EURASIP Journal on Audio, Speech, and Music Processing

METHODOLOGY

Open Access

Deep encoder/decoder dual-path neural network for speech separation in noisy reverberation environments

Chunxi Wang[1], Maoshen Jia[1*] and Xinfeng Zhang[1]

Abstract

In recent years, the speaker-independent, single-channel speech separation problem has made significant progress with the development of deep neural networks (DNNs). However, separating the speech of each interested speaker from an environment that includes the speech of other speakers, background noise, and room reverberation remains challenging. In order to solve this problem, a speech separation method for a noisy reverberation environment is proposed. Firstly, the time-domain end-to-end network structure of a deep encoder/decoder dual-path neural network is introduced in this paper for speech separation. Secondly, to make the model not fall into local optimum during training, a loss function stretched optimal scale-invariant signal-to-noise ratio (SOSISNR) was proposed, inspired by the scale-invariant signal-to-noise ratio (SISNR). At the same time, in order to make the training more appropriate to the human auditory system, the joint loss function is extended based on short-time objective intelligibility (STOI). Thirdly, an alignment operation is proposed to reduce the influence of time delay caused by reverberation on separation performance. Combining the above methods, the subjective and objective evaluation metrics show that this study has better separation performance in complex sound field environments compared to the baseline methods.

Keywords Speech separation, Deep learning, Speech enhancement, SISNR

1 Introduction

Speech separation is widely known as the cocktail party problem [1, 2]. Its goal is to separate the target speaker's speech from complex sound field environments (other speakers, background noise, and reverberation). While human beings have a strong speech separation capability and can recognize the target speaker's speech even in complex environments; however, there is still a significant challenge for machine systems.

Speech separation, as an important front-end processing technique, is widely used in tasks such as hearing prosthesis, mobile communication, robust automatic speech, and speaker recognition. It has received extensive attention from researchers. However, the performance of current speech separation systems still needs to fully meet the requirements of human auditory perception, especially in complex sound field environments.

Speech separation has been studied for decades. In the early stages, following the assumption that speech signals conformed to a specific probability distribution (Gaussian or Laplacian) and that the background noise is stable (the spectral characteristics do not change with time), some methods such as Computational Auditory Scene Analysis (CASA) [3], Independent Component Analysis (ICA) [4] and Non-negative Matrix Factorization (NMF) [5] have been adopted in speech separation. In addition, considering the mask is essential in the process of speech separation, some mask estimation methods have been proposed based on probabilistic mixture models [6, 7] and sparse component analysis [8]. These methods

*Correspondence:
Maoshen Jia
jiamaoshen@bjut.edu.cn
[1] Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China

exhibit good separation performance in low-reverberation environments, but their performance decreases with the increase of reverberation time and/or noise levels.

With the development of deep learning, data-driven approaches have been introduced to speech separation. These methods learn features and patterns directly from the data without making any assumptions or prerequisites about the task domain. Based on this methodology, joint optimization of masking functions and deep recurrent neural networks is proposed for single-channel speech separation in the time domain [9]. Another approach involves operating in the complex domain with simultaneous enhancement of the magnitude and phase spectra to estimate the real and imaginary components of the ideal ratio mask [10].

However, two main difficulties have hindered the development of speech separation. These are the "permutation problem" and the "output dimension mismatch problem."

To solve the above problem, the permutation invariant training (PIT) [11] method is used in the training phase of the speech separation model to solve the uncertainty problem of speaker order in the mixed signals. Specifically, PIT lists all possible permutations and uses the minimum separation error to update the network. In the end, the source labels corresponding to the separated information are obtained.

In addition, deep clustering (DPCL) [12] is adopted to separate speech by calculating the embedding vector for each time–frequency bin and using the K-Means clustering method. Due to the use of permutation-free training, it can handle multiple sound sources simultaneously, achieving speaker-independent speech separation. However, the K-means clustering is utilized in DPCL [12], which requires significant computational resources. To overcome this issue, a deep attractor network (DANet) [13] is proposed to perform mask estimation without clustering. The time–frequency bins corresponding to each source are integrated by creating attractor subpoints in the high-dimensional space of the mixture signal. Compared with DPCL, the computational effort is significantly reduced. A variety of time–frequency mask-based separation methods have been successively proposed [14–16]. In order to achieve sufficient frequency resolution, phase/magnitude decoupling is inevitable for time–frequency decomposition, which results in imperfect reconstruction accuracy of the sources. In contrast, it can effectively avoid this problem when separation is performed in the time-domain. Hence, a long short-term memory time-domain audio separation network (LSTM-TasNet) [17] has been proposed, where a codec architecture is adopted to model the signal in the time-domain. Furthermore, a fully convolutional time-domain audio separation network (Conv-TasNet) [18] was proposed to solve the overfitting of LSTM-TasNet by using a temporal convolutional network (TCN) structure. As an end-to-end time-domain separation network, it can be used for modeling speech signals for a long-term dependency because of a deep one-dimensional dilated convolution block.

However, the input mixture signal is composed of a large number of time steps. In other words, if the receptive field of a one-dimensional convolutional neural network is smaller than the sequence length, it is difficult to achieve utterance-level modeling. Therefore, a dual-path recurrent neural network time-domain separation network (DPRNN-TasNet) [19] has been proposed to model long sequences through iterative intra- and inter-chunk operations. Similarly employing the dual-path strategy, dual-path transformer network (DPTNet) [20] utilizes a transformer module that enables long-term dependency modeling of speech signals. In addition, Wavesplit [21] achieved better speech separation performance by computing speaker vectors within a temporal window and obtaining the global vectors via clustering.

The above models [11–21] have been trained on the WSJ0-2mix [12] dataset through continuous updating and optimization, and the objective evaluation metric (Scale-invariant signal-to-noise ratio improvement (SISNRi)) has been improved continuously. In fact, the WSJ0-2mix is an ideal dataset in an anechoic acoustic environment (without the interference of noise and room reverberation), which contains only clean speech with two speakers. Through training on clean speech dataset, these methods can achieve excellent performance. Nevertheless, when the models [11, 14, 15, 17–21] are trained and tested on datasets with more complex sound field environments, the separation performance will degrade. Even careful optimization of the model's hyperparameters can only provide some relief [22, 23], and the improvement is insignificant.

In complex sound scenarios, it is a challenge to achieve satisfactory separation performance using single-channel information. Therefore, a training model is proposed based on azimuth and distance, using the distinct spatial locations of the speakers captured by a microphone array [24]. Moreover, multiple types of information, such as video information, have been adopted for speech separation [25, 26]. There is a certain improvement in separation performance among these methods but with higher requirements for recording equipment.

In the most recent study, TF-GridNet [27] achieves speaker separation through the utilization of complex spectral mapping, in conjunction with loss functions and DNN architectures. The model once again underscores its significant potential in the domain of time–frequency monaural speech separation.

In summary, motivated by the previous work, we propose a method for single-channel speech separation in complex sound field environments. End-to-end speech separation is performed using only a single microphone capture to obtain information in the presence of noise and reverberant interference. The contributions of the proposed method are summarized as follows:

- A network structure of a deep encoder/decoder dual-path neural network is proposed, which enhances the model's ability to extract speech features. Experimental results demonstrate the feasibility of the method.
- A new loss function known as the stretched optimal scale-invariant signal-to-noise ratio (SOSISNR) is proposed, and experimental results show that it outperforms the scale-invariant signal-to-noise ratio (SISNR) in complex sound field environments.
- Using a multi-objective joint optimization strategy, the loss function was extended based on short-time objective intelligibility (STOI) [28] to match the human auditory system better.
- The alignment operation is proposed to reduce the model's reliance on a priori knowledge of the sound field and to increase the robustness of the model.

The rest of the paper is organized as follows. Sections II and III describe the specific implementation steps of the proposed method and the rationale. Section IV describes the experimental procedure. Finally, the conclusions and analysis of the experiments are presented in section V.

2 Description of the separation model

2.1 Problem formulation

In a multi-source scenario, the signal $y(t)$ recorded by a mono microphone can be modeled in the time-domain as:

$$y(t) = \sum_{i=1}^{I} s_i(t) * r_i(t) + \sigma(t), \tag{1}$$

where $i = 1, 2 \ldots, I$, I represents the number of the sound sources. t represents continuous time, indicating signal's continuous variation along the time dimension. $s_i(t)$ and $r_i(t)$ denote the i th speaker's speech and the room impulse responses (RIRs) between the i th speaker and the microphone, respectively, and "$*$" denotes the convolution operation. $\sigma(t)$ denotes the background noise.

The aim of this paper is to obtain/estimate each speaker's speech $s_i(t)$, $i = 1, 2 \ldots, I$, from a recorded signal contaminated by noise and reverberation.

2.2 Deep encoder/decoder

Currently, the architecture of encoder, decoder and separator [18–20, 29, 30] have been adopted in speech separation.

Specifically, the role of the encoder is to extract features from mixture signals and map them to a feature space of an appropriate dimension. The features are then analyzed and processed by the separator to separate the high-dimensional representation of each component in the mixed speech. Finally, the decoder uses the separated features to reconstruct the original speech signal. The existing works have mainly focused on the separator, with relatively less attention paid on the encoder and decoder part. The linear (shallow) operators are commonly used to extract features, which can limit the expressiveness of the model and may not achieve satisfactory performance in complex sound scenarios.

In order to explore the transformation capabilities of the deep encoder/decoder structure on complex signals, this paper attempts to utilize this structure to focus more on the local features of the speech signal [31]. By concurrently integrating advanced dual-path neural network separation modules, we aspire for the model to exhibit superior separation performance. Therefore, a speech separation method for noisy and reverberant environments is proposed by combining a deep encoder/decoder and a dual-path neural network. The deep encoder/decoder structure is shown in Fig. 1.

