Research on Speech Enhancement Algorithm Based on SA-Unet

Yi Zhang
Software College
Yunnan University
Kunming, Yunnan Province, China

Junhui Liu
Software College
Yunnan University
Kunming, Yunnan Province, China

Qing Duan
Key Laboratory for Software Engineering of Yunnan Province
Yunnan University
Kunming, Yunnan Province, China
qduan@ynu.edu.cn

Ruiqiong Wu
Software College
Yunnan University
Kunming, Yunnan Province, China

Bisen Xie
Software College
Yunnan University
Kunming, Yunnan Province, China

Yun Liao
Software College
Yunnan University
Kunming, Yunnan Province, China

***Abstract*—Most of the traditional speech enhancement algorithms are studied in the speech and audio domain. However, the suppression of noise by these methods is not obvious, and the effect of extracting pure speech is not good. In recent years, deep learning has been effectively developed based on its strong learning ability and is being used by more and more researchers. In view of this, this paper proposes a deep learning network model for research in the time domain, which is a SA-Unet network model that combines self-attention and Unet network structure and uses it for speech separation tasks. The model adds a self-attention mechanism between up-sampling and down-sampling in the Unet structure to enhance the perception between contexts, thereby more accurately separating the noise portion of the speech. Finally, the speech quality assessment scores are objectively obtained through the speech quality evaluation indicators and compared with the traditional speech enhancement algorithms. Experiments prove that SA-Unet has made outstanding contributions to speech enhancement technology.**

Keywords-component; speech enhancement, deep learning, self-attention, Unet, Speech separation

I. INTRODUCTION

Speech enhancement is to extract meaningful pure speech from noisy speech and weaken its noise, so as to improve speech quality and readability[1]. In the real world, speech enhancement technology has been applied in many fields. For instance, in the field of mobile communications, it is used to enhance the quality of a call during transmission; Speech enhancement technology is embedded in the hearing aids device to effectively remove the noisy sounds that are mixed into the human ear[2]. In addition, it can also be used in the preprocessing stage of voiceprint recognition and speech recognition[3].

This paper takes this problem as an opportunity to start the study and proposes a new neural network structure model of SA-Unet (Self-Attention Unet) for speech enhancement. The network fuses the Self-Attention mechanism and Unet's network structure. The Self-Attention mechanism is integrated between Unet down-sampling and up-sampling to extract internal relations, capture the internal structure of speech, reduce the number of parameters, and optimize the network model. This paper applied this network structure to speech enhancement, and proved that the results and time efficiency of the experiments generated by the network are improved compared with the Unet network structure. The structure of this paper is organized as follows: The first part describes the dilemma faced by current speech enhancement technologies and the urgent need for technology in the research field; The second part will briefly introduce what the predecessors have done and combine these work to improve the method proposed in this paper; The third part will introduce the SA-Unet network structure and how it can be applied in the field of speech enhancement; The fourth part introduces the experimental steps of the network structure proposed in this paper and compares it with the experimental results of other methods; The fifth part will summarize the experiments in this paper, express personal opinions, and prospect future work.

II. RELATED WORK

A. Traditional Speech Enhancement Algorithm

Traditional speech enhancement methods include spectral subtraction, Wiener filtering and minimum mean square error (MMSE). Paliwal K[4] and others attempted to use spectral subtraction to compensate for the noise modulation spectrum in

978-1-7281-4689-8/19/$31.00 ©2019 IEEE
DOI 10.1109/ICMCCE48743.2019.00187

mixed noise and weaken the noise in mixed speech. Alam M J[5]et al. successfully suppressed the white noise remaining after speech enhancement by Wiener filtering. Because of the effective suppression of white noise by this method, Wiener filtering has been applied for a long time. EPHRAIM[6] et al. use the method of minimum mean square error (MMSE) to evaluate whether there is noise in the ratio of the energy value in a certain range to the minimum value in this period to achieve the effect of noise reduction.

