Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 2074844, 9 pages
https://doi.org/10.1155/2022/2074844

Research Article

Voice Quality Evaluation of Singing Art Based on 1DCNN Model

Yang Liusong[1] **and Du Hui**[2]

[1]*School of Art and Media, Suqian University, Suqian 223800, Jiangsu, China*
[2]*School of Foreign Studies, Suqian University, Suqian 223800, Jiangsu, China*

Correspondence should be addressed to Yang Liusong; 20141@squ.edu.cn

Received 27 May 2022; Revised 22 June 2022; Accepted 28 June 2022; Published 30 July 2022

Academic Editor: Baiyuan Ding

Traditional speech recognition still has the problems of poor robustness and low signal-to-noise ratio, which makes the accuracy of speech recognition not ideal. Combining the idea of one-dimensional convolutional neural network with objective evaluation, an improved CNN speech recognition method is proposed in this paper. The simulation experiment is carried out with MATLAB. The effectiveness and feasibility of this method are verified by simulation. This new method is based on one-dimensional convolutional neural network. The traditional 1DNN algorithm is optimized by using the fractional processing node theory, and the corresponding parameters are set. Establish an objective evaluation system based on improved 1DCNN. Through the comparison with other neural networks, the results show that the evaluation method based on the improved 1DCNN has high stability, and the error between subjective score and evaluation method is the smallest.

1. Introduction

Signal processing is a general term for processing various types of signals according to various expected purposes and requirements. The processing of analog signals is called analog signal processing, and the processing of digital signals is called digital signal processing. The so-called signal processing refers to the process of processing the signals recorded on a certain medium in order to extract useful information. It is a general term for the process of signal extraction, transformation, analysis, synthesis, and so on. Among the common audio signal types, there is also a music signal. Music signal is affected by music theory, musical instrument pronunciation rules, psychological perception, and other factors, which is different from voice signal in analysis and processing methods. Moreover, the music-level analysis involves many knowledge fields, and knowledge is easy to spread. So in this lesson, we will first introduce some basic concepts of music signals and then take you to see the principles and solutions behind these problems from the cases of music scenes in our real-time audio interaction. Music has no national boundaries. Different regions and languages can be related to each other through music. As the most influential field in the world, music has its unique charm. However, due to the influence of people's subjective consciousness and old ideas, the scientific development of singing art is restricted, and talents in the field of singing are gradually lacking [1]. In 2019, Nygren et al. studied the impact of gender on individual vocal cords on sexual development. By investigating the satisfaction of the research objects on their own voice, everyone has a preliminary evaluation of their own voice from the perspective of gender [2]. In addition, David et al. also discussed the impact of gender on voice function in 2019 and found that people with gender diversity will be limited by a variety of voice function fields, which further supplements the voice quality caused by gender differences [3]. Rachel et al. analyzed the signal processing of singing voice and explored the technology of applying audio signal processing method to singing voice on the basis of previous research, in order to realize large-scale personalized listening experience [4]. For the processing of singing voice, Nanzaka R. led his team to propose a novel vocal music enhancement system, which can make the voice of amateur singers reach the professional level. By using the singing voice of professional singers, the singing voice of amateur singers can be enhanced in a frequency band

representing the distinctive characteristics of professional singers, opening up a new mode in the field of singing art [5]. Kim et al. studied the automatic analysis of pop music of singing voice, constructed a music tag data set, which was specially used for singing voice, and demonstrated the potential application of vocal music tagging system in music retrieval, music thumbtack, and singing evaluation [6].

Saleem et al. used the binary classification method of deep neural network to separate the target speech from the mixed signal in 2020, solved the problem of over smoothing, and carried out spectral variance equalization to match the estimated and clean speech features [7]. In the field of separation of singing voice and accompaniment, members of Lin K. collaborative group put forward a unique neural network method, which adopts the most advanced singing separation system competition of multi-channel modeling, data enhancement, and model mixing, making song evaluation more comprehensive [8]. In the latest research, Medeiros et al. compared the measurement methods of the variation of the predetermined fundamental frequency between singing and speech, and confirmed that the stability of the predetermined fundamental frequency in singing is higher, so as to facilitate the processing of the singing voice in the later research [9]. Lehner et al. proposed a machine learning method for vocal music detection, which can automatically identify the region in the music record of at least one person singing [10]. Based on the denoising self-coding model of mixed magnitude spectrum, Mimilakis et al. studied the mapping function of neural network, estimated the magnitude spectrum of singing voice from the corresponding mixture, and realized the separation of singing voice [11]. Murthy et al. used the new features based on formant structure to segment the vocal cord region and nonsound region, and the accuracy of the developed system in the song music segment detection test is as high as 98% [12]. Meiyanti et al. established a voice recognition research on singing technology based on female voice register in 2018. The detection results show that the average success rate of introducing voice register in real time is 57% [13].

To sum up, with the progress of science and technology, people pay more and more attention to the objective evaluation of sound, and the requirements for the quality of singing are also higher and higher. However, few studies have proposed an excellent method in singing speech recognition. Most of the objective evaluation of singing only focuses on the sound quality, the parameter selection is single, and the indicators in the sound evaluation are not detailed enough. Therefore, this study will optimize CNN with better feature classification ability and apply CNN optimization algorithm to the objective evaluation of singing art voice, in order to achieve the objective selection of singing art talents and protect the singer's voice environment.

