Speech Signal Enhancement Based on Recurrent Neural Network

Cui Dongyan [1]

[1] School of Artificial Intelligence, North China University of Science and Technology ,Tangshan, Hebei, China

E-mail: *cdy_xxz@163.com*

Abstract

In real life, speech signals are often interfered by various types of noise in the surrounding environment, which greatly affects the performance of speech processing systems. Speech enhancement algorithms usually have significant suppression effect on stationary noise, but perform poorly in handling non-stationary noise. In addition, traditional speech enhancement algorithms often focus on specific noise environments, and they are difficult to meet the requirements of complex and changeable application scenarios. To overcome these problems, this article combines the advantages of deep learning, proposing a comprehensive speech enhancement algorithm that can adaptively learn the features of complex noise environments by utilizing data-driven technologies, such as recurrent neural network, to improve the effect and robustness of speech enhancement. Compared with traditional speech enhancement algorithms, this method can better handle non-stationary noise and has stronger universality and adaptability, achieving good results in complex and changeable noise environments.

Keywords: Speech Enhancement; Deep Learning; Denoising; Recurrent Neural Network.

1. Introduction

With the rapid advancement of computer and network technology, voice communication, as an important way of human communication, has been widely used in various fields such as individuals, businesses, and governments. Therefore, researching and developing voice communication technology is of great significance and value. However, in the process of voice communication in actual environments, different types of noise are often generated, which can lead to a decrease in the quality of the voice communication system. With the development of speech signal processing technology, speech enhancement technology has become an important solution for improving the performance of speech signals.

Before the emergence of deep learning, speech enhancement mainly relied on traditional signal processing methods. Typical traditional methods include time-domain method, frequency-domain method (such as spectral subtraction), wavelet transform method, etc. These methods have solved some basic noise suppression problems, but their universality and robustness in practical environments are limited. In 1985, Ephraim et al. further improved the MMSE-STSA algorithm and proposed a new algorithm, the MMSE-LSA algorithm [1]. In 1991, a speech denoising algorithm based on singular value decomposition (SVD) was proposed by Dendrinos and Bakamidis et al. It achieved certain results in speech denoising, but still had defects such as speech distortion [2]. Sanam proposed an adaptive thresholding method in 2013 to enhance the wavelet packet coefficients of noisy speech. This method utilizes the characteristics of Teager energy to adaptively determine the threshold and is applicable to different sizes of noise. It can improve the quality of speech signals in noisy environments and has practical application value in speech processing [3].

With the rise of deep learning technology, especially the breakthrough achievements of deep neural networks in many tasks, researchers have begun to attempt to apply deep learning methods to speech enhancement [4].

In 2018, Valin proposed the first generation model denoising algorithm, the RNNoise model, which was strongly influenced by traditional algorithms in its algorithm design, including manual feature extraction and dividing the model into speech activity detection (VAD), noise spectrum estimation, and spectrum subtraction parts [5]. Dacheng Y proposed the PHASEN algorithm in 2020, which is an algorithm that can accurately estimate signal amplitude and phase information [6]. In 2021, Wang et al. proposed a time-domain speech denoising model based on a two-stage Transformer, which is a model that reduces noise in speech by combining the Transformer model with time-domain information [7-8].

The development and application of deep learning technology have provided new solutions for speech signal processing, which are more accurate and reliable compared to traditional algorithms. This article will design a speech enhancement system based on deep learning technology, analyze the processing effects of different algorithms through experiments, and compare and evaluate them to obtain the optimal algorithm.

2. Principles of Speech Signal Enhancement Based on Deep Learning Algorithms

2.1. Network Architecture and Performance Analysis of RNN

RNN is a recursive neural network whose network architecture is based on the processing of sequential data. Unlike traditional feedforward neural networks, RNNs have a certain memory capability and can process sequence data by storing previous states. They perform well in tasks such as natural language processing, speech recognition, and time series prediction.

