Multiresolution Cochleagram Speech Enhancement Algorithm Using Improved Deep Neural Networks with Skip Connections

chaofeng lan (lanchaofeng@hrbust.edu.cn)
Harbin Industrial University: Harbin University of Science and Technology
https://orcid.org/0000-0003-4667-0635

Chundong Liu
Harbin University of Science and Technology

Lei Zhang
Blood Transfusion General Hospital of Heilongjiang Province Land Reclamation Bureau

Research

Keywords: Speech enhancement, Low SNR, Source-to-distortion ratio, MRCG

Posted Date: February 24th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-229829/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
Multiresolution Cochleagram Speech Enhancement Algorithm Using Improved Deep Neural Networks with Skip Connections
Chaofeng Lan*, Chundong Liu1, Lei Zhang2

Abstract
Deep learning based methods have been a recent benchmark method for speech enhancement. However, these approaches are limited in low signal-to-noise ratios (SNR) conditions, for speech loss and low intelligibility. To address this problem, we improve Multi-Resolution Cochleagram (MRCG), and gammachirp filter bank is used to decompose the speech signal in time and frequency, and the low-resolution signal is denoised by the minimum mean-square error short-time spectral amplitude estimator (MMSE-STSA). Improve Multi-Resolution Cochleagram (I-MRCG) is adopted as the input feature of Skip connections-DNN (Skip-DNN). In this paper, the source to distortion ratio (SDR) is used in the training process, and the logarithm is introduced to observe the iterative process more clearly. Experiments were performed on the TIMIT database with four noise types at four levels of SNR. I-MRCG as the input feature of the Skip-DNN model, the average PESQ is 2.6783, and the average STOI is 0.8752. Compared with MRCG, the PESQ and STOI obtained by MRCG are increased 1.4% and 1.5%, respectively. This shows that MRCG is the input feature of the Skip-DNN model, and the speech enhancement effect after training is better than other features. It can not only solve the problem of speech loss in a low SNR environment, but also obtain more robust speech enhancement. The loss function experiment shows that compared to MSE and SDR, the improved SDR as the loss function of the speech enhancement model has the best enhancement effect.

Keywords: Speech enhancement; Low SNR; Source-to-distortion ratio; MRCG

1 Introduction
Speech enhancement is an extensively studied research area that has many applications, such as, in telecommunication and human-computer interaction. In the field of telecommunication, the received speech signals are often corrupted by background noise, such as in a military field, where communication can be difficult due to the complex battlefield conditions, and the demand for high quality speech. In human computer interaction, some intelligent devices began to use speech instead of the keyboard as input, however, this requires high-performance speech recognition in a real world with noisy background sound environment. In these applications, noisy speech needs to be enhanced to improve their intelligibility and quality for more effective communications or interactions.

Speech enhancement can be performed in supervised or unsupervised manner. The spectral subtraction [1] and the Wiener filtering [2] are commonly used methods in unsupervised speech enhancement. A common problem encountered in these conventional methods is that in the enhanced speech often contains artefacts called “musical noise”. Ephraim and Malah [3] assumed that the noise was stationary Gaussian and used the algorithm based on minimum mean square error to enhance the noisy speech and reduce the level of “musical noise”.

*Correspondence: lanchaofeng@hrbust.edu.cn
Department of School of Electrical and Electronic Engineering, Harbin University of Science and Technology, Harbin 150080, P.R.China
However, most noise signals are non-stationary in real life. Martin [4] proposed a speech enhancement algorithm based on minimum statistics for non-stationary noise signals. Previous studies have shown that unsupervised speech enhancement has a good enhancement performance in environments with high SNR and stationary noise, but a poor performance in environments with low SNR and nonstationary noise [5-6].

