Research Article

Research on Speech Enhancement Algorithm of Multiresolution Cochleagram Based on Skip Connection Deep Neural Network

Chaofeng Lan †,1 YuQiao Wang †,1 Lei Zhang ‡,2 Chundong Liu †,1 and Xiaojia Lin †,1

†College of Measurement and Communication Engineering, Harbin University of Science and Technology, Harbin 150080, China
‡Beidahuang Industry Group General Hospital, Harbin 150088, China

Correspondence should be addressed to Chaofeng Lan; lanchaofeng@hrbust.edu.cn, Lei Zhang; happyzhang68@126.com, and Chundong Liu; lchd840@163.com

Chaofeng Lan, Lei Zhang, and Chundong Liu contributed equally to this work.

Received 20 October 2021; Revised 8 April 2022; Accepted 19 April 2022; Published 9 May 2022

Academic Editor: Bin Gao

Copyright © 2022 Chaofeng Lan et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The speech enhancement effect of traditional deep learning algorithms is not ideal under low signal-to-noise ratios (SNR). Skip connections-deep neural network (Skip-DNN) improves the traditional deep neural network (DNN) by adding skip connections between each layer of the neural network to solve the degradation problem of DNN. In this paper, the Multiresolution Cochleagram (MRCG) features in the gammachirp transform domain are denoised to obtain the improved MRCG (I-MRCG). The noise reduction method adopts the Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator (MMSE-STSA) and takes I-MRCG as the input feature and Skip-DNN as the training network to improve the speech enhancement effect of the model. This paper also proposes an improved source-to-distortion ratio (SDR) loss function. When the loss function uses the improved SDR, it will improve the performance of Skip-DNN speech enhancement model. The experiments in this paper are performed on the Edinburgh dataset. When using I-MRCG as the input feature of Skip-DNN, the average perceptual evaluation of speech quality (PESQ) is 2.9137, and the average short-time objective intelligibility (STOI) is 0.8515. Compared with MRCG as Skip-DNN input features, the improvements are 0.91% and 0.71%, respectively. When the improved SDR is used as the loss function of the speech model, the average PESQ is 2.9699 and the average STOI is 0.8547. Compared with other loss functions, the improved SDR has a better enhancement effect when used as the loss function of the speech enhancement model.

1. Introduction

As one of the core technologies in information science, speech enhancement technology is mainly used to enhance the clarity and intelligibility of speech polluted by noise. In recent years, speech enhancement has been widely used in mobile phones, smart appliances, multimedia applications, military eavesdropping, and other fields [1–4]. At the same time, speech recognition technology also needs speech enhancement technology [5]. In practical application, speech enhancement preprocessing is generally carried out in the speech processing system to improve the anti-interference ability of the system, thereby achieving more effective communication or interaction [6].

Speech enhancement methods can be divided into unsupervised and supervised learning methods. The most commonly used unsupervised learning methods are spectral subtraction and Wiener filtering [7, 8]. A problem in traditional speech enhancement algorithms is that the enhanced speech often contains an artifact called “musical noise.” Virag [9] incorporated the masking properties of the human auditory system into speech enhancement and proposed an algorithm that allows automatic adjustment of parameters in time and frequency, which
effectively reduces “musical noise” but outputs a low SNR. Liu et al. [10] proposed an improved spread-spectrum speech enhancement algorithm. The algorithm dynamically calculates the parameters according to the masking threshold of the key frequency segment of each speech frame and improves the output SNR. Previous studies have shown that unsupervised speech enhancement has good performance in high SNR and stationary noise environments but poor performance in low SNR and nonstationary noise environments.

