REAL-TIME, UNIVERSAL, AND ROBUST ADVERSARIAL ATTACKS AGAINST SPEAKER RECOGNITION SYSTEMS

Yi Xie[1]*, Cong Shi*[1]*, Zhuohang Li*[2]*, Jian Liu*[2]*, Yingying Chen*[1]*, Bo Yuan*[1]

[1]Rutgers University, New Brunswick, NJ, USA, 08901
[2]The University of Tennessee, Knoxville, TN, USA, 37996

ABSTRACT

As the popularity of voice user interface (VUI) exploded in recent years, speaker recognition system has emerged as an important medium of identifying a speaker in many security-required applications and services. In this paper, we propose the first real-time, universal, and robust adversarial attack against the state-of-the-art deep neural network (DNN) based speaker recognition system. Through adding an audio-agnostic universal perturbation on arbitrary enrolled speaker's voice input, the DNN-based speaker recognition system would identify the speaker as any target (i.e., adversary-desired) speaker label. In addition, we improve the robustness of our attack by modeling the sound distortions caused by the physical over-the-air propagation through estimating room impulse response (RIR). Experiment using a public dataset of 109 English speakers demonstrates the effectiveness and robustness of our proposed attack with a high attack success rate of over 90%. The attack launching time also achieves a $100\times$ speedup over contemporary non-universal attacks.

***Index Terms*—** speaker recognition systems, adversarial examples, universal adversarial attack

1. INTRODUCTION

In recent years, voice user interface (VUI) has been integrated into various platforms, such as smartphones and smart appliances, and is shaping up to become the hubs of our increasingly connected lives. With the prevalent usage of VUI, speaker recognition system, which identifies a person from characteristics of voices, could be seamlessly integrated and used for various security-enhanced applications, such as remote voice authentication to prevent fraud in financial services, voice-matched voice assistants that can only respond to the owner's voice, and even suspects identification and criminals detection [1, 2].

Deep network networks (DNNs), with its superiority over current state-of-the-art models (e.g., universal background model-Gaussian mixture model) [3, 4], has been becoming the computation core of the speaker recognition systems. However, recent studies have shown that DNN models are vulnerable to adversarial input in various fields (e.g., computer vision [5], natural language processing [6, 7] and speaker verification [8]). The most related work [8] generates adversarial examples against an end-to-end speaker verification model, which is a binary speaker recognition system that verifies whether the voice is uttered by a claimed speaker or not. However, the adversarial attack against a more complex multi-class speaker recognition model still remains unexplored. Moreover, this attack [8] is *individual attack* (i.e., non-universal) requiring to generate different perturbation for each voice input, which would cost considerable time training perturbations for each individual voice input and thus make real-time attacks impossible.

In this paper, we build the first *real-time*, *universal*, and *robust* targeted adversarial attack on X-vector [9], a state-of-the-art DNN-based multi-class speaker recognition model. The adversarial attack is performed by crafting an audio-agnostic universal perturbation which can be added into any enrolled speaker's any voice input to deceive the speaker recognition system, causing it to output an adversary-desired (targeted) speaker label. The generated universal perturbation uses repeated-playback of fixed-length universal noise to fit different voice input with various lengths. Additionally, unlike the existing digital attack [8] that feeds the adversarial examples to the speaker verification model directly, in this paper we take one step forward to build robust adversarial attacks through estimating the sound distortions introduced by the physical world propagation, which makes the adversarial examples remain effective while being played over-the-air. Experiments on a public dataset of 109 speakers show the effectiveness and robustness of our proposed attack with a high attack success rate of over 90%. The achieved attack launching time is only around $0.015s$, which is $100\times$ speedup over contemporary non-universal attacks.

2. RELATED WORK

Adversarial Attack on Speech Recognition. Recent studies have successfully produced adversarial examples against automatic speech recognition (ASR) system (i.e., speech-to-text), which is the most prevalent application in the audio space. For instance, Vaidya *et al.* [10, 7] generate noise-like adversarial sound making ASR models output adversary-desired text transcriptions. Nonetheless, the generated adversarial examples would be perceived as noises by human, which may draw considerable attention on practical attacks. To solve this problem, Carlini *et al.* [6] propose to craft adversarial samples by adding unnoticeable perturbations into orig-

inal speech, misleading the model to translate the adversarial examples to adversary-desired text. Moreover, CommanderSong [11] can embed any malicious command into regular songs, which could be recognized by ASR systems as malicious commands but still being perceived as common music by human. However, all the aforementioned ASR adversarial attacks are individual attack through solving an optimization problem for each individual input audio, which needs high run-time requirements (e.g., several hours) to compute the adversarial examples per input audio. Alternatively, a more recent work [12] produces a single universal perturbation which can fool ASR systems causing an error in transcription. This work is in the case of untargeted attack, in which the adversary cannot specify the expected speech transcription during the phase of adversary example generation.

