Ye Bai

Bytedance Inc.
43 N 3rd Ring W Rd, Beijing, China
E-mail: baiye@cau.edu.cn

RESEARCH INTERESTS

Speech Recognition, Language Modeling, Multimodal Modeling.

WORK EXPERIENCE

Research Scientist *March 2023 - Present*
Bytedance Inc.
Building large-scale speech understanding systems. Leading a project on large-scale unsupervised pre-trained acoustic models to achieve state-of-the-art speech recognition systems, which support billion-user products of Bytedance. Building industry's first end-to-end speech question-answering system.

Speech Engineer *June 2021 - March 2023*
Kuaishou Technology Co., Ltd
Building and optimizing large-scale speech recognition systems for understanding short videos and live streaming. A typical function is to automatically generate subtitles for short videos. The systems serve multiple billion-user products, including Kuaishou, Kwai, Kuaiying. Also building transducer-based speech recognition systems to support Kuaishou intelligent customer service.

Building virtual human motion generation systems, which are applied to Kuaishou virtual singer "Zhang Fengqin" (about 500 thousands followers).

Research Intern *Apr. 2019 - Jan. 2020*
Bytedance Technology Co., Ltd
Optimizing end-to-end speech recognition models. The implemented algorithms support online speech recognition systems, which serve multiple billion-user products, including Douyin, Tiktok, Jianying.

EDUCATION

Institute of Automation, Chinese Academy of Sciences *Sep. 2016 - June 2021*
Ph.D in Pattern Recognition and Intelligent Systems
Advisor: Prof. Jianhua Tao

China Agricultural University *Sep. 2012 - June 2016*
Bachelor in Communication Engineering

SERVICES

- A member of CCF Technical Committee on Speech Dialogue and Auditory Processing
- Reviewer: ICASSP, INTERSPEECH, Speech Communication, Journal of Signal Processing Letter.
- Assisting to organize INTERSPEECH 2020 as the leading volunteer. Organizing Student Events of INTERSPEECH 2020 as the local coordinator.

SKILLS

Programming: Python, C/C++

Tools: TensorFlow, PyTorch, KALDI, Lingvo

Languages: Chinese, English

SELECTED HONORS/AWARDS

- **Merit Student of University of Chinese Academy of Sciences** *2019*
- **Best Student Paper Candidate of ISCSLP 2018** *2018*
- **Champion of Jingdong Finance Speech Recognition Competition (1/240)** *2018*
  I built an ASR system, which achieved a top-1 score in the competition, based on the telephone dataset of Jingdong Finance in one week. The absolute value of CER is lower than the second team by 2%. The other teams attending this competition included Xiaomi Inc., Cheetah Mobile Inc.

SELECTED PUBLICATIONS

Full list in 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
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
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. Cunhang Fan, Bin Liu, Jianhua Tao, Zhengqi Wen, Jiangyan Yi, **Ye Bai**: Utterance-level Permutation Invariant Training with Discriminative Learning for Single Channel Speech Separation. The 11th International Symposium on Chinese Spoken Language Processing (ISCSLP 2018)

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

PROJECTS

Parallel Streaming/Non-streaming Transducer-based Speech Recognition (July 2021-)
Developing parallel frame-synchronous streaming/non-streaming Conformer-Transducer speech recognition systems. The non-streaming system is used in Kuaishou video understanding and content display, including live streaming subtitle generation, automatic video editing. The streaming system is used for intelligent customer service. I developed vectorized beam-search decoding algorithm for the transducer systems, thus one GPU card can process multiple audio streams in parallel. I also deploy WFST-based hotword biasing function for the system to supporting user-defined word recognition. I also developed heterogeneous computation inference library with colleagues to speed-up the systems.

Dynamic Routing Models with Sparsely-Gated Mixture-of-Experts (MoE) (Fab. 2022-)
Developing MoE based Conformer systems. The system can select forward paths adaptively, thus the model can improve the scale without reduction of the inference speed. The implemented MoE-based ASR model achieves relative character error rate reduction of 8%. Further, I proposed a parameter-efficient MoE model with recursive computation, which uses about 1/3 of parameters of the encoder to achieve similar performance with SOTA conformer model. The work is publicated on INTERSPEECH 2022.

