***ChestLive*: Fortifying Voice-based Authentication with Chest Motion Biometric on Smart Devices**

YANJIAO CHEN, Wuhan University, China

MENG XUE, Wuhan University, China

JIAN ZHANG, Wuhan University, China

QIANYUN GUAN, Wuhan University, China

ZHIYUAN WANG, Wuhan University, China

QIAN ZHANG, Hong Kong University of Science and Technology, China

WEI WANG, Zhejiang University, China

Voice-based authentication is prevalent on smart devices to verify the legitimacy of users, but is vulnerable to replay attacks. In this paper, we propose to leverage the distinctive chest motions during speaking to establish a secure multi-factor authentication system, named *ChestLive*. Compared with other biometric-based authentication systems, *ChestLive* does not require users to remember any complicated information (e.g., hand gestures, doodles) and the working distance is much longer (30cm). We use acoustic sensing to monitor chest motions with a built-in speaker and microphone on smartphones. To obtain fine-grained chest motion signals during speaking for reliable user authentication, we derive Channel Energy (CE) of acoustic signals to capture the chest movement, and then remove the static and non-static interference from the aggregated CE signals. Representative features are extracted from the correlation between voice signal and corresponding chest motion signal. Unlike learning-based image or speech recognition models with millions of available training samples, our system needs to deal with a limited number of samples from legitimate users during enrollment. To address this problem, we resort to meta-learning, which initializes a general model with good generalization property that can be quickly fine-tuned to identify a new user. We implement *ChestLive* as an application and evaluate its performance in the wild with 61 volunteers using their smartphones. Experiment results show that *ChestLive* achieves an authentication accuracy of 98.31% and less than 2% of false accept rate against replay attacks and impersonation attacks. We also validate that *ChestLive* is robust to various factors, including training set size, distance, angle, posture, phone models, and environment noises.

CCS Concepts: • **Security and privacy** → **Authentication**; • **Human-centered computing** → **Ubiquitous and mobile computing**.

Additional Key Words and Phrases: User Authentication, Acoustic Sensing, Chest Motion Biometric

ACM Reference Format:

Yanjiao Chen, Meng Xue, Jian Zhang, Qianyun Guan, Zhiyuan Wang, Qian Zhang, and Wei Wang. 2021. *ChestLive*: Fortifying Voice-based Authentication with Chest Motion Biometric on Smart Devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 4, Article 148 (December 2021), 25 pages. <https://doi.org/10.1145/3494962>

Authors' addresses: Yanjiao Chen, chenyanjiao@whu.edu.cn, Wuhan University, China; Meng Xue, xuemeng@whu.edu.cn, Wuhan University, China; Jian Zhang, jzhang@whu.edu.cn, Wuhan University, China; Qianyun Guan, cainan.guan@whu.edu.cn, Wuhan University, China; Zhiyuan Wang, wang_zhiyuan@whu.edu.cn, Wuhan University, China; Qian Zhang, qianzh@cse.ust.hk, Hong Kong University of Science and Technology, China; Wei Wang, wangw@zju.edu.cn, Zhejiang University, China.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2021 Association for Computing Machinery.

2474-9567/2021/12-ART148 \$15.00

<https://doi.org/10.1145/3494962>

Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., Vol. 5, No. 4, Article 148. Publication date: December 2021.

1 INTRODUCTION

Voice assistants on smartphones use wake-up words to activate the system and authenticate the user. With the help of voice assistants, users can perform various personal businesses, including sending messages, making calls, and checking emails [9]. However, such voice-based authentication systems are vulnerable to replay attacks and impersonation attacks, leading to serious consequences of financial loss and privacy leakage.

To strengthen voice-based authentication, researchers leverage many kinds of biometrics to fortify voice-based authentication systems. Shang et al. [23] examined the vibration of vocal cords to defend against voice spoofing, which requires users to press the phone against the throat. Chen et al. [4] exploited the magnetic field emitted from loudspeakers as the evidence of replay attacks, which needs users to move the phone in a specific trajectory during authentication. Other works [12, 25, 36, 37] have utilized pop noise, lip movement, articulatory gestures, and the vocal system as biometrics for voice-based authentication. However, these systems either restrict the phone position to be exact during authentication or have a limited working distance (usually less than 12cm).

In this paper, we propose *ChestLive*, a reliable multi-factor authentication system with sufficient working distance. *ChestLive* utilizes chest motions caused by speaking as the distinctive biometric to differentiate legitimate users and attackers. In particular, we observe that the chest motion patterns during speaking is vastly different from that during silence. While chest motions in silence have a monotone cyclic pattern, the chest motions during speaking are modulated by the specific voice command. To achieve non-intrusive and contactless monitoring of chest motions, we adopt acoustic sensing with a built-in speaker and microphone of commercial smartphones. When an attempt to wake up the smartphone is detected, the speaker emits inaudible acoustic signals and the microphone collects received signals that contain both the voice commands and the reflected ultrasound signals from the user's body. We apply a low-pass filter and a high-pass filter to separately obtain the low-frequency audible voice signal and the high-frequency inaudible signal reflected by chest motions.

To corroborate voice-based authentication with chest motions, we face several challenges in practice. First, to the best of our knowledge, none of the previous works have explored the correlation between voice commands and corresponding chest motions. It is necessary to extract representative features from both voice commands and chest movement simultaneously for reliable authentication. Second, recovering clear fine-grained chest motion signal is non-trivial since the changes in chest motion due to speaking is subtle and is susceptible to interference. The final challenge is a lack of samples to train a reliable authentication model for a new user. Machine learning algorithms are hungry for training data, but unlike images, sensing data of biometric characteristics is few. Most existing works collect relatively large numbers of sensing data samples under the laboratory environment, but in practice, the enrollment phase for a new user is short, during which a limited number of samples can be gathered.

To solve the above challenges, we first investigate the feasibility of utilizing chest motions as biometrics, and find that the correlation between voice and corresponding chest motions presents unique patterns for different individuals.

Motivated by the observation, we design the architecture of *ChestLive*, which proceeds through data collection, acoustic signal analysis, liveness detection, feature extraction and user authentication. For data collection, the smartphone placed in front of the user transmits inaudible acoustic signals towards the user, and the microphone receives both the (audible) voice and the inaudible signal reflected from the chest simultaneously. To pinpoint the period during which the user speaks the wake words, we leverage an adaptive context attention model (ACAM) [10, 21] to adaptively remove background noise and partition the speaking events. As the changes in chest motions due to speaking is subtle and easily drowned by environmental noises, we develop an acoustic signal analysis framework to obtain clear chest motion signal. Firstly, we transform the ultrasound into the frequency domain by Fast Fourier transform (FFT) to find the main doppler frequency magnitude and its corresponding energy. However, each FFT coefficient contains χ^2 noise. To remove these noises and aggregate the energy of main reflections over time, we divide the received signal into tiny frequency bins, and calculate the cosine distance

Fig. 1. *ChestLive* use scenario.

Fig. 2. Chest motions of 3 users saying "Hey, Siri".

among the frequency bins to discover the main doppler frequency magnitude from the spectrogram. After this, we utilize probabilistic principal component analysis (PPCA) [24] upon the various modulated Doppler signals for noise removal and obtain the primitive chest motion signal. Nonetheless, the primitive signal suffers from static and non-static interference from the environment. To recover the fine-grained chest motion signal, we develop a background model to remove static interference and leverage Empirical Wavelet Transform (EWT) [6] to eliminate non-static interference due to the movements of other body parts. The chest motion signal after interference removal is used to perform liveness detection for quickly screen simple replay attacks launched by electronic speakers only. To defend against impersonation attacks and replay attacks (involve both electronic speakers and impersonators), we correlate the chest motion signal with the voice signal and extract Mel-frequency cepstral coefficients (MFCCs) as features to uniquely identify legal users. For user authentication, our biggest challenge is how to establish a reliable authentication model for a new user with only a few samples collected during enrollment. To tackle this problem, we develop a meta-learning framework, which first builds a general model with massive collected data that has good generalization property, and then fine-tunes the general model to the new user with only a few training samples.

We have implemented *ChestLive* as a smartphone application, which is distributed to volunteers for remote assessment due to the lockdown of COVID-19. The typical use scenario of *ChestLive* is shown in Fig. 1. It is only required that the smartphone is placed in front of the user's chest, e.g., held by hand or put on the desk. No complicated gestures are required and the position of the smartphone is flexible. To demonstrate the effectiveness of *ChestLive*, we have recruited 61 volunteers (20 females, 41 males) aged 20 to 49. It is validated that *ChestLive* can achieve a high authentication accuracy and can defend against replay attacks and impersonation attacks with a low false accept rate. We also validate the robustness of *ChestLive* in terms of distance, direction, noise, phone model, and voice command.

We summarize our key contributions as follows.

- We explore a new two-factor authentication paradigm of leveraging the breathing pattern of the speaker to corroborate voice authentication. The proposed *ChestLive* system is more robust to the displacement of the smartphone than existing approaches.

