A Recursive Network with Dynamic Attention for Monaural Speech Enhancement

Andong Li$^{1,3}$, Chengshi Zheng$^{1,3}$, Cunhang Fan$^{2,3}$, Renhua Peng$^{1,3}$, Xiaodong Li$^{1,3}$

1 Key Laboratory of Noise and Vibration Research, Institute of Acoustics, Chinese Academy of Sciences, Beijing, China
2 NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China
3 University of Chinese Academy of Sciences, Beijing, China

{liandong, cszheng, pengrenhua, lxd}@mail.ioa.ac.cn, cunhang.fan@nlpr.ia.ac.cn

Abstract
For continuous speech processing, dynamic attention is helpful in preferential processing, which has already been shown by the auditory dynamic attending theory. Accordingly, we propose a framework combining dynamic attention and recursive learning together for monaural speech enhancement. Apart from a major noise reduction network, we design a separated sub-network, which adaptively generates the attention distribution to control the information flow throughout the major network. Recursive learning is introduced to dynamically reduce the number of trainable parameters by reusing a network for multiple stages, where the intermediate output in each stage is corrected with a memory mechanism. By doing so, a more flexible and better estimation can be obtained. We conduct experiments on TIMIT corpus. Experimental results show that the proposed architecture obtains consistently better performance than recent state-of-the-art models in terms of both PESQ and STOI scores. The code is provided at https://github.com/Andong-Li-speech/DARCN.

Index Terms: monaural speech enhancement, recursive learning, attention U-Net, dynamic attention

1. Introduction
In real environments, clean speech is often contaminated by background interference, which may significantly reduce the performance of automatic speech recognition [1], speaker verification [2] and hearing aids [3]. Monaural speech enhancement aims to extract the target speech from the mixture when only one microphone is available [3]. In recent years, deep neural networks (DNNs) have shown their promising performance on monaural speech enhancement even in highly non-stationary noise environments owing to its superior capability in modeling complex nonlinearity [5]. Typical DNN-based methods can be categorized into two classes according to estimation targets, where one is masking-based [6] and the other is spectral mapping-based [7].

Conventional DNNs usually adopted fully-connected (FC) layers for noise reduction [6, 7]. To tackle the speaker generalization problem, Chen et al. proposed to utilize stacked long short-term memory (LSTM) [8], which significantly outperformed DNNs. Recently, various convolutional neural networks (CNNs) with complex topology were proposed [9, 10, 11, 12], which could reduce the number of trainable parameters. More recently, Tan et al. combined convolutional auto-encoder (CAE) [13] and LSTM together to propose convolutional recurrent neural network (CRN) [14], where CAE helped to learn temporal-frequency (T-F) patterns and LSTM effectively captured dynamic sequence correlations.

A variety of models with more complex topologies have been proposed recently [10, 11, 12, 14], which have shown improved performance, they still have some limitations for the following two folds. For one thing, the number of the parameters is often partly limited to meet the low-latency requirement, which heavily restricts the depth of the network. For another, the increase of depth is more likely to cause a gradient vanishing problem. Recently, progressive learning was proposed [15, 16], which decomposes the mapping process into multiple stages. Experimental results in [16] have shown that it dramatically decreases the number of trainable parameters and effectively maintains the performance. Based on this conception, recursive learning [17] was proposed by reusing the network for multiple stages, where the output in each stage is linked by a memory mechanism. It further alleviates the parameter burden and deepens the network without introducing extra parameters.

Human tends to generate adaptive attention with dynamic neuron circuits to percept complicated environments [18], which is also described by the auditory dynamic attending theory for continuous speech processing [19]. For example, when a person hears an utterance from the real environments, the more noise components are dominant, the more neuron attentions are needed to figure out the meaning and vice versa. This phenomenon reveals the dynamic mechanism of auditory perception system. Motivated by the physiological phenomenon, we propose a novel network combining dynamic attention and recursive learning together. Different from the previous networks [10, 11, 12, 14] that a single complex network is designed for the task, the framework encompasses a major network and an auxiliary sub-network in parallel, where the one is noise reduction module (NRM) and the other is attention generator module (AGM). The workflow of the framework is as follows: at each intermediate stage, both the noisy feature and the estimation from the last stage are combined into the current input.

