Journal of Physics: Conference Series

PAPER • OPEN ACCESS

A Single Channel End-to-End Speech Enhancement using Complex Operations

To cite this article: Jie Wu *et al* 2022 *J. Phys.: Conf. Ser.* **2218** 012001

View the article online for updates and enhancements.

A Single Channel End-to-End Speech Enhancement using Complex Operations

Jie Wu[1], Hongqing Liu[1,*], Lu Gan[2] and Yi Zhou[1]

[1]School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing, China
[2]College of Engineering, Design and Physical Science, Brunel University, London, U.K

*Corresponding author email: hongqingliu@cqupt.edu.cn

Abstract. This paper investigates the possibility of using complex operations to perform speech enhancement task in time domain. To that end, first, the Hilbert transform is utilized to prepare the complex input in time domain. After that, the complex temporal convolutional network (CTCN) is developed to conduct complex convolutions. By cascading the TCN and the CTCN modules, the final proposed network form an encoder-decoder structure, which performs an end-to-end speech enhancement task. The results demonstrate that utilizing complex information in time domain indeed improves the enhancement performance. Compared to other approaches, the proposed network also demonstrates a superior performance in terms of objective evaluations.

Keywords: Speech Enhancement; Single-channel; Complex network; Time domain.

1. Introduction

Speech enhancement is one of highly expected tasks in modern speech applications. It aims to separate the target voice from mixture speech signal. In recent years, we have all witnessed that deep learning based approaches have outperformed traditional methods in speech enhancement. The noisy signals can be enhanced in both time-frequency domain and time domain. The time domain enhancement networks can be classified into two methods, which are direct regression approaches[1] and adaptive front-end methods[2]-[3]. The regression approach directly learns the regression function from the mixture to rebuild the target speech, and the adaptive front-end method usually uses convolutional encoder and decoder or a U-shaped network, which is similar to Fourier transform and its inversion.

As a different strategy, speech enhancement can also be conducted in frequency domain using spectrogram. The training methods in the frequency domain are mainly classified into two categories, namely mask-based and mapping-based approaches. For the masking methods, ideal binary masking (IBM)[4] and ideal ratio masking (IRM)[5] often use the amplitude between clean speech and mixed speech to perform enhancement. However, they only pay attention to amplitude information the magnitude spectrum and ignore the phase information because it was believed that the phase information was difficult to estimate. Due to that, the phase information was thus used to reconstruct target speech. This bounds the upper limit of network performance, for the phase deviation also increases interferences. Recently, several networks have adopted the phase reconstruction concept and seen certain improvements. For that, the deep complex U-Net[7] that simultaneously utilizes magnitude spectrum and phase spectrum based on a proposed variant of U-Net[8] was developed to process complex spectrogram. Soon after, the deep complex convolution recurrent network

(DCCRN)[9] adapted the complex multiplication rules to the LSTM part of the enhancement module on the basis of DCUNet.

Inspired by the complex network, this paper attempts to design a complex network in time domain, which means it is an end-to-end system. To do that, the real-valued time-domain signal is passed to Hilbert transform to produce the imaginary part of the corresponding real part. A complex TCN structure, termed as complex temporal convolutional network (CTCN), that uses the time modeling capability of TCN to process complex-valued information in the time domain, is developed.In our experiments, it is found that the addition of time-domain imaginary part information improves the speech quality of the reconstructed target signal and achieves a better performance than the other networks on the DNS1 Challenge dataset.

The rest of article is organized as follows. Section II introduces the descriptions of enhancement task. In Section III, we describe the proposed CTCN model. Section IV describes the experimental details and results. Section V summarizes the article.

2. Mathematical Background

2.1. Signal Model and Preparation

The single-channel speech denoising problem can be described as estimating the original clean source x(t), using the noisy mixture y(t), given by:

$$\mathrm{y}(\mathrm{t}) = x(\mathrm{t}) + n(\mathrm{t}) \tag{1}$$

where n(t) is the unwanted noise.

If the denoising network is conducted in frequency domain, one can transform the signal in(1) to spectrogram using short-time Fourier transform (STFT), and the resulting signal is complex that contains both real and imaginary parts. Based on that, the DCCRN network is developed that utilizes complex information[9]. Taking a different strategy, we attempt to perform speech enhancement in time domain by a use of complex network. To that aim, the imaginary signal needs to be generated from the real-valued time domain inputs.

