Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/6/20230904

Full convolutional speech enhance network based on with combined amplitude and phase spectra

Xingda Li

Fu Zhou University, No.2 Wulongjiang Avenue, Fuzhou, China

XINGDA.LI.2022@MUMAIL.IE

Abstract. With the coming of the information age, speech has a pivotal role in modern society, and the study of speech enhancement has become necessary. To implement the often-neglected speech phase spectrum. This research proposes a full convolutional model based on an enhanced U-Net. The standard convolution kernel in the model is replaced with a deformable convolution with adaptive learning bias, a double attention mechanism is added between the encoder and decoder layers, and many residual connections are used to link the different layers, and the phase spectrum is also used as one of the inputs. The simulation experiments show that the added modules all contribute to the speech enhancement effect. Among them, the addition of the deformable convolution has the greatest contribution to the model.

Keyword: deformable convolution, speech enhancement, phase spectrum.

1. Introduction

Voice has been increasingly prevalent since the information era began. Nevertheless, both the damage to the speech information in the process of transmission and the quality of the speech can be significantly affected by the noise that is mixed in while recording. Speech enhancement technology is the extraction of meaningful speech signals from noise-polluted speech signals. Speech enhancement has a great potential for a wide range of applications. For example, the application in hearing aids, the application in noise-canceling headsets, and the improvement of speech quality in voice calls under extreme conditions [1]. Speech enhancement is also a major pre-processing module of speech recognition, where the quality of speech directly determines the accuracy of speech recognition. That is why speech enhancement in modern society has a role that cannot be ignored [2].

Since deep learning has advanced in recent years, time-frequency masking and time-frequency mapping techniques for speech improvement have become widely used. The majority of conventional techniques, however, merely improve the amplitude spectrum of speech and directly utilize the phase spectrum of the noise-containing speech to reconstruct it, ignoring the phase and limiting the benefit of speech enhancement at low signal-to-noise ratios (SNR). Speech enhancement that uses traditional deep learning networks also has the disadvantages of large computation and many parameters, as well as a need to go through splitting and splicing operations to enhance the spectral map, which greatly increases the time-consuming speech enhancement. And the speech features have the characteristics of mutability and disorder, which are hardly extracted by ordinary convolution.

DOI: 10.54254/2755-2721/6/20230904

In purpose to compensate the above shortcomings this paper proposes a fully convolutional CMPADN (Combined amplitude and phase using double-attention mechanism and deformable convolutional network) network based on improved U-Net network. Convolution layers are used instead of fully linked layers in CMPADN networks, which decreases the number of parameters and computational labor and does away with the need to stitch together spectral data. In CMPADN network. 1) The addition of the phase spectrum enables the model to perform better at low SNR. 2) The double-attention mechanism combining spatial attention and channel attention has enabled the model to focus on the extraction of effective information and accelerate the model convergence. 3) The addition of the deformable convolution module facilitates the extraction of irregular features and enhances the feature extraction capability of the model. 4) Insert residual connections between layers to ensure that the model does not diverge and that detailed information is retained

2. Related work

Traditional methods and deep learning are the two broad categories of speech enhancement. Spectral subtraction, Wiener filter etc. can be considered as traditional methods. Yet traditional speech enhancement methods make a series of assumptions on noise and utilize people's a priori knowledge of noise. Therefore, their robustness is worse, and the performance tends to be significantly reduced when encountering real noise, which has a large limitation. To solve these problems, deep learning-based approaches are proposed. Depending on the improvement object, deep learning may be separated into two categories: time domain and frequency domain.

