

Exploring AI in Vishing: Threats and Countermeasures

Kayla Council

Computer Science Department

Hampton University

Hampton, VA, USA

kayla.council@my.hamptonu.edu

Chutima Boonthum-Denecke

Computer Science Department

Hampton University

Hampton, VA, USA

chutima.boonthum@hamptonu.edu

ABSTRACT

This study examines how artificial intelligence (AI) can help with voice phishing (vishing) attacks, with a particular emphasis on deepfake technologies and AI-driven voice synthesis. It examines the strategies used by cybercriminals, assesses the effectiveness of the present defenses, and identifies difficulties in identifying and preventing such attacks. The results show that to combat the increasing complexity of vishing strategies, there is an urgent need for sophisticated detection systems and preventive actions. Future directions include the creation of cooperative policy frameworks to control the misuse of AI and easily accessible solutions for small enterprises.

KEYWORDS

AI, vishing, deepfake, voice synthesis, cybersecurity.

1 Introduction

Voice phishing, or vishing, has become a serious cybersecurity risk. Artificial intelligence (AI) has helped vishing, historically relying on human actors, especially with voice synthesis and deepfake technology. Due to these developments, attackers can now imitate trusted voices with an unprecedented level of realism. AI-driven vishing attacks, which target people, businesses, and governments alike, have increased in frequency and sophistication as these tools have become more widely available.

2 Background

Since its emergence in the early 1900s, voice phishing, or vishing, has remained a persistent threat [12]. Despite advances in cybersecurity, vishing continues to be effective because it exploits human behavior rather than technological vulnerabilities. Cybercriminals often use emotional manipulation or create a sense of urgency to coerce victims into revealing sensitive information, such as credentials, account details, or personal data [12]. They capitalize on predictable employee behaviors, such as politeness and a desire to help. By targeting the human element, these attacks can bypass even the most robust technical defenses, highlighting how social engineering tactics, emphasize the need for increased awareness and training to defend against these sophisticated threats [12].

3 Methods of Vishing Attacks Using AI

Cybercriminals are increasingly using AI technologies to enhance the effectiveness of their vishing attacks. The key AI techniques employed in these attacks are outlined in the subsections below.

3.1 Voice Synthesis

Voice synthesis involves leveraging artificial intelligence to produce natural-sounding, expressive computer-generated speech by processing and learning from text and audio inputs [6]. This technology enables cybercriminals to exploit its capabilities in vishing attacks. AI-driven voice synthesis allows attackers to replicate a target's voice with high accuracy, making it difficult for victims to distinguish between a genuine call and a fraudulent one. Figure 1: Diagram of the Open Voice framework, demonstrating how AI manipulates speech attributes such as style, language, and tone to replicate a reference speaker's voice. A notable example of AI-driven voice synthesis is in a 2019 vishing attack that occurred when the CEO of a UK-based energy company transferred approximately \$243,000 to a Hungarian supplier [2]. He believed he was speaking with the chief executive of the company's German parent, but the voice was synthesized to mimic the executive's accent. The CEO became suspicious when he noticed the call from an Austrian phone number, and he never received the promised reimbursement. These two factors revealed that it was a scam [2].

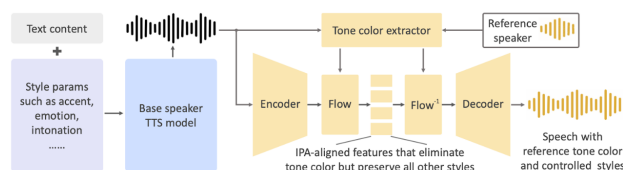


Figure 1: Overview of the Open Voice framework, illustrating how AI in technology modifies speech attributes like style, language, and tone to replicate the voice of a reference speaker [5].

3.2 Deepfake Technology

Deepfake technology is an advanced form of artificial intelligence that manipulates audio, video, or both to replicate a specific individual with remarkable precision [13]. Unlike voice synthesis, which creates lifelike computer-generated audio, deepfake technology focuses on mimicking a particular person's voice in real-time, enabling attackers to impersonate trusted individuals during phone calls [13]. A recent example of deepfake technology misuse occurred in 2024 when a video featuring Elon Musk was circulated online [3]. Figure 2 illustrates this scenario by comparing two video screenshots: one featuring the authentic Elon Musk and

the other showcasing the deepfake version, demonstrating the deceptive potential of such technology. In the video, Musk appeared to endorse a cryptocurrency scam, convincing viewers that it was a legitimate opportunity. The lifelike portrayal successfully deceived many individuals, resulting in substantial financial losses. This incident highlights the growing sophistication of deepfake technology and its potential to be weaponized for fraudulent activities [3].

