A Deep Convolutional Encoder-Decoder Model for Robust Speech Dereverberation

D. S. Wang[1], *Y. X. Zou*[1]*, *W. Shi*[2]
[1]ADSPLab/ELIP/Shenzhen Key Laboratory for IMVR, Peking University Shenzhen Graduate School 518055, China
*{zouyx@pkusz.edu.cn}
[2]Hian Speech Science & Technology Co., Ltd, Shenzhen 518055, China

***Abstract*—Research shows that speech dereverberation (SD) with Deep Neural Network (DNN) achieves the state-of-the-art results by learning spectral mapping, which, simultaneously, lacks the characterization of the local temporal spectral structures (LTSS) of speech signal and calls for a large storage space that is impractical in real applications. Contrarily, the Convolutional Neural Network (CNN) offers a better modeling ability by considering local patterns and has less parameters with its weights sharing property, which motivates us to employ the CNN for SD task. In this paper, to our knowledge, a Deep Convolutional Encoder-Decoder (DCED) model is proposed for the first time in dealing with the SD task (DCED-SD), where the advantage of the DCED-SD model lies in its powerful LTSS modeling capability via convolutional encoder-decoder layers with smaller storage requirement. By taking the reverberant and anechoic spectrum as training pairs, the proposed DCED-SD is well-trained in a supervised manner with less convergence time. Additionally, the DCED-SD model size is 23 times smaller than the size of DNN-SD model with better performance achieved. By using the simulated and real-recorded data, extensive experiments have been conducted to demonstrate the superiority of DCED-based SD method over the DNN-based SD method under different unseen reverberant conditions.**

Keywords—speech dereverberation; deep convolutional encoder-decoder (DCED); spectral mapping; local temporal spectral structures (LTSS); storage space

I. INTRODUCTION

In an enclosed room, due to the sound waves reflecting off walls and other surfaces, the received signals of the microphone inevitably contain reverberation, which denotes the attenuated and delayed duplicates of the original speech signals [1]. The reverberation leads to the degradation of speech quality and intelligibility, which deteriorate the performance of a wide range of speech applications, such as speech recognition [2], speaker verification [3], speaker localization [4] and so on.

To mitigate the adverse impact of the reverberation, many speech dereverberation (SD) techniques have been developed. Traditional SD methods can be considered in two categories [5] the linear prediction (LP) residual processing [6, 7] and the blind channel estimation [8, 9]. In essence, the performance of these methods is largely dependent on the reliable estimation of the inverse filter of the room impulse response (RIR) to deconvolve the reverberant signals. However, the inverse filter is difficult to estimate since the RIR is time-varying. Besides, when the reverberation time becomes longer and the environment changes with distinctive reverberation, the traditional SD methods fail to keep good performance [5]. Recently, the deep learning has been successfully applied and achieved the state-of-the-art results in SD [10, 11], where the Deep Neural Network (DNN) has been employed as a regression model to learn the mapping from the reverberant speech spectral to its anechoic version. Although the performance of DNN-based SD methods precedes that of the traditional SD methods, they suffer from two limitations for further improving their performance: (1) the DNN is treated as a nonlinear transformation without considering the characterization of the local temporal spectral structures (LTSS) of speech signals; (2) the DNN adopted in the previous literatures [10, 11] contains millions of parameters that calls for a large storage space, which is impractical in real applications.

On the contrary, literature studies in image and speech domains show that the Convolutional Neural Network (CNN) is able to characterize the local patterns of speech spectral by using the convolutional kernel and has much less parameters with its weights sharing property [12]. Additionally, the CNN has already proved its better performance in the speech denoising tasks as compared with the DNN [13, 14]. But to the best of our knowledge, CNN has not yet been employed to deal with the SD task. Thus, to solve the above two deficiencies of DNN-based SD method, we attempt to utilize the CNN for the SD task for the first time, where the Deep Convolutional Encoder-Decoder (DCED) is a good candidate due to its powerful modeling capability and less parameters required. It is noted that the DCED structure has been successfully applied for the image segmentation [15], image denoising [16] and so on. Besides, the DCED has been employed for speech enhancement (DCED-SE) [14]. Their preliminary experimental results prove the superiority of DCED-based SE method over the DNN and Recurrent Neural Network (RNN) based SE methods.

