Chinese Journal of Electronics
Vol.32, No.3, May 2023

Monaural Speech Separation Using Dual-Output Deep Neural Network with Multiple Joint Constraint

SUN Linhui, LIANG Wenqing, ZHANG Meng, and LI Ping'an

(*College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China*)

Abstract — Monaural speech separation is a significant research field in speech signal processing. To achieve a better separation performance, we propose three novel joint-constraint loss functions and a multiple joint-constraint loss function for monaural speech separation based on dual-output deep neural network (DNN). The multiple joint-constraint loss function for DNN separation model not only restricts the ideal ratio mask (IRM) errors of the two outputs, but also constrains the relationship of the estimated IRMs and the magnitude spectrograms of the clean speech signals, the relationship of the estimated IRMs of the two outputs, and the relationship of the estimated IRMs and the magnitude spectrogram of the mixed signal. The constraint strength is adjusted through three parameters to improve the accuracy of the speech separation model. Furthermore, we solve the optimal weighting coefficients of the multiple joint-constraint loss function based on the optimization idea, which further improves the performance of the separation system. We conduct a series of speech separation experiments on the GRID corpus to validate the superiority performance of the proposed method. The results show that using perceptual evaluation of speech quality, the short-time objective intelligibility, source to distortion ratio, signal to interference ratio and source to artifact ratio as the evaluation metrics, the proposed method outperforms the conventional DNN separation model. Taking the gender into consideration, we carry out experiments among Female-Female, Male-Male and Male-Female cases, which show that our method improves the robustness and performance of the separation system compared with some previous approaches.

Key words — Monaural speech separation, Joint constraint, Deep neural networks, Dual-output.

I. Introduction

Speech separation [1] is a challenging task, which belongs to a branch of source separation [2]. Speech separation aims to recover the high-quality and high-intelligibility target speech signal from the mixed speech. It's quite simple for a human to distinguish the target speaker speech from the noise or multi-speakers' environment. However, it's difficult for machine to recognize the correct one. In daily life, speech separation can be applied to the frontend of machine translation, advanced hearing aid [3] and automatic speech recognition (ASR) [4], which helps improve their performance. Therefore, it is significant and practical to make machine with the ability to obtain pure source signal from the mixed source signal. The source separation problem [5] can be categorized into multichannel, stereo-channel (binaural) and single-channel (monaural). Among them, monaural speech separation usually refers to the way of recording the mixed speech with a single microphone, which is often chosen as an experimental object due to its practicality.

To handle the tough monaural speech separation, some approaches were proposed over the past several decades. Those methods can be divided into three categories: unsupervised separation method, semi-supervised separation method and supervised separation method. With unsupervised separation method, a prime class of monaural speech separation is known as computational auditory scene analysis (CASA) [6], which extracts speech features to separate interesting speech by masking other interfering sources. Features such as pitch and amplitude modulation spectrum (AMS) are exploited to segregate the voiced component of co-channel speech [7]. Independent component analysis (ICA) [8] is also an unsupervised method used for speech sep-

Manuscript Received Apr. 27, 2022; Accepted Aug. 16, 2022. This work was supported by the National Natural Science Foundation of China (61901227, 62071242) and the China Scholarship Fund (202008320043).
 DOI:10.23919/cje.2022.00.110

aration. Recently, there has been increasing interest in purely supervised approaches. According to the difference of training target, supervised speech separation method [9] can be categorized to Time-Frequency (T-F) masking-based method and mapping-based method.

T-F masking-based method learns a mapping function from the mixed speech features to T-F mask of the clean speech to separate mixed speech. This kind of methods commonly estimates the ideal binary mask (IBM) and ideal ratio mask (IRM) [10], or some other masks such as phase-sensitive mask (PSM) [11] and complex ideal ratio mask (cIRM) [12]. In general, IBM or IRM is often used as a training target to conduct experiments. For IBM, according to the comparison result of signal-to-noise (SNR) and local threshold, a T-F unit is assigned 1 or 0. As for the IRM, a T-F unit is dynamically assigned the ratio of target signal energy and mixture signal energy. Because IBM can be viewed as a two-category classification problem, Gaussian mixture models (GMM) and Bayesian classifier can be used to predict IBM [13]. May and Dau [14] used GMM and well-trained SVM to estimate IBM. With the development of deep learning, researchers pay much attention to the deep neural network (DNN) which can build the nonlinear relationship between input and output. Wang *et al.* [10] first used DNN-based method to get IBM for speech separation. Zhang and Wang [9] used ensemble learning methods to set up a multi-context network to estimate IRM for speech separation. Huang *et al.* [15] constructed a recurrent neural network (RNN) for speech separation, where the IRM was embedded as an extra processing layer to the output layer, and their method outperforms that with non-negative matrix factorization (NMF). Besides T-F mask-based method, mapping-based method is also widely used in speech separation. Mapping-based method builds a regression function from the mixed speech features to the clean speech. Du *et al.* [4] proposed a DNN-based regression model to predict the nonlinear relationship between mixed speech and target speech using a large numbers of training data. Wang *et al.* [16] exploited multi-output DNN to train the gender detector, which helps to build another mapping-based regression network for speech separation according to different gender combinations.

In recent years, speech separation approaches based on deep learning have been rapidly developed. Chauhan *et al.* [17] constructed a model of speech separation and speech recognition based on deep learning, which is an efficient and accurate method to separate mixture speech and recognize emotion of speech simultaneously. The time domain approach based on deep learning has achieved good results on speech separation problems recently. Wan [18] proposed a time domain speech separation algorithm based on fully convolutional network, which makes up for the shortcomings of the traditional T-F domain method. Fan *et al.* [19] proposed an end-to-end approach for speech separation based on 1-demensional convolutional network, which exploits speech waveform as the input of network for preliminary separation, followed by the fusion depth feature for further separation. Besides, many studies have demonstrated that phase information has an important role in speech perception quality. Zheng *et al.* [20] constructed a separation model based on deep learning with novel training target named instantaneous frequency deviation (IFD). This method optimizes magnitude spectrum and phase spectrum obtained by estimated IFD at the same time, which improves the intelligibility of the separated speech.

Among various monaural speech separation methods based on deep learning, the loss function has a great impact on the performance of separation system, because it controls the quality of the model. Kang *et al.* [21] proposed a novel loss function, which is compounded of Mel-scale weighted mean square error, temporal and spectral variations similarities between the original reference speech signal and the estimated speech signal. This method computed the gradients based on non-linear frequency, which can mitigate the excessive smoothing of the estimated speech signal. In addition, Naithani *et al.* [22] also proposed a novel objective function. This objective function optimized the extender short time objective intelligibility (ESTOI) measure and obtained prominent performance. However, these methods mentioned above did not take into accounts the joint relationship between the different source signals to be separated. Our team once considered the relationship between output masks of different source signals and conducted a series of studies with good results [23]. To further improve the accuracy of separating multiple sources, we propose the novel objective function considering the joint relationship between the separated sources from the multiple aspects. From our experiments, the proposed speech separation method obtains better results compared with those of other loss functions.

This article is organized as follows. In Section II, we present the flowchart of the traditional speech separation system, the training targets and the traditional loss functions for speech separation based on DNN model. In Section III, we elaborate three joint-constraint loss functions, and a comprehensive joint-constraint loss function, and present the whole process of speech separation based on the multiple joint-constraint loss function. In Section IV, we present the experimental set-

tings and results analysis. Finally, we summarize this article and describe future work in Section V.

II. Monaural Speech Separation Model

1. Monaural speech separation problem formulation

There are many ways to simulate the mixed signal, e.g., linear instantaneous mixing, linear convolution mixing and non-linear mixing. In the majority cases in speech separation, linear instantaneous mixing is frequently chosen. The monaural mixed signal is often formulated as follows:

$$y(t) = \sum_{i=1}^{n} x_i(t) \tag{1}$$

where $x_i(t)$ and $y(t)$ refer to the i-th target signal and mixed signal combined with $n(n \geq 2)$ branch sources, respectively. The mixed signal with two target signals is commonly used for experiment owing to its convenience and simplicity, and it can be formulated as $y(t) = x_1(t) + x_2(t)$.

The DNN-based dual-output separation model can separate $x_1(t)$ and $x_2(t)$ simultaneously. The overall flowchart of the traditional monaural speech separation based on dual-output DNN is shown in Fig.1 , which aims to simultaneously evaluate the two target signals from the mixed signal. The network has five layers, including input layer, three hidden layers and output layer. Before the training of network, the time-domain speech signal is preprocessed to extract features as the input of the network. The input features of DNN model for speech separation are usually short-time Fourier transform (STFT) magnitude spectrogram, log-power spectrogram (LPS) and some other acoustic features [24] like Mel-Frequency Cepstral Coefficients (MFCC), Multiresolution Cochlea-gram (MRCG), Gammatone Frequency Cepstral Coefficients (GFCC), etc.

