CEC2 E008 Technical Paper

Chengwei Ouyang, Kexin Fei, Haoshuai Zhou, Linkai Li, Congxi Lu

Orka Inc.

Abstract

This paper proposes an end-to-end system for the second Clarity Enhancement Challenge (CEC2). Our system consists of three main parts: a denoising module, a beamforming module for speech enhancement, and an amplification module for hearing loss compensation. The objective evaluation results (HASPI) show that our system significantly outperforms the baseline system on the development set, with an average HASPI score of **0.7673**.

Index Terms: speech enhancement, beamforming, hearing aids, multi-stage

1. Introduction

The 2nd Clarity Enhancement Challenge [1] aims to improve the performance of hearing aids for speech-in-noise. Compared with other speech enhancement challenges, CEC2 requires both denoising and hearing aid enhancement. This paper describes the Orka system, which consists of a denoising module, a beamforming module and an amplification module. The denoising module targets single-channel speech enhancement, while the beamforming module focuses on multi-channel speech enhancement, and the amplification module will amplify the signals based on the listener's hearing audiogram and improve the intelligibility.

For a long time, the speech enhancement algorithms have been performed in the time-frequency (T-F) domain, which relies on *Short Time Fourier Transform* (STFT) to transform a signal from time-domain to time-frequency domain. However, the frequency resolution is constrained by the chosen window size by STFT, which will introduce equal algorithm latency in real time tasks. We propose Orka window in this wok, an asymmetric forward and backward window pair, which can achieve a high frequency resolution while still maintaining a low reconstruction latency.

Recently, the importance of phase attracts multiple researches, and a multitude of phase-aware neural networks are proposed and achieve state-of-the-art performance in speech enhancement [3, 7]. In this work, we utilize the phase information in the beamforming module only. As beamforming relies on the delay of recordings of each channel, corresponding to the phase information in T-F domain, the denoising module processes the magnitude information only while leaving the phase unchanged.

2. Method

The overall architecture of the proposed system is shown in Fig.1. It consists of three parts, namely denoising module, beamforming module and amplification module. Both denoising module and amplification module are operated in the magnitude domain, while the beamforming module is operated in the complex domain. The 6 channels mixed-signals (consists of CH1-CH3 for both left and right ear) and head rotation information is used in this system, and *Short Time Fourier Transform* (STFT) [2] is performed to get the complex spectrum **S(x)** and **S(r)**, consisting of magnitude and phase information. An spectrum feature encoder and decoder are implemented to better refine the spectrogram of different frequency bands for 32kHz sample rate. The denoising module takes the magnitude information of the 6-channel mixed-signals as inputs only while leaving the phase unchanged, and it processes the 6 channels separately, without using the inter-channel information. After the denoising module, the processed magnitude and the original phase of mixed-signals are coupled to re-generate the coarse complex spectrum. For the beamforming module, it takes the coarse complex spectrum **D(x)** and the head rotation information after STFT **S(r)** as inputs, and the inter-channel information is also utilized to construct the 1 channel denoised spectrum **B(x)**. The amplification module takes the magnitude information of the denoised spectrum and the hearing loss audiogram as inputs, and the inverse-STFT (iSTFT) is used to transform the output TF-domain spectrogram to a time-domain waveform, which is the final enhanced signal **y**.

