Practical Adversarial Attacks Against Speaker Recognition Systems

Zhuohang Li[†], Cong Shi[‡], Yi Xie[‡], Jian Liu[†], Bo Yuan[‡], Yingying Chen[‡]

[†]The University of Tennessee, Knoxville, TN, USA, 37996
[‡]Rutgers University, New Brunswick, NJ, USA, 08901
zli96@vols.utk.edu, cs1421@winlab.rutgers.edu, yx238@scarletmail.rutgers.edu, jliu@utk.edu, bo.yuan@soe.rutgers.edu, yingche@scarletmail.rutgers.edu

ABSTRACT

Unlike other biometric-based user identification methods (e.g., fingerprint and iris), speaker recognition systems can identify individuals relying on their unique voice biometrics without requiring users to be physically present. Therefore, speaker recognition systems have been becoming increasingly popular recently in various domains, such as remote access control, banking services and criminal investigation. In this paper, we study the vulnerability of this kind of systems by launching a practical and systematic adversarial attack against *X-vector*, the state-of-the-art deep neural network (DNN) based speaker recognition system. In particular, by adding a well-crafted inconspicuous noise to the original audio, our attack can fool the speaker recognition system to make false predictions and even force the audio to be recognized as any adversary-desired speaker. Moreover, our attack integrates the estimated room impulse response (RIR) into the adversarial example training process toward practical audio adversarial examples which could remain effective while being played over the air in the physical world. Extensive experiment using a public dataset of 109 speakers shows the effectiveness of our attack with a high attack success rate for both digital attack (98%) and practical over-the-air attack (50%).

1 INTRODUCTION

Research interest in voice controllable system (VCS) has grown considerably in recent years as the system provides a convenient way of meeting a user's various daily needs through voice commands. With such a convenient and ubiquitous speech interface, speaker recognition system (a.k.a. voice recognition system), identifying the voice of a given utterance among a set of enrolled speakers, could be seamlessly integrated and thus be used to facilitate a series of security-enhanced voice-based applications. For instance, new features in recent Apple Siri can reliably recognize voices, enabling HomePod to respond to requests from multiple users in shared spaces. Some companies (e.g., Voice Biometrics Group [7]) use speaker recognition technologies to let people gain access to information or give authorization without being physically present. Chase Voice ID [2] exploits remote voice authentication to quickly verify users and prevent fraud when they call the bank's customer service center. It has shown a critical need for deploying voiceprint recognition systems on the top of many existing applications.

Existing studies have demonstrated that applying deep neural networks (DNN) to speaker recognition task has great advantages in terms of its highly scalable embedding performance and the ability of coping with practical interference (e.g., noises and reverberation). For instance, DNN posteriors have been used to derive sufficient statistics for alternative i-vectors calculation allowing to discriminate speakers at triphone level [10]. The researchers also showed that DNN-based solutions could lead to a significant improvement over current state-of-the-art solutions, such as conventional universal background model-Gaussian mixture model (UBM-GMM), on telephone speech [12]. However, DNN model has been shown a serious vulnerability when facing intentionally distorted inputs, such as *adversarial examples*, which could fool the model to make false classifications. Thus, DNN-based speaker recognition systems would be inevitably threatened by the adversarial examples.

In the audio space, Carlini *et al.* [3] recently demonstrated that it is possible to modify the translation output of automatic speech recognition (ASR) system to any adversary-desired text by adding an inconspicuous perturbation to the original voice command. Moreover, CommanderSong [21] can embed any malicious command into regular songs, making people hear them as common music whereas the ASR systems would recognize them as malicious commands. Different from ASR system that mainly focuses on speech-to-text translation, speaker recognition model utilizes embedding methods to extract features that represent voice similarities to distinguish speakers regardless of their speech content. To the best our knowledge, the only existing adversarial attack [9] against speaker recognition system targets for an end-to-end speaker verification model, which is a binary speaker recognition system that gives either accept or reject by verifying whether the voice is uttered by a claimed speaker. By adding an inconspicuous perturbation to the original voice, the model might produce incorrect outputs such as rejecting a legitimate user or vice versa. Additionally, such an attack only considers digital scenarios, in which the generated adversarial sample is directly fed into the speaker recognition system without being played through a speaker.