- We leverage the channel energy of acoustic signals to capture the chest movement, and recover ideal chest motion signals based on probabilistic principal component analysis (PPCA) and empirical wavelet transform (EWT).
- We design a meta-learning based authentication approach to address the cold-start problem of new users. A general neural learning network that is trained with data across numerous existing users can quickly be fine-tuned for the new user with a few new samples collected during enrollment.
- We conduct experiments with 61 volunteers in the wild using off-the-shelf smartphones. Extensive evaluation results show that the meta-learning based authentication approach can reach high accuracy with only a few samples from a new user. *ChestLive* can defend against both replay and impersonation attacks. In addition, *ChestLive* can realize reliable authentication at a working distance of 30cm, and is robust to various influential factors, including different phone angles, phone types, voice commands, user posture and ambient noises.

2 RELATED WORK

In this section, we briefly review state-of-the-art literature on wireless human-body movement detections and voice-based multi-factor authentications.

2.1 Acoustic Sensing Based Motion Detection on Smartphones

RF signals [2, 33] and WiFi signals [1, 11, 34] have been used to monitor chest motions. However, these methods require additional RF or WiFi equipment. In comparison, acoustic sensing can be implemented by the smartphone using built-in speakers and microphones. It has been widely used in tracking human-body movements [13, 17, 22, 27]. Ruan et al. [22] proposed a device-free gesture recognition system that leverages the Doppler effect caused by hand motion. CAT [13] uses Frequency Modulated Continuous Wave (FMCW) and Doppler effect to estimate the distance between an audio transmitter and a receiver. To track the movement of fingers, Nandakumar et al. [17] utilized Orthogonal Frequency Division Multiplexing (OFDM) to monitor the finger's motion. LLAP [27] depended on the phase change of Multiple Continuous Wave to track the hand/finger movement. Pioneering studies in ultrasound have explored dangerous driving behaviors [31], fatigue driving [29], smoking detection [30].

Although acoustic signals have been used to sense motions, these methods cannot be directly applied to chest motion detection because these movements have a large range of motion. There is a handful of studies on acoustic sensing based chest motion detection. Nandakumar et al. [16] estimated the chest motions by Frequency-Modulated Continuous Wave (FMCW) of acoustic signals. Wang et al. [26] improved the robustness and precision of chest motion detection by calculating the correlation of the emitted and the received FMCW of acoustic signals. Nonetheless, these two solutions did not consider the interference from the motions of other body parts, thus are easily affected by light body movements of the user. BreathListener [32] utilizes ultrasound's Energy Spectrum Density (ESD) that is based on Doppler effect to recover driver's breath signal, and utilize Generative Adversarial Network (GAN) [7] to refine the breathing waveform. However, the GAN network can remove or cripple the pattern that caused by speaking voice commands. Furthermore, the Doppler effect caused by chest and abdomen movement while speaking is significantly different from normal breath. Therefore, we design Channel Energy of acoustic signals to capture the chest movement that is correlated with the voice signal for user authentication while speaking, and then de-noise and aggregate CE, and further cancel the static and non-static interferences.

2.2 Voice-based Two-factor Authentication on Smartphones

Voice has been widely used for hands-free authentication. However, voice-based authentication is vulnerable to replay attacks. Therefore, many efforts have been made to strengthen voice authentication with other authentication factors, especially liveness detection. VAuth [5] used accelerometers to collect the vibrations of the user's body surface and matched them with the speech signals for authentication. However, the accelerometers are deployed in an extra device, e.g., eyeglasses, earbuds, necklaces. Shang et al. [23] examined the throat movement and the high-frequency vibrations of the vocal cords to defend against voice spoofing. In order to obtain the throat movement, the smartphone needs to touch the user's throat. Chen et al. [4] exploited the magnetic field emitted from loudspeakers as the evidence of replay attacks, where users are required to move the smartphone in a certain trajectory in front of their mouths when speaking the voice commands. VoicePop [25] extracted the features of pop noises produced by breathing during speaking for liveness detection, but the authentication accuracy dropped to below 80% when the distance between the smartphone and the mouths is more than 5cm. VoiceLive [37] detected live users by their unique vocal systems. VoiceGesture [36] exploited articulatory gestures for liveness detection for voice-based authentication. However, both VoiceLive and VoiceGesture requires the phone to be within 3cm of the mouth. LipPass [12] extracted the Doppler effect of lip movement during speaking for authentication using acoustic signals on smartphones, but the authentication accuracy dropped to below 90% when the distance between the smartphone and the lips is more than 12.5cm.

Our proposed *ChestLive* correlates voice and chest motions during speaking to reach a reliable authentication. *ChestLive* is robust to the displacement of the smartphone and can achieve reliable authentication at a working distance of 30cm. We have addressed the problem of lacking training samples from new registered users with meta-learning. Our experiments with 61 volunteers verify the effectiveness of *ChestLive*.

3 PRELIMINARY

3.1 Speech Production Mechanism

Speech production mainly involves three systems, namely, the respiratory system, the vocalization system, and the voice-forming system. Speech production interrelates with breathing in three phases. Firstly, before each pronunciation, a person usually inhales to obtain the air used for the vocal system to produce the speech. Then, through exhalation, the airflow presses against the glottis. An increase in the glottis pressure increases the speed of the glottis airflow, causing the glottis to close and the vocal cords to move and vibrate. Finally, the vibration of vocal cords produces pulse signals to form sounds of different timbres delivered by the mouth (including mandible, lips, tongue, and soft palate). Speech production involves many organs, including lungs, rib cage, diaphragm, trachea, muscle, and voice-forming organs [28]. Coordinated movements between various organs play an essential role in producing clear and meaningful speech sounds.

3.2 Speaking and Chest Motion

The speech production mechanism has been exploited to corroborate voice authentication. Existing works have associated speeches with throat movement [23], lip movement [12], articulatory characteristics [36], and breathing patterns [20] for authentication. These physiological features are captured by accelerometers, WiFi signals, and acoustic signals. Nevertheless, existing authentication methods are limited by their accessibility and working distance.

We observe that the chest motion is an ideal indicator of speaking, and the chest movement pattern while speaking of different users exhibits distinctive features due to individual differences. The chest movement pattern while speaking is more difficult to imitate than voice or hand gestures since it is mostly determined by users' physiological and behavioral characteristics. Speaking-based authentication is convenient and natural since users are not required to remember or perform complicated gestures. While breathing during silence mainly shows a

Fig. 3. The 39-dimensional MFCC of the voice and the correlation between the voice and the chest motions when the user is saying "Hey, Siri". In each subfigure, the left and the right graphs present "Hey" and "Siri", respectively.

Fig. 4. The 39-dimensional MFCC of the correlation between the voice and the chest motions when different users are saying "Hey, Siri". In each subfigure, the left and the right graphs present "Hey" and "Siri", respectively.

cyclic pattern, breathing during speaking is shaped by the corresponding speech. In particular, speaking changes the pattern of chest motions. Fig. 2 demonstrates that when saying the same wake word "Hey Siri", the breathing pattern of users is different from their normal breathing pattern, and the breathing patterns of different users are diversified.

To verify the feasibility of using chest motions as biometrics for user authentication, we first show that the chest motions are affected by the speech, and then demonstrate that the chest motions of different users are distinctive when saying the same words. To analyze how speech modulates the chest motions, we calculate the correlation between the speech and the corresponding chest motions in the time domain. As the change in the chest motions due to speech is subtle, we extract Mel-frequency cepstral coefficients (MFCCs) from the correlated signal as features. Fig. 3 shows the MFCC features of the voice signal and the correlation signal when the user is saying "Hey Siri". It is shown that when the user is saying different words, e.g., "Hey" and "Siri", the features of voice signals are vastly different, and the features of the correlation signals are also distinctive. This confirms that the chest motions are indeed affected by different speeches. To show that the chest motions of different users are distinctive when saying the same words, we present the MFCC features of three different users when they are speaking "Hey, Siri" in Fig. 4. Note that Fig. 4(b) and Fig. 3(b) are from the same user at different sessions. We can observe that the MFCC features of the correlation signal are different in different users when they are saying the same words, and the MFCC features of the correlation signal are similar in the same user when the user is saying the same words at different sessions. Thus, we can leverage the correlation between the voice and

its corresponding chest motion to authenticate users. By correlating the chest motion with the speech, we can corroborate voice authentication by providing breathing-based liveness detection and enriching the feature space by analyzing the characteristics of chest motions.

Acoustic signals provide a convenient and accessible way to capture the chest motions. Most smartphones are equipped with a speaker that can emit acoustic signals of a relatively wide frequency range and microphones that can sample and analyze the received data. The authentication process is contactless, and our experiments show that the working distance is as large as 30cm. As shown in Fig. 1, users only need to hold the phone in front of their chests in a natural way and speak the voice commands. The speaker transmits an inaudible ultrasound towards the users, and the microphone collects the mixed signal of the voice and the reflected signal by the chest. In the detailed design of *ChestLive*, we will describe how to separate the mixed-signal and achieve a robust authentication.