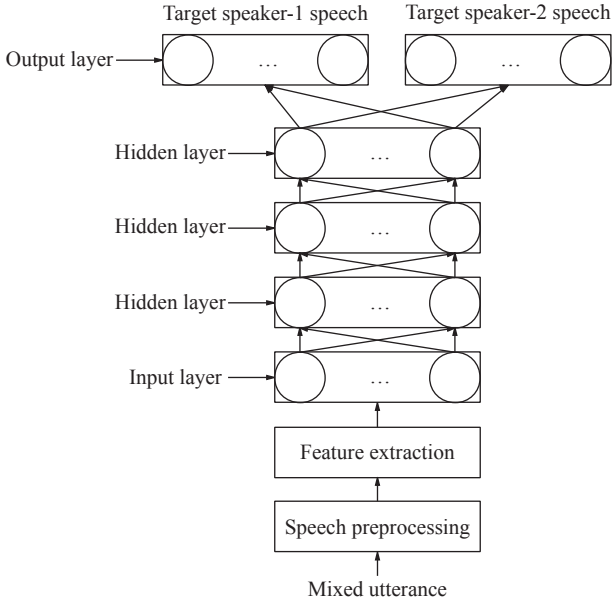


Fig. 1. Overall flowchart based on dual-output DNN for monaural speech separation.

2. Training target for DNN-based model

1) Ideal ratio mask

The ideal ratio mask (IRM) is viewed as the soft form of IBM and can lead to a better performance in many cases. It is frequently chosen as the training target of DNN-based model in supervised speech separation. The formulation of IRM is shown as follows:

$$\boldsymbol{M_i}(t,f) = \left(\frac{\boldsymbol{S}_i(t,f)^2}{\boldsymbol{S}_1(t,f)^2 + \boldsymbol{S}_2(t,f)^2 + \boldsymbol{\varepsilon}} \right)^k, \; i = (1,2) \tag{2}$$

where $\boldsymbol{S}_1(t,f)$ and $\boldsymbol{S}_2(t,f)$ represent magnitude spectrogram of target speaker-1 speech and target speaker-2 speech at time frame m and frequency channel f, respectively. To prevent the denominator from becoming 0, $\boldsymbol{\varepsilon}$ is a minimal positive number. The tunable parameter k is commonly chosen as 0.5. In masking-based methods for speech separation, the estimated T-F mask via DNN model is multiplied with the corresponding elements of the magnitude spectrogram of the mixed speech to obtain the target speech magnitude spectrogram. And then, the time domain waveform originates from the estimated magnitude spectrogram via inverse short-time Fourier transform (ISTFT).

2) Target magnitude spectrum

The target magnitude spectrogram (TMS), $\boldsymbol{S}_1(t,f)$, $\boldsymbol{S}_2(t,f)$ is the corresponding training target of mapping-based approaches. TMS is aimed to learn the linear regression function between target speech and mixed speech via DNN or other prediction models. And then, adding the phase information of the mixed signal into the estimated magnitude spectrogram, we can reconstruct the utterances of two target speakers via ISTFT directly.

3. Loss function for DNN-based model

1) Masking-based method

The magnitude spectrogram of the time domain signal $x_1(t)$ and $x_2(t)$ can be generated via speech preprocessing and STFT, denoted as $\boldsymbol{S}_1(t,f)$ and $\boldsymbol{S}_2(t,f)$, respectively. Then, we can calculate the IRM by (2). The DNN-based model for monaural speech separation is a linear regression model, not a logistic regression model, and the minimum mean square error (MMSE) is often chosen as the criterion of DNN architecture. The loss function reveals the relationship and distance between estimated features and real values, which is mainly used to solve the parameters of DNN.

The loss function of the single output DNN model primarily concentrates on mapping the relation between the estimated ratio value and the original reference ratio value. The loss function with one target source is formulated as follows:

$$\text{Loss}_1 = \frac{1}{2T}\sum_{t=1}^{T} \| \widehat{\boldsymbol{M}}_t - \boldsymbol{M}_t \|^2 \tag{3}$$

where T denotes the total number of time frames. $\boldsymbol{M}_t$ and $\widehat{\boldsymbol{M}}_t$ are the ideal value and the estimated value of ratio mask of the target source at t-th frame, respectively. Here, $\widehat{\boldsymbol{M}}_t$ is the output of DNN model using the feature of mixed signal $\boldsymbol{Y}_t$ as the input. $\widehat{\boldsymbol{M}}_t = f(\boldsymbol{Y}_t)$, where $f(\cdot)$ refers to the complicated function operation of DNN layers. When two target source signals need to be separated and reconstructed, two DNN models are needed for the single output DNN architecture, while only one dual-output DNN architecture is needed, which will cut down the time and computation consumption of speech separation system. The loss function of the dual-output DNN is formulated as follows:

$$\text{Loss}_2 = \frac{1}{2T}\sum_{t=1}^{T}(\| \widehat{\boldsymbol{M}}_{1t} - \boldsymbol{M}_{1t} \|^2 + \| \widehat{\boldsymbol{M}}_{2t} - \boldsymbol{M}_{2t} \|^2) \tag{4}$$

where $\boldsymbol{M}_{it}$ and $\widehat{\boldsymbol{M}}_{it}$ are the original reference IRMs and their estimation values of the i-th speaker at t-th frame, respectively. The loss function (3) can only get the estimated IRM of one target signal at a time, while (4) can get estimated IRM of two target signals at the same time.

2) Mapping-based method

Different from the masking-based method, the mapping-based method like TMS is to establish the nonlinear relationship between the feature spectrogram of the mixture speech signal and the feature spectrogram of the clean speech signal of the target speakers directly. Even if the DNN architecture is similar with that with the masking-based method, it corresponds to the different loss function, which can be formulated as following MMSE issue:

$$\text{Loss}_3 = \frac{1}{2T}\sum_{t=1}^{T}(\| \widehat{\boldsymbol{S}}_{1t} - \boldsymbol{S}_{1t} \|^2 + \| \widehat{\boldsymbol{S}}_{2t} - \boldsymbol{S}_{2t} \|^2) \tag{5}$$

where $\boldsymbol{S}_{1t}$ and $\boldsymbol{S}_{2t}$ indicate the ideal feature spectrogram vectors, and $\widehat{\boldsymbol{S}}_{1t}$, $\widehat{\boldsymbol{S}}_{2t}$ indicate the estimated feature spectrogram vectors of two target sources, respectively. In the DNN training stage, magnitude spectrogram of mixed signal $y(t)$ are also used as the input features of DNN model. By minimizing the loss function, DNN model can be well trained and suitable network parameters can be obtained. In the DNN testing stage, we exploit the magnitude spectrogram of testing mixed utterance of two speakers as the input features of the DNN model, and then we can get the enhanced magnitude spectrogram of two target sources from the output layer, simultaneously. Eventually, we can reconstruct the utterances of two target speakers via ISTFT.

III. Speech Separation Based on DNN with Joint Constraint

The loss function used in traditional dual-output speech separation based on DNN usually only considers the error between the estimated value and the original reference value, which can't well constrain the training of neural networks, resulting in poor separation performance. In order to improve the clarity and decrease the distortion of the separated speech, we propose three novel joint-constraint (JC) loss functions from different aspects for speech separation based on dual-output DNN, which correspondingly considers the relationship between the predicted IRM and the magnitude spectrogram, the relationship between IRMs and the connection between the target magnitude spectrogram and the mixed magnitude spectrogram. Furthermore, we comprehensively consider all factors above and incorporate them into a new joint-constraint loss function. All joint-constraint loss functions proposed in this paper can make the predicted mask close to the ideal mask and decrease the error between the estimated magnitude spectrogram obtained by the predicted mask and the ideal magnitude spectrogram, which improve the intelligibility of the reconstructed speech signal. In the following subsections, we will discuss separately these newly proposed joint-constraint loss functions in detail.

