EchoImage: User Authentication on Smart Speakers Using Acoustic Signals

Yanzhi Ren[1], Zhiliang Xia[1], Siyi Li[1], Hongbo Liu[1], Yingying Chen[2]
Shuai Li[3], Hongwei Li[1]
[1]Dept. of CS, University of Electronic Science and Technology of China [2]WINLAB, Rutgers University, USA
[3]University of Oulu, Finland
[1]{renyanzhi05, hongweili}@uestc.edu.cn [2]yingche@scarletmail.rutgers.edu

2023 IEEE 43rd International Conference on Distributed Computing Systems (ICDCS) | 979-8-3503-3986-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICDCS57875.2023.00048

***Abstract*—The user authentication has drawn increasingly attention as the smart speaker becomes more prevalent. For example, smart speakers that can verify who is sending voice commands can mitigate various types of attacks such as replay attack or impersonation attack. Existing user authentication solutions either cannot be applicable to smart speakers directly or require certain additional user-device interaction or pre-installed infrastructure, which may severely affect the user experience and create extra burdens to users. In this work, we propose a user authentication system EchoImage utilizing acoustic images, which are derived from the smart speaker by emitting beep signals and sensing echoes from the user's body with its microphone array, as the proof for user authentication. Given the acoustic samplings of the reflected beep signal, our system designs a distance estimation component by applying a correlation based technique on the beamformed signal to estimate the distance between the user and microphone array. Our image construction component then constructs a virtual imaging plane using the estimated distance and steers the array towards each grid of the plane to generate an acoustic image of the user. Moreover, we propose a transfer learning-based method to derive efficient features from the constructed images, and employ SVM classifiers for accurate user authentication. Our extensive experiments demonstrate that our system is robust and accurate across various scenarios.**

I. Introduction

Smart speakers, also known as intelligent voice assistants, are becoming increasingly popular in our daily life. As of 2021, the number of global smart speaker shipments was 186 million, and it will surpass 200 million annually in 2023 [1]. Such voice-based systems are evidenced by a wide variety of products from several companies, such as Amazon Alexa [2], Apple HomePod mini [3], Xiaomi Mi [4] and so on. These devices usually wait for the legitimate user's voice commands, and then follow the received command to perform various activities. For example, the smart speaker could help users to place phone calls, schedule appointments, check emails, or control appliances at home. Furthermore, recent studies report that some critical functionalities could even be supported in the near future. For instance, a user-to-user payment feature for restaurants or ticketing is going to be available on Amazon Alexa after 2022 [5].

However, since voice sound propagates in an open channel, smart speakers are vulnerable to various types of attacks. The replay attack [6] could control the speaker illegally via playing a user's pre-recorded command. In impersonation attack [6] or voice synthesizing attack [7], the adversary can generate fake commands that mimic legitimate users' voice to fool the system. The dolphin attack [8] could even produce inaudible malicious commands to control the speaker. The traditional human-voice based authentication schemes [9], [10] cannot thwart these attacks. In addition, the consequences are also serious once the smart speakers are attacked, e.g., result in burglary by opening a smart-door for attackers, or cause financial loss by transferring money to a malicious account.

Because of these potential risks, it is essential to develop an authentication scheme to verify the user's identity for mitigating these attacks. The widely adopted user authentication methods based on password or fingerprint [11], [12] might not be applicable to this case where smart speakers are usually not equipped with the touch screen or fingerprint reader. Recently, captcha based schemes [13] are developed to authenticate users when receiving a voice command on smart speakers. However, some additional actions would be required in these solutions and they could be cumbersome for user experience, especially for senior citizens or people with disabilities. Blue *et al.* [14] designed a 2MA scheme which uses the direction of arrival (DoA) of the voice commands to detect possible remote attacks. However, this scheme assumes that the user is in constant possession of their smartphones for localization. Some developments [15]–[17] focus on developing methods to differentiate live and replayed voice commands using the user's physiological activities or mobility features. However, these techniques either require certain pre-installed infrastructures to the device or cannot be applicable to the cases when the user is static. In addition, they could also not be applied for user authentication scenarios.

In recent years, studies [18], [19] investigate the user authentication/identification problem via active acoustic sensing. These schemes use the reflected beep sound, which contains rich information of the user's face or body, as the proof for authentication/identification. Moreover, microphone arrays nowadays have been widely adopted by most smart speakers [20] to increase the diversity of the collected audio, and these arrays could enable us to visualise intensity images of sound [21], [22]. These observations trigger our idea from active sensing [18] to construct acoustic images using echoes reflected from human's body for user authentication on smart speakers. Specifically, in this paper, we design a user authentication system EchoImage that uses the acoustic images, which are derived from the smart speaker by emitting beep signals with its speaker and sensing echoes from the

2575-8411/23/$31.00 ©2023 IEEE
DOI 10.1109/ICDCS57875.2023.00048

user's body with its microphone array, as the robust proof for user authentication.

However, several unique challenges need to be addressed when developing such a system: First, due to multi-path propagations of sound waves, the received signal is a mixture of echoes which arrive at the microphone array via multiple paths after bouncing various reflectors, making it hard to recognize echoes reflected back from the direction of user's body accurately for acoustic image construction. Second, the received reflected beep signal could also be significantly affected by the distance between the microphone array and user, requiring our system to be adaptive to different sensing distances. Third, the generated acoustic images usually contain too much redundant information of the user, making it a challenging task to extract reliable features from these images for accurate user authentication. Forth, obtaining a large dataset to train a robust classier for user authentication is problematic because it is burdensome for users to collect sufficient acoustic images at almost every possible distance. Last but not least, our proposed system should re-use existing acoustic sensors on smart speakers without requiring any add-on infrastructures or equipments.

To cope with these challenges, our proposed EchoImage consists of three main components: *Distance Estimation*, *Image Construction* and *User Authentication*. The system takes as input the recorded acoustic samplings of the beep signal captured by the microphone array. In the *Distance Estimation* phase, the acoustic data is first processed to steer the look-direction of the array to an arbitrary region of the user's body. A correlation based technique is then adopted to identify the beginning points of echoes from the beamformed signal for accurately estimating the distance between the user and microphone array. During *Image Construction*, our system utilizes the estimated distance to construct a virtual imaging plane and steers the array towards each grid of the plane. The acoustic image is then constructed by assigning a pixel value to each grid, and such value considers the energy of echoes reflected from the direction of corresponding grid. Finally, in *User Authentication*, the feature extraction is performed to derive features from the constructed images by using a pre-trained CNN model, and based on such features our system adopts SVM classifiers to perform user authentication. We summarize our main contributions as follows:

- We propose to use acoustic images, which are derived from the smart speaker by emitting a pre-designed beep signal and sensing its echoes from the user's body via the microphone array, as the proof for user authentication.
- To be adaptive to different distances between the user and microphone array, we design a distance estimation scheme to steer the array's look-direction towards the user's body and detect the beginning points of echoes from the beamformed signal for accurate distance estimation.
- To capture the unique characteristics of the user, we consider the energy of echoes from his/her body and develop an imaging scheme based on MVDR beamforming to get the in-air acoustic image of the user.
- To achieve an accurate user authentication with limited training sample size, we adopt a transfer-learning based model to derive effective features from the constructed image and design a data augmentation scheme for generating synthesized training images.
- Our extensive experimental results show that our proposed system EchoImage is accurate and robust for user authentication under various real world scenarios.

