

XUANJUN CHEN

D12942018@ntu.edu.tw · xjchen.tech · Google Scholar · linkedin.com/in/jun-ntu

SUMMARY

Xuanjun is a Ph.D. candidate at National Taiwan University, advised by [Hung-yi Lee](#) and [Jyh-Shing Roger Jang](#). His research explores the intersection of large language models and speech/audio processing, including foundation models for audio understanding and generation, as well as agentic LLMs. He has published **15 first/co-first** peer-reviewed papers at COLM, TIST, TASLP, ICASSP, INTERSPEECH, ASRU, SLT, etc. His research has been recognized with:

- IEEE SPS Scholarship · Best Student Paper Nomination (ASRU '25, 11/447 papers)

EDUCATION

National Taiwan University (NTU)

- Ph.D. in Communication Engineering, GPA: 4.3/4.3 Sep 2023 – Present
- M.S. in Computer Science and Information Engineering, GPA: 4.19/4.3 Sep 2020 – Jan 2023

National Taiwan University of Science and Technology (Taiwan Tech)

- B.S. in Computer Science and Information Engineering, GPA: 4.11/4.3 Jun 2020
- Certificate of Achievement: Top 5% in CSIE across three semesters Fall '18, Fall '19, Spr. '20

RESEARCH INTERESTS & SKILLS

Research Interests: Large Language Models, Audio Language Models, Audio Generation, Audio Deepfakes
Coding & Toolkit: Python; PyTorch, PyTorch Lightning, Hugging Face Transformers

RESEARCH EXPERIENCE

NVIDIA Academic Grant Program, Lead Student Researcher, Taipei Taiwan Oct 2026 – Mar 2027 (expected)

- **Audio Generation.** Leading research on fine-grained text-to-audio model to enhance audio generation capabilities.

Shanda Group, Research Intern (Remote), Tokyo Japan Apr 2026 – Jul 2026

- **Audio Generation.** Investigated whether LLM agents can automate reinforcement learning for Text-to-Speech models. [1]

National Taiwan University, Research Assistant (Ph.D. & M.S.), Taipei Taiwan Oct 2020 – Present

- Published **15 first/co-first** peer-reviewed papers, with 5 more currently under review.
- **Audio Language Models.** Analyzed the post-training dynamics of large audio language models [2], authored a high-impact survey [7], and served as a core contributor to DeSTA2.5-Audio [3], Dynamic-SUPERB [4], and Codec-SUPERB [5] [6].
- **Retrieval and Agentic LLMs.** Pioneered a dual-loop agentic framework for optimizing benchmark generation [8]; improved retrieval via graph navigation [9] and delta planning [10], alongside a systematic RAG study for historical archives [11].
- **Audio Deepfakes.** Led projects on codec-based deepfakes [12]–[15], singing voice deepfakes [16]–[20], audio-visual deepfake localization and sync detection [21]–[22], adversarial attacks on ASV and active speaker detection [23]–[24], and conventional TTS/VC deepfakes [25], where our system ranked 3rd of 42 teams in ASVspoof 2021 LA; currently authoring an explainability survey.

RESEARCH HIGHLIGHTS

Total citations **600+** as of Oct 2026.

- [Can LLM Agents Automate Reinforcement Learning for Text-to-Speech?](#) ICASSP '27 (Submitted)
TL;DR: An LLM agent reconstructs and improves an RL post-training recipe for zero-shot TTS, cutting failure cases by half.
- [What Does LALM Post-training Learn? Rethinking Audio Understanding and Reasoning Capabilities](#) NAACL '27 (Submitted)
TL;DR: Much of the post-training gain comes from fitting answer formats and textual cues, not from better audio understanding.
- [MetaBench-Harness: Unlocking End-to-End Optimization of Benchmark Harnesses](#) ICLR '27 (Submitted)
TL;DR: A dual-loop framework where a meta-harness refines the harnesses that evolve benchmarks hard enough for frontier LLMs.
- [CodecFake+: Codec-Based Resynthesized Data as a Proxy for Detecting CodecFake Speech](#) TASLP 2026
TL;DR: Re-synthesized speech from 31 codecs plus a codec taxonomy as a training proxy for detecting 17 CoSG deepfake systems.
- [Towards Generalized Source Tracing for Codec-Based Deepfake Speech](#) ASRU 2025 · Best Student Paper Nominee
TL;DR: SASTNet fuses Whisper semantics with Wav2vec2/AudioMAE acoustics to trace CodecFake speech back to its source codec.
- [Localizing Audio-Visual Deepfakes via Hierarchical Boundary Modeling](#) ISCSLP 2026
TL;DR: HBMNet localizes manipulated audio-visual segments via coarse-to-fine boundary modeling, beating prior SOTA.
- [DeSTA2.5-Audio: Toward General-Purpose Large Audio Language Model with Self-Generated Cross-Modal Alignment](#) TASLP 2026[‡]
TL;DR: General-purpose LALM trained on 5M self-generated audio QA pairs; keeps the backbone LLM's language ability.
- [Dynamic-SUPERB Phase-2: A Collaboratively Expanding Benchmark for Spoken Language Models with 180 Tasks](#) ICLR 2025[‡]
TL;DR: 180 community-contributed tasks over speech, music, and audio; no current model handles all of them well.

First/co-first author unless marked [‡] (contributing author).