1. Joint constraint of IRM and magnitude

In DNN-based speech separation, we usually use neural network to estimate time-frequency mask of speech signal, but finally we need to recover the magnitude spectrogram of the target speech signal, which is indirectly obtained through the mask estimated by the trained model. Therefore, the joint constraint of the predicted mask and the magnitude spectrogram can reduce the distortion of the reconstructed magnitude spectrogram of the target signal so that improves the performance of separation. Considering the relationship between IRM and magnitude spectrogram, we propose a joint-constraint loss function which minimizes the weighted sum of the IRM errors and the magnitude spectrogram errors corresponding to IRM estimations of the two speech signals. The loss function can be written as

$$\begin{aligned}\mathrm{JC}_1 =&\frac{1}{2T}\sum_{t=1}^{T}(\| \widehat{\boldsymbol{M}}_{1t}-\boldsymbol{M}_{1t}\|^2+\| \widehat{\boldsymbol{M}}_{2t}-\boldsymbol{M}_{2t}\|^2\\&+\alpha(\| \widehat{\boldsymbol{M}}_{1t}\odot\boldsymbol{Y}(t)-\boldsymbol{S}_{1t}\|^2\\&+\| \widehat{\boldsymbol{M}}_{2t}\odot\boldsymbol{Y}(t)-\boldsymbol{S}_{2t}\|^2))\end{aligned}\tag{6}$$

where $\boldsymbol{S}_{1t}$ and $\boldsymbol{S}_{2t}$ are the ideal magnitude spectrogram of target speaker-1 and target speaker-2, respectively. $\widehat{\boldsymbol{M}}_{1t}$ and $\widehat{\boldsymbol{M}}_{2t}$ are both the output of the network. $\odot$ denotes Hadamard product operator, which means that the matrix on both sides of the symbol is multiplied element by element. The estimated magnitude spectrogram of the two speakers can be obtained by the following formulas:

$$\begin{cases}\widehat{\boldsymbol{S}}_{1t}=\widehat{\boldsymbol{M}}_{1t}\odot\boldsymbol{Y}(t)\\ \widehat{\boldsymbol{S}}_{2t}=\widehat{\boldsymbol{M}}_{2t}\odot\boldsymbol{Y}(t)\end{cases}\tag{7}$$

Since our purpose is to separate the speech signals of two target speakers without distortion as much as possible, the estimated value obtained by multiplying the mask output from the network, by the magnitude spectrogram of the mixed speech from different speakers should be as close as possible to the magnitude spectrogram of our target speaker. That is to say, the smaller the error between $\widehat{\boldsymbol{S}}_{1t}$ and $\boldsymbol{S}_{1t}$, $\widehat{\boldsymbol{S}}_{2t}$ and $\boldsymbol{S}_{2t}$, the higher the reducibility of separated speech and the better the separation performance. The importance of the two speakers is the same, so the two terms at the end of (6) share a common weight α.The range of α is from 0 to 1. When α takes different value, it means different constraint to separation.

Compared with the traditional loss function, the loss function (6) makes full use of the relationship between the mask and the magnitude spectrogram, which increases the effectiveness of the mask constraint. It not only makes the distance between the estimated IRM and the ideal IRM small but also makes the distance between the estimated magnitude spectrogram and the ideal magnitude spectrogram small. The neural network trained with the guidance of this joint-constraint loss function can output more accurate mask and the calculated magnitude spectrogram from the estimated value is closer to the target magnitude spectrogram so that the recovered target speech signal is with less loss.

2. Joint constraint of IRM and sum of mask squares

For traditional dual-output speech separation based on DNN, the loss function usually considers the relationship between the predicted value and the ideal value of a single signal. In fact, there is also a unique relationship between masks corresponding to two different source signals when the mixed speech comes from two clean speech signals. In this paper, we not only take into accounts the two masks separately, but also investigate them together. In the full consideration of the characteristic between masks, we propose a novel joint-constraint loss function which jointly constrains the masks of two speech signals to improve the correctness of model prediction. The second joint-constraint loss function proposed in this paper can be written as

$$\begin{aligned}\mathrm{JC}_2 =&\frac{1}{2T}\sum_{t=1}^{T}(\| \widehat{\boldsymbol{M}}_{1t}-\boldsymbol{M}_{1t}\|^2+\| \widehat{\boldsymbol{M}}_{2t}-\boldsymbol{M}_{2t}\|^2\\&+\beta(\| \widehat{\boldsymbol{M}}_{1t}^2+\widehat{\boldsymbol{M}}_{2t}^2-\mathbf{1}\|^2))\end{aligned}\tag{8}$$

When the mixed speech comes from two pure speech signals, the target IRM vector of each source signal in our research can be expressed as

$$\begin{cases}\boldsymbol{M}_{1t}=\sqrt{\dfrac{\boldsymbol{S}_{1t}{}^2}{\boldsymbol{S}_{1t}{}^2+\boldsymbol{S}_{2t}{}^2+\varepsilon}}\\ \boldsymbol{M}_{2t}=\sqrt{\dfrac{\boldsymbol{S}_{2t}{}^2}{\boldsymbol{S}_{1t}{}^2+\boldsymbol{S}_{2t}{}^2+\varepsilon}}\end{cases}\tag{9}$$

From (9), we can see that at the t-th frame, sum of squares of $\boldsymbol{M}_{1t}$ and $\boldsymbol{M}_{2t}$ is equal to $\mathbf{1}$. Therefore, in an ideal situation, the sum of squares of $\widehat{\boldsymbol{M}}_{1t}$ and $\widehat{\boldsymbol{M}}_{2t}$ should be as close as possible to $\mathbf{1}$. In other words, when the network output is sufficiently accurate, $\boldsymbol{M}_{1t}{}^2+\boldsymbol{M}_{2t}{}^2\approx\mathbf{1}$. The first two terms in (8) are designed to reduce the approximation error between $\widehat{\boldsymbol{M}}_{1t}$ and $\boldsymbol{M}_{1t}$, $\widehat{\boldsymbol{M}}_{2t}$ and $\boldsymbol{M}_{2t}$, which is not sufficient to train deep neural networks in the case of separating multiple speech signals simultaneously. Adding this constraint item to the traditional loss function can mine the joint information between dual-output masks and limit the output value of network to a certain range so that the predicted value closer to the actual value which can effectively make up for above shortcoming. Similar with (6), β is also a constraint factor and its value range is from 0 to 1. It can be adjusted to achieve different constraints.

3. Joint constraint of IRM and magnitude sum

Inspired by the joint-constraint loss functions proposed above, we note that there is also a relationship between the magnitude spectrogram of the target signal and the magnitude spectrogram of the mixed signal when the mixed speech signal is mixed by two clean speech signals. Considering the connection, we propose the third joint-constraint loss function, which not only constrains the predicted mask errors, but also makes the sum of the estimated magnitude spectrogram close to the mixed magnitude spectrogram to improve the

performance of model. The third joint-constraint loss function proposed in this paper can be written as

$$\begin{aligned}\mathrm{JC}_3 =&\frac{1}{2T}\sum_{t=1}^{T}(\| \widehat{\boldsymbol{M}}_{1t}-\boldsymbol{M}_{1t}\|^2+\| \widehat{\boldsymbol{M}}_{2t}-\boldsymbol{M}_{2t}\|^2\\&+\gamma(\| \widehat{\boldsymbol{M}}_{1t}\odot\boldsymbol{Y}(t)+\widehat{\boldsymbol{M}}_{2t}\odot\boldsymbol{Y}(t)-\boldsymbol{Y}(t)\|^2))\end{aligned} \tag{10}$$

In our research, the mixed speech is generated by mixing two pure speeches. The relationship of their magnitude spectrogram can be denoted as following equation:

$$\boldsymbol{S}_{1t}+\boldsymbol{S}_{2t}=\boldsymbol{Y}(t) \tag{11}$$

According to (7) we can obtain that under ideal circumstances, $\widehat{\boldsymbol{S}}_{1t}$ is as close to $\widehat{\boldsymbol{S}}_{1t}$ as possible, and $\widehat{\boldsymbol{S}}_{2t}$ is as close to $\widehat{\boldsymbol{S}}_{2t}$ as possible. Therefore, the error between the sum of estimated target magnitude spectrogram and the mixed magnitude spectrogram is close to $\mathbf{0}$. In other words, if the estimated IRMs are accurate enough, there is $\widehat{\boldsymbol{S}}_{1t}+\widehat{\boldsymbol{S}}_{2t}\approx\boldsymbol{Y}(t)$. Compared with the traditional loss function, the constraint term in (10) can decrease the error between the sum of the separated magnitude spectrogram of the two speakers and the mixed magnitude spectrogram. From this angle, this method can improve the accuracy of the separation network and get more distortion-free speech. Similar with α and β, γ is also a weighting factor used to adjust the constraint of the item and its value range is from 0 to 1.