As shown in Fig. 1, the encoder obtains the single-channel signal $y(t)$ recorded by the microphone, resamples this signal, and transforms it into a matrix dimension as $\mathbf{Y} \in \mathbb{R}^{1\times S}$, where S represents the number of the time steps. A one-dimensional convolution operation is performed in the first layer of the encoder, expressed as follows:

$$\mathbf{E}_1 = Conv(\mathbf{Y}, \mathbf{V}), \tag{2}$$

where $\mathbf{V}$ represents a kernel of size L and stride $L/2$, and $Conv(\cdot)$ represents a one-dimensional convolution operation. $\mathbf{E}_1 \epsilon \mathbb{R}^{K\times S}$ is obtained by one-dimensional convolution of the kernel $\mathbf{V}$, where K is the feature dimension of the signal. The recording signal $y(t)$ is initially mapped with a linear transform. Then, starting from the second layer, the input of the p th layer is the output of the $(p-1)$ th layer. The output of the p th layer can be expressed as:

$$\mathbf{E}_p = PReLU\big(Conv\big(\mathbf{E}_{p-1}, \mathbf{V}_p\big)\big), \tag{3}$$

where p=2, 3, ..., P, p is the index of the encoder layer and P is the number of deep encoder layers ($P = 4$ in this paper). The p th layer has a kernel $\mathbf{V}_p$ of size 3 with stride 1 and a $PReLU$, and $\mathbf{E}_p \epsilon \mathbb{R}^{K\times S}$ represents the output of the p th layer. In this way, the recording signal $y(t)$ is further mapped into a non-linear potential space by stacking the encoding layers. Using the

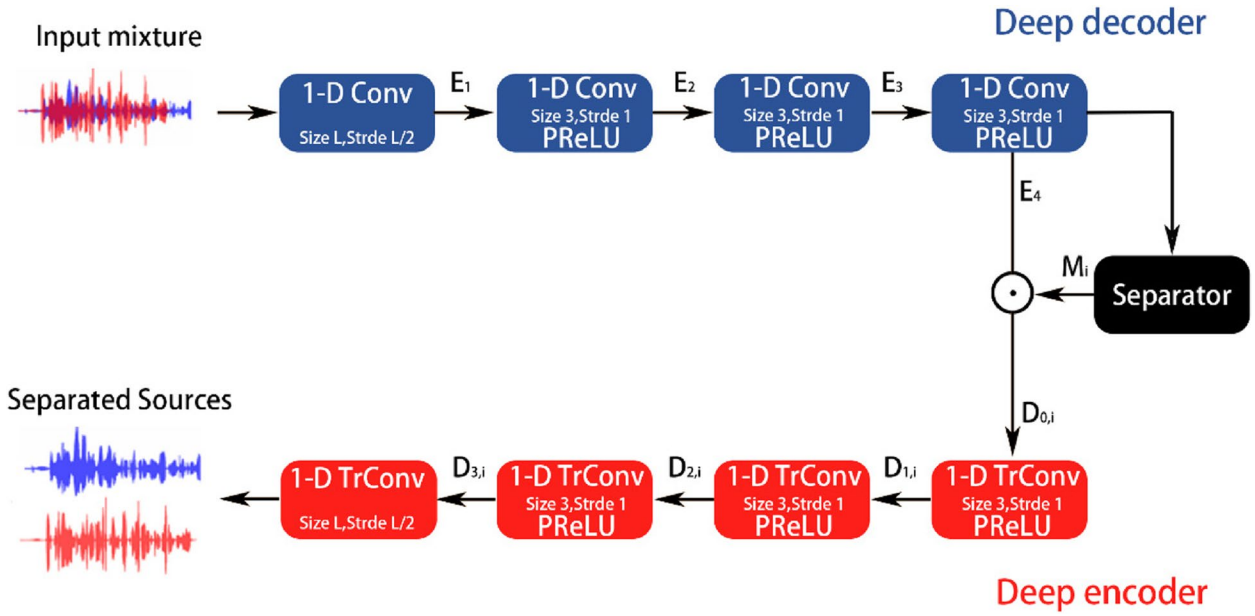

Fig. 1 The deep encoder/decoder structure

output from the deep encoder, the separation module estimates the mask $\mathbf{M}_i$ corresponding to the i th speaker.

The input feature $\mathbf{D}_{0,i}$ of the decoder can be obtained from the output of the deep encoder and the mask $\mathbf{M}_i$ as follow:

$$\mathbf{D}_{0,i} = \mathbf{E} \odot \mathbf{M}_i, \tag{4}$$

where $\mathbf{D}_{0,i} \epsilon \mathbb{R}^{K \times S}$, "0" represents the layer 0 of the decoder, which is the input of the deep decoder. $\mathbf{E}$ is the output of the deep encoder, and i represents the index of the speaker. "$\odot$" represents the element-wise multiplication.

The decoder reconstructs the waveform by transforming the two-dimensional feature $\mathbf{D}_{0,i} \epsilon \mathbb{R}^{K \times S}$. Starting from the first layer of the decoder, the input of the q th layer is the output of the $(q-1)$ th layer. The output of the q th layer can be expressed as:

$$\mathbf{D}_{q,i} = PReLU\left(TrConv\left(\mathbf{D}_{q-1,i}, \mathbf{U}_q\right)\right), \tag{5}$$

where q=1, 2, ..., Q, q is the index of the decoder layer, and $(Q+1)$ is the number of deep decoder layers ($Q = 3$inthispaper). $i = 1, 2 \cdots, I$, I represents the number of the sound sources. $TrConv(\cdot)$ represents the transpose convolution operation. The q th layer of the decoder has a kernel $\mathbf{U}_q$ of size 3 with stride 1 and a $PReLU$, $\mathbf{D}_{q,i} \epsilon \mathbb{R}^{K \times N}$ denotes the output of the q th layer. The fourth layer performs the transpose convolution operation of the kernel $\mathbf{U}$ as follows:

$$\widehat{\mathbf{s}}_i = TrConv\left(\mathbf{D}_{3,i}, \mathbf{U}\right), \tag{6}$$

where $\mathbf{U}$ represents a kernel of size L and stride $L/2$, $\mathbf{D} \epsilon \mathbb{R}^{K \times S}$ is a high-dimensional representation of the decoder reconstructed waveform, and $\widehat{\mathbf{s}}_i \epsilon \mathbb{R}^{1 \times S}$ is the reconstructed speech signal.

2.3 DPRNN separation module

The DPRNN module is used as a separator in this paper, which consists of three operations: segmentation, DPRNN block processing, and overlapping-add [19]. The overall flowchart is shown in Fig. 2:

The output of the deep encoder $\mathbf{E} \epsilon \mathbb{R}^{K \times S}$ is obtained according to formula (3), where K and S can be considered as the feature dimension and the time dimension, respectively. As shown in Fig. 2, in the segmentation stage, the output of the deep encoder is segmented into F chunks of equal size, where each chunk with length H and hop size $\hbar_p$ ($\hbar_p = \frac{H}{2}$), and the first and last chunks are zero-padded (so that all the output of the deep encoder can be processed). Each block can be represented as $\mathbf{T}_f \in \mathbb{R}^{K \times H}$, where $f = 1, 2, \ldots, F$, f is the index of the block, F represents the number of the blocks. The combined information of the F chunks into a three-dimensional tensor $\mathbf{G} = [\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_F] \in \mathbb{R}^{K \times H \times F}$.

In the DPRNN block processing stage, the tensor $\mathbf{G}$ is passed to a stack of C DPRNN blocks. Each block contains intra-block and inter-block. Intra-block processing (local modeling) and inter-block processing (global modeling) are performed iteratively, while keeping the

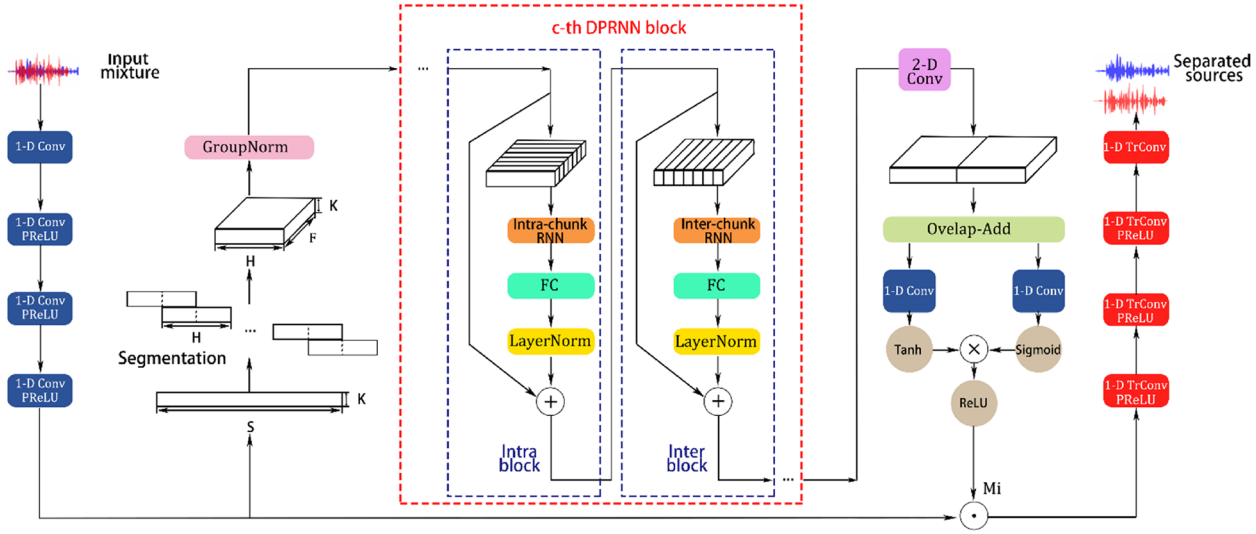

Fig. 2 The overall flowchart

dimensionality of the three-dimensional tensor constant, so that results can be learned for different time dimensions.

