End-to-End Speech Enhancement Using Fully Convolutional Networks with Skip Connections

Dujuan Wang and Changchun Bao
Speech and Audio Signal Processing Lab, Beijing University of Technology
Beijing, 100124, China
E-mail: wangdujuan@emails.bjut.edu.cn, baochch@bjut.edu.cn

***Abstract*— The purpose of speech enhancement is to extract useful speech signal from noisy speech. The performance of speech enhancement has been improved greatly in recent years with fast development of the deep learning. However, these studies mainly focus on the frequency domain, which needs to complete time-frequency transformation and the phase information of speech is ignored. Therefore, the end-to-end (i.e. waveform-in and waveform-out) speech enhancement was investigated, which not only avoids fixed time-frequency transformation but also allows modelling phase information. In this paper, a fully convolutional network with skip connections (SC-FCN) for end-to-end speech enhancement is proposed. Without the fully connected layers, this network can effectively characterize local information of speech signal, and better restore high frequency components of waveform using lesser number of the parameters. Meanwhile, because of existence of skip connections in different layers, it is easier to train deep networks and the problem of gradient vanishing can also be tackled. In addition, these skip connections can obtain more details of speech signal in different convolutional layers, which is beneficial for recovering the original speech signal. According to our experimental results, the proposed method can recover the waveform better.**

I. INTRODUCTION

This Speech enhancement aims to suppress or reduce noise interference for improving quality and intelligibility of speech disturbed by various noise [1]. In the past few decades, researchers have proposed many speech enhancement methods, and those methods effectively improved the intelligibility and quality of noisy speech. Traditional speech enhancement methods contain spectral subtraction [2], Wiener filtering [3], subspace algorithms [4, 5], statistical model-based methods [6]. Later, with the continuous development of the deep learning, the methods based spectral mapping or mask have been proposed and widely used [7, 8, 9, 10, 11]. Although these methods successfully achieved speech enhancement to some extent, there are still some problems. They almost achieve speech enhancement in frequency domain, which rely on the usage of short-time Fourier transform (STFT) and focus only on processing magnitude spectrogram and the phase information of speech signal is not considered. In fact, the operation of the STFT needs many parameters and cannot optimize performance of speech enhancement absolutely. Moreover, this destruction is usually inevitable, which finally results in the distortion of the recovered speech signal.

To avoid transformation between the time and frequency domain, the methods of performing series of operation on the raw waveform attract more and more attention, especially in the area of automatic speech recognition (ASR) [12, 13, 14, 15]. These methods achieve better performance than those approaches that are based on the hand-crafted features (e.g. MFCC and the cepstrum).

Furthermore, the characteristics of speech signal of time domain and frequency domain are different greatly. In frequency domain, the frequency components are represented on the frequency bins, and repeated patterns of the formants can be observed in low to middle frequency bins while consonants can only occur in high frequency bins. However, in time domain, the situation is totally different. A sample of speech signal cannot alone represent any information, and its information has to depend on its neighbor samples, which makes the task of speech enhancement more difficult. And that is why most researches choose to complete the target of speech enhancement in frequency domain rather than in time frequency. In order to use raw waveform to achieve better performance, the convolution neural network (CNN) was preferred [13, 14, 15], because the locally useful acoustic information can be acquired by convolution operation. However, there are few research researches that directly enhance speech signal in time domain. Since the fully convolutional network (FCN) was proposed for enhancing magnitude spectrum of speech signal [16], an end-to-end model of speech enhancement was subsequently proposed by making use of the FCN [17, 18]. In general, this method is defined in terms of a single model and directly operated on raw waveform.

In this paper, in order to further improve the performance of end-to-end speech enhancement, a fully convolutional network with skip connections (SC-FCN) is proposed. There are two benefits to use SC-FCN to conduct speech enhancement. Firstly, comparing with traditional deep neural network (DNN) and CNN [7, 19], this network only contains convolutional layers, due to the lack of fully connected layers, so lesser number of parameters are needed. Meanwhile, it is well-known that neurons in the convolutional layer are connected to a local region in the input data, and neurons in the convolutional column share the parameters, thus local information of speech signal can be effectively characterized

[17]. Moreover, differing from the FCN [16], in this work, skip connections are added between convolutional layers, which not only makes training networks easier but helps more speech details reserved in different convolutional layers so that the speech signal is recovered better.

The rest of this paper is organized as follows. In Section 2, the related works are described. In Section 3, the details of the proposed method are presented. Experimental setup is provided in Section 4, test results are discussed in section 5, and the conclusions are given in section 6.