II. Related Work

There have been widely adopted methods for user authentication/identification on mobile devices. Some devices rely on manual entry of a pre-set password or a sequence of beats [11]. The biometric characteristics such as fingerprints or face ID have also been widely used for user authentication on mobile devices in recent years [12]. Along this direction, several physiological or behavioral feature based authentication schemes, e.g., using finger gestures or slides [23], [24], breathing patterns [25], cardiac features [26], behavioral patterns [27] or keystroke dynamics [28], have also been proposed. However, all of the above schemes might not be applicable to our case where smart speakers are usually not equipped with the touch screen, camera, keyboard or fingerprint reader.

Some studies have been developed for user authentication or liveness detection on smart speakers. For instance, Uzun *et al.* [13] propose a captcha based scheme when the smart speaker receives a voice command from the user. However, it requires people to interact with the device and it could be cumbersome for user experience, especially for senior citizens or people with disabilities. Blue *et al.* [14] designed a 2MA scheme which utilizes the direction of arrival (DoA) of the voice commands to detect possible remote attacks. However, this scheme assumes that the user is in constant possession of their smartphones for localization. Several works have also been proposed to differentiate the live and replayed voice commands by leveraging the user's physiological activities or mobility features [15]–[17]. However, these systems can only be realized with the help of certain pre-installed infrastructures, and it leads to a certain degree of overhead in real deployment.

There are also some studies dedicated for acoustic sensing. In these techniques, the device emits a pre-designed signal and then uses its microphone to capture echoes for authenticating the user's liveness or identity. FormaTrack [19] presents a user identification solution by extracting unique features from echoes bouncing off the user's body. Nowadays, microphone arrays have also been widely adopted by most smart speakers [20] to increase the diversity of the collected audio. Lee *et al.* [29] develop a sonar-based liveness detection system for smart speakers by emitting an inaudible sound and tracking the user's direction using the array. However, these schemes are not easy to be applicable to the user authentication scenario for smart speakers. Moving forward, in [21], [22], the microphone array has been utilized for creating acoustic images to visualise the intensity of sound. However, such systems require a customized array for image generation and

the total cost for such deployment is still too high. In addition, it is also not clear how these methods can deal with the user authentication problem.

The most similar work to our own is by EchoPrint [18]. EchoPrint actively emits acoustic signals from the speaker to 'illuminate' the user's face and authenticate the user by deriving features from the echoes bouncing off the facial contour. However, it is not clear how to apply this authentication method on microphone arrays. Unlike the aforementioned work, we aim to develop an acoustic sensing system that enables smart speaker to verify the user's identity leveraging its existing speaker and microphone array. Our proposed system does not need the user's explicit participation and is also easy-to-use without requiring any dedicated hardware or pre-installed infrastructures.

III. Background

This section presents relevant background for this paper, far-field case, linear frequency modulated chirp and beamforming. The background concept will lead us into the detailed technical design of our proposed system.

A. Far-Field Case

The smart speaker typically holds a single microphone array. In this work, we assume that the sound wave usually propagates along a straight line in the air, so the wave front from the sound source should be curved. However, when the sound source is far enough away from the microphone array to essentially appear as a point in the space (with no discernible dimension or size), the spherical shape of the sound waves has grown to a large enough radius that one can reasonably approximate the wave front as a plane-wave with no curvature. In such case, the following condition should be satisfied [30]:

$$\mathrm{L} \geq \frac{2 \times d^2}{\lambda} \tag{1}$$

where L is the distance between the source and the microphone, λ is the wavelength of the sound wave, and d is the inter-distance between microphones of the array. Intuitively, if the distance between the sound source and the array is relatively large compared to the dimensions of the array, we can approximate the wave front as flat and such assumption does not introduce much error in the calculation. For example, if the frequency of the sound wave is 3000 Hz, and the corresponding $\lambda = 0.11$m. For an array with size of 0.1m, the source can be viewed as the far-field case as long as it is 0.18m from the array.

B. Linear Frequency Modulated Chirp

Linear Frequency Modulated (LFM) chirp is a signal in which the frequency increases (or decreases) over time and it is commonly used as the active transmitted sound source of a radar system. Specifically, suppose f_0 is the center frequency, B is the bandwidth, T is the dispersion time and A is the amplitude, and the LFM signal can then be represented as ($|t| \leq \frac{T}{2}$):

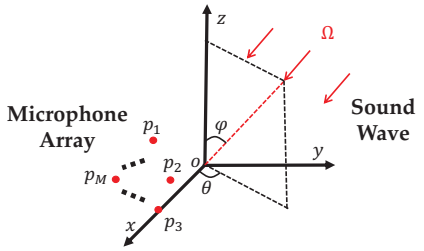


Fig. 1. Illustration of the signal propagation model for the microphone array.

$$\mathrm{s}(t) = A\cos 2\pi(f_0 t + \frac{B}{2T}t^2) \tag{2}$$

C. Signal Propagation Model for Microphone Array

Suppose we have an array consisting of M microphones, and the location of the m-th microphone can be expressed as:

$$p_m = [p_{xm}, p_{ym}, p_{zm}]^T, m = 1, ..., M \tag{3}$$

where p_{xm}, p_{ym} and p_{zm} denote the x, y and z coordinates of the m-th microphone. Thus, the position vectors of M microphones in the array could be represented as:

$$P = \{p_1, ..., p_M\} \tag{4}$$

we next assume that the sound source is in the far field and the sound wave reaches the array with an incident angle of $\Omega = \{\theta, \varphi\}$, where θ and φ denote the azimuth angle and the elevation angle respectively as illustrated in Figure 1. Thus, the sound propagation vector [31] can be expressed as:

$$v(\Omega) = -[\sin\varphi\cos\theta, \sin\varphi\sin\theta, \cos\varphi]^T \tag{5}$$

Thus, to derive the sound propagation delay (relative to origin) on the m-th microphone, we derive the time difference of arrival (TDOA) via computing the inner product of the sound propagation vector $v(\Omega)$ and the position vector p_m [31]:

$$\tau_m(\Omega) = -\frac{v^T(\Omega)p_m}{c} \tag{6}$$

Remind that the propagation delay can also be translated to the phase shift of the signal. Specifically, we let $\omega_0 = 2\pi f_0$ denote the center angular frequency of the narrow-band signals which are received by the array. Thus, the phase shift (relative to origin) for the m-th microphone can be expressed as:

$$\omega_0\tau_{\mathrm{m}}(\Omega) = -\omega_0\frac{v^T(\Omega)p_m}{c} = -\frac{\omega_0}{c}v^T(\Omega)p_m = -k^T(\Omega)p_m \tag{7}$$

where $k(\Omega) = \frac{\omega_0}{c}v(\Omega)$ is defined as the wave number [31].

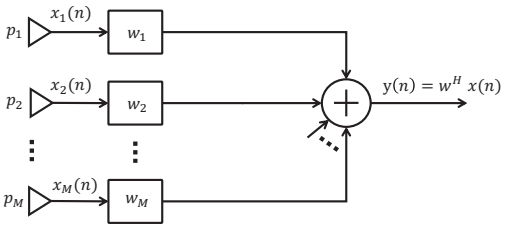


Fig. 2. The flowchart of beamforming.

