KAI ZHEN

http://kaizhen.us • kaizhen723@gmail.com

EMPLOYMENT

Full-Time

Amazon.com, Inc.

- Applied Scientist II
  - Artificial General Intelligence (AGI), Pittsburgh, PA — Apr. 2021 – present
    - Developed and released various on-device speech recognition production models for Alexa
    - Optimized inference efficiency for audio encoders and large language models (LLMs)

Internship

Amazon.com, Inc.

- Applied Scientist Intern
  - Alexa Speech, Pittsburgh, PA — Summer 2020
    - Project: Network Compression for On-Device Speech Recognition
      <Best Internship Poster Presentation>

LinkedIn Corporation

- Machine Learning & Relevance Intern
  - Ads-AI Group, Mountain View, CA — Summer 2019
    - Project: Ads Response Rate Prediction with Language Model Enriched Semantic Features
  - Company Standardization Group, New York City, NY — Summer 2018
    - Project: Relevance Ranking Using Recurrent Neural Network for LinkedIn Resume Builder

Academic Part-Time

Indiana University — Aug. 2015 – Mar. 2021

- Research Assistant: Audio Signal Analysis/Synthesis Technology Based on Machine Learning
  - Published in leading machine learning and speech processing conferences and journals
  - Contributed to 5 US patents as an inventor
- Associate instructor in Department of Computer Science and Intelligent Systems Engineering

EDUCATION

Ph.D., dual major in Computer Sciences and Cognitive Science — May. 2021

- Indiana University, Bloomington, United States
- Committee: Minje Kim (chair, IU Intelligent Systems Engineering), Robert Goldstone (co-chair, IU Cognitive Science), Donald Williamson (IU Computer Science), and Shen Yi (U. of Washington, Speech and Hearing Sciences)
- Dissertation: "Neural Waveform Coding: Scalability, Efficiency and Psychoacoustic Calibration"
  <Winner of the Outstanding Research Award (IU Cognitive Science)>

M.S., major in Computer Science — Jul. 2015

- Tsinghua University, Beijing, China

B.S., major in Software Engineering — Jul. 2012

- Xidian University, Xi'an, China

PROFESSIONAL ACTIVITIES

Conference Reviewer

- IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP): 2019 - 2024
- ISCA Interspeech: 2022 - 2023
- EURASIP European Signal Processing Conference (EUSIPCO): 2022 - 2023
- IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA): 2021 - 2023
- IEEE International Conference on Data Mining (ICDM): 2020
- Association for the Advancement of Artificial Intelligence (AAAI): 2017 - 2018

Journal Reviewer

- European Association for Signal Processing (EURASIP) Journal on Audio, Speech, and Music Processing
- IEEE MultiMedia
- Speech Communication

PUBLICATIONS

International Journal Articles

[J002] **Kai Zhen**, Jongmo Sung, Mi Suk Lee, Seungkwon Beack, and Minje Kim, "Scalable and Efficient Neural Speech Coding: A Hybrid Design", *IEEE/ACM Transactions on Audio, Speech, and Language Processing* ***(IEEE/ACM TASLP), 30 (2021): 12-25.***

- ***This is an extended version of [C001] and [C003].***

[J001] **Kai Zhen**, Mi Suk Lee, Jongmo Sung, Seungkwon Beack, and Minje Kim, "Psychoacoustic Calibration of Loss Functions for Efficient End-to-End Neural Audio Coding," *IEEE Signal Processing Letters* ***(SPL) 27 (2020): 2159-2163.***

- ***We proposed a novel method to compress audio signals without degrading the quality. It can facilitate a faster data transmission and reduce energy cost. See the demo page.***

Referred International Conference Proceedings

[C009] Rupak Vignesh Swaminathan, Grant Strimel, Ariya Rastrow, Harish Mallidi, **Kai Zhen**, Hieu Nguyen, Nathan Susanj, Athanasios Mouchtaris, "*Max-Margin Transducer Loss: Improving Sequence-Discriminative Training Using a Large-Margin Learning Strategy*," in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (**ICASSP**), Seoul, Korea, 14-19 April, 2024.

- ***Is there a better way than minimum word error rate (MWER) for training ASR Transducers? Our paper introduces Max Margin Transducer (MMT). It does a better job of telling apart positive and negative samples among n-best hypotheses.***

[C008] Martin Radfar, Paulina Lyskawa, Brandon Trujillo, Yi Xie, **Kai Zhen**, Jahn Heymann, Denis Filimonov, Grant Strimel, Nathan Susanj, Athanasios Mouchtaris, "*Conmer: Streaming Conformer with no self-attention for interactive voice assistants*," *In Proc. Annual Conference of the International Speech Communication Association (**Interspeech**)*, Dublin, Ireland, August 21-24, 2023.

- ***We proposed an alternative model architecture to the state-of-the-art that can perform better on voice assistants.***

[C007] **Kai Zhen**, Martin Radfar, Hieu Duy Nguyen, Nathan Susanj, Grant Strimel, Athanasios Mouchtaris, "*Sub-8-bit Quantization for On-Device Speech Recognition: A Regularization-Free Approach*", *IEEE Workshop on Spoken Language Technology (**IEEE SLT**)*, Doha, Qatar, January 9-12, 2023.