3.3 Threat Model

We consider replay attacks and impersonation attacks against *ChestLive*.

3.3.1 Replay attacks. Since *ChestLive* considers two authentication factors, namely, voice and chest motions, which are both captured by the microphone, two types of replay attacks may be launched.

- *Simple replay attacks.* In simple replay attacks, the attacker only records and replays the recording of the user's voice, which can be acquired relatively easy and from a distance. This simple attack is proved to be effective against the Automatic Speech Recognition (ASR) systems of many commercial smartphones [35].
- *Advanced replay attacks.* In advanced replay attacks, the attacker is able to record the acoustic signal that contains both the voice and the sensing signal reflected by the chest from a vantage point. This is possible if the attacker is close to the user when the user is performing authentication.
- *Duplicate replay attacks.* In duplicate replay attacks, the attacker is able to obtain the the legitimate recordings from the victim's device, e.g., by hacking the victim's device.

3.3.2 Impersonation attacks. The attacker plays the recording of the legitimate user's voice and, at the same time, mimics the chest motions of the victim user.

4 SYSTEM DESIGN

The overview of *ChestLive* is shown in Fig. 5. There are five main modules, namely, data collection, acoustic signals analysis, liveness detection, feature extraction, and user authentication.

- *Data collection.* To sense the chest motions during speaking without interfering the voice signals, we leverage inaudible acoustic sensing. The speaker transmits inaudible acoustic signals towards the user, and the microphone receives both the (audible) voice and the inaudible signal reflections from the chest simultaneously.
- *Acoustic signal analysis.* To reconstruct a fine-grained chest motion signal for reliable authentication, we develop an acoustic signal analysis framework. Firstly, we transform the signal into the frequency domain and compute the channel energy (CE) of a series of time-frequency bins. By de-noising and aggregating the CE, we build the primitive chest motion signals. Then, we further remove the static and non-static interference using background model and EWT to obtain a clear chest motion signal.
- *Liveness detection.* The chest motion signal enables us to quickly screen simple replay attacks launched by electronic speakers only. If the liveness detection result is negative, we will reject the authentication request immediately without going through the following steps; otherwise, we proceed to determine whether the live person is authentic or not.

Fig. 5. Overview of *ChestLive*.

- *Feature extraction.* We calculate the correlation between the voice signal and the corresponding chest motion signal, and extract MFCC as features.
- *User authentication.* To tackle the problem that a new user has only a few samples collected during enrollment for training, we design a meta-learning framework. We first train a general model with ideal generalization property using massive data from numerous users collected beforehand, and then fine-tune the general model to a new user using only a few samples.

4.1 Data Collection

ChestLive leverages active voice commands and corresponding chest movement to extract unique physical characters of arousing voice assistants for authenticating legitimate users. Therefore, The first step is to extract active voice commands and the chest motion simultaneously from the microphone received signal.

4.1.1 Obtaining Chest Movement Signals. *ChestLive* uses the built-in speaker of the smartphone to send a pilot acoustic signal $S(t) = A_0 \sin(2\pi f t + \varphi)$, where A_0 is the amplitude, f is the frequency, and φ is the initial phase. After being reflected by the user's chest and other objects in the environment, the signals received by the built-in microphone in the same smartphone can be represented as

$$R(t) = \sum_{i \in \Omega_I} A_i \sin(2\pi f_i t + \varphi_i), \quad (1)$$

where Ω_I is the set of all paths of reflected acoustic signals, A_i , f_i and φ_i are the amplitude, frequency and phase of path i , respectively. A_i approximately satisfies $A_i \propto 1/d_i$, where d_i is the propagation distance of path i . In our experiments, we utilize the speaker in the smartphone to emit a pilot acoustic signal at $20kHz$, which is inaudible to human ears. We collect the reflected signals with the microphone in the smartphone at a sampling rate of $44.1kHz$, which is supported by most off-the-shelf smartphones.

To capture the chest movement while speaking and normal breathing, we first extract the ultrasound signals that contain the chest movement from the microphone received signal through a Butterworth high-pass filter whose cut-off frequency is $18k$ Hertz. After we extract the ultrasounds that contain the chest movement, the next question is how to obtain chest motion signals from it. To solve this problem, we refer to the ESD that is utilized in [32] to design Channel Energy (CE), which is the generalization of ESD. In other words, ESD is a special case of CE when the band of frequency is set to 20 Hertz. Specifically, we divided the received signal into tiny frequency bins, and then we take the main doppler frequency magnitude from the spectrogram over time by calculating the cosine distance among the tiny frequency bins.

Now, we describe how to calculate the Channel Energy of the received signal, select the signals that contain chest motion information, and construct the chest motion signal.

Calculate the Channel Energy: The reflected signals mainly consist of three parts, namely, the signals reflected by the chest and abdomen, the signals reflected by other body parts, and the signals reflected by the objects in the environment. The latter two parts that caused noise can be removed in 4.2. Therefore, we divided the received signal into tiny frequency bins to calculate the channel energy. Specifically, we perform 8192-FFT on $R(t)$ to obtain a spectrogram as shown in Fig. 6. Let $FFT(f, t)$ denote the amplitude of a time-frequency bin on the spectrogram. Each time-frequency bin has a frequency range of 5Hz and a time range of 2s (the sample rate is 44.1Khz, thus 8192 samples have a time interval of 83ms). We calculate the single channel energy of each time-frequency bin by aggregating the energy of the FFT coefficients within the 5Hz frequency range of f^* , and the Channel Energy (CE^1) can be presented as,

$$CE_{t,f} = \frac{1}{N} \sum_{f^* \in f} (FFT(f^*, t))^2, \quad (2)$$

where $N = 8192$ is the number of FFT points, f^* is the frequency bin that has 5Hz frequency band, and f is the whole frequency ranges. And CE_{t,f^*} presents single channel energy. For each frequency shift, we obtain a time-series Channel Energy vector as $CE_f = [CE_{t_0,f}, CE_{t_1,f}, CE_{t_2,f}, \dots, CE_{t_K,f}]$.

To get the main doppler frequency magnitude from the spectrogram over time, we compute the cosine distance between the two single channel energy vectors of a frequency shift as

$$Dis_f = 1 - \frac{CE_{t,f_i^*} * CE_{t,f_j^*}}{|CE_{t,f_i^*}| * |CE_{t,f_j^*}|}, \quad (3)$$

where $f_i^*, f_j^* \in f$. Then, we convert the vector-form distance vector Dis_f to a square-form distance matrix DIS_f , where the row i and column j in DIS_f present the distance of E_{t,f_i^*} and E_{t,f_j^*} . After this, we calculate the sum of all rows for each column and select the max column as the reference single channel energy vector. Therefore, we present the cosine distance between the single channel energy vector of a frequency shift and the reference single channel energy vector as DIS_F . We only keep the frequency shifts that have a significant cosine difference, i.e., $DIS_F > \alpha$, which indicates that the frequency shift contains useful information of chest motions. In our experiments, we set α as 0.96. In this way, we can extract the main doppler frequency magnitude from the spectrogram over time.

De-noising and Aggregating CE: Since speaking and breathing mechanism of human is different, the Doppler effect caused by normal breathing and speaking is significantly different. The difference of Doppler effect will scatter the chest movement while speaking or breathing to different frequency bins, which makes the extract of chest motion harder. Therefore, the next question is how to remove the noises that are resulted in the difference of Doppler effect. To remove these noises and aggregating the main reflections over time, we utilize probabilistic principal component analysis (PPCA) [24] upon the various modulated Doppler signals to effectively de-noise the doppler reflections and recover the primitive chest motion. In this way, rich information related to chest motions is preserved, which results in more reliable user authentication.

As shown in Fig. 7, we obtain several single channel energy signals corresponding to different frequency shifts, which are around 19955Hz, 19970Hz, 20020Hz, and 20025Hz with the frequency band of 5Hz. These signals are similar to each other with observable breathing cycles, but they also contain glitches due to unwanted noises. To integrate the useful information of all frequency shifts while removing noises, we resort to probabilistic principal components analysis (PPCA) [24], an advanced version of PCA. PPCA is based on a probability model and is faster than PCA. We leverage PPCA to combine the multiple signals in Fig. 7 into one signal that contains rich information of chest motions but no irrelevant noises.

¹In order to be consistent, our subsequent processing on the signal is still called CE.

Fig. 6. Different frequency bins in 2 seconds.

Fig. 7. The information of chest motions in different frequency bins.

Let $CE_h(t)$ denote the set of Channel Energy of all frequency shifts at time t . $CE_h(t)$ is a $1 \times n$ vector, where n is the number of preserved frequency shifts. Let $CE_l(t)$ denote the output of PPCA given $CE_h(t)$ at time t . CE_l is a scalar.