In this paper, we explore the possibility of conducting an over-the-air adversarial attack in practical scenarios, in which the adversarial examples are played through a loudspeaker to compromise speaker recognition devices. The target multi-class speaker recognition model is built upon the state-of-the-art DNN-based X-vector [16] with 109 speakers in our testing model. We show that by adding an inconspicuous perturbation into the original audio, our attack can deceive the speaker recognition system causing a false prediction. In practice, there are a few challenges we face in launching such an attack: (1) Unlike binary speaker verification system, attacking a multi-class speaker recognition model requires more sophisticated adversarial learning processes to make the adversarial examples to be classified as the adversary-desired speaker; (2) To make the adversarial examples remain effective while being played over the air, the added perturbation needs to be robust enough to survive real-world distortions caused by different audio propagation channels (e.g., multi-path effect), ambient noises sources, and speaker & microphone limitations; and (3) The distortion between the generated adversarial example and the original speech should be as small as possible, making the example stealthy and unnoticeable to human.

In particular, we applied gradient-based adversarial machine learning algorithms to generate adversarial examples for two types of representative attacks: *(1) Untargeted attack* that aims to disable the speaker recognition system by making the audio signals misclassified as incorrect speakers; and *(2) Targeted attack* that is designed to change the classification result to an adversary-desired speaker, which will enable the adversary to pass the authentication with fraudulent identities. To generate untargeted adversarial examples, we adapt fast gradient sign method (FGSM) [6] to compute the adversarial perturbation by taking the derivative of the cross-entropy loss between the output probability distribution and its true label. Whereas the targeted adversarial examples are computed through solving an optimization process to minimaize the cost function of the noise level as well as the distance to the targeted class. In order to effectively launch an over-the-air adversarial attack to compromise speaker recognition devices, we estimate the room impulse response (RIR) and consider the inferred distortions of the recorded sound at microphone end in the adversarial learning phase. Our main contributions are summarized as follows.

- To the best of our knowledge, this is the first work to build a practical and systematic adversarial attack, including both untargeted and targeted attacks, against multi-class speaker recognition system.
- We propose to use the estimated RIR that reflects acoustic channel state information to generate practical audio adversarial examples that can still remain effective while being played over the air in a realistic environment.
- We implement gradient-based adversarial learning algorithms to make the generated adversarial examples unnoticeable to humans and effectively compromise speaker recognition systems.
- Extensive experiments, including both digital attack and practical over-the-air attack, are conducted using a public dataset of 109 English speakers. The results show that our attack achieves a high attack success rate of over 98% for digital attack and 50% for over-the-air attack.

2 RELATED WORK

Attacks on Speech Recognition Systems. Recent studies have demonstrated the potential of spoofing automatic speech recognition (ASR) systems with adversarial attacks. As an initial study, Carlini and Wagner [3] develop an adversarial attack against DNN-based ASR system. By adding an imperceptible perturbation, an audio waveform could be transcribed as any desired target phrase. However, the adversarial example, i.e., an audio waveform combined with a perturbation, would lose its effectiveness after being played over-the-air. This problem is further investigated in a later work [14], which achieves over-the-air attack at simulated environments. In addition, CommanderSong [21] develops a more practical over-the-air attack which stealthily embeds voice commands into songs through adversarial learning. The songs could then be played from a remote loudspeaker to launch attacks. These attacks target only speech recognition models, while few studies explore the vulnerability of speaker recognition systems to the adversarial examples.

Attacks on Speaker Recognition Systems. Traditional attacks against speaker recognition systems could be broadly categorized as replay attack [18], speech synthesis attack [11], impersonation attack [1], and voice conversion attack [8]. In particular, replay attack [18] uses pre-recorded voice samples of the user to spoof the speaker recognition system. Such an attack, however, is less effective in many practical scenarios, such as call center where a speaker uses his voice for both authentication and interaction. The inconsistency of the voice could alert the staff and thwart the actual attack (e.g., request bank transfer). To avoid such inconsistencies, speech synthesis attacks [11] generate a victim's voice by learning an acoustic model from a limited set of voice samples. However, human and synthetic speeches could be differentiated with higher order Mel-cepstral coefficient [4],

