ISSN 2616-5775 Vol. 6, Issue 13: 107-114, DOI: 10.25236/AJCIS.2023.061316

Research on Speech Separation Method Based on Deep Neural Network

Yandi Luo[1,*], Ming Fang[1]

[1]*Shanghai Aerospace Electronic Technology Institute, Shanghai, China*
[*]*Corresponding author: yandiluo@qq.com*

Abstract: *This paper focuses on the single-channel speech separation problem and uses deep neural network technology to deeply analyze the loss function, feature extraction and quality evaluation indexes. By improving the existing separation algorithms, a new method combining the jointly constrained loss function and the integrated optimizer is proposed. The study first examines the background and significance of research in the field of speech separation and outlines the current state of research. Then, the basics of speech separation technology and the underlying principles of deep neural networks are introduced in detail. The newly proposed system structure and quality evaluation metrics are used to compare the separation effects of different methods. In addition, the study improves the performance of the algorithm and enhances the ability of the model to avoid local optimums and improve the training efficiency through the proposed joint constrained loss function and integration optimizer.*

Keywords: *single-channel speech separation; loss function; integration optimizer; deep neural network*

1. Introduction

Voice communication, as a basic form of interpersonal communication, has been integrated into all aspects of daily life with the popularization of smart devices. Numerous technology companies such as Apple, Microsoft, Alibaba, and Baidu are constantly exploring and developing advanced speech technologies. However, the performance of these technologies is often limited in complex real-world environments, such as the interference of noise and reverberation. This highlights the importance of sound source localization, speech event detection, noise reduction and enhancement in improving speech recognition accuracy. Especially in multi-person mixed sound environments, such as bars and conference rooms, recognizing and separating individual sound sources becomes a challenge. This paper focuses on exploring deep learning-based speech separation techniques for single-channel speakers, aiming to improve the separation effect in noisy environments and provide new insights and solution strategies in the field of speech separation.

2. Speech separation system composition based on integration and optimization

2.1 Network model

This paper addresses the problem of blind separation of speech signals, in particular Auditory Scene Analysis (ASA), which involves separating and grouping signals from mixed sounds from the same source. The processing of speech signals includes analog-to-digital conversion, feature extraction, and pre-emphasis processing [1]. Mel Frequency Cepstral Coefficent (MFC) is an important feature extraction method, which is based on the human ear's nonlinear perception of frequency and more closely resembles human hearing [2]. In addition, the frame-wise plus window processing and Fast Fourier Transform (FFT) of the signal are used to extract the energy spectrum of each frame of the signal, which is further converted to the Mel domain processing, and finally the Mel Cepstral Coefficients are obtained by the Discrete Cosine Transform. These processes demonstrate the potential of deep learning in the field of speech processing, and this paper aims to improve the effectiveness of single-channel speech separation and promote the development of speech processing technology through these techniques.

Compared with traditional speech separation algorithms, deep neural networks (dons) can not only learn the deep characteristics of speech, but also establish the nonlinear relationship between input and output, thus making dons an effective tool for solving the single-channel speech separation problem and significantly improving the quality of speech separation. The speech separation task is a process of

ISSN 2616-5775 Vol. 6, Issue 13: 107-114, DOI: 10.25236/AJCIS.2023.061316

predicting the ideal floating-value mask by performing linear regression operations through deep learning. In this process, the minimum mean square error is usually used as the main criterion for the error metric. However, traditional algorithms often fail to adequately consider the amplitude relationship of the speech signal when estimating the minimum mean square error between the network output and the ideal floating-value mask. This leads to the possibility that separated speech may be interspersed with other speech components, which reduces the intelligibility and clarity of the speech [3].

When training deep neural networks (dons), accurate prediction of the true amplitude spectrum of the target speech is required in addition to the predicted ideal floating-value mask. The goal is to make the ideal floating-value mask prediction as accurate as possible, which requires digging deeper into the correlation between the signal amplitude spectrum and the ideal floating-value mask to improve the overall accuracy of the separated speech.