Assuming that CE_l is the latent factor that reflects the true chest motions, according to PPCA, the observed signals CE_h follows a multivariate normal distribution

$$CE_h | CE_l \sim N(CE_l \mathbf{W}, \sigma^2 \mathbf{I}), \quad (4)$$

where $\mathbf{W}$ is a $1 \times n$ vector known as the principal axes, σ^2 is the noise term, and $\mathbf{I}$ is a $1 \times n$ identity matrix.

According to PPCA, CE_l follows the distribution of

$$E_l \sim N(\mu, \mathbf{W} \mathbf{W}^T + \sigma^2), \quad (5)$$

where μ is the mean of CE_h . We can utilize the expectation maximization algorithm [14] to obtain σ^2 and $\mathbf{W}$, and finally compute CE_l .

As shown in Fig. 8, we compare the recovered chest motion signals using ESD [32] and using *ChestLive*. It is shown that *ChestLive* restores more precise chest motions by de-noising and aggregating the main reflections. It seems that the peaks of *ChestLive* at 3.5s, 7.2s, 11.2s, and 11.4s in the bottom figure correspond to the dent of [32] in the top figure. However, different from [32] that focuses mainly on obtaining breathing cycles, *ChestLive* aims to extract fine-grained features from the signal of chest motions for authentication. Therefore, the signal recovered by [32] is not suitable for our purpose.

4.1.2 Obtaining Voice Command Signal. To obtain the chest motion signals corresponding to the voice command, we need to detect the periods during which the user is saying the wake-up words. We conduct voice activity detection (VAD) using an adaptive context attention model (ACAM) [10], which removes background noise adaptively. We train ACAM with the dataset from [10], which is collected in various environments including bus stops, construction sites, parks, and rooms. Since the sampling rate of [10] is 16kHz, we need to down-sample our sampling rate from 44.1kHz to 16kHz. As shown in Fig. 15, ACAM can detect the occurrence of voice activities. However, the detected area also contains redundancies. Therefore, we further pinpoint the periods of wake-up words by [21]. Fig. 16 depicts the finer detected area of the wake-up words "Hey, Siri".

Fig. 8. Chest motion extraction using ESD (top) and *ChestLive* (bottom).

Fig. 9. Static interference elimination.

Fig. 10. Non-static interference elimination based on EWT decomposition.

4.2 Acoustic Signals Analysis: Removing Static and Non-static Interference

The goal of *ChestLive* is to fortify voice-based authentication on smart devices. And, the environment of the user's login attempts is different. Therefore, the next problem is how to remove the signals that are reflected from the environments. To this end, we design a static interference elimination algorithm to eliminate the impact of different environments. Except for the interference from different environments, the received signal may contain the interference from the slight motion by other body parts of the user. Therefore, we design a non-static interference elimination algorithm to remove the interference from other body parts of the user and extract the signals reflected from the user's chest. Specifically, according to Section 4.1.1, the microphone received ultrasound signal contained the noise of other body parts and the objects in the environment. Therefore, it is still a critical problem to extract the chest movement form Channel Energy signals due to the interferences in the environment. Specifically, interference comes from two sources. One is the reflections from static objects in the environment, and the other is the reflections from other body parts of the user that may be moving. To extract the chest motion while speaking and normal breathing, we conduct sophisticated interference elimination. Our interference elimination algorithm is summarized in Algorithm 1.

4.2.1 Static Interference Elimination. Static inference mainly results from reflections of objects in the environment. Since the user may be present in different environments during authentication, the static interference elimination algorithm must be adaptable. Lines 1~10 of Algorithm 1 deal with static interference elimination. Every time the user activates the voice assistant, we build a background model based on the surrounding environment. To adapt to environmental changes (e.g., the user is on the go), we update the model based on the ratio of the absolute value and the maximum value of the current signal sequence to those of the next signal sequence [19]. Finally, the signal is smoothed by a sliding window of 10s width. Fig. 9 illustrates that the signal is adjusted to nearly zero-mean after static interference elimination.

4.2.2 Non-static Interference Elimination Based on EWT. Apart from static interference, the signal may also be reflected by other body parts of the user that may be stationary (e.g., head) or slightly moving (e.g., arm while walking). The normal breathing rate of humans is 10~35 breathing cycles per minute (bps) [15]. Since the breathing rate overlaps with the frequency of movements of other body parts, we cannot remove non-static interference through filtering. Lines 11~14 of Algorithm 1 tackle non-static interference elimination.

To separate the chest motion signal from the signals reflected by other (non-static) body parts, we utilize the idea of empirical wavelet transform (EWT) [6], which decomposes a mixed-signal into multiple signals by

Algorithm 1: Static and Non-static interference elimination.

Input: Signal sequence $CE(t)$, $t \in [1, T]$, the number of measurements used for building the background model k .

Output: $Y(t)$, the signal without static and non-static interference.

- 1: /* Remove static interference*/
- 2: Initialize background model $B = \frac{1}{k} \sum_k CE(k)$, $B_{old} = 0$, and $i = k$.
- 3: **while** $i < T$ **do**
- 4: $\mu = |CE(i) - B_{old}| / \max(CE(i), B_{old})$.
- 5: $B = (1 - \mu) * B + CE(i) * \mu$.
- 6: $B_{old} = CE(i)$.
- 7: $CE(i) = CE(i) - B$.
- 8: $i = i + 1$.
- 9: **end while**
- 10: Calculate the moving average of a 10s time window.
- 11: /* Remove non-static interference*/
- 12: Conduct Empirical Wavelet Transform to get multiple decomposition components.
- 13: Calculate FFT for each component obtained from EWT.
- 14: Find the component $C(i)$ within the frequency range 0.16Hz-0.6Hz.
- 15: Return $Y(t) = C(i)$.

building adaptive filter banks. It is shown that EWT has better performance than Ensemble Empirical Mode Decomposition (EEMD) [3] used in [32]. Therefore, we use EWT to eliminate non-static interference.

EWT consists of three steps, i.e., estimating the components of the spectrum, segmenting and determining boundaries of the range, and defining Meyer filter banks. In the frequency domain, the scaling function ϕ_1 and the empirical wavelets ψ_i is defined as

$$\phi_1(\omega) = \begin{cases} 1, & |\omega| \leq (1-\gamma)\Omega_1, \\ \sin(\frac{\pi}{2}\beta(\gamma, \Omega_1)), & (1-\gamma)\Omega \leq |\omega| \leq (1+\gamma)\Omega_1, \\ 0, & \text{otherwise,} \end{cases} \quad (6)$$

$$\psi_i(\omega) = \begin{cases} 1, & (1+\gamma)\Omega_i \leq |\omega| \leq (1-\gamma)\Omega_{i+1}, \\ \cos(\frac{\pi}{2}\beta(\gamma, \Omega_{i+1})), & (1-\gamma)\Omega_{i+1} \leq |\omega| \leq (1+\gamma)\Omega_{i+1}, \\ \sin(\frac{\pi}{2}\beta(\gamma, \Omega_i)), & (1-\gamma)\Omega_i \leq |\omega| \leq (1+\gamma)\Omega_i, \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

where $0 < \gamma < 1$ is used to control the overlap between consecutive transition phases, $\beta(\gamma, \Omega_i) = \beta(\frac{1}{2}\gamma\Omega_i(|\omega| - (1-\gamma)\Omega))$, $\Omega_i = \frac{(\omega_i + \omega_{i+1})}{2}$, where $1 \leq i \leq N-1$, ω_i and ω_{i+1} are frequencies of boundaries, and N is the total segments. With $\phi_1(\omega)$ and $\psi_i(\omega)$, we can compute the approximation Φ and the detailed coefficient Ψ .

$$\Phi = \int E(\tau)\phi_1(\tau - t)d\tau = \text{IFFT}(E(\omega) * \phi_1(\omega)), \quad (8)$$

$$\Psi = \int E(\tau)\psi_i(\tau - t)d\tau = \text{IFFT}(E(\omega) * \psi_i(\omega)). \quad (9)$$

Fig. 10 demonstrates the decomposition results. The signal after static interference elimination is decomposed to three main components, i.e., $Ewt - C1$, $Ewt - C2$, and $Ewt - C3$. We apply FFT to each decomposition component.

We keep the component whose frequency range mostly lies within 0.16Hz~0.6Hz [32], which corresponds to human breathing rate ($Ewt - C1$ in Fig. 10).

4.3 Liveness Detection: Detecting User Presence

After interference elimination, we perform liveness detection on the obtained chest motion signals to determine whether a live user is present or not. *ChestLive* conducts liveness detection based on the human breath. Specifically, we compute FFT on the chest motion signals and select the peak in the frequency range extracted by a band-pass filter (0.16Hz ~ 0.6Hz). To distinguish whether a live user is present or not, *ChestLive* sets the threshold to be 2.