In 2017 Pascual et al. proposed SEGAN speech enhancement based on GAN network, in 2018 Stoller et al. proposed a time-domain speech separation model for wave-u-net, and in 2019 Pandey et al. implemented TCN (temporal convolutional network) on speech enhancement [3]. Time-domain based methods have the benefits of easy processing methodology, no loss of original signal details such as phase, etc. However, the harder extraction of time domain signal features can be an obstacle to enhance the performance [4]. The method of frequency domain based deep learning is to transform the time domain signal by short time Fourier transform (STFT) to get a complex domain spectrum of speech, and its amplitude spectrum is used as the enhancement object, and then reconstruct the speech together with the noisy phase spectrum. According to the enhancement mode, it can be classified into feature mapping and feature masking [5]. Deep learning makes almost no assumptions about noise; hence it has strong robustness and can have better enhancement performance for all kinds of noise, which is also the mainstream speech enhancement scheme at present. There is a deep neural network (DNN) that improves the feature extraction capability by increasing the depth of the model to obtain better speech enhancement performance, and there is a recurrent neural network (RNN) that enhances the ability to model the temporal dimension in the speech enhancement process.

Two types of models enhance the speech spectrum from different dimensions and both have good performance. Nevertheless, the deep learning approach still has its limitations. The first is that the use of a fully connected layer results in larger model parameters, which places higher demands on the equipment used to train and use the model, also decreases the speed of speech enhancement. Secondly, Since the information in the temporal dimension is not fully utilized, the entire spectral map cannot be improved. Full convolution neural networks (FCN) have gained popularity as a study area for improving speech in recent years, and their use can not only minimize the number of model parameters, but also the spatial information implicit in the speech signal can be significantly captured to enhance the model for feature extraction by applying convolution operations [2]. With the in-depth study of speech, it has also been found that phase also has a non-negligible effect on speech quality, especially when the SNR is low [3].

3. Method

This chapter defines the speech enhancement problem in section 3.1 and introduces the CMPADN network structure in section 3.2

DOI: 10.54254/2755-2721/6/20230904

3.1. Definition of speech enhancement

For pure speech contaminated by noise, we can consider it as the summation of pure speech and pure noise

$$Y(t) = X(t) + N(t) \tag{1}$$

where Y, X, and N stand for the noisy speech time domain functions, pure speech, and noise, respectively. Our goal is to separate X from Y. The following equation is obtained by performing STFT on the speech signal.

$$Y_{t,f} = X_{t,f} + N_{t,f} \tag{2}$$

t and f represent the time and frequency respectively, $Y_{t,f}, X_{t,f}, N_{t,f}$ represent the complex values at the corresponding time and frequency. By selecting the appropriate SFTF parameters, we can get the complex spectrum of suitable size. The amplitude spectrum and phase spectrum of speech can be gotten by taking the mode operation and taking the phase angle operation on it. The amplitude spectrum is typically the only enhancement object used in speech enhancement algorithms. The enhanced amplitude spectrum is then reconstructed with the noisy phase spectrum. This method has the unavoidable limitation that the use of noisy phase spectra results in a waveform with non-removable noise. Along with the in-depth understanding of speech, it is noticed that phase plays a role in low SNR. Therefore, in this experiment, both the phase spectrum and the amplitude spectrum are used as the enhancement objects. Lastly, using the inverse short time Fourier transform, the improved phase spectrum and amplitude spectrum are rebuilt to create the final enhanced speech.

3.2. Network structure

The overall structure of the current CMPADN speech enhancement model is depicted in Figure 1. CMPADN has made improvements in the following areas. 1) The original U-Net network has no connection between the encoder and decoder, and the present model connects the encoder and decoder by the residual convolution module. In deep networks, residual connection can avoid gradient dispersion and can retain more detailed information. The direct connection of different layers has a large feature ambiguity, and the addition of convolution operation can effectively eliminate this ambiguity. 2) In the original U-Net structure, the encoder layer is directly connected by 2*2 maximum pooling. It can reduce the computational effort, but the maximum pooling will simultaneously lose a significant quantity of useful information. So, this experiment uses the designed RBC (Residual Branch convolution) convolution module instead of the original convolution operation, in which the feature map is convolved, pooled, and deconvolved, and the feature map obtained by deconvolution is added with the upper layer input, and then the information lost by pooling is learned by residual convolution. The number of channels of the 5 RBC convolution modules is set to 64, 128, 256, 512 and 1024 respectively. 3) A double attention mechanism is added to the decoder. Each module is described in depth in the paragraphs that follow.