D. Beamforming

Beamforming is a spatial filtering technique which is used in microphone array for directional signal reception [32]. As shown in Figure 2, the key idea behind such technique is to combine weighted signals received by each microphone from the array in such a way that signals experience constructive or destructive interferences, depending on the direction in which the microphone array is listening to, and we also refer it as the look direction of the array.

Minimum Variance Distortionless Beamforming. The Minimum Variance Distortionless (MVDR) beamforming is a widely used optimal beamforming design, and its goal is to minimize the variance of the beamformer output. If we assume that the noise and the desired signal are uncorrelated, the variance of the captured signals is the sum of variances of the desired signal and the noise. Hence, the MVDR solution seeks to minimize the sum, thereby mitigating the effect of the noise. Specifically, if we steer the array's look direction to the incident angle of the sound wave (i.e., $\Omega = \{\theta, \varphi\}$), the weight vector of the MVDR beamforming is [33]:

$$wv_{MVDR} = \frac{\rho_n^{-1} p_s}{p_s^H \rho_n^{-1} p_s} \tag{8}$$

where $p_s = [e^{jk^T(\Omega)p_1}, ..., e^{jk^T(\Omega)p_M}]^T$ and ρ_n is the normalized covariance matrix of the background noise on M microphones.

IV. System Overview

In this section, we provide an overview of our proposed system EchoImage. As illustrated in Figure 3, our proposed system consists of three major components: *Distance Estimation*, *Image Construction* and *User Authentication*. The system takes as input the recorded acoustic samplings of the reflected beep signal captured by the microphone array, and a bandpass filter is applied to remove the environmental noise from the recorded data. The distance estimation component processes the acoustic data to steer the look-direction of array to an arbitrary region of the user's body. A correlation based technique is then adopted to identify the beginning points of echoes from the beamformed signal for estimating the distance between the user and microphone array. Given the estimated distance, in image construction component, our system first constructs a virtual imaging plane and steers the array towards each grid of the plane. The acoustic image is then generated

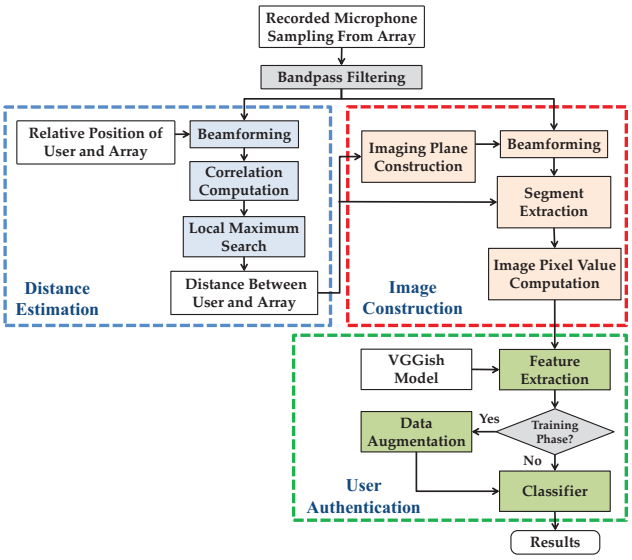


Fig. 3. System flow of our EchoImage.

by assigning a pixel value to each grid, and such value considers the energy of echoes reflected from the direction of the corresponding grid. After that, the user authentication is performed by deriving features from the constructed images using a transfer-learning based model. Based on such features, our system adopts SVM classifiers to authenticate the user's identity. To achieve a robust user authentication with limited training data, a data augmentation scheme is also proposed to generate synthesized training samples based on the acoustic signal propagation model.

V. User Authentication System

In this section, we present the detailed system implementation of our system EchoImage.

A. Parameter Setting of the Beep Signal

When designing the probing beep signal played through the speaker for user authentication, we mainly consider three factors: frequency band, length and time interval.

Frequency Band. For frequency band selection, our proposed system suffers by several limitations. First, according to [21], the grating lobes are manifestations of inadequate spatial sampling for an array with uniformly spaced microphone distances, which are undesirable in beamforming since the the array is as sensitive to waves coming from the directions of grating lobes as for the steering direction. To avoid this phenomenon, the distance between microphones should be smaller than half wavelength of the received signal (i.e., $< \frac{\lambda}{2}$). Thus, giving the inter-microphone distance of the array spanning from 4 cm to 7 cm on most smart speakers, the frequency of our designed beep signal would have to be set below 3000 Hz. Thus, it is not desirable to use higher frequencies (e.g., greater than 20 kHz) to ensure inaudibility to humans as in other state-of-the-art works. Second, our designed beep signal should not be significantly impacted by the environmental noises, which are mostly concentrated below 2000 Hz. In summary, given

the tradeoff of these factors, our system adopts the 2 kHz to 3 kHz bandwidth beep acoustic signal.

Although this selection makes the beep signal audible to humans, the impact of this selection is minimal since our system triggers the user authentication process infrequently. Moreover, we further conduct a post-use survey after our experimental evaluation and the results show that most participants in our usability study did not consider the emitted sound annoying in the authentication process.

Length. The length of beep signal also impacts the accuracy and reliability of our system. The acoustic elements on smart speakers cannot generate or pick up too short beep signals and their corresponding reflections. Thus, it seems reasonable to set a longer length of the beep signal since more energy at each frequency could be collected. However, multi-path distortions could be severe if the duration of the emitted beep signal is too long since reflections from far away users could also be collected during this long sensing process [34]. Therefore, in this work, the length of the beep signal is empirically set as about 0.002s.

Time Interval. The last parameter we consider is the time interval between two consecutive beep signals. This parameter is related to the sensing speed of our system: a longer time interval results in a longer time our system needs to take for authentication, and a shorter interval could cause errors since the reflected signals might overlap with each other. Based on our observations, the reflected sound could still exist even after 0.3s from the beep sound. Therefore we set the interval to be 0.5s in this work.

B. Distance Estimation

The basic idea of our EchoImage is to utilize echoes reflected from the user's body to construct an acoustic image, which contains unique information of the user's identify, as the proof for user authentication. Thus, to accurately identify echoes from the user's body, we first need to estimate the distance between the user and the smart speaker. In the commonly used techniques [35]–[37], a straightforward way is to evaluate the correlations between the received echo and the original beep signal, and then select large peaks in the correlation sequence for distance estimation. However, echoes from human body could be relatively weak such that echoes from surrounding obstacles in other directions could also generate peaks with comparable amplitudes, and this makes such distance estimation process inaccurate and unstable.

To solve this problem, in this paper, we first use MVDR beamforming to steer the array's look-direction to an arbitrary small region on the user's body. In this case, the array only has good sensitivity to the reflected signals from this body region, and the acoustic waves from other directions will be suppressed. Therefore, we can then evaluate the correlations between this beamformed signal and the original signal, and identify peaks in such correlation sequence for accurate and robust distance estimation.