Fig. 11 and Fig. 13 show the time-domain and the frequency-domain signals after all previous steps when there is no live user. As we can see, the time-domain signal is almost flat, and the frequency-domain signal has no dominant peaks. Fig. 12 and Fig. 14 display the time-domain and the frequency-domain signals after all previous steps when there is a live user performing authentication. The patterns in Fig. 12 and Fig. 14 are distinctively different from those in Fig. 11 and Fig. 13. In particular, we also show the ground-truth breathing signals measured by the Neulog belt in Fig. 12 and Fig. 14. It is shown that our extracted signals are very close to the ground-truth signals collected by the belt.

Liveness detection enables us to screen replay attacks performed by electronic speakers in an efficient way. If there is no live user, we can directly reject the authentication request without performing feature extraction and user authentication; otherwise, the system proceeds to the following steps. Note that liveness detection cannot defend against impersonation attacks and advanced replay attacks.

4.4 Feature Extraction: Correlating Chest Motions with Voice

We jointly combine the features of user's voice and the chest motion simultaneously for more reliable authentication. To achieve this goal, we need to synchronize the voice signal and the chest motion signal. More specifically, we are interested in the distinctive chest motions when the user is saying the commands but not the chest motions during silence periods.

We first extract the voice command $V(t)$ from the audible sounds obtained after low-pass filtering. Based on the interval of the extracted voice command, we obtain the corresponding segment of chest motion signals (the decomposition components obtained in Section 4.2). After that, the correlation between the chest motions and the voice commands is calculated.

To extract features that can best characterize both the voice signal and the chest motion signal, we adopt Mel-frequency cepstrum coefficient (MFCC). MFCC captures the key features of human hearing, i.e., different sensitivity to different frequencies of acoustic signals. The linear frequency spectrum is first mapped to the Mel nonlinear frequency spectrum based on auditory perception, and then converted to the cepstrum. MFCC is widely used to process voice signals for speech recognition. As the chest motion signal is derived from different frequency bins of the acoustic signal, MFCC is also suitable to extract the rich features from different frequency components of the chest motion signal. Therefore, we utilize MFCC to extract features for user authentication.

4.4.1 Calculating the Correlation between Chest Motion and Voice Command. After obtaining the starting and the ending points of voice activity from the voice signals, as shown in Fig. 16, we can locate corresponding chest motion signal. Then, we quantify the relationship between the voice signal $V(t)$ and the chest motion signal $C(t)$ by correlation

$$Corr(n) = \frac{1}{N} \sum_{m=0}^{N-1} V(m) * C(m), \quad (10)$$

where N is the number of samples. $Corr(n)$ measures the similarity between $V(t)$ and $C(t)$.

Fig. 11. Time-domain signal when there is no live user.

Fig. 12. Time-domain signal when there is a live user.

Fig. 13. Frequency-domain signal when there is no live user.

Fig. 14. Frequency-domain signal when there is a live user.

4.4.2 Extracting Features. After calculating the correlation between voice command and its corresponding chest motions, we extract features from the correlation coefficients $Corr(n)$, $n \in [1, N]$. To this end, we utilize Mel-frequency cepstral coefficients that contain 13-dimensional features, and its 13-dimensional first-order difference and 13-dimensional second-order difference as features. This results in a 39-dimensional feature vector, which is used to build the user authentication classifier. Fig 3 shows the feature vector of voice command and the feature vector of correlation results of voice command and its corresponding chest motions. In comparison, the latter contains more distinctive features that help better authentication individual users.

4.5 User Authentication: Recognizing New Users Using Meta Learning

Given the extracted features, we train a deep neural network for user authentication. Deep learning is well-known for its amazing performance in image processing and speech recognition. Nonetheless, deep learning is fueled by a huge number of training samples, which are not available in our sensing task. Inspired by the human learning process that draws on prior knowledge to deal with new tasks, meta-learning has been proposed to deal with unseen tasks based on previously learned knowledge.

For user authentication, we adopt a simple but effective meta-learning algorithm, Reptile [18], to address the problem of insufficient training samples. The rationale of Reptile is to first train a general neural network with

Fig. 15. Speaking event detection.

Fig. 16. Chest motion signals and voice signals during speaking.

Fig. 17. Meta-learning model in *ChestLive*. Fig. 18. Use scenarios of *ChestLive*.

good generalization property, and then adapt the general neural network quickly to a new task with only a few new training samples. To achieve this goal, we first train a general neural network on the data of numerous users and average the results for the update directions. More specifically, we iteratively update the set of parameters $\mathbf{P}$ as [18]

$$\mathbf{P} \leftarrow \mathbf{P} + \epsilon \frac{1}{n} \sum_{i=1}^n (\tilde{\mathbf{P}}_i - \mathbf{P}), \quad \text{where } \tilde{\mathbf{P}}_i = \text{SGD}(L_i, \mathbf{P}, k), \quad \epsilon = \frac{1 - \frac{i}{n}}{m}, \quad (11)$$

where n is the number of iterations, m is the meta step size, SGD is the function that executes stochastic gradient descent with k gradient steps on the loss function L_i starting with the parameters $\mathbf{P}$.

Note that user authentication is a binary classification problem (i.e., whether the user is legitimate or not). All individuals other than the legitimate user are illegal, but it is impossible to collect data of all illegal users for training. We show that *ChestLive* is able to distinguish *strangers* whose data is not in the training set by experiments in the next section.

Fig. 17 shows the architecture of meta learning model in *ChestLive*, which consists of 4 convolutional layers, each having a 2-D convolution kernel of 4×4 followed by a ReLU layer and a MaxPooling layer. A final fully-connected layer merges the features and uses a sigmoid function to detect whether the current input sample comes from a legitimate user or not.

5 EVALUATION

In this section, we first present the experiment setup for *ChestLive* and then conduct extensive experiments to answer the following questions:

- (Q1) Whether the user can access *ChestLive* with high accuracy?
- (Q2) Can *ChestLive* prevent the legitimate user from replay attacks?
- (Q3) How much training data is required by our system?
- (Q4) Is *ChestLive* robust under different voice commands?
- (Q5) Is *ChestLive* robust under different situations? For example, under different distances, directions, smartphone types, postures and ambient noises.

5.1 Experiment Setup

We have recruited 61 volunteers (20 females, 41 males) aged from 20 to 49. Due to COVID-19, we were not able to conduct large-scale experiments in a controlled laboratory environment. Instead, we have developed a lightweight application of *ChestLive* and distributed it to our volunteers. We asked the volunteers to perform the experiments using their own smartphones at different times according to their own routines, thus the data of a user is collected at different sessions. There is no constraint for what they should wear such that the volunteers may wear anything in their daily lives during authentication. The volunteers are instructed to send the collected data to us within one week, and most volunteers sent the collected data within three days. Therefore, the training samples and test samples come from different environments, which helps evaluate the robustness of *ChestLive*. Our users owned different types of smartphones including *Samsung A9*, *Sony Xperia XZ2*, *Huawei Mate8*. We test on the voice command of "Hey Siri", "Ok Google", "Alexa" and frequently-used words "Describe", "Guess", "Choice". As shown in Fig. 18, users are only required to hold the smartphone in front of their chests, which is a normal gesture for people to hold smartphones. Our experiments are approved by the Ethics Board of our university.

There are two sets of key parameters to be determined in our experiments. The first is the width of the time-frequency bin for calculating the channel energy. If the frequency range is too wide, the bin will contain too much noise that interfere with the channel energy. If the frequency range is too narrow, the number of bins will be large and the computational complexity will be high. According to our experiments, we finally set the width of the time-frequency bin as 5Hz . The second is the set of parameters for the meta-learning algorithm, i.e., the number of iterations n , the meta step size m , and the number of gradient steps k . We conduct extensive experiments to seek for the optimal parameter settings to ensure good performance and convergence. According to our experiments, we set $n = 2,000$, $m = 0.1$ and $k = 25$.

5.2 Performance Metrics

- True accept rate (TAR). TAR is the probability of successfully identifying a legitimate user as a legitimate user.
- True reject rate (TRR). TRR is the probability of successfully identifying an attacker as an attacker.
- False accept rate (FAR). FAR is the probability of incorrectly treating an attacker as a legitimate user.
- False reject rate (FRR). FRR is the probability of incorrectly treating a legitimate user as an attacker.
- Receiver operating characteristic(ROC) curve. ROC is the metric that presents the performance of an authentication model under various discrimination thresholds. It is plotted by true accept rate and false accept rate.
- Equal error rate (EER). EER is the point where FAR is equal to FRR.
- Area under the curve (AUC). AUC is the metrics that present the ability of an authentication model to distinguish between a legitimate user and an attacker.

Fig. 19. The authentication accuracy of different users using *ChestLive*.

Fig. 20. The receiver operating characteristic (ROC) curve of the authentication accuracy of *ChestLive*.

Fig. 21. The false reject rate of users.

