PENG CHENG

pengcheng326@hotmail.com • (86) 15057169279 • https://pengcheng-tech.github.io/ • Hangzhou, Zhejiang, China

Research Interests: AIGC security, IoT security

PROFESSIONAL EXPERIENCE

Researcher, State Key Laboratory of Blockchain and Data Security, Zhejiang University

June 2024 – Present

Member of the AI Data Security Team

- Conducted research on AI-generated content (AIGC) security
- Developing the *DFscan platform* for multimodal deepfake detection. Leading the technical team in the platform design and development

---

Qiushi Research Fellow, ZJU-Hangzhou Global Scientific and Technological Innovation Center

December, 2023 - May, 2024

Member of the Cyberspace Security Research Institute.

- Conducted research on multimodal data security and privacy protection technologies.
- Developed multimodal deepfake detection systems.

Postdoctoral Researcher, Zhejiang University

January, 2021 - November, 2023

College of Computer Science and Technology

- Researched voice security and privacy protection in human-computer interaction scenarios.
- Postdoctoral supervisor: Prof. Kui Ren (Qiushi Chair Professor, AAAS/ACM/CCF/IEEE Fellow, Dean of the College of Computer Science and Technology of Zhejiang University)

EDUCATION

Ph.D. Degree, Lancaster University

October, 2016 - December, 2020

School of Computing and Communications

- Major: Computer Science
- Thesis: *Acoustic-channel Attack and Defence Methods for Personal Voice Assistants*.
- Supervisors: Prof. Utz Roedig (Full Professor of Computer Science at University College Cork) and Prof. Jeff Yan (Full Professor of Cyber Security at the University of Southampton)

Dual Master's Degree, KU Leuven, Belgium & Tsinghua University, China

September, 2012 - September, 2015

Department of Electrical Engineering, KU Leuven & School of Integrated Circuits, Tsinghua University

- Major: Electrical Engineering (KU Leuven) & Integrated Circuit Engineering (Tsinghua University)
- Supervisor: Prof. Guoqiang Bai (Associate Professor at Tsinghua University)

B.Eng. Degree, Beijing University of Posts and Telecommunications, China

September, 2008 - June, 2012

School of Electronic Engineering

- Major: Electronic Science and Technology

RESEARCH PROJECTS

Total Funding: £7.01M (British Pounds; all figures in GBP)

National Natural Science Foundation of China (NSFC)

- Principal Investigator: "Research on Speech Synthesis Data Compliance Management Technology Based on Intrinsic Characteristics of Audio Signals"，(2025-2028, NSFC General Program Project, £54K)
- Participant: "High-Performance Visual Perception Models Using Deep Learning" (2023–2026, £59K)
- Participant: "Cross-Chain Security in Heterogeneous Blockchain Networks" (2023–2027, NSFC Key Project, £309K)
- Participant: "Research on Voice Attack and Defense Based on the Physical Characteristics of Smart Device Sensing Components" (2022–2025, £64K)

Key R&D Programs of China

- Participant: "Multimodal Network Environment Construction Technology Based on Public Cloud-Network Resources", 2024-2027 (£1.64M)
- Participant: "Aggregation and Transfer of Machine Learning Models" under the *National Science and Technology Innovation 2030 Initiative - New Generation Artificial Intelligence*, 2021 – 2025 (£1.51M)
- Participant: "Security Protection Technology for Industrial Control Programming Platforms Based on Domestic Cryptographic Algorithms", 2022 – 2024 (£1.60M)

Provincial, Municipal, and University-Level Research Projects

- Participant, the Key R&D Programme of Zhejiang Province, 2025 (£272K)
- Participant, Hangzhou Key R&D Program: "Key Technologies and Platform Development for Security Detection of Large AI Models", 2024 – 2027 (£1.83M)
- Principal Investigator, Hangzhou West Innovation Corridor Development Special Fund: "Toolchain for Deep Synthetic Content Analysis Based on Physical Attribute Attribution", 2024 – 2026 (£33K)

Industry Collaboration

- Principal Investigator, **Zhejiang University-Alibaba Collaboration Project**: "Active and Passive Security Protection Technologies for the *Maojing* Voice Interaction System", 2025 – 2026 (£41K)
- Participant, **Zhejiang University-Ant Group Joint Laboratory Industry Collaboration Project**: "Security Risk Detection and Alignment Strategies for Large Model-based AI Agents", 2025 (£65K)
- Participant, **China Southern Power Grid Research Institute Co., Ltd.**: "Research and Development of AI-Driven Automated Security Detection Technologies and Components for Power Systems", 2024 – 2026 (£255K)
- Participant, **CRRC Zhuzhou Electric Locomotive Research Institute Co., Ltd.**: "Deep Learning Algorithm Evaluation and Security Verification Platform", 2024 – 2027 (£127K)
- Participant, **China Southern Power Grid Research Institute Co., Ltd.**: "Research on AI Attack-Defense Library Design and Test Component Development for New Power System Scenarios (2023)", 2023 – 2025 (£163K)