2. Feature Extraction and Recognition of Singing Speech Signal

2.1. Singing Speech Recognition and Processing Method. Singing speech recognition is a special kind of speech recognition, and the biggest difference is that the singing voice has accompaniment and melody. However, the general principle of singing speech recognition remains unchanged, which is to identify continuous speech signals, retrieve keywords, and finally determine the position of words in sentences. The specific implementation process is shown in Figure 1.

As shown in Figure 1, the received voice signal is first converted into electrical signal by voice acquisition equipment such as microphone and then transmitted to the recognition system for front-end processing. Secondly, after the front-end processing, it is necessary to extract the features of the speech signal. Some of the extracted feature parameters are directly measured and estimated. The other part of the parameters constitutes a new pattern. By comparing with the original pattern of the database in the computer, a better matching combination is found. One part turns to measure estimation, and the other part turns to expert knowledge. Finally, all the new patterns get their corresponding recognition results according to the corresponding recognition decision.

Art voice audio is the same as other voice. After receiving the signal, it will be converted into digital signal, and then it will be sampled and quantified. Signal sampling is a discrete transformation process, which can ensure the integrity of the signal and reproduce the original signal. The quantization after sampling is to divide the amplitudes equally and ensure that there is no difference in the characteristics of samples in the same amplitude range. But in the process of signal processing, signal weakening is inevitable. In order to reproduce the original signal, it is necessary to emphasize the singing sound in advance to improve the high-frequency component of the audio and gradually make the signal close to the original signal. Cable connection is also a key factor affecting signal quality. In practice, in order to better ensure the quality of the signal, it is necessary to control various factors involved. In the continuous production process, the continuous phenomenon will occur due to the influence of factors such as improper size or large diameter of the cable. The poor working condition of continuous instruments is the key to these factors.

In order to facilitate the reproduction of the original signal, it is necessary to pre-emphasize the singing voice to improve the high-frequency component of the audio and gradually make the signal close to the original signal [14]. Generally, filter processing is used for pre-emphasis. The commonly used filter is actually FIR digital filter. FIR filter is a high-pass filter, which can improve the high-frequency components. At the same time, it is also convenient for formant detection, which improves the stability of signal in quantization processing [15].

There is no periodicity and no fixed law in singing voice. The sampling value and characteristic parameters of the signal will change irregularly with time, so it has time variability [16]. In signal processing, it is often considered that the voice signal of singing art is stable in a very short time, so the signal also has short-term stability. According to the characteristics of voice signal, the signal can be divided into several voice segments before processing [17]. Then, the signal is divided into several smooth fixed lengths by the

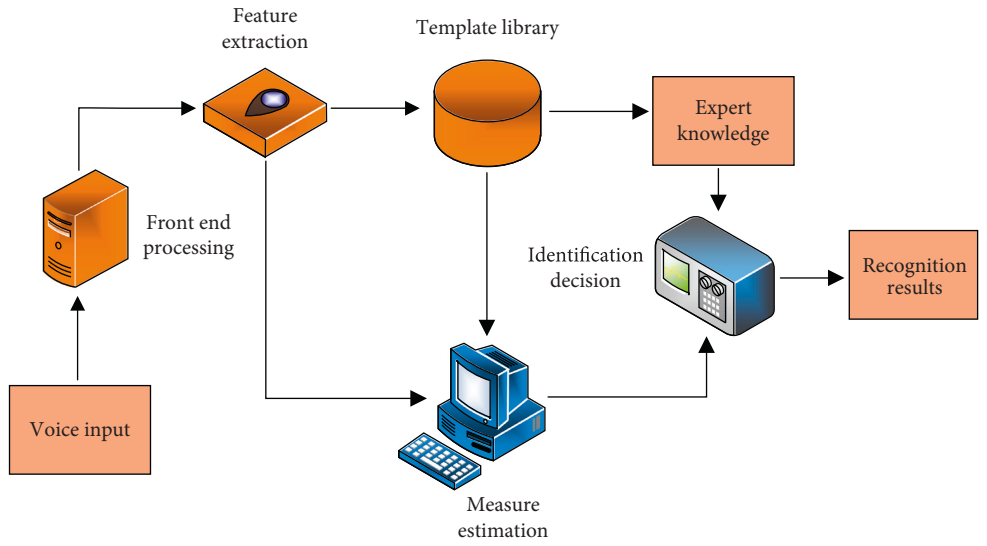

Figure 1: Schematic diagram of speech recognition.

windowing processing function, and the expression of the processing function is shown in

$$s_w(n) = s(n)w(n), \tag{1}$$

$s(n)$ in formula (1) represents the original signal, $w(n)$ is the windowing function used in processing, and the commonly used window functions are rectangular window, Hamming window, and Hanning window [18]. The comparison of the main lobe width and the first side lobe attenuation of the three window functions is shown in Table 1.

In Table 1, $B/\triangle w$ represents the width of main lobe falling, and A/dB represents the attenuation of the first side lobe. It can be seen from Table 1 that the main lobe width of Hanning window is the largest, followed by the Hamming window, and the Hamming window is the largest of the three [19]. Therefore, considering the two indicators, the Hamming window is selected as the window function.