SELECTED PUBLICATIONS

* equal contribution; † co-second author; ‡ contributing author

Audio Language Models and Generation

- [1] X. Chen, et al., “Can LLM Agents Automate Reinforcement Learning for Text-to-Speech?” *ICASSP 2027 (Submitted)*
- [2] H.H. Chen*, X. Chen*, et al., “What Does LALM Post-training Learn? Rethinking Audio Understanding and Reasoning Capabilities,” *NAACL 2027 (Submitted)*
- [3] K.H. Lu, Z. Chen, ... X. Chen†, ... H. Lee, “DeSTA2.5-Audio: Toward General-Purpose Large Audio Language Model with Self-Generated Cross-Modal Alignment,” *TASLP, 2026*
- [4] C.Y. Huang, ... X. Chen†, ... S. Watanabe, H. Lee, “Dynamic-SUPERB Phase-2: A Collaboratively Expanding Benchmark for Measuring the Capabilities of Spoken Language Models with 180 Tasks,” *ICLR 2025*
- [5] H. Wu*, ... X. Chen†, ... H. Lee, “Codec-SUPERB: An In-Depth Analysis of Sound Codec Models,” *Findings of ACL 2024*
- [6] H. Wu, X. Chen, et al., “Codec-SUPERB @ SLT 2024: A Lightweight Benchmark for Neural Codec Models,” *SLT 2024*
- [7] H. Wu, X. Chen, et al., “Towards Audio Language Modeling – An Overview,” *Technical Report, Feb. 2024*

Retrieval and Agentic Large Language Models

- [8] X. Chen, et al., “MetaBench-Harness: Unlocking End-to-End Optimization of Benchmark Harnesses,” *ICLR '27 (Submitted)*
- [9] C.Y. Li*, X. Chen*, et al., “CodaRAG: Connecting the Dots with Associativity Inspired by Complementary Learning,” *ACM TIST, 2026*
- [10] W.C. Chou*, X. Chen*, et al., “Only Ask What You Don’t Know: Grounded Delta Planning for Efficient Multi-step RAG,” *COLM 2026*
- [11] C. Lin*, B.H. Feng*, X. Chen*, et al., “A Preliminary Study of RAG for Taiwanese Historical Archives,” *ROCLING 2025 · Best Paper Award*

Audio Deepfakes: Neural Codec, Singing Voice, Audio-Visual, Adversarial, and Conventional TTS/VC Scenarios

- [12] X. Chen*, et al., “CodecFake+: Codec-Based Resynthesized Data as a Proxy for Detecting CodecFake Speech,” *TASLP, 2026*
- [13] X. Chen*, I.M. Lin*, et al., “Towards Generalized Source Tracing for Codec-Based Deepfake Speech,” *ASRU 2025 · Best Student Paper Nominee*
- [14] X. Chen, et al., “Codec-Based Deepfake Source Tracing via Neural Audio Codec Taxonomy,” *INTERSPEECH 2025*
- [15] X. Chen, et al., “Mitigating Proxy-to-Wild Domain Gap in Deepfake Speech,” *ISCSLP 2026 · Selected Outstanding Paper*
- [16] X. Chen, et al., “Why Is Spectrogram ResNet Still Effective in Singing Voice Deepfake Detection?” *ICASSP '27 (Submitted)*
- [17] Y.A. Hsieh*, X. Chen*, et al., “On Bona Fide Audio Discrepancies in Singing Voice Deepfake Detection,” *ICASSP '27 (Submitted)*
- [18] X. Chen*, C.Y. Hu*, et al., “Joint Fullband-Subband Modeling for High-Resolution SingFake Detection,” *INTERSPEECH 2026 · Long Oral*
- [19] X. Chen, et al., “How Does Instrumental Music Help SingFake Detection?” *ICASSP 2026*
- [20] X. Chen, et al., “Singing Voice Graph Modeling for SingFake Detection,” *INTERSPEECH 2024 · Oral*
- [21] X. Chen, et al., “Localizing Audio-Visual Deepfakes via Hierarchical Boundary Modeling,” *ISCSLP 2026*
- [22] X. Chen, et al., “Multimodal Transformer Distillation for Audio-Visual Synchronization,” *ICASSP 2024*
- [23] X. Chen, et al., “Neural Codec-based Adversarial Sample Detection for Speaker Verification,” *INTERSPEECH 2024*
- [24] X. Chen, et al., “Push-Pull: Characterizing the Adversarial Robustness for Audio-Visual Active Speaker Detection,” *SLT 2022*
- [25] Y.L. Liao*, X. Chen*, et al., “Adversarial Speaker Distillation for Countermeasure Model on Automatic Speaker Verification,” *SPSC 2022*

SELECTED HONORS

Research Honors and Scholarship

- **IEEE Signal Processing Society Scholarship** One of dozens of international recipients 2026
- **NVIDIA Academic Grant Recipient** GPU compute resources via competitive global selection 2026
- **Selected Outstanding Paper (ISCSLP 2026)** The 15th International Symposium on Chinese Spoken Language Processing 2026
- **Best Student Paper Nomination (ASRU 2025)** Selected 11 out of 447 submitted papers (Top 2.5%) 2025
- **Best Paper Award (ROCLING 2025)** The 37th Conference on Computational Linguistics and Speech Processing 2025
- **Google Student Travel Grant** Sponsored by Google Asia Pacific to present at ICASSP 2024 2024
- **CTCI Foundation Science and Technology Scholarship** Bursary Award 2024 & Research Scholarship 2025 2024–2025

Service and Outreach

- **Co-Organizer**, Responsible Speech and Audio Generative AI, Special Session at IEEE ASRU 2025 2025
- **Technical Committee**, Codec Speech Proc. Universal PERFORMANCE Benchmark Challenge, IEEE SLT 2024 2024
- **Reviewer**: AAAI, ICLR, NeurIPS, ICML; ACL, EMNLP, NAACL, COLM; ICASSP, INTERSPEECH, ASRU, SLT, etc. 2023–Now