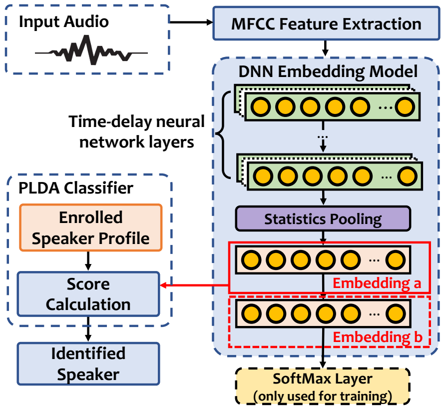


Figure 1: Architecture of X-vector system.

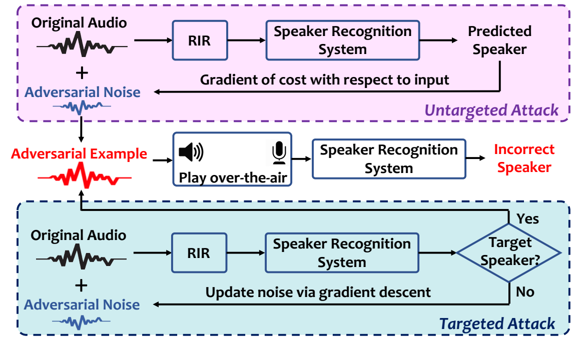


Figure 2: Attack overview.

or deep neural netowrk [20], rendering such attacks ineffective. Furthermore, an adversary could imitate a victim's voice through impersonation/voice conversion [1, 8] by manipulating existing voice samples from other users. However, these attacks could be defended by PLDA and Factor Analysis (FA) techniques which are usually integrated in state-of-the-art DNN-based speaker recognition models [19]. In comparison, our adversarial attack could circumvent the defense mechanisms while stealthily altering the DNN model's outputs without introducing noticeable distortion of speech. The most related study [9] develops an adversarial attack against an end-to-end DNN-based speaker verification model. This work, however, only considers a simple binary classification problem in the digital field without playing the adversarial examples over the air. Differently, in this paper we are the first to explore the vulnerability of state-of-the-art multi-class speaker recognition systems by developing a practical over-the-air adversarial attack.

3 SPEAKER RECOGNITION SYSTEM & THREAT MODEL

3.1 Target Speaker Recognition System

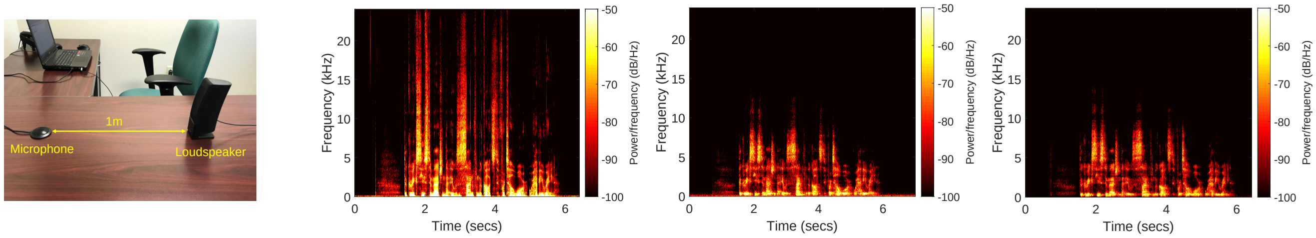
In this work, we choose X-vector architecture [16] as the target speaker recognition system as it shows a superior performance comparing to traditional i-vector models, and has been used as baseline in several follow-up studies [15]. As illustrated in Figure 1, the X-vector system first takes an input audio and divided into frames and extract mel frequency cepstral coefficents (MFCCs) features. The extracted features are then fed into a time-delay neural network (TDNN). At each layer, TDNN computes the activation of frames at the current and neighboring time steps. Subsequently, a statistics pooling layer aggregates the input segment by taking the mean and standard deviation of the output from the last frame-level layer. In addition, two fully-connected layers are used to map the concatenated statistics into embeddings. Finally, the PLDA classifier computes the probability of the voice belonging to each speaker in the enrollment set by comparing the similarity between the *embedding a* taken from the second last hidden layer and the enrolled speaker profile. A prediction is made by choosing the speaker with the highest calculated probability.

