白 烨
快手科技
北京市海淀区上地西路 6 号
邮箱: baiye@cau.edu.cn
电话: 13021981594

研究兴趣

语音识别、语言模型、关键词检测、虚拟人驱动生成模型

工作经历

语音算法专家 (快 Star 招聘计划) *2021 年 7 月 - 现在*
快手科技
构建和优化大规模语音识别系统，支持短视频、直播、客服的内容理解和字幕展示。典型应用的 APP 有快手、快影、Kwai 等。研发虚拟人动作驱动方案和生成模型，应用在快手虚拟歌手“张风琴”，获得数十万粉丝关注。

语音实习生 *2019 年 4 月 - 2020 年 1 月*
字节跳动 *AI Lab.*
优化大规模端到端语音识别模型。所研发的算法应用在抖音、Tiktok、剪映等应用的审核与字幕功能。

教育经历

中国科学院自动化研究所 *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

学术服务

- CCF 语音对话与听觉专委会委员。
- ICASSP, INTERSPEECH, Speech Communication, Journal of Signal Processing Letter 等语音领域会议和期刊审稿人。
- 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。

技能

编程: Python, C/C++

工具: TensorFlow, PyTorch, KALDI, Lingvo

语言: 中文, English

部分荣誉/奖励

- **中国科学院大学三好学生** *2019*
- **ISCSLP 2018 最佳学生论文候选** *2018*
- **京东金融语音识别大赛冠军 (1/240)** *2018*

部分项目介绍

高并行流式/非流式 Transducer 语音识别系统 (2021.7-)
研发帧同步高并行流式/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，实现单 GPU 卡支持多路呼入 (目前的开源实现一般不具备向量化并行解码能力)。针对 Transducer 系统，还进一步地开发了基于 WFST 的热词功能，实现根据用户需求灵活提升特定词汇准确率。

稀疏动态模型 mixture-of-experts (MoE) (2022.2-)
研发基于混合专家模型 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 Interspeech2021.

LAS 语音识别系统优化 (字节跳动 2019.5-2019.12)
优化产品级端到端语音识别系统，支持抖音、剪映、TicTok 审核、字幕。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。相对于极限系统，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。

非自回归快速语音识别
提出了一个基于非递归的前馈神经防落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，提出了跨模态知识迁移，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 INTERSPEECH2020 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
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。相比传统的 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。

语音关键词检测
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 (2022.7-)
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

完整列表请见 Google Scholar: http://dw-z.ink/15K4A

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. Jiangyan Yi, Jianhua Tao, Zhengqi Wen, **Ye Bai**: Adversarial Transfer Learning for Low-Resource Speech Recognition. IEEE/ACM Trans. Audio, Speech & Language Processing

12. Zhengkun Tian, Jiangyan Yi, Jianhua Tao, **Ye Bai**, Zhengqi Wen: Self-Attention Transducers for End-to-End Speech Recognition. Interspeech2019