Chang Zeng (曾 暢 ソウ チョウ)

Generative AI/Deep Learning

Tokyo, Japan | (81)07085400924 | zengchang.elec@gmail.com

| github.com/zengchang233 | scholar.google.com/citations?user=gfGyn49j-MkC&hl=en

Chang Zeng is a Ph.D. candidate with 6 years of speech signal processing/sequence-to-sequence (S2S)/deep learning experiences. He has explored Singing Voice Generation and Text-to-Speech (TTS) in universities and Speech Recognition in the industry. He has published more than a dozen articles at international conferences, including Interspeech and ICASSP. Besides, he is also a contributor to open-source tools such as ASV-Subtools and WeSpeaker. Now he is interested in **Deep Generative AI for language and audio processing**. He would like to dedicate himself to the Deep Generative Modeling field for his long-term future work.

Education

National Institute of Informatics (SOKENDAI) **2020.10 - 2024.03.22 (expected)**

Doctor of Informatics

Sponsored by the Japanese government MEXT scholarship, supervised by Prof. Yamagishi.

The University of Tokyo **2017.10 - 2020.03**

Master of Electrical Engineer and Information System

Supervised by Prof. Minematsu.

Tianjin University **2012.09 - 2016.07**

Bachelor of Engineering

Publications

Featured Publications

1. Wang, C., **Zeng, C (co-first author)**., & He, X. (2023). Xiaoicesing 2: A High-Fidelity Singing Voice Synthesizer Based on Generative Adversarial Network. In *Proc. Interspeech 2023, 5401-5405.* Project page. *(CoreRank A)*
2. Wang X., **Zeng, C (co-first author)**., Chen, J., Wang, C. (2023). CrossSinger: A Cross-Lingual Multi-Singer High-Fidelity Singing Voice Synthesizer Trained on Monolingual Singers. *Accepted by ASRU 2023.* Project page.
3. Zhu, W., Wang, Z., Lin, J., **Zeng, C.**, & Yu, T. (2023) SSI-Net: A MULTI-STAGE SPEECH SIGNAL IMPROVEMENT SYSTEM FOR ICASSP 2023 SSI CHALLENGE. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-2) *(CoreRank B)*
4. **Zeng, C**., Wang, X., Cooper, E., Miao, X., & Yamagishi, J. (2022). Attention back-end for automatic speaker verification with multiple enrollment utterances. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6717-6721). Project page. *(CoreRank B)*
5. **Zeng, C**., Zhang, L., Liu, M., & Yamagishi, J. (2022). Spoofing-Aware Attention based ASV Back-end with Multiple Enrollment Utterances and a Sampling Strategy for the SASV Challenge 2022. In *Proc. Interspeech 2022, 2883–2887. (CoreRank A)*
6. **Zeng, C**., Wang, X., Miao, X., Cooper, E., & Yamagishi, J. (2023). Improving Generalization Ability of Countermeasures for New Mismatch Scenario by Combining Multiple Advanced Regularization Terms. In *Proc. Interspeech 2023, 1998-2002. (CoreRank A)*
7. **Zeng, C**., Miao, X., Wang, X., Cooper, E., & Yamagishi, J. (2022). Joint Speaker Encoder and Neural Back-end Model for Fully End-to-End Automatic Speaker Verification with Multiple Enrollment Utterances. *Computer Speech & Language. (JCR Q1)*
8. Tang, H., Liu, Z., **Zeng, C.**, & Li, X. (2023). Beyond Universal Transformer: block reusing with adaptor in Transformer for automatic speech recognition. *Accepted by ISNN 2024.*
9. Wang, C., **Zeng, C (co-first author)**., Chen, J., & He, X. (2023). HiFi-WaveGAN: Generative Adversarial Network with Auxiliary Spectrogram-Phase Loss for High-Fidelity Singing Voice Generation. Project page. *Accepted by ISNN 2024.*

Under Review

1. **Zeng, C**., Wang, C., & Miao, X. (2024). InstructSing: High-fidelity Singing Voice Generation via Instructing Yourself. *Submitted to ICME 2024.* Demo page. *(CoreRank A)*
2. **Zeng, C.**, & Wang, C. (2024). HAM-TTS: Hierarchical Acoustic Modeling for Token-Based Zero-Shot Text-to-Speech with Model and Data Scaling. *Submitted to ACL 2024.* Demo Page.
3. **Zeng, C.**, & Wang, C. (2024). Few-shot Singing Voice Conversion Based on Latent Diffusion Model. *Submitted to Interspeech 2024.*

Activities

Competitions

- 4th/77 place for VoxCeleb Speaker Recognition Challenge 2019.
- 2nd/110 place for Zhijiang Cup Speech Recognition for Conversational Scenario 2021.
- 4th/42 place for Audio Deep Synthesis Detection Challenge 2022 track1 (Low-quality Fake Audio Detection, LF).
- 5th/27 place for Audio Deep Synthesis Detection Challenge 2022 track2 (Partially Fake Audio Detection, PF).

