# Xintong Wang

xintongwang9709@gmail.com · https://oshindow.github.io/

## Education

**National University of Singapore**, Singapore — August 2024 – June 2028

*Ph.D* in Computer Science
Advisor: Assoc. Prof. Ye Wang
**Research Interests:** Automatic Speech Recognition (ASR), Text-to-Speech (TTS), Speech Language Models

**Beijing Forestry University**, Beijing, China — September 2018 – July 2022

*B.S.* in Mathematics

## Research Projects

**National University of Singapore** Singapore — August 2024 – Present

*Ph.D Student*

- Scalable and Generalizable Speech Assessment — September 2025 – May 2026
  - Developed **Whisper-Pinyin**, a weakly supervised Mandarin MDD framework using human-labeled and pseudo-labeled data from multiple domains
  - Improved generalization to unseen accents and speakers, reducing phone error rate (PER) by **22.3%** in-domain and **48.4%** out-of-domain. **Code:** https://github.com/oshindow/whisper_pinyin.

**National University of Singapore** Singapore — October 2023 – September 2024

*Research Assistant*

- Accent Text-to-Speech (TTS) — March 2024 – September 2024
  - Developed **Joycent**, a diffusion-based accent TTS model for data augmentation in MDD, supporting multi-accent adaptation from only a few seconds of accented speech prompts.
  - Outperformed baselines in naturalness, accentedness, and speaker similarity, showing effectiveness in accented speech generation. **Code:** https://github.com/oshindow/Joycent-code.
- Mandarin Speech Assessment — October 2023 – March 2024
  - Designed **Pitch-Aware RNN-T**, an end-to-end model based on a stateless RNN-T architecture that integrates self-supervised learning (SSL) features with explicit pitch modeling for Mandarin MDD.
  - Achieved a phone error rate (PER) of **27%** on non-native Mandarin speech while trained solely on native Mandarin speech.

## Working Experience

**Xiaobing.ai** Beijing, China — July 2022 – October 2023

*Machine Learning Engineer*

- Cross-lingual Singing Voice Synthesis (SVS) — May 2023 – July 2023
  - Developed **CrossSinger**, the first SVS acoustic model that supports multi-singer cross-lingual singing voice generation while trained solely on monolingual singing data.
  - Synthesized high-fidelity (48 kHz) cross-lingual singing voices for multiple singers, including code-switching scenarios.
- iSTFT-based Singing Voice Vocoder — January 2023 – May 2023
  - Developed an **iSTFT-based vocoder** for high-fidelity (48 kHz) singing voice synthesis, achieving improved naturalness with fast inference speed.
  - Outperformed transposed convolutional network-based vocoders in terms of the naturalness of synthesized singing voices.

- Singing Voice Synthesis (SVS) July 2022 – January 2023
  - Involved in the development of **XiaoiceSing2**, a production-level Generative Adversarial Network (GAN)-based SVS acoustic model that supports multi-singer high-fidelity (48 kHz) singing voice synthesis in Mandarin for **X-Singer**, a real-world singing voice generation platform.

## Internship

**Xiaobing.ai** Beijing, China May 2021 – July 2022

*Speech Recognition Intern*

- Mashup Song Generation System August 2021 – July 2022
  - Processed 30k+ lyric timestamp annotation files to construct a unified MIDI representation database.
  - Developed a rule-based mashup song generation system from a large-scale MIDI database conditioned on predefined chord progressions (e.g., 1-5-6-3-4-1-2-5).
  - Generated MIDI-based mashup song files for the X-Singer singing voice synthesis platform.
- Automatic Speech Recognition (ASR) May 2021 – August 2021
  - Developed Python-based N-gram toolkits that align with the algorithms used in SRILM.
  - Built N-gram language models from large-scale corpora, including data pre-processing, models merging, and pruning, to improve ASR performance.
  - Deployed and maintained **production-level ASR systems**.

## Publications

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

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

[3] **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.* ASRU, 2023

[4] **Xintong Wang**, Chuangang Zhao, "A 2D Convolutional Gating Mechanism for Mandarin Streaming Speech Recognition," In Information vol. 12, no. 4, pp. 165, 2021.

[5] **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

[6] **Xintong Wang**, and Ye Wang, "Joycent: Diffusion-based Accent TTS without Accented Phone Prediction," 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) |

## Repositories

- Transformer-Transducer ☆ 78 17
  A pytorch_lightning reimplementation of the Transducer module from ESPnet.
- Whisper-Pinyin
  Huggingface model: walston/whisper-pinyin 55 downloads
- Joycent
  Huggingface model: walston/whisaid-zh-grl 24 downloads

## Programming Skills

**Frameworks**: Pytorch/Lightning, Transformers, Kaldi, k2/icefall, ESPnet, WeNet, SpeechBrain, CosyVoice
**Languages**: Python, Shell, C++/C