DRC-NET for The 2nd Clarity Enhancement Challenge

Jinjiang Liu, Xueliang Zhang

College of Computer Science, Inner Mongolia University, China

jetliu1994@foxmail.com, cszxl@imu.edu.cn

Abstract

This technique report summarises an end-to-end system based on DRC-NET for the 2nd clarity enhancement challenge. In this hearing aid system, first, the DRC-NET recovers target speech from a noisy mixture sampled by the Behind-The-Ear (BTE) hearing aid microphone array. Then, a personalized amplification filter calculated by NAL-R is applied to the estimated speech signal to adopt audiograms of hearing-impaired listeners. The DRC-NET combines convolutional neural networks (CNN) and RNNs as a basic neural operator in dense U-Net architecture, which helps the system could better modeling the long-term variations and short-term local structure of the speech spectrum. Experimental result shows that the system significantly outperforms the baseline system in the development test set.

Index Terms: the 2nd clarity enhancement challenge, DRC-NET, speech Enhancement, hearing aids

1. Introduction

The 2nd Clarity challenge aims to find optimal multi-channel speech enhancement algorithms for behind the ear (BTE) hearing aids in noisy reverberate environments. The simulated BTE device has 3 microphones (front, mid, and rear) per each ear, spaced 7.6 mm apart. In this challenge, the target talker is approximately in front of listeners, and the listener would slightly steer their head. There is also some additional information that could be utilized, a head rotation signal, 4 short and clean utterances of the talker, and an audiogram of the talker measured previously. In this system, we only use the microphone array signals for the DRC-NET speech enhancement and the audiogram coefficient NAL-R hearing aid amplification [1].

Since target talkers are always present at a small range ahead of the listener, adopting an effective speech enhancement model with good spatial sensitivity as a fixed neural beamformer could be a reliable solution. In our previous works on inplace GCRN, which is a spatial sparsity-based neural network speech enhancement algorithm, we proved that spatial cue exists in each frequency bin, and the speech signal can be effectively picked up by the way of inplace operation. The inplace operation means to avoid down-sampling the frequency, which leads to the loss of spatial information. However, the difficulty in this challenge is the low spatial sparsity due to the reverberation and the multiple interfering sound sources. Therefore, we need a large model, which not only has the inplace characteristic for spatial cue processing, but is also good at speech pattern modeling and reverberation processing. The DRC-NET is our latest work on speech dereverberation, as reverberation is copies of speech that have been reflected with infinite times, DRC-NET uses RNN for long-term variation like speech envelope and reverberation tail modeling as IIR process. and utilize CNN for local fine structure modeling as FIR process,

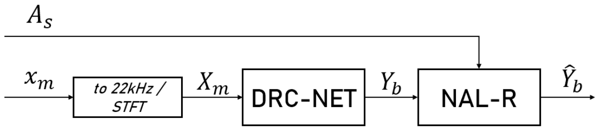

Figure 1: *The proposed DRC-NET based BTE hearing aid system*

2. Methodology

2.1. System overview

The proposed system includes a DRC-NET for multi-channel speech enhancement and a NAL-R module for hearing aid amplification, as shown in Figure 1. The 6 channel noisy speech signal x_m sampled by the BTE microphones in $44.1kHz$, is first sampling to $22.05kHz$ and transformed to time-frequency domain as X_m using Short Time Fourier's transformation (STFT). To meet the acquirement of the 5ms look forward limit of the hearing aid delay, we use 512 sample (23ms) FFT with 5ms (110 samples) frame-shift, and 18 ms left zero padding for the DRC-NET cold start.

2.2. DRC-NET

The DRC-NET is a T-F domain neural network, which is constructed by convolutional layers and DRC blocks in a common dense U-Net structure. Figure 3. shows the structure of DRC-NET with detailed descriptions of intermediate feature size. The major difference from the original DRC-NET, is the input layer does not down-sampling the frequency dimension, which leads to the first and last DRC block being in inplace style for better spatial sensitivity.

