Zhehuai (Tom) Chen

SJTU SpeechLab
Department of Computer Science and Engineering
Shanghai Jiao Tong University
3-502 SEIEE Building, 800 Dongchuan Road, Shanghai, 200240

Phone: +086 15921010742
Email: dian.chenzhehuai@gmail.com
chenzhehuai@foxmail.com
Skype: chenzhehuai@outlook.com

RESEARCH INTERESTS

Weighted finite-state transducers (WFST), static and dynamic search algorithm (Decoder). Discriminative training, end-to-end (E2E), keyword spotting (KWS), robust ASR, language model. Distributed and GPU-based parallel training and inference.

EDUCATION

2014 - 2019, Ph.D., Computer Science, Shanghai Jiao Tong University.
Supervised by Prof. Kai Yu (`https://speechlab.sjtu.edu.cn/members/kai_yu`).

2010 - 2014, B.E., Electronics and Information Engineering, Huazhong University of Science and Technology.

INTERNSHIP

2018, ***Facebook*** Research Internship (mentor: Christian Fuegen).

2018, ***JHU*** Visiting Research Scholar (mentor: Daniel Povey),
remote collaboration with ***NVIDIA*** (mentor: Justin Luitjens).

2017, ***Microsoft*** Research Internship (mentor: Jasha Droppo).

2014 - present, ***AISpeech*** Internship (ASR training & serving platform).

2014, ***HP*** Internship.

AWARDS

2018, ***Google*** Ph.D. Fellowship (World, total 40 Ph.D. students).

2018, ***Best Paper Nomination***, ISCA InterSpeech.

2017, ICASSP 2017 student travel grant.

2016, InterSpeech 2016 student travel grant.

2013, ***Microsoft*** Young Fellows Scholarship (Asia, total 36 undergraduates).

2013, Excellent Student (University, top 2%).

2013, National Undergraduate Electronic Design Contest (National, Second Prize).

PROFESSIONAL EXPERIENCE

End-to-end ASR Research Internship (Facebook, Menlo Park) 2018.9 - 2018.12

Working on end-to-end speech recognition. We advance the end-to-end contextual speech recognition and submit two papers on this topic. We also work out a distributed training framework for sequence-to-sequence models based on PyTorch and Block Momentum.

GPU WFST Decoder Visiting Research Scholar (JHU & NVIDIA) 2018.1 - 2018.4

Working on an extension [1] of the Kaldi toolkit that supports WFST decoding on GPUs, supervised by Daniel Povey (`http://www.danielpovey.com`).

[1] `https://github.com/chenzhehuai/kaldi/tree/gpu-decoder`

The lattice based WFST decoder achieves identical results and significant speedups (**15**-fold for single sequence and **46**-fold with sequence parallelism). We submit a conference paper on this topic. Remote collaboration with NVIDIA. Contributor of Kaldi.

Robust ASR Research Internship (MSR, Redmond) 2017.5 - 2017.7

Research internship in *speech and dialog research group*, Microsoft Research, Redmond, supervised by Jasha Droppo (`https://www.microsoft.com/en-us/research/people/jdroppo/`).

Significantly advancing the state-of-the-art unsupervised single-channel overlapped speech recognition system and publishing a transaction and a conference paper on this topic. CNTK contributor.

Speech Recognition End-to-end ASR and Decoding Framework 2014.8 - present

1. Inference framework in CTC. The proposed PSD framework achieves **5**-fold speedup versus traditional CTC-based system and **30**-fold speedup versus HMM-based system. The framework can be extended to LF-MMI.
2. End-to-end (E2E) speech recognition. Propose modular training strategy for direct acoustics-to-words (A2W) modeling.
3. Parallelize WFST compose, determinize and minimize algorithms and achieve 1-time speedups.
4. Implement a dynamic decoder which interpolates language models on-the-fly. Design a fast n-gram hash map to alleviate on-the-fly composition and rescoring overheads.
5. Sequence discriminative training in KWS. Solve the search space modeling problem in KWS. Sequence discriminative training in both HMM and CTC achieves significant improvement.
6. Confidence measure in CTC. The proposed posterior based confidence measure achieves significant improvement versus traditional methods in both CTC and HMM trained models.
7. LSTM language modeling (LM) and lattice rescoring. Speed up the lattice rescoring significantly. The improvement includes model inference, history clustering and stream parallelization.
8. Human-directed ASR errors are collected from confusion network. BLSTM LM is trained to estimate sentence completion scores, combined with confusion network scores to do correction.
9. Design a parametric model, which can be inferenced with offline decoding records of the whole process, to tune beam dynamically by features in decoding process so as to thoroughly speedup.

