Improved speech reconstruction with the complex-valued full convolution neural network

Weili Zhou*, Ruijie Ji, Jinxiong Lai
School of Electronic and Information Engineering, Foshan University, Foshan, China

ABSTRACT

As a learning goal of deep neural networks, time-frequency masking has become research focus of supervised speech reconstruction. The previous study proves that the complex-value ratio mask (cRM) can simultaneously estimate the amplitude and phase components of the clean signal in the noisy speech. Compared with the other time-frequency masking features, the best performance can be achieved. However, because the imaginary structure is not obvious and the neural network learning is difficult, there is still no accurate estimation method at present. In this paper, we improved the speech reconstruction with the complex-valued full convolution neural network (CFCCN). Based on the theoretical of complex-value neural network, the complex-valued building blocks are designed for CFCCN to handle complex domain operations and estimate cRM. The building blocks include the complex convolution filter, complex activation functions, complex batch-norm, complex pooling and the weight initialization strategies. The experiment results show that in terms of subjective and objective measurements, this work achieves at least improvement of 1.3%-12.5% in contrast to the state-of-the-arts DNN based speech reconstruction methods in challenging conditions, where the environment noises are diverse, and the signals are non-stationary.

Keywords: Speech enhancement, complex-value, CFCCN

1. INTRODUCTION

Speech enhancement is currently widely used in mobile communications, speech recognition front-end modules and hearing aids, etc[1]. When constructing a deep learning model for speech reconstruction, the short-term Fourier transformation (STFT) is used to convert the waveforms into spectrum as the input features. The spectrum is generally represented as a complex form which usually needs to be decomposed into the amplitude and phase component in the real-value network. Therefore, the amplitude spectrum and phase spectrum of clean speech needs to be estimated in speech reconstruction. However, because the imaginary structure is not obvious and the neural network learning is difficult, the phase estimation is often neglected. This leads to most methods only pay attention to the estimation of the amplitude spectrum, and the phase information is reused with the noisy speech when reconstructing the clean speech[2]. The studies have shown that the clean speech reconstructed with the noise signal phase has an error compared to reconstruction with the actual clean speech phase. The lower the signal-to-noise ratio (SNR), the greater the error. The application of the speech reconstruction is often in a noisy environment, and the SNR is relatively low. Therefore, it is urgent to have an effective way to estimate the phase information of the clean speech and improve the reconstruction accuracy and sound quality of the speech reconstruction. This paper improved the speech reconstruction with the complex-valued full convolution neural network (CFCCN) which can simultaneously estimate the amplitude and phase components of the clean signal in the noisy speech.

2. RELATED WORK

There have been a large number of research results on speech enhancement based on deep learning in recent years. Pascual published the SEGAN speech enhancement model, and it has been cited more than 80 times[3]. Rethage proposed end-to-end deep speech enhancement models wavenet in 2019, the noisy speech is employed as the original waveform input without the need to calculate any explicit time-frequency representation, and the performance has made great progress[4]. In 2018, Yoshiki et al. used the ideal time-frequency (T-F) masking and the amplitude spectrum of the target speech as a supervised speech enhancement target, and estimated and synthesized the target speech waveform through a deep neural network. The experiment proved that the separated speech can significantly suppress the noise[5]. Erdogan et

* willychow@163.com

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 1250615
 · 0277-786X · doi: 10.1117/12.2662182

al.[6] further introduced phase information into time-frequency masking in 2019, and proposed the phase sensitive mask (PSM) to reconstruct speech signals. Subsequently, Williamson proposed the complex ratio masking feature, and experimentally verified that complex ratio masking can remove noise more effectively compared to IBM, IRM and PSM. In 2019, an online data-driven noise adaptive updates method for speech enhancement based on variational Bayesian non-negative matrix factorization, which improved the model's robustness to actual unknown noise. On the basis of this research, the speech signal separation with the combination of variational Bayesian non-negative matrix factorization and deep neural network was proposed in 2020, which can better solve the problem of source confusion[7]. With the continuous improvement of the GAN model, GAN-based speech enhancement methods have also emerged recently and achieved good results[8-9].