Academic Activities

- Research assistant at Yamagishi Lab, National Institute of Informatics
- ICASSP 2022 Oral Presentation.
- Interspeech 2022 Oral Presentation.
- Interspeech 2023 Poster Presentation.
- Reviewer of IEEE Open Journal of Signal Processing.
- Organization committee of Joint Workshop of VoicePersonae and ASVspoof 2023.
- Reviewer of ICASSP, ICME, and Interspeech.

Open Source

- WeSpeaker contributor (450 stars)
- ASV-Subtools contributor (561 stars)
- Attention backend

Work Experiences

RevComm **2023.09 - Present**

Intern Speech Researcher

Responsibilities

- Building a contextual speech recognition system for domain adaptation.

Achievements

- Reproduced a tree-constraint pointer generator for the Transformer-based contextual ASR model according to this paper. A GNN-based biasing word predictor was trained to generate an indicative probability to fuse the ASR output probability and biasing word probability. Moreover, a plugin version has been crafted to eliminate the necessity for retraining the Transformer-based ASR model.
- Rescoring the decoding results of the Transformer-based ASR model by the pretrained GPT2 language model. Modeling spaces of the ASR model and GPT2 were aligned by Internal Language Model Estimation.
- The performance on test-clean and test-other datasets of Librispeech was improved by **10%-15%** relatively.
- Further analysis shows the ratio of retrieving biasing words was improved by around **6%** absolutely.

Xiaoice **2022.07 - 2023.07**

Intern Avatar Researcher

Responsibilities

- I am focusing on building a generative model for high-fidelity (48kHz) singing voice generation tasks collaborating with other researchers and engineers.

Achievements

- Improved the Fastspeech2-based XiaoiceSing to **XiaoiceSing2** by utilizing **adversarial training** to generate a mel-spectrogram with high accuracy. The MOS score shows that the quality of synthesized singing voices is quite close to the human level (4.23 vs 4.27). The paper has been accepted by Interspeech 2023;
- Developed a GAN-based vocoder model called **HiFi-WaveGAN** to reconstruct the waveform from the mel-spectrogram by incorporating a novel **Pulse Sequence**. The proposed vocoder beats the widely-used HiFi-GAN model in terms of MOS metric for 48kHz singing voice synthesis task (4.23 vs 3.89);
- Improved XiaoiceSing2 to a cross-lingual (en, jp, cn) multi-singer singing voice synthesizer **CrossSinger** by incorporating language information into the encoder and leveraging a novel **bias eliminator** to remove singer information contained in the encoder. The paper has been accepted by ASRU 2023;
- Improved HiFi-WaveGAN to **InstructSing** by fusing differential digital signal processing (DDSP) and adversarial training. The imperative information generated by the DDSP module is refined by a UNet-based module to instruct and accelerate the speed of the adversarial training. Experiments show that the proposed model can converge within 20,000 training steps, which is **one-tenth** of other neural vocoders.
- Hierarchical modeling for token-based large voice language model. Hierarchical modeling acoustic context information by inserting the HuBert Unit as the intermediate representation for the text module of VALLE. The pronunciation accuracy of synthesized speech is significantly improved.

ALIBABA **2020.04 - 2020.09**

Speech Recognition Researcher (Full-time)

Responsibilities

- In Alibaba, I belonged to a department that is responsible for TAOBAO living. To prevent live broadcasters from violating laws and regulations, I developed two systems with my colleagues.

Achievements

- Developed a large-scale speaker recognition system that aims at identifying the broadcaster in each living room as the exact person;
- Developed a spoken word detection system aims at catching some illegal words spoken by broadcasters;
- Explored SSL for speech representation and developed an end-to-end speech recognition system with knowledge distillation based on ESPNet.

Skills

Tools and Languages	Python, CPP, Shell, Git, MySQL
Toolkits	PyTorch, Pytorch-Lightning, SpeechBrain, WeNet, WeSpeaker, Kaldi, Espnet
Communication	Chinese (native), English (business level), Japanese (N2-125)