Supervised speech enhancement algorithms are divided into shallow model algorithms and deep model algorithms. The nonnegative matrix factorization speech enhancement algorithm of the shallow model algorithm constructs the signal base as the prior information of the enhancement stage under the premise of being independent of the pure speech and the noise. It processes the noise to obtain the enhanced speech [11]. Since the shallow model cannot extract useful features from the data automatically, the processing ability of high-dimensional data is limited, resulting in a poor speech enhancement effect. The deep model algorithm was introduced into the field of speech enhancement. Talbi and Bouhlel [12] proposed a speech enhancement method based on stationary bionic wavelet transform (SBWT) and spectral magnitude minimum mean square error (MMSE) estimation. Combined with LWT and artificial neural network, MMSE estimation of spectral amplitude is used, which has a good noise reduction effect. Mahmmod et al. [13, 14] proposed a new estimator based on a combination of optimal linear and nonlinear Laplacian distribution, and the proposed estimator is based on MMSE perception. It reduces residual noise without destroying the speech signal and regenerates distorted speech components. In recent years, DNN has been widely used in speech enhancement [15]. At present, DNN speech enhancement can be roughly divided into two methods. The first is to seek the mapping between noise and pure speech spectrum [16], and the other is based on masking [17]. DNN-based speech enhancement usually produces excessively smooth speech, resulting in speech distortion and loss of intelligibility, and large DNNs take up large memory and slow training speed [18]. Kim and Shin [19] proposed to exaggerate the training target and use another DNN to estimate the exaggerated residual error in
DNN-based speech enhancement, so that the dynamic range of enhanced speech is closer to clean speech. Tan and Wang [20] proposed two compression pipelines to reduce the model size of DNN-based speech enhancement, reducing the model size without significantly reducing their enhancement performance. Tu and Zhang [21] proposed using Skip-DNN for speech enhancement and achieved a better speech enhancement effect. Wang and Bao [22] used Skip-DNN to study speech enhancement in the mel domain and achieved high speech quality and intelligibility. Cheng et al. [23] integrated the mel domain and gammatone domain and improved the performance of speech enhancement through DNN training.

Compared with the mel filter bank and the gammatone filter bank, the gammachirp filter bank is more in line with the auditory characteristics of the human ears, so this paper will choose I-MRCG in the gammachirp transform domain to extract speech features. To improve the speech enhancement performance, especially the speech enhancement effect under low SNR, this paper uses the I-MRCG feature as the input of the Skip-DNN model. It compares the speech enhancement effect of Skip-DNN under different feature parameters and speech enhancement models. The STOI and PESQ evaluation scores under different SNR conditions are analyzed. This paper proposed an improved SDR loss function. By comparing the speech enhancement effect of the Skip-DNN model using different loss functions, the influence of the improved SDR loss function on the performance of Skip-DNN is analyzed, and the enhancement effect of the proposed model and the impact of the loss function in this paper are studied.

2. Methods

2.1. Gammachirp Filter Model. The gammachirp filter model was first proposed by Irino and Patterson [24] based on the gammatone filter in 1997, and its time-domain expression is expressed as

\[
g(t) = a t^n e^{-2b \text{ERB}(f_c) t} e^{2\pi f_c t + j \ln t + j \phi} u(t),
\]

where \(a\) is the amplitude, \(n\) and \(b\) are used to adjust the distribution of the gamma function, \(f_c\) is the center frequency of the gammachirp filter, \(t\) is the time, and \(\phi\) is the initial phase. Since the initial phase has limited influence on the power spectrum, \(\phi = 0\) is generally taken. \(\text{ERB}(f_c)\) is the equivalent rectangular bandwidth of the filter at frequency \(f_c\). \(e\) is the chirp factor, which has a linear relationship with the sound level intensity \(P_s\).

The frequency-domain function can be obtained by Fourier transform of the gammachirp time-domain function,
Figure 5: Continued.
Figure 5: Continued.
Figure 5: Continued.

(e) STOI score under Volvo noise

(f) PESQ score under Volvo noise
Figure 5: Effect evaluation of different speech enhancement models.
the pure speech signal can be obtained. The MMSE-STSA is a traditional speech enhancement algorithm. Its noise reduction process is as follows: convert the noisy speech signal \( Y(t) \) and the pure speech signal \( S(t) \) to the frequency domain through fast Fourier transform (FFT), and the noisy speech signal and the pure speech signal can be obtained. The \( i \)-th spectral component features are

\[
Y_i = R_ie^{i\theta_i}, \quad S_i = A_ie^{i\theta_i}.
\]

Among them, \( R_i \) and \( A_i \) are the amplitudes of the noisy speech and the pure speech, respectively, and \( \theta_i \) and \( \lambda_i \) are the phases of the two, respectively.

