Sanyuan Chen

sanyuan-chen.github.io

(+86) 130-4511-8932 · sanyuan.chen.ai@gmail.com · sanyuan-chen

Education

Harbin Institute of Technology, Harbin, China — Aug 2019 – Jun 2024 (expected)

Ph.D. Student in Computer Science, SCIR lab

- Microsoft Research Asia Joint-PhD Program
- Supervised by Prof. Xiangzhan Yu and Dr. Ming Zhou

Harbin Institute of Technology, Harbin, China — Aug 2015 - Jun 2019

Bachelor Degree in Computer Science, Honor School

- Supervised by Prof. Wanxiang Che
- Research Intern at DASlab, Harvard University
- Exchange Student at National Chiao Tung University

Research Profile

Research Focus: Language Model Pre-Training for Speech and Audio Processing.
Research Interests: Self-Supervised Learning, Speech and Audio Pre-training, Generative Modeling.
Research Highlights:

- VALL-E and VALL-E X can synthesize multilingual speech with anyone's voice from just 3 seconds of audio, and wins the UNESCO Netexplo Innovation Award 2023 (top 10 out of over 3000 innovations).
- WavLM ranks 1st in the SUPERB leaderboard, advances the state of the art in over 50 speech processing tasks, benchmarks and challenges, including DIHARD III, CHiME 4, VoxSRC 2022, VoiceMOS 2022, and ADD 2023.
- BEATs ranks 1st in the AudioSet, Balanced AudioSet and ESC-50 audio classification leaderboard, and powers winning systems in DCASE 2023 Automated Audio Captioning Challenge and Sound Event Detection Challenge.

Selected Publications

Full list in Google Scholar (**1000+ citations**) (*joint first author)

1. **Sanyuan Chen**, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, Furu Wei. BEATs: Audio Pre-Training with Acoustic Tokenizers. **ICML 2023 oral**.
2. Chengyi Wang*, **Sanyuan Chen***, Yu Wu*, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, Furu Wei. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. **ArXiv Preprint**.
3. Ziqiang Zhang*, **Sanyuan Chen***, Long Zhou*, Yu Wu, Shuo Ren, Shujie Liu, Zhuoyuan Yao, Xun Gong, Lirong Dai, Jinyu Li, Furu Wei. SpeechLM: Enhanced Speech Pre-Training with Unpaired Textual Data. **ArXiv Preprint**.
4. **Sanyuan Chen**, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Micheal Zeng, Furu Wei. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech. **J-STSP**.
5. **Sanyuan Chen**, Yu Wu, Chengyi Wang, Shujie Liu, Zhuo Chen, Peidong Wang, Gang Liu, Jinyu Li, Jian Wu, Xiangzhan Yu, Furu Wei. Why does Self-Supervised Learning for Speech Recognition Benefit Speaker Recognition? **Interspeech 2022**.
6. **Sanyuan Chen**, Yu Wu, Chengyi Wang, Zhengyang Chen, Zhuo Chen, Shujie Liu, Jian Wu, Yao Qian, Furu Wei, Jinyu Li, Xiangzhan Yu. UniSpeech-SAT: Universal Speech Representation Learning with Speaker Aware Pre-Training. **ICASSP 2022**.
7. **Sanyuan Chen**, Yu Wu, Zhuo Chen, Jian Wu, Takuya Yoshioka, Shujie Liu, Jinyu Li, Xiangzhan Yu. Ultra Fast Speech Separation Model with Teacher Student Learning. **Interspeech 2021**.
8. **Sanyuan Chen**, Yu Wu, Zhuo Chen, Jian Wu, Jinyu Li, Takuya Yoshioka, Chengyi Wang, Shujie Liu, Ming Zhou. Continuous Speech Separation with Conformer. **ICASSP 2021**.
9. **Sanyuan Chen**, Yu Wu, Zhuo Chen, Takuya Yoshioka, Shujie Liu, Jinyu Li, Xiangzhan Yu. Don't shoot butterfly with rifles: Multi-channel Continuous Speech Separation with Early Exit Transformer. **ICASSP 2021**.
10. Yutai Hou*, **Sanyuan Chen***, Wanxiang Che, Cheng Chen, Ting Liu. C2C-GenDA: Cluster-to-Cluster Generation for Data Augmentation of Slot Filling. **AAAI 2021**.
11. **Sanyuan Chen**, Yutai Hou, Yiming Cui, Wanxiang Che, Ting Liu, Xiangzhan Yu. Recall and learn: Fine-tuning Deep Pretrained Language Models with Less Forgetting. **EMNLP 2020**

