Stop Deceiving! An Effective Defense Scheme Against Voice Impersonation Attacks on Smart Devices

Wenbin Huang, Wenjuan Tang, *Member, IEEE*, Hongbo Jiang, *Senior Member, IEEE*, Jun Luo, *Senior Member, IEEE*, and Yaoxue Zhang, *Senior Member, IEEE*

***Abstract*—Both *voice communication* and automatic speech verification (ASV) over smart devices are vulnerable to the voice impersonation (VI) attack, which is often launched via imitating a target's voice characteristics to deceive human auditory sense or fool the ASV system. Researchers have designed a number of defense schemes yet without the consideration of universality due to the lack of comprehensive data sets. In this article, we propose a universal defense scheme based on the VI data set collected from a famous TV show named "The Sound." First, we deliver a thorough study on the VI attacks in both auditory and ASV systems to verify the collected simulated voice could spoof the auditory and the ASV system with a notable probability. Second, we propose a quasi-Gaussian distribution (QGD)-based defense scheme with the discovery about specific voice characteristics that are distinct between attackers and targets. Finally, we conduct extensive experimental results on our collected VI data set as well as the auxiliary ASVspoof2017 data set, to indicate the proposed QGD scheme outperforms the state-of-the-art schemes: backpropagation neural network, support vector machine, and Gaussian mixture model, in terms of accuracy.**

***Index Terms*—Defense, impersonation attack, smart devices, speech verification, voice characteristics.**

I. Introduction

VOICE, as a kind of convenient and reliable biometric signal, has been widely used in smart Internet of Things (IoT) devices for human–machine interaction. Ever since Apple launched the voice assistant Siri in 2012 [1], many other manufacturers have followed Apple's footsteps to launch voice assistants, such as Amazon Alexa [2] and Google Home [3]. According to the Grand View Researchs report, the market revenue of these intelligent voice assistant is predicted to achieve $31.8 billion by 2025 [4]. Due to the popularity of voice assistants and the distinctiveness of the voice signal, voice-based identification would be a mainstream way in smart devices for information delivery and command control.

However, different from other human biometrics, human voices are often exposed to the public [5], as people communicate over phone calls in public or upload their videos/audios to the Internet, which brings severe security risks into the voice-based systems. An *attacker* with malicious purpose could easily collect voices from a *target* by eavesdropping/recording, or downloading audio clips from his/her online social networking website [6]. Upon collecting sufficient voice samples, the attacker could imitate the target's voice to launch voice impersonation (VI) attacks on the human auditory system and/or automatic speech verification (ASV) system through smart devices [7], [8], which we term human-aimed VI attack and machine-aimed VI attack, respectively.

Human-Aimed VI Attack: VI could trick the human auditory system by imitating the characteristics of voice, such as tone, rhythm, and style [9]. One interesting example is a famous Chinese TV show named "The Sound" [10], in which imitators dub an existing film without appearing, and viewers guess whether the current dub is performed by originators or imitators. This kind of VI may incur serious security risks if performed by attackers, who pretend to be the target's contacts and aim to deceive the target through imitating the contact's voice. For example, attackers could vary voice content with the dialogue scene to make fake voice calls or send scam voice messages, which deliver fake information to damage the targets safety and property [11], [12]. According to the Ministry of Public Security's reports in 2020, human-aimed VI accounts for a very large proportion among the cases of telephone fraud [13].

Machine-Aimed VI Attack: Smart devices (such as smartphones, Amazon Echo, and Google Home) often adopt ASV systems for identity verification before they can be controlled by voice commands [3], [19]. For example, smartphones with an ASV system could transform its mode from lock screen to work by recognizing whether a wake-up command matches the preset voice [20], [21]. However, the ASV system is fragile and vulnerable to be attacked by imitating the target's specific wake-up command [22], [23]. Once the ASV system is spoofed [24], [25], the attacker can perform a series of malicious operations on the smartphones, such as setting a malicious alarm clock, sending a malicious text message, and even making a malicious bank transfer.

Manuscript received April 29, 2021; revised August 10, 2021; accepted August 22, 2021. Date of publication September 6, 2021; date of current version March 24, 2022. This work was supported by the Natural Science Foundation of Hunan under Grant 2021JJ40119. *(Corresponding author: Wenjuan Tang.)*

Wenbin Huang, Wenjuan Tang, and Hongbo Jiang are with the Department of College of Computer Science and Electronics Engineering, Hunan University, Changsha 410082, China (e-mail: wenbinhuang@hnu.edu.cn; wenjuantang@hnu.edu.cn; hongbojiang2004@gmail.com).

Jun Luo is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore (e-mail: junluo@ntu.edu.sg).

Yaoxue Zhang is with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: zhangyx@tsinghua.edu.cn).

Digital Object Identifier 10.1109/JIOT.2021.3110588

From these two kinds of attack scenarios, we can find that the impersonation attack incurs severe security risks in smart devices. To conquer these risks, many efforts have been contributed. The widely used method is based on spectral features proposed by [26], but the performance is not stable since the speaker model and the background model are not always highly distinct. By revealing the variations of feature space, many works [27], [28] use the fundamental frequency or the format frequency change of the vowels for detection. However, feature space with vowel segmentation requires forced alignment and manual correction. By taking advantage of the unique articulatory gesture when users speak a passphrase, Chen *et al.* [17], Yan *et al.* [18], and Zhang *et al.* [29] proposed to detect the VI attack by constructing a magnetic field and a "fieldprint," which are not applicable when people wear masks. Pursuing the optimal detection accuracy of spoofing voices, Białobrzeski *et al.* [30] and Lavrentyeva *et al.* [31] proposed deep learning and neural network models, which require thousands of pieces of data to train a robust network model. Since the lack of VI data set, the neural network method is difficult to achieve good results.

In addition to the shortcomings mentioned above, these schemes have a common limitation: the scenario they apply is that the voice content during the attack is the same as the collected voice content, which we term text-dependent VI attack. Nevertheless, in general human-aimed VI attacks, voice content varies with the dialogue scene, i.e., text-independent VI attack is unfortunately ignored. Considering the limited resources of smart devices, how to discover the common key points of text-dependent and text-independent VI attack is a critical challenge. Furthermore, different from playback [14], speech synthesis [15], and speech conversion [16] that could generate large data sets, the VI data used for the VI detection scheme proposed in the existing literature are generated by inviting a small number of volunteers to conduct mutual simulation. For instance, Mariéthoz and Bengio [26] invited four volunteers, one as a mimicker and the other three as the mimicked, to collect VI data. In [28], only three imitators were invited to generate nine VI data. At this point, the amount of VI data collected is unrepresentative due to small data set and unprofessional impersonation, which hinders the analysis and design of the VI attack defense. Therefore, it is urgent to design an effective detection scheme for both text-dependent and text-independent VI attacks in the absence of a comprehensive data set.

To fill the gap of VI data set along with effective VI attack defense, in this work, we collect a general data set in multiple languages based on a reality TV show, and design general VI defense schemes based on quasi-Gaussian distribution (QGD). Through nontrivial numerical analysis on this data set, we discover that different people's voices vary greatly in Euclidean space, which is effective for both text-independent and text-dependent VI. To recognize the impersonation voice, we propose the QGD scheme based on the Euclidean distance between the input voice and the registered voice. By training different people's voices, a distribution is verified with numerical analysis of the difference. The impersonation voice could be identified if the input voice does not match the verified distribution. The main contributions of this work are summarized as follows.