These statistical-based methods have been the mainstream algorithms in the field of speech enhancement for a long time. Although the spectrum subtraction method is mature and the algorithm is simple, the scope of application is small, and it cannot be used in a low SNR environment; Wiener filtering cannot handle non-stationary noise signals; Although the minimum mean square error method has made a great breakthrough in suppressing white noise signals, its algorithm is complex and the time efficiency is low.

B. Speech enhancement algorithm based on deep learning

In recent years, with the development of deep learning, many scholars have begun to use the deep learning to study the technology of speech enhancement. Many scholars began to use deep learning to study the technology of speech enhancement[7][8]. YUAN Wen-Hao[9]et al. used deep convolutional neural networks to study on two dimensions of time domain and frequency domain of speech signals, which effectively improved the performance of speech enhancement under unknown noise conditions; Yu Hua[10] et al. improved the accuracy of the speech enhancement algorithm by improving the deep confidence network (DBN), and compared the traditional LOG-MMSE algorithm, which verified that the DBN has higher accuracy than the traditional method; Valentini-Botinhao C[11] et al. used noisy speech as input to the Deep Recurrent Neural Network (DRNN), trained the network model, and output relatively pure speech.

However, the structure of these neural network models is more complicated, the experimental efficiency is lower, and there is still more noise remaining in the experimental results. In order to make up for these shortcomings, this paper will do further research based on previous research.

III. SA-UNET SPEECH ENHANCEMENT ALGORITHM MODEL

A. Introduction to SA-Unet Model

The Unet neural network has been used in image segmentation. For the first time, Stoller D [11]et al. applied the Unet network to speech separation to inhibit the noise portion of speech. The design idea is to use the time domain spectrum of the mixed speech as the input and output of the network, and then separate it. Macartney C [13]et al. used the Unet network for speech enhancement and achieved good results. The model is derived from the Unet network model named after the U-shaped network structure. The Unet network model is designed based on the architectural idea of a full convolutional neural network (FCNN). The network structure is relatively simple. The down-sampling portion on the left side of the network is the encoding process, and the right up-sampling is the decoding process. Therefore, some scholars often refer to the Unet network structure as the encoder-decoder structure. Because the algorithm has been effectively applied in speech enhancement, it has attracted more scholars to explore the research of deep learning models in this field. This paper adds a self-attention mechanism on the basis of predecessors to compensate for the problem that the Unet network does not have high voice quality due to insufficient context information during up-sampling. This network structure consists of three parts. The first part is feature extraction, reading the internal structure of the data. The second part is the self-attention mechanism to interpret the dependence within the speech. The third part is the up-sampling part. After each up-sampling, the number of channels is halved, and then the speech features extracted by the down-sampling are spliced. The network structure called SA-Unet is shown in Figure 1.

