Junchuan ZHAO

Email: junchuan@comp.nus.edu.sg | Tel: +65 90570532 | Homepage: https://danny-nus.github.io/ |
Linkedin: 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 Talking Head Generation.
- **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.

PUBLICATIONS

(* indicates equal contribution)

KSDiff: Keyframe-Augmented Speech-Aware Dual-Path Diffusion for Facial Animation May 2026

Tianle Lyu*, **Junchuan Zhao***, Ye Wang †

2026 IEEE International Conference on Acoustics, Speech, and Signal Processing (**ICASSP 2026**)

- 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.
- Demonstrated state-of-the-art performance on benchmark datasets such as HDTF and VoxCeleb, with improvements in lip synchronization accuracy and head-pose naturalness.

InconVAD: A Two-Stage Dual-Tower Framework for Multimodal Emotion Inconsistency Detection May 2026

Zongyi Li, **Junchuan Zhao**, Francis Bu Sung Lee, Andrew Zi Han Yee

2026 IEEE International Conference on Acoustics, Speech, and Signal Processing (**ICASSP 2026**)

- Proposed InconVAD, a two-stage framework for detecting emotion inconsistency across speech and text modalities, targeting cases where multimodal emotional cues conflict.
- Designed independent uncertainty-aware unimodal towers in the first stage to produce robust emotion predictions without representation collapse under inconsistent signals.
- Introduced a cross-modal inconsistency classifier in the second stage to identify mismatches and selectively integrate consistent cues for reliable multimodal emotion analysis.

- Demonstrated through extensive experiments that InconVAD outperforms existing methods in emotion inconsistency detection, yielding more stable and interpretable predictions.

Disentangling Score Content and Performance Style for Joint Piano Rendering and Transcription Apr 2026
Zeng Wei, **Junchuan Zhao**, and Ye Wang †
The Fourteenth International Conference on Learning Representations (**ICLR 2026**)

- Proposed a unified framework that explicitly disentangles score content and performance style to jointly address piano performance rendering and automatic transcription.
- Designed separate representations for note-level score information and global performance style, enabling expressive rendering while preserving score fidelity.
- Formulated rendering and transcription as mutually supervised sequence modeling tasks, removing the need for note-level alignment or manual style annotation.
- Introduced a diffusion-based style generation module that predicts performance style directly from score content, supporting controllable and flexible rendering.
- Achieved strong results on both rendering and transcription benchmarks, demonstrating effective content–style disentanglement and style-aware performance modeling.

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 †
2025 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.

Preprints

(* indicates equal contribution)

Segment-Aware Conditioning for Training-Free Intra-Utterance Emotion and Duration Control in Text-to-Speech

Jan 2026

Qifan Liang*, Yuansen Liu*, Ruixin Wei*, Nan Lu, **Junchuan Zhao**, Ye Wang †

- Proposed a training-free controllable framework for pretrained zero-shot TTS that enables intra-utterance emotion and duration control without additional model training.
- Introduced a segment-aware conditioning mechanism that modulates emotion and timing at the segment level, allowing fine-grained adjustments within a single utterance.
- Leveraged the conditioning to adjust expressive attributes and timing behavior in pretrained TTS models, avoiding the need for retraining or additional supervision.
- Demonstrated the ability to produce expressive and duration-controlled speech from existing TTS backbones, showing improved control over emotion dynamics and temporal prosody in generated speech.

CoMelSinger: Discrete Token-Based Zero-Shot Singing Synthesis With Structured Melody Control and Guidance

Sep 2025

Junchuan Zhao, Wei Zeng, Tianle Lyu, and Ye Wang †

- Proposed 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.
- Achieved consistent improvements over state-of-the-art SVS systems in pitch accuracy, timbre similarity, and zero-shot robustness on both seen and unseen singers.

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 ↔ 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.

TEACHING EXPERIENCE

CS3244: Machine Learning, National University of Singapore Spring 2026

- Prepared and delivered weekly tutorials, reinforcing key concepts and fostering hands-on learning.
- Supervised group projects, providing feedback on group proposals and grading individual progress reports.

CS4347/5647: Sound and Music Computing, National University of Singapore Fall 2024, 2025

- 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.
- Current vocalist with the NUS Jazz Band, specializing in R&B, Soul, and Jazz.
- Regular performer (vocalist) at the annual Sound and Music Computing (SMC) Lab concerts at NUS.