



Joint constraint algorithm based on deep neural network with dual outputs for single-channel speech separation

Linhui Sun¹ · Ge Zhu¹ · Pingan Li¹

Received: 9 October 2019 / Revised: 5 February 2020 / Accepted: 18 March 2020 / Published online: 12 April 2020
© Springer-Verlag London Ltd., part of Springer Nature 2020

Abstract

Single-channel speech separation (SCSS) plays an important role in speech processing. It is an underdetermined problem since several signals need to be recovered from one channel, which is more difficult to solve. To achieve SCSS more effectively, we propose a new cost function. What's more, a joint constraint algorithm based on this function is used to separate mixed speech signals, which aims to separate two sources at the same time accurately. The joint constraint algorithm not only penalizes residual sum of square, but also exploits the joint relationship between the outputs to train the dual output DNN. In these joint constraints, the training accuracy of the separation model can be further increased. We evaluate the proposed algorithm performance on the GRID corpus. The experimental results show that the new algorithm can obtain better speech intelligibility compared to the basic cost function. In the aspects of source-to-distortion ratio, signal-to-interference ratio, source-to-artifact ratio and perceptual evaluation of speech quality, the novel approach can obtain better performance.

Keywords Deep neural network (DNN) · Single-channel speech separation · Joint constraint · Cost function · Dual outputs

1 Introduction

Single-channel speech separation (SCSS) is the process of separating multiple sources from one channel, which has a wide range of applications in automatic speech recognition (ASR), hearing aids and speaker recognition [1–4]. Because of the excellent ability to model the nonlinear relationship between input features and output targets, deep neural network (DNN) has been widely used in the field of speech separations [5–14]. According to the number of DNN outputs, the DNN-based methods can be divided into two categories: single-output DNN and multi-output DNN. DNN with single output can be used to map the relationship between the mixed signal and the single target source. For example, Han et al. used DNN with single output to directly learn the nonlinear

relationship of magnitude spectrogram between the reverberation and the clean signal, which achieved performance improvement in terms of denoising and de-reverberation [6]. Sun et al. proposed a two-stage method to address monaural source separation problem with the help of single output DNN [7]. This type of DNN maps a specific signal and has a significant separation performance. However, the single output DNN can only separate one voice source at a time, which is time consuming. For the multi-output DNN, the different sources can be separated simultaneously, which costs less time. In [8], Du et al. proposed the dual-output DNN architecture to separate the target and interfering speech at a time, which demonstrated that the dual-output DNN outperforms Gaussian mixture model (GMM) method in semi-supervised mode [1]. Huang et al. used the recurrent neural network (RNN) with dual outputs to separate the mixed signal, where the ideal ratio mask was embedded as an additional processing layer to the original output layer of the network [9]. The method separated different sources at one time and obtained performance improvements compared with the nonnegative matrix factorization (NMF) method. In addition, Wang et al. used multi-output DNN to train the detector for testing gender, and then the signal sources were separated according to different gender combinations [10]. Among different SCSS methods, the cost function has a great influence on the perfor-

✉ Linhui Sun
sunlinhuislh@126.com
Ge Zhu
sunlinhuislh@163.com
Pingan Li
lpa@njupt.edu.cn

¹ College of Telecommunications & Information Engineering,
Nanjing University of Posts and Telecommunications,
Nanjing 210003, Jiangsu, China

mance of speech. In [13], a new cost function was proposed to optimize the extended short time objective intelligibility (ESTOI) measure, which obtained excellent performance. What’s more, Joho et al. used a cost function to solve the problem of blind speech separation (BSS) when the individual signal is non-stationary [15]. In our previous study, we optimized the cost function to obtain the dictionary and got good separation performance [16]. However, the cost functions mentioned above did not take into account the joint relationship between the sources to be separated. To tackle this barrier and separate sources effectively, we propose a new cost function considering the joint relationship between the separated sources. On this basis, a good quality speech separation algorithm with dual output is proposed, which can simultaneously obtain two high-quality separated speech sources. The rest of the paper is organized as follows. Section 2 reviews the mixed modes of signal. Section 3 describes the new cost function for the dual-output DNN model and its solution. The experimental settings and results are displayed in Sect. 4. Finally, we conclude the paper and describe future work in Sect. 5.

2 Mixed mode of signal

In practice, mixed speech is generated by mixing the original signals in one or more ways, such as linear instantaneous mixing, linear convolution mixing and nonlinear mixing. Since the analysis of linear instantaneous mixing is more convenient to description, we regard the mixed signal as the form of linear mixing of multiple signals:

$$y(t) = \sum_{i=1}^m a_i x_i(t), \tag{1}$$

where a_i is a constant coefficient, representing the weight of the i -th target speech signal, and $x_i(t)$, $y(t)$ are the i th target speech signal and the mixture of m ($m \geq 2$) sources, respectively. For simplicity, we consider the situation of mixed signal containing two sources, i.e. $y(t) = a_1x_1(t) + a_2x_2(t)$, although the proposed algorithm can be generalized to more than two sources. Then, the SCSS problem can be formulated to estimate $x_i(t)$ from $y(t)$ obtained by single microphone.

3 Proposed SCSS approach based on joint constraint algorithm

In this section, we first introduce the model architecture of the dual-output DNN for speech separation, and then propose a joint constraint algorithm based on a novel cost function to learn the dual-output DNN model for SCSS.

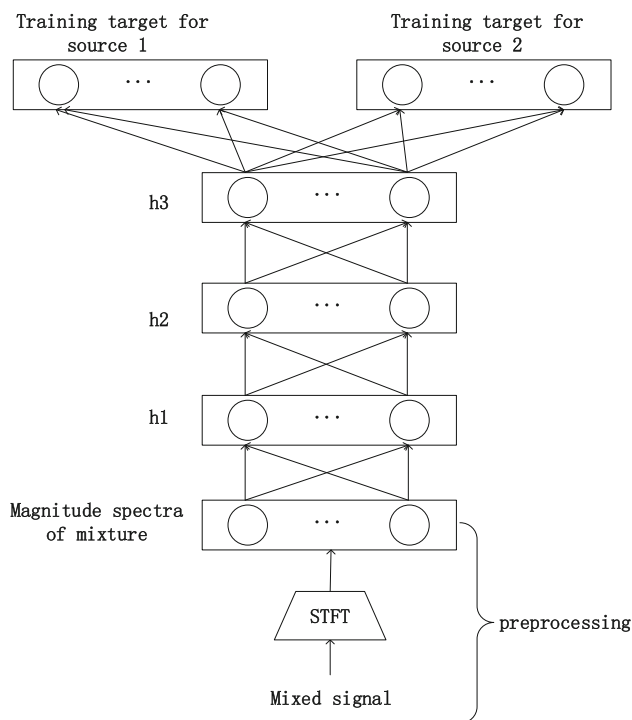


Fig. 1 Dual output DNN for speech separation

3.1 Model architecture

As shown in Fig. 1, our aim is to separate all single speech signals from the mixed speech at the same time, rather than to learn a model that treats a specific source as a target and the other as interference [12].

