

白 烨

邮箱: baiye@cau.edu.cn

电话: 13021981594

教育经历

中国科学院自动化研究所-模式识别国家重点实验室

2016 年 9 月 - 2021 年 6 月

模式识别与智能系统专业, 工学博士

导师: 陶建华研究员

中国农业大学

2012 年 9 月 - 2016 年 6 月

通信工程专业, 工学学士

专业排名第一推免至中科院读博

工作经历

研究科学家

2023 年 3 月 - 2023 年 8 月

字节跳动-语音

研发超大规模语音预训练模型和生成式端到端音频理解系统。训练千万小时量级数据的大规模无监督预训练编码器, 并结合大语言模型构建音频理解系统。所研发的 Seed-ASR 语音识别系统较之前的系统错误率相对降低 20% 40%, 该系统应用在字节跳动包括豆包在内的各种涉及语音交互的应用中。

算法专家

2021 年 7 月 - 2023 年 3 月

快手科技-多媒体内容理解部

研发大规模语音识别系统, 应用在快手等亿级用户视频 APP 的内容理解、字幕展示、语音交互功能。研发虚拟人动作驱动方案和生成模型, 应用在虚拟歌手, 获得数十万粉丝关注。

算法研究实习生

2019 年 4 月 - 2020 年 1 月

字节跳动-AI-Lab 智能语音

研发大规模端到端语音识别模型 LAS。在业内率先将端到端语音识别应用在抖音等亿级用户审核与字幕功能。

研究兴趣概况

我的研究兴趣主要集中在生成式音频理解系统、自监督学习、语言模型、音频伪造检测等音频理解与交互相关的领域。具体上, 在生成式音频理解系统方面, 主要是研究基于自回归模型, 将音频转换为文本 (语音识别、语音翻译、音频描述等) 所涉及到的各种方法和模块, 特别是语音无监督预训练模型 (编码器模型) 和文本无监督预训练模型 (语言模型)。在语音自监督训练方面, 在字节跳动研发基于 teacher-student learning 千万小时级预训练模型。在文本无监督预训练模型方面, 提出语音-文本模态对齐 (alignment), 将语言模型中的知识迁移到语音识别模型中。偏中长期的研究兴趣是研究无监督预训练方法有效的机理。

在音频伪造检测方面, 在国际上率先提出“部分音频伪造”的概念, 并建立了数据集 benchmark 和基线系统, 并定义了评价指标。相比之前“全真全假”音频, “部分音频伪造”伪装性更强。另外, 针对现有生成式语音合成越来越逼真的问题, 提出了 TraceableSpeech, 在结合语音合成系统增加数字水印, 让合成音频可追溯。

部分研究项目介绍

非自回归快速语音识别 (中科院 2020.1-2021.6)

提出了一个基于非递归的前馈神经网络的非自回归语音识别系统 LASO。提出 Position Dependent Summarizer (PDS) 模块来将声学层语义表示转换为 Token 级别语义表示。在推理阶段, 系统直接将

概率最高的 Token 预测出来，而不需要 beam-search，大大减小识别耗时。所提方法在公开数据集获得 6.4% 的 CER，超过当年的 state-of-the-art 自回归模型 (6.7%)，但识别耗时只有其 1/50。进一步地，在语音识别领域首次提出语音-文本模态对齐 (*alignment*) 概念，利用跨模态知识迁移，将纯文本模态模型的知识迁移到纯语音模态，进一步地提升了准确率。相关论文发表在 INTER-SPEECH2020 以及期刊 IEEE/ACM Transactions on Audio Speech and Language Processing。

端到端语音识别系统的外部文本知识集成 (中科院 2020.1-2021.6)

提出了基于老师-学生学习的方法 LST(Learn Spelling from Teachers)，来将外部语言模型的知识集成到端到端语音识别系统中。相比传统 Shallow Fusion，Deep Fusion 等方法，该方法不增加推理代价。进一步地，提出了因果完形填空器语言模型，将整句文本知识压缩到语言模型，然后利用 LST 方法，将整句文本知识迁移到端到端语音识别系统。相关工作发表在 Interspeech2019 以及期刊 IEEE/ACM Transactions on Audio Speech and Language Processing。LST 方法还被悉尼科学院陶大程院士等评价为“一种有效的利用语言知识的手段”。

“部分伪造音频”检测 (中科院 2020.1-2021.6)

提出了“部分音频伪造”的概念，并构建数据集、评价指标和基线系统。相比传统“全真全假”的音频半真半假的“部分音频伪造”隐蔽性更高，危害更大，而技术上定位、溯源难度更大。相关工作发表在 Interspeech2021。