Assuming that the noise does not affect the phase, the estimated value \( \hat{S}_i \) of the pure speech can be approximately equal to the estimated value \( \hat{A}_i \) of the magnitude spectrum. According to the characteristics of the correlation between adjacent frames of the speech signal spectrum, the estimated value \( \hat{A}_i \) of the magnitude spectrum can be expressed. Then, the estimated value \( \hat{S}_i \) of the pure speech can be obtained:

\[
\hat{S}_i = \hat{A}_i = E\{A_i|Y_1, Y_2\} = E\{A_i|Y_1\} = \frac{\sum_n c_n a_p(Y_i|a_i, a) p(a_i, a) da_i da}{\sum_n c_n p(Y_1|a_i, a) p(a_i, a) da_i da_i},
\]

where \( E(\bullet) \) represents the expectation of the parameter, \( p(\bullet) \) is the probability density function, \( p(a_i) \) is the probability density function of \( A_i \), and \( p(a_i, a) \) is the probability of the amplitude–phase joint probability distribution.

Assuming that the noise is the stationary additive white Gaussian noise, then

\[
\hat{A}_i = \Gamma(1.5) \sqrt{\frac{\nu_i}{Y_i}} M(-0.5, 1, -\nu_i) R_i,
\]

where \( \Gamma(\bullet) \) represents the gamma function, \( M(a, c, x) \) represents the confluenze hypergeometric function, \( \nu_i = (\xi_i (1 + \xi_i) / Y_i, \xi_i \) and \( Y_i \) represent the priori SNR and posteriori SNR, respectively.

2.2.2. I-MRCG Feature Extraction. Traditional MRCG consists of local information and global information. Different from traditional MRCG, I-MRCG uses MMSE-STSA to process speech signals. Because the noise energy is high in the environment of low SNR, if MMSE-STSA is used to denoise the local features of traditional MRCG, the noise estimation may be inaccurate, and some inherent speech features may be lost. Therefore, MMSE-STSA is used to denoise the global features of traditional MRCG, which can obtain clear global information on the one hand and retain the inherent details of speech on the other hand. The improved MRCG is spliced by the traditional MRCG, and the MRCG is processed by MMSE-STSA. At this time, the I-MRCG contains both inherent local information and clear global information. The I-MRCG process is shown in Figure 1 [25].

The I-MRCG acquisition process is as follows:

(1) The speech signal is divided into subbands through the gammachirp filter bank, and cochlear images of different resolutions are obtained through different frame lengths. The frame length and frame shift are set to 20 ms and 10 ms, respectively, and then, the

| Table 1: The standard deviations of STOI. |
|------------------------------------------|
| MRCG +DNN | IMRCG +DNN | MRCG +Skip-DNN | MFCC +Skip-DNN | IMRCG +Skip-DNN |
| m109 | 0.1074 | 0.1003 | 0.0898 | 0.0883 | 0.8366 |
| Pink | 0.1664 | 0.1534 | 0.1441 | 0.1342 | 0.1309 |
| Volvo | 0.0393 | 0.0368 | 0.0363 | 0.0388 | 0.0346 |
| Leopard | 0.0823 | 0.0723 | 0.0665 | 0.0660 | 0.0636 |

| Table 2: The standard deviations of PESQ. |
|------------------------------------------|
| MRCG +DNN | IMRCG +DNN | MRCG +Skip-DNN | MFCC +Skip-DNN | IMRCG +Skip-DNN |
| m109 | 0.5529 | 0.5038 | 0.4248 | 0.4369 | 0.4147 |
| Pink | 0.9168 | 0.8034 | 0.7658 | 0.6943 | 0.6660 |
| Volvo | 0.2519 | 0.2659 | 0.2031 | 0.2256 | 0.2114 |
| Leopard | 0.9693 | 0.9331 | 0.8492 | 0.8620 | 0.8416 |

which can be expressed as

\[
G_f = \frac{1}{e^{-a} \cdot \{2\sqrt{b^2 + (f - f_s)^2}\}^\frac{1}{\nu} \cdot e^{i\alpha} \cdot \{2\sqrt{b^2 + (f - f_s)^2}\}^\frac{1}{\nu} \cdot e^{i\theta}},
\]

where \( a \) is a constant, the second term is the Fourier transform of the time domain function of the gammatone filter, denoted as \( G_T(f) \), and the third term is an asymmetric function, denoted as \( H_A(f) \). By normalizing the amplitude, the frequency spectrum of the gammachirp filter can be expressed as

\[
G_C(f) = G_T(f) \cdot H_A(f).
\]