4. Multiple joint constraint

In order to maximize the accuracy of the separated speech, we propose a comprehensive joint-constraint loss function to train the network. This integrated loss function considers the joint relationship between the masks, the relationship between the mask and the magnitude spectrogram and the connection of between the target magnitude spectrogram and the mixed magnitude spectrogram. It makes full use of the advantages of the loss function proposed above and can train a more accurate DNN model so that the predicted value closer to the ideal value. The multiple joint-constraint loss function for monaural speech separation based on dual-output DNN can be defined as follows:

$$\begin{aligned}\mathrm{JC}_4 =&\frac{1}{2T}\sum_{t=1}^{T}(\| \widehat{\boldsymbol{M}}_{1t}-\boldsymbol{M}_{1t}\|^2+\| \widehat{\boldsymbol{M}}_{2t}-\boldsymbol{M}_{2t}\|^2\\&+\alpha(\| \widehat{\boldsymbol{M}}_{1t}\odot\boldsymbol{Y}(t)-\boldsymbol{S}_{1t}\|^2+\| \widehat{\boldsymbol{M}}_{2t}\odot\boldsymbol{Y}(t)-\boldsymbol{S}_{2t}\|^2)\\&+\beta(\| \widehat{\boldsymbol{M}}_{1t}^2+\widehat{\boldsymbol{M}}_{2t}^2-\mathbf{1}\|^2)\\&+\gamma(\| \widehat{\boldsymbol{M}}_{1t}\odot\boldsymbol{Y}(t)+\widehat{\boldsymbol{M}}_{2t}\odot\boldsymbol{Y}(t)-\boldsymbol{Y}(t)\|^2))\end{aligned} \tag{12}$$

where α, β and γ are the regularization coefficients used to characterize the constraint ability of the relation term. They all range from 0 to 1. When $\alpha=0$, $\beta=0$, and $\gamma=0$, the constraint strength of the joint constraint is 0, and (12) is equivalent to (4). The regularization coefficients α, β and γ can be selected according to the experimental results, which are usually used in previous studies. However, this way has a large amount of calculation, and the selected value is not necessarily optimal. Therefore, in order to make the weight coefficients more accurate and the binding force of the relation term more appropriate, we apply the optimization algorithm to solve the regularization coefficients. From (12), we can split it into L_1, L_2 and L_3 as following formulas:

$$\begin{cases}L_1=\dfrac{1}{2T}\displaystyle\sum_{t=1}^{T}\| \widehat{\boldsymbol{M}}_{1t}\odot\boldsymbol{Y}(t)-\boldsymbol{S}_{1t}\|^2\\\qquad\quad+\| \widehat{\boldsymbol{M}}_{2t}\odot\boldsymbol{Y}(t)-\boldsymbol{S}_{2t}\|^2\\L_2=\dfrac{1}{2T}\displaystyle\sum_{t=1}^{T}\| \widehat{\boldsymbol{M}}_{1t}^2+\widehat{\boldsymbol{M}}_{2t}^2-\mathbf{1}\|^2\\L_3=\dfrac{1}{2T}\displaystyle\sum_{t=1}^{T}\| \widehat{\boldsymbol{M}}_{1t}\odot\boldsymbol{Y}(t)+\widehat{\boldsymbol{M}}_{2t}\odot\boldsymbol{Y}(t)-\boldsymbol{Y}(t)\|^2\end{cases} \tag{13}$$

Then the JC_4 method can be expressed as:

$$\mathrm{JC}_4=k_1\mathrm{Loss}_2+k_2L_1+k_3L_2+k_4L_3 \tag{14}$$

As can be seen from (14), the JC_4 method can be regarded as a combinatorial prediction method which integrates Loss_2, L_1, L_2 and L_3 prediction methods. The key to combinatorial prediction is how to properly determine the weighting coefficient of each single predictive method. In order to find the optimal weights of three coefficients α, β and γ in the JC_4 method, we did a series of deductions in the following.

The neural network model using any of Loss_2, L_1, L_2 and L_3 prediction method can be used to predict IRM value $\widehat{\boldsymbol{M}}_t=[\widehat{\boldsymbol{M}}_{1t},\widehat{\boldsymbol{M}}_{2t}]$, that is, to solve the same problem. The actual observation value can be denoted as $\boldsymbol{M}_t=[\boldsymbol{M}_{1t},\boldsymbol{M}_{2t}]$. The predicted value of the i-th method can be denoted as $\widehat{\boldsymbol{M}}_{it},(i=1,2,3,4)$ and the prediction error of the i-th method can be denoted as $\boldsymbol{e}_{it}=\boldsymbol{M}_t-\widehat{\boldsymbol{M}}_{it}$. The weighting coefficient of the i-th method can be denoted as k_i and constrained by $\sum_{i=1}^{P}k_i=1,(P=4)$. Then the predicted value of the combinatorial prediction method can be denoted as $\widehat{\boldsymbol{M}}_t=\sum_{i=1}^{P}k_i\widehat{\boldsymbol{M}}_{it}$. And the prediction error of the combinatorial prediction method can be denoted as $\boldsymbol{e}_t=\boldsymbol{M}_t-\widehat{\boldsymbol{M}}_t$. Therefore, we can get the following for-

mulas:

$$\begin{aligned} \boldsymbol{e}_t &= \boldsymbol{M}_t - \sum_{i=1}^{P} k_i \widehat{\boldsymbol{M}}_{it} = \sum_{i=1}^{P} k_i \boldsymbol{M}_{it} - k_i \widehat{\boldsymbol{M}}_{it} \\ &= \sum_{i=1}^{P} k_i (\boldsymbol{M}_{it} - \widehat{\boldsymbol{M}}_{it}) = \sum_{i=1}^{P} k_i \boldsymbol{e}_{it} \end{aligned} \tag{15}$$

The sum of squares of the prediction error of the combinatorial prediction method can be denoted as

$$\begin{aligned} J &= \sum_{t=1}^{N} \boldsymbol{e}_t^2 = \sum_{i=1}^{P} \sum_{j=1}^{P} k_i k_j \left(\sum_{t=1}^{N} e_{it} e_{jt} \right) \\ &= (k_1 \;\; k_2 \;\; \cdots \;\; k_P) \\ &\quad \cdot \begin{pmatrix} \sum_{t=1}^{N} e_{1t}^2 & \sum_{t=1}^{N} e_{1t} e_{2t} & \cdots & \sum_{t=1}^{N} e_{1t} e_{Pt} \\ \sum_{t=1}^{N} e_{2t} e_{1t} & \sum_{t=1}^{N} e_{2t}^2 & \cdots & \sum_{t=1}^{N} e_{2t} e_{Pt} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{t=1}^{N} e_{Pt} e_{1t} & \sum_{t=1}^{N} e_{Pt} e_{2t} & \cdots & \sum_{t=1}^{N} e_{Pt}^2 \end{pmatrix} \\ &\quad \cdot (k_1 \;\; k_2 \;\; \cdots \;\; k_P)^{\mathsf{T}} \end{aligned} \tag{16}$$

We denote the weighting coefficient vector of the combinatorial prediction method as $\boldsymbol{K} = [k_1, k_2, \ldots, k_P]^{\mathsf{T}}$ and the prediction error vector of the i-th method is denoted as $\boldsymbol{E}_i = [e_{i1}, e_{i2}, \ldots, e_{iN}]^{\mathsf{T}}$. Then the prediction error matrix can be expressed as $\boldsymbol{e} = [\boldsymbol{E}_1, \boldsymbol{E}_2, \ldots, \boldsymbol{E}_P]$. Vectors $\boldsymbol{E}_1, \boldsymbol{E}_2, \ldots, \boldsymbol{E}_P$ are independent of each other. The inverse matrix of is always exist. And there is $\boldsymbol{E}_{ij} = \boldsymbol{E}_{ji} = \boldsymbol{E}_i^{\mathsf{T}} \boldsymbol{E}_j = \sum_{t=1}^{N} e_{it} e_{jt}$. Then the sum of squares of the prediction error of the combinatorial prediction method can be expressed as

$$\begin{aligned} J &= \boldsymbol{e}^{\mathsf{T}} \boldsymbol{e} \\ &= (\boldsymbol{k}_P)^{\mathsf{T}} \begin{pmatrix} E_{11} & E_{12} & \cdots & E_{1P} \\ E_{21} & E_{22} & \cdots & E_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ E_{P1} & E_{P2} & \cdots & E_{PP} \end{pmatrix} (\boldsymbol{k}_P) \\ &= (\boldsymbol{k}_P)^{\mathsf{T}} \boldsymbol{E}_{(P)} (\boldsymbol{k}_P) \end{aligned} \tag{17}$$

where $\boldsymbol{E}_{(P)}$ is the error information matrix of the prediction, which reflects the prediction error information of the neural network model with different loss functions.

Denoting $\boldsymbol{R}_P = [1, 1, \ldots, 1]^{\mathsf{T}}$ as the unit array of $P \times 1$. Then constraint condition of weighting coefficients can be transformed from $\sum_{i=1}^{P} k_i = 1$ into $\boldsymbol{R}_P{}^{\mathsf{T}} \boldsymbol{K}_P = 1$. The optimal combinatorial prediction problem can therefore be expressed as mathematical nonlinear programming problem:

$$\begin{cases} \min J = \min \boldsymbol{e}^{\mathsf{T}} \boldsymbol{e} = \min \boldsymbol{K}_P{}^{\mathsf{T}} \boldsymbol{E}_{(P)} \boldsymbol{K}_P \\ \text{s.t. } \boldsymbol{R}_P{}^{\mathsf{T}} \boldsymbol{K}_P = 1 \end{cases} \tag{18}$$

If a weighting coefficient vector $\boldsymbol{K}_P$ brings the sum of squares of the prediction error of the combinatorial prediction method to a minimum, then the $\boldsymbol{K}_P$ is called the optimal weighting coefficient vector and the corresponding combinatorial method is called the optimal combinatorial prediction method.