To improve clarity and reduce the distortion of separated speech, a new loss function based on joint constraints is proposed in this paper.

$$\text{Loss}_2 = \frac{1}{2T}\sum_{i=1}^{T}\left(||\hat{M}_i - M_i||^2 + \alpha||\hat{M}_i \odot Y_i - S_i||^2\right) \tag{1}$$

This method not only reduces the error between the prediction result and the ideal floating-value masking value, but also minimizes the difference between the amplitude of the generated signal and the amplitude of the actual separated signal. Such an approach makes the reconstructed separate sound closer to the original pure sound, thus effectively improving the quality of speech separation.

Gradient descent algorithms play a key role here as commonly used optimizers. Traditional stochastic gradient descent and adaptive moment estimation optimizers, although widely used, are sensitive to changes in the learning rate and are prone to falling into local optimal solutions. To address these limitations, this paper proposes a novel mono speech separation algorithm based on integrated optimization. The algorithm combines R adaptive moment estimation, which improves the variance problem of adaptive moment estimation by dynamically adjusting the learning rate, and Lookahead algorithm, which optimizes the updating process of "fast weights" and "slow weights" to improve the stability and convergence speed of the learning rate. Stability and convergence speed are shown in Figure 1 [4].

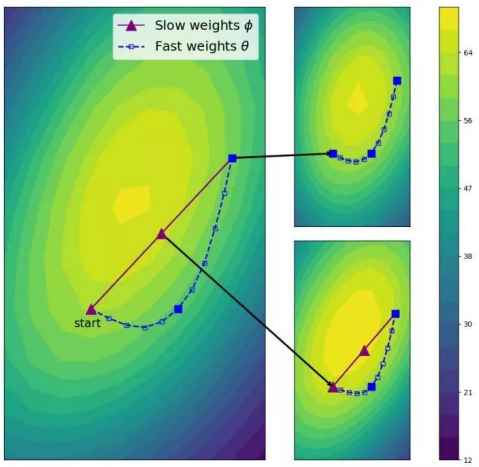

Figure 1: Lookahead optimizer effect diagram

2.2 Separation Algorithm Based on Jointly Constrained Loss Function and Integrated Optimizer

On this basis, a new speech separation method is proposed, which can be divided into three stages: training, separation, and speech reconstruction. In the training process, the features of the input signal are fed into the DNN, and then the optimizer is used to constrain the loss function and correct the weights and biases of each layer, which finally forms a complete DNN speech separation system. In the separation stage, the trained DNN is used for processing. In the speech reconstruction stage, the predicted hybrid speech amplitude spectrum is derived, and then the predicted speech is transformed using short-time Fourier inverse transform to obtain the time-domain signal as in Eq. 2:

ISSN 2616-5775 Vol. 6, Issue 13: 107-114, DOI: 10.25236/AJCIS.2023.061316

$$\hat{S}_1 = \hat{M}_1 \odot Y \tag{2}$$

(1) Training phase:

Step 1: Preprocess the source and mixed signals, then perform a short-time Fourier transform, which is normalized to obtain the amplitude spectrum of the source signal $S_{1\,\mathrm{train}}$ and the amplitude spectrum of the mixed signal Y_{train} .

Step 2: Find the ideal floating value mask IRM .

Step 3: The input feature of the DNN is Y_{train} and take IRM as the output target of the DNN.

Step 4: In the forward propagation phase, the nodes at each level are initialized with weighting and bias.

Step 5: In the backward propagation stage, an optimization algorithm is used to minimize the loss function, and the neurons in each layer are iterated for weighting and bias optimization.

Step 6: Obtain the trained DNN model.

(2) Separation stage:

Step 1: The tested hybrid speech is preprocessed, and then short-time Fourier transformed and normalized to obtain the amplitude spectrum of the hybrid speech Y_{test} .

Step 2: Get the estimated target for the output of the DNN $\hat{M}_1$.

(3) Speech reconstruction stage:

Step 1: Based on Eq. 2, the target speech amplitude spectrum is estimated as $\hat{S}_1$ using $\hat{M}_1$ and the amplitude spectrum of the hybrid speech Y_{test} .

