李浩宇

haoyu.li.cs@sjtu.edu.cn

教育背景

上海交通大学，上海 2023 – 2026

在读硕士研究生 计算机科学与技术 导师：俞凯
GPA: 3.84/4.0

北京交通大学，北京 2019 – 2023

学士 计算机科学与技术
GPA: 3.90/4.0

研究方向

研究兴趣主要包括**语音增强（Speech Enhancement, SE）**、**语音识别（Automatic Speech Recognition, ASR）**、与**语音合成（Text-to-Speech, TTS）**三个方向。在语音增强方向，开展文本线索引导的**盲源语音分离（Blind Speech Separation, BSS）**与**目标说话人提取（Target Speaker Extraction, TSE）**方法研究；在语言识别方向，主要研究关键词检测（Keyword Spotting, KWS）技术，构建**基于 Transducer/CTC 架构的 KWS 系统**，并对基于 ASR 构建的 **KWS 解码算法**进行研发创新；语音合成方面，曾经做过**提升说话人一致性**的研究。

实习经历

思必驰科技股份有限公司，苏州 2024 年 6 月-2024 年 11 月

本地算法组语音算法实习生

围绕**关键词检测（Keyword Spotting, KWS）**与**语音增强（Speech Enhancement, SE）**任务开展研究。

- **文本引导语音分离系统构建**：基于 INTERSPEECH 2024 已录用论文中的文本引导语音分离框架，完成噪声语音数据集仿真构建，基于合成数据训练语音分离模型，并在多人对话场景开展**验证测试**，使系统在真实场景下的拒识率从 5.3% 降低至 4.3%，同时将误唤醒次数降低至基线的 20%。
- **端到端关键词检测算法研发**：研究基于 CTC 架构的 KWS 系统，在信噪比-5dB 至 20dB 的噪声场景下，为 CTC 系统设计实现**鲁棒流式解码算法**；同时优化基于加权有限状态转换器（Weighted Finite State Transducer, WFST）的 KWS 系统。**相关成果形成 ICASSP 2025 论文 2 篇**。
- **工程架构优化**：负责开源语音引擎技术调研（ESPNet、WeNet），维护 PyTorch 训练和测试框架，完成超参数搜索的实验。参与 ICASSP 2025 多通道增强文章的词错误率指标测量。

阿里巴巴科技股份有限公司，北京 2025 年 5 月-2025 年 9 月

未来生活实验室多模态算法实习生

设计流匹配 TTS 系统的说话人自适应对齐方案，解决零样本语音克隆中音色一致性问题。核心创新包括：（1）时序自适应：根据去噪阶段动态调整监督强度；（2）层级自适应：针对不同网络层分配差异化对齐目标。在阿里 10 万小时语音数据上实现 SOTA 说话人相似度，收敛效率提升近 3 倍。**相关成果形成论文 1 篇，正在投稿至 INTERSPEECH 2026**。

科研经历

Text-aware Speech Separation For Multi-talker Keyword Spotting INTERSPEECH 2024

***Haoyu Li**, Baochen Yang, Yu Xi, Linfeng Yu, Tian Tan, Hao Li, Kai Yu*†

- **关键词引导排列消歧**：提出基于关键词信息引导的排列问题消解方案，前端语音分离（Speech Separation）模型在混合含关键词的测试数据场景上，排列错误率由 50% 降为 1%。排列问题的解决不仅使模型推理的计算量减半，还降低了识别模型误唤醒的风险。
- **唤醒音频分离效果增强**：系统引入的关键词线索信息不仅解决了排列问题，还提高了关键词音频的分离效果，在 STOI 和 SI-SNR 两项指标上都超越了基线分离系统。
- **利于前后端适配**：利用前端推理的音频微调后端，可以增强后端在未见数据上的召回率，有利于实际生产环境部署和性能优化。

Detect, Attend and Extract: Keyword Guided Target Speaker Extraction 投稿至 IJCAI 2026

*Haoyu Li**, *Yu Xi**, *Yidi Jiang, Shuai Wang*†, *Kate Knill, Mark Gales, Haizhou Li, Kai Yu*†