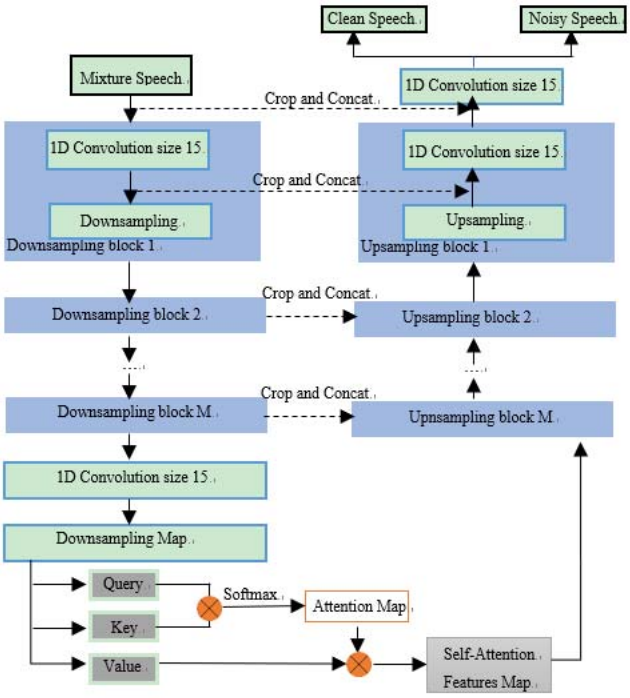

Figure 1. Network structure of SA-Unet

B. SA-Unet for speech enhancement

As can be seen from the above figure, the Self-Attention is added to the original Unet structure diagram to re-adjust the output characteristics of the encoder. This has the advantage of reducing the training model and increasing accuracy. Suppose the input sequence is $\lambda = [\lambda_1, \lambda_2, \lambda_3, \cdots, \lambda_a]$, the length of each sequence is l, the dimension of each vector is d, and the output dimension of Self-Attention is d', then the sequence of query vectors (Query), the input of the key vector sequence (Key) and the value vector sequence (Value) is:

$$Q_i, K_i, V_i = x_i \tag{1}$$

The three sequences in (1) are all from the input of the same sequence, and the above three vectors are converted as follows.

$$Q_i' = Q_i * W_Q \quad (2)$$

$$K_i' = K_i * W_K \quad (3)$$

$$V_i' = V_i * V_K \quad (4)$$

The query vector, the key vector and the value vector are linearly transformed by equations (2), (3), and (4), and the sequence is consistent with the original length. The dimension is converted from the original d to d', where W_Q, W_K, W_v are the weights randomly generated by Q_i, K_i, V_i.To prevent the result from being too large, divide the similarity between the query value and the key value by a scale $\sqrt{d_k}$, where d_k represents the vector dimension of the query and key, and then normalize it with softmax. As follows:

$$\alpha_i = softmax\left(\frac{Q_i' * K_i'^T}{\sqrt{d_k}}\right) \quad (5)$$

Where α_i is the attention weight corresponding to the sequence x_i. Multiply $\alpha_i$by the matrix V_i to get the attention mechanism feature map, as follows:

$$attention(query, key, value) = \alpha_i * V_i' \quad (6)$$

Then, after the formula (6), the attention map can be obtained to extract the features.

The SA-Unet network model is used in speech enhancement. The purpose is to separate the input mixed speech waveform $H \in [-1,1]^{L*C}$ into K speech source waveforms $S_1, S_2, \cdots, S_K$, where $S_k \in [-1,1]^{L*C}$, $k \in \{1,2,\ldots,K\}$. C represents the number of audio channels, and L represents the number of audio samples. In single channel speech, $C = 1$,. $K = 2$,

IV. EXPERIMENT ANALYSIS

A. Datasets

In order to ensure the fairness and impartiality of the experimental results, this experiment uses the same public datasets as [13]. The datasets are the VCTK speech datasets [14], a speech corpus established by the University of Edinburgh's Speech Research Center to better promote the development of speech enhancement and speech recognition research. The speech datasets collected a total of 30 native English testers with male and female voices. Among them, 28 people's speech were used for training datasets, and 2 people's speech were used for testing. The noise datasets are added to pure speech to generate mixed speech datasets according to the method provided in [15]. A total of 11,572 mixed speeches were generated as a training dataset. A total of 824 test speech datasets were used to test the quality of the training model speech.

B. Experimental setup

In this experiment, the environment is Ubuntu16.04, and the GPU is trained on GeForce GTX 1080Ti. The model is trained 2000 times with 512 data each time. In this paper, the mean square error is used as the loss function of the model training. An ADAM optimizer with a learning rate of 0.0001, a decay rate of β1=0.9 and β2=0.999 and a batch size of 16 is used to approximate the optimal value of the network model. In addition, 16 filters are added in each layer, a down-sampling block filter of size 15 and an up-sampled block filter of size 5. After 2000 iterations, the training is ended. Finally, the network model is evaluated using a noise speech test dataset.