Attack Chance on DNN Model. During the training phase, the DNN model leverages gradient descent algorithms to learn the mapping between the input space and the embedding space. However, if the trained model architecture and parameters are known, the gradient information will remain trackable. Therefore, with the direction of the gradient, it is possible for an adversary to add a well-crafted perturbation to the input and consequently manipulate the output prediction. Moreover, by taking advantage of gradient-based algorithms, the computed perturbation can be so subtle as to be unnoticeable to human.

3.2 Threat Model

In this work, we assume a white-box setting, where the adversary has complete knowledge to the speaker recognition model. The adversary also has access to the physical environment where the actual attack will be launched, and is capable of utilizing a speaker and microphone to measure the room impulse response (RIR) to launch over-the-air attacks. Specifically, the flow of our attack is shown in Figure 2, where the adversarial examples are generated from two types of attacks, and played over-the-air to deceive the speaker recognition system at the microphone side. To make the adversarial examples remain effective while being played in the air, we use the estimated RIR to model the sound distortion when generating adversarial examples. The detailed generation flow is presented as follows:

Untargeted Attack. The input audio signal first goes through the measured RIR to simulate the played over-the-air process. The speaker recognition system then takes the distorted signal and makes a prediction. The untargeted adversarial noise is computed by taking the gradient of the cost function (i.e. multi-class cross entropy) with respect to the input audio. Finally, the adversarial example is constructed by adding the computed noise to the original audio.



(a) Experimental setup (b) Played voice (c) Recorded voice (d) Predicted voice (w/ RIR)

Figure 3: Preliminary experiment to verify the effectiveness of the estimated RIR.

Targeted Attack. The targeted attack works in an iterative way: First, the targeted adversarial noise is initialized with zeros. The original audio is then combined with the adversarial noise and modified according to the estimated RIR. A prediction is made by the speaker recognition system. If the prediction does not match the adversary-desired speaker, the noise will be modified according to the gradient descent's direction where the probability of the target class increases. If the prediction and the adversary-desired speaker are matched, the adversarial example is simply the sum of the original audio and the adversarial noise.

4 GENERATING PRACTICAL ADVERSARIAL EXAMPLES

4.1 Room Impulse Response Estimation

In order to infer the sound distortions during playback and recording in a practical speaker-microphone setting, the actual sound propagation can be quantitatively modeled as:

$$y(t) = K[x(t)] \otimes h'(t) + n(t), \quad (1)$$

where $x(t)$ is the played voice, $y(t)$ is the recorded voice, $K[\cdot]$ is the *N-th* order discrete-time Volterra kernel to represent a nonlinear memoryless system, which is usually used to model harmonic distortions caused by the nonlinearity of loudspeaker and microphone. The response $h'(t)$ characterizes the preposition of signals propagating through different paths with various delays, and $\otimes$ denotes the convolution process. Additionally, some noises $n(t)$ uncorrelated with the input signal are added to the output. Since it is difficult to separate the responses of $h'(t)$ from the $K[\cdot]$ in practice, the sound propagation is simplified as:

$$y(t) = x(t) \otimes h(t), \quad (2)$$

where $h(t)$ is a composite response to represent both linear and nonlinear characteristics and can be estimated using audio measurement techniques [5].

Specifically, we first play an excitation signal $x_e(t)$, where the signal frequency varies exponentially with time and play it through a loudspeaker. The signal allows each harmonic distortion at each order pack into a separate impulse response [5] and can be represented as:

$$x_e(t) = sin\Big(\frac{2\pi f_1 T}{ln(\frac{f_2}{f_1})}\big(e^{\frac{t}{T}ln(\frac{f_2}{f_1})} - 1\big)\Big), \quad (3)$$

where f_1, f_2 are the start and stop frequencies of sweeping, respectively, and T is the signal duration. In our RIR estimation, we set $f_1 = 20Hz$, $f_2 = 20kHz$ and $T = 5s$. When being played through a loudspeaker, $x_e(t)$ has a constant magnitude and is followed by a few seconds of silence to avoid sound aliasing caused by multi-path effects/harmonic distortions. With the response $y_e(t)$ recorded by the microphone, the room impulse response (RIR) $h(t)$ can be estimated by convolving it with an inverse filter: $h(t) = y_e(t) \otimes f(t)$, where the filter $f(t)$ is the time-reversal of $x_e(t)$.