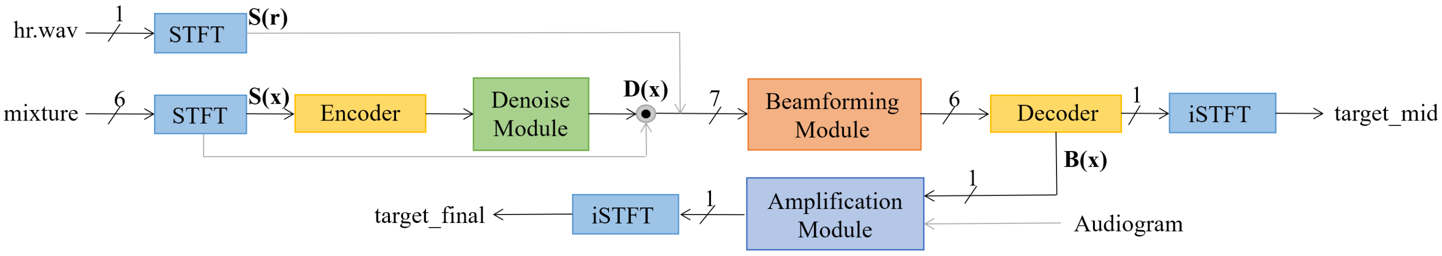

Figure1: *Overall architecture of the end-to-end model.*

2.1. Orka Window

To satisfy the real-time speech enhancement requirement, we propose an asymmetric forward and backward window pair - Orka window. Orka window is composed of a forward window $w_1[n]$ and a backward window $w_2[n]$, whose definition and shape are shown in Eq.1 and Fig.2(a). In most real-time speech processes, the reconstruction latency is half of window size, so the window size has to be kept short in order to achieve low latency. But short window size will restrain frequency resolution and thus the model performance. To solve this problem, we designed a forward and backward window pair to simultaneously achieve a high frequency resolution while still maintaining a low reconstruction latency. The overall latency is independent of window size and can be two times of hop size, which is $2 * R$. Besides, this window pair satisfies the Constant Overlap-Add (COLA) property and thus can achieve perfect reconstruction of original signal. In our experiment, we set $N_1 = 64$, $N_2 = 448$ and $R = 64$ under 32kHz sampling rate, which corresponds to 2ms hop size with 16ms window size. The overall latency of 4ms meets the 5ms latency requirement.

$$w_1[n] = \begin{cases} sin^2(\frac{n\pi}{2N_1}), & if\ 0 \le n < N_1 \\ 1, & if\ N_1 \le n \le N_2 \\ sin(\frac{\pi(N_2+R-n)}{4R}), & if\ N_2 < n \le\ N_2 + R \end{cases}$$

$$w_2[n] = \begin{cases} 0, & if\ 0 \le n < N_2 - R \\ cos^2(\frac{\pi(n-N_2)}{2R}), & if\ N_2 - R \le n \le N_2 \\ sin(\frac{\pi(N_2+R-n)}{2R}), & if\ N_2 < n \le N_2 + R \end{cases}$$

