ASE: Practical Acoustic Speed Estimation Beyond Doppler via Sound Diffusion Field

SHENG LYU, The University of Hong Kong, Hong Kong SAR, China
CHENSHU WU, The University of Hong Kong, Hong Kong SAR, China

Passive human speed estimation plays a critical role in acoustic sensing. Despite extensive study, existing systems, however, suffer from various limitations: First, the channel measurement rate proves inadequate to estimate high moving speeds. Second, previous acoustic speed estimation exploits Doppler Frequency Shifts (DFS) created by moving targets and relies on microphone arrays, making them only capable of sensing the radial speed within a constrained distance. To overcome these issues, we present ASE, an accurate and robust Acoustic Speed Estimation system on a single commodity microphone. We propose a novel Orthogonal Time-Delayed Multiplexing (OTDM) scheme for acoustic channel estimation at a high rate that was previously infeasible, making it possible to estimate high speeds. We then model the sound propagation from a unique perspective of the acoustic diffusion field, and infer the speed from the acoustic spatial distribution, a completely different way of thinking about speed estimation beyond prior DFS-based approaches. We further develop novel techniques for motion detection and signal enhancement to deliver a robust and practical system. We implement and evaluate ASE through extensive real-world experiments. Our results show that ASE reliably tracks walking speed, independently of target location and direction, with a mean error of 0.13 m/s, a reduction of 2.5x from DFS, and a detection rate of 97.4% for large coverage, *e.g.*, free walking in a 4m × 4m room. We believe ASE pushes acoustic speed estimation beyond the conventional DFS-based paradigm and inspires exciting research in acoustic sensing. Code is available at https://github.com/aiot-lab/ASE.

CCS Concepts: • **Human-centered computing** → **Ubiquitous and mobile computing systems and tools**; • **Networks** → **Cyber-physical networks**.

ACM Reference Format:
Sheng Lyu and Chenshu Wu. 2025. ASE: Practical Acoustic Speed Estimation Beyond Doppler via Sound Diffusion Field. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 3, Article 115 (September 2025), 26 pages. https://doi.org/10.1145/3749475

1 INTRODUCTION

Capturing the portraits of indoor human activities is an enduring task in the wide sensing community [31, 48, 78, 85, 97]. Frequently, human subjects are in motion, rendering **passive speed estimation** one of the most fundamental components in human sensing. At the heart of understanding the physical state of moving subjects, speed provides valuable insights into human behaviors and health. With speed profiles, a broad range of applications can be accommodated, such as gait recognition [13, 86], fall detection [23, 35], human activity recognition [20, 44, 99], tracking [10, 40, 48] and fitness tracking [62, 79, 87], *etc.*

Particularly, walking speed estimation plays an important role in well-being monitoring. It is increasingly perceived as the sixth vital sign [17, 41] which is closely associated with and predictive of one's health conditions [79]. Slowing walking speed suggests increased frailty, leading to potential physical and cognitive decline [51, 52]. Moreover, walking speed acts as a biomarker for gait recognition [80] and an effective indicator of risky falls

Authors' Contact Information: Sheng Lyu, The University of Hong Kong, Hong Kong SAR, China, shenglyu@connect.hku.hk; Chenshu Wu, The University of Hong Kong, Hong Kong SAR, China, chenshu@cs.hku.hk.

ACM 2474-9567/2025/9-ART115
https://doi.org/10.1145/3749475

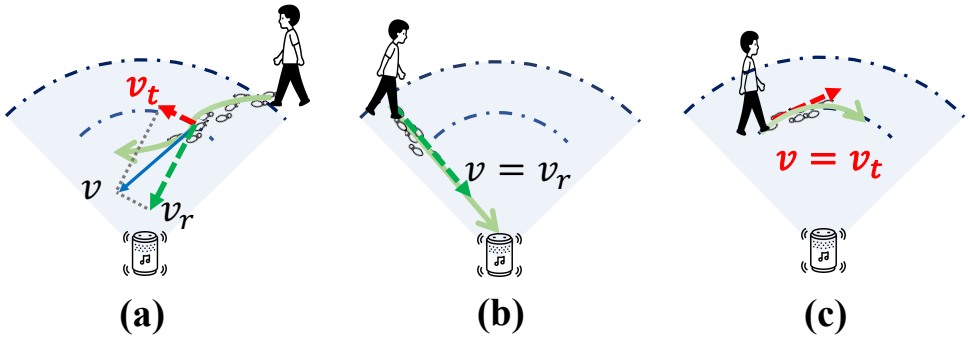


Fig. 1. ASE vs. DFS. Walking speed $\vec{v}$ can be decomposed to radial speed v_r and tangent speed v_t. DFS can only capture radial speed v_r, but fails to capture v_t. Conversely, ASE can capture both v_r and v_t, a complete estimation of v.

[23]. Enabling these applications calls for accurate and robust speed estimation, preferably in a contactless and passive manner.

Passive speed estimation, however, is a long-standing open challenge. Various approaches have been proposed for indoor speed estimation using vision systems [4, 6], wireless signals [76, 79, 94], and acoustic signals [10, 37, 92]. Camera-based approaches, such as the VICON motion capture system [6], usually provide the most accurate speed but require complex and expensive hardware setups and are limited within a functional area. Wireless sensing has recently been extensively studied, yet mostly relies on specialized radars [1, 3, 77, 82, 90, 91], or certain WiFi chipsets.

Acoustic devices, especially smart speakers and IoT audio, are now widely and interactively available in our everyday lives, often as plug-and-play devices with co-located microphones and speakers, making acoustic sensing an increasingly hot topic in recent years [15, 16, 55, 67, 68, 97]. The widespread usage of such audio devices hold substantial potential for enabling significant applications, provided that accurate and robust speed estimation can be achieved through acoustic sensing.

Existing acoustic speed estimation methods mostly rely on the Doppler Frequency Shifts (DFS) caused by target movements to derive the speed [35, 92]. These approaches, however, suffer from three fundamental limitations in acoustic speed estimation:

- The low sound speed imposes an innate limit on the maximum Channel State Information (CSI) rate achievable on an acoustic channel [97], which unfortunately, is insufficient for estimating high speeds (*e.g.*, typical walking speed of around 1 m/s).
- Depending on the specific moving direction and location, Doppler-based approaches can only capture the partial speed projected in a radial direction that creates reflection path length changes and thus frequency shifts [48, 74], as shown in Fig. 1.
- DFS often only utilizes one or a few reflection paths off the target for sensing, leading to performance degradation with the increase of distance, *e.g.*, over 1 m [30, 31, 37, 75, 97], and thereby confining them for short-range sensing.

To overcome these limitations, we pose a crucial research question: *Can we achieve robust and practical acoustic speed estimation beyond DFS-based approaches?* We present ASE, a completely novel pipeline that achieves practical acoustic speed estimation using only a *single* microphone channel. At a high level, it features three distinct components: 1) an innovative modulation scheme that breaks the upper limit of acoustic CSI, 2) a novel theoretical model that can derive the whole speed, rather than solely radial speed, while being independent of target/device locations and moving orientations, and 3) a set of innovative techniques that make ASE more robust in practice.

Regarding the first challenge, we propose a novel Orthogonal Time Delayed Multiplexing (OTDM) scheme. Based on in-depth analysis, we reveal that the CSI rate is inherently limited by the travel speed of sound and the propagation distance. This limitation restricts acoustic devices from estimating high speeds, presenting a fundamental challenge in acoustic speed estimation. Inspired by OFDM, we develop a specialized modulation framework to mix the signals effectively without extra cost. We leverage the separation ability of orthogonal signals to send two signals concurrently, while delaying one sequence for half of the frame time. Through this approach, we can effectively atomize the original frame to its half and boost 2x of the CSI rate.

To surpass Doppler-based speed estimation, we build a comprehensive model that can capture the entire speed. Our model builds upon the acoustic diffusion model, as illustrated in Fig. 3. Specifically, we investigate the fundamental properties of sound diffusion indoors and introduce a concise approach for speed estimation. We draw aspiration from previous physical studies of room acoustics [28], which model the diffusive sound propagation. Conceptually, ASE employs the distinct spatial distribution of the sound pressure field. The analysis of spatial-temporal properties of the sound field shows that the correlation of the sound energy embodies the speed of a moving target. Although the Autocorrelation Function (ACF) has been employed for extracting periodic vital signs [21, 70], building the theoretical and practical bridge between the ACF of sound and speed is new and non-trivial. In ASE, we establish the theoretical relationship in the context of the acoustic diffusion field and model the ACF of acoustic CSI as a function of the moving speed. Different from DFS, the proposed model statistically leverages all the reflection signals, and integrates over all possible directions, thus making speed estimation less dependent on the moving location and direction and building a novel foundation for speed estimation with commodity acoustic devices.

We also incorporate several effective designs to transform ASE into a practical system. We employ pseudo-noise code in ASE with nice orthogonality and tolerance to interference for CSI estimation from inaudible sound signals. By careful modulation and filtering, the sensing signals are made hardly audible compared to previous designs like the widely used chirps [68], which are considerably intrusive to human ears in practice. Moreover, we identify robust motion indicators and devise an effective approach that embraces frequency diversity to significantly boost the weak signals for speed estimation, which largely extends the sensing coverage and improves the speed estimation accuracy.

We prototype ASE on commodity audio devices. We conduct comprehensive experiments to evaluate ASE for diverse walking behaviors in real-world indoor scenarios. The results show that ASE achieves a mean speed accuracy of 0.13 m/s with a 90%-tile error of < 0.2 m/s and an overall detection rate of 97.4% in a 4m $\times$ 4m room, significantly outperforming prior DFS-based approach, which yields a 2.5$\times$ higher mean error under the same conditions. We further conduct case studies on human activity recognition, fall detection, and gait analysis as potential applications by profiling the speeds of different human activities. The superior performance validates ASE and its underlying model as a new paradigm of acoustic speed estimation for many applications.

Contributions: We believe ASE lays a completely new foundation for acoustic speed estimation and offers new insights into the field by making the following core contributions:

- We develop OTDM, the first-of-its-kind modulation scheme that allows acoustic CSI estimation at a high rate exceeding the previous maximum possible rate.
- To the best of our knowledge, we are the first to employ the sound diffusion model for speed estimation and integrate it with the acoustic channel, which fundamentally differs from DFS-based approaches.
- We design and implement ASE system using a single commodity microphone. We incorporate a pipeline of distinct techniques and conduct experiments to validate its superior performance over prior approaches.

In the rest of the paper, we first introduce a universal dilemma in acoustic speed estimation and our OTDM design in §2. Then we present the theoretical model in §3 and the design of ASE in §4. Implementation details and

evaluations are presented in§5, respectively, followed by discussions in §6 and related works in §7. We conclude in §8.

2 OTDM DESIGN

We first discuss the contradiction between high speed estimation and the acoustic CSI rate. After that, we will present our novel solution of Orthogonal Time Delayed Multiplexing (OTDM) scheme to boost the CSI rate.

2.1 High Speed and Low CSI Rate Dilemma

Initially, we will discuss a universal problem of acoustic speed estimation. To profile speed information, a common practice is to estimate Channel State Information (CSI), which characterizes how sound propagates in a Tx/Rx system. The CSI rate F_s refers to how many CSIs we can acquire per second in the system. However, the inadequacy of the acoustic CSI rate imposes a fundamental challenge for speed estimation. Conceptually, the lower CSI sampling rates we acquire, the smaller the maximum frequency shift can be observed. Given a CSI rate of F_s, the maximum frequency shift detectable on the Doppler spectrum is $F_s/2$. Assuming a carrier frequency of f, the maximum measurable speed is

$$v_{\max} = \frac{F_s}{2 \cdot f} \cdot c, \tag{1}$$

which is only 0.4 m/s considering $f = 20$ kHz and $F_s = 49$ Hz. Here c denotes the speed of the sound.When factoring in noise, this obviously falls short of measuring typical indoor walking speeds.

Then another question arises: can we increase the acoustic CSI rate? Unfortunately, slow sound speed imposes an inherent limitation on the maximum achievable CSI rate on an acoustic channel. Assuming the sampling rate of acoustic signal is f_s (note that f_s is very different from the CSI rate F_s), plus x-m propagation path, then the CSI rate is given by

$$F_s = \frac{f_s}{x/c \cdot f_s} = \frac{c}{x}. \tag{2}$$

As we can see, if we want to increase the CSI rate, the only way is to limit the sensing distance. For example, considering the in-air sound speed of 343 m/s and the longest propagation path of 7 meters, the minimum channel measurement interval should be larger than 20 ms , resulting in a CSI rate of 49 Hz. However, even if we narrow this range to 3.5 meters, the maximum possible CSI rate without causing signal mixture only increases to 100 Hz, still insufficient for estimating human walk speed around 1 m/s.

This observation reminds us that the current CSI rate cannot support high speed estimation for daily activities (*e.g.*, walking), and it is infeasible to increase the CSI rate given the sensing distance for the current channel design. To tackle this problem, we present a novel OTDM design in the next section.

2.2 OTDM Transmission Scheme

Based on the above discussion, we find a challenging contradiction for acoustic speed estimation: Given the current channel design, it is impractical to enhance the CSI rate, while not reducing the sensing range or causing signal mixture. The core of this dilemma is the failure to fully utilize the channel capacity. This situation reminds us of concurrent transmission schemes in network data communication, which are achieved through multiplexing modulations to allow for greater resource utilization. These multiplexing schemes can be roughly divided into Time Division Multiplexing (TDM), Frequency Division Multiplexing (FDM), Code Division Multiplexing (CDM), *etc*, depending on which dimension is being reused. Notably, Orthogonal Frequency Division Multiplexing (OFDM) is an extension of FDM that divides the frequency band into closely spaced, orthogonal subcarriers. Each subcarrier can be modulated independently, thereby providing a larger data rate within the same frequency bandwidth.