2.2. Acoustic Parameter Extraction of Singing Speech. The main acoustic parameters of singing art voice include formant, fundamental frequency, range, and average energy. Extracting these acoustic parameters for research can better understand the beauty of sound. The voice of a singing performance is used for audio analysis. Among them, formant plays a decisive role in the depth and emotional color of voice, so the extraction of formant can basically show the singer's personal ability. In this study, AR model detection method is used to extract the first and third formants, and the extraction process is shown in Figure 2.

Figure 2 shows that the sample data are preprocessed first, and the processed signal is then sent to the AR model for detection. In order to ensure the accuracy of formant extraction, the AR model detection results need to be further extracted by LPC spectrum detection, then the formant points in the signal are captured by peak detection method, and finally the formant frequency is obtained. As one of the important standards to reflect the quality of voice, it is also very important to accurately extract the range for voice quality evaluation. The pitch value in the range is calculated by

$$\overline{D} = \frac{1}{N}\sum_{j}^{N} D_j, \tag{2}$$

In equation (2), $\overline{D}$ is the average of all pitches, n is the number of audio samples, D_j is the j pitch, and the standard deviation of all pitches is calculated as shown in

$$\sigma = \sqrt{E\left[\left(D_j - \overline{D}\right)^2\right]}. \tag{3}$$

In equation (3), $E[\ldots]$ is the average pitch, j is [1, 2, ..., n], N is the number of audio samples, and D_j is the *j-th* pitch. The average energy is often used to measure the signal size of singing voice. The calculation method is shown in

$$E_n = \sum_{k=-\infty}^{+\infty} x^2(k)w(n-k), \tag{4}$$

E_n in equation (4) is the energy in a short time, the input signal is represented by $x(k)$, and $w(n-k)$ is the window function.

Formant perturbation is a parameter used to measure the change value of formant in the corresponding period, which can reflect the voice quality of singers and evaluate their technical level. The first formant perturbation is defined by

$$\frac{1}{N-1}\sum_{i=1}^{N}\left|\frac{1}{F_{1i}} - \frac{1}{F_{1(i-1)}}\right|. \tag{5}$$

In equation (5), F_{1i} represents the first formant of i cycles, $F_{1(i-1)}$ represents the first formant of $i-1$ cycles, and N represents the number of audio samples. The third formant perturbation is defined by

Table 1: Comparison of window functions.

—	Rectangular window function	Hanning window function	Hamming window function
B/Δw	0.87	1.68	1.5
A/dB	14	34	45

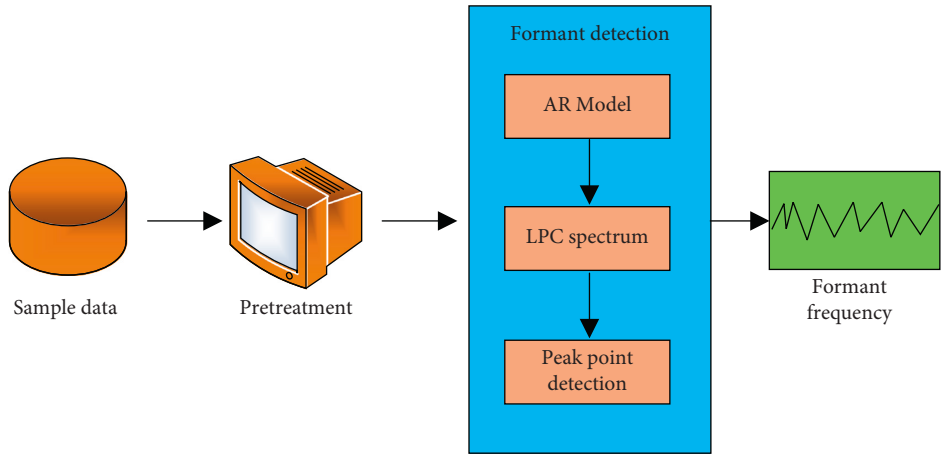

Figure 2: AR model peak detection.

$$\frac{1}{N-1}\sum_{i=1}^{N}\left|\frac{1}{F_{3i}}-\frac{1}{F_{3(i-1)}}\right|. \tag{6}$$

In equation (6), F_{3i} represents the first formant of i cycles, $F_{3(i-1)}$ represents the first formant of $i-1$ cycles, and N represents the number of audio samples.

3. Objective Evaluation of Singing Voice Based on Improved 1DCNN

3.1. Improved 1DCNN Speech Recognition Algorithm. Convolutional neural network is a frequently used technology in the field of image. The difference between speech and image is that speech is a one-dimensional signal. If two-dimensional means similar to image parameters are used to extract speech signal, large errors will inevitably occur. Therefore, in order to preserve the one-dimensional features of speech signal, a convolutional neural network based on one-dimensional vector is proposed. In order to reduce the training time of convolutional neural network and ensure the accuracy of calculation, the fractional order processing node theory is proposed for the training function to reduce the training time.

