International Journal of Web-Based Learning and Teaching Technologies
Volume 19 • Issue 1

A Method for Improving the Pronunciation Quality of Vocal Music Students Based on Big Data Technology

Dan Shen, Harbin University, China*
Wenjia Zhao, Harbin University, China

ABSTRACT

With the development of internet technology, big data has been used to evaluate the singing and pronunciation quality of vocal students. However, current methods have several problems such as poor information fusion efficiency, low algorithm robustness, and low recognition accuracy under low signal-to-noise ratio. To address these issues, this article proposes a new method for evaluating sound quality based on one-dimensional convolutional neural networks. It uses sound preprocessing, BP neural networks, wavelet neural networks, and one-dimensional CNNs to improve pronunciation quality. The proposed 1D CNN network is more suitable for one-dimensional sound signals and can effectively solve problems such as feature information fusion, pitch period detection, and network construction. It can evaluate singing art sound quality with minimum errors, good robustness, and strong portability. This method can be used for the evaluation and diagnosis of voice diseases, helping to improve students' professional abilities.

KEYWORDS

Acoustic Parameters, Artistic Voice, Big Data, Convolutional Neural Network, Quality Evaluation, Quality Improvement

INTRODUCTION

Music is a field that has an extensive influence on human beings and the natural world. People's love for a good voice in the popular sense is inestimable. Today, with the development of science and technology, big data technology has been integrated into people's lives, and it has a wide audience in the field of singing. For example, China's most influential application products such as "Sing Ba" and "National K Song" have as many as 700 million users. Among these products, the popularity of the singing scoring system of the Chinese Academy of Sciences demonstrates people's appreciation for singing and the need for the cultivation and practice of high quality vocal music.

The artistic expression of the voice in opera, stage, film, television, and radio is referred to as artistic voice in the industry (Huang, 2022). There are diversities of styles and evaluation standards, but good voices with many popular styles must have something in common, and this commonality

DOI: 10.4018/IJWLTT.335034 *Corresponding Author

can be extracted as a popular evaluation standard for good voices. The traditional singing evaluation includes standard pitch lines. As long as the pitch of the user's singing is aligned with the pitch line, he can achieve high scores. At the same time, there is an evaluation system that seems to be a parallel world, that is, the evaluation of vocal music teachers or music critics. The evaluation criteria of singing include many aspects, such as air sinking, head cavity resonance, and accurate rhythm. The traditional singing voice evaluation standards are mostly judged by people, which are highly subjective and lack scientific explanation and objective evaluation standards. In addition, due to people's neglect of talent training in the field of singing, the influence of cognitive limitations and uselessness, the research and development of singing science is seriously restricted, resulting in a current lack of professional talent in the field of singing (Hao, 2021). According to the Chinese Music Industry Players Survey Report released in 2017, the talent shortage is one of the three biggest problems facing Chinese music start-up companies, but the cultural industry including the music field, is also affected by the current global economic downturn. As an industry that maintains a high rate of growth, music is the most important part of the cultural industry. Therefore, the cultivation of singing artists is imperative (Yang & Yue, 2020). The process of cultivating singing artists includes many aspects such as the selection of voice talents, voice guidance, singing training, and voice maintenance (Zhao & Jin, 2022). In view of the subjective limitations of the traditional singing voice evaluation process, it is hoped to scientifically define singing voice quality evaluation criteria by means of scientific research and establish a complete set of singing voice objective evaluation systems. The evaluation system further promotes the development of subjects such as vocal music education and voice medicine, as well as the cultivation of talents and the maintenance of voice (Syafitri et al., 2018).