The magnitude spectrum of the filter can be expressed as

\[
|G_C(f)| = |G_T(f)||H_A(f)| = \frac{1}{\sqrt{2\pi b^2 (f - f_s)^2}} \cdot e^{i\lambda}.
\]

2.2. I-MRCG Characteristic Parameters

2.2.1. Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator. MMSE-STSA is a traditional speech enhancement algorithm. Its noise reduction process is as follows: convert the noisy speech signal \( Y(t) \) and the pure speech signal \( S(t) \) to the frequency domain through fast Fourier transform (FFT), and the noisy speech signal and the pure speech signal can be obtained. The \( i \)-th spectral

\[
Y_i = R_i e^{i\beta_i}, \quad S_i = A_i e^{i\theta_i}.
\]
Figure 6: Continued.
Figure 6: Continued.
high-resolution cochlear spectrogram CG1 can be obtained through the power function

(2) The low-resolution cochleagram CG2 acquisition process is roughly the same as CG1, except that the frame length is 200 ms

(3) The cochlear spectrogram CG1′ is obtained by obtaining CG1, but the speech signal at this time is processed by MMSE-STSA noise reduction, and CG1′ is passed through an 11 × 11 mean filter to get CG3

(4) The cochlear spectrogram CG2′ is obtained by obtaining CG2, but the speech signal at this time is processed by MMSE-STSA noise reduction, and CG2′ is passed through an 23 × 23 mean filter to get CG4

(5) The four features are spliced together to obtain I-MRCG

2.3. Skip-DNN Speech Enhancement Model. The skip connection is a commonly used connection form in neural networks. The traditional DNN causes network degradation due to the same weight, resulting in the poor fitting effect of complex features. The skip connection is used to break the symmetry of the network and find some useful features that are masked during the training process. The skip connection can solve the problem of singularity and gradient disappearance caused by the unidentifiable part of the model. It can also solve the problem of loss of detailed information in the transmission process due to different dimensions between layers. The missing information is supplemented by skip connections. The network structure of Skip-DNN is shown in Figure 2.

As can be seen from Figure 2, to prevent the mismatch between the dimensions of the input layer and the hidden layer in the process of skip connection, the number of nodes in some hidden layers is set as the dimension of the input layer. Because the input layer is a one-dimensional vector and the input of the Skip-DNN speech enhancement model
is two dimensions of time and frequency, speech and noise have strong correlations in both dimensions. Therefore, the feature space as the input signal should include context information. That is, the input speech feature parameters should include this frame and its preceding and following frames.

To illustrate the impact of skip connections on the data transmission process, the residual learning unit is shown in Figure 3.

Assume that the input speech signal feature is $Y$, the feature learned by the Skip-DNN is denoted as $I(Y)$, and the feature obtained after $Y$ which is linearly transformed is denoted as $F(Y)$. Through the connection of Skip-DNN in Figure 3, the structure can be concluded that $I(Y) = F(Y) + Y$. In the residual network, because the network is deep, there will be redundant layers, but in Skip-DNN, the number of network layers is not deep, so there will be no redundancy layer. That is, $F(Y)$ is not equal to 0. The method of skip connection enables a certain gradient to be guaranteed when updating parameters in the process of backward propagation, thus solving the problem of gradient disappearance in deep networks.

This paper uses the gammachirp filter bank that can simulate the auditory characteristics of the human ear to divide the frequency domain speech into subbands and obtain multiresolution cochlear features according to different frames. Taking the ideal ratio mask (IRM) as the training target, the nonlinear relationship between the feature parameters and the training target and the speech enhancement model based on Skip-DNN is constructed. The model is shown in Figure 4.

In the training stage, firstly, the multiresolution cochlear feature $Y(t,f)$ is obtained by subbanding the noisy speech $S(t)$ by the gammachirp filter. The specific extraction process is given in Section 2.2.