Adversarial Attack on Speaker Recognition. Different from speech recognition systems, speaker recognition (a.k.a., voice recognition) mainly focuses on extracting individual-dependent voice characteristics through embedding methods to identify speakers' identities regardless of their speech content. It has been shown a growing trend of using DNNs in the embedding layers of speaker recognition model due to its superiority of scalable embedding performance [3, 4]. However, few studies have been conducted to explore the vulnerability of the DNN-based speaker recognition system. To the best of our knowledge, the only related study [8] proposes to build adversarial examples against an end-to-end speaker verification model, which is a binary speaker recognition system. Moreover, this attack is *individual attack*, which requires a long time to craft different perturbation for each voice input. It does not consider any sound distortions caused by practical over-the-air playback either. To bridge the gap in terms of all the aforementioned issues, in this paper we explore the possibility of launching real-time universal, targeted, and robust adversarial attacks against multi-class speaker recognition system, with 109 speakers in our testing model.

3. REAL-TIME, UNIVERSAL, AND ROBUST ADVERSARIAL EXAMPLES

3.1. Target Speaker Recognition Model

In this work, the DNN-embedding-based X-vector system [9] is used as the speaker recognition system since it has shown a significant improvement over standard i-vector models, and has been further studied in many follow-up studies (e.g., [13, 14]). The architecture of X-vector system is shown in Figure 1. Specifically, for an input audio, the system first extracts mel-frequency cepstral coefficents (MFCCs) features using a sliding window. The extracted features are then passed to a time-delay neural network (TDNN) structure [15] that operates on audio frames. The statistics pooling layer takes the output of the final frame-level layer as input, aggregates over the input segment, and computes its mean and standard deviation. Subsequently, hidden layers are used to map the concatenated statistics into final embeddings. In the recognition phase, the probabilistic linear discriminant analysis (PLDA) computes the probability of the input audio belonging to each enrolled speaker with the embedding information and identifies the speaker label with the highest calculated score.

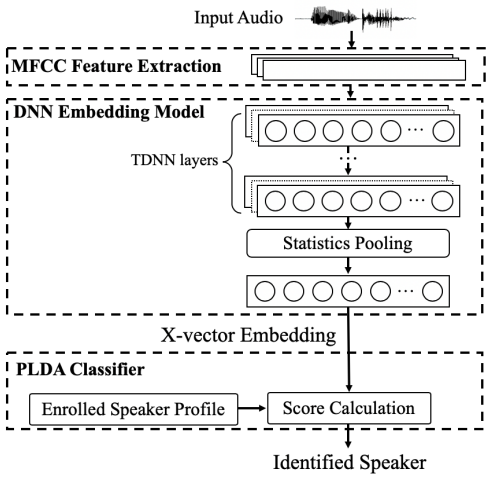

Fig. 1. Targeted speaker recognition model (X-vector).

3.2. Challenges and Threat Model

Challenges. Generating such a real-time, universal, and robust adversarial example against speaker recognition system in practice raises a number of challenges:

(1) Real-time Adversarial Attack. To craft an adversarial noise with respect to the speaker's speech, using conventional optimization-based approach is usually very time-consuming, which makes many practical attack scenarios impossible, such as playing the adversarial noise on a hidden speaker in a real-time manner along with the speaker's voice input.

(2) Universal Targeted Adversarial Example. Using an audio-agnostic universal perturbation to deceive the speaker recognition system, which causes it to misclassify any enrolled speaker's input audio as the adversary-desired speaker, needs to build a universal mapping from the audio sources to the adversary-desired target. The proposed algorithm needs to be general enough to various length audio inputs spoken by different speakers with various accents.

(3) Robust Adversarial Example. The attack performance would be inevitably impacted by the sound distortions due to the attenuation and multi-path effects while playing the adversarial examples over the air. Thus, the generated adversarial perturbation needs to be robust enough to remain effective under this kind of real-world distortions.