2.2.1. Channel-wise GRU

The channel-wise GRU is the GRU processing in the frequency series of a frame, or in the time series of a frequency bin. When using channel-wise GRU processing time sequence of frequency bins, the Channel GRU can effectively extract spatial information inside each frequency bin [2]. Which also have been utilized as spatial filters both in PCG-AIID systems [3], and for IGCRN-based stereo AEC. [4]. When it processing in a frame along the frequency axis it could also effectively model the frequency pattern of speech signal [5] [6]. In this system, we use bidirectional GRU in the frequency dimension and unidirectional GRU in the time dimension.

2.2.2. RC unit and DRC-Block

In this system, the Channel-wise GRUs are cascaded and connected with CNN as an RC unit. RNN can effectively maintain long-term information and CNN could better tackle the local detail. However, the naive stack of many RC units would not

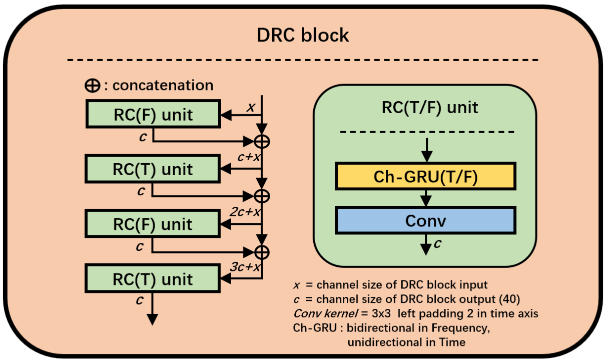
DRC block
⊕ : concatenation
RC(F) unit
RC(T) unit
RC(F) unit
RC(T) unit
x
c
c+x
2c+x
3c+x
RC(T/F) unit
Ch-GRU(T/F)
Conv
c
x = channel size of DRC block input
c = channel size of DRC block output (40)
Conv kernel = 3x3 left padding 2 in time axis
Ch-GRU : bidirectional in Frequency,
unidirectional in Time

Figure 2: *The RC unit and DRC block*

work, due to the gradient vanishing in the backward stage and information loss in the forward stage. Dense connectivity is an effective solution that effectively utilizes many identity mapping to tackle those problems. To be mentioned that, the convectional kernel size is set to 3x3, we left padding 2 zero frame to the beginning of the time dimension to make sure the system is causal.

2.3. NAL-R amplification

After the DRC-NET, the enhanced binaural speech signal are frequency-dependent amplified according to the NAL-R fitting and measured audiograms (both L and R ears) for a specific listener. The audiogram containing audible threshold at $[250, 500, 1000, 2000, 3000, 4000, 6000, 8000]$ Hz was measured by an audiometer. The audiogram is transformed into time domain FIR filter by the NAL-R prescription [1] and then applied to the enhanced speech signal.

2.4. System details

Due to the symmetry of the human head structure, we shared the DRC-NET parameters for the left and right ears in both the train and test stages. Except for the output convolutional layer of DRC-NET, the other convolutional layers will be first followed by the ELU activation function, and then a LayerNormalization on channel and frequency dimension. The MAC for each ear is 39.46G, total parameter number is 1.95M.

3. Experiment

3.1. Databases

The official CEC2 dataset contains 6000 scenes for training, and 2500 and 1500 scenes for validation test and evaluation test respectively. Each scene includes 6 channel BTE device recording, a head rotating signal, and 4 short utterances for the target talker. In this system, we only use the 6-channel mixture as system input for DRC-NET speech enhancement. We did not use extra corpus extending the CEC2 challenge dataset.

3.2. Loss function

The loss function is in T-F domain as follow:

$$L = MAE(S_r - Y_r) + MAE(S_i - Y_i) + MAE(S_a^{0.3} - Y_a^{0.3}) \quad (1)$$

where $MAE(.)$ denotes for mean average error. S is the clean target speech spectrum and Y is the DRC-NET estimated spectrum. The subscripts r, i, and a represent the real and imaginary parts and the magnitude of the spectrum, respectively. In the amplitude loss, the amplitude dynamics is compressed by the power of 0.3 for accurate estimation of medium to high-frequency components, which will be further amplified by NAL-R system in most of hearing loss cases.

