INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Multi-stage Progressive Speech Enhancement Network

Xinmeng Xu[1,2], *Yang Wang*[1], *Dongxiang Xu*[1], *Yiyuan Peng*[1], *Cong Zhang*[1], *Jie Jia*[1], *Binbin Chen*[1]

[1]vivo AI Lab, P.R. China
[2]E.E. Engineering, Trinity College Dublin, Ireland

xux3@tcd.ie, {yang.wang.rj,dongxiang.xu,pengyiyuan,zhangcong,jie.jia,bb.chen}@vivo.com

Abstract

Speech enhancement is a fundamental way to separate and generate clean speech from adverse environment where the received speech is seriously corrupted by noise. This paper applies a novel progressive network for speech enhancement by using multi-stage structure, where each stage contains a channel attention block followed by dilated encoder-decoder convolutional network with gated linear units. In addition, each stage generates a prediction that is refined by a supervised attention block. What is more, a fusion block is inserted between original inputs and outputs of previous stage. Multi-stage architecture is introduced to sequentially invoke multiple deep-learning networks, and its key ingredient is the information exchange between different stages. Thus, a more flexible and robust outputs can be generated. 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.

Index Terms: speech enhancement, encoder-decoder convolutional network, channel attention, supervised attention, cross-stage feature fusion, multi-stage progressive network

1. Introduction

Speech enhancement aims at improving speech intelligibility and quality of a speech signal by removing or attenuating background noise, such as ambient noise, interfering speech, and room reverberation. It is used as preprocessor in speech communication applications to improve their performance in noisy environments. Monaural speech enhancement provides a versatile and cost-effective approach to the problem by utilizing recordings from a single microphone.

Traditional monaural speech enhancement approaches can be grouped into Spectral Subtraction [1], Minimum Mean-square Error (MMSE) estimation [2], Wiener filter (linear MMSE) [3], Kalman filter [4], and subspace methods [5]. In the recent years, Deep Neural Networks (DNNs) have shown their promising performance on monaural speech enhancement even in highly non-stationary noise environments owing to their superior capability in modeling complex non-linearity [6]. DNN-based approaches to speech enhancement generally convert speech signal to Time-frequency (T-F) representation, and extract input features and training targets from it [6]. In addition, training targets are either masking based [7] or mapping based [8]. Masking based methods usually performed better than mapping based methods in terms of speech quality metrics [8, 9, 10].

Various complex models for speech enhancement are recently proposed [11, 12, 13], which evidently improved robustness and effectiveness of speech enhancement. However, the shortcomings of models with complex topologies are also obvious, 1) numerous parameters severely restricted the depth of networks, which brings high cost and increases the difficulty of network training, and 2) the complex structures are more likely caused the problems such as vanishing gradient and overfitting [14]. Recently, multi-stage learning has been proposed [15, 16, 17, 18], which decomposes training process into multiple stages that sequentially apply same or a combination of different models. Based on the effect of multi-stage architecture that the model used in given stage utilizes the outputs from previous stages as input and refines these outputs using attention mechanisms [19], it decreases the number of trainable parameters dramatically and maintains its performance effectively.

Accordingly, being different from the previous networks [11, 12, 13] that a single complex network is designed for task, this paper proposed a novel multi-stage progressive speech enhancement system, with several key components, 1) Each stage comprises a channel attention (CA) block [20] followed by dilated encoder-decoder convolutional networks with Gated Linear Units (GLUs) [11]. 2) A Supervised Attention Module (SAM) is inserted between every two stages to enable progressive learning, which exploits the previous stage prediction to compute attention maps that are in turn used to refine the previous stage predictions before passing them to the next stage. 3) A Cross-stage Feature Fusion (CSFF) mechanism is set for flowing feature information from earlier to later stages, which stabilizes the multi-stage network optimization.[1]

The main contributions of this paper can be summarized as follows:

- Applying multi-stage architecture for speech enhancement. The proposed model allocates the speech enhancement task to multiple sub-networks to progressively reduce noise for noisy speech.
- Applying SAM and CSFF mechanisms to scale the processed speech features from earlier stages. These mechanisms re-inject original speech information and mitigate possible speech information loss in earlier stages.
- Applying the dilated encoder-decoder convolutional networks with GLUs which are adept in the generalization capability for untrained speakers at different SNR levels[21].

The rest of this paper is organized as follows: In Section 2, the proposed method is presents in detail. Section 3 is the dataset and experimental settings. Section 4 demonstrates the results and analysis, and a conclusion is shown in Section 5.

