DCCRN-SUBNET: A DCCRN and SUBNET Fusion Model for Speech Enhancement

Xin yuan, Qun Yang* and Shaohan Liu
Nanjing University of Aeronautics and Astronautics
Nanjing, China
yuanxin_n@126.com {qun.yang, liushaohan}@nuaa.edu.com

2021 7th International Conference on Computer and Communications (ICCC) | 978-1-6654-0950-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCC54389.2021.9674517

Abstract—**Currently, most of the speech enhancement methods can't address the performance degradation problem caused by low signal-to-noise ratios (SNR) and non-stationary noises. For better speech enhancement at the above scenarios, this paper proposes a two-stage method that fuses DCCRN and SubNet. Compared with the single stage-stage networks, two-stage networks have more powerful mapping capabilities. This paper uses complex-valued spectrogram as the training target. In the first stage, the DCCRN takes the magnitude and phase as input and estimates corresponding target of clean speech. By simulating the complex-valued operation, the DCCRN can train the complex target effectively. However, it is still difficult to handle the low SNR and non-stationary noises. This paper uses the SubNet as the second stage network for better speech enhancement. In the second stage, the SubNet further refines the magnitude of target frequency by exploiting the context frequencies. Its input is consisted of magnitude of target frequency and several context frequencies. The output is the estimation of the clean speech magnitude target for the corresponding frequency. The experimental results show that the proposed method obtains better performance than other baseline models in terms of PESQ, STOI and SI-SDR.**

Keywords—speech enhancement, two-stage, complex network, sub-bank

I. INTRODUCTION

Speech enhancement aims to eliminate the noise in the original speech signal and get higher quality output. Compared with traditional methods, the method based on deep learning has achieved better performance. However, how to deal with non-stationary and low level of the signal-to-noise ratios (SNR) noise is still a challenge.

Formulated as a supervised learning problem, noisy speech can be enhanced by neural networks either in time-frequency (TF) domain or in time-domain. In time-domain, the models learn a regression function from the waveform of a speech-noise mixture and generate the target speech [1][2][3]. In the time-frequency domain, the models work on the spectrogram, and their output can be masks, such as IBM [4], IRM [5], cIRM [6], etc., or be approximate signals for target speech.

At present, most of the research is carried out in time-frequency domain because the fine-detailed structures of speech and noise can be more separable with TF representations after the short-time Fourier transform (STFT) [7][8]. In these researches, a complex-valued spectrogram can be decomposed into real and imaginary part in the Cartesian coordinate, while in polar coordinate it can be decomposed into magnitude and phase part. While early studies focused on the magnitude, the later works, such as cIRM [6], show the beneficial improvement of the phase. Recently, with the development of the Neural Network, models with encoder-decoder structure are developed. These models can capture the local time-frequency details of speech signals through the complete convolution network structure and the feature selection ability provided by the encoder-decoder mechanism [9].

The above models are all single-stage networks. Due to their limited mapping ability, they usually can't run well for relatively difficult tasks. To deal with this problem, two-stage models have been proposed [10][11][12]. The main idea of two-stage models is to obtain coarse-grained results from the first network and further refine it with the second network.

This paper proposes a two-stage model to accomplish speech enhancement task in low SNR condition. In our method, we use Deep Complex Convolution Recurrent Network （DCCRN） [13] as the first stage network. The network is carried out in time-frequency domain, takes magnitude and phase as inputs, and uses complex network to improve its modeling ability. Its output is the magnitude and phase target of clean speech corresponding to TF unit. We add an Attention gate [14] to the DCCRN network to strengthen the model's attention to local features. After the first stage, the magnitude and phase target is obtained. However, due to the heavy noise, there is still interference in the magnitude target. Therefore, we use SubNet [15] as our second stage network to refine the magnitude target, for speech has spectral dependencies along the frequency axis [16], especially in magnitude. SubNet is a LSTM network, whose input consists of one frequency, together with several context frequencies. The output is a prediction of the clean speech magnitude target for the corresponding frequency. By fusing SubNet, the model achieves better enhancement performance in low SNR condition. To evaluate our proposed method, we conduct experiments on a severely noisy speech dataset, and the results show that our approach outperforms other speech enhancement methods mentioned in Section 5.