5.3 Meta Learning in User Authentication

To answer the question of **Q1**, we have collected data from 61 volunteers. Each user is asked to speak "Hey, Siri" for 100 times. In total, there are $100 \times 61 = 6,100$ samples in total. We use the whole data (100×14 samples) from 14 users to train an initial model. For each of the remaining 47 users, we use 20 of their samples to fine-tune the model, which results in 47 authentication models. Finally, we use the rest 80 samples of each of the 47 users to evaluate the accuracy of the authentication models.

As shown in Fig. 19, the authentication accuracy of *ChestLive* is high with an average accuracy of 98.31%. The lowest true accept rate comes from volunteer 12. The reason is that volunteer 12 has smaller chest movement pattern than the others. All the true accept rate is over 95.61%. This indicates that *ChestLive* is capable of identifying individual users with a few shot training data, which corroborates the friendly feature of *ChestLive* in user authentication. Furthermore, during the login phase, since the user may not be able to successfully login in a single attempt, it is necessary to evaluate the number of times the user has logged in, which is correlated to the user experience. Fig. 23 shows the cumulative distribution function (CDF) of speaking times during the login phase for all the participants. We can see that the probability of volunteers login into the system by speaking the

Fig. 22. Impact of different phones for enrollment and login on the authentication accuracy of *ChestLive*.

Fig. 23. CDF of the login times for successful login in.

Fig. 24. False reject rate of *ChestLive* when three victims are impersonated by five attackers.

Fig. 25. False accept rate of the five attackers who impersonate three victims.

voice command once is over 93%. And, the probability of volunteers login into the system by speaking the voice command 2 times is over 98%. These results indicate that *ChestLive* can achieve a good user experience.

Figure 20 shows the receiver operating characteristic (ROC) curves that demonstrate the tradeoff between true accept rate and false accept rate about the overall participants and the average EER is 2.92%. *ChestLive* verifies the legitimacy of users with an AUC of 0.9927. Figure 21 shows the false reject rate curve of participants. The average FRR of *ChestLive* is 1.03%, which presents that *ChestLive* can achieve a good user authentication.

To evaluate the performance of *ChestLive* in random attacks, we randomly select 5 *strangers* whose data is not used to train the general neural network. Then, we use the data of the strangers to attack the authentication model of 9 randomly-selected victims. As shown in Table 1, the true reject rate of the strangers is 97.56% and the false accept rate of the strangers is 2.44%. In addition, the authentication accuracies of the 9 victims are 99.15%, 100%, 99.46%, 98.60%, 98.50%, 100%, 100%, 99.48%, 97.62%, respectively. This verifies that *ChestLive* is able to correctly reject unregistered strangers whose data is not used for training.

5.4 Security Analysis

In this section, we evaluate the performance of *ChestLive* in user authentication against replay attacks and impersonation attacks. In these two attack scenarios, we takes more attention to whether an attacker can

Table 1. The performance of *ChestLive* in defending against various attacks

Attack types	TRR	FAR
Random attacks	97.56%	2.44%
Simple replay attacks	100%	0
Advanced replay attacks	100%	0
Duplicate replay attacks	100%	0
Impersonation attacks	98.77%	1.23%

Fig. 26. The roc curve of the three victims.

successfully fool *ChestLive*. Therefore, we adopt true reject rate, false accept rate, ROC curve, and equal error rate as our evaluation metrics.

5.4.1 Replay Attacks. We consider three types of replay attacks: simple replay attacks, advanced replay attacks, and duplicate replay attacks. In simple replay attacks, the attacker only records the voice command signal of the victim user, i.e., the distance between the attacker's recording device and the victim's device is more than 5m away to ensure the attacker can only record the voice command. In advanced replay attacks, the attacker records both the voice command signal and the acoustic sensing signal of chest motions of the victim user, i.e., the distance between attacker's recording device and the victim's device is around 15cm to ensure the attacker can record both the voice command signal and the acoustic sensing signal of chest motions. In duplicate replay attacks, we assume that the attacker can hack into the victim's device to obtain the recorded acoustic signal for authentication. For the three types of replay attacks, we play the recordings of the attacker with a distance of 15cm to the victim's device for 20 times. Meanwhile, the victim's device will emit inaudible acoustic signals for sensing.

The results in Table 1 show that no case of replay attacks is successful. For simple replay attacks, the sensing results contain no chest motion, and the authentication request will be rejected at the liveness detection phase. For advanced replay attacks and duplicate replay attacks, although the recordings contain the chest motion information, it will be drowned by the reflected acoustic signals emitted by the device. In summary, the attacker cannot break *ChestLive* via replay attacks.

5.4.2 Impersonation Attacks. In impersonation attacks, the attacker records and plays the voice command signal of the victim user, and mimics the chest motions of the victim user at the same time. We randomly select three

Fig. 27. Impact of training set size. Fig. 28. EER of *ChestLive* with different training set sizes.

users as victims. Five volunteers act as impersonators. We allow impersonators to practice before the attacks. Each impersonator is shown the chest motion signal of a victim and their own chest motion signal derived by *ChestLive* on the smartphone screen, which guides them to improve their impersonation. Each attacker impersonates each victim for 20 times while the attacker is sitting in a chair, the distance between the smartphones and attacker's chest is around 15cm, and the smartphone is in the front of the chest.

Table 1 shows that impersonation attacks are slightly more effective than replay attacks, but the TRR is still as high as 98.77%. Fig. 26 shows the ROC curve of the three victims under the impersonation attacks of five attackers. The area under the curve of the three victims are 0.9897, 0.9357, 0.9528, respectively. And the equal error rate of the three victims are 2.81%, 8.65%, 11.48%, respectively. Fig. 24 illustrates the average FAR of the three victims. We can observe that victim 3 has high FAR. The possible reason is victim 3 has average heights and weights. Fig. 25 demonstrate and FAR of different impersonators attacking different victims. It is shown that attacker 1 is a good impersonator of victim 1, and attacker 4 is a good impersonator of victim 3. This is because both attacker 1 and victim 1 are females of 22s with similar heights and weights, and both attacker 4 and victim 3 are males of 20s with similar heights and weights. Even so, the unique chest motions of individuals are still hard to imitate by an attacker with similar physiques.

In summary, *ChestLive* is effective in defending against replay attacks and impersonation attacks.

5.5 Impact of Various Factors

In this section, we start with experiments introduction and then answer the problem of Q3-Q5 in Section 5.

5.5.1 Impact of Training Dataset. To address Q3, we randomly select 20 volunteers and the experiments setting is the same as 5.1. Collecting more samples from the legitimate user during enrollment help build a more robust classifier, but is more burdensome. We utilize meta-learning method to address the problem of limited training samples. In other word, we study the performance of *ChestLive* given different numbers of training samples of a new legitimate user. Specifically, after getting the initial model in 5.3, we utilizing 1 ~ 20 recordings from the 20 volunteers to fine-tune the initial model gradually, and then evaluate this authentication model utilizing 80 recordings from the rest data of the 20 volunteers. In other words, we train the authentication model using 1 ~ 20 recordings in the enrollment phase, and then evaluate the authentication accuracy of *ChestLive* using 80 recordings in the login phase.

Fig. 27 shows the authentication accuracy of *ChestLive* with different numbers of training samples. The authentication accuracy of *ChestLive* increases sharply at the start and then stabilizes after 10 training samples are available. Beyond 10 training samples, the authentication accuracy does not improve much. Fig. 28 shows the EER of different training set size. The results show that the authentication performance will increase with the increase of training samples. However, the EER keeps stable when the samples is larger than 10. These indicates

Fig. 29. Impact of distance.

Fig. 30. Impact of phone direction.

Fig. 31. Impact of time-variant stability of the motion.

Fig. 32. Impact of voice commands.

Fig. 33. Impact of postures.

Fig. 34. Impact of ambient noise level.

that we may collect as few as 10 training samples during the enrollment phase to build a reliable authentication model, which imposes a very light burden on users.

5.5.2 Impact of Distance. Since we leverage inaudible acoustic signals to capture chest motions, attenuation in signal magnitude is inevitable. As the distance between the smartphone and the user's chest increases, the signal attenuation intensifies, which will degrade the performance of *ChestLive*. Therefore, to address Q5, we randomly select 20 volunteers from the 61 volunteers to evaluate the performance of *ChestLive* in different distance. Each participant is required to conduct 20 times of authentications at each distance respectively while the participant is sitting in a chair and the smartphone is in the front the chest. Since a person's forearm is usually shorter than 30cm, we consider that the maximum distance between the smartphone and the chest is no more than 30cm.

As shown in Fig. 29, the average authentication accuracy have the trend of slightly increase and then slightly decrease, and the highest authentication accuracy is obtained at the distance of 15cm. When the distance is beyond 15cm, the authentication accuracy of *ChestLive* decreases due to signal attenuation. However, the lowest authentication accuracy can still reach 94.92%, which means that the working distance of our system is much longer than existing authentication method [4], [36], [12], [25], [37]. Specifically, the working rage of both VoiceLive [37] and Zhang et al. [36] are no more than 3cm. And, the authentication accuracy of LipPass [12] rapidly decreases as the distance between the lips and the smartphones exceeds 12cm.

