Suqi Zhang

1155226708@link.cuhk.edu.hk https://saki-77.github.io/

Research Interests

Spatial Audio Generation, Speech Enhancement, Speech/Music Source Separation, Multimodal Audio LLM, Generative Models, Deep Learning

Education

The Chinese University of Hong Kong **Sept. 2024 – Now**
Master of Science in Electronic Engineering, GPA: 3.7/4.0 *Hong Kong*

Xi'an Jiaotong-Liverpool University **Sept. 2020 – July 2024**
Bachelor of Science in Information and Computing Science with Honours, GPA:3.93/4.0 *Suzhou, Jiangsu*

Industry Experience

Audio Research Intern **Mar. 2025 – Now**
Data-VnE(Video and Edge), ByteDance
Topic: Mono to Spatial Audio Generation; Universal Source Separation; Audio LLM

Publication & Submission

[1] Suqi Zhang, Zheqi Dai, Yongyi Zang, Yin Cao, Qiuqiang Kong. "DiffStereo: End-to-End Mono-to-Stereo Audio Generation with Diffusion Transformer", Interspeech, 2025.

[2] Suqi Zhang, Xianjun Xia, Chuanzeng Huang. "Flow2Stereo: Stereo Audio Generation with Flow Matching Models", in submission to ICASSP, 2026.

Research Experience

Universal Source Separation **July 2025 - Now**
Research project advised by Dr. Xianjun Xia

- Aim to design a model to either perform text-guided extraction of a specified source or autonomously separate all sources from mixtures containing arbitrary sources.
- Constructed datasets covering speech, music, and sound events by cleaning existing datasets (e.g., AudioSet) and forming input-output pairs with text description.

Stereo Audio Generation with Flow Matching Models **Mar. 2025 - June 2025**
Research project advised by Dr. Xianjun Xia

- Proposed *Flow2Stereo*, a conditional flow matching model that explicitly predicts stereo background effects from mono audio for realistic stereo rendering.
- Redesigned the mono-to-stereo pipeline to separately generate stereo background effects and integrate them in a residual-style, enhancing spatial realism.
- Achieved superior performance on standard benchmarks over other uni-modal stereo generation methods, demonstrating effectiveness without relying on handcrafted priors.

End-to-End Mono to Spatial Audio Generation **Oct. 2024 - Mar. 2025**
Research project advised by Dr. Qiuqiang Kong

- "DiffStereo: End-to-End Mono-to-Stereo Audio Generation with Diffusion Transformer" accepted by Interspeech 2025.
- Addressed the limitations of traditional mono-to-stereo generation methods by introducing an end-to-end diffusion-based solution without relying on handcrafted features or spatial priors.
- Adopted the diffusion transformer (DiT) directly in frequency domain to model stereo effect distributions, achieving competitive objective ratings and consistently better subjective ratings.

Deep Source Separation for Speech Using Generative Models **Aug. 2023 - June 2024**
Research project advised by Dr. Yin Cao

- Focused on speech enhancement (SE) by enhancing the desired speaker's voice from noisy speech, with an emphasis on recovering parts of the speech that are heavily masked by noise.
- Applied additional distortion methods to diversify the training data, such as analog-to-digital conversion, anti-automatic gain control, and Global System for Mobile Communications (GSM) transmission, making the model more robust in real-world noisy environments.

- Experimented with a GAN-based method, inspired by MetricGAN, combining the lightweight SE model DeepFilterNet2 as the generator and a metric-estimated network as the discriminator. The approach directly optimizes evaluation metrics to improve SE results.
- Trying to utilize SGMSE, a score-based generative diffusion model with its diffusion process based on stochastic differential equations (SDEs), to generate more natural speech by leveraging latent features from DeepFilterNet2 as a condition to guide the SGMSE model in generation.

Source Separation for Speech or Music **June - Sept. 2023**

Summer Undergraduate Research Fellowship Program, advised by Dr. Yin Cao

- Studied courses and textbooks of signal and systems and digital signal processing.
- Conducted a literature review in the field of source separation.
- Deep dived into deep learning libraries for speech processing, i.e., 'Asteroid', to learn models and implementations for speech separation and speech enhancement, such as Demucs and DPRNNTasNet.
- Enhanced hands-on experience of DPRNNTasNet. Created a simplified inference, and integrated the model's structural code into a new package to facilitate the independent use of the model without the need for complex cross-referencing of multiple files.

Convolutional Embeddings for Domain Adaptation in Medical Segmentation **June - Sept. 2022**

Summer Undergraduate Research Fellowship Program, advised by Dr. Erick Purwanto

- Aimed at identifying and segmenting functional tissue units (FTUs) across five human organs.
- Pre-trained EfficientNetB7 with datasets containing human kidney tissue images from *"HuBMAP - Hacking the Kidney Competition"* for better model initialization and generalization ability.
- Pre-processed the training data with data augmentation including Gaussian blurring, rotation, shifting, etc. and performed color normalization to eliminate the color inconsistency within the training set and the test set.
- Performed ensemble learning to leverage multiple models including ResNet50 and EfficientNetB7.
- Introduced multi-modal and convolutional embedding representations that provide embedding for data from different organs and institutions, which improves model's generalization ability and enables the model to adapt to different scenarios. Seven channels are added to the original RGB channels, in which five channels are used to encode information of organ types and two channels are for encoding the information of institutions.
- Achieved an accuracy of 0.81 in *"HuBMAP + HPA - Hacking the Human Body"–2022 Kaggle Competition*, ranked in the top 50 among 1175 teams.

Honors & Awards

Degree of Bachelor Science with Honours (First Class), XJTLU & University of Liverpool *Aug. 2024*

Grand Prize (CNY 2000) and Top 3 in XJTLU Student Research-Led Learning Symposium *Nov. 2023*

First Prize and Top 3 in XJTLU Student Research-Led Learning Symposium *Oct. 2022*

Summer Undergraduate Research Fellowship Winner, School of Advanced Technology *Sept. 2022*

University Academic Achievement Award - Top 10% students *Aug. 2022, 2021*

Third prize in *DJI* RoboMaster Competition - University League *May, 2021*

Relevant Coursework

- Neural Networks & Deep Learning
- Spoken Language Processing
- Audio Signal Processing for Music Applications
- Image Processing and Computer Vision
- Artificial Intelligence
- Advanced OO Programming
- Algorithmic Foundations and Problem Solving
- Decision Computation and Language

Skills & Interests

Skills: PyTorch, Python, Java, C, C++, Matlab, Linux, OpenGL, LaTeX, GitHub
Interests: Speech/music processing, Deep Learning, Generative models