3-2023

VOLERE: Leakage resilient user authentication based on personal voice challenges

Rui ZHANG

Zheng YAN

Xuerui WANG

Robert H. DENG
Singapore Management University, robertdeng@smu.edu.sg

Citation

ZHANG, Rui; YAN, Zheng; WANG, Xuerui; and DENG, Robert H.. VOLERE: Leakage resilient user authentication based on personal voice challenges. (2023). *IEEE Transactions on Dependable and Secure Computing*. 20, (2), 1002-1016.
Available at: https://ink.library.smu.edu.sg/sis_research/7820

VOLERE: Leakage Resilient User Authentication Based on Personal Voice Challenges

Rui Zhang, Zheng Yan, *Senior Member, IEEE*, Xuerui Wang, and Robert H. Deng, *Fellow, IEEE*

Abstract—Voiceprint Authentication as a Service (VAaS) offers great convenience due to ubiquity, generality, and usability. Despite its attractiveness, it suffers from user voiceprint leakage over the air or at the cloud, which intrudes user voice privacy and retards its wide adoption. The literature still lacks an effective solution on this issue. Traditional methods based on cryptography are too complex to be practically deployed while other approaches distort user voiceprints, which hinders accurate user identification. In this article, we propose a leakage resilient user authentication cloud service with privacy preservation based on random personal voice challenges, named VOLERE (VOice LEakage REsilient). It applies a novel voiceprint synthesis method based on a Log Magnitude Approximate (LMA) vocal tract model to fuse original voices of different speaking modes in order to generate a synthesized voiceprint for authentication. Thus, raw voiceprints of users can be well protected. We implement VOLERE and conduct a series of user tests. Experimental results show sound performance of VOLERE regarding authentication accuracy, efficiency, stability, leakage resilience and user acceptance. In particular, its authentication accuracy is reasonably stable regardless user nationality, gender, age, elapsed time, and environment, as well as variance of speaking modes.

Index Terms—User authentication, biometric authentication, voiceprint, privacy preservation, leakage resilience

1 Introduction

Background: As a biological feature commonly owned by human beings (except for those with voice disabilities), voice has enough inter-user differences and individual stability to be applied to authenticate a user. Voiceprint authentication does not require users to remember any passwords. It shows prominent advantages in some specific scenarios, e.g., during driving and in a surgical operation room, when user hands are occupied or when users are inconvenient to show their faces. Because of diffusive transmission of sound, voiceprint authentication has become a valuable authentication method and showed great convenience and usability.

On the other hand, it is highly expected to offer Voiceprint Authentication as a Service (VAaS) to commonly provide user authentication to a number of relying parties, e.g., intelligent voice interaction systems such as vehicle voice assistants and smart home controllers. VAaS universally offered by the cloud supports user authentication anywhere and at any time for various applications, which has obvious advantages such as flexibility, university, economy, usability and generality, thus attracts high attention. Thanks to the benefits provided by the cloud for setting up a common platform for user authentication and the consistency of user authentication applied by many relying parties.

Motivations. But voice privacy leakage is a big problem that retards wide adoption of VAaS. The voiceprints used in authentication are sensitive private information. When using VAaS, voiceprints are inevitably uploaded to the cloud, used to identify a person. But at the same time, once this biometric information is leaked, it may bring serious consequences to the users and their accounts. Unfortunately, the leakage of voiceprint information is almost inevitable in daily social networking in both physical world and cyber space. Nowadays, voice interactive digital services are widely used, which requests special concern on voiceprint leakage resilience. Although there already exist many studies about voiceprint authentication [1], [2], [3], [4], little research has been conducted to ensure voice privacy preservation and leakage resilience in VAaS.

However, existing technologies for leakage resilience of private information are not feasible to be directly applied into voiceprint authentication. Some technologies (e.g., data blur [5], [6] and differential privacy [7]) distort voiceprint features, thus impact user authentication accuracy. Cryptographic based techniques (e.g., attribute-based encryption [8], homomorphic encryption [9], [10] and access control [11], [12]) suffer from high computational complexity and inflexibility. Voiceprint-based user authentication needs to execute pattern matching. Attribute-based encryption cannot directly work for it due to ciphertext decryption before matching. If the cloud is not trusted or attacked, user information could be directly leaked at the cloud, which is not in

- *Rui Zhang and Xuerui Wang are with the State Key Laboratory on Integrated Services Networks, School of Cyber Engineering, Xidian University, Xi'an 710071, China. E-mail: cumt_zhangrui@126.com, wangxueruimay@163.com.*
- *Zheng Yan is with the State Key Laboratory on Integrated Services Networks, School of Cyber Engineering, Xidian University, Xi'an 710071, China, and also with the Department of Comnet, Aalto University, 02150 Espoo, Finland. E-mail: zyan@xidian.edu.cn.*
- *Robert H. Deng is with the School of Information Systems, Singapore Management University, Singapore 188065. E-mail: robertdeng@smu.edu.sg.*

Manuscript received 16 June 2021; accepted 24 Jan. 2022. Date of publication 1 Feb. 2022; date of current version 14 Mar. 2023.
This work was supported in part by the National Natural Science Foundation of China under Grant 62072351, in part by the Academy of Finland under Grants 308087, 335262, and 345072, in part by Shaanxi Innovation Team Project under Grant 2018TD-007, and in part by 111 Project under Grant B16037. (Corresponding author: Zheng Yan.)
Digital Object Identifier no. 10.1109/TDSC.2022.3147504

line with our goal. Using homomorphic encryption makes system design complicated and suffers from high computational complexity. In addition, ciphertext is easy to be hacked by a correlation attack. It is a kind of cryptanalysis attack that studies correlation between known sequences and unknown sequences and (or) recursion relations of unknown sequences to solve an unknown variable. Preventing the correlation attack requests biometric features to be randomly distributed. But unfortunately, correlation of biometric features occurs very frequently, which provides sufficient information to launch the correlation attack [13]. When an attacker knows the statistical characteristics of voiceprints and algorithm details, encryption becomes not so secure since biometric features are not randomly distributed. Therefore, it is urgent to design and implement a voiceprint authentication system that can offer leakage resilience even though voice is leaked in VAaS. But this is not an easy task.

Challenges. First, it is almost impossible to prevent the leakage of voiceprint information. An attacker may acquire such information stored in a server through SQL injection and other means. The voice of a user's daily speech can be easily recorded. Second, it is hard to ensure that voiceprint authentication still works securely even though some voice information has been leaked. This is an urgent and tough issue that should be solved. Third, existing privacy preservation schemes are not suitable for voiceprint authentication due to the difference of protected objects and computation demands. Even though the voiceprint is encrypted or protected in an authentication system, attackers can still get the user's voiceprints outside of the system, making cryptographic solutions totally useless. Last but not the least, variance of applications and services requests a common framework that can offer VAaS in a generic way and with privacy preservation. Any sensitive voice information should not be leaked from the framework.

Contributions. In this paper, we propose VOLERE (VOice LEakage REsilient), a leakage resilient user authentication system based on random personal voiceprint challenges. We notice that when a person speaks in different modes, his/her voices show somehow difference. For example, when the person talks in his/her native language, his/her voice is gentle and smooth. When he/she speaks in a foreign language, his/her speaking tone might become high, speaking speed might slow down, meanwhile his/her pronunciation might not be very standard. By making use of this property, we propose a novel voiceprint synthesis method based on a Log Magnitude Approximate (LMA) vocal tract model. We fuse the voice features of a user in different speaking modes into a new synthesized voiceprint to replace the user's original voiceprint for user registration and authentication, thus ideally enhance user privacy at the cloud by avoiding the usage of true voiceprints. Even though the user's original voiceprints and previously synthesized voiceprints are disclosed, VOLERE is robust against such disclosure and can offer sufficient security. Different from voiceprint obfuscation, VOLERE uses synthesized voiceprint generated from the voices of two different speaking modes of a same person to perform authentication in order to ensure authentication accuracy since the synthesized voiceprint contains more biological features of the person than a single voiceprint.

Concretely, VOLERE is a cloud-based voiceprint authentication system that offers VAaS. By applying the synthesized voiceprints for user authentication based on random personal voice challenges, it is resilient in the face of original normal voice leakage and can successfully resist replay attacks. VOLERE does not request the user to remember any passwords. By following random challenges provided by the cloud, the user says some words with different speaking modes to complete authentication. We implement VOLERE and conduct a series of user tests to verify its performance with regard to authentication accuracy, efficiency, stability, leakage resilience, and user acceptance. Dozens of volunteers from seven countries participated in our user tests. Our experimental results show that VOLERE can achieve an average authentication Equal Error Rate (EER) of 3.57%, and performs very well in all leakage resilience tests. Its average authentication time is about 20 milliseconds. It also gains general acceptance from the participants regarding perceived ease of use, usefulness, interface, playfulness and attitude of usage. In particular, its authentication accuracy is reasonably stable regardless user nationality, gender, age, elapsed time, and authentication environment (no matter quiet or noisy). It is also not impacted by the variance of speaking modes. Specifically, the contributions of this paper are summarized as below:

- We propose VOLERE, a leakage resilient VAaS system, which can provide a common user authentication service based on voice challenges over the cloud with voiceprint privacy preservation.
- We propose a novel voiceprint synthesis method based on an LMA vocal tract model to combine multiple voice features extracted from different speaking modes into a synthesized voiceprint for user authentication.
- We implement VOLERE and conduct a series of user tests to evaluate its performance in terms of authentication accuracy, efficiency, stability, leakage resilience, and user acceptance. The results show VOLERE's excellence and advantages.
- VOLERE achieves user voice leakage resilience to a great extent. Even though original voiceprints in any speaking modes and any synthesized voiceprints of a user are gained by an attacker, the attacker cannot impersonate the user. It can also resist replay attacks and adversarial attacks under our security model.

The rest of this paper is organized as follows. Section 2 briefly reviews technical background and related work. Section 3 describes the system model and security model of VOLERE, followed by preliminary techniques. In Section 4, we present the design of VOLERE with security analysis. Then, we describe VOLERE implementation, user study and experimental results in Section 5. Finally, a conclusion is drawn in the last section.