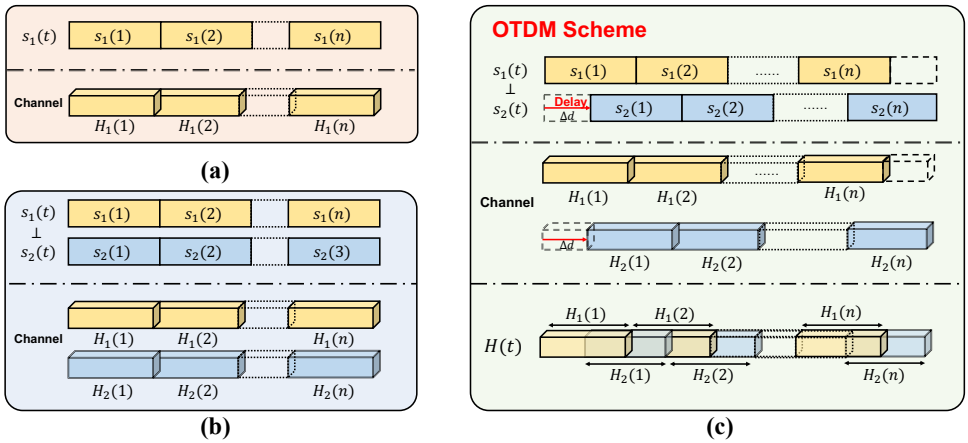


Fig. 2. OTDM Scheme. (a) Given $s_1(t)$, the channel is estimated by frame. (b) Given orthogonal sequences $s_1(t)$ and $s_2(t)$, we can get two channel estimations at the same time. (c) Orthogonal sequences $s_1(t)$ and $s_2(t)$, with $s_2(t)$ delayed by Δd, are modulated into one sequence $s(t)$. $H_1(t)$ and $H_2(t)$ estimated from the two sequences are concatenated into $H(t)$ alternately.

In our context, we care about fully exploiting the *temporal channel capacity* to achieve a higher CSI rate within the same time period. As shown in Fig. 2(a), transmitting a single probing signal yields one channel estimation H_1, which has proven insufficient. The question is, if we can estimate the channel simultaneously, can we get more channel estimations and increase the CSI rate? Thanks to the separability of orthogonal signals, with two orthogonal sequences $s_1(t)$ and $s_2(t)$, we can indeed obtain two channel estimations simultaneously, as shown in Fig. 2(b). However, this alone does not increase the channel rate because the measurements are aligned, resulting in redundant channel information. Inspired by OFDM, if we introduce spacing between the temporal estimations, we can observe the channel at different times and split the channel into more granular frame clips, potentially increasing the CSI rate. To achieve this, we are inspired to design a novel modulation scheme named Orthogonal Time-Delayed Multiplexing (OTDM).

The key idea is that, by using orthogonal signals, we can transmit multiple sensing signals concurrently over the same physical channel (*i.e.*, speaker-microphone pair) and perform channel estimation separately, which brings more samples of the channel measurements. If we further *shift the multiple orthogonal signals by a certain amount of time*, we effectively obtain finer-grained CSI measurements in the time domain, *i.e.*, a higher CSI rate. By OTDM, the duration and interval of each separate sensing signal will not be shortened.

Suppose we have two sequences $s_1(t)$ and $s_2(t)$ with decent orthogonality over time. Then we can transmit them concurrently, while still being able to separate them on the receiver side by using correlation. Previous orthogonal transmission schemes typically transmit synchronized signals. Differently, as shown in Fig. 2, to enhance the CSI rate, we need to delay one sequence by $\Delta d = N_s/2$, where N_s is the default sensing interval as described in §4.1. Then we can acquire two sets of channel estimation: $H_1 \in \mathbb{R}^n$ and $H_2 \in \mathbb{R}^n$. The two CSI measurements, estimated from the two orthogonal sequences respectively, characterize the physical channel independently, yet at slightly shifted times. Therefore, by stacking the two series of CSI measurements, *i.e.*, $[H_1(1), H_2(1), \cdots, H_1(n), H_2(n)]$, we can double the CSI rate from $1/N_s$ to $2/N_s$ without shortening the sequence length (*i.e.*, signal duration). Our extensive experiment will show the effectiveness of OTDM design. We will detail the design for the most common case (*i.e.*, one speaker) in §4.2, while our scheme can generalize to more.

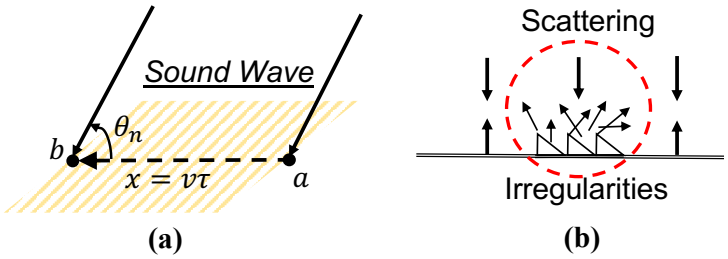


Fig. 3. Sound Diffusion Model. (a) The sound wave traverses while an object is moving $a \rightarrow b$. (b) The sound wave is diffusive in all directions by irregular reflectors.

3 Acoustic Diffusion Speed Model

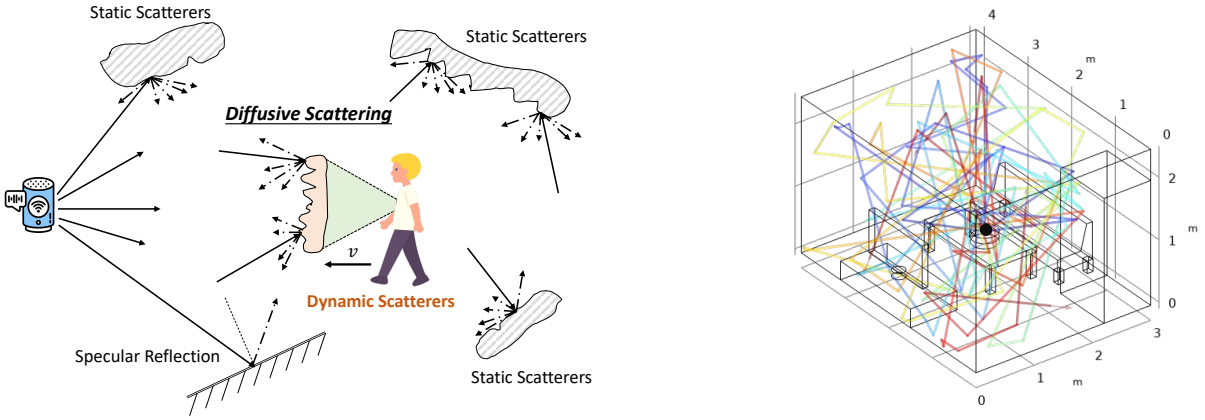
Conventional approaches for speed estimation relying on DFS can only measure the radial speed, making it a location- and direction-dependent solution, as shown in Fig. 1. In this section, we will introduce a novel theoretical model for capturing the complete speed. Our model is primarily inspired by room acoustics [28]. Below, we first briefly review statistical room acoustics and then establish the model that works with channel measurements on commodity audio devices.

3.1 Limitations of DFS

Existing DFS-based acoustic speed estimation implicitly or explicitly relies on a simplistic propagation model. These methods typically involve identifying the primary reflection path between the human body and audio devices and deducing the DFS from the corresponding bins. Consequently, these techniques yield only partial information regarding speed estimation [43, 94], specifically the radial speed v_r between the human and acoustic device pair, neglecting the tangent component v_t, as illustrated in Fig. 1. Essentially, if we only consider one or a few reflection paths, the tangent component will inevitably be lost. Fortunately, we notice that the acoustic signals will be scattered by the numerous reflectors to different directions in the environment, as can be seen from Fig. 3(b) and Fig. 4. If we can efficiently leverage the multi-path reflections, it equivalently creates various estimations of speed along different directions. Conceptually, the reflectors can be regarded as many "virtual speakers", offering numerous speed observations of different tangent components. From a statistical perspective, we can imagine these speed components are synthesized into a complete estimation of speed v, which contains both racial speed v_r and tangent speed v_t. To this end, we explore the propagation properties of sound and derive the speed information from a statistical approach. We will first introduce the acoustic diffusion model in §3.2 and generalize it to CSI in §3.3.

3.2 Sound Diffusion Field

Given a bounded sound propagation scenario (*e.g.*, indoor space), the sound energy is partly reflected/scattered by the obstacles, as shown in Fig. 3. Considering a moving target in space that continuously distorts the sound propagation, a diffuse sound field will be created due to the continuous redistribution of sound energy, especially in an environment with rich diffuseness such as a room. Thus, according to the acoustic wave equation [46], we can express the sound pressure as $\mathbf{p} = p(x,t) = Ae^{j(\omega t - \vec{k}\cdot\vec{r})}$, where A denotes the amplitude, ω is the frequency of the wave, $\vec{k}$ indicates the propagation vector, and c is the propagation speed of sound. r is the position vector, *i.e.*, $\vec{r} = x\vec{i}+y\vec{j}+z\vec{k}$. Assume the space is excited by a limited-band signal, and consider two adjacent observation points a and b separated by a distance of x, as shown in Fig. 3. The sound pressure is expressed as $p_a(t) = A\cos(\omega t - \phi)$ and $p_b(t) = A\cos(\omega t - \phi - kx\cos(\theta))$, respectively, where θ denotes the direction of the incident sound wave,



(a) (b)

Fig. 4. Diffusive Scattering in the room. **(a)** Illustration of diffusion scattering and specular reflections. The acoustic waves not only experience specular reflections, but also diffusive scatterings on any non-flat planes. **(b)** A simulation result of how an acoustic wave is reflected and scattered in a common room. The black point emits the acoustic rays with a frequency range of 17 kHz- 20 kHz. By employing the rich multi-path semantics, these scatterers are creating "virtual speakers", thus creating different views of speed estimation.

A and ϕ are random amplitudes and angles, and k and ω represent the wave number and center frequency of the sound signals. The correlation coefficient of the sound energy over space provides insight into the degree of diffuseness. Specifically, the spatial correlation over $p_a(t)$ and $p_b(t)$ is expressed as

$$\psi_p = \frac{\overline{p_a(t)p_b(t)}}{\sqrt{\overline{p_a(t)^2} \cdot \overline{p_b(t)^2}}}, \tag{3}$$

where ψ_p represents the correlation coefficient, and $\bar{\cdot}$ denotes operation of time average. Since sound is a plane wave, we can derive the time average of $p_a(t)^2$ and $p_b(t)^2$ both as $\frac{A^2}{2}$, and meanwhile, the average of the product of sound pressure is computed as $\overline{p_a(t)p_b(t)} = \frac{1}{\Delta t}\int_t^{t+\Delta t} p_a(t)p_b(t)dt = \frac{A^2}{2}\cos(kx\cos\theta)$. Therefore, by substituting them into Eq. (3), we derive the correlation for the direction of incidence θ as

$$\psi_p\left(x,\theta\right) = \cos\left(kx\cos\theta\right). \tag{4}$$

Then by integrating Eq. (4) over all directions, we can obtain the spatial correlation of the total sound pressure. If we consider that the incident sound waves are distributed in a plane, namely 2D model, we can get

$$\psi_p\left(x\right) = \frac{1}{2\pi}\int_0^{2\pi} \psi\left(x,\theta\right) d\theta = J_0(kx), \tag{5}$$

where $J_0(x) = \frac{1}{2\pi}\int_0^{2\pi}\cos(x\cos\theta)d\theta$ is the 0^{th}-order Bessel function of its first kind. If we consider 3D scattering model, *i.e.*, the sound waves are distributed in a sphere, we can acquire the correlation coefficient as,

$$\psi_p\left(x\right) = \frac{1}{4\pi}\int_0^{\pi}\int_0^{2\pi} \psi\left(x,\theta\right) d\phi d\theta = \frac{\sin(kx)}{kx}, \tag{6}$$

where ϕ is the azimuth angle, varying from 0 to 2π. Further considering a target moving from point a to b at a speed of v, we have $x = v\tau$, where τ is the traveling time. Thus, the above equation can be written as

$$\psi_p(v;\tau) = \frac{\sin(kv\tau)}{kv\tau}. \tag{7}$$

It bridges the spatial correlation of the sound pressure ψ_p with a target's moving speed v, independent of the moving direction and location. Note that in both scenarios, we average diffuseness across *all directions*, in contrast to DFS, that only takes the radial speed into consideration, promising a potential approach for speed estimation beyond the Doppler-based approach.

Understanding Sound Diffuseness in Room Environment: Indoor environments inherently exhibit scattering effects. As established by modern room acoustic studies [14, 28, 61], the level of scattering increases with increasing sound frequency and the roughness of surfaces. In common room settings, these scattering effects are ubiquitous and pervasive. Curved walls, textured surfaces, or objects with uneven shapes do not reflect sound in a single, coherent direction. Instead, they disperse sound energy across a wide range of angles, much like how a crumpled piece of paper scatters light compared to a smooth mirror. This significantly increases the scattering coefficient, meaning more sound energy is diffused rather than specularly reflected. As summarized in Tab. 1, typical scattering coefficients for indoor materials range from 0 (no scattering) to 1 (full scattering). Ordinary objects such as chairs and tables act as scatterers, diffusing incident sound energy throughout the space (see Fig. 3(b)). The scattering becomes more pervasive in complex room settings and with irregular reflectors. This implies that modeling sound diffusion in a room offers a new perspective on its acoustic behavior. By adopting a statistical perspective on the sound field, we can analyze the acoustic environment without considering the specific material types of objects, whether they are on Line of Sight (LoS) or Non-Line of Sight (NLoS) paths. In other words, this statistical approach enables us to estimate overall acoustic properties in a more comprehensive way than the DFS method and simplifies the processing by eliminating the need to separate specific paths from the received signal.