Supervised algorithms have been developed to enhance speech corrupted by non-stationary noise, including shallow neural networks and deep neural networks [7-8]. Non-negative matrix speech enhancement algorithm in the shallow neural network, clean speech and noise are trained separately to obtain good speech enhancement performance [9-10]. Because the number of layers in the shallow neural network is small, the fitting of the test data set is not good and only simple features can be extracted, which leads to poor speech enhancement effect. Recently, DNN have been applied to speech enhancement [11-13]. DNN is utilized to train the time-frequency mask between clean speech and noise, which greatly improved the intelligibility of speech. The mapping relationship between noisy and clean speech spectra is learned by the DNN model. In addition, "dropout" is used to prevent overfitting, and the mini-batch stochastic gradient descent algorithm is used to accelerate the training speed [14-16]. The DNN model is utilized to predict the complex ratio masking and then predicted the amplitude and phase of the speech, which compensated the phase offset caused by noise [17]. Chen proposed to use multi-resolution cochleagram to obtain global and local features of speech and improve the speech enhancement performance in a low SNR [18]. To improve the generalization ability of the DNN model, Chen et al. added noise to the data during the training process to diversify noise and improve the speech enhancement performance [19]. Tu et al, used DNN to predict target and interference respectively, and the accuracy of the estimated target speech was significantly improved in speech recognition [20]. Tu and Zhang proposed a Skip-DNN model that maps from the Mel-frequency power spectra of noisy speech to that of clean speech, which helps resolve the problem of gradient vanishing and improve the performance of speech enhancement [21]. Tseng et al. utilized a sparse non-negative matrix factorization to extract speech features and utilized the DNN model to estimate the ideal binary mask (IBM), which improves the intelligibility of enhanced speech in a low SNR environment [22].

According to the above analysis, the cochlear features with multi-resolution can better reflect the speech signal. Compared with the Mel filter bank and the Gammatone filter bank, the Gammachirp filter bank is more consistent with the auditory feature of the ear, so the I-MRCG in the Gammachirp transform domain will be chosen to extract speech features. Therefore, to improve the speech enhancement performance, especially the speech enhancement effect under low SNR, this paper uses the MRCG feature as the Skip-DNN model input. The effects of Skip-DNN speech enhancement under different feature parameters and speech enhancement models will be compared and analyzed, the quality and intelligibility of the enhanced speech will be evaluated, and evaluate the impact of different loss functions on the training effect of the Skip-DNN based speech enhancement model.

2 Methods

2.1 Speech enhancement based on deep learning

There are often three main components in a deep learning based on speech enhancement algorithm: learning machines, training targets, and acoustic features [23-24]. In the process of feature extraction,
the time-domain speech signal is usually converted into the time-frequency domain, for example, log-power spectra, Mel-frequency cepstrum coefficient, Mel frequency power spectra, and MRCG. The selection of training targets has an important influence on training results. The training target usually uses clean speech power spectrum, IBM or ideal ratio mask (IRM). In the training stage, the Skip-DNN model is trained using a collection of speech data, which are composed of feature parameters and training targets. In the enhancement stage, the noisy speech features are processed by the trained Skip-DNN model to predict the clean speech features. Then the enhanced speech was obtained by speech synthesis, and finally, the speech quality was evaluated by speech evaluation. Based on the above research ideas, this paper proposes the structural block diagram of the Skip-DNN speech enhancement algorithm, as shown in Fig.1.

Fig.1 Diagram of a Skip-DNN based speech enhancement system

The main parts of the speech enhancement system in Fig.1 are described as follows:

In general, we assume the noise is additive in speech enhancement. The expression of noisy is defined as:

$$ Y(t, f) = S(t, f) + N(t, f) $$

where, $Y(t, f)$ denotes noisy signal, $S(t, f)$ denotes clean speech signal and $N(t, f)$ denotes noise signal. The noisy signal in time-frequency domain can be expressed as:

$$ Y(t, f) = S(t, f) + N(t, f) $$

The feature extraction part and the establishment of the Skip-DNN model in Fig.1 have been explained in Sections 2.2 and 2.3.

This paper utilizes IRM as the training target, $IRM \in 0,1$ , it denotes the proportion of clean speech energy in the noisy energy. IRM is expressed as:

$$ IRM(t, f) = \left( \frac{S(t, f)^2}{S(t, f)^2 + N(t, f)^2} \right)^{\beta} $$

where, $S(t, f)^2$ and $N(t, f)^2$ denote speech energy and noise energy within a time-frequency unit, respectively. The tunable parameter $\beta$ scales the mask, and is commonly chosen to $1/2$.

In this study, we use the IRM as our training target. According to (3), we can namely estimate clean speech as follows:

$$ S(t, f) = IRM(t, f) \cdot Y(t, f) $$

Then, the enhanced speech that combines the estimated magnitude spectrum and phase of noisy is obtained as follows:

$$ S(t, f) = |S(t, f)| \cdot e^{i \angle Y(t, f)} $$

where, $\angle Y(t, f)$ denotes the phase of the noisy and $S(t, f)$ denotes the reconstructed clean speech signal.

