Contents lists available at YXpublications

International Journal of Applied Mathematics in Control Engineering

Journal homepage: http://www.ijamce.com

Speaker Speech Separation Combining CNN-BLSTM and Self-Attention Mechanism

Yongqiang Zhang, Haixin Zhu, Wanzhen Zhou

[a] *School of Information and Control Engineering, Qingdao University of Technology*

ARTICLE INFO

Article history:
Received 15 September 2022
Accepted 24 December 2022
Available online 25 December 2022

Keywords:
Speech separation
self-attention mechanism
CNN-BLSTM
time-frequency unit
speech signal

ABSTRACT

Speech separation is to separate the interesting target speech from the mixed speech signal and improve the speech quality. In view of the fact that the time-frequency feature of speech has a global correlation, this paper proposes a method of introducing self-attention mechanism based on CNN-BLSTM to realize the task of speech separation. Combined with the self-attention mechanism, this method optimizes the deep time-frequency features extracted by CNN-BLSTM, so that the time-frequency units dominated by target speech get more attention, so as to better distinguish the time-frequency units of clean speech and noise, so as to improve the effect of speech separation. Experiments are carried out on TIMIT data sets, and two indexes, signal-to-noise ratio (source-to-distortion ratio, SDR) and short-term intelligibility (short-time objective intelligibility, STOI), are used to evaluate. The experimental results show that the overall separation performance of the model with the self-attention mechanism is significantly improved.

1. Introduction

Voice interaction is widely used in daily life, but in the process of interaction, the speech signal mixed with noise will reduce the accuracy of information transmission. In real life, in addition to the voice of the target speaker, there are usually other voices of the speaker [1]. For noisy speech signals, speech separation technology can improve the performance of speech interaction and improve the quality of speech signals. Therefore, how to separate the target speech from noise has become a research focus.

Speech separation methods in the time-frequency domain can be divided into two categories: traditional methods and deep learning methods. The traditional methods are based on signal processing [2-4], based on computational auditory scene analysis[5] and based on shallow models [6-8], but none of them can learn the deeper nonlinear features in speech data automatically. it is highly dependent on feature selection, so it is difficult to meet the application in real life. The deep learning method can effectively learn the deep representation of the original speech data, the requirement of feature selection is not high, and can use the nonlinear mapping ability to achieve the extraction of noisy speech to a pure speech, which is more in line with the real application scene. Therefore, the method based on deep learning has gradually become the mainstream method of speech separation technology.

Due to the speech signal having an obvious Spatio-temporal structure and nonlinear relationship, the deep learning algorithm represented by Deep Neural Network (Deep Neural Network, DNN) has a multi-level nonlinear processing structure and is good at mining deep information in speech data, which significantly improves the performance of speech separation[9]. In 2014, the algorithm of applying DNN to speaker separation task was proposed by HUANG et al for the first time[10]. The masking function is added to the original output layer of the network as an additional processing layer, and the target speech amplitude spectrum is estimated from the mixed speech amplitude spectrum by time-frequency masking. In 2015, HUI et al proposed a CNN model using Maxout as the activation function, which separates speech by estimating the IRM of time-frequency units[11]. In 2018, NAITHANI et al proposed a new loss function based on LSTM, which significantly improves the performance under the application of low delay and optimizes the objective intelligibility[12]. In the same year, EPHRAT et al fused video streams in the AV model to improve separation performance by making use of the high temporal correlation between auditory information and visual information[13]. In 2019, WANG et al.[14] integrated voiceprint recognition to obtain a priori information to improve the speech quality of the target speaker. In 2021, Guo et al.[15] further improved the speech quality and intelligibility of target speech by reducing the structure of LSTM units combined with attention mechanism and

* Corresponding author.
E-mail addresses: zyq@hebust.edu.cn (Y. Zhang)
Doi:

suppressing the influence of noise-dominated time-frequency units on the separation effect.