AGM is adopted to generate the attention set, which is subsequently applied to NRM through the pointwise convolution and sigmoid function. In this way, AGM actually serves as a type of perception module to flexibly adjust the weight distribution of NRM, leading to better performance for noise suppression. To our best knowledge, it is the first time for a dynamic attention mechanism to be introduced for speech enhancement task.

The remainder of the paper is structured as follows. Section 2 formulates the problem. The architecture of the network is illustrated in Section 3. Section 4 is the dataset and experimental settings. Section 5 presents the results and analysis. Section 6 draws some conclusions.
2. Problem formulation and notation

In the time domain, a noisy signal can be modeled as \( x(n) = s(n) + d(n) \), where \( n \) is the discrete time index. With short-time Fourier transform (STFT), it can be further rewritten as:

\[
X_{k,l} = S_{k,l} + D_{k,l},
\]

where \( X_{k,l}, S_{k,l}, \) and \( D_{k,l} \) respectively refer to the noisy, clean, and noise components at the frequency bin index \( k \) and the time frame index \( l \). In this study, the network is deployed to estimate the magnitude of spectrum (MS), which is then incorporated with noisy phase to recover the estimated spectrum. Inverse short-time Fourier transform (iSTFT) is used to reconstruct the waveform in the time domain.

For simplicity of notation, we define the principal notations used in this paper. \( |X| \in \mathbb{R}^{T \times F}, |S| \in \mathbb{R}^{T \times F}, |\hat{S}| \in \mathbb{R}^{T \times F} \), and \( |\tilde{S}| \in \mathbb{R}^{T \times F} \) denote the magnitude of noisy spectrum, the magnitude of clean spectrum, the estimated magnitude of spectrum in the \( l \)th stage, and the final estimated magnitude of clean spectrum, respectively. \( T \) and \( F \) refer to the timestep and the feature length, respectively. As recursive learning is used, the superscript \( l \) denotes the stage index, and the number of stages is notated as \( Q \).

3. Architecture illustration

3.1. Stage recurrent neural network

Stage recurrent neural network (SRNN) is first proposed in [17], which is the core component in recursive learning. It is capable of aggregating the information across different stages with a memory mechanism, which is comprised of two parts, namely two-dimensional convolutional (2D Conv) block and convolutional-RNN (Conv-RNN) block. The first part tries to project the input features into a latent representation, followed by Conv-RNN to update the state in the current stage. Assuming the output of 2-D Conv and Conv-RNN at stage \( l \) are respectively notated as \( \hat{h}^l \) and \( h^l \), the inference of SRNN is formulated as:

\[
\hat{h}^l = f_{\text{conv}}^l \left( |X| \cdot |\hat{S}^{-1}| \right), \quad (2)
\]

\[
h^l = f_{\text{conv,rr}}^l \left( \hat{h}^l, h^{l-1} \right), \quad (3)
\]

where \( f_{\text{conv}} \) and \( f_{\text{conv,rr}} \) refer to the functions of 2-D Conv and Conv-RNN, respectively. \( h^{l-1} \) is the state of the last stage. In this study, ConvGRU [20] is adopted as the unit in Conv-RNN, whose calculation process gives as follows:

\[
z^l = \sigma \left( W_z^l \odot \hat{h}^l + U_z^l \odot h^{l-1} \right), \quad (4)
\]

\[
r^l = \sigma \left( W_r^l \odot \hat{h}^l + U_r^l \odot h^{l-1} \right), \quad (5)
\]

\[
n^l = \tanh \left( W_n^l \odot \hat{h}^l + U_n^l \odot r^l \odot h^{l-1} \right), \quad (6)
\]

\[
h^l = (1 - z^l) \odot \hat{h}^l + z^l \odot n^l, \quad (7)
\]

where \( W \) and \( U \) represent the weight matrices of the cell. \( \sigma(\cdot) \) and \( \tanh(\cdot) \) respectively, denote the sigmoid and the tanh activation functions. \( \odot \) represents the convolutional operator and \( \odot \) is the element-wise multiplication. Note that biases are ignored for notation convenience.