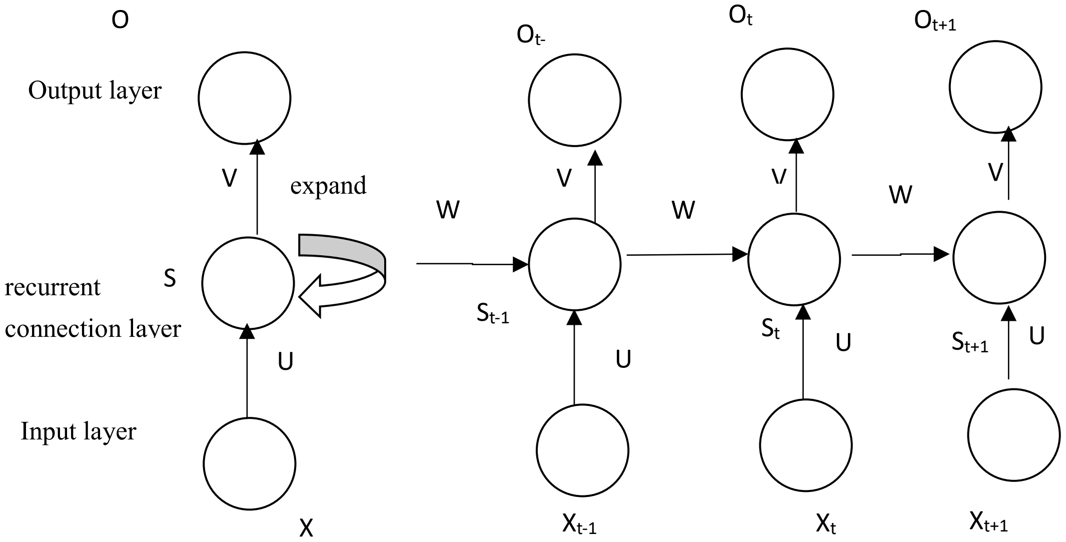


Fig.1 Network architecture of RNN

Using *t* to represent the current input sequence number, *t-1* to represent the sequence number of the previous input, X_t and X_{t-1} to represent the current and previous inputs, respectively, and S_t and S_{t-1} to represent the current and previous hidden states, the calculation formula for hidden states is:

$$S_{(t)} = f(W \times S_{t-1} + U \times X_t) \tag{1}$$

Among them, W and U are the weights of RNN, and f(.) represents the activation function. If O_t is used to represent the current output, the calculation formula is:

$$O_t = g(V \times S_t) \tag{2}$$

Among them, V is the weight of RNN, and g (.) represents the activation function.

The performance of RNN is influenced by multiple factors, including network depth, number of neurons, learning rate, activation function, and optimizer. Usually, increasing the depth of the network and the number of neurons can improve the performance of the model, but at the same time, it will increase training time and computational costs. Properly setting the learning rate and using appropriate activation functions and optimizers can accelerate the training process of the model and improve its accuracy.

2.2. Design of Speech Enhancement Method Based on Deep Learning

The main implementation steps of the RNN based enhancement method design are as follows:

1) Data preprocessing: Speech signal preprocessing to remove noise and noise.

2) Feature extraction: Extract MFCC or other feature parameters of speech signals and normalize them.

3) Build network structure: Adopt an RNN network structure suitable for speech enhancement, usually consisting of an input layer, RNN hidden layer, output layer, etc.

4) Training model: Using preprocessed training datasets, train the model through appropriate loss functions and optimization algorithms to obtain a high-precision speech enhancement model.

5) Model testing: Use test speech data to evaluate the performance of the model, which can be evaluated by calculating indicators such as signal-to-noise ratio (SNR) and speech quality.

6) Model optimization: Based on the test results, optimize and adjust the model to improve the speech enhancement effect.

3. Experimental Simulation and Result Analysis

3.1. Introduction to Experimental Environment

This chapter introduces the experimental environment used. The experiment was implemented using MATLAB programming and the interface was designed using MATLAB GUI, version R2020a. The computer configuration is Intel (R) Core (TM) i5-9300H CPU @ 2.40GHz, with 16GB of RAM.

