e-ISSN:2582-7219

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 7, Issue 10, October 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.521

© 2024 IJMRSET | Volume 7, Issue 10, October 2024| DOI: 10.15680/IJMRSET.2024.0710050

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Adversarial Attacks in Voice Biometrics: Understanding the Attack Mechanisms and Defense Strategies

Sandhya H M, Sahana M P, Mamatha D

Assistant Professor, Department of Computer Science & Application, The Oxford College of Science, Bengaluru, India

PG Student (MCA) Department of Computer Science & Application, The Oxford College of Science, Bengaluru, India

PG Student (MCA) Department of Computer Science & Application, The Oxford College of Science, Bengaluru, India

ABSTRACT: While the development of automatic speaker verification systems continues at a blistering pace, the robustness of these systems against adversarial attacks is a significant concern. To date, little is known about the vulnerabilities of speaker verification systems to adversarial perturbations. Understanding the potential threats of adversarial attacks for speaker verification and developing targeted defense strategies is of great importance for securing voice biometrics in practical applications.

Our work provides the first research effort to systematically analyze adversarial examples in speaker verification. We first describe the design and implementation of such attacks on speaker verification systems. Afterwards, we consider the possible defense strategies upon the current adversarial perturbations. Preliminary results indicate that straightforward applications of adversarial defense methods from image classification to the speaker verification domain are not effective even when the strength of adversarial perturbations is at its minimal level. We discuss limitations of state-of-the-art defense strategies in the speaker verification scenario.

We hope our research can motivate more efforts by the speech and signal processing community to develop defense strategies that can effectively conquer the current adversarial attacks and thus fulfill the potential of voice biometrics in various practical applications.

I. INTRODUCTION

As voice biometrics technology continues to advance and voice recognition becomes a common feature for everyday life applications, securing the voice biometric system against potentially dangerous or fraudulent attacks has drawn more and more attention. Due to inherent limitations in using a voice biometric system for secure purposes such as lack of revocability, forgetting, etc., developing an attack-proof voice biometric system becomes an interesting but challenging problem. Similar to attacks in other machine learning models like computer vision systems, machine learning-based voice biometric systems are also susceptible to adversarial attacks. However, prior literature on adversarial attacks in voice biometric systems is less well-understood than adversarial attacks in other machine learning models such as text-based applications. In this work, we invite the first step to understand adversarial attacks in voice biometrics employing deep learning methods. We provide detailed empirical and analytical investigations of the susceptibility of DNN based voice biometric models to adversarial attacks and defense strategies on the system.

Protecting the privacy and security of voice biometrics is essential for the widespread acceptance and adoption of voice biometrics technology. Similar to attacks in other machine learning models, machine learning-based voice biometric models are also susceptible to adversarial attacks. In this work, we study adversarial attacks in DNN-based voice biometric systems. To understand the attack mechanisms and find the optimal defense strategies, we present empirical and analytical investigations in the DNN-based voice biometric model.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

II. BACKGROUND AND SIGNIFICANCE

In recent years, voice biometrics authentication systems have gained remarkable momentum and are widely used in various applications such as smart personal assistants, smart car access, etc. A voice biometrics system authenticates a person by verifying his/her voice with the enrolled identity. The enrolled voice is recorded during registration and stored as a voiceprint. During verification, a new query speech signal is captured from the person, and the voiceprint is matched to the query speech. Psychological and behavioral factors impact both speech production and how speakers vary their utterances, which can lead to a high data mismatch across the two boundaries. Speech is mostly rich due to the presence of both phonetic and paralinguistic features including content and indexical information, that is, both physiological and physical speaker characteristics. The existence of these rich features, and the shift from text-based systems used in typical challenge competitions to speech biometric systems utilized in human technology, have led to the validity and subsequent acceptance of these systems.

2.1 RESEARCH AIM AND OBJECTIVES

The third chapter includes the following main stages: - Perform research on the effectiveness of the voice accent modifier (VAM) when attacking a multi-user voice biometric system - Develop an effective protection mechanism that could be used to reduce the attack success - Carry out comparative research of metric and conversional learning models for the long-term security of the voice biometric system model.