According to Hilbert transform, an imaginary signal can be produced by its real-valued one, given by:

$$\mathrm{x}_i(\mathrm{t}) = \mathrm{H}[x(\mathrm{t})] = x(\mathrm{t}) * h(\mathrm{t}) \tag{2}$$

where H[·] and * respectively represents the Hilbert transform and convolution, and subscript **i** indicates the imaginary part.

In (2), h(t) is the impulse response of the transform, given by:

$$\mathrm{h}(\mathrm{t}) = (\pi t)^{-1} \tag{3}$$

From the point of view of the frequency spectrum, this transform multiplies the positive frequency part of our original signal by -i, that is, while keeping the amplitude constant, the phase is shifted by -90 degree, and for negative frequency components is 90 degree.

Finally, performing Hilbert transform on both sides of (1) and putting the real and imaginary parts together, one obtains

$$\mathrm{y}_{\mathrm{r,i}}(\mathrm{t}) = x_{r,\mathrm{i}}(\mathrm{t}) + n_{r,i}(t) \tag{4}$$

where **r** and **i** represent the real and imaginary parts of each signal after transformation. The real part represents the original time domain signal, and the imaginary part is the result of the real part signal shifted by a 90 degree.

2.2. Training Target

The model we propose runs in an end-to-end manner and it directly reconstructs the original signal to produce speech enhancement results. In this work, both the real part represented by the clean signal and the imaginary part of the signal after the Hilbert transform are used as training targets. The objective function uses a weighted scale-invariant source-to-noise ratio (SI-SNR). The SI-SNR is

$$\begin{cases} \mathrm{s_{target}} := \langle \hat{\mathrm{s}}, \mathrm{s} \rangle \mathrm{s} \times \|\mathrm{s}\|^{-2} \\ \mathrm{e_{noise}} := \hat{\mathrm{s}} - \mathrm{s_{target}} \\ \mathrm{SI\text{-}SNR} := 10 \log_{10} \|\mathrm{s_{target}}\|^{2} \times \|\mathrm{e_{noise}}\|^{-2} \end{cases} \tag{5}$$

where $\hat{\mathrm{s}}$ and s are the estimator and original clean sources, and $\|\mathrm{s}\|^2$ represents the energy of the signal. By taking the complex signal into account, the final weighted loss function is

$$\mathrm{J} = \mathrm{a} * \mathrm{L_{SI\text{-}SNR}}(X_r) + (1-a) * \mathrm{L_{SI\text{-}SNR}}(X_i) \tag{6}$$

Where **L** function is to calculate the SI-SNR of the estimator and the target. To ensure the real and imaginary parts play the equal importance, empirically, **a** is set to 0.5.

3. Proposed Time Domain Network

3.1. Complex Temporal Convolutional Module

The temporal convolutional network (TCN) module can be used to solve sequentially prediction[2]. Each TCN block includes a 1*1-conv operation, followed by a dilation convolution operation, and finally a skip connection path to avoid the problem of gradient disappearance, as shown in Figure 1. Despite the excellent performance of TCN, it only takes real signal as its input and real convolutions are conducted. To develop our time domain complex network, our building block is also TCN, but it takes complex signal as input and complex convolutions are performed, termed as complex temporal convolutional network module (CTCN), which is provided in Figure 2. Compared with original TCN, CTCN adds an imaginary branch to the ordinary real-valued TCN, and at the same time, introducing complex multiplication in the output part to simulate the correlation between magnitude and phase.

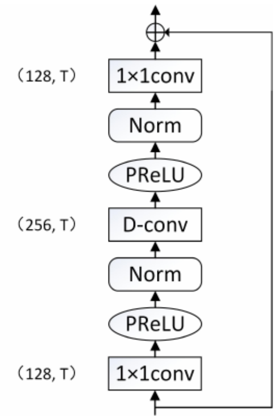

Figure 1. The diagram of each layer in original TCN

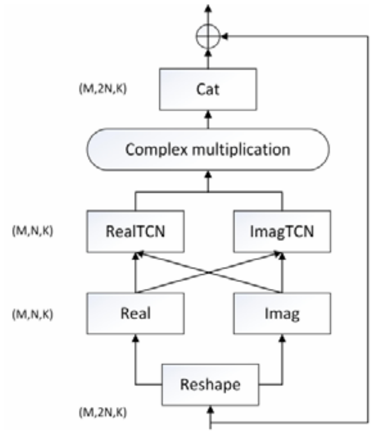

Figure 2. The diagram of each layer in the proposed CTCN

From Figure 2, the CTCN block simulates complex convolutions with real-value TCN, where CTCN consists of two real-value TCN operations, which controls the complex information from inputs.