Specifically, intra-block processing starts with the intra-chunk RNN in the third dimension of tensor **G**, as follows:

$$\mathbf{W}_c = \left[\mathrm{h}_c\left(\mathbf{G}_{c-1}[:,:,j]\right), j = 1, 2, \ldots, F\right], \tag{7}$$

where c=1, 2, …, C, c is the stack index of the DPRNN block, and C is the number of repetitions of the DPRNN block. $\mathbf{G}_0 = \mathbf{G}$, $\mathbf{W}_c \epsilon \mathbb{R}^{X \times H \times F}$ is the output of the intra-chunk RNN, and $\mathrm{h}_c(\cdot)$ represents the mapping function of the intra-chunk RNN. $\mathbf{G}_{c-1} \in \mathbb{R}^{K \times H \times F}$ is the speech feature of the previous layer of the three-dimensional tensor. Further, to ensure that the dimension of the tensor does not change, a linear fully-connected (FC) layer is applied, which is defined as follows:

$$\widehat{\mathbf{W}}_c = \mathbf{J}\mathbf{W}_c + \mathbf{a}, \tag{8}$$

where $\widehat{\mathbf{W}}_c \in \mathbb{R}^{K \times H \times F}$ is the transformed speech feature tensor and $\widehat{\mathbf{W}}_c$ has the same dimension as $\mathbf{G}_{c-1}$. $\mathbf{J} \in \mathbb{R}^{K \times X}$ is the weight tensor of the FC layer, and $\mathbf{a}$ is the bias of the FC layer.

To improve the generalization ability of the model, a layer normalization (LN) was performed on the transformed speech feature tensor $\widehat{\mathbf{W}}_c$. In addition, to avoid model degradation during training, a residual connection is added between the input $\mathbf{G}_{c-1}$ of the intra-chunk RNN and the LN, as shown in formal (9):

$$\widehat{\mathbf{G}}_c = \mathbf{G}_{c-1} + \mathrm{LN}\left(\widehat{\mathbf{W}}_c\right), \tag{9}$$

where $\widehat{\mathbf{G}}_c \in \mathbb{R}^{K \times H \times F}$ is the output of the intra-block processing as well as the input of the inter-block processing, $\mathrm{LN}(\cdot)$ is the layer normalization operation.

The inter-block processing is performed in the second dimension of the three-dimensional tensor $\widehat{\mathbf{G}}_c$, i.e., inter-chunk RNN in the second dimension of tensor $\widehat{\mathbf{G}}_c$, which is shown as follows:

$$\mathbf{R}_c = \left[\mathrm{v}_c\left(\widehat{\mathbf{G}}_c[:,j,:]\right), j = 1, 2, \ldots, H\right], \tag{10}$$

where $\mathbf{R}_c \epsilon \mathbb{R}^{X \times H \times F}$ is the output of the inter-chunk RNN and $\mathrm{v}_c(\cdot)$ represents the mapping function of the inter-chunk RNN. The subsequent operations are similar to those of the intra-block, which is consisted of inter-chunk RNN mapping, the FC and LN layers. The output of the c th DPRNN block is represented as $\mathbf{G}_c \epsilon \mathbb{R}^{K \times H \times F}$. So, the final output of the block processing, i.e., $\mathbf{G}_C \epsilon \mathbb{R}^{K \times H \times F}$, is obtained by repeating C DPRNN blocks.

Finally, the two-dimensional convolutional layer learns the mask of each sound source and performs an overlapping-add operation. The mask $\mathbf{M}_i \epsilon \mathbb{R}^{K \times S}$ corresponding to the i th speaker can be obtained.

3 Training objective

3.1 Joint loss function

In clean speech separation tasks, the loss function is often based on SISNR and utterance-level permutation invariant training (uPIT) [18, 19], with the starting point to get the model to predict signals closer and closer to the original clean signal.

A new loss function, SOSISNR, was designed to cope with single-channel speech separation in complex sound field environments. Meanwhile, the loss function was

extended using STOI to improve the intelligibility of separated speech and make it more compatible with the human auditory system.

Using a multi-objective joint optimization strategy, the joint loss function $\mathcal{L}$ is expressed as follows:

$$\mathcal{L} = -\mathcal{L}_{\text{sosisnr}} - \lambda \cdot \mathcal{L}_{\text{STOI}}, \tag{11}$$

where $\mathcal{L}_{\text{sosisnr}}$ is the loss function SOSISNR, and $\mathcal{L}_{\text{STOI}}$ is the STOI-based loss function. λ is the weight of $\mathcal{L}_{\text{STOI}}$, which is set to 2 in this paper. The derivation and analysis of the two loss functions (i.e., $\mathcal{L}_{\text{sosisnr}}$, $\mathcal{L}_{\text{STOI}}$) are given in Sub-sections III.B and III.C, respectively.

3.2 Stretched optimal scale-invariant signal-to-noise ratio

SISNR, as a time-domain loss function, can be used to measure the similarity between the output of the model and ground truth. It has been widely used in signal processing fields such as speech separation [18, 19] and speech enhancement [32, 33]. The illustration of SISNR and SOSISNR are shown in Fig. 3.

As shown in the orange part, the original clean speech s is first mapped to obtain the target signal s_{target}, in order to attenuate the effect of scale variations due to either the original clean speech s or the estimated signal $\hat{s}$. The target signal s_{target} is defined in formula (12):

$$\mathbf{s}_{target} = \frac{\langle \widehat{\mathbf{s}}, \mathbf{s} \rangle \mathbf{s}}{\|\mathbf{s}\|^2}. \tag{12}$$

And then the distance between the estimated signal $\widehat{\mathbf{s}}$ and the target signal $\mathbf{s}_{target}$, i.e., the noise signal $\mathbf{e}_{noise}$, is defined as follows:

$$\mathbf{e}_{noise} = \widehat{\mathbf{s}} - \mathbf{s}_{target}. \tag{13}$$

Finally, the intensity ratio of the target signal $\mathbf{s}_{target}$ to the noise signal $\mathbf{e}_{noise}$ (SISNR) is defined as:

$$\text{SISNR} = 10log_{10}\frac{||\mathbf{s}_{target}||^2}{||\mathbf{e}_{noise}||^2}. \tag{14}$$

Furthermore, $\mathbf{s}_{target}$ and $\mathbf{e}_{noise}$ can be updated to the following formula:

$$\mathbf{s}_{target} = \frac{|\widehat{\mathbf{s}}|\cos(\theta)}{|\mathbf{s}|}\mathbf{s}, \tag{15}$$

$$\mathbf{e}_{noise} = \widehat{\mathbf{s}} - \frac{|\widehat{\mathbf{s}}|\cos(\theta)}{|\mathbf{s}|}\mathbf{s}, \tag{16}$$

Combing formula (14), (15), (16), the expression of SISNR is further derived as:

$$\begin{aligned}\text{SISNR} &= 10log_{10}\frac{|\widehat{\mathbf{s}}|^2\cos^2(\theta)}{|\widehat{\mathbf{s}}|^2+|\widehat{\mathbf{s}}|^2\cos^2(\theta)-2\frac{|\widehat{\mathbf{s}}|}{|\mathbf{s}|}\cos(\theta)|\widehat{\mathbf{s}}|\ |\mathbf{s}|\ \cos(\theta)} \\ &= 10log_{10}cot^2(\theta)\end{aligned} \tag{17}$$

where θ represents the angle between $\widehat{\mathbf{s}}$ and $\mathbf{s}$. Obviously, as in formula (15), by adjusting the original clean speech $\mathbf{s}$ to a suitable scale, the magnitude of $\mathbf{s}_{target}$ is not associated with the original clean speech $\mathbf{s}$. Instead, $\mathbf{s}_{target}$ is expressed in terms of the estimated signal $\widehat{\mathbf{s}}$ as well as the angle θ trigonometric function, with no change in direction compared to $\mathbf{s}$. As in formula (17), SISNR is only related to the angle θ, which is irrelevant to the magnitude of $\widehat{\mathbf{s}}$ and $\mathbf{s}$. The functional relationship between θ and SISNR is shown as the orange part in Fig. 4.

In separation performance evaluation, the angle information between the separated signal and the original signal is more important than the magnitude

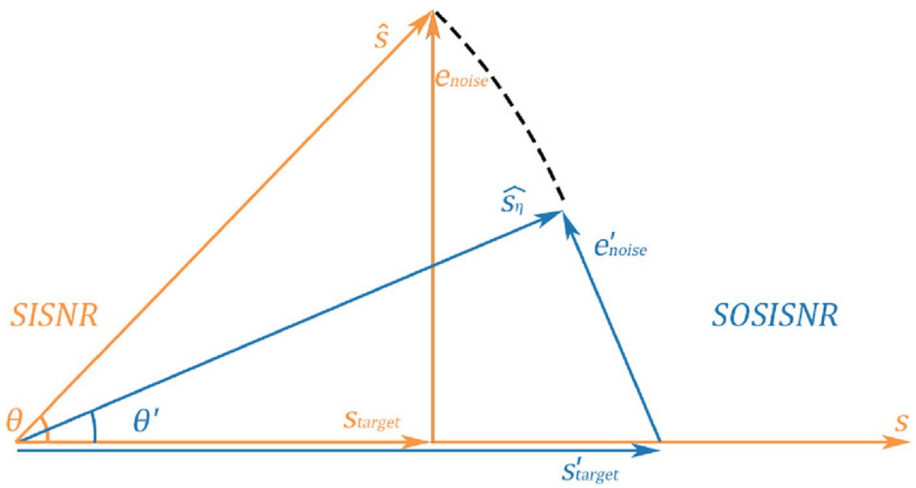

Fig. 3 The illustration of SISNR and SOSISNR

information. Compared to the signal-to-noise ratio (SNR) [34]:

$$\mathrm{SNR} = 10log_{10}\frac{\|\mathbf{s}\|^2}{\|\widehat{\mathbf{s}}-\mathbf{s}\|^2}, \tag{18}$$

SISNR is more suitable for speech separation tasks. Specifically, by regularizing the separated speech signal $\mathbf{s}$, SISNR overcomes the disadvantage that SNR is susceptible to variations in the energy of input signal. In other words, SISNR is able to evaluate the angle between the separated signal and the original signal without being affected by the change of signal energy [35, 36].

