times 4m$ localization area, if we divide the area into $5cm \times 5cm$ cells, it takes $80 \times 80 = 6400$ iterations to cover the overall space. To this end, our solution is as follows: We first iterate several Fresnel ellipses in localization area as the candidate initial ellipse, which will then be combined with the known cluster to form a complete $\mathcal{F}_{ellipse}$. Then we analyze the human gait features between any pair of adjacent steps to determine the optimal trace and corresponding initial ellipse.

As illustrated in Eq. 8, we describe the ellipse cluster as $\mathcal{F}_{ellipse} = \{L_d(t)|L_d(t_0)\}$. We define the candidate initial ellipses as $\{L_d^1(t_0), L_d^2(t_0), \ldots, L_d^i(t_0)\}$ (as shown in the Fig. 6 and the 2^{nd} part of Fig. 11). For each given initial ellipse $L_d^i(t_0)$, we have

$$\mathcal{F}_{ellipse}(i) = [L_d^i(t_0), \mathcal{F}_1, \ldots, \mathcal{F}_k]. \tag{17}$$

According to the multi-modal model in §Section. IV-C, we then calculate the error the same as Eq. 15 in each $\mathcal{F}_{ellipse}(i)$ to obtain the optimal trace $\hat{\mathcal{T}}(i)$ for each $\mathcal{F}_{ellipse}(i)$ (as shown in the 2^{nd} part of Fig. 11). According to the Eq. 17, each initial


① Undetermined $\mathcal{F}_{ellipse}$
Fresnel ellipse cluster
AoA ray
$\mathcal{F}_{ellipse}$
② Iterate optimal ellipse & trace
$\mathrm{argmin}E_t(x_t, y_t)$
Error estimation
candidate initial ellipse
Optimal Trace
step {length direction
Walking parameters
Optimal Ellipse
$L_d^i(t_0)$
③ Determine initial location
Initial location
Search all possible locations
candidate initial location
Iteration

(a) Iterating optimal Fresnel ellipse

(b) Iterating overall space

Fig. 11: Different algorithm of initial location estimation.

ellipse is corresponding to a cluster of ellipses; therefore, the number of $\hat{\mathcal{T}}(i)$ is the same as the number of candidate initial ellipse, and we record all trajectories as

$$TSet_{opt} = (\hat{\mathcal{T}}(1), \hat{\mathcal{T}}(2), \ldots, \hat{\mathcal{T}}(i)). \tag{18}$$

Now, the goal is to determine the optimal candidate trajectory. For this purpose, our insight is that the user tends to maintain a steady step during walking [25]. Consequently, we propose an indicator to measure the quality of each candidate trajectory. First, we segment each trajectory via the footstep, based on which we can obtain the step length and movement direction of each step. The above steps can be easily implemented, because we have described how to identify the user's footstep in §Section. IV-B. Next, we calculate the step length and movement direction as follows:

$$\begin{cases} step_{length} = \left\| \hat{\mathbf{p}}_i - \hat{\mathbf{p}}_{i-1} \right\|, \\ step_{direction} = angle(step_{length}), \end{cases} \tag{19}$$

where the $angle(\cdot)$ denotes the angle operation. Second, we extract the statistic features from footstep variables, including the mean and standard deviation. By leveraging these features, we exclude those improper candidate trajectories with unusual walking habits (including unreasonable step length, unstable cadence and large direction change). Finally, we consider the above features to select the optimal trace from $TSet_{opt}$.

When we finally determine the trace, the corresponding $\mathcal{F}_{ellipse}$ and $L_{\vec{d}}^i(t_0)$ are determined as well. Therefore, the initial position $\vec{l}_u(t_0)$ is the intersection point of the trace and the determined $L_d^i(t_0)$ (as shown by the red cross in the 3^{rd} part of Fig. 11).

Searching for the optimal ellipse significantly improves the computation efficiency. For example, according to Eq. 2 and Eq. 13, it is possible to utilize 200 ellipses to cover the $4m \times 4m$ space. Compared with 6400 iterations, searching the optimal ellipse saves up to 30× execution time, and this gap will become more significant when facing a larger space or finer granularity.