3.2. Attention gate

Attention U-Net (AU-Net) is first proposed in [21] to improve the accuracy in the segmentation-related tasks, where attention gates (AGs) are inserted between the convolutional encoder and the decoder. Compared with a standard U-Net, AU-Net has the capability of automatically suppressing the irrelevant regions and emphasizing the important features. As the spectrum includes abundant frequency components, where formants are often dominant in the low-frequency regions and the regions of the high-frequency have a sparse distribution, it is necessary to discriminate different spectral regions with different weights. The schematic of the AG adopted in this paper is shown in Figure 1. Assuming the inputs of the unit are \( p \) and \( q \), where \( p \) and \( q \) refer to the feature of a decoding layer and its corresponding feature in an encoding layer, respectively. \( W_{p}, W_{q} \) and \( W_{r} \) are 2-D convolutional layers, whose kernel size are \( 1 \times 1 \). \( C_p \) and \( C_q \) refer to the channel numbers of \( p \) and \( q \). Batch normalization is used after each convolutional operation.

![Figure 1](image-url)
As the network is trained for multiple stages, at each of which the emphasis is given to each training stage. We set \( \lambda_l \equiv 1 \), with \( l = 1, \ldots, Q \) in this study, i.e., the same emphasis is given to each training stage.

### 4. Experimental setup

#### 4.1. Dataset

The experiments are conducted on TIMIT corpus [27]. 4856, 800 and 100 clean utterances are selected for training, validation and testing, respectively. Training and validation dataset are created under the SNR levels ranging from -5dB to 10dB with the the interval 1dB whilst we test the model under the SNR conditions of (-5dB, 0dB, 5dB, 10dB). 130 types of noises used in [17] are for training and validation. Another 5 types of noises from NOISEX92 (babble, f16, factory2, m109 and white) are used to explore the generalization capacity of networks. All the collected noises are first concatenated into a long vector. During each mixed process, a random cutting point is generated, which is subsequently mixed with an utterance under an SNR level. As a result, we create 40,000, 4000, 800 noisy-clean pairs for training, validation and testing, respectively.

#### 4.2. Baselines

In this study, four networks are selected as the baselines, namely SLSTM [8], CRN [14], GRN [10] and DCN [12], all of which have achieved state-of-the-art performance recently. For SLSTM, four LSTM layers with 1024 units are stacked, followed by one FC layer to obtain the MS. The input of SLSTM includes the concatenation of the current frame and the previous ten frames. CRN is a type of real-time architecture combining CNN and LSTM. GRN and DCN are typical fully-convolutional network with gating mechanism.

#### 4.3. Parameter setup

All the utterances are sampled at 16kHz. The 20ms Hamming window is applied, with 10ms overlap in adjacent frames. 320-point STFT is adopted, leading to a 161-D feature vector in each frame. All the models are trained with mean-square error (MSE) criterion, optimized by Adam [28]. The learning rate is initialized at 0.001, we halve the learning rate when consecutive 3 validation loss increment happens and the training is early-stopped when 10 validation loss increment happens. All

---

**Figure 2:** The schematic of proposed architecture. (a) Proposed architecture and its unfolding structure. The architecture encompasses two parts in parallel, namely AGM and NRM. (b) The detailed structure of NRM. (c) The connection between AGM and NRM.
Table 1: Experimental results under seen noise cases. **BOLD** indicates the best result for each case. The number of stage Q = 3 for proposed architecture.

| Metric     | PESQ | STOI (in %) |
|------------|------|-------------|
| SNR        | -5dB | 0dB | 5dB | 10dB | Avg. | -5dB | 0dB | 5dB | 10dB | Avg. |
| Noisy      | 1.46 | 1.76 | 2.14 | 2.35 | 1.98 | 62.44 | 71.55 | 81.81 | 89.53 | 76.33 |
| SLSTM      | 2.35 | 2.67 | 2.98 | 3.21 | 2.80 | 80.16 | 86.64 | 91.24 | 94.18 | 88.06 |
| CRN        | 2.42 | 2.74 | 3.06 | 3.34 | 2.89 | 81.45 | 88.09 | 92.60 | 95.57 | 89.03 |
| GRN        | 2.47 | 2.76 | 3.04 | 3.32 | 2.88 | 82.94 | 88.63 | 92.10 | 95.03 | 88.79 |
| DCN        | 2.41 | 2.75 | 3.06 | 3.31 | 2.88 | 80.64 | 87.40 | 92.10 | 95.03 | 88.79 |
| Proposed (Q = 3) | **2.60** | **2.91** | **3.21** | **3.46** | **3.05** | **83.74** | **89.07** | **93.17** | **95.79** | **90.44** |