Step 2: The estimated target speech amplitude spectrum is analyzed against the previously extracted phase spectrum of the mixed signal to produce the estimated target signal spectrum.

Step 3: Processing of the target spectrum using short-time Fourier inversion to produce a time-domain signal.

3. Speech separation system experiment and effect analysis

3.1 Parameter design and data set

The experimental environment for this study consisted of an Intel(R) Xeon(R) CPU E5-2680v4 processor, an NVIDIA TITAN Xp graphics card, 16 GB of RAM, and Python 3.8 and Pytorch 1.5.1 software running in the Anaconda virtual environment and Pycharm IDE. The dataset is derived from the GRID corpus, which covers the English statements of 34 speakers, and 2,000 voices of 4 (2 male and 2 female) of these speakers are selected as the test data for the experiment.The DNN model consists of 512 neurons in each of the input and output layers, and 1,024 neurons in each of the 3 hidden layers, and the training period is set to 100, with an initial learning rate of 0.0001, and the last hidden layer adopts the Sigmoid activation function was used for the last hidden layer, and a dropout rate of 0.2 was set to prevent overfitting. The regularization factor λ was adjusted according to different gender combinations, and it was found that male-female, male-male, and female-female combinations were best separated at λ = 0.5, λ = 0.5 or 0.6, and λ = 0.7, respectively. This setting makes the DNN more efficient and accurate in managing the single-channel speech separation task [5].

In the experiments, different numbers of neurons were set up to explore the effect of the number of neurons in the hidden layer on the performance of the network, including 256, 512, 1024, 2048 and 4096.

ISSN 2616-5775 Vol. 6, Issue 13: 107-114, DOI: 10.25236/AJCIS.2023.061316

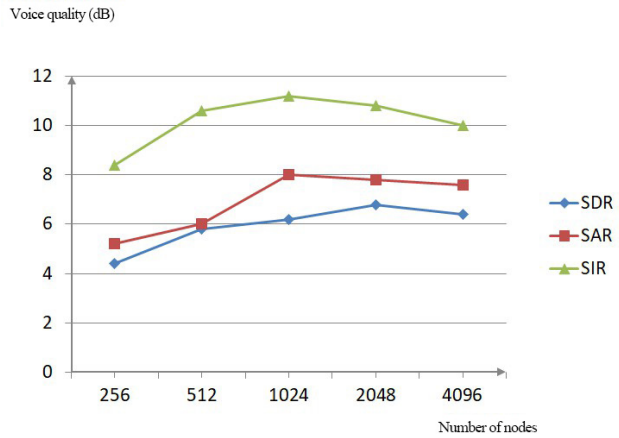

Figure 2: Separation performance as affected by the number of neurons.

As the results in Figure 2 show, the performance of separate speech first improves as the number of neurons increases, but the performance improvement is limited after the number of neurons reaches 1024. When the number of neurons is low, the network is under-trained due to the lack of feature information; however, an excessive number of neurons leads to learning some unnecessary interfering features and thus slight overfitting, especially when the number of nodes exceeds 1024, the SAR and SIR values start to decrease.The SDR value, which is a key indicator of the performance of separated speech, does not improve significantly after 1024 nodes and decreases after decreases at 4096 nodes. Considering the balance between performance and computation, it was finally decided to set the number of neurons in the hidden layer to 1024. This setting ensures the performance of the algorithm while avoiding unnecessary computational burden.

