Hindawi
Security and Communication Networks
Volume 2021, Article ID 6664578, 10 pages
https://doi.org/10.1155/2021/6664578

Research Article

Spoofing Speaker Verification System by Adversarial Examples Leveraging the Generalized Speaker Difference

Hongwei Luo,[1] **Yijie Shen**,[2] **Feng Lin**,[2] **and Guoai Xu**[1]

[1]*National Engineering Laboratory of Mobile Network Security, Beijing University of Posts and Telecommunications, Beijing 100876, China*
[2]*Institute of Cyberspace Research, College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China*

Correspondence should be addressed to Feng Lin; flin@zju.edu.cn

Received 1 December 2020; Revised 25 December 2020; Accepted 29 January 2021; Published 9 February 2021

Academic Editor: Ting Chen

Copyright © 2021 Hongwei Luo et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Speaker verification system has gained great popularity in recent years, especially with the development of deep neural networks and Internet of Things. However, the security of speaker verification system based on deep neural networks has not been well investigated. In this paper, we propose an attack to spoof the state-of-the-art speaker verification system based on generalized end-to-end (GE2E) loss function for misclassifying illegal users into the authentic user. Specifically, we design a novel loss function to deploy a generator for generating effective adversarial examples with slight perturbation and then spoof the system with these adversarial examples to achieve our goals. The success rate of our attack can reach 82% when cosine similarity is adopted to deploy the deep-learning-based speaker verification system. Beyond that, our experiments also reported the signal-to-noise ratio at 76 dB, which proves that our attack has higher imperceptibility than previous works. In summary, the results show that our attack not only can spoof the state-of-the-art neural-network-based speaker verification system but also more importantly has the ability to hide from human hearing or machine discrimination.

1. Introduction

In recent years, the verification system has been employed in many scenarios including entrance guard, online payment, and smart home management. Speaker verification system that offers a convenient and reliable verification is a popular biometrics system. It verifies the person by an utterance that contains the unique biometrics feature called voiceprint. Voiceprint has the advantages of noncontact, high privacy, and low cost compared with other biometrics features used in verification system (e.g., fingerprint [1], face ID [2], and iris features [3]). Therefore, speaker verification has become a promising biometrics technique and received high social acceptance, especially in the area of smart Internet of Things (IoT) such as voice assistant.

One concern of applying speaker verification systems is whether they are secure enough. To explore this latent risk, we first reviewed the speaker verification system to understand how the speaker verification system works. We found that current speaker verification systems can be divided into two types: one is text-dependent speaker verification (TD-SV) and the other is text-independent speaker verification (TI-SV). They have different requirements for inputs. TD-SV requires users to say the same utterance with the one used to enroll, but users can say any utterance for verification in TI-SV. Obviously, TI-SV offers a more convenient speaker verification system than TD-SV. However, we raised a question about the security of the speaker verification system based on TI-SV: Can the speaker verification system based on TI-SV be spoofed by the adversary?

Firstly, we need to find the state-of-the-art speaker verification system based on TI-SV. We found that deep-neural-networks-based speaker verification shows a better performance than conventional solutions through the review. Because it shows a more promising prospect, we set it as our target speaker verification system to explore the latent risk. With the further survey, we raised another question:

Can the speaker verification system based on deep neural networks be spoofed by adversarial examples? To find out whether the possibility of spoofing exists, we firstly perform a comprehensive review of voiceprint identification technology using deep neural networks [4–6]. Among these works, a state-of-the-art embedding vector model with generalized end-to-end (GE2E) loss function [5] proposed by Google performs best and has been applied in many domains (e.g., education and speaker verification). Our work tries to explore the vulnerability in TI-SV based on the GE2E loss function due to the above knowledge. We rebuild this speaker verification system, therefore, with the TIMIT dataset and try to spoof it by adversarial examples. To achieve the spoofing attack, we firstly give two requirements.

1.1. Imperceptible. The adversarial examples need to be similar to the original utterance. In other words, the injected perturbation to the original utterance should be slight enough. Otherwise, the spoofing attack will be discovered with ease by humans or machines, which will cause the spoofing attack to fail.

1.2. Purposeful. The victim should be verified to the special target set by the adversary. Then the spoofing attack is more powerful and destructive.

We consider that the speaker verification system verifies the user by detail waveform rather than the macroscopical waveform. The trait gives us the possibility to spoof the speaker verification system with a slight perturbation. Based on this possibility, we found several technical challenges that still need to be addressed to realize the spoofing attack: (i) How to generate the perturbation for the victim? (ii) How to limit the perturbation slightly enough? (iii) How to evaluate the perturbation, which can measure the influence in utterance? We structure a novel adversarial examples generator for producing effective slight perturbation to spoof the state-of-the-art speaker verification system, which is shown in Figure 1. The adversary utilizes the generator to generate a tiny perturbation to inject into an utterance of the unregistered user; the synthetic utterance will be verified as the targeted registered legal user. In addition, we designed a novel loss function to achieve the imperceptibility that tries to find the slightest perturbation on the premise of the spoofing attack. We proposed two different methods that can evaluate perturbation in maximum noise ratio (MNR) and signal-to-noise ratio (SNR). Finally, we evaluated our spoofing attack on the TIMIT dataset, which contains 630 speakers and 6300 sentences. Based on the dataset, we rebuild the state-of-the-art speaker verification system with GE2E loss function and two identification models, linear discriminate analysis (LDA) and cosine similarity threshold, respectively. In our experiments, the spoofing attack achieved up a high success rate of 82% and a slight distortion of −77 dB in MNR (usually negative) and 76 dB in SNR (usually positive).

To summarize, contributions in our work can be listed as follows:

(1) We proposed a novel multifactor-based attack to spoof the state-of-the-art TI-SV system based on deep learning. Our spoofing attack transforms an illegal utterance's verification result into a legal target with a slight perturbation. Meanwhile, we do not have any utterance from the target. To the best of our knowledge, this is the first exploratory work in spoofing the state-of-the-art TI-SV system based on deep learning.