This paper is organized as follows: Section 2 introduces the related work. Then, the overview of our method and the details of the proposed network architecture are described in Section 3. And in Section 4 experimental details are explained as well as the information about the dataset. We present the results and some analysis in Section 5 and Section 6 concludes the paper.

Authorized licensed use limited to: NATIONAL TAIWAN NORMAL UNIVERSITY. Downloaded on August 26,2024 at 18:03:26 UTC from IEEE Xplore. Restrictions apply.

II. Related Work

For the purpose of speech enhancement and denoising, Daniel et al. [17] propose the Wave-U-Net. The method implements end-to-end speech enhancement in the time-domain, which allows modelling phase information and avoids fixed spectral trans-formations.

Tan et al. [9] combine convolution auto-encoder (CAE) [18] and LSTM together to propose convolution recurrent neural network (CRN), where CAE helped to learn temporal-frequency (T-F) patterns and LSTM effectively captured dynamic sequence correlations. Different from Wave-U-Net, CRN estimate the real and imaginary spectrogram of mixture speech simultaneously with the use of CSM. The CSM [9] possess the full information of a speech signal to reconstruct speech.

The CRN uses complex-valued spectrogram as the training target but trains the model in a real-valued network. That means the method predicts the real and imaginary part respectively. In order to train the complex target more effectively, Hu et al. propose DCCRN [13]. The method effectively combines both the advantages of DCUNET [19] and CRN. When training the model, DCCRN estimates cIRM [6] and is optimized by signal approximation (SA).

For better deal with the challenges brought by the non-stationary and low level of the SNR noise, the two-stage networks were proposed. Hao et al. [15] propose a full-band and sub-band fusion model named FullSubNet. The output of the full-band model is the input of the sub-band model. The FullSubNet can capture the global full-band context while modeling signal stationarity and attending the local spectral pattern. Li et al. [20] propose a method decoupling the optimization tasks of magnitude and phase. In the first stage, the method only optimizes the magnitude and removes the most noise components. In the second stage, only the phase needs to be modified by the network. Meanwhile the network further refines the magnitude. Hao et al. [21] propose a two-stage approach that consists of binary masking and spectrogram inpainting. In the binary masking stage, binary mask is obtained by hardening soft mask and then be used to remove time-frequency points that are dominated by severe noise. In the spectrogram inpainting stage, the method uses a CNN with partial convolution to perform inpainting on the masked spectrogram obtained in the first stage.

III. DCCRN-Subnet

This paper focuses on the denoising task, and the objective is to suppress the non-stationary and low level of the SNR noise. we denote X as the complex-valued STFT spectrogram of noisy speech, and X is defined as

$$X = X_r + jX_i \tag{1}$$

where X_r, X_i denotes the real and imaginary part of the spectrogram, respectively. We denote S, $\tilde{S}$ as the complex-valued STFT spectrogram of clean and estimated speech, respectively. S and $\tilde{S}$ can be defined as

$$S = S_r + jS_i \tag{2}$$

$$\tilde{S} = \tilde{S}_r + j\tilde{S}_i \tag{3}$$

We propose a DCCRN and SubNet fusion model to complete the speech enhancement task. The basic workflow is shown in Fig. 1. With the STFT, we can get the real and imaginary parts of the noisy spectrum, and then feed them into the DCCRN. The DCCRN output the corresponding target of the clean speech. After the first stage, we can obtain a coarse complex target, i.e.,$(\tilde{S}_r^{DCCRN}, \tilde{S}_i^{DCCRN})$, the real and imaginary part of the spectrogram estimated by DCCRN, which is expected to provide complementary information to the following SubNet model. In the second stage, $\tilde{S}_r^{DCCRN}$and X_r are contacted as the input of SubNet. and the network then estimates the refined real target, i.e., $\tilde{S}_r^{SubNet}$, which is stacked with $\tilde{S}_i^{DCCRN}$ to obtain a final output. We adopt the cIRM as our model's learning target, and the spectrogram of the estimated speech can be calculated by multiplicative pattern DCCRN-R. The inverse STFT (ISTFT) is applied to obtain the final real-valued time-domain waveform.

