AIAP Final Project

You-Jia Lai (309553005)
Yao-Ren Lee (309511058)
National Yang Ming Chiao Tung University
AAIP
Group 5

1 Introduction

Speech Enhancement(SE) is the task of taking a noisy speech input and producing an enhanced speech output,which has benefited from deep learning in terms of intelligibility and perceptual quality.DCCRN [1] has outperformed previous models on SE with complex-valued operation and s submitted to the Interspeech 2020 Deep Noise Suppression (DNS) challenge ranked first for the real-time-track. Inspire by it and its previous work CRN [2],we found out that though calculating the complex part can help increasing the accuracy,which come behind is the increase of calculation and parameters.Thus,we **aim to design a model which is more lightweight and remains the complex calculation**.

2 CRN and DCCRN

CRN [2] applied an encoder-decoder architecture with casual convolutions.In the purpose of estimating CRM, CRN only take magnitude as input to calculate,then add the phase part after calculation.Recent research shows that complex spectral mapping can outperform magnitude spectral mapping.Still,it calculate the phase part with real convolution,which is in discrepancy with complex number multiplication.It may cause lack of information during learning.The DCCRN [1] fix this problem by adapting complex CNN and BN layer in both encoder and decoder and also using complex LSTM to replace traditional LSTM layer.By computing both part of input wave,more comprehensive information can be learned.

3 Method

Inspired by TCN [3] temporal convolutional neural network. It purposed the dilated causal convolution. The original purpose of TCN is to evaluate CNN whether can replace RNN for sequence modeling. And making DCCRN more lightweight and also keep the complexed calculation is our purpose. The RNN in DCCRN is to capture the long term information. The dilation causal convolution can use large respective field with dilation to capture long term information. Hence, we design the network which use dilated causal convolution to replace the RNN of DCCRN network architecture. Our network architecture becomes a deep complex dilated causal convolution encoder decoder.

Unlike in RNNs where the predictions for later timesteps must wait for their predecessors to complete, convolutions can be done in parallel. And RNN takes 50 percent of the total parameters. We replace the RNN to convolution and for compensate the loss, we add one convolution layers witch 128 channels to encoder and decoder. The number of parameters from 3.7M to 1.8M and the run time from 0.018s to 0.0079s.

3.1 Training target

When training, our model estimates CRM and is optimized by signal approximation (SA). And we convert the mask to polar coordinate. The Cartesian coordinate representation of mask $M = M_r + jM_i$ can also be expressed in polar coordinates:

$$M_{mag} = \sqrt{M_r^2 + M_i^2}$$
$$M_{phase} = \arctan 2(M_i, M_r)$$

The estimated clean speech $\hat{S}$ can be calculated as below:

$$\hat{S} = Y_{mag} \cdot M_{mag} \cdot \exp^{Y_{phase}+M_{phase}}$$

(Y is noisy speech)

3.2 Loss functions

SI-SNR. We follow DCCRN [1] and use SI-SNR as our loss function. But we think that SI-SNR is only constraint on time domain signals. It lacks the knowledge of human perception. Hence, we want to add some auditory knowledge about human. We add another loss function which is perceptual loss. It will talk about in the next section.

Perceptual loss. Except SI-SNR loss, we calculate log mel spectrogram error [4] to be perceptual loss. The enhanced log mel spectrogram and clean log mel spectrogram calculate the RMSE. **We take the logarithm of them. This is motivated by human hearing. We don't hear loudness on a linear scale.** We refer this article to calculate our log mel spectrogram.

Total loss. The total loss of the purposed framework is:

$$loss = (-SISNR + \alpha \cdot perceptual\ loss)/(1+\alpha), \quad where\ \alpha = 2$$

4 Experiments

4.1 Dataset

In our experiments, we train and test the model on Interspeech2020 DNS challenge dataset.The DNS challenge is published by Microsofton Interspeech2020,intended to promote collaborative research in real-time single-channel Speech Enhancement aimed to maximize the perceptual quality and intelligibility of the enhanced speech. We generate 250 hours training set by mix the speech and noise at random SNR between -5 and 20 dB and use their public testing set as our testing set.For further training,we generate 100hrs of reverb data by adding room impulse response ,using image method [5] .

4.2 Training setup

The window length and hop size are 25 ms and 6.25 ms, and the FFT length is 512. We use Pytorch to train the models, and the optimizer is Adam. The initial learning rate is set to 0.001, and it will decay 0.5 when the validation loss goes up. All the waveforms are resampled at 16k Hz

4.3 Experimental results and discussion

PESQ, parameters and rum time. The inference time runs on a PC with one NVIDIA 2080Ti and the length of input wav is 3.75s. The comparison result we can see in table 2. Althogh DCCRN [1] has the best PESQ, **our model is more lightweight and its inference time is more faster. The inference time is close to non-complexed calculation. Also, our model still has competitive PESQ result**.