- **关键词引导的目标说话人提取新范式**：针对传统目标说话人提取（TSE）系统依赖预注册语音的局限性，提出基于部分文本线索（关键词）的提取新范式，适用于会议记录、语音助手交互等动态场景，无需繁琐的预注册流程。
- **检测-关注-提取（DAE）三阶段框架**：设计 Detect-Attend-Extract 框架，首先检测关键词在混合音频中的存在性和时间位置，然后基于关键词上下文关注目标说话人并生成说话人表征，最后提取目标语音。相比传统基于语音注册的基线系统，仅使用 28.4% 的文本线索即可实现更优性能。
- **ASR-SV 联合训练的关键词引导线索编码器**：提出结合语音识别（ASR）和说话人验证（SV）的多任务学习框架，通过交叉注意力机制实现语音与文本对齐，支持关键词检测、时间定位和说话人表征生成，在 Libri2Mix 数据集上 SI-SNRi 达到 16.45dB，显著优于传统 TSE 基线（12.98dB）。
- **高精度关键词定位**：提出的动态规划算法可实现约 100ms 误差的关键词时间定位，并支持关键词存在性检测（F1-score 达 98.06%），增强了系统在真实场景中的实用性。

Time-Layer Adaptive Alignment for Speaker Similarity in Flow-Matching Based Zero-Shot TTS 投稿至 INTERSPEECH 2026

*Haoyu Li**, *Mingyang Han**, *Yu Xi, Dongxiao Wang, Hankun Wang, Haoxiang Shi, Boyu Li, Jun Song, Bo Zheng, Shuai Wang*†, *Kai Yu*†

- **流匹配零样本语音合成的说话人建模优化**：针对 Flow-Matching（FM）框架缺乏显式说话人监督的问题，提出时序-层级自适应说话人对齐（TLA-SA）损失函数，提升零样本语音合成中的说话人相似度。
- **说话人信息分布的实证分析**：首次系统揭示 FM 去噪过程中说话人信息在时序（去噪步数）和层级（网络深度）维度的非均匀分布特性，发现早期去噪阶段和浅层网络对说话人身份建模更为关键。
- **自适应对齐机制设计**：基于时序和层级的动态权重分配策略，利用预训练说话人编码器对 FM 中间表征进行自适应监督，在 10 万小时工业级数据集上实现 2.9 倍收敛加速，说话人相似度提升 2.2%-3.0%。
- **跨架构泛化验证**：方法在语言模型级联架构（LM-based，如 CosyVoice 2）和端到端架构（LM-free，如 F5-TTS）上均一致有效，并开源了大规模训练数据集。

Streaming Keyword Spotting Boosted by Cross-layer Discrimination Consistency ICASSP 2025

*Yu Xi**, *Haoyu Li**, *Xiaoyu Gu, Hao Li, Yidi Jiang, Kai Yu*†

- **目前缺乏容易维护的流式 KWS 解码算法**：现有基于 CTC 的 KWS 解码算法或依赖于次优的 ASR 算法，或依赖于复杂的解码图算法。实际应用中，我们需要流式且简单的 KWS 解码算法。
- **支持任意唤醒词的流式 CTC KWS 解码算法**：我们为基于 CTC 系统实现了简单且有效的流式 KWS 算法，该算法支持任意唤醒词。算法利用 CTC 在浅层/深层解码输出的不同性质，可以更好地区分正例和负例，从而提升 KWS 系统的召回率。
- **算法对低信噪比场景鲁棒**：我们的实验表明，提出的解码算法在 Hey-Snips 数据集上超越了传统 ASR 解码算法和解码图算法的基线，并且在极低信噪比（-5dB 和 0dB）上依然可用。

MFA-KWS: Effective Keyword Spotting with Multi-head Frame-asynchronous Decoding TASLP

Yu Xi, Haoyu Li, Xiaoyu Gu, Yidi Jiang, Kai Yu†

- **扩展 Transducer 架构的 KWS 系统**：我们在训练过程中，在 Transducer 架构上加入 CTC 损失约束，辅助 RNN-T 损失训练 ASR 模型，模型得以更快地收敛，测试性能也得以提升。
- **多头帧同步解码算法**：结合传统 RNN-T 和 CTC 的流式 KWS 解码算法，我们为 Transducer-CTC 联合架构模型提出多头帧同步解码算法，即同时融合 Transducer 和 CTC 两个头的解码结果作为最终输出。
- **多头帧异步解码算法**：我们提出的系统将传统 RNN-T 替换为其变种，再将 CTC 头的解码算法换为帧异步的算法，最后利用多种不同策略融合两个头的输出。我们设计的多头帧异步解码算法在 Transducer-CTC 联合架构下，在多个 KWS 测试集上达到了目前最优的性能，并达到了 1.47x-1.63x 的加速比。