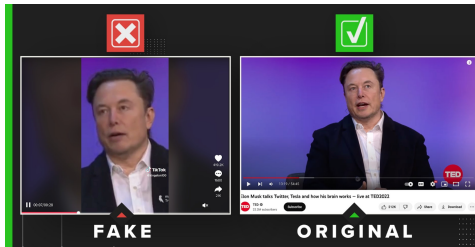


Figure 2: Comparison of authentic and deepfake video screenshots of Elon Musk, highlighting the visual similarities that make deepfakes deceptive and dangerous [4].

3.3 Social Engineering Integration

Social engineering integration enhances the effectiveness of AI-driven vishing attacks by combining advanced technology with psychological manipulation techniques [9]. AI tools generate adaptive scripts and real-time responses that exploit human behavior, increasing the likelihood of success [9]. These tactics are illustrated in Figure 3, which highlights six common social engineering attack methods: baiting, pretexting, tailgating, phishing scams, phone scams, and shoulder surfing. These scripts often incorporate urgency, authority, or emotional appeals to manipulate the victim’s decision-making process [9]. A recent example of this occurred in 2024, when cybercriminals used AI-generated voices to impersonate trusted individuals, such as family members or colleagues [1]. The attackers employed psychological manipulation by creating a sense of urgency or authority, leading victims to disclose sensitive information or transfer funds [1]. The combination of realistic voice cloning and tailored social engineering tactics significantly increased the success rate of these scams, showcasing the growing capability of AI to exploit human behavior effectively [1].



Figure 3: Six common social engineering attack methods including baiting, pretexting, tailgating, phishing scams,

phone scams, and shoulder surfing, often exploited in AI-driven vishing attacks [10].

4 Countermeasures and Detection Techniques

It is crucial to put in place efficient defenses and detection methods to lessen the impact of more complex AI-driven vishing attacks. Technological solutions and human-focused solutions are the two primary types of countermeasures. Given the dynamic nature of these threats, both are essential.

Several technical solutions have been developed to detect AI-generated manipulations whether in voice synthesis or deepfake technology. These solutions are not always apparent to the naked eye, but they play a crucial role in verifying the authenticity of both audio and video content. The primary techniques include:

4.1 AI and Machine Learning Tools

While AI technology can be used to generate deepfakes and synthetic voices, it can also be used to detect them. Machine learning algorithms are designed to identify inconsistencies in both audio and visual components, such as irregular speck patterns in synthetic voices or lighting mismatches in deepfake videos. These tools analyze subtle details that may be difficult for humans to detect, improving the detection of AI-driven manipulation [8]. Figure 1 illustrates a general framework for detecting manipulated media using AI-driven approaches, highlighting the various stages involved in identifying deepfakes.

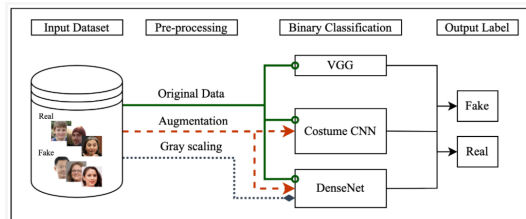


Figure 4: Overview of an AI-driven framework for detecting deepfake media, showcasing key stages in the detection process [10].

4.2 Digital Watermarking

Digital Watermarking is an effective method of detection, which involves embedding invisible markers into audio and video files (see Figure 5). These markers serve to track the origin of the content and verify its authenticity. Since altering these watermarks typically compromises the integrity of the file, digital watermarks provide a dependable way to identify manipulated media, including both deepfake videos and synthetic voices [8].

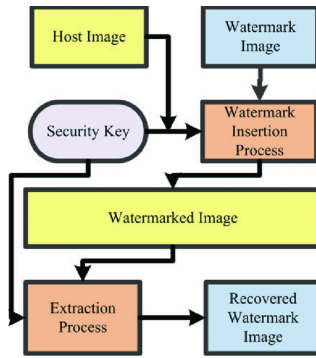


Figure 5: Basic concept of digital watermarking, showing the process from the host image to the recovered watermark image, including the watermark insertion and extraction processes [7].