A preliminary experiment is conducted to verify the effectiveness of the proposed RIR estimation. As shown in Figure 3(a), a loudspeaker and a microphone is placed on a table in a typical office environment. We use the loudspeaker to play a voice human speech sample and record it with the microphone. Figure 3(b) shows the spectrogram of the played voice sample, while Figure 3(c) and 3(d) are the spectrograms of the voice samples recorded by the microphone and predicted one using the estimated RIR, respectively. The mean square error (MSE) between the recorded and predicted voice samples is 0.112, while the error achieves 0.84 between the played and the recorded voice samples. The high similarity between the recorded and predicted voice samples demonstrate the effectiveness of our RIR estimation.

4.2 Adversarial Example Generation

The X-vector system showed in Figure 1 can be viewed as a function $f(\cdot)$, which takes as an input utterance X and outputs a probability vector $P = [p_1, ..., p_i]$, containing the predicted probability scores p_i for each speaker i. The untargeted and targeted adversarial examples can be generated in the following adversarial learning processes, respectively.

Untargeted Adversarial Example. It is constructed by adding a perturbation δ to the original input utterance. To be explicit, we write the adversarial example $X' = X + \delta$. Due to the local linearity of DNN models, a linear perturbation is sufficient to be constructed for an untargeted attack (referred as the fast gradient sign method (FGSM) [6]):

$$\delta = \epsilon sign\left(\nabla_X J(X, y)\right), \quad (4)$$

where $sign(\cdot)$ denotes the signum function, $J(X, y)$ represents the cost function between the input X and corresponding label y, and ϵ is a pre-choosen constant to control the attack

Table 1: Results of digital untargeted attack.

Attack Strength (i.e., ϵ)	No Attack	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}
Speaker Recognition Accuracy (%)	92.81	84.71	41.33	12.11	2.23	1.37
Attack Success Rate (%)	–	8.73	55.47	86.95	97.60	98.52
Average Distortion (dB)	–	−89.06	−69.15	−49.24	−29.33	−9.41

strength. We use the cross-entropy between the predicted probability vector and the true label as the cost function: $J(X, y) = -y \cdot log(P)$. For conventional digital attack, the untargeted adversarial example is generated by:

$$X' = X + \epsilon sign\Big(\nabla_X\big(-y \cdot log(f(X))\big)\Big). \tag{5}$$

To derive practical adversarial examples that remain effective while being played over-the-air, we model the air channel using the estimated RIR h. The input audio X is first convolved with the estimated h to simulate the over-the-air process, and the adversarial example can be generated as:

$$X' = X + \epsilon sign\Big(\nabla_X\big(-y \cdot log(f(X \otimes h))\big)\Big). \tag{6}$$

Targeted Adversarial Example. The adversarial example targeting at label y_t can be generated through solving an optimization problem:

$$\text{minimize } ||\delta||_2, \text{ s.t. } f(X + \delta) = y_t, \tag{7}$$

where $||\cdot||_2$ denotes the L_2 norm. As solving the direct non-linear constrained non-convex problem is difficult, in practice, we solve:

$$\text{minimize } -y_t \cdot log(f(X + \delta)) + c||\delta||_2, \tag{8}$$

where c is a pre-choosen constant which controls the attack strength. Specifically, the first term will be reduced when the predicted distribution is aligning the target label, and the second term penalizes the perturbation magnitude. Gradient descent is applied to find the optimal perturbation δ^*, and an targeted adversarial example is generated by $X' = X + \delta^*$. To be effective under practical settings, the optimization process (i.e., Equation 8) needs to be reformulated to include RIR:

$$\text{minimize } -y_t \cdot log\big(f\big((X + \delta) \otimes h\big)\big) + c||\delta||_2. \tag{9}$$