With the Lagrange multiplier, the sum of squares of the prediction error of the combinatorial prediction method can be expressed as

$$J = \boldsymbol{K}_P{}^{\mathsf{T}} \boldsymbol{E}_{(P)} \boldsymbol{K}_P + \lambda (\boldsymbol{R}_P{}^{\mathsf{T}} \boldsymbol{K}_P - 1) \tag{19}$$

The necessary condition for J to achieve a minimum value is

$$\frac{\partial}{\partial \boldsymbol{K}_P} (\boldsymbol{K}_P{}^{\mathsf{T}} \boldsymbol{E}_{(P)} \boldsymbol{K}_P + \lambda (\boldsymbol{R}_P{}^{\mathsf{T}} \boldsymbol{K}_P - 1)) = 0 \tag{20}$$

In other words,

$$2 \boldsymbol{E}_{(P)} \boldsymbol{K}_P + \lambda \boldsymbol{R}_P = 0 \tag{21}$$

Multiply the two sides of (21) by $\boldsymbol{E}_{(P)}{}^{-1}$ to the left, there is

$$2 \boldsymbol{K}_P + \lambda \boldsymbol{E}_{(P)}{}^{-1} \boldsymbol{R}_P = 0 \tag{22}$$

And then multiply the two sides of (22) by $\boldsymbol{R}_P{}^{\mathsf{T}}$ to the left, there is

$$2 \boldsymbol{R}_P{}^{\mathsf{T}} \boldsymbol{K}_P + \lambda \boldsymbol{R}_P{}^{\mathsf{T}} \boldsymbol{E}_{(P)}{}^{-1} \boldsymbol{R}_P = 0 \tag{23}$$

Because

$$\boldsymbol{R}_p{}^{\mathsf{T}} \boldsymbol{K}_P = 1 \tag{24}$$

substituting (24) into (23), we can solve the Lagrange multiplier λ as

$$\lambda = -\frac{2}{\boldsymbol{R}_P{}^{\mathsf{T}} \boldsymbol{E}_{(P)}{}^{-1} \boldsymbol{R}_P} \tag{25}$$

And then substituting (25) into (22), we can solve the optimal weight vector as

$$\boldsymbol{K}^* = \frac{\boldsymbol{E}_{(P)}{}^{-1} \boldsymbol{R}_P}{\boldsymbol{R}_P{}^{\mathsf{T}} \boldsymbol{E}_{(P)}{}^{-1} \boldsymbol{R}_P} \tag{26}$$

where $*$ denotes the optimal solution symbol. Eventually, we can obtain the optimal weighting coefficient vector $\boldsymbol{K}^* = [k_1{}^*, k_2{}^*, k_3{}^*, k_4{}^*]^{\mathsf{T}}$ using (26). To be consistent with the form of (12), we can multiply $\boldsymbol{K}^*$ with

$\frac{1}{k_1{}^*}$. Then we can get the optimal weighting coefficients α, β and γ of the multiple joint-constraint loss function as follows:

$$[1,\alpha,\beta,\gamma]^{\mathsf{T}}=\left[1,\frac{k_2{}^*}{k_1{}^*},\frac{k_3{}^*}{k_1{}^*},\frac{k_4{}^*}{k_1{}^*}\right]^{\mathsf{T}} \tag{27}$$

The difference between (12) and the traditional dual-output loss function (4) is that three regular terms are added, which improves binding force from different angles to increases the effectiveness of the constraint on the estimated mask. The separation task is experimented between different gender combinations, and the constraint strength of the regularization term is affected by corresponding optimal weighting coefficients. By this method, the accuracy of the DNN separation model is improved, so that the recovered IRM of the separated source is closer to the actual output target, which makes the magnitude spectrogram of the separated signal closer to the magnitude spectrogram of the actual target source signal. It should be noted that all joint-constraint loss functions proposed in this paper are not limited to separating the speech mixed by two signals, and can be easily extended to the case of separating more than two source speech signals.

5. Monaural speech separation with joint constraint

The monaural speech separation based on dual-output DNN with joint constraint includes two stages: training and testing. The training stage aims to learn a DNN with more accurate network parameters based on the multiple joint-constraint cost function. The goal of the testing stage is to obtain the target speech signal from the mixed signal through a series of linear and nonlinear transformations using the trained DNN. We summarized the overall process for the monaural speech separation based on joint-constraint dual-output DNN in Fig.2.

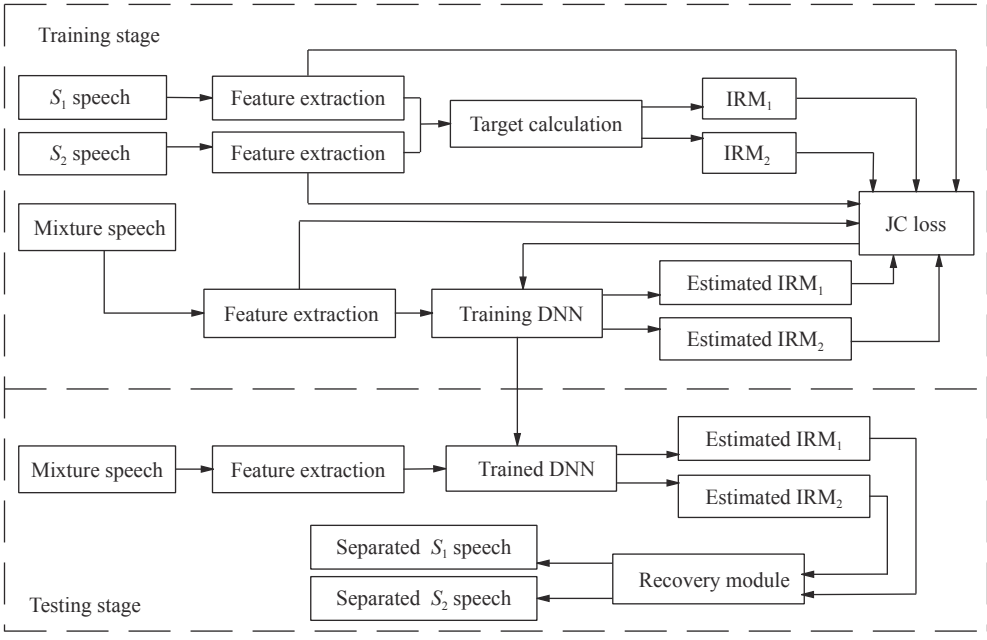


Fig. 2. Block diagram of proposed method.

In the training stage, we perform STFT on the mixed speech and two target speech signals to obtain corresponding magnitude spectrogram $\boldsymbol{Y}(t)$, $\boldsymbol{S}_{1t}$ and $\boldsymbol{S}_{2t}$, respectively. We can achieve corresponding training target $\boldsymbol{M}_{1t}$ and $\boldsymbol{M}_{2t}$ using (9). Then use the mixed magnitude spectrogram as the input of the neural network, to train the network under the guidance of the training target and the constraint of the joint loss function using (12). Finally, a trained network is gotten by adjusting the network parameters many times. In the testing stage, first, perform STFT on the test mixture speech to acquire the magnitude spectrogram feature. Then input this feature into the trained network to get the estimated mask of target speaker-1 and target speaker-2, respectively. Finally, we can obtain magnitude spectrogram of target speaker-1 speech and target speaker-2 speech using (7).

In the recovery module, using the estimated magnitude spectrum and the phase information of the mixed signal, we reconstruct the utterances of two target speakers via ISTFT.

IV. Experimental Results and Analysis

In this part, to evaluate the speech separation per-

formance of our proposed methods, many experiments are conducted on the GRID corpus [25]. First of all, the experimental dataset and the concrete experimental settings are introduced. Secondly, we describe the settings of hyperparameters in DNN architecture. Then we conduct experiments with all joint-constraint loss functions proposed above and compare their performance with that of traditional method and other previous loss-based separation methods.

1. Experimental configurations

1) Database

In our experiments, the training and the testing dataset are from the GRID corpus. There are 18 males and 16 females in the corpus, each person with 1000 clean utterances (each utterance last about one second). The utterances of two males and two females are randomly selected as the experimental data. The speech signals of every two speakers in the dataset are mixed, so that six gender combinations (F1+F2, F1+M1, F1+M2, F2+M1, F2+M2 and M1+M2) can be acquired. F and M stands for the female and male. To generate the training dataset, 700 utterances of each target speaker are randomly selected from GRID corpus, and then the utterances of two target speakers are added as the mixed utterances. As for the testing dataset, each person's 50 utterances are also randomly selected from the remaining utterances in the corpus. In other words, each person's test data and training data are different. The final experimental results are obtained by statistically average to get a general conclusion. All utterances used for experiments are down-sampled from 25 kHz to 16 kHz. The frame length is set to 512 samples and the frame shift is set to 256 samples. The 257-dimensional normalized magnitude spectrogram features obtained by STFT is used for the input of DNN model.