C. Result analysis

In order to evaluate the training effect of SA-Unet network model objectively, the enhanced test set speech is compared with the original test set speech (824 speech), and the enhanced speech quality index is calculated. This article uses the following speech assessment metrics to evaluate speech quality. The higher the score of the assessment metric, the better.

- PESQ (Perceptual Assessment of Voice Quality): Subjective Voice Quality Assessment[16], it is proposed in ITU-TP.862 to objectively assess voice quality. The scope is$[-0.5, 4.5]$.
- CSIG: Mean opinion score (MOS) prediction of signal distortion only for speech signals[17]. The range is$[1, 5]$.
- CBAK: Prediction of the intrusiveness of background noise[17]. The range is$[1, 5]$.
- COVL: A prediction of the overall effect. The range is$[1, 5]$.
- SegSNR: Segmented SNR testing is often used for speech enhancement and speech coding. Calculated as follows:

$$SegSNR = \frac{10}{L}\sum_{l=0}^{L} \lg \frac{\sum_{m=lM}^{lM+M-1} s^2(m)}{\sum_{m=lM}^{lM+M-1} (s(m) - \hat{s}(m))^2}$$

Where L is the number of frames of speech and M is the frame length. The range is $[0, \infty]$

Through the test, the enhanced speech is compared with the test speech. The speech evaluation scores are objectively calculated using the above indicators and compared with other speech enhancement algorithms. The experimental results are shown in Table I:

TABLE I. COMPARISON OF SA-UNET VOICE QUALITY ASSESSMENT SCORES WITH OTHER METHODS

Metric	Noisy	Wiener	Wave-U-Net	SA-Unet
PESQ	1.97	2.22	2.40	2.66
CSIG	3.35	3.23	3.52	3.68
CBAK	2.44	2.68	3.24	3.71
COVL	2.63	2.67	2.96	3.16
SegSNR	1.68	5.07	9.97	11.95

In order to see the effect of speech enhancement more intuitively, a test case is randomly extracted. Through the Python visualization tool, the voice before and after enhancement can be displayed, as shown in Figure 2 and 3.

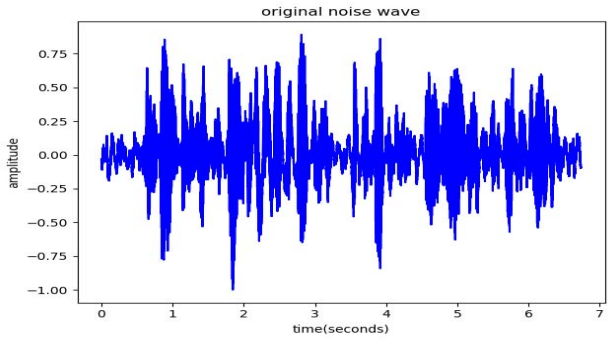

Figure 2 Original noise spectrum

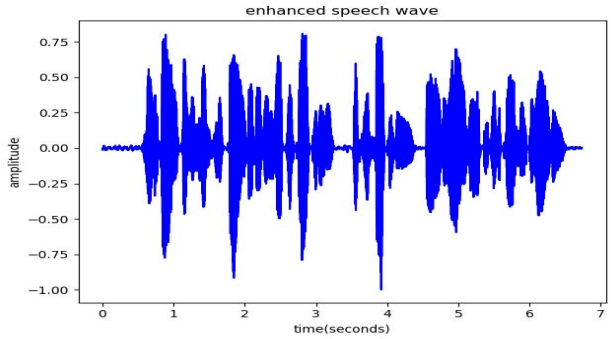

Figure 3 Enhanced speech spectrum

From the perspective of visualization, Figure 2 is a map of the original mixed speech without speech enhancement. The speech map shows a messy and mixed noise signal, which can have a great impact on the reception of pure speech signals. Figure 3 shows the speech spectrum after speech enhancement. It is obvious in the figure that the speech signal is very clear and clean, and there is almost no noise. This signal is very helpful for the reception of communication signals and the improvement of the accuracy of speech recognition.

