Rui-Chen Zheng | 郑瑞晨

Phone: (+86)181-8926-6050 | **Email:** zhengruichen@mail.ustc.edu.cn
Homepage: https://zhengrachel.github.io/

Education

M.S. Student in Information and Communication Engineering *2021.09 - Now*
University of Science and Technology of China *Hefei, Anhui, China*
- Supervised by Prof. Zhen-Hua Ling
- GPA: 3.9/4.3

Bachelor of Electronic Information Engineering *2017.09 – 2021.06*
University of Science and Technology of China *Hefei, Anhui, China*
- Thesis: Method and Practice on Text-to-Speech Without Text
- GPA: 3.89/4.3
- Minor in Business Administration

Research Experience

Speech Reconstruction from Silent Lip and Tongue Articulation by Pseudo Target Generation and Domain Adversarial Training
- Supervised by Prof. Zhen-Hua Ling
- This paper studies the task of speech reconstruction from ultrasound tongue images and optical lip videos recorded in a silent speaking mode, where people only activate their intra-oral and extra-oral articulators without producing sound. We propose to employ a method built on pseudo target generation and domain adversarial training with an iterative training strategy to improve the intelligibility and naturalness of the speech recovered from silent tongue and lip articulation. To be specific, pseudo targets were first generated for silent articulations in order to enable the same supervised training paradigm as when vocalized articulations were used as input. Besides, by adding a domain discriminator behind the encoder, the model is able to learn common features existed in silent and vocalized articulation. Iterative training strategy was conducted to obtain better pseudo targets. Experiments show that our proposed method significantly improves the intelligibility and naturalness of the reconstructed speech in both silent and vocalized speaking mode compared to the baseline.
- Demo-Page: https://zhengrachel.github.io/ImprovedTaLNet-demo/
- Published in ICASSP 2023. Click https://ieeexplore.ieee.org/document/10096920 to view the full paper.

Incorporating Ultrasound Tongue Images for Audio-Visual Speech Enhancement Through Knowledge Distillation
- Supervised by Prof. Zhen-Hua Ling
- Audio-visual speech enhancement (AV-SE) aims to enhance degraded speech along with extra visual information such as lip videos, and has been shown to be more effective than audio-only speech enhancement. This paper proposes further incorporating ultrasound tongue images to improve lip-based AV-SE systems' performance. Knowledge distillation is employed at the training stage to address the challenge of acquiring ultrasound tongue images during inference, enabling an audio-lip speech enhancement student model to learn from a pre-trained audio-lip-tongue speech enhancement teacher model. Experimental results demonstrate significant improvements in the quality and intelligibility of the speech enhanced by the proposed method compared to the traditional audio-lip speech enhancement baselines. Further analysis using phone error rates (PER) of automatic speech recognition (ASR) shows that palatal and velar consonants benefit most from the introduction of ultrasound tongue images.
- Demo-Page: https://zhengrachel.github.io/UTIforAVSE-demo/
- Published in INTERSPEECH 2023. Click https://arxiv.org/abs/2305.14933 to view the full paper.

Incorporating Ultrasound Tongue Images for Audio-Visual Speech Enhancement
- Supervised by Prof. Zhen-Hua Ling
- Compared to the conference paper, this paper adds a new method, namely memory-based audio-lip speech enhancement method, to address the challenge of acquiring ultrasound tongue images during inference. Specifically, this method further proposes the introduction of a lip-tongue key-value memory network into the AV-SE model to better model the alignment between the lip and tongue modalities. The lip-tongue key-value memory network enables the retrieval of tongue features based on readily available lip features, thereby assisting the subsequent speech enhancement task. Experimental results demonstrate that this method further narrows the performance gap between audio-lip and audio-lip-tongue speech enhancement model and exhibits strong generalization performance on unseen speakers and in the presence of unseen noises. Furthermore, phone error rate (PER) analysis of automatic speech recognition (ASR) reveals that while all phonemes benefit from introducing ultrasound tongue images, palatal and velar consonants benefit most.
- Demo-Page: https://zhengrachel.github.io/IUTIforAVSE-demo/
- Published on IEEE/ACM Transactions on Audio, Speech, and Language Processing 2024. Click https://ieeexplore.ieee.org/document/10418525 to view the full paper.

Selected Honors

- Honor Rank for Top 5% Graduates of USTC. *2021.06*
- Huawei Scholarship *2020.12*
- USTC Outstanding Student Scholarship, Gold Award *2019.12 & 2018.12*
- Top-Notch Program Funding *2019.12 & 2018.12*

Skills

Language – Chinese/English

- TOEFL iBT: 106 *2023.11*
- CET-6 607 *2018.12*

Programming Language

- Python/Matlab/C

Teaching Assistant Experience

- Fundamentals of Speech Signal Processing by Prof. Zhen-Hua Ling, USTC *2022 Fall*
- Fundamentals of Speech Signal Processing by Prof. Zhen-Hua Ling, USTC *2021 Fall*
- Computer Programing Design A by Lecturer. Hu Si, USTC *2020 Fall*

Other Experience

Admitted to UCLA CSST Summer Internship Program *2020.07*