2 Background and Related Work

2.1 Speaker Recognition

Speaker recognition (SR), also known as voiceprint recognition (VPR), is a biometric recognition technology based on a speaker's personal information in speech signals. Traditional

VPR mostly uses Mel Frequency Cepstrum Coefficient (MFCC) features and a Gaussian Mixture Model (GMM) model [14], and has achieved excellent results. Later, algorithms based on I-Vector and Deep Neural Network (DNN) were proposed.

On the basis of GMM model, Reynolds *et al.* [14] proposed Gaussian Mixture Model-Universal Background Model (GMM-UBM) to overcome the problem caused by short training speech and small corpus. The GMM-UBM model only needs to be trained once and can be used repeatedly. Inspired by the theory of joint factor analysis, Dehak *et al.* [15] proposed to extract a more compact vector from GMM mean hyper vector, called I-Vector. Up to now, I-vector is still one of the best modeling frameworks in text-independent voiceprint recognition. Researchers' subsequent improvements are usually based on the optimization of I-Vector, including Linear Discriminant Analysis (LDA) [15] and Probabilistic Linear Discriminant Analysis (PLDA) [16] and even Metric Learning [17].

Variani *et al.* [18] studied the application of DNN in a small footprint text-dependent speaker recognition task. A trained DNN is used to extract speech features, i.e., D-Vectors, which are used to compare with an input speaker model for verification. Experimental results showed that this method achieves good performance compared with the commonly used I-Vectors. Snyder *et al.* [19] used data augmentation to improve the performance of DNN-based speaker recognition. In order to distinguish a speaker, an embedding DNN architecture is trained from a variable length corpus. The trained DNN contains several frame layers, a pooling layer and two segment layers. Speaker embeddings can be extracted from the segment layer with a fixed-dimensional matrix, which are called X-Vectors. The training speed of X-Vector is very fast and its recognition accuracy is good.

2.2 Voiceprint Authentication

Jayamaha *et al.* [1] proposed a voice authentication system based on Hidden Markov Model (HMM) to extract certain features from voice waveforms for user authentication. Experimental test showed that the accuracy of this system is not high, only about 86%. But this system can resist replay attack (spoofing attack) to a certain extent. Gałka *et al.* [2] presented an access control solution based on voice, which uses an HMM-GMM method and achieves an EER of 3.4%. Yan and Zhao [3] proposed a voiceprint authentication system based on random auto-challenge to resist forgery attack. It achieves an average recognition rate of 80.6%. Wang *et al.* [4] designed an anti-spoofing voiceprint authentication system called VoicePop based on pop noise detection. The system can defend against spoofing attacks and achieves over 93.5% detection accuracy with EER around 5.4%. However, leakage resilience and privacy issue were not discussed in the above works. Although some of potential risks in voiceprint authentication have attracted attention with some solutions, voice leakage is still an open issue, which requests a feasible solution.

2.3 Leakage Resilience Methods

Some solutions have been proposed to resist private information leakage in biometric authentication. Wang and Hatzinakos proposed a privacy preservation method based on random transformation [20]. Fan and Lin proposed a hybrid privacy preservation method, which adopts multi-factor authentication and embeds a smart card into a biometric authentication system [21], so that biometric data does not need to be shared with a remote server. Pillai *et al.* [22] proposed an iris recognition framework based on random projections and sparse representations. Its algorithm can deal with common distortion in iris image collection. Because random projections and permutations were used to make the proposed algorithm irreversible, attackers cannot obtain user information through simple reverse engineering.

Some researchers have studied leakage resilience in fingerprint authentication. Li and Kot [23] proposed a fingerprint authentication system, which uses data hiding and data embedding technologies to embed private user identity into a fingerprint template. Due to data hiding, attackers cannot obtain the identity and original fingerprint from any stolen templates. Li and Kot [24] also proposed a fingerprint authentication system that collects two fingerprint images. The directional features of one fingerprint are combined with the minutiae of another to form a composite fingerprint template. Thus, even though the template is stolen, a single true fingerprint cannot be exposed. The system achieves excellent accuracy with EER 0.4% and preserves privacy at a high level. However, whether this system can resist replay or spoofing attacks requests further investigation.

Obviously, the above reviewed methods cannot be directly applied into voiceprint authentication due to either the difference of applied biometric features or the difference of system models.

2.4 Countermeasures of Voiceprint Authentication Attacks

As mentioned above, voiceprint authentication is suffering from some potential threats. Replay attack is probably the simplest and most common threat. Some schemes have been proposed to resist replay attacks. Zhang *et al.* proposed VoiceLive [25], a practical liveness detection system based on smartphones. It captures time-difference-of-arrival (TDoA) change of phoneme sounds to the two microphones of the phone to detect replay attacks since such a change does not exist under replay attacks. VoiceGesture [26] uses smartphones as a Doppler Radar to detect Doppler shifts caused by user's articulatory gestures. WiVo [27] captures mouth motions by characterizing the correlation between wireless signal dynamics and a user's voice syllables, thus achieves the goal of liveness detection. ItCaField [28] was proposed to detect loudspeaker-based spoofing attack by constructing "fieldprint", a physical field of acoustic energy created when sound propagates over the air. The fieldprint is similar to voiceprint, but distinctive between human users and loudspeakers. Muhammad *et al.* proposed Void [29] for quick liveness detection, which makes use of the differences in spectral power between live-human voices and voices replayed by speakers. Both of above works achieve high accuracy in liveness detection, over 99%. Unfortunately, neither of them considers privacy preservation and voice leakage resilience in VAaS.

Recently, researchers studied adversarial attack on voiceprint authentication. This attack takes the weakness of deep learning that is not resistant to subtle perturbations. Vmask

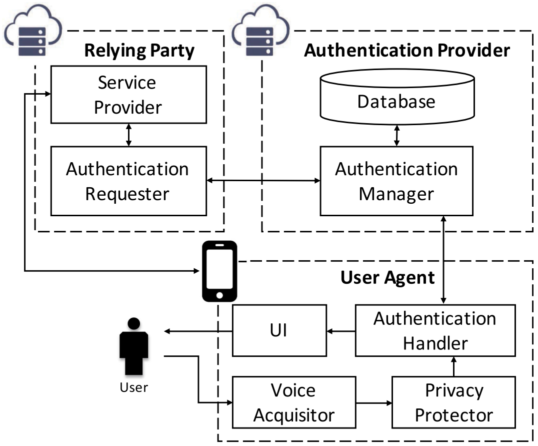


Fig. 1. VOLERE system model.

[30] can mislead classification to recognize an audio record as a target speaker by generating adversarial samples. Chen *et al.* [31] formulated adversarial sample generation as an optimization problem and proposed a novel algorithm to estimate a threshold for speaker recognition and solve the optimization problem.

3 Problem Statement and Preliminary

This section specifies the system model and security model of VOLERE. In addition, we introduce the preliminary techniques of VOLERE.

3.1 System Model

VOLERE consists of three types of entities: User Agent (UA), Relying Party (RP), and Authentication Provider (AP), as shown in Fig. 1. UA is a user-owned personal device, which includes a User Interface (UI) that interacts with the user for voiceprint authentication and RP service access, a voice acquisitor that collects user voices, a privacy protector to generate synthesized voiceprints, and an authentication handler to communicate with AP. RP provides a remote service, e.g., an Internet service. Multiple RPs could exist to offer various online services based on user voiceprint authentication. RP contains an authentication requester, which requests AP for user registration and authentication. AP provides a cloud-based authentication service for multiple RPs and their users. It contains an authentication manager that performs user authentication and a database that stores user authentication related information.

When a user connects to RP and requests its service through UA, RP requests AP to perform user authentication. AP receives the request from RP and then establishes a connection with UA. During authentication, AP sends some challenge codes, which are some words or sentences that are randomly generated to prevent attackers from using previous audio recordings to impersonate a user. When UA receives the challenge codes, it instructs the user (via UI or audio prompt) to repeat the codes with his/her own voices in a required specific speaking mode. The voice data acquisitor (e.g., a microphone) in UA collects the user's voice signals, and the privacy protector processes them by using the voiceprint synthesis method to generate a synthesized voiceprint. The authentication handler in UA then sends the synthesized voiceprint to AP. AP recognizes the speaking contents as challenged and compares the received synthesized voiceprint with a target template stored in its database and sends an authentication result to RP. Then, RP is able to decide whether to provide its service to the user.

3.2 Security Model and Threat Model

VOLERE is a text-independent voiceprint user authentication system to recognize a speaker. During authentication, challenge codes are generated randomly to prevent replay attacks. In addition, VOLERE transforms original voiceprint identification information into anonymous one by using a voice synthesis method to combine voice features of a user in different speaking modes to prevent privacy disclosure.

3.2.1 Security Model

We assume that the data in VOLERE is protected in the process of processing, transmission and storage by applying existing techniques. For example, the communication channel between RP and AP is secure with mutual authentication. While the communications between UA and RP and between UA and AP are protected with Transport Layer Security (TLS). RP provides its service to a user once he/she can pass the authentication offered by AP. Although AP cannot be fully trusted with regard to voiceprint privacy preservation, it should execute VOLERE procedures as designed. Herein, we assume that the integrity of AP database is ensured. Since UA is its user's personal device, we assume that the user trusts its own UA for voice signal collection and synthesized voiceprint generation. In addition, due to profit conflict and business difference, AP and RP will not collude.

3.2.2 Threat Model

We focus on preventing voiceprint leakage at AP and over-the-air. We assume that an adversary knows VOLERE's detailed algorithms and system model. It is quite possible for the adversary to retrieve the identity information of a user stored in AP, including voiceprint information. The adversary may use SQL injection, speech recording or other means to get voiceprint information. He/She can use this information to impersonate a user and gain access to the RP service. In addition, the adversary can also try to use this information to attack other RP services. These services may universally use the proposed generic voiceprint authentication service offered by AP. Herein, we assume that the adversary cannot collect the original recordings of all users' voices in all speaking modes since this is a tough task with a high cost. Since in daily life, people speak in a normal speaking mode. If they are asked to change the mode without a clear purpose, this will obviously cause attention.