Experience

Natural Language Computing Group, Microsoft Research Asia Beijing, China

Joint-PhD Student (Advisor: Dr. Yu Wu, Dr. Shujie Liu, Dr. Zhuo Chen, Dr. Jinyu Li) May 2020 - Present

1. **Language Model Pre-Training for Full Stack Speech Processing**
   - We proposed WavLM, a large-scale self-supervised speech pre-training framework, that achieves state-of-the-art performance in over 50 speech processing tasks.
   - WavLM ranks 1st in the SUPERB and VoxSRC 2021 speaker verification permanent leaderboard.
   - WavLM advances the state of the art in DIHARD III, CHiME 4, VoxSRC 2022, VoiceMOS 2022 challenges.
   - WavLM is integrated into HuggingFace and TorchAudio.
2. **Language Model Pre-Training for Speech Synthesis**
   - We proposed VALL-E, a language modeling approach for text to speech synthesis, that achieves state-of-the-art zero-shot TTS performance with in-context learning capabilities.
   - We proposed VALL-E X, a cross-lingual version of VALL-E that can help anyone speak a foreign language in their own voice without an accent.
   - VALL-E wins the UNESCO Netexplo Innovation Award 2023 (top 10 out of over 3000 innovations).
3. **Language Model Pre-Training for Speech Recognition**
   - We proposed SpeechLM, a textual enhanced speech pre-training model, that achieves 12% relative WER reduction over data2vec with only 400K unlabeled text sentences on the LibriSpeech benchmark.
   - We proposed PBERT, a speech pre-training model with supervision-guided codebooks, that achieves 17% relative WER reduction over HuBERT on the LibriSpeech benchmark.
   - We proposed ILS-SSL, a self-supervised speech pre-training method with intermediate layer supervision, that achieves over 20% relative WER reduction over HuBERT on the LibriSpeech benchmark.
4. **Language Model Pre-Training for Speaker Verification**
   - We proposed UniSpeech-SAT, a speech representation learning method with speaker-aware pre-training, that significantly improved the previous model in speaker identification oriented tasks.
   - Unispeech-SAT outperforms the winner system in the VoxSRC 2021 Speaker Verification Challenge.
5. **Language Model Pre-Training for Audio Understanding**
   - We proposed BEATs, a discrete label prediction based audio pre-training framework, that achieves state-of-the-art performance across various audio and speech understanding tasks.
   - BEATs ranks 1st in the AudioSet, Balanced AudioSet and ESC-50 leaderboard.
   - BEATs is utilized by the winning system in DCASE 2023 Automated Audio Captioning Challenge.
   - BEATs is utilized by all the top 5 winning systems in DCASE 2023 Sound Event Detection Challenge.
6. **Speech Separation of High Accuracy and Low Latency**
   - We proposed a Conformer-based speech separation system with state-of-the-art performance, and 2x inference speed with an early exit mechanism and teacher-student learning method.
   - Our speech separation system ranks 1st in the VoxCeleb Speaker Recognition Challenge 2020.
   - Our speech separation system is shipped in the Microsoft Conversation Transcription Service.

Research Center for Social Computing and Information Retrieval, HIT Harbin, China

Research Student (Advisor: Prof. Wanxiang Che) Jul 2019 - May 2020

1. **Efficient Fine-tuning of Pre-Trained Language Models**
   - We proposed RecAdam, an open-source optimizer for large language model fine-tuning, that achieved state-of-the-art results on the GLUE benchmark, including 10 natural language understanding tasks.
   - RecAdam enables BERT-base to achieve better performance than directly fine-tuning of BERT-large.

Data Systems Laboratory, Harvard University Cambridge, US

Research Student (Advisor: Prof. Stratos Idreos, Prof. Alexander Rush) Jul 2018 - May 2019

1. **Rapid Deep Ensemble Learning**
   - We proposed MotherNets, a fast training approach for large neural network ensembles, that provides a new Pareto frontier for the accuracy-cost tradeoff compared to the state-of-the-art approaches.
2. **Named Tensors for PyTorch**
   - We build the Named Tensors project, which allows users to give explicit names to PyTorch tensor dimensions for easier dimension rearrangement and extra safety. It gains over 400 stars on GitHub and is merged into PyTorch.