Furthermore, we add the Attention gate to the DCCRN, as the Attention gate has the capability of automatically suppressing the irrelevant regions and emphasizing the important features.

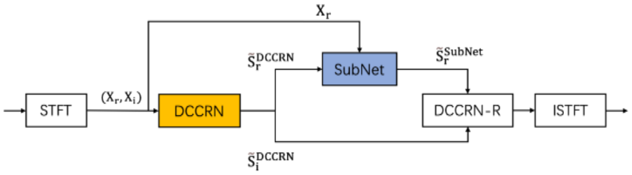

Fig. 1. Two-stage network

A. DCCRN

The DCCRN is shown in Fig. 2. It is an essentially causal convolutional encoder-decoder architecture with two LSTM layers between the encoder and the decoder. The encoder aims to extract high-level features and the decoder aims to reconstruct the low-resolution features to the original size of the input by strided deconvolution operations. LSTM is used to model the temporal dependencies and a dense layer follows the last LSTM. We integrate the Attention gate block into the first skip connection. In our method, DCCRN modifies CRN with complex CNN and uses complex LSTM to replace the traditional LSTM. The complex module models the correlation between magnitude and phase by simulating the complex multiplication.

The complex encoder/decoder block includes complex Conv2d, complex batch normalization [22] and real-valued PReLU [23]. The complex batch normalization and PReLU are implemented as [22]. Complex Conv2d consists of four traditional Conv2d operations. The complex-valued convolutional filter W is defined as $W = W_r + jW_i$, where the real-valued matrices W_r and W_i represent the real and imaginary part of a complex convolution kernel, respectively. Meanwhile, the input complex matrix X is defined as $X = X_r + jX_i$. Finally, the complex output can be calculated by the complex convolution operation $X \circledast W$:

$$F_{out} = (X_r * W_r - X_i * W_i) + j(X_r * W_i + X_i * W_r) \tag{4}$$

Authorized licensed use limited to: NATIONAL TAIWAN NORMAL UNIVERSITY. Downloaded on August 26,2024 at 18:03:26 UTC from IEEE Xplore. Restrictions apply.

where F_{out} denotes the output feature of one complex layer.

Similar to complex convolution, given the real and imaginary parts of the complex input X_r and X_i, complex LSTM output F_{out} can be defined as:

$$F_{rr} = LSTM_r(X_r);\ F_{ir} = LSTM_r(X_i); \tag{5}$$

$$F_{ri} = LSTM_i(X_r);\ F_{ii} = LSTM_i(X_i); \tag{6}$$

$$F_{out} = (F_{rr} - F_{ii}) + j(F_{ri} + F_{ir}) \tag{7}$$

where $LSTM_r$ and $LSTM_i$ represent two traditional LSTMs of real part and imaginary part, and F_{rr} is caculated by $LSTM_r$, with input X_r.

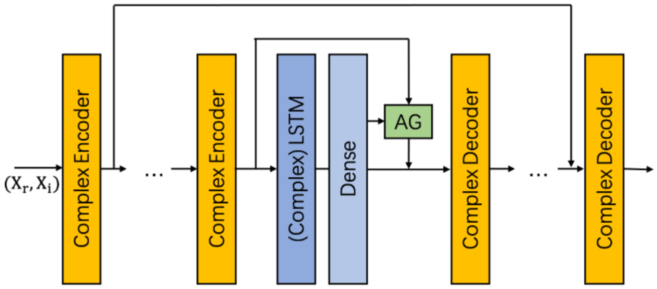

Fig. 2. DCCRN network

B. Attention-gate block

The schematic of Attention gate block is shown in Fig. 3, in which additive attention [14] is used. By using attention mechanism, Attention gate module can automatically suppress irrelevant areas of the input, and highlight salient features useful for specific tasks. In this paper, we add Attention gate block to skip connection, to make better use of context information.