The physical attributes of the singing voice include sound quality, tone intensity, pitch, timbre, tone length, and breathing stability (Atmowardoyo et al., 2020), subjectivity, empiricism, and abstraction (Hsieh et al., 2020). For example, in singing competitions, there are two common evaluation methods of singing level. One is quantitative expression by one or more professionals (O'Brien et al., 2018). This evaluation method is relatively fair, but it is closely related to the singer's singing state and the evaluator's preference, and there are large subjective factors, and the evaluation results have large errors. The second is an evaluation team composed of a large number of audiences, and errors are eliminated to a certain extent according to the comprehensive evaluation of the evaluation team, but the accuracy of the evaluation results is also greatly affected due to the inconsistency or large gap in the professional quality of the audience (Oh & Song, 2021). Therefore, the subjective evaluation method is not only inefficient, but its accuracy is also questioned. In recent years, due to the proposal of convolutional neural networks and the continuous emergence of related achievements in deep learning technology, artificial intelligence technology has continued to develop and mature; it has been successfully applied to the fields of image, speech, and art, and has made breakthroughs in accuracy. Face++'s face recognition technology and iFLYTEK's speech recognition technology are among the top in the world. The application of artificial intelligence in the field of art also includes rendering of art paintings and style transfer technology. It is challenging to apply artificial intelligence technology to the field of voice, such as in the evaluation of voice quality, the fusion of multi-style voice characteristics, and the identification of voice diseases; however, it is very meaningful for the cultivation of singing talent (Kessler, 2018).

Ultimately, because the quality and state of the singing voice are of great significance to the scientific selection, teaching, training and diagnosis of voice diseases, it is necessary to study an effective method to scientifically evaluate them. In this paper, and in consultation with a large number of documents, the voice evaluation parameters are scientifically explained and defined, and a complete set of singing voice evaluation parameters is proposed. For noise-polluted sound signals, a pitch period extraction based on wavelet transformation and fourth-order statistics is proposed. In addition, the algorithm of singing voice evaluation was studied in detail by using a convolutional neural network, a one-dimensional convolutional neural network evaluation algorithm suitable for

singing voice signal was proposed, and the validity and reliability of the algorithm in this paper were verified by experiments.

The innovative contribution of this paper lies in the scientific explanation and definition of voice evaluation parameters, and a set of complete voice evaluation parameters of singing is proposed. A pitch period extraction method based on wavelet transformation and fourth-order statistics is proposed for noise polluted sound signals. The proposed evaluation method can better solve the problems of feature information fusion and utilization and low signal-to-noise ratio. The problems of pitch period detection, one-dimensional convolutional neural network construction, and training efficiency are solved. The method proposed in this paper can be used not only for the evaluation of singing voices, but also for the diagnosis of voice diseases. Through accurate and effective phonetic assessment, students' pronunciation problems can be found and corrected in time, which is helpful to the improvement of students' professional abilities. The proposed method can evaluate the quality level of singing art sound, with small comparison error, good robustness, and strong portability.

LITERATURE REVIEW

Effective vocal music education should be scientific, systematic, and incorporate multiple art forms. Zhang and Liu (2018) argue that dance training is essential for vocal music education, particularly for popular songs and musicals. Separating singing from dance training is inappropriate. Comprehensive vocal music training should include singing techniques, music theory, rhythm, tone quality, and expressiveness. Incorporating other art forms such as dance, theater, and visual arts can enhance the overall artistic level of music performance. A cross-disciplinary approach should be adopted to create diverse and personalized musical performances, promoting a holistic education in vocal music. For example, in opera, vocal performance and dance performance are integrated (Tejedor-García et al., 2020). A good operatic performer must execute a strong combination of songs and dances. In recent years, in some large-scale literary and artistic performances, the perfect combination of song and dance has been praised (Zhou, 2020). In opera, musicals, song and dance dramas, and song and dance parties, the full combination of body and vocal singing makes the performance effect perfect. The perfect combination of vocal music and dance (physical training) is that the song and dance art we see in the literary and artistic performance activities not only meets people's aesthetic needs in many aspects, but also provides people with aesthetic enjoyment that is pleasing to the eyes and consistent with audio-visual enjoyment (Geng, 2021).