II. Related Work

Although speech enhancement based on spectral features have achieved great success, modeling raw waveform for speech enhancement still needs to be investigated. At present, many researches have conducted speech recognition by directly processing raw speech signal [12, 13, 14, 15]. Recently, speech enhancement with raw waveform has begun to attract many attentions. S.-W. Fu has proposed a speech enhancement approach that directly uses the raw waveform [17].

For the characteristics of raw waveform, convolutional operation is considered as the first choice. In early work, CNN was usually chosen as network model to process raw waveform [12, 14, 20]. However, a sample of speech signal and its neighbors are interdependent in time domain, which makes it difficult for fully connected layers to generate high and low frequency parts of raw waveform in the same time [17]. Similarly, it is not easy for the hidden fully connected layers to process raw waveform. Thus, the FCN was considered to map raw waveform [17, 18]. Compared to the traditional CNN, the benefit of the FCN is that all fully connected layers of the CNN are replaced by convolutional layers [17, 21].

Skip connections have been studied for a long time [22, 23, 24, 25, 26]. Skip connections were added to the DNN to achieve very deep networks, which greatly improved the performance of image recognition [23, 24]. In addition, skip connections also played an important role in encoder-decoder networks to achieve image denoising [25]. Additionally, Ming Tu proposed a similar idea that skip connections were added to all layers of the DNN to accomplish speech enhancement [26]. Those studies have achieved experimental goal to some extent by using skip connections.

III. End-to-end Speech Enhancement

As mentioned above, the most research works of speech enhancement were completed in frequency domain, and they focused mainly on processing magnitude spectrum and the phase information was ignored. In fact, the phase is important for the quality and intelligibility of the reconstructed speech. Although the phase components have been taken into consideration by using complex components [27, 28] in later studies, the STFT is still necessary in these methods, which increased computational complexity to some extent. Due to these problems, end-to-end (i.e. waveform-in and waveform out) speech enhancement have become the interest of recent researches. End-to-end speech enhancement model can simply be described in Fig. 1.

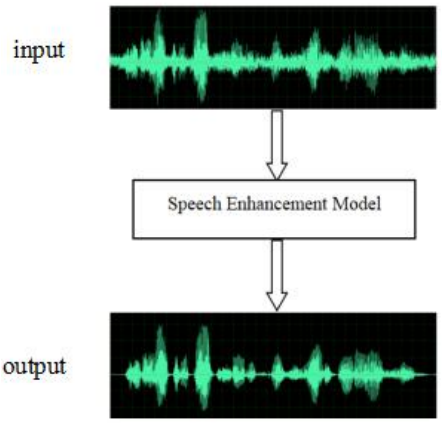

Fig. 1 End-to-end speech enhancement model

As shown in Fig. 1, the input is noisy signal and output is an estimate of clean signal. And end-to-end speech enhancement model is very simple, and the time-frequency domain conversion is unnecessary, which reduces the complexity of computation. The key point of the whole end-to-end speech enhancement system is the design of speech enhancement model since the raw waveform is used. The deep learning model has shown the great advantages in speech enhancement, so how to design better networks for raw waveform to achieve speech enhancement becomes the main topic. In [17], the author proposed raw waveform-based speech enhancement by using FCN. In this paper, in order to further improve quality and intelligibility of the enhanced speech, SC-FCN model is proposed.

Like the FCN, the SC-FCN also only consists of convolutional layers, which is beneficial for the speech enhancement model to efficiently preserve local structures of the features and lesser weights are used. In order to train SC-FCN easier, the dimension of all convolutional layers is one, which not only reduces the number of parameters in each layer, but also reduces the computation cost. The significant benefits of the SC-FCN is its skip connections (e.g. a residual block), as shown in Fig. 2.

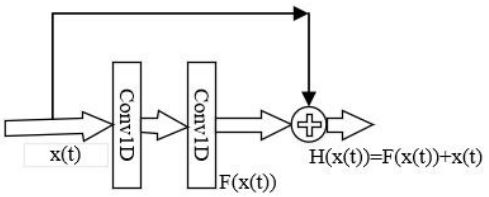

Fig. 2 Skip connection: a residual block

Fig. 2 shows the basic residual block made up of the skip connection. x(t) is the input of the residual block. F(x(t)) and H(x(t)) denote a residual mapping and an underlying mapping, respectively. By adding skip connections, the original

mapping is rewritten as H(x(t)) = F(x(t))+x(t)), and the network fits a residual mapping, that is, F(x(t)) = H(x(t))-x(t), rather than H(x(t)). This makes tiny change become obvious and more sensitive to the output of networks. By adding skip connections to the network, we hope that the network fits a residual mapping instead of directly fitting a desired underlying mapping [24, 25].

