EURASIP Journal on Audio,
Speech, and Music Processing

METHODOLOGY

Open Access

Time-domain adaptive attention network for single-channel speech separation

Kunpeng Wang[1*], Hao Zhou[1], Jingxiang Cai[1], Wenna Li[1] and Juan Yao[1,2]

Abstract

Recent years have witnessed a great progress in single-channel speech separation by applying self-attention based networks. Despite the excellent performance in mining relevant long-sequence contextual information, self-attention networks cannot perfectly focus on subtle details in speech signals, such as temporal or spectral continuity, spectral structure, and timbre. To tackle this problem, we proposed a time-domain adaptive attention network (TAANet) with local and global attention network. Channel and spatial attention are introduced in local attention networks to focus on subtle details of the speech signals (frame-level features). In the global attention networks, a self-attention mechanism is used to explore the global associations of the speech contexts (utterance-level features). Moreover, we model the speech signal serially using multiple local and global attention blocks. This cascade structure enables our model to focus on local and global features adaptively, compared with other speech separation feature extraction methods, further boosting the separation performance. Versus other end-to-end speech separation methods, extensive experiments on benchmark datasets demonstrate that our approach obtains a superior result. (20.7 dB of SI-SNRi and 20.9 dB of SDRi on WSJ0-2mix).

Keywords Speech separation, Adaptive attention, Convolutional block attention, Transformer

1 Introduction

Speech separation, originating from the famous “cocktail party problem” [1, 2] and aiming to estimate the target speech from multi-person mixed speech, is widely applied in such fields as automatic speech recognition, speaker recognition and mobile communications [3, 4]. It is a difficult task to construct an automatic system like the human auditory system which can easily focus, in complex acoustic surroundings, on the target speech and automatically filter out irrelevant sounds. In this paper, due attention is paid to the single-channel speech separation, which is a more challenging task than multi-channel speech separation owing to the lack of spatial location information.

In recent years, important advances achieved in supervised speech separation, especially with the development of deep learning, have greatly promoted the single-channel speech separation technology [5–8]. Some research in deep-learning based speech separation techniques focus on time-frequency (T-F) domain methods [9–15]. These approaches first apply a short-time Fourier transform (STFT) to obtain a T-F representation of the speech, then separate the T-F features of each target. The target waveforms are then reconstructed by applying the inverse STFT to the separated features. However, the T-F-based approach has limited separation performance because it ignores clean phase information and uses noisy phases for time-domain signal reconstruction [16]. To address such an issue, [17] proposed the time-domain speech separation network (TasNet), which, with no time-frequency domain transformation required, directly modeled the time-domain speech signals, trained the time-domain

*Correspondence:
Kunpeng Wang
kwang@swust.edu.cn
[1] School of Information Engineering, Southwest University of Science and Technology, Mianyang, China
[2] Department of Automation, University of Science and Technology of China, Hefei, China

loss with PIT [11] and reconstructed the speech waveform. Most of the current mainstream speech separation explorations are conducted based on TasNet, such as Conv-TasNet [18], FurcaNeXt [19], DPRNN [20], Wavesplit [21], DPTNet [22], and SepFormer [23], all of which have made great progress compared with T-F methods. Directly modeling speech signals in the time domain avoids the problem of increased computational complexity caused by using STFT and reduces the delay in the speech signal processing process. Moreover, the time domain method eliminates the frequency decomposition step, thereby avoiding phase errors when reconstructing the pure speech signal.

More recently, attention-based speech separation methods have received much interest. In [22] and [23], self-attention based networks [24] are applied in single-channel speech separation and have achieved SOTA results, which demonstrate its powerful capacity for long sequence modeling. Worth noting is that the self-attention network is unique in its obtaining the output by calculating entire segments of speech information or more attention is paid to the long-term dependence of the sequences at different time scales. However, owing to some subtle local details (such as temporal or spectral continuity, which refers to the smoothness and consistency features of speech signals in time or frequency domain, and spectrogram structure, timbre, etc.) existing in speech separation [25, 26], the self-attention network may not result in the well extraction of these local information, which led to our motivation of using an additional network to improve the model's ability to capture local details in speech.

Based on this idea, a single-channel speech separation structure based on the time-domain adaptive attention network (TAANet) is proposed, with N identical local and global attention blocks stacked, and each block contains a local attention network and a global attention network. In a local attention network, we used the CBAM structure [27]; the channel attention and spatial attention mechanism are introduced to achieve more complete model description of the details of speech features. The channel attention module filters each channel and pays more attention to the correlation information between different encoders at the same time. The spatial attention module mainly focuses on the short-term context feature information in blocks corresponding to different speakers. In a global attention network, the self-attention mechanism is adopted to explore the global association of speech contexts [22, 24]. Moreover, all the blocks are connected serially, so that the model could adaptively focus on the local information and the global information at each feature level. With this adaptive attention network, the model could effectively extract the local information and the global information of the speech features, thus further promoting the separation performance. Compared with other end-to-end speech separation methods, extensive experiments on benchmark datasets demonstrate that our approach is able to improve the separation performance. In general, the main contributions of this paper are as follows:

- A new end-to-end single-channel speech separation model is constructed, which uses the local attention module and the global attention module to model the speech sequence, and combines the local detail information and global context feature information in the sequence coding, which can effectively improve the Performance of the model on the speech separation task.
- Considering the channel information and spatial information inside the speech sequence feature coding, a local attention module is added to the network. This module introduces the channel attention mechanism and the spatial attention mechanism, which can make the model pay more attention to the local details in the speech block. Make the network more sensitive to speaker characteristics.
- For feature information extraction between long sequence blocks, a global attention module is built, and feed forward is improved, and linear is replaced by GRU with better time series processing ability, so that the network can more effectively model long sequence context.

The rest of this paper is organized as follows: Section 2 reviews some related work. Section 3 describes the details of the proposed method. The experimental settings and results of proposed method are given in Sections 4 and 5, respectively. Section 6 draws some conclusions.

2 Related work

Single-channel speech separation (SS) is a classical task in speech processing, which aims at separating each source from mixed speech to improve the quality and intelligibility of speech data. Considering that the mixed speech has some similarity with the separated speech, the speech separation model does not directly output the speech waveform of each speaker but indirectly obtains the information of each speaker by estimating the mask of each speaker through the separator, and then reduces the speech waveform according to the speaker mask. The block diagram of speech separation is shown in Fig. 1. First of all, the mixed speech signal x is processed by the encoder to obtain

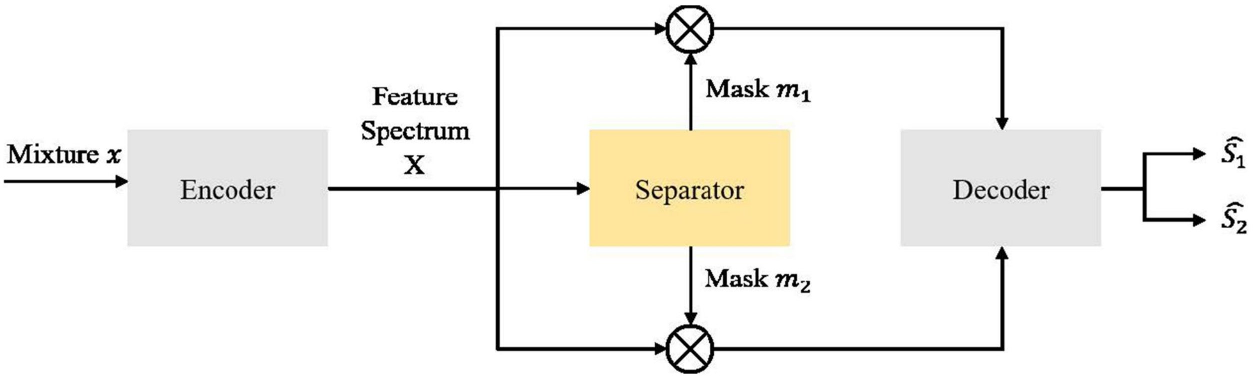

Fig. 1 Block diagram of deep learning based speech separation. Firstly, the encoder converts the speech signal into a feature representation, and then the feature separation and mask estimation are performed by the separator. Finally, the separated speech waveforms are reconstructed by the decoder

the feature spectrum X. The feature spectrum X of the mixed signal is then fed into the separator model to estimate the mask $m_i, i = 1, 2, ..., C$ for each speaker. Estimated masks can be used as training targets for speech separation. The common time-frequency masks include ideal binary mask (IBM) [28, 29], target binary mask (TBM) [30], and ideal ratio mask (IRM) [9, 31], which can significantly improve the intelligibility and perceptual quality of separated speech. Finally, the speaker mask m_i is multiplied element by element with the feature spectrum X and then processed by the decoder to obtain the separated speaker speech waveform $\hat{s}_i$.