Specifically, we consider a coordinate system in which the location of the array center is at the origin, and assume that

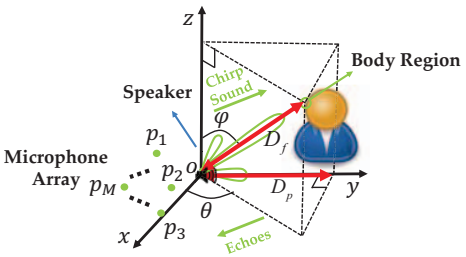


Fig. 4. The illustration of user-array distance estimation.

users intentionally stand directly in front of the array when he/she performs some safety-critical operations as illustrated in Figure 4. EchoImage emits chirp sound signals from the speaker to 'illuminate' the user's body and the received reflected signal is sampled with a frequency of 48 kHz. A 2 to 3 kHz Butterworth bandpass filter is then applied to remove environmental noises in other frequency band. We assume that $\{r_{m,l}(t), 1 \leq l \leq L, 1 \leq m \leq M\}$ represents the received signal for L beeps from M microphones of the array after the bandpass filtering. The MVDR beamforming is applied and its corresponding weights are set according to Equation (8).

We next describe how to determine suitable values for the steered azimuth angle θ and elevation angle φ in Equation (8). In this work, we deem that it is feasible to steer the array to an arbitrary region on the user's upper body for distance estimation. Therefore, we empirically set the θ as $pi/2$ and the φ as a value between $pi/3$ to $2pi/3$, respectively. We also find that this parameter selection could ensure the array always be steered to the user's upper body regardless of his/her height or distance in most of time. If we use $\hat{r}_l(t)$ to represent the beamformed signal for the l-th beep ($1 \leq \mathrm{l} \leq \mathrm{L}$) and let $s(t)$ be the original chirp signal. To identify the beginning points of echoes, the signal $s(t)$ is slid across the beamformed signal $\hat{r}_l(t)$ with a moving window and the correlation sequence is calculated using the matched filter as follows:

$$C_l(t) = \int_{-\infty}^{+\infty} \hat{r}_l(\tau) h(t-\tau) d\tau \tag{9}$$

where $h(t)$ denotes the conjugated and time-reversed version of the original beep signal $s(t)$ (i.e., $h(t) = s^*(-t)$). To capture the overall trend changes of $C_l(t)$, we identify the envelope of $C_l(t)$ using the envelope detection schemes proposed in [38] and denote it as $E_l(t)$. Intuitively, the periodical peaks within envelope $E_l(t)$ can be used as candidates to identify the beginning points of echoes from the user for distance estimation.

However, it is still not applicable to use one $E_l(t)$ directly for distance estimation. The reason is that the derived envelope $E_l(t)$ could be easily impacted by the existence of random interferences, and hence the distance cannot be estimated accurately using merely one beep signal. For this reason, we utilize the fact that the signal reflected back from static object should have a more stable similarity value with the original beep signal [35]. So it is natural for us to think about

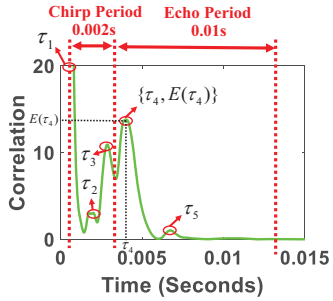


Fig. 5. The illustration of body echo detection using peaks within correlation values derived from a matched filter.

whether we could add envelope $E_l(t)$ from each chirp signal together to enhance stable peaks corresponding to human body and remove other unstable peaks generated by random interferences. Specifically, given the number of beeps L and envelope $E_l(t)$, we compute the sum as follows:

$$\mathrm{E(t)} = \frac{1}{L}\sum_{l=1}^{L} |\mathrm{E}_l(\mathrm{t})|^2 \tag{10}$$

To identify peaks, we search for a set of local maxima from $E(t)$ by varying t and denote it as $MaxSet$ which consists of W local maximum points: $MaxSet = \{\{\tau_w, E(\tau_w)\} \mid 1 \leq w \leq W\}$. For each $\{\tau_w, E(\tau_w)\} \in MaxSet$, it satisfies that $E(\tau_w) > E(t)$ for any $t \in (\tau_w - d, \tau_w + d)$ and $E(\tau_w) > th$, where d is a pre-defined small distance and th is a threshold. Intuitively, the first local maximum point τ_1 in $MaxSet$ should correspond to the reception of chirp signal traveled directly from the speaker to the microphone and the subsequent local maximum points $\tau_2, ...\tau_W$ represent the reception of echoes. After the searching process, the 0.002s period of the received signal after the first local maximum τ_1, which corresponds to the sound which traveled directly from the device's speaker to the microphone, will be detected as chirp period. Next, the 0.01s period after the chirp period will then be denoted as the echo period and the local maximum point $\{\tau_{w'}, E(\tau_{w'})\}$ with the largest value $E(\tau_{w'})$ in echo period will be detected as the reception of echoes from the user's body for distance estimation. Note that the corresponding delay $\tau_{w'}$ could equivalently represent the sound propagation time for the round-trip path between the array and the steered body region on the user's upper body. Thus, if we denote such distance as D_f, it can be derived using the speed of sound c as: $D_f = \frac{\tau_{w'} \times c}{2}$. Given the corresponding azimuth angle θ and elevation angle φ of the body region, the user-array distance D_p can be calculated as: $D_p = D_f \sin\varphi \sin\theta$. Figure 4 shows such geometric relationship for clarity.

Feasibility Study. We provide a feasibility study to show the effectiveness of our distance estimation scheme. Specifically, we place one microphone array with a tripod in an empty room, and invite one volunteer to stand in front of the array with a distance of about 0.6 meters. We steer the array to the user's upper body and the corresponding azimuth angle θ and elevation angle φ are empirically set as $pi/2$ and $pi/3$ respectively according to the relative positions of the array

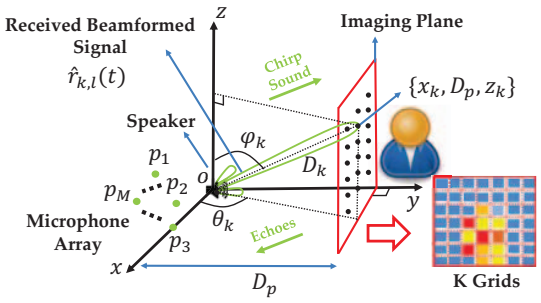


Fig. 6. The illustration of acoustic imaging using beamforming.

and the user. We then collect 20 beep signals for distance estimation purpose and we show how the reception of echoes from the user body is determined using correlation sequence derived from Equation (10) in Figure 5.

From Figure 5, we can observe that the 0.002s period of the received signal after the first peak τ_1 is identified as the chirp period. During this period, the acoustic signal propagates directly from the speaker to the microphones. In addition, we can also observe that the peak with the largest value in the echo period (i.e., $\tau_4 = 0.004$) is detected as the reception of echoes from the user body. The corresponding distances D_f and D_p could be derived as: $D_f = \frac{\tau_4 \times c}{2} = 0.68$ meters and $D_p = D_f \sin pi/2 \sin pi/3 = 0.58$ meters, which is very close to the ground truth (i.e., $D_p = 0.6$m). Such encouraging result confirms the feasibility of using our proposed scheme for the distance estimation.