Secondly, the speech and noise are converted to the time-frequency domain by the gammachirp operation, respectively, expressed as $S(t,f)$ and $N(t,f)$. As the training target of the speech enhancement model in this chapter, IRM calculates the ratio of pure speech and noise energy to obtain $IRM(t,f)$, which can be expressed as

$$IRM(t,f) = \left[ \frac{S^2(t,f)}{S^2(t,f) + N^2(t,f)} \right]^{1/2}, IRM \in [0,1].$$

Among them, $S^2(t,f)$ and $N^2(t,f)$ represent the pure speech energy and noise energy in the time-frequency domain, respectively.

Finally, the useful speech information is extracted autonomously by Skip-DNN; the nonlinear relationship between the feature parameters and the training target and the speech enhancement model is obtained.

The noisy speech $Y(t,f)$ is converted to the time-frequency domain to obtain $Y(t,f)$, the feature parameters are extracted, and the training target $IRM^+$ is estimated through the speech enhancement model obtained by training. The target speech spectrum can be obtained by multiplying it by the noisy speech spectrum. Then, the estimated pure speech signal amplitude and the phase of the noisy speech are reconstructed to obtain the estimated speech, which can be expressed as

$$\hat{S}(t,f) = IRM^+(t,f) \times Y^+(t,f),$$

$$IRM(t,f) = \left[ \frac{S^2(t,f)}{S^2(t,f) + N^2(t,f)} \right]^{1/2}, IRM \in [0,1].$$

Among them, $\hat{S}(t,f)$ represents the target speech amplitude in the time-frequency domain, $\mathcal{L}Y^+(t)$ represents the noisy speech phase, $\hat{Y}(t)$ represents the reconstructed pure speech signal.

2.4. Improved SDR Loss Function. Mean square error (MSE) loss function is the most used in deep learning speech enhancement models [26]. Although MSE is used in many models, MSE simply calculates the similarity between a given target and an estimated target and cannot guarantee the optimal result. The small value of the model training results cannot indicate that the speech enhancement model estimates the high quality of speech. Therefore, in 2017, Venkataramani et al. [27] proposed an SDR loss function suitable for speech enhancement. The loss function SDR is obtained by calculating the ratio of the speech signal to the distorted signal, which is highly correlated with the speech quality.

The SDR loss function is shown in

$$\max_{\hat{y}} \text{SDR}_{\text{loss}}(y, \hat{y}) = \max_{\hat{y}} \frac{\langle y, \hat{y} \rangle^2}{\|y\|^2 \|\hat{y}\|^2} \times \min_{\hat{y}} \frac{\|\hat{y}\|^2}{\langle y, \hat{y} \rangle^2},$$

where $y$ is the actual value and $\hat{y}$ is the predicted value. Since it is a constant relative to the network output, its optimized negative reciprocal is expressed as

$$\text{loss}_{\text{ven}}(y, \hat{y}) = -\frac{\langle y, \hat{y} \rangle^2}{\|\hat{y}\|^2}. \quad (11)$$

It was further improved by Choi et al. [28]. The optimized loss function is shown in

$$\text{loss}_{\text{SDR}}(y, \hat{y}) = -\sqrt{\frac{\text{loss}_{\text{ven}}(y, \hat{y})}{\|y\|^2 \|\hat{y}\|^2}} = -\frac{\langle y, \hat{y} \rangle^2}{\|y\| \|\hat{y}\|}. \quad (12)$$

The value range of the loss function is between $[0,-1]$. This paper takes the negative value of the simplified result formula and then takes the negative logarithm, which can
be expressed as

\[ I\text{-}\text{loss}_{\text{SDR}}(y, \hat{y}) = -\log_{10}\left(-\text{loss}_{\text{SDR}}(y, \hat{y})\right). \] (13)

At this time, the value range of the improved loss function (equation (13)) is \([0, 0]\), which is convenient to observe the change of loss in the iterative process. At the same time, by observing the logarithmic curve, we can see that the closer the logarithmic loss value is to 0, the more stable it is. Therefore, this paper chooses equation (13) as the loss function for network training.