3.2. Convolutional Time-domain Complex Speech Enhancement Network

The proposed network is an essentially causal convolutional architecture with a separation module. First, short segments of the input waveform are converted into intermediate high-dimensional features by an encoder that combines one-dimensional convolution and nonlinear activation functions. After that, the separation module stacked by 1-D dilated convolutional blocks is used to estimate the high-dimensional feature mask corresponding to the source. Finally, the decoder is used to transform the masking characteristics of the enhancement and reconstruct the time-domain waveform.

Specifically, the enhancements block is shown Figure 3. First, we perform Hilbert transform on the time domain signal to obtain the imaginary part information. The real and imaginary parts share the same one-dimensional convolutional encoder for feature modeling, and a speech enhancement module estimates the mask of the enhancement, where the last TCN module are replaced with CTCN module. The complex multiplication rule is used on the CTCN module to combine the real and imaginary parts. The network structure resembles the Conv-Tasnet and therefore, the proposed network is named as CTCN-Tasnet in this work。

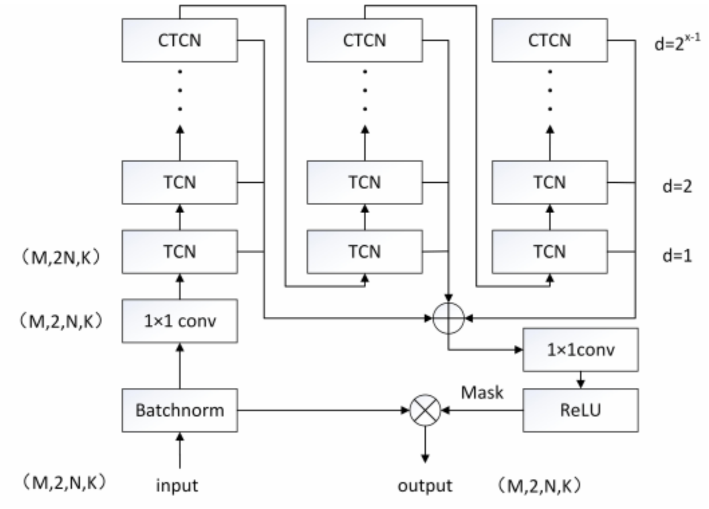

Figure 3. The structure of enhancement block

4. Evaluation Results and Comparisons

4.1. Datesets

In our experiments, we first evaluate the proposed model and several baselines on the Interspeech2020 DNS challenge dataset (DNS1)[9]. We use the script provided by DNS1 to generate training, validation, and evaluation sets, which are 40, 10, and 12.5 hours of utterance, respectively. The mixture in train and verification is generated by randomly selecting utterances from the speech set and noise set, and mixing them at a random signal-to-noise ratio (SNR) of -5 dB to 20 dB. An evaluation set is generated under 5 typical SNR of (0, 5, 10, 15, 20)dB.

4.2. Training Setup and Baselines

The time domain model cuts all audio into 2 ms segments, and the frequency domain model uses the original best configurations of each model.We adopt four state-of-the-art baselines for comparisons, and they are CRN [10], Conv-TasNet [2], DPRNN[3]. All the models are tested with their best configurations based on the suggestions in their papers.

4.3. Results and Analysis

We now compare the proposed model to different models with PESQ [11] and the results are provided in Table 1.

Table 1. PESQs with various numbers of CTCN on DNS1 dataset.

_____Model SNR	0db	5db	10db	15db	Ave
Noisy	1.860	2.167	2.495	2.803	2.331
CRN	2.541	2.832	3.033	3.323	2.932
Conv-TasNet	2.623	2.974	3.151	3.432	3.043
DPRNN	2.682	3.022	3.237	3.504	3.115
DCCRN	2.752	3.043	3.277	3.513	3.144
CTCN-TasNet	**2.801**	**3.097**	**3.326**	**3.537**	**3.184**

From Table 1, It can be found that the performance of the CTCN-TasNet outperforms the baselines of CRN, Conv-TasNet, and DPRNN, which demosnrates the effectiveness of exploring the imaginary information in the signal. In addition, the performance of the CTCN-TasNet network is similar to DCCRN. In the low SNR case, the CTCN-TasNet is superior to the DCCRN, whereas in high SNR scenario, the DCCRN network is slightly better than the CTCN-TasNet. On average, however, they produce the same performance.

