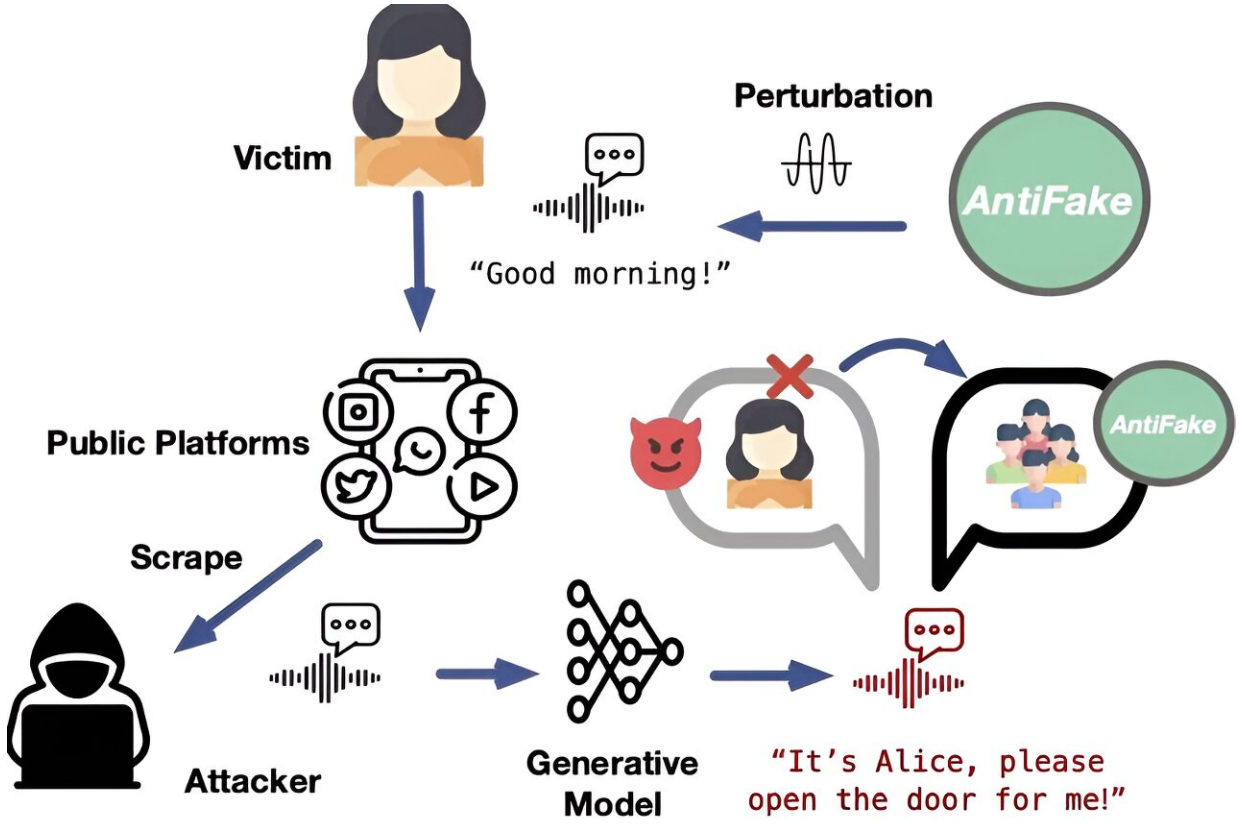
Defending your voice against deepfakes

November 27 2023, by Shawn Ballard



Overview of how AntiFake works. Credit: Ning Zhang, McKelvey School of Engineering, Washington University in St. Louis

Recent advances in generative artificial intelligence have spurred developments in realistic speech synthesis. While this technology has the potential to improve lives through personalized voice assistants and accessibility-enhancing communication tools, it also has led to the

emergence of deepfakes, in which synthesized speech can be misused to deceive humans and machines for nefarious purposes.

In response to this evolving threat, Ning Zhang, an assistant professor of computer science and engineering at the McKelvey School of Engineering at Washington University in St. Louis, has developed a tool called AntiFake, a novel defense mechanism designed to thwart unauthorized speech synthesis before it happens. Zhang presented AntiFake Nov. 27 at the Association for Computing Machinery's Conference on Computer and Communications Security in Copenhagen, Denmark.

Unlike traditional deepfake detection methods, which are used to evaluate and uncover synthetic audio as a post-attack mitigation tool, AntiFake takes a proactive stance. It employs adversarial techniques to prevent the synthesis of deceptive speech by making it more difficult for AI tools to read necessary characteristics from voice recordings. The code is freely available to users.

"AntiFake makes sure that when we put voice data out there, it's hard for criminals to use that information to synthesize our voices and impersonate us," Zhang said. "The tool uses a technique of adversarial AI that was originally part of the cybercriminals' toolbox, but now we're using it to defend against them. We mess up the recorded audio signal just a little bit, distort or perturb it just enough that it still sounds right to human listeners, but it's completely different to AI."

To ensure AntiFake can stand up against an ever-changing landscape of potential attackers and unknown synthesis models, Zhang and first author Zhiyuan Yu, a graduate student in Zhang's lab, built the tool to be generalizable and tested it against five state-of-the-art speech synthesizers. AntiFake achieved a protection rate of over 95%, even against unseen commercial synthesizers. They also tested AntiFake's

usability with 24 human participants to confirm the tool is accessible to diverse populations.

Currently, AntiFake can protect short clips of speech, taking aim at the most common type of voice impersonation. But, Zhang said, there's nothing to stop this tool from being expanded to protect longer recordings, or even music, in the ongoing fight against disinformation.

"Eventually, we want to be able to fully protect voice recordings," Zhang said. "While I don't know what will be next in AI voice tech—new tools and features are being developed all the time—I do think our strategy of turning adversaries' techniques against them will continue to be effective. AI remains vulnerable to adversarial perturbations, even if the engineering specifics may need to shift to maintain this as a winning strategy."

More information: Zhiyuan Yu et al, AntiFake: Using Adversarial Audio to Prevent Unauthorized Speech Synthesis, *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security* (2023). DOI: 10.1145/3576915.3623209

Provided by Washington University in St. Louis