2. Model Architecture

2.1. Multi-stage progressive network

The proposed multi-stage progressive network consists of K stages, where Figure 1 illustrates a three-stage progressive net-

[1] Speech samples are available at: https://xinmengxu.github.io/AVSE/multiStagePN.html

http://dx.doi.org/10.21437/Interspeech.2021-520

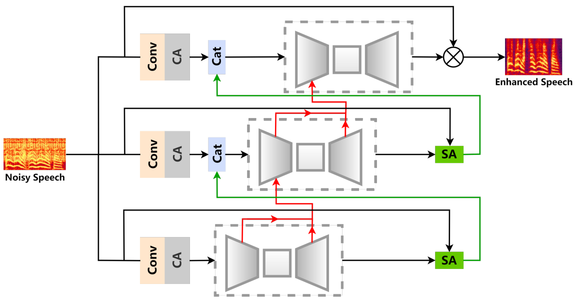

Figure 1: *Schematic diagram of three-stage architecture for multi-stage progressive speech enhancement. The Conv and Cat represent the convolution layer and concatenate layer, respectively. CA denotes the channel attention block, and its structure is shown in Figure 2(d). SA is supervised attention block that diagram block is shown in Figure 2(c). Red arrows represent the cross-stage feature fusion and its diagram block is shown in Figure 2(b). In addition, the dilated encoder-decoder convolutional networks with GLUs (see Figure 2(a)) framed by gray dotted box.*

work. Each stage has parallel training process, which is accessible to the original spectrogram, and comprises a CA block followed by dilated encoder-decoder convolutional networks with GLUs.

The proposed model inserted a supervised attention block between adjacent stages rather than simply cascading multiple stages, which regulates the feature maps, filters out irrelevant information, and rescales the feature maps of previous stages before passing them to the next stage. For K-stage progressive networks where $K \geqslant 3$, a cross-stage feature fusion block enables intermediate features of stage $k-1$ flowing to stage k, where $3 \leqslant k \leqslant K$.

The multi-stage architecture iteratively refines the initial predictions, which caused that the prediction of previous stage may include errors, such as information lost. To overcome this problem, supervised attention block integrates the combination between prediction from stage $k-1$ and the original input for re-injecting the original speech information to stage $k-1$, and cross-stage feature fusion block transfers the intermediate features of stage $k-1$ to stage k for consolidating the intermediate features of stage k.

As the proposed model is trained for multiple stages, each stage generates an intermediate estimation. At any given stage k, where $3 \leqslant k \leqslant K$, the accumulated loss function can be defined as:

$$\mathcal{L} = \sum_{k=1}^{K} \lambda_k \mathcal{L}^{(k)} \tag{1}$$

where λ_k is the weight coefficient for stage k, $\mathcal{L}^{(k)}$ is the loss function for the k^{th} stage, which can be expressed as:

$$\mathcal{L}^{(k)} = \left\| M^k \odot X - S \right\|_2 \tag{2}$$

where M^k is predicted mask in k^{th} stage, X and S denote the Short-time Fourier transform (STFT) magnitude of noisy speech and clean speech. In addition, the operator $\odot$ represents the Hadamard product.

2.2. Component details

The proposed model consists of four components, which are encoder-decoder subnetwork, cross-stage feature fusion block, supervised attention module, and channel attention module, respectively. The block diagram of these components are shown in Figure 2.

2.2.1. *Encoder-decoder subnetwork*

The proposed model uses encoder-decoder structure that based on the standard U-Net [22], shown in Figure 2(a), for subnetwork in each stage. The encoder contains several stacked 2-D convolutional layers, each of which is followed by batch normalization [23] and exponential linear unit (ELU) [24]. The dilation is applied to the layers along the frequency direction. The bottleneck consists of several GLUs [25] to perform time-dilated convolutions. The decoder is the mirror representations of the encoder except all the convolution layers are replaced to deconvolution layers. Skip connections are introduced to compensate for information loss during the encoding process.

2.2.2. *Cross-stage feature fusion*

A multi-stage progressive speech enhancement system with three or more stages ($K \geqslant 3$) is constructed by inserting a cross-stage feature fusion block (the schematic is shown in Figure 2(b)) between two stages. Note that each input feature from stage k is passed through a 1×1-convolution and a ReLU operation, after which a batch normalization is performed. After that, two outputs of the encoder branch and decoder branch are added, and its result is again sent through a 1×1-convolution block.

The cross-stage feature fusion block makes our model less vulnerable to the information loss due to encoder-decoder architecture, enriches the features of stage $k+1$, and stabilizes the network optimization procedure.

