# EchoSensor: Fine-grained Ultrasonic Sensing for Smart Home Intrusion Detection

JIE LIAN and CHANGLAI DU, University of Louisiana at Lafayette, USA
JIADONG LOU, University of Delaware, USA
LI CHEN, University of Louisiana at Lafayette, USA
XU YUAN, University of Delaware, USA

This article presents the design and implementation of a novel intrusion detection system, called EchoSensor, which leverages speakers and microphones in smart home devices to capture human gait patterns for individual identification. EchoSensor harnesses the speaker to send inaudible acoustic signals (around 20 kHz) and utilizes the microphone to capture the reflected signals. As the reflected signals have unique variations in the Doppler shift respective to the gaits of different people, EchoSensor is able to profile human gait patterns from the generated spectrograms. To mine the gait information, we first propose a two-stage interference cancellation scheme to remove the background noise and environmental interference, followed by a new method to detect the starting point of walking and estimate the gait cycle time. We then perform the fine-grained analysis of the spectrograms to extract a series of features. In the end, machine learning is employed to construct an identifier for individual recognition. We implement the EchoSensor system and deploy it under different household environments to conduct intrusion detection tasks. Extensive experimental results have demonstrated that EchoSensor can achieve the averaged Intruder Gait Detection Rate (IDR) and True Family Member Gait Detection Rate (TFR) of 92.7% and 91.9%, respectively.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing;**

Additional Key Words and Phrases: Ultrasonic sensing, smart home, intrusion detection, individual identification

**ACM Reference format:**
Jie Lian, Changlai Du, Jiadong Lou, Li Chen, and Xu Yuan. 2023. EchoSensor: Fine-grained Ultrasonic Sensing for Smart Home Intrusion Detection. *ACM Trans. Sensor Netw.* 20, 1, Article 10 (October 2023), 24 pages.
https://doi.org/10.1145/3615658

## 1 INTRODUCTION

Home security systems are essential to a family for the protection of properties and the maintenance of household safety from potential break-ins. Different technologies and solutions for

This work was supported in part by NSF under Grants 2019511, 2146447, and in part by the BoRSF under Grants LEQSF(2019-22)-RD-A-21. Any opinion and findings expressed in the article are those of the authors and do not necessarily reflect the view of funding agencies.

Authors' addresses: J. Lian, and L. Chen, University of Louisiana at Lafayette, Lafayette, LA 70503; emails: c00441055@louisiana.edu, li.chen@louisiana.edu; C. Du, University of Louisiana at Lafayette, Virginia Tech, Blacksburg, VA 24061; email: cdu@vt.edu; J. Lou and X. Yuan, University of Delaware, Newark, DE 19716; emails: loujd@udel.edu, xyuan@udel.edu.

1550-4859/2023/10-ART10 $15.00
https://doi.org/10.1145/3615658

mining biometric information to implement unsolicited intrusion detection have been proposed [6, 28, 30, 41, 46, 47], mainly relying on surveillance cameras, floor sensors, wearable sensors, and Wi-Fi devices. By extracting desirable features from captured data (e.g., video from cameras, foot pressure from floor sensors, stride and velocity reported by wearable sensors, and **channel state information** (**CSI**) from Wi-Fi devices), the unique patterns for an individual can be spotted. Some solutions have been commercialized in the professional home security systems to conduct the effective intrusion detection tasks. However, the Parks Associates report [11] shows that only 27% of the U.S. households have adopted home security systems, for which the most commonly cited reason is the high cost of such professional systems incurred by the purchase of expensive devices or service fees. Undoubtedly, many people do not envision the security as a necessary expense and are still reluctant to invest in home security. Surveillance camera is the most popular home security solution, in addition to extra cost for purchasing the devices, many serious privacy concerns have been raised by the community on the private and sensitive video data, whereas a remote attacker may have the chance to gain access to the video captured by the camera [38] or infer the action of the victim from the encrypted video stream [17, 18]. The newly appeared Wi-Fi CSI-based human identification solution is a viable method that probably can cut down the deployment cost to some extent. However, the uneven distribution of the CSI Fresnel zone and the noise of electromagnetic signals itself make the obtained Wi-Fi signals unstable [44]. In addition, Wi-Fi CSI-based solutions keep sending CSI measurement packets, which inevitably occupy Wi-Fi channel resources and impact the performance of nearby Wi-Fi devices [20]. In comparison, a self-installing software-based system at a negligible expense is highly desired, especially given the proliferation of smart home devices.

The functional components of appliances in smart homes offer new opportunities to explore their sensing capabilities for home security, especially in authentication with biometrics. As a promising type of such components, the audio system equipped with internal speaker and microphone can be well leveraged to implement new smart home sensing. As an example, voice assistant systems, equipped with speakers and microphones, have played important roles in smart homes by detecting voice commands and making timely responses. According to a report [33] from Google, voice assistant systems are available on more than 400 million devices, including Google Home, Amazon Echo, Apple HomePod, Aristotle, DingDong, Rokid, Mi AI Speaker, smartphones, headphones, TVs, smartwatches, and others. The popularity of voice systems promotes the recent research interest in exploring the potential sensing capabilities of built-in speakers and microphones, which are "always-on" in these systems, to serve new needs in smart homes. A critical direction of such an exploration, which has been pervasively attracting research community's interest, is to enable the sensing capability of voice assistant devices using acoustic signals.

Many research efforts have been made to explore the sensing abilities of acoustic signals on objects, shapes, and environmental dynamics. For example, in [40], Sonar has been developed to gather information about distant objects such as range, angle, and velocity using sound propagation. In **Internet-of-Things** (**IoT**) applications, the inaudible acoustic signal at 20 kHz has been utilized to sense human activities. Specifically, in [9, 24], the acoustic signal is used to estimate human motion and gesture, while in [35], the surface of mobile devices has been sensed to enable touchscreen functions. Moreover, acoustic signals have also been used to capture mouth movements in order to combat voice replay attacks [19, 48]. However, all existing solutions require the device to be close to the body part, making them unsuitable in smart homes for intrusion detection. Despite some research efforts [8, 25] extending the acoustic sensing capabilities to cover a larger range of distance, they only apply to coarse-grained detection while relying on the known environment (i.e., floor plan), failing to meet the requirements of smart home environments.

To advance the acoustic sensing in IoT applications, we propose EchoSensor, a human authentication system which uses the built-in speaker and microphone on the ubiquitous audio, voice, or other devices in smart homes to capture human biometric information (i.e., gait) for the purpose of intrusion detection. Specifically, EchoSensor controls the speaker to transmit inaudible acoustic signals around 20 kHz and leverages the microphone to capture the reflected signals. When the acoustic signal is bounced off from a walking human, variations in the frequency domains will be observed in the reflected signals, due to the Doppler effect, which may enclose the rich gait patterns representing the unique information of this person. Although there exist some solutions to explore the gait-based human identification using acoustic signals, they leverage the professional devices [15, 42, 50], i.e., Doppler radar or acoustic sensor, which have the powerful acoustic signal transmitting function and strong signal resolution capability to ensure the capture of high-quality Doppler effect spectrogram, without mining the fine-grained gait features. Such solutions cannot be applied to our proposed EchoSensor system, as we only rely on the readily available audio or voice devices at smart home, rather than purchasing extra professional devices to implement the individual recognition for the intrusion detection purpose.

The goal of this article is to explore the biometric information of a walking human through inaudible acoustic sensing, which is used as the basis for the design and implementation of EchoSensor. To achieve this goal, we generate the spectrogram of reflected signals with Doppler shift and develop signal processing techniques to mitigate interference/noise. With the pure acoustic signals acquired as a result, clear gait patterns can be mined by EchoSensor for the use of individual recognition. In particular, we first develop a solution to let EchoSensor be capable of detecting the starting point of a walking man and identifying his gait cycle time to differentiate between the independent gaits. With walking steps separated into a set of two-gait samples, the unique and fine-grained features of an individual, i.e., cepstral coefficient vector, leg contour curve, spectrum signature, and torso speed and gait cycle time pairs, are extracted via a series of signal processing techniques. To realize the eventual goal of intrusion detection, all family members of a household will enroll in EchoSensor, with their gait features extracted and trained through the machine learning classifier, resulting in a detector for them. Once an individual is sensed, his gait patterns with fine-grained features are mined by EchoSensor. Then, the detector can automatically classify him as a family member or an intruder.

Compared to other home security systems based on surveillance cameras [28], floor sensors [30], wearable sensors [6], or Wi-Fi signals [41, 46, 47], EchoSensor possesses a set of salient features, including but is not limited to: *First*, it is a software-based system that can be self-installed in existing home audio or voice devices, which is hard to be notified and compromised by an unsolicited intruder. *Second*, it leverages the speakers and microphones equipped in existing home appliances, without incurring extra cost. The fine-grained features extracted by EchoSensor are sufficient to distinguish an individual. *Third*, it is based on acoustic signals rather than radio frequency signals, thus relieving the competition for radio spectrum resource in smart homes.

We implement EchoSensor on Huawei p20 with Android 8.1 and a speaker of Bose Soundlink Revolve. Beyond our proof-of-concept implementation, EchoSensor is also envisioned to work effectively on many other smart home audio and voice devices, such as Google Home, Amazon Echo, Apple HomePod, and so on, which are suitable for the home scenario of intrusion detection because they locate at different places, covering the majority of regions in a house. We have evaluated the performance of EchoSensor in 10 households involving a total of 24 participants in terms of intruders and family members detection. Experimental results show that EchoSensor can achieve the averaged **Intruder Gait Detection Rate** (**IDR**) and **True Family Member Gait Detection Rate** (**TFR**) of 92.7% and 91.9%, respectively. The intrusion alarm accuracy can reach 98% and 100%, respectively, with only 5 and 10 detected two-gait samples.