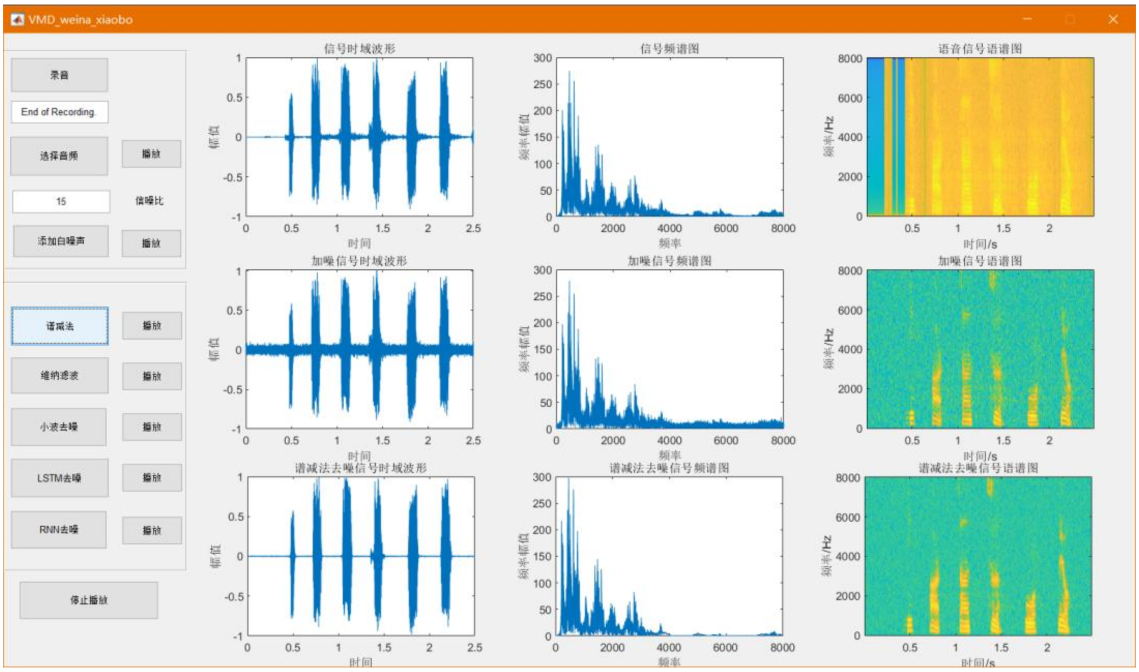
3.2. Analysis of Experimental Results

The analysis of experimental results using traditional methods is shown in the Figure 2.

In traditional algorithms, this experiment used three methods: spectral subtraction, Wiener filtering, and wavelet denoising. To evaluate the performance of these methods, we applied them to recorded speech data, enhanced and denoised them, and evaluated them based on the evaluation metrics mentioned earlier. The experimental results are shown in Table 1 below:

Tab.1 Data on various measurement indicators of traditional speech enhancement methods

Method	SNR	PESQ	RMSE	NSR
Spectral Subtraction	15.29	2.584	0.036	0.187dB
Wiener filtering	14.05	2.256	0.034	0.124dB
Wavelet denoising	14.23	2.138	0.037	0.176dB



End of Recording

15 信噪比

Fig.2 Simulation image of traditional speech enhancement method

According to Table 1, among traditional algorithms, spectral subtraction has the best noise reduction effect, with an SNR value of 15.29dB. At the same time, the PESQ value is also higher than the other two methods, reaching 2.584. The SNR value of the Wiener filtering method is 14.05dB, indicating that its noise reduction effect is slightly inferior to spectral subtraction. Although wavelet denoising method is slightly inferior to Wiener filtering and spectral subtraction in terms of denoising effect, it surpasses Wiener filtering method in NSR value, reaching 0.176.

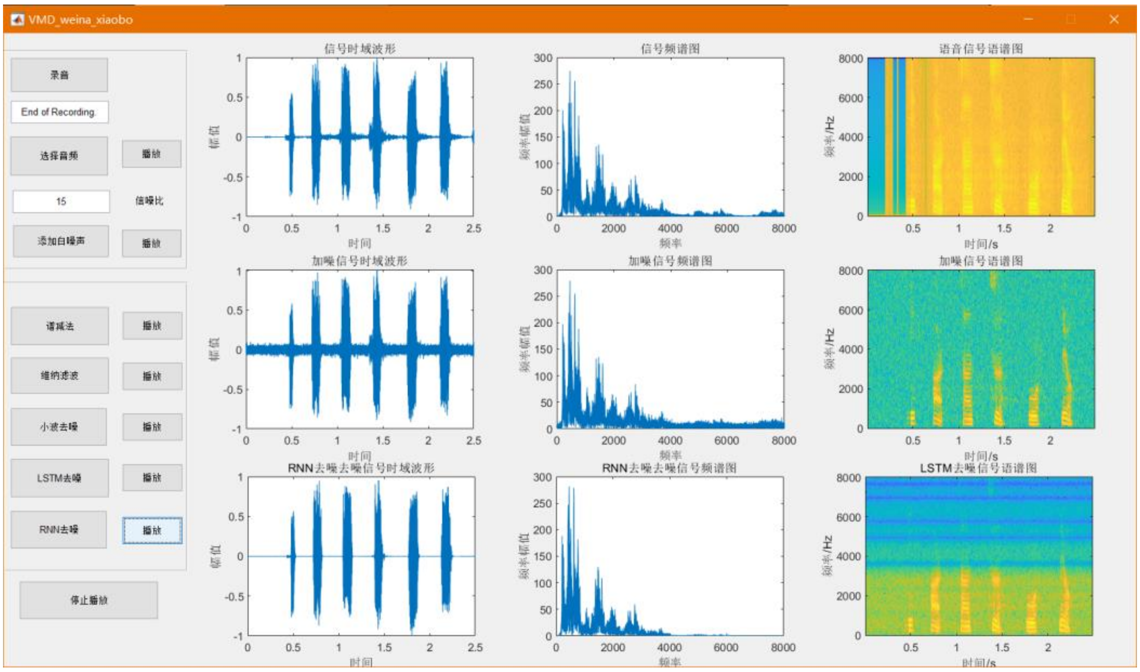
3.3. Analysis of Experimental Results of Enhancement Method Based on RNN

