Xintong Wang

xintong.wang@u.nus.edu · (+65) 9120 7974

https://github.com/oshindow · https://oshindow.github.io/

Highlights

Industry Experience: 2+ years of research experience as a Research Intern and Machine Learning Engineer at Xiaoice (formerly Microsoft), with work spanning automatic speech recognition, singing voice synthesis, and neural vocoders.
Research Focus: Speech understanding and generation, including ASR, L2 speech assessment, accent TTS, speech representation learning, and spoken language models.
Publication Record: First-author work published at *IEEE ASRU* and *Interspeech*, including an *Interspeech 2024 Oral* paper.

Education

National University of Singapore (NUS) Singapore, Singapore
Ph.D. in Computer Science, Advisor: Prof. Ye Wang *Aug. 2024 -- Present*

Beijing Forestry University (BJFU) Beijing, China
B.S. in Mathematics and Applied Mathematics, Advisor: Prof. Jing Wang *Sep. 2018 -- Jul. 2022*

Experience

Research Assistant Singapore, Singapore
National University of Singapore (Sound and Music Computing Lab) *Oct. 2023 -- Sep. 2024*

- Designed an an end-to-end model combining self-supervised speech representations with explicit pitch modeling for Mandarin mispronunciation detection and diagnosis.
- Achieved a 27.3% PER on non-native Mandarin speech while training exclusively on native speech, which led to Pitch-Aware RNN-T, an *Interspeech 2024 Oral* paper.

Machine Learning Engineer Beijing, China
Xiaoice (Waveland Team) *Jul. 2022 -- Oct. 2023*

- Developed, a multi-singer, multi-lingual singing voice synthesis acoustic model that supports cross-lingual singing voice synthesis using only monolingual singing data.
- Unified cross-lingual phonetic representations using the International Phonetic Alphabet and applied adversarial learning to reduce timbre-lingual entanglement, resulting in CrossSinger, an *ASRU 2023* paper.
- Developed an iSTFT-based neural vocoder for efficient, high-fidelity singing voice synthesis, improving perceptual naturalness over transposed-convolution-based vocoders

Research Intern Beijing, China
Xiaoice (Waveland Team) *May 2021 -- Jul. 2022*

- Built N-gram language models from a corpus of 50M+ sentences, improving ASR performance across all nine internal test scenarios and achieving up to a 33% relative CER reduction.
- Developed an automated pipeline that converted over 30K timestamped lyric annotations into structured MIDI files.
- Generated chord-conditioned mashups by linking compatible chorus sections from approximately 80% of the Chinese pop songs in the database.

Research Projects

Full-Duplex Spoken Language Models Singapore, Singapore

National University of Singapore, Ongoing Ph.D. Research *Jan. 2026 -- Present*

- Investigating full-duplex SLMs that compress redundant acoustic tokens into dense representations, with the goal of improving speech modeling while preserving low-latency interaction.
- Fine-tuned SLM model with role-playing prompts for controllable spoken dialogue generation.

Generalizable L2 Speech Assessment Singapore, Singapore

National University of Singapore, Ph.D. Research *Jun. 2025 -- Dec. 2025*

- Developed Whisper-Pinyin, a semi-supervised framework that combines human-annotated and pseudo-labeled data for Mandarin mispronunciation detection and diagnosis.
- Trained the model on over 100 hours of accented speech, achieving relative phone error rate reductions of 22.3% on in-domain testsets and 48.4% on out-of-domain testsets.

Accented Speech Synthesis Singapore, Singapore

National University of Singapore, Ph.D. Research *Jan. 2025 -- Jun. 2025*

- Designed Joycent, a diffusion-based accent TTS model focused on accent-timbre disentanglement and layer-specific accent conditioning.
- Constructed a 138-hour, 9-accent Mandarin corpus and evaluated the model under unseen-speaker, and cross-accent settings. Joycent improved accent transfer while preserving naturalness and speaker similarity, with the clearest gains for unseen speakers in cross-accent synthesis.

Publications

[1] **Xintong Wang**, Chuangang Zhao, "A 2D Convolutional Gating Mechanism for Mandarin Streaming Speech Recognition," In Information, 2021.

[2] **Xintong Wang**, Chang Zeng, Jun Chen, and Chunhui Wang, "CrossSinger: A Cross-Lingual Multi-Singer High-Fidelity Singing Voice Synthesizer Trained on Monolingual Singers," In *Proc. IEEE ASRU, 2023*

[3] **Xintong Wang**, Mingqian Shi, and Ye Wang, "Pitch-Aware RNN-T for Mandarin Chinese Mispronunciation Detection and Diagnosis," In *Proc. Interspeech, 2024, Oral*

[4] Junchuan Zhao, **Xintong Wang**, and Ye Wang, "Prosody-Adaptable Audio Codecs for Zero-Shot Voice Conversion via In-Context Learning," In *Proc. Interspeech, 2025*

[5] **Xintong Wang**, Junchuan Zhao, and Ye Wang, "Joycent: Multi-Accent TTS via Disentangled Accent Modeling and Layer-Specific Conditioning," *arXiv preprint arXiv: 2606.16417*

[6] **Xintong Wang**, Niven Jia Hao Ang, and Ye Wang, "Whisper-Pinyin: A Multi-domain Weakly Supervised Learning Framework for Mandarin Mispronunciation Detection and Diagnosis," Under Review

Academic

Teaching Assistant:

CS3263	Foundations of Artificial Intelligence	AY2026/27 (Semester 1)
SWS3027	Sound and Music Computing	AY2025/26 Summer School
CS3244	Machine Learning	AY2025/26 (Semester 2)
CS3244	Machine Learning	AY2024/25 (Semester 2)

Programming Skills

Programming: Python, C/C++, Bash, LaTeX
Machine Learning: PyTorch, PyTorch Lightning, Hugging Face Transformers
Speech Processing: Kaldi, k2/icefall, ESPnet, WeNet, SpeechBrain, CosyVoice