Research on singing voice primarily includes two areas: (1) research on the acoustic evaluation parameters of the singing voice signal, and (2) research on the establishment of an objective evaluation method for the singing voice (Lange & Costley, 2020). In China, since the late 1970s, research on voice has been conducted by scholars and experts all over the country (Wang & Wu, 2021). In the early days, according to Foote & McDonough (2017), researchers in the process of exploring how to evaluate singing voice and teaching vocal music and found that factors affecting the objective evaluation results of the entire artistic voice include fundamental frequency and average energy. Yu & Xiong searched for evaluation parameters from the perspective of spectrum analysis. This objective evaluation method only obtained the relationship between vibrato and formants, the method was limited, and the recognition rate was low (Yu & Xiong, 2022). In the latest research on objective evaluation of voices, Professor Yuan Jian of the Xi'an Conservatory of Music used the Radial Basis Function (RBF) method to build a model network for objective evaluation of voices and transformed qualitative evaluations into quantitative score vectors. The output is thus scored (Jing, 2022).

In this paper, a one-dimensional convolutional neural network method is used to establish an objective evaluation model network for voice. From a probabilistic perspective, RBF does not have the same good probabilistic characteristics as Softmax. As the number of layers of the network increases, both the complexity of the network structure and the weight parameters that need to be

trained increase accordingly. However, one-dimensional convolutional neural networks have three prominent features: local connection, weight sharing, and pooling operations. Therefore, the network has a simple structure, fewer weight parameters, good training effects, and high classification accuracy. Voice audio samples are one-dimensional signals, and the evaluation model of one-dimensional convolutional neural networks can more objectively reflect their characteristics.

METHODOLOGY

Recognition Method of Artistic Voice

Pre-Emphasis Technology

There are many reasons for the weakening of the signal, but the high frequency is much more affected in the oral lip than the low frequency part (Jiang, 2019). In order to facilitate the analysis of the most original signal, the voice should be pre-emphasized to increase the high-frequency components, so that the signal is closer to the original signal. In the experiment, the pre-emphasis is usually carried out by means of a filter, an FIR digital filter is generally selected, and its transfer function is shown in formula (1):

$$H(z) = 1 - \mu z^{-1} \tag{1}$$

Windowing and Framing of Artistic Voice

Artistic voice has no periodicity, nor is it a regular signal. Both the sampling value and the characteristic parameters will change irregularly with time. But it is customary to think that voice can be regarded as a stable signal in a very short period of time, which is generally 10~30ms (Saito & Akiyama, 2017). Therefore, voice can be identified as having two basic characteristics of time-varying and short-term stability (Zou et al., 2021). According to the above characteristics of voice, before processing the voice, it is often divided into several segments to ensure that these voice segments have short-term stability; this operation is called framing (Wu et al., 2022).

The window processing function expression is shown in formula (2)

$$s_w(n) = s(n)w(n) \tag{2}$$

In formula (2), s(n) represents the original signal, and w(n) represents the windowing function used. Through the movement of the window function in the original signal, the voice can be effectively divided into several smooth fixed lengths, so as to realize the segmentation of the original voice (Jiang et al., 2021). This process is called windowing. The window function directly affects the characteristics of several voice segments, and the selection of the window function will directly affect the objective evaluation of the voice (Fouz-González, 2020).

Commonly used window functions generally include rectangular window, Hamming window and Haining window. The expression of the window function is as follows:

The rectangular window is

$$w(n) = \begin{Bmatrix} 1, 0 \le n \le wlen - 1 \\ 0, \text{n is other value} \end{Bmatrix} \tag{3}$$

Hamming window is

$$w(n) = \left\{ \begin{array}{l} 0.54\text{-}0.46\cos\left[\dfrac{2\pi \mathrm{n}}{(wlen-1)}\right], 0 \le n \le wlen-1 \\ 0, \mathrm{n} \text{ is other value} \end{array} \right\} \tag{4}$$

Haining window is

$$w(n) = \left\{ \begin{array}{l} 0.5\left[1\text{-}\cos(\dfrac{2\pi \mathrm{n}}{wlen-1})\right], 0 \le n \le wlen-1 \\ 0, \mathrm{n} \text{ is other value} \end{array} \right\} \tag{5}$$

