A FULLY CONVOLUTIONAL NEURAL NETWORK FOR COMPLEX SPECTROGRAM PROCESSING IN SPEECH ENHANCEMENT

Zhiheng Ouyang[1]*, Hongjiang Yu*[1]*, Wei-Ping Zhu*[1] *and Benoit Champagne*[2]

[1]Department of Electrical and Computer Engineering, Concordia University, Canada
[2]Department of Electrical and Computer Engineering, McGill University, Canada

ABSTRACT

In this paper we propose a fully convolutional neural network (CNN) for complex spectrogram processing in speech enhancement. The proposed CNN consists of one-dimensional (1-d) convolution and frequency-dilated 2-d convolution, and incorporates a residual learning and skip-connection structure. Compared with the state-of-the-art, the proposed CNN achieves a better performance with fewer parameters. Experiments have shown that the complex spectrogram processing is effective in terms of phase estimation, which benefits the reconstruction of clean speech especially in the female speech case. It is also demonstrated that the model yields a convincing performance with small memory footprint when the number of parameters is limited.

Index Terms— speech denosing, complex spectrogram, phase processing, frequency dilation, fully convolutional neural network

1. INTRODUCTION

Recent studies on speech enhancement have resorted to deep learning as a primary tool to develop a data-driven method. In particular, the fully-connected deep neural network (for the purpose of simplicity, we call it DNN in this paper) has been widely adopted and investigated as a non-linear mapping function between noisy features and clean ones [1–4]. Most recently, some researchers have attempted to replace DNN by CNN [5, 6] to provide a more flexible architecture. Park and Lee [5] proposed a fully convolutional network for speech denosing, where CNN is employed to extract the features for the reconstruction of the clean speech. Rethage et al. [6] applied WaveNet [7], a CNN for speech synthesis, to directly estimate clean speech in the time-domain.

It should be mentioned that in most of the existing algorithms, the speech spectral phase remains unchanged. Yet studies have shown that employing spectral phase can further improve the perceptual quality of speech [8–10]. Specifically, Krawczyk and Gerkmann [10] showed that the perceptual evaluation of speech quality (PESQ) could be improved by around 0.2 when using the combination of noisy magnitude and estimated phase for speech reconstruction. While it may be beneficial to process noisy phase for a better denoising performance, yet it is difficult to directly estimate the true phase of clean speech from noisy phase using deep learning, possibly due to the wrapping effect and the lack of phase structure in human speech [3, 11].

Besides, some speech enhancement approaches have been developed based on the processing of complex noisy spectrogram, in which the noisy phase is implicitly processed. Williamson et al. [3] proposed a DNN-based masking technique to estimate complex masking from a spectral feature set. Fu et al. [11] employed a CNN to estimate clean complex spectrograms directly from noisy spectrograms. Though performance is improved compared with DNN-based magnitude-processing method, no further evidence is given to show the effectiveness of phase estimation through complex spectrogram processing.

In this paper, we propose a new CNN structure for complex spectrogram processing. Compared with the previous work [3, 11], the proposed CNN is fully convolutional, and consists of 1-d convolution and frequency-dilated 2-d convolution. Frequency dilation is employed to produce a large receptive field with small filters. Hence the proposed CNN can be configured with fewer parameters while still achieving a competitive performance. We verify the effectiveness of phase processing through complex spectrogram estimation and demonstrate the improved perceptual quality of the processed speech. Thanks to the fully convolutional architecture, the proposed CNN still yields a good performance when the number of parameters is limited and the memory footprint is kept relatively small, which leads to a memory-efficient model that facilitates the implementation of the proposed algorithm on embeded devices.

2. ALGORITHM DESCRIPTION

2.1. Complex Spectrogram

Consider a noisy speech signal $x(t) = s(t)+n(t)$, where $s(t)$ and $n(t)$ are the clean speech and additive noise, respectively, and t is the discrete-time index. The complex spectrogram of $x(t)$ is defined as its short time Fourier transform (STFT) over consecutive frames, i.e., $X(k, l) = \text{STFT}\{x(t)\}$. The com-

plex spectrogram can be expressed as $X(k,l) = X_r(k,l) + jX_i(k,l)$, where $X_r(k,l) = \mathrm{Re}\{X(k,l)\}$ and $X_i(k,l) = \mathrm{Im}\{X(k,l)\}$. The task of complex spectrogram processing is to estimate the complex spectrogram $S(k,l)$ of the clean signal $s(t)$, either directly from $X(k,l)$, or from other features obtained from $x(t)$. With the estimated complex spectrogram $\hat{S}(k,l) = \hat{S}_r(k,l) + j\hat{S}_i(k,l)$, the inverse short time Fourier transform (iSTFT) is applied to obtain the estimated clean signal, i.e., $\hat{s(t)} = \text{iSTFT}\{\hat{S}(k,l)\}$. It has been shown that the estimation of complex spectrogram is strongly related to the segmental signal-to-noise ratio (SSNR) [11].