With the further improvement of deep network models, the performance of SS using sequence prediction has improved significantly, and deep learning based SS uses a data-driven approach to learn better separation models and greatly compensate for the shortcomings of traditional methods. Traditional speech separation algorithms, in realistic situations, have difficulty in meeting the assumptions, and thus the separation performance is greatly reduced. At the same time, the traditional linear model has difficulty in capturing the highly nonlinear characteristics of speech signals, and the large computational volume and high computational complexity make it difficult to meet the requirements of real-time applications. Compared with traditional speech separation methods, deep learning-based speech separation methods are driven by big data, and current models can be trained on large-scale data sets, thus obtaining better results under given conditions and being able to model speech signals using the high nonlinearity of neural networks. In general, single-channel SS in the field of deep learning can be divided into two categories: time-frequency domain SS methods and time-domain SS methods.

2.1 Time-frequency domain SS methods

Deep clustering (DPCL) [10] constructs the affinity matrix with an ideal binary mask (IBM) and makes the error function of the target and estimated affinity matrix as the target to train a deep neural network, which first maps the time-frequency units of mixed speech into an embedding space, then executes a clustering algorithm, and finally generates a binary mask based on each clustering label to separate the target speech. DPCL++ [32] optimizes the overall structure of DPCL; DPCL++ pays more attention to temporal contextual information during training. The joint enhancement layer in the DPCL++ network refined signal estimation through clustering and enhancement phases. Finally, end-to-end training was used to maximize signal fidelity, thus further improving the separation performance of speech signals. Based on these, the deep attractor network (DANet) solves the source permutation problem that previously existed in DPCL by creating attractors in the embedding vector to aggregate the T-F units of each source [13, 33, 34].

Phase-enhanced speech separation methods are an important research direction in the T-F domain. Williamson et al. [35] proposed a supervised monophonic speech separation method that simultaneously enhances the magnitude spectrum and phase spectrum by operating in the complex frequency domain, using a deep neural network to estimate the real and imaginary components of the ideal ratio mask defined in the complex domain, correcting the reconstructed speech phase to improve the speech separation. Tan et al. [36] proposed a gated convolutional recurrent network (GCRN) for complex spectral mapping to enhance both the amplitude and phase response of noisy speech, and the GCRN performed well in objective speech intelligibility and quality for complex spectral mapping. Guochen Yu et al. [37] developed a dual-branch joint amplitude and phase estimation framework (DBT-Net) for single-channel speech enhancement task, aiming to recover coarse and fine-grained regions of the entire spectrum in parallel. The DBT-Net includes an amplitude estimation branch and a spectral purification branch, where the amplitude estimation branch aims to filter out the main noise components in the amplitude domain. The complex spectral purification branch is carefully designed to restore lost spectral details and implicitly estimate the phase information in the complex-valued spectral domain.

Although the time-frequency domain methods have good performance in some evaluation metrics, there are still some problems [18]: (1) STFT and iSTFT are a fixed transformation, and the obtained T-F features are not necessarily the most suitable for the speech separation task; (2) the phase information of the features in the T-F domain is difficult to model, and the time-frequency domain methods usually only utilize the assignment information in the features for feature extraction and separation. Therefore, there is a performance limit for time-frequency domain SS methods.

2.2 Time-domain SS methods

Unlike time-frequency domain SS, the time-domain approach processes the speech signal directly in time-domain through an encoder-decoder network, replacing STFT and ISTFT with encoder-decoder, reducing the computational cost of SS and the minimum delay required for the output. Based on this, Luo et al. have proposed Conv-TasNet [18] and DPRNN [20], which use temporal convolutional networks [38, 39] and dual-path RNNs for feature separation, respectively, and effective improvements are achieved in single-channel SS tasks compared to time-frequency based speech separation methods., making such methods attract much attention.

Inspired by transformer networks in NLP, speech separation based on a self-attention mechanism has attracted widespread interest. In [22], an improved transformer structure is applied to a dual-path speech separation network, which enables the separation model to effectively process long speech sequences globally and improve the separation performance. In [25], the self-attention mechanism is used in U-Nets to extract high-level, large-granularity contexts. In this paper, we also employ the self-attention to learn global features in speech signals. Consistent with [22] for extracting global feature information, the same multi-headed attention mechanism is used in this paper because of its superior performance in extracting speech contextual features. Notably, our approach models the speech signal using two different attention networks simultaneously, both local detail information and global contextual information of speech are considered, thus better extracting useful information from the speech signal.

3 Proposed method

There are three parts involved in time-domain adaptive attention based single-channel speech separation: encoding and segmentation, time-domain adaptive attention network, and overlap-add and decoding (as shown in Fig. 2a). In detail, the adaptive attention networks composed of N identical local and global attention blocks, with each block including a local attention network and a global attention network which jointly process the speech signals. Meanwhile, the model has the capability to capture the features of frame-level and utterance-level by adaptively adjusting the proportion of local attention and that of global attention in multiple attention layers.

3.1 Problem description

The single-channel speech separation can be described as estimating the target speech $s_i(t)$ from the mixed speech. The mixed speech $x(t) \in \mathbb{R}^{1\times T}$ can be expressed as:

$$x(t) = \sum_{i=1}^{C} s_i(t) \tag{1}$$

where C represents the number of target speakers, and $s_i(t) \in \mathbb{R}^{1\times T}, i = 1, 2, ...C$ represents the target speech.

3.2 Encoding and segmentation

As mentioned in the previous section, for the supervised speech separation approaches, the mixed speech $x \in \mathbb{R}^{1\times T}$ is first converted to a feature spectrum $X \in \mathbb{R}^{E\times L}$ by STFT. As in TasNet [17], we use a 1D convolutional layer to replace the traditional STFT:

$$X = ReLU(Conv1d(x)) \tag{2}$$

where $ReLU(\cdot)$ denotes the element-wise rectified linear unit to ensure non-negative output. E and L respectively represent the number of feature vectors in the spectrum X and the length of each vector, of which are related to the parameter settings of the 1D convolutional layer.

Following [20], the output X of encoder is divided into S overlapped segments during the segmentation, with each segment having a length of R and a hop size of $R/2$. The first and last segments are zero-padded to ensure the same size of each segment. Then, these S segments are concatenated into a 3D tensor $W = [w_1, w_2, ...w_S] \in \mathbb{R}^{E\times S\times R}$, where $w_1, w_2, ...w_S \in \mathbb{R}^{E\times R}$ are the 2D segments.