3. Results

3.1. Experimental Parameter Settings. The experimental pure speech data comes from the Edinburgh dataset [29], and the noise data comes from the NoiseX-92 database. During the experiment, 70 pure speech samples of the Edinburgh dataset were used as the training set, 30 pure speech samples were used as the test set, and four kinds of noises (m109, pink, Volvo, and leopard) were used. All clean speech is mixed with the above four kinds of noise at 3 SNR levels (-6 dB, 0 dB, and 6 dB) to build a multicondition stereo database. As for signal analysis, all clean and noisy waveforms are resampled to 8 kHz. A total of 1200 noisy speeches (about 6000 s) are constructed, of which 840 are used as training data, and the remaining 360 are used as test data.

The parameter settings of the neural network: the number of iterations is 60, and the stochastic gradient descent algorithm is selected to improve the training process of the network. The number of hidden layer nodes is set to 1024, and the dropout rate is 0.2. Since ReLU helps solve the convergence problem of deep neural networks, ReLU is used as the activation function in this experiment. The voice feature of the input layer is set to 3 frames, and the voice feature of the output layer is set to 1 frame. In the processing of the following experimental data, both the speech signal and the noise signal are selected according to the above criteria.

In this paper, m109, pink, Volvo, and leopard are introduced into the pure speech signal as background noise. Among them, Section 3.2 analyzes the speech enhancement effect of noisy speech signals under different signal-to-noise ratios, different noise environments, and different models, and the evaluation results are the evaluation scores of each noise environment. Section 3.3 analyzes the speech enhancement effects under different SNRs, different noise environments, and different loss functions.

3.2. Comparison between Different Speech Enhancement Models. This experiment trains five kinds of speech enhancement models under the condition of the SNR which is -6 dB, 0 dB, and 6 dB, namely, MRCG and DNN combination, IMRCG and DNN combination, MRCC and Skip-DNN combination, MFCC [30] and Skip-DNN combination, and I-MRCG and Skip-DNN combination. The noise in the test phase is m109, pink, Volvo, and leopard noise environment. The evaluation indicators are STOI and PESQ. The STOI and PESQ score is shown in Figure 5. To illustrate the results more clearly, the standard deviations of STOI for different models in different environments are shown in Table 1. The standard deviations of PESQ for different models in different environments are shown in Table 2.

It can be seen from Table 1 that the STOI of the speech enhancement model combined with Skip-DNN and I-MRCG for the four kinds of noise is higher than other models. It can be seen from Table 2 that the PESQ of the speech enhancement model of MRCG and Skip-DNN is slightly higher than that of Skip-DNN and I-MRCG in the Volvo noise environment, and the PESQ of the speech enhancement model of Skip-DNN and I-MRCG in the other noise environment is higher than others. It can be seen from Figure 5 that I-MRCG and Skip-DNN speech enhancement models have better effects on processing pink noise under the environment of low SNR. STOI is 7.2%, 4.4%, 2.8%, and 0.7% higher than MRCG and DNN combination, I-MRCG and DNN combination, MRCC and Skip-DNN combination, and MFCC and Skip-DNN combination, respectively. PESQ is 15.6%, 8.8%, 7.6%, and 2.4% higher than other models, respectively. For other noisy environments, the speech enhancement models of I-MRCG and Skip-DNN are also improved compared with other models.

3.3. Comparison between Different Loss Functions. To analyze the influence of different loss functions on Skip-DNN, in the speech enhancement model based on Skip-DNN, the performances of MSE, SDR, and improved SDR three loss functions under different SNR and different background noise environments were compared, with the PESQ and STOI as performance indicators. The results are shown in Figure 6.

As shown in Figure 6, only when the background noise is Volvo and the SNR is -6 dB, the combined Skip-DNN and improved SDR speech enhancement model STOI is lower than the Skip-DNN and MSE combined model. The speech enhancement model PESQ combined with Skip-DNN and improved SDR is higher than other models.

4. Conclusions

This paper proposes an I-MRCG speech feature processed by MMSE-STSA. The low-resolution features after noise reduction of MMSE-STSA are spliced with the high-resolution features of traditional MRCG. The obtained 1-MRCG has not only detailed local information but also clear global information. The network uses Skip-DNN training and solves the problem of gradient disappearance in the training process through skip connections. During the training process, the loss function adopts SDR. It optimizes by taking the negative value and the negative logarithm, resulting in an improved SDR loss function.

