Technical Report

Tong Lei[1,2], Zhongshu Hou[1,2], Yuxiang Hu[2], Wanyu Yang[2], Tianchi Sun[1,2], Xiaobin Rong[1,2], Dahan Wang[1,2], and Jing Lu[1,2]

[1]Key Laboratory of Modern Acoustics Institute of Acoustics, Nanjing University, Nanjing 210093, China

[2] NJU-Horizon Intelligent Audio Lab, Horizon Robotics, Nanjing 210038, China.

1. Introduction

This is the technical report for our submission to the second Clarity Enhancement Challenge (CEC2). We submitted six sets of results from different processing strategies, with ID E016, E021, E022, E035, E037, and E039 respectively.

2. Our proposed system

The diagram of our proposed system is illustrated in Figure 1, which mainly includes an online speech dereverberation pre-processing, a multi-channel target speech enhancement model for each ear, a post-processing model, and several linear processing units. The online speech dereverberation module employs the state-of-the-art (SOTA) rule-based method called online weighted prediction error (online WPE) [1]. For the multi-channel enhancement model, we use the causal EaBNet [2], followed by post-processing based on the modified multi-scale temporal frequency convolutional network with axial self-attention net (MTFAA) [3] or based on the Taylor superimposition [4].

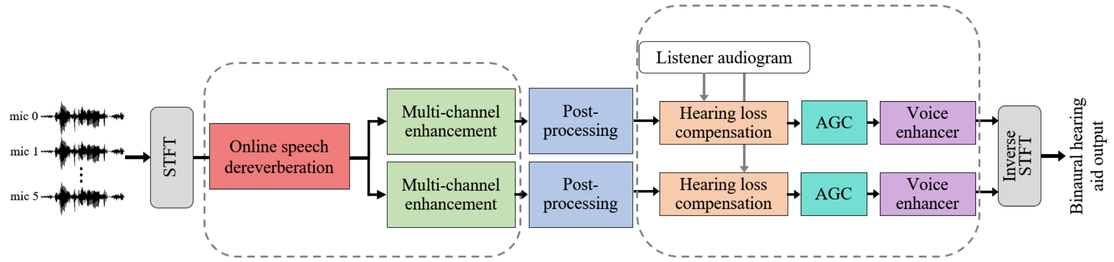
mic 0
mic 1
mic 5
STFT
Online speech dereverberation
Multi-channel enhancement
Multi-channel enhancement
Post-processing
Post-processing
Listener audiogram
Hearing loss compensation
Hearing loss compensation
AGC
AGC
Voice enhancer
Voice enhancer
Inverse STFT
Binaural hearing aid output

Figure 1: *Block diagram of our proposed system.*

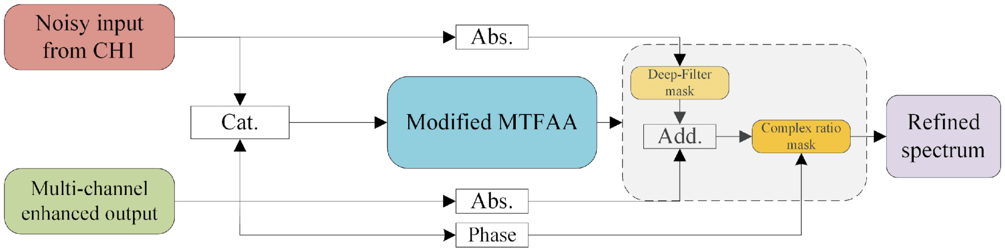

Figure 2: *Post-processing based on modified MTFAA.*

The post-processing module based on the modified MTFAA is illustrated in Figure 2, where we modify MTFAA by replacing the Main-Net with two dual-path recurrent neural networks (DPRNNs) [5]. To refine the speech outputted by the multi-channel enhancement model, the spectrogram of the noisy signal, concatenated with the prior enhanced speech of the left/right ear, is sent to the modified MTFAA. A Deep-Filter mask [6] is applied to the magnitude of the noisy spectrum, whose output is added to the magnitude of the multi-channel enhancement output. Then the compensated magnitude with the phase spectrum of the multi-channel enhancement output is further processed by a complex ratio mask.

The linear processing includes automatic gain control (AGC), the voice enhancer, and the hearing loss compensation processing, whose gain is calculated by the revised Prescription of Gain/Output (POGO Ⅱ) [7] using the binaural audiograms.

The six systems are listed below. All of them are evaluated using objective metrics. We prefer E022, E039, and E035 to be evaluated using subjective listening experiments with hearing-impaired listeners respectively.

- E022: EaBNet_drb + post process (modified MTFAA) + HLC with AGC
- E037: EaBNet_drb + post process (modified MTFAA) + HLC with AGC and voice enhancer
- E039: EaBNet_drb + post process (TaylorBeamformer) + HLC with AGC
- E016: EaBNet_drb + post process (TaylorBeamformer) + HLC with AGC and voice enhancer
- E035: EaBNet_drb + HLC with AGC
- E021: EaBNet_drb + HLC with AGC and voice enhancer

3. Key features

All the modules are implemented and cascaded in the short-time Fourier transform (STFT) domain, with a 220-sample Hanning window and a 110-sample hop-size at the sampling rate of 44.1 kHz. The total time latency is 220/44100 ms, less than the requirement (5 ms) of this challenge.

We trained the neural networks in our system only on the core database and did not use the head rotation data.

4. Experimental result

The average HASPI scores of our six systems on the 2500 samples of the development set are presented in Table 1.

Table 1: *HASPI scores of our six systems.*

SYSTEM ID	E021	E035	E016	E039	E037	E022
HASPI	0.696	0.675	0.721	0.701	0.734	0.712

5. Reference

[1] R. Ikeshita, K. Kinoshita, N. Kamo, and T. Nakatani, "Online speech dereverberation using mixture of multichannel linear prediction models," *IEEE Signal Processing Letters*, vol. 28, pp. 1580-1584, 2021.

[2] A. Li, W. Liu, C. Zheng, and X. Li, "Embedding and beamforming: All-neural causal beamformer for multichannel speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 6487-6491, 2022.

[3] G. Zhang, L. Yu, C. Wang, and J. Wei, "Multi-scale temporal frequency convolutional network with axial attention for speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 9122-9126, 2022.

[4] A. Li, S. You, G. Yu, C. Zheng, and X. Li, "Taylor, Can You Hear Me Now? A Taylor-Unfolding Framework for Monaural Speech Enhancement," *arXiv preprint arXiv:2205.00206*, 2022.

[5] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 46-50, 2020.

[6] W. Mack and E. A. Habets, "Deep filtering: Signal extraction and reconstruction using complex time-frequency filters," *IEEE Signal Processing Letters*, vol. 27, pp. 61-65, 2019.

[7] D. Schwartz, P. Lyregaard, and P. Lundh, "Hearing aid selection for severe-to-profound hearing loss," *Hearing Journal*, vol. 41, no. 2, pp. 13-17, 1988.