In fact, SISNR is not necessarily the optimal choice for speech separation in complex acoustic environments. As shown in the orange part of Fig. 3, the $\mathbf{e}_{noise}$ of SISNR is not necessarily orthogonal to $\mathbf{s}$. Using SISNR may mislead the model in complex environments, which often causes the model to fall into a local best [35]. Meanwhile, as the orange curve shown in Fig. 4, there are several extreme points (i.e., $0, \pm\pi$) over a period ($-\pi \sim \pi$). The closer the angle θ is to $-\pi$ or π, the larger the SISNR value. This means that $\widehat{s}$ and s are similar under such angles. Obviously, such a conclusion is incorrect [36].

In order to have only one correct extreme point in the range of $-\pi$ to π, we reconstructed a phase-corrected signal $\widehat{\mathbf{s}}_\eta$. The goal is to double the period of the SISNR to reduce the error extreme points. At the same time, in order to prevent the training from falling into the local optimum, we define a new target signal in the vector $\mathbf{s}$ direction. The new target signal $\mathbf{s}'_{target}$ is obtained by scaling the original signal to make its amplitude independent of the original signal.

Based on the above derivation, we propose a loss function, SOSISNR. The illustration of SOSISNR is shown in the blue part of Fig. 3. Specifically, $\widehat{\mathbf{s}}_\eta$ is reconstructed based on the spatial relationship between the estimated signal $\widehat{\mathbf{s}}$ and the original speech$\mathbf{s}$. This is achieved by keeping the magnitude of the reconstructed estimated signal $\widehat{\mathbf{s}}_\eta$ equal to that of the original clean speech $\widehat{\mathbf{s}}$ ($|\widehat{\mathbf{s}}_\eta| = |\widehat{\mathbf{s}}|$) and halving the angle between the $\widehat{\mathbf{s}}$ and $\mathbf{s}$ $\left(\theta' = \frac{\theta}{2}\right)$, where θ' represents the angle $\widehat{\mathbf{s}}_\eta$ and$\mathbf{s}$.

Meanwhile, the new target signal $\mathbf{s}'_{target}$ is defined as follows:

$$\mathbf{s}'_{target} = \alpha\mathbf{s}, \tag{19}$$

where α represents the scale adjust factor. Substituting formula (19) into formula (14), SISNR$\prime$ is obtained:

$$\mathrm{SISNR}' = 10log_{10}\frac{||\alpha\mathbf{s}||^2}{||\widehat{\mathbf{s}}_\eta-\alpha\mathbf{s}||^2}. \tag{20}$$

To simplify the calculation, an intermediate function $f(\cdot)$ is defined as follows:

$$f(\alpha) = \frac{\|\alpha\mathbf{s}\|^2}{\|\widehat{\mathbf{s}}_\eta-\alpha\mathbf{s}\|^2}. \tag{21}$$

In order to obtain the scale adjustment factor α corresponding to the maximum value of $f(\alpha)$, the derivative of $f(\alpha)$ is calculated below:

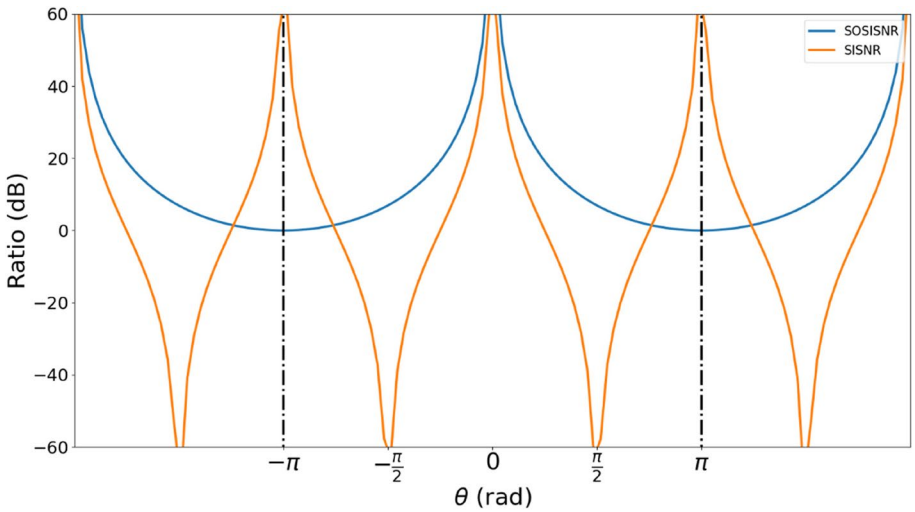

Fig. 4 The functional relationship between θ and ratio

$$\frac{df(\alpha)}{d\alpha} = \frac{2\alpha \mathbf{s}^2\left(\widehat{\mathbf{s}}_\eta^2 - \alpha\widehat{\mathbf{s}}_\eta\mathbf{s}\right)}{\left(\widehat{\mathbf{s}}_\eta - \alpha\mathbf{s}\right)^4} = 0 \tag{22}$$

Therefore, α can be obtained:

$$\alpha = \frac{\left|\widehat{\mathbf{s}}_\eta\right|^2}{\langle\widehat{\mathbf{s}}_\eta, \mathbf{s}\rangle}. \tag{23}$$

After adjusting the period of SISNR and recalculating the scale adjust factor α, with (23) and (19), the reconstructed target signal $\mathbf{s}'_{target}$ can be rewritten as:

$$\mathbf{s}'_{target} = \frac{\left|\widehat{\mathbf{s}}_\eta\right|^2}{\langle\widehat{\mathbf{s}}_\eta, \mathbf{s}\rangle}\mathbf{s} = \frac{\left|\widehat{\mathbf{s}}_\eta\right|}{|\mathbf{s}|cos\theta'}\mathbf{s}. \tag{24}$$

Substituting formula (24) into (13), the noise $\mathbf{e}'_{noise}$ corresponding to the reconstructed target signal $\mathbf{s}_{target}{}'$ is given as follows:

$$\mathbf{e}'_{noise} = \widehat{\mathbf{s}}_\eta - \frac{\left|\widehat{\mathbf{s}}_\eta\right|}{|\mathbf{s}|cos\theta'}\mathbf{s}. \tag{25}$$

Furthermore, the loss function SOSISNR is defined by combined formula (24) and (25) as follows:

$$\begin{aligned} \text{SOSISNR} &= 10log_{10}\frac{\|\mathbf{s}'_{target}\|^2}{\|\mathbf{e}'_{noise}\|^2} \\ &= 10log_{10}csc\left(\theta'\right) \\ &= 10log_{10}\frac{2}{1-\cos^2(\theta)}. \end{aligned} \tag{26}$$

The functional relationship of θ and SOSISNR value are shown in the blue part in Fig. 4. There is only one extreme point of SOSISNR with θ in the range from $-\pi$ to π. SOSISNR reaches its maximum value only when the angle θ is zero, which reflects the similarity of $\widehat{\mathbf{s}}$ and $\mathbf{s}$ correctly. Also, unlike SISNR $\epsilon(-\infty, +\infty)$, SOSISNR ranges from 0 to positive infinity.

3.3 STOI-based loss function

Although SISNR can well reflect the correlation between the estimated speech $\widehat{s}$ and the original clean speech s. However, as a distance-based loss function, SISNR cannot directly reflect the effect of signals on human hearing [37]. In contrast, STOI is a commonly used objective evaluation metric [23] (STOI $\epsilon[0, 1]$, a higher value representing better speech intelligibility), which is closely related to human auditory perception [38]. Moreover, it analyzes speech segments as a whole [39], which is more conducive to learning long-range context dependencies. Based on this motivation, the STOI-based loss function $\mathcal{L}_{\text{STOI}}$ has been extended to a joint loss function [40].

Taking the estimated speech $\hat{\mathbf{s}}$ and the original clean speech $\mathbf{s}$ as input, $\mathcal{L}_{\text{STOI}}$ can be obtained as follows:

a. Removal of silent frames from estimated speech $\widehat{\mathbf{s}}$ and original clean speech $\mathbf{s}$.
b. The short-time Fourier transform (STFT) is used to obtain the corresponding representation in the time–frequency domain.
c. Perform a one-third octave band analysis.
d. To compensate for the global level difference and improve the stability of the STOI, normalization and clipping are performed.
e. Measure Intelligibility. The intermediate intelligibility $\zeta_{b,n}$ is defined as the spectral correlation coefficients between the two temporal envelopes:

$$\zeta_{b,n} = \frac{\left(\widehat{\mathbf{s}}_{b,n} - m_{\widehat{\mathbf{s}}_{b,n}}\right)^T\left(\mathbf{s}_{b,n} - m_{\mathbf{s}_{b,n}}\right)^T}{\|\widehat{\mathbf{s}}_{b,n} - m_{\widehat{\mathbf{s}}_{b,n}}\|_2\|\mathbf{s}_{b,n} - m_{\mathbf{s}_{b,n}}\|_2} \tag{27}$$

where b and n are the indexes of the one-third octave and the short-time temporal envelope vectors, respectively. $b = 1, 2, \ldots, B$, and $n = 1, 2, \ldots, N$. B and N are the numbers of one-third octave bands and the short-time temporal envelope vectors, respectively. $\widehat{\mathbf{s}}_{b,n}$ and $\mathbf{s}_{b,n}$ represent the short-time spectrogram vector of the estimated speech $\widehat{\mathbf{s}}$ and the original clean speech $\mathbf{s}$, respectively. $m(\cdot)$ is the sample mean of the corresponding vector.

Ultimately, $\mathcal{L}_{\text{STOI}}$ can be obtained by averaging the intermediate intelligibility of all bands and short-time temporal envelope vectors:

$$\mathcal{L}_{\text{STOI}} = \frac{1}{BN}\sum_{b=1}^{B}\sum_{n=1}^{N}\zeta_{b,n}. \tag{28}$$

Choosing the appropriate loss function is crucial for training and optimizing deep learning models. When using the joint loss function, it is necessary to ensure that each loss function is appropriate that their numerical range and symbol selection can ensure the effectiveness of parameter update and optimization. This is to avoid the problem of gradient disappearance, which occurs when the gradients of different loss functions cancel each other out. When the gradient disappears, the network cannot effectively perform back propagation and update and cannot be optimized in the right direction. The value of $\mathcal{L}_{\text{STOI}}$ ranges from 0 to 1 in formula (11), so it is necessary to ensure that the other loss function in the joint loss function $\mathcal{L}$ is non-negative. SOSISNR just meets this condition.