(2) We consider imperceptibility as a key metric in the spoofing attack. Hence, our spoofing attack improves imperceptibility by a novel loss function. The result shows that the imperceptibility of our spoofing attack is much better than previous works.

(3) We evaluated our spoofing attack on the state-of-the-art TI-SV system based on deep learning with the TIMIT dataset including 630 speakers. The result shows that our spoofing attack achieves up a high success rate and high imperceptibility.

2. Related Work

This section will introduce previous works about spoofing attacks in speaker verification systems and adversarial examples in the audio domain.

2.1. Attack on Speaker Verification Services. Many conventional methods were used to realize speaker verification systems before deep learning, including two models that work best: *i*-vector [7] and Gaussian mixture model-universal background model (GMM-UBM) [8]. The two models acquired effective results in realizing speaker verification systems, but there is still some vulnerability that was found by adversaries. For instance, two genetic algorithms were used to produce a new utterance, which will be identified as the target in both *i*-vector and GMM-UBM, with a target utterance from the dataset [9]. Moreover, adversaries can also use voice conversion to transform the original utterance to the target utterance, and the transformed utterance will have similar features to the target utterance in speaker verification systems [10]. The deep-learning-based verification systems also can be spoofed (e.g., the spoofing attack in *d*-vector [11]). In our work, we aimed at the state-of-the-art TI-SV and proposed a novel attack method to spoof the state-of-the-art speaker verification system.

2.2. Audio Adversarial Examples. Adversarial examples have been utilized in different domains as spoofing attack methods and reported effective results in recent research [12–14], including the audio domain. Speech-to-text is an important problem in the audio domain; an effective solution has been universally utilized, which is called connectionist temporal classification [15] (CTC). With the development of technology, a spoofing attack that is based on adversarial examples has been proposed, one generator will produce a perturbation to make the new utterance sounds like the original one, and Carlini and Wagner [16]

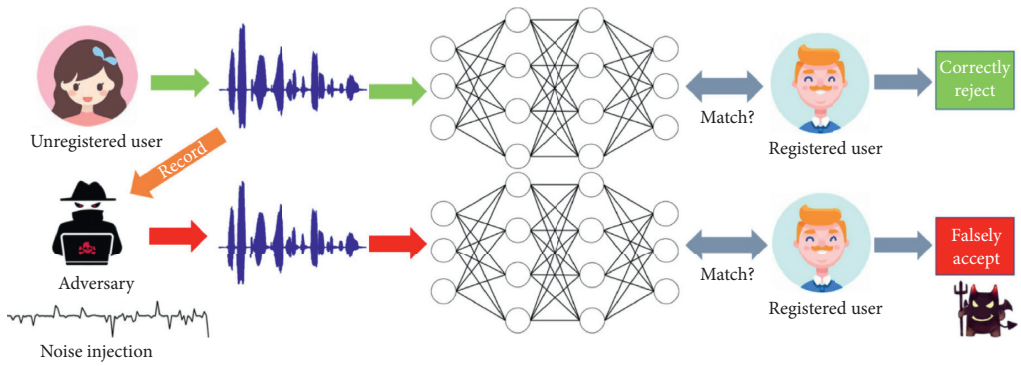

Figure 1: The adversary injects a special and slight perturbation into the original utterance, which does not change the waveform but influences the verification result.

proposed utilizing this method to spoof the speech-to-text system. The spoofing attack [17] to the speech-to-text system is reported as more imperceptible. Aiming at intelligent voice assistants, Qin et al. [18] proposed an attack to produce an incorrect command that cannot be detected by people. Our work is based on adversarial examples to attack the state-of-the-art speaker verification system. Compared with previous works, our target is the speaker verification system rather than the speech-to-text system and our attack is more imperceptible.

3. Background

In this section, we will elaborate on technologies and related concepts that were used in our work.

3.1. Speaker Verification Systems. First of all, we will briefly introduce two different forms of speaker verification systems, TD-SV and TI-SV. There are several works in TD-SV [19, 20], but TD-SV required users to say the fixed utterance in both enrollment and verification. The constraint makes TD-SV not able to be utilized in continued verification and high user experience requirement situation. Hence, TI-SV was proposed to mitigate this constraint. TI-SV focuses on finding the features from speakers independently; also several works [21, 22] were proposed by the benefit of these advantages. TI-SV is more practical and convenient than TD-SV. To this end, it is a meaningful work to explore the attack possibility on TI-SV.

Due to this point, our work focuses on exploring vulnerabilities in TI-SV, and we first reviewed the technologies of TI-SV. Through the review, we found that several works realized TI-SV based on *i*-vector [7] and GMM-UBM [8]. Further, we found that, with the rapid development of deep learning networks, it was also used in speaker verification to extract the voiceprint to representing the speaker's identity. The deep-learning-based TI-SV performs better than previous solutions based on *i*-vector and GMM-UBM. Because of the higher efficiency of the deep-learning-based TI-SV, our attack is targeting deep-learning-based TI-SV. We analyzed the main process of these solutions. For an utterance, the voiceprint extractor employs an embedding network to extract feature vectors which can represent the identity of the speaker from a processed utterance. After the process, the deep-learning-based TI-SV verifies feature vectors by different methods. We found that the loss function is a critical part that will influence the final performance directly during our experiment. Hence, we reviewed the deep-learning-based TI-SV, and we found the state-of-the-art loss function, tuple-based end-to-end [6] (TE2E) loss function, and generalized end-to-end [5] (GE2E) loss function, which have been widely used in real life and achieved great performance. Next, we will briefly illustrate GE2E and its prior work TE2E.