In this paper, STOI and PESQ are utilized to evaluate speech quality. Among them, STOI is an objective evaluation method of speech intelligibility. It uses a short time unit and eliminates the mute unit. The STOI value is in the range of 0 to 1. The larger
the STOI value, the higher the intelligibility. PESQ are objective evaluation methods for speech quality. The evaluation range of PESQ is between 1.0 and 4.5. The higher the score, the better the speech quality.

2.2 Speech Feature Analysis

The speech signal is short-time stationary for a length of approximately 10-30ms. To mitigate the nonstationary and time-varying effects, the speech signal is often processed in frames of shorter segments. The human ear is like a set of filter banks focusing only on signals in specific frequency bands. In the processing process, the voice signal can be divided into frames and frequencies to the Fourier transform domain. I-MRCG using different frame lengths to obtain four cochleagram with different resolutions is based on the cochleagram of the filter bank.

The I-MRCG acquisition process is as follows:

1. The time-domain signal is segmented through the 4-order and 64-channels Gammatone filter banks to obtain the 64 sub-band speech signals, set the corresponding frame length to 20ms with a frameshift of 10ms. Thus, a high-resolution cochleagram CG1 was obtained, with a feature size of 465*64. At the same time, CG1 is linearly transformed by power function, to make it more consistent with the auditory feature of the ear.

2. The process of obtaining low-resolution cochleagram CG2 is the same as CG1. The noisy speech signal is denoised through MMSE-STSA, and then step (1) is performed, except that the frame length is changed to 200ms, and the feature size of CG2 is 465 * 64.

3. CG1 was smoothed by adding a mean filter with a window length of 11×11 to obtain a low-resolution cochleagram CG3. If the window goes beyond the given cochleagram, the outside units take the value of zero.

4. CG4 is computed in a similar way to CG3, except that window length of the smoothing filter is 23 × 23.

5. CG1, CG2, CG3, and CG4 are fused and stitched with the underlying features along the time domain direction, and then combined with their first and second-order differences to obtain I-MRCG. The I-MRCG process is shown in Fig. 2.
2.3 The Skip-DNN Architecture

In speech enhancement, Skip-DNN is often used to learn the nonlinear relationship between input features and training targets, and its hierarchical structure can independently extract useful information from speech features. Skip-DNN is based on DNN by adding skip connections between layers to reduce information loss, so it can better solve the problem of part of the useful energy in a low SNR environment. The Skip-DNN speech enhancement structure is shown in Fig. 3.

![Fig. 3. The architecture of Skip-DNN speech enhancement](image)

As shown in Fig. 3, the Skip-DNN structure is mainly composed of the input layer, output layer, and hidden layer. Some of the hidden layers should have the same number of nodes as the input layer to resolve the dimensions mismatch between the hidden layer and the input layer. The commonly used loss function in the speech enhancement model based on deep learning is the mean square error. The mean square error calculates the similarity between the given target and the target estimated by the neural network, which can be expressed as:

\[
MSE(y, y') = \frac{1}{N} \sum_t \sum_f (y(t, f) - y'(t, f))^2
\]  

(6)

Among them: \( t \) and \( f \) respectively represent time and frequency; \( N \) represents sample point.

Because the mean square error is only a simple calculation of the similarity between a given target and an estimated target, in 2010, Vincent et al. [25] proposed a more detailed quantification method to evaluate the quality of speech and hearing based on the way the SNR perceives speech quality. Later, it was used as a loss function to measure the prediction performance of speech enhancement models based on deep learning. The SDR can be expressed as:

\[
SDR = 10 \log_{10} \frac{S_{\text{target}}^2}{\|e_{\text{interf}} + e_{\text{noise}} + e_{\text{artif}}\|^2}
\]

(7)

Among them: \( S_{\text{target}} \) means target speech; \( S_{\text{interf}} \) means non-target interfering speech; \( e_{\text{noise}} \) means noise; \( e_{\text{artif}} \) means other sounds introduced by the algorithm, for example, "musical noise".