2.2.3. *Supervised attention module*

The schematic diagram of supervised attention module is shown in Figure 2(c). The supervised attention module is set between every two stages, which provides the original input to re-inject the original features, and which generates attention maps to suppress informative features at current stage and only allow useful ones to flow to the next stage for further processing.

The output features of stage $k-1$ is the input of supervised attention module, and is fed to a simple 1×1-convolution block to generate a residual feature map. The residual feature map is then added to the original feature map, and its result is then processed by a 1×1-convolution block followed by sigmoid activation to generate an attention mask which are used to re-calibrate the transformed input processed by a 1×1-convolution, resulting in attention-guided features. These attention-guided features are multiplied with residual feature map, and then added to input of supervised attention module. The augmented feature is the output of supervised attention module, and is passed to the stage k.

2.2.4. *Channel attention module*

Every stage of proposed model are consist of a channel attention block (see Figure 2(d)) that takes transformed input (obtained after a convolutional layer) as incoming feature, and the channel attention block uses three 1×1-convolution blocks to form the query and key-value pair. The proposed model first computes

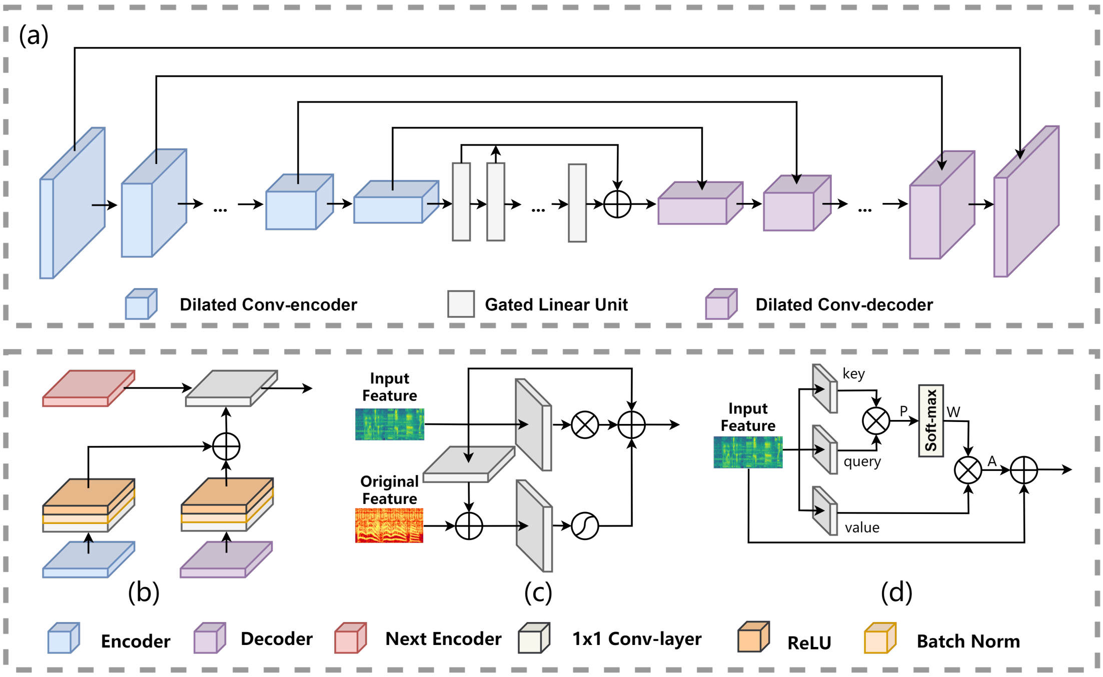

Figure 2: *(a) Block diagram of dilated encoder-decoder convolutional networks with GLUs. (b) Cross-stage feature fusion between two stages. (c) Supervised attention module. (d) Channel attention module, it computes the attention mask* **W** *and apply it to value,* **A** *is the attention component.*

the weight **P** for getting the attention component **A**, given by

$$\mathbf{P} = \frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{F}} \tag{3}$$

where **Q**, **K** are represented query and key respectively, and F denotes the number of frequency bins. The soft-max function (.) is used to generate attention mask $\widehat{\mathbf{W}} = \{\widehat{W}_{i,j}\}$,

$$\widehat{W}_{i,j} = \frac{exp(W_{i,j})}{w_j} \tag{4}$$

where

$$w_j = \sum_{i=1}^{F} exp(W_{i,j}) \tag{5}$$

The attention component **A** is determined by

$$\mathbf{A} = \widehat{\mathbf{W}}\mathbf{V} \tag{6}$$

where **V** denotes the index of value. Consequently, the channel attention block output **Y** is

$$\mathbf{Y} = \mathbf{X} + \delta\mathbf{A} \tag{7}$$

where **X** is the input of channel attention block, and δ is a scale with initial value zero which used to allow the network to first rely on the cue in the local channel **X** and then gradually assign more weight to the non-local channels using back-propagation to reach its optimal value [26].