In the training stage, we use the magnitude spectra of mixed training signal as the features to train the model. By short-time Fourier transform (STFT) of overlapping windowed frame, the magnitude spectra of $y(t)$ and $x_i(t)$ ($i = 1, 2$) can be extracted, denote as $Y(t, f)$ and $X_i(t, f)$ ($i = 1, 2$) at time t and frequency f , respectively. The corresponding training target we used is the ideal ratio mask (IRM), which can describe the distribution of the sources accurately. The ideal ratio mask can be formulated as follows:

$$M_i(t, f) = \frac{X_i(t, f)}{a_1X_1(t, f) + a_2X_2(t, f) + \varepsilon}, \quad (i = 1, 2), \tag{2}$$

where $X_i(t, f)$ and $M_i(t, f)$ are the spectrogram and IRM of i -th source, respectively. ε is a minimal positive number to prevent denominator from becoming zero. In the separation stage, multiplying the estimated ratio mask (RM) obtained from well-trained DNN model with the amplitude spectrum of the mixed test signal, we can obtain the estimations of the individual acoustic signals [14], which can be described by the following formula:

$$\begin{aligned} \widehat{X}_{1t} &= \widehat{M}_{1t} \odot Y_t \\ \widehat{X}_{2t} &= \widehat{M}_{2t} \odot Y_t \end{aligned} \tag{3}$$

where $\widehat{X}_{1t}, \widehat{X}_{2t}, Y_t$ are the estimated spectra vectors of two separated sources and the mixed signal vector at t -th frame, respectively. \widehat{M}_{it} is the estimated value of i -th IRM, and \odot means Hadamard product.

It can be clearly seen from Eq. (3) that the accuracy of the estimated RM has a direct impact on separation performance. Therefore, we propose a novel cost function which exploits the joint information of the dual output to obtain more accurate estimated RM.

3.2 New cost function

For the single-output DNN, the cost function mainly focuses on the mapping relationship between the input mixed signals and target source [5]:

$$J_{1t} = \frac{1}{2T} \sum_{t=1}^T \left\| \mathbf{M}_t - \widehat{\mathbf{M}}_t \right\|_2^2, \tag{4}$$

where T indicates the number of time frames. \mathbf{M}_t and $\widehat{\mathbf{M}}_t$ are the IRM vector of target source and the corresponding estimation vector at t -th frame, respectively. Here, $\widehat{\mathbf{M}}_t = f(Y_t)$ is the output of DNN. Ref. [10] trained the dual-output DNN parameter with the basic cost function, which is used to learn the relationship between different sources and corresponding estimates:

$$J_2 = \frac{1}{2T} \sum_{t=1}^T (\|\mathbf{M}_{1t} - \widehat{\mathbf{M}}_{1t}\|_2^2 + \|\mathbf{M}_{2t} - \widehat{\mathbf{M}}_{2t}\|_2^2), \tag{5}$$

Obviously, Eq. (4) can only obtain the ratio mask of one speech signal at a time, while Eq. (5) can get two speech signal ratio masks. However, Eq. (5) ignores the joint relationship between the separated sources. Therefore, for the dual-output DNN, we propose a joint constraint algorithm by using a new cost function. It not only utilizes the nonlinear relationship between the input mixed signals and individual sources, but also considers the joint relationship between the separated sources:

$$\begin{aligned} J_3 &= \frac{1}{2T} \sum_{t=1}^T \left(\left\| \mathbf{M}_{1t} - \widehat{\mathbf{M}}_{1t} \right\|_2^2 + \left\| \mathbf{M}_{2t} - \widehat{\mathbf{M}}_{2t} \right\|_2^2 \right. \\ &\quad \left. + \lambda \left\| (\mathbf{M}_{1t} + \mathbf{M}_{2t}) - (\widehat{\mathbf{M}}_{1t} + \widehat{\mathbf{M}}_{2t}) \right\|_2^2 \right), \end{aligned} \tag{6}$$

where $0 \leq \lambda \leq 1$ is a regularization parameter. According to Eq. (2), we can obtain

$$\begin{aligned} \mathbf{M}_{1t} &= \frac{X_{1t}}{a_1 X_{1t} + a_2 X_{2t} + \epsilon}, \\ \mathbf{M}_{2t} &= \frac{X_{2t}}{a_1 X_{1t} + a_2 X_{2t} + \epsilon}. \end{aligned} \tag{7}$$

For clarity, we consider that the mixed signal is added by the source signal in equal proportion, i.e. $a_1 = a_2 = 1$, although the proposed algorithm can be easily extended to cases where the weight of target source is different. Hence, $\mathbf{M}_{1t} + \mathbf{M}_{2t} = 1$ at t -th frame since the element in ϵ is the same minimal positive number. $\widehat{\mathbf{M}}_{1t}$ and $\widehat{\mathbf{M}}_{2t}$ are the estimations of \mathbf{M}_{1t} and \mathbf{M}_{2t} . There is a joint relationship between $\widehat{\mathbf{M}}_{1t}$ and $\widehat{\mathbf{M}}_{2t}$: each element in the sum of $\widehat{\mathbf{M}}_{1t}$ and $\widehat{\mathbf{M}}_{2t}$ approaches 1 when the estimates are completely accurate. Compared with the basic cost function, we add a constraint term to exploit the relationship between the dual outputs of DNN. The first two items of Eq. (6) penalize the predicted error of the IRM over the corresponding estimations, and the third item takes advantage of the joint relationship between $\widehat{\mathbf{M}}_1(t, f)$ and $\widehat{\mathbf{M}}_2(t, f)$ to train the DNN model. Compared with MaxDiffer [9], we use the joint relation term instead of the maximum difference term, and compare the separation performance between them in the experiment. The specific solution of proposed algorithm is shown in next subsection.