5.5.3 Impact of Phone Angles. Different people have different habits of holding their smartphones, i.e., people prefer to hold their smartphones in the left hand or the right hand. Therefore, to address Q5, we randomly select 20 volunteers from the 61 volunteers to evaluate the performance of *ChestLive* in different angles. Each participant is required to conduct 20 times of authentications at each angle respectively while the participant is sitting in a chair, the distance between the smartphones, and the user's chest is around 15cm.

We study the impact of three different angles, i.e., front slightly left by the left hand, central front by either left or right hand and front slightly right by the right hand. Specifically, the concrete angles of both right and left are 45-degree angles in front of users. The performance of *ChestLive* under different phone angles is shown in Fig. 30. It is shown that 0-degree angles in front of users have the best performance than 45-degree angles in front of users for the two users, and front slightly right by the right hand has slightly better performance on the left. This is because that the ultrasound signal attenuation on the 0-degree angles in front of users is the least compared to the 45-degree angles, and the participants are used to play on the smartphone by the right hand, which makes the body habitually tilts to the left/right a little. However, all the performance is higher than 96.96%, which indicates that phone angles do not have a significant influence on the performance of *ChestLive* and *ChestLive* is user-friendly to both left-handed and right-handed users.

5.5.4 Impact of Smartphone Type. A user may enroll in the system on a smartphone and later log into the system on other smartphones. Microphones of different smartphones have different frequency selectivity, different hardware, and different positions. Thus, it is essential for us to study the performance of *ChestLive* on different smartphones. We randomly select 20 volunteers from the 61 volunteers to evaluate the performance of *ChestLive* on different smartphones. Each participant is required to conduct 20 times of authentications at different smartphones respectively while the participant is sitting in a chair, the distance between the smartphones and the user's chest is around 15cm, and the smartphone is in the front of the chest.

We randomly choose the smartphone types of *Xiaomi 10 (Mi10)*, *Xiaomi 6 (Mi6)*, *Nova 4E (4E)*, *Honor 30S (S30)*, *Huawei P30 (P30)*, *Huawei P40 (P40)*, *Honor V10*, *Meizu 16th*, *Mate 10*, *Mate 30*, *Redmi K30*. The authentication accuracy of *ChestLive* with different smartphones for enrollment and authentication shows in Fig. 22. It is shown that the volunteer who enrolls on 4E and logins on P30 only slightly decreases, enrolls on P40 and logins on P30 only slightly decreases, and enrolls on Honor V10 and logins on Mate10 only slightly decreases. This is because both the 4E, P30, P40, Honor V10, and Mate10 are from the same manufacturer. The volunteer who enrolls on Mi6 and logins on Meizu 16th only slightly decreases, enrolls on Mi10 and logins on S30 only slightly decreases, enrolls on Redmi K30 and logins on Mate30 only slightly decreases, and enrolls on Mate30 and logins on Meizu 16th only slightly decreases. This is because although their manufacturers are different, their manufacturing time is similar. However, the volunteer who enrolls on Mi6 and logins on Nova 4E has the largest decreases. This is because the 4E and Mi6 are not only from different manufacturer but has two years difference in production, which cause different hardware, and different positions. However, the absolute average value of authentication accuracy difference caused by the different of enrollment/login device is 2.01%, which confirms that *ChestLive* is robust and compatible with different smartphones.

5.5.5 Impact of Voice Command. Different voice assistants have different default commands, and users can customize commands for their own use. We explore the three most popular voice assistants, i.e., Siri, Alexa, and Google assistant, whose commands are "Hey, Siri", "Alexa" and "Okay, Google". We also consider three customized commands "Guess", "Describe" and "Choice". We randomly select 20 volunteers from the 61 volunteers to evaluate the performance of *ChestLive* in different voice command. And, the distance between the smartphones and user's chest is around 15cm, and the smartphone is in the front of the chest. Moreover, for each voice command, the volunteer speaks the command for 20 times for enrollment, and performs 80 times of authentications. Fig. 32 illustrates the authentication accuracy of different voice commands, which is all higher than 95.61%. This shows that *ChestLive* can suit different voice commands without performance degradation.

5.5.6 Impact of Time-variant Stability of the Motion. Users' physical conditions may affect the stability of speaking motions. Thus, we evaluate the performance of *ChestLive* when users access their phones being hungry, cold, or tired. To evaluate the performance of *ChestLive* when the user is hungry, we collect the data before the participants have lunch/dinner. To evaluate the performance of *ChestLive* when the user feels cold, we collect data of participants wearing T-shirt in a room with an air-conditioning controlled temperature of 20°C. To evaluate the performance of *ChestLive* when the user feels tired, we collect data when the user feels tired (self-reported). We randomly select 20 volunteers from the 61 volunteers, and each participant is instructed to conduct 20 times of authentications at each of the three physical conditions. The participants are asked to keep the smartphone in front of their chests at around 15cm apart.

Fig. 31 shows that physical conditions do have an impact on the authentication accuracy of *ChestLive*. Specifically, the authentication accuracy of *ChestLive* is the highest (up to 99.09%) when the user's physical condition is normal. In comparison, the average authentication accuracy of *ChestLive* when the users are hungry, cold, or tired are 97.93%, 98.06%, and 98.05%, respectively. And, the decline in authentication accuracy is within 1%, which also validates the time-variant stability of *ChestLive*.

5.5.7 Impact of Posture. Users may want to access their phones with different postures, which may affect the chest motions. Therefore, we evaluate the performance of *ChestLive* when users hold the phone at different postures, i.e., walking, standing, sitting, lying down, as shown in Fig. 18. We randomly select 20 volunteers from the 61 volunteers, and each participant is instructed to conduct 20 times of authentications at each of the postures in Fig. 18. The participants are asked to hold the smartphone in front of their chests at around 15cm apart.

Fig. 33 shows that postures do have an impact on the authentication accuracy of *ChestLive*. Specifically, the authentication accuracy of *ChestLive* has the best performance when the user is sitting on chairs. In this case, the human pose will not affect the sensing of chest motions. The authentication accuracy of *ChestLive* during sitting on chairs is up to 99.15%. Similar to sit on chairs, the authentication accuracy of *ChestLive* also is high when the user is lying on beds. In this case, the chest motion is affected by the amplitude of chest motion. The authentication accuracy of *ChestLive* during lying on beds falls to 98.28%. Similar to is lying on beds, the authentication accuracy of *ChestLive* during sitting on sofas falls to 96.69%. This is because that the chest motion is affected by the posture of sitting on sofas. Since the user shaking slightly unconsciously while standing, the authentication accuracy falls to 97.62%. The authentication accuracy is relatively low when the user is walking. In this case, even though the mobile phone and the chest may be relatively still, the movement of other body parts, e.g., arm swing, will affect the sensing of chest motions. Most previous works on smartphone authentication do not consider the case of walking. The authentication accuracy of *ChestLive* during walking is over 96.11%, which we will consider improving in our future work.

5.5.8 Impact of Ambient Noise. As we use acoustic sensing that is sensitive to noises, we evaluate the performance of *ChestLive* under different ambient noises. We play white noise at different volumes on our laptops. In general, the average road traffic noise heard at a distance of 25 m is 65 dB, the noise at library and office is 50 dB and 55 dB respectively [8]. We randomly select 20 volunteers from the 61 volunteers to evaluate the performance of *ChestLive* in different ambient noise. Each participant is required to conduct 20 times of authentications at different ambient noise respectively while the participant is sitting in a chair, the distance between the smartphones and user's chest is around 15cm, and the smartphone is in the front of the chest.

The authentication accuracy of *ChestLive* in different noise levels shows in Fig. 34. It is shown that the average authentication accuracy decreases with increased ambient noise. Moreover, the accuracy of *ChestLive* is 97.84%, 96.96%, 96.56% in the different noise levels. At 65 dB, the authentication accuracy is still as high as 96.56%, which shows that *ChestLive* is robust to different environments.

5.5.9 Discussion. There are other uncontrollable factors that may affect the stability of chest motions. The acoustic signals may be interfered by user activity, clothing and environmental noises. Normally, users will perform authentication when they are stationary or perform mild activities (e.g., walking), thus we may mitigate the impact of activities through noise removal. Our experiment results have shown that the performance will slightly degrade as environmental noise is relatively small. If a burst of environmental noise overwhelms the voice and the chest motion signals, an extra authentication attempt may be required. The clothes will move with the chest but with smaller displacement. In the future, we will further improve *ChestLive* to be sensitive to clothed chest motions.