Motivated by the work of DCED-SE model, in this study, we proposed a DCED-SD model composed of the symmetric convolutional encoding and decoding layers, which are the basic components of the DCED. However, we have modified the DCED which is able to better handle the SD task, that is, behind the decoder, one convolution layer is added and flattened to be fully connected (FC) with the final output layer, which essentially is a trade-off between the reduction of the number of parameters and the improvement of the efficiency of the model. Since the extra convolution layer reduces the dimension of the output of the decoder, and the FC layer is able to maintain the global characteristics for better estimation of the anechoic spectrum. By taking the reverberant and anechoic

spectrum of speech signals as training pairs under different reverberant conditions, the proposed DCED-SD model is shown to be well-trained with less convergence time compared to that of the DNN-SD model. Additionally, the size of DCED-SD model is 23 times smaller than that of the DNN-SD model, which makes the DCED-SD model promising to be applied in the embedded system, for example, the hearing aids [14]. Through extensive experiments under various unseen reverberant conditions, in comparison with the DNN-based SD method, the better performance has been achieved by our proposed DCED-based SD method in terms of the Perceptual Evaluation of Speech Quality (PESQ) [17] and Short-Time Objective Intelligibility (STOI) [18] measures.

II. Problem Formulation

In the reverberant environment, the signal captured by the microphone can be modeled as

$$y(t) = h(t) * x(t) \tag{1}$$

where $x(t)$ is the original speech signal, $h(t)$ is the RIR from the source to the microphone. To make the solution for SD task clearly interpretative, it is noted that in this paper, we only consider the reverberation excluding noise. For a 16kHz sampled signal, we take the 320-points short-time Fourier transform (STFT), then the STFT of $y(t)$ and $x(t)$ can be represented as

$$\mathbf{Y}(m) = \left[Y(1,m), Y(2,m), ..., Y(161,m)\right]^T \tag{2}$$

$$\mathbf{X}(m) = \left[X(1,m), X(2,m), ..., X(161,m)\right]^T \tag{3}$$

where m is the frame index, 161 is the total number of frequency bin, $\{Y(k, m),\ 1 \le k \le 161\}$ and $\{X(k, m),\ 1 \le k \le 161\}$ are the log spectral magnitudes (LSM) of $y(t)$ and $x(t)$ respectively. Given the reverberant spectral $\mathbf{Y}(m)$, the goal of the SD is to estimate the anechoic LSM vector $\mathbf{X}(m)$. Following the previous DNN-based SD methods [10, 11], the SD problem can be solved by the following mean square error (MSE) based optimization formulation

$$\min_{\boldsymbol{W}} \frac{1}{M} \sum_{m=1}^{M} \left\| \mathbf{X}(m) - f\left(\mathbf{S}(m) \middle| \boldsymbol{W}\right) \right\|_2^2 \tag{4}$$

where M is the total number of frames, $f(\cdot)$ denotes the nonlinear mapping between the reverberant spectrum and anechoic spectrum, $\boldsymbol{W}$ is the weight of DNN, and $\mathbf{S}(m)$ is a LSM matrix incorporating the n_T neighboring frames of the mth frame $\mathbf{Y}(m)$ as follows

$$\mathbf{S}(m) = \left[\mathbf{Y}(m - n_T), ..., \mathbf{Y}(m), ..., \mathbf{Y}(m + n_T)\right] \tag{5}$$

where n_T is determined to be 5 empirically in our work, which denotes that 11 frames of reverberant spectral are used. In [10, 11], $\mathbf{S}(m)$ is directly vectorised and used as the input of DNN. It is worthy to point out that the vectorised $\mathbf{S}(m)$ is unable to characterize the topology of the input [19], for example, the local temporal spectral structures (LTSS) of speech. Besides, the DNN is totally made up of FC layers, which leads to large storage space. Instead, in this paper, $\mathbf{S}(m)$ is viewed as an image and used as the input of CNN to fully explore the LTSS of speech for more abstract features representation, which is helpful for the better estimation of the anechoic spectrum and reducing the quantity of model parameters to a large extent.