3.3 Learning DNN model by solving the cost function

There are two essential steps in learning DNN model: forward propagation and backpropagation (BP) [17]. Forward propagation is the process of obtaining the results of the output layer by calculating the weights and biases layer by layer. BP algorithm is a process of reversely adjusting weights and biases under the constraints of cost function. We use BP with the gradient descent (GD) for training DNN model. For simplicity of description, we analyze the signal of t -th frame as a representative. Then the final result can be obtained by accumulation. Hence, the function in Eq. (6) at t -th frame can be stated as:

$$\begin{aligned} J(\mathbf{W}, \mathbf{b}; \mathbf{x}, \mathbf{y}) &= \frac{1}{2} \left(\left(\left\| \mathbf{M}_{1:512} - \widehat{\mathbf{M}}_{1:512}^L \right\|_2^2 + \left\| \mathbf{M}_{513:1024} - \widehat{\mathbf{M}}_{513:1024}^L \right\|_2^2 \right. \right. \\ &\quad \left. \left. + \lambda \left\| \widehat{\mathbf{M}}_{1:512}^L + \widehat{\mathbf{M}}_{513:1024}^L - 1 \right\|_2^2 \right) \right), \end{aligned} \tag{8}$$

where \mathbf{W}, \mathbf{b} are the weight vector and bias vector, respectively. For the third term of Eq. (6), the $\widehat{\mathbf{M}}_{1t}, \widehat{\mathbf{M}}_{2t}$ is the first 512 and the last 512 output nodes of the actual DNN output in t -th frame, since the number of output nodes in our experiment are 1024. We use $\widehat{\mathbf{M}}_{m:n}^L$ to represent $\widehat{\mathbf{M}}_i$

$(f, g)(i = 1, 2)$, and the corresponding ideal ratio mask is expressed as $\widehat{M}_{m:n}^L$. The subscript of $\widehat{M}_{m:n}^L$ presents the m th neuron node to the n th neuron node of the output layer network, where L is the number of layers in DNN. In other words, $m:n$ is 1:512 and 513:1024 respectively, which means both $\mathbf{M}_{1:512}$ and $\mathbf{M}_{513:1024}$ are 512-dimensional vectors. \widehat{M}^l as the actual output of l th layer of DNN can be regarded as $\sigma(\mathbf{z}^l) = \sigma(\mathbf{W}^l \widehat{M}^{l-1} + \mathbf{b}^l)$. Here \mathbf{z}^l is the neuron output state matrix of l -th DNN layer, and the element z in the matrix can be obtained by the interaction of the previous layer of neurons x_i with the weight and bias, namely $z = \sum_{(i=1)}^{(s_l)} W_i x_i + b_i$, where s_l is the number of nodes in l -th layer. The activation function we used in DNN layers is the sigmoid function $\sigma(z) = 1/(1 + e^{-z})$. In the proposed cost function, the network takes the magnitude spectra of mixed signal as input, while the corresponding training target $[\mathbf{M}_{1:512}, \mathbf{M}_{513:1024}]$ is the concatenation of ideal ratio mask from different sources. Next, for the neuron node i of output layer L , we calculate the $\partial J(\mathbf{W}, \mathbf{b}; \mathbf{x}, \mathbf{y}) / \partial Z_i^l$ of output layer according to the following formula:

$$\begin{aligned} \delta_i^l &= \frac{\partial J(\mathbf{W}, \mathbf{b}; \mathbf{x}, \mathbf{y})}{\partial Z_i^l} \\ &= \frac{1}{2} \frac{\partial}{\partial Z_i^{l-1}} \left(\left\| \mathbf{M}_{1:512} - \widehat{M}_{1:512}^L \right\|_2^2 + \left\| \mathbf{M}_{513:1024} - \widehat{M}_{513:1024}^L \right\|_2^2 \right. \\ &\quad \left. + \lambda \left\| \widehat{M}_{1:512}^L + \widehat{M}_{513:1024}^L - 1 \right\|_2^2 \right) \\ &= \frac{\partial}{\partial Z_i^l} \frac{1}{2} \sum_{j=1}^{512} \left((M_j - \widehat{M}_j^l)^2 + \lambda (\widehat{M}_j^l + \widehat{M}_{j+512}^l - 1)^2 \right) \\ &\quad + \frac{1}{2} \frac{\partial}{\partial Z_i^l} \sum_{j=513}^{s_l} (M_j - \widehat{M}_j^l)^2 \\ &= \frac{1}{2} \sum_{j=1}^{512} \frac{\partial}{\partial Z_i^l} \left((M_j - \widehat{M}_j^l)^2 + \lambda (\widehat{M}_j^l + \widehat{M}_{j+512}^l - 1)^2 \right) \\ &\quad + \frac{1}{2} \sum_{j=513}^{s_l} \frac{\partial}{\partial Z_i^l} (M_j - \widehat{M}_j^l)^2 \\ &= \frac{1}{2} \sum_{j=1}^{512} \frac{\partial}{\partial Z_i^l} \left((M_j - \sigma(z_j^l))^2 + \lambda (\sigma(z_j^l) + \sigma(z_{j+512}^l) - 1)^2 \right) \\ &\quad + \frac{1}{2} \sum_{j=513}^{s_l} \frac{\partial}{\partial Z_i^l} (M_j - \sigma(z_j^l))^2 \\ &= \begin{cases} -(M_i - \sigma(z_i^l)) \cdot \sigma'(z_i^l) + \lambda (\sigma(z_i^l) + \sigma(z_{i+512}^l) - 1) \\ \quad \cdot \sigma'(z_i^l), 1 \leq i \leq 512 \\ -(M_i - \sigma(z_i^l)) \cdot \sigma'(z_i^l) + \lambda (\sigma(z_i^l) + \sigma(z_{i-512}^l) - 1) \\ \quad \cdot \sigma'(z_i^l), 513 \leq i \leq 1024 \end{cases} \end{aligned} \tag{9}$$