3. PROPOSED METHOD

3.1 Overview of CFCCN

As shown in Figure 1, the architecture of the CFCCN is divided into the following stages: 1) Design phase: The building blocks include the complex convolution filter, complex activation functions, complex batch-norm, complex pooling and the weight initialization strategies; 2) training phase: we use international standard database to build cRM calibration data sets, network training sets and network training sets and testing set. We train the CFCCN model based on cRM training data; 3) Testing stage: The cRM of the input noisy speech spectrum is estimated using the trained CFCCN model. Then we apply the cRM mask matrix to the noisy sample to reconstruct the spectrogram of the speech component..

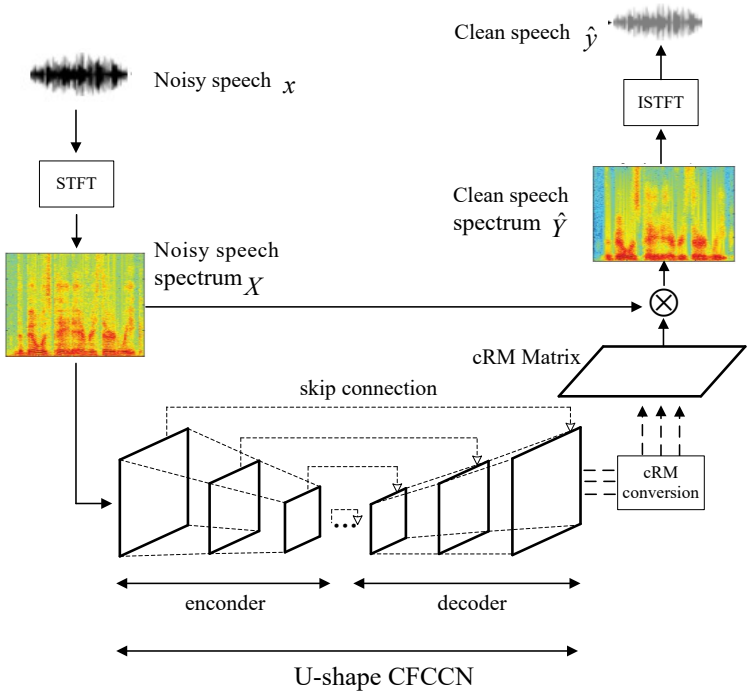

Figure 1. The architecture of the CFCCN.

3.2Complex building blocks for CFCCN

We design the complex building blocks for CFCCN which are similar as that proposed by the complex value neural network. Full detail can be found[10].

(1) Complex-valued convolutional filter

The Complex-valued convolutional filter matrix is obtained by W = A + iB:

$$\mathbf{W}*h=(\mathbf{A}*x-\mathbf{B}*y)+i(\mathbf{A}*x+\mathbf{B}*y) \tag{1}$$

The real and imaginary parts of the filter matrix is defined as follow:

$$\begin{bmatrix}\Re(\mathbf{W}*h)\\ \Im(\mathbf{W}*h)\end{bmatrix}=\begin{bmatrix}\mathbf{A} & -\mathbf{B}\\ \mathbf{B} & \mathbf{A}\end{bmatrix}*\begin{bmatrix}x\\ y\end{bmatrix} \tag{2}$$

(2) Activation function

The leaky CReLU is used to replace the ReLU

$$\mathrm{CReLU}(z)=\mathrm{ReLU}(\Re(z))+i\mathrm{ReLU}(\Im(z)) \tag{3}$$

(3) Complex batch-norm

The batch-norm is defined as follow:

$$\overset{\square}{x}=(\mathbf{V})^{-\frac{1}{2}}(x-E[x]) \tag{4}$$

where V is calculated:

$$V=\begin{pmatrix}V_{rr} & V_{ri}\\ V_{ir} & V_{ii}\end{pmatrix}=\begin{pmatrix}\mathrm{Cov}(\Re\{x\},\Re\{x\}) & \mathrm{Cov}(\Re\{x\},\Re\{x\})\\ \mathrm{Cov}(\Im\{x\},\Re\{x\}) & \mathrm{Cov}(\Im\{x\},\Im\{x\})\end{pmatrix} \tag{5}$$

(4) Weight initialization strategies

The same steps[10] is used to compute the variance of the complex weight parameters.