In recent years, deep learning method has shown great potential in solving the task of speaker separation, but it can not well take into account the global correlation of speech time-frequency features, which will affect the speech quality after separation to a certain extent. In order to solve the above problems, this paper proposes a method of introducing self-attention mechanism into the CNN-BLSTM model, which can effectively pay attention to the global correlation of input speech features in the time-frequency domain. Finally, the quality of the speech after separation and reconstruction is measured by experiments, and the separation performance of the method is evaluated objectively.

2. Data processing

2.1 Time-frequency decomposition

The daily collected speech signal is a time-domain waveform signal, so it is difficult to find the frequency characteristics of the speech signal. In this paper, short-time Fourier transform (STFT) is used for time-frequency decomposition, and the one-dimensional time-domain signal is transformed into a two-dimensional time-frequency signal by framing, windowing and fast Fourier transform. STFT such as Formula (1).

$$STFT(t,f) = \int_{-\infty}^{+\infty} [z(\mu)g(\mu - t)]\mathrm{e}^{-j2\pi f\mu} d\mu \quad (1)$$

In the formula, g *(t)* denotes the window function, and $\boldsymbol{STFT}(\boldsymbol{t},\boldsymbol{f})$ is a one-dimensional time-domain signal, the t-frame of the time frame and the short-time Fourier transform coefficient of the f-band.

2.2 Feature extraction

The selection of features has a great influence on the final separation effect of mixed speech. In this paper, STFT is used to decompose the speech signal in time-frequency. Each time-frequency obtained contains complex real and imaginary parts, both of which are used as inputs to the model.

2.3 Separate target

In this paper, the ideal floating masking (Complex ideal ration mask, cRM) in complex domain considering the phase information of speech signal is used as the training target[16]. CRM has real and imaginary parts, so estimate them separately, such as (2) and (3).

$$M_r = \frac{X_r Y_r + X_i Y_i}{X_r^2 + X_i^2} \quad (2)$$

$$M_i = \frac{X_r Y_i - X_i Y_r}{X_r^2 + X_i^2} \quad (3)$$

Where the real parts of X_r and Y_r are the real parts of X and Y respectively, and the imaginary parts of X_r and Y_r are the imaginary parts of X and Y, respectively. X is the STFT coefficient of the mixed speech at the time frame t and the time-frequency unit of the frequency f, and Y is the STFT coefficient of the target speech in the corresponding time-frequency unit.

M_r and M_i are usually between -1 and 1, which makes it more difficult to estimate. Therefore, this article uses the sigmoid function to compress the value between [0,1].

After obtaining the cRM, combined with the phase information of the original mixed speech, the speech waveform signal is reconstructed by inverse short-time Fourier transform.

3. Model building

3.1 Separate model framework

This section uses the SACNN-BLSTM method to separate the mixed speech of two speakers, and constructs a speech separation framework, as shown in fig.1. Take the features of the mixed speech after STFT as the input of inflated convolution, extract the local correlation of the input features, connect the bi-directional LSTM network, obtain the long-term dependence of speech time-frequency features, and then take the output of the BLSTM layer as the input of the self-attention layer, pay attention to the important global correlation, calculate the weight for the time-frequency, and assign values to the important time-frequency features, so as to get a more accurate classification. After the estimated two speakers' independent complex domain time-frequency masking (cRM) is obtained and multiplied by the input mixed, the respective spectrograms of the two speakers are obtained, and then the inverse Fourier transform is performed to get a clean speech waveform signal.