可追溯的生成语音 (中科院 2024.1-2024.6)

针对语音合成越来越逼真，合成音鉴别越来越难的问题，提出了一种“可追溯的语音”，结合合成系统端到端训练数字水印模块。该方法可以在将生成质量影响控制到很小的情况下，将水印信息加入到音频中。相关工作发表在 Interspeech2024。

部分系统项目介绍

音频理解系统 (字节跳动 2023.3-2024.8)

研发超大规模语音预训练模型和生成式端到端音频理解系统。在超千万小时规模的无监督音频数据上开展预训练，主要负责研发基于 teacher-student learning 架构的无监督预训练编码器模型。研发的编码器模型应用到 Seed-ASR——一种生成式语音识别系统，该系统相对之前的语音识别模型，相对错误率下降 20% 40%。该系统主要应用于字节跳动包括豆包在内的各种涉及语音交互的应用中。技术报告已公开。

稀疏动态模型 mixture-of-experts (MoE) (快手 2022.2-2023.3)

研发快手基于混合专家模型 MoE 的 Conformer 模型，构建大规模稀疏动态模型的语音识别系统。稀疏动态模型在推理时自适应地选择合适的前馈子模型路径，所以可以在保持模型推理速度的情况下，水平扩展模型的规模，提升最终的准确率。本人实现基于大规模混合专家模型 MoE 的语音识别模型，在大规模数据上相对提升 8%。同时还结合递归计算和 MoE 适配，研发了模型参数高效的 MoE-Conformer 模型，使用 1/3 的 Encoder 参数量实现了和 SOTA 模型接近的效果，该工作发表在 Interspeech2022。

高并行流式/非流式 Transducer 语音识别系统 (快手 2021.7-2023.3)

研发快手帧同步高并行流式/非流式 Conformer-Transducer 语音识别系统，其中非流式系统支持内容理解业务 (如字幕、剪辑等)，流式系统支持语音交互业务 (如客服等)。Transducer 是一种帧同步语音识别系统，具备性能好，天然支持流式的特点，适用范围广泛。本人研发了 Transducer 语音识别系统，实现了向量化高并行 Transducer 解码算法，并推进后续异构计算加速；实现了高并行流式 Chunk Transducer 解码算法，单 GPU 卡支持多路呼入。针对 Transducer 系统，还进一步地开发了基于 WFST 的热词功能，可根据用户需求灵活提升特定词汇准确率。

LAS 语音识别系统优化 (字节跳动 2019.5-2019.12)

研发产品级 LAS 端到端语音识别系统并超越传统 CTC 系统，是业内率先上线的端到端语音识别系统，支持抖音、剪映、TicTok 的审核、字幕业务。主要优化有：

1. 实现 **LST 算法**。基于老师-学生学习，将大规模语言模型中的知识集成到 LAS 系统。相对于基线系统，CER 相对下降 10%。该算法由本人提出。
2. 实现 **MWER 训练**。实现最小词错误率 (MWER) 训练准则，最小化 LAS 系统期望词错误率。相

比基于 LST 训练的模型, CER 进一步相对下降 10%。

3. 实现 **CLAS**。实现 Neural Biasing 机制来让 LAS 系统解码出用户定制词汇 (如人名, 歌曲名)。

4. 实现 CTC 强制对齐, 来生成字幕时间戳。

语音关键词检测 (中科院 2019.5-2020.1)

提出共享权值自注意力机制来进行语音关键词检测。该模型使用 TDNN 和自注意力机制作为基本模块, 来构建 DeepKWS 系统。我们基于 self-attention 的输入相同。所以 Attention 在同一空间这一特点, 提出共享权值的方法, 基于共享权值的方法, 模型整体参数量大大减小。所提方法准确率接近当年 SOTA 模型 ResNet, 但参数量只有其 1/20。该工作发表在 Interspeech2019。

虚拟人动作生成系统 (快手 2022.7-2023.3)

构建虚拟人驱动系统, 使虚拟人根据音乐进行歌唱动作或者舞蹈, 以使虚拟人短视频制作摆脱对中之人的依赖。本人实现驱动部分, 项目分为两块: 1) 基于模板的半自动方法, 该方法根据音乐的 BPM, 选择动作模板并进行拼接。基于此方法, 将虚拟人短视频制作周期从 2 周减少到 2 小时。2) 实现基于 VAE、VQVAE 的动作生成模型, 根据音乐自动地生成舞蹈动作。