3.4 Alignment operation and utterance-level permutation invariant training

This paper focuses on speech separation in complex sound field environments. The input mixture speech

signals in the training process are heterogeneous, with different levels of reverberation. In order to make the network cope with complex environments, this paper proposes the alignment operation to reduce the time delay error caused by reverberation. Specifically, this method performs time alignment on the estimated speech $\widehat{\mathbf{s}}$ and the original clean speech $\mathbf{s}$, so that the network can learn the corresponding relationship between them more accurately and improve the separation performance.

After obtaining the joint loss function $\mathcal{L}$ according to formula (11), the loss function needs to be modified in order to achieve the best separation performance. The alignment operation is proposed so that the loss function can better reflect the separation performance of the model, allowing the model to learn more effective information. The alignment operation is shown in Fig. 5.

As shown in Fig. 5, the specific method is to obtain $\mathbf{s}^{\tau}$ by cyclically shifting the original clean speech $\mathbf{s}$. Here τ is an integer value, which represents the number of shift samples. Then, the original clean speech $\mathbf{s}$ is replaced by the shifted original speech $\mathbf{s}^{\tau}$. The joint loss function $\mathcal{L}$ is calculated by inputting estimated speech $\widehat{\mathbf{s}}$, and each shifted original speech $\mathbf{s}^{\tau}$ into the formula (11). Comparing all the loss values with different τ, the minimum loss function value is found as shown in formula (29):

$$\mathcal{L}_A\left(\widehat{\mathbf{s}}, \mathbf{s}\right) = \min_{\tau} \mathcal{L}\left(\widehat{\mathbf{s}}, \mathbf{s}^{\tau}\right) \tag{29}$$

where $\widehat{\mathbf{s}} \epsilon \mathbb{R}^{1\times S}$ and $\mathbf{s}^{\boldsymbol{\tau}} \epsilon \mathbb{R}^{1\times S}$ represent the estimated speech and the shifted original speech, respectively, and S represents the length of the tensor after sampling processing. The value of τ ranges from 1 to S.

The introduction of this operation enables the model to reduce its reliance on a priori knowledge of sound field environments. It enables the model to cope with different levels of reverberation (offsetting the time delay caused by reverberation) and to significantly reduce the workload of aligning annotations to different speech signals. At the same time, at each epoch of the network training, the operation can be propagated forward to provide timely and appropriate feedback to the model.

Furthermore, to solve permutation ambiguity during training [39], combined with the alignment operation, the utterance-level permutation invariant training was introduced. It can guide the model to train a speaker-independent separation model. The updated loss function is as follows:

$$\mathcal{L}_{PIT^A} = \sum_{i=1}^{I} \min_{\gamma(i)\epsilon\Gamma} \mathcal{L}_A\left(\widehat{\mathbf{s}}_{(i)}, \mathbf{s}_{\gamma(i)}\right) \tag{30}$$

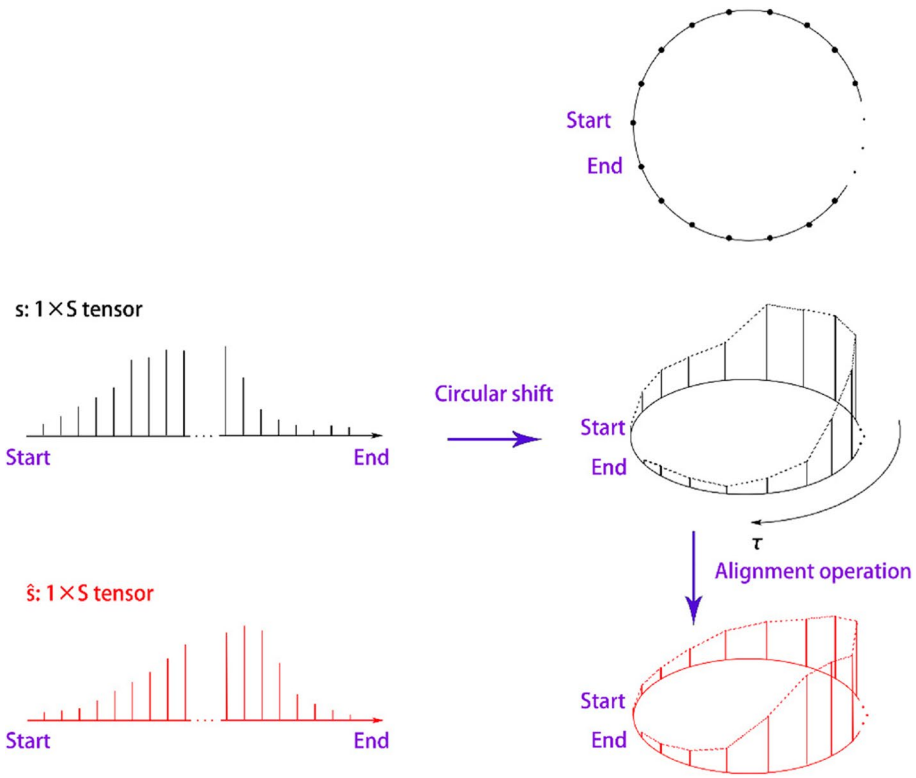

Fig. 5 Illustration of alignment operation

where i represents the speaker index, $i = 1, 2 \cdots, I$, and I represents the number of speakers. $\widehat{s}_{(i)}$ is the i th estimated speaker speech. $s_{\gamma(i)}$ represents the original clean speech of the $\gamma(i)$ th speaker, $\gamma(i)$ represents the possible index of the original clean speech corresponding to the i th estimated speaker speech. Γ is the set of all possible permutations for all I speakers.

4 Experimental settings

4.1 Dataset

The complex scenario where the speech source of interest is disturbed by other speakers, noise, and reflection components simultaneously is simulated for the experiment. We simulated RIRs using an image method [41–43] for a rectangular room with dimensions of 7 m×5 m×3 m, with the microphone placed in the center of the room (3.5 m×2.5 m×1.5 m). Reverberant utterances from different speakers with reverberation time (T_{60}) of 100 ms, 200 ms, and 300 ms were randomly generated by varying the sound absorption coefficient of the walls. The speech signal from the si_tr_s dataset of the Wall Street Journal dataset (WSJ0) [12] is chosen as the source signal. In addition, two reverberant utterances from different speakers were randomly selected and mixed with an SNR between -5 dB to 5 dB to generate 30 h of training data. Similarly, the validation and test set (from WSJ0 si_dt_05 and si_et_05) were generated in the same way to produce 10 h and 5 h of data, respectively. A spatial version of the WSJ0-2mix dataset [12] was generated in this way. Based on this, we paired the spatial version of WSJ0-2mix with noisy audio (including noise background scenes such as restaurants, cafés, and bars) from the WSJ0 Hipster Ambient Mixtures (WHAM!) dataset [44]. Then, we generated randomly mixed speech in the WSJ0-2mix dataset with noise at three SNR levels of 5 dB, 10 dB, and 15 dB. This process was designed for speech separation tasks in environments with varying levels of background noise. All speech signals were resampled at 8 kHz.

4.2 Training setup

In the first layer of the deep encoder and the last layer of the deep decoder, the kernel size L is set to 2, and the hop size is $L/2$. In the separation module, the number of overlapping stacks C of DPRNN blocks is set to be 6, and BLSTM [45] with 128 hidden units in each direction is used as intra- and inter-block RNN.

In the STOI-based loss function, the time–frequency spectrum of the speech signal is obtained by STFT with the Hanning window length and the hop size are 1024 and 256, respectively.

The network is trained for 100 epochs on 4-s-long segments with an initial learning rate of $2e^{-4}$. If no better results are obtained on the validation set for 3 consecutive epochs, the learning rate will be halved. If the best model is not updated for 10 consecutive epochs, training will be stopped early. Adam [46] is used as the optimizer, and a gradient clipping with a maximum L_2-norm of 5 is applied during training. To ensure fairness, all models are trained using PyTorch profiler [47] on 2 NVIDIA GeForce RTX 4090 GPU devices.

4.3 Evaluation metrics

In the experiments, four objective evaluation metrics and one subjective evaluation metric were used to evaluate the separation performance of our proposed method and the baseline methods.

The objective evaluation metrics include SISNRi [33], signal distortion ratio improvement (SDRi) [48, 49], perceptual evaluation of subjective quality (PESQ) [50], and STOI [28]. These objective metrics are obtained by comparing the model output speech with the original clean speech. The SISNRi and SDRi are energy ratios which can be used to measure the similarity between signals. PESQ scores ranged from -0.5 to 4.5 and STOI scores ranged from 0 to 1. The higher the value, the better the quality of the separated signal.

The MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA) [51, 52] is chosen as the subjective evaluation. The MUSHRA is conducted by asking a number of experienced listeners to rate the quality of the separated mixtures. The value of MUSHRA ranges from 0 to 100, and the higher the value, the better the quality of the separated signal.

5 Experimental results and analysis

The proposed method in this paper was compared with three baseline methods (Conv-TasNet, DPRNN-TasNet, and DPTNet). During the experiments, the three baseline methods have successfully recurred with the same specific structure and hyperparameter settings as in [18–20].

The proposed method and its ablation experiments (The proposed method without the deep encoder/decoder and the loss function without $\mathcal{L}_{\text{STOI}}$) were compared with the baseline method. In addition, we conducted comparative experiments by replacing the proposed method's SOSISNR with the original SISNR.

The sizes of the aforementioned models, their computational complexities for 4-s segments, and objective evaluation metrics are shown in Table 1.