V. Implementation and Evaluation

In this section, we introduce our experimental methodology from both hardware and software perspectives.

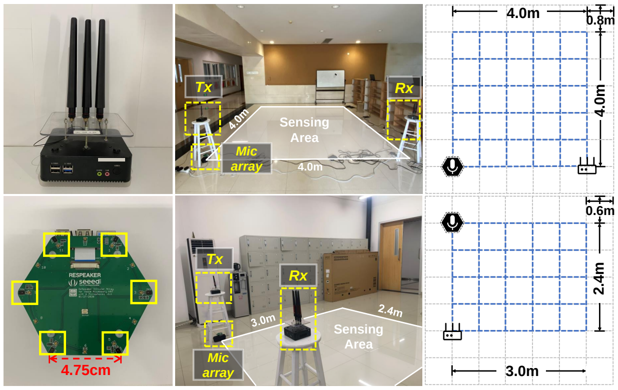


Fig. 12: Experimental Setup.

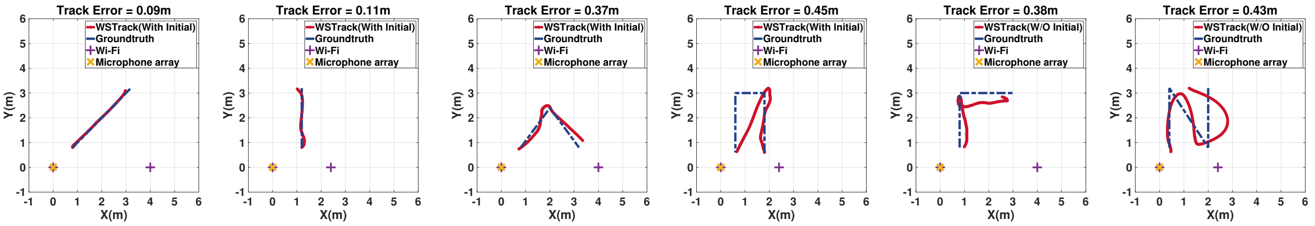
A. Implementation

Hardware: We use two off-the-shelf devices equipped with Intel 5300 NIC to collect Wi-Fi information, where the transmitter and receiver own one antenna and three antennas, respectively. In terms of sound, we use a Seeed Respeaker circular array with six microphones, which meets most of the commercial smart speakers such as Google Home and Amazon Echo. The microphone and the Wi-Fi transmitter are in the same location to make up the "voice assistants", which supports both acoustic and Wi-Fi functions. As shown in Fig. 12, the microphone array presents a regular hexagon shape with a side length of $4.75cm$. The sampling rate of the microphone is 48kHz, which covers the whole frequency range of the footstep sound. Both microphone array and Wi-Fi are connected to the laptop for data collection.

Software: We implement our system on a laptop with Intel i7-11800H CPU and 16G RAM. We use SSH tool based on linux-802.11n of Ubuntu 14.04.3 to control the transmission of packets. Finally, we write and execute the code in Matlab.

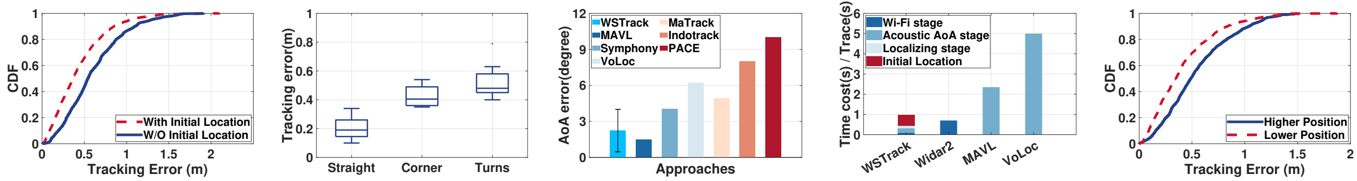
B. Overall Performance

We conduct experiments in two typical indoor scenarios to evaluate the performance of *WSTrack* (background noise of approximately $40dB$). As shown in Fig. 12, the experiment scenarios consist of an open space and an office, ranged $4m \times 4m$ and $2.4m \times 3m$ respectively. We employ five volunteers to collect trajectories without pre-training, and the trajectories include both simple (*e.g.*, straight and diagonal



