Ye Bai

Kuaishou Technology Co., Ltd
Shangdi West Road No.6, Beijing, China
E-mail: baiye@cau.edu.cn
Phone: +86-13021981594

RESEARCH INTERESTS

Speech Recognition, Language Modeling, Keyword Spotting.

WORK EXPERIENCE

Speech Engineer *June 2021 - Present*
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.

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

- 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

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**, 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**, 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
4. **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
5. **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
6. **Ye Bai**, Jiangyan Yi, Jianhua Tao, Zhengqi Wen, Bin Liu, Voice Activity Detection Based on Time-Delay Neural Networks, APSIPA2019
7. **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)
8. **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)
9. 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)
10. Jiangyan Yi, Jianhua Tao, Zhengqi Wen, **Ye Bai**: Adversarial Transfer Learning for Low-Resource Speech Recognition. IEEE/ACM Trans. Audio, Speech & Language Processing
11. 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)

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

RESEARCHES

Non-Autoregressive Architectures for Fast ASR

We propose a non-recurrent feedforward neural network based non-autoregressive system for low-latency ASR. We propose 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, we propose a cross-model knowledge transferring method to use the knowledge in large-scale pretrained language models. The extended version paper 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 propose 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 propose 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. The extended version paper is published on the journal IEEE/ACM Transactions on Audio Speech and Language Processing.

Shared Weight Self Attention for Keyword Spotting

We propose 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.

Chinese Language Model Adaptation Dataset

We built a Chinese text dataset for language model adaptation. We present a series of language model adaptation experiments based on pretraining-finetuning scheme. This work is published on ISCSLP 2018. The extended journal paper is published on Journal of Signal Processing Systems.

TDNN Based Voice Activity Detection

We propose a small-footprint time-delay neural network based voice activity detection system. Compared with baseline DNN system and LSTM system, relative reduction of EER is 41.26%. This work is published on APSIPA 2019.

CTC Based Keyword Spotting

We propose to use Connectionist Temporal Classification based acoustic model for keyword spotting. This work is published on ISCSLP 2016.

PROJECTS

Optimizing LAS Based ASR Systems (2019.5-2019.12)

Optimizing production-level end-to-end ASR system, which is trained on 10k hours of speech data. The system outperforms a well trained DFSMN-CTC based system on the same dataset and serves for . The main work includes:
1. Implement **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. Implement **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. Implement **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. Implement CTC based forced alignments for generating timestamps.

Transformer Based ASR System (2019.10-2020.1)

A PyTorch based Speech-Transformer system. The features include:
1. **Minimal Dependency**. The system does not depend on external software for feature extraction or decoding. Users just install PyTorch deep learning framework.
2. **Good Performance**. The system includes advanced algorithms, such as Label Smoothing, SpecAug, LST, and achieves good performance on AISHELL1. The baseline CER on AISHELL1 test is 6.6%, which is better than ESPnet.
3. **End to End**. The feature extraction and tokenization are online. The system directly processes wav files. Thus, the procedure is much simplified.

The project is released at https://github.com/by2101/OpenASR.

Optimizing KALDI Based ASR Systems (2017.7-)

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. (2019.5-)

Develop neural network based keyword spotting systems for low-resource setting. We propose a BERT-like unsupervised method for the low-resource keyword spotting.

Voice Activity Detection Systems (2017.7-2017.12)

Implement a time-delay neural network based VAD system as the front end of the ASR system for extracting speech in the audio stream.