De-AntiFake: Rethinking the Protective Perturbations Against Voice Cloning Attacks

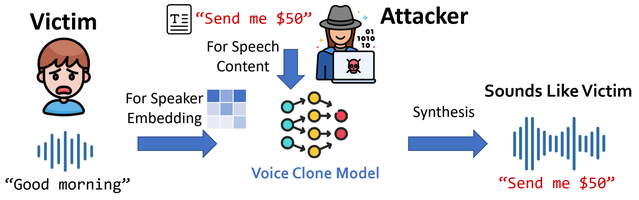
Wei Fan, Kejiang Chen, Chang Liu, Weiming Zhang, Nenghai Yu

University of Science and Technology of China

ICML International Conference On Machine Learning

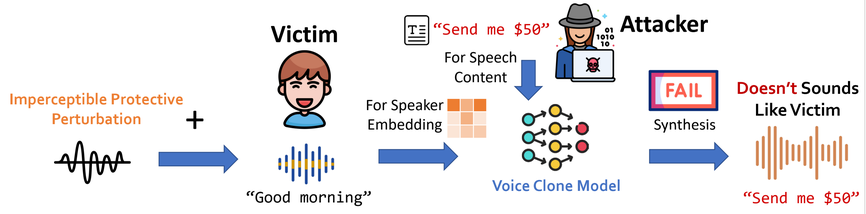
Motivation

❶ Voice Cloning Attacks Bring Security Risks



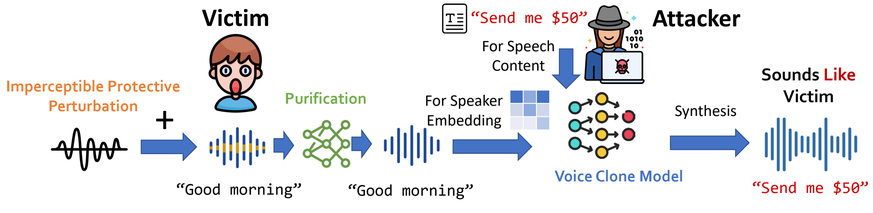
😈 Deceive human and speaker verification, carry **security and privacy risks**

❷ Current Active Defenses Against Voice Cloning



🛡️ Relying on **imperceptible adversarial perturbations** to prevent voice cloning models from generating victim-like speech

❸ But If the Attackers Try to *Purify* the Audio...



😟 If existing defenses are vulnerable to purification, they **may provide a false sense of security**

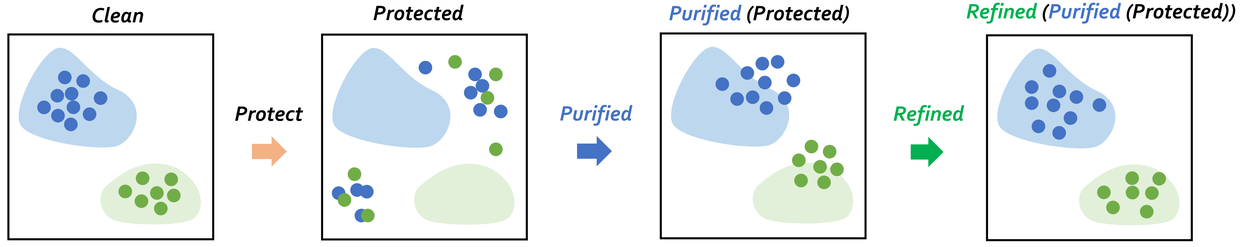
Our Work

1. Analyzing the Impact of Purification Attacks on Existing Perturbation-based Voice Cloning Defenses.

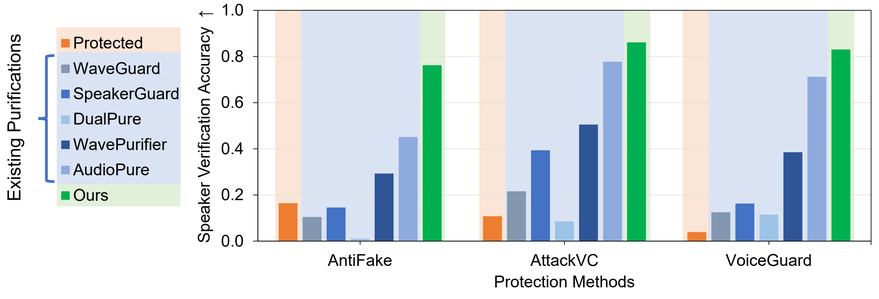
- ✓ **Conclusion:** Existing defenses are shown to be vulnerable to purification (at least **45.1%** protected samples can be bypassed).
- ✓ **Observation:** Existing purification introduces *distortions in voice cloning model embedding spaces*, which degrades voice cloning performance.

2. Proposing *PhonePuRe*: A Novel Purification Attack to Further Reveal the Brittleness of Existing Voice Cloning Defenses.

- ✓ **PhonePuRe: Two-Stage Framework** (**Pu**rification + **Phone**me-Guided **Re**finement)
  - ❑ Purification Stage: Preliminarily mitigate noise (unconditional diffusion).
  - ❑ Phoneme-Guided Refinement Stage: *Mitigate distortions in voice cloning model embedding spaces* (conditional diffusion).



- ✓ **Our Method Outperforms Existing Purification Methods, Increasing the Attack Success Rate: 45.1% → 76.2%.**



⚠️ Our work **reveals vulnerabilities** in existing voice cloning defenses, underscores the need for **more robust defenses to protect our voice**

Code and audio samples: de-antifake.github.io
Contact: range@mail.ustc.edu.cn

Project Page | Code