For the layer of $l = L - 1, L - 2, \dots, 2$, $\frac{\partial J(\mathbf{W}, \mathbf{b}; \mathbf{x}, \mathbf{y})}{\partial z_i^l}$ can be calculated by the formula as follows:

$$\begin{aligned} \delta_i^{l-1} &= \frac{\partial J(\mathbf{W}, \mathbf{b}; \mathbf{x}, \mathbf{y})}{\partial Z_i^{l-1}} \\ &= \frac{\partial}{\partial Z_i^{l-1}} \frac{1}{2} \left(\left\| \mathbf{M}_{1:512} - \widehat{M}_{1:512}^L \right\|_2^2 + \left\| \mathbf{M}_{513:1024} - \widehat{M}_{513:1024}^L \right\|_2^2 \right. \\ &\quad \left. + \lambda \left\| \widehat{M}_{1:512}^L + \widehat{M}_{513:1024}^L - 1 \right\|_2^2 \right) \\ &= \frac{1}{2} \frac{\partial}{\partial Z_i^{l-1}} \left(\sum_{j=1}^{512} (M_j - \widehat{M}_j^l)^2 + \sum_{j=513}^{s_l} (M_j - \widehat{M}_j^l)^2 \right. \\ &\quad \left. + \sum_{j=1}^{512} \lambda (\widehat{M}_j^l + \widehat{M}_{j+512}^l - 1)^2 \right) \\ &= \frac{1}{2} \sum_{j=1}^{512} \frac{\partial}{\partial Z_i^{l-1}} \left((M_j - \widehat{M}_j^l)^2 + \lambda (\widehat{M}_j^l + \widehat{M}_{j+512}^l - 1)^2 \right) \\ &\quad + \frac{1}{2} \sum_{j=513}^{s_l} \frac{\partial}{\partial Z_i^{l-1}} (M_j - \widehat{M}_j^l)^2 \\ &= - \sum_{j=1}^{512} (M_j - \sigma(z_j^l)) \cdot \frac{\partial \sigma(z_j^l)}{\partial Z_i^{l-1}} + \lambda \sum_{j=1}^{512} (\sigma(z_j^l) + \sigma(z_{j+512}^l) - 1) \\ &\quad \cdot \frac{\partial}{\partial Z_i^{l-1}} (\sigma(z_j^l) + \sigma(z_{j+512}^l)) - \sum_{j=513}^{s_l} (M_j - \sigma(z_j^l)) \cdot \frac{\partial \sigma(z_j^l)}{\partial Z_i^{l-1}} \\ &= \begin{cases} \sum_{j=1}^{512} -(M_j - \sigma(z_j^l)) \cdot \sigma'(z_j^l) \cdot \frac{\partial z_j^l}{\partial Z_i^{l-1}} + \lambda \sum_{j=1}^{512} (\sigma(z_j^l) \\ \quad + \sigma(z_{j+512}^l) - 1) \cdot \sigma'(z_j^l) \cdot \frac{\partial z_j^l}{\partial Z_i^{l-1}}, 1 \leq j \leq 512 \\ \sum_{j=513}^{s_l} -(M_j - \sigma(z_j^l)) \cdot \sigma'(z_j^l) \cdot \frac{\partial z_j^l}{\partial Z_i^{l-1}} + \lambda \sum_{j=513}^{s_l} (\sigma(z_j^l) \\ \quad + \sigma(z_{j-512}^l) - 1) \cdot \sigma'(z_j^l) \cdot \frac{\partial z_j^l}{\partial Z_i^{l-1}}, 513 \leq j \leq 1024 \end{cases} \\ &= \begin{cases} \sum_{j=1}^{512} \frac{\partial z_j^l}{\partial Z_i^{l-1}} \cdot \sigma'(z_j^l) (- (M_j - \sigma(z_j^l)) + \lambda (\sigma(z_j^l) \\ \quad + \sigma(z_{j+512}^l) - 1)), 1 \leq j \leq 512 \\ \sum_{j=513}^{s_l} \frac{\partial z_j^l}{\partial Z_i^{l-1}} \cdot \sigma'(z_j^l) ((M_j - \sigma(z_j^l)) + \lambda (\sigma(z_j^l) \\ \quad + \sigma(z_{j-512}^l) - 1)), 513 \leq j \leq 1024 \end{cases} \\ &= \sum_{j=1}^{s_l} \frac{\partial z_j^l}{\partial Z_i^{l-1}} \cdot \delta_j^l \cdot \sigma'(z_j^l) \\ &= \sum_{j=1}^{s_l} \left(\delta_j^l \cdot \frac{\partial}{\partial Z_i^{l-1}} \sum_{k=1}^{l-1} \sigma(z_k^{l-1}) \cdot W_{jk}^{l-1} \right) \\ &= \sum_{j=1}^{s_l} (\delta_j^l \cdot W_{ji}^{l-1}) \cdot \sigma'(z_i^{l-1}) = \left(\sum_{j=1}^{s_l} (\delta_j^l \cdot W_{ji}^{l-1}) \right) \cdot \sigma'(z_i^{l-1}) \end{aligned} \tag{10}$$

By replacing $l - 1$ with l , we can obtain $\delta_i^l = \sum_{j=1}^{s_{l+1}} (W_{ji}^l \cdot \delta_j^{l+1}) \cdot \sigma'(z_i^l)$. And the derivative of the weight \mathbf{W} , bias \mathbf{B} of l -th layer can be obtained by the gradient descent method:

$$\begin{aligned} \frac{\partial J(\mathbf{W}, \mathbf{b}; \mathbf{x}, \mathbf{y})}{\partial W_{ji}^l} &= \frac{\partial J(\mathbf{W}, \mathbf{b}; \mathbf{x}, \mathbf{y})}{\partial Z_j^l} \cdot \frac{\partial Z_j^l}{\partial W_{ji}^l} = \delta_j^{l+1} \widehat{M}_i^l \\ \frac{\partial J(\mathbf{W}, \mathbf{b}; \mathbf{x}, \mathbf{y})}{\partial b_j^l} &= \frac{\partial J(\mathbf{W}, \mathbf{b}; \mathbf{x}, \mathbf{y})}{\partial Z_j^l} \cdot \frac{\partial Z_j^l}{\partial b_j^l} = \delta_j^{l+1} \end{aligned} \tag{11}$$

Table 1 Joint constraint algorithm for SCSS

Task: Separate mixed signal using joint constraint algorithm based on DNN with dual output.