PUBLICATIONS

Five Most Relevant/Representative Publications:

- Ba, Z., Zhong, J., Lei, J., **Cheng, P.**[*], Wang, Q., Qin, Z., Wang, Z., Ren, K. *SurrogatePrompt: Bypassing the Safety Filter of Text-to-Image Models via Substitution*. Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS 2024), 1166–1180.
- **Cheng, P.**, Wang, Y., Huang, P., Ba, Z., Lin, X., Lin, F., Lu, L., Ren, K. *ALIF: Low-Cost Adversarial Audio Attacks on Black-Box Speech Platforms Using Linguistic Features*. IEEE Symposium on Security and Privacy (SP 2024), 1628–1645.
- Ba, Z., Wen, Q., **Cheng, P.**[*], Wang, Y., Lin, F., Lu, L., Liu, Z. *Transferring Audio Deepfake Detection Capability Across Languages*. Proceedings of the ACM Web Conference (WWW 2023), 2033–2044.

- **Cheng, P.**, Wu, Y., Hong, Y., Ba, Z., Lin, F., Lu, L., Ren, K. (2023). *UniAP: Protecting Speech Privacy With Non-Targeted Universal Adversarial Perturbations*. IEEE Transactions on Dependable and Secure Computing (TDSC, JCR Q1), 21(1), 31–46.
- **Cheng, P.**, Roedig, U. (2022). *Personal Voice Assistant Security and Privacy—A Survey*. Proceedings of the IEEE, 110(4), 476–507.

Other Publications (Reverse Chronological Order)

- Liu, Q., Zhang, Y., Ba, Z., Shuai, C., **Cheng, P.**, Zheng, T., Wang, Z. *Attack-Resistant Watermarking for {AIGC} Image Forensics via Diffusion-based Semantic Deflection*. To appear at the 14th International Conference on Learning Representations (ICLR 2026).
- Ba, Z., Yi, L., **Cheng, P.***, Li, Q., Wang, Q., Lu, L. *Beyond Content: A Comprehensive Speech Toxicity Dataset and Detection Framework Incorporating Paralinguistic Cues*. To appear at the 40th AAAI Conference on Artificial Intelligence (AAAI 2026).
- Dong, Z., Shuai, C., Ba, Z., **Cheng, P.**, Qin, Z., Wang, Q., Ren, K. (2025). *WMCopier: Forging Invisible Watermarks on Arbitrary Images*. The 39th Conference on Neural Information Processing Systems (NeurIPS 2025).
- REN, K., LIN, F., BA, Z., LIU, Z., **Cheng, P.** (2025) *Deepfake detection: key challenges and technical approaches*. Computing Magazine of the CCF. 1(2): 8−15.
- Huang, P., Pan, K., Wang, Q., **Cheng, P.**, Lu, L., Ba, Z., Ren, K. (2025). *SecHeadset: A Practical Privacy Protection System for Real-time Voice Communication*. Proceedings of the ACM MobiSys 2025.
- Ba, Z., Gong, B., Wang, Y., Liu, Y., **Cheng, P.***, Lin, F., Lu, L., Ren, K. (2024). *Indelible "Footprints" of Inaudible Command Injection*. IEEE Transactions on Information Forensics and Security (TIFS, JCR Q1).
- Huang, P., Wei, Y., **Cheng, P.**, Ba, Z., Lu, L., Lin, F., Wang, Y., Ren, K. (2024). *Phoneme-Based Proactive Anti-Eavesdropping with Controlled Recording Privilege*. IEEE Transactions on Dependable and Secure Computing (TDSC, JCR Q1).
- Huang, P., Wei, Y., **Cheng, P.**, Ba, Z., Lu, L., Lin, F., Zhang, F., Ren, K. *InfoMasker: Preventing Eavesdropping Using Phoneme-Based Noise*. Network and Distributed System Security Symposium (NDSS 2023).
- **Cheng, P.**, Sankar, M. S. A., Bagci, I. E., Roedig, U. Adversarial Command Detection Using Parallel Speech Recognition Systems. Computer Security - ESORICS 2021 International Workshops, 238–255.
- **Cheng, P.**, Bagci, I. E., Roedig, U., Yan, J. (2020). *SonarSnoop: Active Acoustic Side-Channel Attacks*. International Journal of Information Security (IJIS, JCR Q2), 19(2), 213–228. **Finalist for the "Most Innovative Research" Pwnie Award at Black Hat USA 2019**.
- **Cheng, P.**, Bagci, I. E., Yan, J., Roedig, U. *Smart Speaker Privacy Control—Acoustic Tagging for Personal Voice Assistants*. IEEE Security and Privacy Workshops (SPW 2019), 144–149.
- **Cheng, P.**, Bagci, I. E., Yan, J., Roedig, U. *Towards Reactive Acoustic Jamming for Personal Voice Assistants*. Proceedings of the 2nd International Workshop on Multimedia Privacy and Security (2018), 1–13.