Table 2: Experimental results under unseen noise cases. **BOLD** indicates the best result for each case. The number of stage Q = 3 for proposed architecture.

| Metric     | PESQ | STOI (in %) |
|------------|------|-------------|
| SNR        | -5dB | 0dB | 5dB | 10dB | Avg. | -5dB | 0dB | 5dB | 10dB | Avg. |
| Noisy      | 1.47 | 1.83 | 2.13 | 2.51 | 1.98 | 60.15 | 72.51 | 81.23 | 89.77 | 75.92 |
| SLSTM      | 2.07 | 2.50 | 2.82 | 3.12 | 2.63 | 74.23 | 84.52 | 89.69 | 93.51 | 88.49 |
| CRN        | 2.20 | 2.60 | 2.92 | 3.21 | 2.73 | 77.10 | 86.19 | 91.03 | 94.79 | 87.28 |
| GRN        | 2.29 | 2.63 | 2.89 | 3.16 | 2.74 | 78.10 | 86.92 | 91.28 | 94.78 | 87.77 |
| DCN        | 2.20 | 2.62 | 2.94 | 3.22 | 2.74 | 76.05 | 85.88 | 90.73 | 94.48 | 86.78 |
| Proposed (Q = 3) | **2.37** | **2.75** | **3.06** | **3.34** | **2.88** | **79.16** | **87.54** | **91.61** | **95.02** | **88.33** |

5. Results and analysis

This section evaluates the performance of different models with perceptual evaluation speech quality (PESQ) [29] and short-time objective intelligibility (STOI) scores [30].

5.1. Objective results

Tables 1 and 2 summarize the results of different models for seen and unseen noise cases, respectively. From the two tables, one can observe the following phenomena. First, CRN, GRN, DCN, and the proposed model consistently outperform SLSTM in both seen and unseen noise cases. This is because SLSTM solely considers the sequence correlations but neglecting the implicit T-F patterns, which is crucial for speech recovery. Moreover, stacked LSTM tends to cause an attenuation effect due to the gradient vanishing problem, which limits the performance. Second, compared with the baselines, the proposed architecture obtains notable improvements in both metrics. For example, when going from CRN to the proposed model, PESQ is improved by 0.16 and STOI is improved by 1.01% on average for seen cases. A similar trend is also observed for unseen cases, indicating that the proposed model has a good noise generalization capability. Third, we observe that GRN and DCN can achieve close performance. This can be explained as both networks have similar topology, where dilation convolutions combined with the gating mechanism are applied for sequence modeling.

5.2. Impact of stages

We study the impact of the number of stages Q, which is given in Figure 3. One can get that with the increase of Q, both values of PESQ and STOI are consistently improved when Q ≤ 3. This indicates that SRNN can effectively refine the performance of the network by a memory mechanism. We also find that when Q increases from 3 to 5, PESQ value slightly degrades whilst STOI is still improved. This is because the distance-based loss like MSE is utilized, the loss function and the optimization process cannot guarantee consistent optimization for both metrics, which is consistent with the previous study in [17].

5.3. Parameter comparison

Table 3 summarizes the number of trainable parameters for different models. The unit is million. **BOLD** denotes the lowest trainable parameters.

| Model   | SLSTM | CRN  | GRN  | DCN  | Proposed |
|---------|-------|------|------|------|----------|
| Para. (million) | 36.81 | 17.58 | 3.13 | 2.92 | 1.23 |

Table 3: The number of trainable parameters among different models. The unit is million. **BOLD** denotes the lowest trainable parameters.

6. Conclusions

In a complicated scenario, a person usually dynamically adjusts the attention to the change of the environments for continuous speech. Based on this neural phenomenon, we propose a framework combining dynamic attention and recursive learning. To adaptively control the information flow of the noise reduction network, a separate sub-network is designed to update the attention representation in each stage and is subsequently applied to the major network. As a recursive paradigm is taken for training, the network is reused for multiple stages. As a result, we achieve a refined estimation stage by stage. Experimental results show that, compared with previous state-of-the-art strong models, the proposed model achieves consistently better performance while further decreasing the parameter burden.
7. References

[1] A. Graves, A. R. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in 2013 IEEE international conference on acoustics, speech and signal processing. IEEE, 2013, pp. 6645–6649.