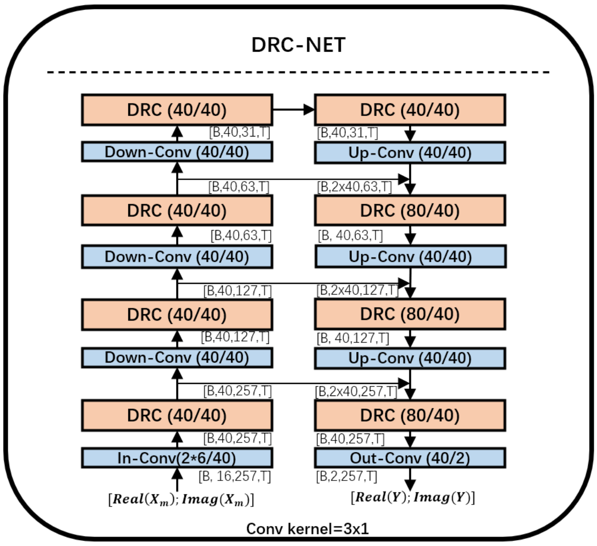

Figure 3: *The DRC-NET architecture*

3.3. Training procedure

The model is trained by AdamW optimizer on a 4 NVIDIA TITAN Xp platform using PyTorch DDP backend, The initial learning rate is set to 1, and gradually tuned down with a decay rate of 0.7 if validation loss does not improve within 15 epochs.

4. Evaluation and Results

The challenge adopts HASPI [7] as the objective quality evaluation. The evaluation results on development test set is shown in Table 1 below.

System	HASPI
Unprocessed	0.1615
NAL-R baseline	0.2493
DRC-NET+NAL-R	0.7118

Table 1: *The evaluation result on development test set*

The results shows that the DRC-NET significantly improved the NAL-R-based baseline system, with a 0.4625 improvement in the terms of HASPI score.

5. Conclusion

In this challenge, we proposed a DRC-NET-based BTE hearing aid speech enhancement system. The DRC-NET is trained as a very powerful neural fixed beamformer. The evaluation shows the system can pick up target speech in extremely noisy reverberant environments.

6. References

[1] D. Byrne and H. Dillon, "The national acoustic laboratories (nal) new procedure for selecting the gain and frequency response of a hearing aid," *Ear and hearing*, vol. 7, pp. 257–65, 09 1986.

[2] J. Liu and X. Zhang, "Inplace Gated Convolutional Recurrent Neural Network for Dual-Channel Speech Enhancement," in *Proc. Interspeech 2021*, 2021, pp. 1852–1856.

[3] J. Li, Z. Yuanyuan, D. Luo, Y. Liu, G. Cui, and Z. Li, "The pcg-aiid system for l3das22 challenge: Mimo and miso convolutional recurrent network for multi channel speech enhancement and speech recognition," 05 2022, pp. 9211–9215.

[4] C. Zhang, J. Liu, and X. Zhang, "Lcsm: A lightweight complex spectral mapping framework for stereophonic acoustic echo cancellation," *arXiv preprint arXiv:2208.07277*, 2022.

[5] J. Liu and X. Zhang, "DRC-NET: Densely connected recurrent convolutional neural network for speech dereverberation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 166–170.

[6] X. Le, H. Chen, K. Chen, and J. Lu, "DPCRN: Dual-Path Convolution Recurrent Network for Single Channel Speech Enhancement," in *Proc. Interspeech 2021*, 2021, pp. 2811–2815.

[7] "The hearing-aid speech perception index (haspi) version 2," *Speech Communication*, vol. 131, pp. 35–46, 2021.