we give the calculation process of Attention-gate unit as follows:

$$\varphi = W_p \sigma_l (W_f f + W_g g + b_g) + b_p \tag{8}$$

$$w = W_s(\sigma_2(\varphi(f, g\,;\, \Phi_{att}))) \tag{9}$$

$$\hat{f} = wf \tag{10}$$

where Φ_{att} is a set of parameters of Attention gate. f is the feature of an encoding layer, i.e., lower-level feature in skip connection. g is the feature corresponding to f, which contains contextual information to determine focus regions of f. By this way, only the related feature can be merged by concatenation operation. W_g , W_f and W_p are the linear transformations computed using channel-wise 1x1 convolutions. b_g and b_p are the bias terms. σ_1 is Leaky ReLu and σ_2 is Sigmoid activation function, respectively. w are the attention coefficients computed from Attention gate. Resampling of attention coefficients is done by using nearest interpolation. Finally, the output of attention gate unit is the element-wise multiplication of f and w.

After the output of attention gate is obtained, it is concatenated with the feature from the corresponding layer along the channel dimension to generate the input of the decoding layer. The object of Attention-gate here is to increase estimation accuracy and improve the performance of speech enhancement task. We will show its effect in the following section.

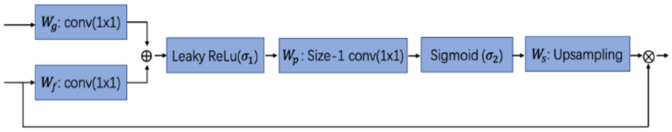

Fig. 3. Attention gate block

C. SubNet

The SubNet predicts the frequency-wise clean speech magnitude target according to the local spectral mode encoded in the noisy sub-band signal. The detailed method is described as follows: we define $X_r(t,f)$ as the real part of a time-frequency point at time t and frequency f. We take a $X_r(t,f)$ and its adjacent 2 × N time-frequency points as a local unit. N represents the number of neighbor frequencies considered on each side. For boundary frequencies, with $f - N < 0$ or $f + N > F - 1$, circular Fourier frequencies are used. We concatenate the local unit and the output of the DCCRN, which is denoted as $F_{DCCRN}(X_r(t,f))$. We then take the $\hat{X}_r(t,f)$ as the input of the SubNet. The $\hat{X}_r(t,f)$ is denoted as follows:

$$\hat{X}_r(t,f) = \begin{bmatrix} X_r(t,\ f-N), \dots, X_r(t, f\ -\ 1), X_r(t,f), \\ X_r(t,f\ +\ 1), \dots, X_r(t,\ f\ +\ N), F_{DCCRN}(X_r(t,\ f)) \end{bmatrix}^{\mathrm{T}} \tag{11}$$

$$\hat{X}_r(t,f) \in \mathbb{R}^{2N+2} \tag{12}$$

The noisy sub-band spectra (composed of 2N + 1 frequencies) provides the local spectral pattern. With the ability of modeling the long-distance context dependency along the frequency axis, the SubNet can learn the local spectral pattern to further refine the magnitude. Limited by the convolution kernel size, the DCCRN can't exploit long-distance context frequencies directly. Thus, although the local pattern is present in the input of the DCCRN as well, the model can't learn it well.

Consequently, the SubNet can learn some different information relative to the DCCRN. Meanwhile, the DCCRN captures the local time-frequency details and the output of the DCCRN provide some complementary information not seen by the SubNet. In this way, the two models complement each other.

D. Training target

In this paper, we adopt the complex Ideal Ratio Mask(cIRM) as our models' target and the model is optimized by signal approximation (SA). Given the complex-valued STFT spectrogram of clean speech S and noisy speech X, cIRM can be defined as

$$cIRM = \frac{X_r S_r + X_i S_i}{X_r^2 + X_i^2} + j\frac{X_r S_i - X_i S_r}{X_r^2 + X_i^2} \tag{13}$$

Authorized licensed use limited to: NATIONAL TAIWAN NORMAL UNIVERSITY. Downloaded on August 26,2024 at 18:03:26 UTC from IEEE Xplore. Restrictions apply.