Speech Synthesis Speech Synthesis using HMM & DNN 2014.3-2014.7

Develop HMM & DNN Speech Synthesis systems and analyze the performance gap between them.

PUBLICATIONS

Zhehuai Chen, Mahaveer Jain, Yongqiang Wang, Michael Seltzer, Christian Fuegen, Joint Grapheme and Phoneme Embeddings for Contextual End-to-End ASR, 20th Annual Conference of International Speech Communication Association (InterSpeech), Graz, Austria, 2019.

Zhehuai Chen, Mahaveer Jain, Yongqiang Wang, Michael Seltzer, Christian Fuegen, End-to-end Contextual Spebech Recognition using Class Language Models and a Token Passing Decoder, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019.

Zhehuai Chen, Wenlu Zheng, Yongbin You, Yanmin Qian, Kai Yu. Label Synchronous Decoding for Speech Recognition. Chinese Journal of Computers, 2019.

Zhehuai Chen, Justin Luitjens, Hainan Xu, Yiming Wang, Daniel Povey, Sanjeev Khudanpur, A GPU-based WFST Decoder with Exact Lattice Generation, 19th Annual Conference of International Speech Communication Association (InterSpeech), 2018 [**Best Paper Nomination**].

Zhehuai Chen, Linguistic Search Optimization for Deep Learning Based LVCSR, in Doctoral Consortium, 19th Annual Conference of International Speech Communication Association (InterSpeech), 2018.

Zhehuai Chen, Jasha Droppo, Sequence Modeling in Unsupervised Single-channel Overlapped Speech Recognition, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, Canada, 2018.

Zhehuai Chen, Qi Liu, Hao Li, Kai Yu, On Modular Training of Neural Acoustics-to-word Model for LVCSR, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, Canada, 2018.

Zhehuai Chen, Yanmin Qian, Kai Yu, Sequence Discriminative Training for Deep Learning based Acoustic Keyword Spotting. Speech Communication, vol. 102, 100-111, 2018.

Zhehuai Chen, Jasha Droppo, Jinyu Li, Wayne Xiong, Progressive Joint Modeling in Unsupervised Single-channel Overlapped Speech Recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 1, pp. 184-196, Jan. 2018. doi: 10.1109/TASLP.2017.2765834.

Zhehuai Chen, Yanmin Qian, and Kai Yu. A unified confidence measure framework using auxiliary normalization graph, IScIDE, 2017.

Zhehuai Chen, Yimeng Zhuang, Kai Yu. Confidence Measures for CTC-based Phone Synchronous Decoding. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, USA, 2017.

Zhehuai Chen, Yimeng Zhuang, Yanmin Qian, Kai Yu. Phone Synchronous Speech Recognition with CTC Lattices. IEEE/ACM Transactions on Audio, Speech and Language Processing, vol. 25, no. 1, pp. 86-97, Jan. 2017. doi: 10.1109/TASLP.2016.2625459.

Zhehuai Chen, Wei Deng, Tao Xu, Kai Yu. Phone Synchronous Decoding with CTC Lattice. 17th Annual Conference of the International Speech Communication Association (InterSpeech), San Francisco, America, 2016.

Zhehuai Chen, Kai Yu, An Investigation of Implementation and Performance Analysis of DNN Based Speech Synthesis System. 12th IEEE International Conference on Signal Processing (ICSP), Hangzhou, 2014.

Mingkun Huang, Yongbin You, **Zhehuai Chen**, Yanmin Qian, Kai Yu. Knowledge Distillation for Sequence Model, 19th Annual Conference of International Speech Communication Association (InterSpeech), 2018.

Yue Wu, Tianxing He, **Zhehuai Chen**, Yanmin Qian and Kai Yu. Multi-view LSTM Language Model with Word-synchronized Auxiliary Feature for LVCSR, CCL, 2017.

Da Zheng, **Zhehuai Chen**, Yue Wu, Kai Yu, Directed Automatic Speech Transcription Error Correction Using Bidirectional LSTM. International Symposium on Chinese Spoken Language Processing (ISCSLP), Tianjin, China, 2016.

Bo Chen, **Zhehuai Chen**, Jiachen Xu, Kai Yu. An Investigation of Context Clustering for Statistical Speech Synthesis with Deep Neural Network. 16th Annual Conference of the International Speech Communication Association (InterSpeech), Dresden, Germany, 2015.