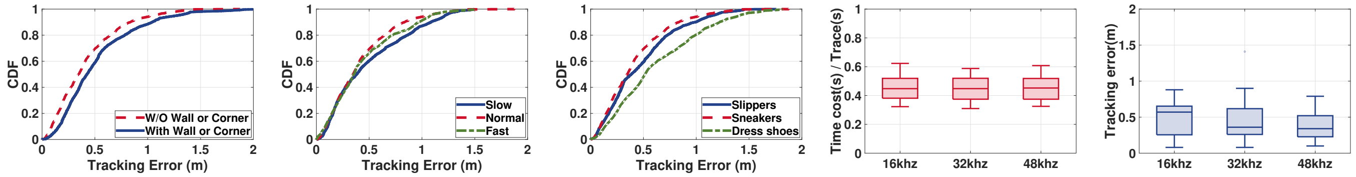
(a) Trajectory 1 (b) Trajectory 2 (c) Trajectory 3 (d) Trajectory 4 (e) Trajectory 5 (f) Trajectory 6

Fig. 13: Tracking examples of different trajectory With & W/O initial location.



(a) Tracking Accuracy (b) Impact of Path Complexity (c) AoA Accuracy (d) Computation Efficiency (e) Impact of Height

Fig. 14: Overall performance.



(a) Impact of Position (b) Impact of Pace Cadence (c) Impact of Footwear (d) Impact of Sampling Rate (e) Impact of Sampling Rate

Fig. 15: Impacts of diverse parameters.

lines) and complex (*e.g.*, corner, letter N) trajectories. Fig. 13 shows some tracking results of *WSTrack*.

Tracking Accuracy. The overall tracking performance of *WSTrack* is demonstrated in Fig. 14a. With the only one link, *WSTrack* achieves an average tracking error of $0.37m$ and $0.54m$ with and without the initial position, which is better than the state-of-art passive localization work by using only acoustic or Wi-Fi solutions (*e.g.*, [14] at $0.75m$ and [6] at $0.69m$) and even comparable to the work with multiple links ([19] at $0.35m$). Fig. 14b demonstrates performance of *WSTrack* under different types of paths, and we can achieve a median tracking error at $0.48m$ even if the trajectory has no less than two sharp turns. Consequently, the proposed method is robust to the environment.

AoA Accuracy. In terms of AoA, we compare *WSTrack* with several classic localization methods. As illustrated in Fig. 14c, the *WSTrack* can achieve a median error of less than $2.23°$, which is superior to other works (*e.g.*, [6] at $4°$, [5] at $6°$, [26] at $4.9°$, [19] at $8°$ and [27] at $10°$ for acoustic solution). It's worth noting that MAVL [4] achieves a better accuracy for their LOS AoAs. However, it takes more execution time to derive the AoA result. It is because we need to perform multiple calculations for the moving target.

Computation Efficiency. We conduct experiments to measure the execution time of different stages in *WSTrack*, and the performance is illustrated in Fig. 14d. Without the initial position determination stage, it takes about $0.46s$ for *WSTrack* to deal with a one-second trajectory, where Wi-Fi signal processing, sound signal processing and tracking stages cost $0.1s$, 0.21 and $0.15s$, respectively. Hence, it is possible to realize real-time tracking. Compared with the previous work, it takes $0.7s$ for Widar 2.0 [14] to deal with a one-second trajectory. MAVL and VoLoc [4], [5] focus on the static sound-source localization, so we only consider the execution time of AoA estimation. Such a process will take $2.35s$ and $5s$ respectively, hence, *WSTrack* saves 5× and 10× execution time for multi-modal localization. Besides, the execution time of initial location determination depends on the granularity of the Fresnel ellipses iteration. We select different determinations including 1, 2, and 3 Fresnel ellipses each time, the execution times for a one-second trajectory are respectively $0.54s$, $0.29s$ and $0.20s$. Meanwhile, the median error of coarse-grained iterations (3 Fresnel ellipses) is only $0.03m$ worse than that of fine-grained iterations (1 Fresnel ellipse).

