HiddenTag: Enabling Person Identification Without Privacy Exposure

QIU Chen[1], DAI Tao[2], GUO Bin[1], YU Zhiwen[1], LIU Sicong[1]

(1. Northwestern Polytechnical University, Xi'an 710072, China;
2. Chang'an University, Xi'an 710064, China)

Abstract: Person identification is the key to enable personalized services in smart homes, including the smart voice assistant, augmented reality, and targeted advertisement. Although research in the past decades in person identification has brought technologies with high accuracy, existing solutions either require explicit user interaction or rely on images and video processing, and thus suffer from cost and privacy limitations. In this paper, we introduce a device-free personal identification system – HiddenTag, which utilizes smartphones to identify different users via profiling indoor activities with inaudible sound and channel state information (CSI). HiddenTag sends inaudible sound and senses its diffraction and multi-path reflection using smartphones. Based upon the multi-path effects and human body absorption, we design suitable sound signals and acoustic features for constructing the human body signatures. In addition, we use CSI to trigger the system of acoustic sensing. Extensive experiments indicate that HiddenTag can distinguish multi-person in 10 – 15 s with 95.1% accuracy. We implement a prototype of HiddenTag with an online system by Android smartphones and maintain 84% – 90% online accuracy.

Keywords: person identification; acoustic sensing; CSI; smart home

DOI: 10.12142/ZTECOM.202103002

https://kns.cnki.net/kcms/detail/34.1294.TN.20210819.1211.001.html, published online August 19, 2021

Manuscript received: 2021-06-15

Citation (IEEE Format): C. Qiu, T. Dai, B. Guo, et al., "HiddenTag: enabling person identification without privacy exposure," *ZTE Communications*, vol. 19, no. 3, pp. 03 – 12, Sept. 2021. doi: 10.12142/ZTECOM.202103002.

1 Introduction

Numerous applications are enabled with the realization of smart living environments and Internet of Things (IoT). Person identification is essential for smart home services, such as real-time recommendations on TV and human-machine interactions in video games[1 – 2]. Based on the personalized applications, users can obtain desirable services pervasively[3 – 5]. Therefore, an accurate, light-training and real-time person identification approach is needed.

Existing person identification mechanisms have many limitations that prevent them from being adopted pervasively. One of the biggest limitations is that they are often intrusive to users' privacy. Camera and computer vision based solutions can recognize different persons effectively, but unfortunately users' faces, gestures and other information may be exposed to others[6 – 8]. For example, monitoring a person's face when she/he is sitting on the sofa and walking in the hallway may cause privacy concerns.

Moreover, many person identification methods need a user to do extra work to help recognize the user. For example, smart speakers such as Amazon Echo and Google Home can identify users by their voiceprints. This approach requires users to speak to trigger recognition[9], which is a reactive solution. We therefore ask the question: can we identify users with-

out asking them to do any additional work and preserve their privacy?

To this end, we introduce HiddenTag, a new device-free person recognition system without pre-installed infrastructure or additional sensors. Only with the built-in smartphone, as shown in Fig. 1, the acoustic sender provides the high frequency sound (18 – 21 kHz) from which people cannot hear. When a user enters the smart environment, the user can keep the normal activities, such as walking, standing, and other types of human activities, and all these can be profiled by an acoustic receiver on off-the-shelf mobile devices. Based on the multi-path effects and bodies absorption in experimental scenarios, HiddenTag constructs the acoustic signatures for different persons. We design high frequency based features and enrich these features by utilizing sweeping and multi-tone techniques. Besides, we explore channel state information (CSI) to detect the human body and trigger the person identification approach. By leveraging machine learning models, our system recognizes different users efficiently in smart home environments. Case studies show the online identification can reach 90.2% accuracy and the corresponding offline group achieves 96.0% accuracy.

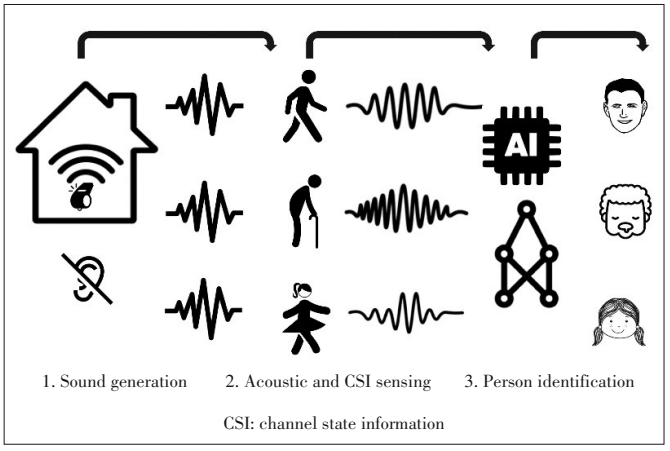
In addition, we pre-trained some common types of noises in the learning model and made HiddenTag more robust to noises. According to the collected historical data and temporal correlation feature, our system further calibrates some errors by using the proposed prediction model. Related SmartThings such as smart LED bulbs and media players are able to provide personalized services based on classification results. This paper makes the following contributions: 1) To the best of our knowledge, HiddenTag is the first high frequency (18 – 21 kHz) acoustic sensing solution for person identification; 2) HiddenTag has introduced sweeping and multi-tone techniques to enrich feature spaces. Adding common types of noises makes HiddenTag robust to real environment noises; 3) HiddenTag is implemented both online and offline. The proposed offline system achieves 96.2% accuracy with four users and the corresponding online system reaches 90.2% accuracy.

The rest of the paper is organized as follows. Section 2 introduces the system design. Experiments and simulations are shown in Section 3. Section 4 further discusses the evaluations. We provide related work and comparison in Section 5. Conclusions and future work are in Section 6.