Training Stage:

Input: Training speech signals of corresponding speaker $x_{1-train}$, $x_{2-train}$ and mixed speech signal y_{train} .

Step1: Preprocess the target and mixed speech signals with STFT, obtained the spectrum Y_{train} and the ideal ratio masks M_1 , M_2 .

Step2: Use Y_{train} and $[M_1, M_2]$ as the input and training target of DNN, respectively.

Step3: Train the DNN with the constraint algorithm based on the new cost function.

Step4: Find appropriate parameters.

Output: Well-trained DNN.

Separation Stage:

Input: mixed test speech signal y_{test} and the well-trained DNN.

Step1: Preprocess the test signals with STFT and obtain the spectrum Y_{test} .

Step2: Obtain the estimated ratio masks \hat{M}_1 , \hat{M}_2 with Y_{test} and well-trained DNN.

Step3: Calculate magnitude spectra by Eq. (3).

Step4: Synthesize separated speech signals.

Output: The separated speech signals \hat{x}_1, \hat{x}_2 .

At the end of the training, with the learning rate α , we can update the weights and bias by the following formula until the number of iterations is met:

$$\begin{aligned} W_{ji}^l &= W_{ji}^l - \alpha \frac{\partial J}{\partial W_{ji}^l} \\ b_j^l &= b_j^l - \alpha \frac{\partial J}{\partial b_j^l} \end{aligned} \quad (12)$$

In the separation stage, we can obtain the predicted ratio masks \hat{M}_i ($i = 1, 2$) of i -th sources by the well-trained DNN model, and then the estimated magnitude spectra of individual speech can be received using Eq. (3). With the amplitude estimation of the speaker and phase of the mixed signal, we can recover the signal from the frequency domain to the time domain via inverse STFT. Finally, the estimated speech signal's waveform is synthesized by an overlap add method.

3.4 Joint constraint algorithm for SCSS

On the basis of the new loss function, the whole process of the joint constraint algorithm is described as Table 1. In the training stage, the target and mixed speech signals are preprocessed for obtaining the magnitude spectra and IRMs. Then, the BP algorithm is used to solve the proposed cost function, and the DNN parameters are obtained. Thus, the training of DNN is completed. With the joint constraint of input and outputs of DNN, a good separation effect can be obtained. In the separation stage, the estimated ratio masks can be obtained with the well-trained DNN and the test signal.

Then the estimated magnitude spectra can be calculated by Eq. (3), and we can synthesize separated speech signals.

4 Experiment and results analysis

In this section, we evaluate the system performance of speech separation by conducting experiments on the GRID corpus [18]. Firstly, we introduce the dataset and experimental settings. Secondly, the impact of regularization parameter is discussed. Thirdly, the comparison of performance between the proposed algorithm and the basic DNN cost function is displayed. Finally, we compare the performance of the proposed method with other SCSS approaches based on different combinations of training targets and the number of outputs.

4.1 Dataset and experimental setup

1) Datasets

We perform the SCSS experiment on the GRID corpus, from which both the training set and the test set are selected. This dataset contains of 18 males and 16 females, each person has 1000 clean utterances. 500 utterances are also randomly selected from GRID corpus for each speaker as a training set. The test data are randomly selected 50 sentences from the remaining 500 utterances, and the final results are obtained by averaging.

All utterances used in the training set and test set are down-sampled to 25 kHz, and the input features are obtained by STFT. The frame length used for the extraction is 512 samples, and frame shift is 256 samples.

2) Parameter settings

The dual-output DNN framework used in experiments is 512-1024-1024-1024-1024, which denotes 512 nodes for the input layer, 1024 nodes for each of the three hidden layers and 1024 nodes for output layer. Because of 512 nodes for the output layer of single source, the size of nodes for the output layer of dual source is 1024. The number of epoch and the batch size in experiments are set to 50 and 128, respectively. The learning rate of the first 10 epoch is set to 0.1, and then decreases by 10% every epoch at last 40 epoch. As for the regularization parameter, we set it to 0.5 when separating male + male and female + male utterances, and 0.6 when separating female + female signal. The reasons for the selection of regularization parameters are analyzed in Sect. 4.2.

3) Evaluation metrics

In this paper, we used several metrics to evaluate the performance of separation, Perceptual Evaluation of Speech Quality (PESQ) score [19], signal-to-distortion ratio (SDR),

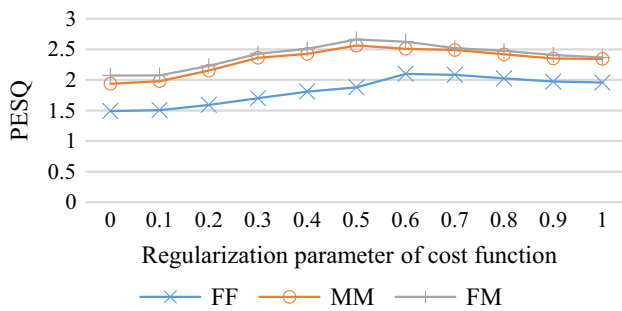


Fig. 2 Average separation performance with respect to regularization parameter

signal-to-interference ratio (SIR) and sources-to-artifacts ratio (SAR) [20]. The higher the metrics are, the less the separated speech distortion is.

4.2 Influence of regularization parameter

To obtain the optimal result in our algorithm, we study the effects of regularization parameter on the performance of different gender combinations. From the corpus, we randomly select 2 males and 2 females, where a total of three gender combinations (F + F, F + M and M + M) can be obtained. F and M are female and male, respectively. And the regularization parameter was changed from 0 to 1 with an increment of 0.1. The effect of regularization parameter on PESQ is shown in Fig. 2. We can see that when λ is less than 0.5, the speech intelligibility of all the three gender combinations is improved with the increase in λ . When the range of λ is from 0 to 0.3, the PESQ increased rapidly, and slowly increased between 0.4 and 0.5. Because of the difference of diverse gender combinations, the best separation performance of λ is not the same. Among them, M + M separation and F + M separation have the best performance when λ is set to 0.5. For F + F separation, the highest separation performance can be achieved when λ is set to 0.6. However, when λ is greater than 0.6, the speech intelligibility of the separated signal begins to decline slowly. Even in this attenuation, the separation performance is better when λ is set to 1 than when λ is set to 0, which proves the effectiveness of the proposed algorithm. Therefore, for different gender combinations, we choose different regularization parameters. Separating M + M and F + M mixed speech signals, λ is set to 0.5, and separating mixed signal in F + F, λ is set to 0.6.