5 ATTACK EVALUATION

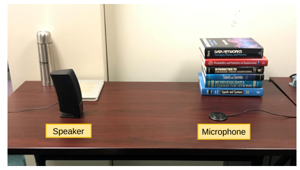
5.1 Experimental Methodology

Dataset and Basedline Model. We evaluate our attack on the dataset CSTR VCTK Corpus [17], which contains total 44217 utterances spoken by 109 English speakers with various accents. The dataset is splitted into training and testing sets with a ratio of 4 to 1. The MFCC features are 30 dimensional MFCCs and derived with a frame length of $25ms$. A pre-trained X-vector model provided in Kaldi [13] is used for embedding extraction. Regarding the computing environment, a NVIDIA DGX-1 server with 4×Tesla V100 GPU (32GB memory) is used.

Evaluation Metrics. (1) *Speaker Recognition Accuracy*: The percentage of utterances being correctly recognized by the baseline model. (2) *Attack Success Rate*: The ratio

Table 2: Results of digital targeted attack.

Attack Strength (i.e., c)	0.4	0.2	0.1	0.05
Attack Success Rate (%)	77.64	86.05	93.27	96.01
Average Distortion (dB)	−34.22	−32.43	−29.66	−25.94



(a) Office

(b) Apartment

Figure 4: Experimental setups for the practical over-the-air adversarial attack.

of the number of succeeded attacks to the total number of attack attempts. (3) *Distortion Metric*: We quantify the relative loudness of the introduced perturbation with respect to the original utterance in decibel: $D(\delta, X) = 20\ log_{10}\left(\frac{max(\delta)}{max(X)}\right)$.

Attack Settings. For the digital untargeted attack, the speaker recognition accuracy on clean dataset is recorded as benchmark and compared with that on the adversary-perturbed dataset. The digital targeted attack is evaluated by attempting to generate adversarial examples for every speaker to target at all other 108 speakers. The practical attack is tested in two real-world scenarios, as shown in Figure 4, where we play and record the adversarial examples generated by the digital and practical attack, respectively, and compare the attack success rate.

5.2 Evaluation of Digital Attacks

Table 1 shows the effectiveness of the untargeted attacks under different ϵ settings, where a larger ϵ value could enable stronger attack but lead to more significant distortions. Specifically, the baseline model could achieve 92.81% speaker recognition accuracy when no attack is present. Under untargeted attacks, the attack success rate grows with the attack strength and reaches 97.6% when $\epsilon = 10^{-2}$, where the baseline model could only correctly recognize 2.23% of the utterances.

Table 2 presents the results of our digital targeted attack. We report attack success only when the resulting speaker matches the desired targeted speaker. Similar to the observations in untargeted attack, the attack success rate increases with the attack strength (i.e., c). Specifically, when $c = 0.2$, our attack can achieve a 86.05% attack success rate while keeping the average distortion at −32.43 dB, which is approximately the difference between the ambient noise in a quite room and a person talking [3].

5.3 Evaluation of Practical Attacks

As shown in Figure 4, for each scenario, we use a loudspeaker to play 10 digital/practical adversarial examples and the voice samples are recorded by the microphone. An untargeted attack is reported success if a speaker is misclassified, while

Table 3: Attack success rate of practical over-the-air targeted attack.

	Playing digital adversarial examples	Playing practical adversarial examples
Office	0%	50%
Apartment	10%	50%

an targeted attack is considered as success only when the recorded voice sample is recognized as the targeted speaker. For untargeted attacks, both digital and practical attacks in the two environments can achieve 100% attack success rate. This is because the speaker recognition, even for the state-of-the-art X-vector, could be impacted by various environmental interferences (e.g., multipath, ambient noises) and thus mis-classified, which makes the untargeted attack less challenging. Table 3 shows the attack success rates of the targeted attacks through playing digital adversarial examples and practical ones which are generated by integrating RIR into the training process. We can observe that the practical attack can achieve a 50% attack success rate in both environments, while only one digital adversarial example succeed in spoofing the baseline model in the apartment. This is because the proposed RIR estimation could precisely characterizes the acoustic channel state information, making it possible to launch over-the-air adversarial attacks.