3.1.1. Tuple-Based End-to-End. TE2E outputs embedding vectors for one evaluation utterance and enrollment utterances by long short-term memory (LSTM); and the centroid of the enrollment embedding vectors is calculated. The centroid can be represented by c_k, where k represents the k^{th} speaker, and it is mean value vector of all utterances from the k^{th} speaker. The evaluation embedding vector is represented by e_j, where j represents the j^{th} speaker. TE2E quantifies the distance between c_k and e_j by cosine similarity with the formula

$$s = w \times \cos\left(e_j, c_k\right) + b, \tag{1}$$

where w and b are the parameters that will be learned during training. Based on these definitions, TE2E loss is defined as follows:

$$L_T\left(e_j, c_k\right) = \delta\left(e_j, c_k\right)\sigma(s) + \left(1 - \delta\left(e_j, c_k\right)\right)(1 - \sigma(s)), \tag{2}$$

$$\sigma(x) = \frac{1}{\left(1 + e^{-x}\right)}, \tag{3}$$

where if j equals k, then $\delta(e_j, c_k) = 1$; otherwise, $\delta(e_j, c_k) = 0$. Although TE2E can work well in both TI-SV and TD-SV, it has several disadvantages. Firstly, TE2E gets a scalar representing the similarity between embedding vector e_j and single centroid c_k; this makes the network not able to capture features from other enrollment speakers. Therefore, they further proposed the GE2E model.

3.1.2. Generalized End-to-End. Compared with TE2E, GE2E builds the similarity matrix between $e_{(ji)}$ and all c_k rather than a single c_k with an $N \times M$ input, where N represents N speakers, M represents M utterances for each speaker, $j \in \{0, 1, \ldots, N-1\}$, $i \in \{0, 1, \ldots, M-1\}$, and $k \in \{0, 1, \ldots, N-1\}$. When calculating c_k, if $j == k$, then calculate c_k with other $M-1$ utterances; otherwise, calculate c_k with M utterances. Then GE2E defined two different loss functions, the softmax loss function is defined by Equation (4), and the contrast loss function is defined by Equation (5).

$$L\left(e_{ji}\right) = -S_{ji,j} + \log \sum_{k=1}^{N} \exp\left(S_{ji,k}\right), \tag{4}$$

$$L\left(e_{ji}\right) = 1 - \sigma\left(S_{ji,j}\right) + \max_{\substack{1<k<N \\ k \neq j}} \sigma\left(S_{ji,k}\right), \tag{5}$$

where $\sigma = (1/(1+e^{-x}))$, S is the similarity matrix, and $S_{ji,k}$ represents the similarity between the i^{th} utterance of the j^{th} speaker and the center of the k^{th} speaker. e_{ji} is the i^{th} utterance of the j^{th} speaker. The softmax loss function performs well on TI-SV and contrast loss function performs well on TD-SV. Because GE2E considers global features rather than local features, it can extract more unique features than TE2E.

Since the GE2E loss function has better performance than the TE2E loss function, our work utilized the GE2E loss function with softmax loss function to realize the TI-SV as our attack target.

3.2. Adversarial Examples. As a learning-based spoofing attack method, adversarial examples were proposed to inject tiny perturbation, which will lead the network to output incorrect results with high confidence. It was first proved effective in the image domain. The difference between the attacked image and the original image cannot be distinguished by the human eyes, but the attacked image will be classified into different results with the original image. In the audio domain, adversarial examples attack also exists. For instance, a generator is utilized to produce adversarial examples to spoof speech-to-text systems [16]. Here, we employed adversarial examples to generate the perturbation that cannot be perceived by ears to realize our attack.

4. Attack Model

In this section, we will illustrate the target and constraints of our spoofing attack.

As the final goal, the adversary wants to transform the illegal utterance to be identified into legal identity, which has been set as the target victim, hereafter A, by the adversary. To achieve this target, for any given utterance x which is from anyone other than the victim, hereafter B, where x belongs to B, the adversary tries to generate a slight perturbation δ and introduced δ into the audio waveform x as a new audio waveform x', where x' equals $x + \delta$. The adversary wants x' to be verified as A, which can spoof the speaker verification system in mobile phone, online systems, and so on.

We assume a black-box setting where the adversary only knows the output scores and the identified result of the speaker verification system, without the detailed structure of speaker verification system that is required in white-box setting. In addition, we assume that our adversarial examples δ can be directly introduced into the waveform without any noise (e.g., ambient noise when we play them over the air). These constraints are reasonable as they also appear in prior work [16]. We prefer to prove the possibility of this attack rather than the practical application. To make the work more confident, we also discussed several advanced attacks in Section 9 to overcome these constraints on this basis, which will improve the practical ability of our work. Note that our spoofing attack injects noise into the real utterance; thus, the antispoofing method by living detection [23, 24] is not effective aiming at this attack.

5. Attack System Design

In this section, we will start with the adversarial examples' requirements in our work. Next, we will design two different systems by unique loss functions and describe them.

5.1. Adversarial Examples' Requirements. As an effective attack, the adversarial examples in our experiments need to satisfy several requirements. In an adversarial example's generative process, we will inject perturbation into a given original utterance to generate a new utterance; we define the new utterance with the following equation:

$$x' = x + \delta, \tag{6}$$

where δ represents the perturbation that will be injected into the utterance and x represents the original utterance. To make the attack effective, we need to restrict the process with the three following points: (1) x' must be in an available range which let the waveform be able to be recovered as an utterance; (2) δ needs to be as slight as possible; (3) the speaker verification system will recognize x' as the special target that the adversary sets before the attack. To better describe our requirements for adversarial examples, we formulate our requirements with the following formulations:

$$\begin{aligned} & x + \delta \in [-1, 1], \\ & \min \quad (\delta), \\ & s.t. \quad C(x + \delta) = t, \end{aligned} \tag{7}$$

where $C(\cdot)$ represents the classification and t represents the target classification result. We make several designs in our adversarial example's generator to satisfy the above requirements. We will elaborate these designs in the remainder of the section.

5.2. The Clip Function. We first design the solution for requirement (1). It is difficult to limit the generator to generate δ, which will make x' in an available range. But we found that we can set any value, which is out of the available range,

as the minimum or maximum value for the available range. It hardly affects the auditory effect. Thus, we designed a special clip function and the generator needs each clip point in δ by the function to keep the new utterance available; the clip function is defined as follows:

$$\text{clip}(x+\delta) = \min(\max(x+\delta, -1), 1). \tag{8}$$

The above function satisfies requirement (1) well through our experiments. Next, we need to design special loss functions to satisfy requirements (2) and (3).