Optimizing LAS Based ASR Systems (Bytedance. May 2019 - Dec. 2019)
Optimizing production-level end-to-end ASR systems, which supports content understanding and subtitle generation of billion-user applications including Douyin, Jianying, TikTok. The main work includes:
1. Implemented **LST algorithm**. Integrating knowledge from external text into the LAS system via teacher-student learning. The CER relative reduction is 10% compared with the baseline. This algorithm is proposed by myself.
2. Implemented **MWER training**. Implement minimum word error rate (MWER) loss to minimize expected word error rate of the LAS system. Compared with the system trained with LST, the CER reduced by 10%.
3. Implemented **CLAS**. Implement a biasing mechanism to guide the LAS system to decode out-of-vocabulary words (such as person names, song titles). It can improve the performance for bad cases. The relative reduction of CER compared with the baseline is 10%.
4. Implemented CTC based forced alignments for generating timestamps.

Non-Autoregressive Architectures for Fast ASR (Feb. 2020 - Feb. 2021)
We proposed a non-recurrent feedforward neural network based non-autoregressive system for low-latency ASR. We proposed a Position Dependent Summarizer (PDS) module which represents semantic corresponding to each token position. At the inference stage, the system selects the most likely token at each position instead of beam-search, so that the inference time cost is much reduced. The proposed system achieves CER 6.4% performance on public dataset AISHELL-1, which outperforms state-of-the-art autoregressive system (6.7%). And the decoding latency is 1/50 of the autoregressive transformer model. This work is published on INTERSPEECH2020.

Further, I proposed a cross-modal knowledge transferring method to use the knowledge in large-scale pretrained language models. This work is published on the journal IEEE/ACM Transactions on Audio Speech and Language Processing.

Integrating Knowledge into End-to-End ASR Systems from External Text-Only Data
We proposed a teacher-student learning based method called LST (Learn Spelling from Teachers), to integrate external knowledge into an end-to-end ASR system. First, the knowledge is represented into a language model. Then, the knowledge is distilled into the end-to-end system. Compared with fusion based methods, the method does not increase complexity during inference. This work is published on INTERSPEECH2019.

To further integrate the whole context in a sentence (both the left context and the right context of a word), we proposed a self-attention based language model called Casual clOze completeR (COR), which estimates the probability of a word given the left context and the right context. Then we use COR as the teacher language model to train the ASR system. Therefore the ASR system uses both the left context and the right context. This work is published on the journal IEEE/ACM Transactions on Audio Speech and Language Processing.

Shared Weight Self Attention for Keyword Spotting
We proposed to share weights of the self-attention mechanism for keyword spotting. We use time-delay neural networks and self-attention as the basic block to build a DeepKWS system. We found that the inputs of self-attention are the same, and the core operation of self-attention is dot-product. So the attention inputs can be in the same space. So we propose to share the weights of the self-attention. This reduces the footprint of the model but does not influence the performance. The performance of the model is close to the state-of-the-art ResNet model, but the number of parameters is 1/20. This work is published on INTERSPEECH2019.

Optimizing KALDI Based ASR Systems (July 2017 - July 2020)
My work includes:
1. **Customization**. I simplified KALDI system and transplanted it to Windows and Android platforms.
2. **Training Acoustic Models**. Optimizing acoustic models on production-level speech datasets.
3. **Training Language Models**. Optimizing and customizing language models for users.
The systems are applied with The State Grid Corporation of China, CRRC Corporation, and Institute of Information Engineering, Chinese Academy of Sciences, etc.

Wake Word Spotting, Huawei Inc. (July 2019 - Feb. 2020)
Develop neural network based keyword spotting systems for low-resource setting. We propose a BERT-like unsupervised method for the low-resource keyword spotting.

Motion Generation for Virtual Human (July 2022 -)
Building motion generation systems to generate motions of virtual human with music, thus we can make videos of virtual human without actors. I developed motion driving part of the whole pipeline, including: 1) template-based semi-automatic methods, which generates motions by selecting motion unit, in terms of BPM of music; 2) VAE/VQVAE-based motion generation methods generate dance motions with music. Based on the methods, we can make without actors and the time cost to make a video of virtual human is reduced from 2 weeks to 2 hours.