We specially pay attention to the leakage resilience capability of VOLERE. According to the above assumptions, there are some possible ways of privacy disclosure that VOLERE may encounter:

- When the adversary obtains any voiceprint information of a single speaking mode, he/she may try to use this information to disguise a user by performing a replay attack.

- When the adversary obtains past synthesized voiceprint information from VOLERE, he/she may try to use this information to impersonate a user by performing a replay attack.
- When the adversary obtains voiceprint information from the database of AP, he/she may try to use this information to disguise a user by performing a replay attack.
- When the adversary gets synthesized voiceprint information from VOLERE, he/she may turn the target to other systems that apply authentication based on the voiceprint of one speaking mode of users.

3.3 Preliminary

Based on the observation of differences in different speaking modes, we designed VOLERE. We notice that speech characteristics (e.g., speed and pitch) show somehow difference when a person speak in different modes, e.g., speaking in a mother tone or a foreign language, or singing. In order to confirm this observation, we carried out a small user test with two participants. One of them provided 10 audio records in English and 10 audio records in Chinese. The other provided 10 audio records in routine speaking mode and 10 audio records in singing. Then, we tried to recognize a person by matching the audios recorded in different speaking modes. The success rate of voiceprint matching is less than 60% although the records came from the same person. This implies that the features of different speaking modes show somehow discrepancy even through they are from the same person. VOLERE attempts to make use of this property to preserve voiceprint privacy and ensure voiceprint authentication safety even some of voiceprint data could been leaked. On one hand, VOLERE uses synthesized voiceprint originated from the voices of two different speaking modes of a same person to perform authentication. This approach can ensure authentication accuracy since the synthesized voiceprint contains more biological features of the person than a single voiceprint. It could be a good way to overcome the negative impact caused by temporal voice change. More importantly, voiceprint synthesis can hide original voiceprint information, thus greatly help in preserving voice privacy. In what follows, we briefly introduce the fundamental techniques used in VOLERE.

3.3.1 Feature Extraction

Feature extraction is the premise of speaker recognition and verification. The commonly extracted features for speech recognition include Mel Frequency Cepstrum Coefficient (MFCC) [32], Linear Prediction Coefficient (LPC), Linear Prediction Cepstrum Coefficient (LPCC) [33], Line Spectrum Frequency (LSF), etc. Herein, we choose LPCC and MFCC because they can reflect the features of a person's vocal track and auditory characteristics of human ears, respectively. These two features meet our design demand of voiceprint synthesis. On the other hand, extracting these two features is cost efficient. The accuracy of voiceprint recognition based on them is also high. Based on our experimental tests, using other features cannot achieve sound accuracy as using LPCC and MFCC.

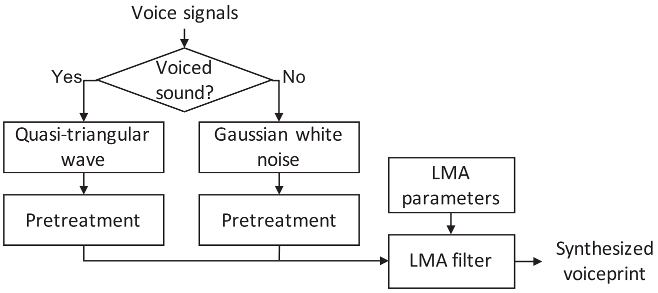


Fig. 2. A procedure of voiceprint synthesis.

3.3.2 LMA Vocal Tract Model

We use voiceprint synthesis to protect voice privacy at AP and enhance leakage resilience. In our design, the two speaking modes for user authentication are specified by the system during registration. Example options include normal voice, imitation of the elderly, imitation of a child and imitation of the opposite gender of a user. We collect voice signals from a user in different speaking modes, extract corresponding features from them, respectively, and fuse the extracted features to form a new synthesized voiceprint. Then, these two sets of features can cover up each other to achieve the goal of protecting the original voiceprint of the user. A voice synthesis algorithm based on an LMA vocal tract model can generate high quality voice with a small set of parameters [34]. VOLERE makes use of this model to generate synthesized voiceprints, as shown in Fig. 2 and described below.

1) Fundamental tone generation. For a speech, there is a big difference between its voiced part and its unvoiced part. Normally, vowels are typical voiced sounds, and their waveforms usually show obvious periodicity. In contrast, vocal cords do not vibrate and their waveforms do not have periodicity when we make an unvoiced tone. So, we can divide pronunciation into voiced part and unvoiced part, and use an artificial synthesis method to simulate a fundamental tone. In the LMA vocal tract model, the voiced part is simulated by quasi-triangular waves, while the unvoiced part is usually simulated by Gaussian white noises.
2) Voiceprint synthesis. After the voiced and unvoiced parts of the fundamental tone are generated, they are fed into an LMA filter to adjust the vocal tract parameters so as to achieve the purpose of changing the tone color of the voice.
3) The frequency response of the LMA filter $H_L(n, c)$ is: $H_L(n, c) = \exp(\sum_{l=0}^{L} c_l n^{-l})$, where c is the cepstral coefficient of the voice signals, which could be LPCC-related or MFCC-related. The former describes the vocal characteristics of the speaker, and the latter describes the auditory characteristics of human ears. Parameter c_m is determined by the characteristics of a speech. If it is a voiced sound, use the LPCC-related coefficients (refer to Appendix B.1 on LPCC feature extraction), which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TDSC.2022.3147504. Otherwise, use the

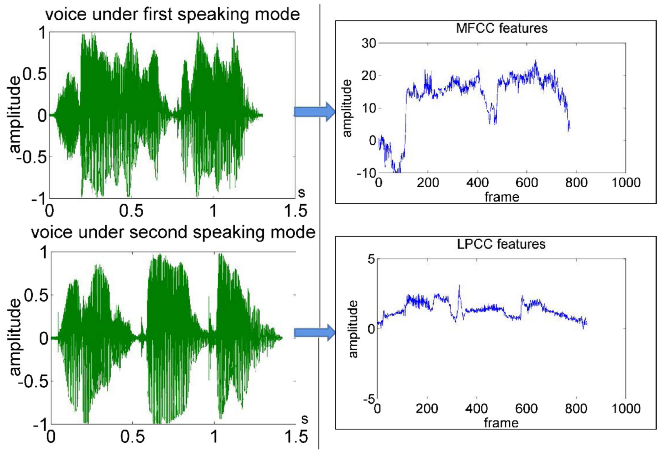


Fig. 3. Feature extraction.

MFCC-related coefficients (refer to Appendix B.2 on MFCC feature extraction), available in the online supplemental material.

3.3.3 Speaker Recognition Based on I-Vector

Speaker recognition confirms whether a speech is said by a specific person. I-Vector is a method based on HMM [35]. Appendix C, available in the online supplemental material, introduces the details of speaker recognition based on I-Vector. It considers both speaker information and channel information, and does not request a big size of training data. It has such advantages as high accuracy and low cost. VOLERE adopts I-Vector for speaker recognition. During user authentication, AP counts the cosine distance between the I-Vectors of a test speech and the I-Vectors of a target template as a final score for making a decision. In Section 5, we also test the authentication accuracy of VOLERE based on X-Vector, which can achieve even higher accuracy than I-Vector.

4 VOLERE Design

VOLERE is a text-independent voiceprint user authentication system. In each authentication process, challenge codes are generated randomly to prevent replay attacks. In addition, VOLERE achieves privacy protection by transforming raw voiceprint identity information into anonymous one (i.e., synthesized voiceprint). For this purpose, we propose a voice synthesis method to prevent privacy disclosure. In this section, we first describe the voice synthesis method. Then, we describe the design of VOLERE, followed by security analysis.

4.1 Voice Synthesis Method

The proposed voice synthesis method is based on the LMA vocal tract model [34]. It makes use of basic syllables and a small set of parameters to synthesize speech. We collect user voices in different speaking modes, extract corresponding features from them respectively, and fuse the extracted features to form a newly synthesized voiceprint. With this approach, on one hand, the user's original voiceprint does not need to be uploaded to AP. On the other hand, the synthesized voiceprint uploaded to AP is pre-processed, so that the voice features in two speaking modes cover up each other to hide original voiceprints. Basically, it is very difficult to extract the original voiceprints from the synthesized one. Thus, we can achieve the purpose of voiceprint privacy preservation at AP and leakage resilience. The voice synthesis method includes two parts: *Feature Extraction* and *Voiceprint Synthesis*.

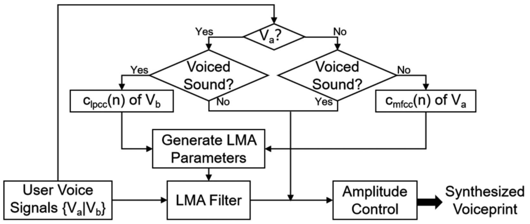


Fig. 4. Voiceprint synthesis.