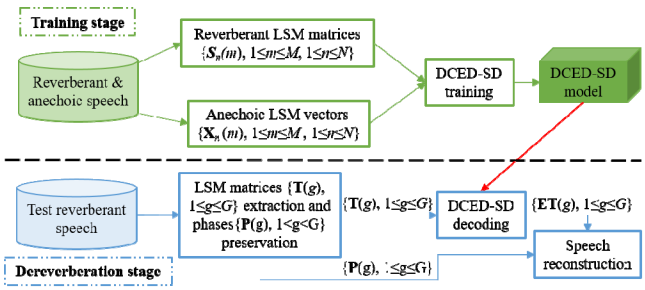

Figure. 1. The Flow Diagram of the DCED-based SD Method

III. DCED-based Seech Dereverberation

Among different types of CNN architecture, the DCED has demonstrated its powerful modeling capability together with much less parameters compared to other CNN structures, which manifests the significant potential of the DCED for solving the SD task. Thus, in this study, we firstly illustrate the DCED-based SD method in details. Then the specific design of the proposed DCED-SD model is introduced substantially.

A. Our proposed DCED-based SD Method

The flow diagram of our proposed DCED-based SD method is given in Fig. 1, which involves the training stage and the dereverberation stage, respectively.

In the training stage, a large set of training data formed under various reverberant conditions, consisting of pairs of reverberant and anechoic speech represented by LSM features, are constructed for the training of DCED-SD model, where the details of training data preparation can be found in Sect. IV. To improve the generalization capability of the DCED-SD model, we use the l_2 norm of the weights as the regularization term added to the MSE to form the final objective function, then the optimization problem (4) is converted into

$$\min_{\boldsymbol{W}} \frac{1}{N \sum_{n=1}^{N} M_n} \sum_{n=1}^{N} \sum_{m=1}^{M_n} \left\| \mathbf{X}_n(m) - f\left(\mathbf{S}_n(m) \middle| \boldsymbol{W}\right) \right\|_2^2 + \lambda \left\| \boldsymbol{W} \right\|_2^2 \tag{6}$$

where λ is the regularization coefficient, N is the total number of the reverberant/anechoic speech used for the construction of the training data, M_n is the total number of LSM matrices for the nth reverberant speech, $\mathbf{X}_n(m)$ is the mth LSM vector of the nth anechoic speech, and $\mathbf{S}_n(m)$ is the corresponding mth LSM matrix of the nth reverberant speech.

In the dereverberation stage, the LSM matrices $\{\mathbf{T}(g),\ 1 \le g \le G\}$ of the test reverberant speech are firstly extracted, where G is the total number of LSM matrices. Meanwhile, since our ears are insensitive to small phase distortions or global spectral shifts [20], the phase vectors $\{\mathbf{P}(g),\ 1 \le g \le G\}$ of the reverberant speech are preserved for the speech reconstruction. With the test LSM matrices as input, the estimated anechoic LSM vectors $\{\mathbf{ET}(g),\ 1 \le g \le G\}$ are decoded by the well-trained DCED-SD model. Finally, the anechoic speech can be resynthesized with $\{\mathbf{ET}(g),\ 1 \le g \le G\}$ and corresponding $\{\mathbf{P}(g),\ 1 \le g \le G\}$ via the inverse STFT. Obviously, the core of our proposed method lies in the DCED-SD model, which requires a proper architecture having the powerful modeling ability and small storage space.