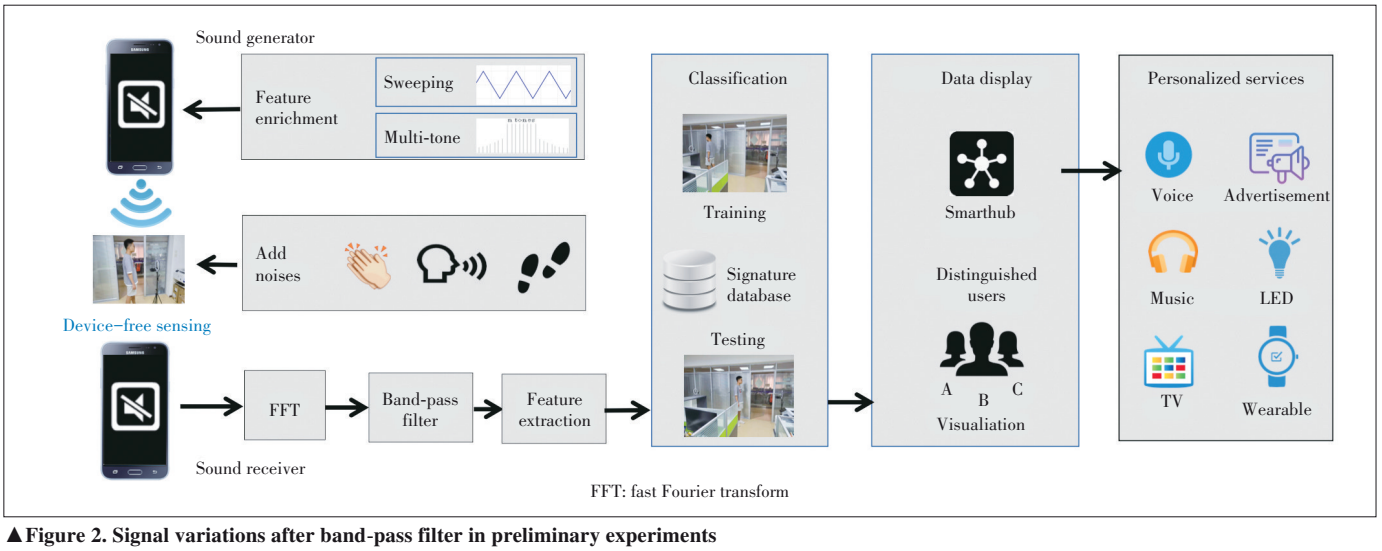
2 System Design

2.1 Overview of Our Approach

• HiddenTag employs existing smartphones without complex hardware modifications. The procedure of HiddenTag is illustrated in Fig. 2, where HiddenTag is a device-free system based on acoustic sensing. Off-the-shelf smartphones send high frequency (18 – 21 kHz) sound signals via speakers. The sound emitter can select one from the following three models: single-tone model, multi-tone model, and sweeping model. Af-



▲Figure 1. Concept view of HiddenTag



▲Figure 2. Signal variations after band-pass filter in preliminary experiments

fected by the user's indoor activities, the acoustic signals are changed in the propagation channels. Receivers of HiddenTag sense the varied acoustic information by microphones. By leveraging a band-pass filter, our system only processes the sound in the frequency range of 18 kHz and 21 kHz, which cannot be heard by human beings.

• In the training phase, based on feature engineering, the system trains different users and labels corresponding data. In the testing phase, HiddenTag adopts a band-pass filter to reduce noises. Classifiers based on the machine learning model are built to identify different users in smart home environments. Further, HiddenTag implements various personalized services (smart LED, music, smart TV, etc.) relying on the results of person identification.

2.2 Preliminaries

The fundamental idea of HiddenTag is that users can be recognized by their acoustic signatures. When the recorder receives acoustic signals, different degradations occur at different frequencies due to frequency selective fading. Additionally, multi-path effects, diffraction, and reflection also impact the acoustic signals. Once users walk in an indoor environment, walking activities and human bodies cause unique multi-path effects and body absorption. Fig. 3 illustrates the causes of such attenuation.

To verify this perspective, we conduct preliminary experiments in an empty room, the size of which is 5 m×5 m. We employ two Huawei Mate 30 smartphones as the acoustic sender and the receiver. The heights of the sender and receiver are 75 cm. The acoustic sender generates sounds frequencies from 18 kHz to 21 kHz. The sampling frequency is 48 kHz. In sweeping mode, the sweeping period is 0.02 s.

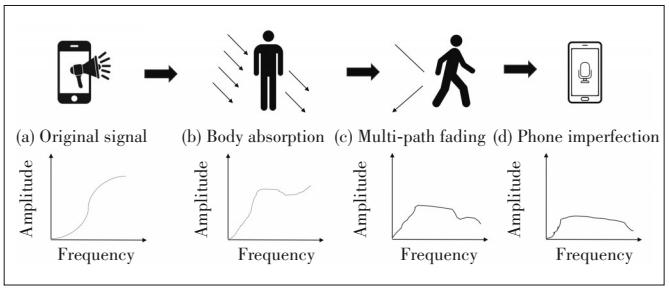
We record acoustic data for three control groups. In the beginning, the experimental room is empty. In the following two groups, User 1 and User 2 enter the room and walk around the sender and receiver.

As shown in Fig. 4, the x-axis indicates the time of the experiment and the y axis refers to the range of sound frequency. We conclude that for each control group, the power distributions on the different spectrums are different, and less power is distributed on the spectrum when there are users compared with the empty group.

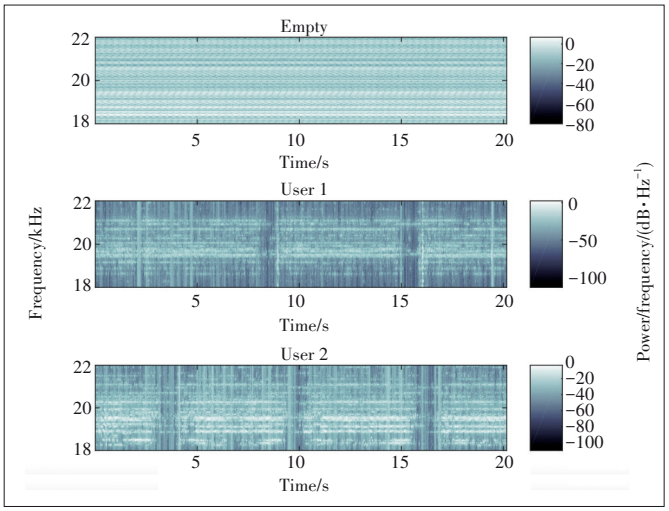
Therefore we design an approach that leverages the different signatures to identify users in the following.

