Hindawi
Discrete Dynamics in Nature and Society
Volume 2020, Article ID 2196893, 10 pages
https://doi.org/10.1155/2020/2196893

Research Article

Speech Separation Using Convolutional Neural Network and Attention Mechanism

Chun-Miao Yuan, Xue-Mei Sun, and Hu Zhao

School of Computer Science and Technology, TianGong University, Tianjin 300387, China

Correspondence should be addressed to Xue-Mei Sun; sunxuemei@tiangong.edu.cn

Received 14 May 2020; Accepted 8 July 2020; Published 25 July 2020

Academic Editor: Jianquan Lu

Speech information is the most important means of human communication, and it is crucial to separate the target voice from the mixed sound signals. This paper proposes a speech separation model based on convolutional neural networks and attention mechanism. The magnitude spectrum of the mixed speech signals, as the input, has its high dimensionality. By analyzing the characteristics of the convolutional neural network and attention mechanism, it can be found that the convolutional neural network can effectively extract low-dimensional features and mine the spatiotemporal structure information in the speech signals, and the attention mechanism can reduce the loss of sequence information. The accuracy of speech separation can be improved effectively by combining two mechanisms. Compared to the typical speech separation model DRNN-2 + discrim, this method achieves 0.27 dB GNSDR gain and 0.51 dB GSIR gain, which illustrates that the speech separation model proposed in this paper has achieved an ideal separation effect.

1. Introduction

Voice information plays an increasingly important role in our lives, and voice communication becomes more and more frequent, such as using chatting software to send voice messages, using voice to control mobile phone applications, making mobile phone calls for voice communication, recognizing the singers from songs [1], and identifying singer's information, lyrics, and song style [2, 3]. The goal of speech separation is to separate the mixed speech into two original speech signals. In signal processing, speech separation is a basic task with a wide range of applications, including mobile communication, speaker recognition, and song separation. There are many potential values for separating mixed speech. Nowadays, speech separation plays a more and more important role in speech processing, and more and more devices need to carry out speech separation task.

Although humans can easily perform speech separation, it is very challenging to build an automatic system that matches the human auditory system. Therefore, speech separation has always been an important research direction of speech processing. The early speech separation method is very limited in the ability of mining nonlinear structure information, and the performance of monaural speech separation has been unsatisfactory. With the development of deep neural networks in recent years [4–6], good results have been achieved in various fields. Compared with traditional speech methods, there are many advantages for deep neural network-based speech separation models. The main contribution of this paper is to apply the convolution neural network to the speech separation tasks, use the multilayer nonlinear processing structure of the convolution neural network to mine the structure information in the speech signals, automatically extract the abstract features, integrate the attention mechanism to reduce the loss of the sequence information, and finally achieve the monaural speech separation.

2. Relative Research

Speech separation problem has been widely studied worldwide. The traditional methods of monaural speech

separation are speech enhancement, which analyze the data of the mixed speech and the noise and then estimate the target speech through the interference estimation of the mixed speech. From the viewpoint of signal processing, many speech enhancement methods propose a power spectrum of estimated noise or an ideal Wiener wave recorder, such as spectral subtraction [7] and Wiener filter [8]. There are many other ways to achieve speech enhancement (e.g., through Bayesian model [9] or time-frequency masking methods [10]). Some researchers tried to use computer auditory scene analysis (CASA) to achieve speech separation, which is the use of computer technology to allow computers to imitate the processing of human auditory signals for modeling, so as to have the ability to perceive sound, process sound, and interpret sound from complex mixed sound sources such as humans. The basic computational goal of auditory scene analysis is to estimate an ideal binary mask to achieve speech separation based on auditory masking of the human ears [11–13]. Speech enhancement technology generally assumes that the interference speech is stable, so its separation performance will be severely degraded when the actual interference noise is nonstationary. Computational auditory scene analysis, with better generalization performance, has no assumptions about noise. Meanwhile, the computational auditory scene analysis heavily relies on speech pitch detection, which is very difficult under the interference of background sounds.

Speech separation is designed to isolate useful signals from disturbed speech signals, a process that can be expressed naturally as a supervised learning problem [14]. A typical supervised speech separation system learns a mapping function from noisy features to separation targets (e.g., ideal masking or magnitude spectrum of speech interested) through supervised learning algorithms such as deep neural networks. Many supervised speech separation methods have been proposed in recent years [15–18]. The learning models for supervised speech separation are mainly divided into two kinds. (1) Traditional methods such as model-based methods and speech enhancement methods. (2) Newer methods using DNNs (Deep Neural Networks). Due to the speech generation mechanism, the input features and output targets of the speech separation show obvious spatiotemporal structure. These characteristics are very suitable for modeling with deep models. Many deep models are widely used in speech separation. Sun et al. [19] proposed two-stage approach with two DNN-based methods to address the problem of the performance of current speech separation methods. New training targets in complement of existing magnitude training targets were trained through neural network methods to compensate for phase of target in order to achieve better separation performance by the authors of [20]. Zhou et al. [21] designed a separation system on Recurrent Neural Network (RNN) with long short-term memory (LSTM), which effectively learns the temporal dynamics of spatial features. Supervised speech separation does not require spatial orientation information of the sound sources, and there are no restrictions on the statistical characteristics of noise. It shows obvious advantages and a fairly bright research prospective under the conditions of monaural, nonstationary noise, and low signal-to-noise ratio [22, 23].

Deep Recurrent Neural Network (DRNN) among them is a representative of deep models and has been widely used in speech separation. DRNN has strong learning ability in speech separation. RNN series of units, such as LSTM [24]/GRU (Gated Recurrent Unit, GRU) [25], all of whose hidden states are calculated according to the Markov model. The previous hidden state will save some previous information, while the magnitude spectrum of the mixed speech is a relatively long sequence, and sequence information will be lost (after all, the capacity of the cell state is limited in practice), which will affect the separation of mixed speech and reduce the accuracy of estimated speech.