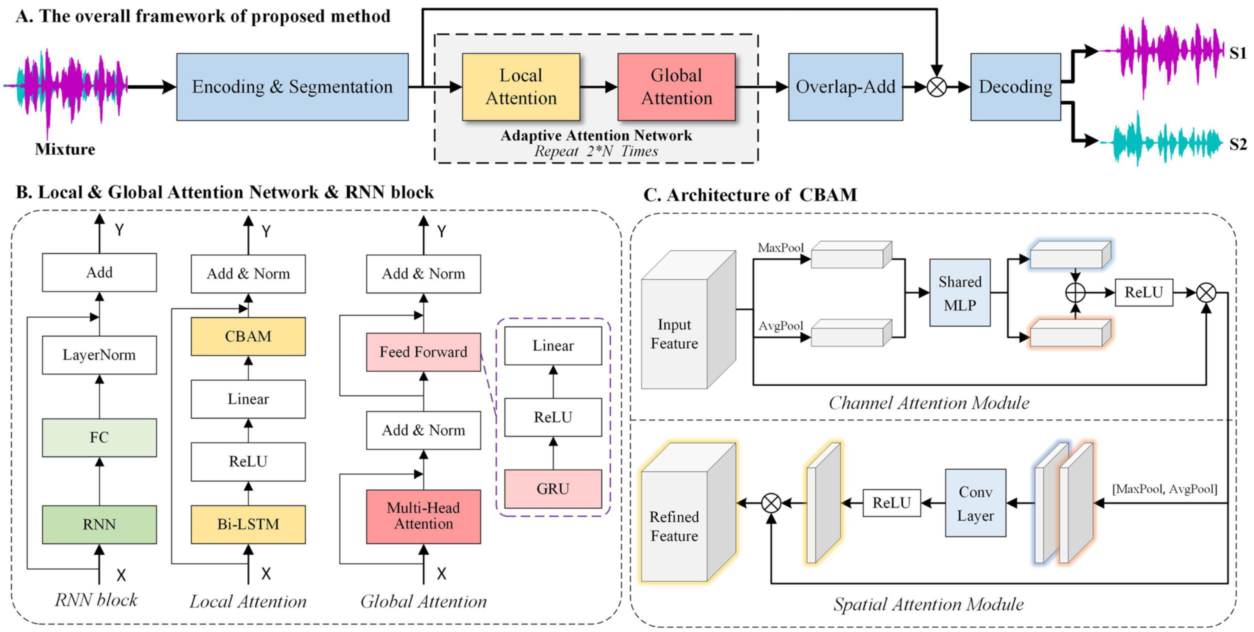

Fig. 2 Overview of the time-domain adaptive attention network for single-channel speech separation. To better extract the local information and global information in speech features, two different attention networks are used in this model: (1) the CBAM is introduced in local attention networks to focus on subtle details of the speech signals (frame-level features). (2) In the global attention networks, the transformer based on self-attention mechanism is used to explore the global associations of the speech contexts (utterance-level features). (3) In the ablation experiment, we replaced the local attention module and/or global attention module with the RNN block

3.3 Time-domain adaptive attention network

As shown in Fig. 2b, a time-domain adaptive attention network is used as a separator comprising N cascaded local and global attention blocks with same structure. The convolutional block attention module (CBAM) [27] is added into Bi-LSTM as a local attention network. LSTM is capable of ignoring irrelevant information and focusing on the key information in mixed speech and, in combination with CBAM, these local information can be further highlighted. CBAM enables the model to pay more attention to the details of feature map through channel attention and spatial attention and has been widely applied in image recognition, speech recognition, speech enhancement, and other fields [40–44]. Through this structure, this local attention network can effectively focus on the frame-level details in speech. In the global attention network, the transformer (with a structure like that in [22]) is used to model the speech at the utterance level. The self-attention based transformer have been proved to be superior in context modeling in many tasks [45–47].

3.3.1 Local attention network

An additional local attention network is introduced so that it could focus on the detailed features of the speech signals, which is a different practice from what is applied in the current self-attention based speech separation [22, 23]. For the convenience of reference, the input and output of the local attention network are represented by $X^{LA} \in \mathbb{R}^{E\times S\times R}$ and $Y^{LA} \in \mathbb{R}^{E\times S\times R}$ respectively. First, a Bi-LSTM, ReLU and a Linear layer are used to fulfill the feature extraction operation, which is defined as follows:

$$H^{LA} = w_1\left(ReLU\left(f_{bi\text{-}lstm}\left(X^{LA}\right)\right)\right) + b_1 \tag{3}$$

where $H^{LA} \in \mathbb{R}^{E\times S\times R}$ represents the output of the linear layer. w_1 and b_1 refers to the weights and bias in the linear layer, and $f_{bi\text{-}lstm}(\cdot)$ denotes the processing function of the Bi-LSTM layer. Then these features are inputted into the CBAM to complete the refining of the feature information. Finally, the refined features and the input features X^{LA} are added to obtain the output of the local attention network:

$$H^{LA'} = f_{cbam}\left(H^{LA}\right) \tag{4}$$

$$Y^{LA} = LN\left(X^{LA} + H^{LA'}\right) \tag{5}$$

where $H^{LA'} \in \mathbb{R}^{E\times S\times R}$ represents the refined features by CBAM. $f_{cbam}(\cdot)$ is an attention weighting operation including channel attention and spatial attention, which is defined in [27]. And $LN(\cdot)$ denotes layer normalization used to normalize the input data of each layer.

CBAM is able to serially generate attentional feature information in both the speech mask channel and spatial dimensions, and then both feature information is multiplied with the previous original input features for adaptive feature correction to output the final feature information. The equation for $f_{cbam}(\cdot)$:

$$X_c = \sigma(MLP(AvgPool(X)) + MLP(MaxPool(X))) \tag{6}$$

$$X_s = \sigma(f^{a\times b}([AvgPool(X');MaxPool(X')])) \tag{7}$$

In Eq. (6), where $X \in \mathbb{R}^{E\times S\times R}$ is the input speech block, $AvgPool$ stands for average pooling, $MaxPool$ stands for maximum pooling, σ denotes the sigmoid function, and $X_c \in \mathbb{R}^{C\times S\times R}$ is the output of the channel attention. In Eq. (7), where $X' \in \mathbb{R}^{N\times S\times R}$ is the channel feature map, $f^{a\times b}$ represents a convolution operation with the filter size of a × b, and $X_s \in \mathbb{R}^{E\times S\times R}$ is the output of the spatial attention.

3.3.2 Global attention network

Similar to the case in the local attention network, the input and output of the global attention network are represented by $X^{GA} \in \mathbb{R}^{E\times S\times R}$ and $Y^{GA} \in \mathbb{R}^{E\times S\times R}$ respectively. First, we apply multi-head attention followed by a residual connection and layer normalization:

$$H^{GA} = LN\left(X^{GA} + f_{mha}\left(X^{GA}, X^{GA}, X^{GA}\right)\right) \tag{8}$$

where $H^{GA} \in \mathbb{R}^{E\times S\times R}$ is a mid-product that contains the global related information. $f_{mha}(\cdot)$ is defined in [24], which consists of multiple scaled dot-product attention modules.

MHA context-aware modeling of long sequences of speech signals allows elements in speech sequences to interact directly, facilitating information transfer and enabling the model to learn sequential information of speech sequences without speech location encoding. The equation for $f_{mha}(\cdot)$ [22]:

$$Head_i = Attention(Q_i, K_i, V_i) = softmax\left(\frac{Q_i K_i^T}{\sqrt{d}}\right)V_i \tag{9}$$

$$MultiHead = Concat(Head_1, \dots, Head_h)W^O \tag{10}$$

$$Mid = LayerNorm(X + MultiHead) \tag{11}$$

$$FFW = ReLU(Mid\,W_1 + b_1)W_2 + b_2 \tag{12}$$

$$Output = LayerNorm(Mid + FFW) \tag{13}$$

Here, $X \in \mathbb{R}^{l\times d}$ is the input with length l and dimension d, $Q_i, K_i, V_i \in \mathbb{R}^{l\times d/h}$ are the mapped queries, keys, and values, and $W^O \in \mathbb{R}^{d\times d}$ is parameter matrices. FFW denotes the output of the position-wise feedforward network, $W_1 \in \mathbb{R}^{d\times d_{ff}}, W_2 \in \mathbb{R}^{d_{ff}\times d}, b_1 \in \mathbb{R}^{d_{ff}}, b_2 \in \mathbb{R}^{d}$, and $d_{ff} = 4 \times d$.

The global attention network finally employs a feed forward network (FFW) to further achieve feature extraction:

$$Y^{GA} = LN\left(H^{GA} + f_{ffw}\left(H^{GA}\right)\right) \tag{14}$$

where $f_{ffw}(\cdot)$ represents the processing of FFW, which is composed of a GRU, ReLU, and linear layer (as shown in Fig. 2b).

3.4 Overlap-add and decoding

The output of the separator is used to generate a mask for each speaker. The speaker mask can be obtained through overlap-add [20], and the network output feature mask are stitched according to the temporal dimension to obtain a pure speech mask for each speaker and multiplied by the encoder output X, thus reconstructing the speech waveform by a transposed convolutional layer:

$$s_m = ReLU\left(OverlapAdd\left(Y^{GA}\right)\right) \tag{15}$$

$$s_c = Conv1d\text{- }Transpose(s_m * X) \tag{16}$$

where $s_m \in \mathbb{R}^{C\times E\times L}$ represents the speaker mask and $Conv1d\text{-}Transpose(\cdot)$ stands for the 1D convolution operation in decoder, which is the same kernel size and step length as those in encoder. $s_c \in \mathbb{R}^{C\times T}$ represents the source speech of C speakers.