4. EXPERIMENT

4.1 Database

The TIMIT corpus[11] which contains a total of 6,300 sentences is used to provide the clean speech samples. The noise samples are selected from the NOISEX-92[12] and DEMAND[13] datasets and used to synthesis the training noisy speeches. The DEMAND and NOISEX-92 include 38 types of real-world noise, and mainly divided into "inside" category and "open air" category. 1000 clean samples are selected in TIMIT and corrupted by the noise signal in NOISEX-92 and DEMAND, thus in total 8000 noisy speech are synthesised from -5dB to 10dB signal to noise ratio (SNR). 8000 noisy speech are used to train the model, and 2000 samples are used in testing.

4.2 Experimental setup

CFCCN consists of 10 convolutional layers and the Adam optimizer is used in the model optimization. The epoch number is 500. The dropout parameter is 0.5 and batch norm is applied to achieve faster convergence of the network. The speech samples are sampled at 16 kHz, and the short-time Fourier transform (STFT) number is set to 512 to produce 32-ms frame length data with 50% overlap. The Time-Frequency features are extracted to build the 257 dimensions input vector and then fed to the network. The state-of-the-arts speech reconstruction algorithms include NAAGN[8], SEGAN[3], DAGAN[9] are the baseline methods for comparison. The waveforms and spectrograms are taken as the subjective measure. The improvement of gSIR and gSDR are used as the objective evaluation metrics. The higher the improvement demonstrates a better performance of the method.

4.3 Result and discussion

Figure 2 shows the subjective comparison of different algorithms. Figure 2a is the clean speech Figure 2b is the clean speech corrupted by white noise at 5dB. Figures 2c-2f illustrates the quasi-clean speech reconstructed of SEGAN, NAAGN, DARGAN[3, 8-9] and CFCCN respectively.

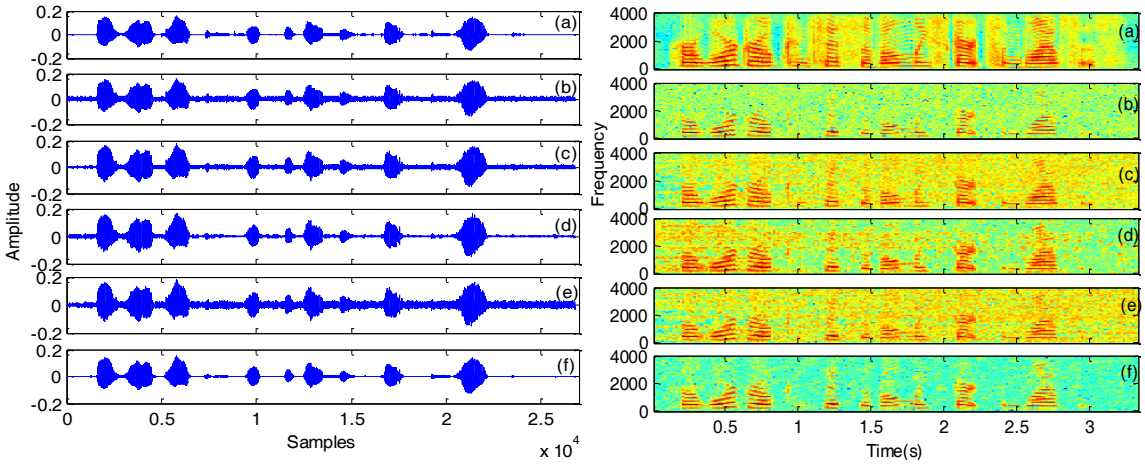

Figure 2. Subjective comparisons of different algorithms.