Convolutional neural network (CNN) has been widely used in deep learning since it was firstly proposed by Yann Lecun et al. [26] in 1998. CNN has its natural advantages in two-dimensional signal processing, and its powerful modeling capabilities have been verified in tasks such as image recognition. CNN at present has been applied to speech separation [27, 28] and has achieved the best separation performance under the same conditions, which has exceeded the DNN-based speech separation systems. Luo et al. [29] proposed Conv-TasNet, which was a fully convolutional time-domain audio separation network to solve the problems of time-frequency masking. Wang et al. [30] applied CNN and modified its loss function to solve the imbalance between classification accuracy, hit rate, and error rate.

Both the traditional and DNN-based separation models have achieved good results, but they all have corresponding shortcomings. Convolutional neural network can make use of the spatial connection of input data so that each element can learn local features without learning global features, and then at a higher-level these local features will be combined together to get global features. Weight sharing can reduce the parameters between different neurons, reduce the calculation time, and improve the speed of the model. By using a variety of convolution filters, multiple feature maps can be obtained, which can detect the same type of features in different positions and ensure the invariance of displacement and deformation to a certain extent. Therefore, this paper proposes a method based on convolutional neural network to solve the problem of the loss of long sequence information of mixed speech. Combined with the attention mechanism, our model can better focus on the timing sequence step, which contributes the most, and solve the problem of insufficient memory of the temporal model to a certain extent, so as to improve the speech separation effect.

3. System Modeling

The purpose of the speech separation is to separate the target voice from the background sound. The overall structure of the speech separation system is shown in Figure 1. The first process of speech separation is to obtain the magnitude spectrum and phase information of mixed speech through STFT (Short-time Fourier Transform) [31]. The estimated targets using speech magnitude spectrum have been shown to suppress noise significantly and improve the speech

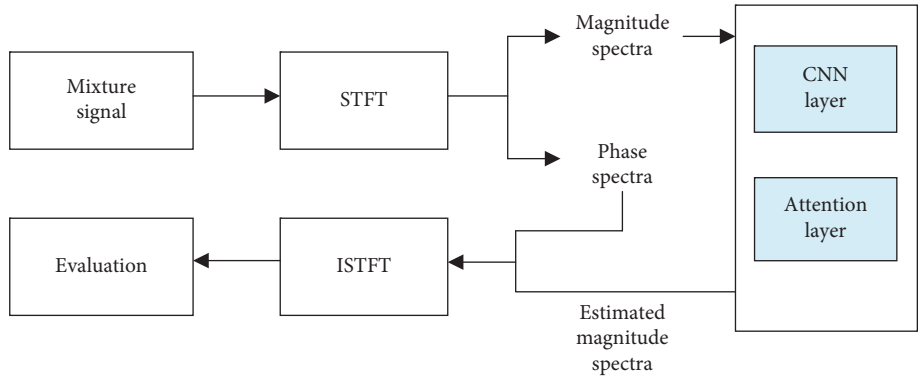

Figure 1: The modeling of speech separation.

intelligibility and its perceived quality. Then, the magnitude spectrum information is used as the input of the speech separation model. The magnitude spectrum is trained by convolutional neural network, and the region of interest of speech is extracted by the attention module. Finally, through the overlap-add method [32], the target speech is obtained by the combination of the magnitude spectrum information estimated by the separation model and the previous phase information. The speech separation ability of the model is evaluated based on the comparison between the estimated speech and the pure speech.

4. Proposed Methods

Deep models, with multilevel nonlinear structures, good at mining structural information in data, which can automatically extract abstract feature representations, have been widely used in image, video, and speech processing. This paper presents a monaural speech separation model based on the convolutional neural network and attention mechanism; we name it CASSM. There are two main modules in this model: convolutional neural network module and attention mechanism, both of which are jointly optimized to complete the task of speech separation. In the convolutional neural network module, it can extract features from the magnitude spectrum of the mixed speech, reduce the dimension of the magnitude spectrum, and model the current and contextual information effectively. Although convolutional neural networks can extract the magnitude spectrum features of mixed speech signals effectively, they cannot model sequence information in the same way. The attention mechanism RNN is good at processing sequence information; however, there are shortcomings in extracting information features versus convolutional neural networks. After the training of neural network, it can recognize the difference of different signal sources at different frame levels effectively; the attention mechanism-based neural network can recognize the importance of each part of the magnitude spectrum with the help of context information [33], thus considerably improving the effect of speech separation.

4.1. Network Structure. The network structure of the model is shown in Figure 2. The magnitude spectrum of the mixed speech is used as the input of the neural network, the estimated magnitude spectrum is obtained by training the neural network, and the original phase is combined to get the estimated speech by ISTFT (inverse short-time Fourier transform).

The amplitude spectrum of mixed speech has high dimensionality when it acts as the input of neural network. CNN is good at mining the spatiotemporal structure information of the input signal and can extract the local feature information of the magnitude spectrum, which can improve the performance of speech separation.

The magnitude spectrum of the mixed speech is used as the network input and then forward propagates along two paths, one of which is used as the input of the convolutional layer, while the other forms the combination sequence of the input layer with the low-dimensional sequence that has been processed by the convolution. The input information transmits through two convolutional filters with different sizes and important information about the mixed speech magnitude spectrum is able to be extracted.