C. *Parameter Study*

Impact of Microphone Height. To acquire high-quality audio, existing works usually stipulate the height of microphone array (*e.g.*, [5] on a desk for human voice and [27] on the floor for structure-borne footsteps). We conduct experiments at different heights to verify the performance of *WSTrack*, and the candidate heights include a lower one ($10cm$ high without contacting the floor) and a higher one (around $70cm$ above the ground to simulate the table or cabinet at home). Fig. 14e illustrates that the median tracking error increases from $0.35m$ to $0.49m$ after moving the array at a higher position. Hence, we can realize accurate tracking even with microphones at different heights.

Impact of Wi-Fi Deployment. Some of the acoustic localization arts utilize DoAs of echo reflected by walls, which requires a short distance between the microphone array and the wall, *e.g.*, VoLoc [5] ($0.2m - 0.8m$ to the wall) and Symphony [6] ($0.35m$ to the wall). Whereas *WSTrack* addresses this limitation, because it operates without using echo. Fig. 15a shows the performance of *WSTrack* with different distances between the microphone array and the wall ($< 0.5m$ and $> 1m$), and the median tracking error is $0.35m$ and $0.42m$ respectively. During the experiments, when the array and user both are near the corner or in a narrow corridor, the footsteps usually cause echo reverberation on the walls, which explains the increasing tracking error. Despite this phenomenon, *WSTrack* achieves high tracking accuracy no matter whether the array is near the wall, which enables *WSTrack* to meet diverse scenarios.

Impact of Pace Cadence. As discussed in §Section. IV-B2, the footstep cadence will influence the performance of AoA estimation. In order to verify the impact of footstep cadence, we require the volunteers to walk at three different paces: fast ($3.5 \sim 4$ steps/s), normal ($2 \sim 2.5$ steps/s) and slow (below 1.5 steps/s). The experimental results are shown in Fig. 15b, where the user's walking cadence introduces almost no impact on tracking performance.

Impact of Footwear. We also study the influence of different shoes on the performance of *WSTrack*. As illustrated in Fig. 15c, we can observe that the sneakers (default) and slippers achieve a more accurate performance.

Impact of Sampling Rate. According to the general settings, *WSTrack* collects footstep signals with $48khz$ sampling rate. We also select $16khz$ and $32khz$ to evaluate the performance of *WSTrack* under different sampling rates. Fig. 15d and Fig. 15e illustrate that the sampling rate has little impact on execution time. When the sampling rate decreases significantly, the tracking accuracy only decreases slightly. Moreover, since the sampling rates of commercial acoustic devices are usually larger than $16khz$, our proposed method can work under practical scenarios.

VI. Related Work

Position tracking is a pivotal part in the field of wireless sensing, whose basic idea is to analyze the user's kinematic pattern through radio frequency signals, so as to complete the tracking of the target. At present, the localization is mainly based on WiFi, acoustic and other signals. This section will introduce some well-known localization methods.

A. Wi-Fi Localization

Early works on Wi-Fi localization rely on the Received Signal Strength Indicator (RSSI) [28]–[34], and the main concerns are that RSSI performance is unstable and building a fingerprint database is time-consuming and labor-intensive. Therefore the arts explore to utilize the stable CSI as raw features [26], [35]–[38]. Widar [20] starts from the fine-grained information of human motion and proposed PLCR to estimate the user's velocity and position. IndoTrack [19] leverages the MUSIC algorithm to extract Doppler velocity and AoA for trajectory tracking. Furthermore, Widar 2.0 [14] takes multi-dimensional parameters into account to avoid cumulative errors caused by single parameters, and proposes a joint estimation algorithm to achieve single-link tracking at the decimeter level. Inspired by these works, *WSTrack* fuses Wi-Fi DFS and acoustic AoA for localization, which realizes accurate tracking with a single WiFi link and avoids the cumulative error caused by single-modal information.