Our contributions can be summarized as follows:

- To the best of our knowledge, EchoSensor is the first work to demonstrate the feasibility of moderate-range (5 meters) gait-based user authentication using speakers and microphones on pervasively existing audio or voice devices in smart homes.
- The new solutions based on the signal processing techniques are developed in EchoSensor to perform the fine-grained analysis so as to acquire the unique human gait patterns. In particular, EchoSensor can detect the starting point of a walking human and identify his gait cycle time to differentiate independent gaits.
- We develop new techniques to identify and extract a series of fine-grained features from low-quality Doppler signals to represent a person's walking patterns. With these features, a machine learning classifier is trained and individual recognition is implemented automatically.
- Extensive experiments are conducted to verify the performance of EchoSensor. Experimental results demonstrate that EchoSensor can achieve high accuracy in intrusion detection in smart homes.

## 2 RELATED WORK

**Acoustic Sensing.** Recently, acoustic sensing has attracted considerable interest. *Sonar* is a prominent exemplary system and has been extensively explored to gather information about distant objects (e.g., range, angle, or velocity) by using sound propagation [40]. With the proliferation of IoT devices, some research efforts have focused more on IoT applications, by taking advantage of the ubiquitous availability of speakers and microphones on audio or voice devices. Promising results have been presented by leveraging the near-ultrasound signal at around 20 kHz for human activity recognition. For instance, in [9, 24], acoustic systems are developed to estimate human motion and gesture by generating tones within 18~22 kHz from speakers; the reflected signals from the human body are captured by microphones for analysis. In [35], the *VSkin* system is proposed to enable the sensing of touch gestures on other surfaces of a mobile device, out of the touchscreen area.

Furthermore, some works have proposed to exploit the usage of a smartphone's speaker and microphone to capture mouth movements to defend voice replay attacks [19, 48]. In particular, Zhang et al. [48] analyzed the Doppler shifts of reflected signals caused by articulatory gestures of users to achieve liveness detection, while Lu et al. [19] further analyzed the uniqueness of Doppler profiles caused by different users' mouth movement patterns for the purpose of user authentication. BreathPrint [4] is proposed to utilize a microphone to capture the breathing patterns of users for authentication. Although the aforementioned systems can detect human motion like gestures and mouth movements, they have the limitations that the device must be close to the target body, making them unsuitable for the general purpose of intrusion authentication in smart homes.

On the other hand, some recent works aim at expanding the applications of acoustic signals by breaking the distance constraint. For example, in [8], the *Sonar sensor* equipped on the smartphone can measure the distance from the smartphone to an object by capturing the elapsed time between the initial pulse and its reflected pulse. In [25], CovertBand is proposed to track users and differentiate various categories of motions by capturing acoustic reflections from the human body. These systems demonstrate the possibility of extending the sensing capability of acoustic signals to a wider range, but they rely on a known environment (i.e., floorplans) and are designed for coarse-grained motion detection. In addition, [45] is proposed to identify human gait patterns via acoustic sensing; however, it assumes the ideal environments without considering the interference, noise, and reflections pervasively existing in the real home environment, making it unsuitable for home applications. Hence, the biometric recognition in a large distance range remains an open problem, especially for serving the smart home intrusion detection purpose. Nevertheless, the aforementioned solutions demonstrate the possibility of extending the sensing capability of acoustic signals to a larger distance range.

**Gait-based User Identification.** Recent works have demonstrated that gait can be utilized as a biometric signature for person identification or authentication. Traditional methods to quantify gait features include video-based and sensor-based methods. In [2, 36], the authors used video cameras to record people walking and extract the gait information. Some other works leverage various sensors to capture the gaits signature, such as floor sensors [12], rotation sensors [22], and accelerometer-based sensors [7, 32, 49]. Moreover, Ngo et al. [26] used the largest inertial sensor-based gait database, to evaluate and compare different sensor-based gait identification approaches. Pan et al. in [31] deployed a geophone on the floor and identified people passing by from detected structural vibration. However, all sensor-based methods require people to carry additional devices on the body or to deploy these sensors in the environment, which are not convenient to use while incurring additional cost. On the other hand, radio signal-based solutions have been proposed for gait identification. For example, [51] has developed an mm-wave radar by employing the FMCW as the transmitted signals and relying on the point clouds to enable multiple people identification via gait information. Considering the relatively lower sound speed and limited bandwidth, it is not suitable for us to employ the FMCW signals for gait identification in the acoustic domain. The reason is that the adjacent acoustic FMCW chirps should have a large duration to minimize interference caused by environmental reflection on the acoustic signals. However, a large interval will result in a low time resolution, thereby limiting the amount of information that can be used. Also, the limited bandwidth will constrain FMCW signals' performance. Some works [41, 46, 47] have proposed convenient, low-cost, and effortless approaches by leveraging the variation of Wi-Fi CSI to detect gait patterns. But they still need additional devices to receive the CSI signal. In addition, such a category of solutions requires continuously sending CSI measurement packets, which inevitably occupies radio resources and impacts the performance of nearby Wi-Fi devices [20] in smart homes.

**Gait-based Human Identification with Acoustic Signals.** Some research works have explored the gait-based human identification through acoustic signals using the professional devices [1, 15, 42, 50], i.e., acoustic sensor or Doppler radar. For example, [1] built a gait-based identification system by using the acoustic sensor to capture the footstep sound so that human gait can be identified. In [13, 42, 50], the gait-based human identification has been implemented by utilizing Doppler signals, which requires professional devices, such as Doppler radar, with powerful acoustic signal transmission function and strong signal resolution capability. Hence, these works can directly acquire the high-quality Doppler effect spectrogram without the need for further analysis of the received signals for human identification. In our proposed application scenarios, these solutions are not applicable, since we only rely on the ubiquitously available smart home devices, where a sequence of fine-grained analysis for the received coarse Doppler signals has to be conducted.

## 3 PRELIMINARIES

We present the preliminary knowledge and necessary background of acoustic sensing and gait-based user identification.

**Acoustic Sensing.** The principle of acoustic sensing is to capture the frequency change of a sound wave in response to a moving object. The phenomenon of frequency change caused by the object reflection is defined as the Doppler effect or Doppler shift, where the shifted frequency is determined by the source frequency and the velocity of the moving object, with a proportional increase with them. In our application, the original source (i.e., speakers) and listener (i.e., microphone) are stationary, thus in absence of any motion, there is no frequency change. When a user moves in

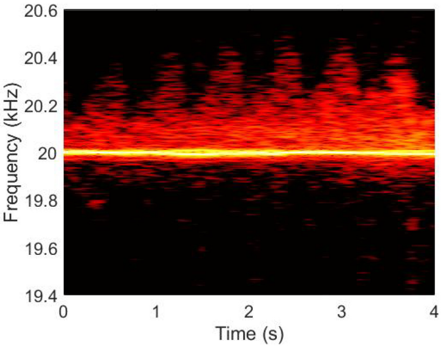


Fig. 1. Spectrogram of a moving person.

front of a speaker, it will reflect the sound waves played by the speaker and cause a frequency shift [27], which can be measured as follows:

$$f_r = f_t \cdot \left( \frac{c + v \cdot \cos\theta}{c - v \cdot \cos\theta} \right), \tag{1}$$

where $f_r$ is the perceived frequency at the microphone, $f_t$ is the original frequency from the speaker, $c$ is the speed of sound in air, $v$ is the velocity of a moving object, and $\theta$ is the angle between the object motion and the beam of ultrasonics.

Figure 1 shows the spectrogram of Doppler shift signals reflected by a walking human toward a speaker 5 meters away. Notably, the spectrogram in the figure is preprocessed by removing various types of interference, which will be elaborated in Section 5.3. Due to the Doppler effect, there are evident energy variations near 20 kHz, which will be used to detect human gaits later.

**Parameters of Interest for the Human Gait.** The research on human gait targets the qualitative and quantitative evaluation of various factors that characterize human walking behaviors. The factor of interest varies according to the field of research. For security purposes, the research interest mainly concentrates on distinguishing and identifying persons by capturing a series of general characteristics of their silhouette and the movements at different body segments while walking [10, 23]. The parameters which most clearly characterize the human gait can be summarized as follows: (1) *Velocity* includes the torso and leg speed; (2) *Short step length* represents the linear distance between two successive placements of the same foot; (3) *Long step* or *stride length* represents the linear distance between the placements of both feet; (4) *Cadence* or *rhythm* represents the number of steps per time unit; (5) *Step width* represents the linear distance between two equivalent points of both feet; (6) *Step angle* represents the direction of the foot during the step; (7) *Gait phases* represent the body posture which may be bending or symmetric.

## 4 FEASIBILITY STUDY

The limited transmission power and resolution capability of existing home devices will result in low-quality received Doppler signals. It is thus in doubt whether they are sufficient to identify the unique human gait patterns. Hence, we conduct the feasibility study of acoustic-based user identification with low-quality Doppler signals.

### 4.1 Gait Patterns Detection

It is necessary to explore whether there is a series of noticeable gait patterns in the low-quality reflected signals of a walking person so as to design the gait-based authentication system. To

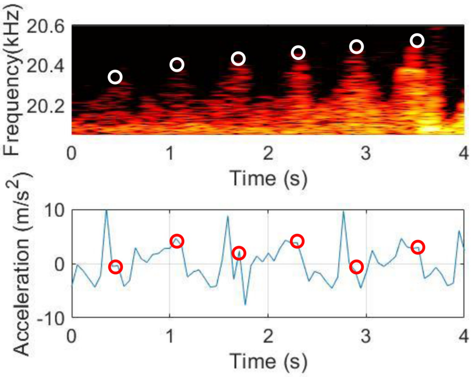


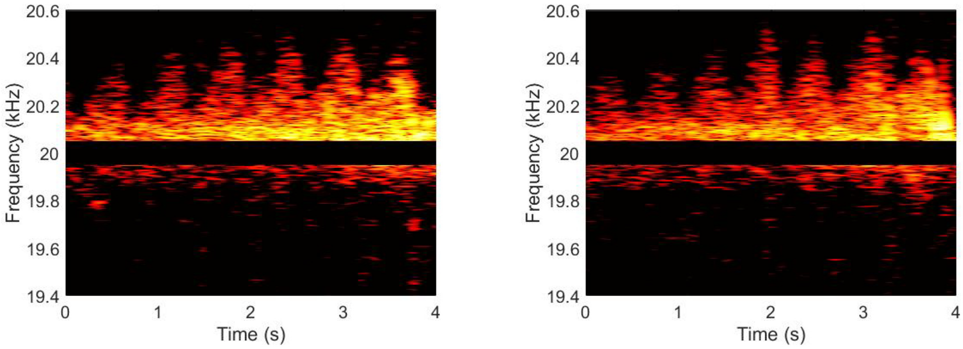
Fig. 2. Detecting gait patterns: frequency changes through acoustic sensing and the ground truth.

capture these patterns, the gait cycle and the constructed gait profile from a person's walking activity have to be taken into account by using purely observed acoustic signals. To demonstrate the feasibility of gait pattern detection, we conducted an experiment to let a person walk in a room toward a speaker (i.e., Bose soundlink Revolve) from 5 meters away. Connected to a smartphone (Huawei P20 with Android 8.1) via Bluetooth, the speaker is controlled to emit inaudible signals at 20 kHz. The Huawei P20, working as a microphone, is placed on the floor and bundled with the speaker to capture the reflected signals from the moving human body.

