Junchuan Zhao

(+65) 90570532 ◇ junchuan@u.nus.edu ◇ Homepage ◇ LinkedIn

EDUCATION

National University of Singapore (NUS) Jan. 2024 – Dec. 2027 (Expected)
- **Ph.D.** *in Computer Science*
  - *Advisor: Prof. Ye Wang, Sound and Music Computing (SMC) Lab*

National University of Singapore (NUS) Aug. 2022 – Dec. 2023
- **M.Sc.** *in Computer Science, specialization in AI* GPA: *4.75/5.00*
  - *Advisor: Prof. Ye Wang*

Beijing University of Posts and Telecommunications (BUPT) Sep. 2018 – Jun. 2022
- **B.Sc.** *in Telecommunication Engineering with Management* GPA: *91.44/100*
  - *Professional Ranking: 4/319*

Queen Mary University of London (QMUL) Sep. 2018 – Jun. 2022
- **B.Sc.** *in Telecommunication Engineering with Management (Dual Degree Program)*
  - *First Class Degree*

PUBLICATIONS

* indicates equal contribution

[1] **J. Zhao**, W. Zeng, T. Lyu, and Y. Wang. "CoMelSinger: Discrete Token-Based Zero-Shot Singing Synthesis With Structured Melody Control and Guidance." *IEEE Transactions on Audio, Speech and Language Processing* **(TASLP)**, 2026. DOI: 10.1109/TASLPRO.2026.3664643

[2] W. Zeng, **J. Zhao**, and Y. Wang. "Bridging Piano Transcription and Rendering via Disentangled Score Content and Style." *The Fourteenth International Conference on Learning Representations* **(ICLR)**, 2026. arXiv:2509.23878

[3] T. Lyu*, **J. Zhao***, and Y. Wang. "KSDiff: Keyframe-Augmented Speech-Aware Dual-Path Diffusion for Facial Animation." *IEEE International Conference on Acoustics, Speech, and Signal Processing* **(ICASSP)**, 2026. arXiv:2509.20128

[4] Z. Li, **J. Zhao**, F. Lee, and A. Yee. "InconVAD: A Two-Stage Dual-Tower Framework for Multimodal Emotion Inconsistency Detection." In *IEEE International Conference on Acoustics, Speech, and Signal Processing* **(ICASSP)**, 2026. arXiv:2509.20140

[5] **J. Zhao**, X. Wang*, and Y. Wang. "Prosody-Adaptable Audio Codecs for Zero-Shot Voice Conversion via In-Context Learning." *26th Annual Conference of the International Speech Communication Association* **(Interspeech)**, 2025. DOI: 10.21437/Interspeech.2025-464

[6] **J. Zhao**, L. Chetwin, and Y. Wang. "SPSinger: Multi-Singer Singing Voice Synthesis with Short Reference Prompt." *IEEE International Conference on Acoustics, Speech, and Signal Processing* **(ICASSP)**, 2025. DOI: 10.1109/ICASSP49660.2025.10888907

[7] **J. Zhao**, L. Chetwin, and Y. Wang. "SinTechSVS: A Singing Technique Controllable Singing Voice Synthesis System." *IEEE/ACM Transactions on Audio, Speech and Language Processing* **(TASLP)**, 2024. DOI: 10.1109/TASLP.2024.3394769

PREPRINTS

[1] **J. Zhao**, M. Vu, and Y. Wang. "Hierarchical Decoding for Discrete Speech Synthesis with Multi-Resolution Spoof Detection." 2026. arXiv:2603.05373

[2] B. Zhang*, **J. Zhao***, A. Madhukumar, Y. Wang, and I. McLoughlin. "CodecFlow: Efficient Bandwidth Extension via Conditional Flow Matching in Neural Codec Latent Space." 2026. arXiv:2603.02022

[3] Q. Liang*, Y. Liu*, R. Wei*, N. Lu, **J. Zhao**, and Y. Wang. "Segment-Aware Conditioning for Training-Free Intra-Utterance Emotion and Duration Control in Text-to-Speech." 2026. arXiv:2601.03170

RESEARCH EXPERIENCE

Discrete Speech Synthesis with Spoof-Guided Decoding Dec. 2025 – Mar. 2026
Advisor: Prof. Ye Wang, SMC Lab, National University of Singapore

- Proposed *MSpoofTTS*, a training-free hierarchical decoding framework for codec-based TTS that improves generation quality without retraining.
- Introduced a multi-resolution spoof detection mechanism that evaluates codec token sequences across different temporal granularities to identify unnatural patterns.
- Developed a progressive spoof-guided inference strategy that performs candidate pruning and reranking based on authenticity scores during decoding.