This article improves on the SDR loss function proposed by Venkataramani [26], where \( SDR_{\text{loss}} \) can be expressed as:

\[
\max_{y} SDR_{\text{loss}}(y, \tilde{y}) = \max_{y} \frac{\langle y, \tilde{y} \rangle^2}{\|y\|^2 - \langle y, \tilde{y} \rangle^2} = \min_{y} \frac{\|\tilde{y}\|^2}{\langle y, \tilde{y} \rangle^2}
\]

(8)

\[
loss_{\text{SDR}}(y, \tilde{y}) = \frac{\langle y, \tilde{y} \rangle^2}{\|y\|^2}
\]

(9)

Among them: \( y \) represents the actual value; \( \tilde{y} \) represents the predicted value.

During the experiment, it can be simplified to:

\[
SDR_{\text{loss}}(y, \tilde{y}) = \frac{\langle y, \tilde{y} \rangle^2}{\|\tilde{y}\|^2} = -10 \log_{10} \frac{\|\tilde{y}\|^2}{\|y\|^2}
\]

(10)

The value range of \( SDR_{\text{loss}} \) is between \( 0 \) and \(-1\). In this paper, the logarithm of the simplified result is taken, and the improved loss function can be defined in the range of \( -\infty \) to \( 0 \), which is more convenient to observe the change of loss in the process of iteration. At the same time, by observing the logarithmic curve, it can be seen that the closer the logarithmic loss value is to 0, the more stable it will be.

\[
SDR_{\text{loss}}(y, \tilde{y}) = -10 \log_{10} \left( \frac{\|y\|^2}{\|\tilde{y}\|^2} \right)
\]

(11)

3 Results and discussion

3.1 Selection and setting of data set

The experimental clean speech data comes from the TIMIT, and the noise data comes from the NoiseX-92 database. During the experiment, 105 speech from the training set of the TIMIT is used as clean samples, the test set contains 45 pure speech samples, and the noise samples with 4 noise types, including f16, babble,
factory, white are used for synthesizing noisy speech. More specifically, in our experiments, all the clean speech is mixed with the above-mentioned noises at 4 levels of SNR (-5 dB, 0 dB, 5 dB, and 10 dB) to construct a multi-condition stereo database. As for signal analysis, all the clean and noisy waveforms are resampled to 8 kHz. A total of 600 noisy speech (about 1200 s) are constructed, of which 420 are used as training data and the rest 180 are used as test data.

In our experiments, all the clean speech is mixed with the above-mentioned noises at 4 levels of SNR (-5 dB, 0 dB, 5 dB, 10 dB) to construct a multi-condition stereo training set. Parameter setting of neural networks: the number of iterations is 30, The stochastic gradient descent algorithm is selected to improve the training process of the network. The number of nodes in the hidden layer is set as 1024, and the discard rate is 0.2, Because ReLU helps to resolve the convergence problem of deep neural networks, this experiment uses ReLU as the activation function. The input layer speech feature is set to 3 frames, and the output layer speech feature is set to 1 frame. During the processing of the following experimental data, the speech signal and the noise signal are selected according to the above criteria.

In this paper, f16, babble, factory, white are introduced into the pure speech signal as background noise. Among them, sections 3.1.2 analyze the speech of noisy speech signals under different SNR, different noise environments, and different models. Enhancement effect, the evaluation result is the evaluation score of each noise environment.

3.1.2 Evaluation of Speech Enhancement Effects of Different Network Models under Different Noise Environments

In order to analyze the influence of different noises on the enhancement effect of different models, in this experiment, under the condition that the SNR is -5 dB, 0 dB, 5 dB and 10 dB, 5 kinds of network models are trained, namely combined MFCC and DNN, combined MRCG and DNN, combined MFCC and Skip-DNN, combined MRCG and Skip-DNN, and I-MRCG and Skip-DNN speech enhancement model.