We use relu as an activation function. The output of the two convolutions is superimposed and processed by the maximum pooling layer, which provides strong robustness, increases the invariance of the current information, speeds up the calculation speed, and prevents overfitting. The sequence continues through several convolutional layers that can be effectively modeled for current as well as contextual information to obtain a low-dimensional embedding sequence. The low-dimensional and high-dimensional embedded sequences are superimposed into a combined one, which passes through two-tier high-speed networks to extract higher-level features.

Still, there are many shortcomings in processing serialized information for speech separation although convolutional neural networks can extract features from the magnitude spectrum of mixed speech, reduce the dimension of the magnitude spectrum, and model the current and contextual information effectively. Comparatively, the attention mechanism RNN is much skilled in processing serialized information, helpful in strengthening the dependence between the magnitude spectrum, and can jointly fulfill the task of speech separation with the convolutional neural network module.

The attention mechanism is used for the input of the mixed speech magnitude spectrum so that the model pays different attention to the information features from different

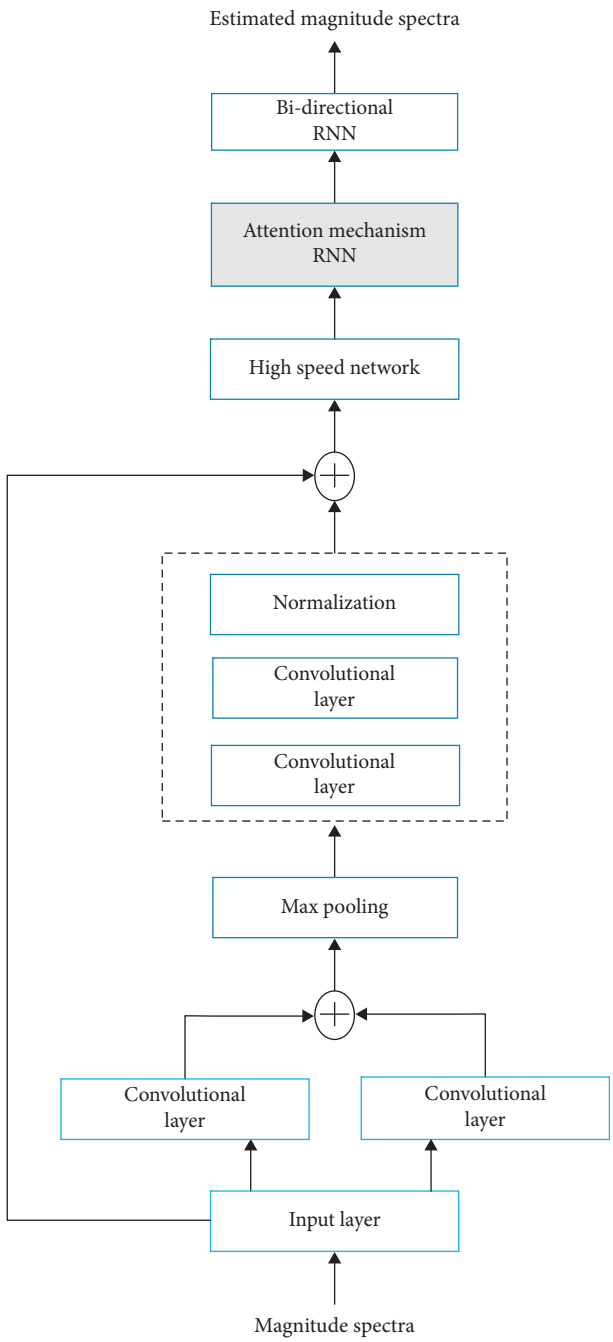

Figure 2: Network architecture of proposed CASSM.

moments. The traditional recurrent neural network uses only the information of state [t−1] when calculating the state [t] at the t moment. Since the magnitude spectrum is a long sequence, there will be a loss of information although the previous state [t−1] contains some previous information. The attention mechanism uses the previous information to calculate the state information at the current moment. The attention mechanism used in this paper is based on a layer of RNN (LSTM unit) with an attention length of 32.

Long Short-Term Memory (LSTM) is a specific recurrent neural network (RNN) architecture that was designed to model temporal sequences and their long-range dependencies more accurately than conventional RNNs. LSTM RNNs are more effective than DNNs and conventional RNNs for acoustic modeling, considering moderately sized models trained on a single machine [34]. LSTM RNN makes more effective use of model parameters than the others considered, converges quickly, and outperforms a deep feedforward neural network having an order of magnitude more parameters.

A bidirectional recurrent neural network and two feedforward layers are used to form a separation module after the attention mechanism RNN.

4.2. Short-Term Fourier Transform. The spectrum of speech signals has its significant role in speech, through which, some nontrivial speech features can be obtained. Stationary signals refer to the signals that the distribution law does not change over time, while nonstationary signals are signals that change over time. The mixed speech signals belong to the nonstationary signals, and the speech signals have the characteristics of nonstationary one.

The Fourier transform (FT) is ubiquitous in science and engineering. For example, it finds application in the solution of equations for the flow of heat, for the diffraction of electromagnetic radiation, and for the analysis of electrical circuits. The concept of the FT lies at the core of modern electrical engineering and is a unifying concept that connects seemingly different fields. The STFT is a fine resolution to the problem of determining the frequency spectrum of signals with time-varying frequency spectra. Fourier transform is the process of conversion between time domain and frequency. The Fourier transform is in its essence, a reversible transform that can transform the original signals and the transformed signals into each other. The characteristics of the Fourier transform are not suitable for analyzing nonstationary signals, which cannot directly represent speech signals. The speech signals change slowly with the change of time, which even can be regarded as unchanged in a short period of time. Analysis of the frequency domain information in a short time is the short-time Fourier transform.