4.3 Performance comparison

In this part, we first compare our proposed algorithm with the dual-output DNN-based speech separation using the basic cost function (noted as Basic). Secondly, the performance of dual-output DNN with different training targets is evaluated.

Then, we assess the results of our method with single-output DNN using different training targets. The comparators can be noted as: AMS-single, AMS-dual (i.e. the basic function), IRM-single, and IRM-dual. IRM and amplitude spectra (AMS) are the training targets of DNN, and dual and single are the output numbers of DNN, respectively.

(1) Comparison with basic cost functions

To evaluate the effectiveness of the proposed method, we have done a series of experiments compared with the cost functions as Eq. (5) and cost function in [9]. We randomly select 2 males and 2 females from the GRID corpus, where a total of six gender combinations (F1 VS F2, F1 VS M1, F1 VS M2, F2 VS M1, F2 VS M2 and M1 VS M2) can be obtained. F1 and F2 are females. M1 and M2 are males. The SDR, SAR, SIR and PESQ values of 50 sentences about six gender combinations are tested, and all of our results are averaged. From the results shown in (a) to (d) of Fig. 3, we can see that compared with the basic dual output DNN (Basic), the SDR, SAR, SIR and PESQ values of proposed method have generally increased under different gender combinations. This is because the proposed method explores the joint relationship between different sources. Specifically, it can be observed from Fig. 3(a)–(d) that the proposed method improves 0.77 dB in SDR, 0.69 dB in SAR, 0.51 dB in SIR and 0.46 in PESQ compared with the basic cost function in separating mixed signals of female–female. We can observe that the optimization in terms of SDR and SIR is not obvious in F1 VS F2, because the pitch frequency is similar in female speech sounds, and the difference between female signals is smaller than other gender mixtures. As we can see from Fig. 3, in terms of male–male separation, SDR, SAR, SIR and PESQ were 1.72 dB, 1.42 dB, 0.86 dB and 0.4, respectively, higher than the method with basic cost function. What’s more, the performance of the male–male signal separated by our method is almost the same as that in male–female obtained by the basic algorithm. Even when SAR and PESQ are used as measurement indicators, our method is slightly higher than the female–male in basic function. It can be seen from the overall trend in the figure that the separation performance is best when signals are mixed up by male and female speech signal. This is because there is the significant discrimination contained in the signals between male and female, such as in the amplitude and pitch frequencies. Compared with the method in [9] (MaxDiffer), the performance of the proposed algorithm is still outstanding. It can be observed from Fig. 3 that the proposed algorithm improves SDR, SAR, SIR and PESQ 0.57 dB, 0.48 dB, 0.16 dB and 0.284 compared with the MaxDiffer in separating mixed signals of female–female. In addition, the proposed algorithm is 0.44 dB, 1.2 dB, 0.35 dB and 0.198 higher in SDR, SAR, SIR and PESQ than the method using MaxDiffer when separating

Fig. 3 Performance comparison for different separation approaches. **a** SDR, **b** SAR, **c** SIR and **d** PESQ

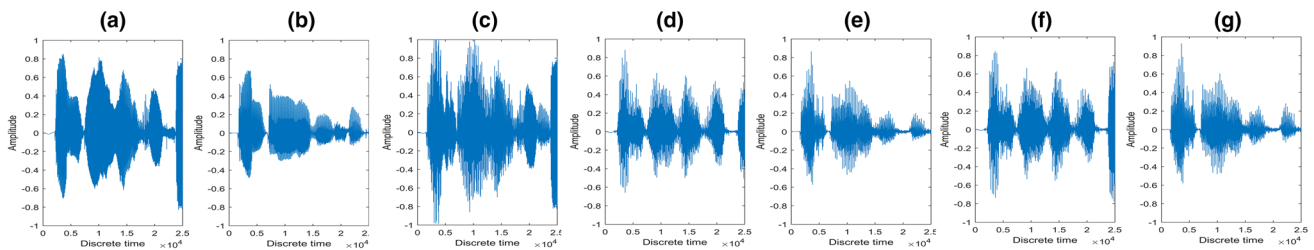
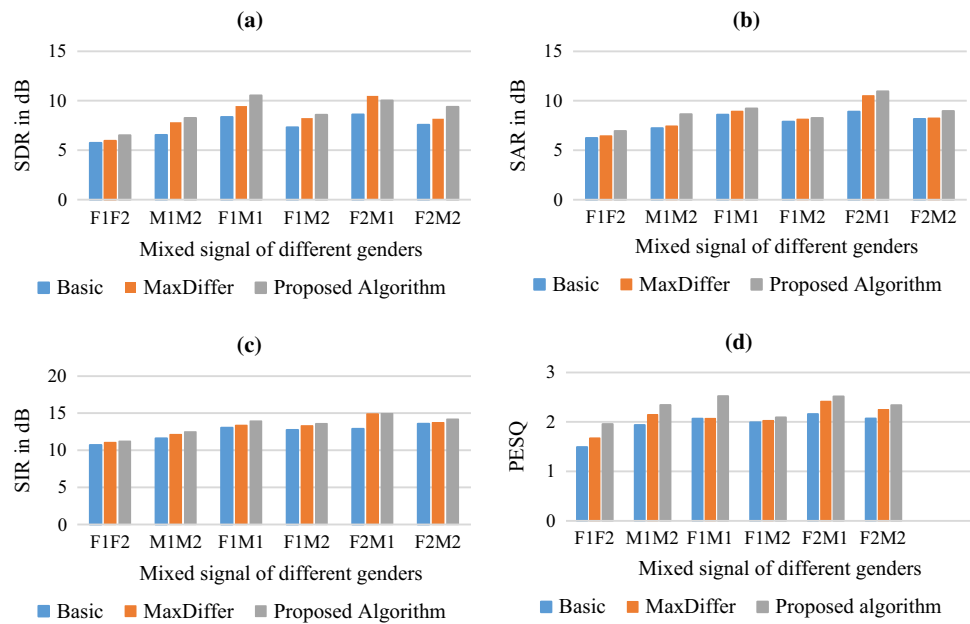