The top plot in Figure 2 presents the spectrogram of Doppler shift over time when a person is walking toward the speaker. The $y$-axis represents the captured frequency shifts of the reflected signals, while the $x$-axis represents time. Note that the acoustic data shown in the figure is preprocessed to remove various types of interference. To capture the ground truth of gait patterns, we also let this person carry another smartphone (Galaxy S9 plus with Android 9) in the pocket of her right leg to collect the accelerometer and gyrometer data. The ground truth of the gait patterns is also shown in Figure 2 (the bottom plot), which represents that the acceleration values on the right leg change periodically in two steps period. The periodical change in right leg speed is caused by the alternating movement of the left and right legs. The peaks in top plot of Figure 2 will alternately be generated by the movement of the left and right legs. To be specific, the first, third, and fifth peaks are caused by the movement of the right leg. From the two plots, we can see that these peaks likely reach the maximum (i.e., crest point) when the acceleration value is positively reduced to 0, while the frequency shift likely reaches the minimum point (i.e., trough point) when the acceleration value is negatively increased to 0. This experiment clearly demonstrates that the captured Doppler frequency changes can reflect a person's gait patterns to a certain extent, and acoustic sensing is feasible to capture human gait patterns.

### 4.2 Gait Pattern Distinction among Persons

We continue to verify whether the captured gait patterns are distinguishable among different individuals. For this purpose, we conducted another experiment by letting two persons walk in turn from the starting point to the speaker. Figure 3 presents the spectrograms of Doppler shifts of the two persons. With a visual inspection, it is clear that there are significant differences between the step shapes, step lengths, and gait velocities of the two subjects. In particular, the upper spectrogram in Figure 3(b) showing 7 full peaks means this person walks 7 steps in 4 seconds. The upper spectrogram in Figure 3(a) showing 6 full peaks means this person walks 6 steps in 4 seconds. Such visually observed differences, along with other underlying differences, can be exploited to build a unique gait profile for each person, which will be sufficient to identify an individual.



(a) Doppler shift of the first person.

(b) Doppler shift of the second person.

Fig. 3. The spectrograms of Doppler shifts from two persons.

## 5 SYSTEM DESIGN

In this section, we propose our design of EchoSensor, which leverages the built-in speaker and microphone in smart home audio devices to enable acoustic sensing capabilities. With the speaker emitting inaudible ultrasonic signals and the microphone receiving reflected signals, fine-grained analysis of reflected signals is employed to extract the enclosed biometric information for individual recognition and intrusion detection. In our design of EchoSensor, a number of technical challenges were encountered to be addressed, summarized as follows:

*First*, the reflected acoustic signals are typically more coarse and of lower quality with much underlying interference and noise, compared to those from professional devices. As most acoustic signals suffer from fast attenuation, the interference and noise are grossly detrimental to Doppler shift measurement and may totally destroy the embedded biometric patterns. It is necessary yet challenging to develop effective interference cancellation or mitigation solutions toward the acquisition of pure acoustic signals so as to mine the accurate biometric information.

*Second*, as EchoSensor aims to only rely on the reflected signals for human recognition without the help of additional equipment for signal resolution, it is significant to determine the useful information that can characterize human walking patterns. Thus, we need to develop solutions to automatically identify the starting point of a human walking and gait-cycle time from the Doppler effect signals, which are challenging, especially when the Doppler effects are buried in strong interference and noise.

*Third*, to recognize an individual, the effective fine-grained features that can represent an individual's behaviors shall be identified, extracted, and well utilized. However, EchoSensor can only rely on the Doppler signals to acquire such a set of features, where a series of signal processing techniques have to be developed to conduct the fine-grained analysis.

### 5.1 Threat Model

The goal of an attacker is to intrude into the victim's home without being detected by EchoSensor. We assume a strong intruder has the ability to transmit an ultrasonic signal, but could not directly access the EchoSensor system.

**No device access.** We assume an attacker cannot immediately control or turn off the audio or voice devices after entering the house. However, he may be aware of the existence of EchoSensor and knows its principle that Doppler gait signal will be captured for intrusion detection. He could choose any specific action to reduce the possibility of being detected, such as mimicking the victim's gait patterns.

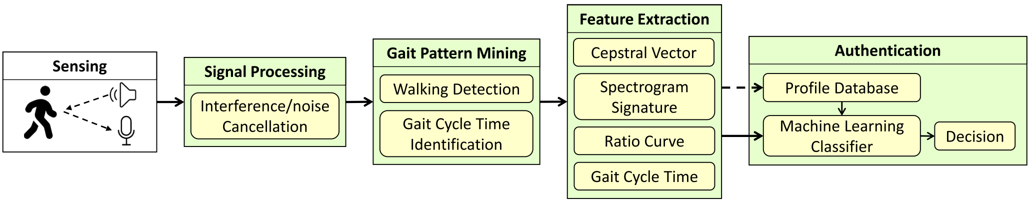


Fig. 4. The workflow of EchoSensor.

**No owner interaction.** We assume an intruder could not ask the victim to turn off the system. The EchoSensor is "always-on" when the victims leave their houses. Also, the intruder is assumed to be present during the victims' absence. That means the victims will not meet the intruder.

**Replay and saturate attack.** We assume the intruder has the knowledge for signal processing. He could process the ultrasonic signal and has the ability to play the ultrasonic signal to EchoSenor to obfuscate the system. For example, he may be the victims' friend and record their gait signals to conduct a replay attack when the victims are not at home. The intruder may also play the high-energy signal in front of the EchoSensor to conduct the saturate attack.

## 5.2 System Overview

Figure 4 exhibits the system design of EchoSensor, consisting of five components: *Sensing, Signal Processing, Gait Pattern Mining, Feature Extraction,* and *Recognition*. In the *Sensing module*, EchoSensor programs the built-in speaker in audio devices to emit inaudible acoustic signals at 20 kHz, and uses the microphone to collect Doppler effect signals with the sampling frequency of 44.1 kHz. Such a sampling frequency ensures it is twice higher than the highest Doppler shift frequency, so the Doppler signal can be entirely reconstructed from the recorded signal [34].

The received signals pass to the *Signal Processing module* through a high-pass Butterworth filter [21]. The interference and noise will be canceled or mitigated here so that the expected pure acoustic signals can be acquired. To mine the gait patterns, EchoSensor has two sub-modules in the *Gait Pattern Mining module*, i.e., walking detection and gait cycle time identification, where the former one is to determine the starting point of a human walking while the latter determines the gait time so that each independent gait can be differentiated. After this stage, the human walking steps can be separated into a set of two-gait samples, which will be sent to the *Feature Extraction module*. In this module, we focus on four categories of features for extraction, i.e., cepstral coefficient vector of each two-gait sample, spectrum signature, ratio curve between the leg speed curve and torso speed curve, and gait cycle time. In the end, we employ a machine learning classifier to recognize an individual in the *Recognition module*. We next illustrate the detailed design of each module.

## 5.3 Interference Cancellation

After receiving the reflected acoustic signals at the microphone, we generate the spectrogram of the Doppler shift for analysis. Figure 5 is an example of the spectrogram from reflected signals, where the **short-time Fourier transform** (**STFT**) is employed to process them. During the STFT process, the original signal is sliced into a set of small bins, with each bin having a 0.3 seconds duration. The overlapping between each two consecutive bins is set to 95%. We then multiply each bin with the Hamming window and apply the 8,192 point **Fast Fourier transformation** (**FFT**) on each bin to compose the spectrogram. As the human walking speed is estimated up to a maximum of 1.5 m/s [37], and the maximum leg speed is up to 5 m/s, the reflected signal will have a maximum shift of 500Hz, according to Equation (1). Thus, in Figure 5, we only need to consider the

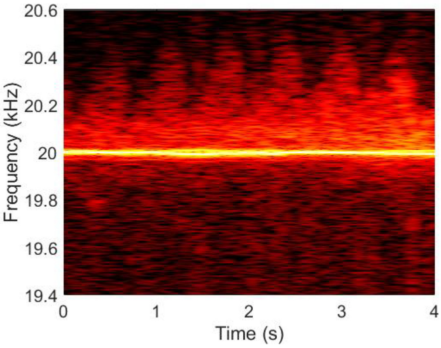


Fig. 5. Spectrogram of Doppler effect from the gait movement.

spectrogram between 19.5 kHz and 20.5 kHz, sufficiently covering all the Doppler shifts bounced off from a moving body. We also observe the blurred shape of the Doppler effect between 20 kHz and 20.5 kHz, drowned in the noise and sub-harmonics. Hence, the next critical step is to mitigate the interference and noise so as to locate the signal of interest.

Since the built-in speaker and microphone are omnidirectional, signals received by the microphone include not only desired reflection from the target, but also noise. Two major sources of noises are identified: direct transmission noise and background reflection noise, where the former one is the original audio played by the speaker while the late one is caused by the static reflector in the environment. Both of them overlap and interfere with the Doppler effect in time domain which greatly distort the quality of Doppler effect spectrogram. To minimize the impact of interference, we leverage the fact that both direct transmission and background reflection do not generate Doppler shift as we assume (1) the speaker and microphone remain static, thus there is no relevant movement between them; (2) there is no moving object in the environment, so the environment is considered to be static. These assumptions guarantee that the aforementioned two types of noises do not change over time, allowing us to safely subtract them from the spectrogram of received signal.