4.3 Forensic Analysis

Forensic analysis focuses on identifying artifacts left behind by manipulation software. When analyzing synthetic voices, experts look for signs such as unnatural speech rhythms, tone inconsistencies, or digital artifacts that suggest tampering. In the case of deepfakes, forensic tools can detect irregular pixel patterns or visual inconsistencies in the video. These methods are essential for uncovering signs of manipulations that may not be immediately apparent through casual observation [8].

In addition, human-focused solutions, such as employee awareness programs and training, are crucial in preventing vishing attacks. By educating employees to identify common social engineering tactics, organizations can strengthen their defenses by enabling staff to respond appropriately to suspicious interactions [1]. Vishing simulations and phishing awareness training can help employees recognize manipulative tactics, such as urgent requests or emotional appeals [1]. Furthermore, establishing verification procedures, such as call-back protocols or requiring additional forms of authentication, can provide an extra layer of protection for sensitive information [1].

5 Analysis and Findings

This section examines the key observations from the research, focusing on how advancements in AI, particularly voice synthesis and deepfake technology, have enhanced the effectiveness of vishing attacks, as well as the countermeasures developed to address these threats.

5.1 Analysis of Vishing Methods Using AI

The finding highlights how cybercriminals have adopted AI-driven voice synthesis and deepfake technology to enhance their schemes. The 2019 vishing attack on a UK-based energy company underscored the effectiveness of voice synthesis in mimicking trusted individuals, allowing attackers to exploit vulnerabilities. Similarly, the misuse of deepfake technology in creating deceptive videos, such as the Elon Musk cryptocurrency scam, illustrates the alarming precision and reach of this technology.

By integrating social engineering tactics with AI-generated voices, attackers manipulate victims by leveraging emotional triggers like urgency and authority. These sophisticated methods demonstrate how AI amplifies the success of traditional social engineering. Posing significant challenges for detection and prevention.

5.2 Evaluation of Countermeasures and Detection Techniques

The countermeasures discussed in this research reveal a paired approach: technological and human-focused solutions. AI and machine learning tools provide significant promise in identifying manipulation, with algorithms capable of detecting inconsistencies in audio and video. Digital watermarking further strengthens authenticity checks by embedding traceable markers into media files, creating an additional layer of defense. Forensic analysis enhances detection capabilities by identifying subtle artifacts left behind by manipulation tools, offering a deeper level of analysis.

On the human-focused side, employee training programs and vishing simulations demonstrate the importance of awareness in mitigating these threats. Such initiatives empower individuals to identify manipulation attempts, while verification protocols help safeguard sensitive interactions.

5.3 Findings and Implications

The research findings emphasize the dual role of AI in both advancing and combating vishing attacks. While technological countermeasures provide robust defenses, they require continuous refinement to keep pace with evolving threats. The importance of integrating human-focused strategies cannot be overstated, as attackers often exploit behavioral vulnerabilities. These findings highlight the need for a holistic approach, combining advanced detection systems with comprehensive training and policy frameworks. Moreover, organizations must prioritize accessibility in countermeasure deployment, ensuring that even small enterprises can adopt these tools to safeguard against sophisticated vishing attacks.

6 Future Work

As AI continues to evolve, the sophistication of vishing attacks is expected to grow, requiring further development in both technological and human-focused countermeasures. Future work should prioritize enhancing existing detection systems, incorporating more advanced machine learning algorithms to better identify manipulations that may be undetectable by current methods. Additionally, research into improving the accessibility of countermeasures for smaller enterprises, which may lack the resources of larger organizations, is crucial to ensuring broad protection against these threats. There is also a need for ongoing education and training programs that evolve alongside new AI capabilities, ensuring individuals can recognize and respond to increasingly realistic vishing tactics. Collaboration between industry stakeholders and policymakers could facilitate the creation of standardized frameworks that support both the development and deployment of these defense mechanisms, encouraging more coordinated efforts to combat AI-driven vishing attacks on a global scale.

7 Conclusion

This study has examined the growing threat of AI-driven voice phishing (vishing) attacks, particularly the role of voice synthesis and deepfake technologies in amplifying the effectiveness of these schemes. Cybercriminals have increasingly adopted sophisticated AI techniques, enabling them to mimic trusted voices and manipulate victims using psychological tactics. As a result, detecting and preventing these attacks has become significantly more challenging. While technological solutions, such as AI-powered detection tools, digital watermarking, and forensic analysis, show promise in identifying manipulated content, human-focused strategies, such as training programs and verification protocols, remain essential for mitigating the risks of vishing.