3. Experiment setup

3.1. Datasets

In order to evaluate performance of the proposed model, experiments are conducted on TIMIT database [27]. There 4000 clean utterances are selected for training set and 800 are selected for validation set, which are created under the SNR levels ranging from -5dB to 10dB. Test set contains 100 utterances under the SNR condition of -5dB, 0dB, 5dB and 10dB.

Noise signals are from self-made dataset which collected from real world environment and categorized into 12 types: room, car, instrument, engine, train, human speech, air-brake, raining, street, mic-noise, ring-bell, and music.

3.2. Training and network parameters

The sampling rate of all the utterances is 16 kHz. The proposed model uses STFT to extract the spectrum from each utterance. A Hamming window with 512 bins and overlap interval of 256 bins are used. The model is optimized by Adam [28]. The initial learning rate is set to 0.001, and is decreased by 50% once when consecutive 3 validation stagnates, and the training is early-stopped when 10 consecutive validation loss increments happened.

3.3. Baselines

In order to evaluate the performance of proposed model, several baseline systems are set up for comparison. Baseline systems are describes briefly below.

- **CRN:** A convolutional recurrent network based speech enhancement system [29]. The encoder of CRN consists of five 2D convolution layers with {16, 32, 64, 128, 256} output channels. The hidden LSTM units are 256, and a Dense layer with 1280 units is after the last LSTM, and five 2D deconvolution layers with output channels {128, 64, 32, 16, 1}.
- **GRN:** A gated residual networks with dilated convo-

Table 1: *Performance of baselines and proposed model.* ***BOLD*** *indicates the best result for each column.*

Metric		PESQ					STOI (in %)				
SNR		-5dB	0dB	5dB	10dB	Avg.	-5dB	0dB	5dB	10dB	Avg.
Noisy speech		1.28	1.63	2.04	2.51	1.87	61.21	70.85	82.13	88.24	75.61
CRN		2.08	2.24	2.79	2.99	2.53	76.32	81.43	88.35	90.47	84.14
GRN		2.14	2.62	2.93	3.27	2.74	77.14	86.98	91.32	94.71	87.53
DACRN(3-stage)		2.55	2.84	3.08	3.35	2.96	81.45	88.09	92.76	94.18	89.12
Proposed	1-stage	2.21	2.67	2.97	3.04	2.72	77.10	84.52	89.69	91.32	86.55
	2-stage	2.48	2.79	3.01	3.38	2.92	80.64	87.60	91.28	94.42	88.49
	3-stage	2.67	**2.94**	3.28	3.51	3.10	**82.95**	88.76	93.20	95.14	90.01
	4-stage	2.60	2.88	**3.30**	**3.52**	3.08	82.78	88.94	92.99	95.20	89.98
	5-stage	**2.71**	2.90	3.29	3.50	**3.12**	82.94	**89.13**	**93.00**	**95.36**	**90.11**

lutions based speech enhancement system [11]. The GRN consists of a frequency-dilated module that contains four 2D convolutional layers with dilation rates $\{1, 1, 2, 4\}$ and with output channels $\{16, 16, 32, 32\}$, and a time-dilated module which is a set of gated linear units with dilation rates, and two succeeding groups repeat the same pattern $\{1, 2, 4, 8, 16, 32\}$.

- **DACRN:** A recursive network for speech enhancement [30] which apply multi-stages and the combination of dynamic attention and recursive learning for speech enhancement.

4. Results and analysis

The speech enhancement systems are evaluated using the following evaluation metrics: Perceptual Evaluation Speech Quality (PESQ) [31] and Short Term Objective Intelligibility (STOI) [32].

4.1. Objective results

Table 1 illustrates that the proposed model produces state-of-the-art results in terms of PESQ and STOI scores as discussed above by comparing against three mentioned speech enhancement systems. According to the table, one can observe the following phenomena:

1) The GRN performed better than CRN. Both models have a common property is that they are work well when exposed to a large number of speakers, because LSTM and GLU layer can leverage longer term information. However, the GLUs show more robust generalization capability for unseen speakers at different SNR levels when comparing between GRN and CRN.
2) The 3-stage DACRN outperforms GRN, although both models have similar structure that are applied GLUs as components. This can be explained that multi-stage architecture performs better than single network architecture.
3) The proposed model with 3 stages has better performance than DACRN, which indicating that the proposed model has better noise generalization capability than the recent multi-stage architecture based speech enhancement system.

