陈文熙

(+86) 189-9837-4728 · 1029713857@sjtu.edu.cn · 微信号 @worst_chan · GitHub @cwx-worst-one

个人总结

本人在校成绩优秀、乐观向上，工作负责、自我驱动力强、热爱尝试新事物。在校期间主要从事音频自监督学习、音频分类、复杂音频理解、音频-语言大模型应用、端到端语音对话系统等相关研究，对深度学习在音频领域的发展趋势以及大语言模型与音频领域的结合有浓厚兴趣。

教育背景

上海交通大学, 计算机科学与技术, *在读博士研究生* 2025.09 - 2030.06

跨媒体语言智能实验室（Cross Media Language Intelligence Lab, X-LANCE），导师：陈谐教授

上海交通大学, 计算机科学与技术（IEEE 试点班）, *工学学士* 2021.09 - 2025.06

- **GPA: 4.07/4.30; 核心学积分: 93.63/100; 排名: 5/32**
- **专业课程**: 程序设计原理与方法 (96); 数据结构 (97); 计算机网络 (96); 自然语言处理 (97) 等
- **荣誉奖项**: 荣昶科技创新奖学金 (2024), 上海交通大学本科生 C 等优秀奖学金 (2022-2024)
- **英语成绩**: CET4: 608; CET6: 515

实习经历

微软亚洲研究院 | 通用人工智能组 (GAI Group), *算法实习生* 2024.09-2025.06

在刘树杰博士与李锦宇博士的指导下，参与端到端语音对话模型的高效与高性能算法设计，并系统地研究了语音大模型相比于文本大模型“降智”现象的成因，同时提出了针对性的缓解策略。

- 提出可控音色的单阶段语音对话系统 SLAM-Omni，创新性引入语义分组生成与历史提示策略，显著提升训练与推理效率。实验表明，在低资源场景下，SLAM-Omni 在声学质量与回答内容上均显著优于同规模模型。相关开源数据集在 Hugging Face 上下载量超过一万，论文被 **ACL 2025 Findings** 录用。
- 系统研究语音大模型在推理阶段“降智”现象的成因，发现其主要源于 SFT 阶段对大模型通用能力的削弱。提出内容增强型 SFT 策略与模型融合方法，有效保留模型通用知识，使训练后的语音大模型在性能上接近级联系统。相关成果拟投稿至 **ICASSP 2026（CCF-B）**。

项目经历

EAT: Self-Supervised Pre-Training with Efficient Audio Transformer 2023.08-2023.12

- **科研内容**: 提出基于掩码式自蒸馏的音频自监督模型 EAT，创新地结合帧级别与整体音频级别目标，增强了模型的音频表征学习能力。设计逆向块状掩码、高掩码率（80%）与轻量 CNN 解码器结构，有效提高预训练效率和表征质量。
- **科研贡献**: 作为课题主要研究者，独立提出并完善 EAT 模型框架，主导核心实验设计与论文撰写。
- **科研结果**: EAT 在 AudioSet (AS-2M, AS-20K)、ESC-50、SPC-2 等多个音频分类任务实现 SOTA 表现，较经典模型（BEATs、Audio-MAE）预训练速度提升超过 10 倍。相关研究成果论文以第一作者身份被 **IJCAI 2024 (CCF-A)** 录用。

IEEE ICME 2024 Grand Challenge 2024.02-2024.03

- **科研内容**: 参加 IEEE ICME 2024 “域转移下的半监督声学场景分类”全球挑战赛，结合自监督预训练模型 EAT、自学习半监督策略及 Test-time Adaptation 推理技术，以实现更高效的跨域场景分类。通过对目标数据进行权重化预训练，并在标记数据上迭代生成伪标签进行模型微调，有效缓解了域迁移问题，在预测集与训练集差异较大的情况下表现显著优于基准方法。
- **科研贡献**: 参赛队伍队长，主要负责训练框架的搭建和优化，和组员共同完成模型训练、微调、推理任务，共同撰写最后的技术报告和 ICME workshop 论文。
- **科研结果**: 最终在挑战赛中取得了 75.2% 的声学场景分类准确率，相比 baseline（60.0%）提升 15.2%，位列全球第二名，与第一名（科大讯飞团队）相差仅 0.6%。相关研究成果形成的两篇论文被 **IEEE ICME 2024 (CCF-B) Workshop** 录用。