Sigmoid function is the default activation function of neuron, because sigmoid function is continuous and differentiable everywhere in the definition domain. At the same time, it can be interpreted as the probability of occurrence of events, so it becomes the activation function of neurons. In this paper, sigmoid is used as a training function to simulate the characteristics of biological neurons in convolutional neural network.

$$f(x)=\frac{1}{1+e^{-x}}, \tag{7}$$

x in equation (7) represents the training time, and the first derivative of the training function is expressed as

$$f'(x)=\frac{e^{-x}}{\left(1+e^{-x}\right)^{2}}. \tag{8}$$

It can be seen from equation (8) that when x is 0, the function has a maximum value, and there is no steep waveform of the function, which means that the fast convergence cannot be obtained. The 0.5th derivative of the training function is shown in

$$\begin{aligned} D^{0.5}f(x) &= \frac{1}{T_{(0.5)}}\int_{0}^{x}\frac{f^{(1)}(t)}{(x-t)^{0.5}} \\ &= \frac{1}{\sqrt{\pi}}\int_{0}^{x}\frac{e^{-t}}{(x-t)^{0.5}\left(1+e^{-t}\right)}\mathrm{d}t. \end{aligned} \tag{9}$$

From equation (9), it can be seen that the convergence speed is much faster than the first derivative when x is close to 0 or 1, which can greatly reduce the training time of the network.

The speech signal will be imported into 1DCNN structure after pre-emphasis. The speech signal is divided into multiple local speech by adding windows and frames, and multiple segmented speech is connected back and forth. This process is called long-term feature, and then speech recognition is carried out, as shown in Figure 3.

As shown in Figure 3, long-term features are imported into 1DCNN as input, and local convolution is performed in convolution layer to extract swimming information.

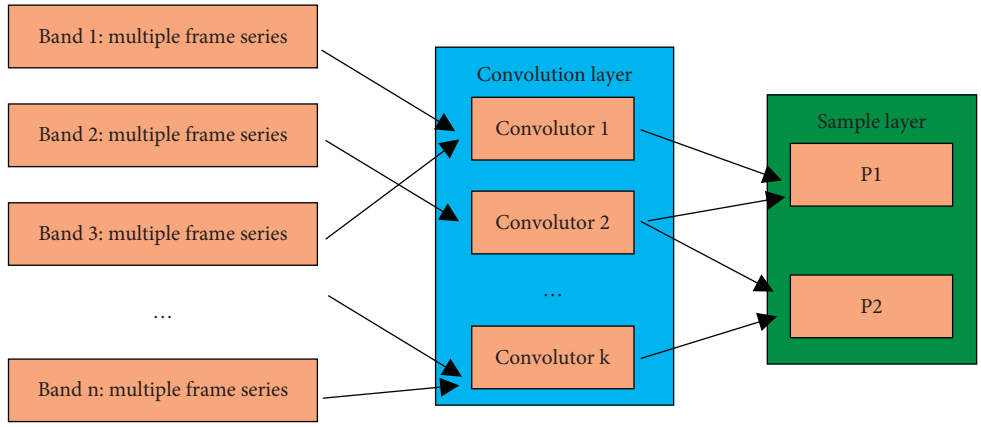

Figure 3: Convolution neural network recognition block diagram.

Convolution operation is performed on different filter bands by the same convolutor, and the output value obtained by convolution is taken as the output of the convolutor. The output formula is shown in

$$T_{i,k} = \theta \sum_{b=1}^{n-1} X_{b,k} Y_{b+i}^{T} + a_k. \tag{10}$$

In equation (10), n is the width of the convolver, Y_{b+i}^{T} is the input eigenvector of group i, $X_{b,k}$ is the weight of group k, and the network offset is set to a_k.

3.2. Evaluation System of Artistic Voice Recognition. With the interdisciplinary development, the intelligent objective evaluation method is gradually recognized by people in the music industry. This research will use the compact combination method to combine the wavelet theory and neural network, so as to carry out the objective evaluation of singing art voice. The neural network training process is based on the idea of error back-propagation. In order to get the minimum mean square error of actual output and predicted value, gradient descent search method is used to train the neural network. The calculation method of the output pattern vector of the sample in the training process is shown in equation (11), in which the number of neurons in the input layer is n, the number of neurons in the hidden layer is m, and the number of neurons in the output layer is N.

$$y_j(t) = f\left(\sum_{j=0}^{N} w_{kj} \psi_{(a,b)} \left(\sum_{j=0}^{m} w_{ik} x_i(t)\right)\right). \tag{11}$$

In equation (11), $f(\ldots)$ is the training function, w_{kj} is the weight from the input layer to the hidden layer, w_{ik} is the weight from the hidden layer to the output layer, the wavelet function is $\psi_{(a,b)}$, and the input mode vector is $x_i(t)$. Take the output mode vector into equation (12) to obtain the error function.

$$E = \frac{1}{2} \sum_{j=1}^{N} \left(y_j(t) - o_j\right)^2. \tag{12}$$

In equation (12), E is the error function of the training process, $y_j(t)$ is the output mode vector, and the expected output is o_j. Objective evaluation of singing art voice is carried out through neural network, and the evaluation process is shown in Figure 4.

As shown in Figure 4, the extracted acoustic feature parameters are normalized first to avoid the influence of scale and dimension. Then, it is input into the learning samples for data classification, which is one of the contents of the initial network establishment. At the same time, the initial weights are input through the subjective evaluation of professional teachers to construct the initial network. After training, the trained neural network can evaluate the input samples to be tested and finally get the evaluation results.