2.3 Sending Inaudible Sound

As we explore the inaudible sound that can be generated from built-in smartphones, choosing parameters for sound generation is a challenge. According to our experimental results and literature, only generating single-tone acoustic signals between 18 kHz and 21 kHz is difficult to support accurate person identification because of the limited information. The feature space is constrained by a fixed sending frequency.



▲ **Figure 3. Signal variations after band-pass filter in preliminary experiments**



▲ **Figure 4. Time-spectral comparison for different ambient mediums**

$$S(t) = A \cdot \sin(2\pi f_0(t) \cdot t) . \quad (1)$$

As shown in Eq. (1), $S(t)$ is the amplitude value of the sin wave, and f_0 is the frequency of the sound wave that we send. If f_0 is a fixed value, the value of $S(t)$ can only reflect the wave at a certain frequency, which means that we do not utilize the inaudible sound on smartphones efficiently. Therefore, we introduce two other models, namely the sweeping model and the multi-tone model, to improve the identification accuracy by enriching feature space.

1) Sweeping model: We propose periodic frequency sweeping from 18 kHz to 21 kHz and set sampling frequency as 48 kHz. Consequently, the frequencies change quickly and cover all the frequencies from 18 kHz to 21 kHz in a short time period. This selection makes the generated sound inaudible for most people, but enriching the feature space for acoustic sensing.

$$f_0(t) = f_l + (f_u - f_l) \times \Delta t / t_d . \quad (2)$$

Different from Eq. (1) where f_0 is a fixed value, the value of f_0 is determined by Eq. (2) in the sweeping model. f_u and f_l indicate the upper bound and lower bound of the sweeping range. t_d is the duration of each sweeping period. Δt is the increment of the current time. As a result, the feature space of sweeping mode includes the data information from different

frequencies.

2) Multi-tone model: Even though the sweeping model includes different sound frequencies in a certain time period, for a specific time point, it can only emit a fixed frequency. In this subsection, we propose a multi-tone model. The sender provides more than one sound wave at the same time. The sender emits inaudible sound waves composed of multiple frequencies. Each component of the synthetic sound represents one sound wave at the designed frequency. Consequently, the multi-tone model enables the opportunity to cover more frequencies simultaneously. However, if HiddenTag emits sound at different frequencies, the distributed power on each frequency will decrease. We will apply the three models and compare the results in the section of performance evaluation.

In general, the multi-tone model can enrich feature space by increasing the number of tones. However, the increasing number of tones will reduce the power assigned to each tone. If the power distributed on each tone is too low, the identification results will decrease when we apply support vector machine (SVM) classification. Fig. 5 shows the result of FFT for the 3 sound generation models.

2.4 Receiving Sound

The process of receiving sounds is introduced as follows.

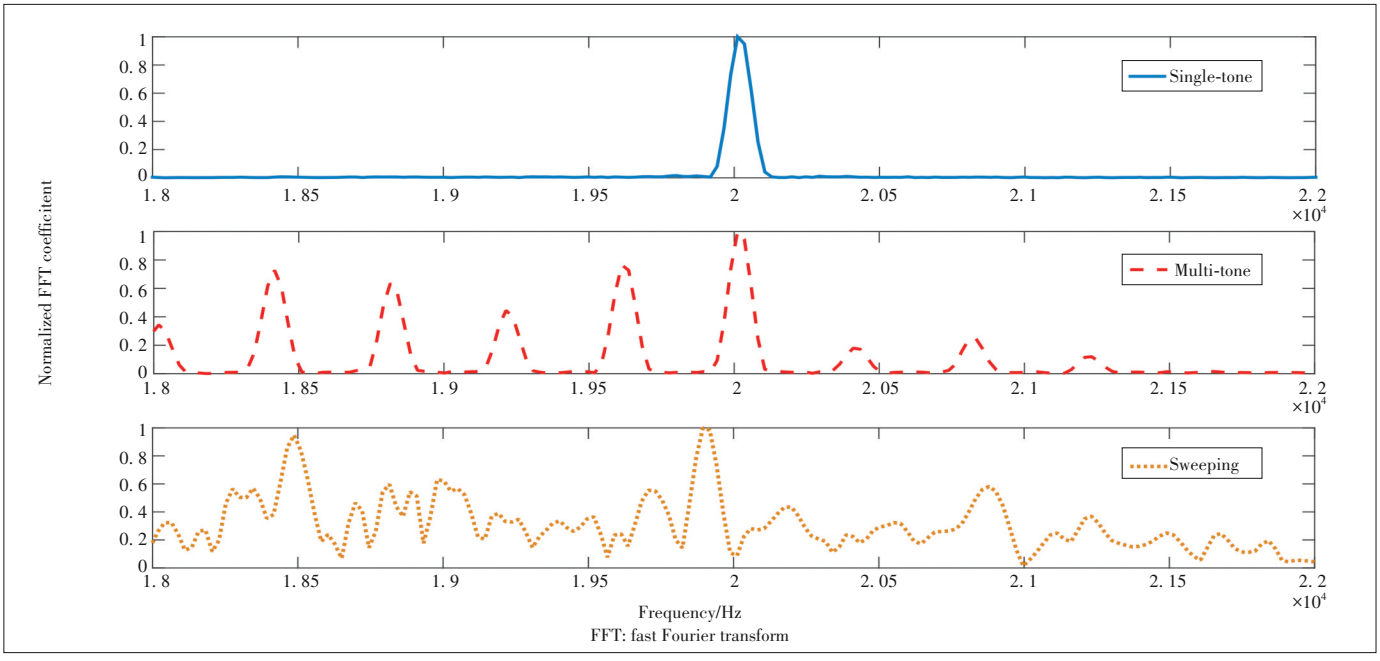
1) Sensing trigger: In HiddenTag, a sensing trigger is needed for person identification. Sensing trigger in our system detects users in a certain area rather than the whole home space. That is, HiddenTag should not recognize users everywhere except for the targeted sensing areas in the smart home. When a user enters the targeted area, HiddenTag will be turned on to collect acoustic data. Otherwise, the HiddenTag remains inactive. This switch can save the energy of smartphones and avoid high frequency acoustic signals when they are unnecessary. In our system, we adopt WiFi CSI signals[10 - 11], which are accurate and pervasive RF signals in smart homes, as the sensing trigger. Once our system detects CSI variations between wireless routers and receivers, HiddenTag will start acoustic sensing in the experimental area where the receiver locates.