B. Acoustic Localization

We can realize admirable performance with acoustic signal such as the human speech. Early, [39]–[41] perform acoustic localization with the ultrasonic sound; however, the ultrasonic devices are not usually available for smart home scenarios. Moreover, [42] leverages RSSI and sound signal transmission from cell phones to joint localization, but the proposed approach requires peer devices assistance. Recently, [4]–[6] propose voice localization techniques based on the voice assistant, and the basic idea is to leverage the echo reflected by the wall to localize the human. Compared with the aforementioned excellent works, *WSTrack* explores the detection of a user's pace for multi-modal tracking. The proposed method can track silent users based on the voice assistants. Hence, the voice assistants benefit from the ability to localize users, especially, we can analyze the user's habits from the historical trajectory to provide better services.

PACE in 2021 [27] is the most relative work compared with *WSTrack*. PACE proposes a system based on footstep localization that innovatively considers airborne and structural sound propagation. For this purpose, PACE utilizes an adversarial network to identify users by analyzing spectral characteristics of footsteps. However, PACE suffers from the following limitation: It is necessary to place the equipment on the ground to enable both airborne and structural sound propagation. In practice, the microphone arrays are not always placed at such low heights. In contrast, *WSTrack* can realize accurate localization with microphone array at different heights. Hence, it is suitable to deploy our proposed method in a wider scenario.

VII. Conclusions

In this paper, we present *WSTrack*, the first device-free localization system based on Wi-Fi and sound fusion. *WSTrack* achieves sub-meter level tracking without additional labor cost. We design a multi-modal model to joint estimate the user's location, and propose a method to determine the initial position by searching the optimal Fresnel ellipse. We conducted experiments on COTS devices to evaluate the performance. The results demonstrate that *WSTrack* is able to achieve tracking with a median error of $0.37m$ and $0.54m$ with and without initial location, the median AoA error is 2.23°, which is superior to state-of-the-art approaches.

Acknowledgment

The work in this paper is supported by National Natural Science Foundation of China (No. 62102282, 62002259, 61922055).

References

[1] S. Kennedy, H. Li, C. Wang, H. Liu, B. Wang, and W. Sun, "I can hear your alexa: Voice command fingerprinting on smart home speakers," in *Proc. of IEEE CNS*. IEEE, 2019, pp. 232–240.

[2] Y. Xie, F. Li, Y. Wu, and Y. Wang, "Hearfit: Fitness monitoring on smart speakers via active acoustic sensing," in *Proc. of IEEE INFOCOM*, 2021, pp. 1–10.

[3] H. Hu, L. Yang, S. Lin, and G. Wang, "A case study of the security vetting process of smart-home assistant applications," in *Proc. of IEEE SPW*. IEEE, 2020, pp. 76–81.

[4] M. Wang, W. Sun, and L. Qiu, "Mavl: Multiresolution analysis of voice localization," in *USENIX NSDI*, 2021, pp. 845–858.

[5] S. Shen, D. Chen, Y.-L. Wei, Z. Yang, and R. R. Choudhury, "Voice localization using nearby wall reflections," in *Proc. of ACM Mobicom*, 2020, pp. 1–14.

[6] W. Wang, J. Li, Y. He, and Y. Liu, "Symphony: localizing multiple acoustic sources with a single microphone array," in *Proc. of ACM SenSys*, 2020, pp. 82–94.

[7] C. Wu, Z. Yang, Z. Zhou, K. Qian, Y. Liu, and M. Liu, "Phaseu: Real-time los identification with wifi," in *Proc. of IEEE INFOCOM*, 2015, pp. 2038–2046.

[8] X. Zheng, J. Wang, L. Shangguan, Z. Zhou, and Y. Liu, "Design and implementation of a csi-based ubiquitous smoking detection system," *IEEE/ACM Transactions on Networking*, vol. 25, no. 6, pp. 3781–3793, 2017.