3.5 Network optimization

The optimization process of our proposed method integrates two major parts, as shown in Algorithm 1. (1) Model training: input mixed speech signal $x(t)$ and clean speaker speech $s_1(t)$, $s_2(t)$ of training set, and train the encoder, separator, and decoder models simultaneously. (2) Model testing: input mixed speech signal $x(t)$ and clean speaker speech $s_1(t), s_2(t)$ of test set, use the model for source separation to obtain the separated speech $\hat{s_1}(t), \hat{s_2}(t)$, and calculate the separation index. A pytorch implementation of our TAANet can be found at "http://www.msp-lab.cn:1436/msp/TAANet".

```
Input: Mixed speech signal $x(t)$ and clean speaker speech
       $s_1(t)$, $s_2(t)$. Set the max epoch: $I_{max}$.
Output: The trained TAANet model.
(1) Model training:
for $epoch = 1, \ldots, I_{max}$ do
    for $iter = 1, \ldots, dataloader$ do
        1. Encoder transforms speech to feature domain:
           $X = Encoder(x)$;
        2. Segmentation of features into a 3D tensor:
           $W = Segmentation(X)$;
        3. Feature extraction using local and global
           attention networks (Repeat N times):
           $Y^{LA} = LocalNet(W)$,
           $Y^{GA} = GlobalNet(Y^{LA})$;
        4. Feature separation and mask estimation:
           $s_m = RELU(OverlapAdd(Y^{GA}))$;
        5. Separated speech reconstruction:
           $s_c = Conv1d\text{-}Transpose(s_m * X)$;
        6. Calculate the separation loss, refer to Eq.(17) to
           (19);
        7. Back propagation, calculating the gradient $\Delta L$;
        8. Update the model parameters $w$: $w = w - \alpha \Delta L$,
           where $\alpha$ is the learning rate.
    Saving the current model.
(2) Model Testing:
  1. Initialize the model and load parameters;
  2. Input the mixed speech into model for separation;
  3. Output the separated speech and calculate the
   performance.
```

Algorithm 1 Time-domain adaptive attention network for single-channel speech separation

4 Experiment setup

4.1 Dataset

4.1.1 WSJ0-2mix

For evaluating the performance of proposed method, we used the publicly available WSJ0-2mix corpus [10], a data set widely used in exploring speaker speech separation. It contains 20,000, 5000, and 3000 two-speaker mixtures in its 30 h training, 10 h validation, and 5 h test sets, respectively. All these mixtures are generated from the Wall Street Journal (WSJ0), with the sampling frequency of 8 kHz.

4.1.2 WHAM! and WHAMR!

Although significant progress has been made in single-channel speech separation techniques, most methods have been researched on clean speech datasets that are not fully representative of real scenarios. To evaluate the speech separation model under more realistic and challenging conditions, we conduct some experiments using the WHAM! [48] and WHAMR! [49] datasets. WHAM! is composed of pure WSJ0-2mix speech data mixed with noise recorded in natural scenes and with a random SNR of −6 to 3 dB. While WHAMR! introduces additional reverberation in the same noise conditions. This task is more challenging because the model needs to perform source separation, denoising, and dereverberation at the same time.

4.1.3 Grid-2mix

To verify the generalization performance of proposed method, the GRID-2mix is used in the experiment, which is generated from the GRID data set [50]. Worth noting is that the GRID data set contains video and audio data, but only audio data are used in the experiment. The GRID-2mix contains 34 different speakers' speeches, with the mixed speech duration of 20 h, of which the proportion of *training*/*validation*/*test* is 7/2/1, with the sampling frequency of 8kHz and the random SNR of −5 to 5 dB.

4.2 Parameter settings

4.2.1 Model configuration

In most of the experiments, the encoder contains 256 convolutional filters, with a kernel size of 4 and stride of 2. The number of local and global attention blocks and the length of segment in segmentation processing are 6 and 200, respectively. Particularly, we set the number of local and global attention blocks to 6 in the ablation experiment. In addition, we detail the inputs and outputs of the middle layers of the model. For example, when we input a speech signal $x \in \mathbb{R}^{1\times32000}$ with length of 4s and sample rate of 8k into the model, the corresponding middle layer tensor information is shown in Table 1. Therein, b and C denote the batch size and the number of speakers in the mixed speech, respectively.

4.2.2 Training details

In the training stage, the TAANet is trained on the data set for 100 epochs, with a batch size of 1 and an initial learning rate of $4e^{-4}$. Adam [51] is used as an optimizer because its hyperparameters are well interpreted and suitable for scenarios with large-scale data and parameters. A fixed step decay is used for the learning rate reduction strategy, that is, the learning rate become 0.98

Table 1 The tensor size of the middle layer of the model when input a speech with length of 4*s* and sampling rate of 8*k*

Module	Layers	Input size	Output size
Encoder	Conv1d	[b, 1, 32000]	[b, 256, 15999]
	GroupNorm	[b, 256, 15999]	[b, 256, 15999]
	Conv1d	[b, 256, 15999]	[b, 64, 15999]
	Segmentation	[b, 64, 15999]	[64, 200, b*162]
Separator	LocalAttention	[64, 200, b*162]	[64, 200, b*162]
	GlobalAttention	[64, 200, b*162]	[64, 200, b*162]
Decoder	OverlapAdd	[64, 200, b*162]	[b*C, 256, 15999]
	Conv1d-Transpose	[b*C, 256, 15999]	[b, C, 32000]

times of the original after every 2 epochs. When the loss of the validation for 10 epochs presents no decrease, the training stops early to reduce the overfitting of the model.

4.3 Loss function

The final output of the end-to-end speech separation model is the time-domain waveform of the clean signals, with the scale-invariant source-to-noise ratio (SI-SNR) [17] usually used as an evaluation index. SI-SNR is defined as:

$$s_{target} = \frac{< \hat{s}, s > s}{||s||^2} \tag{17}$$

$$e_{noise} = \hat{s} - s_{target} \tag{18}$$

$$SI\text{-}SNR = 10 \log_{10} \frac{||s_{target}||^2}{||e_{noise}||^2} \tag{19}$$

therein, $\hat{s}$, the estimated target speech and s, the clean target speech are required to be normalized with a mean value of 0 to ensure scale invariance. The larger the SI-SNR value means the better the speech separation performance, but gradient descent is used to train the model during the training process, so the actual loss function is the inverse of SI-SNR. In addition, utterance-level permutation invariant training (uPIT) [12] is used in the model training process.

5 Results

5.1 Results on WSJ0-2mix

First, we compare the scale-invariant signal-to-noise ratio improvement (SI-SNRi) and signal-to-distortion ratio improvement (SDRi) [52] performance of proposed method and other methods.

The methods compared in the experiments are as follows: DPCL++ [32], an improved deep clustering speech separation method; uPIT-BLSTM-ST [12], an utterance-level permutation invariant training speech separation method; tasNet [17], a speech separation method based on time-domain mask; Conv-TasNet [18], a time-domain speech separation method based on convolutional networks; DeepCASA [53], a deep computer auditory scene analysis speaker-independent speech separation method; FurcaPa [54] for speech separation using deep attention gated extended time domain convolutional networks; FurcaNeXt [19], an end-to-end monophonic speech separation based on dynamic gated extended time convolutional networks; DPRNN [20], a speech separation task using dual path recurrent neural networks; SVOICE [55], a gated recurrent neural network for unknown number of speaker speech separation; DPTNet [22], which constructs a transformer-based two-path speech separation model; and SepFormer [23], which learns short- and long-term dependencies by using transformer's multi-scale approach. The baseline method of TAANet is derived from DPTNet, in which the global attention module refers to the attention mechanism in DPTNet. Unlike DPTNet, TAANet introduces a local attention module to pay more attention to the detail features within speech segments.

As shown in Table 2, the TAANet achieves the best performance on the WSJ0-2mix test set with SI-SNRi of 20.7 dB and SDRi of 20.9 dB, respectively. The data shown in the table are from their papers. What is more, compared with DPRNN and DPTNet, the SI-SNRi of our model are improved by 1.9 and 0.5 dB, respectively. Thanks to the adaptive attention network, our model can focus on both the local features of the speech signals and the global ones simultaneously, thus further promoting the speech separation performance.

As shown in Fig. 3, compared with the baseline methods, the test results of TAANet on the WSJ0-2mix dataset achieve better performance on SDRi, SI-SNRi, PESQ, and STOI. From (a), (b), (c), (d), the median of TAANet (red line position) is better than the baseline methods, and the lower limit of TAANet under the four indicators is also better than the baseline methods. Among SDRi, SI-SNRi, and PESQ, TAANet has achieved the best performance. In the performance of most test samples, TAANet can achieve better speech separation.