DCASE Challenge 2024 Task 6: Automated Audio Captioning 2024.03-2024.06

- **科研内容**: 参加国际声学领域权威赛事 DCASE 2024 的“自动音频字幕生成（AAC）”任务，提出了一种融合自监督模型 EAT 与大语言模型 Vicuna 的高效音频字幕生成框架，并采用低秩适配器（LoRA）

策略实现快速微调。此外，创新性地引入基于对比式预训练模型 CLAP 的字幕筛选机制，提升生成字幕与原始音频的语义一致性，最终单系统和融合系统的各项客观评测指标均超过往年方法。
- **科研贡献**: 担任参赛团队队长，主导模型架构设计和优化，带领团队共同完成系统的训练、微调、推理及融合工作，负责技术报告与相关论文的撰写。
- **科研结果**: 最终提交的系统在主评测指标 FENSE 上位列全球第三，在其他客观指标（如 SPIDEr）上均排名第一。相关研究成果论文以第一作者身份被 **ICASSP 2025 (CCF-B)** 录用。

科研论文

- EAT: Self-Supervised Pre-Training with Efficient Audio Transformer
  Wenxi Chen, Yuzhe Liang, Ziyang Ma, Zhisheng Zheng, Xie Chen
  IJCAI 2024
- SLAM-Omni: Timbre-Controllable Voice Interaction System with Single-Stage Training
  Wenxi Chen, Ziyang Ma, Ruiqi Yan, Yuzhe Liang, Xiquan Li, Ruiyang Xu, Zhikang Niu, Yanqiao Zhu, Yifan Yang, Zhanxun Liu, Kai Yu, Yuxuan Hu, Jinyu Li, Yan Lu, Shujie Liu, Xie Chen
  ACL 2025 *(Findings)*
- SLAM-AAC: Enhancing Audio Captioning with Paraphrasing Augmentation and CLAP-Refine through LLMs
  Wenxi Chen[*], Ziyang Ma[*], Xiquan Li, Xuenan Xu, Yuzhe Liang, Zhisheng Zheng, Kai Yu, Xie Chen
  ICASSP 2025
- SimulS2S-LLM: Unlocking Simultaneous Inference of Speech LLMs for Speech-to-Speech Translation
  Keqi Deng, **Wenxi Chen**, Xie Chen, Philip C. Woodland
  ACL 2025
- DRCap: Decoding CLAP Latents with Retrieval-Augmented Generation for Zero-shot Audio Captioning
  Xiquan Li, **Wenxi Chen**, Ziyang Ma, Xuenan Xu, Yuzhe Liang, Zhisheng Zheng, Qiuqiang Kong, Xie Chen
  ICASSP 2025 *(Oral)*
- Towards Reliable Large Audio Language Model
  Ziyang Ma, Xiquan Li, Yakun Song, **Wenxi Chen**, Chenpeng Du, Jian Wu, Yuanzhe Chen, Zhuo Chen, Yuping Wang, Yuxuan Wang, Xie Chen
  ACL 2025 *(Findings)*
- Emovoice: LLM-based Emotional Text-to-Speech Model with Freestyle Text Prompting
  Guanrou Yang, Chen Yang, Qian Chen, Ziyang Ma, **Wenxi Chen**, Wen Wang, Tianrui Wang, Yifan Yang, Zhikang Niu, Wenrui Liu, Fan Yu, Zhihao Du, Zhifu Gao, Shiliang Zhang, Xie Chen
  ACM MM 2025
- EmoBox: Multilingual multi-corpus speech emotion recognition toolkit and benchmark
  Ziyang Ma, Mingjie Chen, Hezhao Zhang, Zhisheng Zheng, **Wenxi Chen**, Xiquan Li, Jiaxin Ye, Xie Chen, Thomas Hain
  Interspeech 2024

竞赛获奖

- IEEE ICME 2024 挑战赛“Semi-supervised Acoustic Scene Classification under Domain Shift”第二名 2024.03
- DCASE Challenge 2024 第六赛道“Automated Audio Captioning”第三名 2024.06