1) *Feature Extraction.* The voice features of LPCC and MFCC required for generating a synthesized voiceprint are extracted (refer to Appendix B), available in the online supplemental material. In both registration and authentication, we collect voices from a user in two different speaking modes SM_a and SM_b, denoting as V_a and V_b. We extract the LPCC features $c_{lpcc}(n)$ of user voices in V_b and the MFCC features $c_{mfcc}(n)$ in V_a, shown in Fig. 3.
2) *Voiceprint Synthesis.* As shown in Fig. 4, a voiceprint synthesis algorithm based on the LMA filter is used to fuse multiple extracted voice features into a synthesized voiceprint. The input of the LMA filter includes two parts. One is the LMA parameters. The other is the fundamental tone generated by concatenating the two speech voices of a user in two different speaking modes. Herein, LMA parameters are generated by the LPCC features or MFCC features extracted in Step 1. We determine whether a voice frame is a voiced part or not by calculating the energy of corresponding fundamental tones. As mentioned above, LPCC and MFCC describe voices from different perspectives. LPCC assumes that there is a linear prediction structure in voice signals, which can accurately describe the periodic characteristics of voiced sounds. On the other hand, the features of adjacent frames in MFCC are almost independent. Since MFCC describes the frequency characteristics of each frame, it can well describe the unvoiced sounds. If an inputting frame of the fundamental tone of V_a is a voiced part, the previously extracted LPCC features of V_b are fused into this frame in the LMA filter as $H_L(x(n), c_{lpcc}(n))$, so that this frame has the characteristics of voiced sound appeared in V_b. If an inputting frame of the fundamental tone of V_b is a invoiced part, the MFCC features of V_a are merged into this frame in the LMA filter as $H_L(x(n), c_{mfcc}(n))$ to make the frame has the characteristics of voiceless sound appeared in V_a. If an inputting frame of V_a is an invoiced part or if an inputting frame of V_b is a voiced part, we keep it as original. With this way, we can embed the feature of one speaking mode into another in order to get the synthesized voiceprint, which contains features owned by the two speaking modes but hides raw voiceprints of both

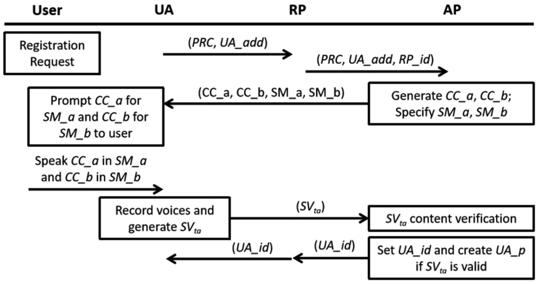

User
UA
RP
AP
Registration Request
(PRC, UA_add)
(PRC, UA_add, RP_id)
Prompt CC_a for SM_a and CC_b for SM_b to user
(CC_a, CC_b, SM_a, SM_b)
Generate CC_a, CC_b; Specify SM_a, SM_b
Speak CC_a in SM_a and CC_b in SM_b
Record voices and generate SV_ta
(SV_ta)
SV_ta content verification
(UA_id)
(UA_id)
Set UA_id and create UA_p if SV_ta is valid

Fig. 5. User registration procedure.

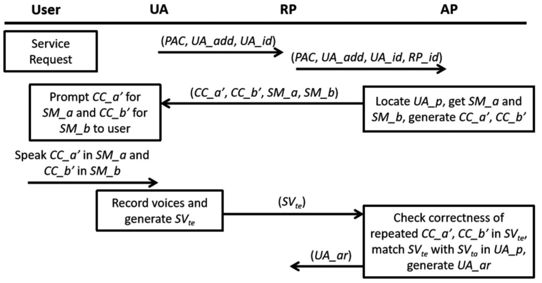

User
UA
RP
AP
Service Request
(PAC, UA_add, UA_id)
(PAC, UA_add, UA_id, RP_id)
Prompt CC_a' for SM_a and CC_b' for SM_b to user
(CC_a', CC_b', SM_a, SM_b)
Locate UA_p, get SM_a and SM_b, generate CC_a', CC_b'
Speak CC_a' in SM_a and CC_b' in SM_b
Record voices and generate SV_te
(SV_te)
Check correctness of repeated CC_a', CC_b' in SV_te, match SV_te with SV_ta in UA_p, generate UA_ar
(UA_ar)

Fig. 6. User authentication procedure.

modes. Herein, we use MFCC and LPCC as two typical features to illustrate how to generate synthesized voiceprints. Note that other features or combined features (u, v) can be applied to generate synthesized voiceprints. In this case, $c_{lpcc}(n)$ and $c_{mfcc}(n)$ will be replaced by $c_u(n)$ and $c_v(n)$ in Fig. 4 to represent a general model of voiceprint synthesis.

4.2 Registration Procedure

Fig. 5 shows the procedure of user registration in VOLERE.

1) *Registration request.* If a user wants to access RP, the user's UA raises a Personal Registration Command (PRC) and sends PRC and UA address (UA_add) to RP;
2) *Registration forward.* When RP receives the PRC, it packages it with its own ID (RP_id) and send them to AP;
3) *Registration challenge.* When AP receives the registration request from RP with regard to UA, it randomly generates two sets of voice challenge codes (CC_a, CC_b) from dataset A (e.g., two short sentences: "who are you?" and "how are you doing?") and sends them to UA with two specified speaking modes (SM_a, SM_b);
4) *Target synthesized voiceprint generation.* UA informs the user (SM_a, SM_b). The user speaks the challenge codes (CC_a, CC_b) accordingly by following the prompts provided by UA. UA records user voices and then extracts different features from the two audio records, respectively. For example, UA specifies that speaking mode SM_a is to imitate the voice of the user's opposite gender, and mode SM_b is the user's normal speaking mode. UA extracts $LPCC$ features from the second record, and $MFCC$ features from the first one. Then, UA fuses these two features into the user's spoken voices of CC_a concatenating CC_b and generates a synthesized voiceprint (denoted SV_{ta}) by using the voice synthesis method before sending it to AP.
5) *User profile creation.* When AP receives the synthesized voiceprint, it verifies if the user correctly repeats CC_a and CC_b as challenged through speech recognition [35] and checks if it is a new user of RP (i.e., whether UA_id can be provided). If the above verification is positive, AP creates a user profile (UA_p) that contains a newly created unique user ID (UA_id) linked to RP_id, SM_a, SM_b, and SV_{ta} (a target template). Then AP sends UA_id to RP and UA. Note that, if AP finds that the user has registered already or the registration suffers from some problems, it sets UA_id as null.
6) *Registration notification.* If UA_id is not null, RP treats registration successful and keeps UA_id. Otherwise, the registration fails.

4.3 Authentication Procedure

Fig. 6 illustrates the procedure of user authentication in VOLERE.

1) *Authentication request.* If a user wants to access a RP service, he/she initiates UA to send a Personal Authentication Command (PAC) with UA_add and UA_id to RP.
2) *Authentication forward.* RP forwards PAC to AP together with UA_add, UA_id and RP_id.
3) *Authentication challenge.* When AP receives the above authentication request, it locates the corresponding user profile (UA_p) to get the two speaking modes that the user used during registration (SM_a and SM_b). AP randomly generates two sets of challenge codes (CC_a', CC_b') from another dataset B different from A and sends them to UA together with SM_a and SM_b.
4) *Test synthesized voiceprint generation.* The user speaks the challenge codes (CC_a', CC_b') accordingly, which are recorded by UA. UA then extracts different features from the two audio records under SM_a and SM_b, respectively. Concretely, UA extracts LPCC features from the second record and MFCC features from the first one. Then, UA fuses these features into the user's spoken voices of CC_a' and CC_b', generates a test synthesized voiceprint (denoted SV_{te}) and sends SV_{te} to AP.
5) *User verification.* AP checks if the user correctly repeats the contents of CC_a' and CC_b' as challenged for resisting replay attack and verifies whether SV_{te} can match with SV_{ta} by performing speaker recognition based on I-Vector. Then an authentication result (UA_ar) is generated and sent to RP.
6) *Service provision.* If UA_ar is positive, RP allows the user to access its service. Otherwise, RP rejects service access.

User re-registration (i.e., SV_{ta} update) can be performed by combining the authentication procedure and the registration procedure together. After the user passes authentication, AP allows him/her to register again and replaces the old SV_{ta} with a new one. User revocation at a RP can be performed by first verifying user Legitimacy before deleting the link of the user profile to the corresponding RP at AP. If the underlying RP is the last one linked to the user profile, the user profile can be totally removed at AP. In VOLERE registration, AP specifies speaking modes to the user with random selection. It is also possible for the user to indicate his/her preferred speaking modes at UA and inform them to AP. Thus, VOLERE can offer sound usability and flexibility with regard to speaking mode selection.

4.4 Security Analysis

Proposition 1. *If an attacker only knows the voiceprint of a user in one mode, he/she cannot impersonate the user.*

Proof. Suppose the attacker acquires the original voiceprint V_a (or V_b). According to the previous description of the synthesized voiceprint $SV = Synthesized(V_a, V_b)$, $mfcc(SV) = mfcc(V_a|H_L(V_b, mfcc(V_a))) = mfcc(V_a|exp(\sum_{l=0}^{L} mfcc(V_a)_l {V_b}^{-l}))$. Obviously, $mfcc(V_a) \neq mfcc(SV)$ and $mfcc(V_b) \neq mfcc(SV)$. Even if we get V_a or V_b, we cannot disguise SV to replace legitimate users.

There are some ways for an attacker to collect more than one mode of user voices. One is to collect data uploaded by users in social networks, but the probability of these data meeting the requirements of user authentication is relatively low. The other is to induce user voices directly, but the purpose is too clear, which is easy to arouse users' vigilance. These two methods are not superior in terms of efficiency and cost.