5.3. Generalized Relevancy Based Attack. Based on the clip function, we designed loss functions to satisfy requirements (2) and (3). As we introduced in Section 4, we need our attack to be able to work under the black-box setting; in other words, the adversary can only get the classification results and the confidence of each speaker from the speaker verification system. The current work needs to estimate a special θ for the loss function [25]; θ will deeply influence the attack performance. Our generator excludes θ's influence, which lets us need not choose special θ before the training. We mainly designed our generalized relevancy loss function based on GE2E loss function; the loss function was represented as follows:

$$\text{Loss}_{\text{initial}} = -\text{Score}_t + \omega \times \log \sum_{k=1}^{N} \exp(\text{Score}_k), \tag{9}$$

where Score represents the confidence score feedback from the speaker verification system between x' and the speaker, t represents the target speaker number, k represents the k^{th} speaker number, and ω represents the reciprocal of the speaker quantity. This loss function will expand the distance between x' and the nontarget identities and shrink the distance between x' and the target identity, while the training process will not rely on special θ.

5.4. Multifactor Based Attack. In our experiment, the generalized relevancy based loss function, which is elaborated in the last section, can guide the generator to generate adversarial examples that can attack the speaker verification system successfully. However, the distortion of adversarial examples will be beyond our tolerance; people will hear obvious noise in adversarial samples and find the attack easily. They are also easily recognized by machines. To this end, we designed a stronger generator with another loss function designed by us, which can achieve the spoofing attack imperceptibly. Hence, we divide our loss function into two parts: one is to achieve the attack and the other is to limit the amplitude perturbation. In particular, we limit the distortion with a special design in the loss function. Note that because the first goal of our generator is to generate a new utterance that can spoof the speaker verification system, we only utilize the part which can limit the noise in the loss function, after the new utterance x' can be verified as the target identity. We utilized the multifactor based loss function to realize spoofing attack. We found that the distortion of successful adversarial examples is much less than before, while the total success rate is close to the result reported by generalized relevancy loss function based attack. We designed this loss function as follows:

$$\text{Loss} = (1-L) \times \text{Loss}_{\text{initial}} + L \times \frac{A}{\log(x) - \log(\delta)}, \tag{10}$$

where L equals 0 when x' was rejected by the speaker verification system; otherwise, L equals 0, and A is a constant. We designed our generator with this loss function. When we initially train the generator, L is set as 0; it will try its best to achieve attack first when x' has been accepted by the speaker verification system; L will be reset as 1; then the generator begins to reduce the distortion, which may cause x' to be rejected by the speaker verification system; then L will be reset as 0. Two different requirements will compete through the whole training process. Thus, the generator can find the slightest perturbation inject into the original utterance to realize the attack and we will get adversarial examples that cannot be recognized by humans or machines.

6. Experiment Setup

6.1. Data Partition and Experimental Environment. In this section, we will describe the design of our experiments. To prove the efficiency of our attacks, we examined our spoofing attack with an open dataset TIMIT which has been used in many other voice-related works [26, 27]. This dataset includes 630 speakers and 6300 sentences. Firstly, we utilized 462 speakers from the TIMIT training set to train the embedding network which is based on the GE2E loss function. Then we extract the embedding vector from the testing set by the embedding network for further speaker verification. We deploy our attacks locally and randomly select 4 illegal speakers and 5 legal speakers from the testing set. For each illegal speaker, we selected 5 sentences and acquired 20 (4×5) sentences; then we trained our generator to produce perturbation for each sentence targeting 5 legal speakers, respectively. Through this process, we obtained 100 (5×20) adversarial examples for spoofing attacks. Then we test these adversarial examples on the speaker verification system. Two different classifications were employed in our work to show the performance of spoofing attack on machine learning solutions and similarity threshold solutions; we also test two loss functions that were introduced in Section 5 to show the performance of different spoofing attack. We conducted the experiments on a server with Ubuntu 16.04 and Intel Xeon CPU E5-2678 v3 @ 2.50 GHz with 125 G RAM. We set the learning rate η as $1e-2$, A as 20, and the epoch as 500.

6.2. Metrics. We employed different metrics for evaluating the above results. For the verification part, we employed false acceptance rate (FAR), false rejection rate (FRR), and average classification error (ACE). They were defined with the following equations:

$$\text{FAR} = \frac{\text{FP}}{\text{FP} + \text{TN}},$$
$$\text{FRR} = \frac{\text{FN}}{\text{TP} + \text{FN}}, \quad (11)$$
$$\text{ACE} = \frac{\text{FAR} + \text{FRR}}{2},$$

where TP represents the amount of correctly classified positive samples, TN represents the amount of correctly classified negative samples, FP represents the amount of incorrectly classified positive samples, and FN represents the amount of incorrectly classified negative samples. Except that, we also employ equal error rate (EER) as metrics which is the error rate when FAR equals FRR. Then, for attack part, we utilized success rate (SR) as the metric, and it is defined as follows:

$$\text{SR} = \frac{\text{success\_attack\_count}}{\text{attack\_count}}. \quad (12)$$

These metrics are widely utilized in evaluating performance of verification.

It is easy to evaluate the performance of classification and verification, but it is difficult to evaluate distortion directly. We need a quantitative method to calculate the distortion. We evaluate an utterance in decibels (dB) and a universal description of relative volume. The following equations represent the MNR of the adversarial samples:

$$\text{dB}(x) = \max_i 20 \cdot \log_{10}(x_i), \quad (13)$$

$$\text{dB}_x(\delta) = \text{dB}(\delta) - \text{dB}(x), \quad (14)$$

where δ represents the perturbation we inject into the utterance. Equation (14) represents the MNR of the attack utterance [16]. We can also write the equation about MNR as follows:

$$\text{MNR} = \text{dB}_x(\delta) = \text{dB}(\delta) - \text{dB}(x). \quad (15)$$

Since the perturbation is smaller than the original utterance, the result will be a negative number. The result is smaller, and the distortion is tinier. Although this result can describe the maximum distortion well, it cannot tell the overall distortion. So, we also introduced another metric SNR which is used in previous work [25]; it can be defined by the following equations:

$$\text{SNR} = 10 \log_{10}\left(\frac{P_x}{P_\delta}\right), \quad (16)$$

where P_x is the signal power of the original utterance and P_δ is the signal power of the injected perturbation. When SNR is large enough, the human nearly cannot hear the noise in the utterance.