Short-time Fourier transform (STFT) is a commonly used tool for processing speech signals. There is only one difference between the Fourier transform and the short-time Fourier transform. The short-time Fourier transform is to divide the speech signals into sufficiently small fragments so that the speech signals can be seen as a stable one. The short-time Fourier transform formula is defined as

$$\mathrm{STFT}(t,\omega) = \int_{-\infty}^{\infty} s(x)\gamma(x-t)e^{-j\omega x}dx. \quad (1)$$

In which, $\gamma(t)$ is the length M window function, $s(x)$ denotes an input signal at time x to be transformed, and STFT (t, ω) is a two-dimensional complex function, which represents the magnitude and phase that change over time and frequency.

In supervised speech separation, feature extraction is an indispensable process, and the selection of features will affect the speech separation model training. From the point of the extracted basic units, the features of speech separation are mainly divided into time-frequency unit-level features and frame-level

ones. Frame-level features are extracted from a frame of signals. This level of features has a larger granularity and can capture the spatiotemporal structure of speech, especially the frequency band correlation of speech. It is more general and holistic and has clear characteristics of speech perception.

The feature extraction in this paper is based on the frame level. The short-time Fourier transform is used to frame and window the speech signals. In each frame, the magnitude spectrum and phase in this paper are obtained by 1024-point STFT. The speech separation model can better display the information of speech by using frame-level features and can mine the space-time structure of speech in all directions during training. The magnitude spectrum obtained by using the short-time Fourier transform in this paper can make model training easier and accelerate the model convergence.

4.3. Attention Mechanisms. In recent years, the attention mechanism has been widely used by many scholars in deep learning models in different fields, such as speech recognition and image recognition. Due to the limitations of information processing, when much information is processed, people usually choose some of the important parts for processing, while ignoring other parts of the information. This idea is the origin of the attention mechanism. When people are reading, they will pay more attention to and deal with some important vocabulary and ignore some less important parts, which will speed up the reading and improve reading efficiency. Attention mechanism now has become a widely used tool; thus, attention mechanism is one of the crucial important algorithms in deep learning algorithms that need the most attention and understanding. However, attention mechanism is currently not widely used in speech separation; therefore, this paper attempts to introduce attention mechanism into the process of speech separation, focusing on the region of interest in speech feature information, so as to improve the accuracy of speech separation.

Strictly speaking, attention mechanism is an idea rather than an implementation of some models, which can be implemented in a variety of ways. Due to the bottleneck of information processing, attention mechanism requires to decide which part may be the focus, allocate the limited information processing resources to more important parts, and specially focus on the key information.

There are many variants of the attention mechanism, and the attention mechanism used in this paper is AttentionCellWrapper in TensorFlow [35], which is a general attention mechanism. This attention structure can be used through a one-way recurrent neural network. While processing the input of each step, it will consider the output of the previous N steps and add the previous historical information to the prediction of the current input through a mapping weighting method:

$$\begin{aligned} \mathscr{U}_i^t &= \mathscr{V}^T \tan h\left(W_\alpha h_t + W_\beta o_i\right), \\ a_i^t &= \text{soft}\max\left(\mathscr{U}_i^t\right), \\ h_t' &= \sum_{i=1}^{t-1} a_i^t o_i. \end{aligned} \tag{2}$$

In which, the vector $\mathscr{V}$ and matrix W_α, W_β are the learnable parameters of the model in the formula, h_t is the matrix of the current hidden-layer state of the LSTM and the LSTM unit connection, and o_i is the output of the ith LSTM unit. The attention weights a_i^t of each moment are normalized calculated by the softmax function. Finally, connecting h_t with h_t' becomes the predicted new hidden state and is also fed back to the next step of the model.

It becomes possible for the simple timing model to use the attention mechanism by using this general attention structure designed by TensorFlow, which can make the entire model more focus on the timing step that contributes the most, and solve the timing model with memory problems to some extent.

4.4. Time-Frequency Masking Calculation Layer. Considering the constraints of TF (time-frequency) masking between the forced input mixed speech signals and the estimated speech signals as well as the benefits of smooth separation, a time-frequency masking calculation layer is used in the speech separation model to jointly optimize the TF model with the entire deep learning model. The time-frequency masking calculation layer is defined as

$$\begin{aligned} \widetilde{y}_1 &= \frac{\left|\widehat{y}_1\right|}{\left|\widehat{y}_1\right| + \left|\widehat{y}_2\right|} \odot z, \\ \widetilde{y}_2 &= \frac{\left|\widehat{y}_2\right|}{\left|\widehat{y}_1\right| + \left|\widehat{y}_2\right|} \odot z, \end{aligned} \tag{3}$$

where $\widetilde{y}_1$ and $\widetilde{y}_2$ are the magnitude spectrums of the estimated speech signals; $\widehat{y}_1$ and $\widehat{y}_2$ are the output magnitude spectrums of the two feedforward networks of the speech separation module; and z is the magnitude spectrum of the input mixed speech signals.

The masking effect is a phenomenon that occurs when the human ears perceive speech. A louder sound will mask a lower one. If the difference between the two sound frequencies is small, the effect of the masking effect becomes more obvious. The masking effect is of great significance in speech processing.

4.5. Loss Function. The loss function is used to describe the difference between the predicted value and the target value. The loss function is generally positive, and the size of the loss function reflects the quality of the model. When the model performs speech separation, the value of the loss function is relatively small if the model speech separation works well. The size of the loss function directly reflects the speech separation effect of the model. The loss function of the training network model in this paper is defined as [36]

$$\begin{aligned} J_{\text{DIS}} = &\left\|y_1 - \widetilde{y}_1^2\right\| + \left\|y_2 - \widetilde{y}_2^2\right\| \\ &- \gamma\left\|y_1 - \widetilde{y}_1^2\right\| - \gamma\left\|y_2 - \widetilde{y}_2^2\right\|, \quad \gamma > 0, \end{aligned} \tag{4}$$

where y_1 and y_2 are the ground truth of magnitude spectrum, $\widetilde{y}_1$ and $\widetilde{y}_2$ are the estimated magnitude spectrum of speech, and $0 \le \gamma \le 1$ is a regularization parameter.