The results show that compared to Conv-TasNet, DPRNN-TasNet, and DPTNet, the proposed method brings a 5.0 dB, 2.1 dB, and 0.2 dB increase in SISNRi as well as a 4.9 dB, 1.9 dB, and 0.1 dB increase in SDRi, respectively. It is demonstrated that the proposed method can perform better speech separation in complex sound field environments. Meanwhile, the STOI

Table 1 Performance comparison

Separator network	Model size (M)	MACs (G)	SISNRi (dB)	SDRi (dB)	STOI
Conv-TasNet	5.1	20.8	7.4	7.9	0.71
DPRNN-TasNet	2.6	85.0	10.3	10.9	0.81
DPTNet	2.6	208.4	12.2	12.7	0.84
Proposed method	**3.8**	**189.3**	**12.4**	**12.8**	**0.87**
Pro w/o deep encoder/ decoder	2.6	-	11.3	11.8	0.85
Pro w/o $\mathcal{L}_{\text{STOI}}$	3.8	-	12.3	12.7	0.84
Pro w/o SOSISNR w SISNR	3.8	-	11.7	12.0	-

shows that the proposed method brings 22.5%, 7.4%, and 3.6% auditory impression improvement compared to the three baseline methods. It proves that the speech separated by the proposed method is more compatible with the human auditory system.

The effect of different configurations on the separation performance of the proposed method was analyzed by means of ablation experiments (e.g., Table 1, rows 6 and 7). Firstly, the introduction of the deep encoder/decoder brings an increase in SISNRi of 1.1 dB and an increase in SDRi of 1.0 dB to the model. This illustrates the greater potential of the deep encoder/decoder for the better transformation of complex signals. This result shows the possibility of combining a deep encoder/decoder with more advanced separation modules to achieve better separation performance. Secondly, the model is optimized using a joint loss function by introducing a loss function related to human auditory, $\mathcal{L}_{\text{STOI}}$. This effectively improves speech intelligibility and better matches the target of the training model to the human auditory system.

Further, the comparative experiments (e.g., Table 1, row 8) indicate that replacing the original SISNR with SOSISNR brings a 0.7 dB increase in SISNRi and 0.8 dB in SDRi for the model. This demonstrates that SOSISNR contributes to enhancing the performance of the speech separation system in noisy and reverberant environments compared to SISNR.

The proposed method, along with three baseline methods, underwent further testing in a real-world environment. Within a room measuring 4.5 m×3.5 m×2.8 m, a subset of data (20 recordings) was randomly selected from the WSJ0-2mix test set and captured using microphones (specifically, the measurement condenser microphone ECM8000) paired with sound cards (Depusheng md22). The room was characterized by an estimated T_{60} reverberation time of 400 ms and a SNR of 8.3 dB.

The background noise consisted of external noise and vibrations within the recording room, stemming from incomplete sound insulation material isolation. The results of the real-world testing were averaged, and the objective evaluation metric SISNRi is shown in Fig. 6:

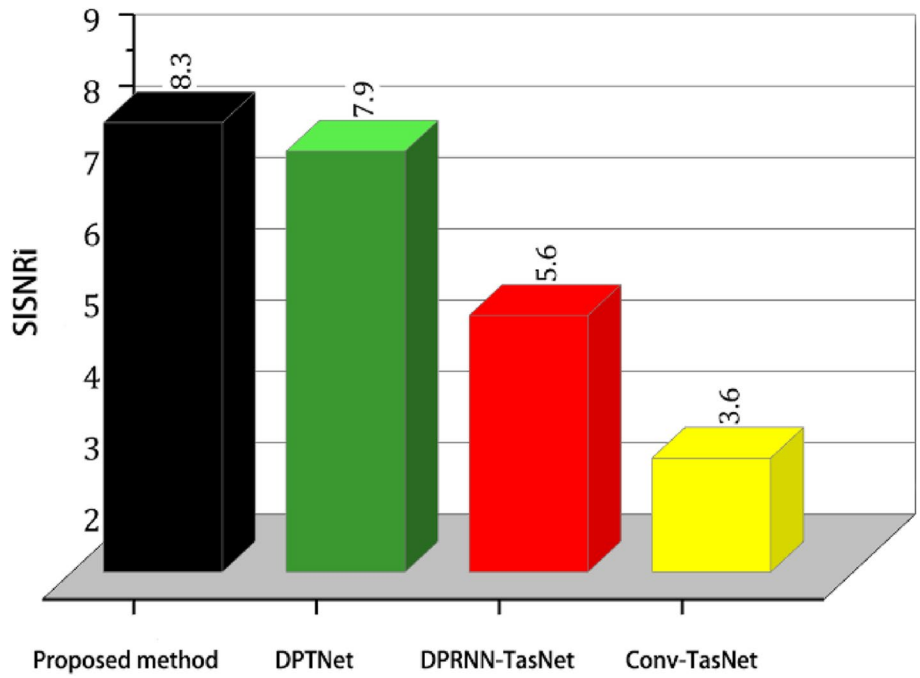

Fig. 6 Results of SISNRi in a real-world environment

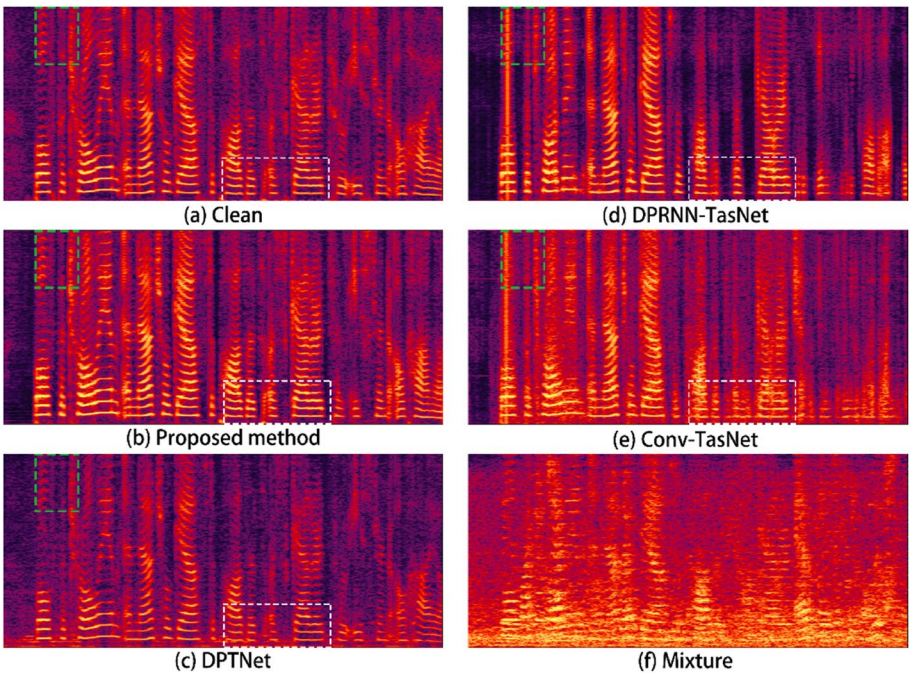

Fig. 7 Spectrogram visualization

The experimental results indicate that in a real-world environment, both the proposed method and the baseline methods inevitably exhibited some decrease in separation performance. However, the proposed method outperformed the three baseline methods in terms of performance within the real-world setting.

To visualize the performance of the proposed method against the three baseline methods, the separation performance of a randomly sampled segment of a mixture of speech is presented via a speech spectrogram, as shown in Fig. 7. This mixture (Fig. 7f) is disturbed by the background noise of the café with an SNR of -5 dB and

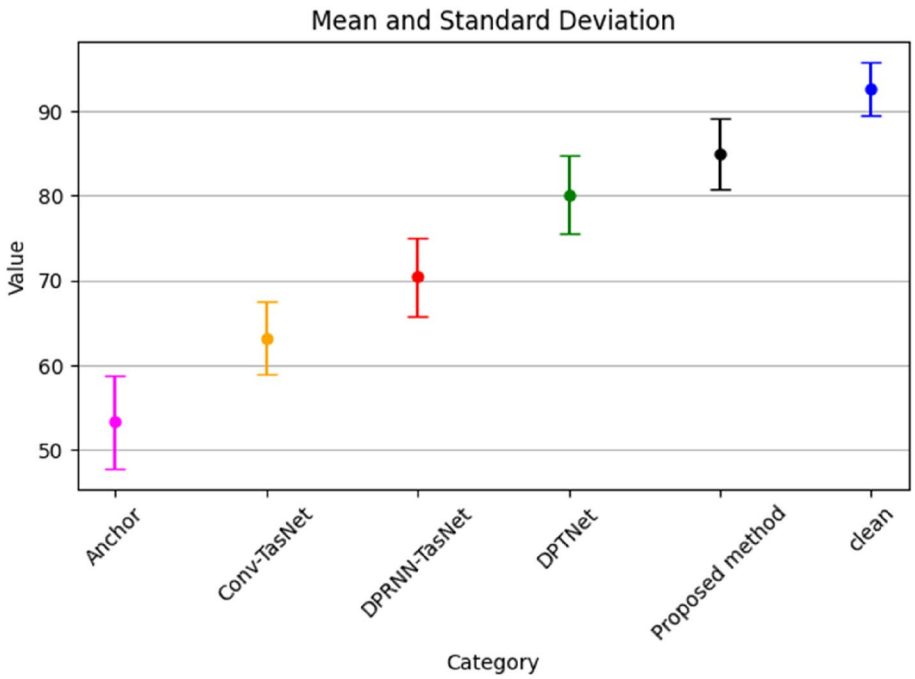

Fig. 8 Results of the MUSHRA listening test with 95% confidence intervals

room reverberation with a T_{60} of 300 ms. Figure 7a is the spectrogram of the original clean signal. Figure 7b, c, d, and e demonstrate the spectrograms of a separated source processed by four methods, respectively.

From Fig. 7, it can be found that all four methods can separate the sound sources. However, the separation performance of the proposed method, DPTNet, and DPRNN-TasNet is significantly better than that of Conv-TasNet. This is due to the fact that Conv-TasNet uses a fixed context length, which results in its lack of long-term tracking of the speaker and generalization to complex sound field environments. Furthermore, as shown by the highlighted white and green dashed boxes (Fig. 7a–d), the proposed method provides a better restoration of the harmonic components. This indicates that our method achieves better separation performance.

