

PENG HUANG
Ph.D. candidate, Zhejiang University, China
College of Computer Science and Technology
IoT Security, Privacy Preserving Techniques
No. 2118 Pinlan Road, Hangzhou, China, 310000

+86-13970448752
penghuang@zju.edu.cn
Personal Webpage
Github

EDUCATION

Zhejiang University, Hangzhou, China — Sep. 2020 – Dec. 2025

Ph.D. Candidate of Computer Science and Technology (Advised by Prof. Zhongjie Ba)
Research Areas: IoT Security, AI Security, Speech Signal Processing

Zhejiang University, Hangzhou, China — Sep. 2016 – Jun. 2020

B.E. of Information Science and Electrical Engineering

INTERNSHIP

Huawei's Independent Cyber Security Lab (ICSL) — 2024.08 - 2024.12

Project: Black-box audio adversarial attacks targeting on edge LLM agents.

PUBLICATIONS

- **[In Submission]** Your input is real: Remember Real Voice Characteristics Against Malicious Machine-crafted Injections. *Tiantian Liu, Changjin Yu, **Peng Huang**, Liang Xiao, Feng Lin, Zhongjie Ba, and Kai Lin.* ACM CCS.
- **[MobiSys'25]** SecHeadset: A Practical Privacy Protection System for Real-time Voice Communication. ***Peng Huang**, Kun Pan, Qinglong Wang, Peng Cheng, Zhongjie Ba, Li Lu, and Kui Ren. MobiSys, 2025. (Acceptance rate: 18.0% = 42 / 233)*
- **[NDSS'23]** InfoMasker: Preventing Eavesdropping Using Phoneme-Based Noise. ***Peng Huang**, Yao Wei, Peng Cheng, Zhongjie Ba, Li Lu, Feng Lin, Fan Zhang, and Kui Ren. NDSS, 2023. (Acceptance rate: 16.2% = 94 / 581)*
- **[TDSC]** Phoneme-Based Proactive Anti-Eavesdropping with Controlled Recording Privilege. ***Peng Huang**, Yao Wei, Peng Cheng, Zhongjie Ba, Li Lu, Feng Lin, Yang Wang, and Kui Ren. IEEE TDSC, 2024.*
- **[Okland'24]** ALIF: Low-Cost Adversarial Audio Attacks on Black-Box Speech Platforms using Linguistic Features. *Peng Cheng, Yuwei Wang, **Peng Huang**, Zhongjie Ba, Xiaodong Lin, Feng Lin, Li Lu, and Kui Ren. IEEE S&P, 2024. (Acceptance rate: 17.8% = 258 / 1449)*
- **[Okland'24]** Text-CRS: A Generalized Certified Robustness Framework against Textual Adversarial Attacks. *Xinyu Zhang, Hanbin Hong, Yuan Hong, **Peng Huang**, Binghui Wang, Zhongjie Ba, and Kui Ren. IEEE S&P, 2024. (Acceptance rate: 17.8% = 258 / 1449)*
- Forged Voice Detection in Real-World Fraud Communication Scenarios. *Z. Xu, P. Cheng, Z. Ba, **P. Huang**, K. Ren. Journal of Cyber Security, 2025.*

RESEARCH EXPERIENCE

• Implicit NSFW Image Detection via Knowledge Extraction from Vision Language Models and State Machine Planning — 2025 – Now

– Extract detection logic for sensitive images using multimodal large models, and perform structured rule modeling with state machines to enable adaptive, high-precision detection of implicit sensitive content across domains, with interpretable attribution.

• Unsupervised Detection of Voice Injection Attacks via Multi-view Fusion and Dynamic Memory — 2025 – Now

– Investigate unified detection of in-band and out-of-band voice injection attacks by designing an unsupervised detection framework that leverages multi-view feature fusion and global-local representation learning.

Incorporate a dynamic memory library to model the distribution of genuine speech, enabling robust anomaly detection without attack samples and achieving over 97% accuracy across diverse scenarios.

- **Privacy-preserving Voice Communication** 2023 – 2024
  - This research focuses on building a system for protecting voice privacy in real-time voice communication. Rather than traditional encryption, we propose to protect voice privacy based on voice obfuscation, which maintains its effectiveness despite communication channel distortions and has been validated in various real-world scenraios.

- **Adversarial Attacks Against ASR** 2022 – 2024
  - This research focuses on developing adversarial attacks against black-box automatic speech recognition (ASR) systems. We manipulate target audio waves at the MFCC feature level and then re-synthesize them using pretrained vocoders. Our approach exploits the robustness paradox of modern ASR systems - while they're designed to recognize speech under various distortions, this same robustness makes them vulnerable to carefully crafted adversarial examples. This work has been included in *garak*, a LLM vulnerability scanner, which has 4.5k+ stars on GitHub.
  - I propose another attack targets on the language model in ASR systems. The attack is based on the observation that ASR systems often utilize language models to predict next language units, which could be utilized to reduce the attack audio energy to make adversarial examples more secretly.

- **Defense of Hidden Voice Recorders** 2021 – 2023
  - This research focuses on preventing unauthorized voice recording by hidden recorders. We propose a system that generates a novel phoneme-based noise to distorted the target voice. The noise is transmitted via inaudible ultrasound, which would not affect users' normal communications. We optimize the ultrasound transmission system to increase the effective range. Besides, we also provide recording privilege for authorized users.

- **Modeling Speaker Fingerprinting with Volterra Series** 2020 – 2021
  - The research focuses on fingerprinting hardware speakers based on the microphones' onset self-oscillation curve. We propose a challenge-response scheme to extract robust microphone fingerprints from the curves of different frequencies.

Professional Services

- **Reviewer**
  IEEE TIFS, IEEE IoT-J, ACM TIoT.

Award

- ACM MobiSys Raising Star, 2026
- ACM MobiSys Student Travel Grant, 2026
- Outstanding Graduate Student of the year, Zhejiang University, 2023
- Undergraduate Academic Scholarship, Zhejiang University, 2016-2019
- Zhejiang Provincial Government Scholarship, Zhejiang University, 2017

Technical Skills

Coding: Machine & Deep Learning (Python), Signal Processing & Simulation (MATLAB), Writing (LATEX)
MISC: Linux, Development Boards.