Yuancheng Wang

✉ yuanchengwang@link.cuhk.edu.cn · 🎓 Google Scholar · Homepage

Education

The Chinese University of Hong Kong, Shenzhen, Shenzhen Sep. 2023 – Present

Ph.D. Candidate Computer Science and Technology; Advisor: Prof. Zhizheng Wu; Expected Graduation: May 2026

- *Research interests*: Speech synthesis, speech representation learning, audio generation and editing, speech foundation models, and discrete diffusion models.
- First author of 8 papers submitted to top-tier AI conferences (ICML, NeurIPS, ICLR, ACL, IEEE SLT), with 2 currently under review; accumulated 760+ Google Scholar citations.
- Founder and core contributor & maintainer of the open-source audio, music, and speech generation toolkit *Amphion*, with 9,300+ stars.
- First author of foundation models for speech generation: *NaturalSpeech 3* (ICML 2024 Oral, Top 1%) and *MaskGCT* (ICLR 2025). Released *FACodec*, the core component of *NaturalSpeech 3*; *MaskGCT* gained 3,000+ stars within the first week of release.
- Recipient of the 2024 Duan Yongping Research Scholarship (Top 2%).

The Chinese University of Hong Kong, Shenzhen, Shenzhen Sep. 2019 – May 2023

B.Sc. Computer Science and Technology

- Awarded the Bowen Scholarship (2019–2023).
- Dean's list, School of Data Science.
- First Prize, Guangdong University Student Mathematics Competition (2020).

Selected Publications

First-/Co-First Author Publications:

1. **Yuancheng Wang**, et al. Metis: A Foundation Speech Generation Model with Masked Generative Pre-training. *submitted to NeurIPS 2025, under review. 5,5,5,5*
2. **Yuancheng Wang**, et al. TaDiCodec: Text-aware Diffusion Speech Tokenizer for Speech Language Modeling. *submitted to NeurIPS 2025, under review. 5,5,4,4*
3. **Yuancheng Wang***, Xueyao Zhang* et al. Advancing Zero-shot Text-to-Speech Intelligibility across Diverse Domains via Preference Alignment. ***ACL 2025 main***.
4. **Yuancheng Wang**, et al. MaskGCT: Zero-Shot Text-to-Speech with Masked Generative Codec Transformer. ***ICLR 2025***. (*over 3,000 GitHub Stars in a week, ranked #1 on GitHub trending for multiple days*)
5. Zeqian Ju*, **Yuancheng Wang***, Kai Shen*, Xu Tan*, et al. NaturalSpeech 3: Zero-Shot Speech Synthesis with Factorized Codec and Diffusion Models. ***ICML 2024***. *(Oral, Top 1%)*
6. Junyi Ao*, **Yuancheng Wang***, et al. SD-Eval: A Benchmark Dataset for Spoken Dialogue Understanding Beyond Words. ***NeurIPS 2024***.
7. **Yuancheng Wang**, et al. AUDIT: Audio Editing by Following Instructions with Latent Diffusion Models. ***NeurIPS 2023***.
8. Xueyao Zhang*, Liumeng Xue*, Yicheng Gu*, **Yuancheng Wang***, et al. Amphion: An Open-Source Audio, Music and Speech Generation Toolkit. ***IEEE SLT 2024***. *(over 9,000 GitHub Stars)*

Selected Collaborative Publications:

1. NVSpeech: An Integrated and Scalable Pipeline for Human-Like Speech Modeling with Paralinguistic Vocalizations. *submitted to AAAI 2026.*
2. AnyEnhance: A Unified Generative Model with Prompt-Guidance and Self-Critic for Voice Enhancement. ***IEEE TASLP.***

3. DualCodec: A Low-Frame-Rate, Semantically-Enhanced Neural Audio Codec for Speech Generation ***Interspeech 2025.***
4. Emilia: An Extensive, Multilingual, and Diverse Speech Dataset for Large-Scale Speech Generation. ***IEEE SLT 2024.*** *(Ranked #1 on Hugging Face trending for audio datasets)*
5. FoleyCrafter: Bring Silent Videos to Life with Lifelike and Synchronized Sounds. *(over 600 GitHub Stars)*
6. RALL-E: Robust Codec Language Modeling with Chain-of-Thought Prompting for Text-to-Speech Synthesis.

Full publication list available at: Google Scholar.

Internships

Meta, Superintelligence Labs May 2025 – Sep. 2025

Research Scientist Intern, California, USA

- Developed diffusion-based ultra-low-bitrate speech tokenizers and investigated their scaling laws, achieving a 6.25 Hz token rate for speech compression, understanding, and generation. Supported research on the Llama 4 speech model; related work is planned for submission to ICLR 2026.