The findings underscore the importance of a multi-faceted approach to defending against AI-driven vishing attacks. This includes not only strong technological defenses but also ongoing education to equip individuals with the knowledge and skills to recognize manipulation. As AI continues to evolve, the need for continued innovation in detection systems and the development of accessible countermeasures for small enterprises will be critical in ensuring broad protections. By encouraging collaboration between technology providers, policymakers, and businesses, we can strengthen our collective defenses against the growing threat of AI-enhanced vishing and better safeguard sensitive information in the digital age.

REFERENCES

- [1] Emily Astranova and Pascal Issa. 2024. Whose Voice Is It Anyway? AI-Powered Voice Spoofing for Next-Gen Vishing Attacks. In Google Cloud Blog. Retrieved from <https://cloud.google.com/blog/topics/threat-intelligence/ai-powered-voice-spoofing-vishing-attacks>.
- [2] Jesse Damiani. 2019. A Voice Deepfake Was Used To Scam A CEO Out Of \$243,000. In Forbes. Retrieved from <https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/>.
- [3] Incode. 2024. Top 5 Cases of AI Deepfake Fraud From 2024 Exposed. In Incode. Retrieved from https://incode.com/blog/top-5-cases-of-ai-deepfake-fraud-from-2024-exposed/?utm_source=chatgpt.
- [4] Kelly Jones. 2022. Video of Elon Musk saying Twitter employees got clips of Trump saying 'you're fired' is fake. In 13's News Now. Retrieved from <https://www.13newsnow.com/article/news/verify/technology-verify/elon-musk-twitter-employees-fired-deepfake-fact-check/536-5b4abaca-fea2-4a4d-86e8-96afa56b9c66>.
- [5] Mercy. (n.d.). How to build real-time voice cloning pipelines. In Mercy. Retrieved from <https://www.mercity.ai/blog-post/how-to-build-real-time-voice-cloning-pipelines>.
- [6] Moveworks. 2024. What is Voice Synthesis? In Moveworks. Retrieved from <https://www.moveworks.com/us/en/resources/ai-terms-glossary/voice-synthesis#:~:text=Voice%20synthesis%20is%20using%20artificial,Previous%20Next%20Pause%20Play>.
- [7] M. K. Pandey, Girish Parmar, Rajeev Gupta, and Afzal Sikander. 2020. Lossless Color Image Watermarking Based on Lifting Scheme and GWO for Copyright Protection. In Researchgate. Retrieved from https://www.researchgate.net/figure/Basic-concept-of-digital-watermarking_fig1_339470274.
- [8] Ben Potaracke. 2024. Mastering Deepfake Detection: Strategies for Identifying the Latest Digital Deception. In Locknet. Retrieved from <https://www.locknetmanagedit.com/blog/cybersecurity/deepfake-detection#:~:text=Machine%20learning%20algorithms%20can%20be,often%20tell%20signs%20of%20deepfakes>.
- [9] Marc Schmitt and Ivan Flechais. 2024. Digital Deception: Generative Artificial Intelligence in Social Engineering and Phishing. In Springer Nature Link. Retrieved from <https://link.springer.com/article/10.1007/s10462-024-10973-2#citeas>.
- [10] Emma Soderlund. (n.d.). Keep an Eye Out for These 6 Social Engineering Techniques Targeting Employees. In CyberPilot. Retrieved from <https://www.cyberpilot.io/cyberpilot-blog/keep-an-eye-out-for-these-6-social-engineering-techniques-targeting-employees>.
- [11] Maryam Taeb and Hongmei Chi. 2022. Comparison of Deepfake Detection Techniques through Deep Learning. In MDPI. Retrieved from <https://www.mdpi.com/2624-800X/2/1/7>.
- [12] TraceSecurity. 2024. Vishing. In Tracesecurity. Retrieved from <https://www.tracesecurity.com/services-software/security-awareness/vishing>.
- [13] Kinza Yasar, Nick Barney, and Ivy Wigmore. 2024. What is Deepfake Technology? In TechTarget. Retrieved from <https://www.techtarget.com/whatis/definition/deepfake>.