To achieve ASE, however, we need to investigate how to obtain ψ_p on commodity audio devices.

3.3 Acoustic Speed Estimation

In this section, we investigate CSI measured on commodity audio devices and extend Eq. (7) to the ACF of acoustic CSI.

Channel Power: Indoor space proves to be an environment with rich diffuseness [28], where the above acoustic diffusion model applies. However, it is infeasible to directly measure the sound pressure without professional equipments. It is crucial to efficiently estimate sound pressure on common devices. In ASE, we approximate the sound pressure as the power of the sound diffusion field. Sound pressure can be considered as the equivalent of the power of CSI [29]. CSI is the linear aggregation of multi-path components, which can then be decomposed into two parts, *i.e.*, the static part Λ_S attributed to static scatterers and the dynamic part Λ_D contributed by dynamic scatterers, typically from human subjects, as shown in Fig. 4. To sum up, we can model the relationship between channel power and sound pressure as

$$\begin{aligned} G(f,t) &= |H(f,t)|^2 + \epsilon(f,t) \approx |p(f,t)|^2 + n(f,t) \\ &= \left| \sum_{i\in\Lambda_D} p_i(f,t) + \sum_{i\in\Lambda_S} p_i(f,t) \right|^2 + n(f,t), \end{aligned} \tag{8}$$

where $H(f,t)$ represents CSI at time t and subcarrier f. Practically, the static components, including the direct path between the speaker and microphone, are cancelled by removing the time average from $H(f,t)$. $\epsilon(t,f)$ and $n(t,f)$ are noise terms, with a variance of $\Sigma_N(f)$. $p(f,t)$ is the sound pressure, which is further decomposed as the dynamic and static components. $p_i(f,t)$ is the sound pressure contributed by the i^{th} scatterer, which can

Table 1. Examples of different scattering coefficients indoor [14, 61]

Room Components	Scattering Coefficient
Cabinet	0 — 0.2 — 1
4 Chairs	0 — 0.45 — 1
1 table + 3 chairs + PC	0 — 0.45 — 1
Audience	0 — 0.7 — 1
Carpets	0 — 0.3 — 1
Irregular Books	0 — 0.5 — 1

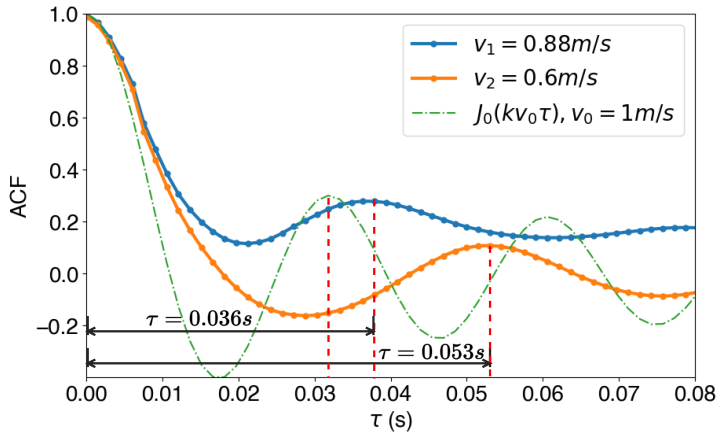


Fig. 5. ACF curves for two different speeds and the theoretical function for $v_0 = 1m/s$.

also be deemed as mutually independent of $p_j(f,t)$, $\forall i \neq j$, in sound diffusion field. Next, we model the spatial properties of $G(f,t)$ for speed estimation.

Speed Estimation Model: To learn the spatial distribution of channel power $G(f,t)$, we compute the ACF function, *i.e.*,

$$\psi_G(f,t,\tau) = \frac{\mathbb{E}\left[G(f,t)G^H(f,t+\tau)\right]}{\mathbb{E}\left[G(f,t)G^H(f,t)\right]} \triangleq \frac{\mathcal{R}_1}{\mathcal{R}_2}. \tag{9}$$

Given that $p_i(f,t)$ for any $i \in \Lambda_D$ and $p_j(f,t)$ for any $j \in \Lambda_S$ are mutually independent, and considering that sound pressures induced by static scatterers are statistically characterized by a zero mean, *i.e.*, $\mathbb{E}[p_j(f,t)] = 0, \forall j \in \Lambda_S$, it follows that any term involving a static component from Λ_S in a product will be canceled in expectation. Specifically, $\forall i \in \Lambda_D$ and $\forall j \in \Lambda_S$, we have $\mathbb{E}[p_i(f,t)p_j(f,t)] = 0$. Let $\boldsymbol{P_D}(t) = \sum_{i\in\Lambda_D} p_i(f,t)$, we have

$$\begin{aligned}\mathcal{R}_1 &\approx \mathbb{E}\left[\boldsymbol{P_D}(t)\boldsymbol{P_D^H}(t+\tau)\right] + \mathbb{E}\left[\boldsymbol{N}(t)\boldsymbol{N}^H(t+\tau)\right] \\ &= \boldsymbol{\Theta_D^H} \cdot \boldsymbol{\Psi_D} + \Sigma_N(f) \cdot \mathbf{I_n},\end{aligned} \tag{10}$$

where $\boldsymbol{N}(t) = \left[n(f_1,t), n(f_2,t), \cdots, n(f_{N_f},t)\right]$, N_f is the number of subcarriers. $\mathbf{I_n} \in \mathbf{R}^{N_\tau}$ is identity matrix, N_τ is the number of time lag for ACF. $\boldsymbol{\Psi_D} \in \mathbf{R}^{N_\tau \times N_D}$ is the ACF matrix of sound pressure incurred by dynamic scatterers, where N_D is the number of dynamic scatterers, *i.e.*, $\boldsymbol{\Psi_D} = \boldsymbol{\Psi_D}(\boldsymbol{\tau}; \boldsymbol{v})$, $\boldsymbol{\tau} \in \mathbf{R}^{N_\tau}$, $\boldsymbol{v} \in \mathbf{R}^{N_D}$. $\boldsymbol{\Theta_D^H} \in \mathbf{R}^{N_D \times N_f}$ is the frequency gain to normalize ACF matrix $\boldsymbol{\Psi_D}$. Similarly, we can compute $\mathcal{R}_2$ as

$$\mathcal{R}_2 = \boldsymbol{\Theta_D^H} \cdot \mathbf{I_{N_D}} + \Sigma_N(f). \tag{11}$$

$\psi_G(f,t,\tau)$ is irrelevant with t and can be seen as linear combination of $\mathbf{\Psi_D}$. At τ_j, $\mathbf{\Psi_D}$ is composed of $\psi_p \in \mathbf{R}^{N_D}$, *i.e.*,

$$\begin{aligned}\mathbf{\Psi_D}|_{\tau=\tau_j} &= \mathbf{\Psi_D}(\tau=\tau_j;\boldsymbol{v}) \\ &= \psi_p(v_i;\tau_j),\ i=1,\cdots,N_D.\end{aligned} \tag{12}$$

Considering a single moving target, we can assume that all the dynamic scatterers contributed by the target share approximately the same speed v_i as the walking speed v. The assumption holds in practice as for human targets, the major reflection energy from the torso dominates that from limbs. Notably, prior DFS-based work implicitly adopted similar assumptions, *e.g.*, hand gesture works usually neglect body motions [32, 93, 97]. To this end, we get $\mathbf{\Psi_D}|_{\tau=\tau_j} = \psi_p(v;\tau_j)$. With this, we can compute Eq. (9) as

$$\psi_G(f,\tau) = \widehat{\mathbf{\Theta}_D^H} \cdot \mathbf{\Psi_D}|_\tau = \widehat{\mathbf{\Theta}_D^H} \cdot \psi_p(v;\tau), \tag{13}$$

where $\widehat{\mathbf{\Theta}_D^H}$ is the broadcast aggregation of $\mathbf{\Theta}_D^H$ and $\Sigma_N(f)$.

With the above Eq. (13), we bridge the ACF of the CSI and that of sound pressure as a function of speed, for the first time, offering a completely different approach to acoustic speed estimation by calculating the ACF of the CSI measured on commodity audio devices. Practically, with the CSI time series as input, we can use the sample ACF $\tilde{\psi_G}(f,\tau)$, where a noise term $\mu(f,\tau)$ will be added to $\psi_G(f,\tau)$.

Fig. 5 shows the ACF of CSI under different speeds. As can be seen, the shape of $\psi_G(f,\tau)$ well resembles $\psi_p(v;\tau)$ as expected. Thus, to derive the speed v, we can find a reference point to align the calculated $\tilde{\psi_G}(f,\tau)$ to the theoretical $\psi_p(v,\tau)$. In our work, we use the first peak of $\psi_G(f,\tau)$, yet the first valley or their combination will also work. Let x_0 denote the reference point on $\psi_p(v,\tau)$, then the speed can be acquired to solve the optimization problem:

$$\begin{aligned} &\min_v \left| x_0 - \frac{\lambda(f)}{2\pi} v\tau_s \right|, \\ \text{s.t.}\quad & x_0 = \min_x \left\{ x : \frac{d\psi_p(x)}{dx} = 0 \wedge \frac{d^2\psi_p(x)}{dx^2} < 0 \right\}, \\ & \tau_s = \min_\tau \left\{ \tau : \frac{\partial\psi_G(f,\tau)}{\partial\tau} = 0 \wedge \frac{\partial^2\psi_G(f,\tau)}{\partial\tau^2} < 0 \right\}, \end{aligned} \tag{14}$$

where $\lambda(f)$ represents the wavelength of sound related to its subcarrier frequency f and τ_s is the counterpart of x_0 on $\psi_G(f,\tau)$, *i.e.*, the delay corresponding to the first peak. This is independent from either 2D or 3D model we choose and allows an efficient approach for speed estimation, as we only need to calculate the ACF of the CSI and localize τ_s.

Multipath Effect: The previous practice separates different paths from the CSI for modeling the multipath propagation. The multipath effect can be harmful for those approaches, as the multipath components will be hard to extract in complicated scenarios. However, our sound diffusion model does not involve decomposing each multipath, instead aggregates the speed information from all the multipaths. Intuitively, we analyze the acoustic channel from a *statistical* perspective, where the multipath information is aggregated together for speed estimation. In other words, we leverage the diffusiveness of the room to infer the speed. To this end, it is practical for realistic room environments and robust to multipath effects. At the same time, we do not need to focus on the specific material properties within the room, as our method integrates information holistically, ensuring that no single material acts as a bottleneck.

Remarks: ASE is the first to employ the *sound diffusion model* for speed estimation and *comprehensively* integrate it with the acoustic channel. The proposed model fundamentally advances acoustic speed estimation by exploring sound pressure theory and leveraging all multipath reflections. As shown in Fig. 1, Eq. (5) and (6), we leverage all

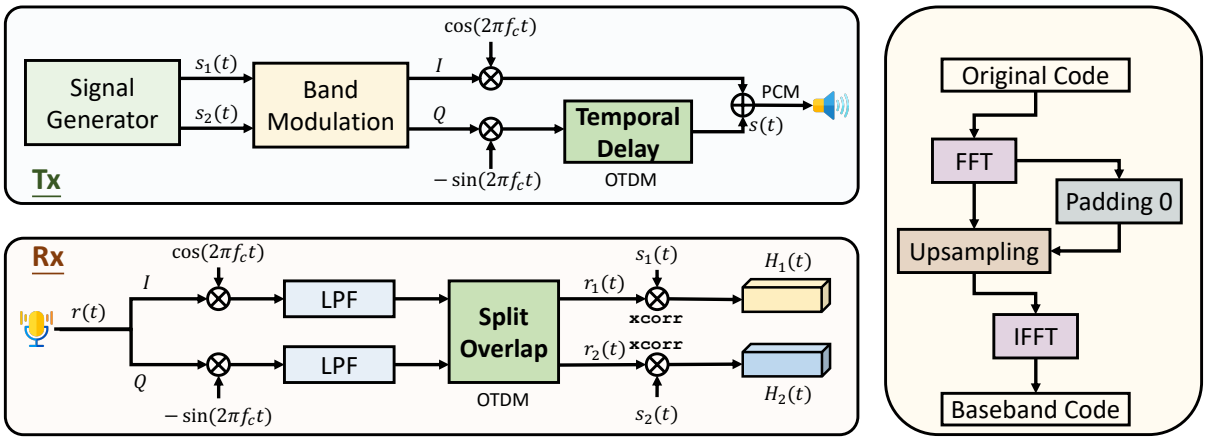


Fig. 6. Transmitter and Receiver Design of A SE. Tx: Band modulation (§4.1) and I/Q modulation (§4.2); Rx: Receiver Demodulation (§4.2) and Channel Estimation (§4.1).

the multipath components and integrate over all directions. This approach diverges completely from traditional DFS methods, marking a new foundation for acoustic speed estimation.

4 ASE DESIGN

This section presents a pipeline of novel techniques that translate our theoretical model into a practical system for robust and accurate speed estimation.