$$where\ R > 0\ is\ the\ hop\ size,\ and\ 0 < N_1 < N_2 - R.$$

Equation 1

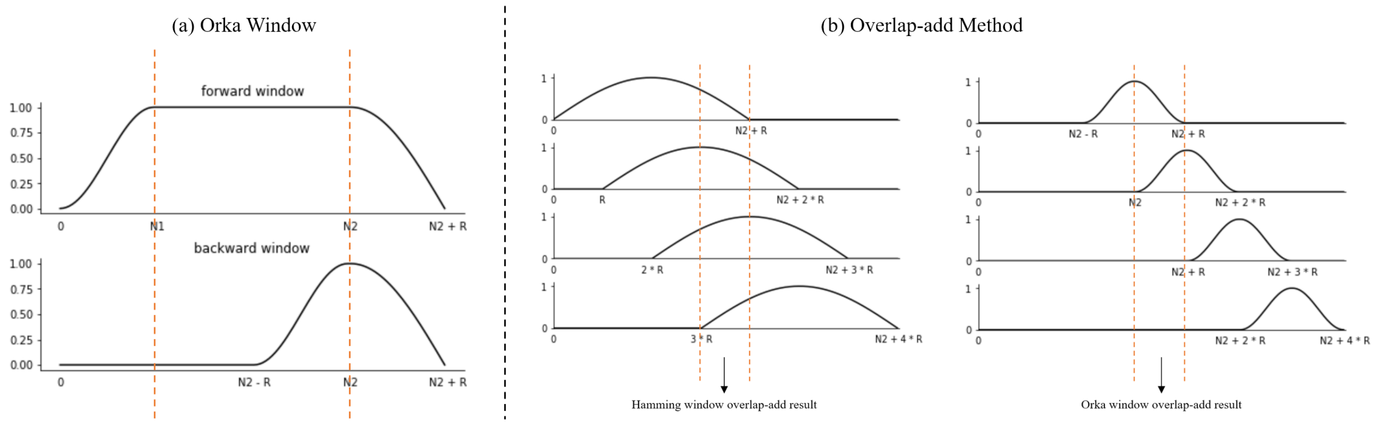

Figure 2: *Orka Window.*

2.2. Spectrum Feature Encoder/Decoder

In 32kHz speech enhancement situation, the number of points of STFT is doubled compared with that of 16kHz. Several researches are focused on converting physical frequency to psychoacoustic frequency based on human perception and avoiding musical noise in the output, like Bark spectrum in RNNoise [9], ERB band [11] in PercepNet [10]. Especially for super wide band (32kHz), the number of points of STFT is doubled compared with that of 16kHz, which will double the network complexity but with minor improvement on performance. However, the phase information will be discarded through this spectrum compression, and the features based on the human perception may not suitable for the neural network. It is pointed out that the lower frequency band contains higher energies and tonalities, while the higher frequency band tends to have lower energy components. Inspired by power compression in [2], we propose refining the energies for each band by multiplying with learnable parameters, and a conv2d layer with a kernel size of 1 is used to refine the inter-channel information.

Inspired by the encoder/decoder block of S-DCCRN [2], we propose an spectrum feature encoder and decoder to better refine the spectrogram of different frequency bands and channels on 256-dimensional STFT features. As shown in Fig. 3, a conv2d with a kernel size of 1 is applied to extract high-dimensional information, while a dilated dense block [12] is employed to capture long-term contextual information from time scale and a conv2d is adopted to decrease the dimensions and extract local features. LayerNorm and PReLU activation are used after each convolution for better performance.

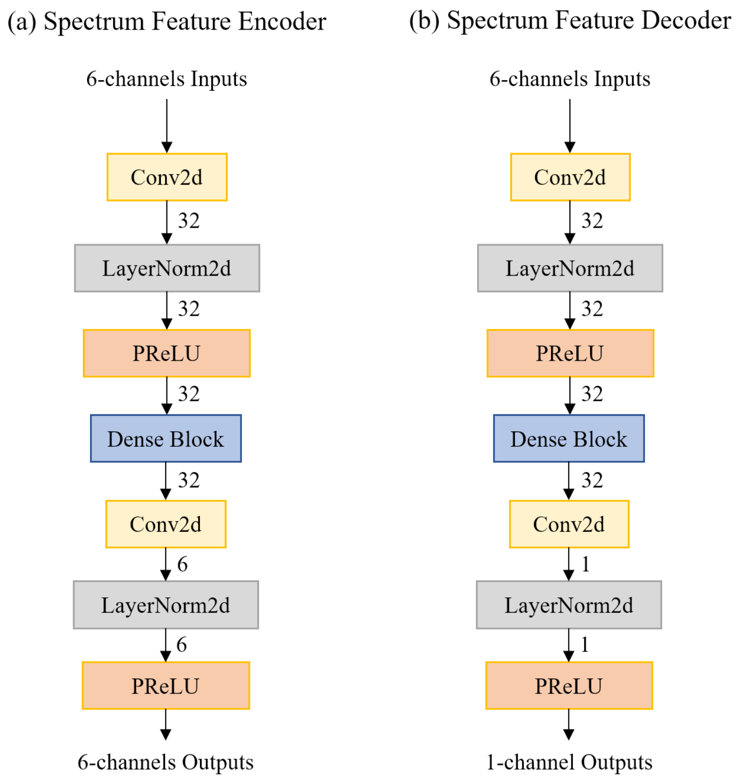

Figure 3: Spectrum Feature Encoder/Decoder

2.3. Denoising Module

Considering the beamforming process is sensitive to phase, we decouple the inputs of encoder into magnitude and phase and transfer the channel dimension into batches to reduce interference between phases, thus the denoising module can be regarded as a single-channel enhancement network. The convolutional encoder-decoder topology has achieved promising performance in speech enhancement applications [3, 7], thus we employed this architecture in all three stages. Only the magnitude information is fed into the denoising module to implement noise removal, leaving the phase information unchanged. The DN-Net in SDD-Net [3] is chosen as our backbone network of denoising module, where the channel size of each encoder and decoder layer is set to 64, and number of TCM layers and dilation rate are set to 5 and [1, 2 ,4 ,8, 16, 32] respectively. The overall architecture of the denoising module is illustrated in Fig. 4.