ByteDance Apr. 2024 – Sep. 2024

Research Intern, Shenzhen, China

- Built a benchmark dataset for spoken dialogue understanding. Contributed to *SD-Eval*, accepted at NeurIPS 2024.

Microsoft Research Asia Nov. 2022 – Jul. 2023

Research Intern, Machine Learning Group (Advisor: Xu Tan), Beijing, China

- Built a large-scale zero-shot TTS system using speech codecs with attribute disentanglement and masked generative modeling. Resulted in *NaturalSpeech 3*, accepted as an Oral presentation at ICML 2024 (Top 1%).
- Conducted research on text-guided audio generation and editing with latent diffusion models. Contributed to *AUDIT*, accepted at NeurIPS 2023.

Research Projects

Speech Synthesis

NaturalSpeech 3: Large-Scale Speech Synthesis with Attribute Disentanglement Lead Researcher

- *NaturalSpeech 3: Zero-Shot Speech Synthesis with Factorized Codec and Diffusion Models*
- **Overview**
  - Designed a novel speech codec to disentangle prosody, content, timbre, and acoustic details.
  - Proposed a unified synthesis framework with factorized diffusion, achieving both robustness and efficiency, while significantly improving similarity, quality, prosody richness, and controllability.
  - Scaled non-autoregressive TTS to 1B parameters, trained on 200K hours of large-scale data.
  - Achieved near-human-level synthesis on the multi-speaker LibriSpeech benchmark.
- **Impact** Widely recognized in academia and industry; covered by Microsoft Research Asia official accounts, Machine Heart, and discussed on Reddit and Hacker News. Adopted as a benchmark by Alibaba, Tencent, ByteDance, Meta, Kayutai, and others.
- **Open Source** Released the core component *FACodec* in the *Amphion* toolkit, which has become a widely used baseline for speech disentanglement research.

MaskGCT: Masked Generative Modeling for Zero-Shot TTS Lead Researcher

- *MaskGCT: Zero-Shot Text-to-Speech with Masked Generative Codec Transformer*
- **Overview**

  - Proposed the first non-autoregressive TTS model without phoneme duration prediction.
  - Introduced a two-stage masked generative modeling (discrete diffusion) framework, eliminating the need for explicit phoneme-level alignment while enhancing similarity, quality, and controllability, especially for accents, emotions, tongue-twisters, and code-switching.
  - First large-scale non-autoregressive TTS model trained on the open-source Emilia dataset, outperforming prior methods in zero-shot synthesis.
- **Impact** Cited 80+ times within one year; adopted as a benchmark by Meta, StepFun, Alibaba, ByteDance, Microsoft, Minmax, Bilibili, and Tencent. Featured by multiple media outlets; highlighted by ModelScope as "the TTS open-source model that topped GitHub trending."
- **Open Source** Released in Amphion; received 3,000+ GitHub stars in the first week and topped GitHub trending for several days.

Metis: Foundation Model for Speech Generation with Masked Generative Pre-training Lead Researcher

- *Metis: A Foundation Speech Generation Model with Masked Generative Pre-training*
- **Overview**
  - Proposed a foundation speech generation model with masked generative (discrete diffusion) pre-training on 300K hours of multilingual data.
  - Unified framework across multiple tasks: TTS, voice conversion, speech enhancement, target speaker extraction, and lip-to-speech; outperformed task-specific expert models on multiple benchmarks.
  - Submitted to NeurIPS 2025; received unanimous high reviewer scores (5,5,5,5).
- **Demo** metis-demo.github.io
- **Open Source** Released foundation and fine-tuned models: huggingface.co/amphion/Metis

Post-Training Alignment

INTP: Post-Training Alignment Framework and Dataset for TTS Lead Researcher

- *Advancing Zero-shot Text-to-Speech Intelligibility across Diverse Domains via Preference Alignment*
- **Overview**
  - Addressed stability issues in complex scenarios (code-switching, cross-lingual, repeated words, tongue-twisters) by proposing a post-training alignment paradigm.
  - Built the INTP dataset: 250K preference pairs (2,000+ hours) covering diverse domains.
  - Extended DPO to multiple TTS paradigms (autoregressive, masked generative, discrete diffusion, flow matching), improving models including CosyVoice2, MaskGCT, and F5-TTS.
  - Studied generalization properties of alignment: cross-lingual, metric, and weak-to-strong transfer.
- **Demo** intalign.github.io
- **Open Source** Released the INTP dataset (1,000+ downloads): huggingface.co/datasets/amphion/INTP