6 CONCLUSION

In this paper, we present *ChestLive*, a multi-factor authentication system that is secure, robust, and supports adequate working distance. *ChestLive* combines the unique features of chest motions during speaking with voice commands to differentiate legitimate users and attackers. We derive channel energy of the acoustic signals that sense the chest movement during speaking, and then de-noise and aggregate the channel energy to build the primitive chest motion signal. We further remove the static and non-static interference using background model and EWT to obtain a clear chest motion signal. To tackle the problem of limited training samples, we develop a meta-learning framework, which first builds a general model with massive collected data that has good generalization property, and then fine-tunes the general model to the new user with only a few training samples. Our extensive experiments under practical settings show that *ChestLive* achieves high accuracy in user authentication, and attains low false accept rate and low false reject rate when facing various types of attacks under different conditions, confirming the effectiveness and robustness of *ChestLive*.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for constructive comments. Yanjiao's research is supported by the National Natural Science Foundation of China under Grant 61972296 and Wuhan Application Advanced Funding under Grant 2019010701011419. Zhang's work was partially supported by the RGC under Contract CERG 16204418, 16203719, R8015 and the Guang-dong Natural Science Foundation No. 2017A030312008.

REFERENCES

- [1] Heba Abdelnasser, Khaled A. Harras, and Moustafa Youssef. 2015. UbiBreathe: A Ubiquitous non-Invasive WiFi-based Breathing Estimator. In *ACM International Symposium on Mobile Ad Hoc Networking and Computing - MobiHoc '15*.
- [2] Fadel Adib, Hongzi Mao, Zachary Kabelac, Dina Katabi, and Robert C. Miller. 2015. Smart Homes that Monitor Breathing and Heart Rate. In *ACM Conference on Human Factors in Computing Systems - CHI '15*.
- [3] K Ajay Babu, Barathram Ramkumar, and M Sabarimalai Manikandan. 2019. Empirical Wavelet Transform Based Lung Sound Removal from Phonocardiogram Signal for Heart Sound Segmentation. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- [4] Si Chen, Kui Ren, Sixu Piao, Cong Wang, Qian Wang, Jian Weng, Lu Su, and Aziz Mohaisen. 2017. You can hear but you cannot steal: Defending against voice impersonation attacks on smartphones. In *International Conference on Distributed Computing Systems (ICDCS)*.
- [5] Huan Feng, Kassem Fawaz, and Kang G Shin. 2017. Continuous authentication for voice assistants. In *23rd Annual International Conference on Mobile Computing and Networking*.
- [6] Jerome Gilles. 2013. Empirical wavelet transform. *IEEE transactions on signal processing* 61, 16 (2013), 3999–4010.
- [7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*.
- [8] Chenyu Huang, Huangxun Chen, Lin Yang, and Qian Zhang. 2018. BreathLive: Liveness detection for heart sound authentication with deep breathing. *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–25.
- [9] Meike Escherich James Moar. 2020. Voice Assistant Market: Player Strategies, Monetisation & Market Size 2020-2024. <https://www.juniperresearch.com/researchstore/content-digital-media/voice-assistants-market-research-report/subscription/player-strategies-monetisation-market-size>.

- [10] Juntae Kim and Minsoo Hahn. 2018. Voice activity detection using an adaptive context attention model. *IEEE Signal Processing Letters* 25, 8 (2018), 1181–1185.
- [11] Jian Liu, Yan Wang, Yingying Chen, Jie Yang, Xu Chen, and Jerry Cheng. 2015. Tracking Vital Signs During Sleep Leveraging Off-the-shelf WiFi. In *ACM International Symposium on Mobile Ad Hoc Networking and Computing - MobiHoc '15*.
- [12] Li Lu, Jiadi Yu, Yingying Chen, Hongbo Liu, Yanmin Zhu, Yunfei Liu, and Minglu Li. 2018. Lippass: Lip reading-based user authentication on smartphones leveraging acoustic signals. In *INFOCOM Conference on Computer Communications*. IEEE.
- [13] Wenguang Mao, Jian He, and Lili Qiu. 2016. CAT: high-precision acoustic motion tracking. In *International Conference on Mobile Computing and Networking*.
- [14] Todd K Moon. 1996. The expectation-maximization algorithm. *IEEE Signal processing magazine* 13, 6 (1996), 47–60.
- [15] Jacopo P Mortola. 2004. Breathing around the clock: an overview of the circadian pattern of respiration. *European journal of applied physiology* 91, 2-3 (2004), 119–129.
- [16] Rajalakshmi Nandakumar, Shyamnath Gollakota, and Nathaniel Watson. 2015. Contactless sleep apnea detection on smartphones. In *International conference on mobile systems, applications, and services*.
- [17] Rajalakshmi Nandakumar, Vikram Iyer, Desney Tan, and Shyamnath Gollakota. 2016. Fingerio: Using active sonar for fine-grained finger tracking. In *CHI Conference on Human Factors in Computing Systems*.
- [18] Alex Nichol and John Schulman. 2018. Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999* 2, 2 (2018), 1.
- [19] Massimo Piccardi. 2004. Background subtraction techniques: a review. In *IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583)*.
- [20] Swadhin Pradhan, Wei Sun, Ghufan Baig, and Lili Qiu. 2019. Combating Replay Attacks Against Voice Assistants. *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–26.
- [21] L Rabiner and M Sambur. 1977. Voiced-unvoiced-silence detection using the Itakura LPC distance measure. In *ICASSP'77. IEEE International Conference on Acoustics, Speech, and Signal Processing*.
- [22] Wenjie Ruan, Quan Z Sheng, Lei Yang, Tao Gu, Peipei Xu, and Longfei Shangguan. 2016. AudioGest: enabling fine-grained hand gesture detection by decoding echo signal. In *ACM international joint conference on pervasive and ubiquitous computing*. 474–485.
- [23] Jiacheng Shang, Si Chen, and Jie Wu. 2018. Defending against voice spoofing: A robust software-based liveness detection system. In *International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*. IEEE.
- [24] Michael E Tipping and Christopher M Bishop. 1999. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61, 3 (1999), 611–622.
- [25] Qian Wang, Xiu Lin, Man Zhou, Yanjiao Chen, Cong Wang, Qi Li, and Xiangyang Luo. 2019. Voicepop: A pop noise based anti-spoofing system for voice authentication on smartphones. In *IEEE INFOCOM Conference on Computer Communications*.
- [26] Tianben Wang, Daqing Zhang, Yuanqing Zheng, Tao Gu, Xingshe Zhou, and Bernadette Dorizzi. 2018. C-FMCW based contactless respiration detection using acoustic signal. *Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 1–20.
- [27] Wei Wang, Alex X Liu, and Ke Sun. 2016. Device-free gesture tracking using acoustic signals. In *International Conference on Mobile Computing and Networking*.
- [28] Wikipedia. 2020. Speech. <https://en.wikipedia.org/wiki/Speech>.
- [29] Yadong Xie, Fan Li, Yue Wu, Song Yang, and Yu Wang. 2019. D 3-Guard: Acoustic-based drowsy driving detection using smartphones. In *IEEE INFOCOM Conference on Computer Communications*. 1225–1233.
- [30] Yadong Xie, Fan Li, Yue Wu, Song Yang, and Yu Wang. 2021. HearSmoking: Smoking Detection in Driving Environment via Acoustic Sensing on Smartphones. *IEEE Transactions on Mobile Computing* (2021).
- [31] Xiangyu Xu, Hang Gao, Jiadi Yu, Yingying Chen, Yanmin Zhu, Guangtao Xue, and Minglu Li. 2017. ER: Early recognition of inattentive driving leveraging audio devices on smartphones. In *IEEE INFOCOM Conference on Computer Communications*.
- [32] Xiangyu Xu, Jiadi Yu, Yingying Chen, Yanmin Zhu, Linghe Kong, and Minglu Li. 2019. BreathListener: Fine-grained Breathing Monitoring in Driving Environments Utilizing Acoustic Signals. In *International Conference on Mobile Systems, Applications, and Services*.
- [33] Zhicheng Yang, Parth H. Pathak, Yunze Zeng, Xixi Liran, and Prasant Mohapatra. 2016. Monitoring vital signs using millimeter wave. In *ACM International Symposium on Mobile Ad Hoc Networking and Computing - MobiHoc '16*.
- [34] Fusang Zhang, Daqing Zhang, Jie Xiong, Hao Wang, Kai Niu, Beihong Jin, and Yuxiang Wang. 2018. From fresnel diffraction model to fine-grained human respiration sensing with commodity wi-fi devices. *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–23.
- [35] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. 2017. Dolphinattack: Inaudible voice commands. In *ACM SIGSAC Conference on Computer and Communications Security*.
- [36] Linghan Zhang, Sheng Tan, and Jie Yang. 2017. Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication. In *ACM SIGSAC Conference on Computer and Communications Security*.
- [37] Linghan Zhang, Sheng Tan, Jie Yang, and Yingying Chen. 2016. Voicelive: A phoneme localization based liveness detection for voice authentication on smartphones. In *ACM SIGSAC Conference on Computer and Communications Security*.