To better analyze the model performance, we used paired *t*-test method for a more thorough statistical analysis of the experimental results. The paired *t*-test algorithm works as follows:

Table 2 Comparison with other methods on WSJ0-2mix. Best values are marked with bold font

Model	# Param	SI-SNRi (dB)	SDRi (dB)	PESQ	STOI
DPCL++ [32]	13.6M	10.8	-	-	-
uPIT-BLSTM-ST [12]	92.7M	-	10.0	-	-
TasNet [17]	-	13.2	13.6	-	-
Conv-TasNet [18]	5.1M	15.3	15.6	-	-
DeepCASA [53]	12.8M	17.7	18.0	-	-
FurcaPa [54]	-	-	18.2	-	-
FurcaNeXt [19]	51.4M	-	18.4	-	-
DPRNN [20]	2.6M	18.8	19.0	3.63	0.97
SVOICE [55]	7.5M	20.1	20.4	-	-
DPTNet [22]	2.7M	20.2	20.6	3.75	**0.98**
SepFormer [23]	26M	20.4	20.5	-	-
TAANet	5.4M	**20.7**	**20.9**	**3.80**	**0.98**

Table 3 Paired *t*-test on SDRi

i	Model	$\overline{D}$	S_D	$\hat{S_D}$	*t*	*p*	α
1	DPRNN-TAANet	1.94	5.37	0.104	17.52	4.32×10^{-65}	0.05
2	DPTNet-TAANet	0.59	2.87	0.056	10.62	8.28×10^{-26}	0.05

Table 4 Paired *t*-test on SI-SNRi

i	Model	$\overline{D}$	S_D	$\hat{S_D}$	*t*	*p*	α
1	DPRNN-TAANet	1.86	5.50	0.107	17.34	7.58×10^{-64}	0.05
2	DPTNet-TAANet	0.61	2.95	0.058	10.52	2.25×10^{-25}	0.05

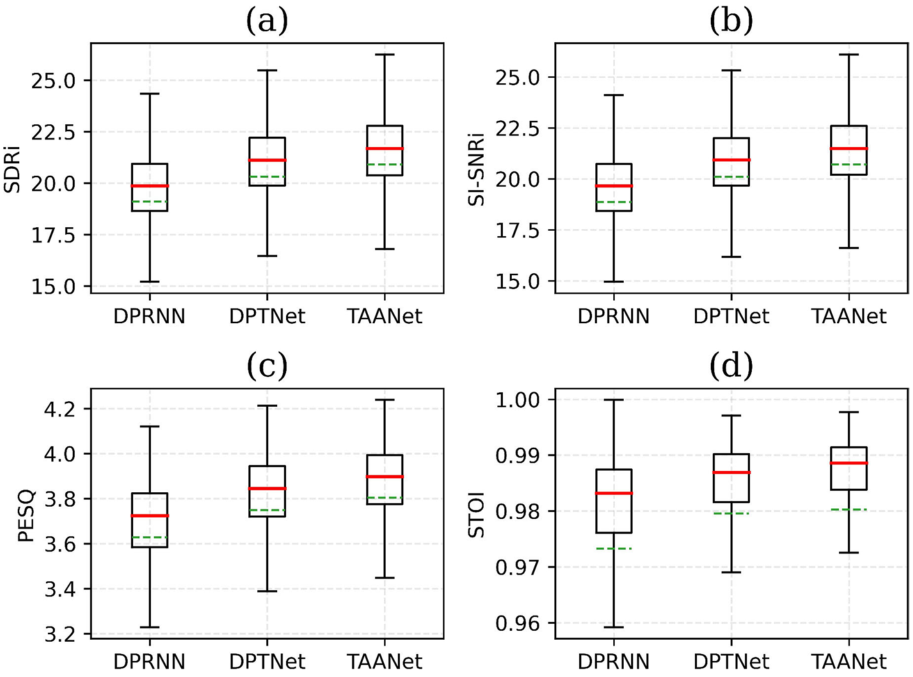

Fig. 3 Box plot of experimental results of DPRNN, DPTNet, and TAANet on WJS0-2mix dataset. Comparison of SDRi (**a**), SI-SNRi (**b**), PESQ (**c**), and STOI (**d**) our method TAANet achieved better performance

Step 1: Given X_1 and X_2 are the test results from different models on the test set, and calculate the difference value D between the two sample sets, $D = X_1 - X_2$.

Step 2: Calculate the mean $\overline{D}$ of the difference values D, $\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$, n represents the sample size and D_i denotes the difference value of the ith sample.

Step 3: Calculate the standard error S_D and standard deviation of sample difference values $\hat{S_D}$. $S_D = \sqrt{\frac{\sum_{i=1}^{n}(D_i - \overline{D})^2}{n-1}}$, $\hat{S_D} = \frac{S_D}{\sqrt{n}}$.

Step 4: Calculate of t-values: $t = \frac{\overline{D}}{\hat{S_D}}$.

The t value is used to find the corresponding probability (p) from the table of normal distribution, which is then compared with α.

Based on the above formula, we first performed significance tests (using 95% confidence intervals) for the SDRi indicators of DPRNN and TAANet. Assuming that there is no significant difference between the two sets of data, a null hypothesis test is proposed for the difference values of the two samples. The corresponding values are calculated: $\overline{D} = 1.84$, $S_D = 5.37$, $\hat{S_D} = 0.104$, $t = 17.52$, $p = 4.32 \times 10^{-65}$. Similarly, we performed significance tests (using 95% confidence intervals) for the SDRi metric for DPTNet and TAANet. The corresponding values are calculated: $\overline{D} = 0.59$, $S_D = 2.87$, $\widehat{S_D} = 0.056$, $t = 10.62$, $p = 8.28 \times 10^{-26}$.

Second, we performed significance tests (using 95% confidence intervals) for the SI-SNRi metric for DPRNN and TAANet. The calculation results are as follows: $\overline{D} = 1.86$, $S_D = 5.50$, $\widehat{S_D} = 0.107$, $t = 17.34$, $p = 7.58 \times 10^{-64}$. Finally, we performed significance tests (using 95% confidence intervals) on the SI-SNRi metrics for DPTNet and TAANet. The results are as follows: $\overline{D} = 0.61$, $S_D = 2.95$, $\widehat{S_D} = 0.058$, $t = 10.52$, $p = 2.25 \times 10^{-25}$.

As shown in Tables 3 and 4, all $p < 0.05$ can be obtained, which means that there is a significant difference between the experimental results of DPTNet and TAANet, DPRNN, and TAANet. The significance analysis results indicate that TAANet outperforms the baseline methods.

Further, the WSJ0-2mix data set is used for the investigations into the impact of different hyperparameters on the performance of the model, and the results are shown in Table 5. The *Window*, *Chunk size*, and *Layers* denote the kernel size of encoder's convolutional layer, the length of segment in segmentation processing, and numbers of local and global attention blocks in TAANet respectively. It can be found that the performance is improved by reducing the size of the filters in encoder, but the shape of the input is also increased at the same time, which leads to more training time costs. The number of layers has a great influence on the performance of the model, which indicates that the cascade of multiple local and global attention blocks is crucial in that it helps the model to adaptively focus on the local information and global information at different layers. Multiple experiments have demonstrated that the best results are obtained when $Window = 4$, $Chunk\ size = 200$, and $Layers = 8$.

Table 5 The effect of different configurations on WSJ0-2mix. Best values are marked with bold font

# Param	Window	Chunk size	Layers	SI-SNRi	SDRi
2.7M	8	150	4	19.0	19.2
2.7M	4	200	4	19.6	19.8
2.7M	2	250	4	19.8	20.1
4.0M	8	150	6	19.2	19.4
4.0M	4	200	6	20.6	20.8
5.4M	4	200	8	**20.7**	**20.9**

5.2 Ablation study

In this section, we perform some ablation experiments on the WSJ0-2mix dataset to demonstrate the effectiveness of local attention networks and global attention networks in TAANet. As can be seen from Table 6, both our local and global networks improve the separation performance compared to the dual-path RNN base model [20]. During the experiment, following DPRNN, we replaced the local attention module and/or global attention module with RNN block. The RNN block consists of an RNN layer, an FC layer, and a layer norm. In particular, the improvement of the global attention network is relatively large (from 19.2 dB to 20.3 dB). When the model employs both local and global attention networks, the speech separation SI-SNRi is improved to 20.6 dB. Ablation study results show that the global network based on self-attention mechanism is effective on the speech separation task; however, it is better in modeling the global of the sequence signal. With the local attention network, the model can focus on the detailed information that can be easily overlooked. Therefore, with our proposed local and global attention network, the model is able to focus on both frame-level and utterance-level pieces of information in the signal, thus improving the speech separation performance.