Speech Representation

TaDiCodec: Ultra-Low Bitrate Speech Codec Lead Researcher

- *TaDiCodec: Text-aware Diffusion Speech Tokenizer for Speech Language Modeling*
- **Overview**
  - Achieved extreme compression with a single-codebook 6.25 Hz speech tokenizer, using a novel diffusion autoencoder without multi-stage training or adversarial losses.
  - Introduced text guidance into end-to-end codec training, greatly improving reconstruction quality and downstream synthesis, especially in code-switching and complex scenarios.
  - Submitted to NeurIPS 2025; received strong reviewer recommendations (5,5,4,4).
- **Demo** tadicodec.github.io
- **Open Source** Released code and models (0.5B, 3B, 4B autoregressive; 0.6B discrete diffusion): Diffusion-Speech-Tokenizer

DiSTok: Scaling Laws for Low-Bitrate Speech Codecs Lead Researcher

- *Diffusion Autoencoders are Scalable Discrete Speech Tokenizers*

- **Overview**
  - Led at Meta internship; explored scaling discrete codecs (12.5 Hz) for Llama 4 speech models.
  - Proposed DiSTok with CTC semantic supervision and diffusion reconstruction, achieving balance between compression, reconstruction quality, and representation.
  - Introduced lightweight diffusion heads, shortcut finetuning, and chunk-wise autoregressive diffusion to accelerate decoding.
  - Expanded codebook to 65,536 and trained on 2M hours of speech.
  - Submitted to ICLR 2026.

Multimodal Understanding and Generation

AUDIT: Instruction-Driven Audio Editing with Latent Diffusion — Lead Researcher

- *AUDIT: Audio Editing by Following Instructions with Latent Diffusion Models*
- **Overview**
  - Proposed a new task of natural-language-guided audio editing.
  - Built a large-scale framework for audio generation and editing with latent diffusion, leveraging synthetic paired data for high-quality supervised training.
- **Impact** Widely adopted as a benchmark after the rise of diffusion models; cited 80+ times; used by Meta, Microsoft, Adobe, Google, Sony, and Tencent.

SD-Eval: Multidimensional Benchmark for Spoken Dialogue — Co-First Author

- *SD-Eval: A Benchmark Dataset for Spoken Dialogue Understanding Beyond Words*
- **Overview**
  - Proposed SD-Eval, a benchmark focusing on paralinguistic (emotion, accent, age) and environmental factors, with 7,303 utterances (8.76h) from 8 public datasets.
  - Built a 1,052h training set with 724K utterances; conducted systematic evaluations across models using objective, subjective, and LLM-based metrics.

Open-Source Tooling

Amphion: Open-Source Toolkit for Audio, Music, and Speech Generation — Founder and Core Contributor

- *Amphion: An Open-Source Audio, Music, and Speech Generation Toolkit*
- **Overview**
  - Founder and core contributor of Amphion, with 9,300+ GitHub stars; largest code contributor.
  - Integrated mainstream audio and speech generation models, including AudioLDM, VALL-E, NaturalSpeech 2/3, FACodec, MaskGCT, and Metis.

Academic Service

Reviewer

- Conferences: International Conference on Learning Representations (ICLR), International Conference on Machine Learning (ICML), Conference on Neural Information Processing Systems (NeurIPS), Conference on Language Modeling (COLM)
- Journals: IEEE Transactions on Audio, Speech, and Language Processing (TASLP)

Teaching Assistant

- Fall 2024: CSC1001 “Introduction to Computer Science: Programming Methodology,” The Chinese University of Hong Kong, Shenzhen
- Spring 2025: DDA4300 “Optimization in Data Science and Machine Learning,” The Chinese University of Hong Kong, Shenzhen

Volunteer

- IEEE Spoken Language Technology Workshop (SLT), 2024

Invited Talks

Towards Natural and Efficient Speech Synthesis: Perspectives on Modeling, Alignment, and Representation Jun. 2025

Xmart Youth Forum (Online), SJTU X-LANCE Lab
Invited talk at the Xmart Youth Forum organized by the X-LANCE Lab, Shanghai Jiao Tong University [Slides] [Video]

Speech Generation with Masked Generative Modeling Apr. 2025

NUS Speech and Music AI Workshop, Singapore
Invited talk hosted by Prof. Ye Wang's group, National University of Singapore

MaskGCT: Zero-Shot Text-to-Speech with Masked Generative Codec Transformer Dec. 2024

OpenMMLab Community Open Talk (Online)
Joint invitation by OpenMMLab and SpeechHome [Video]

NaturalSpeech 3: Speech Disentanglement and Zero-Shot TTS in the Era of Big Data Mar. 2024

SpeechHome AI Tech Salon (Online)

Honors and Awards

- Duan Yongping Research Scholarship, 2024
- Bowen Scholarship, 2019–2023
- Dean's Commendation Award, School of Data Science
- First Prize, Guangdong University Student Mathematics Competition, 2020
- First Prize, National High School Mathematics Competition