Visaulize spectrogram and compare different SNR. We also visualize the noisy and enhanced spectrogram. Visualization results show in 1 and 2. The situation of our model is similar to other

Table 1: PESQ on DNS challenge public test set (simulated data only)

Model	PESQ
Noisy	1.582
DCCRN [1]	**2.747**
Ours w/ SI-SNR loss	2.438
Ours w/ SI-SNR and perceptual loss	2.677

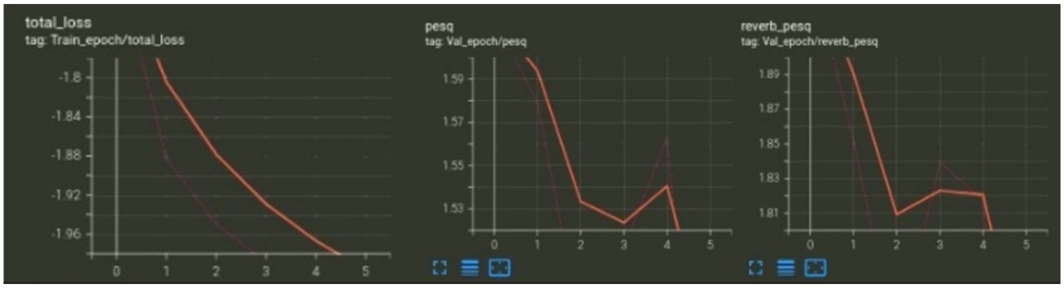
Table 2: Comparison of the number of parameters

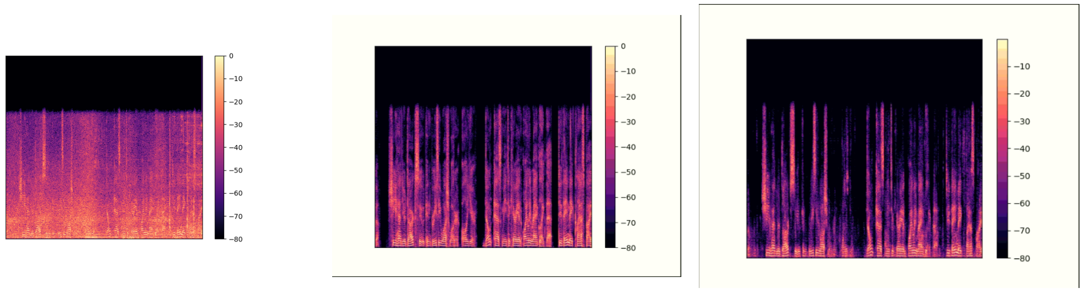
Model	Params	Run time
CRN [2]	17.58M	**0.007979s**
DCCRN-E [1]	3.74M	0.01894s
Ours	**1.88M**	**0.007978s**

speech enhancement methods. In lower SNR, the model performance is not good compared to higher SNR.

Reverb testing set

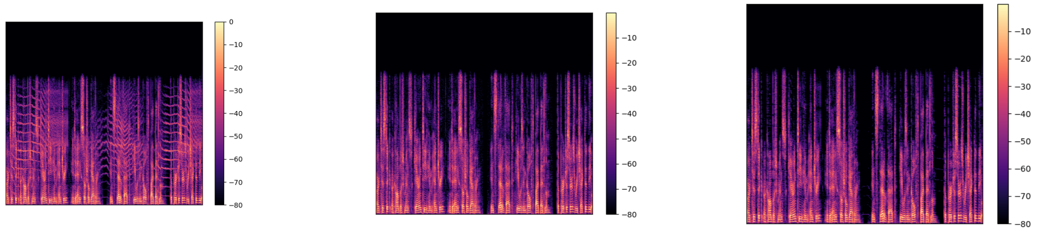
By the figure showing loss and pesq,we can see that although the loss is decreasing,we still got bad pesq score on DNS testset,we infer this may cause by the gap of our data and the testset,which lead our model to overfitting on the reverb data generated by us.





((a)) noisy ((b)) clean ((c)) enhanced

Figure 1: Visualization results of low SNR spectrogram



((a)) noisy ((b)) clean ((c)) enhanced

Figure 2: Visualization results of high SNR spectrogram

5 Conclusion

Our perceptual loss can benefit our model training. And our model is more lightweight and more faster than DCCRN. Also, it still has competitive PESQ result

References

[1] Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, and Lei Xie, "Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement," *arXiv preprint arXiv:2008.00264*, 2020.

[2] Ke Tan and DeLiang Wang, "A convolutional recurrent neural network for real-time speech enhancement," *Interspeech*, 2018.

[3] Shaojie Bai, J Zico Kolter, and Vladlen Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[4] Seo-Rim Hwang, Joon Byun, and Young-Cheol Park, "Performance comparison evaluation of speech enhancement using various loss functions," 2021.

[5] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," 1979.