4.1 Baseband Sequence Selection

Transmitted Sequence: CSI is by default not available in acoustic sensing, and we send certain signals to probe the channel. Normally, there are three main types of waveform in acoustic sensing [9]: pure-tone, Pseudo-Noise (PN) code, and FMCW. While FMCW signals are widely used in acoustic sensing [12, 31], they are not standard impulse signals and thus introduce CIR estimation errors [38, 65]. As for the sole impulse signal, the energy of the short-time signal would fade out quickly. Therefore, we seek the PN sequence for CIR estimation in ASE. PN sequence consists of equally spaced Dirac impulses, the signs of which alternate with specific rules. It is a noise-like signal with statistical randomness. We choose the Kasami sequence [26] as our sensing waveform, mainly due to its superior orthogonality and tolerance to interference. Particularly, the mutual orthogonality of Kasami sequences is critical to our OTDM design, as detailed in §2.

Modulation to Inaudible Band: ASE uses only the inaudible sounds for sensing and modulates the sensing signals on the acoustic band of 17 kHz to 24 kHz, the pseudo-ultrasonic band supported by most commodity devices today. Since the PN code, including Kasami, has a spreading spectrum over the whole band, we should modulate the Kasami sequence to our desired band. The modulation process is shown in Figure 6. The conversion of the full-band signal into a band signal is achieved by either temporal [93] or spectral [57] interpolation. In ASE, we use the frequency domain interpolation. We perform N-point FFT to obtain the frequency domain information, where N is the original length of the sequence. Zero Padding is performed between positive and negative frequencies, creating a new sequence of length N_s. Then we perform IFFT to obtain the temporal signal. The band of the interpolated signal is thus transformed to $\frac{N}{N_s} f_s$. For a single-sequence signal, we can multiply it by the carrier signal to move the frequency band towards the central frequency f_c. In ASE, we choose $f_c = 20.25$kHz. We use 63-point Kasami sequences and modulate them to $N_s = 512$, achieving a bandwidth of 5.9kHz. The transmission signal is coded via PCM and decoded at the receiver's end.

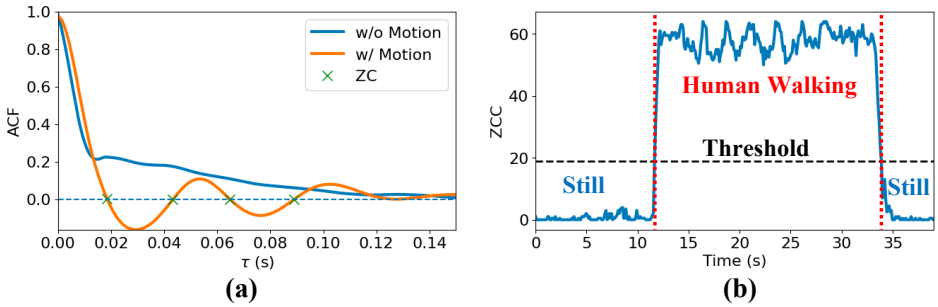


Fig. 7. Illustration of Motion Detection. (a) ACF w/ and w/o movement. ACF without motion has considerably lower ZC counts. (b) Extracted the ZCC feature to detect motion.

4.2 Transmission Scheme

In this part, we will elaborate on our transmission scheme design. We will then describe the channel estimation principles.

Single-Speaker OTDM: We stack two orthogonal signals as a complex signal and pass it into the I/Q modulator for transmission. Specifically, we can treat the orthogonal sequences $s_1(t)$ and $s_2(t)$ as the real and imaginary parts of a complex signal, *i.e.*, $s(t) = s_1(t) + js_2(t)$. To delay $s_2(t)$, we shift the sequence by padding zeros to the front of $s_2(t)$. Practically, however, a speaker can only transmit a single sequence of real signals. To transmit the resultant complex signal over a speaker, we pass it through an I/Q modulator to convert it into a real one. As illustrated in Fig. 6, we additionally multiply $s_1(t)$ and $s_2(t)$ with a carrier signal of orthogonal phase and obtain the modulated real signal: $s(t) = s_1(t)\cos(2\pi f_c t) - s_2(t)\sin(2\pi f_c t)$. With such modulation, the two components remain orthogonal to each other, while the modulated signal can be transmitted through a single speaker. At Rx, the received signal is first demodulated by the carrier signal with a low-pass filter (LPF). We can then separate the overlapped signal and estimate the CSI for each sequence at the receiver's side, as shown in Fig. 6. By doing so, we can obtain two separated signals $r_1(t)$ and $r_2(t)$. Then, channel estimation will be performed as follows.

Channel Estimation: To capture the channel properties, we generate the Kasami sequence $s(t)$ of length N_s as the probe signal and transmit it over a speaker. The signal interacts with the environment before arriving at the microphone, where it undergoes scattering in the sound diffusion field. At the receiver end, we acquire $r(t)$ and separate it into $r_1(t)$ and $r_2(t)$, respectively. The CIR $h(t)$ is then estimated as the correlation between them, *i.e.*, $h_i(t) = r_i(t) * s(t), i \in \{1, 2\}$. Then CSI $H_i(f, t)$ is obtained by transforming $h_i(t)$ into the frequency domain via FFT. $H_1(f, t)$ and $H_2(f, t)$ will be then merged using the OTDM scheme to acquire the boosted channel estimation $H(f, t)$, as illustrated in §2.

Synchronization between Microphone and Speaker Pairs: While the microphone and speaker are connected to the same controller and should theoretically be temporally aligned, hardware imperfections can introduce small time discrepancies between the Tx and Rx pair. Traditional FMCW tracking algorithms usually require precise synchronization to correctly determine the start time of IF signals. However, our approach minimizes the impact of synchronization issues. At a high level, any discrepancies manifest as phase offsets in the asynchronized Channel State Information (CSI). This is not a significant problem for ASE because we do not introduce blank intervals and instead focus on extracting channel power to determine speed. Specifically, the asynchronous CSI would be

$$H_d(f) = H(f) \cdot \exp(-j \cdot 2\pi f \cdot \tau) \tag{15}$$

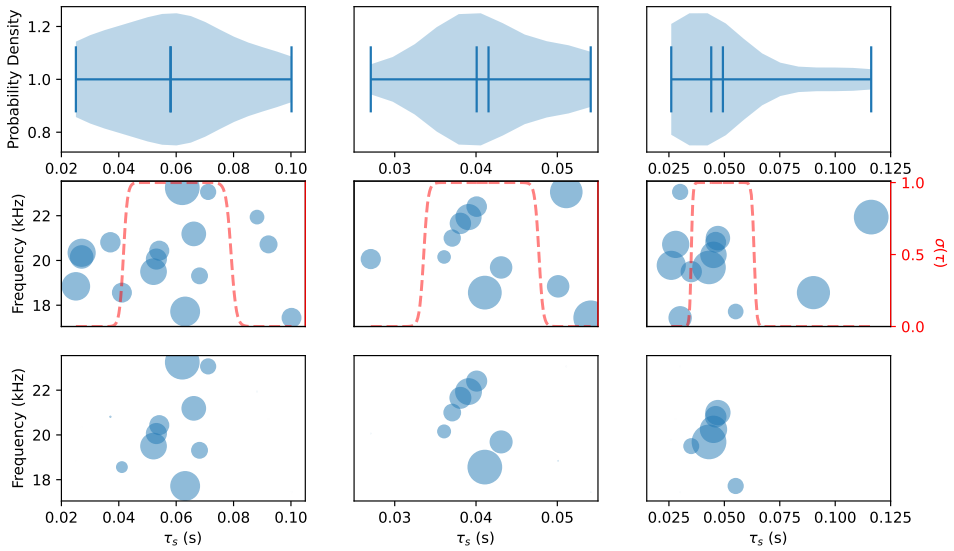


Fig. 8. Peak prominence weighting examples. Top: Violin plots of the distribution of the delays of the first peak. Middle: Original weight distribution, where the circle size indicates the weights. The red dashed line represents the fitted $\sigma(\tau)$. Bottom: Adjusted prominence weights.

where τ is the temporal offset. And the channel power, as denoted in (8) can be written as

$$\begin{aligned} G_d(f) &= |H_d(f)|^2 + \epsilon(f,t) \\ &= |H(f)|^2 + \epsilon(f,t) = G(f), \end{aligned} \tag{16}$$

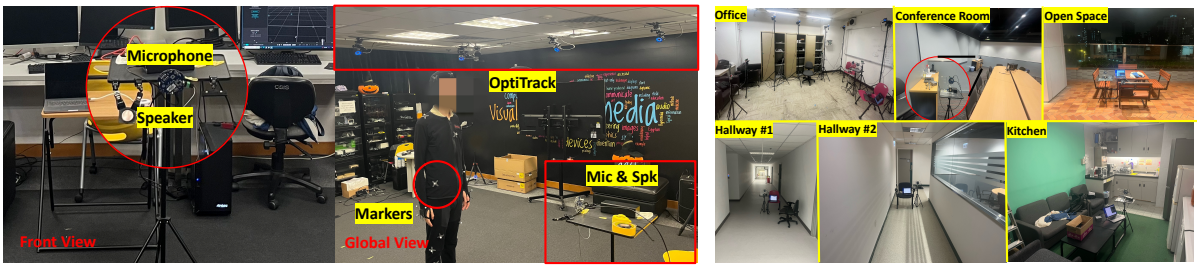
which is equivalent to the synchronized version. Therefore, the temporal offsets do not affect the magnitude of the channel power, ensuring reliable speed acquisition even without precise synchronization.

4.3 Speed Estimation

Now we present how to estimate speed from the CSI time series, given the model described in §3. We devise a robust motion indicator to detect movements. We perform speed estimation when motion is detected. We further enhance frequency diversity to achieve robust speed estimation.

Motion Indicator: A robust motion indicator can help ensure only valid measurements are used for speed estimation, reducing the system overhead while improving estimation accuracy. Therefore, we propose a novel motion indicator called Zero Crossing Count (ZCC) in ASE. As described in Sec. 3, ACF of channel power $\psi_G(f,\tau)$ is the function of moving speed v. When there is no motion, the ACF curve would be overall flat, resulting in no obvious peaks or zero crossings. In contrast, in the case of motion, the ACF will exhibit peaks and valleys, resulting in significantly more zero crossings, as shown in Fig. 7. To identify motion, we can perform Zero Crossing (ZC) analysis on the ACF and count dominant ZCs. Fig. 7 shows an example of the ACF and ZCC values, indicating a clear difference between motive and still states. Hence, we can find a threshold to reliably detect motion from the ZCC values.

Weighted Subcarrier Combining: Proposed model accounts for all multipath components for speed estimation. Recall Eq. (13), however, the theoretical model estimates speed from a single subcarrier. Given multiple subcarriers available, we can further embrace frequency diversity to boost speed estimation. Instead of performing speed estimation on each subcarrier and fusing the estimates, we propose to properly combine $\psi_G(f,t)$ calculated



(a) Overall Setting. (b) Different Environments.

Fig. 9. Experiment Setting.

on each subcarrier to enhance the SNR of the speed signal. Formally, we aim to obtain the boosted ACF as $\hat{\psi}_G(\tau) = \sum w(f)\psi_G(f,\tau)$, where

$$\begin{aligned} \max_{w} \quad & \hat{\psi}_G(\tau_s) = \sum_i w(f_i)\psi_G(f_i, \tau_{s_i}), \\ \text{s.t.} \quad & \sum_i w(f) = 1, w(f) \geq 0, \forall f \in \mathbb{F}, \end{aligned} \tag{17}$$

where $\mathbb{F}$ denotes the subcarrier set. τ_s is the first peak of $\hat{\psi}_G(\tau_s)$. The key is to find optimal weights $w(f)$ for effective combining. The desired goal is to obtain prominent peaks in the ACF for accurate and robust speed estimation. Therefore, we propose a novel weighting method based on the original peak prominence. The idea is that larger weights should be given to the ACF that features more prominent first peaks. Denote $\kappa(t)$ as the prominence of the detected first peak in $\psi(f,\tau)$. Intuitively, we can simply set $w(f) = \kappa(t)$ and combine all subcarriers. However, as shown in Fig. 8, we observe that some ACF may have prominent peaks largely deviated from the centered delay of the majority of peaks. This would enlarge the noise and lower the combined first peak. To this end, we design a compensatory weight decaying algorithm. Specifically, we calculate the mean and median lags of all the identified peaks, and their lower or upper quartiles. We keep the prominence for peaks whose delays between the mean and median untouched, reduce the weights for those between the mean/median and the quartiles, and discard those falling outside the quartiles. Therefore, we use a sigmoid function as the decaying curve: $\sigma(\tau) = 1/(1+\exp(-a(\tau-b)))$, where τ is the peak value and a and b are two parameters that can be fitted. The adjusted prominence values are used as the weights for calculation in Eq. (17).

Frequency Alignment: The weighted combining will enhance the speed signals and reduce the noise, only if the signals are synchronized. However, significant subcarrier frequency variations misalign the ACFs across subcarriers, particularly the first peak, for the same speed. Take two subcarriers on 20 kHz and 24 kHz as an example. Considering a speed $v = 0.5m/s$, the peaks will appear at a delay $\tau = x_0\lambda(f)/2\pi v$. The wavelength $\lambda(f)$ depends on the carrier frequency, and will be 1.715 cm for 20 kHz and 1.429 cm for 24 kHz, leading to two considerably different delays for the first peaks in the respective ACFs. Therefore, we need to eliminate the subcarrier frequency differences and align all the ACFs. We scale the ACF in the lag domain and align them with respect to the first peaks. Particularly, we choose a virtual frequency as a reference, denoted as f_{ref}, and scale the ACF on all subcarriers to the same f_{ref}, *i.e.*, $\psi(f,\tau) \rightarrow \psi(f_{ref},\tau')$. Recall Eq. (14), we obtain the aligned first peak as,

$$v(f,\tau_s) = \frac{x_0 c}{2\pi f \tau_s} = \frac{x_0 c}{2\pi f_{ref}\tau_s'} = v(\tau_s'), \tag{18}$$

where $\tau_s' = \frac{f}{f_{ref}}\tau_s$ represents the aligned first peak. Practically, it is implemented by interpolation. Then the weighted combining can be done on the aligned ACF.