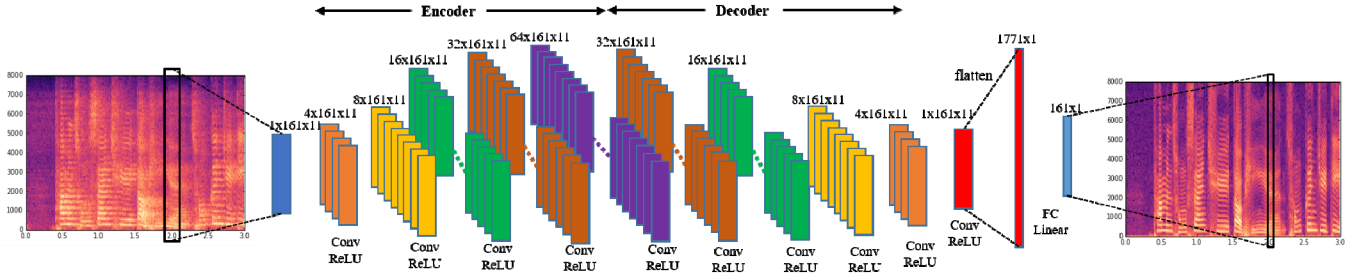

Figure. 2. Our Proposed Architecture of the DCED-SD model

B. *The Architecture of our proposed DCED-SD model*

Our proposed architecture of the DCED-SD model is shown in Fig. 2, where the DCED-SD model comprises the encoder, decoder, one convolution layer and one FC layer.

The encoder is repetitions of a convolution layer and a ReLU activation [21] layer, where the filters' number increases gradually. Correspondingly, the decoder is symmetric with gradually decreasing number of filters. Considering the width of the input spectral (e.g. 11 frames are set in our work) for SD task is small, in our design of the DCED model, the pooling layers are removed, since the pooling layers reduce the width of features quickly to be 1, which is unbeneficial for capturing the features along the time axis of the input spectral.

In addition, in order to simultaneously reduce parameters' quantity of the model and improve the modeling capability, we take a trade-off strategy in the DCED model design in the following two aspects: (1) To reduce the number of parameters, one convolution layer is added following the decoder aiming at reducing the dimension of the output of the decoder; (2) To maintain the global characteristics of spectral for better estimation of the anechoic spectrum, the output of the last convolution layer is flattened and connected with the final output (clean spectral) to form a FC layer.

IV. Experiments and Analysis

A. *Experimental Settings*

To build a large-scale training dataset for the DCED-SD model training, we randomly select 540 sentences from the THCHS-30 database [22] as the clean speech where each has a length of 9 seconds with a sampling rate of 16KHz. Then the reverberant speech is generated by convolving the clean speech with 45 different RIRs, which are simulated under different reverberant conditions that are illustrated in Table I. The RIRs are simulated by varying the reverberation time T_{60}, room size, distance and the direction-of-arrival (DOA) from the source to the microphone [23]. Thus, 23400 (540×45) reverberant sentences can be obtained, resulting in a collection of 60-hour training data, where we randomly selected 5-hour, 10-hour, 25-hour, and 50-hour training subsets for the experiments. The test data contains 200 sentences randomly chosen from the THCHS-30 database, three simulated RIRs (T_{60} is set to be 0.3s, 0.6s and 0.9s respectively) and three real-recorded RIRs from the MARDY database [24] that corresponds to three different distances (1m, 2m, 3m) between the microphone array and the center loudspeaker. It is noted that neither the sentences nor the RIRs are used in the training data.

