Zhiyuan Zhao

approximetal@gmail.com approximetal.github.io

Overview

Profile

- 8+ years in speech/audio generative systems (ASR, TTS, VC, speech LLMs, and large-scale dataset development).
- 4 publications (2 first-author), 15 patents (7 first-inventor), and 50K+ dataset downloads on Hugging Face.

Interests

- Controllable generative audio/music systems with natural-language interfaces.
- Audio-language alignment and multimodal speech/audio foundation models for generation and evaluation.

Education

King's College London **2016 - 2017**
M.Sc. in Robotics London, UK · Merit

Central South University **2012 - 2016**
B.Sc. in Electronic Information Science and Technology Changsha, China

Open-Source Projects

Mar 2026 - Present **SVLive:** Controllable Music Generation / Live Coding / Synthesizer Programming

- Browser-based live coding music playground (*Strudel* + *Vital* + *LLM*), supporting real-time editable synthesis.
- Support for 2,300+ synthesizer presets with lazy rendering and programmable sound selection.
- Exploration of natural-language control over synth parameters, music structure, and audio-feedback alignment.

Nov 2025 - Jan 2026 **LEMAS:** Large-scale Multilingual Speech Corpus & Benchmark Models

- 150,000+ hour multilingual speech corpus with word-level timestamps across 10 languages, built via a scalable high-quality data pipeline.
- Benchmark models for speech synthesis (*LEMAS-TTS*) and speech editing (*LEMAS-Edit*) trained on the proposed dataset.

Sep 2025 - Oct 2025 **VoicePlatform:** Enhancement / Recognition / Synthesis / Editing / Translation

- Multimodal AI platform integrating six capabilities: audio processing(denoise, upsampling and separation), speech recognition, synthesis, editing, video editing, and translation.
- Simple and clear interface for exploring and controlling multiple audio/video AI modules in a unified system.

Apr 2024 - Feb 2025 **OpenVideo:** High-quality Text-to-Video Dataset and generation framework

- 672+ hour / 106K+ clip / 720p text-to-video dataset sourced from *Pexels-Raw* for generative video research.
- Dataset tooling for annotation, validation, and migration to ensure quality control and consistent video-text alignment.

Industry Experience

Mar 2026 - Present · **XPENG Robotics** · Senior Research Engineer

- Post-training **Omni speech-interaction LLMs** for embodied robotics: instruction following, spoken dialogue robustness, and real-time human-robot interaction.
- Built **speech/LLM data curation, evaluation, and failure-analysis pipelines** for iterative model improvement across speech, LLM, and robotics teams.

Mar 2022 - Mar 2026 · **International Digital Economy Academy (IDEA)** · Speech Technology Lead / Algorithm Engineer

- Created **LEMAS**, a **150K+ hr, 10-language** speech corpus with word-level timestamps; released **LEMAS-TTS** and **LEMAS-Edit** benchmark models for generative speech research. Project · HF
- Led multimodal speech-agent research and engineering; built a **real-time avatar dialogue system** with zero-shot voice cloning, low-latency inference, multi-concurrency support, and character-level voice interaction.
- Trained **zero-shot multilingual TTS/VC and speech editing models** covering **7 languages** (ZH/EN/DE/FR/PT/ES/IT); built cleaning and alignment pipelines for **100K+ hr** speech data.
- Built an end-to-end **AI video dubbing / speech editing pipeline** (ASR → LLM translation → TTS/VC → subtitle rendering); 1-min video processed in **~3-5 min**, matching top industry solutions.
- Trained and deployed **Vietnamese ASR** (Conformer, streaming+offline), significantly outperforming Azure and iFlytek on internal benchmarks; also deployed ZH/EN/VI TTS.
- Built multilingual data collection, quality inspection, alignment, ASR, denoising, TTS/VC, and editing pipelines for ZH, RU, ID, TH, VI, and more.