As far as we know, MFCC features can be reversed to generate high fidelity audio. Based on our survey, light weighted methods, such as Librosa toolkit [36], can only generate audio with the same text content, but its efficiency is not ideal. It takes about 15 minutes to generate a 30-second .wav audio file. HTS toolkit [37] and other methods using machine learning to build user model usually need to collect a large amount of data and spend long time in model training and learning. So, it is hard to raise an attack by using these toolkits. □

Proposition 2. *VOLERE does not subject to replay attack when SV_{ta} is not compromised.*

Proof. AP randomly generates challenge codes (CC_a' and CC_b') in every time of authentication and asks the user to speak the same words as specified in CC_a' and CC_b'. AP verifies if the repeated codes in SV_{te} are the same as the challenged through speech recognition. An attacker cannot use old SV_{te} to replay since the current challenge codes are different from the old ones even though the features extracted from the old SV_{te} can match with SV_{ta}. □

Proposition 3. *If SV_{ta} is compromised, an attacker cannot replay SV_{ta} or use a slight variation of SV_{ta} to impersonate a user.*

Proof. During authentication, VOLERE verifies if the words repeated by the user (or attacker) is the same as the underlying challenges that are randomly generated by AP. The challenge codes used in authentication are different from those in registration since they are selected from different datasets. Even if SV_{ta} is compromised, the words recognized from it should be different from the ones recognized from SV_{te}. Therefore, replay attack will not happen if SV_{ta} is compromised. □

Proposition 4. *Reverse engineering of SV_{ta} or SV_{te} to recover the LPCC and MFCC parameters of a user cannot allow an attacker to impersonate the user. In other words, it is difficult to extract original voiceprint from a synthesized one.*

Proof. In VOLERE, LPCC and MFCC of different speaking modes are integrated to form a new voiceprint. For the purpose of privacy preservation, LPCC and MFCC parameters in the original audio are protected by UA that is trusted. The possibility of the attacker to obtain the original feature parameters from the fused voiceprint is discussed below. Suppose that the attacker obtains the synthesized voiceprint SV_{te} or SV_{ta}, because their generation process is the same, so only SV_{te} is taken as an example for discussion herein. At this point, $SV_{te} = Synthesized(V_a, V_b)$, According to the previous description, $mfcc(SV_{te}) = mfcc(V_a|H_L(V_b, mfcc(V_a))) = mfcc(V_a|exp(\sum_{l=0}^{L} mfcc(V_a)_l {V_b}^{-l}))$. Obviously, $SV_{te} \neq V_a$ and $SV_{te} \neq V_b$. The attacker cannot extract the features of the original audios through the synthetic voiceprint.□

Proposition 5. *An attacker cannot implement an over-the-air adversarial attack in VOLERE.*

Proof. Generally, when an attacker conducts an adversarial attack against a target user's voiceprint $x_{ta}(n)$, the attacker first generates an adversarial sample δx. Then, δx will be added into an audio recording $x(n)$, which causes the system to recognize $x'(n) = x(n) + \delta x$ as $x_{ta}(n)$. In VOLERE, even if the attacker collects the voiceprint information SV_ta of the target user in advance, it is still hard to conduct an over-the-air adversarial attack.

First, the attacker tries to generate an adversarial sample δx and adds δx into both V_a and V_b. According to the above, the adversarial attack should add δx into any audio recording SV to get SV', which causes the system to recognize SV' as SV_{ta}. However, SV is not an over-the-air played audio. According to our system design, $SV = Synthesize(V\_a, V\_b)$. If the attacker adds δx into V_a and V_b through an over-the-air method, then he/she gets $SV'' = Synthesize(V\_a + \delta x, V\_b + \delta x)$. Obviously, $SV'' \neq SV'$, which cannot make VOLERE identify it as the target user.

Second, it is hard to generate adversarial δx_a and δx_b. Herein δx_a and δx_b can be added into V_a and V_b separately to get $SV''' = Synthesize(V\_a + \delta y, V\_b + \delta y)$, which is subject to $SV''' = SV_{ta}$. According to the assumption in Section 3.2.2, it is difficult for an attacker to collect the voices of a user in all speaking modes. Herein, we assume that the attacker can obtain SV_{ta} and the recording of a certain speaking mode, for example V_a. Since V_b is still unknown for the above equation, a definite solution cannot be obtained. That is to say, the attacker cannot solve δx_a and δx_b from the above equation. □

In short, VOLERE ensures its security through random personal voice challenges, preserves original voiceprint privacy by applying the synthesized voiceprint, and achieves user authentication by matching the voice features extracted from SV_{te} with those in SV_{ta}. Even though original voiceprints in any speaking modes and any synthesized voiceprints of a user can be gained by an attacker, the attacker cannot intrude VOLERE. Thus, VOLERE achieves user voice leakage resilience.

We are also aware of the recent spring up of adversarial attacks. Most of existing researches on adversarial attacks in voiceprint recognition only consider white box test [38], [39], where an adversary has full knowledge of a target speaker's recognition model as well as its parameters. Notably, user voiceprint data are usually protected during network transmission and at storage in many practical applications. However, black box attack is rarely mentioned in existing works. Obviously, there is a big difference between the black box attack and the white box attack. The impact of the difference on attack success rate is still unknown.

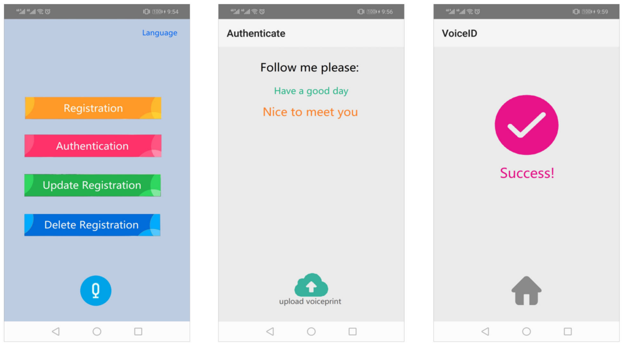
Researchers have studied how to defend against adversarial attacks. Commonly used methods include adversarial training, gradient vanishing or random transformation, but they all have different degrees of defects, such as difficult to defend and easy to break [40]. For further improving the security of VOLERE, we can introduce a secret shared between AU and AP to obscure SV at AU before sending it to AP and remove the obscuration at AP before speaker recognition. Since the attacker does not know the secret, it cannot generate the same SV, assuming that it cannot get a target user's UA or access this UA. To overcome distrust of AP, a trusted execution environment (e.g., Intel Software Guard Extensions (SGX)) can be employed to execute the above process. In addition, we can also set the threshold of user challenge response time as a small value in order to resist black box attacks. Normally, the adversarial learning attack takes longer time than VOLERE authentication time, refer to the next section.

5 User Studies and Experimental Results

We implemented VOLERE and tested its performance in terms of authentication accuracy, efficiency, stability, leakage resilience, and user acceptance. We adopted the following evaluation metrics: *False Negative Rate* (FNR); *False Positive Rate* (FPR); *Detection Error Tradeoff* (DET) curve that explains the relationship between FNR and FPR under various detection thresholds; *Equal Error Rate* (EER) that is defined as the rate when FNR equals FPR; *Accuracy* is the probability that the system correctly accepts an eligible user or rejects a non-eligible one; *Efficiency* that is reflected by the time consumed in an authentication process; *Time Stability* indicates if VOLERE performs stable with time flying and is not impacted much by a user's voice change within an expected long period of time; *Leakage resilience* indicates the probability that the system successfully resists attacks when some user private information is leaked; *User Acceptance* that concerns perceived ease of use, usefulness, playfulness, UI and attitude of accepting a system, which represent a user's willingness to accept the system.

5.1 VOLERE Implementation

We implemented VOLERE with Java language. UA was prototyped in Huawei Honor 6X, an android phone, which contains a number of functions for user registration, authentication, re-registration and revocation. We also used this model phone to do all user tests. RP and AP were implemented as two services offered by two servers running Windows 10 Operating Systems. UA connects RP and AP through WiFi. MYSQL was used to manage the databases in RP and AP. We used a set of 50 daily expressions as challenge codes. All of them are no more than 10 words. The expressions are divided into two subsets. The first subset contains 10 expressions for registration. The second one contains the other 40 for authentication. In addition, the speaking modes were randomly selected from four modes: normal voice, imitation of the elderly, imitation of a child and imitation of opposite gender. As shown in Fig. 7, the UA was developed as a mobile app. It can receive user commands, inform RP, receive the challenge codes from AP, instruct its user to speak, collect user voices, process them and send synthesized voiceprints to AP for registration or authentication, as well as access RP services if the user authentication is successful.



(a) Homepage of LeakVoiAuth (b) Authentication process (c) The result of authentication

Fig. 7. UI of VOLERE UA.

In our prototype, we adopted Baidu speech recognition toolkit [41] for speech recognition and used VOLERE synthesized voiceprints for speaker recognition. As shown in Fig. 7b, when the user sends out an authentication request through UA, AP will generate two groups of challenge codes CC_a' (e.g., "have a good day"), CC_b' (e.g., "nice to meet you") and the corresponding speaking modes: SM_a and SM_b. UA will prompt the user how to repeat the challenge codes. When the challenge code lights up (displayed as orange words), it is the time for the user to read it. UA records the two audio segments read by the user, generates a synthesized voiceprint and uploads it to the AP for verification. If the verification is positive, UA gets the result shown in Fig. 7c. Note that the user can select an appropriate language from the upper right corner of the interface shown in Fig. 7a. The two groups of challenge codes shown in Fig. 7b are randomly generated according to the language selected by the user. In our test, the languages can be selected include English, Chinese, and Finnish.

5.2 User Study Participants and Dataset

5.2.1 Preliminary Training and Test

Before our user tests, we carried on preliminary parameter training and testing to get a proper I-Vector model, including its Universal Background Model (UBM) and a threshold ($\theta = 1.4650$), for speaker recognition. Herein, we used two corpuses, TIMIT and AISHELL. TIMIT is an English corpus that contains recordings of 630 speakers, each speaker reading 10 sentences. AISHELL is a Chinese corpus with 50 speakers, each speaker read over 200 sentences. Both TIMIT and AISHELL have a sampling rate of 16kHz and a sampling format of 16 bits. The first 530 speakers of TIMIT and the first 160 speakers of AISHELL are organized as training sets, respectively. We used a training set to train a UBM [14] . Then, we used the remaining data of the corresponding corpus to form a testing set, and tested with the GMM-UBM method and the I-Vector [15] method, respectively. The experimental results show that using I-Vector can always achieve high enough speaker recognition accuracy over 99.5%, better than GMM-UBM, especially with fewer samples. Notably, the selection of samples had a certain impact on EER. When the user samples used in training are abundant, EER is low. In general, I-vector can achieve high enough recognition accuracy with fewer samples than GMM-UBM. Thus, we use I-vector to illustrate VOLERE design and implementation in this paper.

5.2.2 *User Study*

We conducted our user studies in both Finland and China. We directly recruited the first group of participants in universities. The first group includes 23 participants, most of them are between 20 and 40 years old, with similar technical background. Next, we recruited the second group of participants based on the first group by using a snowballing method. The second group includes 42 participants. Most of them were relatives and friends of the first group, and they were more widely distributed in terms of age, occupation and so on. Totally, there were 65 participants from seven countries, in which there are 42 males and 23 females. The participants were aging from 14 to 78 years old and they all owned an Android phone. Most of them had fingerprint authentication experiences, but only 6 participants experienced voiceprint authentication. We performed our user tests in three environments: quite place (12 tests), office (31 tests), and canteen (22 tests). Refer to Appendix D, available in the online supplemental material, for detailed background information of participants. Compared with existing related work [3], [4], which recruited less than 20 participants and each participant only provided 10 utterances, our user study scale is larger. The amount of the participants in our user test is also more than [2]. In particular, our participants were from seven countries and each of them provided 22 utterances.