V. CONCLUSIONS

This paper proposes a new SA-Unet network model structure for single-channel speech enhancement. The model structure is obtained by transforming the Unet network model. Due to the successful application of Unet network in speech separation, this paper adds Self-Attention mechanism to increase the context prediction ability and optimize the network model. Through a large number of test experiments, it is proved that the enhanced speech is more readable. Compared with the current mainstream speech enhancement algorithms, it shows outstanding results. Next, you can adjust the size and parameters of the network model further, and you may have better results. The next step is to apply SA-Unet to multi-channel speech enhancement, which may have good results.

REFERENCES

[1] P.C.Loizou, Speech Enhancement: Theory and Practice, 2nded. Boca Raton, FL, USA: CRC Press, Inc., 2013.

[2] Park S R, Lee J A Fully Convolutional Neural Network for Speech Enhancement[J]. 2016.

[3] Pascual S, Bonafonte A, Serrà, Joan. SEGAN: Speech Enhancement Generative Adversarial Network[J]. 2017.

[4] Paliwal K, Kamil Wójcicki, Schwerin B Single-channel speech enhancement using spectral subtraction in the short-time modulation domain[J]. Speech Communication, 2010, 52(5):450-475.

[5] Alam M J, O"Shaughnessy D. Perceptual improvement of Wiener filtering employing a post-filter[J]. Digital Signal Processing, 2011, 21(1):54-65.

[6] EPHRAIM. Speech enhancement using a minimum mean square error short-time spectral amplitude estimator[J]. IEEE Trans, Acoust. Speech Signal Process. 1984, 32(6):1109-1121.

[7] Nugraha A A, Liutkus A, Vincent E. Multichannel audio source separation with deep neural networks[J]. IEEE/ACM Transactions on Audio Speech & Language Processing, 2016, 24(9):1652-1664.

[8] Huang P S, Kim M, Hasegawa-Johnson M, et al. Joint Optimization of Masks and Deep Recurrent Neural Networks for Monaural Source Separation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 23(12):2136-2147.

[9] YUAN Wen-Hao, SUN Wen-Zhu, XIA Bin, OU Shi-Feng. Improving Speech Enhancement in Unseen Noise Using Deep Convolutional Neural Network. Acta Automatica Sinica, 2018, 44(4): 751-759.

[10] Yu Hua, Tang Yufeng, Zhao Li. An Advanced Speech Enhancement Algorithm Based on Deep Belief Network[J]. Journal of Data Acquisition & Processing, 2018, 33(05):29-36.

[11] Valentini-Botinhao C, Wang X, Takaki S, et al. Speech Enhancement for a Noise-Robust Text-to-Speech Synthesis System Using Deep Recurrent Neural Networks[C]// Interspeech 2016.

[12] Stoller D, Ewert S, Dixon S. Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation[J]. 2018.

[13] Macartney C , Weyde T . Improved Speech Enhancement with the Wave-U-Net[J]. 2018.

[14] Valentini-Botinhao, Cassia. (2017). Noisy speech database for training speech enhancement algorithms and TTS models, 2016 [sound]. University of Edinburgh. School of Informatics. Centre for Speech Technology Research (CSTR). https://doi.org/10.7488/ds/2117.

[15] Valentini-Botinhao C , Wang X , Takaki S , et al. Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech[C]// Isca Speech Synthesis Workshop. 2016.

[16] P.862.2: Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs, ITU-T Std. P.862.2, 2007.

[17] Y. Hu and P. C. Loizou, “Evaluation of objective quality measures for speech enhancement,” IEEE Trans. on Audio, Speech, and Language Processing, vol. 16, no. 1, pp. 229–238, Jan 2000.