Threat Model. In this work, we consider the white box threat model where the adversary has full knowledge of the target speaker recognition model as well as its parameters. In order to build a robust adversarial attack considering the sound distortions in the room where the attack will be launched, we assume the adversary has access to the room's layout. As shown in Figure 2, we aim to find a single audio-agnostic universal perturbation that can be applied on arbitrary enrolled speakers' input audio to mislead the speaker recognition sys-

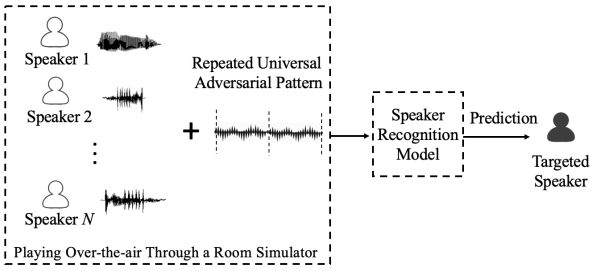

Fig. 2. Threat model of the proposed attack.

tem causing it output the specific adversary-desired speaker label. Additionally, we expect to build a more robust adversarial perturbation that can remain effective while being played over-the-air in acoustic room simulated environments.

3.3. Real-time, Universal and Robust Adversarial Attacks

Most of the existing targeted adversarial attacks would fool DNN-based systems through building different adversary perturbation for each individual input. Differently, in this paper we explore how to build a single universal perturbation that can be directly applied to arbitrary speaker's any utterance, making the speaker recognition system output the adversary-desired speaker label. Such a universal perturbation would greatly shorten the attack launching time, making real-time attacks possible.

To clearly present the steps of our perturbation generation, we model the target speaker recognition system, X-vector, as a function $F(x)$, which takes as input an utterance x and outputs a predicted speaker label. We define $P(x)$ as the function of all DNN layers (including PLDA) to compute the probabilities of classifying x as each of the profiled speakers. We can recognize the voice as the speaker with highest calculated probability, $F(x) = Argmax(P(x))$. Therefore, to launch a universal targeted adversarial attack, where targeted speaker label is t, we aim to find a perturbation δ that could achieve $F(x+\delta) = Argmax(P(x+\delta)) = t$ for arbitrary x.

To build such a universal attack, we need to find a general solution that can make the generated perturbation effective for all the utterances regardless of their speakers, accents, speech content and length. To overcome the issue of varying utterance length, we dynamically construct the universal perturbation δ based on the length of the input utterance x:

$$\delta = Crop([\triangle\delta \frown ... \frown \triangle\delta], x), \tag{1}$$

where $\triangle\delta$ is a short-length adversarial perturbation (e.g., $1s$ in our work), and $[\triangle\delta \frown ... \frown \triangle\delta]$ is a vector constructed by repeating $\triangle\delta$. $Crop(\cdot,\cdot)$ crops the first input to the length of the second input. With this process, the derived perturbation δ could be applied to the audio input with any length.

To minimize the distortion between the adversarial example and the original voice, δ would be clipped to a pre-defined range. The generated adversarial example with the clipped δ could be formulated as:

$$x' = x + Clip_{\epsilon}(\delta), \tag{2}$$

where $Clip_{\epsilon}(\delta)$ is the function to perform element-wise clipping of δ. Values of δ outside the interval $[-\epsilon, \epsilon]$ would be clipped to the interval edges, and ϵ is our pre-defined attack strength.

Moreover, to preserve the effectiveness of the adversarial example while being played over the air, we first mimic the sound distortions during playback and recording by estimating room impulse response (RIR), r, which characterizes the acoustic propagation (e.g., reverberations) in a room environment. The details of how to estimate RIR (i.e., r) based on the room setting are provided in Section 3.4. Then, we could iteratively derive the targeted adversarial example through the following objective function:

$$Argmax(P(x' * r)) = t, \tag{3}$$

where t is the targeted speaker label, $*$ denotes the convolution operation, and $x' * r$ is the estimated adversarial example recorded by the microphone. It is important to note that the estimated RIR represents a certain mapping from the played sound to the recorded sound as per specific location of the loudspeaker and microphone in the room. To make the generated adversarial examples robust in various environmental settings, we estimate multiple RIRs $\mathbf{r}$ in various environments. To make the adversarial perturbation survive all these environments, we randomly select one RIR in $\mathbf{r}$ for each training step when updating the perturbation based on each training utterance. In addition, as directly solving the non-linear constrained non-convex problem is difficult, we iteratively solve the following optimization problem[7]:

$$\textbf{minimize}\, \max(\max\{P(x' * r)_i : i \neq t\} - P(x' * r)_t, -\kappa), \tag{4}$$

where $\{P(x' * r)_i : i \neq t\}$ represents the output probabilities of all speakers except the targeted speaker, while $P(x' * r)_t$ denotes the predicted probability to the targeted speaker. κ is a configurable parameter which represents attack confidence and is set to 0 in our implementation. To generate the universal perturbation, we iteratively modify the trainable sequence, $\triangle\delta$, which is used for constructing δ, with the entire training dataset until satisfying the desired attack success rate. For each training utterance, if the predicted probability of the targeted class is larger than other classes, the update of the perturbation $\triangle\delta$ is skipped on the next sample.

3.4. Room Impulse Response Estimation

Acoustic propagation in a room is commonly considered as a linear and time-invariant system. Thus the recorded signal $R(x)$ could be presented as a deterministic function of the played signal x: $R(x) = x * r$, where r is the estimated room impulse response (RIR), and $*$ denotes the convolution operation. To simulate the play-over-the-air process in the physical world, we take the RIR generated by an acoustic room simulator [16] into account in the adversarial example training phase. Specifically, the simulator can adjust several parameters, including the size of a 3D shoe-box room, the location of the audio sources and microphones, and the reverberation

Table 1. Results of universal targeted attack.

Attack Strength	Noise Level	Min. Attack Success Rate	Max. Attack Success Rate	Avg. Attack Success Rate
ϵ=0.05	-18.84dB	98.47%	100%	99.95%
ϵ=0.03	-23.27dB	95.31%	99.91%	98.40%
ϵ=0.01	-33.96dB	53.32%	95.48%	83.82%

rate. Optimization with the simulated RIR would increase the robustness of the generated adversarial example, and consequently enable over-the-air attack in practice.

4. EXPERIMENTAL RESULTS

4.1. Experimental Methodology

Dataset. We evaluate our proposed attack on an English multi-speaker corpus provided in CSTR voice cloning toolkit (VCTK) [17]. In total, the dataset contains 44217 utterances spoken by 109 speakers with various accents. The dataset is divided into a training and a testing set with a ratio of 4:1.

Baseline Model. In our TensorFlow-implemented X-vector system [9], 30-dimensional MFCC features with a frame length of $25ms$ are extracted. A pre-trained X-vector DNN embedding model provided in Kaldi [18] is used in the model. The baseline model achieves a classification accuracy of 92.8% on 8896 testing utterances from 109 speakers.

Evaluation Metrics. *(1) Attack Success Rate*: The ratio between the number of succeeded attacks and the total number of attack attempts; *(2) Noise Level*: We quantify the relative noise level of the perturbation δ with respect to the original audio x in decibels (dB): $D(\delta, x) = 20log_{10}(\frac{max(\delta)}{max(x)})$.

4.2. Attack Evaluation

Effectiveness of Universal Targeted Attack. To evaluate the effectiveness of our proposed universal targeted attack, we alternatively choose one of the 109 enrolled speakers as the targeted speaker and the rest 108 speakers as victims. In total, we generated 109 universal adversarial perturbations, trying to make the speaker recognition system classify the victims' utterances as the targeted speakers. As shown in Table 1, by adjusting attack strength ϵ, the noise level ranges from $-18.84dB$ to $-33.96dB$. As discussed in the previous study [6], such noise level is considered to be quasi-imperceptible to humans. For instance, $-33.96dB$ is comparatively the difference between a person talking and the ambient noise in a quiet room. For each ϵ value, the minimum, maximum, and average attack success rate among all attack attempts targeting on 109 speakers are calculated. We can observe that when the noise level is $-18.84dB$, a high average attack success rate of 99.95% can be reached. When the noise level decreases to -33.96 dB, the average attack success rate still remains over 80%, which illustrates the effectiveness of our proposed universal targeted attack.