C. Image Construction

We propose to use echoes reflected from the user body to construct an acoustic image for user authentication. However, accurately identifying echoes from the user's body is a challenging task since the received signal is a mixture of numerous echoes which are reflected back from both human body and other surrounding objects. To cope with this problem, we propose a beamforming based technique to accurately identify echoes reflected from each part of the user's body, and then use them to construct an acoustic image of the user for authentication.

To perform the acoustic imaging, we utilize the derived user-array distance D_p and consider a virtual 2-D square imaging plane with a distance of D_p from the microphone array. In such case, this virtual plane would just 'contain' the surface of the user's body. We first adjust the coordinate system so that the imaging plane is parallel to the x-o-z plane as illustrated in Figure 6. We then equally divide the imaging plane into K grids, and assume the coordinate of the center of the k-th grid is represented as $\{x_k, D_p, z_k\}$. Thus, the corresponding azimuth angle θ_k and elevation angle φ_k can be derived as follows:

$$\theta_k = \arccos \frac{x_k}{\sqrt{x_k^2 + D_p^2}} \tag{11}$$

$$\varphi_k = \arccos \frac{z_k}{\sqrt{x_k^2 + D_p^2 + z_k^2}} \tag{12}$$

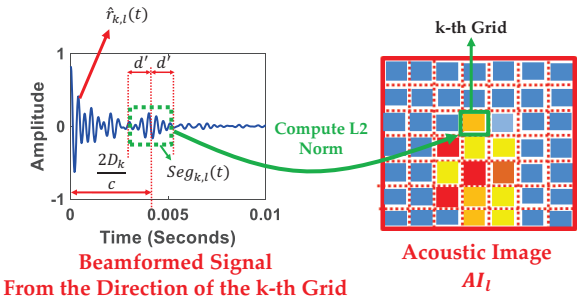


Fig. 7. The illustration of pixel value computation.

The MVDR beamforming is then performed by steering the array to the k-th grid according to Equation (8) using parameters θ_k and φ_k, and such process is repeated to 'scan' each grid on the imaging plane. Specifically, similar as in Section V-B, we use $r_{m,l}(t)$ to denote the received signal for the l-th beep signal from the m-th microphone of the array, and then use $\hat{r}_{k,l}(t)$ to represent the beamformed reflected signal from the direction of the k-th grid on the imaging plane.

To identify echoes created by the user's body, from Figure 6, we can observe that the path length of sound reflected from the user's body in the direction of the k-th grid should be approximately equal to the distance between the k-th grid and the origin point (i.e., $D_k = \sqrt{x_k^2 + D_p^2 + z_k^2}$). However, such relationship does not hold for echoes from other obstacles in the surrounding environments. Thus, inspired by these observations, we take a segment $Seg_{k,l}(t)$ from the beamformed signal $\hat{r}_{k,l}(t)$ which corresponds to direction of the k-th grid and it satisfies: $Seg_{k,l}(t) = \{\hat{r}_{k,l}(t), \frac{2D_k}{c} - d' \leq t \leq \frac{2D_k}{c} + d'\}$ where d' is a pre-defined safeguard distance. This segment could represent the echoes from the human body in the k-th grid direction and we set the corresponding pixel value $P_{k,l}$ as the $L2$ norm of the $Seg_{k,l}(t)$, which is illustrated in Figure 7. In this work, we repeat the above process to scan each grid in an increasing order to construct the final acoustic image AI_l (with size of K grids) from the l-th beep signal for input to the user authentication.

Feasibility study. We provide a feasibility study on how the acoustic images change when reflected echoes are collected from different users. Specifically, we first set the number of grids on the imaging plane as 32400 (i.e., $K = 180 \times 180 = 32400$) and the size of each grid as 0.01 m by 0.01 m. We then let the distance between the user (i.e., user A or B) and the microphone array be 0.7 m and collect 2 beep signals for acoustic image generation purpose. The images are then derived from the received signal and presented in Figure 8. From Figure 8, we can intuitively observe that the acoustic images of a particular user are very similar, while those images between two users differ significantly. This observation confirms the initial feasibility of adopting acoustic images generated by our proposed image construction scheme for accurate user authentication.

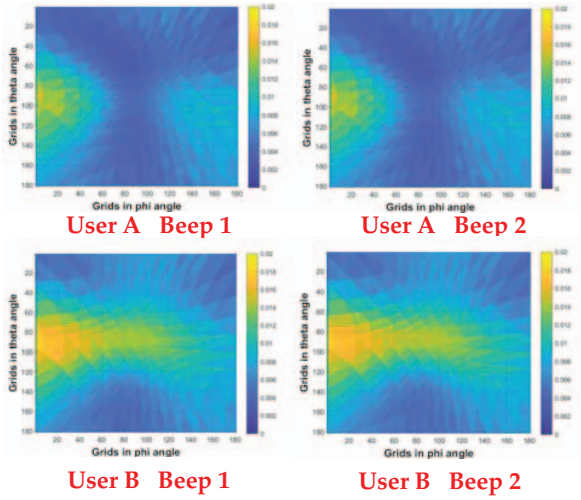


Fig. 8. An illustration of acoustic images of user A and B.

D. Feature Extraction

The acoustic image AI_l, which is characterized by uneven energy enhancements or attenuations in different grids, can represent the unique features of the user. We propose to first extract features from the derived acoustic image AI_l, and then employ classifiers for user authentication. However, traditional feature extraction schemes usually derive features by manual observing unique patterns from the image. These extracted features contain too much redundant information, and they cannot capture the image patterns accurately.

Recently, deep learning approaches based on artificial neural networks have shown great success in a variety of applications and they have also been proven to be effective for feature extraction [39]. However, deep learning models usually need large amounts of data for training and its performance will decrease rapidly when the training set is limited [39]. For these reasons, in this work, we use the concept of transfer learning, which is a machine learning technique with which we are able to transfer knowledge from one previously trained model to another to solve a new problem. Specifically, we adopt the pre-trained VGGish model [40] as the network for feature extraction. The VGGish model is a popular image recognition architecture and it consists of 13 convolutional layers, 5 pooling layers and 3 fully connected layers as illustrated in Figure 9. To transfer the pre-trained model, we resize the image to match the input of VGGish model and keep the pre-trained parameters of the first 18 layers frozen and use the 5-th pooling layer as the output of our feature extractor. The $7 \times 7 \times 512 = 25088$ features are extracted as the input for the classifier.

E. User Authentication

Given the extracted features from the VGGish model, we employ multiple Support Vector Machine (SVM) [41] classifiers for user authentication as illustrated in Figure 10.