Selection should also consider other factors, such as the length of the window function, which is not as long as possible, nor as short as possible (Vijayan et al., 2018). The frequency characteristic of the window function is:

$$\Delta f = \frac{1}{NT_S} \tag{6}$$

Acoustic Parameter Extraction of Voice

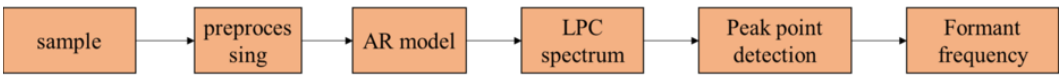
Voice is a special art with four basic characteristics: pitch, intensity, timbre, and length. The acoustic meaning of pitch is the frequency with which the vocal cords vibrate. According to the principle of acoustics, "long, large, thick, and loose objects vibrate slowly and have a low frequency; on the contrary, short, small, thin, and compact objects vibrate fast and have a high frequency." The acoustic meaning of sound intensity is the amplitude of the vibration of the vocal cords. The process is: "The vibration of vocal cords is caused by the movement of breath."(Manson, 2013) The acoustic meaning of timbre is the uniqueness of vocal cords. This is the personality of the voice, and the personality of the voice is the timbre. The acoustic meaning of pitch is the length of time the vocal cords continue to vibrate (Wang & Liu, 2022).

There are many methods for extracting formants, such as the bandpass filter combination method, the spectrum reciprocal method, the LPC (Linear Predictive Coding) detection method and the AR(Augmented Reality) model detection method. After calculation and comparison, the AR method was selected in this experiment to extract the first and third formants (Sun, 2019). Because it provides a good vocal tract model (provided that the artistic voice samples are basically free of noise). The extraction process is shown in Figure 1.

The experimental steps are:

(1) Preprocess the original noisy signal, extract useful information segments from the processed signal, and combine these information segments into a voice segment.

Figure 1. Peak detection of AR model



In order to solve the defects of traditional signal noise reduction technology and improve the reliability and practicability of fan fault detection technology, we propose a noise reduction pretreatment technology based on the pre-whitening method. The correlation between the sudden change of signal energy in a certain frequency band and the signal is removed by pre-whitening the collected signal. The noise signal is uniformly distributed in each frequency band to optimize the subsequent results of signal extraction and separation, thereby improving the reliability of signals required for signal detection and fault diagnosis.

(2) Arrange the extracted voice segments and filter out the DC(Direct Current) components in them, set the signal-to-noise ratio to superimpose the noise, and finally, use the DWT(Discrete Wavelet Transform) wavelet transform to reconstruct the target signal with the obtained low-frequency coefficients structure (Zhang & Tsai, 2021).
(3) Perform pitch detection on the reconstructed artistic voice segment.

The common sound range estimation method is to take the maximum and minimum values of the pitch of songs and song scores and use the mean and standard deviation to improve the accuracy. Therefore, in order to obtain the sound range, we must first obtain the value of the pitch. The calculation of the pitch D is described below.

Pitch is determined by the vibration frequency of the object (vocal cord). That is, the more times the object (vocal cord) vibrates in a certain period of time, the higher the pitch; conversely, the lower the number of vibrations, the lower the pitch. Lin Tao said in *Beijing Phonetics Experiment Record*, "tone and intonation are the performance of pitch." However, pitch is abstract, and its value can only be obtained from the pitch curve. Therefore, D is often used to describe pitch in experiments, which can reflect the value of pitch, so the definition of pitch D is:

$$D = 12 * \log_2(F \div F_0) \tag{7}$$

The average value D of all pitches in the singing voice and score is shown in formula (8):

$$\overline{D} = \frac{1}{N}\sum_{j=1}^{N} D_{\mathrm{j}} \tag{8}$$

The standard deviation of all pitches in the singing voice and score is shown in formula (9):

$$\sigma = \sqrt{E\left[\left(D_J - \overline{D}\right)^2\right]}, \mathrm{j} = 1,2,\ldots N \tag{9}$$

Formant perturbation is defined as the rate of change between the formant of one cycle and the formant of the next cycle. The formant perturbation is often used to measure the formant variation in the corresponding period, and its essence is to reflect the quality of the singing voice or the skill level of the singer.

The perturbation of the first formant is a measure of the rate of change of the first formant between adjacent periods, and the perturbation of the third formant is a measure of the rate of change of the third formant between adjacent periods.