1) We extract the impersonation audio from the famous show named "The Sound" to construct a real VI data set. Based on the VI data set, we conduct a series of VI attacks on smart devices, and the results show that the collected impersonation data could deceive the auditory and ASV system with a notable probability.
2) We make an important discovery that the voice features of the same people are densely distributed in the Euclidean space, and the voice features of different people are sparsely distributed in the Euclidean space. Based on this discovery, we propose an effective QGD defense scheme to combat text-independent and text-dependent VI attacks.
3) We verify the effectiveness of the QGD defense scheme on the VI data set and ASVspoof2017 data set, and compare it with three state-of-the-art algorithms: back-propagation neural network, support vector machine (SVM), and Gaussian mixture model (GMM). The extensive experimental results demonstrate its high accuracy.

The remainder of this article is organized as follows. Section II presents the related work. The VI attack is presented in Section III. We describe the collected data set in Section IV. We implement the defense scheme for VI attack in Section V. The performance of our scheme is described in Section VI. Finally, we conclude the work in Section VII.

II. Related Work

The VI attack attracted many research that focused on playback [32]–[34], speech conversation [35], [36], and speech synthesis [37], [38].

These impersonation attacks are launched by leaving traces of the physical properties of the recording and playback devices, or signal processing phonetic feature artifacts from synthesis or conversion systems. Based on that, Chen *et al.* [17] and Yan *et al.* [18] proposed to distinguish speakers (either humans or loudspeakers) by constructing a magnetic field and "fieldprint." Because the magnetic field and "fieldprint" are mainly used to distinguish the physical components of loudspeakers from human organs, these methods do not work well when the genuine speaker and impersonator are live human beings. Moreover, they only considered text-dependent attacks, and ignored text-independent attacks that are more harmful.

To distinguish genuine voice and impersonation voice, Mariéthoz and Bengio [26] proposed the method based on mel-frequency cepstral coefficients (MFCCs) cooperated with GMM to build speaker model and background model. The speaker model corresponds to the claimed speaker, and the background model theoretically represents all possible speakers except the claimed speaker. When unknown voice samples are presented to the system during the test, the likelihood ratio test is used for making a binary decision, that is, whether the

given voice samples belong to the claimed speaker model or the background model.

Other research studies detected VI effectively based on the variations of feature space from different users, such as the change of fundamental frequency [27] or the formant frequencies of the vowels [28]. A disguise detector is developed in [28] to detect impersonation voice. The rationale behind the disguise detector is that impersonators practice less of the impersonated voices, and thus, exhibit greater variation in acoustic parameters under disguise. Specifically, the disguise detector uses quadratic discriminant on the first two formants to quantify the amount of acoustic variation on a vowel-by-vowel basis. However, these methods require vowel segmentation by forced alignment and manual correction, and they only consider the text-dependent VI attack.

Additionally, Zhang *et al.* [29] proposed a liveness detection system for voice spoofing attack detection on smartphones. When a user speaks the wake-up command to a smartphone, the advanced mobile audio hardware is used to perceive and extract the user's specific pronunciation and gesture characteristics. However, pronunciation and gesture features are affected by many factors, such as age-related organ changes and facial features such as beards, which can lead to a higher rate of false positives. In addition, under the current situation of COVID-19, this method is not applicable when people wear masks for voice verification.

Driven by the development of deep learning, Białobrzeski *et al.* [30] and Lavrentyeva *et al.* [31] applied deep learning for voice spoofing attack detection. Białobrzeski *et al.* [30] proposed to detect spoofing voice by using robust Bayesian and light neural networks. This method is motivated by the hypothesis that Bayesian models are robust to overfitting and the generalization problem could be improved by several regularization techniques. However, the neural network is fragile and susceptible to small perturbations. Moreover, the deep learning-based method requires thousands of pieces of label data to train the neural network model. In the absence of large-scale VI data set, we establish a backpropagation neural network, which does not show good performance in our experiment.

From the above-related work, we find that there are still many shortcomings in the research of VI attack detection. They need forced alignment followed by manual correction and cannot provide reasonable performance. Additionally, many of them only consider the text-dependent VI attack and ignore the text-independent VI attack. To fill this gap, we propose a general VI spoofing detection scheme.

III. Voice Impersonation Attack

In this section, we show how attackers can impersonate the target's voice biometric information to perform spoofing phone calls or fool the ASV system, which we call human-aimed VI attack and machine-aimed VI attack, respectively.

A. Human-Aimed VI Attack

The attacker's goal is to pretend to be a target's contact by mimicking the voice characteristics of the target's contact, and then to conduct fake voice calls or send scam voice messages to deceive the target. The attack model is shown in Fig. 1.

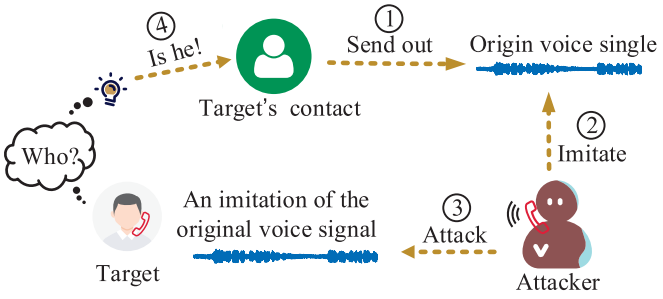


Fig. 1. Attack model for voice communications.

As depicted in Fig. 1, the attacker attacks the target by imitating the target contact's voice timbre and rhythm. This attack is easy to succeed without showing the face.

B. Machine-Aimed VI Attack

The attack model in this scenario is the attacker imitates the target's voice to speak specific command to wake up the target's smart device. We carry out a series of tests to wake up smartphones with impersonated voices, and the results unexpectedly show the impersonated voice can successfully deceive popular smartphones with a noticeable probability.

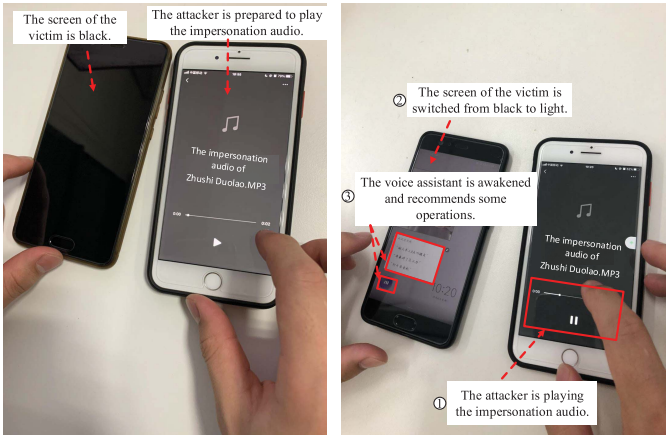
Specifically, we use popular and modern phones; iPhone 8 Plus to play the impersonated voice, and HUAWEI P10 Plus, Xiaomi 10, and Xiaomi 10 Pro as the target devices, which support voice verification, voice command execution, as well as user-defined wake-up commands in 8 bytes.

1) We take the specific attack command of *Zhushi Duolao* on HUAWEI P10 Plus as an example to illustrate the attack process as shown in Fig. 2. First, we take iPhone 8 Plus to play the genuine audio three times as an enrollment for the speaker model. After speaker enrollment, the impersonation audio is played. We can see from this figure, the phone is locked before attack, while after the attacker plays the impersonated voice, the phone is wakened up and recommends some operations.
2) We select six genuine audios and their impersonation audios to conduct VI attack on the three attacked devices. The impersonation audio is played 50 times and 101 times, respectively, and the successful waken up rate can be seen in Table I.

IV. Data Collection

We extract the VI data set from a famous variety program named "The Sound" [10], in which impersonators dub a film without appearing, and viewers guess whether the current dub is performed by the originator or an impersonator. Languages included in the VI data set we collected include Chinese, English, Russian, and Chinese dialects (Sichuan dialect, Cantonese, etc.).