Aiming to better demonstrate the respective contributions of local attention networks and global attention networks in TAANet, we have processed some samples by using different models and visualized the separation results. These samples are a mixture of the voices of two male speakers. Separation for speakers of the same gender is more challenging and can better demonstrate the model's handling of signal details in the speech separation process. As shown in Fig. 4, the first and second row represent the speech spectrogram of mixed speech and clean speech, and the fourth, fifth, sixth, and seventh rows represent the speech spectrogram of baseline, local, global, and local and global attention networks minus the clean speech spectrogram, respectively. Comparing (c) and (e), or (d) and (f), with the red boxes marking the parts, the local attention network has smaller residuals in the speech spectrogram, meaning it more concerned with detailed features. And comparing (d) with (h) as a whole, (h) has a smaller overall residual than (c). This means that it is more

Table 6 Ablation experiment results on WSJ0-2mix

Model	SI-SNRi	SDRi
Base model	19.2	19.4
Local only	19.5	19.8
Global only	20.3	20.6
Local and global (TAANet)	20.6	20.8

concerned with global features and proves that the global network can effectively focus on utterance-level features. Contrasting the fifth and sixth rows, the global attention network performs better overall; however, it ignores certain detailed information in the speech signal (frame-level), such as temporal or spectral continuity, spectral structure, and timbre. When both local and global attention networks are used (refer to (i) and (j)), the deviation between separated spectrogram and clean spectrogram is minimal. This phenomenon suggests that our model is capable of combining the advantages of both types of attention and focusing on both frame-level and utterance-level information, thus further demonstrating the effectiveness of the proposed TAANet.

To further demonstrate the role of each module, we conducted a significance statistical analysis on the ablation experiments. We used the Friedman test to evaluate the performance of the ablation models and ranked them. The tested models include the base model, local only, global only, and TAANet, with a total of 2621 samples in the test set. The algorithm process for Friedman test [56] is as follows (assuming the comparison of performance among k models on N samples):

Step 1: Find r_i^j – the rank of the model j on the $i-th$ sample, and compute the average rank R of model j: $R = \frac{1}{N}\sum_i r_i^j$;

Step 2: The null hypothesis states that all models have the same performance, and compute the Friedman statistic: $\chi_F^2 = \frac{12N}{k(k+1)}(\sum_j R_j^2 - \frac{k(k+1)^2}{4})$, it is asymptotically χ^2 distributed with $k-1$ degrees of freedom;

Step 3: If χ^2 exceeds the critical value, then reject the null hypothesis, otherwise accept it. When the null hypothesis is rejected, a post hoc test is used to determine the nature of the difference.

Given the result of the Friedman test, we conducted the Holm test as a test to compare model TAANet and other ablation models. The test statistics for comparing the two models used the method is as follows: $z = (R_1 - R_2)/\sqrt{\frac{k(k+1)}{6N}}$. The z value is used to find the corresponding probability (p) from the table of normal distribution, which is then compared with an appropriate α.

Testing on SDRi, $R_{basemodel} = 1.6475$, $R_{localonly} = 2.2232$, $R_{globalonly} = 2.8718$, $R_{taanet} = 3.2575$, and with $\alpha = 0.05$, $k = 4$, $N = 2621$, the standard error is $SE = \sqrt{\frac{4 \cdot (4+1)}{6 \cdot 2621}} = 0.03566$.

Testing on SI-SNRi, $R_{basemodel} = 1.6468$, $R_{localonly} = 2.2163$, $R_{globalonly} = 2.8764$, $R_{taanet} = 3.2587$, and with $\alpha = 0.05$, $k = 4$, $N = 2621$, the standard error is $SE = \sqrt{\frac{4 \cdot (4+1)}{6 \cdot 2621}} = 0.03566$.

Table 7 Friedman test on the SDRi for ablation experiment

i	Model	$z = (R_{taanet-R_i})/SE$	p	$\alpha/(k-1)$
1	Base model	$(3.2575 - 1.6475)/0.03566 = 45.15$	0.0000	0.0167
2	Local only	$(3.2575 - 2.2232)/0.03566 = 29.01$	0.0000	0.0250
3	Global only	$(3.2575 - 2.8718)/0.03566 = 10.81$	6.95×10^{-27}	0.0500

Table 8 Friedman test on the SI-SNRi for ablation experiment

i	Model	$z = (R_{taanet} - R_i)/SE$	p	$\alpha/(k-1)$
1	Base model	$(3.2587 - 1.6468)/0.03566 = 45.15$	0.0000	0.0167
2	Local only	$(3.2587 - 2.2163)/0.03566 = 29.22$	0.0000	0.0250
3	Global only	$(3.2587 - 2.8764)/0.03566 = 10.72$	1.85×10^{-26}	0.0500

As shown in Tables 7 and 8, all p are less than significance level $\alpha = 0.05$. The Holm procedure rejects all hypothesis. This shows that TAANet performs significantly better than ablation models at the significance level $\alpha = 0.05$.

5.3 Results on WHAM! and WHAMR!

In this part of the experiment, we evaluate the speech separation performance of the model in scenes with ambient noise and reverberation. The results are reported in Table 9, where we compare the performance of several methods for SI-SNR and SDR on the WHAM! and WHAMR! datasets. The data in Table 9 are from replications of what has been reproduced for these methods in published papers. We found some of the methods in Table 2 to be not reproduced in published papers and so do not appear in Table 9. Our model has achieved a superior result: the SI-SNR and SDR improvement under the two data sets are 15.5 dB, 15.8 dB and 12.0 dB, 11.2 dB, respectively. This result suggests that the proposed method also performs well under noise and reverberation conditions.

5.4 Results on Grid-2mix

Relative experiments on the GRID-2mix are conducted for the purpose of proving the generalization performance of the proposed method. The data in GRID-2mix are shorter than those in WSJ0-2mix, which makes separation more difficult with GRID-2mix. In this experiment, DPRNN and DPTNet are used as the baseline methods to be compared with the proposed method. From Table 10, TAANet demonstrates the best performance on GRID-2mix (SI-SNRi of 16.0 dB and SDRi of 16.8 dB), which is the proof that our approach does a good job of source separation and shows better generalization performance.