To further evaluate the subjective evaluation metric of the proposed method, the MUSHRA listening test was conducted with the participation of 20 experienced listeners. The proposed method and three baseline methods were used to process 18 randomly selected mixture signals from the test set. According to the MUSHRA specifications, each experiment included a hidden reference and a 3.5 kHz low-pass filter anchor. The results of the MUSHRA listening test with 95% confidence intervals are shown in Fig. 8:

From Fig. 8, it can be seen that the MUSHRA scores of the proposed method are higher than those of the baseline methods, which means that the proposed method provides a better auditory experience for listeners.

These methods were further evaluated at different noise reverberation levels. This was done by generating test sets for nine levels of different acoustic environments using the same methodology as in subsection IV.A. The proposed method and the two baseline methods were tested using STOI and PESQ as objective evaluation metrics. The average results of the nine test sets are shown in Figs. 9 and 10, where "re" represents the T_{60} with units of "ms" and "n" represents the SNR with units of "dB."

As shown in Figs. 9 and 10, the STOI and PESQ of the proposed method are higher than that of the three

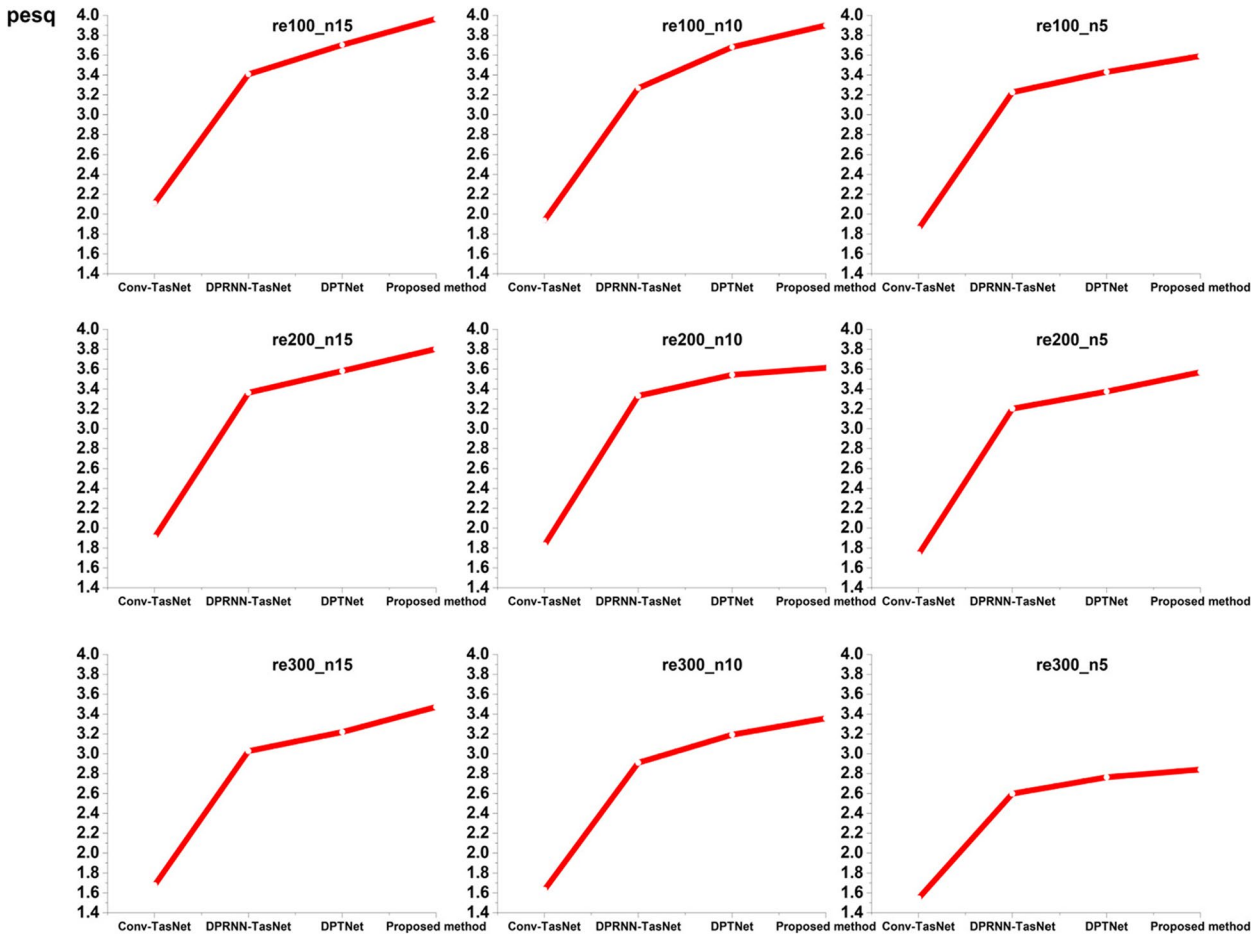

Fig. 9 Results of PESQ

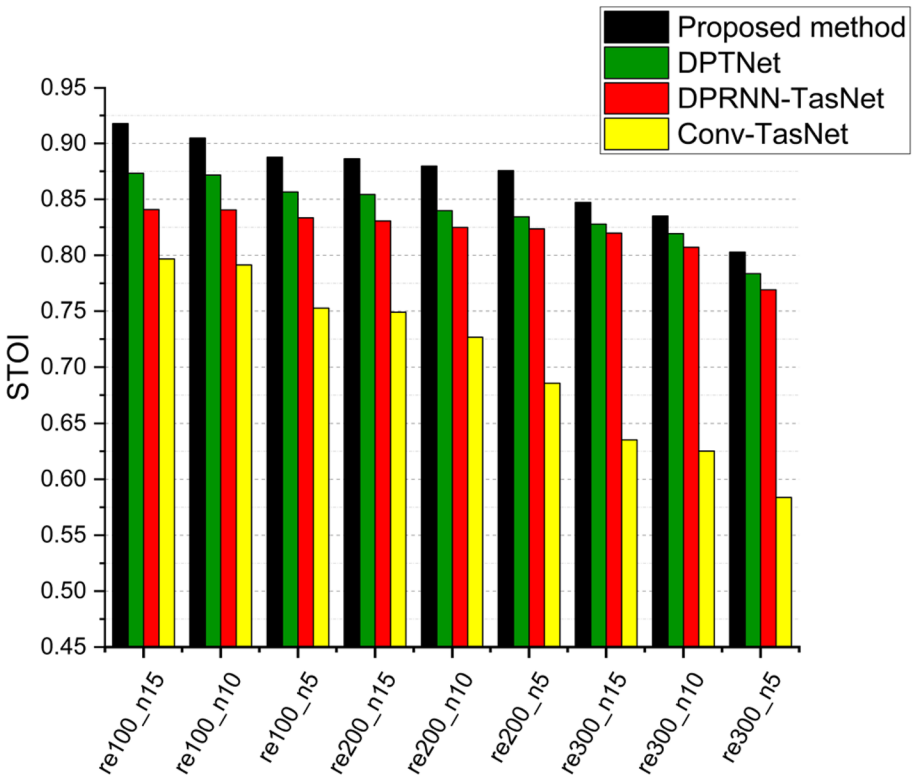

Fig. 10 Results of STOI

baseline methods in complex environments. The experimental results show an inevitable degradation of the system performance with increasing T_{60} and SNR. However, even in more complex sound fields, the proposed method still achieves a robust improvement over the two baseline methods.

6 Conclusion

This paper proposes a deep encoder/decoder dual-path neural network that can better model complex signals. The network can separate the clean speech of each speaker from a mixture with noise and reverberation. In addition, a new loss function, SOSISNR, is proposed to further improve the performance of the model. The joint loss function is extended with the STOI-based loss function to make the model more compatible with the human auditory system.

The alignment operation is proposed to reduce the sensitivity of the model to the utterance starting points and to increase the robustness of the model. Combined with the above operations, the subjective and objective evaluation metrics show that this study has better separation performance in complex sound field environments and shows superiority in various scenarios. At the same time, the model maintains a relatively small model size, which is not demanding on the recording equipment and has wide applicability.

In the future, the model can be improved by incorporating more advanced separation modules. The generalization performance of the proposed method on other unseen datasets needs to be further tested. Simultaneously, in complex scenarios, the model can be further extended to handle situations involving three or more speakers.

Abbreviations

DNNs	Deep neural networks
SOSISNR	Stretched optimal scale-invariant signal-to-noise ratio
SISNR	Scale-invariant signal-to-noise ratio
STOI	Short-time objective intelligibility
CASA	Computational Auditory Scene Analysis
ICA	Independent Component Analysis
NMF	Non-negative Matrix Factorization
PIT	Permutation invariant training
DPCL	Deep clustering
DANet	Deep attractor network
LSTM-TasNet	Long short-term memory time-domain audio separation network
Conv-TasNet	Fully-convolutional time-domain audio separation network
TCN	Temporal convolutional network
DPRNN-TasNet	Dual-path recurrent neural network time-domain separation network
DPTNet	Dual-path transformer network
SISNRi	Scale-invariant signal-to-noise ratio improvement

RIRs	Room impulse responses
FC	Fully-connected
LN	Layer normalization
uPIT	Utterance-level permutation invariant training
SNR	Signal-to-noise ratio
STFT	Short-time Fourier transform
WSJ0	Wall Street Journal dataset
T_{60}	Reverberation time
WHAM!	WSJ0 Hipster Ambient Mixtures!
SDRi	Signal distortion ratio improvement
PESQ	Perceptual evaluation of subjective quality
MUSHRA	MUlti Stimulus test with Hidden Reference and Anchor

Acknowledgements
This work was supported by the National Natural Science Foundation of China under Grants (No. 61971015) and the Beijing Natural Science Foundation (No. L223033).

Authors' contributions
Wang C. performed the whole research and wrote the paper. Jia M. provided support to the writing and experiments. All authors read and approved the final version of the paper.

Funding
This work was supported by the National Natural Science Foundation of China under Grants (No. 61971015) and the Beijing Natural Science Foundation (No. L223033).