5. Simulation Experiments and Analysis

In this paper, experiments were performed on a Windows 10 Professional 64-bit computer with hardware configuration of i5, 8GB RAM, and GT1060X 6G graphics card. The program was written in PyCharm software using Python.

The MIR-1K dataset [37] was used to evaluate the model. The dataset has 1,000 mixed music clips encoded at a 16 KHz sampling rate, with a duration from 4 to 13 seconds. The clips were extracted from 110 Chinese karaoke songs performed by both male and female amateurs. The singing voice and background music were mixed into a music fragment with a signal-to-noise ratio of 0 dB. In this dataset, 175 clips were used for the training sets, and 825 were used for the testing sets.

Speech separation usually uses three parameters of BSS-EVAL [38] to verify the performance of the model: SDR (Signal to Distortion Ratio), SIR (Signal to Interference Ratio), and SAR (Signal to Artifact Ratio). SDR is the ratio between the power of mixed speech signals and the difference value of the mixed speech signals and the target speech signals. SIR is the total power ratio between the target speech signals and the interference signals. SAR stands for artifacts introduced by processing speech signals. SDR calculates how much total distortion exists in the separated sound. A higher SDR value indicates a smaller overall distortion of the speech separation system. SIR directly compares the degree of speech separation between nontarget sound and target sound. A higher SAR value indicates a smaller introduced error of the speech separation system. A higher value of the three parameters indicates a better effect of the speech separation model.

In this experiment, in order to evaluate the effect of the speech separation model, the global NSDR (GNSDR), global SIR (GSIR), and global SAR (GSAR) were used to evaluate the overall effect, all of which were the mean value of the tested fragments based on their lengths calculated. The normalized SDR (NSDR) [39] is defined as

$$\mathrm{NSDR}(\hat{o}, o, m) = \mathrm{SDR}(\hat{o}, o) - \mathrm{SDR}(m, o). \tag{5}$$

In which, o is the original pure signal, $\hat{o}$ is the estimated speech signal, and m is the mixed speech signals. NSDR is used to estimate the SDR improvement of mixed speech signal m and speech signal $\hat{o}$.

We conducted multiple comparative experiments to comprehensively evaluate the validity and reliability of the proposed model. Firstly, we tested the influence of LSTM and GRU gating units on the effect of speech separation. Secondly, we verified the improvement of the separation effect of the traditional speech separation models by the attention mechanism. Then, we compared the separation effect between DRNN-2 + discrim proposed by Huang et al. and CASSM proposed in this paper. Finally, we explored the effect of attention length on the effect of speech separation.

5.1. The Influence of LSTM and GRU on Speech Separation Model. The recurrent neural network has the function of short-term memory, but if the training sequence is long enough, it is difficult for the network to transmit the more important information from the previous to the later process. Therefore, if the speech problem is processed, the recurrent neural network will miss some important information, resulting in a reduction in processing power and a negative impact on speech processing results. Another problem is that, in the later backpropagation phase, the recurrent neural network will encounter the problem of gradient disappearance, which is the weight value of updating each layer in the network. Gradient disappearance means that the gradient value will continue to decrease as the training proceeds, and when the gradient drops to a very small amount, the neural network will not continue to learn, and the training process will stop, which may affect the training results.

In order to solve these problems, two methods, LSTM unit (Long Short-Term Memory unit) [24] and GRU (Gated Recurrent Unit) [25], have been proposed through the efforts of many researchers. Both types of units have internal mechanisms, gates, which can regulate the flow of information to achieve the function of memory. These two gating units have different characteristics. In our experiments, the effect from the two types of gating units, LSTM and GRU, on the results of the speech separation model was evaluated.

Firstly, the effect on the separation model was reflected by using the different gating units of the three-layer unidirectional recurrent neural network layer, RSSM (RNN-based Speech Separation Model). The speech separation results of each model were described by three parameters: GNSDR, GSIR, and GSAR. The experimental results of the two gating units are shown in Figure 3.

The effect of different gating units was analyzed from the experimental results in the Figure 3. The speech separation model using LSTM gating units obtained a gain of 0.07 dB on the GNSDR, a gain of 0.39 dB on the GSIR and a gain of −0.09 dB on the GSAR relative to the GRU.

The ASSM (Attention-based Speech Separation Model) was mainly used to analyze the effect of attention mechanism on speech separation and to verify the effectiveness of attention mechanism in speech separation. Next, the ASSM was used to evaluate the effect from the two gated units, LSTM and GRU, on the results of the speech separation model in the attention mechanism. The ASSM was mainly used to analyze the effect of attention mechanism on speech separation and verify the effectiveness of attention mechanism in speech separation. ASSM was composed of fully connected layer, attention mechanism RNN, and bidirectional RNN.

The experimental results in Figure 4 showed three indicators of speech separation. In the experiment, the ASSM using the LSTM unit was 0.05 dB higher in GNSDR and 0.37 dB in GSIR than that of using GRU unit. From the experimental results of the two groups, it can be concluded that LSTM is superior to GRU on GNSDR and GSIR. Therefore, LSTM was employed in our proposed model in terms of the attention mechanism.