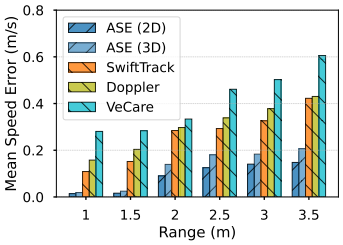


Fig. 10. MSE-DW.

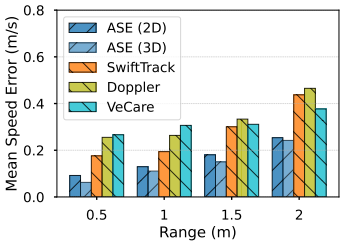


Fig. 11. MSE-CW.

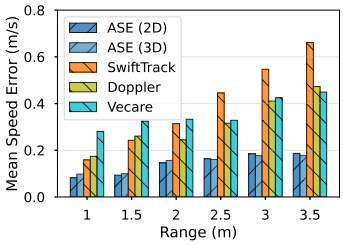


Fig. 12. MSE-RW.

5 EVALUATIONS

5.1 Implementation

Hardware: We implement ASE with programmable smart speaker prototype, *i.e.*, a UMA-8-SP USB microphone [5], and AS05308AS-R speaker[2]. Notably, we only **use one microphone channel** in our experiments. The microphone and the speaker are co-located. The hardware is connected to a power hub and MacBook Pro 2021.
Software: We implement the algorithms of ASE in Python and Matlab. Specifically, we develop a sound playing and recording program with Python and implement the pipeline of processing algorithms with MATLAB. We apply a 1-s window to compute ACF with a step of 0.1s. We perform motion detection, based on ZCC, to determine whether a user is moving and perform speed estimation when motion is detected. Before weighted subcarrier combining, we also perform outlier detection to sift out some abnormal ACFs, which occasionally appear as either a near-linear trend or a zig-zag spike.
Data Collection: As shown in Fig. 9, we mainly conduct our experiment in a 4×4m room equipped with the OptiTrack [4], which serves as the ground truth. It is a camera-based precise motion capture system. We derive speed from the solved whole-body skeleton coordinates from OptiTrack. To calibrate and fuse the data across cameras, participants wear specialized clothes with visual markers. The experiment is conducted in a regular office environment and is subject to various background noises, including footsteps in the corridor and the sound of the room air conditioner. During the experiment, only the test participant is walking while others in the room are stationary. We have gained IRB approval from our university for data collection. In total, we have collected over 700 minutes of moving data traces. We evaluate the performance of ASE under various settings.
Metrics: We use Mean Speed Error (MSE) between the estimated speed and the ground truth and detection rate (*i.e.*, how often ASE can reliably detect a speed) as the main evaluation metrics. Our results demonstrate the effectiveness of our approach in accurately estimating speed and detecting movement in a daily living environment. We present detailed results below.

5.2 Overall Performance

Our evaluation of ASE considers the coverage and orientation of the audio devices by involving three different scenarios: direct walk (*i.e.*, a straight line towards the devices), circle walk (*i.e.*, around the devices with varying radii) and random walk (*i.e.*, zig-zag walk, back-and-forth walk and run, *etc*).
Direct Walk (DW): We place the device at the center of one wall in the room. Participants are asked to walk towards the audio device from 1m to 3.5m. We then calculate the mean speed error in Fig. 10 and detection rate in Fig. 13. Our results demonstrate that ASE can estimate speeds reliably up to a distance of 3.5 m, with an average speed error of 0.08 m/s. When the object is close, the mean speed error is 0.016 m/s, which increases to 0.14m/s at 3.5m. Our system achieved an average detection rate of 99.0% in the case of direct walking, with a slight degradation to 95.0% at a distance of 3.5m.

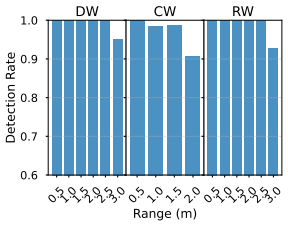


Fig. 13. Detection Rate.

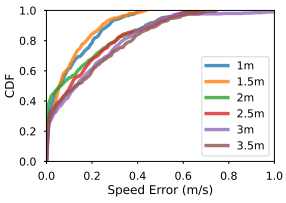


Fig. 14. CDF vs. Distances.

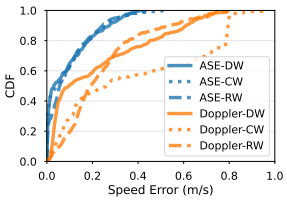


Fig. 15. CDF vs. Walk Means.

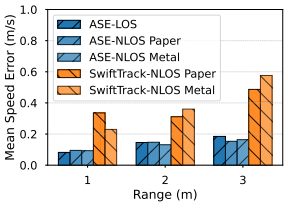


Fig. 16. MSE of NLOS Case.

Circle Walk (CW): We conduct experiments in a circle-walk scenario. Participants were instructed to walk around the device at varying radii between 0.5m and 2m. The mean error and detection rate are depicted in Fig. 11 and Fig. 13 respectively. While DFS often struggles to estimate speed on a circular path, ASE achieves a mean speed error of 0.13m/s within 1 meter, 2 times lower than that of DFS.

Random Walk (RW): We comprehensively evaluate the performance of ASE in random walk scenarios. To do so, we let participants move freely within the experimental room, and summarize the results in Fig. 12 and Fig. 13. As seen, our approach achieved an average detection rate of 98.8%, with a perfect detection rate of 100% within 3m. The total average mean error is 0.13m/s, which lowers to 0.08m/s when participants are walking within a shorter range. We also portray the CDF in Fig. 14, which shows a 90%-tile error of less than 0.2m/s. The results demonstrate ASE's robust performance in realistic environments for practical applications.

2D vs 3D model: We evaluate both 2D and 3D models in different scenarios, as illustrated in Fig. 10, Fig. 11 and Fig. 12. The 3D model shows a lower mean speed error of 2.2 cm/s in circle walk scenarios compared to the 2D model. In random walk scenarios, the 3D model is more effective at longer distances. This evaluation underscores the feasibility of both models and their applicability to different scenarios.

5.3 Comparative Study

We focus on acoustic-based speed estimation and compare ASE with three baselines: 1) SwiftTrack [97]: An acoustic speed algorithm for practical fast motion tracking, 2) DFS: A widely used method, implemented based on CAT [37], and 3) VeCare [96]: A recent work of a statistical acoustic sensing framework for child presence detection. The former two are DFS-based speed estimation approaches. For a fair comparison, we transmit modulated Zadoff-Chu (ZC) signals as in [97] with the same frame length, *i.e.*, 10ms. We use unmixed CSI and extracted the frequency shift peak in the DFS spectrum for Doppler. We compare baselines under scenarios in §5.2, and also test the Non-Line-of-Sight (NLoS) condition. We apply the same settings as VeCare. We will first compare SwiftTrack and Doppler and leave VeCare as the last part.

Different Walk Means: We compare ASE with SwiftTrack and Doppler in three means: direct walk, circle walk and random walk. Our results are shown in Fig. 10, Fig. 11 and Fig. 12 respectively. While ASE achieves a median error of 0.08m/s in direct walking, SwiftTrack and Doppler yield errors of 0.26m/s and 0.30m/s, respectively. Moreover, ASE significantly surpasses SwiftTrack and Doppler by 68.9% and 101% for circle walk and by 176.7% and 146.2% for random walk. We further depict the CDF of ASE against Doppler in Fig. 15. It shows DFS exhibits poor performance in circle walk while ASE retains consistent performance across various walking means. It is plausible for worse performance of SwiftTrack against Doppler in random walk, given its speed estimation algorithm is tailored for gesture tracking.

Sensing Distance: A unique advantage of ASE is the enlarged sensing coverage, and we verify this by evaluation over different distances, as illustrated in Fig. 10, Fig. 11 and Fig. 12. While the error generally increases with respect to distances for all the methods, ASE consistently outperforms the baseline methods at all distances, with more significant performance gains at larger distances. As shown in Fig. 10, ASE achieves a mean speed error less than 0.15m/s at 3.5 m, while SwiftTrack and Doppler soar to 0.42m/s and 0.43m/s, respectively. Overall, both

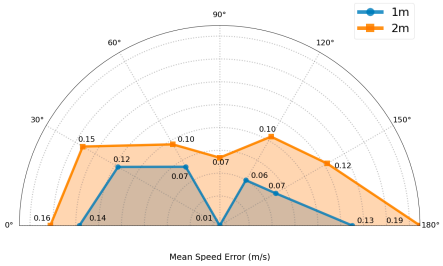


Fig. 17. Impact of different directing angles.

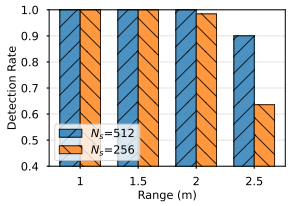


Fig. 18. Impact of sequence lengths.

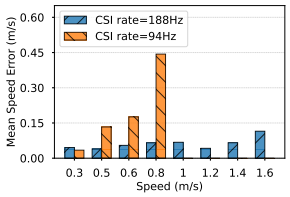


Fig. 19. Impact of different CSI rates.

SwiftTrack and Doppler degrade significantly beyond the distance of 1.5m. ASE obtains superior performance because our model leverages all multipath signals, which get more observations across the room and expand the range to accommodate room-scale sensing.

LoS vs. NLoS: We assess our system in NLoS settings with obstructions like a metal panel or a paper bag in front of the speaker and microphone. Despite sound's known vulnerability to absorption [59], ASE remains effective in NLoS scenarios, exhibiting an average error rate of 0.13m/s and a standard deviation of merely 0.003m/s, as illustrated in Fig. 16. In contrast, the performance of SwiftTrack drops by 14.4% and 11.2% under metal and paper obstructions, respectively. We benefit from the effective use of multipath reflections, which is particularly helpful in NLoS scenarios.

Comparison with VeCare: VeCare includes a statistical framework for presence detection, which mentions speed components. We implement the algorithm and evaluate in three scenarios. As shown in Fig. 10, Fig. 11 and Fig. 12, the performance of VeCare is much worse than our ASE. Notably, the performance of ASE is 149.9% better than VeCare in random walk scenarios. VeCare, which extrapolates Wi-Fi sensing methods without comprehensive modeling, is not targeted for speed estimation and does not include specific designs. These results highlight the novelty and superiority of our speed estimation techniques and theoretical model.

5.4 Benchmark Study

In this section, we present a detailed analysis of various impact factors in ASE. We begin by investigating the impact of different movement directions on the accuracy of our approach. Next, we examine the performance of our system extending to dual speakers to conduct the experiment. We also study the impact of various acoustic factors, including sound level, sound amplitude, and interference sources.

Different Orientations: We evaluate the impact of different walking orientations at various angles with respect to the audio devices. We evaluate different angles, from 0° to 180° with an increase of 30°, where 90° is defined as facing directly towards the device. As illustrated in Fig. 17, the smallest speed error is observed at 90°. The errors increase as the orientation shifts away from 90° towards both sides, yet are still acceptable even when at 0° or 180°.

Sound Level: We modulate the transmission signal to the inaudible band to minimize interference with human hearing for daily use as detailed in Sec. 5.1, rendering the sound nearly inaudible. To evaluate the impact of different sound levels, we transmit signals at sound pressure levels of 33.4dB, 35.3dB, 36.7dB, 38.2dB, and 41.7dB, respectively. As shown in Fig. 21, the results indicate that the mean speed error decreases as the sound volume increases, as does the detection rate. Notably, our system's maximum sound pressure of 41.7dB is significantly lower than most previous acoustic sensing studies [39, 68, 93, 96]. Despite the sound leakage problem on commodity devices [30], the sound noise of our system can be fairly neglected in commonly used settings.

Amplitude of Kasami Sequence: Besides the volume of the device, the sound power is also determined by the amplitude of the original sequence. We hereby experiment with different amplitudes, shown in Fig. 22. This is not

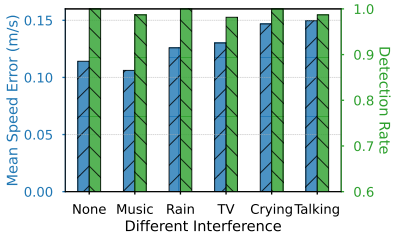


Fig. 20. Impact of different interference sources.

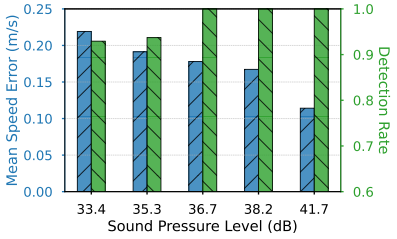


Fig. 21. Impact of different sound volume levels.

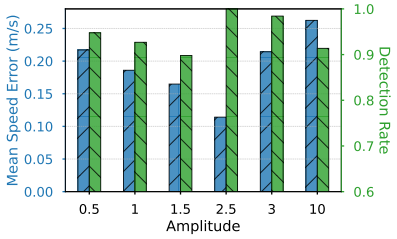


Fig. 22. Impact of amplitudes of original sequence.

the final amplitude of the transmitted signal, as it must go through band modulation and PCM before transmission. As seen, the performance decreases first and then increases as amplitudes increase, and an amplitude of 2.5 strikes the best for both speed accuracy and detection rate. Amplitudes smaller than 2.5 will result in low sound power and thus large errors, while overlarge amplitudes will lead to distortions of the signal [30] due to the hardware limitation. In our case, sounds with amplitudes larger than 2.5 would be clipped off. Hence, we choose 2.5 as the amplitude of the original PN sequence.