Fig. 4 Schematics of F–F signals synthesis and decomposition. **a** Target (F1). **b** Target (F2). **c** Mixed (F1 + F2). **d** Estimation separated by the DNN-basic (F1). **e** Estimation separated by the DNN-basic (F2).

f Estimation separated by the proposed algorithm (F1). **g** Estimation separated by the proposed algorithm (F2)

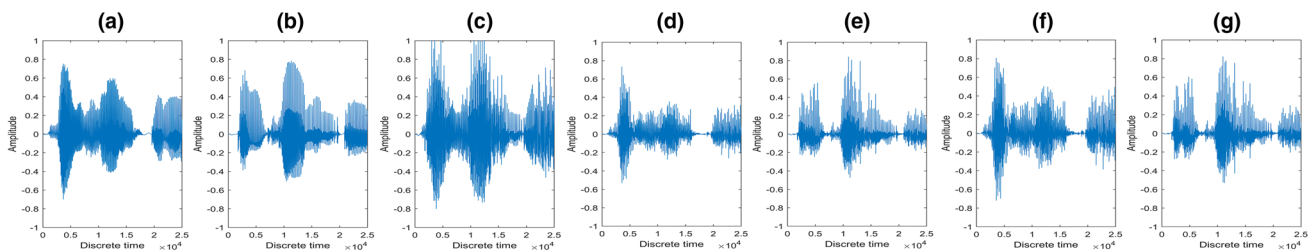


Fig. 5 Schematics of M–M signals synthesis and decomposition. **a** Target (M1). **b** Target (M2). **c** Mixed (M1 + M2). **d** Estimation separated by the DNN-basic (M1). **e** Estimation separated by the DNN-basic (M2).

f Estimation separated by the proposed algorithm (M1). **g** Estimation separated by the proposed algorithm (M2)

male–male signal. Although in terms of F2–M1 separation, the proposed method is slightly inferior to the method using the maximum difference term in SDR and SIR, the overall trend of separating the female–male signals is still better than MaxDiffer.

We can see from these results that the algorithm proposed has superiorities in the intelligibility of speech and the quality of separation. We randomly extract the utterances of the

different gender pairs from the corresponding test data, as shown in Figs. 4, 5, and 6, where (a) and (b) are schematic representations of the pure speech in the time domain, and (c) represents the spectra of the mixed speech. (d) and (e) are estimated sources separated by the DNN-basic method. To give a fair comparison with (d) and (e), the normalized version of our algorithm is shown in (f) and (g). It is obvious

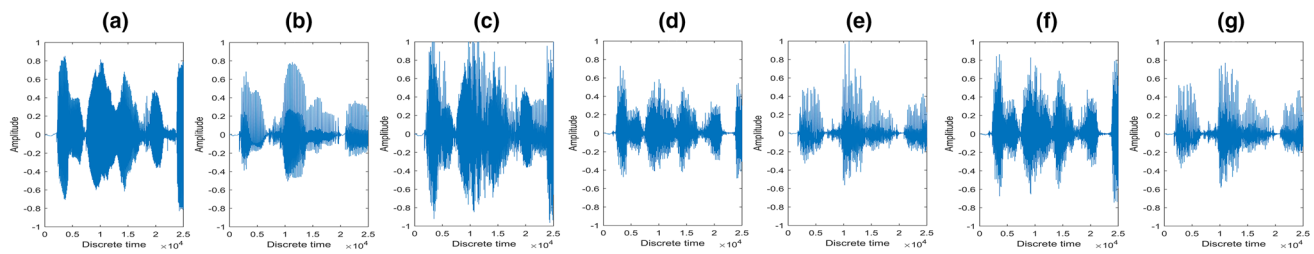


Fig. 6 Schematics of F–M signals synthesis and decomposition. **a** Target (F). **b** Target (M). **c** Mixed (F + M). **d** Estimation separated by the DNN-basic (F). **e** Estimation separated by the DNN-basic (M). **f** Estimation separated by the proposed algorithm (F). **g** Estimation separated by the proposed algorithm (M)

Table 2 Performance comparisons of dual-output DNN with different targets

Combination	Gender	SDR (dB)	SAR (dB)	SIR (dB)	PESQ
AMS-dual	F + F	3.7093	3.3435	7.1954	1.219
	F + M	4.7461	4.8484	9.8097	1.763
	M + M	4.0484	4.6266	7.8512	1.702
IRM-dual (basic)	F + F	5.7422	6.2422	10.7022	1.49
	F + M	7.8955	8.3877	13.1102	2.071
	M + M	6.541	7.2256	11.619	1.937
Proposed algorithm	F + F	6.515	6.9355	11.2191	2.098
	F + M	9.581	9.343	14.296	2.662
	M + M	8.2652	8.6468	12.4819	2.562

Table 3 Performance comparisons of different targets and output numbers

Combination	Gender	SDR(dB)	SAR(dB)	SIR(dB)	PESQ
AMS-single	F + F	4.7381	4.9802	9.4581	1.521
	F + M	5.8347	5.954	11.5716	1.858
	M + M	5.025	5.467	9.9059	1.675
IRM-single	F + F	6.7343	7.2066	11.308	2.281
	F + M	9.6159	10.0293	14.9217	2.931
	M + M	8.4084	8.7147	12.8915	2.744
Proposed algorithm	F + F	6.515	6.9355	11.2191	2.098
	F + M	9.581	9.343	14.296	2.662
	M + M	8.2652	8.6468	12.4819	2.562

that the amplitude of the speech obtained by our algorithm is closer to the original signal.

(2) Evaluations of dual-output DNN with different training targets

To evaluate the performance of dual-output DNN with different training targets in three gender mixtures, we conduct experiments on three different gender combinations, namely F + F, F + M, M + M. From the results displayed in Table 2, we can see that the separation performance of IRM-targeted DNN is better than amplitude spectrum-targeted DNN in general. This can be explained that the mapping-based approach works well at low frequencies, but loses some details in medium and high frequencies, which are important for speech intelligibility and speech quality [21]. The proposed algorithm is outstanding in different gender combinations compared with other dual-output methods. Especially in separating mixed signals of males, the proposed method is 1.72 dB in SDR, 1.19 dB in SIR, 1.42 dB in SAR and 0.63 in PESQ higher than the IRM-dual method. In addition, compared with AMS dual output, the advantages are more obvious. For example, in terms of F–M separation, the optimizations of SDR, SAR, SIR and PESQ are 4.83 dB, 4.63 dB, 4.49 dB and 0.9, respectively. It can be seen from

the results that the separation performance and speech intelligibility of the separated signal are significantly improved by the proposed method.