Before the test, we gave participants a brief introduction to our system and asked them to try it two or three times to make sure they know how to use the system. We also signed privacy agreements with all participants to allow us to save their voice data for research purpose and subsequent experiments. During the user study and test, we asked each participant to do the following: 1) Register his/her synthesized voiceprint into the system; 2) Perform voiceprint authentication tests for 10 times; 3) Re-register and update his/her voiceprint in the system; 4) Perform voiceprint authentication tests for 10 times again.

The data (including authentication time, authentication result, synthesized voiceprints, etc.) of each participant's 20-times authentication tests were recorded for evaluating the performance of VOLERE. We also stored each segment of the voices of the participants for further analyzing the capability of leakage resilience of VOLERE. After the participants finished the user test, we also interviewed them with a questionnaire (shown in Appendix E, available in the online supplemental material) to collect their feedback on usage experience.

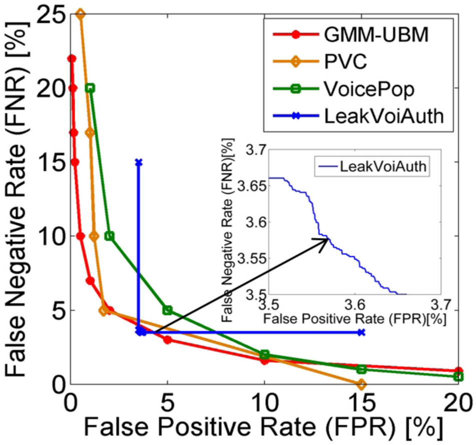


Fig. 8. DET curves of VOLERE and other systems.

5.3 Experimental Results and Analysis

5.3.1 *Accuracy and Efficiency*

For evaluating accuracy and efficiency, we used all registered 65 synthesized voiceprints as matching targets. During the user tests, each of the 65 participants performed 20 authentication operations. In total, we recorded 1,300 synthesized voiceprints for authentication. Each of these voiceprints was checked to judge its validity through speech recognition and speaker recognition. In the speech recognition, AP checks if the user correctly repeats the contents of challenge codes, so as to prevent replay attacks to a certain extent. In the speaker recognition, AP checks whether the voiceprint features in the audio are consistent with the registered ones. In addition, we also tested the authentication accuracy of VOLERE under ambient noise impact.

Fig. 8 shows the DET curve of average authentication accuracy. We observe that VOLERE achieves an EER of 3.57% in authentication. We find that the curve of VOLERE is close to a right angle. In fact, its error rate is mainly raised by the speech recognition regarding random challenges, which causes an error rate about 3.5% in our test. It has existed before the speaker recognition with synthesized voiceprint. As a result, no matter how the parameters of the speaker recognition change, the total error rate cannot be reduced to 0. If we only consider the synthesized voiceprint authentication accuracy, it can reach over 99.8%. Obviously, if we increase the size of our corpus and adopt a better voice content recognition toolkit than the Baidu speech recognition, the EER of VOLERE will be greatly reduced.

We also compared VOLERE with several existing systems: GMM-UBM [2], PVC [3], and VoicePoP [4]. It is clear to see that VOLERE has a superior performance on accuracy. In addition, the average authentication time consumed by VOLERE was also tested. Reading a challenge code costs 1.86 seconds on average. Matching the

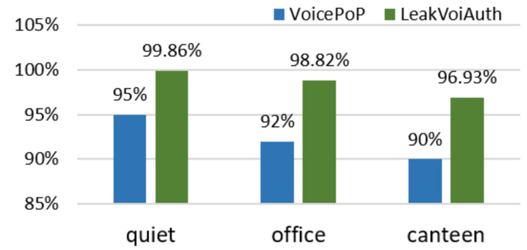


Fig. 9. Authentication accuracy with ambient noise impact.

recording with registered voiceprints costs less than 20 milliseconds on average. Because the user test was carried out in a local network, the time spent for data transmission is not representative. In VOLERE, the total amount of transmitted data is usually less than 200 KB. In general, it takes less than 5 seconds for a user to complete a single authentication. These results show that VOLERE is highly accurate and also efficient.

In order to make sure that a good model was chosen by VOLERE, we tested its performance by using other features that can represent voiceprint instead of MFCC in speech synthesis in the LMA filter. We carried out two tests. We used the recording data of randomly selected 10 participants as the dataset of these tests. In the first test, we used Perceptual Linear Predictive (PLP) coefficients [42] to replace MFCC and got average authentication accuracy as 94.5%, which was slightly lower than using MFCC. In the second test, we used LPCC to replace MFCC for speech synthesis. Average authentication accuracy reduced to only 89.0%. Although the number and offset of datasets may impact testing results, it is obvious that using other features cannot achieve sound accuracy as using LPCC and MFCC.

We further tested the effects of different levels of environmental noise on the accuracy of VOLERE and compared with VoicePoP in three environments: quiet place, office room and canteen. Fig. 9 shows our testing result. It can be seen that both systems are affected by noises. However, VOLERE still maintains over 96.9% accuracy even in the noisiest canteen environment. This shows that VOLERE has better authentication accuracy than VoicePoP regarding the influence of environmental noises.

In order to study the impact of a series of factors on authentication accuracy, we further categorized all participants into different groups according to their profiles. The factors we took into account includes gender, age (11-20, 21-40, 41-60), and nationality. As shown in Fig. 10, VOLERE works better for female, but gender has no much impact on authentication accuracy. About age, VOLERE works best for the age range 21-40, but normally performs well for different age ranges. For nationalities, British voiceprint authentication shows the best, Chinese is the runner up. We think the results are impacted by two reasons. First, whether the participants are native speakers impacts. Participants from other countries than British, Finish and Chinese used English in the user studies. The authentication accuracy of these groups was lower than that of the British group. Second, the used training set may affect. As mentioned earlier, we used two corpora for pre-training and testing. One of the corpora is in English and the other is in Chinese. From the results, although Finnish users use their mother tongue to participate in the user study, the accuracy of this group is lower than that of British and Chinese. This may be caused by some bias in the dataset used for model training. Although nationality impacts the accuracy of VOLERE a bit, the authentication accuracy regarding all nationalities can reach above 95.0%.

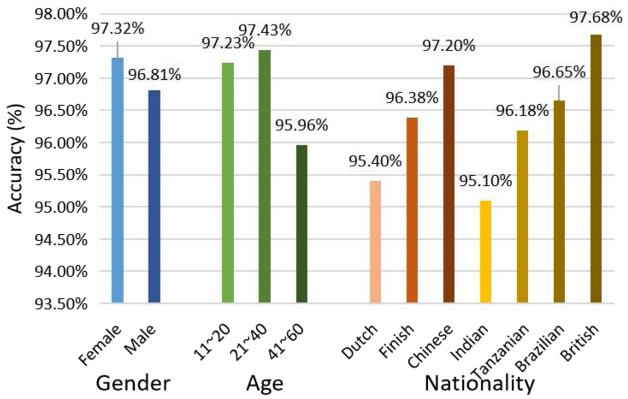


Fig. 10. Authentication accuracy impacted by different factors.

TABLE 1
Authentication Accuracy in Different Speaking Modes

Group	SM_a	SM_b	Accuracy
1	Normal voice	Imitation of the elderly	96.4%
2	Normal voice	Imitation of a child	96.5%
3	Normal voice	Imitation of opposite gender	96.5%

In order to explore the impact of different speaking modes on system performance, we conducted the following experiment. We invited the same 5 participants to join three experiments. In each experiment, each participant conducted one registration operation and ten authentication operations. In all experiments, the first speaking mode of the registration and authentication was appointed as normal speaking mode, while the second speaking mode was appointed as imitation of opposite gender, the elderly and a child, respectively. The average authentication accuracy of VOLERE is shown in Table 1. We can see that there is no significant difference among the three experiments. Therefore, different speaking modes do not have much influence on the accuracy of VOLERE. Randomly assigning speaking modes works stable.

5.3.2 *Time Stability*

Time stability is another important factor that should be considered in voiceprint authentication. We hope that after a reasonably long period of time, a user's voiceprint does not change much compared with his/her originally registered voiceprint. This kind of stability enables a user authentication system to run normally without false rejection. As a result, such a system can ensure authentication accuracy and offer sound user experience. We conducted another user study with 10 of the participants who had participated in the previous user study executed six months ago. According to the contact information left by the participants in the first user test, we randomly selected a part of them and sent user test invitations. There were 10 previous participants who agreed to join in the test. The test process was the same as that of the first test, but we did not repeat the introduction of the system because they still remembered the first test. Since the participants had no further contact with VOLERE in the past six months, we allowed them to explore the system freely to familiarize themselves with it before the test officially started. In this new user study, the participants were asked to repeat the authentication operation for 20 times, and we used the voiceprint registered half year ago for speaker recognition. In this test, the authentication accuracy of VOLERE reaches 96.5%, with a FNR of 3.5%. The results of the test indicates that a long period of time could affect authentication accuracy a bit. But it can still be maintained at a high level, which is acceptable.

For further improving authentication accuracy and minimizing time impact, we tried X-Vector for speaker recognition. X-Vector uses Deep Neural Network (DNN) to improve the performance of speaker recognition. It can map variable-length recordings to fixed-dimensional embedding, and performs better with a large-scale training dataset than I-Vector. We used the 200 recordings in the new user study as a testing set. The testing result based on X-Vector achieves an accuracy of 97.9%, which is better than the result based on I-Vector. This shows that it is feasible to use the X-Vector to optimize the VOLERE system for achieving better time stability. But X-Vector is suitable for a big dataset. When the dataset is not big enough, I-Vector performs better than X-Vector, e.g., when the amount of utterance is less than 50 or the time of each utterance is less than 0.5 seconds.