Table I. Configurations used for RIR simulation

T_{60} (s)	0.2 to 1 with 0.2 step
Room size (m)	7×5×3 (small), 12×10×3 (medium), 17×15×3(large)
Position of the microphone	in the center of the room with the height to be 1.5m
Distance (m)	[1, 1.5, 2] for the small room; [1, 2, 4] for the medium room; [1, 3, 6.5] for the large room
Direction of arrival (°)	elevation and azimuth are randomly sampled from 0~180 and 0~360

The architecture of the DCED-SD model used in our experiment is illustrated in Fig 1, where the number of filters for each convolution layer is 4, 8, 16, 32, 64, 32, 16, 8, 4 and 1 respectively, and the size of all filters is fixed at (3, 3). Besides, the DNN-based SD method is used as the baseline, and the architecture of the DNN is the same as [10], which contains 1771(161×11)-dimensional input, 3 hidden layers with 1600 nodes per layer and 161-dimensional output. For a fair comparison, the l_2 norm regularization term is added to the objective function of both DCED and DNN, and the regularization coefficient is chosen to be 0.001 through validation experiments. In addition, an adaptive learning rate method, namely, adadelta [25] is applied for training the DCED-SD and DNN-SD model. To evaluate the performance of SD, the Perceptual Evaluation of Speech Quality (PESQ) and Short-Time Objective Intelligibility (STOI), reflecting a high correlation with subjective evaluation scores and a comparison between the envelopes of segregated speech and clean speech, are used as the metrics.

B. *Performance Comparison with Different Training Data Size*

In this section, experiments are conducted by using different training data size, where the simulated RIR with 0.6s reverberation time is chosen for the test. Fig. 3 presents the average PESQ and STOI performance comparison with different training data size. It can be seen that the performance of both DCED-based and DNN-based SD method goes up as the training data size increases, indicating that sufficient training data enables the models to obtain good generalization capability. Meanwhile, using the training data of same size, the DCED-based SD method outperforms the DNN-based SD method with higher average PESQ and STOI, which demonstrates the superiority of DCED-based method. In addition, we found that only 10 epochs are required for training

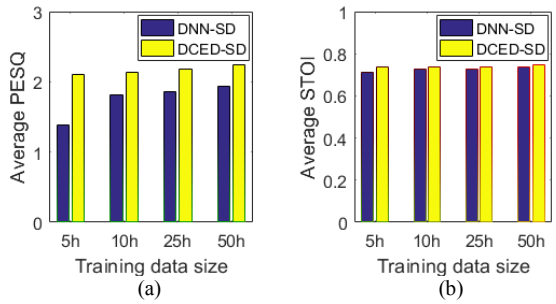

Figure. 3. (a) The average PESQ versus training data size; (b) The Average STOI versus training data size

Table II. Performance comparison with different T_{60}

T_{60} (s)	0.3		0.6		0.9	
Metrics	PESQ	STOI	PESQ	STOI	PESQ	STOI
Rev	2.37	0.75	1.96	0.67	1.77	0.57
DNN-50	2.23	0.74	1.99	0.74	1.87	0.69
DCED-50	**2.47**	**0.78**	**2.25**	**0.75**	**2.09**	**0.72**

Table III. Performance comparison with real-recorded RIR

Distance	1m		2m		3m	
Metrics	PESQ	STOI	PESQ	STOI	PESQ	STOI
Rev	**2.67**	**0.85**	2.21	0.67	2.09	0.58
DNN-50	1.85	0.74	1.80	0.72	1.77	**0.66**
DCED-50	2.41	0.78	**2.32**	**0.73**	**2.29**	**0.66**

DCED-SD model, which consumes less divergence time as compared to that of DNN-SD model well-trained with 50 epochs. Besides, the parameters' quantities of the DCED-SD model and DNN-SD model are 333637 and 8216161 respectively, as a result, the size of the DCED-SD model is 23 times smaller than that of the DNN-SD model, which makes the DCED-SD model more applicable in practice.

C. Performace Comparison with Different Simulated RIRs

To explore the influence of reverberation time on the performance of dereverberation, we use three different kinds of simulated RIRs with T_{60} set to be 0.3s, 0.6s and 0.9s. The DCED-SD model and DNN-SD model trained by the 50-hour data are tested, which are termed as DCED-50 and DNN-50 respectively for short. To show the performance improvement by DCED-50, the PESQ and STOI of the reverberant (Rev) speech are also recorded, which is shown in Table II. Compared to the reverberant speech, DCED-50 increases the quality of speech especially when the T_{60} is 0.9s, where PESQ and STOI improvements of 0.32 and 0.15 can be achieved. Besides, the DCED-50 outperforms the DNN-50 under all reverberant conditions. These results demonstrate that our proposed DCED-SD model has more powerful regression and generalization capability by characterizing the LTSS of speech.