TABLE I. OBJECTIVE RESULTS ON THE TEST DATA WITH SEEN AND UNSEEN NOISE

Method	**Seen**			**unseen**		
	SI_SDR	***STOI***	***PESQ***	***SI_SDR***	***STOI***	***PESQ***
Noisy	1.67(2.51)	0.67(0.72)	0.99(1.23)	1.53(2.49)	0.66(0.71)	0.96(1.21)
Wave-U-Net	10.12(11.77)	0.75(0.79)	2.08(2.38)	3.01(5.41)	0.70(0.74)	1.35(1.49)
CRN	10.87(11.99)	0.78(0.82)	2.13(2.42)	3.29(5.53)	0.71(0.75)	1.38(1.51)
FullSubNet	11.91(13.49)	0.89(0.91)	2.37(2.60)	4.17(6.38)	0.74(0.78)	1.48(1.62)
DCCRN	10.56(12.09)	0.85(0.88)	2.06(2.47)	3.38(5.66)	0.72(0.76)	1.40(1.54)
DCCRN-SubNet	11.59(12.85)	0.87(0.89)	2.22(2.53)	3.85(6.09)	0.73(0.77)	1.44(1.58)
DCCRN-SubNet-AG	11.72(12.96)	0.88(0.90)	2.29(2.56)	3.99(6.11)	0.74(0.77)	1.47(1.58)
DCCRN-E-SubNet-AG	16.41(17.52)	0.94(0.92)	2.68(2.73)	4.96(7.14)	0.74(0.78)	1.54(1.67)

where X_r and X_i denote the real and imaginary parts of the noisy complex spectrogram, respectively. The real and imaginary parts of the clean complex spectrogram are represented by S_r and S_i.

We use two multiplicative patterns [13] for DCCRN. The estimated clean speech $\tilde{S}$ can be calculated as below.

$$DCCRN - R : \tilde{S} = (X_r \cdot \widetilde{M}_r) + j(X_i \cdot \widetilde{M}_i) \qquad (14)$$

$$DCCRN - E : \tilde{S} = \tilde{X}_{mag} \cdot \widetilde{M}_{mag} \cdot e^{\tilde{X}_{phase} + \widetilde{M}_{phase}} \qquad (15)$$

DCCRN-R estimates the mask of the real and imaginary part of $\tilde{X}$, respectively. Besides, DCCRN-E performs in polar coordinates.

The loss function of model training is SI-SNR [24], which is defined as:

$$s_{target} := (< \tilde{s}, s > \cdot\, s)/\|s\|_2^2 \qquad (16)$$

$$e_{noise} := \tilde{s} - s_{target} \qquad (17)$$

$$SI - SNR := 10log10(\frac{\|s_{target}\|_2^2}{\|e_{noise}\|_2^2}) \qquad (18)$$

where s and $\tilde{s}$ are the clean and estimated time-domain waveform, respectively. $< \cdot, \cdot >$ denotes the dot product between two vectors and $\| \cdot \|_2$ is Euclidean norm (L2 norm).

IV. EXPERIMENT

A. Dataset

Our experiments were carried out on the dataset of aishell2 and NOISEX-92 [25]. We selected 1500 utterances as clean datasets from ashell2, including 16 speakers (8 males and 8 females). We split the datasets into 1250, 125 and 125 utterances as training, verification and evaluation sets respectively. The noise dataset is from NOISEX-92. We use eight types as seen noises (babble, buccaneer1, destroyer engine, f16, factory1, machinegun, pink and white) and the other four as unseen noises (buccaneer2, destroyerops, factory2 and hfchannel). Speech noise mixing used in training and verification is generated by selecting utterances from speech sets and noise sets, and mixing them at a SNR between -5 dB and 15 dB. The common evaluation set is generated at five SNR (- 5dB, 0 dB, 5 dB, 10 dB, 15 dB). Further， we selected utterances from common evaluation set to form the low-SNR evaluation set, in which SNR is less than 10 dB. Using the method above, we prepare 40,000 noisy-clean pairs for training set, 4000 for common test set, 2400 for low-SNR test set and 4000 for validation set. All signals are sampled at 16kHz.