The benefits of employing complex spectrogram estimation are twofolds. Firstly, in contrast to conventional magnitude estimation, by processing complex spectrogram we are estimating the magnitude and phase at the same time, while avoiding the difficulty of phase estimation with neural networks. Secondly, thanks to the similarity between real and imaginary spectrograms [3, 11], it is possible to use a single neural network to estimate them jointly.

2.2. Dilated 2-d and 1-d frequency convolution

The input to the CNN consists of a limited number of successive spectrogram frames. However, the frequency dimension is usually several hundreds and requires a larger size of receptive field in order to exploit the contextual information. Thus, as a common practice, it is necessary to increase the size of the filters in the frequency dimension, e.g., Fu *et al.* [11] used a filter with size of 25 in frequency in their implementation.

Dilated convolution has been successfully applied in various contexts including imaging segmentation [12] and speech synthesis [7]. In dilated convolution, whenever a filter weight is applied to the input, a fixed number of input values are skipped, which makes the size of receptive field larger than that of the filter. Stacking dilation convolution results in an exponential expansion of the receptive field.

Figure 1 shows the frequency-dilated 2-d convolution. By stacking this dilated convolution with an increasing dilation factor, one could keep the size of filter relatively small, while obtaining a large receptive field in frequency. For example, stacking 7 layers of such a frequency-dilated convolution with a filter size of 3 gives a receptive field with size 255, which is enough to cover a spectrogram with 500-point discrete Fourier transform (DFT).

The 1-d convolution is a special case where the filter is only of 1-dimension. In Fig. 2, the 1-d convolution is applied along frequency axis. It is more efficient than 2-d convolution when the goal is to increase the size of receptive field along frequency axis only. Hence it has been used in some recent works [5, 13].

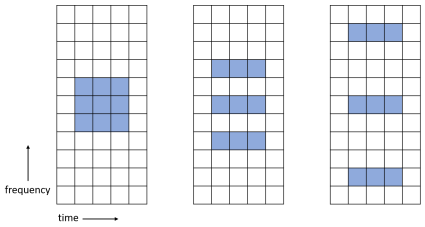

Fig. 1. Frequency-dilated convolution. The filter size is 3×3. From left to right, the dilation factor for frequency is 1, 2, 4, respectively. The dilation factor for time remains as 1

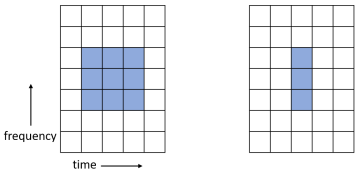

Fig. 2. Left: 2-d convolution. Right:1-d convolution along frequency axis

2.3. Network Architecture

Inspired by WaveNet [7], we propose a convolutional network for complex spectrogram estimation as shown in Fig. 3. It is fully convolutional and consists of a set of 2-d convolutional layers (denoted as Conv2d) and 1-d convolutional layers (Conv1d). The Conv2d layer uses both frequency-dilated 2-d convolution and regular 1-d convolution, while the Conv1d layer only uses regular 1-d convolution along the frequency axis. It is possible to combine real and imaginary spectrograms as the input [11], yet we find treating them separately may lead to a better performance.

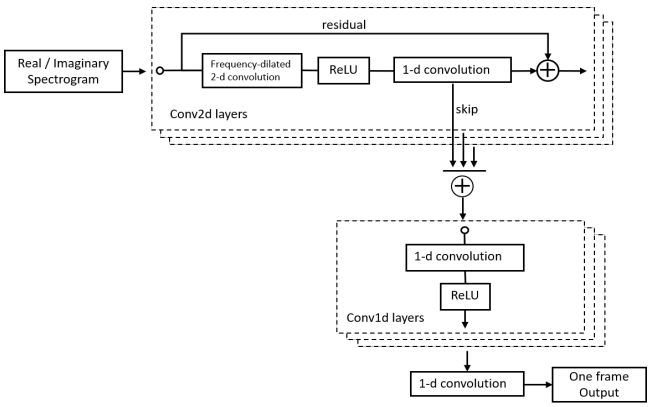

Fig. 3. Proposed CNN architecture

The input to the CNN consists of 13 frames. Stacking Conv2d layers allows the receptive field of time to increase linearly. When the size of the receptive field along time axis equals or exceeds the number of input frames, the central

frame of the output of Conv2d layers will contain the information from all input frames. Hence it is extracted as the input to Conv1d layers to produce a single-frame output.