The mathematical definition of the first formant perturbation is shown in formula (10)

$$\frac{1}{N-1}\sum_{i-1}^{N}\left|\frac{1}{F_{1i}}-\frac{1}{F_{1(i-1)}}\right| \tag{10}$$

The mathematical definition of the third formant perturbation is shown in formula (11)

$$\frac{1}{N-1}\sum_{i-1}^{N}\left|\frac{1}{F_{3i}}-\frac{1}{F_{3(i-1)}}\right| \tag{11}$$

The definition of average energy is the representation of a semaphore in an identical environment. Average energy is often used to measure the relative magnitude of the singing signal. See formula (12) for the mathematical definition of average energy.

$$E_n=\sum_{k=-\infty}^{+\infty}x^2(k)w(n-k) \tag{12}$$

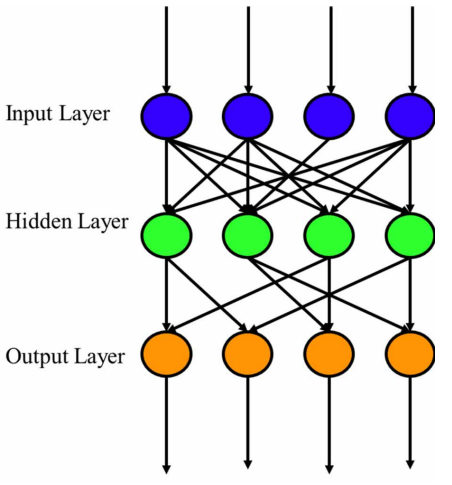
Objective Evaluation Method of Artistic Voice

In the MatlabR2016a environment, two network structures, BP neural network and wavelet neural network, are established for the objective evaluation of singing, paving the way for the comparison of the one-dimensional convolutional neural network . The purpose is to find a more objective, fair, and accurate objective evaluation method of voice.

Evaluation Method of Voice Recognition Based on BP Neural Network

BP neural network has been widely used in various fields, and it is also the most studied artificial intelligence network model. The BP neural network model is a feedback network model proposed by Rumelhart et al. in 1985, and its structure is shown in Figure 2. Theoretically, the three-layer BP neural network model can be used to realize any continuous image.

Figure 2. BP network structure



The steps of objectively evaluating artistic voice by BP neural network:

1. Normalization of data

The formula is

$$\hat{x} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{13}$$

Normalize the extracted acoustic feature parameters, because this can eliminate the influence of scale and dimension;

2. Classification of data

Under the subjective evaluation of the professional vocal music teacher and the president of the Artistic Voice Association as judges, the scores were given as the target value of the training network output;

3. Establishment of BP neural network

The establishment process of the entire network includes the design of the input layer and the output layer, followed by the determination of the number of neurons in the hidden layer, and finally the determination of the training function in the hidden layer;

4. Network training and learning process
5. The prediction process is the evaluation process

Evaluation Method of Voice Recognition Based on Wavelet Neural Network

Since the 1980s, the wavelet theory has been used in many fields, and it has also been expanded significantly. At present, with the rise of artificial intelligence neural networks, the combination of interdisciplinary subjects is increasingly pursued. As wavelet theory and neural networks are combined, the question of whether the wavelet characteristics and the advantages of neural networks can be combined has aroused great thought by scholars. There are two main research ideas: 1) Using the wavelet theory to preprocess the signal. Because the wavelet transform has the local characteristics of time and frequency, the feature extraction of the signal is realized by the wavelet, and then the extracted feature vector is input into the network, and 2) Using wavelet neural network (WNN) or wavelet network. The training function is replaced by a wavelet function, because the wavelet has a zoom property.

The objective evaluation process of singing artistic voice by wavelet neural network:

1. Data normalization, using the following formula:

$$\hat{x} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{14}$$

After the acoustic features are extracted, the acoustic features are transformed from one space to another. In this space, the more characteristic parameters are more consistent with a certain probability

distribution, and the dynamic range of the characteristic parameter value range is compressed. This reduces the mismatch between training and test environment and improves the robustness of the model, which is actually the operation.