5.2. CASSM Performance Verification Experiment. This experiment mainly aimed to verify the effect of the speech separation model proposed in this paper. Since the results of

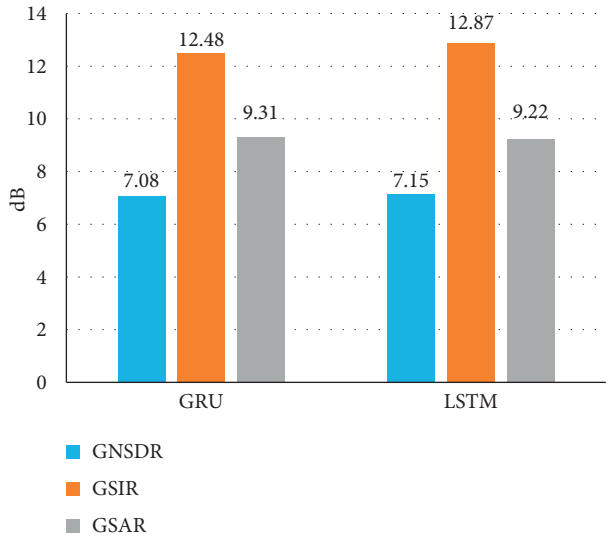

Figure 3: The comparison for the effect of GRU and LSTM in RSSM.

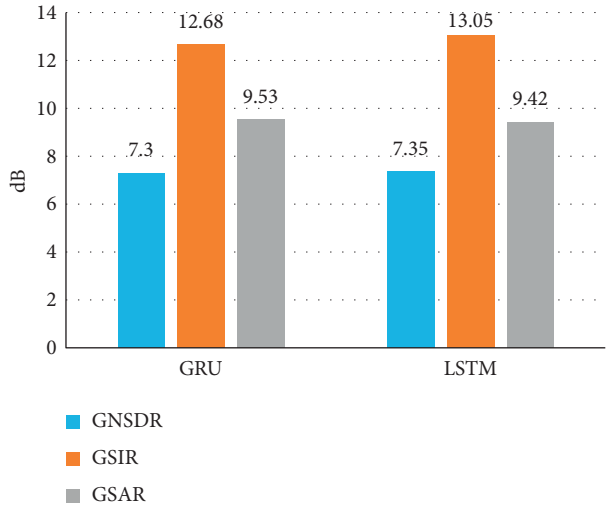

Figure 4: The comparison for the effect of GRU and LSTM in ASSM.

the speech separation model based on recurrent neural network were not very satisfactory, inspired by attention mechanism and convolutional neural network, this paper proposed a speech separation model based on attention mechanism and convolutional neural network, CASSM, and the effectiveness of the model would be verified through the following experiments.

Huang et al. [36] applied the recurrent neural network to speech separation and proposed a speech separation variant model based on deep recurrent neural network. The speech separation effect varies from different models employed. In previous research, masking of monochannel signal separation and the joint optimization of deep recurrent neural network have achieved better speech separation performance. In our experiment, a comparison was made among these models: CASSM, the DRNN-2 + discrim model with the best separation effect achieved by Huang et al. [36], RNMF (Robust Low-Rank Nonnegative Matrix Factorization), and MLRR (Multiple Low-Rank Representation).

The attention mechanism used was LSTM unit with an attention length of 32. The experimental results are shown in Figure 5. Compared with MLRR and RNMF, the value of CASSM was 3.87 dB and 2.75 dB higher in GNSDR, 7.96 dB and 5.93 dB higher in GSIR, and −1.06 dB and -0.39 dB lower in GSAR. Experiments demonstrated that convolutional neural network plus attention mechanism had greater advantages over traditional speech separation models. Convolutional neural network can more fully extract sound spectrum features. Attention mechanism can strengthen the dependence between magnitude spectrum features and improve the speech separation performance. Compared with DRNN-2 + discrim, the improvement of CASSM was 0.27 dB higher in GNSDR and 0.51 dB higher in GSIR. Experiments illustrated that there was still a gap between DRNN-2 + discrim and CASSM in terms of processing amplitude spectra. In DRNN-2 + discrim, the original magnitude spectrum of the mixed speech was directly used as the input features, while the CNN module of CASSM used the combined sequence formed by the high-dimensional magnitude spectrum of the mixed speech as the input features. Meanwhile, a better speech separation effect in the experiment was achieved since the attention mechanism module of CASSM had reduced the loss of sequence information.

The "discrim" denotes the models with discriminative training

5.3. Ablation Experiment. In this section, we conducted ablation experiments to verify the effectiveness of our proposed model. Firstly, we removed the convolutional neural network module in our model and observed the speech separation results of only attention module; then, we continued to remove the attention module and observed the speech separation result of the model with only recurrent neural network.

In our experiment, the attention mechanism used was LSTM unit with an attention length of 32, and RNN was a unidirectional RNN with 1000 hidden layers. The effect of each model was analyzed from the experimental results. After removing the CNN module, the GNSDR, GSIR, and GSAR of the model decreased by 0.37 dB, 0.66 dB, and 0.18 dB, respectively, while the attention module was removed, the GNSDR, GSIR, and GSAR of the model declined by 0.57 dB, 0.72 dB, and 0.42 DB, respectively. On the whole, the speech separation performance was substantially decreased.

The results of ablation experiments show that, after removing CNN and attention mechanism, respectively, the speech separation ability of the model has decreased to varying degrees (Figure 6).