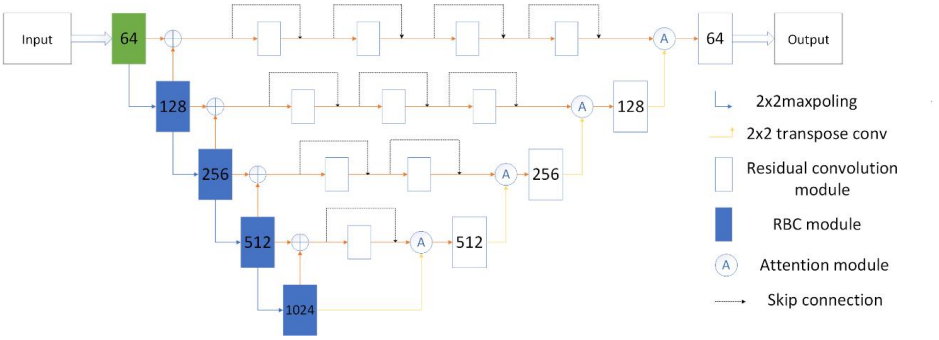

Figure 1. General structure of the model.

DOI: 10.54254/2755-2721/6/20230904

3.2.1. Encoder layer. In the neural network structure, the encoder layer is responsible for feature extraction. The number of channels is increased to improve the extraction of deep features. The convolution kernel of the convolution neural network extracts features, and the receptive field changes as the size of the convolution kernel changes. Information is gathered more readily the larger the receptive field. However, as the convolutional kernel increases, the computational effort also increases dramatically. Therefore, a multilayer stepwise convolution method is proposed. Two layers of 5*5 convolution kernels work better than one layer of 7*7 [4], and the former has a smaller computational effort and a higher ability to extract nonlinear features than the latter. The multi-layer convolution used in the encoder layer applies this principle. We believe that as the number of convolution layers increases, more abstract and advanced features are extracted. The process is similar to the way the human brain processes information.

3.2.2. Deformable convolution. In the selection of the convolution kernel, this experiment also makes an innovation, unlike the image features, the features of the speech signal are more abstract. This experiment switches from the standard convolution to the deformable convolution for better feature extraction [5]. The convolution kernel of standard convolution is fixed. We translate and slide this fixed size convolution kernel on the input feature map, multiply the values on the feature map with the convolution kernel and then sum up to get the convolution result. However, the feature information contained in different positions of the feature map is different. Consequently, it is required to adaptively change the convolution kernel's size and form at various points. This is achieved by deformable convolution, which allows the use of different shapes of convolution kernels at different locations. It is implemented by adding a bias to each sample point of the standard convolution, and the magnitude of the bias is learned by the neural network. As shown in Figure 2.