Availability of data and materials
Not applicable.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 11 May 2023 Accepted: 29 September 2023
Published online: 12 October 2023

References

1. A.W. Bronkhorst, The cocktail party phenomenon: a review of research on speech intelligibility in multiple-talker conditions. Acta Acust. Acust. **86**(1), 117–128 (2000)
2. S. Haykin, Z. Chen, The cocktail party problem. Neural Comput. **17**(9), 1875–1902 (2005)
3. P. Comom, C. Jutten, *Handbook of blind source separation: independent component* (Academic Press, Elsevier, Burlington, Analysis and Applications, 2010)
4. S.F. Boll, Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on Acoustics Speech and Signal Processing **27**(2), 113–120 (1979)
5. K. Yoshii, R. Tomioka, D. Mochihashi, M. Goto, *Beyond nmf: time-domain audio source separation without phase reconstruction," ISMIR* (2013), pp.369–374
6. Y. Jia, Q. Yang, M. Jia, W. Xu, C. Bao, Multiple sound source separation via ideal ratio masking by using probability mixture model. J. Signal Process. **37**(10), 1806–1815 (2021)
7. X. Chen, W. Wang, Y. Wang, X. Zhong, A. Alinaghi, Reverberant speech separation with probabilistic time frequency masking for b-format recordings. Speech Communications. **68**, 41–54 (2015)
8. M. Jia, J. Sun, C. Bao et al., Separation of multiple speech sources by recovering sparse and non-sparse components from B-format microphone recordings. Speech Commun. **96**, 184–196 (2018)
9. P.-S. Huang, M. Kim, M. Hasegawa-Johnson, P. Smaragdis, Joint optimization of masks and deep recurrent neural networks for monaural source separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, Dec. **23**(12), 2136–2147 (2015)
10. D.S. Williamson, Y. Wang, D. Wang, Complex ratio masking for monaural speech separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing **24**(3), 483–492 (2016)
11. D. Yu, M. Kolbæk, Z.H. Tan, J. Jensen, Permutation invariant training of deep models for speaker-independent multi-talker speech separation," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (IEEE, New Orleans, LA, USA, 2017), pp. 241-245
12. J. R. Hershey, Z. Chen, J. Le Roux and S. Watanabe, "Deep clustering: discriminative embeddings for segmentation and separation," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (IEEE, Shanghai, China, 2016), pp. 31–35
13. Z. Chen, Y. Luo and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (IEEE, New Orleans, LA, USA, 2017), pp. 246–250
14. M. Kolbæk, D. Yu, Z.-H. Tan, J. Jensen, Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing, Oct. **25**(10), 1901–1913 (2017)
15. Y. Luo, Z. Chen, N. Mesgarani, Speaker-independent speech separation with deep attractor network. IEEE/ACM Transactions on Audio, Speech, and Language Processing **26**(4), 787–796 (2018)
16. Z. -Q. Wang, J. L. Roux and J. R. Hershey, "Alternative objective functions for deep clustering," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (IEEE, Calgary, AB, Canada, 2018), pp. 686–690
17. Y. Luo and N. Mesgarani, "TaSNet: time-domain audio separation network for real-time, single-channel speech separation," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (IEEE, Calgary, AB, Canada, 2018), pp. 696–700
18. Y. Luo, N. Mesgarani, Conv-TasNet: surpassing ideal time–frequency magnitude masking for speech separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, Aug. **27**(8), 1256–1266 (2019)
19. Y. Luo, Z. Chen and T. Yoshioka, "Dual-Path RNN: efficient long sequence modeling for time-domain single-channel speech separation," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (IEEE, Barcelona, Spain, 2020), pp. 46–50
20. J. Chen, Q. Mao, and D. Liu, "Dual-path transformer network: direct context-aware modeling for end-to-end monaural speech separation," *Interspeech* 2020 (ICSA, Shanghai, China, 2020), pp. 2642–2646
21. N. Zeghidour, D. Grangier, Wavesplit: end-to-end speech separation by speaker clustering. IEEE/ACM Transactions on Audio, Speech, and Language Processing **29**, 2840–2849 (2021)
22. C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi and J. Zhong, "Attention is all you need in speech separation," *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (IEEE, Toronto, ON, Canada, 2021), pp. 21–25
23. T. Cord-Landwehr, C. Boeddeker, T. Von Neumann, C. Zorilă, R. Doddipatla and R. Haeb-Umbach, "Monaural source separation: from anechoic to reverberant environments," *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*, Bamberg, Germany, 2022, pp. 1-5
24. H. Taherian, K. Tan, D. Wang, Multi-channel talker-independent speaker separation through location-based training. IEEE/ACM Transactions on Audio, Speech, and Language Processing **30**, 2791–2800 (2022)
25. K. Tan, Y. Xu, S.-X. Zhang, M. Yu, D. Yu, Audio-visual speech separation and dereverberation with a two-stage multimodal network. IEEE Journal of Selected Topics in Signal Processing **14**(3), 542–553 (2020)
26. D. Michelsanti et al., An overview of deep-learning-based audio-visual speech enhancement and separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing **29**, 1368–1396 (2021)
27. Z. -Q. Wang, S. Cornell, S. Choi, Y. Lee, B. -Y. Kim and S. Watanabe, "TF-GridNet: making time-frequency domain models great again for monaural speaker separation." *2023 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), (IEEE, Rhodes Island, Greece, 2023)
28. C.H. Taal, R.C. Hendriks, R. Heusdens, J. Jensen, An algorithm for intelligibility prediction of time–frequency weighted noisy speech. IEEE Trans. Audio Speech Lang. Process. **19**(7), 2125–2136 (2011)
29. C. Xu, W. Rao, E.S. Chng, H. Li, SpEx: multi-scale time domain speaker extraction network. IEEE/ACM Transactions on Audio, Speech, and Language Processing **28**, 1370–1384 (2020)

30. M. W. Y. Lam, J. Wang, D. Su and D. Yu, "Effective low-cost time-domain audio separation using globally attentive locally recurrent networks," *2021 IEEE Spoken Language Technology Workshop (SLT)*, (IEEE, Shenzhen, China, 2021), pp. 801-808
31. B. Kadıoğlu, M. Horgan, X. Liu, J. Pons, D. Darcy and V. Kumar, "An empirical study of Conv-Tasnet," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (IEEE, Barcelona, Spain, 2020), pp. 7264–7268
32. Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "Dccrn: deep complex convolution recurrent net-work for phase-aware speech enhancement," *Interspeech* 2020 (ICSA, Shanghai, China, 2020), pp. 2472–2476
33. S. Lv, Y. Fu, M. Xing, J. Sun, L. Xie, J. Huang, Y. Wang, and T. Yu, "S-DCCRN: super wide band DCCRN with learnable complex feature for speech enhancement," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (IEEE, Singapore, Singapore, 2022), pp. 7767–7771
34. Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey, "SDR half-baked or well done?," in Acoustics, Speech and Signal Processing (ICASSP), *2019 IEEE International Conference on. IEEE*, 2019, pp. 626 630
35. C. Ma, D. Li and X. Jia, "Optimal scale-invariant signal-to-noise ratio and curriculum learning for monaural multi-speaker speech separation in noisy environment," *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, (IEEE, Auckland, New Zealand, 2020), pp. 711-715
36. Y. Sun, L. Yang, H. Zhu, and J. Hao, "Funnel deep complex U-Net for phase-aware speech enhancement," *Interspeech* 2021 (ICSA, Brno, Czech Republic, 2021), pp. 161–165
37. A. Li, W. Liu, X. Luo, C. Zheng and X. Li, "ICASSP 2021 deep noise suppression challenge: decoupling magnitude and phase optimization with a two-stage deep network," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (IEEE, Toronto, ON, Canada, 2021), pp. 6628–6632
38. H. Zhang, X. Zhang and G. Gao, "Training supervised speech separation system to improve STOI and PESQ directly," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (IEEE, Calgary, AB, Canada, 2018), pp. 5374–5378
39. S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, H. Kawai, End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing, Sept. **26**(9), 1570–1584 (2018)
40. Y. Zhu, X. Xu, Z. Ye, FLGCNN: a novel fully convolutional neural network for end-to-end monaural speech enhancement with utterance-based objective functions. Appl. Acoust. **170**, 107511 (2020)
41. J.B. Allen, D.A. Berkley, Image method for efficiently simulating small-room acoustics. J. Acoustical Soc. Amer. **65**, 943–950 (1979)
42. R. Cheng, C. Bao, Z. Cui, MASS: microphone array speech simulator in room acoustic environment for multi-channel speech coding and enhancement. Appl. Sci. **10**(4), 1484 (2020)
43. R. Scheibler, E. Bezzam and I. Dokmanić, "Pyroomacoustics: a python package for audio room simulation and array processing algorithms," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (IEEE, Calgary, AB, Canada, 2018), pp. 351–355
44. G. Wichern et al.,"WHAM!: extending speech separation to noisy environments," *Interspeech* 2019 (ICSA, Graz, Austria, 2019), pp. 1368–1372
45. S. Honchreiter, J. Schmidhuber, Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
46. D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," *2014,arXiv preprint arXiv*: 1412.6980
47. Pytorch ,"Profiler,"https://pytorch.org/tutorials/recipes/recipes/profiler.html, 2020, Accessed: 2020–10–21
48. Y. Isik, J. L. Roux, Z. Chen, S. Watanabe and J. R. Hershey , "Single-channel multi-speaker separation using deep clustering," *Interspeech 2016 (ICSA, San Francisco, USA, 2016)*, pp.545-549
49. E. Vincent, R. Gribonval, C. Fevotte, Performance measure- ment in blind audio source separation. IEEE transactions on audio, speech, and language processing **14**(4), 1462–1469 (2006)
50. Perceptual evaluation of speech quality (PESQ), *An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. Rec. ITU-T P. 862, 2001* (International Telecommunications Union, Geneva, Switzerland, 2001)
51. BS.1534, Int. Telecomm. Union, *Method for the subjective assessment of intermediate quality levels of coding systems* (1997)
52. B. Series, Method for the subjective assessment of intermediate quality level of audio systems. International Telecommunication Union Radiocommunication Assembly. (2014)

Publisher's Note