Interferences: We also conduct experiments to evaluate the performance under different interference sources. We choose common sources of interference in indoor environments, including music, rain, TV broadcasting, baby crying, and human talking, illustrated in Fig. 20. This indicates that our system is robust to noise and interference, and can accurately estimate the speed even in the presence of interference.

Sequence Lengths: We then study the influence of different sequence lengths N_s. As discussed above, the sequence length impacts the sensing coverage and CSI rate. When the sequence length is reduced, the CSI rate would increase, but the sensing coverage would diminish simultaneously. To investigate it, we conduct experiments using sequence lengths of 512 and 256 and mainly examine the detection rate to study the coverage. As shown in Fig. 18, using both lengths achieves great performance for small coverage, *e.g.*, within 1.5 m. However, the detection rate for the sequence length of 256 drops significantly at larger distances of 2m and 2.5m. Particularly, at a distance of 2.5m, the detection rate drops to 63.6% for $N_s = 256$ while the rate still remains at 99% for $N_s = 512$. ASE defaults to $N_s = 512$ to strike a balance between sensing coverage and CSI rate.

CSI Rate: We now study how CSI rates will impact the performance for different speeds. To control the moving speed, we use a programmable rail track with a plate on it. We test with speeds of 0.3m/s to 1.6m/s, and compare the performance with a CSI rate of 94 Hz and 188 Hz, respectively. We apply the OTDM scheme to achieve the CSI rate of 188 Hz. As shown in Fig. 19, a low CSI rate cannot measure large speeds and will produce large errors. The mean errors for speeds of 0.6m/s and 0.8m/s with a CSI rate of 94Hz are 0.18m/s and 0.44m/s, respectively. Comparatively, the corresponding errors with that of 188Hz are 0.055m/s and 0.066m/s, respectively. Notably, our system maintains a relatively low average error of 0.11m/s even at 1.6m/s. Overall, the results clearly demonstrate the need for a sufficient CSI rate for speed estimation and justify the effectiveness of the proposed OTDM scheme. In ASE, we can hold speed up to 1.6m/s using single-speaker, adequate for capturing indoor walk speed.

Different Subcarrier Combining Weight: We test different subcarrier combining weight methods. Our results show that using prominence as the weight leads to a mean speed error of 0.13m/s, whereas using motion statistics simply [96] results in an error of 0.19m/s. Besides, we compare different weight decaying algorithms. We observe a 15% decrease in mean speed error when applying sigmoid weight decay. These results confirm the effectiveness of our subcarrier combination weight in ASE.

Different Device Height: To investigate the impact of varying the height of the transmitter and receiver pair on system performance, we conducted experiments using ASE with different hardware heights, as shown in Figure 23. By utilizing a tripod, we controlled the height range from 80 to 115 cm. The results indicate that ASE retains

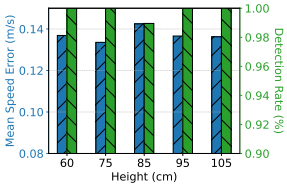


Fig. 23. Different Height.

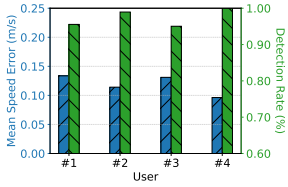


Fig. 24. Different Users.

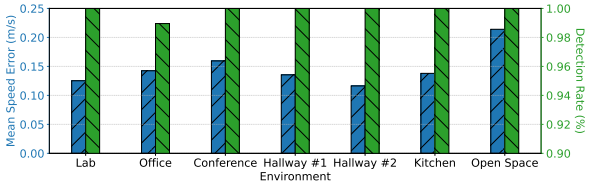


Fig. 25. Different Environment and Locations.

consistent accuracy under changes in heights, with a mean of 0.13 m/s and a standard deviation of 0.002 m/s. Despite these variations, the average detection rate remains high at 99.8%. This suggests that the height of the devices does not significantly affect the overall performance of ASE.

Different Users: We also examine how ASE performs for different users. We recruit four volunteers and let them walk freely with varying speeds, orientations, and trajectories. The results are shown in Fig. 24. Our approach achieves 80% tile error of 0.2m/s. The detection rates of three users are 95.55%, 98.95% 95.03% and 100%, respectively. Accordingly, the mean error of four users is 0.12m/s with an average detection rate of 97.4%. These findings demonstrate the effectiveness and accuracy of ASE in estimating the speed of human movement in indoor environments. Our approach achieves high detection rates and low error rates across multiple users with varying movement patterns and speeds. These results have important implications for the use of ASE in real-world applications, such as surveillance, security, and healthcare, where accurate estimation of human movement speed is critical.

Different Environments and Locations: To evaluate ASE's robustness across various settings, we conduct experiments in multiple environments, as illustrated in Fig. 9b. We deploy devices in a lab, office, conference room, hallways, kitchens, and an open space. These locations differ in geometry, building materials, floor heights, and ambient noise levels. Except for the lab and office, we leverage the depth camera to get the range information and then convert it to speed. During the tests, participants walk randomly within each environment. We compute the mean speed error and detection rate for all scenarios, as summarized in Fig. 25. In indoor environments, the average speed error is 0.136 m/s with a standard deviation of 0.01 m/s, which aligns with previous benchmark results. The corresponding detection rate averages 0.998, with a standard deviation of 0.003. These results demonstrate that ASE consistently performs well across varied indoor settings, proving its robustness. We also evaluate ASE in an open space, where the mean speed error increases to 0.21 m/s, although the detection rate remains high. This outcome is expected given the relative scarcity of reflectors in open environments, which can impact speed estimation accuracy. Nevertheless, the consistently high detection rate highlights the robustness of our motion detection algorithm. Future work can explore the development of advanced speed estimation methods tailored for open areas.

Runtime Analysis: In this section, we evaluate the runtime performance on various platforms. We first test the algorithm's performance on a MacBook 2021 (with Apple M1 Pro chip), analyzing a 10-second audio segment using a 1-second sliding window with a 0.1-second step. The algorithm demonstrates exceptional efficiency, achieving a total runtime of 2.299 seconds, well below the 10-second data duration, which confirms its capability for real-time processing. The decoding of OTDM takes 0.0999 seconds, and the ACF part finishes in 0.831 seconds. Peak prominence and weight computations are notably fast at 0.098 seconds and 0.067 seconds, respectively. Additionally, memory usage is a modest 387.56 MB, well within the capacity of modern systems, making this algorithm highly suitable for efficient, real-time processing. Additionally, we deploy our system on an embedded device, *i.e.*, Raspberry Pi Compute Module 4, which is equipped with 4-core ARM-v8 processors. For the same 10-second audio segment with identical windowing parameters, the algorithm processes the respiration data with an overall runtime of 3.912 seconds, still significantly faster than the data duration, ensuring real-time performance even on resource-constrained hardware. The decoding takes 0.384 seconds, and ACF calculations

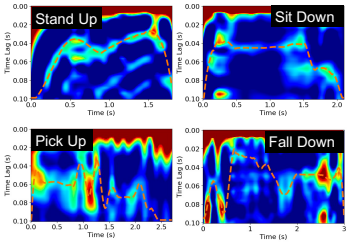


Fig. 26. Speed profile of different human activities.

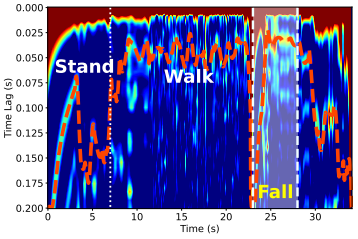


Fig. 27. Speed profile of continuous activities.

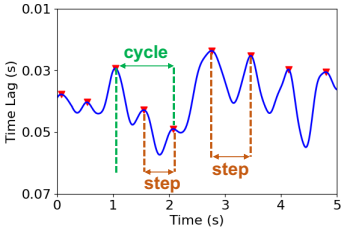


Fig. 28. Gait step and Cycle.

total 0.440 seconds. Memory usage on the Raspberry Pi is optimized at 318.57 MB, demonstrating the algorithm's ability to operate efficiently within the limited resources of an embedded system. These results highlight the algorithm's robustness and adaptability, enabling real-time speed analysis across both high-performance and embedded platforms.

5.5 Case Study

In this section, we present a case study on three applications of human walking speed: activity recognition, fall detection, and gait recognition, to showcase various ASE-enabled applications.

Human Activity Recognition: Speed profiles act as the core of human activity recognition [16, 44, 50, 54, 85]. As shown in Fig. 26, the speed profiles (the dashed lines) generated by ASE can effectively distinguish different daily activities such as standing up, sitting down, and picking up objects. For instance, the speed profile of standing up shows a slower increase compared to sitting down. Picking up an object, however, forms a 'valley' in the profile, reflecting the bending and rising motion. Fig. 27 provides an example of recognizing different activities through continuous monitoring. We keep it as a future direction to recognize different activities with ASE deployed on commodity smart speakers.

Fall Detection: Fall detection is a vital yet challenging task. Passive speed estimation provides an effective way to detect falls [22, 25, 35, 99], mainly because falls induce a distinct speed pattern compared to normal daily activities. As shown in Fig. 26 and Fig. 27, a fall induces a unique pattern in the ACF matrix. The speed changes rapidly from a lower rate to a significantly higher rate within a brief period, yielding a large instantaneous acceleration. Following speed suggests further movement on the ground, before eventually decreasing to zero.

Based on these observations, we employ deep learning for proof-of-concept fall detection. As the ACF matrix contains speed information, we leverage pretrained MobileNet-v2 [53] as the ACF matrix encoder to classify the fall events. To increase data availability, we develop a recording platform that connects multiple off-the-shelf audio devices to a server via MQTT [36]. In total, we have 305 samples, with 175 representing fall events. We achieve the F-1 score of 0.92, demonstrating the effectiveness of ASE for feature extractions.

Gait Analysis: Gait analysis is important in healthcare, especially for patients with Parkinson's disease [7, 13, 87]. Intuitively, the walking speeds estimated by ASE allow us to perform gait analysis. As a proof-of-concept, we recognize gait steps and cycles by detecting the peak in the speed profile, as shown in Fig. 28. Our analysis of 10 traces reports an average gait cycle time of 1.00s with a variance of 1.2ms, demonstrating the potential of ASE for gait applications.

6 Discussions

We conclude some discussions and limitations for further exploration.

Multiple Speeds: ASE assumes the human body shares the same speed v for walk speed estimation. The rationale behind our assumption is that we target room-scale sensing with relatively large coverage of several meters. In

our proposed cases, the body speed dominates limb speeds. Moreover, a similar assumption is also implicitly adopted by many other works [32, 43, 97], *i.e.*, hand-gesture works usually neglect body motions. As the next step, we are actively exploring multi-speed ASE. Potentially, Eq. (12) suggests the potential to decompose the ACFs to account for varying speeds, which indicates an opportunity for future research to refine speed estimation by considering multiple speed scenarios.

Multi-target Scenario: Current ASE is not designed for multi-person scenarios. In most cases, indoor walking speed downstream tasks, such as fall detection and gait monitoring, involve a single person. Audio devices enable interaction in real-time and alert when abnormal conditions are detected. Therefore, ASE can already foster many applications. Yet it is also potentially extend it to multi-person sensing in the future, *e.g.*, by separating multipath signals for different ranges. We also look forward to future work addressing the issue of multi-person movement.

Multiple Speakers: In ASE, we use OTDM with a single speaker and two channels, offering practical room-scale sensing. While this setup is common, OTDM can be extended to multiple speakers with advanced modulation scheme. Future work can explore OTDM+OFDM design for higher CSI rates.

Wi-Fi Speed Estimation: Numerous studies have utilized Wi-Fi for velocity capture [33, 43, 79]. However, the acoustic method offers several distinct advantages. Typically, commodity acoustic devices come with a co-located microphone and speaker, allowing sensing with a single device, *e.g.*, Amazon Echo Dot, without any hardware/firmware changes. In contrast, Wi-Fi sensing generally necessitates two separate transceivers and relies on specific chipsets (e.g., Intel 5300, Qualcomm Atheros NICs) that require firmware adjustments. In addition, acoustic devices bring side benefits by inherently making a voice interface available [72, 84, 95, 100]. For instance, besides running acoustic fall detection, a smart speaker can facilitate emergency calls directly. Therefore, it is imperative to enable accurate speed estimation for acoustic sensing. This work is orthogonal to existing Wi-Fi-based approaches.

Phase-based Estimation: Recent studies [33, 97] have employed phase difference to address the challenges posed by high speed scenarios. They implicitly acquire speed by calculating acceleration or displacement. In contrast, our proposed framework, OTDM plus sound diffusion model, directly estimates speed without the need for further transformations. While our current approach utilizes channel amplitude, it is promising to extend this design to incorporate phase-based speed estimation, which we intend to explore in future work.