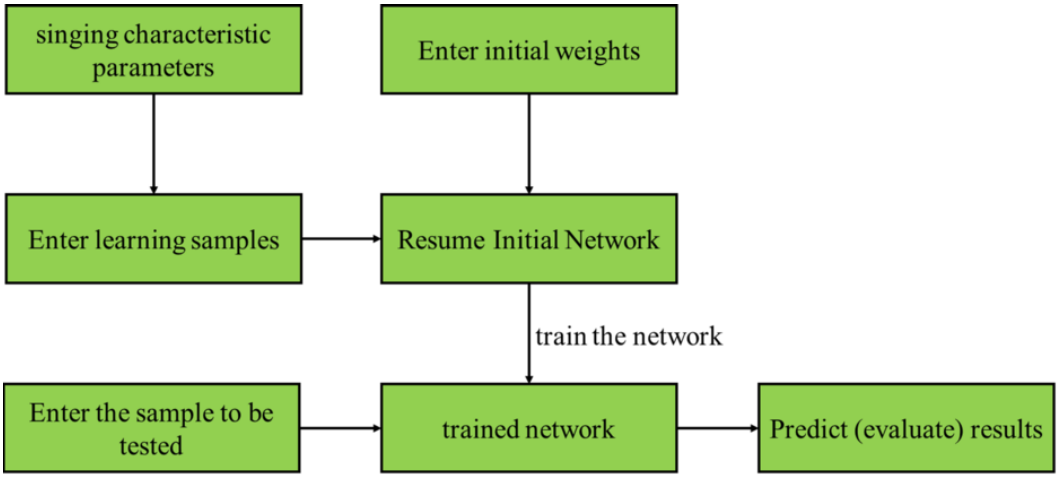
2. The classification of data, the scores are given respectively under the subjective evaluation of the professional vocal music teacher and the president of the Artistic Voice Association as judges, as the target value of the output of the training network, and this is used as the output sample;
3. The establishment of the wavelet neural network, the establishment of the whole network includes the design of the input layer and the output layer, followed by the determination of the number of neurons in the hidden layer, and finally the determination of the training function in the hidden layer;
4. Network training and learning process;
5. The prediction process is the evaluation process, as shown in the Figure 3.

Voice Recognition Based on One-Dimensional Convolutional Neural Network

The steps of improving one-dimensional convolutional neural network for objective evaluation of artistic voice are:

(1) Input the original speech signal and perform windowing and framing preprocessing. In this paper, the length of the original voice signal is 30s, and the frame length is 30ms; that is, a complete voice signal can be divided into 1000 frames of short-term voice signals.
(2) Feature parameter extraction. Extract eight characteristic parameters for a frame of short-term speech signal – the first formant, the first formant perturbation, the third formant, the third formant perturbation, the fundamental frequency, the sound range, the fundamental frequency perturbation, and the average energy.
(3) Construct the one-dimensional input signal of the convolutional neural network. In order to improve the robustness and fault tolerance of features, this paper combines 10 frames of short-term speech signals into a long-term signal, the features of a long-term signal are also set to eight parameters, and the size is eight features of 10 frames of short-term speech signals the respective mean values of the parameters.

Figure 3. The objective evaluation process of singing voice with wavelet neural network



(4) Normalization of data. The formula used is shown in formula (15)

$$\hat{x} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{15}$$

The extracted acoustic feature parameters are normalized, as this removes scale and dimension effects.

(5) Classification of data. Under the subjective evaluation of professional vocal music teachers and the president of the Artistic Voice Association as judges, the scores were given as the target value of the output of the training network.
(6) The establishment of one-dimensional convolutional neural network. The establishment process of the entire network includes the design of the input layer and the output layer, followed by the determination of the number of neurons in the hidden layer, and finally the determination of the training function in the hidden layer.
(7) Network training and learning process.
(8) The prediction process is the evaluation process.