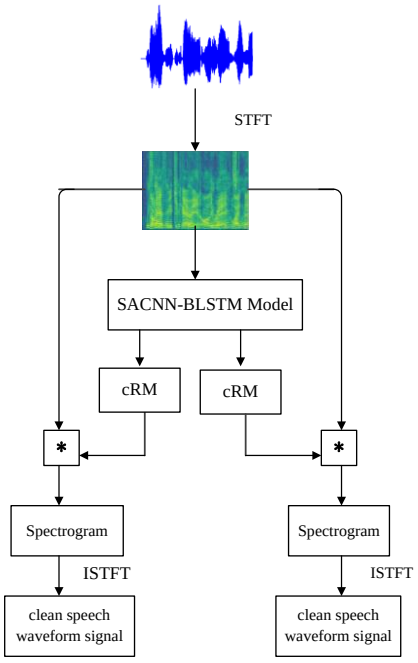

Fig. 1 Separate model framework

3.2 SACNN-BLSTM model

The SACNN-BLSTM model is formed by integrating the attention mechanism into the compound model formed by the combination of Dilated CNN and BLSTM. The network structure diagram is shown in fig.2. In this model, Dilated CNN has the characteristics of weight sharing and local connection, which can capture the local correlation of data features while reducing the number of network parameters. Moreover, Dilated CNN can increase the receptive field without increasing the number of parameters. The BN layer normalizes the calculated results of the convolution layer in batch to speed up the speed of network training and convergence. There is relevance before and after the time series data. BLSTM can effectively obtain the context information of the data through the memory function, and realize the feature extraction of long-time dependent data.

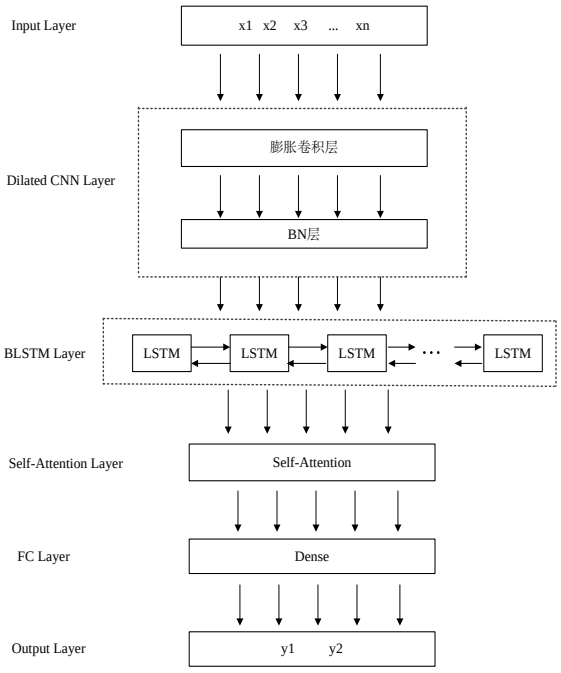

Fig. 2 SACNN-BLSTM model network structure

The self-attention mechanism[17] can further optimize the features of the BLSTM output, calculate the similarity between the input speech time-frequency units, use the Softmax function to normalize the score and give weight to the features, and pay attention to the global correlation of the frequency latitude. Due to the result of one operation is not accurate and can not well distinguish the time-frequency characteristics between the two signals, the multi-head attention mechanism is adopted in the SACNN-BLSTM model, the self-attention with 8 heads is used for parallel operation, and the calculated weights of 8 times are superimposed to obtain more accurate weight coefficients. The input matrix is mapped to different subspaces, and the data is encoded in the form of scaled point product, so as to realize the repeated calculation and weight superposition of the similarity between the input time-frequency units. so that there is a clear distinction between the speech signals of two different speakers. Therefore, the self-attention mechanism is introduced into the CNN-BLSTM network structure for modeling, and a more accurate time-frequency masking after separation will be obtained.

4. Experiment and result analysis

4.1 Experimental data

The pure speech data set used in this experiment is the TIMIT benchmark corpus. The speech of two different speakers is randomly selected from the TIMIT as clean speech, and the two voices are randomly mixed to generate 11000 mixed speech, of which 8000 are used as the training set and 3000 as the verification set. The training set does not coincide with the test set, and the test set randomly selects two clean voices to generate 100 mixed speech and tests the model performance and generalization ability.