部分发表论文

在国际顶级期刊、会议发表论文 30 余篇, 完整列表请见 Google Scholar: <https://sourl.cn/EKJM8t>

1. **Ye Bai**, Jiangyan Yi, Jianhua Tao, Zhengkun Tian, Zhengqi Wen, Shuai Zhang: Fast End-to-End Speech Recognition Via Non-Autoregressive Models and Cross-Modal Knowledge Transferring From BERT. IEEE/ACM Trans. Audio, Speech & Language Processing, 2021
2. **Ye Bai**, Jiangyan Yi, Jianhua Tao, Zhengqi Wen, Zhengkun Tian, Shuai Zhang: Integrating Knowledge Into End-to-End Speech Recognition From External Text-Only Data. IEEE/ACM Trans. Audio, Speech & Language Processing, 2021
3. **Ye Bai**, Jie Li, Wenjing Han, Hao Ni, Kaituo Xu, Zhuo Zhang, Cheng Yi and Xiaorui Wang, Parameter-Efficient Conformers via Sharing Sparsely-Gated Experts for End-to-End Speech Recognition, Interspeech2022
4. **Ye Bai**, Jiangyan Yi, Jianhua Tao, Zhengkun Tian, Zhengqi Wen and Shuai Zhang, Listen Attentively, and Spell Once: Whole Sentence Generation via a Non-Autoregressive Architecture for Low-Latency Speech Recognition, Interspeech2020
5. **Ye Bai**, Jiangyan Yi, Jianhua Tao, Zhengkun Tian and Zhengqi Wen, Learn Spelling from Teachers: Transferring Knowledge from Language Models to Sequence-to-Sequence Speech Recognition, Interspeech2019
6. **Ye Bai**, Jiangyan Yi, Jianhua Tao, Zhengqi Wen, Zhengkun Tian, Chenghao Zhao and Cunhang Fan, A Time Delay Neural Network with Shared Weight Self-Attention for Small-Footprint Keyword Spotting, Interspeech2019
7. **Ye Bai**, Jiangyan Yi, Jianhua Tao, Zhengqi Wen, Bin Liu, Voice Activity Detection Based on Time-Delay Neural Networks, APSIPA2019
8. **Ye Bai**, Jianhua Tao, Jiangyan Yi, Zhengqi Wen, Cunhang Fan, Jianhua Tao: CLMAD: A Chinese Language Model Adaptation Dataset. The 11th International Symposium on Chinese Spoken Language Processing (ISCSLP 2018)
9. **Ye Bai**, Jiangyan Yi, Hao Ni, Zhengqi Wen, Bin Liu, Ya Li, Jianhua Tao: End-to-end keywords spotting based on connectionist temporal classification for Mandarin. The 10th International Symposium on Chinese Spoken Language Processing (ISCSLP 2016)
10. Jiangyan Yi, Jianhua Tao, Zhengqi Wen, **Ye Bai**: Adversarial Multilingual Training for Low-Resource Speech Recognition. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)

11. Zhengkun Tian, Jiangyan Yi, Jianhua Tao, **Ye Bai**, Zhengqi Wen: Self-Attention Transducers for End-to-End Speech Recognition. Interspeech2019
12. Jiangyan Yi, **Ye Bai**, Jianhua Tao, Haoxin Ma, Zhengkun Tian, Chenglong Wang, Tao Wang, Ruibo Fu: Half-truth: A partially fake audio detection dataset. Interspeech2021
13. Junzuo Zhou, Jiangyan Yi, Tao Wang, Jianhua Tao, **Ye Bai**, Chu Yuan Zhang, Yong Ren, Zhengqi Wen: TraceableSpeech: Towards Proactively Traceable Text-to-Speech with Watermarking. Interspeech2021

授权专利

1. 白烨, 王晓瑞, 语音识别方法、装置、电子设备及计算机可读介质. 专利号: ZL202210891670
2. 白烨, 温正棋, 基于前馈神经网络的低延时语音识别模型及训练方法. 专利号: ZL201910866156
3. 温正棋, 白烨, 一种基于知识迁移的序列到序列语音识别模型训练方法. 专利号: ZL202010988191

学术服务

- CCF 语音对话与听觉专委会委员。
- ICASSP, INTERSPEECH, IEEE/ACM Trans. Audio, Speech & Language Processing, Speech Communication, Journal of Signal Processing Systems, Artificial Intelligence Review 等人工智能/语音领域顶级期刊和会议审稿人。
- INTERSPEECH 2020 大会组织本地志愿者组织人。Interspeech 2020 学生会议本地联络员。