The VI data set contains 454 sets of multilingual data, including the genuine audio of a classic film that appears in "The Sound," and the impersonation audio of a segment



(a) (b)

Fig. 2. Attacker is conducting the VI attack. (a) Before the attack. (b) After the attack.

TABLE I
VI ATTACKS CONDUCTED ON DIFFERENT DEVICES

HUAWEI P10 Plus			
No.	**Commands**	**success rate**	
		50 times	***101 times***
1	*Zhushi Duolao.*	70%	77.2%
2	*Airen Xiangsha.*	52%	60.4%
3	*Qinzhe Weichou.*	48%	47.5%
Xiaomi 10			
No.	**Commands**	**success rate**	
		50 times	***101 times***
1	*Ruren Yinshui.*	56%	54.5%
2	*Zhushi Duolao.*	44%	47.5%
3	*Zhanluan Weiping.*	42%	39.6%
Xiaomi 10 Pro			
No.	**Commands**	**success rate**	
		50 times	***101 times***
1	*Ruren Yinshui.*	42%	51.4%
2	*Cangsheng Liluan.*	46%	44.5%
3	*Qinzhe Weichou.*	36%	41.5%

performed by the impersonator. Each set contains four audio pieces, which are *genuine audio 1* (G1), *genuine audio 2* (G2), *genuine audio 3* (G3), and impersonation audio of genuine audio 1 (IG1), respectively. Among them, G1, G2, and G3 are three sentences from different content of the same person, extracted from the simulated movie. The IG1 has the same content as G1 but the speaker is different. To guarantee the quality of the extracted audios, the impersonation audio and genuine audio are selected from clear segments with similar sentence lengths ranging from 3 to 7 s. To avoid the effects of different emotions on voice, the voice is collected from the same emotional state.

V. QGD DEFENSE SCHEME

In this section, we demonstrate our QGD scheme for detecting VI attacks in detail.

A. Overview

The proposed QGD scheme consists of the major modules: signal processing, feature extraction, distribution matching, and decision logic, as shown in Fig. 3. In which, the upper

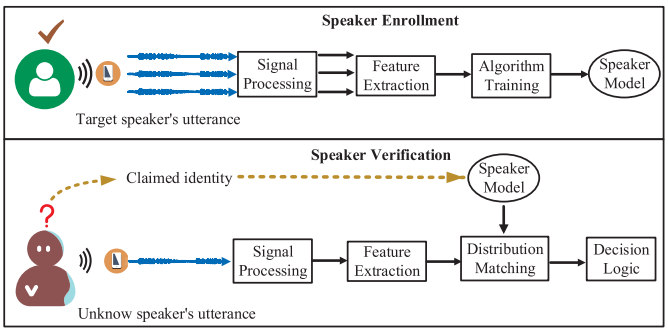


Fig. 3. Modular representation of QGD.

and lower panel represent the enrollment process and the verification process, respectively. In the enrollment process, with three voices that first conducted signal processing in time and frequency domain, high-quality voices are obtained. Then, a number of acoustic features are extracted from the processed signals, utilized for calculating the Euclidean distance and training a speaker model. In the verification process, the voice input by an unknown speaker is preprocessed with feature extraction in a similar way as the speaker enrollment. Then, the Euclidean distance of the features is matched with the speaker model trained in the enrollment. The speakers' identity is verified if the unknown voice obeys the speaker model; otherwise, it is rejected.

B. Signal Processing

To obtain a high-quality voice for verification, we first execute signal processing, including preweighting, framing, and windowing. Since silent pauses, interrupts, as well as the transient background noises could affect the effectiveness of feature extraction and they do not reflect the speaker's identify features, we first use a preweighed filter to remove these traditional nonspeech parts of the signal.

After generating a signal with almost no pauses among words and phrases, we segment the signal into 20-ms time frames, with a 50% overlap between successive ones. The frame length of 20 ms is shorter in duration than typical phonemes, which guarantees the signal within the frame to be stable. The overlap ensures even the shortest phonemes can be used for identity recognition.

We multiply each frame signal by a Hamming window function to reduce the influence of spectrum leakage before extracting the signals features. Although window functions reduce the amplitude of early or late phonemes that appear in frames, this problem can be solved by overlapping between frames. Finally, we obtain a clean speech signal containing identity features for next step of feature extraction.

C. Feature Extraction

To imitate the nonlinear mechanism of human auditory, we take the method of MFCC to extract the features of voice. MFCC computation starts from fast Fourier transform (FFT) of the input sampled audio and computes the discrete Fourier transform. To optimize the FFT performance, we set the FFT

input length to be the next power of 2 from the original signal length. For example, when the audio sample rate is 48 kHz, an FFT with 1024-point is performed on a 20-ms frame, which contains 960 data points, resulting in a spectral frequency resolution of 48 kHz/1024 = 46.875 Hz. The output is passed through a filterbank-based upon Mel scale, given as $\text{mel}(x) = 1125\log(1 + x/700)$. For an input voice signal with N samples, the filters are given by

$$f_m = \left(\frac{N}{F_s}\right)\text{mel}^{-1}\left(\text{mel}(f_l) + m\frac{\text{mel}(f_h) - \text{mel}(f_l)}{M+1}\right) \quad 0 \le m \le M+1 \tag{1}$$

where M is the total number of filters, F_s is the sampling frequency, and f_l and f_h are the lowest and highest frequencies of the filter. The normalized triangular filters [40], $H_m(k)$ are then given by

$$H_m(k) = \begin{cases} 0, & k < f_{m-1} \\ \frac{2(k-f_{m-1})}{(f_{m+1}-f_{m-1})(f_m-f_{m-1})}, & f_{m-1} \le k \le f_m \\ \frac{2(f_{m+1}-k)}{(f_{m+1}-f_{m-1})(f_{m+1}-f_m)} & f_m < k \le f_{m+1} \\ 0, & k > f_{m+1} \end{cases} \quad 1 \le m \le M. \tag{2}$$

Next, the energy output of filterbank [41] is given by

$$S_m = \log\left[\sum_{k=0}^{N-1}|X_k|^2 H_m(k)\right] \quad 1 \le m \le M. \tag{3}$$

MFCC is the discrete cosine transform (DCT) of the M filter outputs, as

$$\text{c}(n) = \sum_{m=0}^{M-1} S_m \cos\left[\pi n\left(m - \frac{1}{2}\right)/M\right] \quad 0 \le m \le M. \tag{4}$$

This filter procedure reduces the number of coefficients in the compression of the information. The sample averaging also reduces the variance of FFT within each filter. Finally, the energy vector is logarithmically compressed, and then the DCT is performed. DCT has two main purposes: first, it separates the content of the slowly changing spectral envelope (such as the vocal tract) from the rapidly changing speech excitation signal. In this way, MFCC only preserves the low-order coefficients related to the channel. Second, due to the spectral features of speech and the overlapping properties of filters, the elements of log filter bank vectors are correlated, such that the elements of eigenvectors are decorrelated. The obtained decorrelation coefficient is suitable for further analysis.

D. Distribution Matching

To distinguish genuine voice and impersonation voice effectively, we propose a similarity measurement method based on the basic principle that the greater the spatial distance between voices, the greater the difference between them. By measuring the spatial distance between voice features, this method can detect both text-independent and text-dependent VI attack.

The similarity measurement method we proposed is to calculate the Euclidean distance [42] between two voice signals according to the features extracted in Section V-C. For the VI data set we collected contains four audio pieces (G1, G2, G3, and IG1) in each set, and the Euclidean distance between two pairs, such as the Euclidean distance between G1 and IG1, is calculated by

$$\text{Dist}(\text{G1}, \text{IG1}) = \sqrt{\sum_{i=1}^{n}(\text{G1}_i - \text{IG1}_i)^2} \tag{5}$$

where G1_i and IG1_i represent the ith frame feature of G_1 and IG_1, respectively.

