DeFake: Using Adversarial Audio Perturbations to Prevent Malicious Voice Cloning

Motivation. Voice cloning technology, while innovative, harbors significant potential for misuse, including identity theft, fraud, and spreading misinformation, which can have far-reaching and detrimental impacts on individuals and society. Therefore, it is essential to establish regulations and technologies to ensure that this powerful tool is employed responsibly and for the betterment of communication and technology, not for deceit or harm.

Proposed Solution - DeFake. Building on the key observation is that voice cloning has to rely on consumers' pre-existing speech samples, which are generally collected from the public space (such as social media), we propose defake, a protective mechanism to add carefully crafted perturbations to voice samples to hinder the cyber criminal's cloning process. DeFake's technical appproach is based on theory of adversarial robustness, a fundamental theoretical weakness of modern AI systems, which simply cannot be patched, unlike conventional software vulnerabilities. Protective noises are added to the original voice in a manner that the processed sample still sounds like the victim to humans, when it is used for speech synthesis by the attacker, the resulting synthetic speech would resemble others' voices rather than the victim's. Consequently, the threat is mitigated while the usability of the speech sample is maintained.

Administrability and Feasibility to Execute. DeFake works by eliminating usable sample for malicious voice cloning, addressing the prevention or authentication intervention point. Theoretically, by perturbing voice inputs to escape the domain of the victim's voices, DeFake inherently obstructs cloning. It can be deployed on various platforms across different technologies. To demonstrate the feasibility, we have developed a prototype and conducted extensive experiments, including user studies, to validate the practicality and efficacy of the system.

Increased Company Responsibility, Reduced Consumer Burden. DeFake can be deployed by both companies and individual consumers using an app since it works by perturbing audio, and this digital processing can be on upstream or downstream. However, upstream actors (such as social media or streaming companies) that publish audio samples are better entities for implementing our proposed mechanism for both performance and security. This technology can be packaged and invoked via voice commands, such as Alexa and Siri, making it accessible to different populations.

Resilience. DeFake is based on adversarial robustness, a theoretical weakness of modern AI systems that cannot be easily patched. Therefore, the high-level construction of the protection will remain the same, even though the concrete techniques for attack and defense will continue to evolve as the AI community continues to introduce new breakthroughs in various capabilities. As new attack mechanisms and new voice cloning technologies are expected to be developed, they can be integrated into the DeFake protection system to further improve the generalizability of the protective noises. To test the resiliency, we conducted a series of experiments to test how attackers can use advanced techniques to invalidate or remove the perturbations. We test two latest techniques in the existing research, and results show that DeFake introduces a dilemma where stronger filtering can break the protection provided by DeFake, but it will simultaneously undermine audio significantly to make it not useful for speech synthesis.