음성 향상을 위한 U-net과 컨볼루션 순환 네트워크의

아리얄 니산, 이상웅*
가천대학교, *가천대학교
nisanaryal123@gmail.com, *slee@gachon.ac.kr

Comparison of U-net and Convolution Recurrent Network for Speech Enhancement

Nisan Aryal, Sang-Woong Lee*
Gachon Univ., *Gachon Univ.

Abstract

Speech enhancement is the task of reducing the noise and improving the quality of the speech. U-net and convolutional recurrent networks are two primary architectures used for the enhancement, and the magnitude of short-time fourier transform (STFT) is used as an input representation. However, a noisy phase is used for the reconstruction in the STFT, limiting the network's performance. Recently, short-time discrete cosine transform (STDCT) has been introduced to overcome this problem. In this paper, we have compared U-net and convolutional recurrent network (CRN) performance for STDCT. Our result shows that U-net outperforms CRN slightly in the performance. However, CRN is lighter and faster than that of U-net.

Keywords: speech enhancement, STFT, STDCT, U-net, CRN

Ⅰ. Introduction

The objective of speech enhancement is to enhance the speech quality by removing or suppressing the noise. Speech enhancement is a fundamental task in signal processing and has wide applications in real world, such as mobile communications and hearing aids. Recently, the demand for the video calling and video conference has significantly increased. This has result to the necessity of real time speech enhancement.

Traditionally, STFT is used as an input representation for speech enhancement [1]. However, in this approach, only the magnitude is enhanced and the noisy phase is used for the reconstruction of the audio. This results in poor performance and limits the quality of the audio. In order to overcome this problem, complex network-based architectures [2, 3] has been proposed. Although complex networks perform outperform than that of the traditional approach, complex networks are computationally heavy. Thus, they are hard to train and are slow in comparison to real-valued networks. In order to solve these problems, STDCT [4] has been recently introduced. Unlike fourier transform, cosine transform gives real value. Thus, STDCT results in a real value representation and the problem of noisy phase is avoided. Similarly, as STDCT is a real representation, complex network is not required to enhance the speech. The real convolution or normal convolutional neural network can be used with STDCT for effective results.

Likewise, there are two main architectures used in speech enhancement. They are U-net [5] and CRN. Both of these networks are based on the encoder-decoder concept. The encoder creates a bottleneck and a latent representation of the input is created, likewise, the decoder takes this latent representation as the input and construct the output. The difference between the U-net and CRN is that there is a recurrent neural network (RNN) present in the bottleneck of the encoder-decoder architecture in CRN, whereas RNN is not present in U-net. The main goal in speech enhancement task is to predict a mask with the help of this architectures and multiply the input representation to create a clean or enhanced speech. Although these two architecture is widely used, they have not been compared in a similar scenario.

In this paper, we have used the STDCT as the input representation and compared the U-net and CRN for speech enhancement.

Ⅱ. Methods

The u-net architecture has four convolutional blocks in the encoder and the decoder. The encoder convolutional block consists of two-layer of convolution, batch normalization, and relu. The decoder convolutional block consists of transpose convolution to upscale the input feature maps, followed by two layers of convolution, batch normalization, and relu. The output of the U-net is passed through a final convolution layer, which is followed by the tanh activation

function to create a mask. Finally, element-wise multiplication is done between the input representation and the mask created by the u-net architecture.

Similarly, CRN consist of seven convolutional blocks in the encoder and the decoder. Each encoder block consists of a convolution layer, followed by batch normalization and prelu. There are two layers of LSTM between the encoder and decoder layers. The decoder block consists of transpose convolution, batch normalization, and prelu. Similar to that of the U-net, the tanh activation function follows the output of the CRN to create a mask, and element-wise multiplication is done between the input and the mask.

Ⅲ. Experiments

1.Experimental setup and dataset

The models are trained in pytorch framework with adam optimizer [6] and scale-invariant signal to distortion ratio [7] as the loss function. Voice bank-DEMAND dataset [8] is used as the dataset. The audio is resampled to 16 kHz before the training, and STDCT is used as the input representation. Finally, we have used wideband pesq for the evaluation. The model is trained in a one-second audio frame cropped randomly from the training data, and testing is done using the whole length.

2.Experimental result

Table 1 Experimental results

Method	Pesq
U-net	2.68
CRN	2.67

The models were trained in the setting mention above, and early stopping is done to find the optimum convergence point. The pesq value is calculated by using the whole length of the audio. From Table1, we can conclude that U-net architecture is slightly better than CRN. The U-net architecture has a pesq of 2.68, and the CRN is slightly behind with a pesq of 2.67.

Table 2 Execution time and total parameters

Method	Execution time in seconds	Total parameters in millions
U-net	0.113	17.2
CRN	0.012	2.2

Table 2 shows the execution time and the total parameters in the two architectures. In order to calculate the execution time, one-second audio is used. Table 2 shows that CRN is 9.41 times faster than U-net. Similarly, while comparing the total parameters, U-net has 17.2 million parameters, whereas CRN has 2.2 million parameters. U-net has more than seven times the parameters than that of CRN.

Ⅳ Conclusions

In this paper, we have studied the U-net and CRN architecture for speech enhancement. The experiment was performed using STDCT. Our result shows that when trained on the similar setup, U-net outperforms CRN by a small margin. However, CRN is the faster and lighter than that of U-net architecture.

ACKNOWLEDGMENT

This work was supported by the Technology development Program (S2735244) funded by the Ministry of SMEs and Startups(MSS, Korea) and by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (2020-0-01907, Development of Smart Signage Technology for Automatic Classification of Non-face-to-face Examination and Patient Status Based on AI)

References

[1] Zhao, H., et al. “Convolutional-recurrent neural networks for speech enhancement”. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018.

[2] Yanxin, H., et al. “DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement”. Interspeech, 2020.

[3] Choi, H. S., et al. “Phase-aware speech enhancement with deep complex u-net”. International Conference on Learning Representations, 2018.

[4] Geng, C., and Wang, L. “End-to-end speech enhancement based on discrete cosine transform”. In 2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), 379-383, 2020.

[5] Ronneberger, O., Fischer, P., and Brox, T. . “U-net: Convolutional networks for biomedical image segmentation”. International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015.

[6] Kingma, D. P., and Ba, J. “Adam: A method for stochastic optimization”. arXiv preprint arXiv:1412.6980, 2014.

[7] Le Roux, J., et al. “SDR-half-baked or well done?. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing” (ICASSP), 2019.

[8] Valentini-Botinhao, C. “Noisy speech database for training speech enhancement algorithms and TTS models”, 2017.