To improve the objective evaluation system of singing voice of 1DCNN, we need to design five parts. First, we need to design the input layer of 1DCNN. The input layer is composed of characteristic parameters, including the first formant, the first formant perturbation, the third formant, the third formant perturbation, fundamental frequency, range, fundamental frequency perturbation, and average energy. A complete speech signal is divided into 1000 frames, and feature parameters are extracted from each frame. In order to avoid the influence of noise, 10 frames of short-term signals are taken as a group to form a long-term signal. The characteristic parameters of the long-term signal are the average characteristic parameters of the short-term signals in the group. A total of 100 long-term signals are arranged to form one-dimensional eigenvectors, which are input to convolution operation of convolution layer. The number of neurons in input layer is 800. The network model architecture that deletes the maximum pool layer has a hierarchy problem. Therefore, first, reduce the number of elements of the feature map to be processed, and second, introduce the hierarchical structure of the spatial filter by making the observation window of the continuous convolution layer larger and larger (i.e., the proportion of the window covering the original input is larger and larger).

The second part is to design the convolution layer. A convolution layer is designed to avoid overfitting. 100 convolution cores are set in the convolution layer, and the edge feature information is retained by zero filling operation.

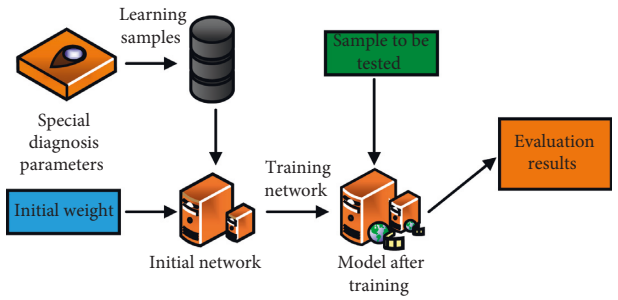

Figure 4: Voice evaluation process.

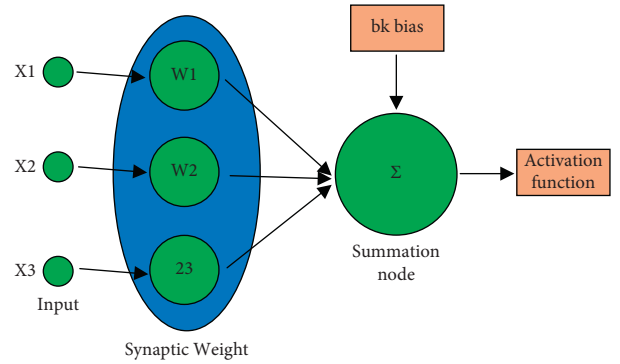

Figure 5: Principle of local convolution.

The convolution layer adopts local convolution, which can share weights. The specific local convolution process is shown in Figure 5.

As shown in Figure 5, the local convolution process requires the input of three neuron nodes and the output of only one neuron node. Among them, three synaptic weights are the values of convolution kernel. In addition, another parameter in the process is the offset BK, and four parameters converge at the junction node to sum up the input activation function.

The third part is the design of pooling layer. The pooling layer needs to sample the average or maximum of one-dimensional eigenvectors of convolution results to reduce the number of output nodes and avoid overfitting. According to the characteristics of speech signal, the maximum pooling method is used to reduce the estimation error caused by convolution layer parameter error and can retain more texture information of speech signal. If the step size of pooling region is set to 3, the tail of eigenvector needs to be zeroed, and the number of neurons is set to 2.67×104.

The fourth part designs the full connectivity layer and sets up 1024 full connectivity layer neurons. Finally, the output layer is designed and 500 output layer neuron nodes are set. The neuron nodes in the output layer are connected with all the neuron nodes in the full connection layer. The neuron of output layer corresponds to the category, and the category is set as the basis, with 5 points and 10 points as the full score. The activation function is used to get the probability of the output value in all categories, and the category value with the largest output probability.

The training of objective evaluation of the singing voice of the improved 1DCNN is the basis of ensuring high accuracy in practical application. In the training process of this network, the data of 1DCNN are set first, including training, verification, and test data. The training set is the parameter involved in the gradient descent process, and the data in the verification set are used to test the accuracy of the model, which can be improved by manually adjusting the parameters. The test set is used after the model training, which is the data set of the accuracy of the final model. 800 samples are set as the training set, and the number of samples in the verification set and the test set is 100, respectively.

Secondly, the parameters of 1DCNN are set, the learning rate is set to 0.1, the super-parameter batch size is the block size divided into the training process, which is set to 50, and epoch is a round-trip process of the data set in CNN, which is set to 10. Then, the forward operation of the training is carried out, and the network is run in the positive order. The influence of the previous neuron node on the subsequent layer of neuron nodes is calculated, and the forward weight and offset are calculated.

$$\text{net}_{h1} = i_1 \times w_1 + i_2 \times w_2 + b_1 \times 1. \tag{13}$$

In equation (13), h_1 is the neuron to be calculated, i_1 and i_2 are the values of the input neuron, w is the weight between the values of the input neuron and h_1, and b_1 is the offset.

The deviation of forward operation is very large, so it needs to use directional propagation to adjust and gradually reduce the deviation between the real value and the output value. In CNN, we need to train the data in each epoch, take 50 voice signals as a group, and continuously train and adjust the weight.