4.2 Data preprocessing

In order to generate a speech separation model independent of the speaker. Firstly, the pure speech signal in the TIMIT data set is divided into 3 seconds. The speech signals of two different people are randomly mixed and used as the mixed speech needed in the experiment. The mixed audio is sampled with the sampling frequency of 16KHZ. As long as the single channel is used, the length of the 25ms Hanning window is 10ms and the step size is 10ms. Each time-frequency of the short-time Fourier transform (STFT) contains a complex real imaginary part as input.

4.3 Hyperparameter setting

The Learning rate, Epochs, Batch size, Optimizer and Dropout settings for the SACNN-BLSTM model are shown in Tab.1.

Tab.1 Hyperparameter setting

Hyperparameter	Set up
Learning rate	0.001
Epochs	100
Batch size	8
Optimizer	Adam
Dropout	0.5

4.4 Evaluation criteria

The use of different evaluation indicators has different effects on the evaluation results. The evaluation index of speech separation effect is mainly divided into two categories: signal aspect and perception aspect[18].

In the aspect of the signal, signal-to-noise ratio (signal-to-distortion ratio, SDR) is generally used, which is one of the commonly used performance standards and reflects the overall separation performance. The larger the SDR value, the better the separation effect. The formula is defined as formula (4).

$$SDR = 10log_{10}\left(\frac{\|s_{target}\|^2}{\|e_{interf} + e_{noise} + e_{artif}\|^2}\right) \quad (4)$$

In the formula，s_{target} is the real target voice，e_{interf} is the error disturbed by other sources，e_{noise} is due to the error caused by the noise，e_{artif} is the interference error of the system itself.

In terms of perception, short-term objective intelligibility (short-time objective intelligibility, STOI) is generally used as an evaluation index[19]. The value range of STOI index is between [0,1]. The higher the value is, the better and clearer the intelligibility of the speech is. The calculation formula is shown in (5).

$$STOI = \frac{1}{TF}\sum_{\sigma=1}^{F}\sum_{\tau=1}^{T} r_{m,k}(t) \quad (5)$$

In the formula, T is the total number of frequency bands; F is the total number of Fram; r is the local correlation coefficient of the time-frequency point.

4.5 Results and analysis

This section implements a multi-speaker separation algorithm based on deep learning, compares the advantages and disadvantages of the proposed algorithm (SACNN-BLSTM) and the traditional deep neural network algorithm (CNN-BLSTM) from two aspects of signal and perception, and uses bold to mark better experimental results. Among them, SDR is mainly used to analyze the overall performance of the speech separation algorithm, and STOI is used to measure speech articulation after separation and reconstruction.

Tab. 2 Average SDR of different methods

Model	SDR(dB)	STOI
CNN-BLSTM	10.74	0.96
SACNN-BLSTM	**11.50**	**0.96**

It can be seen from Tab.2 that the SDR of the SACNN-BLSTM

model is significantly higher than that of the baseline model, and on this basis, the STOI values are basically the same, which shows that it is feasible to use the SACNN-BLSTM model for multi-speaker separation tasks. This model can improve the overall separation performance of the model without changing the short-term intelligibility of speech, and significantly improve the quality of clean speech signal after separation and reconstruction to a certain extent.

Tab. 3 Average SDR of different gender mixtures

Model	CNN-BLSTM		SACNN-BLSTM	
Gender	SDR(dB)	STOI	SDR(dB)	STOI
Male- Male	11.48	0.97	**11.51**	**0.97**
Female-Female	9.85	0.95	**10.58**	**0.95**
Male-Female	11.11	0.96	**12.08**	**0.96**

Tab.3 lists the SDR and STOI values of the SACNN-BLSTM model and the CNN-BLSTM model proposed in this paper, respectively, in the case of mixed speech of male and male speakers, the mixed speech of male and female speakers, and mixed speech of female and female speakers. In any case, the separated speech obtained by the SACNN-BLSTM model is significantly higher than that obtained by the CNN-BLSTM model. It is proved that the method proposed in this paper is feasible and effective.