For an unknown voice sample IG1 input, if it is relatively matched with G1, their Euclidean distance is very close. We select a set of voice samples from the data set and choose eight frames randomly for demonstration. Figs. 4 and 5 show the distribution of acoustic features of human-aimed attacks and machine-aimed attacks, respectively.

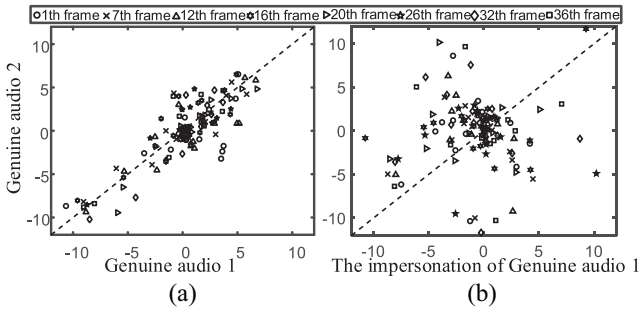


Fig. 4. Text-independent feature parameter distribution of different voices for human-aimed attack. (a) Distribution of text-independent phonetic features between the genuine speaker. (b) Distribution of text-independent phonetic features between the impersonator and genuine speaker.

The main risk of the VI attack aimed at humans is the text-independent characteristic, that is, the attacker can vary the voice content according to the dialogue scene. In Fig. 4, we compare the difference in the parameter feature distribution of text-independent voices between the impersonator and the genuine speaker. In comparison of Fig. 4(a) and (b), we observe that under the text-independent attack, the feature space distribution of the genuine speaker's different voices is denser than the feature space distribution between impersonator and the genuine speaker.

While for the machine-aimed VI attack, it is text dependent. The attacker aims to spoof the ASV system by imitating the specific wake-up command of the target. In Fig. 5, we compare the difference in the parameter feature distribution of text-dependent voices between the impersonator and the genuine speaker. Fig. 5(a) shows the distribution of the phonetic feature between the text-dependent voices from the genuine speaker is close to the 45th line while the impersonator and the genuine speaker's are much sparse as shown in Fig. 5(b).

According to the Euclidean distance calculated in this section, we obtain a very interesting discovery: although the impersonation voice can spoof the auditory and ASV system successfully, the feature distribution between the genuine and impersonation voices is more scattered in the Euclidean space no matter they are text independent or text dependent.

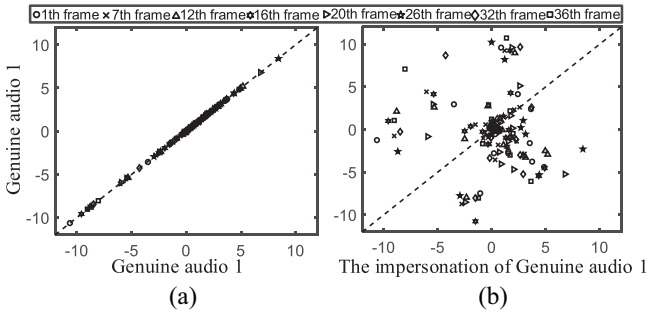


Fig. 5. Text-dependent feature parameter distribution of different voices for machine-aimed attack. (a) Distribution of text-dependent phonetic features between the genuine speaker. (b) Distribution of text-dependent phonetic features between the impersonator and genuine speaker.

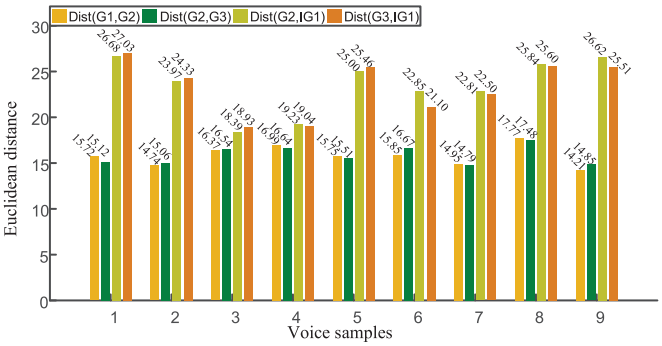


Fig. 6. Euclidean distance of voice signals for human-aimed attack.

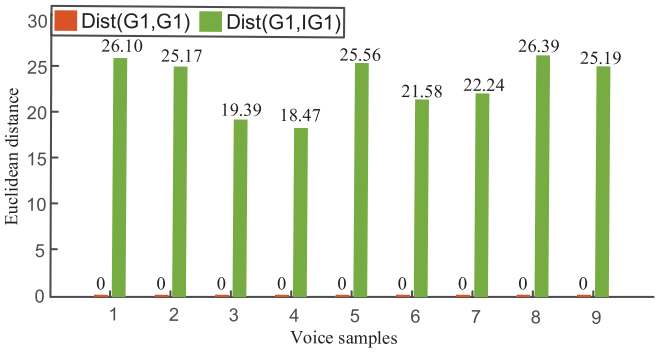


Fig. 7. Euclidean distance of voice signals for machine-aimed attack.

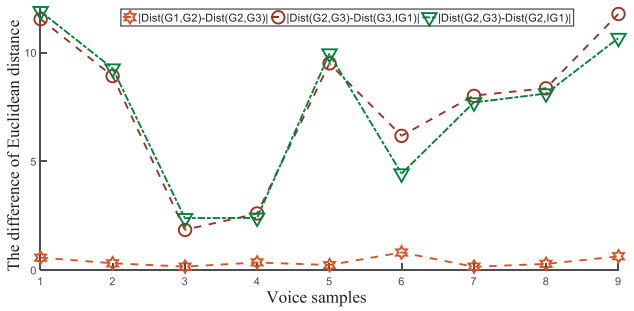


Fig. 8. Euclidean distance gap for human-aimed attack.

E. Algorithm Training

To explore the difference between genuine voice and impersonation voice more visualized, we randomly select nine sets of voice samples in the VI data set for the graphical expression of Euclidian distance as shown in Figs. 6 and 7.

For the text-independent human-aimed attack, we display Euclidean distance between the IG1 and G2, G3 in Fig. 6. The results show that although the attacker can impersonate the target's speech habits and the rhythm to achieve the purpose of the auditory deception, the Euclidean distance between the imitate voice and genuine voice is larger. As shown in Fig. 6, the Dist(G2, IG1) and Dist(G3, IG1) are larger than Dist(G1, G2) and Dist(G2, G3).

The Euclidean distance calculated for the VI attack aimed at machine is shown in Fig. 7. Because the genuine speaker's voice is text dependent, the Dist(G1, G1) almost equals 0. But for impersonation voice, although it is text dependent, it still shows a very large Euclid distance.

Based on the analysis demonstrated from Figs. 6 and 7, we discover that although impersonation voice can successfully deceive the auditory system and machine's ASV system, Euclidean distances are significantly different. For the text-dependent VI aimed at machines, we can distinguish impersonation voice by calculating the Euclidean distance between the input voice and the voice samples registered in the speaker enrollment phase. If the Euclidean distance equals 0, we can recognize the input voice as a genuine voice; otherwise, recognize it as an impersonation voice.

While for the text-independent VI aimed at human, since the voice contents are not totally the same, and the Euclidean distance cannot be 0, how to detect such attack is a critical challenge. Fortunately, as shown in Fig. 6, we observe the Euclidean distance between genuine voice and impersonation voice is much larger than the Euclidean distance between genuine voice, and the Euclidean distance between genuine voices is very small. On top of this pivotal observation, we are naturally driven to answer the following question.

Whether the Euclidean distances among genuine voice and/or impersonation voice subject to a law?