6 CONCLUSION

In this paper, we explore the vulnerability of speaker recognition systems with a practical and systematic adversarial attack against *X-vector*. We apply several gradient-based adversarial machine learning algorithms to generate digital adversarial examples for both untargeted and targeted attacks. In order to design more practical adversarial examples that remain effective when being played over the air, we integrate the estimated room impulse response into the adversarial example generation. Extensive experiment in both digital and real-world settings demonstrate the effectiveness of the proposed attacks.

REFERENCES

[1] Talal B Amin, James S German, and Pina Marziliano. 2013. Detecting voice disguise from speech variability: Analysis of three glottal and vocal tract measures. In *Proceedings of Meetings on Acoustics 166ASA*, Vol. 20.

[2] Chase Bank. 2019. Security as unique as your voice. (Oct 2019). https://www.chase.com/personal/voice-biometrics

[3] Nicholas Carlini and David Wagner. 2018. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)*. IEEE, 1–7.

[4] Lian-Wu Chen, Wu Guo, and Li-Rong Dai. 2010. Speaker verification against synthetic speech. In *2010 7th International Symposium on Chinese Spoken Language Processing*. IEEE, 309–312.

[5] Angelo Farina. 2000. Simultaneous measurement of impulse response and distortion with a swept-sine technique. In *Audio Engineering Society Convention 108*. Audio Engineering Society.

[6] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

[7] Voice Biometrics Group. 2019. Voice Biometrics Group. (Oct 2019). https://https://www.voicebiogroup.com/

[8] Tomi Kinnunen, Zhi-Zheng Wu, Kong Aik Lee, Filip Sedlak, Eng Siong Chng, and Haizhou Li. 2012. Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4401–4404.

[9] Felix Kreuk, Yossi Adi, Moustapha Cisse, and Joseph Keshet. 2018. Fooling end-to-end speaker verification with adversarial examples. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1962–1966.

[10] Yun Lei, Nicolas Scheffer, Luciana Ferrer, and Mitchell McLaren. 2014. A novel scheme for speaker recognition using a phonetically-aware deep neural network. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1695–1699.

[11] Johan Lindberg and Mats Blomberg. 1999. Vulnerability in speaker verification-a study of technical impostor techniques. In *Sixth European Conference on Speech Communication and Technology*.

[12] Mitchell McLaren, Yun Lei, and Luciana Ferrer. 2015. Advances in deep neural network approaches to speaker recognition. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 4814–4818.

[13] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The Kaldi Speech Recognition Toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.

[14] Yao Qin, Nicholas Carlini, Garrison Cottrell, Ian Goodfellow, and Colin Raffel. 2019. Imperceptible, Robust, and Targeted Adversarial Examples for Automatic Speech Recognition. In *International Conference on Machine Learning*. 5231–5240.

[15] Suwon Shon, Hao Tang, and James Glass. 2018. Frame-level speaker embeddings for text-independent speaker recognition and analysis of end-to-end model. In *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 1007–1013.

[16] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5329–5333.

[17] Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al. 2017. CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit. *University of Edinburgh. The Centre for Speech Technology Research (CSTR)* (2017).

[18] Zhizheng Wu, Sheng Gao, Eng Siong Cling, and Haizhou Li. 2014. A study on replay attack and anti-spoofing for text-dependent speaker verification. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*. IEEE, 1–5.

[19] Zhizheng Wu, Tomi Kinnunen, Eng Siong Chng, Haizhou Li, and Eliathamby Ambikairajah. 2012. A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case. In *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 1–5.

[20] Hong Yu, Zheng-Hua Tan, Yiming Zhang, Zhanyu Ma, and Jun Guo. 2017. DNN filter bank cepstral coefficients for spoofing detection. *Ieee Access* 5 (2017), 4779–4787.

[21] Xuejing Yuan, Yuxuan Chen, Yue Zhao, Yunhui Long, Xiaokang Liu, Kai Chen, Shengzhi Zhang, Heqing Huang, XiaoFeng Wang, and Carl A Gunter. 2018. Commandersong: A systematic approach for practical adversarial voice recognition. In *27th {USENIX} Security Symposium ({USENIX} Security 18)*. 49–64.