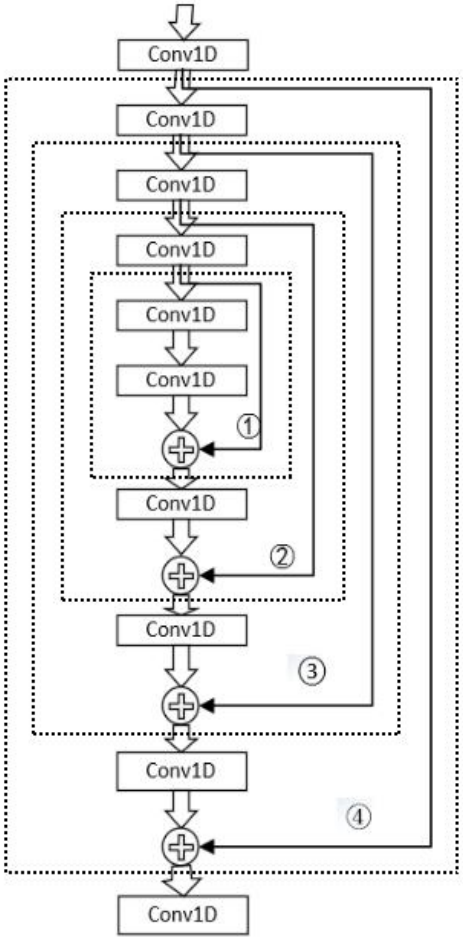

Fig. 3 The architecture of SC-FCN

More details about the proposed network can be found in Fig. 3. As shown in Fig. 3, the proposed fully convolutional network with skip connections only consists of convolutional layers, and there are no max pooling layers in the network like WaveNet [29]. Differing from the architecture of the basic residual network in [24], the residual block is nested each other. Since the networks fit a residual mapping rather than an underlying mapping by adding skip connections, it easier to optimize the networks and obtain better performance. This implies that it is possible to optimize the networks while fitting the residual mapping of residual mapping, and the performance of speech enhancement can be better improved. In our network, nine convolutional layers are used, meanwhile a total of four skip connections are used and four residual blocks (e.g. ① ② ③ ④) are obtained.

Similar to Fig. 2, four blocks in Fig. 3 can be defined as follows:

$$H(x_1(t)) = F(x_1(t)) + x_1(t) \quad (1)$$

$$H(x_2(t)) = F\{H(x_1(t))\} + x_2(t) \quad (2)$$

$$H(x_3(t)) = F\{H(x_2(t))\} + x_3(t) \quad (3)$$

$$H(x_4(t)) = F\{H(x_3(t))\} + x_4(t) \quad (4)$$

where $x_1(t)$, $x_2(t)$, $x_3(t)$ and $x_4(t)$ denote the inputs of four residual blocks, respectively. F is a residual mapping function, and H expresses the identity mapping.

IV. Experimental Setup

In this section, we will describe the experimental setup, database and network parameters.

A. Data Set

To evaluate the performance of speech enhancement using the fully convolutional network with skip connections, the TIMIT database [31] is chosen as the training and test sets. For the training set, 4620 sentences from different speakers in the TIMIT database are used as the training set of clean speech, and three noise types (Babble, F16, Factory) are used as the training set of noise. Meanwhile, three noise types and 4620 sentences are artificially mixed at four different SNR levels (-5dB, 0dB, 5dB, 10dB). We randomly selected 4620 sentences from the mixed speech to build 4-hour training set of noisy speech. For the test set, the clean speech consists of 201 sentences from TIMIT database and three seen noise types (Babble, F16, Factory) are mixed as test sets. In addition, extra three unseen noise types (office, street, volvo) are also chosen to combine with clean speech of test set to obtain noisy speech. The noisy speech of the testing set is also formed at four different SNR levels (-5dB, 0dB, 5dB, 10dB).

All signals are down sampled to 8 kHz. In our experiments, the raw waveform is processed in a frame-wise manner, and 512 samples are extracted from raw waveform to form a frame as used in [17].