2) Hyperparameters settings of DNN

The dual-output DNN framework used in experiments is 257-1024-1024-1024-514, which means that there are 512 nodes in the input layer, 1024 nodes in the each of three hidden layers and 514 nodes in the output layer. Because the number of single-source output nodes is 257, the number of output layer nodes of DNN model with dual-output is 514 (257×2). For the input layer and all the hidden layers, a random dropout is applied to prevent overfitting and dropout rate is set to 0.2. For each hidden unit, we choose rectified linear unit (ReLU) as the activation function which can refrain the gradient vanishing problem during the training process. Because the two outputs of model are both range from 0 to 1, we choose sigmoid function as the activation function for output unit. The total amounts of training epoch is set to 50 and the mini-batch size is set to 128. The learning rate is initialized as 0.01. Stochastic gradient descent method is used for optimization.

3) Evaluation metrics

The separated speech signals for each approach are evaluated with five objective metrics, including the perceptual evaluation of speech quality (PESQ) [26], the short-time objective intelligibility (STOI) score [27], the signal-to-interference ratio (SIR), the signal to distortion ratio (SDR), and the sources-to-artifacts ratio (SAR) [28]. The PESQ algorithm provides a subjective MOS prediction for objective speech quality assessment by comparing the separated speech signal and the original reference speech signal. The score ranges from −0.5 to 4.5 and the higher score, the better speech quality. STOI measures objective intelligibility by calculating the short-time correlation of the spectral energy of the separated speech signal and the original reference speech signal. Its value ranges from 0 to 1 and the higher value means the better speech intelligibility. SDR measures the distortion of speech signal, SAR indicates the suppression of speech separation to the interference error of the system and SIR indicates the suppression of speech signal to the interference error of other speech signals. The unit of these three indicators is dB and their values are all positively correlated with the separation system performance.

2. Experiments on proposed algorithm

In this part, we conduct a series of experiments with the proposed speech separation methods based on dual-output DNN. First, we investigate the influence of regularization coefficients in the joint-constraint loss functions. Then we compare our proposed methods (noted as JC_1, JC_2, JC_3 and JC_4, respectively) with the dual-output DNN-based speech separation using the basic loss function based on masking (noted as Basic-IRM), mapping (noted as Basic-TMS) method and other previous method.

1) Influence of regularization parameter

First of all, in order to acquire the supreme results of our JC_1 algorithm, we research the influence of regularization coefficient on the speech separation performance of different gender combinations. Two males and two females are randomly selected from the GRID corpus, which can produce six different gender combinations. We average the results of the 4 cross-gender combinations as the result of the F-M. F-F and M-M represent the gender combination of F1+F2 and M1+M2, respectively. The regularization coefficient α is changed from 0 to 1, incrementing by 0.1 at a time. Experimental results of different regularization coefficients of JC_1 are shown in Fig.3(a). When α is equal to 0, the result represents the performance of the traditional loss function. It can be seen from the results that when α is less

than 0.5, the PESQ scores of three gender combinations increase in varying degrees with the increase of α. The PESQ score increase rapidly when the value of α is from 0 to 0.3, and increase slowly when the value of α is from 0.4 to 0.5. Due to the difference in the speech characteristics of different gender combinations, the value of optimal coefficient is different. For M-M separation and F-M separation, the separation system has the best performance when α is set to 0.5. For F-F case, the separation system has the optimal performance when α is set to 0.7. When α is larger than 0.7, the curve of speech intelligibility of the separated signal shows a downward trend with the increase of α, which means that the error of the estimated signal and the original reference signal increases. In addition, the separation performance is better when α is set to 1 than when α is set to 0, which proves the effectiveness of the proposed joint-constraint algorithm. Therefore, for different gender combinations of experiments with JC_1, we choose different regularization coefficient. Separating M-M and F-M mixed speech signals, α is set to 0.5, and separating F-F mixed signal, α is set to 0.7. Similarly, we did the same experiments with JC_2 and JC_3, respectively. The effect of regularization coefficients of JC_2 and JC_3 on PESQ is shown in Fig.3(b) and (c), respectively. For M-M and F-M cases, the separation system has the optimal performance when β and γ is set to 0.5 in JC_2 and JC_3. For F-F case, the separation system has the optimal performance when β and γ is set to 0.6 in JC_2 and JC_3. In conclusion, separating M-M and F-M mixed speech signals based on DNN with JC_2 or JC_3, β or γ is set to 0.5, and separating mixed signal in F-F case, β or γ is set to 0.6. The same settings are used in the following comparative experiments.

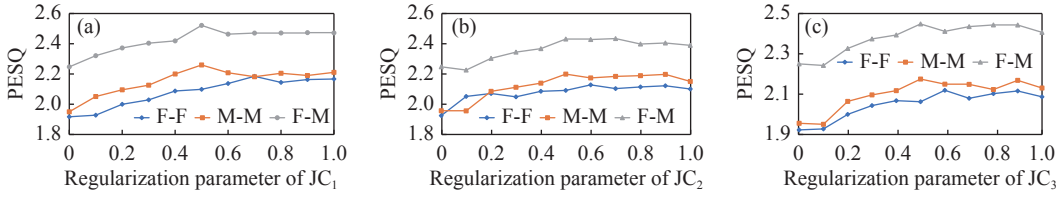


Fig. 3. Average separation performance with regularization parameter. (a) JC_1; (b) JC_2; (c) JC_3.

2) Separation performance of JC_1, JC_2 and JC_3

In this part, we analyze experimental results of speech separation based on dual-output DNN with JC_1, JC_2 and JC_3 by comparing with those of Basic-TMS and Basic-IRM. The PESQ scores, STOI, SAR, SDR and SIR values are tested. The experimental results are listed in Table 1. From the results, we obtain some findings. Firstly, the performance of JC_1, JC_2 and JC_3 all are better than that of traditional methods, which proves that the joint constraint of estimated value and speech signal feature from multiple aspects between different source signals is significant. Secondly, utilizing the relationship between mask and magnitude spectrogram has the greatest impact on the separation performance. This can be explained that we finally need to reconstruct the magnitude spectrogram of target speech signal for speech separation. Therefore, the combined constraint of mask and magnitude spectrogram in JC_1 can make the predicted value more accurate, so that the reconstructed speech signal is closer to the pure speech signal compared with the JC_2 and JC_3 methods. Thirdly, compared with TMS, the IRM way is more effective for speech separation, especially in solving the same gender combination problem. In addition, the proposed method can achieve better performance in the F-M separation cases. The system performance on cross-gender combination outperforms those on the same gender combination. It is because the same gender possesses high similarity the speech signals, which lead to the separation difficulty.

Table 1. Separation performance comparison of dual-output DNN with different methods

Method	Gender	PESQ	STOI(%)	SIR(dB)	SDR(dB)	SAR(dB)
Basic-TMS	F-F	1.922	78.50	5.7213	5.1264	7.9683
	M-M	2.132	81.32	6.2013	4.9532	7.7462
	F-M	2.164	83.32	9.7698	7.0325	9.1639
Basic-IRM	F-F	1.917	79.67	5.7624	4.9391	8.5022
	M-M	1.95	82.49	6.1779	4.9779	7.8573
	F-M	2.247	84.25	10.2349	7.2884	9.2441
JC_1	F-F	2.167	85.46	7.9618	5.5091	8.9258
	M-M	2.221	84.32	9.1738	6.3517	8.6679
	F-M	2.473	87.78	12.8367	8.3247	9.9247
JC_2	F-F	2.125	85.33	7.9016	5.5695	8.1933
	M-M	2.198	84.21	8.9157	5.6508	8.1669
	F-M	2.446	87.66	13.6825	7.2952	9.6075
JC_3	F-F	2.119	85.30	9.045	5.9041	7.6908
	M-M	2.171	84.28	9.994	6.3496	8.1946
	F-M	2.447	87.95	13.8387	8.6556	9.3879

Specifically, it can be observed from Table 1 that the JC_1 method improves 0.25 in PESQ, 5.79% in STOI, 2.2 dB in SIR, 0.57 dB in SDR, and 0.42 dB in SAR in the F-F separation compared with the Basic-IRM. In terms of M-M separation, the JC_1 method improves 0.27 in PESQ, 1.83% in STOI, 3 dB in SIR, 1.37 dB in SDR and 0.81 dB in SAR respectively compared

with the Basic-IRM. Additionally, in F-M case where 0.23 in PESQ, 3.53% in STOI, 2.6 dB in SIR, 1.04 dB in SDR and 0.68 dB in SAR are improved. Comparing with the Basic-TMS method, the performance of the JC_1 method keeps excellent. We can see from Table 1 that the JC_1 method improves 2.24 dB in SIR, 0.38 dB in SDR and 0.96 dB in SAR in the F-F separation compared with the Basic-TMS. In addition, the JC_1 method obtains 0.31, 3.07 dB, 1.29 dB, and 0.76 dB increment in PESQ, SIR SDR, and SAR in the separation of F-M case. For M-M separation, the JC_1 method is also better than Basic-TMS. Similarly, from the Table 1, we can obviously see that the JC_2 and JC_3 methods also obtain varying degrees of improvement on different gender-combinations compared with Basic-IRM and Basic-TMS methods, which indicates the validity of our proposed joint constraints.