D. Performace Comparison with real-recorded RIRs

To further investigate the generalization capability of our proposed DCED-SD model, three different real-recorded RIRs, which corresponds to three different distances (1m, 2m, 3m) between the microphone array and the center loudspeaker, are selected from the MARDY database to construct the test data. It is noted that the reverberation becomes severe as the distance increases, and performance results are illustrated in Table III. It can be seen that the PESQ and STOI scores are not boosted when the reverberation is weak (distance=1m), partly because mild reverberation does not lead to significant sound quality degradation [11]. When the reverberation becomes severe, the DCED-50 improves the quality of the reverberant speech and outperforms the DNN-50, which further shows the superiority of our proposed DCED-based SD method. Besides, one sentence is randomly selected from the test set to show the SD effect as shown in Fig 4, where the RIR corresponds to the distance of 3m. We can observe that the low and middle frequencies of the enhanced spectrogram via DNN-50 were blurred. Contrarily, as compared to the reverberant speech, the DCED-50 has a better restoration with 0.23 and 0.1 improvement of PESQ and STOI scores, respectively, which shows the efficiency and robustness of our proposed DCED-based SD method.

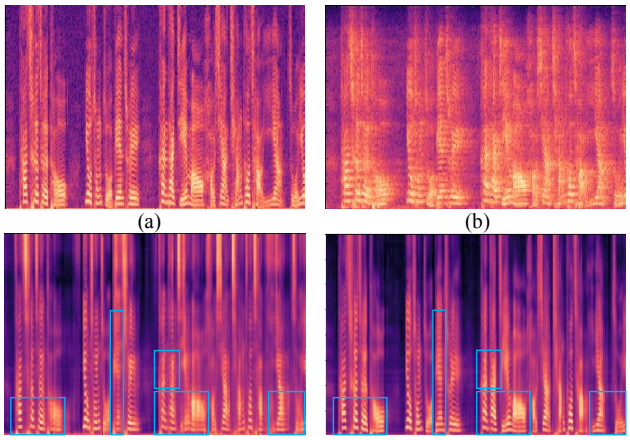

Figure. 4. Spectrogram of one randomly selected sentence from the test set, the real-recorded RIR corresponds to 3m between the microphone and the center loudspeaker: (a) anechoic speech (PESQ=4.5, STOI=1); (b) reverberant speech (PESQ=1.89, STOI=0.59); (c) dereverberation by DNN-50 (PESQ=1.84, STOI=0.63); (d) dereverberation by DCED-50 (PESQ=2.12, STOI=0.69)

V. CONCLUSION

In this paper, we firstly propose a new DCED-SD model for SD by learning a mapping from the LSM of reverberant speech to that of anechoic speech. As compared to the state-of-the-art DNN-based SD method, the proposed DCED-based SD method is able to utilize the LTSS of speech for more efficient modeling to yield a better performance under various simulated and real-recorded RIRs conditions, where higher PESQ and STOI scores, better restoration capability can be achieved. Meanwhile, because of the weights sharing property, the DCED-SD model has a much smaller size which makes the DCED-SD model promising for practical applications. Our future work focuses on the optimization of the DCED architecture for further improvements of SD and other speech applications.

ACKNOWLEDGMENT

This work is supported by Shenzhen basic research project entitled "Study on speech emotion recognition of children in natural environment with multi-modalities".

REFERENCES

[1] Lebart, Katia, Jean-Marc Boucher, and P. N. Denbigh. “A new method based on spectral subtraction for speech dereverberation.” Acta Acustica united with Acustica 87.3 (2001): 359-366.