B. Network Setup

In order to further improve quality and intelligibility of speech enhancement, we designed a network called SC-FCN based on recent success of the FCN model [17]. The architecture of our proposed model is shown in Figure 3. The proposed network has nine convolutional layers with zero padding and preserve the same size as the input. Four skip connections are added to few stacked layers in the network, and every skip connection is passed through different convolutional layers, and our experiment showed that this configuration can work very well. In our experiments, all of convolutional layers consist of 28 filters and the dimension of a filter is 29*1. Moreover, the last layer only has 1 filter.

To compare our proposed approach, the CNN and FCN are chosen as the reference network models. The inputs of the CNN and FCN are the same as the input of the SC-FCN. The CNN has six convolutional layers and three fully connected layers (1024 nodes), whereas the FCN contains nine convolutional layers. Note that the determination of both the

number of training epoch and model structure all depend on the error of the validation set.

In addition, Leaky rectified linear units (LeakyReLU) [32] is used as the activation function for all the models. The Adaptive moment estimation (Adam) [33] is chosen as the learning optimizer to train network models.

C. *Evaluation Metrics*

The performance of speech enhancement for each approach was evaluated by measuring the perceptual evaluation of speech quality (PESQ) [34] and the short-time objective intelligibility (STOI) [35]. The score of the PESQ varies from -0.5 to 4.5. The higher score indicates the better quality. In addition, the score of the STOI is between 0 and 1. Similarly, the intelligibility of the recovered speech is better when the score of the STOI is higher.

V. RESULTS AND DISCUSSION

The average PESQ and STOI scores of the proposed FCN with skip connections (i.e. SC-FCN), CNN and FCN are presented in Fig. 4 and Fig. 5.

According to Fig. 4 and Fig. 5, we can observe that the performance of the SC-FCN is consistently better than the CNN and FCN, both in the average PESQ and STOI, which proves that the proposed method is more effective in improving the quality and intelligibility of speech than the other two networks. Especially, the average PESQ and STOI scores of the SC-FCN significantly outperform that of the CNN. This further indicates the convolution operation is more advantageous than fully connected layers in processing the raw waveform to achieve speech enhancement.

In order to further observe the performance of different methods for different noise types at four different SNR levels (-5dB, 0dB, 5dB, 10dB), the PESQ and STOI scores of the enhanced speech results are presented in Table I and Table II.

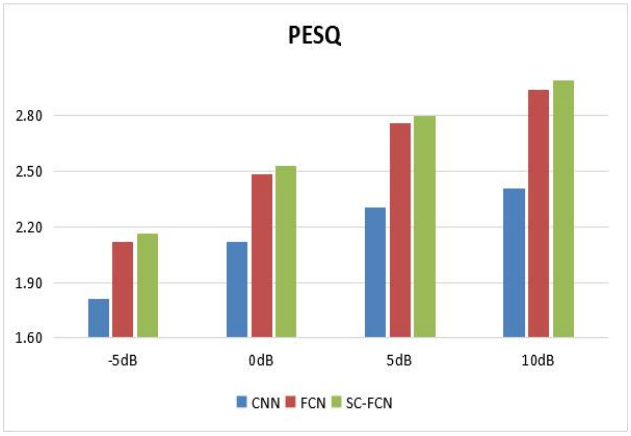

Fig. 4 The average PESQ of different methods at four different SNR levels.

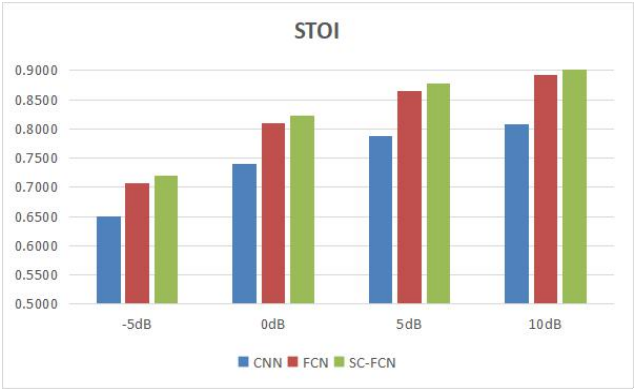

Fig. 5 The average STOI of different methods at four different SNR levels.

Table I: Performance comparison of the PESQ under different methods.