3) Separation performance of JC_4

In this part, firstly, to obtain the optimal results of JC_4 algorithm, we research the effects of regularization coefficients α, β and γ of JC_4 on the speech separation performance. JC_4 is the integration of JC_1, JC_2 and JC_3. The regularization coefficients α, β and γ in JC_4 all range from 0 to 1. If two of them are 0, JC_4 degenerates to JC_1, JC_2 or JC_3. The JC_4 method is equivalent to the Basic-IRM method if all of them are 0. We evaluate the separated speech performance of JC_4 with different values of α, β and γ using PESQ scores, STOI and SIR. According to the analysis in the section of influence of regularization parameter, JC_1 has the optimal separation performance when α is nearby to 0.5 for different gender combinations. According to the analysis in the section of separation performance of JC_1, JC_2 and JC_3, JC_1 has better separation performance than JC_2 and JC_3.Therefore, the value of in JC_4 is larger than that of β and γ. In the following experiments, α is set to 0.5, and β and γ are changed from 0 to 0.5 with an increment of 0.1 to find more optimal value. Experimental results are listed in Table 2. From the table, we can see that the PESQ scores, STOI and SIR values are affected in varying degrees under different joint constraint. Specifically, with $\alpha = 0.5$ and $\gamma = 0$, the speech intelligibility is improved when β is from 0 to 0.4, and decreased when β is from 0.4 to 0.5. Therefore, α and β are set to 0.5 and 0.4, respectively. When γ is less than 0.2, the speech intelligibility is improved with the increase of γ. When γ is bigger than 0.2, the curve of speech intelligibility of the separated speech signal shows a downward trend with the increase of γ, which means that the error of the estimated value and the original reference value increases. The separation performance is better when α, β and γ are set to other value combinations than when α, β and γ are all set to 0, which proves the effectiveness of the proposed method. From the experimental results, we can see that the weighting coefficients has a great influence on speech separation performance. Therefore, it is significant to solve the optimal weighting coefficients to further improve the system performance.

Table 2. PESQ, STOI and SIR of JC_4 with various values of α, β and γ

Method	α	β	γ	PESQ	STOI(%)	SIR(dB)
JC_4	0	0	0	2.247	84.25	10.2349
	0.5	0.1	0	2.273	84.56	10.6283
	0.5	0.2	0	2.324	85.01	11.4047
	0.5	0.3	0	2.388	86.22	12.3751
	0.5	0.4	0	2.375	85.03	12.1061
	0.5	0.5	0	2.343	84.97	12.0911
	0.5	0.4	0.1	2.392	86.36	13.2109
	0.5	0.4	0.2	2.413	86.65	13.5892
	0.5	0.4	0.3	2.401	85.57	13.3911
	0.5	0.4	0.4	2.412	85.36	12.2692
	0.5	0.4	0.5	2.317	85.69	12.1532
	k_2'	k_3'	k_4'	**2.516**	**89.14**	**14.5208**

Secondly, to evaluate the effectiveness of the optimal weighting coefficient vector we solve based on the optimization idea, we compare the separated speech performance of the optimal weighting coefficient combination with the different weighting value combinations of α, β and γ using PESQ scores, STOI and SIR. Firstly, using the prediction results of the training set, we calculate the prediction error $\boldsymbol{E}_i$ at each frame of model that separately with Loss_2, L_1, L_2 and L_3 loss function. We further calculate the statistical average of the prediction error of all frames and then the prediction error information of four methods is spliced into the matrix $\boldsymbol{e}$. Then we solve the optimal weighting coefficient vector $\boldsymbol{K}^* = [{k_1}^*, {k_2}^*, {k_3}^*, {k_4}^*]$ using (26). The corresponding optimal weighting coefficient vector of the JC_4 can be obtained by (27), and it can be denoted as: $[\alpha, \beta, \gamma]^{\mathsf{T}} = [\frac{{k_2}^*}{{k_1}^*}, \frac{{k_3}^*}{{k_1}^*}, \frac{{k_4}^*}{{k_1}^*}]^{\mathsf{T}} = [{k_2}', {k_3}', {k_4}']^{\mathsf{T}}$. From the results in Table 2, we can obtain that applying optimal weighting coefficients combination method obtains the better separation performance. Specifically, the optimal weighting coefficient combination method obtains 0.27 increment in PESQ, 5% increment in STOI and 4.29 dB increment in SIR compared with the traditional method. And the optimal weighting coefficient combination method obtains 0.13 increment in PESQ, 3% increment in STOI and 2.15 dB increment in SIR compared with the case of $\alpha, \beta, \gamma = \{0.5, 0.3, 0\}$ and obtains 0.1 increment in PESQ, 2% increment in STOI and 0.93 dB increment in SIR compared with the case of $\alpha, \beta, \gamma = \{0.5, 0.3, 0\}$. Obviously, from the results we can see that the optimal weighting coefficient vector we solve based on the optimization idea using nonlinear

programming is more effective compared with the weighting value combinations selected empirically.

Thirdly, to evaluate the speech separation performance of dual-output DNN based on the multiple joint-constraint loss function, we conduct a series of experiments with JC_4 method on different gender combinations and compare the performance with the Basic-IRM, JC_1, JC_2, JC_3 and MaxDiffer [15]. Fig.4 summarizes SIR, SDR and SAR values of reconstructed speech by different separation methods for two-speaker mixed speech in the speech-independent situation. From the results shown in Fig.4(a), (b) and (c), we can see that the JC_4 is excellent in different gender combinations compared with other separation methods. For example, comparing with MaxDiffer method, JC_4 improves SIR, SDR and SAR 1.2 dB, 0.9 dB and 1.8 dB in separating mixed signal of F-M, and in the M-M separation, JC_4 obtains 1.5 dB, 3.6 dB and 0.6 dB increment in SIR, SDR and SAR. It can be seen from the results that speech intelligibility of the separated signal is significantly improved by the JC_4 method.

In addition, to visually see the separation performance of each algorithm, we randomly select the testing utterances waveforms of F-M separation pairs (which is more intuitional) to display the effect on the intelligibility and the quality of separated speech signals which are shown in Fig.5, where (a) and (b) are waveforms of the original reference speech in T-F domain, and (c) refers to the waveform of the mixture speech. (d) and (e) are estimated waveforms separated by Basic-IRM approach. (f) and (g) are estimated waveforms separated by JC_4 method. It is apparent that the shape of the speech waveforms recovered by JC_4 algorithm is the closest to the original reference signal.

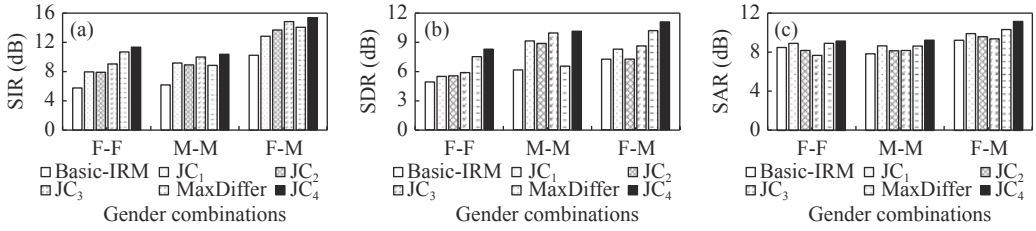


Fig. 4. Separation performance comparison for different approaches. (a) SIR; (b) SDR; (c) SAR.