As such, in our *Signal Processing Module*, we propose a two-stage scheme to eliminate all the existing interference and noise, in order to acquire a clear Doppler shift spectrogram.

- **Stage 1** As discussed, the Doppler effect can only exist in the range of $20 \pm 0.5$kHz, considering the human walking speed. Hence, we pass the received signal to a high-pass Butterworth filter with a cut-off frequency of 19 kHz to obtain the spectrogram that is only located within 19.5 kHz~20.5 kHz.
- **Stage 2** This stage tries to remove the noise from direct transmission and background. Since these two types of noise stay static over time, their spectrograms would be static (i.e., no Doppler effect). This allows us to eliminate interference by subtracting the spectrograms of noise from that of received signals. To acquire the background noise, we put the speaker in an empty room to generate the ultrasonic signal, where there is no moving object in front. The microphone will record the background noises that are reflected from different directions, which will be stored in EchoSensor for the use of subtraction. That is, EchoSensor eliminates the power from these static reflectors by simply subtracting the output of the FFT in background noise from the FFT of reflected signals. This spectral subtraction method can eliminate all interference/noise from the static reflector in the environment. Theoretically, the direct transmission noise can also be subtracted in this process. However, the built-in speakers in smart home devices are not designed to generate

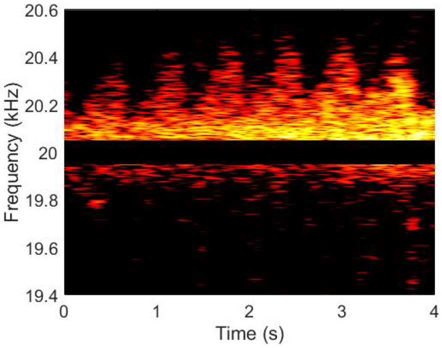


Fig. 6. The resulted spectrogram after applying two-stage cancellation.

such a high-frequency signal, which will cause the frequency of emitted ultrasonic waves to fluctuate slightly over time and thus cannot be eliminated by spectral subtraction. Such a phenomenon has been verified in our extensive experiments, where we have consistently observed that the frequency generated by the speaker fluctuates from 19.96 kHz to 20.04 kHz. Thus, we can set the received signal energy values between 19.95 kHz and 20.05 kHz to zeros for safely removing all the direct transmission noise without impacting the Doppler effect.

Figure 6 shows the resulted spectrogram after applying our two-stage scheme to Figure 5. We can see the clear waveform of Doppler effect after removing all interference/noise. The waveform of the Doppler effect varies over time, representing the human is moving because it has different instantaneous speeds at different moments.

### 5.4 Walking Detection

After getting the clear Doppler effect waveform, we need to determine the starting point of human walking so as to identify a person's gait. This step is important as accurately detecting the walking activity can also ensure that EchoSensor is initiated only when a person is detected to be walking. We now discuss how we can detect walking activity using acoustic data. Although the direct transmission and background noise are eliminated in Section 5.3, there may still exist some underlying interference. We keep track of such noise level threshold as $N_t$, which is initialized by calculating the average energy level between 19.5 kHz~20.5 kHz at time $t_0$, the time after initiating the speaker for 0.5 seconds. The reason for choosing 19.5 kHz~20.5 kHz is that the Doppler effect only locates within this range. $N_t$ is updated by Exponential Moving Average algorithm [14] in silent time period, i.e.,

$$N_t = (1 - \alpha)N_{t-1} + \alpha E_t, \tag{2}$$

where $E_t$ is the current environment noise level (i.e., averaged energy level between 19.5 kHz~20.5 kHz) at time slice $t$, $N_{t-1}$ is the noise level threshold at time slice $t-1$, and $\alpha$ is set to 0.1 according to prior work on detecting the start of walking using Wi-Fi signals [41]. We have observed there is still some impulse noise in the spectrogram. To avoid the misunderstanding of the impulse noise as the beginning of the Doppler signal, we set the detecting threshold value as three times the noise level. Such a threshold value is set up based on our observation that it is higher than any impulse noise we have observed and is also beyond the energy level of the Doppler signal. The Doppler signal caused by human motion is detected when the energy level between 19.5 kHz~20.5 kHz is above the detecting threshold value.

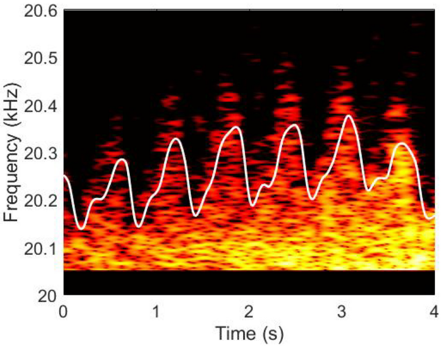


Fig. 7. Upper contour of the marked spectrogram.

### 5.5 Gait Cycle Time Estimation

We next try to separate the spectrogram of Doppler effect into a set of two-gait samples. The gait cycle time, defined as the time duration between two consecutive events when the right heel touches the ground, can be roughly estimated by visual inspection on the spectrogram of Figure 6. We elaborate our design of an estimation strategy that can automatically calculate the gait cycle time.

Our estimation is based on the upper contour from the leg reflection. As a person walks toward the speaker, the generated Doppler effect will be within $f_{\min} \sim f_{\max}$, i.e., 20 kHz~20.5 kHz. Likewise, when this person walks away from the speaker, $f_{\min}$ and $f_{\max}$ will be 19.5 kHz and 20 kHz, respectively. To identify each point in the leg upper contour, we define a function $P(f, t)$ as follows:

$$P(f, t) = \frac{\sum_{f_{\min}}^{f} F(f, t)}{\sum_{f_{\min}}^{f_{\max}} F(f, t)}, \tag{3}$$

where $F(f, t)$ represents the Doppler effect energy on the frequency $f$ at a certain time $t$ on the spectrogram, $\sum_{f_{\min}}^{f} F(f, t)$ represents the cumulative energy from $f_{\min}$ to a frequency $f$, and $\sum_{f_{\min}}^{f_{\max}} F(f, t)$ represents the total cumulative energy of the Doppler signal at time $t$. Hence, $P(f, t)$ represents the ratio of cumulative energy that is lower than frequency $f$ over the total Doppler effect energy at time $t$. An appropriate value of $f$ at time $t$ in Doppler effect spectrogram needs to be obtained so as to identify a point in the upper contour. We set the threshold value of $P(f, t)$ as 95% to derive the frequency $f$, following the suggestion in previous work [39] which explored human movement model using Doppler radar signal.

Human walking involves an acceleration phase and a uniform speed phase. To get a steady walking, we should remove the acceleration phase while keeping the uniform speed phase as the sample. That is, we remove the first two-gait cycle of a walking sample, which represents the start of walking. As a result, we obtain the upper contour in Figure 7 shown as the white curve, by applying the aforementioned method to the spectrogram in Figure 6. It is clearly observed that our method tracks the variation of leg speed during each walking period. We continue to process the leg contour curve by subtracting the frequency on the curve with the averaged frequency over all points. The resulted new curve is then passed to a low-pass filter with the cut-off frequency of 6 Hz to get a smooth curve. The valleys represent the lowest leg speed of a gait cycle, indicating a heel touches the ground, and the gait cycle time is the time duration between two consecutive touches on the ground of the right heel.

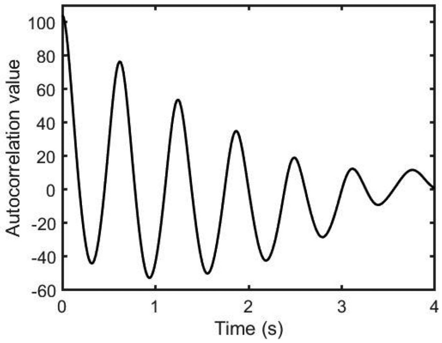


Fig. 8. Autocorrelation of the leg contour curve.

However, since FFT may cause frequency leakage, some noise power will render the valley offset on the spectrogram. This may lead to some errors in the estimated gait cycle time, as the time corresponding to the valley does not necessarily represent the beginning of a gait cycle. To minimize the adverse impact of frequency leakage, we use the autocorrelation of the leg contour curve to robustly estimate gait cycle time, which is calculated by

$$R(\tau) = \sum_{t} (F(t) - \mu)(F(t - \tau) - \mu), \tag{4}$$

where $\mu$ is the averaged frequency of the entire leg contour, $F(t)$ is the frequency of leg contour at time $t$, and $\tau$ represents the time shift. The autocorrelation is taken over a period of steady walking, which means that we take the autocorrelation on the smooth curve obtained after the processing of the low-pass filter. Through autocorrelation, we can obtain a better estimation of the gait cycle time than directly searching for valleys on the leg contour curve. Each peak in the autocorrelation function indicates that the contour curve is similar to its original version when shifting for a time period of $\tau$. A gait cycle contains two footsteps where the leg speed changes twice, thus the leg contours are similar when shifting $L/2$, where $L$ is the gait cycle time. Figure 8 shows the contour curve after autocorrelation, where the first peak appears at the point $\tau = 0.62$ seconds. Thus, the estimated gait cycle time $L = 2\tau$ is 1.24 seconds. A gait cycle is also called as two-gait sample or instance in the following of this article.

### 5.6 Feature Extraction

Fine-grained feature extraction is a critical step in EchoSensor to capture the unique biometric information of each individual. In the denoised spectrogram (Figure 6), we extract four types of features related to the gait pattern.