Preprints/Under Review

- Gong, B., **Cheng, P.***, Ba, Z., Lu, L., Tang, J., Hou, J., Xue, S., Ren, K. (2025) *Bridging the Synthesis-Detection Gap: A Chinese Audio Deepfake Dataset and Industrial-Scale Retrieval-Augmented Detection*.
- Ba, Z., Zhang, Y., **Cheng, P.***, Gong, B., Zhang, X., Wang, Q., Ren, K. (2025). *Robust Watermarks Leak: Channel-Aware Feature Extraction Enables Adversarial Watermark Manipulation*. arXiv:2502.06418
- Gong, B., Shuai, C., Wen, Q., **Cheng, P.**, Wang, Q., Ba, Z., Wang, Z., Ren, K. (2025). *Test-Time Adaptation for Audio Deepfake Detection*.
- Ba, Z., Fu, H., Yang, Y., Chen, H., Wang, Q., **Cheng, P.**, Qin, Z., Ren, K. (2025). *JudgeRail: Har- nessing Open-Source LLMs for Fast Harmful Text Detection with Judicial Prompting and Logit Rectification*.
- Wen, Q., **Cheng, P.**, Ba, Z., Yi, L., Qin, Z., Lu, L., Wang, Q., Ren, K. (2024). *CLINDA: A Cross-lingual Domain Adaptation Framework for Challenging Audio Deepfake Detection Tasks across Languages*.
- Lei, J., Wang, Q., **Cheng, P.**, Ba, Z., Qin, Z., Wang, Z., Liu, Z., Ren, K. (2023). *Masked Diffusion Models Are Fast and Privacy-Aware Learners*. arXiv:2306.11363

Thesis

- **Cheng, P.** (2020). *Acoustic-Channel Attack and Defence Methods for Personal Voice Assistants*. Ph.D. Dissertation, Lancaster University.Granted Patents

STUDENT SUPERVISION AND MENTORSHIP

- Co-supervising **5 Ph.D. students**, **7 Master's students** and **one undergraduate student** at Zhejiang University on research projects in AIGC security and multimodal privacy
- Co-supervised a Zhejiang University Master's student to win the **National Graduate Scholarship (China)-the highest-level scholarship awarded to Master's students in China** (Oct 2024).
- Successfully mentored **1 PhD student**, **5 Master's students**, and **5 undergraduates** to degree completion, guiding thesis design, research execution, and publication strategies.
- Teaching Assistant for lab/practical sessions: *CS4615: Computer Systems Security*, University College Cork, Ireland; *SCC110: Software Development*, Lancaster University, UK.

HONORS & AWARDS

- **Third Place** | *IJCAI 2025 Deepfake Detection Challenge* (2025), **Advisor**, Track: Audio-Visual Detection and Localization (DDL-AV)
- **National Grand Prize (Top-Tier Award)** | *19th "Challenge Cup" National Competition for Extracurricular Academic Science and Technology Works* (2024), **Advisor** (Ranked 2nd among 3 advisors), Project: *"Multimodal AI Audit Matrix: Deepfake Detection and NSFW Content Regulation Platform"*
- **Top 5 Nationwide** | *3rd China Artificial Intelligence Competition* (2021), **Primary Advisor** (Ranked 1st among 2 advisors), Track: *Audio Deepfake Detection Under Open-Speaker Scenarios*.
- **Top 5 Nationwide** | *3rd China Artificial Intelligence Competition* (2021), **Primary Advisor** (Ranked 1st among 2 advisors), Track: *Speaker-Specific Audio Deepfake Detection*.
- **Finalist for "Most Innovative Research" Pwnie Award** | *Black Hat USA 2019*, **First Author** (1st of 4 contributors) for research titled *"SonarSnoop: Active Acoustic Side-Channel Attacks"*.
- **Postdoctoral Excellence Grant (Second Class)** | Zhejiang Provincial Department of Human Resources and Social Security (Aug 2021).
- **Ph.D. Scholarship** | Faculty of Science and Technology, Lancaster University, UK (2016–2020).