Codec-Based Bandwidth Extension via Flow Matching Nov. 2025 – Feb. 2026
Advisor: Prof. Ye Wang, SMC Lab, National University of Singapore

- Introduced *CodecFlow*, a neural codec-based BWE framework that reconstructs speech directly in latent space, enabling efficient high-fidelity extension without waveform or spectrogram modeling.
- Introduced a voicing-aware conditional flow matching module to map low- to high-bandwidth codec embeddings, improving high-frequency reconstruction and perceptual quality.
- Developed a structure-constrained residual VQ scheme based on DAC to stabilize latent alignment and reduce representation mismatch across codec resolutions.

Expressive Speech-Aligned Talking Head Generation Aug. 2025 – Nov. 2025
Advisor: Prof. Ye Wang, SMC Lab, National University of Singapore

- Proposed *KSDiff*, a keyframe-augmented, speech-aware dual-path diffusion framework for audio-driven facial animation that jointly models expression and head-pose motions.
- Designed a Dual-Path Speech Encoder (DPSE) to disentangle raw audio features into expression-related and head-pose-related components, enabling more precise motion control.
- Introduced Keyframe Establishment Learning (KEL) to predict salient motion keyframes with intense dynamics, improving motion fidelity and synchronization.

Zero-Shot Singing Voice Synthesis with Melody Control Apr. 2025 – Sep. 2025
Advisor: Prof. Ye Wang, SMC Lab, National University of Singapore

- Introduced *CoMelSinger*, a discrete token-based zero-shot SVS framework that enables explicit and structured melody control while preserving in-context learning capability.
- Identified prosody leakage in prompt-based discrete SVS and addressed it via contrastive learning and pitch-aware regularization, reducing redundant melody cues from acoustic prompts.
- Introduced a lightweight Singing Voice Transcription (SVT) module to provide frame-level pitch and duration supervision, improving pitch accuracy and temporal alignment.

Zero-Shot Voice Conversion with Prosody Adaptation Oct. 2024 – Feb. 2025
Advisor: Prof. Ye Wang, SMC Lab, National University of Singapore

- Developed 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.

Multi-Singer Singing Voice Synthesis with Short Prompts Jan. 2024 – May 2024
Advisor: Prof. Ye Wang, SMC Lab, National University of Singapore

- 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.

Singing Voice Synthesis with Controllable Singing Techniques Apr. 2023 – Sep. 2023
Advisor: Prof. Ye Wang, SMC Lab, National University of Singapore

- Introduced *SinTechSVS*, an end-to-end SVS system with explicit control over singing techniques, integrating a frame-level Singing Technique Annotator (STA), a diffusion-based SVS model with attention-based STLS conditioning, and a Transformer-based Singing Technique Recommender (STR) for technique prediction from music scores.
- 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.

INTERNSHIP EXPERIENCE

Research Assistant, Tsinghua University, NLP Lab Jun. 2023 – Oct. 2023
Mentor: Dr. Yuan Yao, Tsinghua University

- Conducted a comprehensive literature review on multimodal large language models (audio ↔ text), generalized audio understanding, and neural audio synthesis.
- Reviewed and categorized representative works on audio LLMs and audio-text models to guide architectural design.
- Explored training a unified model architecture for joint text and audio modeling, with a focus on improving cross-modal alignment between textual and acoustic representations.

MISCELLANEOUS

Languages	Mandarin (native), English (professional proficiency)	
Skills	Python, Java, JavaScript, MATLAB	
Services	Reviewer for TASLP, ToMM, ICASSP, Interspeech, ACM MM, ISMIR	
	Event chair and performer at NUS SMC Concert	2022 – 2025
	Jazz Vocalist at NUS Jazz Band	2022 – 2026
Teaching	CS3244 Machine Learning	AY2025/26
	CS5647 Sound and Music Computing	AY2024/25, AY2025/26
Awards	SoC Honour List of Student Tutors	AY2024/25
	SoC's Teaching Fellowship Scheme (TFS)	2025
	NUS Research Scholarship	2023 – 2027
	Queen Mary University of London Undergraduate College Prize (14/600)	2022
	Outstanding College Student in Beijing	2022