The noisy speech in the test stage is the speech in f16, factory, babble and white noise environment. The enhancement effect of noisy speech is shown in Fig.4.
It can be seen from Fig. 4 that when the background noise is f16, whether it is speech quality or speech intelligibility evaluation, almost all the speech enhancement models combined with I-MRCG and Skip-DNN have the best speech enhancement effect. The average STOI is 5.3%, 4.8%, 2.2%, and 1.7% higher than the combined MFCC and DNN, combined MRCG and DNN, combined I-MRCG and DNN, and combined MRCG and Skip-DNN, respectively. The average PESQ is 11.6%, 11.0%, 5.1%, and 3.3% higher than other models, respectively; When the background noise is a factory, it is almost always best to combine I-MRCG and Skip-DNN. The average...
STOI is 5.5%, 4.9%, 3.3%, and 0.8% higher than other models, respectively. The average PESQ is 11.5%, 11.0%, 7.0% and 2.8% higher than other models respectively; When the background noise is babble, it is almost always best to combine I-MRCG and Skip-DNN. The average STOI is 7.2%, 6.3%, 3.7%, and 1.9% higher than other models, respectively. The average PESQ is 18.6%, 16.8%, 11.4%, and 6.5% higher than other models, respectively; When the background noise is white, it is almost always best to combine I-MRCG and Skip-DNN. The average STOI is 6.0%, 5.3%, 3.4%, and 1.5% higher than other models, respectively. The average PESQ is 8.2%, 8.1%, 5.2%, and 2.5% higher than other models, respectively. It can be seen that under different conditions, using I-MRCG as the feature input of different speech enhancement models can obtain better enhanced speech.

3.1.3 Comparison between different loss functions

Comparing the three loss functions of MSE, SDR and improved SDR under different SNR and different background noise environments to enhance the performance in the Skip-DNN based speech enhancement model, and PESQ and STOI are used as performance indicators. The specific results are shown in Table 1 and Table 2

| Noise type | SNR | Noisy | Skip-MSE | Skip-SDR | Skip-Improve SDR |
|------------|-----|-------|----------|-----------|------------------|
| f16        | -5dB| 0.5774| 0.7833   | **0.7862**| **0.7856**       |
|            | 0dB |       |          |           |                  |
|            | 5dB |       |          |           |                  |
|            | 10dB|       |          |           |                  |
| factory    | -5dB| 0.5439| 0.7123   | 0.7095    | **0.7157**       |
|            | 0dB | 0.6617| 0.8384   | 0.8407    | **0.8421**       |
|            | 5dB | 0.7692| 0.9117   | 0.9134    | **0.9148**       |
| babble     | -5dB| 0.6605| 0.8240   | 0.8253    | **0.8260**       |
|            | 0dB | 0.779 | babble   | 0.9134    | **0.9327**       |
|            | 5dB | 0.8583| 0.9636   | 0.9644    | **0.9649**       |
| white      | -5dB| 0.6619| 0.7971   | 0.8075    | **0.8081**       |
|            | 0dB | 0.8836| 0.9709   | 0.9722    | **0.9728**       |

Table 2 Average PESQ scores for three loss function

| Noise type | SNR | Noisy | Skip-MSE | Skip-SDR | Skip-Improve SDR |
|------------|-----|-------|----------|-----------|------------------|
| f16        | -5dB| 1.3427| 2.0949   | 2.1210    | **2.1270**       |
|            | 0dB | 1.5674| 2.5379   | 2.5446    | **2.5674**       |
|            | 5dB | 1.8857| 2.8607   | 2.8799    | **2.9947**       |
|            | 10dB| 2.2340| 3.3581   | 3.3853    | **3.4027**       |
| factory    | -5dB| 1.2431| 1.8978   | 1.8670    | **1.8972**       |
|            | 0dB | 1.5336| 2.3778   | 2.3899    | **2.3994**       |
|            | 5dB | 1.8328| 2.7725   | 2.7871    | **2.7974**       |
| babble     | -5dB| 1.3590| 1.7451   | 1.7452    | **1.7635**       |
|            | 0dB | 1.7010| 2.2861   | 2.2816    | **2.2967**       |
|            | 5dB | 2.0057| 2.9584   | 2.9842    | **2.9969**       |
|            | 10dB| 2.3219| 3.2827   | 3.3069    | **3.3232**       |
| white      | -5dB| 1.2656| 2.0264   | 2.0324    | **2.0206**       |
|            | 0dB | 1.4542| 2.6318   | 2.6610    | **2.6755**       |
|            | 5dB | 1.7384| 3.1435   | 3.1808    | **3.2069**       |
|            | 10dB| 2.0786| 3.3911   | 3.4265    | **3.4417**       |

It can be seen from Table 1 and Table 2 that only when the background noise is f16, the SNR is -5dB and the background noise is white, the SNR is 0dB, the target obtained by combining Skip-DNN and the improved SDR speech enhancement model. The intelligibility of speech is lower than the other two
methods. Only when the background noise is factory and white, and the SNR is -5dB, the auditory quality of the target speech obtained by combining Skip-DNN and the improved SDR speech enhancement model is lower than the other two methods. Therefore, when the loss function is selected as improved SDR, the performance of Skip-DNN based speech enhancement model can be improved.