IMPACTS

Industry Contributions

1. Vulnerabilities in Text-to-Image AI Models

- Proposed *Surrogate Prompt, a prompt attack method* to bypass safeguards of commercial large-scale AI models (e.g., Midjourney, Stability.ai), enabling systematic generation of policy-violating content. Findings were **acknowledged and taken into consideration for security improvement** by these leading vendors.

2. Attacks on AIGC Watermarking Systems

- Developed *WMCopier*, an adversarial watermark forgery method leveraging diffusion inversion techniques to compromise commercial watermarking systems. **Recognized by Amazon's Responsible AI Team for identifying critical vulnerabilities**, leading to collaborative improvements in their defensive frameworks. Received official thank-you letter.

3. Open-Source Contributions

- Developed and open-sourced *ALIF, the first black-box adversarial attack framework* leveraging linguistic features to generate **imperceptible adversarial audio** that effectively bypasses human perception while compromising speech AI systems. **Adopted by NVIDIA** for integration into their **official AI security toolkit** to advance red teaming capabilities (*under active development*).

Media Recognition

- **Developed** *SonarSnoop: Active Acoustic Side-Channel Attacks*, a novel methodology exploiting smart device speakers and microphones to extract sensitive data. Garnered attention from IT media (*Motherboard, ZDNet, Sophos*). Praised by **Bruce Schneier** (renowned cryptographer) and **Prof. Ross Anderson** (Well-known security researcher) on X for its novelty.

GRANTED PATENTS

- Qian, Y., Zhang, X., Wang, Q., Ba, Z., **Cheng, P**., et al. "Review System, Method, Computer Device, and Medium for Image Sensitive Elements", *China, Patent No. CN106610969A[P]*, Granted: 02-Apr-2025.

- Ba Zhongjie, Wu, Y., **Cheng, P**., et al. “Privacy Protection Method and Device Using White-Box Speech Adversarial Examples”, *China, Patent No. CN2022109965917[P]*, Granted: 03-Dec-2024.
- Ba, Z., Zheng, Q., **Cheng, P**., et al. “Enhanced Deepfake Image Detection Method and Device Based on Generative Adversarial Networks (GANs)”, *China, Patent No. CN2024100814592[P]*, Granted: 21-Jun-2024.
- Ba, Z., Wang, Y., **Cheng, P**., et al. “Security Evaluation Method for Speech Recognition Models via Semantic Space Perturbation”, *China, Patent No. CN116758899B*, Granted: 13-Oct-2023.
- Huang, P., Ba, Z., **Cheng, P**., et al. “Privacy Protection Method, System, and Medium for Voice Communication Based on Voice Obfuscation”, *China, Patent No. CN119449493B*, Granted: 08-Jul-2025.
- Gong, B., Huang, P., Ba, Z., **Cheng, P**., et al. “Cross-Domain Detection Method and Device for Deep Synthetic Audio Based on Self-Supervised Auxiliary Tasks”, *China, Patent No. CN119479611B*, Granted: 29-Apr-2025.

ACADEMIC SERVICES

Journal Editorial Roles

- **Special Issue Initiator and Guest Editor**: Special Issue on “Intelligent Voice Security and Defense Technologies”, *Journal of Cyber Security*, 2025.

Conference Program Committees

- **Program Committee Member**: ACM Web Conference (WWW 2025), AAAI Conference on Artificial Intelligence (AAAI 2026).

Journal Reviewer Roles

- **English Journals**: *Proceedings of the IEEE*, *ACM Transactions on Internet of Things (TIOT)*, *IEEE Internet of Things Journal (IoT-J)*.
- **Chinese Journal**: *Journal of Information Network Security*.

OTHER ACADEMIC ACTIVITIES

- Contributed to the preparation of a €720,839 grant proposal (*Science Foundation Ireland – SFI*) on *Security and Privacy of Personal Voice Assistants* under Prof. Utz Roedig (2019). The proposal is anchored in my doctoral research.
- Visiting Scholar at the Department of Computer Science, University College Cork (UCC), Ireland (2019–2020).
- Participated in the *São Paulo Advanced Science School (ESPCA)* on Smart Cities, hosted by the University of São Paulo (2017).
- Selected as **one of 75 global top graduate students and postdoctoral researchers** (sponsored by the *São Paulo Research Foundation – FAPESP*).

GOOGLE SCHOLAR PAGE

https://scholar.google.com/citations?user=hz3kfCEAAAAJ&hl=en