5.3.3 Leakage Resilience

In order to verify the leakage resilience capability of VOLERE, we conducted a series of experiments as described below. The experiments were based on the voice data collected in our user tests. The dataset contains 1,430 groups of audios, that is, 2 registration audios and 20 authentication audios of 65 participants. Each group of audios contains two original audios in different speaking modes and one synthesized audio. For impersonation attack, we took the synthesized audios collected in registration as a target set. For each target, we took its corresponding 10 groups of authentication audios as a test set, which contains 10 synthesized audios (i.e., positive cases in our test) and 20 original audios (i.e., negative cases). For replay attacks, we also took the synthesized audios collected in registration as a target set, and the synthesized audios collected in authentication as a test set. The positive cases in this test set contain the same challenge codes as the challenged, while the negative cases in this test set contain different challenge codes from the challenged.

Although we analyzed VOLERE security in Section 4.4, we did further verification through experiments. We took the user data that the attacker may obtain as a test set and used the attacker's impersonation target as a training set, which is usually composed of user voiceprints. A user model can be obtained by training, which is matched with the test set to judge if a tested user voiceprint is legitimate. We input these two sets of data into the implemented VOLERE system and judged VOLERE's capability of leakage resilience according to testing results. Specifically, four types of attacks were simulated.

- *Experiments on impersonation attack.* Assuming that an attacker obtains the original voiceprint of a user in any one of two applied speaking modes. The attacker tries to impersonate the user with it. Accordingly, we studied VOLERE's leakage resilience capability by matching all participants' original voiceprints in any speaking modes with their synthesized voiceprints. The training set consists of all SV_{te} and SV_{ta}. The test set consists of all V_a and V_b. After the I-vector model of attacking target is established by using the training set, the data in the test set is input in turn for voiceprint recognition to check whether it can match the target voiceprint. Successful matching means that the attacker can disguise as a legitimate target user by using the audio in a certain speaking mode.
- *Experiments on replay attack with* SV_{te}. Assuming that an attacker obtained any SV_{te} and tried to impersonate the user with it. Accordingly, we studied VOLERE's leakage resilience capability by matching the content of SV_{te} with any other SV_{te} in our dataset through speech recognition. In this case, the training set consists of all SV_{te}, and the test set consists of all SV_{te}. After building the target user model with the training set, the data are extracted from the test set to match it. Successful matching means that the attacker can use SV_{te} to implement replay attack.
- *Experiments on replay attack with* SV_{ta}. Assuming that an attacker obtains any SV_{ta} and tries to impersonate a user with it. Accordingly, we studied VOLERE's leakage resilience capability by matching the content of SV_{ta} with any one of SV_{te} through speech recognition. The training set consists of all SV_{te} and the test set consists of all SV_{ta}. After building the target user model with the training set, the data in the test set is matched with one of them. Successful matching means that the attacker can use SV_{ta} to implement replay attack.
- *Experiments on reverse engineering.* Assuming that an attacker intercepts a user's synthesized voiceprint from the system. The attacker tries to use it to impersonate a legitimate user in various systems. Two situations may happen.

  First, the attacker could use the user's old synthesized voiceprint to match his/her new synthesized voiceprint. We simulated this situation to study our system's leakage resilience capability by matching the user's newly re-registered voiceprint with his/her previously registered one in order to test the effectivity of synthesized voiceprint update (i.e., reregistration).

  Second, the attacker could use a stolen synthesized voiceprint to gain access to an ordinary system. In this case, the stolen synthesized voiceprint is matched with the user's original voiceprint. Accordingly, we studied our system's leakage resilience capability by extracting LPCC and MFCC features from the user's synthesized voiceprint and matching the extracted features with the features of his/her original voiceprint.

For Case 1 and Case 4, the EERs of VOLERE are 0%. For Case 2 and Case 3, FPRs are 3.7% and 4.3%, respectively. This result shows that the leakage resilience capability of VOLERE is pretty good. We found that false authentication was mainly caused by the mistakes of speech recognition. If

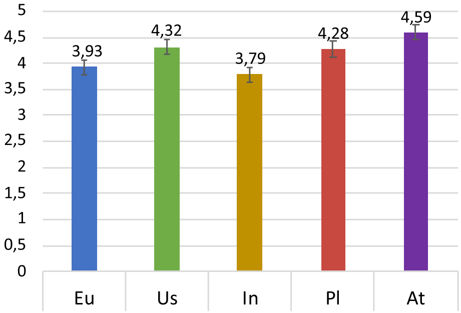


Fig. 11. User acceptance test result.

we improve our speech recognition model by adopting a better toolkit than Baidu speeach recognition, (e.g., by training the speech recognition model with a big dataset), FPR can be further reduced, which is very possible in practice since AP is a cloud server. Note that, it is impossible to compare VOLERE with existing works (e.g., GMM UBM [2], PVC [3], and VoicePoP [4]) regarding leakage resilience since they did not consider this issue. To our best knowledge, VOLERE is the first non-crypto voiceprint user authentication method with leakage resilience.

5.3.4 User Acceptance

We also interviewed the participants with a questionnaire designed based on Technology Acceptance Model (TAM) [43], refer to Appendix E, available in the online supplemental material. It consists of 15 statements in terms of five aspects, i.e., perceived ease of use (Eu), usefulness (Us), user interface (In), playfulness (Pl) and attitude of usage in future (At). Each aspect contains three statements with similar meanings. We calculated the average score of three statements as the score of their corresponding aspect. The measurement of questionnaire is designed based on a 5-point Likert scale. Each of the statements is measured by a score from 1 to 5, i.e., from totally disagree to totally agree.

The result of user acceptance test is shown in Fig. 11. We can see that the average values of users' evaluation on the five factors of TAM are all above 3.7, which implies that the usage experiences of the participants were positive. Specifically, Us, Pl and At were above 4.2, while At exceeded 4.5. This implies that the participants preferred accepting VOLERE and they thought VOLERE is very useful and playful with fun. However, the score about In was not very good, which encourages us to further improve the UI of VOLERE. Considering the prototype of VOLERE was developed by university students, this low score is acceptable. Perceived ease of use is the second lowest score, which should be further improved.

We notice that subjective bias does exist in the questionnaire setting. In addition, there are other problems, such as participants' acquiescence or rejection of extreme scores. The former can be balanced by setting negative bias control group. The latter can be reduced by setting a larger score range. Limited by our test scale and test process, we did not set up such a complex questionnaire. The purpose of this test is to gain usage feedback from the participants regarding their acceptance on VOLERE.

We notice that applying dual speaking modes may cause trouble to some people, including the assumption of binary gender. It is certainly worth noting that system design needs to respect different cultures and concepts, which is consistent with our goal of improving system usability. The prototype system was implemented for proof-of-concept and performance test. Towards practical deployment and real applications, further optimization is essential. There are many alternatives of speaking modes, such as singing, speaking dialects, etc. The user can also use two normal voiceprints to generate a synthesized voiceprint, which is also works but not so secure. By adding a secret during synthesizing, we can also make synthesized voiceprints different from their originals.

TABLE 2
Comparison With Existing Works

Reference	Ac	Ef	TS	LR & PP	UAc	LD	AA	VAaS
VoiceLive [25]	¿99%	-	-	-	-	✓	-	-
VoiceGesture [26]	¿99%	-	-	-	-	✓	-	-
WiVo [27]	99.1%	0.32s	90.3% after 24h	-	-	✓	-	-
ItCaField [28]	99.16%	-	-	-	-	✓	-	-
Void [29]	99.6%	0.035s	-	-	-	96.2%	-	-
Vmask [30]	-	-	-	-	-	-	67.5%	-
FakeBob [31]	-	13.4min	-	-	-	-	99%	-
VOLERE	96.43%	0.02s	96.5%	✓	4.2/5	✓	✓	✓

Ac: Accuracy. Ef: Efficiency. TS: Time Stability.
LR: Leakage Resilience. PP: privacy preservation at the cloud. UAc: User Acceptance.
LD: Liveness Detection. AA: Adversarial Attack success rate.
-: Not mentioned, not considered or not supported.
✓: Supported or resisted.

5.4 Comparison and Discussion

We further compare the performance of VOLERE with existing works according to the evaluation metrics proposed in Section 5, shown in Table 2. We can see that VOLERE outperforms other systems in terms of privacy preservation and leakage resilience with sound accuracy and efficiency. Meanwhile VOLERE shows great time stability. Considering the total authentication time of VOLERE (<5 seconds), VOLERE can offer secure VAaS under adversarial attack FakeBob [31] since it costs too much time (13.4 minutes) to complete a black box attack.

6 Conclusion

In this paper, we proposed VOLERE, a leakage resilient user authentication system with privacy preservation based on random personal voice challenges. We analyzed VOLERE's security and resilience under different private voice information leakages and adversarial attacks. The user tests based on a prototype showed that VOLERE achieves high authentication accuracy with an average EER 3.57%. Its authentication accuracy is reasonably stable regardless user nationality, gender, age, elapsed time, and authentication environments, as well as speaking modes. It also has sound efficiency. It worked very well in the tests of leakage resilience. About usage experiences, VOLERE performed positively on user acceptance. In the future, we will further improve VOLERE usability and security by optimizing its UI design and embedding a more effective liveness detection mechanism into it.

References

[1] R. M. M. Jayamaha, M. R. Senadheera, T. N. C. Gamage, K. P. B. Weerasekara, G. A. Dissanayaka, and G. N. Kodagoda, "Voizlock-human voice authentication system using hidden Markov model," in *Proc. 4th Int. Conf. Inf. Autom. Sustainability*, 2008, pp. 330–335.

[2] J. Gałka, M. Masior, and M. Salasa, "Voice authentication embedded solution for secured access control," *IEEE Trans. Consum. Electron.*, vol. 60, no. 4, pp. 653–661, Nov. 2014.

[3] Z. Yan and S. Zhao, "A usable authentication system based on personal voice challenge," in *Proc. Int. Conf. Adv. Cloud Big Data*, 2016, pp. 194–199.

[4] Q. Wang *et al.*, "VoicePop: A pop noise based anti-spoofing system for voice authentication on smartphones," in *Proc. IEEE Conf. Comput. Commun.*, 2019, pp. 2062–2070.

