AI-Powered Voice Spoofing for Next-Gen Vishing Attacks

By Mandiant

Published: 2024-07-23 · Archived: 2026-04-06 01:27:51 UTC

Written by: Emily Astranova, Pascal Issa

Executive Summary

- AI-powered voice cloning can now mimic human speech with uncanny precision, creating for more realistic phishing schemes.
- According to news reports, scammers have leveraged voice cloning and deepfakes to steal over HK$200 million from an organization.
- Attackers can use AI-powered voice cloning in various phases of the attack lifecycle, including initial access, and lateral movement and privilege escalation.
- Mandiant's Red Team uses AI-powered voice spoofing to test defenses, demonstrating the effectiveness of this increasingly sophisticated attack technique.
- Organizations can take steps to defend against this threat by educating employees, and using source verification such as code words.

Introduction

Last year, Mandiant published a blog post on threat actor use of generative AI, exploring how attackers were using generative AI (gen AI) in phishing campaigns and information operations (IO), notably to craft more convincing content such as images and videos. We also shared insights into attackers' use of large language models (LLMs) to develop malware. In the post, we emphasized that while attackers are interested in gen AI, use has remained relatively limited.

This post continues on that initial research, diving into some new AI tactics, techniques, and procedures (TTPs) and trends. We take a look at AI-powered voice spoofing, demonstrate how Mandiant red teams use it to test defenses, and provide security considerations to help stay ahead of the threat.

Growing AI-Powered Voice Spoofing Threat

Gone are the days of robotic scammers with barely decipherable scripts. AI-powered voice cloning can now mimic human speech with uncanny precision, injecting a potent dose of realism into phishing schemes. We are reading more stories on this threat in the news, such as the scammers that reportedly stole over HK$200 million from a company using voice cloning and deepfakes, and now the Mandiant Red Team has incorporated these TTPs when testing defenses.

Brief Overview of Vishing

Unlike its traditionally email-based counterpart, vishing (voice phishing) uses a voice-based approach. Rather than sending out an email with the hopes of garnering clicks, threat actors will instead place phone calls directly to individuals in order to earn trust and manipulate emotions, often by creating a sense of urgency.

Like traditional phishing, a threat actor's goal is to deceive individuals into divulging sensitive information, initiating malicious actions, or transferring funds using social engineering tactics. These deceptive calls often impersonate trustworthy entities such as banks, government agencies, or tech support, adding an extra layer of authenticity to the scam.

The rise of powerful AI tools such as text generators, image creators, and voice synthesizers has sparked a wave of open-source projects, making these technologies more accessible than ever before. This rapid development is putting the power of AI into the hands of a wider audience, fueling the potential for more convincing vishing attacks.

AI-Powered Voice Spoofing in the Attack Lifecycle

Modern voice cloning involves recording and processing audio and training a model. Training the model relies on a powerful combination of open-source libraries and algorithms, of which there are many popular choices today. When these initial steps are completed, attackers may take additional time to understand speech patterns of the individual being impersonated, and even write a script before conducting operations. This helps create an extra layer of authenticity, and the attack is more likely to be successful.

Next, attackers may use AI-powered voice spoofing in different stages of the attack lifecycle.

Initial Access

There are various ways a threat actor can gain initial access using a spoofed voice. Threat actors can impersonate executives, colleagues, or even IT support personnel to trick victims into revealing confidential information, granting remote access to systems, or transferring funds. The inherent trust associated with a familiar voice can be exploited to manipulate victims into taking actions they would not normally take, such as clicking on malicious links, downloading malware, or divulging sensitive data. Although voice-based trust systems are seldom used, AI-spoofed voices can also potentially bypass voice-based authentication systems used for multi-factor authentication or password resets, granting unauthorized access to critical accounts.

Lateral Movement and Privilege Escalation

Threat actors can leverage AI voice spoofing to hop from system to system, impersonating trusted individuals to manipulate their way to higher access levels. There are a few ways this may unfold.

One method of lateral movement is chaining impersonations. Imagine an attacker initially gaining access by impersonating a helpdesk employee. After establishing communication with a network administrator, the attacker could subtly record the administrator's voice during the interaction. This captured audio can then be used to train a new AI voice spoofing model, allowing the attacker to seamlessly impersonate the administrator and initiate communication with other unsuspecting targets within the network. This chaining of impersonations enables the attacker to move laterally, potentially gaining access to more sensitive systems and data.

Another method is during the initial access phase, threat actors might discover readily available voice recordings on a compromised host, such as voicemails, meeting recordings, or even training materials. These recordings can be leveraged to train AI voice-spoofing models, allowing the attacker to impersonate specific individuals within the organization without needing to interact with them directly. This can be particularly effective for targeting high-value individuals or bypassing systems that rely on voice biometrics for access control.

Mandiant Red Team Proactive Case Study

In late 2023, Mandiant conducted a controlled red team exercise with a client, using AI voice spoofing to gain initial access to their internal network. This case study highlights the effectiveness of this increasingly sophisticated attack technique.