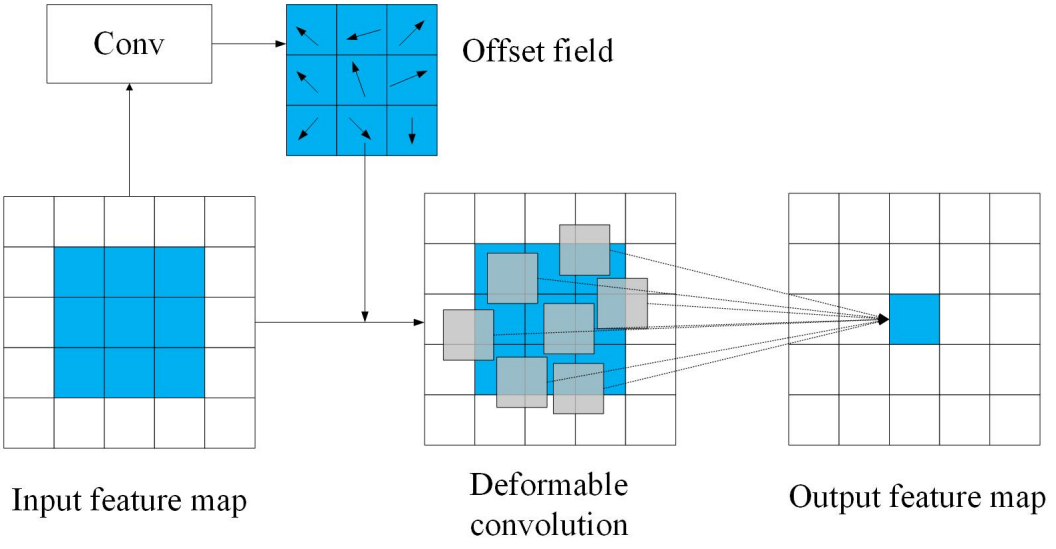

Figure 2. Illustration of deformable convolution.

The formula for the deformable convolution is as follows:

$$y(p_0) = \sum_{p_n \epsilon R} w(p_n) * x(p_0 + p_n + \Delta p_n) \#(3)$$

$p_n$denotes sampling points of the deformable convolution, $w(p_n)$denotes the weight of the convolution kernel on p_n, R ∈ {(-1,1), (0,1)…(1,1)}. *x, y* are the input features and output features respectively, Δp_n is the coordinate offset. It is the learning of bias that gives the deformable convolution a more powerful feature extraction capability compared to the standard convolution.

DOI: 10.54254/2755-2721/6/20230904

3.2.3. Residual branch convolution module. Figure 3 depicts the layout of the RBC (Residual Branch Convolution) convolution module. The input after two layers of deformable convolution is summed with the input after residual convolution to obtain Output2. Transpose Output2 to get Output1, Output2 is the output of this layer, while Output1 is used as one of the inputs of the upper layer that is added to output2 of the upper layer. The residual connection is also added in the design of the convolution module, in order to retain more valid information. After considering the computation and feature extraction capability, this experiment decided to add two layers of 5*5 deformable convolution to the RBC convolution module, BN layer is chosen for normalization and ReLu as the activation function.

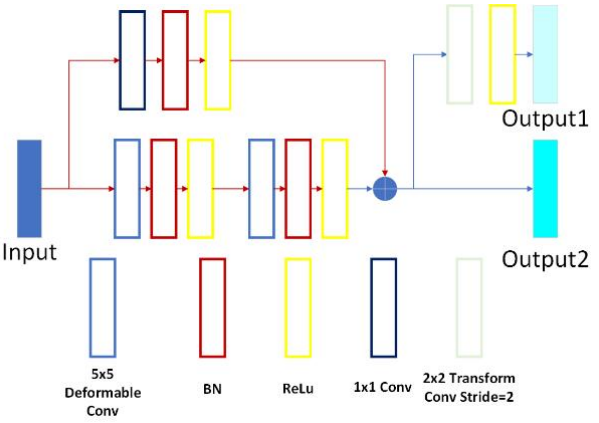

Figure 3. Structure of the RBC module.

3.2.4. Residual convolution module. Figure 4 depicts the architecture of the created residual convolution module. A convolution operation with residual connection is used. Two layers of 5*5 deformable convolution are chosen for the convolution operation part. The main role of this module is to resolve the feature ambiguity between layers by convolution operations and to learn the information that is lost due to pooling operations. Four, three, two, and one residual convolution modules are placed in the first, second, third, and fourth layers of the model to resolve the semantic gap between different convolutional layers.

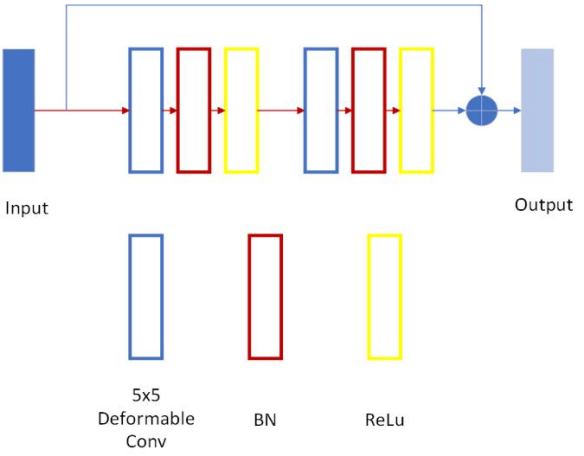

Figure 4. Structure of the residual convolution module.