The *first* feature is the cepstral coefficient vector [29], which takes into account of the normalized energy over the whole frequency band. To reduce the inclusion of noise, we consider the area below the leg contour curve and above the frequency line of 20.05 kHz in Figure 7, since the massive amount of energy above the leg contour is mostly incurred by frequency leakage. With the measured gait cycle time (in Section 5.5), the Doppler signal in this area will be cut into a set of small pieces, each of a two-gait sample length covering the peak stride and mid stride that alternately appear [13]. Each two-gait sample $h[n]$ can be represented as follows [5]:

$$h[n] = u[n] \otimes v[n], \tag{5}$$

where $n$ represents a specific point on the discrete signal, $u[n]$ represents the pure Doppler signal generated by human motion, and $v[n]$ represents the vocal tract information which models the source of most interference. We apply the FFT to a two-gait sample as follows:

$$H\left(e^{j\omega}\right) = U\left(e^{j\omega}\right) V\left(e^{j\omega}\right), \tag{6}$$

where $\omega$ represents the angular frequency of the signal, $H(e^{j\omega})$, $U(e^{j\omega})$, and $V(e^{j\omega})$ represent the Fourier transformation of $h[n]$, $u[n]$, and $v[n]$, respectively. Then, a power spectrum is calculated by entry-wise multiplication of the FFT coefficients and conjugates. A log operator is applied to the power spectrum to build a log power spectrum, i.e.,

$$\log\left(\left|H\left(e^{j\omega}\right)\right|^2\right) = \log\left(\left|U\left(e^{j\omega}\right)\right|^2\right) + \log\left(\left|V\left(e^{j\omega}\right)\right|^2\right), \tag{7}$$

which shows that the power spectrum of Doppler signal and vocal tract information are linearly added together in the log power spectrum. We can separate the Doppler signal from the vocal tract information through **Inverse FFT** (**IFFT**):

$$c_h[n] = c_u[n] + c_v[n]. \tag{8}$$

As we only consider the real cepstrum, taking IFFT to log power spectrum is equivalent to performing FFT. Hence, the cepstrum can be viewed as the frequency decomposition of the log power spectrum. The left side is the low-frequency component, representing the frequency decomposition of the log power spectrum envelope, including the biometric information. The right side consists of the high-frequency components which represent the frequency decomposition of the log power spectrum details for the vocal tract information. Therefore, in Equation (8), $n$ represents a specific point on the cepstrum, $c_h[n]$ represents the cepstrum of the received signal, $c_u[n]$ is the left side of the cepstrum coming from the Doppler signal, and $c_v[n]$ is the right side of the cepstrum coming from vocal tract information. We pick up the top 100 points on the cepstrum's left side to form a cepstral coefficient vector, which can describe the energy change of Doppler shift in each two-gait sample.

The *second* feature is the spectrogram signature [41]. For each two-gait sample, we apply the 128 points FFT where the length of the Hamming window is equal to that of the respective two-gait sample. As a result, a 128 point FFT sequence will be built for each two-gait sample. The amplitude of each point represents the energy level on a specific frequency. We continue to normalize the energy of each FFT sequence into the same scale (i.e., between 0 and 1). After normalization, each sample is used to calculate the normalized energy by deriving the mean magnitudes of every five adjacent FFT points. This will result in a total of 124 energy points, among which we drop the last three and select 60 points with the interval of one to serve as the spectrogram signature of each sample.

The *third* feature is the ratio curve between the leg and torso speed curves. The leg and torso speed curves can be respectively calculated by using the frequency contours of the leg and the torso. Notably, the leg frequency, as discussed before, represents the Doppler shift frequency caused by leg movement, which quantifies the moving speed of the leg in each two-gait cycle. The torso frequency contour can be obtained from Equation (3). That is, we set the threshold value of $P(f, t)$ as 50% to derive the frequency [39], allowing us to get torso frequency of each two-gait sample. After getting the leg and torso frequency curves, we can apply Equation (1) to get the leg speed curve and torso speed curve, respectively, which together can characterize the motion of limbs during walking. Instead of directly using the leg speed curve and torso speed curve as features, we calculate the ratio values between each speed value on the leg speed curve and that on the torso speed curve, to get the ratio curve. Since the torso and leg are likely in the same direction even if a person changes walking directions, such a ratio curve will be relatively robust to the direction change. We save the values on this ratio curve into a fixed length vector to serve as our third feature.

The gait-cycle time, which quantifies the duration of a gait cycle, will be used as the *fourth* feature.

In the end, we obtain four categories of features from each two-gait cycle to represent an individual's unique gait pattern. We would like to note that the specific features of leg speed, step length, cadence, step width, angle, and gait phases are not directly extracted. However, these factors have significant effects on the Doppler signal. For example, the leg speed and step length directly affect the shape of the contours, and cadence impacts the length of contours. The step width and step angle affect the energy distribution of the Doppler signal, thereby signifying the contour pattern. In addition, the two-step period contains the periodic change in the gait phases. Therefore, we can conclude that the extracted features contain leg speed, step length, cadence, step width, angle, and gait phases information.

### 5.7 Recognition

Based on the features extracted from aforementioned steps, we are ready to use the detected gait patterns to serve the purpose of intrusion detection and individual recognition. Essentially, our system is based on gait recognition to distinguish between the gait patterns of intruders and family members.

The received acoustic signal will be processed by EchoSensor at each module with detailed steps as aforementioned to extract the features. Here, the **Principal Component Analysis** (**PCA**) [43] is applied to help reduce the dimensionality of the features by extracting the principal components and reducing the dimension of features from 222 to $N$. These features will be trained in the *Machine Learning Classifier module* using the **Support Vector Machine** (**SVM**) to acquire a unique model for each family member, so that any two-gait instance can be classified into two classes of self-gait and non-self-gait. With reduced dimensions, SVM can act with shorter training time and higher accuracy, compared with using the original data. We propose to use the LibSVM tool [3] with the **Radial Basis Function** (**RBF**) kernel in the training phase. The optimal values of parameters $n$ and $g$ for the RBF kernel are selected through grid search, which takes less than $10s$. It is worth noting that, to maintain the robustness of EchoSensor, we can enable it to periodically collect the two-gait instances of each family member so as to capture a rich set of gait patterns. Finally, the SVM calculates the fitness probability of a newly appeared two-gait instance belonging to the enrolled family members or not.

## 6 PERFORMANCE EVALUATION

In this section, we conduct experiments to evaluate the performance of EchoSensor in the real home environments.

### 6.1 Experiments

EchoSensor is implemented on a Huawei P20 smartphone with Android 8.1 and a Bluetooth speaker Bose soundlink Revolve. The smartphone works as the tone generator and controls the Bose soundlink Revolve speaker via Bluetooth connection to emit the ultrasonic signals at 20 kHz.[1] The built-in microphone on smartphone is used as the recorder. We set the sampling frequency as 44.1 kHz according to the Nyquist sampling theorem. The recorded Doppler data is sent to a laptop for further processing, with Matlab employed as the signal processing and machine learning tools.

**Data Collection.** Our experiments were conducted in 10 real family houses with a total of 24 participants. The number of family members in each household ranges from 1 to 4. There are

[1] The sound pressure at 1 m from the speakers will reduce to 36 dB, which will not cause side effects such as headache, dizziness, and pressure or pain in the ears.

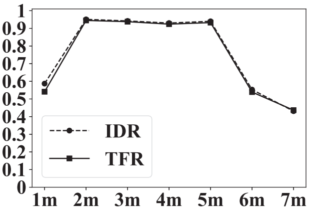


Fig. 9. IDR and TFR at different distances in home 1.

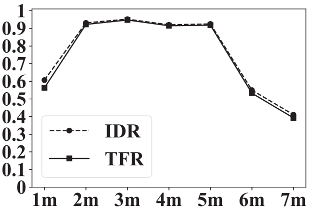


Fig. 10. IDR and TFR at different distances in home 2.

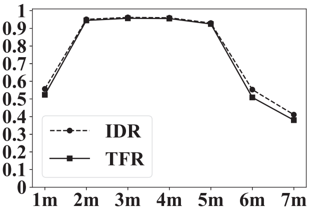


Fig. 11. IDR and TFR at different distances in home 3.

10 females and 14 males, with their ages ranging from 20 to 50. The experiments were approved by our University's IRB. Initially, the background noise was recorded at each house with no person moving. Each participant walked in his house toward and away from the speaker (up to 7 meters away) at his natural pace, repeatedly 40 times. In each house, two strangers were introduced as the intruders, each walking 20 round trips. Notably, each participant walked in their natural manner, not necessarily walking along the straight line to the speaker. The reflected Doppler signal was recorded by the smartphone and processed on the laptop, with 200 two-gait instances collected from each participant.

**Evaluation Metrics.** To measure the sensing capability of EchoSensor in detecting intruders and family members, we use two stringent and fine-grained metrics of gait instances identification accuracy: IDR and TFR. Specifically, IDR represents the ratio of a true intruder's two-gait instances detected over his total ones. TFR represents the ratio of all detected true family members' two-gait instances over all detected ones claimed to belong to family members. Notably, in the real-world deployment, a threshold (e.g., 50% or more) for the IDR can be set to safely trigger the intrusion alarm, as will be elaborated in Section 6.3.

## 6.2 Operational Distance

We first conduct experiments to identify the best operational distance of EchoSensor. For our collected data, we roughly truncate the two-gait instances at seven distance ranges, i.e., $(0, 1m)$, $(1m, 2m), \ldots, (6m, 7m)$. For each family member, we train seven SVM classifiers corresponding to each distance range. Figures 9–11 show the IDR and TFR at different distance ranges in three households, where $2m$ represents the distance range of $(1m, 2m)$ and the same for others. We can see that EchoSensor achieves the best gait instance detection accuracy at the distance range of 1 to 5 meters, with IDR and TFR both more than 92%. When the distance is more than 5 meters, both IDR and TFR quickly drop, with only 50% accuracy at 7 meters. The reason is that the energy of the reflected Doppler signals is too weak to be detected if the participant is far away from the speaker. Interestingly, we also observe that both IDR and TFR perform worse within 1 meter. This is due to the fact that within a close distance, most signals are reflected from the feet movement, hindering the microphone from capturing sufficient gait patterns for identification.

In summary, we can conclude that the appropriate operational distance for EchoSensor ranges from 1 to 5 meters for the accurate gait recognition, which is used as the basis to perform the following set of experiments. Figure 12 also affirms our conclusion, which exhibits the averaged IDR and TFR values for all 10 families. We believe such an operational distance is sufficient for EchoSensor to work in the smart home environments, where the audio devices typically are placed in the center of a room. Moreover, there may be multiple audio devices in each home, placed at different positions or rooms, which can all install EchoSensor and work together to cover the house's majority area.