2) Fast Fourier transform (FFT): Modern smartphones are able to generate sound waves with frequencies from 20 kHz to 22 kHz. There is an interesting phenomenon: most people cannot hear the sound between 18 kHz and 22 kHz. Considering that the users in the smart home do not suffer from the hearable noises, we can leverage such sound to identify different users. We use two smartphones in which the FFT converts time domain signals into representation in the frequency domain. That is, the FFT takes a block of time-domain data and returns the frequency spectrum of the data. Based on applying FFT and inverse fast Fourier transform (IFFT), we obtain data from both the time domain and frequency domain.

3) Band-pass filtering: In order to reduce the noises from the background and focus on the high acoustic frequency range, we adopt a band-pass filter. A band-pass filter passes signals with frequencies in a certain range and attenuates signals with frequencies out of that range. We keep the sound signals in the frequency range between 18 kHz and 21 kHz. The order of the band-pass filter is 9.

2.5 Launch Machine Learning Engine

1) Constructing acoustic features: Designing suitable feature space is important and challenging for high frequency



▲Figure 5. Three models of sound generation

sound. Different from most speech recognition works, classical features such Mel frequency cepstral coefficient (MFCC) [12] and AFTE[13] do not work well in our system. In HiddenTag, we explore classical features in statistics and extract them from both the time domain and frequency domain.

The features are calculated for a time window, the size of which can be adjusted based on the system's recommendations. In each time window, Table 1 shows the main features adopted in our system.

Specifically, we introduce power spectral entropy and crest factor in detail. In specific, entropy is a common measurement of disorder within a macroscopic system. In HiddenTag, spectral entropy is defined as following steps. First, we compute the spectrum $X(\omega_i)$ of the received signal. Next, we calculate the power spectral density (PSD) of the received signal via squaring its amplitude and normalizing it by the number of bins.

$$P\left(\omega_i\right) = \frac{1}{N}|X_{w_i}|^2 . \tag{3}$$

Then, we normalize the calculated PSD so that it can be viewed as a probability density function (PDF).

$$p_i = \frac{P(\omega_i)}{\sum_i P(\omega_i)} . \tag{4}$$

The power spectral entropy can be now calculated using a standard formula for an entropy calculation.

$$\mathrm{PSE} = -\sum_{i=1}^{n} p_i ln p_i . \tag{5}$$

Crest factor is a feature indicating the ratio of peak values to the effective value for a waveform. For example, crest factor 1 indicates no peaks and higher crest factors indicate peaks. In our system, as shown in Eq. (6), the crest factor refers to the peak amplitude (x_{peak}) of the waveform divided by the root mean square (RMS) value (x_{RMS}) of the waveform. Let C_{dB} denote the crest factor and RMS denote the square root of mean square (the arithmetic mean of the squares of a set of numbers), we have:

$$C_{dB} = 20\log_{10}\frac{x_{\mathrm{peak}}}{x_{\mathrm{RMS}}} . \tag{6}$$

2) Handling noises: Although we have used band-pass filters to reduce the noises which are not in the target range, there are other noises distributed on the frequency area from 18 – 21 kHz. These noise samples may reduce the classification accuracy of HiddenTag. Considering common noises in smart home environments include speaking, clapping, and some background noises, our system can add to or remove four types of noises (background, clapping, speaking and door knocking) from the dataset automatically when we train classification models. Besides, we can assign different ingredients to each type of noise. Once the noises occur in the testing phase, since the training model includes common noises, our system is confident in handling such a problem.

3) Classification: HiddenTag leverages SVM as the classification algorithm. Before implementing SVM in the proposed system, we should consider two problems. Which type of kernel shall be adopted? How to set the value of the penalty parameter? In our datasets, since the number of features is larger than that of observations, according to characterizations of common kernels, we select linear kernels for our SVM approach. Additionally, a low-value penalty parameter in SVM tends to make the decision surface smooth, while a high penalty parameter tries to train all samples correctly by giving the model freedom to select more samples as support vectors. We need to select the penalty parameter in SVM to achieve optimal results. HiddenTag adopts grid search to choose the penalty parameter. Besides, since our system aims to identify users in a short time period, the observation samples are limited. According to the features of common kernels used in SVM (linear kernel, radial basis function (RBF) kernel, etc.), we adopt linear kernels to obtain the optimized classification results.

4) Calibrate exceptions by prediction: Even if HiddenTag is able to identify different users, there still exists the probability of recognizing users incorrectly. Based on our observations, if the proposed system identifies users successfully for most cases, when some exceptions happen, we can calibrate the errors by historic information. In our system, as shown in Algorithm 1, we introduce an approach to avoid exceptions by leveraging the historical information. In each round, when we identify a user, our system not only counts the classification result from SVM in the current round, but also adds the previous results with a certain proportion (α). The parameter α can be adjusted according to the feedback of test cases.

▼**Table 1. Main features extracted in HiddenTag**

Features	Explanation
Crest factor	The value indicates how extreme the peaks are in a waveform
Energy	The energy of the signal in the time domain
Entropy of energy	The entropy of energy in the time domain
Spectral centroid	The center of the gravity of the frequency domain spectra
Spectral spread	The average spread of the spectrum in relation to its centroid
Spectral roll-off	The frequency below 90% of the magnitude distribution of the spectrum is concentrated
Spectral flux	The squared difference between two successive spectral frames
Spectral entropy	The entropy of the spectral energies
Spectral flatness	The ratio of the geometric mean to the arithmetic means of a power spectrum
Zero crossing rate	The rate of sign-changes along with a signal

Algorithm 1. Calibration algorithm for exceptions in HiddenTag

Require: α – between 0 and 1; $P_i'(j)$ – classification result

of user i in time period j before calibration
Ensure: U_{max} – the user with maximal prediction probability (identification result); n is the number of users, m is the number of time rounds
for int $i = 0; i < n; i$++ **do**
  for int $j = 0; j < m; j$++ **do**
    predict user by current round result and historic data
    $P_i(j) = P_i(j-1) \times \alpha + P_i'(j)$
  end for
end for
U_{max} is the user with maximal prediction probability
    select the user with the highest confidence in SVM
Return U_{max}

2.6 Applications of Personal Identification at Smart Home

Because HiddenTag can distinguish users in a smart home with convincing accuracy, we implement more applications via SmartHome Hub to provide personalized services. Our system integrates smart LED and speakers to show the identification results. For the installed smart LED, it will be assigned with different colors to different users. Once the user is identified, the corresponding color will be shown on the bulbs. The speaker can play personalized music for different users. If the user's web account is associated with HiddenTag, the preference music will be played on the smart speaker once the user is recognized. The system does not require explicit user interactions, such as login to an account, recognizing, recalling, or executing users' preferences. More applications can be integrated through SmartHome Hub based on the results of person identification.