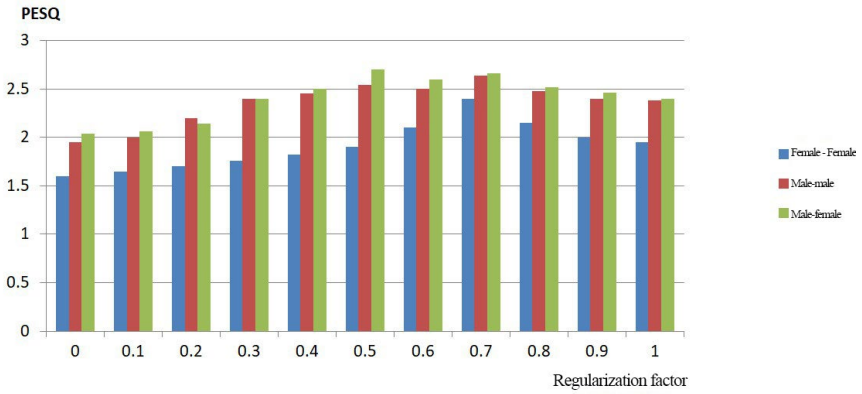

Figure 3: Comparison of regularization coefficient's effect on PESQ

As shown in Figure 3, the effect of the regularization coefficient (λ) in the range of 0 to 1 is explored for speech separation tasks with different gender combinations. The experiment covers three combinations of male-male, male-female, and female-female extracted from the GRID corpus, in which the regularization coefficient is adjusted by increasing 0.1 each time, and the PESQ value is used as the performance evaluation index. The results show that in the male-female combination, the PESQ value shows the largest increase as λ increases when λ does not exceed 0.5. In contrast, the optimal regularization coefficient for female-female and male-male combinations is 0.7, respectively. This reflects the differences in the regularization coefficient requirements for different gender combinations, guiding different regularization coefficient settings in the gender speech separation experiments: λ is set to 0.7 in the mixed separation experiments for female-female and male-male combinations, whereas λ is set to 0.5 for male-female combinations [6].

3.2 Comparison experiments with traditional algorithms

The loss function for the joint constraints in this method is first compared to the traditional loss function. From the thirty-four presenters, four presenters were selected, two males denoted as M1 and M2, and two females denoted as F1 and F2.These presenters can be grouped into six different combinations: F1-F2, F1-M1, F1-M2, F2-M1, F2-M2, and M1-M2.The experimental comparisons were

ISSN 2616-5775 Vol. 6, Issue 13: 107-114, DOI: 10.25236/AJCIS.2023.061316

made by randomly choosing 50 speeches as test speeches, and comparing them, and averaging the obtained The results were averaged to compare the four metrics, as shown in Figures 4 through 7.

Four speech effect metrics, SDR, SAR, SIR, and PESQ, are higher than those of the DNN based on the conventional loss function. This suggests that the floating-value masking can be better constrained when computing the relationship between the amplitude spectrum of the output signal. In the female-female type mixing experiments, the separation effect of the method increases by 0.69 dB, 1.43 dB, 0.71 dB, and 0.42 in SDR, SAR, SIR, and PESQ, respectively. The results show that regardless of the method used for female-female mixing, the separation effect is not as good as that of the results of other types of mixing, and compared with the conventional method, the method is not significantly more effective. This is since both female voices are extremely high and close in frequency, so their features are remarkably similar and mixing them together will increase the difficulty of separation.