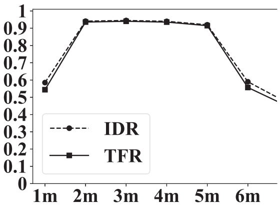


Fig. 12. Averaged IDR and TFR at different distances.

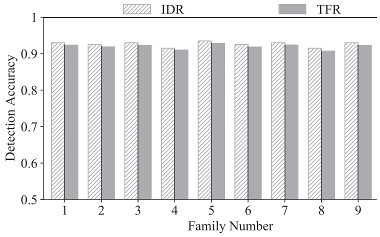


Fig. 13. Averaged IDR and TFR in all 10 families.

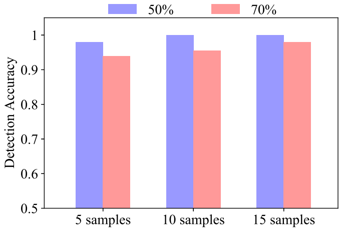


Fig. 14. Alarm accuracy under different numbers of samples and thresholds.

## 6.3 Overall Performance

Having identified the appropriate operational distances, we next present the performance of EchoSensor in all 10 households. In each household, we use each family member's data between 1 to 5 meters to train his SVM classifier. All family members' data will be considered as positive samples. Meanwhile, we take another three participants who are neither family members nor intruders as the negative samples in each SVM classifier's training. For recognition, we set a detection threshold as 0.5, which means that EchoSensor classifies the gait sample as a family member when the corresponding fitness probability is higher than 0.5; as an intruder, otherwise. The SVM classifier is trained with 5-fold cross-validation, where one-fifth of the family members' two-gait instances are used for testing and the remaining ones are used for training.

Figure 13 illustrates the IDR and TFR of EchoSensor in each home. The maximum IDR and TFR values are achieved at home 5, with 93.5% and 92.9%, respectively. The minimum IDR and TFR are 91.5% and 90.7%, respectively, at home 8. The averaged IDR and TFR values across all 10 families are 92.7% and 91.9%, respectively, demonstrating that EchoSensor can detect the subtle gait patterns of both intruders and family members with high accuracy.

*Notably, the IDR and TFR here represent the ratios of the detected two-gait instances to validate the robustness of our system. In the real-world deployment, a lower threshold for IDR (say 50%) is sufficient to trigger the intrusion detection alarm safely.* To claim an intruder, we can set an alarm threshold, which specifies the percentage of two-gait samples detected that is unrecognized. We then define an alarm accuracy as the possibility of one individual truly identified as an intruder given a set of his two gait samples. Figure 14 presents the alarm accuracy with various number of detected two-gait samples when setting the alarm threshold to be 50% and 70%. We observe that alarm accuracy reaches 98% and 94% with the thresholds of 50% and 70%, respectively, with 5 two-gait samples. This result demonstrates that 5 two-gait samples are sufficient to trigger an alarm. When using 15 two-gait samples, both values of alarm accuracy increase to 100%, given alarm threshold values of 50% and 70%, respectively. Despite a fully reliable detection, it may not be necessary to use 15 two-gait samples for the practical concern of space limitation. We prefer to use 5 two-gait samples to set up our intrusion detection alarm system in practice.

## 6.4 Impact of Extracted Features

To signify the effectiveness of four features, we conduct another set of experiments, each including three features, with Figure 15 showing the respective performance. When all four features are included, EchoSensor achieves the highest values of IDR and TFR with 93% and 91.9%, respectively. When one of the cepstral vector, spectrogram signature, leg contour, and torso speed and gait cycle time pair is excluded, the resulted IDRs are 85.5%, 86.5%, 88%, and 80.5%, respectively, and the TFRs decrease to 84.7%, 85.7%, 87.3%, and 79.5%, respectively. We observe the leg contour has relatively

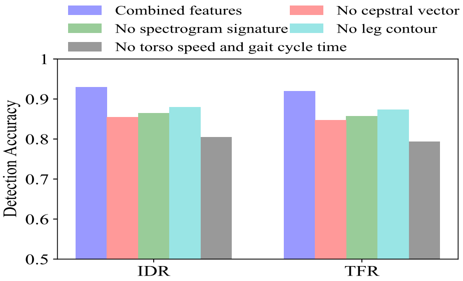


Fig. 15. Different features.

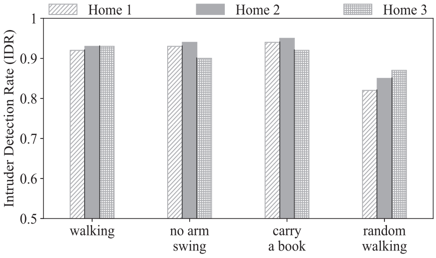


Fig. 16. Different actions.

slight impact on the detection accuracy. However, it is important to help us determine gait-cycle time as discussed in Section 5.5.

### 6.5 Impact of Person Actions

We then consider some unusual behaviors of an intruder when he invades a house. Four different actions are considered: (1) walking at normal pace; (2) keeping arms stable to avoid touching something accidentally; (3) carrying a book stolen from the house; (4) randomly walking in front of the Echosensor to look for objects in the victim's home. We conducted our experiments in three homes, where two intruders walked with four different actions. Figure 16 shows the averaged IDRs for each action in each home. It is seen that EchoSensor can still achieve more than 90% of IDRs under two unusual actions (no arm swing and carrying a book) in all three homes. This demonstrates that EchoSensor is robust to different actions of an intruder: the gait pattern may be impacted by different actions, but it is still distinctive from family members and thus can be detected. When the intruder is walking in the house, the EchoSensor can still achieve more than 80% of IDRs in all three homes, leading to a higher alarm accuracy as we have discussed in Section 6.3.

### 6.6 Impact of Mimicry and Saturate Attack

For mimicry attack, we consider the scenarios where an intruder may be familiar with the family members and try to mimic their walking patterns to obfuscate the system. Four participants were asked to try their best to walk in the same way, such as walking at the same speed and swinging their arms in the same amplitude. We select three participants to be the family members and the remaining one to be the intruder. Figure 17 shows the results of IDRs and TFRs under four different combinations, treated as four families. We obverse that the mimicry attack has a negligible impact on the intrusion detection accuracy of EchoSensor system. This demonstrates that EchoSensor is able to capture rich fine-grained information in gait patterns and is robust to mimicry attacks.

For saturate attack, we also choose one participant as the intruder and the other three as family members, with the total of four combinations, considering as four families. In each family, the intruder carries a speaker playing 20 khz ultrasonic sound to obfuscate the system. Figure 18 shows the results of IDRs and TFRs in four families under the saturate attack. It is observed that EchoSensor can still maintain the high IDRs and TFRs, demonstrating its robustness to saturate attack.

### 6.7 Impact of Device Placement

As the devices may be deployed at different places, such as on the floor, on a coffee table, or on a chair, we now examine the impact of device placement, especially from the perspective of height, on the performance of EchoSensor. Figure 19 shows the results of IDRs and TFRs at four different heights (0 m, 0.2 m, 0.5 m, and 1 m). With a low height of 0 m and 0.2 m, the IDR and TFR can

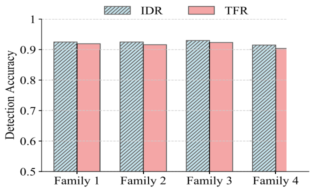


Fig. 17. Mimicry attack.

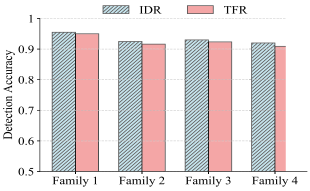


Fig. 18. Saturate attack.

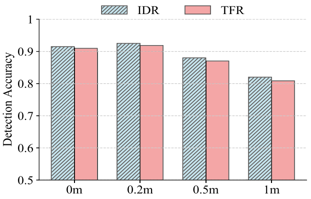


Fig. 19. Different heights.

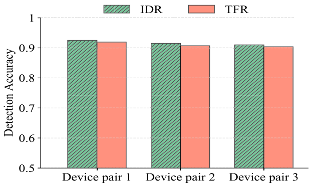


Fig. 20. Different devices.

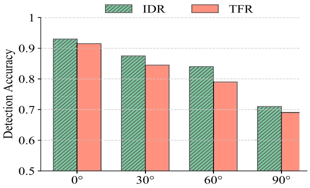


Fig. 21. Different angles.

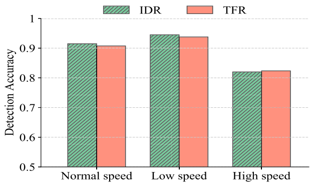


Fig. 22. Different speeds.

be higher than 90%. When the height grows to 0.5 m and 1 m, the IDR drops to 88% and 82%, respectively, and the TFR drops to 87% and 80.8%, respectively. Notably, even when the IDR and TFR at 1 m reduce to 80%, EchoSensor can still perform well in triggering intrusion alarm as we have discussed in the end of Section 6.3, by setting the threshold of intrusion detection alarm as 70% or 80%.

## 6.8 Impact of Different Devices

We next evaluate the performance of EchoSensor implemented on different devices. Three different smartphones and Bluetooth speakers are used to implement our experiments, where we pair them as the Huawei P20 and Bose soundlink (device pair 1), Galaxy s9 plus and Amazon Echo (device pair 2), iPhone 8 and Apple Homepod (device pair 3). Four participants are involved in this experiment, with two treated as the family members and two as the intruders. For each participant, 200 two-gait samples are collected from each implementation. As shown in Figure 20, EchoSensor achieves averaged IDRs of 92.5%, 91.5%, 92% and TFRs of 91.9%, 90.7%, 90.4%, respectively, for implementations on three devices. Clearly, despite slight differences in the IDRs, EchoSensor exhibits good performance regardless of the types of speakers.