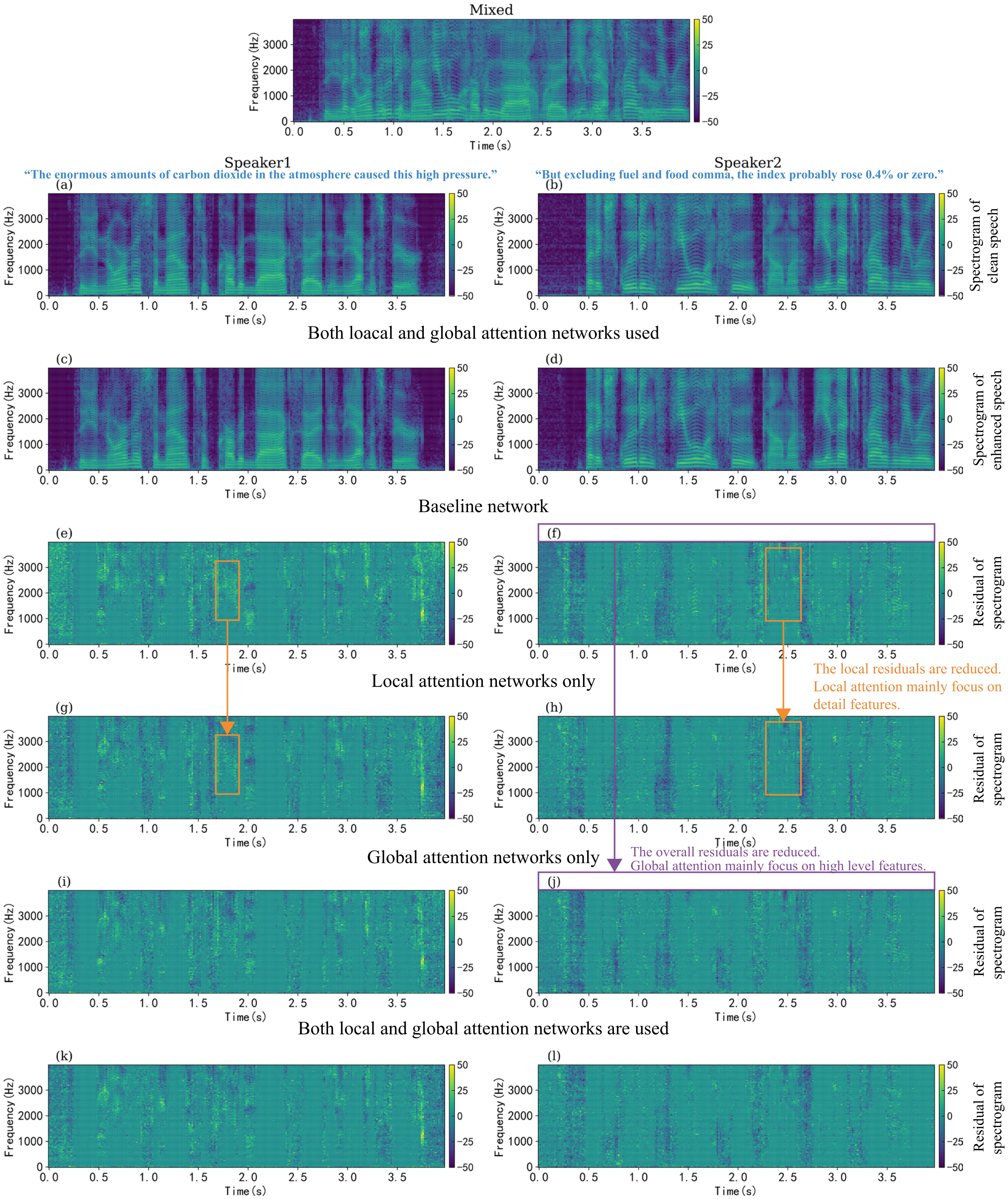

Fig. 4 Visualization results of the spectrogram of the separated samples. First row: spectrogram of mixed speech. Second row: spectrogram of clean speech. The fourth, fifth, sixth, and seventh rows represent the separation results by using baseline, local, global, and local and global attention networks, respectively, and then subtracting the spectrogram of clean speech

Table 9 Results on WHAM! and WHAMR!. Best values are marked with bold font

Model	WHAM		WHAMR	
	SI-SNRi	SDRi	SI-SNRi	SDRi
Chimera++ [57]	9.9	-	-	-
Conv-TasNet [49, 58]	12.7	-	8.3	-
Leamable fbank [58]	12.9	-	-	-
DPRNN [55]	13.9	-	10.3	-
SVOICE [55]	15.2	-	12.2	-
Wavesplit [21]	15.4	15.8	12.0	11.1
TAANet	**15.5**	15.8	12.0	**11.2**

Table 10 Comparison with baseline methods on GRID-2mix. Best values are marked with bold font

Model	SI-SNRi	SDRi
DPRNN	14.7	15.5
DPTNet	15.5	16.3
TAANet	**16.0**	**16.8**

6 Conclusion

In this paper, a time-domain adaptive attention network (TAANet) is proposed to realize single-channel speech separation. The TAANet is different from the existing speech separation methods directly using self-attention network in that the speech signal modeling is completed by using local attention networks and global attention networks, so that the model can simultaneously focus on frame-level and utterance-level features. With this adaptive structure, our model is able to take advantage of both local and global attention networks to better accomplish speech signal modeling. Extensive experiments have proved the excellent performance of this adaptive attention network. In the future, we would like to try different ways to connect the local and global attention blocks, and this can effectively reduce the model size. Furthermore, the attempt to use dynamic mixing technology for data augmentation will further improve the separation performance.

Abbreviations

TAANet	Time-domain adaptive attention network
STFT	Short-time Fourier transform
SOTA	State-of-the-art
SS	Speech separation
IBM	Ideal binary mask
TBM	Target binary mask
IRM	Deal ratio mask
DPCL	Deep clustering
uPIT	Utterance-level permutation-invariant training
CBAM	Convolutional block attention module
CAM	Channel attention module
SAM	Spatial attention module
SI-SNR	Negative scale-invariant source-to-noise ratio

Acknowledgements

The authors thank the associate editor and the anonymous reviewers for their constructive comments and useful suggestions.

Authors' contributions

In this paper, H. Zhou and K. Wang designed and programmed the proposed algorithm and wrote this paper; J. Cai and W. Li participated in the algorithm designation and tested the proposed method. J. Yao participated in the algorithm design, algorithm programming, and paper writing processes. All authors have read and approved the final manuscript.

Funding

This work is supported in part by National Key R &D Program of China under Grant 2021YFB1715000 and in part by Sichuan Science and Technology Program under Grant 2021YFG0315. We would like to thank Professor Hongsen He for meaningful discussion.

Availability of data and materials

The datasets analyzed during this article are available on the Internet.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 22 October 2022 Accepted: 31 March 2023
Published online: 11 May 2023

References

1. E.C. Cherry, Some experiments on the recognition of speech, with one and with two ears. J. Acoust. Soc. Am. **25**, 975–979 (1953)
2. S. Haykin, Z. Chen, The cocktail party problem. Neural Comput. **17**, 1875–1902 (2005)
3. D. Wang, J. Chen, Supervised speech separation based on deep learning: an overview. IEEE/ACM Trans. Audio Speech Lang. Process. **26**, 1702–1726 (2018)
4. M. Zhu, Z. Huang, X. Wang, X. Wang, C. Wang, H. Zhang, G. Zhao, S. Chen, G. Li, Automatic speech recognition in different languages using high-density surface electromyography sensors. IEEE Sensors J. **21**(13), 14155–14167 (2021). https://doi.org/10.1109/JSEN.2020.3037061
5. P. Dadvar, M. Geravanchizadeh, Robust binaural speech separation in adverse conditions based on deep neural network with modified spatial features and training target. Speech Commun. **108**, 41–52 (2019). https://doi.org/10.1016/j.specom.2019.02.001
6. Z.X. Li, Y. Song, L.R. Dai, I. McLoughlin, Listening and grouping: an online autoregressive approach for monaural speech separation. IEEE/ACM Trans. Audio Speech Lang. Process. **27**(4), 692–703 (2019). https://doi.org/10.1109/TASLP.2019.2892241
7. J. Byun, J.W. Shin, Monaural speech separation using speaker embedding from preliminary separation. IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 2753–2763 (2021). https://doi.org/10.1109/TASLP.2021.3101617
8. C. Fan, J. Tao, B. Liu, J. Yi, Z. Wen, X. Liu, End-to-end post-filter for speech separation with deep attention fusion features. IEEE/ACM Trans. Audio Speech Lang. Process. **28**, 1303–1314 (2020). https://doi.org/10.1109/TASLP.2020.2982029
9. Y. Wang, A. Narayanan, D. Wang, On training targets for supervised speech separation. IEEE/ACM Trans. Audio Speech Lang. Process. **22**, 1849–1858 (2014)
10. J. Hershey, Z. Chen, J.L. Roux, S. Watanabe, in *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, Deep clustering: Discriminative embeddings for segmentation and separation (IEEE, Shanghai 2016), pp. 31–35