3 Evaluation

3.1 Experiment Setup

HiddenTag includes an Android application and a module-view-control (MVC) based website to process acoustic data and recognize users. All the devices are deployed in a smart home environment. We use a Huawei Mate 30 smartphone as the controller, sender, and receiver. A proposed mobile application plays inaudible sound (18 – 21 kHz) on senders. It supports three models: single-tone, multi-tone, and sweeping frequency. Initially, we choose a multi-tone model for our experiments. Our system has generated 15 tones which are distributed from 18 – 21 kHz uniformly. The speaker's volume is set to 100%. The distance between the sender and receiver is 3 m. The area between the sender and receiver is empty. The sender and receiver are placed 75 cm above the floor. After receiving the varied acoustic signals by human activities, received acoustic data will be transmitted to the Dell T3640 server via WiFi. Based on the Python Scikit-Learn library, HiddenTag classifies different users via SVM. The c (penalty parameter) value is selected by grid search.

In our evaluation, we seek to answer three questions: Does HiddenTag identify different users successfully? Since there are often more than 3 family members living in a home environment, how many distinct users that our system can identify? What factors can affect the experimental results?

3.2 Evaluation Metrics

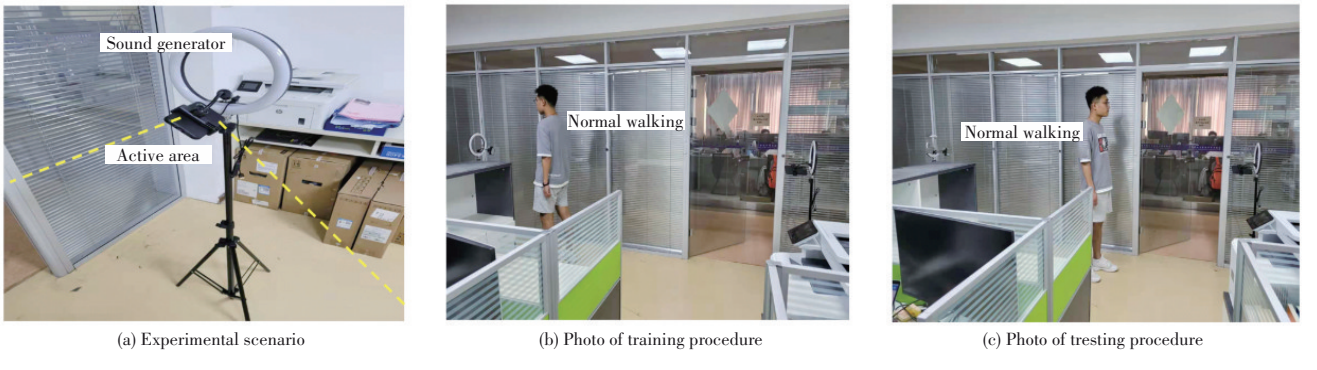
In the offline analysis, we use accuracy in a confusion matrix to describe person identification results. For online test results, we define that accuracy is the success rate for our recognition.

3.3 Case Study

We divide the case study into two phases: the training phase and the testing phase.

In the training phase, when each user enters the experimental environment, our system will detect user activities and start to profile the user. A user walks normally between the sender and the receiver. The user can also turn around and stand shortly. This training procedure lasts for 60 s. After the training procedure, the user leaves the experimental room.

When the user returns to the room, once she/he walks into the same experimental area, HiddenTag starts to recognize the user and shows the confidence of user recognition. This step is the testing phase. Fig. 6 illustrates the experimental environ-



(a) Experimental scenario (b) Photo of training procedure (c) Photo of tresting procedure

▲**Figure 6. Experimental scenario and case study**

ment and corresponding case study.

In this subsection, we observe the group with four users as shown in Table 2. Four users participated in the experiment. Each user was trained and tested separately. Table 3 is the confusion matrix for the classification. As shown in Fig. 7 (a), testing accuracy can achieve 96.1%. We thus conclude that the time length of training influences the accuracy. The longer time of training obtains better results. However, considering our application scenario should limit training procedure to a certain time length, we choose 60 s in our implementation.

Then, we focus on the number of users in our case study. We extend our experiments from 4 users to 10 users. After changing the number of users, based on Fig. 7 (b), we notice our system still achieves an accuracy of more than 90%. Although the system performs better when the system includes fewer users, HiddenTag can still process 10 users with acceptable accuracy.

▼Table 2. Information of four volunteers

User	A	B	C	D
Height/cm	176	177	174	163
Weight/kg	65	80	70	55
Age	25	31	33	40
Gender	M	M	F	F