7 Related Works

Acoustic Sensing: Acoustic sensing has been extensively explored recently and realize multiple applications, including indoor localization [19, 34, 38, 40], fine-grained gesture tracking [32, 37, 42, 57, 67, 75, 88, 92, 97], vital signs monitoring [27, 31, 47, 66, 68, 69, 73], silent speech enhancement [18, 56, 89], sound source localization [15, 55, 63, 64, 71], and communication [11, 60, 98]. The latest work VeCare [96] introduces statistical acoustic sensing for presence detection by extrapolating statistical WiFi sensing. ASE significantly differs from VeCare in multiple important aspects: 1) VeCare neither constructs the comprehensive model nor considers the fundamental differences between acoustic and electromagnetic waves. Conversely, ASE integrates a novel sound diffusion model with the acoustic channel for speed estimation for the first time. 2) VeCare mainly targets motion detection and breathing rate estimation, but does not address speed estimation (despite limited preliminary explorations). Differently, ASE not only establishes the theoretical model but also builds practical techniques for location-independent and large-coverage speed estimation, achieving remarkably better performance. 3) Additionally, we propose the first-of-its-kind OTDM design to increase the CSI rate, enabling previously challenging high speed estimation in a unique manner. Other than the key differences, we believe ASE complements VeCare towards a comprehensive framework for motion detection, vital sign monitoring, and speed estimation, in a distinct paradigm differing from the prevalent literature in acoustic sensing.

Speed Estimation: Speed is vital information in human sensing, and has been exploited in various applications like fall detection [24, 35, 45, 76], interactive games [49], gait monitoring [62, 79, 87], and localization [40, 48]. Speed estimation, however, is a challenging and enduring task. Camera-based approaches, such as VICON [6], require professional devices and specialized calibration, which cannot be used in ubiquitous settings. In wireless sensing, speed is generally derived using DFS, regardless of Wi-Fi [48, 58, 83], acoustic [37, 37, 55, 57, 71, 92, 97], or mmWave radar [8, 81] signals. Recently, some work [33, 97] also acknowledge the problems of low CSI rate and have proposed using phase difference to estimate acceleration or displacement, indirectly inferring speed. However, they do not jump out of the DFS framework, which inherently limits observations to radial speed components. Moreover, inferring speed from displacement or acceleration introduces additional transformations, thereby increasing cumulative error [37]. In contrast, ASE introduces a novel and explicit speed estimation framework that statistically aggregates all paths by modeling the sound diffusion field, thereby overcoming these limitations. Despite the advances in statistical wireless sensing [79, 94], sound waves and EM waves differ in nature, and acoustic speed estimation incurs different challenges. We build the acoustic diffusion speed estimation model plus OTDM design, a new way of acoustic speed estimation.

8 Conclusions

In this paper, we present ASE, an end-to-end system for acoustic speed estimation. We identify two research problems in the acoustic speed estimation: insufficient CSI rate to capture the speed and the oversimplified signal model that fails to capture the entire speed. To this end, we propose a novel OTDM scheme to achieve a higher CSI rate for speed estimation. Inspired by the sound diffusion field, we establish a comprehensive theoretical model that enables full speed estimation from the spatial correlation of sounds. We extensively evaluate ASE on commodity devices, which achieves an average accuracy of 13 cm/s for normal walking speed estimation. Overall, the proposed ASE goes beyond the DFS-based paradigm and can open up new directions in acoustic sensing.

Acknowledgments

We sincerely thank all the anonymous reviewers for their valuable feedback throughout the submission process. We also wish to express our gratitude to the staff members at HKU who generously provided the initial space for our experiments. In addition, we are grateful to the lab members who assisted with proofreading the manuscript. This work is supported in part by NSFC under Grant No. 62222216, Hong Kong RGC GRF under Grant No. 17212224 and Healthy Longevity Catalyst Awards under Grant No. HLCA/E-712/22.

References

[1] 2021. X4M03 – laonuri.com. https://www.laonuri.com/product/x4m03/. https://www.laonuri.com/product/x4m03/
[2] 2022. Speakers & Receivers | AS05308AS-R. https://puiaudio.com/product/speakers-and-receivers/AS05308AS-R. https://puiaudio.com/product/speakers-and-receivers/AS05308AS-R
[3] 2023. IWR1843 data sheet, product information and support | TI.com. https://www.ti.com/product/IWR1843. https://www.ti.com/product/IWR1843
[4] 2023. Motion Capture Systems. http://optitrack.com/index.html. http://optitrack.com/index.html
[5] 2023. USB Audio Streaming: UMA-8-SP USB mic array. https://www.minidsp.com/products/usb-audio-interface/uma-8-sp-detail. https://www.minidsp.com/products/usb-audio-interface/uma-8-sp-detail
[6] 2023. Vicon | Award Winning Motion Capture Systems. https://www.vicon.com/. https://www.vicon.com/
[7] MD Akhtaruzzaman, Amir Akramin Shafie, and Md Raisuddin Khan. 2016. Gait analysis: Systems, technologies, and importance. *Journal of Mechanics in Medicine and Biology* 16, 07 (2016), 1630003.
[8] Alejandro Blanco, Pablo Jiménez Mateo, Francesco Gringoli, and Joerg Widmer. 2022. Augmenting mmWave localization accuracy through sub-6 GHz on off-the-shelf devices. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*. 477–490.
[9] Chao Cai, Rong Zheng, and Jun Luo. 2022. Ubiquitous acoustic sensing on commodity iot devices: A survey. *IEEE Communications Surveys & Tutorials* 24, 1 (2022), 432–454.

[10] Huijie Chen, Fan Li, and Yu Wang. 2017. EchoTrack: Acoustic device-free hand tracking on smart phones. In *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*. 1–9. doi:10.1109/INFOCOM.2017.8057101

[11] Tuochao Chen, Justin Chan, and Shyamnath Gollakota. 2022. Underwater messaging using mobile devices. In *Proceedings of the ACM SIGCOMM 2022 Conference*. 545–559.

[12] Haiming Cheng and Wei Lou. 2021. Push the Limit of Device-Free Acoustic Sensing on Commercial Mobile Devices. In *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*. 1–10. doi:10.1109/INFOCOM42981.2021.9488703 ISSN: 2641-9874.

[13] Ting-Hui Chiang, Yi-Juan Su, Huan-Ruei Shiu, and Yu-Chee Tseng. 2020. 3D Gait Tracking by Acoustic Doppler Effects. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 3146–3149.

[14] Adil Faiz, Joël Ducourneau, Adel Khanfir, and Jacques Chatillon. 2012. Measurement of sound diffusion coefficients of scattering furnishing volumes present in workplaces. In *Acoustics 2012*.

[15] Tingchao Fan, Huangwei Wu, Meng Jin, Tao Chen, Longfei Shangguan, Xinbing Wang, and Chenghu Zhou. 2023. Towards Spatial Selection Transmission for Low-end IoT devices with SpotSound. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*.

[16] Andrea Ferlini, Dong Ma, Robert Harle, and Cecilia Mascolo. 2021. EarGate: gait-based user identification with in-ear microphones. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. 337–349.

[17] Stacy Fritz and Michelle Lusardi. 2009. White paper:"walking speed: the sixth vital sign". *Journal of geriatric physical therapy* 32, 2 (2009), 2–5.

[18] Yongjian Fu, Shuning Wang, Linghui Zhong, Lili Chen, Ju Ren, and Yaoxue Zhang. 2022. SVoice: Enabling Voice Communication in Silence via Acoustic Sensing on Commodity Devices. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*. 622–636.

[19] Zhihui Gao, Ang Li, Dong Li, Jialin Liu, Jie Xiong, Yu Wang, Bing Li, and Yiran Chen. 2022. Mom: Microphone based 3d orientation measurement. In *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 132–144.

[20] Arindam Ghosh, Amartya Chakraborty, Dhruv Chakraborty, Mousumi Saha, and Sujoy Saha. 2019. UltraSense: A non-intrusive approach for human activity identification using heterogeneous ultrasonic sensor grid for smart home environment. *Journal of Ambient Intelligence and Humanized Computing* (2019), 1–22.

[21] Yanbin Gong, Qian Zhang, Bobby H.P. NG, and Wei Li. 2022. BreathMentor: Acoustic-based Diaphragmatic Breathing Monitor System. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (July 2022), 53:1–53:28. doi:10.1145/3534595

[22] Yuqian Hu, Feng Zhang, Chenshu Wu, Beibei Wang, and KJ Ray Liu. 2020. A WiFi-based passive fall detection system. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1723–1727.

[23] Yuqian Hu, Feng Zhang, Chenshu Wu, Beibei Wang, and KJ Ray Liu. 2021. DeFall: Environment-independent passive fall detection using WiFi. *IEEE Internet of Things Journal* 9, 11 (2021), 8515–8530.

[24] Yuqian Hu, Feng Zhang, Chenshu Wu, Beibei Wang, and K. J. Ray Liu. 2020. A WiFi-Based Passive Fall Detection System. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Barcelona, Spain, 1723–1727. doi:10.1109/ICASSP40776.2020.9054753

[25] Sijie Ji, Yaxiong Xie, and Mo Li. 2022. SiFall: Practical Online Fall Detection with RF Sensing. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*. 563–577.

[26] Tadao Kasami. 1966. *WEIGHT DISTRIBUTION FORMULA FOR SOME CLASS OF CYCLIC CODES:*. Technical Report. Defense Technical Information Center, Fort Belvoir, VA. doi:10.21236/AD0632574

[27] Maruchi Kim, Anran Wang, Srdjan Jelacic, Andrew Bowdle, Shyamnath Gollakota, and Kelly Michaelsen. 2023. A Low-power wearable acoustic device for accurate invasive arterial pressure monitoring. *Communications Medicine* 3, 1 (2023), 70.

[28] Heinrich Kuttruff. 2000. Room acoustics. (2000). Publisher: Spon press UK.

[29] Robert B Leighton and Matthew Sands. 1965. *The Feynman lectures on physics*. Addison-Wesley Boston, MA, USA.

[30] Dong Li, Shirui Cao, Sunghoon Ivan Lee, and Jie Xiong. 2022. Experience: practical problems for acoustic sensing. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*. 381–390.

[31] Dong Li, Jialin Liu, Sunghoon Ivan Lee, and Jie Xiong. 2022. LASense: Pushing the Limits of Fine-grained Activity Sensing Using Acoustic Signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 1 (March 2022), 21:1–21:27.

[32] Dong Li, Jialin Liu, Sunghoon Ivan Lee, and Jie Xiong. 2022. Room-Scale Hand Gesture Recognition Using Smart Speakers. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*. 462–475.

[33] Wenwei Li, Ruiyang Gao, Jie Xiong, Jiarun Zhou, Leye Wang, Xingjian Mao, Enze Yi, and Daqing Zhang. 2024. WiFi-CSI Difference Paradigm: Achieving Efficient Doppler Speed Estimation for Passive Tracking. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 2 (2024), 1–29.

[34] Jie Lian, Jiadong Lou, Li Chen, and Xu Yuan. 2021-01-01. EchoSpot: Spotting Your Locations via Acoustic Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (Sept. 2021-01-01), 113:1–113:21. doi:10.1145/3478095

[35] Jie Lian, Xu Yuan, Ming Li, and Nian-Feng Tzeng. 2021. Fall Detection via Inaudible Acoustic Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (Sept. 2021), 1–21. doi:10.1145/3478094

[36] Roger A Light. 2017. Mosquitto: server and client implementation of the MQTT protocol. *Journal of Open Source Software* 2, 13 (2017), 265.

[37] Wenguang Mao, Jian He, and Lili Qiu. 2016. CAT: high-precision acoustic motion tracking. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*. ACM, New York City New York, 69–81. doi:10.1145/2973750.2973755

[38] Wenguang Mao, Wei Sun, Mei Wang, and Lili Qiu. 2020. DeepRange: Acoustic Ranging via Deep Learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (Dec. 2020), 1–23. doi:10.1145/3432195

[39] Wenguang Mao, Mei Wang, and Lili Qiu. 2018. AIM: Acoustic Imaging on a Mobile. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, Munich Germany, 468–481. doi:10.1145/3210240.3210325

[40] Wenguang Mao, Zaiwei Zhang, Lili Qiu, Jian He, Yuchen Cui, and Sangki Yun. 2017. Indoor follow me drone. In *Proceedings of the 15th annual international conference on mobile systems, applications, and services*. 345–358.

[41] Addie Middleton, Stacy L Fritz, and Michelle Lusardi. 2015. Walking speed: the functional vital sign. *Journal of aging and physical activity* 23, 2 (2015), 314–322.

[42] Rajalakshmi Nandakumar, Vikram Iyer, Desney Tan, and Shyamnath Gollakota. 2016. FingerIO: Using Active Sonar for Fine-Grained Finger Tracking. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, San Jose California USA, 1515–1525. doi:10.1145/2858036.2858580

[43] Kai Niu, Xuanzhi Wang, Fusang Zhang, Rong Zheng, Zhiyun Yao, and Daqing Zhang. 2022. Rethinking Doppler effect for accurate velocity estimation with commodity WiFi devices. *IEEE Journal on Selected Areas in Communications* 40, 7 (2022), 2164–2178.

[44] Xiaomin Ouyang, Xian Shuai, Jiayu Zhou, Ivy Wang Shi, Zhiyuan Xie, Guoliang Xing, and Jianwei Huang. 2022. Cosmo: contrastive fusion learning with small data for multimodal human activity recognition. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*. 324–337.

[45] Sameera Palipana, David Rojas, Piyush Agrawal, and Dirk Pesch. 2018. FallDeFi: Ubiquitous Fall Detection using Commodity Wi-Fi Devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (Jan. 2018), 1–25. https://dl.acm.org/doi/10.1145/3161183

[46] Allan D Pierce. 2019. *Acoustics: an introduction to its physical principles and applications*. Springer.