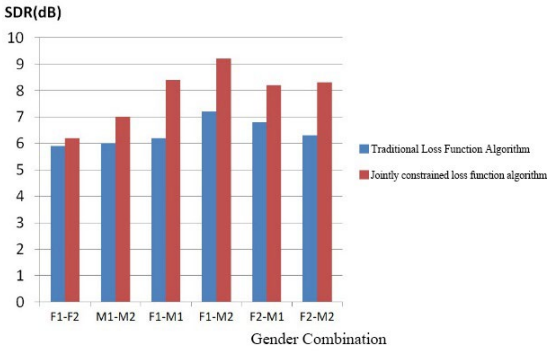

Figure 4: Comparison of SDR separation metrics with conventional algorithms

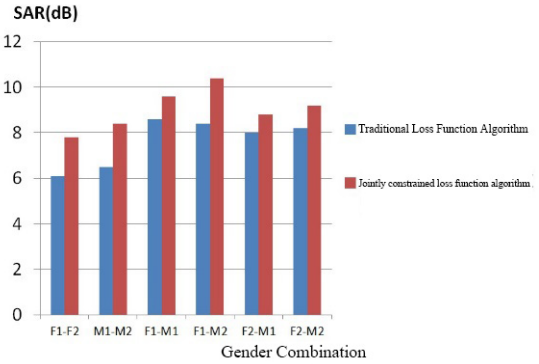

Figure 5: Comparison of SAR separation metrics with conventional algorithms

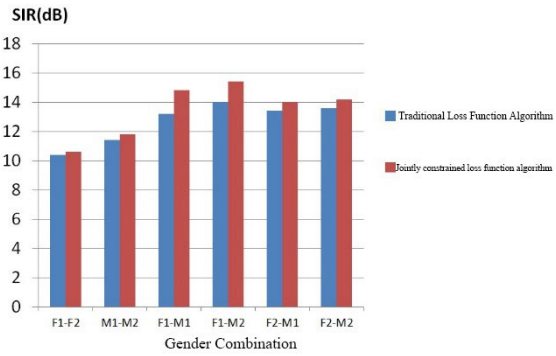

Figure 6: Comparison of SIR separation metrics with conventional algorithms

ISSN 2616-5775 Vol. 6, Issue 13: 107-114, DOI: 10.25236/AJCIS.2023.061316

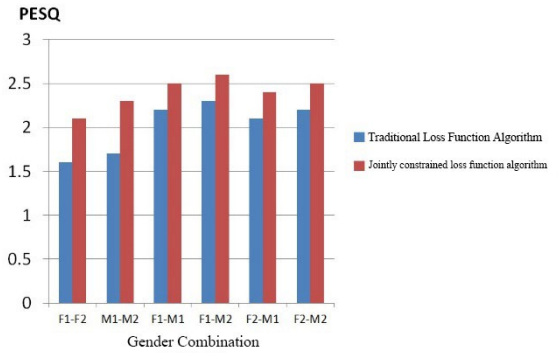

Figure 7: Comparison of PESQ separation metrics with conventional algorithms

Because of this, during the male-female separation experiments, the signals are very different, and the characteristics of each speech are very different, so the proposed method has a great improvement in performance compared to the traditional method, for example, in the separation experiment of F1-M2, the four metrics, SDR, SAR, SIR, and PESQ, are greatly improved, which are respectively increased by 1.89 dB, 1.38 dB, 2.32 dB, and 0.49. This is since the difference between F1 and M2 is greater than that between F1 and M1, and therefore, the results of speech separation are better. Although the performance of separation will not be improved in the case of male-male mixture, it is always better than the female-female group. The results show that the difference in gender influences the variability of the experimental separation results, but overall, the method is superior to the traditional separation method. Through the study of the method, it is proved that the method can improve the degree of intelligibility and clarity of the language. The results show that the separated speech obtained by the method proposed in this paper has a significant similarity in appearance to the pure original signal, as well as a reduction in bursts, indicating a better loss function under joint constraints.

Using radar as a built-in optimizer for lookahead, it not only accelerates the convergence, but also shortens the learning cycle of the separation model and effectively prevents the loss function from converging towards local optimality, improving the accuracy of the separation algorithm. The ideal floating-value mask is treated as the object of training, and it is evaluated using various gradient algorithms such as SGD, Adam, radar, and so on. The inputs, training set, validation set, and test set of the network are the same as set up above. All the gradient algorithms except SGD algorithms are admirably adapted. In SGD, the learning rate is set to 0.1 for the first ten epochs, and each subsequent cycle epoch is reduced by 10%. Momentum was set to 0.5 for the first five epochs and 0.9 for the remaining epochs. Table 4.1 shows the results of the comparative experiment with four metrics and the results of the separation task for the male-female group were averaged for ease of representation [7].