Robustness Analysis Using Room Simulator. An acoustic room simulator toolkit [16] is used to simulate the audio propagation in a room environment. Specifically, a modeled room with a size of $5m \times 5m \times 3m$ is used, and 120 locations of the loudspeaker and the microphone are chosen randomly in the room for RIR estimation. For the estimated RIRs, 100 locations are used to build the universal, targeted and robust adversarial perturbation, and the rest 20 locations are used for testing. Table 2 summarizes the results of our practical universal adversarial perturbation. We can observe that the universal adversarial perturbations trained with RIRs still remain effective after the over-the-air simulation. In particular, the practical universal perturbation generated with a noise level of $-18.84dB$ can still achieve an average attack success rate of 90.19%. For comparison, we test the adversarial perturbation of the same noise level and without RIR in the simulated room environment. However, the average attack success rate decreases significantly to 1.33%. This shows that our approach can efficiently improve the robustness of the generated adversarial examples.

Table 2. Results of robust universal targeted attack using acoustic room simulator.

	Noise Level	Min. Attack Success Rate	Max. Attack Success Rate	Avg.Attack Success Rate
Without RIR	-18.84dB	0.7%	3.52%	1.33%
With RIR	-18.84dB	74.68%	98.05%	90.19%
	-23.27dB	66.54%	96.81%	86.17%
	-33.96dB	54.48%	90.83%	78.25%

Speedup on Attack Time. Unlike conventional individual attacks that require to build adversarial perturbation for each individual voice input, our proposed universal attack could generate a single perturbation that makes arbitrary speaker's utterances to be identified as the adversary-desired speaker. Thus, simply playing the pre-generated universal perturbation nearby the victim speaker becomes possible for launching adversarial attacks. For showing the possibility of launching real-time attacks, we compare the attack launching time of using the conventional individual targeted attack method [6] and our proposed universal attack for a given audio signal. Particularly, the conventional targeted attack requires at least $15s$ to deploy, measured on a Tesla V100 GPU with $32GB$ memory, while our proposed universal method only takes an average of $0.015s$, which results in a $100\times$ speedup.

5. CONCLUSION

This paper proposes a real-time, universal, and robust targeted adversarial attack against speaker recognition system. The proposed attack builds a universal perturbation that can be added into any enrolled speaker's voice input to fool the system causing it to output any adversary-desired speaker label. The robustness of the adversarial perturbations is also greatly improved by using an acoustic room simulator to estimate the sound distortions associated with playing the audio over-the-air. Evaluation on a public dataset of 109 speakers shows the effectiveness and robustness of our proposed attack.

Acknowledgments This research is supported in part by the National Science Foundation grants CNS1801630 and CCF1909963, the Army Research Office grant W911NF-18-1-0221 and the Air Force Research Laboratory grant FA8750-18-2-0058.

6. REFERENCES

[1] Google, "Voice match and media on google home," https://support.google.com/googlenest/answer/7342711?hl=en, Sep. 2019.

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

[3] Yun Lei, Nicolas Scheffer, Luciana Ferrer, and Mitchell McLaren, "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, 2014, pp. 1695–1699.

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

[5] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

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

[7] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou, "Hidden voice commands," in *25th {USENIX} Security Symposium ({USENIX} Security 16)*, 2016, pp. 513–530.

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

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

[10] Tavish Vaidya, Yuankai Zhang, Micah Sherr, and Clay Shields, "Cocaine noodles: exploiting the gap between human and machine speech recognition," in *9th {USENIX} Workshop on Offensive Technologies ({WOOT} 15)*, 2015.

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

[12] Paarth Neekhara, Shehzeen Hussain, Prakhar Pandey, Shlomo Dubnov, Julian McAuley, and Farinaz Koushanfar, "Universal adversarial perturbations for speech recognition systems," *arXiv preprint arXiv:1905.03828*, 2019.

[13] David Snyder, Daniel Garcia-Romero, Gregory Sell, Alan McCree, Daniel Povey, and Sanjeev Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5796–5800.

[14] Desh Raj, David Snyder, Daniel Povey, and Sanjeev Khudanpur, "Probing the information encoded in x-vectors," *arXiv preprint arXiv:1909.06351*, 2019.

[15] David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *Interspeech*, 2017, pp. 999–1003.

[16] Robin Scheibler, Eric Bezzam, and Ivan Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 351–355.

[17] Veaux Christophe, Yarnagishi Junichi, and MacDonald Kirsten, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," *The Centre for Speech Technology Research (CSTR)*, 2016.

[18] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011, number CONF.