ZHIYUAN PENG

Google Scholar,Homepage, Github, Linkedin
+1 919-438-7234 ⋄ jerrypeng1937@gmail.com

EDUCATION

North Carolina State University *2024 - current*
Post-Doc in Computer Science
Research interests: **Large language models (LLMs), Tool-augmented LLMs, AI Safety**

The Chinese University of Hong Kong *2017 - 2023*
Ph.D. in Electronic Engineering
Research interests: **Speech Processing/Recognition, Speaker Recognition, Bayes. Adaptation**
Coursework: DeepLearning | BigData | ProbabilisticModel | SpeechProcessing | DataMining

Harbin Institute of Technology, China *2013 - 2017*
Bachelor of Electronic and Information Engineering, GPA: 90.91/100, Rank 1st
Coursework: C/C++ | FPGA/Verilog | Circuit | Networking | DigitalSignalProcessing | OS

SKILLS

Language: **Python** (**8years**) | **Bash/Linux** (**7years**) | C/C++ (4 years) | SQL|Perl|Java/Javascript
Tools: LangChain/OpenAI | HuggingFace | **PyTorch/Tensorflow**(**6years**)//Keras | Azure/AWS | Git | Docker/Hadoop | Fairseq/Espnet/Kaldi(**5years**)/PyKaldi/Wenet | scikit-learn
Others: MCU and Embedded System Development | FPGA/VerilogHDL/Vivado/SystemHDL

PROJECT EXPERIENCE

LLMs Dec. 2022 - current

- Convenience: Open-sourced Gentopia for easy config. of LLM-powered agents
- Efficiency: QLoRA on *Llama2* LLM for network packet diagnosis | ReWOO to reduce high token use
- Interpretability: Auto-translate numerical features into semantic descript. for root cause analysis
- Scalability: Actively working on augmenting LLMs to master massive APIs (with knowledge graph)

Speech 2017 - current

- Efficiency: Sparsely-shared LoRA for ASR adaptation | Layer Pruning for ASR acceleration
- Extensibility: Prompt tuning for target-speaker ASR | wav2vec2 and RoBERTa for speech disorder assessment | Multi-task learning for language recognition
- Robustness: Bayes. regularize PLDA for adaptation | Model twining against adversarial attacks

Besides, I also had a touch in software defect prediction, using Longformer, and learned token pruning (LTP) for inference acceleration on long code sequences.

WORK EXPERIENCE

Meituan Sept. 2021 - May 2022
Research Intern

- Accelerating wav2vec2 by fbank2vec for Transformer-based self-supervised pre-training of ASR
- Bayesian backend adaptation for speaker verification(Bayes PLDA, Coral)
- Large-scale knowledge distillation for light-weight speaker verifier

PUBLICATIONS

ReWOO: Decoupling Reasoning from Observations for Efficient Augmented Language Models, B. Xu, ***Z. Peng***, *B. Lei, S. Mukherjee, Y. Liu, D. Xu, submitted to ICLR 2024, demo, code*

Sparsely Shared LoRA on Whisper for Child Speech Recognition, *W. Liu, Y. Qin, **Z. Peng**, T. Lee, ICASSP 2024*

Extending Whisper with Prompt Tuning to Target-Speaker ASR, *H. Mao, **Z. Peng**, M. Shao, J. Li, J. Liu, T. Lee, ICASSP 2024*

Co-evolving Data-driven and NLU-driven Synthesizers for Generating Code in Domain Growth and Data Scarcity, *J. Gu, Z. Nan, **Z. Peng**, X. Shen, D. Xu, EMNLP 2023 Workshop*

CoMFLP: Correlation Measure based Fast Search on ASR Layer Pruning, *W. Liu, **Z. Peng**, T. Lee, Interspeech 2023*

Covariance Regularization for Probabilistic Linear Discriminant Analysis, ***Z. Peng**, M. Shao, X. He, K. Ding, T. Lee, G. Wan, ICASSP 2023*

Unifying Cosine and PLDA Back-ends for Speaker Verification, ***Z. Peng**, X. He, K. Ding, T. Lee, G. Wan, Proc. Interspeech 2022*

Label-free Knowledge Distillation with Contrastive Loss for Light-weight Speaker Recognition , ***Z. Peng**, X. He, K. Ding, T. Lee, G. Wan, ISCSLP 2022*

Pairing Weak with Strong: Twin Models for Defending against Adversarial Attack on Speaker Verification, ***Z. Peng**, X. LI, T. Lee, Proc. Interspeech 2021*

Exploiting Pre-Trained ASR Models for Alzheimer's Disease Recognition Through Spontaneous Speech, *Y. QIN, W. LIU, **Z. Peng**, SI Ng, J. LI, H. Hu, T. Lee, NCMMSC 2021*

Mixture Factorized Auto-encoder for Unsupervised Hierarchical Deep Factorization of Speech Signal, ***Z. Peng**, S. Feng, and T. Lee, in Proc. ICASSP 2020*

Adversarial Multi-task Deep Features and Unsupervised Back-end Adaptation for Language Recognition, ***Z. Peng**, S. Feng, and T. Lee, in Proc. ICASSP 2019*

Combining Adversarial Training and Disentangled Speech Representation for Robust Zero-Resource Subword Modeling, *S. Feng, T. Lee, and **Z. Peng**, in Interspeech 2019*

Child Speech Disorder Detection with Siamese Recurrent Network using Speech Attribute Features, *J. Wang, Y. Qin, **Z. Peng** and T. LEE, in Interspeech 2019*

SEMINAR TALKS

Large-scale Pairwise Classification and its Application in Speaker Verification *May 2019*

- The back-end for speaker verification is to perform similarity scoring of embeddings. Pairwise SVM is a potential alternative to the standard PLDA scoring back-end.
- Developed both Cython and C++ implementations for PSVM.

Introduction to Probabilistic Graphical Model: Variational Inference *, May 2018*

- The standard training method of GMM-ivector extractor has two individual EM training phases that may result in sub-optimal solutions. Variational inference can be adopted to jointly train both GMM and ivector extractor.
- Developed the C++ implementation for variational inference of GMM-ivector extractor.

HONORS

National Second Pride of Electronic Design (Frequency meter with Verilog) *2015*

National scholarship *2014, 2015*

Best Tutor Award *2015*