Our work utilized the above two works as our evaluation metrics for evaluating distortion. They, respectively, represented the maximum distortion which shows waveform alert in detail and the total influence of the distortion which shows waveform alert in total.

7. Evaluation

In this section, we will evaluate our spoofing attack on the state-of-the-art TI-SV based on deep learning.

7.1. Performance without Attack. We need to study the performance of our speaker verification system on the TIMIT dataset which can prove that the spoofing result under attack is caused by our spoofing attack rather than the system's poor performance. Thus, we first run an evaluation to prove that the speaker verification system can verify the identities of users before we evaluate the performance of our spoofing attack. We train the embedding vector extractor by the training data from the TIMIT dataset. After training, we randomly select 100 people for evaluating the performance of the speaker verification system. When the LDA was used, FAR = 5%, FRR = 5%, and ACE = 5%. Figure 2 shows the ROC curve when it uses cosine similarity. The EER equals 5% and the area under curve (AUC) can reach 0.99. Meanwhile, we found that when we set the threshold value at 0.59 the TI-SV has the best performance through the experiments. These results show whichever classification is used by the TI-SV, and the identities of users can be correctly verified. Then we can examine our spoofing attack on the TI-SV based on these results.

7.2. Performance with Spoofing. In this section, we will first evaluate the distortion of our generator, which can generate adversarial examples with different loss functions. Firstly, we produce adversarial samples by cosine similarity score and randomly show three waveforms from the same utterance; the result is shown in Figure 3; the yellow waveform is generated by generalized relevancy based attack, the red waveform is generated by multifactor based attack, and the black waveform is the original waveform. We can observe that the waveform generated by our multifactor based attack is more similar to the original waveform than the waveform generated by our generalized relevancy based attack. These waveforms show, on the intuitive, that the multifactor based attack will have better performance than generalized relevancy based attack on imperceptibility. Beyond that, we need to describe the distortion on a quantitative level. Firstly, a distribution diagram for MNR in Figure 4 is shown, and the ordinate value represents the quantity in this range. The average MNR for our multifactor based attack is −33 dB, and it is −18 dB for generalized relevancy based attack. The best-reported MNR for our multifactor based attack and generalized relevancy based attack is −77 dB and −22 dB, respectively. Our multifactor based attack is more imperceptible than generalized relevancy based attack in this metric. Besides, our best performance is also better than −45 dB, which was reported by previous work [16]. The MNR describes the distortion under the maximum scene, and we still need to describe the distortion under the global scene. Thus, a comparison of SNR was made between two attacks.

Figure 5 shows the distribution of SNR, and the ordinate value represents the quantity in this range. The average SNR

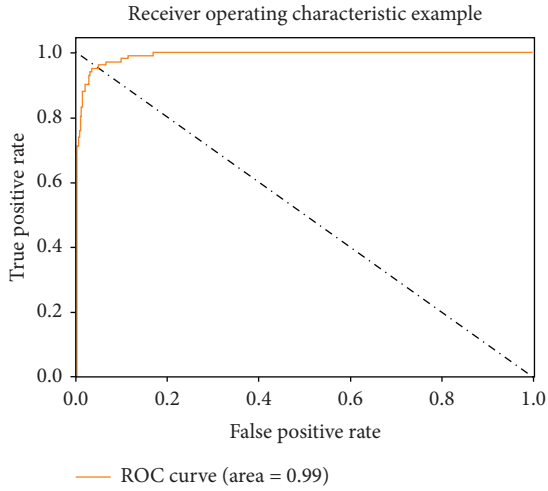

Figure 2: The ROC curve of the GE2E based TI-SV.

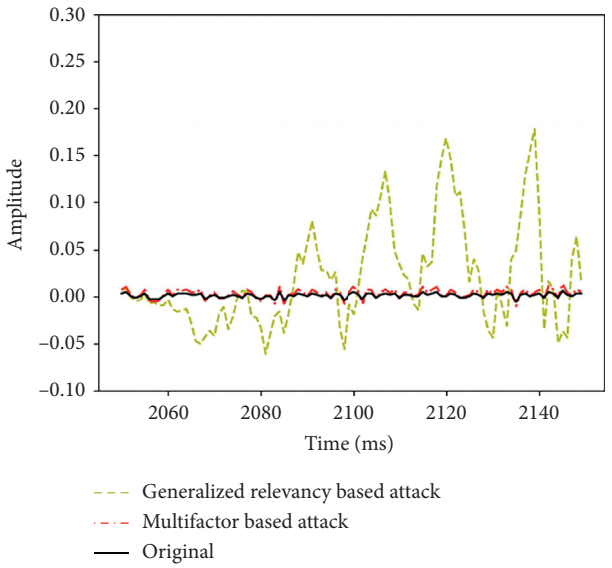

Figure 3: The different waveforms from the same utterance: (a) the yellow waveform is generated by generalized relevancy based attack, (b) the red waveform is generated by multifactor based attack, and (c) the black waveform is the original waveform.

for our multifactor based attack is 31 dB, and it is 17 dB for generalized relevancy based attack; meanwhile, the best result of SNR for multifactor based attack is 76 dB, which is larger than 26 dB that was reported by generalized relevancy based attack. The best performance is better than 31 dB in the previous work too. The above results can prove that our multifactor based spoofing attack will have better performance compared with previous work in terms of imperceptibility. In other words, our attack owns higher imperceptibility, which is a key feature for adversarial audio samples. After evaluating the imperceptibility of generalized relevancy based attack and multifactor based attack, we need to evaluate the performance of the SR in different classifications.