[9] K. Yang, X. Zheng, J. Xiong, L. Liu, and H. Ma, "Wiimg: Pushing the limit of wifi sensing with low transmission rates," in *Proc. of IEEE SECON*, 2022, pp. 1–9.

[10] L. Xu, X. Zheng, X. Li, Y. Zhang, L. Liu, and H. Ma, "Wicam: Imperceptible adversarial attack on deep learning based wifi sensing," in *Proc. of IEEE SECON*, 2022, pp. 10–18.

[11] R. Xiao, J. Liu, J. Han, and K. Ren, "Onefi: One-shot recognition for unseen gesture via cots wifi," in *Proc. of ACM SenSys*, 2021, pp. 206–219.

[12] Y. Wang, K. Wu, and L. M. Ni, "Wifall: Device-free fall detection by wireless networks," *IEEE Transactions on Mobile Computing*, vol. 16, no. 2, pp. 581–594, 2016.

[13] Y. Cao, H. Chen, F. Li, S. Yang, and Y. Wang, "Awash: handwashing assistance for the elderly with dementia via wearables," in *Proc. of IEEE INFOCOM*, 2021, pp. 1–10.

[14] K. Qian, C. Wu, Y. Zhang, G. Zhang, Z. Yang, and Y. Liu, "Widar2. 0: Passive human tracking with a single wi-fi link," in *Proc. of ACM MobiSys*, 2018, pp. 350–361.

[15] Y. Xie, J. Xiong, M. Li, and K. Jamieson, "md-track: Leveraging multi-dimensionality for passive indoor wi-fi tracking," in *Proc. of ACM Mobicom*, 2019, pp. 1–16.

[16] C. R. Karanam, B. Korany, and Y. Mostofi, "Tracking from one side: Multi-person passive tracking with wifi magnitude measurements," in *Proc. of ACM IPSN*, 2019, pp. 181–192.

[17] R. H. Venkatnarayan, M. Shahzad, S. Yun, C. Vlachou, and K.-H. Kim, "Leveraging polarization of wifi signals to simultaneously track multiple people," in *Proc. of ACM IMWUT*, 2020, pp. 1–24.

[18] J. Xiong, K. Sundaresan, and K. Jamieson, "Tonetrack: Leveraging frequency-agile radios for time-based indoor wireless localization," in *Proc. of ACM Mobicom*, 2015, pp. 537–549.

[19] X. Li, D. Zhang, Q. Lv, J. Xiong, S. Li, Y. Zhang, and H. Mei, "Indotrack: Device-free indoor human tracking with commodity wi-fi," in *Proc. of ACM IMWUT*, 2017, pp. 1–22.

[20] K. Qian, C. Wu, Z. Yang, Y. Liu, and K. Jamieson, "Widar: Decimeter-level passive tracking via velocity monitoring with commodity wi-fi," in *Proc. of ACM Mobihoc*, 2017, pp. 1–10.

[21] D. Wu, D. Zhang, C. Xu, Y. Wang, and H. Wang, "Widir: walking direction estimation using wireless signals," in *Proc. of ACM IMWUT*, 2016, pp. 351–362.

[22] H. Wang, D. Zhang, J. Ma, Y. Wang, Y. Wang, D. Wu, T. Gu, and B. Xie, "Human respiration detection with commodity wifi devices: do user location and body orientation matter?" in *Proc. of ACM IMWUT*, 2016, pp. 25–36.

[23] J. Chen, Z. Wang, D. Tuo, Z. Wu, S. Kang, and H. Meng, "Fullsubnet+: Channel attention fullsubnet with complex spectrograms for speech enhancement," in *Proc. of IEEE ICASSP*, 2022, pp. 7857–7861.

[24] X. Tong, H. Wang, X. Liu, and W. Qu, "Mapfi: Autonomous mapping of wi-fi infrastructure for indoor localization," *IEEE Transactions on Mobile Computing*, 2021.

[25] W. Wang, A. X. Liu, and M. Shahzad, "Gait recognition using wifi signals," in *Proc. of ACM IMWUT*, 2016, pp. 363–373.