3.2.5. Double attention mechanism. The attention mechanism was first applied to machine translation, and it is based on the way the human brain processes information. Our brain receives a lot of information every day, especially visual information. In order for us to quickly distinguish between the

DOI: 10.54254/2755-2721/6/20230904

images The brain does not process all the information it receives, it automatically filters out the important parts to focus on and chooses to ignore the unimportant ones. This allows us to achieve both efficient and fast results. Attention mechanisms in neural networks play the same role. There are two main categories of attention processes: spatial attention mechanisms and channel attention mechanisms. This experiment uses many residuals and jump connections, which retains perfect detailed information but also causes redundancy of information, so introducing the attention mechanism can remove redundant information and improve the effect of speech enhancement. In purpose of achieving better attention effect, this experiment adopts a double attention mechanism combining two attention mechanisms, which divides the weights of channels and spaces by global information. Figure 5 depicts its structure. First, different channels in the feature map are given different weights, and then different weights are also given on the space of the feature map. In this way, the information is filtered both spatially and channel-wise, which can greatly reduce the computational effort. In the channel attention. Proper channel interaction can be useful for maintaining performance and reducing model complexity [6]. So we change the fully connected layer after the average pooling to a one-dimensional convolution with k of 5. The appropriate local channel interaction ensures both the performance and reduces the computation.

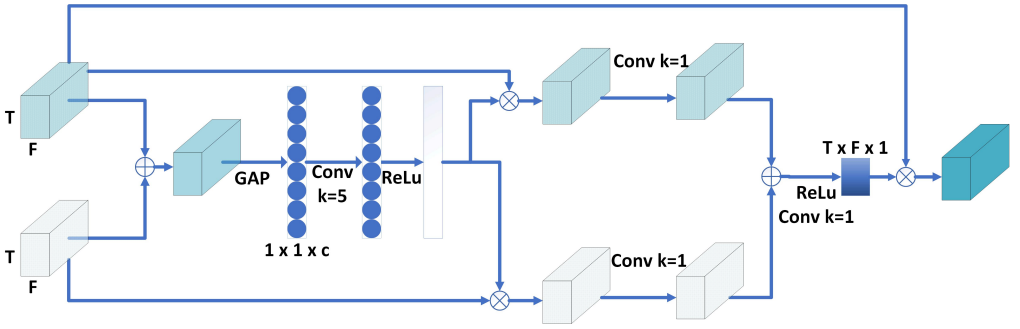

Figure 5. Structure of the double attention mechanism.

4. Experiment

This section provides a detailed description of the experiment-related configuration and data processing, as well as a comparative analysis of the simulation results.

4.1. Experiment-related configuration

We perform simulation experiments in python pytorch 1.6 framework. In the experiments, the mean square error function is selected as the loss function. By comparison, Adam (adaptive moment estimation) with better performance is finally chosen as the optimizer. After pre-experiments, the learning rate is set to 0.001 and the size of epoch is given as 50.

4.2. Data sets and data processing

This experiment selects 400 pure and noisy speech pairs from Valentina database [7] as the training data set, which contains 10 different types of noise, 2 kinds of artificially generated noise (noise similar to the speech waveform and babble) and 8 kinds of natural noise including: home noise, office noise, cafeteria noise, restaurant noise, subway station noise, car noise, train noise, and street noise. Basically, all possible noise scenarios are covered. The test set is a mixture of pure speech from the TIMIT database [8] and noise from Noisex92[9]. In this paper, a dataset containing 100 pairs of pure and noisy speech pairs was created as the test set for this experiment. In the test set, three different SNR of -5dB,0dB,5dB are included.

All speech in this experiment is sampled at 16KHz. The parameters of STFT are set as win-length of 512, hop-length of 128, n_fft of 512, and 'hann' window. The final results are evaluated by using Short Time Objective Interpretability STOI), Perceptual Evaluation of Speech Quality (PESQ).