## 6.9 Impact of Walking Direction and Speed

We exam the performance of EchoSensor when the walking direction and the orientation of microphone have a certain angle. Figure 21 shows the results of IDRs and TFRs when such an angle varies from 0, 30, 60, to 90 degrees. Both IDR and TFR degrade with the growth of angles. When the angle increases to 60 degrees, they drop to 84% and 79%, respectively. When the angle increases to 90 degrees, they drop to 71% and 69%, respectively. The reason is that a larger angle will reduce the Doppler reflection, which leads to fewer Doppler signals captured by the microphone and thus lower detection accuracy. However, the leg and arm movement still could be extracted to a certain extent, making the intruder's patterns distinctive from these of the family members.

We further evaluate EchoSensor when intruders walk at different speeds. In Figure 22, compared with normal speed, both IDR and TFR increase (from 91.5% to 94%, and from 90.7% to 93%) when the intruders walk slowly. The reason is that high-quality spectrogram can be obtained from slow walking, which allows EchoSensor to extract more clear patterns. In contrast, when walking fast, the performance of IDR and TFR drop to 82% and 82.1%, respectively.

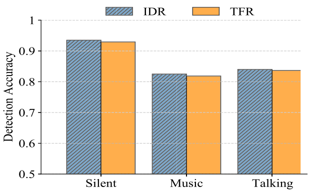


Fig. 23. Different noise.

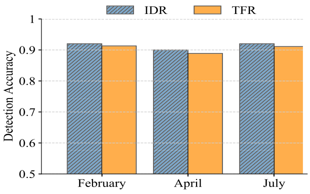


Fig. 24. Different time.

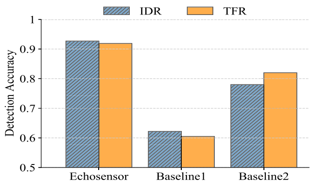


Fig. 25. EchoSensor v.s Baselines.

## 6.10 Impact of Noises

We next evaluate the robustness of EchoSensor when there exists a certain level of noise in smart home. Theoretically, the low-frequency noises at home, such as music playing and people talking, can be simply removed by low-pass filtering. However, due to the frequency leakage [16], these audible noises may impact the performance of EchoSensor. Hence, we conducted the following experiment: the two-gait samples of family members were recorded in a silent environment, while an intruder's two-gait samples were recorded in three scenarios: silent, music playing, and people talking with loud footstep noise. In noisy environments, both the IDRs and TFRs drop below 85%, as shown in Figure 23, because the frequency leakage of such sounds significantly blurs the spectrogram. However, they remain higher than 90% when we choose 5 two-gait samples as a group, which demonstrates our system is reliable in a noisy environment.

## 6.11 Impact at Different Time

We then evaluate the time consistency of the gait pattern regarding our EchoSensor's performance. We collect 100 gait samples from one user (family member) in February 2022 to build the EchoSensor classifier. His gait samples are then collected in April and July for testing, while one user from dataset described in Section 6.1 is considered as the intruder. Due to changes in the weather, the clothes of that user change at different times. Figure 24 shows the performance at each month. In each month, the IDR is 92%, 90%, and 92%, and the TFR is 91.33%, 88.9%, and 91.13%. The stable result shows the gait pattern is resilient to time and clothing changes.

## 6.12 Comparison with Other Approaches

The previous gait identification system was not aimed at intrusion detection, and their system working flow differs from ours. Specifically, the existing acoustic-based gait detection solution [13, 42, 50] did not extract the fine-grained features for identification. Instead, they relied on professional devices and neural network models to compute the features by using the whole spectrogram as input. Reference [45] implemented with the COTS device; however, they do not consider the reflection in the different environments and do not contain the interference cancellation part. We call the [13, 42, 50] as baseline 1 while [45] as baseline 2. We implement their approaches on the same devices as that of EchoSensor without using professional devices. We used their original setting and trained the gait identification system to compare with EchoSensor. From Figure 25, we can observe that EchoSensor can achieve the averaged IDR and TFR of 92.7% and 91.9%, while baseline 1 and baseline 2 are only of 62.2% and 60.5%, and of 78% and 82%, respectively. The major performance dropping for the two baselines is due to the missing of interference Cancellation process or the fine-grained feature extraction process. This experiment demonstrates the importance and necessity of fine-grained signal processing and feature extraction developed in EchoSensor for accurate intrusion detection.

# 7 DISCUSSION

In this section, we conduct security analysis and discuss the potential improvement of EchoSensor.

## 7.1 Security Analysis

We analyze three types of attacks: mimicry attack, saturate attack, and replay attack.

**Mimcry attack.** This attack requires the intruder to imitate the gait pattern of the victim, for example, the same speed and the same behavior when walking However, it is extremely difficult to accurately imitate the family members' gait pattern in practice. Our experiment in Section 6.6 validated that EchoSensor is robust to this attack.

**Saturate attack.** The intruder is assumed to know the principle of EchoSensor and carries a device playing the high-energy ultrasonic signal to saturate the microphone when entering the house, making her gait signal be drown in noise. EachoSensor can capture the mixed signal, including both the reflected Doppler signals from the intruder's gait and the playing ultrasonic signal. However, this playing ultrasonic signal will be treated as the noise for removal in EchoSensor, leaving only intruder's Doppler signal for recognition. Our experiment in Section 6.6 validated that saturate attack does not work in EchoSensor.

**Replay attack.** The intruder has the chance to record the Doppler gait signal secretly from victims without being noticed. Our EchoSensor is vulnerable to this type of attack since the recorded signal has similar characteristics of victim's Doppler shift signal, which can fool the EchoSensor. But notably, when an intruder enters into the house to replay the audio, EchoSensor may capture his gait patterns and trigger an alarm before he performs replay attack.

## 7.2 Limitations and Discussions

EchoSensor provides a proof of concept to successfully demonstrate the feasibility of employing the existing smart home devices to conduct acoustic sensing for gait patterns to serve the purpose of intrusion detection. Yet, this system has some limitations to be improved in our future work.

First, EchoSensor's IDR and TFR are sensitive to the change in walking direction and speed, where the captured Doppler signals will change accordingly. Specifically, when the walking direction shifts over 60 degrees, the IDR and TFR of EchoSensor will significantly degrade. The arbitrate walking direction indeed limits EchoSensor's practical deployment based on the current design. The more robust solutions are expected to make EchoSensor more practical. To overcome such a limitation, two possible solutions can be explored. First, we can enrich the training set to include data samples from different directions and cover all possible scenarios. Second, when conducting the detection, we can treat the multiple consecutive gait samples as a group rather than a single gait sample for identification. It is expected that a group of gait samples collected from a user walking in an irregular trajectory will be collected in different directions. Although gait samples at a high angle cannot be recognized properly, those with a low angle can still be recognized. We can set a threshold when the number of recognized gait samples belonging to a family member exceeds the threshold. We will treat the whole group of samples to be from the same family member. Notably, a small threshold is sufficient for accurate detection. In addition, we can also leverage multiple audio devices to cover all possible angles in each room. This is feasible in current smart home environments, given there typically exist multiple voice devices.

Second, when the device's height is more than 0.5 meters, most of the Doppler signals will be reflected from the arm swing, which drowns the Doppler signals reflecting from the gaits. Under this scenario, new signal processing solutions that can remove the strong Doppler signals from the arm swing while extracting the weak Doppler signals from the gaits are needed.

Third, the detection range of EchoSensor can be up to 5 meters, which works in general house environments. Notably, the intrusion mostly happens at the entry door or window, so our EchoSensor can be deployed at a specific location, which is no more than 5$m$ from the entry door or windows. Hence, 5$m$ detection range should be sufficient for EchoSensor's practical deployment in most houses. Definitely, the longer detection distance can further enhance our EchoSensor's capability. To further extend its applications into the building with larger rooms or halls, new signal processing techniques have to be developed to remove the interference from weak Doppler signals, and more fine-grained features need to be extracted to mine the useful information.

Fourth, if an intruder's moving speed is less than 0.33 m/s, EchoSensor cannot detect him. To cover the low movement case, we can set a more narrow range to mitigate the direct transmission noises. In fact, our empirical studies have exhibited that setting a narrow range from 19.99 kHz and 20.01 kHz is sufficient to remove the majority of noises. Corresponding to this narrow range, if an intruder's moving speed is above 0.08 m/s, EchoSensor can always detect him with high performance.

In the end, the current EchoSensor only works for detecting one person. When multiple people appear, their Doppler signals will mix together, making their gaits patterns not distinguishable. In our future work, we plan to employ multiple devices or multiple tone frequencies to enrich the Doppler signals to include much richer gait information that can be distinguished for individual recognition, in the appearance of multiple people at the same time. Another working direction is to implement a powerful beamforming algorithm so that we can directly separate the signal from different targets by the direction difference. Microphone array can also be utilized to enhance the received signals of interests and allow us to develop more powerful signal processing technology for us. In addition, deep learning approaches will also be explored to boost classifier's recognition capability in our future work. All these potential technologies will be explored in our future works.

## 8 CONCLUSION

In this article, we proposed a novel intrusion detection system, called EchoSensor, that only relies on the speaker and microphone, ubiquitously available in smart home audio devices for individual gait recognition. Through the design of a two-stage interference cancellation scheme, EchoSensor can acquire the pure signals of Doppler shift, for the mining of enclosed gait information. New signal processing techniques have been developed and applied to the Doppler effect spectrogram to perform the fine-grained analysis of gait patterns by detecting the starting point of walking, estimating gait cycle time, and extracting effective features so that each individual's unique gait patterns can be distinguished. With the extracted fine-grained features, EchoSensor employs machine learning classifiers for conducting individual recognition tasks. Implemented on existing smart home devices without extra hardware deployment or modification, EchoSensor is economical compared with the state-of-the-art home security systems. We deployed our EchoSensor system in the real home environment and conducted extensive experiments to evaluate its performance from different perspectives. Experimental results demonstrated that EchoSensor, economical, and lightweight, can achieve competitive intrusion detection accuracy in smart home environment.

## REFERENCES

[1] M. Umair Bin Altaf, Taras Butko, and Biing-Hwang Fred Juang. 2015. Acoustic gaits: Gait analysis with footstep sounds. *IEEE Transactions on Biomedical Engineering* 62, 8 (2015), 2001–2011.

[2] Imed Bouchrika, Michaela Goffredo, John Carter, and Mark Nixon. 2011. On using gait in forensic biometrics. *Journal of Forensic Sciences* 56, 4 (2011), 882–889.