RESULTS, ANALYSIS, AND DISCUSSION

Sound Pre-Processing

Most of the energy is concentrated in the low frequency range. This may cause the signal-to-noise ratio of the high frequency end of the message signal to drop to an unacceptable level. However, because the energy of the higher frequency component in the message signal is small, there is rarely enough amplitude to generate the maximum frequency offset, so the signal amplitude that generates the maximum frequency offset is mostly caused by the low-frequency component of the signal. Thus, the higher the frequency, the smaller the energy contained. In this way, the spectrum needs to be de balanced. A measure called pre-emphasis and de-emphasis is adopted. The central idea is to effectively process the signal by using the difference between the signal characteristics and the noise characteristics. That is, before the noise is introduced, an appropriate network (pre-emphasis network) is used to artificially emphasize (enhance) the high-frequency component of the transmitter input modulation signal. Figure 4 shows a comparison of the pitch spectrum of a voice sample |a| before and after pre-emphasis. As can be seen from the figure, in the absence of pre-emphasis, the amplitude of the fundamental line is large, which will affect the calculation of the spectral envelope; while in the pre-emphasis experiments, the high-pass filter suppresses the amplitude of the fundamental line, thereby increasing the amplitude of the high-frequency components and reducing the turbulent amplitude of the spectrum. In the analysis of the actual situation, the pre-emphasis technology is usually adopted; that is, the high-pass filter is connected after the signal is sampled, which can reduce the influence of the glottal pulse and obtain more channel parameters.

In Figure 5, wlen is the frame length, and inc is the frame shift. The length of the frame shift is often within half of the frame length, that is, inc<×wlen. Because such a value can ensure the smoothness of overlapping frames, it ensures the short-term stability of artistic voice segments. Only by being able to accurately identify artistic voices can we transmit them to big data for accurate sound analysis and effectively evaluate the pronunciation of vocal music students.

BP Neural Network

Enter the audio data into the network structure that was previously saved, forecast the audio data, and get the final output. The objective evaluation result of artistic voice based on BP neural network

Figure 4. Comparison of pitch spectrum of voice sample |a| before and after pre-emphasis

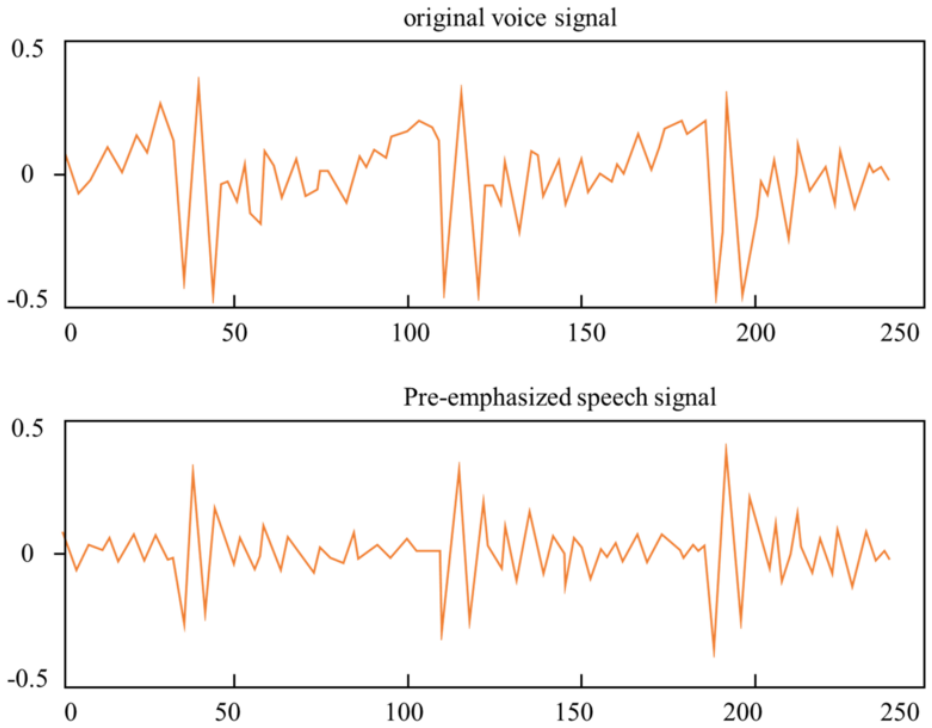
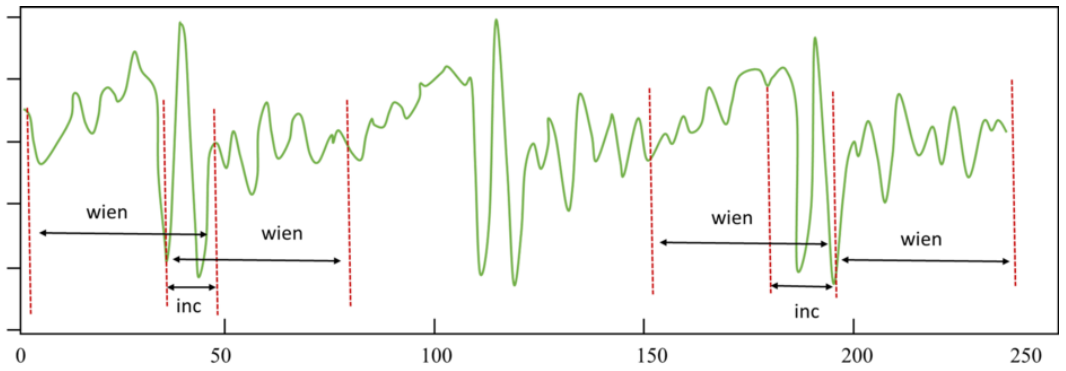