The exercise began with obtaining client consent and crafting a custom realistic social engineering pretext. The Red Team opted to impersonate a member of the client's security team, requiring a natural voice sample. After reviewing the pretext with the client, the client provided explicit permission to use their voice for this exercise.

Next, we obtained the necessary audio data to train a model, and achieved a passable level of realism. Open-source intelligence (OSINT) played a crucial role in the next phase. By gathering employee data (job titles, locations, phone numbers), the Red Team identified potential targets most likely to recognize the impersonated voice and possess the necessary permissions for our objectives. With a curated target list, the team initiated spoofed calls via VoIP services and number spoofing.

After facing voicemail greetings and other initial hurdles, the first unsuspecting victim answered with a trusting "Hey boss, what's up?". The Red Team had reached a security administrator who reported to the person whose voice was spoofed. Leveraging the pretext of a "VPN client misconfiguration," the Red Team exploited the opportune timing of a recent global outage impacting the client's VPN provider. This carefully chosen scenario instilled a sense of urgency and increased the victim's susceptibility to our instructions.

Due to the trust in the voice on the phone, the victim bypassed security prompts from both Microsoft Edge and Windows Defender SmartScreen, unknowingly downloading and executing a pre-prepared malicious payload onto their workstation. The successful detonation of the payload marked the completion of the exercise, showcasing the alarming ease with which AI voice spoofing can facilitate the breach of an organization.

Security Considerations

This type of exploitation is social in nature, and currently technical detection controls are limited. Available mitigations center around three major principles: awareness, source verification, and future technical considerations.

Awareness

Educate employees, particularly those who control money and access, on the existence and methodologies of AI vishing attacks. Consider adding AI enhanced threats to security awareness training. With such effective and accessible mimicry available to threat actors, everyone should now adopt a healthy dose of skepticism when dealing with phone calls, especially if they fall under one or more of the following cases:

- The caller is saying things that sound too good to be true.
- The call is from an untrusted number/entity.
- The caller tries to enforce questionable authority.
- The caller is out of character for the source.

Employees in trusted positions should be extremely wary of high urgency calls that demand immediate action, especially when the caller asks or gives financial or access oriented information, such as requesting a one-time password (OTP). Employees should be empowered to hang up and report suspicious calls, especially if they believe AI vishing is involved. It is likely another employee is about to receive the same attack.

Source Verification

When possible, cross-reference the information with trusted sources. This includes hanging up and calling back at a number previously validated for the source. The caller can be asked to send a text message from a previously validated number or ask them to send an email or an enterprise chat message.

Train employees to spot audio inconsistencies, such as sudden variation of background noise, which could be a symptom of the threat actor not spending enough time cleaning the audio. Look for unusual speech patterns, like a completely different vernacular than what the source typically uses. Watch for unnatural inflections, fillers not commonly used by the source, strange clicks, pauses or abnormal repetition. Pay attention to voice timbre (tone) and cadence as well.

Establish code words for executives and critical staff that deal with sensitive and/or financial information. Do this out of band so there is no trace within the enterprise to limit exposure in case of a breach. The code words can then be used to validate individuals in case of doubt.

If possible, let unknown numbers go to voicemail. Apply the same vigilance to voice calls that you would otherwise apply to emails. Report any suspicious calls for wider awareness.

Future Technical Considerations

Today, organizations can, at best, implement traditional security measures to protect audio conversations within the organization, like using separate networks for VoIP channels as well as implementing authentication and transmission encryption for the same. However, this does not resolve attacks made against employees' personal phones.

Going forward, organizations should consider protecting all audio assets, implementing technologies such as digital watermarking that are subtle enough to be imperceptible to the human ear, but easily detected by AI technologies.

Eventually, mobile device management tools will offer technologies to help verify callers. In the meantime, organizations should consider requiring all sensitive conversations to occur over enterprise chat channels, where strong authentication is required, and identities are not easily spoofed.

Research and tools are actively being developed to help in detecting deepfakes. While they have inconsistent accuracy today, they can still provide value in identifying deepfakes in voicemail or offline voice notes. The

detection capabilities will improve over time and eventually be adopted into supportable enterprise tooling. For additional reading, consider the active research going into real-time detection, such as DF-Captcha, which suggests a simple application to queue human prompts implemented using challenge response to validate the identity of the party on the other line.

Conclusion

In this blog post, we explored how modern AI tools can help create more convincing vishing attacks. The alarming success of Mandiant's vishing underscores the urgent need for heightened security measures against AI voice-spoofing attacks. While technology offers powerful tools for both attackers and defenders, the human element remains the critical vulnerability. The case study we shared should serve as a wake-up call, urging organizations and individuals alike to take proactive steps.

Mandiant started leveraging AI voice-spoofing attacks in its more complex Red Team Assessments and Social Engineering Assessments to demonstrate the impact such an attack could have on an organization. As threat actors' use of this technique increases in frequency, it is imperative that defenders plan and take precautions.

Posted in

- Threat Intelligence