Table 2. PESQ under different future frames on the DNS datasets with SNR = 0 dB.

Model	Causality (look head)	DNS1 PESQ	DNS3 PESQ
Noisy		1.86	2.51
Conv-TasNet	nocausal	2.62	3.02
Conv-TasNet	21.25ms	2.55	2.96
Conv-TasNet	1.25ms	2.42	2.86
CTCN-TasNet	nocausal	**2.80**	**3.07**
CTCN-TasNet	21.25ms	**2.72**	**3.05**
CTCN-TasNet	1.25ms	**2.57**	**2.99**

To further show the performance improvement of the CTCN-TasNet over Conv-TasNet, we conduct the experiments on the DNS1 and DNS3 datasets at the same time allowing to access different future frames. For the TCN network, by applying causal convolution, the network can be easily modified to access limited future frames. We let the first few layers be non-causal, while the remaining layers are causal. In Conv-TasNet and CTCN-TasNet, when L = 40 (convolution kernel size) and N = 512 (channels of encoder and decoder) adopting 1/2 overlapping coding, the number of non-causal layer is three, and the number of causal layer is five and the rest of the hyperparameter settings are the same as the paper. By calculations, the future information accessed by the algorithm is (40*8+40/2)/16000=21.25 ms. Table 2 shows PESQ of different methods on DNS datasets with SNR = 0 dB.

As we can see from Table 2, the proposed CTCN-TasNet outperforms Conv-TasNet in all the cases considered. On the DNS1 dataset, CTCN-TasNet achieves an average improvement of 0.17 in terms of PESQ, whereas on the DNS3, CTCN-TasNet achieves an average of 0.09 in terms of PESQ. These results show that accessing future frame information can effectively promote the result of the network, but network still needs to be tuned for different scenarios.

5. Conclusion

This paper proposes a new way to process time-domain speech waveforms using complex operations. Given only the real waveform is available, Hilbert transform is utilized to construct the corresponding imaginary part. With both the real and imaginary signals in time domain, the CTCN module is developed to efficiently perform complex convolutions. The proposed CTCN-TasNet outperforms other networks in terms of PESQ for a smaller model size. It is of interest to point out that this transform is independent of the subsequent networks used and can be applied to different structures, which is our future work.

References

[1] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, “End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, pp. 1570–1584, 2018.

[2] Y. Luo and N. Mesgarani, “Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation,” IEEE Transactions on Audio, Speech, and Language Processing, pp. 1–1, 2019.

[3] L. Yi, C. Zhuo, and Y. Takuya, “Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation,” ICASSP, pp. 46–50, 2019.

[4] A. Narayanan and D. Wang, “Ideal ratio mask estimation using deep neural networks for robust speech recognition,” Acoustics, Speech and Signal Processing, pp. 7092–7096, 2013.

[5] X. Li, J. Li, and Y. Yan, “Ideal ratio mask estimation using deep neural networks for monaural speech segregation in noisy reverberant conditions,” INTERSPEECH, pp. 1203–1207, 2017.

[6] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “An experimental study on speech enhancement based on deep neural networks,” IEEE Signal Processing Letters, pp. 65–68, 2014.

[7] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” MICCAI, 2015.

[8] 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 phaseaware speech enhancement,” INTERSPEECH, pp. 2472–2476, 2020.

[9] C. K. R. A., G. Vishak, C. Ross, B. Ebrahim, C. Roger, D. Harishchandra, M. Sergiy, A. Robert, A. Ashkan, B. Sebastian, R. Puneet, S. Sriram, and G. Johannes, “The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results,” in INTERSPEECH, 2020, pp. 2492–2496.

[10] K. Tan and D. Wang, “A convolutional recurrent neural network for real-time speech enhancement,” Interspeech, pp. 3229–3233, 2018.

[11] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs,” in 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221), vol. 2. IEEE, 2001, pp. 749–752.