Table 1. gSIR comparison: CFCCN vs. baselines.

Datasets / Model	NOISEX-92	DEMAND
SEGAN[3]	7.2	7.8
NAAGN[8]	7.3	7.5
DARGAN[9]	7.9	8.3
CFCCN	8.4	9.2

Table 2. gSDR comparison: CFCCN vs. baselines.

Datasets / Model	NOISEX-92	DEMAND
SEGAN[3]	6.2	6.0
NAAGN[8]	6.5	5.8
DARGAN[9]	6.9	6.5
CFCCN	7.4	7.3

According to waveforms, Figure 2f obtains a cleaner signal than the other comparison, and is closer to the Figure 2a. In terms of the spectrograms, Figures 2c-2e are still noisy compared with Figure 2a. On the contrary, the voice parts of Figure 2f are of less artifacts and distortions, and the noises appear to be suppressed. The results show that Figure 2f obtains a cleaner signal than the other comparison, and is closer to Figure 2a. Furthermore, Tables 1-2 demonstrate the comparison of gSIR, gSDR. It shows that the CFCCN obtains the best performance on NOISEX-92 and DEMAND dataset due to its effectiveness of complex-valued networks for speech reconstruction and phase correction. SEGAN[3] and NAAGN[8] outperform DARGAN[9] on average, mainly because the deployed dynamic attention mechanism has more speech coherence and fidelity.

5. CONCLUSION

This work improved the speech reconstruction with CFCCN. Based on the theoretical of complex-value neural network, the complex-valued building blocks are designed for CFCCN to handle complex domain operations and estimate cRM. The experiment results show that in terms of subjective and objective measurements, this work achieves at least improvement of 1.3%-12.5% in contrast to the state-of-the-arts DNN based speech reconstruction methods.

ACKNOWLEDGMENTS

This work is supported by the Natural Science Foundation of Guangdong Province (2019A1515111148), Guangdong Province Colleges and Universities Young Innovative Talent Project (2019KQNCX168).

REFERENCES

[1] Wang, D., "Deep learning reinvents the hearing aid," IEEE Spectrum, 54(3), 32-37(2017).
[2] Zhou, W. L., "Sparse representation-based quasi-clean speech construction for speech quality assessment under complex environments," IET Signal Processing, 11 (4), 486-493(2017).
[3] Pascual, S., Bonafonte, A., and Serra, J., "SEGAN: Speech enhancement generative adversarial network," Proc. Interspeech, 77-82(2018).
[4] Rethage, D., "A wavenet for speech denoising," Proc. ICASSP, 423-426(2019).
[5] Masuyama, Y., "Consistency-aware multi-channel speech enhancement using deep neural networks," arXiv preprint: arXiv:2002.05831, (2018).
[6] Erdogan, H. and Roux, J. L., "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," Proc. ICASSP, 708-712(2019).
[7] Zhou, W., Zhu, Z. and Liang, P., "Speech denoising using Bayesian NMF with online base update," Multimedia Tools and Applications, 78(2), 15647-15664(2019).
[8] Deng, F., Jiang, T., Wang, X. R. and Zhang, C., "NAAGN: Noise-aware attention-gated network for speech enhancement," Proc. Interspeech, 2457-2461(2020).
[9] Li, A., Zheng, C., Peng, R. and Fan, C., "Dynamic attention based generative adversarial network with phase post-processing for speech enhancement," arXiv preprint arXiv:2006.07530, (2020).
[10] Pal, C. J., "Deep complex networks," Proc. ICLR, 1-10(2018).
[11] "TIMIT speech corpus", https://catalog.ldc.upenn.edu/, (2020).
[12] "NOISEX-92 database", http://www.auditory.org/mhonarc/2006/msg00609.html, (2020).
[13] Thiemann, J., Ito, N. and Vincent, E., "The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings," The Journal of the Acoustical Society of America, 133(5), 3591-3591(2013).