The second part consists of the following main stages: - Define the requirements for constructing the physical non-intrusive attack - Simulate the changes in the performance of the voice biometric system in the case of an overall large attack - Participate in the analysis of the performed attack in the frequency and time domain of the voice audio stream based on the performed attack on the neural network model of the multiuser voice biometric system - Compare the effectiveness of the performed attack on the conversional and metric-trained voice

This study also aims to: - Identify the main stages and methods for constructing an attack under investigation - Simulate the performance of an artificial neural network for a multi-user voice biometric system in order to compare metric and conversional learning - Attack the model of the artificial neural network with a physical non-intrusive attack - Analyze the changes occurring in the voice of the subject and the requirements for constructing the physical non-intrusive attack - Determine the impact of the performed attack on different types of voice audio streams - Develop an effective protection mechanism to reduce the effectiveness of the research attack - Carry out comparative research of metric and conversional learning

As shown in the previous section, the major goal of this work is to demonstrate that it is feasible to construct an adversarial attack against a machine learning model via a non-intrusive physical channel. This statement allows for the exploration of attacking machine learning models outside of the digital or analog channels. The most understudied and unprotected attack in voice biometrics is the physical one, especially the nonintrusive physical one.

III. EFINITION AND SCOPE

Adversarial attacks belong to an ill-specific feature tactic because urban systems are built on computing features or symbols presented in the digital domain. Digital representations have a lot of adjacent realms within an areacollected under adversarial perturbations to complex loss of knowledge and experience full mistakes and therefore create non-negligible negative impacts. The study of adversarial attacks in the field of voice biometrics is of great importance. On the one hand, it demonstrates the weakness of the neat machine learning-based voice recognition functions that differ from the established methods. On the other hand, confidence in operating a growing number of voice-activated applications is fragile, especially in the context of advisors who can exploit people who use speech recognition functions. As a result, deduced attacks serve as a digital test and certification mechanism for the robustness and generalization character of mission-critical voice biometric systems. Therefore, the study can produce transferable countermeasures that can be used to mitigate adversarial attacks against unknown voice biometric systems.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

3.1 APPLICATIONS

Note that although there has been a considerable amount of work in security and privacy of voice biometrics, it has been largely focused on developing new methods for privacy-preserving techniques. For instance, one of the objectives while developing the raw signal representations for voice biometrics was to eliminate the human-demographical related features from the input data without trading off the primary representation learning capabilities needed for the speaker recognition task. Existing privacy-preserving voice biometric methods are exclusively designed to maintain the privacy-preserving functionalities while enforcing the internal discrimination in the latent representations required by the primary task. Whereas, in this work, we focus on detecting the adversarial overlap by explicitly optimizing the feature-based perturbations in the latent voice representations.

IV. DVERSARIAL ATTACKS IN VOICE BIOMETRICS

To the best of our knowledge, there is no systematic study of adversarial attacks and adversarial defense mechanisms in the domain of voice biometrics. Attacks on biometrics can be categorized into two different groups: identification-centric attacks and classification-centric attacks. First, a successful attack only requires the model to fail to authenticate during the identification task. These attacks do not aim to fool the target model and enable an attacker to be considered as a different identity B against the state-of-the-art speaker identification models, and hence, be undetectable and be authorized. Second, an attack is successful if the trigger, embedded adversarial audio, consistently misleads the speech classifier from classifying the payload sample to the target class A. Adversarial attacks in voice biometrics need to be sophisticated and transparent as the physical person or a real-time voice-based system such as a voice activated lock needs to justify the need for its operation. However, a black-box version of the voice-based commercialized system and the information about target speaker(s) are not always available, so an attacker can only assume limited prioritized knowledge about the operational system. Although the common biometric methods are based on the speaker identification models, also known as the verification model, and the source separation models (if existing any) would be affected by the existence of adversarial audio.

We explain the process of adversarial attacks on biometric systems in general, and in voice biometrics specifically, detailing existing works in adversarial attacks on biometric systems in general (irrespective of biometric modality), and in voice biometrics. Adversarial attacks have recently gained significant attention in various areas of computer vision, natural language processing, and speech domains. These works demonstrate that the vulnerabilities in modality-specific machine learning models have the potential to become a major privacy and security issue, and an adversary can easily manipulate the input to deceive the target machine learning model. Adversarial attacks in biometrics, specifically in voice and face biometrics, could pose a significant challenge to the real-world deployment of voice and face-based security-critical applications. In comparison to fingerprint and iris-based biometric modalities, the adversarial vulnerabilities of voice and face-based biometric modalities have newly gained attention and have shown promising results.