SNR (dB)	**Methods**								
	Babble			F16			Factory		
	CNN	FCN	SC-FCN	CNN	FCN	SC-FCN	CNN	FCN	SC-FCN
-5	1.6207	1.8378	**1.9099**	1.7341	1.9403	**2.0264**	1.9432	2.2103	**2.2512**
0	1.8868	2.2173	**2.2766**	2.0975	2.3538	**2.4278**	2.2094	2.5411	**2.5809**
5	2.0542	2.5182	**2.5634**	2.3514	2.7077	**2.7598**	2.3709	2.7957	**2.8323**
10	2.1370	2.7228	**2.7715**	2.4913	2.9671	**3.0177**	2.4618	2.9689	**3.0145**
	Office			Street			Volvo		
	CNN	FCN	SC-FCN	CNN	FCN	SC-FCN	CNN	FCN	SC-FCN
-5	1.7095	**1.9357**	1.9086	1.8358	2.1509	**2.1626**	2.0211	2.6391	**2.7256**
0	2.0858	**2.3654**	2.3583	2.1522	2.5265	**2.5577**	2.2673	2.9074	**2.9614**
5	2.3191	2.6909	**2.7109**	2.3414	2.8048	**2.8433**	2.4099	3.0430	**3.0891**
10	2.4395	2.9088	**2.9384**	2.4444	2.9864	**3.0450**	2.4638	3.1071	**3.1604**

Table II: Performance comparison of the STOI under different methods.

SNR (dB)	Methods								
	Babble			F16			Factory		
	CNN	FCN	SC-FCN	CNN	FCN	SC-FCN	CNN	FCN	SC-FCN
-5	0.5406	0.5920	**0.6099**	0.6089	0.6451	**0.6679**	0.6738	0.7321	**0.7383**
0	0.6459	0.7209	**0.7405**	0.7260	0.7796	**0.7987**	0.7557	0.8254	**0.8328**
5	0.7118	0.8025	**0.8229**	0.7973	0.8625	**0.8767**	0.7978	0.8750	**0.8832**
10	0.7455	0.8444	**0.8660**	0.8297	0.9021	**0.9130**	0.8155	0.8976	**0.9058**
	Office			Street			Volvo		
	CNN	FCN	SC-FCN	CNN	FCN	SC-FCN	CNN	FCN	SC-FCN
-5	0.6442	0.6868	**0.6909**	0.6835	0.7445	**0.7557**	0.7537	0.8367	**0.8544**
0	0.7470	0.8088	**0.8159**	0.7677	0.8362	**0.8479**	0.8003	0.8834	**0.8949**
5	0.7957	0.8703	**0.8786**	0.8050	0.8824	**0.8921**	0.8174	0.9017	**0.9110**
10	0.8151	0.8963	**0.9049**	0.8185	0.9010	**0.9099**	0.8227	0.9080	**0.9167**

As shown in Table I and Table II, the PESQ and STOI of the SC-FCN are higher than CNN and FCN for different noise types (three seen noise: Babble, F16, Factory; three unseen noise types: office, street, Volvo) at four SNR levels (-5dB, 0dB, 5dB, 10dB) except for the PESQ of office noise at -5 dB and 0 dB. Specifically, the performance of the enhanced speech by using the SC-FCN and FCN is much more effective than CNN. It is mainly because characteristics of the signal in time domain is greatly different from that in frequency domain, that is, the feature in time domain must depend on its neighbors. However, the fully connected layers are weak in processing raw waveform, and the high frequency components are often missed. By comparing experimental results between the SC-FCN and FCN, we can find that the SC-FCN can better improve the quality and intelligibility of the enhanced speech by adding skip connections to networks. This indicates it's effective for the denoising by adding some skip connections to the network to force the network to learn residual signal.

VI. Conclusions

In this paper, an approach based on the fully convolutional network with skip connections was proposed to process raw waveform to further improve performance of end-to-end speech enhancement. Inspired by the FCN, the skip connections were added to the FCN. By adding skip connections, the network fitted a residual mapping instead of fitting an underlying mapping, which makes the learning process easier and convergence faster. The experimental results showed that the proposed network achieved the better performance in end-to-end speech enhancement.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Grant No. 61831019, No. 61471014 and No. 61231015).

References

[1] P. C. Loizou, Speech Enhancement: Theory and Practice, 2nd ed. Boca Raton, FL, USA: CRC Press, Inc., 2013.

[2] M. Berouti, R. Schwartz, and J. Makhoul, " Enhancement of speech corrupted by acoustic noise," in Proc. of the Int. Conf. On Acoustics, Speech, and Signal Processing (ICASS), vol. 4, Apr 1979, pp. 208–211.

[3] J. Lim and A. Oppenheim, " All-pole modeling of degraded speech, " IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. 26, no. 3, pp. 197–210, Jun 1978.