For the single-user scenario, when a user registers to our system, the user is required to stand in front of the smart

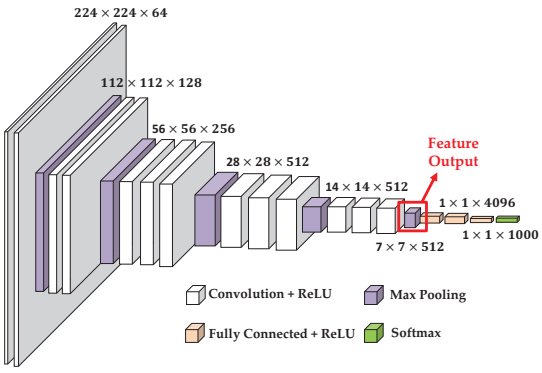


Fig. 9. The illustration of feature extraction using VGGish model.

speaker, our system then extracts the user's unique features from generated acoustic images as the training data. In this case, since we only have this legitimate user's training data while lack of other spoofers' data, we apply a special version of SVM classifier, namely Support Vector Domain Description (SVDD) [42], to train a classifier using only one-class data (i.e., the legitimate user's training data) for distinguishing the legitimate user from other spoofers. For the multiple-user (e.g., n users) scenario, in the register phase, our system extracts features from n legitimate users as training data and we adopt two classifiers for user authentication. Specifically, our system first takes all legitimate users' training data as a whole, and train a SVDD classifier to distinguish all legitimate users from other spoofers. Moreover, after this spoofer detector, a n-class SVM classifier is then trained based on n legitimate users' data to further verify the legitimate user's identity.

In the user authentication phase, the extracted features obtained from acoustic images are input to our proposed authentication model. For the single-user scenario, the SVDD classifier directly outputs a predictive label to determine whether the user is a legitimate user or a spoofer. For the multiple-user scenario, the SVDD classifier first verifies whether the user is one of the legitimate users. If the predictive label is positive, our system would feed features to the second SVM classifier, which will further verify the identity of the user. Otherwise, the label is negative, indicating the presence of spoofer for this particular user. Figure 10 shows this authentication procedure for clarity.

F. Data Augmentation

Ideally, to achieve a robust performance, sufficient training data is required for constructing SVM classifiers. Thus, a user should stand in front of the smart speaker at various distances so as to collect sufficient acoustic images for training purpose during his/her registration. However, this may impose large burdens to the user, and it is also hard to tell when sufficient data has been collected. In this work, to populate the training data, we propose a data augmentation technique for generating 'synthesized' training samples based on the sound propagation model.

The basic idea of our data augmentation scheme is to transform real collected acoustic images into synthesized ones

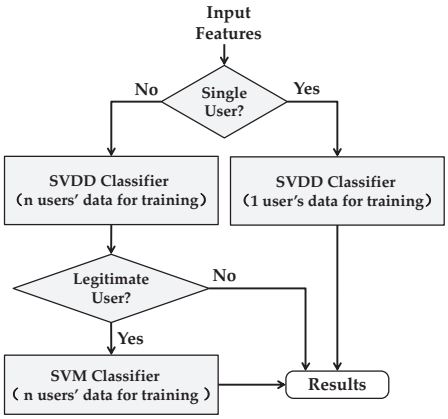


Fig. 10. The illustration of SVM classifiers for user authentication.

User ID	Gender	Age	Occupation
1-5	Male	10-20	Undergraduate Student
6	Female	10-20	Undergraduate Student
7-15	Male	20-30	Graduate Student
16-19	Female	20-30	Graduate Student
20	Male	30-40	Faculty, Staff and Engineer

TABLE I
DEMOGRAPHICS OF SUBJECTS IN THE EXPERIMENT.

by assuming different distances between the user and the smart speaker. Specifically, based on the coordinate system as illustrated in Figure 6, we first consider two acoustic images, and their distances to the origin are denoted as D_p and D'_p, respectively. As described in Section V-C, the coordinates of the k-th grid on these two images could be represented as $\{x_k, D_p, z_k\}$ and $\{x_k, D'_p, z_k\}$ respectively, and their corresponding distances between the k-th grid and the origin point can be calculated as:

$$D_k = \sqrt{x_k^2 + D_p^2 + z_k^2} \tag{13}$$

$$D'_k = \sqrt{x_k^2 + {D'_p}^2 + z_k^2} \tag{14}$$

From the sound propagation inverse-square law [43], we can transform the pixel value $P_{k,l}$ of the first image to the corresponding pixel value $P'_{k,l}$ of the second image as follows:

$$P'_{k,l} = \left(\frac{D_k}{D'_k}\right)^2 P_{k,l} \tag{15}$$

Thus, based on the above transformation model, we can transform the pixel value $P_{k,l}$ in one real acoustic image (with a distance of D_p to the array) to the pixel value $P'_{k,l}$ of any new synthesized image (with a distance of D'_p to the array), thus augmenting our training data.

VI. Performance Evaluation

In this section, we conduct experiments to evaluate the performance of our authentication system EchoImage.

A. Experimental Setup

We conduct experiments with 20 volunteers (ranging from 18 to 39 years old) to evaluate the effectiveness of our EchoImage. The demographics of volunteers are detailed in Table I. A size of 20 volunteers is also typical for user authentication/identification studies [44]. Among the 20 volunteers, 12 of them register with our authentication system while the rest 8 volunteers act as spoofers.

We then implement the prototype of our proposed system on top of a ReSpeaker [45], which is a commercial circular array equipped with six microphones uniformly distributed at the circumference of a circle with an adjacent distance of about 5 cm. An omni-directional speaker is then placed besides the array and this setup is similar as most commercial smart speakers [2]–[4]. During the experiment, we set the bandwidth of the beep signal as 2 to 3 kHz with a length of 0.002s as described in Section V-A. We then connect a laptop to the system to control the transmission and reception of acoustic signals, and the acoustic data is also written into a sound file stored in the laptop during the user authentication process.

1) Experiment Environments: We conduct our experiments under three representative environments. Specifically, without loss of generality, we choose one laboratory room, one conference hall and one outdoor place to conduct experiments. In the experiment, we first keep each place quiet (about 30 dB) to conduct data collection for training. In the testing phase, we either keep the rooms quiet or play music (or people chatting noise and traffic noise) using a computer to collect data for testing. The volume of the computer is set with normal (about 50 dB) and it was placed about 1 to 2 meters away from the microphone array. In addition, residents could also behave normally (e.g., studying or even passing through the test locations) during the whole data collection process. Thus, our experiments include the interference from environmental noises or human activities and it can represent the real-life usage of our authentication system.

To evaluate the consistency of acoustic images extracted by our system over a period of time, we repeat the experiment on different days to complete the data collection for each location: the 0 to 2 days (Session 1), 3 to 7 days after the first day (Session 2), and 8 to 10 days after the first day (Session 3). Specifically, we take 200 chirps from Session 1 as our training set to construct the acoustic images for each user, and then use the rest of 300 chirps collected from Session 1 and 3 to test the performance of our proposed system.

2) Metrics: We use recall, precision, accuracy and F-measure to evaluate the effectiveness of our system. Specifically, we let tp, tn, fp and fn denote the total number of true positives, true negatives, false positives and false negatives, respectively. These metrics are then defined as follows:

Recall: it is defined as $tp/(tp+fn)$, which represents the ratio of the number of correctly identified acoustic images associated with the intended user to the total number of acoustic images associated with the intended user.

Precision: it is defined as $tp/(tp+fp)$, which represents the ratio of the number of correctly identified acoustic images associated with the intended user to the total number of acoustic images which are identified as the intended user.