Table 1: Comparison of separation performance under different gradient algorithms

Optimizer	Gender mix	SDR/dB	SAR/dB	SIR/dB	PESQ
SGD	Female - Female	5.23	5.76	8.22	1.76
	M-F	6.74	6.98	10.95	2.01
	M-M	5.55	6.58	9.36	1.86
Adam	Female - Female	5.35	6.61	10.95	1.89
	M-F	7.75	8.53	11.61	2.18
	M-M	6.62	7.68	10.65	2.03
Radam	Female - Female	5.46	6.88	10.86	1.99
	M-F	8.01	8.86	12.45	2.24
	M-M	6.68	7.93	10.76	2.16
Integration Optimizer	Female - Female	6.54	7.67	11.27	2.11
	M-F	8.67	9.64	14.58	2.51
	M-M	7.16	8.64	11.95	2.22

From Table 1, the separation results are better with the integration optimizer compared to the other optimizers. This is caused by the shortcomings of other optimization algorithms, for example, SGD will fall into a local optimum when the momentum parameter is too small, while the optimal solution will be missed when the momentum parameter is too large; Adam will cause some problems when the fast learning rate increases, such as the loss function fails to converge. The final separation results show that the average separation of the integrated optimizer is best under the male-female gender combination condition, with an increase in the SDR, SAR, SIR, and PESQ performance metrics by 1.93 dB, 2.66 dB,

ISSN 2616-5775 Vol. 6, Issue 13: 107-114, DOI: 10.25236/AJCIS.2023.061316

3.63 dB, and 0.5, respectively. In addition, even though random adds a rule restriction to Adam by modifying it, Adam and Adam are similar and only slightly better at speech separation. In addition, in the performance of speech separation, this paper also compares the training time of the network, the training period of the other models is more than 7 hours, but the training time of the integration optimizer is 1.5 hours shorter, which shows that the integration optimizer not only improves the accuracy, but also shortens the learning time of the model.

For the experimental comparison of different training targets, as shown in Figure 8, the ideal floating-value mask and the target magnitude spectrum are not overly superior in performance to the target magnitude spectrum although there is a difference between them, and the target magnitude spectrum has a higher SAR value than the ideal floating-value mask by 0.65 dB in the female-female separation test, but it is the distortion ratio of the SDR and the PESQ that are the most important measures of the system's performance. From the experimental point of view, the use of the ideal floating-value mask has a better separation effect than the target magnitude spectrum, the reason for this is that the ideal floating-value mask takes values in the range of 0~1 interval, while the target magnitude spectrum takes values in the range of the full-frequency domain, and therefore the prediction error of the ideal floating-value mask is smaller than that of the target magnitude spectrum. In addition, in the target magnitude spectrum experiment, the female-female combination has a weaker separation effect compared with the male-male combination, which indicates that the target magnitude spectrum is more sensitive to the effect role of gender. For this reason, this thesis improves the performance of deep neural network speech separation by optimizing its loss function based on the model of ideal floating-value masking and adopting the training method of joint constraints.