[47] Kun Qian, Chenshu Wu, Fu Xiao, Yue Zheng, Yi Zhang, Zheng Yang, and Yunhao Liu. 2018. Acousticcardiogram: Monitoring Heartbeats using Acoustic Signals on Smart Devices. In *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*. IEEE, Honolulu, HI, 1574–1582. doi:10.1109/INFOCOM.2018.8485978

[48] Kun Qian, Chenshu Wu, Yi Zhang, Guidong Zhang, Zheng Yang, and Yunhao Liu. 2018. Widar2. 0: Passive human tracking with a single Wi-Fi link. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*. 350–361.

[49] Kun Qian, Chenshu Wu, Zimu Zhou, Yue Zheng, Zheng Yang, and Yunhao Liu. 2017. Inferring motion direction using commodity wi-fi for interactive exergames. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 1961–1972.

[50] Hossein Raeis, Mohammad Kazemi, and Shervin Shirmohammadi. 2021. Human activity recognition with device-free sensors for well-being assessment in smart homes. *IEEE Instrumentation & Measurement Magazine* 24, 6 (2021), 46–57.

[51] Line Jee Hartmann Rasmussen, Avshalom Caspi, Antony Ambler, Jonathan M Broadbent, Harvey J Cohen, Tracy d'Arbeloff, Maxwell Elliott, Robert J Hancox, HonaLee Harrington, Sean Hogan, et al. 2019. Association of neurocognitive and physical function with gait speed in midlife. *JAMA network open* 2, 10 (2019), e1913123–e1913123.

[52] Andrea L Rosso, Joe Verghese, Andrea L Metti, Robert M Boudreau, Howard J Aizenstein, Stephen Kritchevsky, Tamara Harris, Kristine Yaffe, Suzanne Satterfield, Stephanie Studenski, et al. 2017. Slowing gait and risk for cognitive impairment: the hippocampus as a shared neural substrate. *Neurology* 89, 4 (2017), 336–342.

[53] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4510–4520.

[54] Dakun Shen, Ian Markwood, Dan Shen, and Yao Liu. 2018. Virtual safe: Unauthorized walking behavior detection for mobile devices. *IEEE Transactions on Mobile Computing* 18, 3 (2018), 688–701.

[55] Sheng Shen, Daguan Chen, Yu-Lin Wei, Zhijian Yang, and Romit Roy Choudhury. 2020. Voice localization using nearby wall reflections. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. ACM, London United Kingdom, 1–14. doi:10.1145/3372224.3380884

[56] Ke Sun and Xinyu Zhang. 2021. UltraSE: single-channel speech enhancement using ultrasound. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. ACM, New Orleans Louisiana, 160–173.

[57] Ke Sun, Ting Zhao, Wei Wang, and Lei Xie. 2018. VSkin: Sensing Touch Gestures on Surfaces of Mobile Devices Using Acoustic Signals. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*. ACM, New Delhi India, 591–605. doi:10.1145/3241539.3241568

[58] Li Sun, Souvik Sen, Dimitrios Koutsonikolas, and Kyu-Han Kim. 2015. Widraw: Enabling hands-free drawing in the air on commodity wifi devices. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*. 77–89.

[59] Xiaoning Tang and Xiong Yan. 2017. Acoustic energy absorption properties of fibrous materials: A review. *Composites Part A: Applied Science and Manufacturing* 101 (2017), 360–380.

[60] Beitong Tian, Lingzhi Zhao, Bo Chen, Mingyuan Wu, Haozhen Zheng, Deepak Vasisht, Francis Y Yan, and Klara Nahrstedt. 2025. AquaScope: Reliable Underwater Image Transmission on Mobile Devices. *arXiv preprint arXiv:2502.10891* (2025).
[61] Pro Sound Training. 2015. Acoustical Scattering. https://www.prosoundtraining.com/2015/04/03/acoustical-scattering/.
[62] M. Umair Bin Altaf, Taras Butko, and Biing-Hwang Juang. 2015. Acoustic Gaits: Gait Analysis With Footstep Sounds. *IEEE Transactions on Biomedical Engineering* 62, 8 (Aug. 2015), 2001–2011. doi:10.1109/TBME.2015.2410142
[63] Bandhav Veluri, Justin Chan, Malek Itani, Tuochao Chen, Takuya Yoshioka, and Shyamnath Gollakota. 2023. Real-time target sound extraction. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.
[64] Bandhav Veluri, Malek Itani, Tuochao Chen, Takuya Yoshioka, and Shyamnath Gollakota. 2024. Look once to hear: Target speech hearing with noisy examples. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–16.
[65] Ville Viikari, Kimmo Kokkonen, and Johanna Meltaus. 2008. Optimized signal processing for FMCW interrogated reflective delay line-type SAW sensors. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* 55, 11 (2008), 2522–2526. doi:10.1109/TUFFC.961
[66] Haoran Wan, Shuyu Shi, Wenyu Cao, Wei Wang, and Guihai Chen. 2021. RespTracker: Multi-user Room-scale Respiration Tracking with Commercial Acoustic Devices. In *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*. IEEE, Vancouver, BC, Canada, 1–10. doi:10.1109/INFOCOM42981.2021.9488881
[67] Anran Wang and Shyamnath Gollakota. 2019. MilliSonic: Pushing the Limits of Acoustic Motion Tracking. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–11. doi:10.1145/3290605.3300248
[68] Anran Wang, Jacob E. Sunshine, and Shyamnath Gollakota. 2019. Contactless Infant Monitoring using White Noise. In *The 25th Annual International Conference on Mobile Computing and Networking*. ACM, Los Cabos Mexico, 1–16. doi:10.1145/3300061.3345453
[69] Lei Wang, Tao Gu, Wei Li, Haipeng Dai, Yong Zhang, Dongxiao Yu, Chenren Xu, and Daqing Zhang. 2023. DF-Sense: Multi-user Acoustic Sensing for Heartbeat Monitoring with Dualforming. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*. 1–13.
[70] Lei Wang, Wei Li, Ke Sun, Fusang Zhang, Tao Gu, Chenren Xu, and Daqing Zhang. 2022. Loear: Push the range limit of acoustic sensing for vital sign monitoring. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–24.
[71] Mei Wang and Wei Sun. 2021. MAVL: Multiresolution Analysis of Voice Localization. *In Proc. of NSDI* (2021).
[72] Qian Wang, Xiu Lin, Man Zhou, Yanjiao Chen, Cong Wang, Qi Li, and Xiangyang Luo. 2019. Voicepop: A pop noise based anti-spoofing system for voice authentication on smartphones. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 2062–2070.
[73] Tianben Wang, Daqing Zhang, Yuanqing Zheng, Tao Gu, Xingshe Zhou, and Bernadette Dorizzi. 2018. C-FMCW Based Contactless Respiration Detection Using Acoustic Signal. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (Jan. 2018), 1–20. doi:10.1145/3161188
[74] Wei Wang, Alex X Liu, Muhammad Shahzad, Kang Ling, and Sanglu Lu. 2015. Understanding and modeling of wifi signal based human activity recognition. In *Proceedings of the 21st annual international conference on mobile computing and networking*. 65–76.
[75] Wei Wang, Alex X. Liu, and Ke Sun. 2016. Device-free gesture tracking using acoustic signals. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking (MobiCom '16)*. Association for Computing Machinery, New York, NY, USA, 82–94. doi:10.1145/2973750.2973764
[76] Yuxi Wang, Kaishun Wu, and Lionel M. Ni. 2017. WiFall: Device-Free Fall Detection by Wireless Networks. *IEEE Transactions on Mobile Computing* 16, 2 (Feb. 2017), 581–594.
[77] Ziqi Wang, Zhe Chen, Akash Deep Singh, Luis Garcia, Jun Luo, and Mani B Srivastava. 2020. Uwhear: through-wall extraction and separation of audio vibrations using wireless signals. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 1–14.
[78] Chenshu Wu, Beibei Wang, Oscar C Au, and KJ Ray Liu. 2022. Wi-Fi Can Do More: Toward Ubiquitous Wireless Sensing. *IEEE Communications Standards Magazine* 6, 2 (2022), 42–49.
[79] Chenshu Wu, Feng Zhang, Yuqian Hu, and K. J. Ray Liu. 2021. GaitWay: Monitoring and Recognizing Gait Speed Through the Walls. *IEEE Transactions on Mobile Computing* 20, 6 (June 2021), 2186–2199.
[80] Chenshu Wu, Feng Zhang, Yuqian Hu, and KJ Ray Liu. 2020. GaitWay: Monitoring and recognizing gait speed through the walls. *IEEE Transactions on Mobile Computing* 20, 6 (2020), 2186–2199.
[81] Chenshu Wu, Feng Zhang, Beibei Wang, and KJ Ray Liu. 2020. mmTrack: Passive multi-person localization using commodity millimeter wave radio. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 2400–2409.
[82] Binbin Xie, Minhao Cui, Deepak Ganesan, Xiangru Chen, and Jie Xiong. 2023. Boosting the Long Range Sensing Potential of LoRa. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*. 177–190.
[83] Jie Xiong and Kyle Jamieson. 2013. {ArrayTrack}: A {Fine-Grained} indoor location system. In *10th USENIX Symposium on Networked Systems Design and Implementation (NSDI 13)*. 71–84.
[84] Chenhan Xu, Tianyu Chen, Huining Li, Alexander Gherardi, Michelle Weng, Zhengxiong Li, and Wenyao Xu. 2022. Hearing Heartbeat from Voice: Towards Next Generation Voice-User Interfaces with Cardiac Sensing Functions. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*. 149–163.

[85] Huatao Xu, Pengfei Zhou, Rui Tan, and Mo Li. 2023. Practically Adopting Human Activity Recognition. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*.

[86] Wei Xu, ZhiWen Yu, Zhu Wang, Bin Guo, and Qi Han. 2019. Acousticid: gait-based human identification using acoustic signal. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–25.

[87] Wei Xu, ZhiWen Yu, Zhu Wang, Bin Guo, and Qi Han. 2019. AcousticID: Gait-based Human Identification Using Acoustic Signal. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (Sept. 2019), 1–25.

[88] Yilin Yang, Xin Li, Zhengkun Ye, Yan Wang, and Yingying Chen. 2023. BioCase: Privacy Protection via Acoustic Sensing of Finger Touches on Smartphone Case Mini-Structures. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*. 397–409.

[89] Luca Jiang-Tao Yu, Running Zhao, Sijie Ji, Edith CH Ngai, and Chenshu Wu. 2025. USpeech: Ultrasound-Enhanced Speech with Minimal Human Effort via Cross-Modal Synthesis. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 9, 2 (2025), 1–31.

[90] Kuang Yuan, Mohamed Ibrahim, Yiwen Song, Guoxiang Deng, Robert A. Nerone, Suvendra Vijayan, Akshay Gadre, and Swarun Kumar. 2024. ToMoBrush: Exploring Dental Health Sensing Using a Sonic Toothbrush. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 3, Article 139 (Sept. 2024), 27 pages. doi:10.1145/3678505

[91] Kuang Yuan, Dong Li, Hao Zhou, Zhehao Li, Lili Qiu, Swarun Kumar, and Jie Xiong. 2025. WindDancer: Understanding Acoustic Sensing under Ambient Airflow. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 2, Article 61 (June 2025), 25 pages. doi:10.1145/3729469

[92] Sangki Yun, Yi-Chao Chen, and Lili Qiu. 2015. Turning a Mobile Device into a Mouse in the Air. In *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, Florence Italy, 15–29. https://dl.acm.org/doi/10.1145/2742647.2742662

[93] Sangki Yun, Yi-Chao Chen, Huihuang Zheng, Lili Qiu, and Wenguang Mao. 2017. Strata: Fine-Grained Acoustic-based Device-Free Tracking. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, Niagara Falls New York USA, 15–28. https://dl.acm.org/doi/10.1145/3081333.3081356

[94] Feng Zhang, Chen Chen, Beibei Wang, and K. J. Ray Liu. 2018. WiSpeed: A Statistical Electromagnetic Approach for Device-Free Indoor Speed Estimation. *IEEE Internet of Things Journal* 5, 3 (June 2018), 2163–2177. doi:10.1109/JIOT.2018.2826227

[95] Huanle Zhang, Wan Du, Pengfei Zhou, Mo Li, and Prasant Mohapatra. 2016. DopEnc: Acoustic-based encounter profiling using smartphones. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*. 294–307.

[96] Yi Zhang, Weiying Hou, Zheng Yang, and Chenshu Wu. 2023. VECARE: Statistical Acoustic Sensing for Automotive In-Cabin Monitoring. In *USENIX NSDI*.

[97] Yongzhao Zhang, Hao Pan, Yi-Chao Chen, Lili Qiu, Yu Lu, Guangtao Xue, Jiadi Yu, Feng Lyu, and Haonan Wang. 2023. Addressing Practical Challenges in Acoustic Sensing To Enable Fast Motion Tracking. In *Proceedings of the 22nd International Conference on Information Processing in Sensor Networks*. 82–95.

[98] Yongzhao Zhang, Yezhou Wang, Lanqing Yang, Mei Wang, Yi-Chao Chen, Lili Qiu, Yihong Liu, Guangtao Xue, and Jiadi Yu. 2023. Acoustic Sensing and Communication Using Metasurface. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*. 1359–1374.

[99] Yue Zheng, Yi Zhang, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. 2019. Zero-effort cross-domain gesture recognition with Wi-Fi. In *Proceedings of the 17th annual international conference on mobile systems, applications, and services*. 313–325.

[100] Suping Zhou, Jia Jia, Zhiyong Wu, Zhihan Yang, Yanfeng Wang, Wei Chen, Fanbo Meng, Shuo Huang, Jialie Shen, and Xiaochuan Wang. 2021. Inferring emotion from large-scale internet voice data: A semi-supervised curriculum augmentation based deep learning approach. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 6039–6047.