[2] Pearson, J., et al. "Robust distant-talking speech recognition." Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on. Vol. 1. IEEE, 1996.

[3] Castellano, Pierre J., S. Sradharan, and David Cole. "Speaker recognition in reverberant enclosures." Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on. Vol. 1. IEEE, 1996.

[4] DiBiase, Joseph H., Harvey F. Silverman, and Michael S. Brandstein. "Robust localization in reverberant rooms." Microphone Arrays. Springer Berlin Heidelberg, 2001. 157-180.

[5] Naylor, Patrick, and Nikolay D. Gaubitch, eds. Speech dereverberation. Springer Science & Business Media, 2010.

[6] Brandstein, Michael S., and Scott M. Griebel. "Nonlinear, model-based microphone array speech enhancement." Acoustic signal processing for telecommunication. Springer US, 2000. 261-279.

[7] Yegnanarayana, Bayya, and P. Satyanarayana Murthy. "Enhancement of reverberant speech using LP residual signal." IEEE Transactions on Speech and Audio Processing 8.3 (2000): 267-281.

[8] Huang, Yiteng, Jacob Benesty, and Jingdong Chen. "Optimal step size of the adaptive multichannel LMS algorithm for blind SIMO identification." IEEE Signal Processing Letters 12.3 (2005): 173-176.

[9] Gannot, Sharon, and Marc Moonen. "Subspace methods for multimicrophone speech dereverberation." EURASIP Journal on Applied Signal Processing 2003 (2003): 1074-1090.

[10] Han, Kun, et al. "Learning spectral mapping for speech dereverberation and denoising." IEEE Transactions on Audio, Speech, and Language Processing 23.6 (2015): 982-992.

[11] Wu, Bo, et al. "A Reverberation-Time-Aware Approach to Speech Dereverberation Based on Deep Neural Networks." IEEE/ACM Transactions on Audio, Speech, and Language Processing 25.1 (2017): 102-111.

[12] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.

[13] Fu, Szu-Wei, Yu Tsao, and Xugang Lu. "SNR-Aware Convolutional Neural Network Modeling for Speech Enhancement." Proceedings of the Interspeech, San Francisco, CA, USA (2016): 8-12.

[14] Park, Se Rim, and Jinwon Lee. "A Fully Convolutional Neural Network for Speech Enhancement." arXiv preprint arXiv:1609.07132 (2016).

[15] Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." arXiv preprint arXiv:1511.00561 (2015).

[16] Mao, Xiao-Jiao, Chunhua Shen, and Yu-Bin Yang. "Image denoising using very deep fully convolutional encoder-decoder networks with symmetric skip connections." arXiv preprint (2016).

[17] Rix, Antony W., et al. "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs." Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on. Vol. 2. IEEE, 2001.

[18] Taal, Cees H., et al. "An algorithm for intelligibility prediction of time–frequency weighted noisy speech." IEEE Transactions on Audio, Speech, and Language Processing 19.7 (2011): 2125-2136.

[19] Sainath, Tara N., et al. "Deep convolutional neural networks for LVCSR." Acoustics, speech and signal processing (ICASSP), 2013 IEEE international conference on. IEEE, 2013.

[20] Xu, Yong, et al. "A regression approach to speech enhancement based on deep neural networks." IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 23.1 (2015): 7-19.

[21] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." Nature 521.7553 (2015): 436-444.

[22] Wang, Dong, and Xuewei Zhang. "THCHS-30: A free Chinese speech corpus." arXiv preprint arXiv:1512.01882 (2015).

[23] Habets, Emanuel AP. "Room impulse response generator." Technische Universiteit Eindhoven, Tech. Rep 2.2.4 (2006): 1.

[24] Wen, Jimi YC, et al. "Evaluation of speech dereverberation algorithms using the MARDY database." in Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC. 2006.

[25] Zeiler, Matthew D. "ADADELTA: an adaptive learning rate method." arXiv preprint arXiv:1212.5701 (2012).