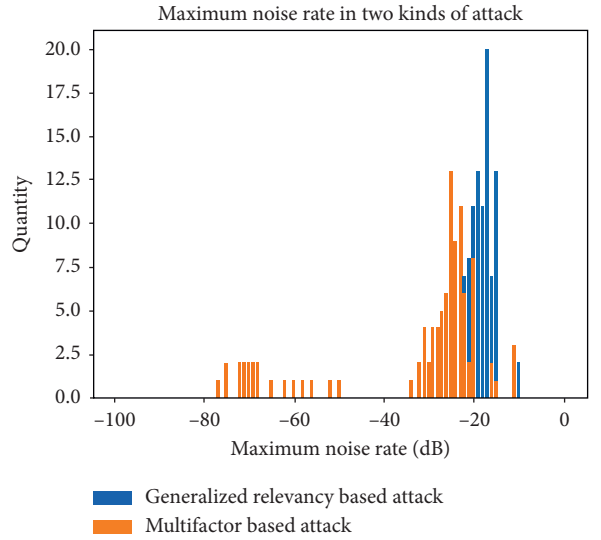

Figure 4: Maximum noise rate for generalized relevancy based attack and multifactor based attack. The ordinate value represents the quantity in this range.

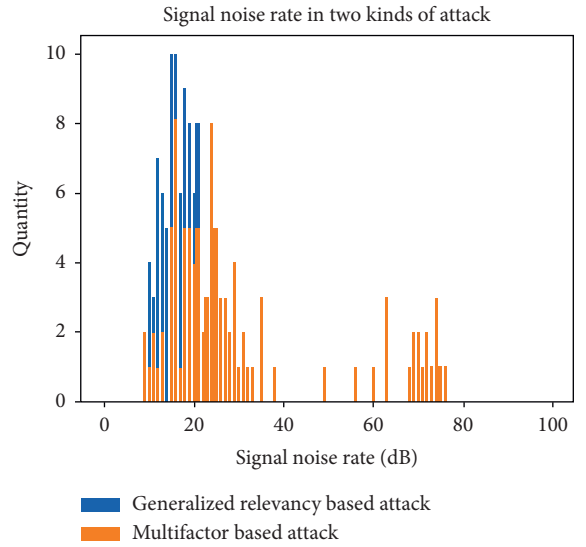

Figure 5: Signal noise rate for generalized relevancy based attack and multifactor based attack. The ordinate value represents the quantity in this range.

7.2.1. Linear Discriminate Analysis. We first tried to spoof the TI-SV, which uses the LDA to verify the identity of the user. We use the enrollments to train a two-class model used to distinguish the target speaker and nontarget speaker, which can realize a speaker verification task. The result is shown in SR1 of Table 1. Through the result, the SR of generalized relevancy based attack can achieve up to 83%. Meanwhile, the SR of multifactor based attack achieved up to 80%. This result proved that the two spoofing attacks can spoof the TI-SV with a similar SR when using LDA to achieve speaker verification.

7.2.2. Cosine Similarity. Since LDA and other machine learning classifications need training before using, which is

Table 1: The SR of spoofing attack. SR1 aims at LDA classification and SR2 aims at cosine similarity threshold classification. C1 represents the generalized relevancy based attack and C2 represents the multifactor based attack.

	C1	C2
SR1 (%)	83	80
SR2 (%)	86	82

inconvenient for enrolling a new user, as a more practical speaker verification system, the TI-SV can utilize cosine similarity and set a fixed threshold to verify the speaker's identity, which does not need training. We set the threshold value at 0.59 which has the best performance in Section 7.1. SR2 in Table 1 shows the results when we use generalized relevancy based attack and multifactor based attack. We can learn from the result that our multifactor based attack still has close performance to that of the generalized relevancy based attack. The SR for multifactor based attack is 82% and the SR for generalized relevancy based attack is 86%.

7.2.3. Summary. Through these results, we can learn that our multifactor based attack aimed at the state-of-the-art TI-SV has a high SR and it is more imperceptible, whatever the TI-SV model is based on machine learning classifications or threshold classifications. Our spoofing attack's SR is lower than that in the previous works; it is due to the fact that to our spoofing attack's target is the TI-SV, and it is more difficult to extract voiceprint than TD-SV that is targeted by previous works. Our multifactor based attack will have a slighter distortion with the original utterance, which is more imperceptible than the generalized relevancy based attack and previous works.

8. Attack Characteristics

The attack has two distinct characteristics: imperceptibility and target-independent ability. We will further analyze this part in the following paragraphs.

8.1. Imperceptibility. The imperceptibility will be changed continually during the procession of generation. It is important to analyze the procession of generation for showing the advantage of multifactor based attack on imperceptibility. We experiment for our attack to analyze the imperceptibility of the perturbation during the procession of generation. We randomly selected all sentences (20 sentences) in the illegal set and a random target in the legal set to observe the procession of generation. Figure 6 shows the average signal noise rate (dB) for each epoch. The result shows that the change of SNR is coincident between the multifactor based attack and the generalized relevancy based attack during the first phase of the procession. However, the multifactor based attack begins to increase the SNR, while the generalized relevancy based attack decreases the SNR persistently in the second phase. Note that the multifactor based attack not only stops the decreasing of the SNR but also explores the highest SNR for the successful attack. We can observe that the final SNR of multifactor based attack is larger than the end of the first phase and much larger than the final SNR of generalized relevancy based attack.

8.2. Target-Independent Ability. We consider that the voiceprint is only dependent on the biometric differences rather than the content. Thus, the factor of the physiological structure will determine the voiceprint. The previous work [28] has proven that the gender is an important factor for voice. Given that view, we analyze the target-independent ability by studying the influence of gender. We randomly selected ten sentences from different males and ten sentences from different females and partitioned them into two groups; each group includes five sentences from different males and five sentences from different females. We enrolled in the speaker verification system with users in one group. After that, we utilized sentences from females to attack the enrolled males; meanwhile, the same operation was done for sentences from males. Note that we use the cosine similarity for verification, since Section 7.1 has proven that it has the same efficacy as that of LDA. Table 2 shows the SR, average signal noise rate (ASNR), and average maximum noise rate (AMNR) after finishing the above experiment. The result shows that gender has little impact on the success rate and the perturbation of the attack. Our multifactor based attack has target-independent ability.

9. Discussion

In this section, we will discuss some advanced attack methods and defense methods for speaker verification systems.