4.1 TYPES OF ADVERSARIAL ATTACKS

Transformation attacks aim to transmute a known voice signal into another known voice of a specific target voice. The goal of this attack then is to render the recognized identity of the voice signal to be changed such that it appears as a target voice when performed by the voice biometric system. In impersonation attacks, an adversary aims to spoof the voice biometric system by emitting speech in the desired target voice identity. Unlike transformation attacks, adversaries might not know any particular voice signal for these types of attacks. They can perform transduction-based transformation from speech to the target voice's spectral parameters. Furthermore, it is experienced that the cost of impersonation attacks in general is more than transformation attacks, as impersonation certainly requires imitation of the voice of the target user.

4.2 ATTACK MECHANISMS

Attacks on voice biometrics have been evolving into three types. That voice steganography is an alternative existing covert channel for securely storing these hash-like sequences has yet to be explored. The aim is to develop new models that sustain adversarial perturbations by using generative and compensatory methods. Attacked voice-based and speech-

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

based authentication systems. This UG voice malware iteration utilizes CodecGAN to embed speech-critical control signals deeply into codec artifacts.

The robustness, compensation, and resilience from adversarial attacks task intends to incentivize the creation of more secure and improved speaker verification models for biometric security systems. Given its binary classification nature, the thresholding on the similarity score derived from the speaker embedding's directly affects the EER on the synthetic and clean evaluation sets. Their codecGAN simultaneously estimates a codec artifact and adversarially purifies this artifact to reconstruct clean speech.

VocCat effectively simulates a codec attack by generating and superimposing codec artifact-derived noise onto clean speech frames.

Traditional adversarial attack mechanisms like Universal Adversarial Perturbation (UAP) and fast gradient sign method (FGSM) have recently been extended to target voice biometrics. Fernandes et al. demonstrated that visually imperceptible perturbations can still lead to incorrect speaker recognition in the least-likelyclass-adv setting, even if UAP was crafted using adversarial speech samples that are not temporally aligned with the original clean speech samples. Note that the probability mass is distributed logarithmically across the correct and wrong classes.

V. IMPACT OF ADVERSARIAL ATTACKS

The self-attack scenario is the scenario where the same voice was registered during enrollment and during verification. This is also known as the zero-effort attack. It is clear that the person was able to successfully pass the verification step using the attack. It is instructive to understand why the attack was successful and which other threats it would apply. The areas directly threatened in this scenario: someone can access a device with a voice-based verification system that is already using his/her own voice and, potentially, with advantages such as vocal command systems, control systems, or voice-based payment systems. In this context, it is crucial that the system is robust against such simple attacks, without requiring additional efforts or complex interactions, such as texting a password to an anchor system. The analogy of precisely defining the threat is to the effectiveness of other biometric systems, such as a fingerprint, where the physical barrier is not straightforward to overcome without being malicious.[4][5]

5.1 SECURITY RISKS

Given that voice biometrics is the most convenient form of biometric authentication, deploying the technology in vulnerable environments presents inherent risks. One could easily spoof a target speaker to log into the accounts by playing a previously recorded authentic voice. Without liveness detection or proper anti-spoofing measures, anyone who would like to imitate a target speaker's voice with a voice imitation tool could gain unauthorized access. However, the limitations of SV systems have identified their own inherent vulnerabilities. An infamous vulnerability of the protection mechanism is sometimes considered to be a security issue explicitly. In voice-activated systems, adversarial voice attacks exploiting the weaknesses of the deep learning-based VAD were reported. For SV systems, another concern is that automated VAD solutions ending with ASR may mis-transcribe, and consequently misclassify the speaker identity. As a result, anyone who speaks in potentially critical conditions, such as field workers under stress, will have difficulty receiving support from SV systems.

5.2 ETHICAL IMPLICATIONS

The transformed text into voice is expected to obtain legitimate access to the user's private data, lock the victims into prisons, create several social and banking frauds, change the user's profiles, or purchase anything online. In this case, a particular names list or specific random passwords do not provide any protection mechanism due to the personalized nature of biometric systems. Therefore, the researchers must adhere to the principled responsible study measures while developing secure voice biometrics recognition models. Moreover, the multi-modal template protection mechanism may be proposed as a possible protection measure for the vulnerable voice biometrics model. The biometrics data usage alongside with a date built validation system can be utilized to prevent the misuse and mistaken usages of voice biometric data with speech samples.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

VI. INTRUSION DETECTION SYSTEM

Since the development of IDSs is expansive and has become an active field of research, it is essential to study the detection of adversarial inputs in voice biometrics. We explore leveraging an ensemble of voice activity detector, a template classifier, and a DNN-based classifier model to detect adversarial attacks on speaker recognition systems.