[3] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2, 3 (2011), 27.

[4] Jagmohan Chauhan, Yining Hu, Suranga Seneviratne, Archan Misra, Aruna Seneviratne, and Youngki Lee. 2017. BreathPrint: Breathing acoustics-based user authentication. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 278–291.

[5] Gilles Degottex. 2010. *Glottal Source and Vocal-tract Separation*. Ph.D. Dissertation.

[6] Davrondzhon Gafurov. 2007. A survey of biometric gait recognition: Approaches, security, and challenges. In *Proceedings of the Annual Norwegian Computer Science Conference*. Annual Norwegian Computer Science Conference Norway, 19–21.

[7] Davrondzhon Gafurov, Kirsi Helkala, and Torkjel Søndrol. 2006. Biometric gait authentication using accelerometer sensor. *JCP* 1, 7 (2006), 51–59.

[8] Daniel Graham, George Simmons, David T. Nguyen, and Gang Zhou. 2015. A software-based sonar ranging sensor for smart phones. *IEEE Internet of Things Journal* 2, 6 (2015), 479–489.

[9] Sidhant Gupta, Daniel Morris, Shwetak Patel, and Desney Tan. 2012. Soundwave: Using the Doppler effect to sense gestures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1911–1914.

[10] Ju Han and Bir Bhanu. 2006. Individual recognition using gait energy image. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28, 2 (2006), 316–322.

[11] Spencer Ives. 2018. Parks Associates Predicts about 27 Percent of U.S. Households to have Security by 2021. Retrieved March 15, 2019 from http://www.securitysystemsnews.com/article/parks-associates-predicts-about-27-percent-us-households-have-security-2021

[12] Jam Jenkins and Carla Ellis. 2007. Using ground reaction forces from gait analysis: Body mass as a weak biometric. In *Proceedings of the International Conference on Pervasive Computing*. Springer, 251–267.

[13] Kaustubh Kalgaonkar and Bhiksha Raj. 2007. Acoustic Doppler sonar for gait recogination. In *Proceedings of the 2007 IEEE Conference on Advanced Video and Signal Based Surveillance*. IEEE, 27–32.

[14] A. J. Lawrance and P. A. W. Lewis. 1977. An exponential moving-average sequence and point process (EMA1). *Journal of Applied Probability* 14, 1 (1977), 98–113.

[15] Hoang Thanh Le, Son Lam Phung, and Abdesselam Bouzerdoum. 2018. Human gait recognition with micro-doppler radar and deep autoencoder. In *Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR'18)*. IEEE, 3347–3352.

[16] 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.

[17] Hong Li, Yunhua He, Limin Sun, Xiuzhen Cheng, and Jiguo Yu. 2016. Side-channel information leakage of encrypted video stream in video surveillance systems. In *Proceedings of the IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications*. IEEE, 1–9.

[18] Jinyang Li, Zhenyu Li, Gareth Tyson, and Gaogang Xie. 2020. Your privilege gives your privacy away: An analysis of a home security camera service. In *Proceedings of the IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 387–396.

[19] Li Lu, Jiadi Yu, Yingying Chen, Hongbo Liu, Yanmin Zhu, Yunfei Liu, and Minglu Li. 2018. Lippass: Lip reading-based user authentication on smartphones leveraging acoustic signals. In *Proceedings of the IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 1466–1474.

[20] Yongsen Ma, Gang Zhou, and Shuangquan Wang. 2019. WiFi sensing with channel state information: A survey. *ACM Computing Surveys* 52, 3 (2019), 46.

[21] Roger G. T. Mello, Liliam F. Oliveira, and Jurandir Nadal. 2007. Digital Butterworth filter for subtracting noise from low magnitude surface electromyogram. *Computer Methods and Programs in Biomedicine* 87, 1 (2007), 28–35.

[22] Soumik Mondal, Anup Nandy, Pavan Chakraborty, and G. C. Nandi. 2012. Gait-based personal identification system using rotation sensor. *Journal of Emerging Trends in Computing and Information Sciences* 3, 2 (2012), 395–402.

[23] Alvaro Muro-De-La-Herran, Begonya Garcia-Zapirain, and Amaia Mendez-Zorrilla. 2014. Gait analysis methods: An overview of wearable and non-wearable systems, highlighting clinical applications. *Sensors* 14, 2 (2014), 3362–3394.

[24] 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, 1515–1525.

[25] Rajalakshmi Nandakumar, Alex Takakuwa, Tadayoshi Kohno, and Shyamnath Gollakota. 2017. Covertband: Activity information leakage using music. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 87.

[26] Thanh Trung Ngo, Yasushi Makihara, Hajime Nagahara, Yasuhiro Mukaigawa, and Yasushi Yagi. 2014. The largest inertial sensor-based gait database and performance evaluation of gait-based personal authentication. *Pattern Recognition* 47, 1 (2014), 228–237.

[27] G. Nicholas. 2010. *College Physics: Reasoning and Relationships*. Brooks/Cole.

[28] Mark S. Nixon and John N. Carter. 2006. Automatic recognition by gait. *Proceedings of the IEEE* 94, 11 (2006), 2013–2024.

[29] Alan Oppenheim and Ronald Schafer. 1968. Homomorphic analysis of speech. *IEEE Transactions on Audio and Electroacoustics* 16, 2 (1968), 221–226.

[30] Robert J. Orr and Gregory D. Abowd. 2000. The smart floor: A mechanism for natural user identification and tracking. In *Proceedings of the CHI'00 Extended Abstracts on Human Factors in Computing Systems*. ACM, 275–276.

[31] Shijia Pan, Ningning Wang, Yuqiu Qian, Irem Velibeyoglu, Hae Young Noh, and Pei Zhang. 2015. Indoor person identification through footstep induced structural vibration. In *Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications*. ACM, 81–86.

[32] Yanzhi Ren, Yingying Chen, Mooi Choo Chuah, and Jie Yang. 2013. Smartphone-based user verification leveraging gait recognition for mobile healthcare systems. In *Proceedings of the 2013 IEEE International Conference on Sensing, Communications and Networking (SECON)*. IEEE, 149–157.

[33] Chandra Rishi and Huffman Scott. 2018. How Google Home and the Google Assistant Helped you Get More Done in 2017. Retrieved March 15, 2019 from https://www.blog.google/products/assistant/how-google-home-and-google-assistant-helped-you-get-more-done-in-2017

[34] Claude Elwood Shannon. 1998. Communication in the presence of noise. *Proceedings of the IEEE* 86, 2 (1998), 447–457.

[35] 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 ACM Annual International Conference on Mobile Computing and Networking (MobiCom'18)*. 591–605.

[36] Thiago Teixeira, Deokwoo Jung, Gershon Dublon, and Andreas Savvides. 2009. PEM-ID: Identifying people by gait-matching using cameras and wearable accelerometers. In *Proceedings of the 2009 3rd ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC)*. IEEE, 1–8.

[37] Inc TranSafety. 1997. Study compares older and younger pedestrian walking speeds. *Road Management and Engineering Journal* (1997). https://web.archive.org/web/20090703084118/http://www.usroads.com/journals/p/rej/9710/re971001.htm. Access date March 15, 2019.

[38] Junia Valente, Keerthi Koneru, and Alvaro Cardenas. 2019. Privacy and security in Internet-connected cameras. In *Proceedings of the 2019 IEEE International Congress on Internet of Things (ICIOT'19)*. IEEE, 173–180.

[39] Ph Van Dorp and F. C. A. Groen. 2008. Feature-based human motion parameter estimation with radar. *IET Radar, Sonar and Navigation* 2, 2 (2008), 135–145.

[40] Ashley D. Waite. 2002. *Sonar for Practising Engineers*. John Wiley and Sons.

[41] Wei Wang, Alex X. Liu, and Muhammad Shahzad. 2016. Gait recognition using wifi signals. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 363–373.

[42] Yingxue Wang, Yanan Chen, Md Zakirul Alam Bhuiyan, Yu Han, Shenghui Zhao, and Jianxin Li. 2018. Gait-based human identification using acoustic sensor and deep neural network. *Future Generation Computer Systems* 86 (2018), 1228–1237.

[43] Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems* 2, 1-3 (1987), 37–52.

[44] Dan Wu, Daqing Zhang, Chenren Xu, Hao Wang, and Xiang Li. 2017. Device-free WiFi human sensing: From pattern-based to model-based approaches. *IEEE Communications Magazine* 55, 10 (2017), 91–97.

[45] 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.

[46] Yunze Zeng, Parth H. Pathak, and Prasant Mohapatra. 2016. WiWho: Wifi-based person identification in smart spaces. In *Proceedings of the 15th International Conference on Information Processing in Sensor Networks*. IEEE.

[47] Jin Zhang, Bo Wei, Wen Hu, and Salil S. Kanhere. 2016. Wifi-id: Human identification using wifi signal. In *Proceedings of the 2016 International Conference on Distributed Computing in Sensor Systems (DCOSS)*. IEEE, 75–82.

[48] Linghan Zhang, Sheng Tan, and Jie Yang. 2017. Hearing your voice is not enough: An articulatory gesture-based liveness detection for voice authentication. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 57–71.

[49] Yuting Zhang, Gang Pan, Kui Jia, Minlong Lu, Yueming Wang, and Zhaohui Wu. 2015. Accelerometer-based gait recognition by sparse representation of signature points with clusters. *IEEE Transactions on Cybernetics* 45, 9 (2015), 1864–1875.

[50] Zhaonian Zhang, Philippe O. Pouliquen, Allen Waxman, and Andreas G. Andreou. 2007. Acoustic micro-Doppler radar for human gait imaging. *The Journal of the Acoustical Society of America* 121, 3 (2007), EL110–EL113.

[51] Peijun Zhao, Chris Xiaoxuan Lu, Jianan Wang, Changhao Chen, Wei Wang, Niki Trigoni, and Andrew Markham. 2019. mid: Tracking and identifying people with millimeter wave radar. In *Proceedings of the 2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS'19)*. IEEE, 33–40.

Received 17 February 2022; revised 14 January 2023; accepted 25 July 2023