B. Preprocessing and Training Setup

The spectrum representation is obtained by applying 512-point STFT with Hanning window size of 512 and hop size of 256 to audio files sampled at 16kz. By removing half of the symmetry, only 257 points STFT are remained. After that, the last STFT point is also moved to produce the input dimension of the power of 2. The network is trained with Adam optimizer, the batch size is 24, and the initial learning rate is 0.0001. For the DCCRN, there are six encoder layers and six decoder layers. The channel numbers are 32, 64, 128, 256, 256 and 256. The stride values are all (2, 1) and kernel sizes are all (5, 2).

V. RESULTS AND DISCUSSION

To evaluate the performance of the proposed DCCRN-SubNet method, Wave-U-Net [17], CRN [9], FullSubNet [15] and DCCRN [13] are used as the reference methods. All of them use the same training set and STFT configuration. Meanwhile, to evaluate the contribution of Attention gate block, the proposed DCCRN-SubNet is also used as the reference approach. In order to get the best performance of the method, we conducted experiment on DCCRN-SubNet-Ag with the multiplicative pattern E [13]. We use three metrics to evaluate the performance of the methods. They are SI-SDR (the scale-invariant speech distortion ratio [26]), PESQ (Perceptual evaluation of speech quality [27]) and STOI (the Short-Time Objective Intelligibility [28]). For all of the metrics, the higher value corresponds to the better quality of the enhanced speech.

The experimental results are reported in Table 1. The numbers in parentheses are the scores for the common test set. The numbers in front of the brackets are the scores for the low-SNR test set. It is shown that FullSubNet obtains the best metrics and the DCCRN ranks the second place in both seen and unseen noise cases among four baselines. On the other hand, the scores for the low-SNR test set are lower, which shows the negative

Authorized licensed use limited to: NATIONAL TAIWAN NORMAL UNIVERSITY. Downloaded on August 26,2024 at 18:03:26 UTC from IEEE Xplore. Restrictions apply.

impact of low SNR. Compared with the DCCRN, the proposed DCCRN-SubNet obtains notable improvements in all metrics, especially for the low-SNR test set. This is because that SubNet can exploit context dependencies along the frequency axis which are essential for speech enhancement. For DCCRN focuses more on local pattern along the time axis, the SubNet provides complementary information.

With Attention gate, the DCCRN-SubNet-Ag can get higher objective metrics compared with the DCCRN-SubNet in seen noise cases. This proves that Attention gate identify the relatively important features in specific frequency domain successfully. Meanwhile, the method can maintain the performance in the unseen noise cases.

Finally, with the multiplicative pattern of DCCRN-E, the DCCRN-E-SubNet-AG achieves the best results. These results testify that the proposed fusion model successfully integrates the virtues of DCCRN and SubNet.

VI. Conclusion

In this paper, we propose a DCCRN and SubNet fusion model, named DCCRN-SubNet, for better speech enhancement in low SNR and non-stationary noise. The DCCRN takes the complex-valued spectrogram as input and output the corresponding magnitude and phase target. By exploiting the long-distance dependency along the frequency axis directly, the SubNet further refine the magnitude target. In this way, the method enhances the noisy speech well. The experimental results show that, compared with previous state-of-the-art methods, the proposed model achieves consistently better performance.

References

[1] Rethage D., Pons J., Serra X., “A Wavenet for Speech Denoising,” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), ICASSP 2018, pp. 5069–5073.

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

[3] Hao, X. , Su, X. , Wang, Z. , Zhang, H. , & Batushiren, “UNetGAN: A robust speech enhancement approach in time domain for extremely low signal-to-noise ratio condition,” INTERSPEECH, pp. 1786–1790, 2019.

[4] Wang D., “On ideal binary mask as the computational goal of auditory scene analysis,” Speech Separation by Humans and Machines, pp. 181–197, 2005.

[5] Sun, L. , Du, J. , Dai, L. R. , & Lee, C. H, “Multiple-target deep learning for lstm-rnn based speech enhancement,” 2017 Handsfree Speech Communications and Microphone Arrays (HSCMA), pp. 136–140.

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

[7] Srinivasan S., Roman N., Wang D.L., “Binary and ratio time-frequency masks for robust speech recognition,” Speech Communication, vol. 48, no. 11, pp. 1486–1501, 2006.