The experiment compares the performance of the Skip-DNN speech enhancement model with MRCG, MFCC, and I-MRCG. The combination of Skip-DNN and 1-MRCG achieves the best effect. For the pink noise environment with SNR of -6 dB, the speech enhancement effect based on Skip-DNN and 1-MRCG is greatly improved, and STOI score is 0.712 and PESQ score is 2.0102. By comparing the loss functions of MSE, SDR, and improved SDR, it is
concluded that the improved SDR can better improve the performance of Skip-DNN.

**Data Availability**

All data included in this study are available from the corresponding author upon request.

**Disclosure**

Part of the contents of this manuscript has been submitted as a preprint according to the following link: https://assets.researchsquare.com/files/rs-229829/v1_covered.pdf?c=1631855582.

**Conflicts of Interest**

The authors declare that there is no conflict of interest regarding the publication of this article.

**Authors’ Contributions**

Chaofeng Lan contributed to the conception of the study and contributed significantly to the analysis and manuscript preparation. Yuqiao Wang made important contributions to making adjustments to the structure of the paper, revised the paper, edited the manuscript, and polished the writing of the paper. Lei Zhang polished, edited, and checked the article during the first submission process and also made important contributions to the process of major revisions and adjustments of this manuscript. Chundong Liu performed the experiment and the data analyses and wrote the manuscript. Xiaojia Lin made important contributions to the process of major revisions and polishing of the writing of this manuscript. All authors reviewed the manuscript. Chaofeng Lan, Lei Zhang, and Chundong Liu contributed equally to this work.

**Acknowledgments**

This work was supported by the Natural Science Foundation of Heilongjiang Province (No. LH2020F033) and the National Natural Science Youth Foundation of China (No. 11804068).

**References**

[1] J. Gnanamanickam, Y. Natarajan, and K. R. S. Preethaa, “A hybrid speech enhancement algorithm for voice assistance application,” *Sensors*, vol. 21, no. 21, p. 7025, 2021.

[2] N. Radha, R. B. Jananie, and A. A. Silviya, “Enhancing speech quality using artificial bandwidth expansion with deep shallow convolution neural network framework,” *Fluctuation and Noise Letters*, vol. 21, no. 1, p. 2250008, 2021.

[3] N. Hu and Y. Bi, “Multimodal Intelligent Acoustic Sensor-Assisted English Pronunciation Signal Acquisition and Phonetic Calibration,” *Journal of Sensors*, 2022.

[4] N. Shankar, G. S. Bhat, I. M. S. Panahi, S. Tittle, and L. M. Thibodeau, “Smartphone-based single-channel speech enhancement application for hearing aids,” *Journal of the Acoustical Society of America.*, vol. 150, no. 3, pp. 1663–1673, 2021.

[5] C. H. Fan, J. Y. Yi, J. H. Tao, Z. Tian, B. Liu, and Z. Wen, “Gated recurrent fusion with joint training framework for robust end-to-end speech recognition,” *IEEE-ACM Transactions on Audio Speech and Language Processing*, vol. 29, pp. 198–209, 2021.

[6] K. Meemann and R. Smiljanic, “Intelligibility of noise-adapted and clear speech in energetic and informational maskers for native and nonnative listeners,” *Journal of Speech, Language, and Hearing Research: JSLHR*, vol. 65, no. 4, pp. 1263–1281, 2022.

[7] C. Li, T. Jiang, and S. Wu, “Single-channel speech enhancement based on improved frame-iterative spectral subtraction in the modulation domain,” *China Communications*, vol. 18, no. 9, pp. 100–115, 2021.

[8] T. T. Wang, H. Y. Guo, X. Yan, and Z. Yang, “Speech signal processing on graphs: the graph frequency analysis and an improved graph Wiener filtering method,” *Speech Communication*, vol. 127, pp. 82–91, 2021.

[9] N. Virag, “Single channel speech enhancement based on masking properties of the human auditory system,” *IEEE Trans. on Speech and Audio Processing*, vol. 7, no. 2, pp. 126–137, 1999.

[10] Y. H. Liu, D. M. Zhou, and Z. J. Jiang, “Improved spectral subtraction speech enhancement algorithm,” *Advanced Materials Research*, vol. 2542, p. 536, 2013.

[11] L. T. Xu, Z. L. Wei, S. F. A. Zaidi, B. Ren, and J. C. Yang, “Speech enhancement based on nonnegative matrix factorization in constant-Q frequency domain,” *APPLIED ACoustics*, vol. 174, article 107732, 2021.