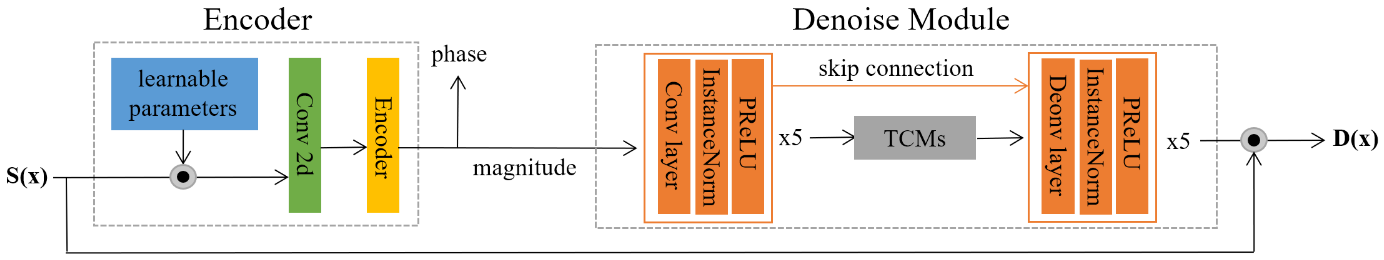

Figure 4: *Denoise Module.*

2.4. Beamforming Module

Motivated by the Minimum-Variance-Distortionless-Response (MVDR) [6], we designed a multi-channel Beamforming model to fully use the advantages of six microphones and head rotation information to locate the position of target speaker. As the sample rate is set to 32kHz in this work, it is hard to model different frequency bands by the same kernel for convolution and transposed convolution. Inspired by the sub-band and full-band processing (SAF) module in S-DCCRN [2], which achieves state-of-the-art performance in super wide band speech enhancement, we extend it to a multi-channel speech enhancement model as our beamforming module. Both sub-band and full-band module's architectures are similar to DCCRN [7], where the key point is to better utilize the real and image part of the spectrum for better speech quality and fine denoising results. The sub-band module takes a complex group convolution with group size 2 to model low-frequency points and high-frequency points separately, while the full-band module receives both the outputs of sub-band module and the original spectrogram as inputs and refines the certain unsmooth connections at the boundary of low frequency and high frequency which is introduced by sub-band module. After the sub-band and full-band processing, the six-channel outputs will be fed into the spectrum feature decoder to get the final one-channel enhanced spectrogram **B(x)**.