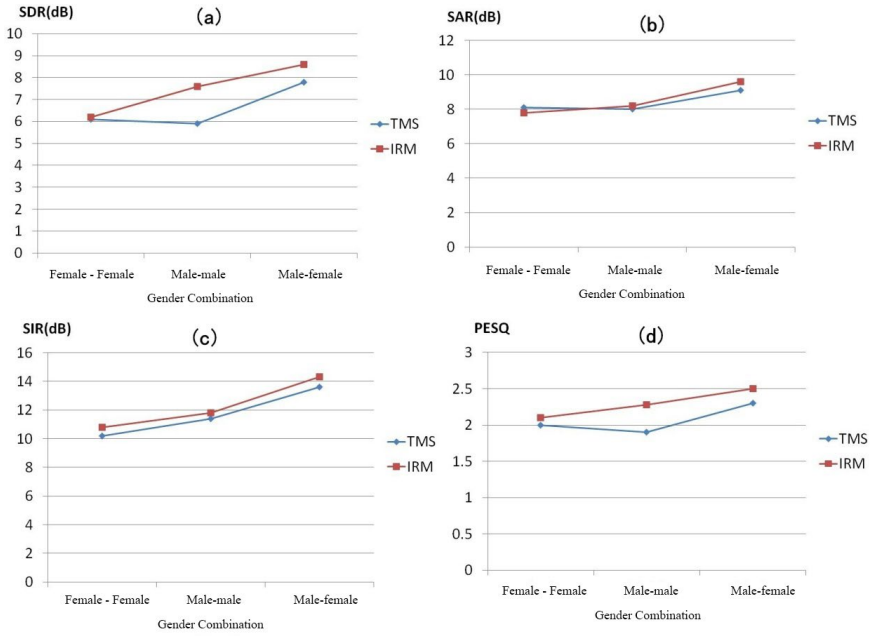

Figure 8: Comparison between different training objectives

4. Conclusions

In this paper, based on the traditional loss function algorithms and optimizers, a separation method based on a joint constrained loss function and an integrated optimizer is proposed, which introduces the basic principles of feature extraction and deep neural networks, including common network structures and basic features, and then outlines common algorithms and models, and proposes a system structure, and compares the separation effects of each method with quality evaluation indexes; since the traditional separation algorithm's loss function is only related to the mean square error between the pure signal and the separated signal, a joint constrained loss function is proposed, which integrates the relationship between the ideal floating-value mask, as well as the inclusion of an integrated optimizer, which can adaptively adjust the hyper-parameters to avoid falling into local optimums, and improve the efficiency of training.

In the experimental part, comparing the effects of the network layered structure and the number of layered neurons, the batch size and the regularization coefficient of the loss function on the separation

ISSN 2616-5775 Vol. 6, Issue 13: 107-114, DOI: 10.25236/AJCIS.2023.061316

performance, and selecting the appropriate hyperparameters, followed by comparing it with the traditional loss function, and then comparing it with the different optimizers, the joint constrained loss function and the integrated optimizer achieved a better separation effect and improved the separation efficiency.

Subsequently, the phase information of the signal can be considered and studied in the complex domain, which can further improve the separation effect and the subjective evaluation of the quality of the human ear hearing is even better. There is also the process of speech separation will require speech enhancement, noise removal and other processes, can be in-depth research in these areas, and speech separation system combined. Multi-channel speech separation is also a research hotspot, this paper only focuses on the single-channel speech separation problem, and then can try to enter the multi-channel speech separation problem.

References

[1] Feng Qibin. Research on single-channel hybrid speech separation algorithm based on computational auditory scene analysis[D]. Taiyuan University of Technology, 2019.
[2] Zhenlei Li, Fei Yang, Na Li et al. Response characteristics of acoustic emission Meier inverse spectral coefficients for loaded rupture of prefabricated cracked sandstone[J]. Journal of China University of Mining and Technology,2023,52(04):713-726.DOI:10.13247/j.cnki.jcumt.20220667
[3] Qinghua Wang ,Jiyun Sun, Jianhua Hu et al. Prediction of adjustment parameters of slant mill piercing machine based on deep neural network[J/OL]. Forging Technology, 2023,(11):73-78+103 [2023-11-30] https://doi.org/10.13330/j.issn.1000-3940.2023.11.012.
[4] Tian Le, Cao Langcai. DMU improvement algorithm for interactive dynamic influence diagram based on lookahead [J]. Systems Engineering and Electronics, 2014, 36(06):1201-1206.
[5] LIN Yujian, WEI Yunlong, CHEN Qiqi et al. A sigmoid function optimization method for embedded computing platform [J]. Small Microcomputer Systems, 2021, 42(10):2053-2058.
[6] SONG Daiyue, LI Kaizhuang, CHEN Qianqian. Optimization of regularization coefficients for ISAR high-resolution imaging algorithm[J]. Electronic Information Countermeasures Technology, 2023, 38(06): 68-75.
[7] YE Xiaowen, ZHANG Yin'e, ZHOU Qi. Image restoration method for generative adversarial networks based on improved reconstruction loss function [J/OL]. Journal of Gannan Normal University,1-6[2023-11-30]https://doi.org/10.13698/j.cnki.cn36-1346/c.2023.06.018.