Table 1 shows a configuration of the proposed CNN. The Conv2d layers are stacked 6 times with frequency dilation increased by a factor of 2 (i.e., 1, 2, 4, 8, 16, 32), which yields a receptive field of size 253 in frequency and size 13 in time. The Conv1d layers are stacked 2 times, followed by the output layer, which is simply 1-d convolution for different channels (denoted as 1d-real and 1d-imag in the table) to separately produce the real-valued frame and the imaginary-valued frame.

The Conv2d layers also adopt a residual learning and skip-connection structure [7, 14] to ease the training of a deeper network. The residual path provides the next Conv2d layer with lower dimensional data from the previous layer which may be lost during the convolution process [5], while the skip connection provides the Conv1d layer with the data processed at the current Conv2d layer.

Table 1. The Network Configuration (Config.1). The height of the filter is the size along the frequency axis, and the width is the size along the time axis. The channel refers to the depth, or the number of feature maps of convolution.

Layer name	Filter name	Height	Width	Channel
Conv2d	dilated 2d	5	3	48
	1d-skip	1	1	48
	1d-residual	1	1	48
Conv1d	1d	3	1	96
Output	1d-real	3	1	1
	1d-imag	3	1	1

3. EXPERIMENTS

The experiment is conducted using TIMIT database [15], in which 780 utterances from both female and male speakers are used for the training and 90 utterances used for testing. Four typical non-stationary noises (babble, street, factory and restaurant) are randomly truncated and used for both training and testing stages.The sampling rate is set to 16 kHz. The SNR levels for training and testing stages are mismatched, with -5 dB, 0 dB, 5 dB, 10 dB for training and -6 dB, 0 dB, 6 dB, 12 dB for testing stage.

3.1. Comparison with previous models

The proposed CNN is compared with two other complex-spectrogram processing methods: CIRM [3] and RI-CNN [11]. CIRM is a DNN-based method that estimates a complex mask from a set of spectral features. RI-CNN is a CNN model that consists of convolution layers and fully-connected layers, and takes complex spectrogram as input. For a fair comparison, all networks are trained and tested with the database and the SNR level described above. Aforementioned SSNR and PESQ are used as performance metrics.

All models use 500-point DFT with 50% overlap. Apart from the DFT length, both reference methods are implemented with the configuration described in the original papers, which makes the number of parameters for RI-CNN, CIRM and the proposed CNN to be $775K$, $3.87M$ and $243K$, respectively. Table 2 shows the result obtained from each network. Clearly, the proposed CNN outperforms RI-CNN, while achieving a comparable performance to CIRM but with around 16 times fewer of parameters.

3.2. Benefit to phase processing

To further investigate whether complex spectrogram processing is beneficial to phase estimation, we combine the clean magnitude with either noisy phase or estimated phase from estimated complex spectrogram to synthesize the speech. We have compared the proposed CNN with the two other complex-spectrogram processing methods. The average PESQ scores for both female and male speech are shown in Fig. 4.

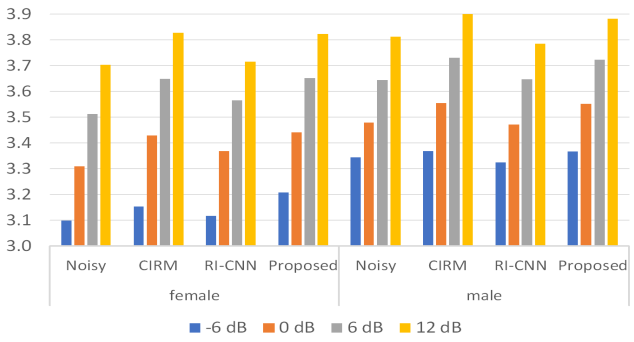

Fig. 4. Average PESQ score on female and male speech by replacing phase

Again, the proposed method shows a comparable performance with CIRM. RI-CNN is found to be the least effective in phase processing with complex spectrogram estimation. For female speech, a maximal improvement of 0.15 is observed. Yet for male speech, the improvement is less significant, possibly indicating that the perceptual quality of female speech is more prone to phase distortion than that of male, and thus benefited more from phase processing.

It is worth mentioning that, current complex spectrogram estimation algorithms may be less effective than algorithms like [10], which estimates phase directly and could improve the PESQ score without processing magnitude. For complex spectrogram estimation algorithms, however, when using the combination of noisy magnitude and estimated phase from the

Table 2. PESQ and SSNR score of different models

metrics	**PESQ**				**SSNR**			
SNR	**-6 dB**	**0 dB**	**6 dB**	**12 dB**	**-6 dB**	**0 dB**	**6 dB**	**12 dB**
unprocessed	1.296	1.674	2.124	2.549	−12.454	−8.046	−2.722	2.994
CIRM	1.740	2.267	2.706	3.071	**−0.874**	**2.242**	5.042	7.504
RI-CNN	1.723	2.018	2.477	2.711	−2.891	0.188	2.710	4.415
proposed	**1.861**	**2.337**	**2.741**	**3.079**	−1.723	2.083	**5.629**	**8.948**

processed complex spectrogram, the improvement on PESQ is rather limited.