Intrusion detection systems (IDSs) are designed to detect malicious activities in an automated manner using various data processing techniques. The IDS uses the information collected from logs or network traffic to detect abnormal or unusual activities, which might indicate a malicious event or attack. Detection of adversarial attacks on voice biometric systems has not been studied much. presented a deep learning-based acoustic event classification model (AECM), which labels the hidden states of ASV (Automatic Speaker Verification) and chooses the adversarial data frame correctly and efficiently, distinguishing it from others in actual applications. Another work shows that in-text independent digital voice assistant where the backdoor is triggered by a voice sample that is classified as an intentionally designed phrase can be a concern.

6.1 BIOMETRIC SYSTEM ENHANCEMENT

The enhanced MFCC feature extraction expands the adversarial perturbation space with voice biometric-unfit feature dimensions detected as those approximate zero mean and have almost zero plot area in perfectly positive and negative classification of own- and foreign-machine voice samples in denominators of alternate hardmin rationale. The increase in the number of zero means reduces the adherence of the MFCC extracted feature to a reasonable domain for an attack, as clearly seen from the alternately negative and positive proportion of perturbed near-zero-mean MFCC signals in three experiments. Design parameters are proposed in the form of mean and window size, ensuring a reasonable domain for just a small proportion of adversecross-correlation (adverse-cc), significantly outperforming synthesis verification, with respect to an equal error rate, as close to the main just-cross-correlation, close as the two approaches compared. To overcome system performance impediments of forced classification errors with a large number of biometrically unfit signal dimensions, far fewer than half of all 48 frame-by-length 999 MFCCs 135 and 392 are proposed to be design-wise-mean zero and window size 17 and were significantly improved in three experiments to almost three orders of magnitude normalized importance order. With the proposed conversation privacy-preserving biometric transformation, similar to the one based on total scatter of classification data, signal ADs can be made with MDM increasing unfavorable system operation conditions of decreased just-cross-correlation with almost zero forced errors.

VII. CASE STUDIES

A defense strategy is also proposed to prevent such identity falsification attacks for voice biometric systems.

The adversarial attack is a kind of machine learning (ML) security problem. It challenges the robustness of an ML model and the countless applications that use ML models. The vulnerability to adversarial attack is widely investigated among object recognition, such as face recognition. These results motivated us to study the adversarial attacks in another biometric technology, voice recognition. Our study confirms the vulnerability of voice biometrics to the adversarial attacks. To further study the adversarial attack in voice biometrics, this study takes the automatic speaker verification (ASV) model as the example. Since facial recognition is also one of the fastest developing areas in biometrics, in this study, some explorations based on the importance of the speech content, defense strategies, and transferability are transplanted from voice biometrics to facial recognition. These explorations will hopefully not only deepen our understanding of adversarial attacks in other biometric technologies but also construct a feasible defense strategy.

.

7.1 REAL-WORLD EWXAMPLES OF ADVERSARIAL ATTACKS

We additionally provided detailed information on the number of training recordings that included the subjects speaking in their own voices. It is worth noting that small numbers for deployment, number, and sample size in the training set are allowed as objectives because they are not central to this communication. The previous investigation was made clear in supplemental information, and the different types of recordings are clearly displayed in the legend of the base pair. This information highlights various factors about conditional adversarial training results with real activists and journalists. It

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

is important to mention that frequent voice biometric counterfeits have a slightly exaggerated best-of-family side-count for counter-real speech samples in the training set. However, it is worth noting that no real known speech issues have been identified. Additionally, there are recordings of digital broadcasts and YouTube interview videos of the targets available for analysis. Furthermore, it is important to clarify that the examples mentioned were not implicated in any attacks prior to the dates the recordings were made. All exchanges involving voice contact under mention were identified and the audio was sent to Amazon Mechanical Turk. A minimum time of activity and frequency of human feedback was ensured during the analysis process. It is important to acknowledge that information about miscommunication and undetermined biases is remotely triggered when creating voice biometric counterfeits in real time. To delve deeper into the topic, let's discuss the two main types of voice biometric presentations that were considered: test-time counterfeits and training-time counterfeits. Some of these counterfeits were also given the opportunity to introduce personal audio samples. This allowed for the normalization of the voice biometric target outside of prior knowledge adversarial understanding.