[8] Narayanan A., Wang D., “Ideal ratio mask estimation using deep neural networks for robust speech recognition,” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), ICASSP 2013, pp. 7092–7096.

[9] Tan K., Wang D., “A convolutional recurrent neural network for real-time speech enhancement,” INTERSPEECH, vol. 2018, pp. 3229–3233.

[10] Huang, Z. , Yu, W. , Zhang, W. , Feng, L. , & Xiao, N., “Gradual network for single image de-raining,” Proc. of ACMM, pp. 1795–1804, 2019.

[11] Ala, B. , My, C. , Cza, B. , & Xla, B., “Speech enhancement using progressive learning-based convolutional recurrent neural network,” Applied Acoustics, vol. 166, pp. 107347, 2020.

[12] Li, A. , Zheng, C. , Fan, C. , Peng, R. , & Li, X., “A recursive network with dynamic attention for monaural speech enhancement” INTERSPEECH, pp. 2422-2426, 2020.

[13] Hu, Y. , Liu, Y. , Lv, S. , Xing, M. , & Xie, L., “DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement,” INTERSPEECH, pp. 2472-2476, 2020.

[14] Oktay O., Schlemper J., Folgoc L. L., et al., “Attention u-net: Learning where to look for the pancreas,” arXiv preprint arXiv:1804.03999, 2018.

[15] X Hao, X Su, Horaud, R. , & X Li., “FullSubNet: a full-band and sub-band fusion model for real-time single-channel speech enhancement,” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), ICASSP 2021, pp. 6633–6637.

[16] Quatieri, T.F., “Discrete-Time Speech Signal Processing”, Principles and Practice. 1st ed. Upper Saddle River, NJ: Prentice Hall, 2002.

[17] Daniel S., Sebastian E., Simon D., “Wave-U-Net: A multi-scale neural network for end- to-end audio source separation,” Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR), ISMIR 2018, pp. 334-340.

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

[19] Choi H.S., Kim J.H., Huh J., et al., “Phase-aware speech enhancement with deep complex u-net,” arXiv preprint arXiv:1903.03107, 2019.

[20] Li, A. , Liu, W. , Luo, X. , Zheng, C. , & Li, X., “ICASSP 2021 Deep Noise Suppression Challenge: Decoupling magnitude and phase optimization with a two-stage deep network,” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), ICASSP 2021, pp. 6628–6632.

[21] Hao, X. , Su, X. , Wen, S. , Wang, Z. , & Chen, W., “Masking and Inpainting: A two-stage speech enhancement approach for low snr and non-stationary noise,” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), ICASSP 2021, pp. 6959–6963.

[22] Trabelsi C., Bilaniuk O., Zhang Y., et al., “Deep complex networks,” arXiv preprint arXiv:1705.09792, 2017.

[23] He, K. , Zhang, X. , Ren, S. , & Sun, J., “Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification,” Proceedings of the IEEE international conference on computer vision, 2015, pp. 1026–1034.

[24] Luo Y., Mesgarani N., “Tasnet: time-domain audio separation network for real-time, single-channel speech separation,” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), ICASSP 2018, pp. 696–700.

[25] Varga A., Steeneken H. J., “Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems,” Speech Communication, vol. 12, no. 3, pp. 247–251, 1993.

[26] Roux, J. L. , Wisdom, S. , Erdogan, H. , & Hershey, J. R., “SDR – Half-baked or well done?” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), ICASSP 2019, pp. 626–630.

[27] Rix, A. W. , B Ee Rends, J. G. , Hollier, M. P. , & Hekstra, A. P., “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), ICASSP 2001, pp. 749–752

[28] Taal, C. H , Hendriks, R. C. , Heusdens, R. , & Jensen, J., “A short-time objective intelligibility measure for time-frequency weighted noisy speech,” IEEE international conference on acoustics, speech and signal processing (ICASSP), ICASSP 2010, pp. 4214–421

Authorized licensed use limited to: NATIONAL TAIWAN NORMAL UNIVERSITY. Downloaded on August 26,2024 at 18:03:26 UTC from IEEE Xplore. Restrictions apply.