11. D. Yu, M. Kolbæk, Z. Tan, J. Jensen, in *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP,* Permutation invariant training of deep models for speaker-independent multi-talker speech separation (IEEE, New Orleans, 2017), pp. 241–245
12. M. Kolbaek, D. Yu, Z. Tan, J. Jensen, Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. **25**, 1901–1913 (2017)
13. Y. Luo, Z. Chen, N. Mesgarani, Speaker-independent speech separation with deep attractor network. IEEE/ACM Trans. Audio Speech Lang. Process. **26**, 787–796 (2018)
14. Z. qiu Wang, J.L. Roux, J. Hershey, in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation (IEEE, Calgary, 2018), pp. 1–5
15. J. Park, J. Hong, J.W. Choi, M. Hahn, Determinant-based generalized sidelobe canceller for dual-sensor noise reduction. IEEE Sensors J. **22**(9), 8858–8868 (2022). https://doi.org/10.1109/JSEN.2022.3162619
16. A. Pandey, D. Wang, in *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, Tcnn: Temporal convolutional neural network for real-time speech enhancement in the time domain (IEEE, Brighton UK, 2019), pp. 6875–6879
17. Y. Luo, N. Mesgarani, in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, Tasnet: time-domain audio separation network for real-time, single-channel speech separation (IEEE, Calgary, 2018), pp. 696–700
18. Y. Luo, N. Mesgarani, Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. IEEE/ACM Trans. Audio Speech Lang. Process. **27**, 1256–1266 (2019)
19. Z. Shi, H. Lin, L. Liu, R. Liu, J. Han, in *International Conference on Multimedia Modeling*, Furcanext: End-to-end monaural speech separation with dynamic gated dilated temporal convolutional networks (Springer, Daejeon, 2020), pp. 653–665
20. Y. Luo, Z. Chen, T. Yoshioka, in *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, Dual-path rnn: Efficient long sequence modeling for time-domain single-channel speech separation (IEEE, Barcelona, 2020), pp. 46–50
21. N. Zeghidour, D. Grangier, Wavesplit: End-to-end speech separation by speaker clustering. IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 2840–2849 (2021)
22. J. Chen, Q. Mao, D. Liu, in *ISCA Interspeech,* Dual-path transformer network: direct context-aware modeling for end-to-end monaural speech separation (ISCA, Shanghai, 2020), pp. 2642–2646
23. C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, J. Zhong, in *Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, Attention is all you need in speech separation (IEEE, Toronto, 2021), pp. 21–25
24. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, u. Kaiser, I. Polosukhin, in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17),* Attention is all you need (MIT Press, California, 2017), pp. 6000–6010
25. M.W.Y. Lam, J. Wang, D. Su, D. Yu, in *Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, Sandglasset: a light multi-granularity self-attentive network for time-domain speech separation (IEEE, Toronto, 2021), pp. 5759–5763
26. M.W.Y. Lam, J. Wang, D. Su, D. Yu, in *2021 IEEE Spoken Language Technology Workshop (SLT)*. Effective low-cost time-domain audio separation using globally attentive locally recurrent networks. (2021), pp. 801–808
27. S. Woo, J. Park, J.Y. Lee, I.S. Kweon, in *European Conference on Computer Vision (ECCV),* CBAM: convolutional block attention module (Springer, Munich, 2018), pp. 3–19
28. G. Hu, D. Wang, in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (WASPAA),* Speech segregation based on pitch tracking and amplitude modulation (IEEE, New Paltz, 2001), pp. 79–82
29. D. Wang, On ideal binary mask as the computational goal of auditory scene analysis, in *Speech Separation by Humans and Machines*, Springer, 181-197 (2005)
30. U. Kjems, J.B. Boldt, M.S. Pedersen, T. Lunner, D. Wang, Role of mask pattern in intelligibility of ideal binary-masked noisy speech. J. Acoust. Soc. Am. **126**(3), 1415–1426 (2009)
31. Y. Li, D. Wang, On the optimality of ideal binary time-frequency masks. Speech Commun. **51**, 230–239 (2009)
32. Y. Isik, J.L. Roux, Z. Chen, S. Watanabe, J. Hershey, in *ISCA Interspeech,* Single-channel multi-speaker separation using deep clustering (ISCA, San Francisco, 2016), pp. 545–549
33. Z. Chen, Y. Luo, N. Mesgarani, in *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, Deep attractor network for single-microphone speaker separation (IEEE, New Orleans, 2017), pp. 246–250
34. J.R. Hershey, J.L. Roux, S. Watanabe, S. Wisdom, Z. Chen, Y.Z. Isik, in *New Era for Robust Speech Recognition, Exploiting Deep Learning,* Novel deep architectures in speech processing, Springer, 2017, pp. 135–164
35. D.S. Williamson, Y. Wang, D. Wang, Complex ratio masking for monaural speech separation. IEEE/ACM Trans. Audio Speech Lang. Process. **24**, 483–492 (2016)
36. K. Tan, D. Wang, Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement. IEEE/ACM Trans. Audio Speech Lang. Process. **28**, 380–390 (2020)
37. G. Yu, A. Li, H. Wang, Y. Wang, Y. Ke, C. Zheng, Dbt-net: Dual-branch federative magnitude and phase estimation with attention-in-attention transformer for monaural speech enhancement. IEEE/ACM Trans. Audio Speech Lang. Process. **30**, 2629–2644 (2022)
38. C. Lea, R. Vidal, A. Reiter, G. Hager, in *Computer Vision - ECCV 2016 Workshops*. Temporal convolutional networks: a unified approach to action segmentation. (2016)
39. C.S. Lea, M.D. Flynn, R. Vidal, A. Reiter, G. Hager, in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Temporal convolutional networks for action segmentation and detection (IEEE, Hawaii, 2017), pp. 1003–1012
40. K. Wang, J. Cai, J. Yao, P. Liu, Z. Zhu, Co-teaching based pseudo label refinery for cross-domain object detection. IET Image Process. **15**, 3189–3199 (2021)
41. S. Xu, E. Fosler-Lussier, in *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, Spatial and channel attention based convolutional neural networks for modeling noisy speech (IEEE, Brighton, 2019), pp. 6625–6629
42. S. Yadav, A. Rai, in *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, Frequency and temporal convolutional attention for text-independent speaker recognition (IEEE, Barcelona, 2020), pp. 6794–6798
43. S. Wang, S. Fernandes, Z. Zhu, Y. Zhang, AVNC: Attention-based VGG-style network for COVID-19 diagnosis by CBAM. IEEE Sensors Journal. **22**, 17431–17438 (2021)
44. S. Zhao, T.H. Nguyen, B. Ma, in *Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, Monaural speech enhancement with complex convolutional block attention module and joint time frequency losses (IEEE, Toronto, 2021), pp. 6648–6652
45. L. Dong, S. Xu, B. Xu, in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition (IEEE, Calgay, 2018), pp. 5884–5888
46. Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q.V. Le, R. Salakhutdinov, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL),* Transformer-xl: attentive language models beyond a fixed-length context (ACL, Florence, 2019), pp. 2978–2988
47. P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, J. Shlens, in *Neural Information Processing Systems (NeurIPS),* Stand-alone self-attention in vision models (MIT Press, Vancouver, 2019), pp. 68–80
48. G. Wichern, J.M. Antognini, M. Flynn, L.R. Zhu, E. McQuinn, D. Crow, E. Manilow, J.L. Roux, in *INTERSPEECH*. Wham!: extending speech separation to noisy environments. (2019)
49. M. Maciejewski, G. Wichern, E. McQuinn, J.L. Roux, in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Whamr!: Noisy and reverberant single-channel speech separation. (2020) pp. 696–700
50. M. Cooke, J. Barker, S. Cunningham, X. Shao, An audio-visual corpus for speech perception and automatic speech recognition. J. Acoust. Soc. Am. **120**(5 Pt 1), 2421–2424 (2006)
51. D.P. Kingma, J. Ba, Adam: a method for stochastic optimization. (2015). CoRR arXiv:1412.6980

52. E. Vincent, R. Gribonval, C. Févotte, Performance measurement in blind audio source separation. IEEE Trans. Audio Speech Lang. Process. **14**, 1462–1469 (2006)
53. Y. Liu, D. Wang, Divide and conquer: a deep casa approach to talker-independent monaural speaker separation. IEEE/ACM Trans. Audio Speech Lang. Process. **27**, 2092–2102 (2019)
54. Z. Shi, H. Lin, L. Liu, R. Liu, J. Han, A. Shi, in *ISCA Interspeech,* Deep attention gated dilated temporal convolutional networks with intra-parallel convolutional modules for end-to-end monaural speech separation (ISCA, Graz, 2019), pp. 3183–3187
55. E. Nachmani, Y. Adi, L. Wolf, in *Proceedings of the 37th International Conference on Machine Learning (ICML),* Voice separation with an unknown number of multiple speakers (ACM, California, 2020), pp. 7164–7175
56. X. Liu, X. Feng, W. Pedrycz, Extraction of fuzzy rules from fuzzy decision trees: an axiomatic fuzzy sets (afs) approach. Data Knowl. Eng. **84**, 1–25 (2013)
57. Z.Q. Wang, J.L. Roux, J.R. Hershey, in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, Alternative objective functions for deep clustering (IEEE, Calgary, 2018), pp. 686–690
58. M. Pariente, S. Cornell, A. Deleforge, E. Vincent, in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Filterbank design for end-to-end speech separation. (2020), pp. 6364–6368

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.