In addition, Table 1 also demonstrated that the number of stage K influences the performance of speech enhancement system.

Table 2: *Per-stage PESQ and STOI scores for 7-stage DACRN and proposed model.* ***BOLD*** *indicates the best result of each column.*

	DARCN		Proposed	
Stages	PESQ	STOI	PESQ	STOI
1	2.70	86.68	2.72	86.55
2	2.90	88.45	2.92	88.49
3	**2.96**	89.12	3.10	90.01
4	2.93	89.10	3.08	89.98
5	2.94	**89.63**	3.10	90.11
6	2.85	88.94	**3.12**	**90.23**
7	2.81	89.32	3.11	90.08

4.2. Impact of stages

Table 2 demonstrates the PESQ and STOI scores for 7-stage DACRN and proposed model, and the following phenomena can be observed:

- Both values of PESQ and STOI scores are improved when stage K is from 1 to 3.
- When $3 \leqslant K \leqslant 7$, the PESQ score of DACRN is consistently decreased whilst the PESQ value of proposed model still improved but its score's rate of improvement is gradually decreased.

This observation reveals that it limits the performance of multi-stage models by adding more and more stages, and the fluctuation of performance measure is negligible when $K \geqslant 3$. This saturation phenomenon is in line with our anticipation, since in principle multi-stage model could be useful with an upper bound for performance improvement.

5. Conclusions

This paper proposed a multi-stage progressive monaural speech enhancement network. The model adopts multi-stage architecture, which applies dilated encoder-decoder convolutional networks with GLUs as branch of each stage, and SAM, CSFF mechanisms as connection blocks between every two stages. According to the experiment results, it improves performance of speech enhancement systems by involving multi-stage architecture, and the proposed connection blocks is a key component of the multi-stage model. The proposed model performs better quality of enhanced speech than the state-of-the-art models.

6. References

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

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

[4] K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," in *ICASSP '87. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 12, 1987, pp. 177–180.

[5] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 251–266, 1995.

[6] 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.

[7] 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.

[8] F. Bao and W. H. Abdulla, "A new ratio mask representation for CASA-based speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 7–19, 2018.

[9] Z. Chen, Y. Huang, J. Li, and Y. Gong, "Improving mask learning based speech enhancement system with restoration layers and residual connection." in *INTERSPEECH*, 2017, pp. 3632–3636.

[10] A. Narayanan and D. Wang, "Investigation of speech separation as a front-end for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 826–835, 2014.

[11] 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.

[12] 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.

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

[14] X. Wang, Y. Qin, Y. Wang, S. Xiang, and H. Chen, "Reltanh: An activation function with vanishing gradient resistance for SAE-based DNNs and its application to rotating machinery fault diagnosis," *Neurocomputing*, vol. 363, pp. 88–98, 2019.

[15] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European conference on computer vision*. Springer, 2016, pp. 483–499.

[16] 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.

[17] Y. A. Farha and J. Gall, "MS-TCN: Multi-stage temporal convolutional network for action segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3575–3584.

[18] X. Hao, X. Su, S. Wen, Z. Wang, Y. Pan, F. Bao, and W. Chen, "Masking and inpainting: A two-stage speech enhancement approach for low SNR and non-stationary noise," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6959–6963.

[19] X. Zeng, W. Ouyang, and X. Wang, "Multi-stage contextual deep learning for pedestrian detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 121–128.

[20] B. Tolooshams, R. Giri, A. H. Song, U. Isik, and A. Krishnaswamy, "Channel-attention dense u-net for multichannel speech enhancement," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 836–840.

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

[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*. PMLR, 2015, pp. 448–456.

[24] D. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[25] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *International conference on machine learning*. PMLR, 2017, pp. 933–941.

[26] J. Lin, A. J. van Wijngaarden, K.-C. Wang, and M. C. Smith, "Speech enhancement using multi-stage self-attentive temporal convolutional networks," *arXiv preprint arXiv:2102.12078*, 2021.

[27] 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, p. 27403, 1993.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015*, Y. Bengio and Y. LeCun, Eds., 2015.

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

[30] A. Li, C. Zheng, C. Fan, R. Peng, and X. Li, "A recursive network with dynamic attention for monaural speech enhancement," *arXiv preprint arXiv:2003.12973*, 2020.

[31] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 2, May 2001, pp. 749–752 vol.2.

[32] 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.