[4] M. Dendrinos, S. Bakamidis, and G. Carayannis, " Speech enhancement from noise: A regenerative approach, " Speech Communication, vol. 10, no. 1, pp. 45–57, 1991.

[5] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement, " IEEE Trans. on speech and audio processing, vol. 3, no. 4, pp. 251–266, 1995.

[6] Y. Ephraim, " Statistical-model-based speech enhancement systems," Proceedings of the IEEE, vol. 80, no. 10, pp. 1526–1555, Oct 1992.

[7] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," IEEE Signal Process. Lett, vol. 21, no. 1, pp. 65–68, Jan. 2014.

[8] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "Dynamic noise aware training for speech enhancement based on deep neural networks," in Proc. Annu. Conf. Int. Speech Commun. Assoc., 2014, pp. 2670–2674.

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

[10] M. Kolbæk, Z.-H. Tan, and J. Jensen, " Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems, " IEEE/ACMTrans. Audio, Speech, Lang. Process, vol. 25, no. 1, pp. 153–167, Jan. 2017.

[11] D.Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," arXiv:1708.07524, 2017.
[12] D. Palaz and R. Collobert, "Analysis of cnn-based speech recognition system using raw speech as input," in INTERSPEECH, 2015, pp. 11-15.
[13] D. Palaz, M. M.-. Doss, and R. Collobert, "Convolutional neural networks-based continuous speech recognition using raw speech signal," in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 4295-4299.
[14] P. Golik, Z. Tüske, R. Schlüter, and H. Ney, "Convolutional neural networks for acoustic modeling of raw time signal in LVCSR," in INTERSPEECH, 2015, pp. 26-30.
[15] W. Dai, C. Dai, S. Qu, J. Li, and S. Das, "Very deep convolutional neural networks for raw waveforms," arXiv preprint arXiv:1610.00087, 2016.
[16] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," arXiv preprint arXiv:1609.07132, 2016.
[17] S. Fu, Y. Tsao, X. Lu and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Kuala Lumpur, 2017, pp. 006-012.
[18] S. Fu, T. Wang, Y. Tsao, X. Lu and H. Kawai, "End-to-End Waveform Utterance Enhancement for Direct Evaluation Metrics Optimization by Fully Convolutional Neural Networks," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 9, pp. 1570-1584, Sept. 2018.
[19] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in INTERSPEECH, 2013, pp. 436-440.
[20] Palaz D, Collobert R, Doss M M. Estimating Phoneme Class Conditional Probabilities from Raw Speech Signal using Convolutional Neural Networks[J]. Computer Science, 2013.
[21] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431-3440.
[22] C. M. Bishop. Neural networks for pattern recognition. Oxford university press, 1995.
[23] Rupesh K Srivastava, Klaus Greff, and J¨urgen Schmidhuber, " Training very deep networks, " in Advances in neural information processing systems, 2015, pp. 2377–2385.
[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.
[25] Xiao-Jiao Mao, Chunhua Shen, and Yu-Bin Yang, " Image denoising using very deep fully convolutional encoder-decoder networks with symmetric skip connections, " arXiv preprint arXiv:1603.09056, 2016.
[26] M. Tu and X. Zhang, "Speech enhancement based on Deep Neural Networks with skip connections," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017, pp. 5565-5569.
[27] D. S. Williamson, Y. Wang, and D. Wang, " Complex ratio masking for joint enhancement of magnitude and phase, " in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process, 2016, pp. 5220–5224.
[28] D. S. Williamson, Y. Wang, and D. Wang, " Complex ratio masking for monaural speech separation, " IEEE/ACM Trans. Audio, Speech, Lang. Process, vol. 24, no. 3, pp. 483–492, Mar. 2016.
[29] A. V. D. Oord et al., "Wavenet: A generative model for raw audio," arXiv:1609.03499, 2016.
[30] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In ICML, 2010.
[31] 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.
[32] A. L. Maas, A. Y. Hannun, and A. Y. Ng, " Rectifier nonlinearities improve neural network acoustic models," in Proc. Int. Conf. Mach. Learn., 2013.
[33] D. Kingma and J. Ba, " Adam: A method for stochastic optimization," arXiv:1412.6980, 2014.
[34] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codes," in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2001, pp. 749–752.
[35] C.H.Taal,R.C.Hendriks,R.Heusdens,andJ.Jensen, "An algorithm for intelligibility prediction of time frequency weighted noisy speech,"IEEE Trans. Audio, Speech, Lang. Process, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.