4 Conclusion

In this paper, the feature input is MRCG. Skip-DNN speech enhancement model with SDR as loss function. It has the features of carrying more voice information, including global and local speech features and clearer training process. Experimental research shows that the speech enhancement model is Skip-DNN, I-MRCG and other features are used as feature input, and IMRCG has the highest voice quality and intelligibility; The Skip-DNN speech enhancement model with SDR as the loss function has better enhancement effect, the loss value of training set and verification set is more stable, and the training process is clearer.

From the above, it can be seen that the Skip-DNN speech enhancement model using the input feature MRCG is compared with other models. This network model can obtain the speech enhancement effect with higher speech quality and speech intelligibility. Similarly, the enhancement effect of noisy in a low SNR environment.

Abbreviations

I-MRCG: Improved multi-resolution cochleagram features; Skip-DNN: Skip connections deep neural network; SDR: Source to distortion ratio; SNR: SNR; PESQ: Perceptual evaluation of speech quality; STOI: Short-term objective intelligibility; MMSE-STSA: Minimum mean-square Error short-time spectral amplitude; IRM: Ideal Ratio mask

Acknowledgements

This work was supported by the national natural science youth foundation of china (No.11804068) and the natural science foundation of Heilongjiang Province (No. LH2020F033).

Availability of data and materials

Please contact author for data requests.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

1 School of Electrical and Electronic Engineering, Harbin University of Science and Technology, Harbin, 150080, P.R.China
2. Blood Transfusion, General Hospital of Heilongjiang Province Land Reclamation Bureau, Harbin, 150086, P.R.China

References:

1. S F Boll, Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on Acoustics, Speech, and Signal Processing. 27(2), 113-120(1979). https://ieeexplore.ieee.org/document/1163209?arnumber=1163209
2. J S Lim, A V Oppenheim, Enhancement and bandwidth compression of noisy speech. Proceedings of the IEEE. 67(12), 1586-1604(2005). https://ieeexplore.ieee.org/document/1455809
3. Y Ephraim, D Malah, Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Transactions on Acoustics, Speech, and Signal Processing. 32(6), 1109-1121(1984). https://ieeexplore.ieee.org/document/1164453
4. R Martin, Spectral subtraction based on minimum statistics. Paper presented in Proc. Euro. Signal Processing Conf, 1994. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1
5. A M Reddy, B Raj, Soft mask methods for single-channel speaker enhancement. IEEE Transactions on Audio, Speech, and Language Processing. 15(6), 1766-1776(2007).
https://dl.acm.org/doi/10.1109/TASL.2007.901310

6. W Felix, G Juergen, W Martin, Feature enhancement by deep LSTM networks for ASR in reverberant multisource environments. COMPUTER SPEECH AND LANGUAGE. 4(28), 888-902(2014).
https://www.sciencedirect.com/science/article/abs/pii/S0885230814000114

7. T G Kang, K Kisoo, J W Shin, NMF-based Target Source Separation Using Deep Neural Network. Published in IEEE Signal Processing Letters, Feb 2015.
https://ieeexplore.ieee.org/document/6892992

8. Y Wang, D L Wang, Towards Scaling Up Classification-Based Speech Separation. Published in IEEE Transactions on Audio, Speech, and Language Processing, July 2013.
https://ieeexplore.ieee.org/document/6473841

9. P Smaragdis, Convolutional Speech Bases and Their Application to Supervised Speech Separation. Published in: IEEE Transactions on Audio, Speech, and Language Processing, January 2007.
https://ieeexplore.ieee.org/document/4032795

10. D D Lee, H S Seung, Learning the parts of objects by non-negative matrix factorization. Nature. 401(10), 788-791(1999). https://www.nature.com/articles/44565#citeas

11. Y Jiang, D L Wang, R S Liu, Z M Feng, Binaural classification for reverberant speech segregation using deep neural networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 22, 2112-2121(2014).
https://ieeexplore.ieee.org/document/6914580

12. D S Williamson, Y Wang, D L Wang, Complex ratio masking for monaural speech separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 24, 483-492(2016).
https://ieeexplore.ieee.org/document/7364200