[5] O. Sarwar, B. Rinner, and A. Cavallaro, "Design space exploration for adaptive privacy protection in airborne images," in *Proc. 13th IEEE Int. Conf. Adv. Video Signal Based Surveillance*, 2016, pp. 159–165.

[6] J. Yu, B. Zhang, Z. Kuang, D. Lin, and J. Fan, "iPrivacy: Image privacy protection by identifying sensitive objects via deep multi-task learning," *IEEE Trans. Inf. Forensics Secur.*, vol. 12, no. 5, pp. 1005–1016, May 2017.

[7] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth, "The reusable holdout: Preserving validity in adaptive data analysis," *Science*, vol. 349, no. 6248, pp. 636–638, 2015.

[8] J. Liao, C. Jiang, and C. Guo, "Data privacy protection based on sensitive attributes dynamic update," in *Proc. 4th Int. Conf. Cloud Comput. Intell. Syst.*, 2016, pp. 377–381.

[9] W. Ding, Z. Yan, and R. H. Deng, "Privacy-preserving data processing with flexible access control," *IEEE Ann. History Comput.*, vol. 17, no. 02, pp. 363–376, Mar./Apr. 2020.

[10] W. Ding, Z. Yan, and R. H. Deng, "Encrypted data processing with homomorphic re-encryption," *Inf. Sci.*, vol. 409, pp. 35–55, 2017.

[11] K. Fan, Q. Tian, J. Wang, H. Li, and Y. Yang, "Privacy protection based access control scheme in cloud-based services," *China Commun.*, vol. 14, no. 1, pp. 61–71, 2017.

[12] S. Donghong, L. Wu, R. Ping, and L. Ke, "Reputation and attribute based dynamic access control framework in cloud computing environment for privacy protection," in *Proc. 12th Int. Conf. Nat. Comput. Fuzzy Syst. Knowl. Discov.*, 2016, pp. 1239–1245.

[13] X. Zhou, S. D. Wolthusen, C. Busch, and A. Kuijper, "Feature correlation attack on biometric privacy protection schemes," in *Proc. 5th Int. Conf. Intell. Inf. Hiding Multimedia Signal Process.*, 2009, pp. 1061–1065.

[14] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[15] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 788–798, May 2011.

[16] P. Matějka *et al.*, "Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2011, pp. 4828–4831.

[17] J. Wang, K.-C. Wang, M. T. Law, F. Rudzicz, and M. Brudno, "Centroid-based deep metric learning for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 3652–3656.

[18] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 4052–4056.

[19] D. Snyder, D. Garcia-Romero , G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5329–5333.

[20] Y. Wang and D. Hatzinakos, "On random transformations for changeable face verification," *IEEE Trans. Syst., Man, Cybern., B (Cybern.)*, vol. 41, no. 3, pp. 840–854, Jun. 2011.

[21] C.-I. Fan and Y.-H. Lin, "Provably secure remote truly three-factor authentication scheme with privacy protection on biometrics," *IEEE Trans. Inf. Forensics Secur.*, vol. 4, no. 4, pp. 933–945, Dec. 2009.

[22] J. K. Pillai, V. M. Patel, R. Chellappa, and N. K. Ratha, "Secure and robust iris recognition using random projections and sparse representations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 9, pp. 1877–1893, Sep. 2011.

[23] S. Li and A. C. Kot, "Privacy protection of fingerprint database," *IEEE Signal Process. Lett.*, vol. 18, no. 2, pp. 115–118, Feb. 2011.

[24] S. Li and A. C. Kot, "Fingerprint combination for privacy protection," *IEEE Trans. Inf. Forensics Secur.*, vol. 8, no. 2, pp. 350–360, Feb. 2013.

[25] L. Zhang, S. Tan, J. Yang, and Y. Chen, "VoiceLive: A phoneme localization based liveness detection for voice authentication on smartphones," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2016, pp. 1080–1091.

[26] L. Zhang, S. Tan, and J. Yang, "Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2017, pp. 57–71.

[27] Y. Meng *et al.*, "WiVo: Enhancing the security of voice control system via wireless signal in IoT environment," in *Proc. 18th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, 2018, pp. 81–90.

[28] C. Yan, Y. Long, X. Ji, and W. Xu, "The catcher in the field: A fieldprint based spoofing detection for text-independent speaker verification," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2019, pp. 1215–1229.

[29] M. E. Ahmed, I.-Y. Kwak, J. H. Huh, I. Kim, T. Oh, and H. Kim, "Void: A fast and light voice liveness detection system," in *Proc. 29th USENIX Secur. Symp.*, 2020, pp. 2685–2702.

[30] L. Zhang, Y. Meng, J. Yu, C. Xiang, B. Falk, and H. Zhu, "Voiceprint mimicry attack towards speaker verification system in smart home," in *Proc. IEEE Conf. Comput. Commun.*, 2020, pp. 377–386,

[31] G. Chen *et al.*, "Who is real bob? Adversarial attacks on speaker recognition systems," 2019, *arXiv:1911.01840*.

[32] V. Tiwari, "MFCC and its applications in speaker recognition," *Int. J. Emerg. Technol.*, vol. 1, no. 1, pp. 19–22, 2010.

[33] H. Gupta and D. Gupta, "LPC and LPCC method of feature extraction in speech recognition system," in *Proc. 6th Int. Conf.-Cloud Syst. Big Data Eng. (Confluence)*, 2016, pp. 498–502.

[34] R.-H. Wang, Q. Liu, and D. Tang, "A new chinese text-to-speech system with high naturalness," in *Proc. 4th Int. Conf. Spoken Lang. Process.*, 1996, vol. 3, pp. 1441–1444.

[35] M. Gales and S. Young, *The Application of Hidden Markov Models in Speech Recognition*. Boston, MA, USA: Now Publishers Inc, 2008.

[36] "Librosa, a python package for music and audio analysis," Dec. 2020. [Online]. Available: https://librosa.org/doc/main/index.html

[37] "HMM/DNN-based speech synthesis system," Dec. 2020. [Online]. Available: http://hts.sp.nitech.ac.jp/?Publications

[38] Z. Li, C. Shi, Y. Xie, J. Liu, B. Yuan, and Y. Chen, "Practical adversarial attacks against speaker recognition systems," in *Proc. 21st Int. Workshop Mobile Comput. Syst. Appl.*, 2020, pp. 9–14.

[39] Y. Xie, C. Shi, Z. Li, J. Liu, Y. Chen, and B. Yuan, "Real-time, universal, and robust adversarial attacks against speaker recognition systems," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 1738–1742.

[40] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 274–283. [Online]. Available: http://proceedings.mlr.press/v80/athalye18a.html

[41] "Baidu speech recognition software," Jan. 2021. [Online]. Available: https://ai.baidu.com/tech/speech/asr

[42] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoustical Soc. America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[43] Z. Yan, Y. Chen, and Y. Shen, "A practical reputation system for pervasive social chatting," *J. Comput. Syst. Sci.*, vol. 79, no. 5, pp. 556–572, 2013.

[44] Y. Yujin, Z. Peihua, and Z. Qun, "Research of speaker recognition based on combination of LPCC and MFCC," in *Proc. IEEE Int. Conf. Intell. Comput. Intell. Syst.*, 2010, vol. 3, pp. 765–767.

[45] S. Nagarajan and R. Sankar, "Efficient implementation of linear predictive coding algorithms," in *Proc. IEEE Southeastcon Eng. New Era'*, 1998, pp. 69–72.

[46] R. Zheng, S. Zhang, and B. Xu, "Text-independent speaker identification using GMM-UBM and frame level likelihood normalization," in *Proc. Int. Symp. Chin. Spoken Lang. Process.*, 2004, pp. 289–292.

Rui Zhang received the BSc degree in computer science and technology from the China University of Mining and Technology, Xuzhou, China, in 2016. Currently she is working toward the PhD degree with Xidian University, Xi'an, China, major in cyber security. Her research interests include information security, user authentication and privacy preservation.

Zheng Yan (Senior Member, IEEE) received the doctor of science degree in technology from the Helsinki University of Technology, Finland. She is currently a professor with Xidian University, China and a visiting professor and Finnish academy research fellow with Aalto University, Finland. Her research interests include trust, security, privacy, and data analytics. She served and is serving as an area/associate editor of the *IEEE Internet of Things Journal*, *Information Fusion*, *Information Sciences*, *IEEE Access*, *Journal of Network and Computer Applications*, etc. She served as a general chair or program chair for more than 30 international conferences including IFIP Networking 2021 and IEEE TrustCom 2015. She is a founding steering committee co-chair of IEEE Blockchain conference. She received several awards, including the Distringuished Inventor Award issued by Nokia (2020), Aalto ELEC Impact Award (2021), 2017 Best Journal Paper Award issued by IEEE Communication Society Technical Committee on Big Data and the Outstanding associate editor of 2017/2018 for *IEEE Access*.

Xuerui Wang received the BE degree from Computer Science and Technology, Xidian University, Xi'an, China, 2018. She is currently working toward the master's degree with Xidian University, Xi'an, China. Her main research interests include attack and defense in authentication systems.

Robert H. Deng (Fellow, IEEE) is currently a AXA chair professor of cybersecurity, director of the Secure Mobile Centre, and deputy dean for Faculty and Research, School of Computing and Information Systems, Singapore Management University, Singapore. His research interests include areas of data security and privacy, network security, and applied cryptography. He received the Outstanding University Researcher Award from National University of Singapore, Singapore, Lee Kuan Yew Fellowship for Research Excellence from SMU, and Asia-Pacific Information Security Leadership Achievements Community Service Star from International Information Systems Security Certification Consortium. He serves/served on the editorial boards of *ACM Transactions on Privacy and Security*, *IEEE Security & Privacy*, *IEEE Transactions on Dependable and Secure Computing*, *IEEE Transactions on Information Forensics and Security*, *Journal of Computer Science and Technology*, and Steering Committee chair of the ACM Asia Conference on Computer and Communications Security. He is a fellow of Academy of Engineering Singapore.