According to the analysis in Tab.3, the mixed speech formed by the clean speech mixing of speakers of different genders is separated by the same separation model, and the quality of the separated speech signal is different. Among them, the mixed speech formed by the clean speech of the female speaker and the clean speech of the other speakers of the same sex, the speech signal obtained after the separation model has the worst overall separation performance and short-term intelligibility in the mixing of male and male speakers, male and female speakers and female and female speakers.

In the SACNN-BLSTM model, the difference between the mixed speech of female and female speakers and the separated speech signal of male and female speakers with the highest SDR value is 1.5 (dB), and that of the separated speech signal of male and male speakers with the highest STOI value is 0.02. Among them, the mixed speech combination with the best overall separation performance is male and female, and the SDR is as high as 12.08 (dB). The mixed speech combination with the best short-term intelligibility of separated and reconstructed speech is male and male, with a value of 0.97. It can be seen that both signal optimization and perception can be considered in the future, that is, the values of SDR and STOI are the highest at the same time.

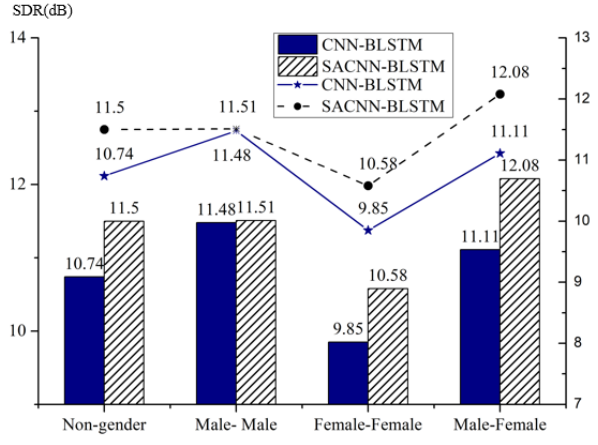

Fig. 3 SDR value analysis of different genders

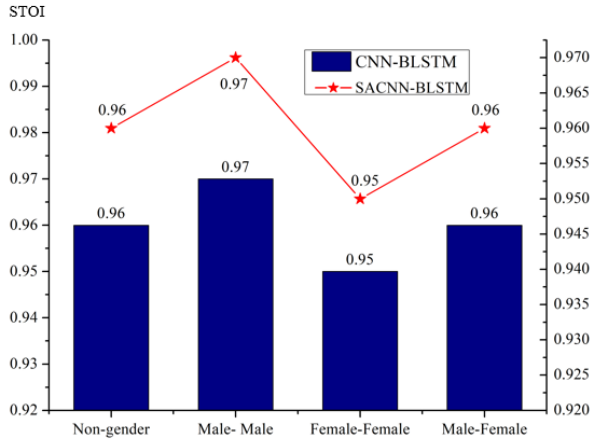

Fig. 4 Analysis of STOI values of different genders

Combined with fig.3 and fig.4, we can see that the overall separation performance of the SACNN-BLSTM model is always significantly higher than that of the CNN-BLSTM model in the case of male and male speakers, male and female speakers, female and female speakers, or comprehensive speakers. This shows that the self-attention mechanism optimizes the deep time-frequency features extracted by the neural network, so that the target speech-dominated time-frequency unit gets more attention, thus suppressing the noise-dominated time-frequency unit, so as to get a better separation effect. Therefore, the integration of self-attention mechanism in the field of speech separation is effective.

5. Conclusion

Considering the time-frequency characteristics of speech signal, this paper introduces the self-attention mechanism is introduced into the CNN-BLSTM. Through the self-attention mechanism to deal with the features extracted by the CNN-BLSTM, we can capture the global correlation of speech in the time-frequency domain and pay more attention to the time-frequency information dominated by target speech. Through the comparative analysis of the evaluation indexes of signal and perception, it is shown that this method further improves the overall separation performance of multi-speaker speech signals. However, the effect of mixed speech separation between girls and girls is relatively poor, and there is some room for improvement. In the next stage of research, the network structure can be improved according to the harmonic characteristics of girls' voice to improve the performance of speech separation.