3.3. Limiting parameters

Some recent works that utilize CNN for speech processing have considered a situation where the number of parameters is limited [5, 16]. Thus in the third experiment, we have configured the model in a parameter-controlled manner. In addition, the memory footprint of the proposed CNN is also considered. While it is rather implementation dependent, a rough measure could be *(size of spectrogram)*×*(2d convolution channel)*×*(1d convolution channel of Conv2d layers)*×*(size of float)*. Two configurations of the proposed CNN with parameters at the level of $100K$ and $50K$ are tested for the overall denoising performance.

Table 3. A network configuration where the number of parameters is $97K$ (Config.2)

Layer name	Filter name	Height	Width	Channel
Conv2d	dilated 2d	5	3	32
	1d-skip	1	1	24
	1d-residual	1	1	24
Conv1d	1d	5	1	64
Output	1d-real	17	1	1
	1d-imag	17	1	1

By stacking 6 Conv2d layers and 2 Conv1d layers, the configuration shown in Table 1 has $243K$ parameters, and the memory footprint is around 29 megabytes (MB). Meanwhile, two configurations shown in Table 3 and Table 4 keep the number of Conv2d and Conv1d layers unchanged, but uses fewer filter channels to reduce the number of parameters and the memory footprint at the same time. The config.2 shown in Table 3 has $97K$ parameters, and the memory footprint is 10 MB. The config.3 in Table 4 further reduces the parameters and the memory footprint to $50K$ and 6 MB, respectively. Figure 5 illustrates the overall performance for all three configurations. While the model with $97K$ parameters still produces a good overall result, the one with $50K$ suffers more loss on PESQ score. Generally speaking, config.2 seems to reach a good balance between the denosing performance and memory efficiency.

Table 4. A network configuration where the number of parameters is $50K$ (Config.3)

Layer name	Filter name	Height	Width	Channel
Conv2d	dilated 2d	5	3	32
	1d-skip	1	1	16
	1d-residual	1	1	16
Conv1d	1d	1	1	48
Output	1d-real	17	1	1
	1d-imag	17	1	1

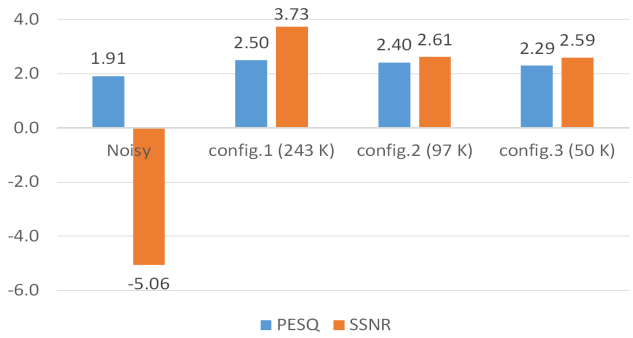

Fig. 5. Performance comparison with different model configurations

4. CONCLUSION

In this study, we have proposed a fully convolutional neural network with frequency-dilated 2-d convolution for complex spectrogram processing. Through a number of experiments, we have demonstrated that the proposed CNN-based method performs very well for complex spectrogram estimation, and that it is also beneficial to phase estimation. We have also paid attention to the memory efficiency of the proposed CNN by considering limited number of parameters and memory footprint, leading to a trade-off between the model complexity and the achievable performance.

References

[1] Y. Xu, J. Du, L. Dai, and C. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, Jan 2015.

[2] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, Dec 2014.

[3] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, March 2016.

[4] Z. Ouyang, H. Yu, W.-P. Zhu, and B. Champagne, "A deep neural network based harmonic noise model for speech enhancement," *Proc. Interspeech 2018*, pp. 3224–3228, 2018.

[5] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," *CoRR*, vol. abs/1609.07132, 2016. [Online]. Available: http://arxiv.org/abs/1609.07132

[6] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 5069–5073.

[7] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016. [Online]. Available: http://arxiv.org/abs/1609.03499

[8] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Communication*, vol. 53, no. 4, pp. 465–494, 2011.

[9] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 55–66, 2015.

[10] M. Krawczyk and T. Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1931–1940, Dec 2014.

[11] S. Fu, T. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," in *IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, Sept 2017, pp. 1–6.

[12] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *CoRR*, vol. abs/1511.07122, 2015. [Online]. Available: http://arxiv.org/abs/1511.07122

[13] K. Tan, J. Chen, and D. Wang, "Gated residual networks with dilated convolutions for supervised speech separation," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 21–25.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[15] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

[16] T. N. Sainath and C. Parada, "Convolutional neural networks for small-footprint keyword spotting," in *16th Annual Conference of the International Speech Communication Association*, 2015.