- ***This is an improved version of [C006]. The neural efficiency optimization method is model-agnostic, which can be easily adopted in multiple locales, device types, and architectures.***

[C006] **Kai Zhen**, Hieu Duy Nguyen, Raviteja Chinta, Nathan Susanj, Athanasios Mouchtaris, Tariq Afzal, and Ariya Rastrow, *"Sub-8-Bit Quantization Aware Training for 8-Bit Neural Network Accelerator with On-Device Speech Recognition," in Proceedings of Annual Conference of the International Speech Communication Association (Interspeech),* Incheon, Korea, September 18-22, 2022.

- ***We proposed a quantization mechanism to compress speech recognition models by more than 5 times without hindering the accuracy. This helps Alexa run faster to improve customers' experience. The method in this work is also relevant to [C001] during my PhD studies.***

[C005] **Kai Zhen**, Hieu Duy Nguyen, Feng-Ju (Claire) Chang, Athanasios Mouchtaris, *"Sparsification via Compressed Sensing for Automatic Speech Recognition,"* *in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (**ICASSP**)*, Toronto, ON, Canada, June 6-12, 2021.

- ***This is the outcome of my internship with Amazon Alexa. We used a compressed sensing inspired optimization method to prune / compress the model weights by more than 70%. It improves runtime efficiency.***

[C004] Haici Yang, **Kai Zhen,** Seungkwon Beack, Minje Kim, *"Source-Aware Neural Speech Coding for Noisy Speech Compression,"* *in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (**ICASSP**)*, Toronto, ON, Canada, June 6-12, 2021.

- ***This work harmonized speech enhancement and coding in the latent space to generate source-specific representations.***

[C003] **Kai Zhen**, Mi Suk Lee, Jongmo Sung, Seungkwon Beack, and Minje Kim, "Efficient And Scalable Neural Residual Waveform Coding with Collaborative Quantization," *in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (**ICASSP**)*, Barcelona, Spain, May 4-8, 2020.

- ***On top of [C001], we introduced collaborative quantization to better allocate bit resources between linear predictive coding and residual coding.***

[C002] **Kai Zhen**, Mi Suk Lee, Minje Kim. "A Dual-Staged Context Aggregation Method towards Efficient End-To-End Speech Enhancement," *in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (**ICASSP**)*, Barcelona, Spain, May 4-8, 2020.

- ***We proposed a hybrid model to better remove the background noise (such as the one from humming birds and ocean waves) from the speech signals. The enhanced speech sounds more clear and natural. See the demo page.***

[C001] **Kai Zhen**, Jongmo Sung, Mi Suk Lee, Seungkwon Beack, and Minje Kim, "Cascaded Cross-Module Residual Learning towards Lightweight End-to-End Speech Coding," *in Proceedings of Annual Conference of the International Speech Communication Association (**Interspeech**)*, Graz, Austria, September 15-19, 2019.

- ***In this paper, we demonstrated a lightweight and scalable design to compress speech signals significantly without losing the intelligibility. This work is not only applicable to speech signals but also models for Alexa speech services. See the demo page.***

Patents

[P005] Kim, Minje, Mi Suk Lee, Seung Kwon Beack, Jongmo Sung, Tae Jin Lee, Jin Soo Choi, and **Kai Zhen**. "Apparatus and method for speech processing using a densely connected hybrid neural network." U.S. Patent Application 17/308,800, filed November 11, 2021.

- ***The innovation from this patent leads to [C002].***

[P004] Mi Suk Lee, Seung Kwon Beack, Jongmo Sung, Tae Jin Lee, Jin Soo Choi, Minje Kim, **Kai Zhen**, "Method and apparatus for processing audio signal," *US Patent App. 17/156,006*, 2021.

- ***The innovation from this patent leads to [J001].***

[P003] Minje Kim, **Kai Zhen**, Mi Suk Lee, Seung Kwon Beack, Jongmo Sung, Tae Jin Lee, Jin Soo Choi. "Residual coding method of linear prediction coding coefficient based on collaborative quantization, and computing device for performing the method." U.S. Patent Application 17/098,090, filed May 13, 2021.

- ***The innovation from this patent leads to [C003].***

[P002] Mi Suk Lee, Jongmo Sung, Minje Kim, **Kai Zhen**,, "Audio signal encoding method and audio signal decoding method, and encoder and decoder performing the same," U.S. Patent Application No. 16/543,095

- ***The innovation from this patent leads to [C001].***

[P001] Minje Kim, Aswin Sivaraman, **Kai Zhen**, Jongmo Sung, et al, "Audio signal encoding method and apparatus and audio signal decoding method and apparatus using psychoacoustic-based weighted error function", *US Patent Application*, US 2019 / 0164052 A1.

- ***With this innovation, we applied psychoacoustics to the training of speech enhancement models. By using psychoacoustic masks, it allows the occurrence of some noise that is inaudible to humans.***

SKILLS

Speech and audio processing: speech and audio coding, enhancement, recognition, source separation
Model compression and optimization: quantization, sparsification
Machine learning: recurrent neural network transducer, transformer, language modeling
Programming: Python, C, C++, MATLAB

HONORS, AWARDS & SCHOLARSHIP

Outstanding Research Award Apr. 2021

- Given by Cognitive Science Program at Indiana University

Top-Rated Intern Poster Aug. 2020

- Among 17 interns receiving the highest rate out of more than 180 participants

Summa Cum Laude Jul. 2012

- Graduate with honor from Xidian University

China National Scholarship Nov. 2010, Nov. 2011

- For the effort on maintaining top-tier GPA and mathematical contest in modeling (MCM)