(3) Comparisons with single-output DNN using different training targets

In this part, the single-output DNN is trained for mapping the relationship between mixed signal features and target signal features. As shown in Table 3, the performance of the proposed algorithm is better than the AMS-single and slightly lower than the IRM-single in general. It may be explained that the single-output DNN is trained for one signal, and the training parameters are more suitable for the specified signal. Specifically, compared with single output, the dual output results of each gender combination on SDR, SAR and SIR decreased. However, the performance of our proposed algorithm is slightly different from that of single-output DNN on SDR, SAR, SIR and PESQ where the maximum reduction is 0.22 dB, 0.69 dB, 0.63 dB and 0.269, respectively. For SDR in F + F, SAR in M + M, and SIR in F + M, the results are almost the same as the single-output separation method. The results clearly show the effectiveness of the proposed algorithm. In addition, different sources can be separated simultaneously by the dual-output method, while

DNN based on single-output can only separate one signal at a time, which is time consuming.

5 Conclusion

In this work, we propose a joint constraint algorithm based on a new cost function which is used to train the dual output deep neural network for the single-channel speech separation problem. The new cost function can exploit the joint information between the sources which are needed to be separated. In order to verify the proposed algorithm performance, we compare it with the methods using state-of-the-art cost function in the dual-output DNN. The experimental result shows that the new method yields significant performance improvements over them. It also indicates that the novel cost function estimates the corresponding ideal output value more accurately and exploits the relationship between the outputs. In this joint constraint, the training accuracy of the separation method can be further increased, and its performance is close to that of single-output DNN. What's more, we further explore the effect of regularization parameter on the intelligibility of the separated sources. The separation performance of the same gender needs to be further improved, so in the future work, we will further reduce the distortion of separated same-gender signals.

Acknowledgements This work is supported by the National Natural Science Foundation of China (Nos. 61901227, 61671252) and the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (No. 19KJB510049).

References

- Du, J., Tu, Y., Dai, L., Lee, C.: A regression approach to single-channel speech separation via high-resolution deep neural networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **24**(8), 1424–1437 (2017)
- Corey, R. M., Singer, A. C.: Dynamic range compression for noisy mixtures using source separation and beamforming. In: *Proc. IEEE Workshop Appl. Signal Processing Audio and Acoustics*, pp. 289–293. New Paltz, NY, USA (2017)
- Chang, J., Wang, D.: Robust speaker recognition based on DNN/i-Vectors and speech separation. In: *Proceedings of IEEE International Conference Acoustics Speech Signal Processings*, pp. 5415–5419. New Orleans, LA, USA (2017)
- Narayanan, A., Wang, D.L.: Investigation of speech separation as a front-end for noise robust speech recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **22**(4), 826–835 (2014)
- Zhang, X.L., Wang, D.L.: A deep ensemble learning method for monaural speech separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **24**(5), 967–977 (2016)
- Han, K., Wang, Y., Wang, D.L., et al.: Learning spectral mapping for speech dereverberation and denoising. *IEEE/ACM Trans. Audio Speech Lang. Process.* **23**(6), 982–992 (2015)
- Sun, Y., Wang, W., Chambers, J., et al.: Two-stage monaural source separation in reverberant room environments using deep neural networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **27**(1), 125–139 (2019)
- Tu, Y., Du, J., Xu, Y., Dai, L., et al.: Speech separation based on improved deep neural networks with dual outputs of speech features for both target and interfering speakers. In: *Proc. Int. Symp. Chin. Spoken Lang. Process.*, Singapore, pp. 250–254. Singapore (2014)
- Huang, P.-S., Kim, M., Hasegawa-Johnson, M., et al.: Joint optimization of masks and deep recurrent neural networks for monaural source separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **23**(12), 2136–2147 (2015)
- Wang, Y., Du, J., Dai, L.R., et al.: A gender mixture detection approach to unsupervised single-channel speech separation based on deep neural networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **25**(7), 1535–1546 (2017)
- Zhang, X., Wang, D.L.: Deep learning based binaural speech separation in reverberant environments. *IEEE/ACM Trans. Audio Speech Lang. Process.* **25**(5), 1075–1084 (2017)
- Grais, E.M., Roma, G., Simpson, A.J.R., et al.: Two stage single channel audio source separation using deep neural networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **25**(9), 1773–1783 (2017)
- Naithani, G., Nikunen, J., Bramsløw, L., et al.: Deep neural network based speech separation optimizing an objective estimator of intelligibility for low latency applications. In: *Proc. IEEE Workshop Acoustic Signal Enhancement*, pp. 386–390. Tokyo, Japan (2018)
- Weninger, F., Eyben, F., Schuller, B.: Single-channel speech separation with memory-enhanced recurrent neural networks. In: *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3709–3713. Florence, Italy (2014)
- Joho, M., Lambert, R. H., Mathis, H.: Elementary cost functions for blind separation of non-stationary source signals. In: *Acoustics, Speech, & Signal Processing*, vol. 5, pp. 2793–2796. Salt Lake City, UT, USA (2001)
- Sun, L., Xie, K., Gu, T., et al.: Joint dictionary learning using a new optimization method for single-channel blind source separation. *Speech Commun.* **106**, 85–94 (2019)
- Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. *Nature* **323**(6088), 533–536 (1986)
- Cooke, M., Barker, J., Cunningham, S., et al.: An audio-visual corpus for speech perception and automatic speech recognition. *J. Acoust. Soc. Am.* **120**(5), 2421–2424 (2006)
- Rix, A., Beerends, J., Hollier, M., et al.: Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In: *Proc. ICASSP*, pp. 749–752. Salt Lake City, UT, USA (2001)
- Vincent, E., Gribonval, R.: Fevotte C 2006 Performance measurement in blind audio source separation. *IEEE Trans. Audio Speech Lang. Process.* **14**(4), 1462–1469 (2006)
- Wang, Y., Narayanan, A., Wang, D.L.: On training targets for supervised speech separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **22**(12), 1849–1858 (2014)

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