▼Table 3. Confusion matrix of four-volunteer experiment

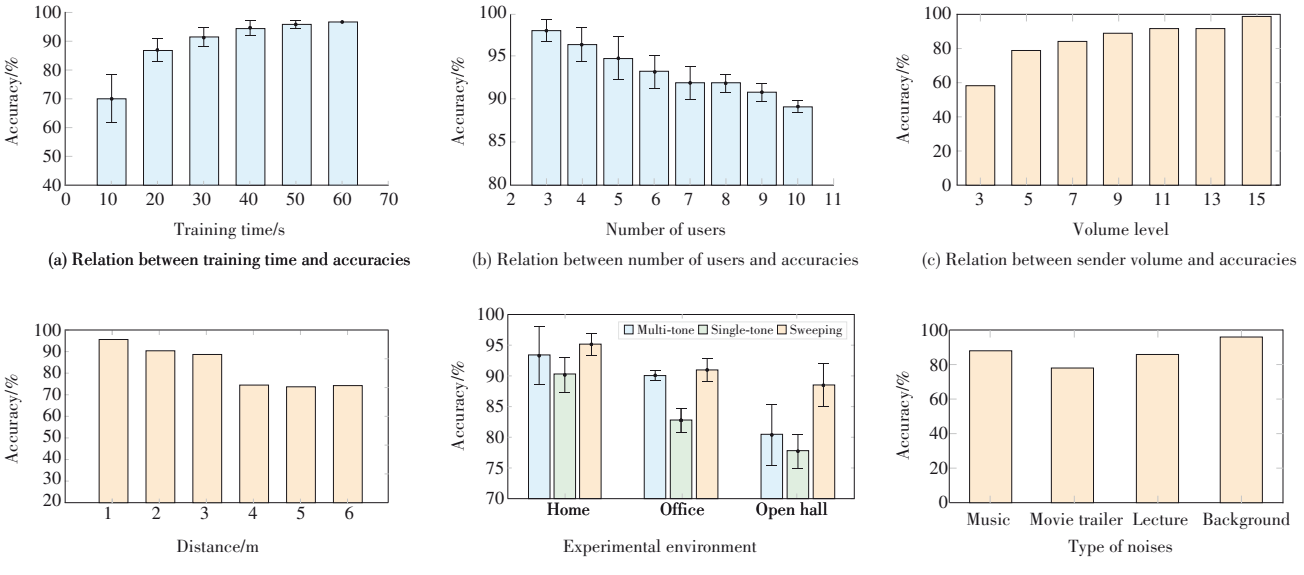
Actual/Classified	A	B	C	D
A	93.0%	1.0%	0.0%	6.0%
B	1.0%	98.0%	0.0%	1.0%
C	0.0%	0.0%	96.0%	4.0%
D	5.0%	16.0%	0.0%	79.0%

Different volumes of the sender sound will change the sound signal strength and identification accuracy. We did the control group experiments to detect which percentage is the best volume for our experiment. Fig. 7 (c) shows the improvement with increasing volume.

Additionally, the distance between the sender and receiver is another factor that affects recognition results. According to existing experimental settings, we only adjust the distance between the sender and receiver. Fig. 7 (d) shows that with closer distance, the group achieves better accuracies. Only when the distance is too short to profile walking activities (within 1 m), the accuracy will decrease.

Then, we compare three sound generation models and discuss which one is the best model for the proposed system. In Fig. 7 (e), we conduct other two control groups by using a single-tone model and a sweeping model. For the single-tone model, we set the frequency of the sound to 20 kHz. For the sweeping model, we sweep frequency from 18 kHz to 21 kHz once per second. We compare the three techniques in different scenarios (smart homes, offices and open halls), and come to the conclusion that sweeping and multi-tone models outperform single-tone models. Because multi-tone and sweeping models increase accuracies by enriching feature space. The multi-tone model is subjected to power decrease and thus needs a power amplifier to improve performance.

Additionally, based on our observations, the errors of the proposed system mainly occur in the first or second frame. Within the time increasing, the errors will decrease sharply. This phenomenon is caused by two reasons. First, the acoustic signature of each user cannot form in a very short time period.



(a) Relation between training time and accuracies
(b) Relation between number of users and accuracies
(c) Relation between sender volume and accuracies
(d) Relation between distance and accuracies
(e) Experiments by different acoustic models
(f) Experiments with different types of noises

▲Figure 7. Experimental results of evaluations

Once a user has walked 3 – 5 gait cycles, our system can recognize the user based on the acoustic signature. Second, as illustrated in Algorithm 1, the results will be calibrated by historical data. The beginning frames do not have the capability of enhancing accuracies by counting the results in previous rounds.

4 Diving into Depth

In this section, we further analyze HiddenTag based on these factors: online performance, experimental environment, and noise handling.

4.1 Pushing Offline to Online

In order to deploy HiddenTag in a real platform, we develop an online system to show the real-time identification results. HiddenTag adopts Node.js and Python Flask to display the real-time accuracies. The time delay of classification results can be controlled from 2 s to 5 s. Table 4 illustrates the comparison between online and offline results in the same experimental scenario. As shown in Eq. (7), for a certain user, the online accuracy is the ratio that times of successful identifications (N_s) divided by the total times of identifications (N_t).

$$Acc_0 = \frac{N_s}{N_t} \times 100\% . \tag{7}$$

Although online accuracy is lower than offline accuracy, it still reaches 85.0% – 90.0%. Next, we extend the online experiments from a room to other scenarios with different layouts and materials. We test HiddenTag in a conference room and a coffee room. The two experiments have achieved online accuracies of 87.5% and 90.5%. The case studies have proved HiddenTag can work normally in different environments.

There are two reasons that the accuracy is lower in the online system. First, in offline classification, SVM is able to choose optimal parameters by brute-force searching. However, it is difficult to optimize all the parameters in a short time period due to computational limitations in an online system. Second, the experimental environment is changing between online training and online testing. We discuss this issue in the following subsection.

4.2 Environment Changing

Our experiments are conducted in the same environment. Unfortunately, the same experiment scenario is always changing due to variations of environmental factors, such as temperature and humidity. To assess its impact, a user walked in the experimental scenario by a five-minute interval. Table 5 shows that the same user is classified by HiddenTag for four times in different time slots. Even if one user enters the room for four times, each event can be identified as different users with 62.25% accuracy. To eliminate the environmental changing, the system needs to continuously collect long-term training data implicitly. Relying on a larger dataset that includes more environments variations, we can identify people even if the environment changes sharply.