DOI: 10.54254/2755-2721/6/20230904

4.3. Experimental results and evaluation

We select four models for comparison with this experimental model, which name is wave-u-net [10], VAE [11], CLED-cSA [12], and CRN [13]. Table 1 and Table 2 shows two metric scores (PESQ, STOI) at all SNR of different models. The wave-u-net is a time domain convolution model. The PESQ scores of CMPADN model are 0.27, 0.34, 0.32 higher than wavr-u-net at SNR of -5dB,0dB, and 5dB, respectively. VAE is a deep learning adversarial network model based on the frequency domain. The PESQ scores of CMPADN model are 0.23, 0.30, and 0.31 higher than VAE at SNR of -5dB, 0dB, and 5dB, respectively. CRN is a circular convolution-based model. The PESQ scores of CMPADN model are 0.31, 0.23, and 0.21 higher than CRN at SNR of -5dB, 0dB, and 5dB, respectively. CLED-cSA is a modified circular convolution model. The PESQ scores of this model are 0.26, 0.24, and 0.01 higher than those of CLED-cSA for SNR of -5dB,0dB, and 5dB, respectively. Through this series of comparative analyses, it can be learned that the upper limit of enhancement of the time-domain-based U-Net speech model is limited. The VAE selected with deep adversarial network has better performance at low SNR. The CRN has a better showing in general, but the performance is greatly reduced at low SNR. CLED-cSA based on improved circular convolution has better performance compared to CRN. However, it is still not as good as the method proposed in this paper, especially at low SNR. The preliminary analysis can conclude that the proposed method has a significant contribution to speech enhancement, and the excellent performance at low SNR can be inferred to be the result of the addition of phase spectrum.

Table 1. PESQ comparison of different models under different SNR conditions.

SNR\model	Noisy	wave-u-net	VAE	CLED-cSA	CRN	CMPADN
-5dB	1.32	1.68	1.72	1.59	1.64	1.95
0dB	1.65	2.03	2.07	2.13	2.14	2.37
5dB	2.03	2.47	2.48	2.78	2.58	2.79

Table 2. STOI Comparison of Different Models Under Different SNR conditions.

SNR\model	Noisy	wave-u-net	VAE	CLED-cSA	CRN	CMPADN
-5dB	58.76	67.89	69.12	63.56	66.15	73.53
0dB	69.84	78.37	78.65	80.15	79.23	83.65
5dB	79.63	83.23	83.24	88.34	86.06	89.79

To further determine the magnitude of the contribution of each module to the speech enhancement effect. This paper performed ablation experiments and proposed the CMPAN model without deformable convolution module, the ADN model without the phase spectrum enhancement, and the CMPDN model without the double attention mechanism. Table 3 presents the simulation findings. Adding the deformable convolution module makes the PESQ score 0.276 higher on average. Adding dual attention module makes the PESQ score 0.043 higher on average. And it discovers that the dual attention method and the deformable convolution module both enhance the model's performance and work well together, especially the deformable convolution improves the speech enhancement effect. It can be inferred that the powerful feature extraction capability of deformable convolution plays a role. During the testing phase, it was found that the double attention mechanism can effectively improve the enhancement effect without significantly increasing the complexity of model, also can speed up the convergence of the model.

The analysis of the double attention mechanism shows that it is its function of removing the large amount of useless information due to residual connections that results in. It is the removal of useless information that allows the model to extract useful information more quickly, thus speeding up the convergence of the model. Table 3 also shows that the presence or absence of phase spectrum enhancement also affects the performance of the model. At SNR of -5 dB, 0 dB, and5 dB, the addition of phase spectrum increases the model PESQ score by 0.17, 0.13, and 0.11, respectively. It can see that the lower the SNR the greater the effect of the phase spectrum. Confirming that speech

DOI: 10.54254/2755-2721/6/20230904

enhancement efficacy at low SNR is affected by the phase spectrum. Through the ablation experiments, it can be seen that the combined enhancement of deformable convolution, double attention mechanism and phase spectrum all contribute significantly to the performance of the model.