To further answer this question, we employ the difference analysis method to explore the general law of internal implication by calculating the Euclidean distance gap between text-independent voices as follows:

$$|(\text{Dist}(\text{G1, G2}) - \text{Dist}(\text{G2, G3})| \tag{6}$$

$$|(\text{Dist}(\text{G2, G3}) - \text{Dist}(\text{G2, IG1})| \tag{7}$$

$$|(\text{Dist}(\text{G2, G3}) - \text{Dist}(\text{G3, IG1})| \tag{8}$$

where Dist(G1, G2), Dist(G2, G3), Dist(G2, IG1), and Dist(G3, IG1) are the Euclidean distance we calculate in Section V-D. Equation (6) represents the Euclidean distance gap between the text-independent genuine voices. Equations (7) and (8) represent the Euclidean distance gap between the text-independent genuine voice and the impersonation voice.

We can analyze from the Euclidean distance gap demonstrated in Fig. 8 that the gap between the text-independent genuine voices tends to be stable while the Euclidean distance gap between the text-independent genuine voice and the

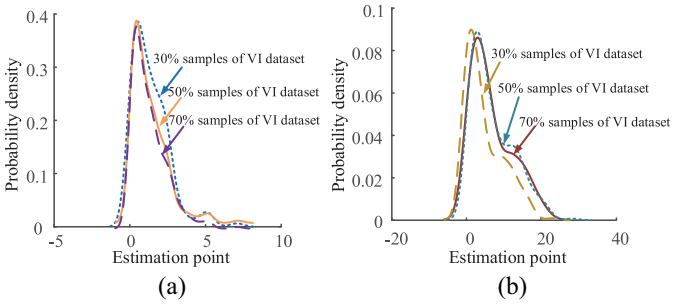


Fig. 9. Discrete density function estimation on VI data set. (a) Euclidean distance gap between genuine voices. (b) Euclidean distance gap between genuine and impersonation voice.

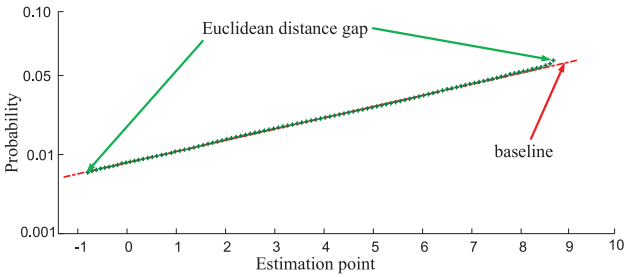


Fig. 10. Normal probability plot for Gaussian property verifies the Euclidean distance gap between genuine voices.

impersonation voice is more floating. Based on this pivotal observation and analysis, we put forward a conjecture.

The stable state of the Euclidean distance gap obeys a certain distribution, and can be used to distinguish the genuine voice and impersonation voice.

To prove this conjecture, we take the method of the probability density function [43] to further investigate the stable state of the Euclidean distance gap between the text-independent genuine voices. The probability density estimation based on the Euclidean distance gap is performed as

$$\hat{f}(y) = \frac{1}{n}\sum_{i=1}^{n} w(y - y_i; h) \tag{9}$$

where $\{y_1, \ldots, y_n\}$ denote the Euclidean distance gap we calculated. w is the kernel (a symmetric probability density function), and $h > 0$ is a smoothing parameter (know as the bandwidth).

Based on the vast amount of VI data set we collected, we perform probability density estimation on different numbers of voice samples. We take 30%, 50%, and 70% of the voice samples in VI data set to evaluate the probability density estimation, and the result is shown in Fig. 9.

With the above numerical analysis of distribution estimation by calculating the probability density, we observe distinctly from Fig. 9(a) that the Euclidean distance of the text-independent genuine voices is subject to Gaussian-like distribution. With the increasing number of samples, the tendency to approximate Gaussian distribution is even more pronounced. Meanwhile, the distribution of Euclidean distance gap of the text-independent genuine voice and impersonation voice is verified in Fig. 9(b). There are two main differeces from Fig. 9(a). First, the peak value of Fig. 9(b) is much higher than the peak value of Fig. 9(a). Second, if we regard both Fig. 9(a) and (b) as Gaussian distribution, the mean value of Fig. 9(b) is much larger than Fig. 9(a).

To further verify the Gaussian-like property of the Euclidean distance gap, we take the normal probability plot [44], a graphical technique to identify substantive departures from normality. The core idea of normal probability plot is that if the data approximate a normal distribution, the value chosen for the resulting image would be close to a straight line. We mark the straight line as baseline in Fig. 10. It can be observed from Fig. 10 that most of the Euclidean distance gaps are distributed on the baseline. Hence, we draw an interesting law: the Euclidean distance gap distribution of the text-independent genuine voices is subject to the Gaussian distribution.

F. Decision Logic

We provide two different decision strategies for two kinds of VI attack detection under two different scenarios.

For text-dependent machine-aimed VI attack detection, we provide QGD-T by calculating the Euclidean distance between the input voice and the registered voice. If the Euclidean distance equals 0, we can recognize the input voice as a genuine voice; otherwise, recognize it as an impersonation voice. For the specific wake-up command is G, the input voice from an unknown speaker is IG, if

$$\mathrm{Dist}(G, IG) = \sqrt{\sum_{i=1}^{n}(G_i - IG_i)^2} = 0 \tag{10}$$

we see IG as a genuine voice, otherwise, it is an impersonation voice. G_i and IG_i represent the ith frame feature of G and IG, respectively.

For text-independent human-aimed VI attack detection, we answer the proposed questions according to the QGD we proved in Section V-E at first. First, we discover the Euclidean distance gap between the text-independent genuine voices tends to be stable. Then, this stable state is proved to obey QGD. Based on the Gaussian-like property of the Euclidean distance gap between the text-independent genuine voices and the core idea of the distribution matching in Section V-D, we provide QGD-I. If an unidentified input voice IG1 does not obey the QGD between the registered voices G1 and G2, then IG1 is considered to be impersonation voice. The QGD-I is formally expressed as, if

$$|\Big(\mathrm{Dist}(\mathrm{G1}, \mathrm{G2}) - \mathrm{Dist}(\mathrm{G2}, \mathrm{IG1})| \approx \mathrm{QGD}\Big(\mu, \sigma^2\Big)\Big) \tag{11}$$

then we see IG1 as an impersonation voice. *QGD* is the Gaussian-like distribution we proved above, and μ and σ^2 are the mean value and the variance of the Gaussian-like distribution, respectively.

VI. Experimental Analysis

In this section, we evaluate our proposed QGD scheme in terms of VI detection accuracy based on our collected VI data

TABLE II
DETAILS OF THE ASVSPOOF 2017 DATA SET

Subset	Speakers	Type of Utterances	
		Genuine	*Replayed*
Training set	10	1507	1507
Development set	8	760	950
Evaluation set	24	1298	12,008

set and authoritative ASVspoof2017 data set [45], as well as compare with three state-of-art algorithms: 1) backpropagation (BP); 2) SVM; and 3) GMM, to evaluate the performance.

A. Experimental Setup

The experiments are run on a Microstar workstatio equipped with Intel Core i7-10700 CPU of 2.90-GHz processor, 16 Cores, and 16-GB RAM. The software used is MATLAB 2016A with three voice toolboxes: 1) voicebox [50] to process the voice sample; 2) MSR Identity Toolbox [51] for GMM training; and 3) LIBSVM Toolbox [52] to train the SVM model.