NTC-KWS: Noise-aware CTC for Robust Keyword Spotting ICASSP 2025

Yu Xi, Haoyu Li, Hao Li, Jiaqi Guo, Xu Li, Wen Ding, Kai Yu†

- **提升基于解码图的 KWS 算法的鲁棒性**：本文针对噪声鲁棒的 KWS 任务，改进基于 WFST 的 KWS 框架，通过引入动态噪声建模机制，减弱了小型 KWS 模型对噪声的过拟合。
- **引入通配符建模噪声**：Noise-aware CTC (NTC) 系统引入了两种通配符，自环和旁路，分别建模语音中由于噪声导致的插入错误和替换/删除错误。

- **系统性能**：NTC 系统在加噪和不加噪的 Hey-Snips 数据集上超越了目前最佳的端到端的基线系统；并且在越低信噪比下，噪声建模带来的提升越明显。

个人技能

- **编程语言**: Python, C/C++, Shell
- **平台/工具包**: Linux, WeNet 相关框架（包括 WeSpeaker, WeSep 和 WeKws）, ESPnet, PyTorch, NeMo, K2
- **学生工作**：班级学习委员，学院志愿者协会宣传部副部长
- **兴趣爱好**: 跑步，旅行，滑雪，台球，FPS

♡ 获奖情况

国家奖学金	2020.10 & 2021.10
北京交通大学三好学生	2020.10 & 2021.10
北京交通大学程序设计竞赛二等奖	2021.5
北京交通大学优秀毕业生	2023.06

论文列表

Text-aware Speech Separation For Multi-talker Keyword Spotting INTERSPEECH 2024

Haoyu Li, *Baochen Yang, Yu Xi, Linfeng Yu, Tian Tan, Hao Li, Kai Yu*†

Detect, Attend and Extract: Keyword Guided Target Speaker Extraction 投稿至 IJCAI 2026

Haoyu Li*, *Yu Xi****, *Yidi Jiang, Shuai Wang*†*, Kate Knill, Mark Gales, Haizhou Li, Kai Yu*†

Time-Layer Adaptive Alignment for Speaker Similarity in Flow-Matching Based Zero-Shot TTS 投稿至 INTERSPEECH 2026

Haoyu Li*, *Mingyang Han****, *Yu Xi, Dongxiao Wang, Hankun Wang, Haoxiang Shi, Boyu Li, Jun Song, Bo Zheng, Shuai Wang*†*, Kai Yu*†

Streaming Keyword Spotting Boosted by Cross-layer Discrimination Consistency ICASSP 2025

*Yu Xi****, ***Haoyu Li****, *Xiaoyu Gu, Hao Li, Yidi Jiang, Kai Yu*†

MFA-KWS: Effective Keyword Spotting with Multi-head Frame-asynchronous Decoding TASLP

*Yu Xi, **Haoyu Li**, Hao Li, Xiaoyu Gu, Kai Yu*†

NTC-KWS: Noise-aware CTC for Robust Keyword Spotting ICASSP 2025

*Yu Xi, **Haoyu Li**, Hao Li, Jiaqi Guo, Xu Li, Wen Ding, Kai Yu*†

Masked Self-distilled Transducer-based Keyword Spotting with Semi-autoregressive Decoding ASRU 2025

*Yu Xi, Xiaoyu Gu, **Haoyu Li**, Jun Song, Bo Zheng, Kai Yu*†

TDT-KWS: Fast And Accurate Keyword Spotting Using Token-and-duration Transducer ICASSP 2024

*Yu Xi, Hao Li, Baochen Yang, **Haoyu Li**, Hainan Xu, Kai Yu*†

Neural Directed Speech Enhancement with Dual Microphone Array in High Noise Scenario ICASSP 2025

*Wen Wen, Qiang Zhou, Yu Xi, **Haoyu Li**, Ziqi Gong, Kai Yu*†