4. Experimental Results and Analysis

4.1. Comparison of 1DCNN and 2DCNN Evaluation Methods. In this study, 100 students of music major in a university were selected as the experimental subjects, so the subjects did not appear any disease within 3 months. The professional recording studio is selected as the recording environment, the noise is less than 45 dB, the voice acquisition equipment is Levitt LCT 940 professional recording microphone, the Fireface UCX computer sound card is used, the accompaniment is Roland electronic piano, the voice signal recording software is CoolEdit, and the recording processing is MIDI computer. The simulation environment is Intel (R) core (TM) i3-2310 m CPU 2.10 GHz (4 CPUs), and MATLAB r2016a is used for simulation.

Under 1DCNN and 2DCNN evaluation methods, the comparison of the influence of the number of time rule frames on the recognition rate is shown in Figure 6.

It can be seen from Figure 6 that the same network model will have a certain impact on the recognition rate when the number of time rule frames is different, and the greater the number of time rule frames, the higher the recognition rate of the network. And it is not difficult to see that under the same time rule frame number level, 1DCNN has a higher recognition rate in the evaluation of singing art voice. In addition, by comparing the convergence rates of the

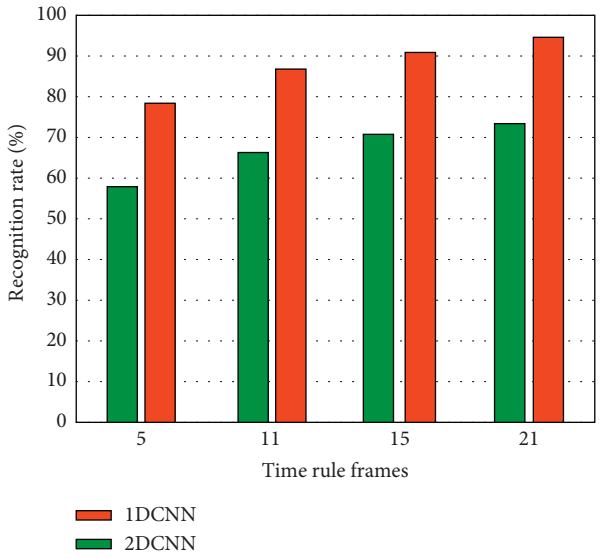

Figure 6: Influence of time rule frame number on recognition rate.

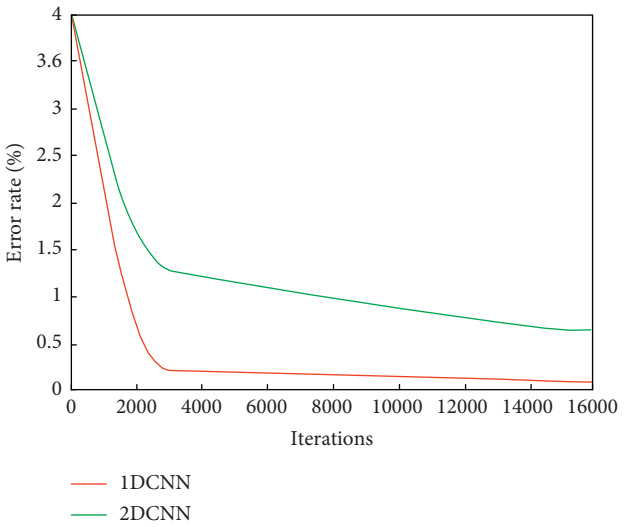

Figure 7: Error convergence trend.

two networks, the stability of the network is analyzed, as shown in Figure 7.

As can be seen from Figure 7, with the increase of the number of iterations, the error rates of the two networks also decrease. When the number of iterations is less than 3000, the slope of the convergence curve of 1DCNN is significantly larger than that of 2DCNN, which indicates that the convergence speed of 1DCNN is significantly faster than that of 2DCNN. The results show that 1DCNN has lower error rate and faster convergence speed than 2DCNN. The comparison of the experimental results of the two network models is shown in Table 2.

It can be seen from Table 2 that under the premise of the same convolver, the accuracy of one-dimensional convolutional neural network is as high as 86.3%, which is 5.9% higher than that of two-dimensional convolutional neural network, and the test time is only 125 S, which is far lower than 181 s of two-dimensional convolutional neural network. The results show that the one-dimensional convolutional neural network can obtain high accuracy and spend less time in evaluation.

Table 2: 1DCNN and 2DCNN experimental result comparison.

Algorithm	Number of convolutor	Accuracy (%)	Time (s)
1DCNN	100	86.3	125
2DCNN	100	80.4	181

4.2. Comparison of Different Neural Network Evaluation Methods. In general evaluation neural network, we can evaluate the neural network through some indicators. Improve our neural network through evaluation. The methods of evaluating neural network and machine learning are similar. Common methods include error, accuracy, R^2 score, etc. Through literature search, it is concluded that wavelet neural network and BP neural network are the more effective and commonly used objective evaluation methods at present. The objective evaluation results and subjective evaluation scores of the two networks and 1DCNN are compared, as shown in Figure 8.