9.1. Universal Perturbation. Current spoofing attacks need the generator to generate perturbations for each utterance, even targeting the same target. We hope our generator can produce a universal perturbation for a special target, which can use only one perturbation to realize the spoofing attack to a special victim. We can learn from some work that the voiceprint will exist in the tiny waveform [29], so it is possible to generate a stabilized perturbation that can include the whole voiceprint for one victim.

9.2. Over-the-Air Injection. Current spoofing attacks aimed at speaker verification systems can only work in the data layer or offer utterance for replaying attacks [25]. These restrictions limited the spoofing attack's range, which makes this attack unable to be utilized. An advanced attack could inject a slight perturbation when a legal speaker is verifying; the adversary can lead the speaker verification system into an incorrect permission space and all the legal user's operations in this space will be unsafe. Some attack in computer vision can spoof classifications by only changing one pixel's content [30]. If we can realize the spoofing attack by only changing one or several points in the utterance, we will have the ability to repeat playing the short perturbation over the air to attack the system when a legal speaker is verifying. Even more, we

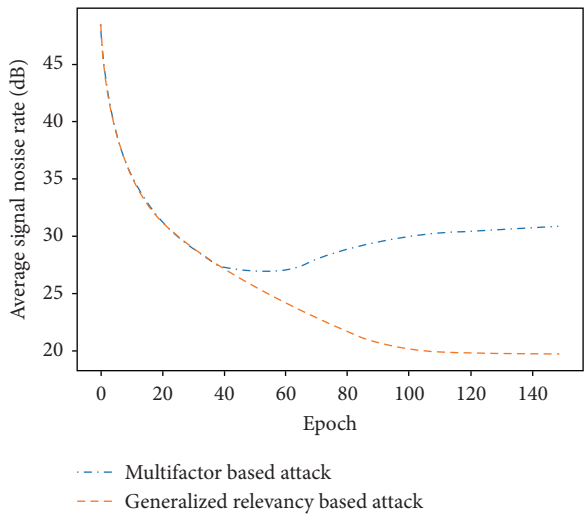

Figure 6: Signal noise rate for generalized relevancy based attack and multifactor based attack during the procession of generation.

Table 2: The successful rate, signal noise rate, and maximum noise of the analysis of the gender. "Male" and "female" represent the genders of the sentence's owner that is utilized to attack.

	Male	Female
SR (%)	88	84
SNR (dB)	38	30
MNR (dB)	−35	−31

can use ultrasound to complete injection instead of audible sound by the technology proposed in previous work [25]. It will further enhance the imperceptibility of the spoofing attack, which makes the attack more practical.

9.3. Mitigation of Multifactor Based Attack. Since our attack is effective, while it is hard to be detected by human or current speaker verification systems, we will discuss several possible defense methods as follows: A detection method is to train a detector using normal utterances and adversarial utterances. In the computer vision domain, they can employ a detector to detect the adversarial samples [31], although this work has a high false positive rate, and it is not robust when the adversary is aware of this defense. It also has the ability to distinguish between normal utterances and adversarial utterances.

Another method is adding transformation for the input (e.g., bit-depth reduction and JPEG compression [32] for images). We can mitigate the attack by applying input transformation such as a bit-depth reduction. Because this process will reduce the information in an utterance including injected perturbation, our attack will not succeed.

10. Conclusion

In this paper, we explore the vulnerability of the speaker verification system which will affect the security of users' economics, privacy, and even safety. We first conduct imperceptible audio adversarial examples to attack the state-of-the-art deep-learning-based TI-SV by our generalized relevancy based attack and multifactor based attack. We evaluate generalized relevancy based attack and multifactor based attack in two patterns to verify the speaker including both machine learning method and threshold method. The multifactor based attack can achieve 82% SR, when the distortion in the utterance only has MNR −77 dB and SNR 76 dB, which are much better than the values in the previous works. Our work also gives out several advanced attacks with a theoretical foundation, which will influence real life a lot. Due to the fact that our effective attack reveals the vulnerability of the speaker verification system, we also proposed several defense methods to mitigate the insecure problems of speaker verification systems.

Data Availability

The voice data used to support the findings of this study have been deposited in the TIMIT repository (https://catalog.ldc.upenn.edu/LDC93S1).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 61972348, in part by the National Key Research and Development Program of China under Grant 2018YFB0803600, and in part by the Leading Innovative and Entrepreneur Team Introduction Program of Zhejiang under Grant 2018R01005.

References

[1] Q. Zheng, A. Kumar, and G. Pan, "Contactless 3d fingerprint identification without 3d reconstruction," in *Proceedings of the 2018 International Workshop on Biometrics and Forensics, IWBF 2018*, pp. 1–6, IEEE, Sassari, Italy, June 2018.

[2] X. Wei, H. Wang, B. Scotney, and H. Wan, "Selective multi-descriptor fusion for face identification," *International Journal of Machine Learning and Cybernetics*, vol. 10, no. 12, pp. 3417–3429, 2019.

[3] Y. Chen, C. Wu, and Y. Wang, "T-center: a novel feature extraction approach towards large-scale iris recognition," *IEEE Access*, vol. 8, pp. 32365–32375, 2020.

[4] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018*, pp. 5329–5333, IEEE, Calgary, Canada, April 2018.

[5] L. Wan, Q. Wang, A. Papir, and I. Lopez-Moreno, "Generalized end-to-end loss for speaker verification," in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018*, pp. 4879–4883, IEEE, Calgary, Canada, April 2018.

[6] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016*, pp. 5115–5119, IEEE, Shanghai, China, March 2016.

[7] N. Dehak, L. J. Rodríguez-Fuentes, and E. Lleida, "I-vector representation based on GMM and DNN for audio classification," in *Proceedings of the Odyssey 2016: The Speaker and Language Recognition Workshop*, ISCA, Bilbao, Spain, June 2016, http://www.isca-speech.org/archive/Odyssey_2016/abstracts/Najim.html.

[8] D. A. Reynolds, "Universal background models," in *Encyclopedia of Biometrics*, S. Z. Li and A. K. Jain, Eds., Springer US, New York, NY, USA, pp. 1349–1352, 2009.