Table 3. PESQ Comparison of under ablation experiment different SNR conditions.

SNR\model	Noisy	ADN	CMPAN	CMPDN	CMPADN
-5dB	1.32	1.78	1.69	1.91	1.95
0dB	1.65	2.24	2.11	2.32	2.37
5dB	2.03	2.68	2.48	2.74	2.79

5. Conclusion

Through the discussion of the existence models, it has been discovered that the majority of models do not include the usage of phase spectrum in the process of improving speech, and to enhance the use of speech phase spectrum. A fully convolutional speech enhancement model on the basis of the improved U-Net model with joint phase spectrum and amplitude spectrum is proposed. In this paper, the complex spectrum of speech signal transformed by STFT is mode-taking and phase-angle-taking operations to obtain the amplitude spectrum and phase spectrum respectively. To solve the problem that most models neglect the enhancement of phase spectrum. And the U-Net model is improved by 1) introducing a deformable convolution module in it, which improves the enhancement ability of the model through the stronger extraction ability of abstract features by deformable convolution 2) adding a double attention mechanism to fuse spatial attention and channel attention. It can filter the information of the feature map and deliver effective information to accelerate the convergence of the model. 3) Add a large number of residual connections to keep useful information, enhance the nonlinear extraction ability of the model and allow effective fusion of different information. Simulation experiments are performed on the TIMIT database. The experimental results demonstrate the role of phase spectrum at low SNR and the effectiveness of the deformable convolution and double attention mechanisms. In the future, we can further improve the processing of the data, use more advanced transformations, and further explore the potential of phase spectrum.

References

[1] Ronneberger, O., Fischer, P., & Brox, T. U-net: Convolutional networks for biomedical image segmentation. 2015 Inter. Conf. Med. Ima. Compute. Assis. Inter.329(2):315-327.

[2] Fu, S.-W., Tsao, Y., Lu, X., & Kawai, H. Raw waveform-based speech enhancement by fully convolutional networks. 2017 Asia-Pacif. Sig. Infor. Proc. Assoc. Ann. Sum. Conf. 221-232.

[3] Paliwal, K., Wójcicki, K., & Shannon, B. The importance of phase in speech enhancement. 2011, Spe. Comm., 53(4), 465-494.

[4] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. Rethinking the inception architecture for computer vision. 2016 Conf. Compute. Vis. Pat. Rec. 754-763.

[5] Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., & Wei, Y.. Deformable convolutional networks. 2017 Inter. Conf. Com. Vis. 872-883.

[6] Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., & Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. 2020 Conf. Compute. Vis. Pat. Rec.,989-998.

[7] Valentini-Botinhao, C., Wang, X., Takaki, S., & Yamagishi, J. Speech Enhancement for a Noise-Robust Text-to-Speech Synthesis System Using Deep Recurrent Neural Networks. Interspeech, 2016 Conf. Compute. Vis. Pat. Rec.1-11.

[8] Garofolo, J. S. Timit acoustic phonetic continuous speech corpus. Linguistic Data Consortium, 1993, Inter speech 53(4), 465-494.

[9] 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. 1993 Spe. Comm., 12(3), 247-251.

[10] Macartney, C., & Weyde, T. Improved speech enhancement with the wave-u-net. 2018 arXiv

DOI: 10.54254/2755-2721/6/20230904

preprint arXiv:1811.11307.

[11] Bando, Y., Sekiguchi, K., & Yoshii, K. Adaptive Neural Speech Enhancement with a Denoising Variational Autoencoder. 2018 Conf. Compute. Vis. Pat. Rec., 2(2), 236-247.

[12] Strake, M., Defraene, B., Fluyt, K., Tirry, W., & Fingscheidt, T. Separated noise suppression and speech restoration: LSTM-based speech enhancement in two stages. 2019 Appl. Sig. Proc. Aud. Ac. 33(29), 122-134.

[13] Tan, K., & Wang, D. A convolutional recurrent neural network for real-time speech enhancement. 2018, Inter speech, 13(22), 152-164.