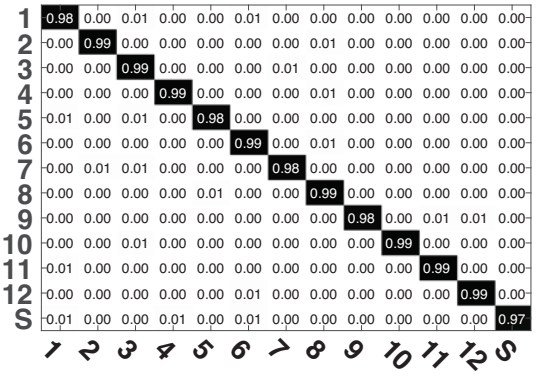


Fig. 11. Confusion matrix of our system.

Accuracy: it is defined as $(tp+tn)/(tp+tn+fp+fn)$, which represents the ratio of the number of correctly identified acoustic images to the total number of acoustic images.

F-measure: the harmonic mean of the precision and recall and it is define as:

$$F = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{16}$$

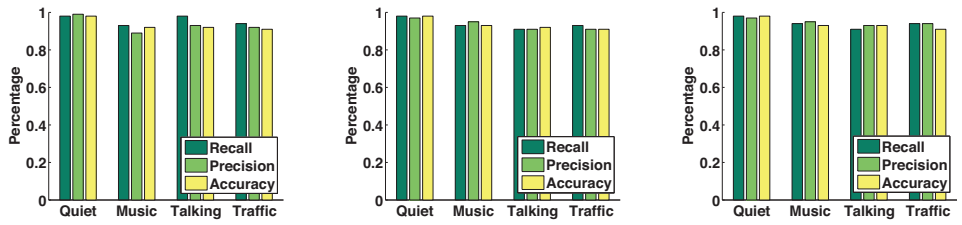
B. Overall Performance

We first evaluate the overall performance of our user authentication system. In this scenario, we conduct the experiments in a quiet laboratory room and set the distance between the microphone array and the user as 0.7 m. Figure 11 shows the confusion matrix of our system for 12 registered users and 8 spoofers. We can observe that our system can achieve over 0.98 accuracy in identifying the registered users and 0.97 accuracy in spoofer detection. These results validate the effectiveness of our system in user authentication.

C. Robustness to Experimental Environments

We next study the robustness of our system when experiments are conducted under different environments. The distance between the microphone array and the user is set as 0.7 m and 8 users are involved in this experiment. Specifically, we evaluate the system performance when experiments are conducted in the laboratory room, conference hall and outdoor place under different environmental noises, and the results are illustrated in Figure 12.

Figure 12(a) to (c) present the recall, precision and accuracy under different environments with various background noises. We can observe that the overall performance is over 0.9 across all environments. This shows that our proposed system can achieve a satisfactory performance for user authentication even if under noisy environments. This is because our system applies a bandpass filter on the received acoustic data and only use 2 kHz to 3 kHz sound for user authentication. Thus, our system will not be significantly impacted by the environmental noises which are mostly below 2 kHz. Further, we can also



(a) Laboratory Room (b) Conference Hall (c) Outdoor Place

Fig. 12. Robustness study under different experimental environments.

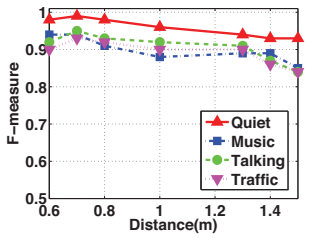
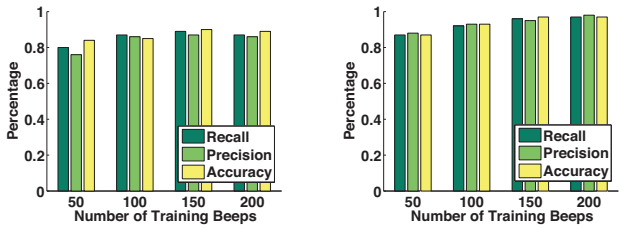


Fig. 13. Performance comparison under different distances between microphone array and user.



(a) Without data augmentation (b) With data augmentation

Fig. 14. Performance comparison of data augmentation with different number of training images.

observe that better performance can be achieved in a quiet environments. This is natural because in a relatively quiet environment, it is still easier to identify the reflected beep sound from the noise. Overall, these results show that our system is effective in user authentication and robust across different noisy environments.

D. Impact of Distance Between Microphone Array and User

In the third set of experiments, we evaluate the effectiveness of our system under different distances between the microphone array and the user. Such distance is an important metric since it is related to the sensing ranging of our system. Specifically, we conduct experiments in the laboratory room by varying the distance between the array and the user from 0.6 m to 1.5 m, and the results are illustrated in Figure 13.

Figure 13 presents the F-measure by varying the distance between the device and the user from 0.6 to 1.5 meters. We observe that the F-measure decreases significantly when the distance is larger than 1 m. This is natural because if the distance between the array and the user becomes relatively large, the echoes become weak and hard to be picked up by the array's microphones and eventually the performance will decrease. In addition, we also observe that our system can achieve over 0.95 F-measure when the distance is less than 1 m under a quiet environment. This result validates that the distance actually has little impact on the accuracy of our system for authentication when it is less than a certain threshold.

E. Impact of Data Augmentation

We then evaluate how our proposed data augmentation scheme can improve the performance of our system by generating synthesized training samples when the training set is limited. Specifically, we first set the distance between the user and the array as 0.7 m to generate acoustic images in the training phase, and then evaluate the performance of our system in Figure 14 with testing images collected from various distances (from 0.6 m to 1.5 m) under two scenarios including one without data augmentation and another with data augmentation. The legends 'recall', 'precision' and 'accuracy' in Figure 14 denote the recall, precision and accuracy of our system, respectively.

Figure 14 (a) and (b) present the recall, precision and accuracy with/without data augmentation when the number of training beeps varies. We can observe that the overall value of recall, precision and accuracy increases when the data augmentation is used, especially when the training images are very limited (e.g., less than 100). These observations show that our data augmentation technique could populate the training data and achieve a robust performance even if the training data is collected from a fixed distance, which could significantly reduce the training burden of the user. In addition, we can also observe that the overall performance becomes stable with more than 100 training samples, which can be collected within one minute when registering a new user.

VII. Conclusion

In this paper, we present an active sensing system EchoImage which enables smart speakers to authenticate users without requiring any additional user-device interaction or pre-installed infrastructure. The proposed system utilizes users' acoustic images, which are derived from the smart speaker by emitting beep signals and sensing echoes from the user's body with its microphone array, as the proof for user authentication. Our distance estimation component applies a correlation based technique on the beamformed signal to estimate the distance between the user and microphone array. To capture the characteristics of the user, our image construction component

then constructs a virtual imaging plane using the estimated distance and steers the array towards each grid of the plane to generate an acoustic image of the user. Moreover, our user authentication component proposes a transfer learning-based method to derive efficient features from the constructed images, and employs SVM classifiers for accurate user authentication. Extensive experiments are conducted to demonstrate that our system is robust and accurate across various scenarios.

Acknowledgments: This work is supported in part by the National Natural Science Foundation of China under Grants 62271124, 61802051, 62172080, 61772121 and 61728102, Natural Science Foundation of Sichuan Province under Grants 2023NSFSC0488 and 2023NSFSC0478, the Fundamental Research Funds for Chinese Central Universities under Grant ZYGX2015J056 and ZYGX2020ZB027, Sichuan Science and Technology Program under Grants 2020JDTD0007 and 2020YFG0298.