[9] Q. Li, H. Zhu, Z. Zhang, R. Lu, F. Wang, and H. Li, "Spoofing attacks on speaker verification systems based generated voice using genetic algorithm," in *Proceedings of the 2019 IEEE International Conference on Communications, ICC 2019*, pp. 1–6, IEEE, Shanghai, China, May 2019.

[10] Z. Wu and H. Li, "Voice conversion and spoofing Attack on speaker verification systems," in *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2013*, pp. 1–9, IEEE, Kaohsiung, Taiwan, November 2013.

[11] E. Variani, X. Lei, E. McDermott, I. Lopez-Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014*, pp. 4052–4056, IEEE, Florence, Italy, May 2014.

[12] M. Zha, G. Meng, C. Lin, Z. Zhou, and K. Chen, "Rolma: a practical adversarial attack against deep learning-based LPR systems," in *Proceedings of the Information Security and Cryptology-15th International Conference, Inscrypt 2019*, Nanjing, China, December 2019.

[13] M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar, "Can machine learning be secure?," in *Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security, ASIACCS 2006*, F. Lin, D. Lee, B. P. Lin, S. Shieh, and S. Jajodia, Eds., , March 2006.

[14] B. Biggio, I. Corona, D. Maiorca et al., "Evasion attacks against machine learning at test time," in *Proceedings of the Machine Learning and Knowledge Discovery in Databases-European Conference, ECML PKDD 2013*, H. Blockeel, K. Kersting, S. Nijssen, and F. Zelezný, Eds., pp. 387–402, Springer, Prague, Czech Republic, September 2013.

[15] A. Graves, S. Fernández, F. J. Gomez, and J. Schmidhuber, Cohen, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the Machine Learning, Twenty-Third International Conference (ICML 2006)*, A. W. Moore, Ed., pp. 369–376, ACM, Pittsburgh, Pennsylvania, USA, June 2006.

[16] N. Carlini and D. A. Wagner, "Audio adversarial examples: targeted attacks on speech-to-text," in *Proceedings of the 2018 IEEE Security and Privacy Workshops, SP Workshops 2018*, pp. 1–7, IEEE Computer Society, San Francisco, CA, USA, May 2018.

[17] X. Liu, K. Wan, Y. Ding, X. Zhang, and Q. Zhu, "Weighted-sampling audio adversarial example attack," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, pp. 4908–4915, 2020.

[18] Y. Qin, N. Carlini, G. Cottrell, I. Goodfellow, and C. Raffel, "Imperceptible, robust, and targeted adversarial examples for automatic speech recognition," in *Proceedings of the International Conference on Machine Learning. PMLR*, pp. 5231–5240, Long Beach, CA, USA, June 2019.

[19] H. Fujimura, N. Ding, D. Hayakawa, and T. Kagoshima, "Simultaneous flexible keyword detection and text-dependent speaker recognition for low-resource devices,," in *Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods, ICPRAM 2020*, M. D. Marsico, G. S. di Baja, and A. L. N. Fred, Eds., pp. 297–307, Scite Press, Valletta, Malta, February 2020.

[20] W. Wang, Y. Zhang, J. Xu, and Y. Yan, "Multiple temporal scales based speaker embeddings learning for text-dependent speaker recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019*, pp. 6311–6315, IEEE, Brighton, UK, May 2019.

[21] R. Jahangir, Y. W. Teh, N. A. Memon et al., "Text-independent speaker identification through feature fusion and deep neural network," *IEEE Access*, vol. 8, pp. 32187–32202, 2020.

[22] F. Zhao, H. Li, and X. Zhang, "A robust text-independent speaker verification method based on speech separation and deep speaker," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019*, pp. 6101–6105, IEEE, Brighton, UK, May 2019.

[23] Q. Wang, X. Lin, M. Zhou et al., "Voicepop: a pop noise based anti-spoofing system for voice authentication on smartphones," in *Proceedings of the IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pp. 2062–2070, IEEE, Paris, France, May 2019.

[24] 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 *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, ser. CCS '19*, pp. 1215–1229, Association for Computing Machinery, New York, NY, USA.

[25] G. Chen, S. Chen, L. Fan et al., "Who is real bob? adversarial attacks on speaker recognition systems," in *Proceedings of the 2020 IEEE Symposium on Security and Privacy (SP)*, IEEE, San Francisco, CA, USA, May 2021.

[26] A. Bhowmick, A. Biswas, and M. Chandra, "Performance evaluation of psycho-acoustically motivated front-end compensator for TIMIT phone recognition," *Pattern Analysis and Applications*, vol. 23, no. 2, pp. 527–539, 2020.

[27] M. T. S. Al-Kaltakchi, W. L. Woo, S. S. Dlay, and J. A. Chambers, "Comparison of i-vector and GMM-UBM approaches to speaker identification with TIMIT and NIST 2008 databases in challenging environments," in *Proceedings of the 25th European Signal Processing Conference, EUSIPCO 2017*, pp. 533–537, IEEE, Kos, Greece, September 2017.

[28] Y. Michalevsky, D. Boneh, and G. Nakibly, "Gyrophone: recognizing speech from gyroscope signals," in *Proceedings of the 23rd USENIX Security Symposium*, K. Fu and J. Jung, Eds., pp. 1053–1067, USENIX Association, San Diego, CA, USA, August 2014, https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/michalevsky.

[29] Y. Wang, Y. Wang, and T. Tan, "Combining fingerprint and voiceprint biometrics for identity verification: an experimental comparison," in *Proceedings of the Biometric Authentication, First International Conference, ICBA 2004*, D. Zhang and A. K. Jain, Eds., pp. 663–670, Springer, Hong Kong, China, July 2004.

[30] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 5, pp. 828–841, 2019.

[31] Q. Guo, J. Ye, Y. Hu et al., "A multivariant partition-based method for audio adversarial examples detection," *IEEE Access*, vol. 8, pp. 63 368–463 380, 2020.

[32] S. Z. Li and A. K. Jain, Eds., *Encyclopedia of Biometrics*p. 875, Springer US, New York, NY, USA, 2009.