Acknowledgments

The work has been partly supported by Natural Science Foundation of Hebei Province (F2018208116); Hebei Science and Technology Support Plan Project (16210312D); Soft Science Funding Project of Shijiazhuang Science and Technology Bureau (195790055A).

References

Cherry E C. Some experiments on the recognition of speech. withone and with two ears [J]. Journal of the Acoustical Society of America. 1953.25(5) : 975-979.

Boll S F . Suppression of acoustic noise in speech using spectral subtraction[J]. Acoustics Speech & Signal Processing IEEE Transactions on, 1979, 27(2):113-120.

Chen J, Benesty J, Huang Y, et al. New insights into the noise reduction Wiener filter[J]. IEEE Transactions on Audio Speech and Language Processing, 2006, 14(4):1218-1234.

Gerkmann T, Hendrikes R C. Unbiased MMSE-Based Noise Power Estimation With Low Complexity and Low Tracking Delay[J]. IEEE Transactions on Audio Speech & Language Processing, 2012, 20(4):1383-1393.

Weintraub, Harold. Assembly and propagation of repressed and derepressed chromosomal states[J]. Cell, 1985, 42(3):705-711.

Guo X W, Diao M F, Zheng C S, et al. Research on distributed speech Separation method based on complex Gaussian mixture Model[J/OL]. Journal of Signal Processing:1-11[2021-03-09].

Kun H, DeLiang W. Towards Generalizing Classification Based Speech Separation.[J]. IEEE Trans. Audio, Speech & Language Processing,2013,21(1).

Mysore G J, Smaragdis P . A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics[C]// IEEE International Conference on Acoustics. IEEE, 2011.

Liu W J, Nie S, Liang S, et al. Research status and Progress of speech Separation Technology based on Deep Learning[J]. ACTA AUTOMATICA SINICA, 2016,42(06):819-833.

Huang P S，Kim M，Hasegawa-Johnson M，et al.Deep learning for monaural speech separation.In: Proceedings of the 2014 IEEE International Conference on Acoustics，Speech，and Signal Processing. Florence: IEEE，2014.1562-1566.

Hui L, Cai M, Guo C, et al. Convolutional maxout neural networks for speech separation[C]// IEEE International Symposium on Signal Processing & Information Technology. IEEE, 2015.

Naithani G, Nikunen J, Bramslw L, et al. Deep neural network based speech separation optimizing an objective estimator of intelligibility for low latency applications[C]// 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC). IEEE, 2018.

Ephrat A, Mosseri I, Lang O, et al. Looking to listen at the cocktail party [J]. ACM Transactions on Graphics (TOG), 2018, 37(4):

Wang Q, Muckenhirn H, Wilson K, et al. VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking; proceedings of the Interspeech 2019, F, 2019 [C].

Donald S. Williamson, Yuxuan W, DeLiang W. Complex ratio masking for monaural speech separation[J]. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP),2016,24(3):

Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need[J]. arXiv, 2017.

Wang D L, Chen J T. Supervised Speech Separation Based on Deep Learning: An Overview.[J]. IEEE/ACM transactions on audio, speech, and language processing,2018,26(10):

Taal C H, Hendriks R C, Heusdens R, et al. An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech[J]. IEEE Transactions on Audio Speech & Language Processing, 2011, 19(7):2125-2136.

Zhang Yongqiang born in 1981, is a PhD candidate, an associate professor, and a master's supervisor. He was graduated from Anhui University of Technology and received master's degree of Computer Application Technology. Main research areas: artificial intelligence, Internet of things technology.

Zhu Haixin born in 1997, Studying in Computer Technology at Hebei University of Science and Technology. The main research directions are: Speech signal processing, artificial intelligence.

Zhou Wanzhen born in 1966, PhD, professor. He was graduated from the Ordnance Engineering College of the Chinese People's Liberation Army in 2009 with a doctorate degree in engineering. He studied at Harbin Institute of Technology with a master of science degree. His main research directions are: network and database, Internet of Things.