▼Table 4. Comparison between online and offline results

	Sweeping	Single-Tone	Multi-Tone
Offline	95.2%	91.0%	93.1%
Online	90.0%	80.0%	85.0%

▼ Table 5. Confusion matrix of identifying the same user in different time periods

		Classified			
		1st	2nd	3rd	4th
Actual	1st	48.0%	6.0%	3.0%	43.0%
	2nd	11.0%	74.0%	8.0%	7.0%
	3rd	1.0%	5.0%	74.0%	20.0%
	4th	40.0%	0.0%	7.0%	53.0%

4.3 Noise Handling

We simulate typical noises when conducting HiddenTag in an experimental scenario. In this experiment, we use a Huawei Mate 30 smartphone to play 3 audio files including a song named "Amazing Grace", the trailer of Game of Throne 8, and a lecture talk of a machine learning class on Coursera. The playing smartphone is close to the receiver (30 cm). By adopting the proposed noise handling method introduced in Section 2.5, Fig. 7 (f) shows that our system still reaches acceptable accuracies even if it encounters different types of noises. HiddenTag can work normally under some types of noises, but the accuracies decrease when it encounters some noises, such as the noises made by the working elevator and starting of the heater.

5 Related Work and Comparison

Existing person identification approaches broadly rely on computer vision and image techniques. By analyzing users' faces and fingerprints, researchers and engineers have provided numerous solutions to user recognition. As a classical face recognition approach, Turk and Pentlend leverage Eigenfaces to define the face space and identify people[14]. Recent researchers use the deep network to enhance recognition accuracy[6, 15 – 16]. DeepID3[6] designs a high-performance deep convolution network and adds supervision to early convolutional layers, and it represents the state-of-the-art technology on YouTube Faces benchmarks. Voice recognition is another type of common approach to identify a user. MUDA et al.[17] explore MFCC and dynamic time warping (DTW) techniques to recognize users. In addition, biometrics techniques, such as fingerprint and retina, are other common types of person identification[18 – 21]. Unfortunately, all of these methods face privacy concerns. Although these approaches can recognize users by biometric information, the key personal and private information has to be exposed.

Recently, some alternative methods have been proposed to

identify persons. Researchers adopt wireless sensing to identify persons, gestures, and even micro-activities[22 - 23]. By classifying variations of WiFi signals, WiWho[23] leverages CSI to describe the user's walking behaviors and identify users in WiFi environments. However, wireless sensing methods often need specific devices, such as the emitter and the receiver with CSI drivers, which are not common in smart home environments.

Acoustic sensing has been a hot topic recently, and lots of corresponding applications, such as speech recognition, indoor localization are implemented in smart homes[24 - 27]. GEIGER et al.[28] presented a system for identifying humans by their walking sound, by leveraging MFCC and Hidden Markov Model, which has reached the offline identification rate of 65.5% for 155 subjects. This approach depends on the sounds of footsteps. Once the shoes and floors are changed, the system might not work normally. This method does not consider noise handling and online accuracies in a real environment. Actually, there is no solution for person identification area by acoustic sensing without human voice or step sound.

As shown in Table 6, different from the existing solutions, HiddenTag is a device-free and highly accurate person identification approach. By using built-in smartphones, we can recognize users only by profiling the common indoor activities at home and in office environments.

6 Conclusions and Future Work

HiddenTag represents the first device-free system that employs inaudible acoustic sensing to achieve accurate person identification. Through this process without any hardware modification, we gain important insights: 1) acoustic information with frequencies from 18 - 21 kHz can profile human indoor activities and recognize users in smart home environments; 2) sweeping frequency and multi-tone models can improve SVM classification for acoustic datasets by enriching features; 3) online and offline identification accuracy can reach more than 90% in simplified testing and training procedures which are close to normal activities in the environments similar to smart homes. We believe HiddenTag's salient advantages will enable a myriad of personalized services in smart homes, including smart voice assistants, augmented reality, energy saving, and various pervasive applications.

Moving forward, we are aiming to further improve the identification accuracies by leveraging other machine learning techniques such as recurrent neural networks and generative adversarial networks and enrich the acoustic features by leveraging transfer learning. In addition, we aim to extend single person identification to multi-person with more walking patterns.

References

[1] CLIFFORD B R, BULL R. The psychology of person identification [M]. London, UK: Routledge, 2017. DOI: 10.4324/9781315533537
[2] BADRINARAYANANN V A , SIERRA J J , MARTIN K M . A dual identification framework of online multiplayer video games: The case of massively multiplayer online role playing games (MMORPGs) [J]. Journal of business research, 2015, 68(5): 1045-1052
[3] FANG B Y, CO J, ZHANG M. DeepASL: enabling ubiquitous and non-intrusive word and sentence-level sign language translation [C]//The 15th ACM Conference on Embedded Network Sensor Systems. Delft, Netherlands: ACM, 2017: 1 - 13. DOI: 10.1145/3131672.3131693
[4] ALI K , LIU A X , WEI W , et al. Keystroke Recognition Using WiFi Signals [C]// ACM MobiCom. Paris, France: ACM, 2015
[5] YI J, LEE Y. Heimdall: mobile GPU coordination platform for augmented reality applications [C]//The 26th Annual International Conference on Mobile Computing and Networking. London, United Kingdom: ACM, 2020: 1 - 14. DOI: 10.1145/3372224.3419192
[6] LSUN Y, LIANG D, WANG X G, et al. DeepID3: face recognition with very deep neural networks [J]. Computer Science, 2015
[7] WANG M, DENG W H. Deep face recognition: a survey [EB/OL]. [2021-03-20]. https://export.arxiv.org/pdf/1804.06655
[8] ZHANG Z Y. Microsoft Kinect sensor and its effect [J]. IEEE multimedia, 19(2): 4 - 10, 2012
[9] LÓPEZ G, QUESADA L, GUERRERO L A. Alexa vs. Siri vs. Cortana vs. Google assistant: a comparison of speech-based natural user interfaces [C]//The AHFE 2019 International Conference on Human Factors and Systems Interaction. Washington, USA: AHFE, 2019. DOI: 10.1007/978-3-319-60366-7_23
[10] HALPERIN D, HU W J, SHETH A, et al. Tool release [J]. ACM SIGCOMM computer communication review, 2011, 41(1): 53. DOI: 10.1145/1925861.1925870
[11] WANG X Y, GAO L J, MAO S W. CSI phase fingerprinting for indoor localization with a deep learning approach [J]. IEEE Internet of Things journal, 2016, 3(6): 1113 - 1123. DOI: 10.1109/JIOT.2016.2558659
[12] LOGAN B. Mel frequency cepstral coefficients for music modeling [EB/OL]. [2021-03-20]. http://citeseerx. ist. psu. edu/viewdoc/summary? doi=10.1.1.11.9216
[13] SMCKINNEY M, BREEBAART J. Features for audio and music classification [EB/OL]. [2021-04-02]. https://ismir2003. ismir. net/presentations/McKinney. pdf
[14] TURK M, PENTLAND A. Eigenfaces for recognition [J]. Journal of cognitive neuroscience, 1991, 3(1): 71 - 86. DOI: 10.1162/jocn.1991.3.1.71
[15] XIAO T, LI H S, OUYANG W L, et al. Learning deep feature representations with domain guided dropout for person re-identification [C]//IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016:

▼**Table 6. Comparison between HiddenTag and other classical approaches**

	Information Type	Training Cost	Hardwares Required	Privacy Level	Accuracy
DeepID3	Image	High	Cameras	Low	92% offline
WiWho	CSI	Normal (100 s)	Special CSI devices	High	80% - 92% offline
Step sound	Normal sound	Low (3 s)	Built-in smartphones	High	65% offline
HiddenTag	High frequency sound (18 - 21 kHz)	Normal (60 s)	Built-in smartphones	High	96% offline, 85% - 90% online

CSI: channel state information

1249 – 1258. DOI: 10.1109/CVPR.2016.140
[16] YU H X, ZHENG W S. Weakly supervised discriminative feature learning with state information for person identification [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020: 5527 – 5537. DOI: 10.1109/CVPR42600.2020.00557
[17] MUDA L, BEGAM M, ELAMVAZUTHI I. Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques [EB/OL]. [2021-04-02]. https://arxiv.org/abs/1003.4083
[18] BRUNELLI R, FALAVIGNA D. Person identification using multiple cues [J]. IEEE transactions on pattern analysis and machine intelligence, 1995, 17(10): 955 – 966. DOI: 10.1109/34.464560
[19] PERALTA D, TRIGUERO I, SANCHEZ-REILLO R, et al. Fast fingerprint identification for large databases [J]. Pattern recognition, 47: 588 – 602, 2014. 10.1016/j.patcog.2013.08.002
[20] RAO G S, NAGARAJU C, REDDY L, et al. A novel fingerprints identification system based on the edge detection [J]. International journal of computer science and network security, 8: 394 – 397, 2008
[21] TROJE N F, WESTHOFF C, LAVROV M. Person identification from biological motion: effects of structural and kinematic cues [J]. Perception & psychophysics, 2005, 67(4): 667 – 675. DOI: 10.3758/BF03193523
[22] WANG Z, YU Z W, LOU X Y, et al. Gesture-radar: a dual Doppler radar based system for robust recognition and quantitative profiling of human gestures [J]. IEEE transactions on human-machine systems, 2021, 51(1): 32 – 43. DOI: 10.1109/THMS.2020.3036637
[23] ZENG Y Z, PATHAK P H, MOHAPATRA P. WiWho: WiFi-based person identification in smart spaces [C]//15th ACM/IEEE International Conference on Information Processing in Sensor Networks. Vienna, Austria: IEEE, 2016: 1-12. DOI: 10.1109/IPSN.2016.7460727
[24] MOON Y, KIM K J, SHIN D H. Voices of the Internet of Things: an exploration of multiple voice effects in smart homes [M]//Distributed, ambient and pervasive interactions. Cham: Springer International Publishing, 2016: 270 – 278. DOI: 10.1007/978-3-319-39862-4_25
[25] TUNG Y C, SHIN K G. EchoTag: accurate infrastructure-free indoor location tagging with smartphones [C]//The 21st Annual International Conference on Mobile Computing and Networking. Paris, France: ACM, 2015: 525 – 536. DOI: 10.1145/2789168.2790102
[26] YANG Z J, WEI Y L, SHEN S, et al. Ear-AR: Indoor acoustic augmented reality on earphones [C]//The 26th Annual International Conference on Mobile Computing and Networking. London, United Kingdom: ACM, 2020: 1 – 14. DOI: 10.1145/3372224.3419213
[27] ZHOU B, ELBADRY M, GAO R P, et al. BatMapper: acoustic sensing based indoor floor plan construction using smartphones [C]//The 15th Annual International Conference on Mobile Systems, Applications, and Services. Niagara Falls, USA: ACM, 2017: 42 – 55. DOI: 10.1145/3081333.3081363
[28] GEIGER J T, KNEIßL M, SCHULLER B W, et al. Acoustic gait-based person identification using hidden Markov models [C]//The 2014 Workshop on Mapping Personality Traits Challenge and Workshop. Istanbul, Turkey: ACM, 2014: 25 – 30. DOI: 10.1145/2668024.2668027

Biographies

QIU Chen (qiuchen@nwpu.edu.cn) received the Ph.D. degree in computer science from Michigan State University, USA in 2017. He is currently an associate professor with Northwestern Polytechnical University, China. His research interests include pervasive computing, mobile computing, and applied machine learning. He is a member of the IEEE.

DAI Tao received his B.S., M.S. and Ph.D. degrees in software engineering from Xi'an Jiaotong University, China in 2008, 2011 and 2020, respectively. He is currently a lecturer at the School of Economics and Management, Chang'an University, China. He was a visiting student at the School of Computer Science, Carnegie Mellon University, USA from September 2018 to September 2019. His main research interests include natural language processing, information retrieval, and machine learning.

GUO Bin received the Ph.D. degree in computer science from Keio University, Japan in 2009 and then went to the French National Institute of Telecommunications for postdoctoral research. He is a professor with Northwestern Polytechnical University, China. His research interests include ubiquitous computing, mobile crowd sensing, and HCI. He is a senior member of the IEEE.

YU Zhiwen received the Ph.D. degree from Northwestern Polytechnical University, China. He is currently a professor and Dean with the School of Computer Science, Northwestern Polytechnical University. His research interests include pervasive computing and human-computer interaction. He is a senior member of the IEEE.

LIU Sicong received the B.S., M.S. and Ph.D. degrees from Xidian University, China in 2013, 2016, and 2020 respectively. From 2017 to 2018, she was a visiting scholar at Rice University, USA. She is currently an associate professor with Northwestern Polytechnical University, China. Her research interests include mobile computing system, mobile and embedded deep learning design, and automated deep model optimization.