Auxiliary Data Set: For the playback attack is a tricky text-dependent VI attack, which completely replicates the victim's voice characteristics and voice content by recording the voice with the microphone and playing it by the speaker, so as to deceive the auditory system and the ASV system. To further verify the effectiveness of the proposed QGD scheme against VI attacks, we utilize an authoritative playback attack data set ASVspoof2017, derived from the RedDots corpus [45] as the other evaluation data set. It consists of training set, development set, and evaluation set, as shown in Table II.

Contrast Schemes: To investigate the effectiveness of our proposed QGD scheme, we compare it with three popular VI attack detection schemes: 1) feature model-based training scheme; 2) deep learning-based scheme; and 3) machine learning-based scheme. The first is GMM proposed in [26], which recognizes the speaker by building the speaker model and background model and has been widely used in the ASV system. The second one is BP neural network, which plays a core role in neural network models and is widely used in voice recognization [30], [31]. It recognizes impersonation voice by two processes: 1) forward propagation and 2) backpropagation [46]. The third one is SVM, which separates the normal from the abnormal by training a hyperplane. Osowska and Osowski [47] and Chauhan *et al.* [48] utilized the SVM to classify the extracted features for authentication and achieve good performance.

Performance Metrics: We use true reject rate (TRR) and false reject rate (FRR) to evaluate the performance of the proposed QGD scheme. TRR measures the proportion of impersonation audios that are correctly distinguished, and is calculated by TRR$=$ TP$/$(TP $+$ FA). FRR measures the proportion of genuine that are wrongly identified as impersonations, and is calculated by $FRR = \mathrm{FP}/(\mathrm{FP}+\mathrm{TA})$. TR, TA, FR, and FA represent true reject (identify impersonation as impersonation), true accept (identify genuine as genuine), false reject (identify genuine as impersonation), and false accept (identify impersonation as genuine), respectively. From these definitions, we can see that a smaller FRR and higher TRR could demonstrate better detection accuracy.

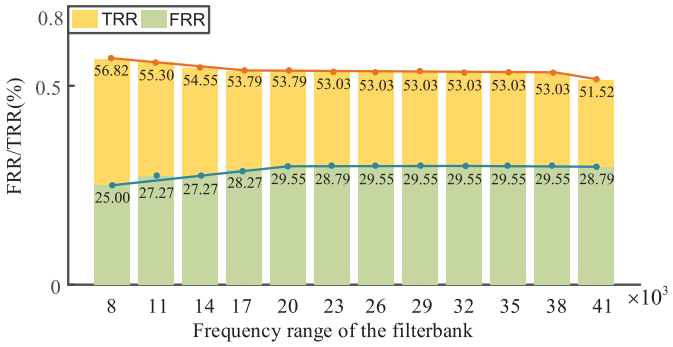


Fig. 11. Performance affect by frequency range of the filterbank on the data set VI.

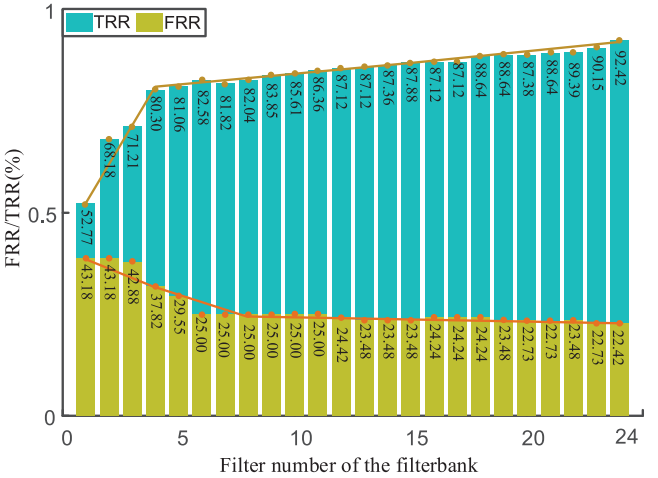


Fig. 12. Performance affect by the filter number of the filterbank on the data set VI.

B. Factors Affecting VI Detection

In the detailed experimental process, we find how to extract effective features significantly that affect the VI detection accuracy. Specifically, we analyze that the key point of feature extraction is a set of filters to simulate the nonlinear features. From (1), it can be obtained that there are mainly two factors affecting the filter: 1) the number of filters and 2) the frequency of the filterbank. To explore the influence of these two factors on the detection accuracy, we conducted parametric scanning to analyze the performance with different filter parameters. Discovered by the effective parameter selection scheme [49], the number of filters in a filterbank varies from 1 to 24 at an interval of 1. Since the human ear perceives the frequency at 20–20 kHz and is most sensitive around 4 kHz, we set 8 kHz as the low frequency boundary of the filterbank according to Nyquist's theorem. The frequency range of the filterbank varies from 8 to 41 kHz at an interval of 3 kHz. We conduct parameter scanning on both the VI data set and ASVspoof2017 data set using the similar parameters.

We scan the parameters on the VI data set as shown in Figs. 11 and 12, and they represent the performance affected by the filterbank frequency and the filter number, respectively.

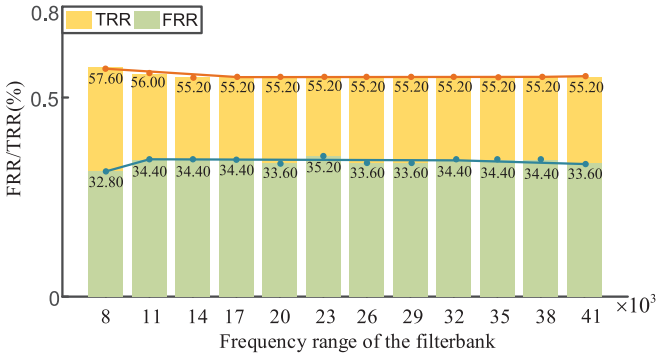


Fig. 13. Performance affect by the frequency range of the filterbank on the data set ASVspoof2017.

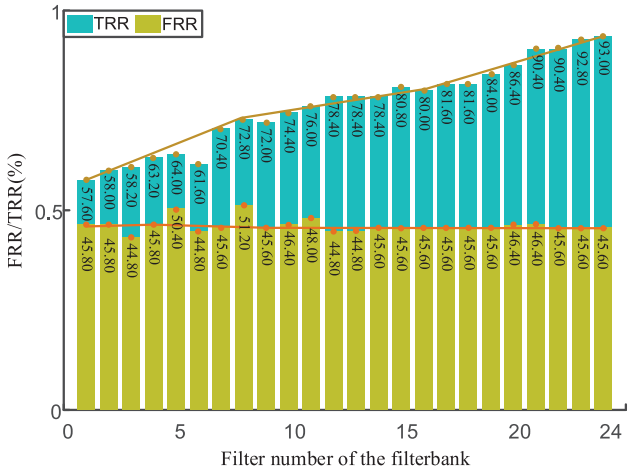


Fig. 14. Performance affect by the filter number of the filterbank on the data set ASVspoof2017.

We set the filter number that equals 1 as the default setting. It is obvious in Fig. 11 that with the filterbank frequency increases, the TRR degrades and the FRR increases. Based on the performance criteria in Section VI-A, we select the highest TRR with lowest FRR to detect VI. Therefore, we set the filterbank frequency equal to 8 kHz. While for the filter number scanned in Fig. 12, with filter number increases, the TRR increases and the FRR degrades. Based on the performance criteria, we set the filter number that is equal to 24.

Figs. 13 and 14 illustrate the effects of frequency and filter number on the ASVspoof2017 data set. Similarly, with the filterbank frequency increases, the TRR degrades and the FRR increases. According to the performance criteria, the higher TRR and lower FRR state the performance is better. Similarly, the best filterbank frequency is 8 kHz, and the best filter number is 24.

C. Experimental Results

After determining the filterbank frequency and filter number that affect the performance of VI detection, we conduct the QGD scheme and the three contrast algorithms on both our collected VI data set and ASVspoof2017 data set. The experimental results are described in detail as follows.