In order to demonstrate the effectiveness of the RNN based enhancement method, this chapter conducted experiments and evaluated it using the aforementioned evaluation metrics. The experimental results are shown in the Table 2:

Tab.2 Data of various indicators for LSTM based enhanced speech enhancement

Method	SNR	PESQ	RMSE
RNN	19.48dB	2.824	0.019

According to the data shown in Table 2, we can see that the RNN based enhancement method performs well in the field of short-term speech enhancement. Its SNR value reached 19.48dB, far higher than the performance of traditional algorithms, which means that the speech signal-to-noise ratio during subtitles has been significantly improved. At the same time, the PESQ value of this method also reached 2.824, surpassing all traditional methods. From the perspective of speech quality, it achieved excellent enhancement effects. In addition, the RMSE value is relatively small, at 0.019, indicating that the enhancement effect is quite good. The value of NSR is also relatively large, indicating that the denoising effect is also very significant.



End of Recording.

15 信噪比

Fig.3 Simulation data of speech enhancement based on RNN

According to Figure 3, overall, the RNN based speech enhancement method performs exceptionally well in the field of short-term speech enhancement. From a technical perspective, this method has shown good results in improving signal-to-noise ratio, speech quality, noise reduction effect, and enhancement effect. Therefore, in practical applications, RNN based speech enhancement methods have broad application prospects and are worthy of further in-depth research.

4. Conclusion

Based on the experimental results, we can see that the speech signal enhancement technology based on RNN performs outstandingly in various indicators, with excellent performance in evaluation indicators such as SNR, PESQ, RMSE, and NSR. Compared with traditional algorithms, it has significant advantages and has achieved good enhancement and noise reduction effects, which can better adapt to various noise scenarios and effectively improve the quality of speech signals. Therefore, we can conclude that RNN based speech enhancement methods perform well in the field of short-term speech enhancement and have broad application prospects.

References

[1] Ephraim Y, Malah D. Speech enhancement using a minimum-mean square error log-spectral amplitudeestimator. *IEEE Transactions on Acoustics Speech and Signal Processing*, 1985, 33(2): 443-445.

[2] Dendrinos M, Barkamidis S, Carayannis G. Speech enhancement from noise: Aregenerative approach. *Speech Communication*, 1991, 10(1): 45-57.

[3] Sanam T F,Shahza C.Noisy speech enhancement based on an adaptive threshold and a modified hardthresholding function in wavelet packet domain. *Digital Signal Processing*, 2013, 23(3): 941-951.

[4] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation // *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015: 3431-3440.

[5] Valin J. A hybrid DSP/deep learning approach to real-time full-band speech enhancement// *The 20th International Workshop on Multimedia Signal Processing*, IEEE, 2018: 1–5.

[6] Dacheng Y, Chong L, Zhiwei X, et al. A Phase-and-Harmonics-Aware Speech Enhancement Network //*Proceedings of the AAAI Conference on Artificial Intelligence*,2020, 34(05): 9458-9465.

[7] Wang K, He B, Zhu W P. TSTNN: Two-stage Transformer based neural network for speech enhancement in the time domain //*ICASSP'21.IEEE International Conference on Acoustics Speechand Signal Processing*. *IEEE*, 2021: 7098-7102.

[8] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need // *NIPS'17. International Conference on Neural Information Processing Systems. ACM*, 2017: 6000–6010.