13. S Srinivasan, N Roman, D L Wang, Binary and ratio time-frequency masks for robust speech recognition. Speech Communication. 48(11), 1486-1501(2006).
https://www.sciencedirect.com/science/article/abs/pii/S0167639306001129

14. Y Xu, Research on speech enhancement method based on deep neural network, University of Science and Technology of China, 2015.
https://kns.cnki.net/kcms/detail/detail.aspx?dbcode=CDFD&dbname=CDFDLAST2015&filename=1015616008.nh&v=OHRDS5WJWj2b+jKx7ficiLYDhmuTh1x8Ayj%25mmd2BKUJz6UUb%25mmd2Bd6QtbW4%25mmd2FuaFLatNN

15. Y Xu, J Du, L R Dai, C H Lee, A regression approach to speech enhancement based on deep neural networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 23(1), 7-19(2015).
https://ieeexplore.ieee.org/document/6932438

16. Y Xu, J Du, L R Dai, C H Lee, An experimental study on speech enhancement based on deep neural networks. IEEE Signal Processing Lett. 21(1), 66-68(2014). https://ieeexplore.ieee.org/document/6665000

17. Y Wang, D Wang, A structure-preserving training target for supervised speech separation. Paper presented at 2014 IEEE International Conference on Acoustics, Speech and Signal Processing, Florence, Italy, 4-9 May 2014.
https://ieeexplore.ieee.org/document/6854777#footnotes

18. J T Chen, Y X Wang, D L Wang, A feature study for classification-based speech separation at very low signal-to-noise ratio. Paper presented at 2014 IEEE International Conference on Acoustics, Speech and Signal Processing, Florence, Italy, 14 July 2014. https://ieeexplore.ieee.org/document/6854965/

19. J T Chen, Y X Wang, D L Wang, Noise perturbation for supervised speech separation. Speech Communication. 78, 1-10(2016). https://www.sciencedirect.com/science/article/abs/pii/S0167639315001405

20. Y Tu, J Du, Y Xu, L Dai, C Lee, Deep neural network based speech separation for robust speech recognition. Published in 2014 12th International Conference on Signal Processing, Hangzhou, China, 19-23 Oct. 2014.
https://ieeexplore.ieee.org/document/7015061?tp=&arnumber=7015061&ranges%3D2013_2015_p_Publication_Year%26queryText%3Ddeep%20neural%20network%20in%20speech=

21. M Tu, X Zhang, Speech enhancement based on Deep Neural Networks with skip connections. Paper presented at 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, New Orleans, LA, USA, 5-
22. H W Tseng, M Hong, Z Q Luo, Combining sparse NMF with deep neural network: A new classification-based approach for speech enhancement. Paper presented at 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, South Brisbane, QLD, Australia, 19-24 April 2015. https://ieeexplore.ieee.org/document/7178350

23. W H Shi, X W Zhang, R X Zhang, W Han, Special Lectures on Deep Learning Theory and Its Applications (4) Lecture 8 The Application of Deep Learning Methods in Speech Enhancement. Military Communication Technology. 37(03), 98-104(2016). https://kns.cnki.net/kcms/detail/detail.aspx?dbcode=CJFD&dbname=CJFDLAST2016&filename=JSTY201603021&v=fgNytVa6QrSPGnMaPHpz2qMtxEbwAw7zQMAvvn0uv0aRrm8owrnCdYqGPOpu0

24. W J Liu, S Nie, S Liang, X L Zhang, Deep learning based speech separation technology and its developments. Acta Automatica Sinica. 42(6), 819-833(2016). http://en.cnki.com.cn/Article_en/CIFDTotal-MOTO201606003.html

25. Vincent E, Rémi G, Cédric F. Performance measurement in blind audio source separation. IEEE Transactions on Audio Speech and Language Processing. 14(4): 1462-1469(2006). https://ieeexplore.ieee.org/document/1643671/similar

26. Choi H S, Kim J H, Huh J, et al. Phase-aware Speech Enhancement with Deep Complex U-Net. International Conference on Learning Representations, Montreal, Canada, 2019. https://arxiv.org/abs/1903.03107v1
Figures

Figure 1

Diagram of a Skip-DNN based speech enhancement system

Figure 2

Improve MRCG extraction flowchart
Figure 3

The architecture of Skip-DNN speech enhancement
Figure 4

Effect evaluation of various speech enhancement models under four SNRs