It can be seen from Figure 8 that the highest score of experts for 100 experimental subjects is 9.8, and the lowest is 6.9. The deviation between BP neural network score and expert score is serious. The coincidence degree of wavelet neural network score and expert subjective evaluation result curve is much higher than that of BP neural network score. The score curve of improved 1DCNN evaluation method is almost consistent with that of expert. It shows that the improved 1DCNN evaluation method is closer to the subjective evaluation of experts. The two neural networks are compared with 1DCNN for objective evaluation and subjective evaluation error values, and the comparison results are shown in Figure 9.

As can be seen from Figure 9, since wavelet neural network is based on wavelet concept, it has more advantages in the whole network structure, and the learning and training process of wavelet neural network is simpler than that of BP neural network. Therefore, in the objective evaluation, although the BP neural network evaluation error for most of the samples is small, the maximum error value of BP neural network is 3.87, the error value between the sample numbers 41~71 fluctuates the most, the average error value during the period is as high as 2.76, and the total average error value is 1.48. The maximum error value of wavelet neural network is 1.12, there is no area with large error fluctuation, and its average error value is 0.84, so on the whole, the error value of wavelet neural network is smaller and more stable, and its performance is better. It is not difficult to see that the error value of the improved 1DCNN evaluation method is almost zero, and there is no error fluctuation in 100 sample tests. The maximum error value is only 0.34, and the average error value is 0.21, which is far less than the average error value of BP neural network evaluation of 1.48 and wavelet neural network evaluation of 0.84. The above results show that the stability of the improved 1DCNN

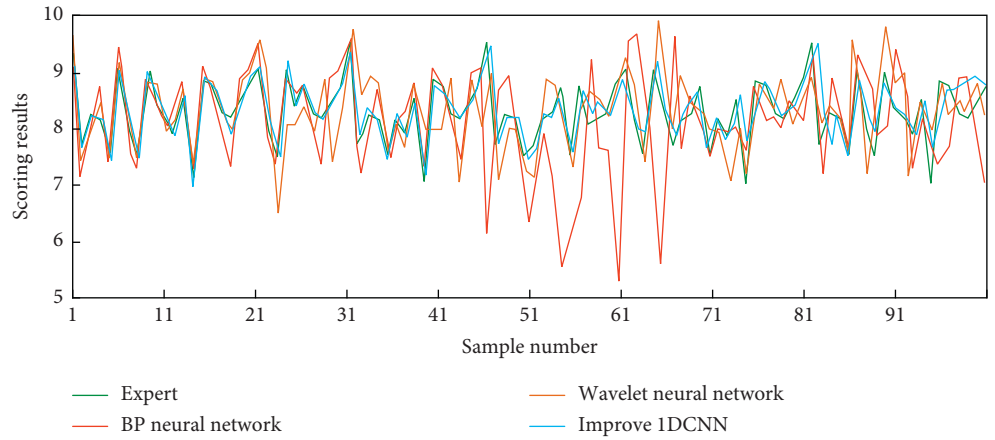

Figure 8: Comparison of a different neural network expert scoring.

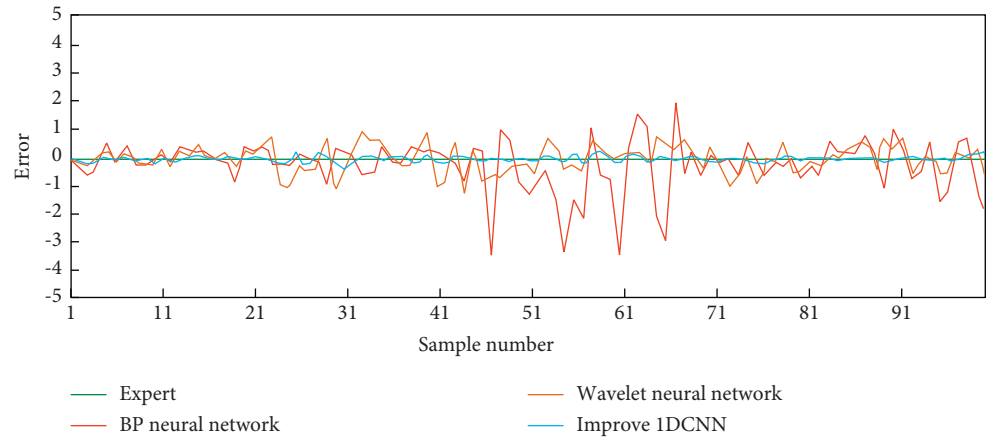

Figure 9: Error comparison of three neural networks.

evaluation method is better than other network evaluation methods, and the error range is very small, which can be ignored.

5. Conclusion

For how to recognize and objectively evaluate the singing voice, this paper proposes an improved CNN speech recognition method, which combines the idea of one-dimensional convolutional neural network with objective evaluation, and uses MATLAB for simulation experiment. The effectiveness and feasibility of the method are verified by simulation. This new method is established on the basis of one-dimensional convolutional neural network. The traditional 1DNN algorithm is optimized by fractional order processing node theory, and the corresponding parameters are set to build an objective evaluation system based on the improved 1DCNN. The final test analysis shows that, compared with 2DCNN, the improved 1DCNN has higher recognition rate, and the convergence speed has been improved, which shows that it is effective to use 1DCNN as the basic algorithm of singing speech recognition. At the same time, in order to verify the superiority of the method, through comparing with other neural networks, the results show that the evaluation method based on the improved 1DCNN has higher stability, and the error between the subjective score and the evaluation method is the smallest. Therefore, it is undeniable that the improved 1DCNN has superior performance and more scientific objective evaluation. To sum up, for the recognition of singing voice, the improved 1DCNN method can effectively process the voice signal with small distortion, and on this basis, it can also objectively evaluate the voice quality of singers, which promotes the scientific development of music talent selection. At the same time, the combination of music and computer technology can also promote interdisciplinary cooperation.

