

Junchuan ZHAO

Email: junchuan@u.nus.edu | Tel: +65 90570532 | Homepage: <https://danny-nus.github.io/> |

Linkedin: [linkedin.com/in/junchuan-zhao-6367951a5/](https://www.linkedin.com/in/junchuan-zhao-6367951a5/)

EDUCATION BACKGROUND

National University of Singapore (NUS)

Jan 2024 — Expected Dec 2027

- **Ph.D. in Computer Science.**
- Member of the **Sound and Music Computing Lab**, advised by Prof. Wang Ye.
- **Research Focus:** Speech and Singing Voice Synthesis, Voice Conversion, Voice Cloning, Neural Audio Codecs, and Expressiveness Control in Generative Models.
- **Relevant Courses:** Advanced Topics in Machine Learning, Topics in Media, Deep Learning with Language Applications.

National University of Singapore (NUS)

Aug 2022 — Dec 2023

- **MSc in Computer Science (AI specialization);** GPA: 4.75/5.0.
- **Relevant Courses:** NNs and Deep Learning, AI Planning and Decision Making, Sound and Music Computing, Uncertainty Modeling in AI.

Beijing University of Posts and Telecommunications (BUPT)

Sep 2018 — Jun 2022

- **BSc in Telecommunication Engineering with Management.**
- Cumulative GPA: 91.44/100; Professional Ranking: 4/319.
- **Relevant Courses:** Discrete Signal Processing, Multimedia Fundamentals, Advanced Transforms Methods.

Queen Mary University of London (QMUL)

Sep 2018 — Jun 2022

- **BSc in Telecommunication Engineering with Management;** First Class Degree.

RESEARCH & PROJECT EXPERIENCE

Research Assistant, Tsinghua University

Jun 2023 — Sep 2023

Advisor: Prof. Zhiyuan Liu

- Conducted a comprehensive literature review on multimodal large language models (audio $\leftrightarrow$ text), generalized audio understanding, and neural audio synthesis.
- Reviewed and categorized representative works on audio LLMs and audio-text models to guide architectural design.
- Led the design of a unified model architecture capable of processing and generating both text and audio modalities.

Research Assistant, Beijing University of Posts & Telecommunications

Jan 2022 — Jun 2022

Advisor: Prof. Shengchen Li

- Proposed SimulNotes-Codebook (SCB), a symbolic encoding scheme that captures pitch-wise (simultaneous) note structures beyond conventional pitch-time-duration (P-T-D) formats.
- Developed a baseline LSTM-VAE model and enhanced it with a multi-resolution, multi-scale encoder-decoder architecture to improve long-term temporal modeling, achieving performance on par with state-of-the-art systems.
- Introduced two metrics—SNNHS and SNPHS—to evaluate note-pattern distribution; objective evaluations (PR, MP, MD, EN, CHE, CC) demonstrated improved pitch and temporal structure modeling.

Final Year Project (FYP), Beijing University of Posts & Telecommunications

Aug 2021 — Jun 2022

Advisor: Prof. Shengchen Li

- Compared different style-transfer techniques for style-controllable music generation, drawing insights from both image and music domains.
- Designed a generation system based on MuseGAN and ClariNet; improved output quality by temporally unrolling piano-rolls, allowing the RNN to model joint distributions of simultaneous notes.
- Developed a VAE-based generator using piano-roll representations and performed comprehensive evaluations across architectures.

PUBLICATIONS

(* indicates equal contribution)

Prosody-Adaptable Audio Codecs for Zero-Shot Voice Conversion via In-Context Learning

Aug 2025

Junchuan Zhao*, Xintong Wang*, and Ye Wang †

In 26th Annual Conference of the International Speech Communication Association (**Interspeech 2025**)

- Introduced a Prosody-Aware Codec Encoder (PACE) that explicitly disentangles prosody from content and timbre, enabling fine-grained control over expressive variations.
- Integrated PACE with the pretrained VALL-E X backbone, leveraging its in-context learning ability to deliver high-quality speech while preserving speaker identity—even for unseen speakers.
- Aligned PACE-generated codes with VALL-E X codes by training PACE to predict the nine VALL-E X audio-code types, ensuring seamless compatibility between modules.
- Outperformed baseline VC systems in speech quality, timbre similarity, and prosody controllability, achieving zero-shot voice conversion that maintains both speaker identity and prosodic consistency.