[26] X. Li, S. Li, D. Zhang, J. Xiong, Y. Wang, and H. Mei, "Dynamic-music: accurate device-free indoor localization," in *Proc. of ACM IMWUT*, 2016, pp. 196–207.

[27] C. Cai, H. Pu, P. Wang, Z. Chen, and J. Luo, "We hear your pace: Passive acoustic localization of multiple walking persons," in *Proc. of ACM IMWUT*, 2021, pp. 1–24.

[28] A. Rai, K. K. Chintalapudi, V. N. Padmanabhan, and R. Sen, "Zee: Zero-effort crowdsourcing for indoor localization," in *Proc. of ACM Mobihoc*, 2012, pp. 293–304.

[29] A. Goswami, L. E. Ortiz, and S. R. Das, "Wigem: A learning-based approach for indoor localization," in *Proc. of ACM CoNEXT*, 2011, pp. 1–12.

[30] X. Tian, R. Shen, D. Liu, Y. Wen, and X. Wang, "Performance analysis of rss fingerprinting based indoor localization," *IEEE Transactions on Mobile Computing*, vol. 16, no. 10, pp. 2847–2861, 2016.

[31] Y. Wen, X. Tian, X. Wang, and S. Lu, "Fundamental limits of rss fingerprinting based indoor localization," in *Proc. of IEEE INFOCOM*, 2015, pp. 2479–2487.

[32] D. Li, J. Xu, Z. Yang, Y. Lu, Q. Zhang, and X. Zhang, "Train once, locate anytime for anyone: Adversarial learning based wireless localization," in *Proc. of IEEE INFOCOM*, 2021, pp. 1–10.

[33] J. Wang, N. Tan, J. Luo, and S. J. Pan, "Woloc: Wifi-only outdoor localization using crowdsensed hotspot labels," in *Proc. of IEEE INFOCOM*, 2017, pp. 1–9.

[34] C. Wu, Z. Yang, C. Xiao, C. Yang, Y. Liu, and M. Liu, "Static power of mobile devices: Self-updating radio maps for wireless indoor localization," in *Proc. of IEEE INFOCOM*, 2015, pp. 2497–2505.

[35] X. Tong, H. Li, X. Tian, and X. Wang, "Wi-fi localization enabling self-calibration," *IEEE/ACM Transactions on Networking*, vol. 29, no. 2, pp. 904–917, 2021.

[36] K. Jiokeng, G. Jakllari, A. Tchana, and A.-L. Beylot, "When ftm discovered music: Accurate wifi-based ranging in the presence of multipath," in *Proc. of IEEE INFOCOM*, 2020, pp. 1857–1866.

[37] S. Tan, L. Zhang, Z. Wang, and J. Yang, "Multitrack: Multi-user tracking and activity recognition using commodity wifi," in *Proc. of ACM CHI*, 2019, pp. 1–12.

[38] X. Tong, Y. Wan, Q. Li, X. Tian, and X. Wang, "Csi fingerprinting localization with low human efforts," *IEEE/ACM Transactions on Networking*, vol. 29, no. 1, pp. 372–385, 2020.

[39] R. Want, A. Hopper, V. Falcao, and J. Gibbons, "The active badge location system," *ACM Transactions on Information Systems*, vol. 10, no. 1, pp. 91–102, 1992.

[40] M. Minami, Y. Fukuju, K. Hirasawa, S. Yokoyama, M. Mizumachi, H. Morikawa, and T. Aoyama, "Dolphin: A practical approach for implementing a fully distributed indoor ultrasonic positioning system," in *Proc. of ACM IMWUT*. Springer, 2004, pp. 347–365.

[41] P. Lazik, N. Rajagopal, O. Shih, B. Sinopoli, and A. Rowe, "Alps: A bluetooth and ultrasound platform for mapping and localization," in *Proc. of ACM SenSys*, 2015, pp. 73–84.

[42] H. Liu, Y. Gan, J. Yang, S. Sidhom, Y. Wang, Y. Chen, and F. Ye, "Push the limit of wifi based localization for smartphones," in *Proc. of ACM Mobicom*, 2012, pp. 305–316.