However, the improved 1DCNN method still has some problems in dealing with distorted sound signals. It is necessary to analyze its limitations and future research directions in the future.

Data Availability

The experimental data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declared that they have no conflicts of interest regarding this work.

References

[1] Z. Song, "English speech recognition based on deep learning with multiple features," *Computing*, vol. 102, no. 3, pp. 663–682, 2020.

[2] U. Nygren, M. Södersten, U. Thyen et al., "Voice dissatisfaction in individuals with a disorder of sex development," *Clinical Endocrinology*, vol. 91, no. 1, pp. 219–227, 2019.

[3] A. David and R. Rube, "Voice Function in Gender-Diverse People Assigned Female at Birth: Results From a Participant-Centered Mixed-Methods Study and Implications for Clinical Practice," *Journal of Speech Language Hearing Research*, vol. 62, no. 9, pp. 3320–3338, 2019.

[4] R. M. Bittner, A. Demetriou, S. Gulati et al., "An Introduction to Signal Processing for Singing-Voice Analysis: High Notes in the Effort to Automate the Understanding of Vocals in Music," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 82–94, 2019.

[5] R. Nanzaka, T. Kitamura, T. Takiguchi, Y. Adachi, and K. Tai, "Spectrum Enhancement of Singing Voice Using Deep Learning," *2018 IEEE International Symposium on Multimedia (ISM)*, vol. 1, pp. 167–170, 2018.

[6] K. L. Kim, J. Lee, S. Kum, C. L. Park, and J. Nam, "Semantic Tagging of Singing Voices in Popular Music Recordings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1656–1668, 2020.

[7] N. Saleem and M. Khattak, "Deep neural networks based binary classification for single channel speaker independent multi-talker speech separation," *Applied Acoustics*, vol. 167, Article ID 107385, 2020.

[8] K. W. Lin, B. T. Balamurali, E. Koh, S. Lui, and D. Herremans, "Singing voice separation using a deep convolutional neural network trained by ideal binary mask and cross entropy," *Neural Computing and Applications*, vol. 32, no. 4, pp. 1037–1050, 2018.

[9] B. Medeiros, J. Cabral, A. R. Meireles, and A. Baceti, "A comparative study of fundamental frequency stability between speech and singing," *Speech Communication*, vol. 128, no. 6, pp. 15–23, 2021.

[10] B. Lehner, J. Schluter, and G. Widmer, "Online, Loudness-Invariant Vocal Detection in Mixed Music Signals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 8, pp. 1369–1380, 2018.

[11] S. I. Mimilakis, K. Drossos, E. Cano, and G. Schuller, "Examining the Mapping Functions of Denoising Autoencoders in Singing Voice Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 266–278, 2020.

[12] Y. Murthy and S. G. Koolagudi, "Classification of Vocal and Non-vocal segments in Audio Clips using Genetic Algorithm based Feature Selection (GAFS)," *Expert Systems with Applications*, vol. 106, pp. 77–91, 2018.

[13] R. Meiyanti, A. Subandi, N. Fuqara, M. A. Budiman, and A. P. Siahaan, "The recognition of female voice based on voice registers in singing techniques in real-time using hankel transform method and MacDonald function," *Journal of Physics: Conference Series*, vol. 978, no. 1, Article ID 12051, 2018.

[14] Z. Zhao, X. Li, H. Liu, and C. Xu, "Improved Target Detection Algorithm Based on Libra R-CNN," *IEEE Access*, vol. 8, Article ID 114044, 2020.

[15] S. Souli, R. Amami, and S. B. Yahia, "A robust pathological voices recognition system based on DCNN and scattering transform," *Applied Acoustics*, vol. 177, no. 5786, Article ID 107854, 2021.

[16] T. S. Le, J. An, Y. Huang, Q. Vo, and Y. J. Kim, "Ultrasensitive anti-interference voice recognition by bio-inspired skin-attachable self-cleaning acoustic sensors," *ACS Nano*, vol. 13, no. 11, Article ID 13293, 2019.

[17] R. Yamashita, M. Nishio, R. K. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into imaging*, vol. 9, no. 4, pp. 611–629, 2018.

[18] Y. Zhang, T. H. Chang, L. Jing, K. Li, H. Yang, and P. Y. Chen, "Heterogeneous, 3D Architecturing of 2D Titanium Carbide (MXene) for Microdroplet Manipulation and Voice Recognition," *ACS Applied Materials & Interfaces*, vol. 12, no. 7, pp. 8392–8402, 2020.

[19] M. Alves, E. Krüger, B. Pillay, K. Lierde, and J. Linde, "The Effect of Hydration on Voice Quality in Adults: A Systematic Review," *Journal of Voice*, vol. 33, no. 1, pp. e13–e28, 2019.