REFERENCES

[1] "Smart speakers - Statistics and Facts," https://www.statista.com/topics/4748/smart-speakers/, 2022.
[2] "Amazon Echo," https://www.amazon.com/All-New-Echo-4th-Gen/dp/B07XKF5RM3, 2022.
[3] "Apple HomePod mini," https://www.apple.com/shop/buy-homepod/homepod-mini, 2022.
[4] "Xiaomi Mi," https://xiaomi-mi.us/, 2022.
[5] "Amazon Pay Is Coming To Alexaąŕs Skills." https://www.pymnts.com/amazon/2017/payments-are-coming-to-alexas-skills/, 2022.
[6] Z. Wu, S. Gao, E. S. Cling, and H. Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2014.
[7] P. L. De Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga, "Evaluation of speaker verification security and detection of hmm-based synthetic speech," *IEEE Transactions on Audio, Speech, and Language Processing*, 2012.
[8] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu, "Dolphinattack: Inaudible voice commands," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '17, 2017.
[9] H. Feng, K. Fawaz, and K. G. Shin, "Continuous authentication for voice assistants," in *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*, 2017.
[10] D. A. Reynolds, "An overview of automatic speaker recognition technology," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002.
[11] B. Hutchins, A. Reddy, W. Jin, M. Zhou, M. Li, and L. Yang, "Beat-pin: A user authentication mechanism for wearable devices through secret beats," in *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, 2018.
[12] D. Maltoni, D. Maio, A. K. Jain, and S. Prabhakar, *Handbook of Fingerprint Recognition*. Springer-Verlag, 2009.
[13] E. Uzun, S. P. H. Chung, I. Essa, and W. Lee, "rtcaptcha: A real-time captcha based liveness detection system," in *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, 2018.
[14] L. Blue, H. Abdullah, L. Vargas, and P. Traynor, "2ma: Verifying voice commands via two microphone authentication," in *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, ser. ASIACCS '18, 2018.
[15] C. Huang, H. Chen, L. Yang, and Q. Zhang, "Breathlive: Liveness detection for heart sound authentication with deep breathing," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2018.
[16] S. Pradhan, W. Sun, G. Baig, and L. Qiu, "Combating replay attacks against voice assistants," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2019.
[17] D. Tang, Z. Zhou, Y. Zhang, and K. Zhang, "Face flashing: a secure liveness detection protocol based on light reflections," in *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, 2018.
[18] B. Zhou, J. Lohokare, R. Gao, and F. Ye, "Echoprint: Two-factor authentication using acoustics and vision on smartphones," in *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '18, 2018.
[19] A. Kalyanaraman, D. Hong, E. Soltanaghaei, and K. Whitehouse, "Forma track: Tracking people based on body shape," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2017.
[20] Y. Meng, J. Li, M. Pillari, A. Deopujari, L. Brennan, H. Shamsie, H. Zhu, and Y. Tian, "Your microphone array retains your identity: A robust voice liveness detection system for smart speakers," in *Proceedings of the 31st USENIX Security Symposium (USENIX Security 22)*, 2022.
[21] J. Grythe, "Acoustic camera and beampattern," in *Technical Note Norsonic*, 2015.
[22] I. Dokmanic and I. Tashev, "Hardware and algorithms for ultrasonic depth imaging," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.
[23] J. Sun, R. Zhang, J. Zhang, and Y. Zhang, "Touchin: Sightless two-factor authentication on multi-touch mobile devices," in *Proceedings of CNS*, 2014.
[24] M. Shahzad, A. X. Liu, and A. Samuel, "Secure unlocking of mobile touch screen devices by simple gestures: you can see it but you can not do it," in *Proceedings of ACM MobiCom*, 2013.
[25] J. Chauhan, Y. Hu, S. Seneviratne, A. Misra, A. Seneviratne, and Y. Lee, "Breathprint: Breathing acoustics-based user authentication," in *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, 2017.
[26] T. Zhao, Y. Wang, J. Liu, and Y. Chen, "Your heart won't lie: Ppg-based continuous authentication on wrist-worn wearable devices," in *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, 2018.
[27] T. Zhu, Z. Qu, H. Xu, J. Zhang, Z. Shao, Y. Chen, S. Prabhakar, and J. Yang, "Riskcog: Unobtrusive real-time user authentication on mobile devices in the wild," *IEEE Transactions on Mobile Computing*, 2020.
[28] A. Acar, H. Aksu, A. S. Uluagac, and K. Akkaya, "WACA: Wearable-assisted continuous authentication," in *2018 IEEE Security and Privacy Workshops (SPW)*, 2018.
[29] Y. Lee, Y. Zhao, J. Zeng, K. Lee, N. Zhang, F. H. Shezan, Y. Tian, K. Chen, and X. Wang, "Using sonar for liveness detection to protect smart speakers against remote attackers," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2020.
[30] S. Shen, D. Chen, Y.-L. Wei, Z. Yang, and R. R. Choudhury, "Voice localization using nearby wall reflections," in *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, 2020.
[31] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing*. Prentice Hall, 1993.
[32] B. Van Veen and K. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Magazine*, 1988.
[33] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Springer, Inc., 2008.
[34] Y.-C. Tung and K. G. Shin, "Echotag: Accurate infrastructure-free indoor location tagging with smartphones," in *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, 2015.
[35] Z. Wang, S. Tan, L. Zhang, and J. Yang, "Obstaclewatch: Acoustic-based obstacle collision detection for pedestrian using smartphone," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2018.
[36] S. Pradhan, G. Baig, W. Mao, L. Qiu, G. Chen, and B. Yang, "Smartphone-based acoustic indoor space mapping," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2018.
[37] B. Zhou, M. Elbadry, R. Gao, and F. Ye, "Batmapper: Acoustic sensing based indoor floor plan construction using smartphones," in *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, 2017.
[38] Y. Ren, C. Wang, Y. Chen, J. Yang, and H. Li, "Noninvasive fine-grained sleep monitoring leveraging smartphones," *IEEE Internet of Things Journal*, 2019.
[39] D. Liang and E. Thomaz, "Audio-based activities of daily living (adl) recognition with large-scale acoustic embeddings from online videos," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2019.
[40] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. Weiss, and K. Wilson, "Cnn architectures for large-scale audio classification," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[41] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning, Data Mining Inference, and Prediction*. Springer, 2001.
[42] D. M. Tax and R. P. Duin, "Support vector domain description," *Pattern Recognition Letters*, 1999.
[43] F. Ning, Y. Liu, C. Zhang, J. Wei, X. Shi, and J. Wei, "Acoustic imaging with compressed sensing and microphone arrays," *Journal of Computational Acoustics*, 2017.
[44] R. Liu, C. Cornelius, R. Rawassizadeh, R. Peterson, and D. Kotz, "Vocal resonance: Using internal body voice for wearable authentication," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2018.
[45] "ReSpeaker Core v2.0," https://wiki.seeedstudio.com/cn/ReSpeaker, 2018.