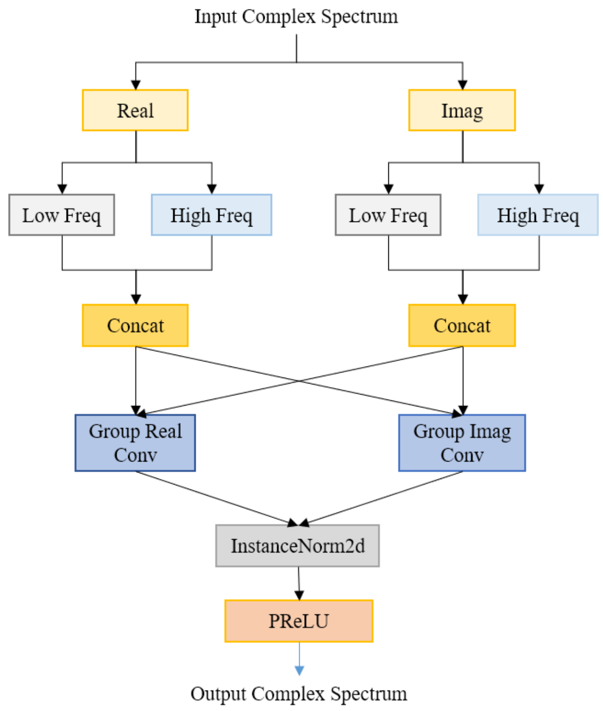

Figure 5: Sub-band Module

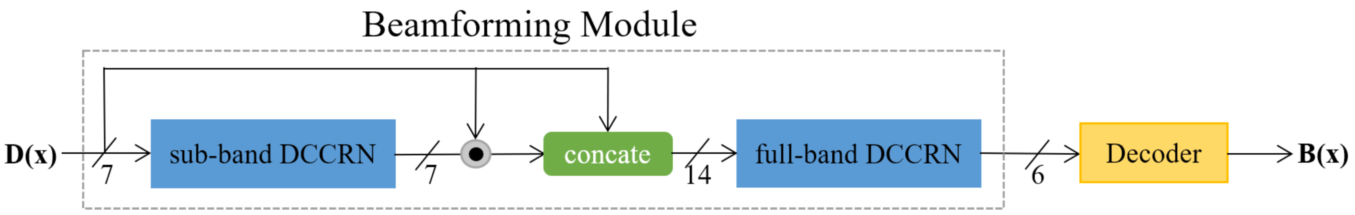

Figure 6: *Beamforming Module.*

2.5. Amplification Module

In the objective evaluation stage, the DN-Net with the same config as denoising module is used as Amplification Module to compensate the hearing loss by different audiograms and maximize the intelligibility for hearing impaired listeners. It takes the outputs of Beamforming Module and listeners' audiogram as inputs, while a fully connected layer is used to combine the spectrum features and audiogram. A self-implemented differentiable HASPI loss function is used to train the Amplification Module, which significantly improves the model's performance on objective HASPI score.

In the subjective evaluation stage, a wide dynamic range compression (WDRC) algorithm is implemented to compensate hearing loss for different hearing impaired listeners. It will only compensate for the frequency band below 8kHz to get a higher compensation scope. To further enhance the speech quality, we slightly mix 2% noisy signal to the enhanced signal, thus the inputs of the WDRC algorithm will be the sum of 98% enhanced signal and 2% noisy signal.

3. Experiments

3.1. Database and Training Setups

Our model is trained on dataset generated by the provider render files [4], including 6000*20 simulated scenes in total, using no other external data, pre-existing tools, software and models. For each scene in the training set, we kept the room setting and the corresponding target speaker unchanged, randomly combining the interference and SNR settings to generate 20 sets of data, in which way our model could be resistant to the given noise. The silence frames are removed by the dataset labels to make the model focus on denoising, and our experiments indicate this brings an increase of **0.0243** in HASPI score. All audiogram data provided by CEC2 is used to train the Amplification Module.

The PyTorch framework [5] is used to implement the system with NVIDIA 3090. All signals are resampled to 32kHz for training. The training process can be divided into two stages. In the first stage, the Denoising Module and Beamforming Module are trained using channel 1 of target signals as ground truth and SNR Loss as loss function. In the second stage, the model of the first stage is loaded as a pre-trained model and Amplification Module is added to train the final model, while the target anechoic signals are used as ground truth and the differentiable HASPI loss is used as loss function. We employ the convolutional encoder-decoder topology in all three stages for its promising performance in speech enhancement applications [3, 7]. The network architectures are described as follows.

Denoising Module: the number of channels of denoising module is {64, 64, 64, 64, 64}, and the convolution kernel size is set to (5, 2) for the first layer and (3, 2) for the others. Stride is set to (2, 1). InstanceNorm and PReLU are processed after each convolution and skip connection is utilized to mitigate the information loss during the mapping process. Between the encoder and decoder, the temporal convolutional modules (TCMs) [13] are utilized to capture the long-term temporal dependencies, while the number of TCM layers and dilation rate are set to 5 and [1, 2, 4, 8, 16, 32] respectively.