5.4. Effect of Attention Length on Speech Separation. This simulation experiment explored the effect of attention length on speech separation through a speech separation model

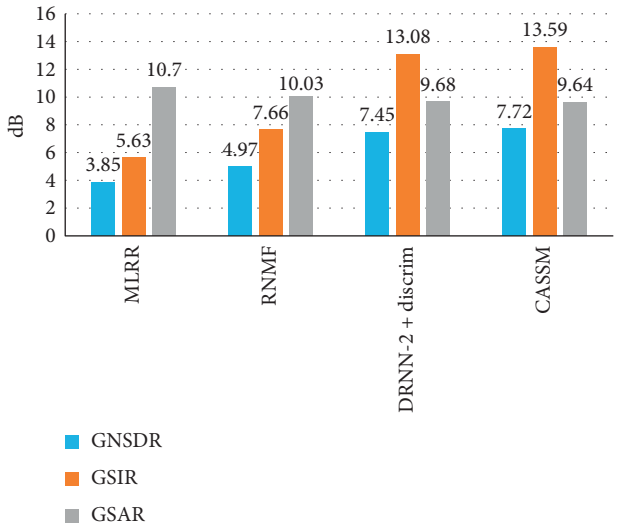

Figure 5: The comparison for the effect of MLRR, RNF, DRNN-2 + discrim, and CASSM.

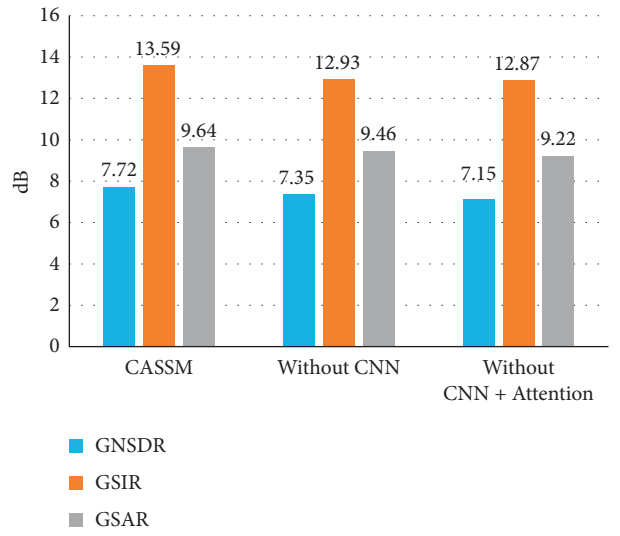

Figure 6: The comparison for the effect of ablation experiments.

based on convolutional neural network and attention mechanism. In this experiment, the attention mechanism used in this paper was based on a one-way RNN (hidden layers are 1000 layers, LSTM units); the length of attention was 8, 16, and 32, respectively.

The experimental results in Figure 7 showed that the speech separation model with attention length of 16 was 0.09 dB higher in GNSDR, 0.2 dB in GSIR, and 0.03 dB higher in GSAR than the speech separation model with attention length of 8. Compared with the speech separation model with attention length of 16, the speech separation model with attention length of 32 was 0.06 dB higher in GNSDR, 0.37 dB higher in GSIR, and 0.1 dB lower in GSAR.

The experimental results indicated that the attention length of the attention mechanism had taken a relatively large effect on speech separation. The effect of speech

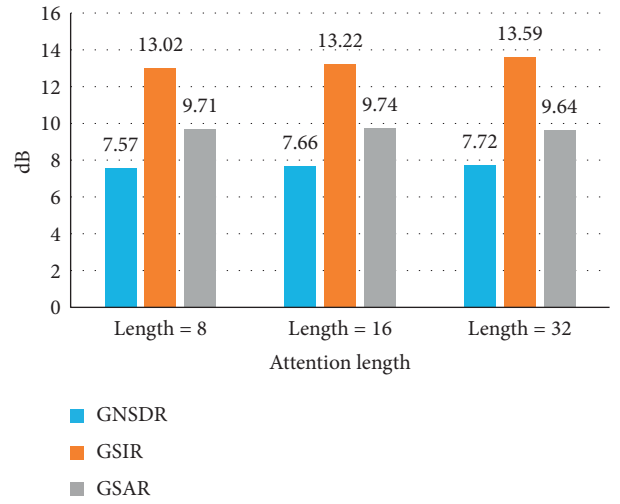

Figure 7: The comparison for the effect of different attention lengths.

separation is getting better and better with the increase of attention length. It can be concluded from the simulation experiment that the increase of attention length could improve the effect of speech separation.

The simulation results also illustrated that the training time would increase with the increase of attention length and the demand for video memory would relatively increase; therefore, a reasonable choice of attention length is highly recommended. There are still some limitations and shortcomings in the experiment. Due to the limitation of equipment, the attention length of the speech separation model could only be adjusted to 32. If it continues to increase, the device may report an insufficient memory. Therefore, research and discussion on this indicator in future experiments can be carried out.

6. Conclusion

This paper proposed and implemented a speech separation model based on convolutional neural network and attention mechanism. Convolutional neural network can effectively extract low-dimensional features and mine the spatiotemporal structure information in speech signals. Attention mechanism can reduce the loss of sequence information. The accuracy of speech separation can be effectively improved by combining two mechanisms. The simulation experiments illustrated that the model had greater advantages over the speech separation model based on recurrent neural network, and the effect of speech separation can be improved through joint optimization of convolutional neural network and attention mechanism.

Data Availability

All data have been included in this article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Program for the Science and Technology Plans of Tianjin, China, under Grant no. 19JCTPJC49200.

References

[1] T. L. Nwe and H. Li, "Exploring vibrato-motivated acoustic features for singer identification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 2, pp. 519–530, 2007.

[2] W. H. Tsai and C. H. Ma, "Triangulation-based singer identification for duet music data indexing," in *Proceedings of the IEEE International Congress on Big Data*, Anchorage, AK, USA, July 2014.

[3] M. I. Mandel, G. E. Poliner, and D. P. W. Ellis, "Support vector machine active learning for music retrieval," *Multimedia Systems*, vol. 12, no. 1, pp. 3–13, 2006.

