Qiaolin Wang

New York, NY qw2443@columbia.edu (646) 528-0549 Homepage

Research Interests

I enjoy studying models that can **Perceive, Reason, and Speak** like humans:
- Multimodal Large Language Models (MLLMs) • Audio-Visual Understanding • Speech Synthesis

Education

Columbia Univeristy *Sept 2024 – Dec 2005*
M.S in Electrical Engineering
- **Concentration:** Speech and Language Processing (Advisor: Prof. Nima Mesgarani)

Wuhan Univeristy *Sept 2020 – June 2024*
B.Eng in Computer Science
- **Core Course:** Computer Systems, Artificial Intelligence, Intelligent Speech Processing

Publications

Layer-wise Minimal Pair Probing Reveals Contextual Grammatical-Conceptual Hierarchy in Speech Representations

Linyang He*, ***Qiaolin Wang****, Xilin Jiang, and Nima Mesgarani

EMNLP 2025 SAC Highlight [pdf]

SightSound-R1: Cross-Modal Reasoning Distillation from Vision to Audio Language Models

Qiaolin Wang, Xilin Jiang, Linyang He, and Junkai Wu, Nima Mesgarani

Submitted to ICASSP 2026 [arXiv]

Research Experience

AVMeme Exam: A Multimodal Multilingual Multicultural Benchmark for LLMs' Multimedia Knowledge and Thinking Beyond Text [Ongoing]
Columbia University New York, NY
Research Assistant *Oct 2025 – Present*
- Built a large-scale audio-visual meme benchmark to test MLLMs' multimodal and cultural understanding
- Spearheaded evaluation of **15** models (Audio, Video, Omni, and commercial LLMs) with fine-grained QA categorization across **1000** memes in multiple languages and cultures
- Aiming for public release and preprint submission by Dec 2025

SightSound-R1: Cross-Modal Reasoning Distillation from Vision to Audio Language Models
Columbia University New York, NY
Research Assistant *Jun 2025 – Sept 2025*
- Proposed SightSound-R1, a novel framework to distill reasoning from Vision to Audio LLMs
- Engineered an audio-focused Chain of Thought (CoT) generation prompt for Qwen2.5-VL-32B with test-time scaling and used a GPT-4o fact-checker to filter visual hallucinations
- Implemented a two-stage training strategy (SFT + GRPO) to distill the verified CoT into Qwen2-Audio-7B
- Improved the LALM's reasoning on unseen datasets, outperforming baselines and achieving **66.1%** on MMAU-Test-Mini (Sound) and **59.5%** on MUSIC-AVQA

Layer-wise Minimal Pair Probing Reveals Contextual Grammatical-Conceptual Hierarchy in Speech Representations
Columbia University New York, NY
Research Assistant *Feb 2025 – May 2025*
- Conducted speech minimal-pair probing with **116k** pairs from BLiMP/COMPS across **71** linguistic tasks
- Probed **16** models (S3M, ASR, AudioLLM, Codec) with layer-wise linear classifiers on frozen representations
- Observed **syntax** > **morphology** > **concepts**: Speech models capture **form** more strongly than **meaning**

- Found **mean pooling** outperformed single-token extraction, yielding more stable speech representations
- Exposed **temporal asymmetry**: grammatical evidence in S3M/S2T-ASR **peaks 500–600 ms pre-onset**, whereas in AudioLLMs and Whisper it **accumulates through onset and beyond**

Snoring Sound Dataset Annotation
Wuhan University — Wuhan, China
Research Assistant — *Jan 2023 – Mar 2023*

- Analyzed Polysomnography (PSG) data from **40** patients to identify respiratory events and sleep stages
- Synchronized clinical PSG signals with ≈ **170** hours of audio using PSG4/Sleepware and Audition
- Labeled snores relative to respiratory events and cross-referencing sleep stages, categorizing snores based on their temporal relation to events
- Contributed to the foundational dataset published at *Interspeech 2023* [pdf]

Work Experience – Speech AI

Research Engineer Intern, Wiz.AI – Singapore — Mar 2024 – Aug 2024

- Developed a multi-task Speech Large Language Model to understand both content and emotion
- Engineered a prompt strategy that integrates dialogue history to enhance Chain of Thought reasoning
- Utilized a window-level query mechanism to capture fine-grained emotional features from raw speech
- Finetuned a cross-modal alignment module between speech (Q-former) and text (Vicuna LLM)
- Implemented a contrastive-learning loss on emotion embedding, achieving **SOTA 74.48%** on IEMOCAP and **62.61%** on MELD for SER

Machine Learning Engineer Intern, Wiz.AI – Singapore — July 2023 – Dec 2023

- Spearheaded a systematic evaluation of voice cloning models and automated the deployment pipeline
- Benchmarked open and closed-source voice-cloning solutions on speech quality and naturalness
- Fine-tuned a StyleTTS2-based voice cloning model on LibriSpeech and proprietary dataset
- Built a speaker verification framework with WeSpeaker, PyAnnote, and SpeechBrain for evaluation
- Achieved competitive speaker similarity (≈ **0.84**) to the commercial model using only **245** hours of data (**1/40** of the original dataset)

Projects

Speech Synthesis Implementation on Game Avatars [**Colab**] [**Bilibili**]
Wuhan University — Wuhan, China
Independent Developer — *Jun 2022 – Sept 2022*

- Developed a high-fidelity, Text-to-Speech (TTS) model for the character "Paimon" from Genshin Impact
- Engineered a pipeline using ECAPA-TDNN for speaker classification and Whisper for transcription
- Built and annotated a multi-speaker dataset of ≈ **48000** clips (**15** hrs.) from **50** Genshin Impact characters
- Finetuned a VITS-based speech synthesis model using a curated set of "Paimon" audio clips
- Launched the project as a technical demo on Bilibili, attracting over **600,000** views and deploying the model on Google Colab for public inference

Technologies

Languages: Python, Java, C, C++ **Tools:** Linux, Git, PyTorch