Jun 2019 - Feb 2022 · **UBTECH Robotics Corp.** · Senior Speech Synthesis Algorithm Engineer

- Researched and deployed **neural TTS/VC models**: Seq2Seq, Transformer, GAN acoustic models, and HiFiGAN/WaveRNN/WaveGlow vocoders.
- Built **few-shot cross-lingual multi-speaker expressive VC/TTS** with MOS and ASR-based objective evaluation, supporting custom voices across languages and emotions.
- Delivered **lightweight bilingual TTS on ARM** (C++ inference, **<300ms latency**) for a dictionary-scanning pen product.
- Responsible for robot voice synthesis featured at **CCTV 2022 Spring Festival Gala** sketch "Name Card."

Apr 2018 - May 2019 · **SF Technology Co., Ltd.** · Optimization Algorithm Engineer

- Built ML-based logistics optimization for parcel-volume forecasting, trunk-line planning, and dynamic route adjustment.
- Deployed dynamic vehicle reallocation system for peak periods (Singles' Day, Spring Festival): hundreds of daily plans auto-generated, **~75% reduction** in manual workload.

Sep 2017 - Jan 2018 · **Samsung Telecom R&D Center** · Software Development Intern

- Bixby voice assistant testing, NLP/word-segmentation analysis, and multilingual data annotation.

Publications

[1] **LEMAS: A 150K-Hour Large-scale Extensible Multilingual Audio Suite with Generative Speech Models**
Zhiyuan Zhao, Lijian Lin, Ye Zhu, Kai Xie, Yunfei Liu, Yu Li
arXiv:2601.04233, 2026 PDF

[2] **MelTok: 2D Tokenization for Single-Codebook Audio Compression**
Jingyi Li*, Zhiyuan Zhao*, Zhisheng Zhang, Yunfei Liu, Lijian Lin, Ye Zhu, Jiahao Wu, Qiuqiang Kong, Yu Li
arXiv:2510.01903, 2025 PDF

[3] **STFTCodec: High-Fidelity Audio Compression through Time-Frequency Domain Representation**
Tao Feng*, Zhiyuan Zhao*, Yifan Xie, Yuqi Ye, Xiangyang Luo, Xun Guan, Yu Li
IEEE ICME, 2025 PDF

[4] **Improving Model Stability and Training Efficiency in Fast Speed High Quality Expressive Voice Conversion System**
Zhiyuan Zhao, Jingjun Liang, Zehong Zheng, Linhuang Yan, Zhiyong Yang, Wan Ding, Dongyan Huang
ACM ICMI, 2021 PDF

* equal contribution.

Patents

First inventor:

[1] Audio editing method, system, terminal, and storage medium — CN202510616027.1
[2] Training method and synthesis method for a speech synthesis model — CN202410346345.6
[3] Data quality inspection method, device, equipment, and storage medium — CN202310010360.9
[4] Cross-lingual audio conversion method, computer equipment, and storage medium — CN202011516681.9
[5] Cross-lingual speech conversion method, computer equipment, and storage medium — CN202011588869.4
[6] Training method of a speech synthesis model, electronic equipment, and medium — CN202011448005.2
[7] Voice conversion method, device, equipment, and storage medium — CN201980003287.4

Co-inventor:

[8] Interactive digital-actor video generation method, system, device, and computer storage medium — CN202510603049.4
[9] Speech synthesis method, device, terminal equipment, and storage medium — CN202111630461.3
[10] Speech emotion recognition method, terminal device, and computer-readable storage medium — CN202111615879.7
[11] Model training method, device, terminal equipment, and computer-readable storage medium — CN202111682661.3
[12] Adaptive sample dataset generation method, device, and equipment — CN202111632906.1
[13] Speech denoising method, device, electronic equipment, and computer-readable storage medium — CN202111682647.3
[14] Encryption method, device, terminal equipment, and computer-readable storage medium — CN202211152830.7
[15] Voice conversion method, system, device, and storage medium — CN201980003189.0