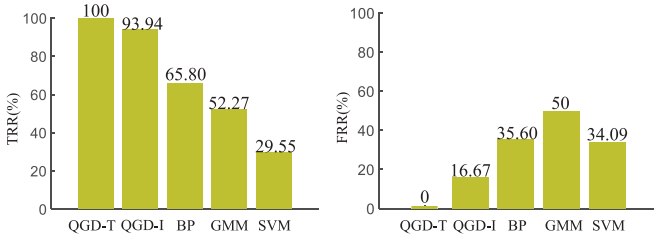


Fig. 15. Detection accuracy on VI data set.

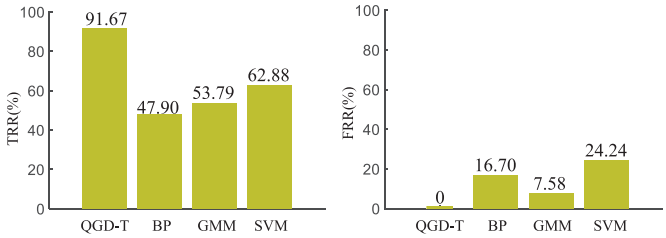


Fig. 16. Detection accuracy on ASVspoof2017 data set.

The VI data set we collected contains 454 sets of voice samples. Based on the principle of data set partition [39], we take 70% of the data set as the training set and 30% as the test set, i.e., 317 sets of voice samples are used for training and 137 sets of voice samples are used for test. The results are shown in Fig. 15, where the QGD-I represents the text-independent VI detection and the QGD-T represents the text-dependent VI detection, respectively.

From Fig. 15, we can demonstrate that our proposed QGD scheme can achieve high detection accuracy. For the QGD-T scheme on the text-dependent VI detection, we achieve 100% TRR and 0% FRR. While for the text-independent VI detection, our QGD-I scheme achieves the highest TRR 93.94% and 16.67% FRR. The BP, GMM, and SVM achieve the similar accuracy in both text-independent VI detection and text-dependent VI detection, of which the TRRs are 65.80%, 52.27%, and 29.55%, while the FRRs are 35.60%, 50.00%, and 34.09%, respectively.

The playback voice is consistent with the content of real voice in playback attack, i.e., the playback attack is the text-dependent spoofing attack. Therefore, we only take the QGD-T detection strategy for playback attack detection, and the results are shown in Fig. 16. As shown in Fig. 16, the QGD-T scheme achieves 91.67% TRR, while the TRRs of BP, GMM, and SVM are 47.90%, 53.79%, and 62.88%, respectively; our QGD-T scheme achieves 0% FRR, while the FRRs of BP, GMM, and SVM are 16.70%, 7.58%, and 24.24%, respectively.

As can be seen from Fig. 16, although the playback attack is a text-dependent spoofing attack conducted by the same people, the proposed QGD scheme still achieves a high detection accuracy compared with other algorithms. We analyze the reason that the addition of intermediate recording devices causes some differences and which could be recognized by the QDG to distinguish the playback voice from the genuine voice.

In summary, from the above figures, we can observe that our proposed QGD scheme achieves high accuracy in both text-independent and text-dependent VI detection, and is robust to playback attack detection. GMM shows similar TRR in VI and ASVspoof2017 data set, but the FRR in the VI data set is very high. In the context of limited data set, both deep learning (BP neural network) and machine learning (SVM) methods do not show good performance compared with QGD methods. Although the performance of the BP neural network is better than that of GMM and SVM in the VI data set, the performance of playback attack detection is poor due to its poor portability.

VII. Conclusion

In this article, we proposed an effective VI defense scheme based on our collected data set sourcing from a real TV show, and the extensive experimental results demonstrate high accuracy. The collected data set is the first VI data set extracting from real human voices, which could support further analysis as a basic data foundation. Meanwhile, the proposed scheme discovers voice feature distinctiveness among different people through numerical analysis, and combats the VI attack on smart devices. In future work, we would further analyze the efficiency of the proposed scheme on a smart device platform.

References

[1] Apple. (2019). *Homekit—Apple Developer*. [Online]. Available: https://developer.apple.com/homekit/

[2] Amazon. (2019). *Alexa*. [Online]. Available: https://www.alexa.com/

[3] *Google Assistant*. (2021). [Online]. Available: https://assistant.google.com/

[4] G. V. Research. (2018). *Voice and Speech Recognition Market Size, Share and Trends Analysis Report*. [Online]. Available: https://www.grandviewresearch.com/press-release/global-voicerecognition-industry

[5] K. Delac and M. Grgic, "A survey of biometric recognition methods," in *Proc. Int. Symp. Electron. Mar.*, 2004, pp. 184–193.

[6] *AudioBoom*. (2021). [Online]. Available: https://audioboom.com/

[7] M. Shirvanian and N. Saxena, "Wiretapping via mimicry: Short voice imitation man-in-the-middle attacks on crypto phones," in *Proc. CCS*, 2014, pp. 868–879.

[8] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Commun.*, vol. 66, pp. 130–153, Feb. 2015.

[9] Z. Elisabeth, "Impersonation reproduction of speech," Lund Univ., Lund, Sweden, Working Paper, 2001.

[10] *Youtube*. (2020). [Online]. Available: https://www.youtube.com/watch?v=rTxFjeLbgNU

[11] W. Tang, J. Ren, K. Deng, and Y. Zhang, "Secure data aggregation of lightweight e-healthcare IoT devices with fair incentives," *IEEE Internet Things J.*, vol. 6, num. 5, pp. 8714–8726, Oct. 2019.

[12] W. Tang, J. Ren, and Y. Zhang, "Enabling trusted and privacy-preserving healthcare services in social media health networks," *IEEE Trans. Multimedia*, vol. 21, no. 3, pp. 579–590, Mar. 2019.

[13] (2020). *CCTV-13 2020, Voice Impersonation*. [Online]. Available: http://tv.cntv.cn/video/C10318/9484e1499ce34d2580a6a78870e5afbc

[14] M. Singh, J. Mishra, and D. Pati, "Replay attack: Its effect on GMM-UBM based text-independent speaker verification system," in *Proc. UPCON*, 2016, pp. 619–623.

[15] A. A. Gritsenko, T. Salimans, R. van den Berg, J. Snoek, and N. Kalchbrenner, "A spectral energy distance for parallel speech synthesis," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Assoc., 2020.

[16] S. Berrak, Y. Junichi, K. Simon, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 132–157, 2020.

[17] S. Chen *et al.*, "You can hear but you cannot steal: Defending against voice impersonation attacks on smartphones," in *Proc. ICDCS*, 2017, pp. 183–195.

[18] 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. CCS*, 2019, pp. 1215–1229.

[19] *WeChat, Voiceprint*. (2015). [Online]. Available: http://thenextweb.com/apps/2015/03/25/wechat-onios-now-lets-you-log-in-using-just-your-voice/

[20] H. Melin, "Automatic speaker verification on site and by telephone: Methods, applications and assessment," Ph.D. dissertation, KTH, Stockholm, Sweden, 2006.

[21] R. Togneri and D. Pullella, "An overview of speaker identification: Accuracy and robustness issues," *IEEE Circuits Syst. Mag.*, vol. 11, no. 2, pp. 23–61, 2nd Quart., 2011.

[22] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu, "Dolphinattack: Inaudible voice commands," in *Proc. CCS*, 2017, pp. 103–117.

[23] W. Tang, K. Zhang, J. Ren, Y. Zhang, and X. Shen, "Flexible and efficient authenticated key agreement scheme for BANs based on physiological features," *IEEE Trans. Mobile Comput.*, vol. 18, no. 4, pp. 845–856, Apr. 2019.