[2] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted gaussian mixture models,” Digital signal processing, vol. 10, no. 1-3, pp. 19–41, 2000.

[3] H. Dillon, Hearing aids. Hodder Arnold, 2008.

[4] P. C. Loizou, Speech enhancement: theory and practice. CRC press, 2013.

[5] D. Wang and J. Chen, “Supervised speech separation based on deep learning: An overview,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 10, pp. 1702–1726, 2018.

[6] Y. Wang, A. Narayanan, and D. Wang, “On training targets for supervised speech separation,” IEEE/ACM transactions on audio, speech, and language processing, vol. 22, no. 12, pp. 1849–1858, 2014.

[7] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “A regression approach to speech enhancement based on deep neural networks,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 1, pp. 7–19, 2014.

[8] J. Chen, Y. Wang, S. E. Yoho, D. Wang, and E. W. Healy, “Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noise,” The Journal of the Acoustical Society of America, vol. 139, no. 5, pp. 2604–2612, 2016.

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

[10] K. Tan, J. Chen, and D. Wang, “Gated residual networks with dilated convolutions for monaural speech enhancement,” IEEE/ACM transactions on audio, speech, and language processing, vol. 27, no. 1, pp. 189–198, 2018.

[11] A. Pandey and D. Wang, “TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain,” in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 6875–6879.

[12] S. Pirhosseinloo and J. S. Brumberg, “Monaural speech enhancement with dilated convolutions,” Proc. Interspeech 2019, pp. 3143–3147, 2019.

[13] V. Badrinarayanan, A. Handa, and R. Cipolla, “Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling,” arXiv preprint arXiv:1505.07293, 2015.

[14] K. Tan and D. Wang, “A convolutional recurrent neural network for real-time speech enhancement,” in Interspeech, 2018, pp. 3229–3233.

[15] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, “SNR-Based Progressive Learning of Deep Neural Network for Speech Enhancement,” in INTERSPEECH, 2016, pp. 3713–3717.

[16] A. Li, C. Zheng, and X. Li, “Convolutional Recurrent Neural Network Based Progressive Learning for Monaural Speech Enhancement,” arXiv preprint arXiv:1908.10768, 2019.

[17] A. Li, C. Zheng, R. Peng, and X. Li, “A Time-domain Monaural Speech Enhancement with Recursive Learning,” arXiv preprint arXiv:2003.09815, 2020.

[18] S. Anderson, T. White-Schwoch, A. Parbery-Clark, and N. Kraus, “A dynamic auditory-cognitive system supports speech-in-noise perception in older adults,” Hearing research, vol. 300, pp. 18–32, 2013.

[19] M. R. Jones, “Time, our lost dimension: toward a new theory of perception, attention, and memory,” Psychological review, vol. 83, no. 5, p. 323, 1976.

[20] N. Ballas, L. Yao, C. Pal, and A. Courville, “Delving deeper into convolutional networks for learning video representations,” arXiv preprint arXiv:1511.06432, 2015.

[21] O. Oktay, J. Schlemper, L. L. Folgo, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz et al., “Attention u-net: Learning where to look for the pancreas,” arXiv preprint arXiv:1804.03999, 2018.

[22] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, pp. 234–241.

[23] S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” in International Conference on Machine Learning, 2015, pp. 448–456.

[24] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, “Fast and accurate deep network learning by exponential linear units (elus),” arXiv preprint arXiv:1511.07289, 2015.

[25] H. Noh, S. Hong, and B. Han, “Learning deconvolution network for semantic segmentation,” in Proceedings of the IEEE international conference on computer vision. 2015, pp. 1520–1528.

[26] H. Zheng, Z. Yang, W. Liu, J. Liang, and Y. Li, “Improving deep neural networks using softplus units,” in 2015 International Joint Conference on Neural Networks (IJCNN). IEEE, 2015, pp. 1–4.

[27] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, “DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1.” NASA STIReport n, vol. 93, 1993.

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

[29] J.-T. Recommendations, “Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codices,” Rec. ITU-T P. 862, 2001.

[30] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “A short-time objective intelligibility measure for time-frequency weighted noisy speech,” in 2010 IEEE international conference on acoustics, speech and signal processing. IEEE, 2010, pp. 4214–4217.