VIII. CURRENT REASERACH TRENDS

The trend to steer the research for the defense strategies is seen in three new directions currently being pursued. Some researchers are focusing on the audio characteristic attributes that would protect speech from adversarial attack. This can be a speech that cannot easily fool a DNN model in speaker verification. For instance, trying to design a speech that prevents voice content from being manipulated by the attackers. This also tries to design speech that is not vulnerable to adversarial attack for universal and targeted exploits. It is clear that developing speech that is not vulnerable to attack properties for a specific attack method is not the first direction to be taken, and it involves intensive research that is especially responsible for the challenge of protecting speech for new and emerging attack trends.

8.1 EMERGING TECHNOLOGIES

With deep learning technology across artificial intelligence models, the adversarial utilization of a single model can be expanded to art-to-face visual authentication, infancy-to-end fingerprint fingerprint vision robust deep learning voice tensor Inspiron AI plant adversarial processing environment, voice. In order to increase the counterfeit detection capability contained in the defensive system, it is exposed to the intensified adversarial attack. The new robust processing environment relies on a variety of deep learning models. This makes the process of attack consumption resources easier to conduct and functional limits the success of attacks. Detect counterfeit and robust deep learning environment. Mitigate the limitations of physical security control laws, such as open access areas, which serve as critical access control markers. Working close to the security layer provides a deeper level of security agencies, real-time threat analysis, and enhanced measuring traffic knowledge of adversarial attempt.

IX. CONCLUSION AND FUTURE DIRECTION

The state-of-the-art speaker discriminators can only offer a partial view due to their generic design. Hence, we assume the knowledge of the D-vector extractor which can capture speaker-identity information to facilitate the execution of the proposed attacks. Therefore, our defense is dependent on both the design of the defender and the adversary. To demonstrate the necessity for this design, we first generate both targeted and non-targeted black-box attacks. These transformations only rely on the fundamental assumption of a white-box attack, which means the adversary does not necessarily know the D-vector extractor. Once the defender is in place, we integrate it with the current speaker discriminator to make the generation of voice-masking more challenging.

In this paper, we proposed targeted adversarial attacks for automatic speaker verification systems and unsupervised domain adaptation as a countermeasure to defend against these attacks. The proposed attacks demonstrate the vulnerability of the existing feature extractors used in speaker verification pipelines. Our proposed UDA-based defense ensures that the feature extractor trained on a voice anti-spoof dataset captures speaker-dependent information so that even if the adversary knows it, masking cannot make the enrollment and test trials from the same speaker to have different statistics. We demonstrate the existence of such a strategy in the VCTK dataset by conducting multiple masking experiments. On the SITW dataset, our proposed UDA-based defense records an equal error rate of less than 4.8% with masking, which is far better compared to the non-defense case.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

REFERENCES

1. H. Tan, L. Wang, H. Zhang, J. Zhang et al., "Adversarial attack and defense strategies of speakerrecognition systems: A survey," Electronics, 2022. mdpi.com
2. A. Jati, C. C. Hsu, M. Pal, R. Peri, W. AbdAlmageed et al., "Adversarial attack and defense strategies fordeep speaker recognition systems," Computer Speech & Language, Elsevier, 2021. sciencedirect.com
3. S. Joshi, J. Villalba, P. Å»elasko, et al., "Study of pre-processing defenses against adversarial attacks onstate-of-the-art speaker recognition systems," in IEEE Transactions, 2021. [PDF]
4. Z. Li, C. Shi, Y. Xie, J. Liu, B. Yuan, Y. Chen, "Practical adversarial attacks against speaker recognitionsystems," â€¦ on mobile computing systems â€¦, 2020. acm.org
5. G. Chen, S. Chenb, L. Fan, X. Du, Z. Zhao, et al., "Who is real bob? adversarial attacks on speakerrecognition systems," in IEEE Symposium on, 2021. [PDF]
6. D. Serafino, "Adversarial Machine Learning applied to Automatic Speech Recognition systems," 2022. polito.it
7. A. Annavarapu, S. Borra, R. Thanki, "Progression in Biometric Recognition Systems and its Security,"Recent Patents on â€¦, 2022. [HTML]
8. M. Hernandez-de-Menendez, "Biometric applications in education," International Journal onâ€¦, Springer,2021. springer.com
9. S. Arora and M. P. S. Bhatia, "Challenges and opportunities in biometric security: A survey," InformationSecurity Journal: A Global â€¦, Taylor & Francis, 2022. [HTML]

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com