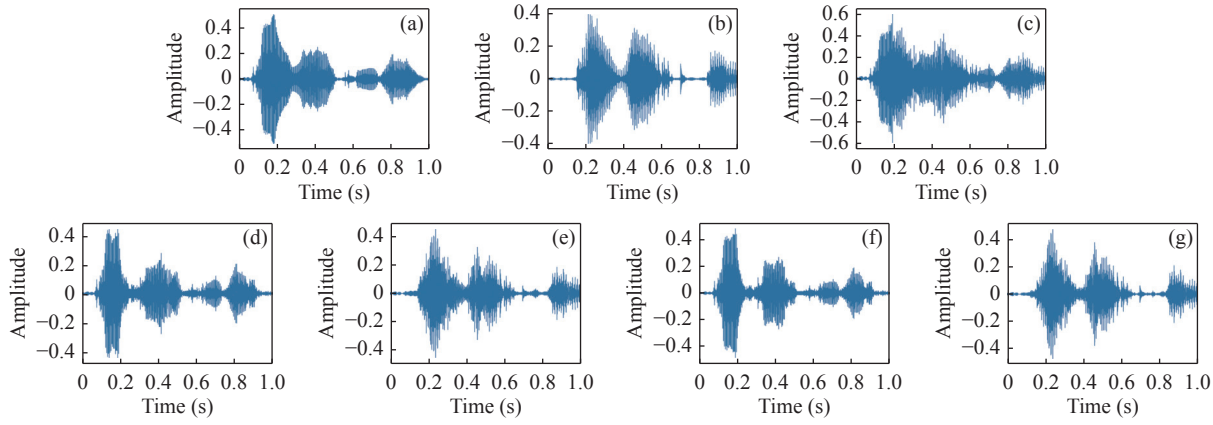


Fig. 5. Waveforms of speech. (a) Target (F); (b) Target (M); (c) Mixed (F+M); (d) and (e) Estimations of F and M separated by Basic-IRM; (f) and (g) Estimations of F and M separated by JC_4 method.

V. Conclusions

In this work, we propose three joint-constraint loss functions and incorporate them into a comprehensive joint-constraint loss function which is used to train the dual-output deep neural network for the monaural speech separation problem. The novel JC_4 loss function exploits the relationship between masks, the connection between masks and magnitude spectrogram and the relationship between target magnitude spectrogram and mixed magnitude spectrogram, so that it can estimate the corresponding output value more precisely. In addition, we solve the optimal weighting coefficients of the multiple joint-constraint loss function based on the optimization idea, which further improves the performance of the separation system. In order to verify the effectiveness of the proposed method, we compare it with the traditional separation method, JC_1, JC_2, JC_3 and other previous method using the dual-output DNN. The experimental results show that the separation performance of the proposed method obtained generally improvement compared with the traditional and other ad-

vanced methods. It also indicates that our proposed method for speech separation is more excellent in handling the different gender combination case. The separation performance in the same-gender mixed separation case is a little terrible comparing with the cross-gender case. So, our future work is to focus on the separation in the same-gender mixed separation case and design more effective and precise network to sort out the tricky problem. Besides, the method proposed in this paper is studied in the context of linear mixing of speaker's speech signal. But in the actual environment, the utterance is often mixed with reverberation and noise. It is worth studying to design a robust algorithm that can de-reverberation and de-noise simultaneously to further improve the separation system performance.

References

[1] J. Du, Y. H. Tu, Y. Xu, *et al.*, "Speech separation of a target speaker based on deep neural networks," in *Proceedings of the 2014 12th International Conference on Signal Processing*, Hangzhou, China, pp.473–477, 2014.

[2] A. Belouchrani, K. Abed-Meraim, J. F. Cardoso, *et al.*, "A blind source separation technique using second-order statistics," *IEEE Transactions on Signal Processing*, vol.45, no.2, pp.434–444, 1997.

[3] M. Yu, A. Rhuma, S. M. Naqvi, *et al.*, "A posture recognition-based fall detection system for monitoring an elderly person in a smart home environment," *IEEE Transactions on Information Technology in Biomedicine*, vol.16, no.6, pp.1274–1286, 2012.

[4] J. Du, Y. H. Tu, L. R. Dai, *et al.*, "A regression approach to single-channel speech separation via high-resolution deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol.24, no.8, pp.1424–1437, 2016.

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

[6] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed., Springer, Boston, MA, USA, pp.181–197, 2005.

[7] G. N. Hu and D. L. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol.18, no.8, pp.2067–2079, 2010.

[8] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Networks*, vol.13, no.4-5, pp.411–430, 2000.

[9] X. L. Zhang and D. L. Wang, "A deep ensemble learning method for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol.24, no.5, pp.967–977, 2016.

[10] Y. X. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol.22, no.12, pp.1849–1858, 2014.

[11] H. Erdogan, J. R. Hershey, S. Watanabe, *et al.*, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proceedings of 2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, South Brisbane, QLD, Australia, pp.708–712, 2015.

[12] D. S. Williamson, Y. X. Wang, and D. L. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol.24, no.3, pp.483–492, 2016.

[13] G. Kim, Y. Lu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *The Journal of the Acoustical Society of America*, vol.126, no.3, pp.1486–1494, 2009.

[14] T. May and T. Dau, "Computational speech segregation based on an auditory-inspired modulation analysis," *The Journal of the Acoustical Society of America*, vol.136, no.6, pp.3350–3359, 2014.

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

[16] Y. N. Wang, J. Du, L. R. Dai, *et al.*, "A gender mixture detection approach to unsupervised single-channel speech separation based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol.25, no.7, pp.1535–1546, 2017.

[17] M. S. Chauhan, R. Mishra, M. I. Patel, *et al.*, "Speech recognition and separation system using deep learning," in *Proceedings of 2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems*, Chennai, India, pp.1–5, 2021.

[18] S. Wan, "Research on speech separation and recognition algorithm based on deep learning," in *Proceedings of 2021 IEEE International Conference on Power, Intelligent Computing and Systems*, Shenyang, China, pp.722–725, 2021.

[19] C. H. Fan, J. H. Tao, B. Liu, *et al.*, "End-to-end post-filter for speech separation with deep attention fusion features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol.28, pp.1303–1314, 2020.

[20] N. J. Zheng and X. L. Zhang, "Phase-aware speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol.27, no.1, pp.63–76, 2019.

[21] T. G. Kang, J. W. Shin, and N. S. Kim, "DNN-based monaural speech enhancement with temporal and spectral variations equalization," *Digital Signal Processing*, vol.74, pp.102–110, 2018.

[22] G. Naithani, J. Nikunen, and L. Bramsløw, *et al.*, "Deep neural network based speech separation optimizing an objective estimator of intelligibility for low latency applications," in *Proceedings of the 2018 16th International Workshop on Acoustic Signal Enhancement*, Tokyo, Japan, pp.386–390, 2018.

[23] L. H. Sun, G. Zhu, and P. A. Li, "Joint constraint algorithm based on deep neural network with dual outputs for single-channel speech separation," *Signal, Image and Video Processing*, vol.14, no.7, pp.1387–1395, 2020.

[24] M. Delfarah and D. L. Wang, "Features for masking-based monaural speech separation in reverberant conditions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol.25, no.5, pp.1085–1094, 2017.

[25] M. Cooke, J. Barker, S. Cunningham, *et al.*, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of Amer-*

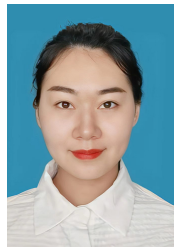
ica, vol.120, no.5, pp.2421–2424, 2006.

[26] A. W. Rix, J. G. Beerends, and M. P. Hollier, *et al.*, "Perceptual evaluation of speech quality (PESQ) - A new method for speech quality assessment of telephone networks and codecs," in *Proceedings of 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, UT, USA, pp.749–752, 2001.

[27] C. H. Taal, R. C. Hendriks, R. Heusdens, *et al.*, " An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol.19, no.7, pp.2125–2136, 2011.

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

SUN Linhui is currently an Associate Professor at Nanjing University of Posts and Telecommunications, China. She received the B.S. degree from Jilin University, China, in 2002, and received the M.S. and Ph.D. degrees both from Nanjing University of Posts and Telecommunications, China, in 2005 and 2013, respectively. Her research interests mainly include speech signal processing and modern speech communication. (Email: sunlh@njupt.edu.cn)



LIANG Wenqing is currently pursuing the M.S. degree at Nanjing University of Posts and Telecommunications, China. She received the B.S. degree from Shandong Technology and Business University, China, in 2020. Her major research interests include speech processing and modern speech communication. (Email: 1020010511@njupt.edu.cn)

ZHANG Meng is currently pursuing the M.S. degree at Nanjing University of Posts and Telecommunications, China. She received the B.S. degree from Xi'an University of Posts and Telecommunications, China, in 2020. Her major research interests include speech processing and modern speech communication. (Email: 1020010619@njupt.edu.cn)

LI Ping'an is currently a Senior Engineer at Nanjing University of Posts and Telecommunications, China. He received the B.S. degree from Nanjing University of Aeronautics and Astronautics, China, in 2000, and received the M.S. degree from Nanjing University of Posts and Telecommunications, China, in 2006. His research interests include signal processing. (Email: lpa@njupt.edu.cn)