[24] L. Zhang, Y. Meng, J. Yu, C. Xiang, F. Brandon, and H. Zhu, "Voiceprint mimicry attack towards speaker verification system in smart home," in *Proc. INFOCOM*, 2020, pp. 377–386.

[25] Q. Yan, K. Liu, Q. Zhou, H. Guo, and N. Zhang, "SurfingAttack: Interactive hidden attack on voice assistants using ultrasonic guided waves," in *Proc. NDSS*, 2020, pp. 1–18.

[26] S. Mariéthoz and S. Bengio, "Can a professional imitator fool a GMM-based speaker verification system?" IDIAP Res. Inst., Martigny, Switzerland, Rep. IDIAPŰRR 05-61, 2005.

[27] J. Eichhorn, R. Kent, D. Austin, and H. Vorperian, "Effects of aging on vocal fundamental frequency and vowel formants in men and women," *J. Voice Found.*, vol. 35, no. 5, pp. 644.e1–644.e9, 2017.

[28] T. B. Amin, P. Marziliano, and J. S. German, "Glottal and vocal tract characteristics of voice impersonators," *IEEE Trans. Multimedia*, vol. 16, no. 3, pp. 668–678, Apr. 2014.

[29] L. Zhang, T. Sheng, and J. Yang, "Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication," in *Proc. CCS*, 2017, pp. 57–71.

[30] R. Białobrzeski, M. Kosmider, M. Matuszewski, M. Plata, and A. Rakowski, "Robust Bayesian and light neural networks for voice spoofing detection," in *Proc. Interspeech*, 2019, pp. 1028–1032.

[31] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin, "Audio replay attack detection with deep learning frameworks," in *Proc. Interspeech*, 2017, pp. 82–86.

[32] J. Yang, L. Liu, and Q. He, "Discriminative feature based on FWMW for playback speech detection," *Electron. Lett.*, vol. 55, no. 15, pp. 861–864, 2019.

[33] J. P. Lesso, "Detection of loudspeaker playback," U.S. Patent 11 051 117, 2020.

[34] D. Li *et al.*, "Multiple phase information combination for replay attacks detection," in *Proc. Interspeech*, 2018, pp. 656–660.

[35] Z. Wu, E. S. Chng, and H. Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2012, pp. 1700–1704.

[36] A. Kargathara , K. Vaidya, and C. K. Kumbharana, "Analyzing desktop and mobile application for text to speech conversation," in *Rising Threats in Expert Applications and Solutions*. Singapore: Springer, 2020, pp. 331–337.

[37] L. Chen, W. Guo, and L. Dai, "Speaker verification against synthetic speech," in *Proc. Int. Symp. Chin. Spoken Lang. Process.*, 2010, pp. 309–312.

[38] P. L. De Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga, "Evaluation of speaker verification security and detection of HMM-based synthetic speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 20, no. 8, pp. 2280–2290, Oct. 2012.

[39] H. Liu and M. Cocea , "Semi-random partitioning of data into training and test sets in granular computing context," *Granular Comput.*, vol. 2, no. 4, pp. 357–386, 2017.

[40] S. Sigurdsson, K. B. Petersen, and T. Lehn-Schiøler, "Mel frequency cepstral coefficients: An evaluation of robustness of MP3 encoded music," in *Proc. ISMIR*, 2006, pp. 286–289.

[41] K. K. Bhuvanagiri and S. K. Kopparapu, "Modified mel filter bank to compute MFCC of subsampled speech," 2014. [Online]. Available: arXiv:1410.7382.

[42] D. Vasilyev and A. Rashich, "SEFDM-signals Euclidean distance analysis," in *Proc. EExPolytech*, 2018, pp. 75–78.

[43] E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Stat.*, vol. 33, no. 3, pp. 1065–1076, 1962.
[44] J. Filliben, "The probability plot correlation coefficient test for normality," In *Technometrics*, vol. 17, no. 1, pp. 111–117, 1975.
[45] T. Kinnunen *et al.*, "RedDots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research," in *Proc. ICASSP*, 2017, pp. 5395–5399.
[46] Y. Kong, L. Jia, and J. Zhang, "Research on voice print recognition based on wavelet analysis and BP-GA," *Trans. Comput. Sci. Eng.*, to be published.
[47] A. Osowska and S. Osowski, "Voice command recognition using statistical signal processing and SVM," in *Proc. Int. Work-Conf. Artif. Neural Netw.*, 2019, pp. 65–73.
[48] N. Chauhan, T. Isshiki, and D. Li, "Speaker recognition using LPC, MFCC, ZCR features with ANN and SVM classifier for large input database," in *Proc. ICCCS*, 2019, pp. 130–133.
[49] L. Lu, L. Liu, M. J. Hussain, and Y. Liu, "I sense you by breath: Speaker recognition via breath biometrics," *IEEE Trans. Dependable Secure Comput.*, vol. 17, no. 2, pp. 306–319, Mar./Apr. 2017.
[50] M. Brookes. (1997). *Voicebox: Speech Processing Toolbox for MATLAB*. Accessed: Mar. 2011. [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html
[51] S. O. Sadjadi, M. Slaney, and L. Heck, *MSR Identity Toolbox*, Microsoft, Seattle, WA, USA, 2013.
[52] C. Chang and C. Lin, "LIBSVM: A library for support vector machines," in *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011.

Wenbin Huang is currently pursuing the Ph.D. degree with the College of Computer Science and Electronics Engineering, Hunan University, Changsha, China.

His research interests include Internet of Things Security, network security, and sensor network security.

Wenjuan Tang (Member, IEEE) received the Ph.D. degree from Central South University, Changsha, China, in 2019.

She is an Associate Professor with the College of Computer Science and Electronic Engineering, Hunan University, Changsha. From 2016 to 2018, she was a Visiting Scholar with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. Her current research interests include security and privacy in network computing and Internet of Things.

Hongbo Jiang (Senior Member, IEEE) received the Ph.D. degree from Case Western Reserve University, Cleveland, OH, USA, in 2008.

He is currently a Full Professor with the College of Computer Science and Electronic Engineering, Hunan University, Changsha, China. He was a Professor with the Huazhong University of Science and Technology, Wuhan, China. His research concerns computer networking, especially algorithms and protocols for wireless and mobile networks.

Prof. Jiang is serving as an Editor for IEEE/ACM TRANSACTIONS ON NETWORKING, an Associate Editor for IEEE TRANSACTIONS ON MOBILE COMPUTING, and an Associate Technical Editor for *IEEE Communications Magazine*.

Jun Luo (Senior Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from Tsinghua University, Beijing, China, in 1997 and 2000, respectively, and the Ph.D. degree in computer science from the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland, in 2006.

From 2006 to 2008, he worked as a Postdoctoral Research Fellow with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. In 2008, he joined the faculty of the School of Computer Science and Engineering, Nanyang Technological University, Singapore, where he is currently an Associate Professor. His research interests include mobile and pervasive computing, wireless networking, machine learning and computer vision, as well as applied operations research. More information can be found at https://personal.ntu.edu.sg/junluo/.

Yaoxue Zhang (Senior Member, IEEE) received the B.Sc. degree from the Northwest Institute of Telecommunication Engineering, Xi'an, China, in 1982, and the Ph.D. degree in computer networking from Tohoku University, Sendai, Japan, in 1989.

He is currently a Professor with the Department of Computer Science and Technology, Tsinghua University, Beijing, China. He has published over 200 technical papers in international journals and conferences, as well as nine monographs and textbooks. His research interests include computer networking, operating systems, ubiquitous/pervasive computing, transparent computing, and big data.

Prof. Zhang is a Fellow of the Chinese Academy of Engineering.