Beamforming Module: the number of channels for the sub-band module and full-band module are {32, 64, 64, 64, 128, 128} and {64, 64, 64, 64, 128, 128} respectively. The convolutional kernel size and stride are set to (5, 2) and (2, 1) respectively. One LSTM layer with 256 hidden units is inserted between encoder and decoder, and InstanceNorm and PReLU are proposed after each convolution and skip connection is utilized to mitigate the information loss during the mapping process.

Amplification Module: the network architecture of amplification module is the same as denoising module, but a fully connected layer is adopted to summarize the frequency points and audiogram information before fed into the amplification module.

3.2. Results and Discussions

The final model has 7M training parameters, requiring approximately 86G FLOPs per second. The CEC2 provides a baseline system achieving the average HASPI score of 0.2493 on the development set, we submit 4 audio compression packages processed by our system with different setups. As is shown in Table1, the average HASPI scores using generated data in Sec 3.1 and head rotation signal are increased by **0.0243** and **0.0005** separately. The proposed system achieves an average HASPI score of **0.7673**.

Table1: *Ablations for HASPI on development set*

Setups	NAL-R baseline	Raw system	+Generated data	+hr.wav	**Proposed system**
HASPI	0.2493	0.7421	0.7664	0.7426	**0.7673**

4. Conclusion

In this work, we proposed a pure deep learning real time low latency hearing aid system, consisting of a Denoise Module, a Beamforming Module and an Amplification Module. With the 5ms latency constraint, our proposed model outperforms other state-of-the-art speech enhancement models like DCCRN [7], SDD-Net [3]. Orka window also plays an important role in this experiment as it simultaneously achieves high frequency resolution and low reconstruction latency. In the future, we will also enable the proposed system in more complicated environments and decrease the computational complexity.

Reference

[1] S. Graetzer, M. Akeroyd, J. P. Barker, T. J. Cox, J. F. Culling, G. Naylor, E. Porter, and R. V. Munoz, “Clarity: Machine learning challenges to revolutionise hearing device processing,” *Interspeech*, 2021.

[2] S. Lv, Y. Fu, M. Xing, et al., “S-DCCRN: Super wide band dccrn with learnable complex feature for speech enhancement,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7767-7771.

[3] A. Li , W. Liu, X. Luo, et al., “A Simultaneous Denoising and Dereverberation Framework with Target Decoupling,” in *Interspeech*, 2021.

[4] https://github.com/claritychallenge/clarity

[5] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga et al., “Pytorch: An imperative style, high-performance deep learning library,” *Advances in Neural Information Processing Systems*, vol. 32, pp. 8026–8037, 2019.

[6] J. Benesty, J. Chen and Y. Huang, “A generalized MVDR spectrum,” *IEEE Signal Processing Letters*, 12(12), pp.827-830.

[7] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang and L. Xie, “DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement,” arXiv preprint arXiv:2008.00264

[8] Griffin, D. and Lim, J., 1984. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on acoustics*, speech, and signal processing, 32(2), pp.236-243.

[9] Valin, J.M., 2018, August. A hybrid DSP/deep learning approach to real-time full-band speech enhancement. In *2018 IEEE 20th international workshop on multimedia signal processing (MMSP)* (pp. 1-5). IEEE.

[10] Valin, J.M., Isik, U., Phansalkar, N., Giri, R., Helwani, K. and Krishnaswamy, A., 2020. A perceptually-motivated approach for low-complexity, real-time enhancement of fullband speech. *arXiv preprint arXiv:2008.04259*.

[11] B.C.J. Moore. An introduction to the psychology of hearing. Brill, 2012.

[12] Ashutosh Pandey and DeLiang Wang, “Densely connected neural network with dilated convolutions for real-time speech enhancement in the time domain,” in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 6629–6633.

[13] Y. Luo and N. Mesgarani, “Conv-tasnet: Surpassing ideal time– frequency magnitude masking for speech separation,” IEEE/ACM Trans. Audio. Speech, Lang. Process., vol. 27, no. 8, pp. 1256– 1266, 2019.