[4] M. Hermans and B. Schrauwen, "Training and analyzing deep recurrent neural networks," in *Proceedings of the International Conference on Neural Information Processing Systems*, Curran Associates Inc., Daegu, South Korea, November 2013.

[5] A. Krizhevsky, I. Sutskever, and G. Hinton, *ImageNet Classification with Deep Convolutional Neural Networks*, Curran Associates Inc., Red Hook, NY, USA, 2012.

[6] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," https://arxiv.org/abs/1409.0473.

[7] 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.

[8] J. Chen, J. Benesty, Y. Huang et al., "New insights into the noise reduction Wiener filter," *IEEE Transactions on Audio Speech & Language Processing*, vol. 14, no. 4, pp. 1218–1234, 2006.

[9] Y. Laufer and S. Gannot, "A bayesian hierarchical model for speech enhancement with time-varying audio channel," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 225–239, 2019.

[10] S. Samui, I. Chakrabarti, and S. K. Ghosh, "Time-frequency masking based supervised speech enhancement framework using fuzzy deep belief network," *Applied Soft Computing*, vol. 74, pp. 583–602, 2019.

[11] R. Cusack, J. Decks, G. Aikman, and R. P. Carlyon, "Effects of location, frequency region, and time course of selective attention on auditory scene analysis," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 30, no. 4, pp. 643–656, 2004.

[12] Y. Chen, "Single channel blind source separation based on NMF and its application to speech enhancement," in *Proceedings of the 2017. IEEE 9th International Conference on Communication Software and Networks (ICCSN)*, pp. 1066–1069, Guangzhou, China, May 2017.

[13] Z. Li, Y. Song, L. Dai, and I. McLoughlin, "Listening and grouping: an online autoregressive approach for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 4, pp. 692–703, 2019.

[14] Y. X. Wang, A. Narayanan, and D. L. 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.

[15] L. I. U. Wen-Ju, S. NIE, S. Liang et al., "Deep learning based speech separation technology and its developments," *Acta Automatica Sinica*, vol. 42, no. 6, pp. 819–833, 2016.

[16] 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.

[17] J. Zhou, H. Zhao, J. Chen, and X. Pan, "Research on speech separation technology based on deep learning," *Cluster Computing*, vol. 22, no. S4, pp. 8887–8897, 2019.

[18] L. Hui, M. Cai, C. Guo, L. He, W. Zhang, and J. Liu, "Convolutional maxout neural networks for speech separation," in *Proceedings of the IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, IEEE, Abu Dhabi, UAE, December 2015.

[19] Y. Sun, W. Wang, J. Chambers, and S. M. Naqvi, "Two-stage monaural source separation in reverberant room environments using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 125–139, 2019.

[20] C. P. Wang and T. Zhu, "Neural network based phase compensation methods on monaural speech separation," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1384–1389, Shanghai, China, July 2019.

[21] L. Zhou, S. Lu, Q. Zhong, Y. Chen, Y. Tang, and Y. Zhou, "Binaural speech separation algorithm based on long and short time memory networks," *Computers, Materials & Continua*, vol. 63, no. 3, pp. 1373–1386, 2020.

[22] P. Haffner, L. Bottou, and Y. Bengio, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[23] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Proceedings of the 2014 IEEE Global Conference on Signal and Information Processing*, pp. 577–581, IEEE, Atlanta, GA, USA, December 2014.

[24] J. Schmidhuber and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[25] K. Cho, B. Van Merrienboer, C. Gulcehre et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, Doha, Qatar, October 2014.

[26] Y. Wang and D. Wang, "A deep neural network for time-domain signal reconstruction," in *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4390–4394, IEEE, South Brisbane, Australia, April 2015.

[27] Z. Shi, H. Lin, L. Liu et al., "Furcax: end-to-end monaural speech separation based on deep gated (de) convolutional neural networks with adversarial example training," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Brighton, UK, May 2019.

[28] A. J. R. Simpson, G. Roma, and M. D. Plumbley, "Deep karaoke: extracting vocals from musical mixtures using a convolutional deep neural network," in *Proceedings of the 12th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2015)*, Springer-Verlag, New York, NY, USA, August 2015.

[29] Y. Luo and N. Mesgarani, "Conv-TasNet: surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[30] J. X. Wang, S. B. Li, H. M. Jiang, and X. X. Bian, "Speech separation based on CHF-CNN," *Computer Simulation*, vol. 36, no. 5, pp. 279–283, 2019.

[31] D. Gabor, "Theory of communication. Part 1: the analysis of information," *Journal of the Institution of Electrical Engineers—Part III: Radio and Communication Engineering*, vol. 93, no. 26, pp. 429–441, 1946.

[32] J. Allen, "Short term spectral analysis, synthesis, and modification by discrete Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 3, pp. 235–238, 1977.

[33] Z. Yang, D. Yang, C. Dyer et al., "Hierarchical attention networks for document classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1480–1489, San Diego, CA, USA, June 2016.

[34] M. Delfarah and D. Wang, "Deep learning for talker-dependent reverberant speaker separation: an Empirical Study," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1839–1848, 2019.

[35] J. Cheng, L. Dong, and M. Lapata, "Long short-term memory-networks for machine reading," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, TX, USA, November 2016.

[36] P. S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2136–2147, 2015.

[37] C. L. Hsu and J. S. Jang, "On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 310–319, 2010.

[38] C. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[39] R. Gribonval, P. Philippe, and F. Bimbot, "Adaptation of bayesian models for single-channel source separation and its application to voice/music separation in popular songs," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 5, pp. 1564–1578, 2007.