[12] M. Talbi and M. S. Boulhed, “A new speech enhancement technique based on stationary bionic wavelet transform and MMSE estimate of spectral amplitude,” *Security and Communication Networks*, vol. 2021, Article ID 9968275, 10 pages, 2021.

[13] B. M. Mahmmod, A. R. Ramli, S. H. Abdulhussain, S. A. Hadad, and W. A. Jassim, “Low-distortion MMSE speech enhancement estimator based on Laplacian prior,” *IEEE Access*, vol. 5, pp. 9866–9881, 2017.

[14] B. M. Mahmmod, S. H. Abdulhussain, M. A. Naser, M. Alsabaah, and J. Mustafa, “Speech enhancement algorithm based on a hybrid estimator,” *Materials Science and Engineering*, vol. 1090, article 012102, 2021.

[15] S. Abdullah, M. Zamani, and A. Demosthenous, “Towards more efficient DNN-based speech enhancement using quantized correlation mask,” *IEEE Access*, vol. 9, pp. 24350–24362, 2021.

[16] G. Li, R. M. Hu, R. Zhang, and X. H. Wang, “A mapping model of spectral tilt in normal-to-Lombard speech conversion for intelligibility enhancement,” *Multimedia Tools and Applications*, vol. 79, no. 27–28, pp. 19471–19491, 2020.

[17] F. Farias and R. Coelho, “Blind adaptive mask to improve intelligibility of non-stationary noisy speech,” *IEEE Signal Processing Letters*, vol. 28, pp. 1170–1174, 2021.

[18] N. Furnon, R. Serizel, I. Illina, and S. Essid, “DNN-based mask estimation for distributed speech enhancement in spatially unconstrained microphone arrays,” *IEEE-ACM Transactions on Audio Speech and Language Processing*, vol. 29, pp. 2310–2323, 2021.

[19] H. Kim and J. W. Shin, “Target exaggeration for deep learning-based speech enhancement,” *Digital Signal Processing*, vol. 116, article 103109, 2021.
M. Tu and X. Zhang, "Speech enhancement based on deep neural networks with skip connections," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, 2017.

D. J. Wang and C. C. Bao, "An ideal Wiener filter correction-based cIRM speech enhancement method using deep neural networks with skip connections," in *14th IEEE International Conference on Signal Processing (ICSP)*, Beijing, China, 2018.

J. Cheng, R. Liang, and Z. Li, "DNN-based speech enhancement with self-attention on feature dimension," *Multimedia Tools and Applications*, vol. 79, no. 43-44, pp. 32449–32470, 2020.

T. Irino and R. D. Patterson, "A time-domain, level-dependent auditory filter: the gammachirp," *The Journal of the Acoustical Society of America.*, vol. 101, no. 1, pp. 412–419, 1997.

C. F. Lan, C. D. Liu, and L. Zhang, "Multiresolution cochleagram speech enhancement algorithm using improved deep neural networks with skip connections," https://assets.researchsquare.com/files/rs-229829/v1_covered.pdf?c=1631855582.

S. R. Hwang, B. Joon, and Y. C. Park, "Performance comparison evaluation of speech enhancement using various loss functions," *The Journal of the Acoustical Society of Korea*, vol. 40, no. 2, pp. 176–182, 2021.

S. Venkataramani, J. Casebeer, and P. Smaragdis, "End-to-end source separation with adaptive front-ends," in *IEEE Asilomar Conference on Signals, Systems, and Computers*, pp. 684–688, Pacific Grove, CA, USA, 2018.

H. S. Choi, J. H. Kim, and J. Huh, "Phase-aware speech enhancement with deep complex U-Net," in *International Conference on Learning Representations.*, pp. 37–57, 2019.

B. C. Valentini, X. Wang, S. Takaki, and J. Yamagishi, "Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks," *Proceedings of Interspeech*, pp. 352–356, 2016.

S. Debnath and P. Roy, "Audio-visual automatic speech recognition using PZM, MFCC and statistical analysis," *International Journal of Interactive Multimedia and Artificial Intelligence.*, vol. 7, no. 2, pp. 121–133, 2021.