SPSinger: Multi-Singer Singing Voice Synthesis with Short Reference Prompt

Apr 2025

Junchuan Zhao*, Chetwin Low*, and Ye Wang †

IEEE International Conference on Acoustics, Speech and Signal Processing (**ICASSP 2025**)

- Proposed SPSinger, a zero-shot multi-singer SVS system that synthesizes high-quality singing voices from music scores and short reference prompts.
- Introduced the Latent Prompt Adaptation Model (LPAM) to enable short-prompt inference by extracting local timbre features directly from music scores and global timbre representations.
- Implemented a novel pitch shift mechanism within LPAM to align score pitch range with the reference singer's range, improving pitch accuracy.
- Achieved superior performance over SOTA SVS systems in both objective and subjective evaluations, demonstrating accurate singer style imitation and strong zero-shot generalization.

SinTechSVS: A Singing Technique Controllable Singing Voice Synthesis System

Apr 2024

Junchuan Zhao*, Chetwin Low, and Ye Wang †

IEEE/ACM Transactions on Audio, Speech, and Language Processing (**TASLP 2024**)

- Introduced SinTechSVS, an end-to-end SVS system with explicit control over seven Chinese singing techniques. It integrates a frame-level Singing Technique Annotator (STA), a diffusion-based SVS model enhanced with an attention-based STLS module for technique conditioning, and a Transformer-based Singing Technique Recommender (STR) that predicts technique sequences from music scores to reduce manual effort.
- Proposed a data-efficient annotation framework using transfer learning and a singing technique classifier, addressing the scarcity of high-quality labeled data and enabling scalable STA training.
- Developed two evaluation metrics—Style Reclassification Accuracy (SR-Acc) and Style Match Rate (SMR)—to assess controllability from both objective and subjective perspectives.
- Experimental results show that SinTechSVS achieves high-quality synthesis in both unconditional and technique-conditioned modes, accurately reproducing singing styles and outperforming baselines in synthesis quality and control.

An Improved Time Series Network Model Based on Multitrack Music Generation

Jul 2022

Junchuan Zhao

Proceeding of 2021 International Conference on Wireless Communications, Networking and Applications (**WCNA 2021**)

- Proposed an improved time-series network for multi-track music generation, extending MuseGAN with a context generator to model inter-track dependencies.
- Introduced correction and modification mapping modules to refine temporal and structural consistency in generated music.
- Experimental results showed enhanced preservation of musical style and structure compared to traditional GAN-based methods.

ONGOING PROJECTS

- Prompt Disentanglement and Discrete Token Modeling for Zero-Shot Singing Voice Synthesis** Jul 2025
- Extended the masked generative transformer-based TTS model MaskGCT to develop a discrete token-based singing voice synthesis system, incorporating explicit control over pitch and duration.
 - Addressed the challenge of prosody disentanglement by enabling pitch control solely from the input score while minimizing interference from speech prompts, using disentangled representation learning.
 - Proposed an inference-time refinement strategy to enhance synthesis quality, leveraging auxiliary acoustic predictors and confidence-aware decoding to mitigate artifacts and improve timbre consistency.

TEACHING EXPERIENCE

- CS4347/5647: Sound and Music Computing, National University of Singapore** Fall 2024
- Designed and implemented course assignments, including both theoretical questions and practical coding tasks.
 - Prepared and delivered weekly tutorials, reinforcing key concepts and fostering hands-on learning.
 - Assessed assignments and projects, providing detailed and constructive feedback to enhance student progress.
 - Delivered the Week 9 lecture on Generative Models for Text-to-Speech (TTS) and Singing Voice Synthesis (SVS).
 - Received the **TFS Award 2025 — 2026** for outstanding teaching performance, based on excellent student feedback and high faculty evaluation scores.

AWARDS & HONORS

- SoC's Teaching Fellowship Scheme (TFS)** 2025 — 2026
Issued by the School of Computing, National University of Singapore
- NUS Research Scholarship** 2023 — 2027
Issued by the School of Computing, National University of Singapore
- Outstanding College Student in Beijing** 2022
Conferred by Beijing University of Posts & Telecommunications
- Queen Mary University of London Undergraduate College Prize (14/600)** 2022
Conferred by Queen Mary University of London
- Interdisciplinary Contest in Modeling Meritorious Winner (7%)** 2020
Issued by Consortium for Mathematics and its Applications (COMAP)
- Merit Student Awards** 2019/2020
Issued by Beijing University of Posts & Telecommunications

SKILLS

- Programming Languages and Packages**
- Python (Huggingface, Lightning, SpeechBrain, PyTorch, TensorFlow, Librosa), C/C++, Java, Javascript, Matlab.
- Music Performance**
- Over 10 years of classical piano training and performance experience.
 - Over 10 years of choir experience in both singing and conducting. Former conductor of the AiYue Chorus (BUPT); bass vocalist in the Beijing Queer Chorus.
 - Current vocalist with the NUS Jazz Band, specializing in R&B, Soul, and Jazz.
 - Regular solo performer at the annual Sound and Music Computing (SMC) Lab concerts at NUS.