Figure 5. Schematic diagram of framing



is obtained by reverse normalization. The actual output is compared with the target value to obtain its error relative to the target value.

Since each person in this database has 10 voice samples, eight are used for training, one is used for verification, and one is used for testing, so the test in this paper can obtain the scoring results of 100 people. According to the scoring results and errors obtained by the BP neural network, compared with the subjective scoring (expert scoring), the two evaluation results are similar, indicating that the BP neural network method is feasible, but the error is large.

It can be seen from Figure 6 that the BP neural network scoring result is similar to the expert scoring trend, but the error fluctuates greatly, indicating that its accuracy is not high. Therefore, we need to replace the wavelet neural network in the next step to further evaluate the students' voices in order to achieve the best results.

Objective Evaluation Based on Wavelet Neural Network

According to the scoring results and errors obtained by the wavelet neural network, compared with the subjective scoring (expert scoring), the two evaluation results are similar, indicating that the method of the wavelet neural network is feasible, however the error is large. According to the experimental results, the wavelet neural network is obtained. The comparison chart with the expert scoring results is shown in Figure 7.

It can be seen from Figure 7 that the scoring results of the wavelet neural network are similar to those of the experts, indicating that the method is feasible, the error fluctuation is smaller than that of the BP neural network, and it is basically stable, indicating that compared with the BP neural network, the wavelet neural network has better stability and accuracy. The wavelet neural network can simulate the expert's scoring, and to a certain extent, it can help the students' vocal music learning well and provide students with better vocal music guidance. Using the wavelet multiresolution analysis method,

Figure 6. Expert and neural network scoring errors of students' voices

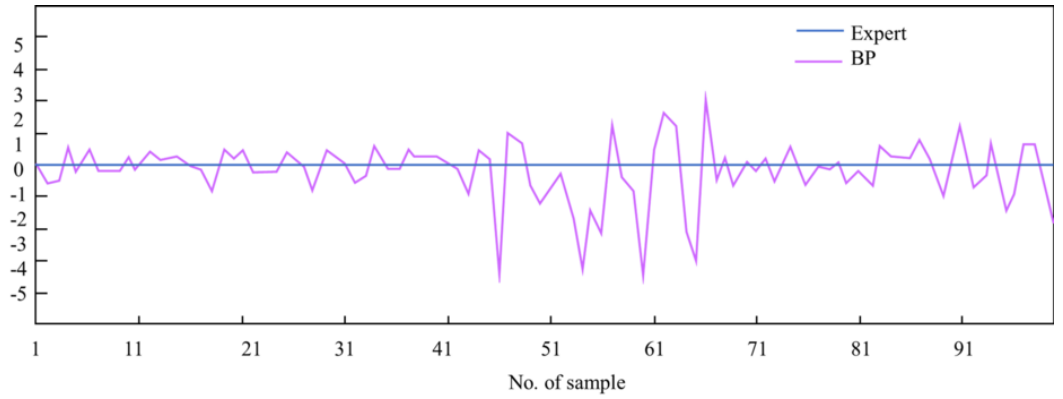
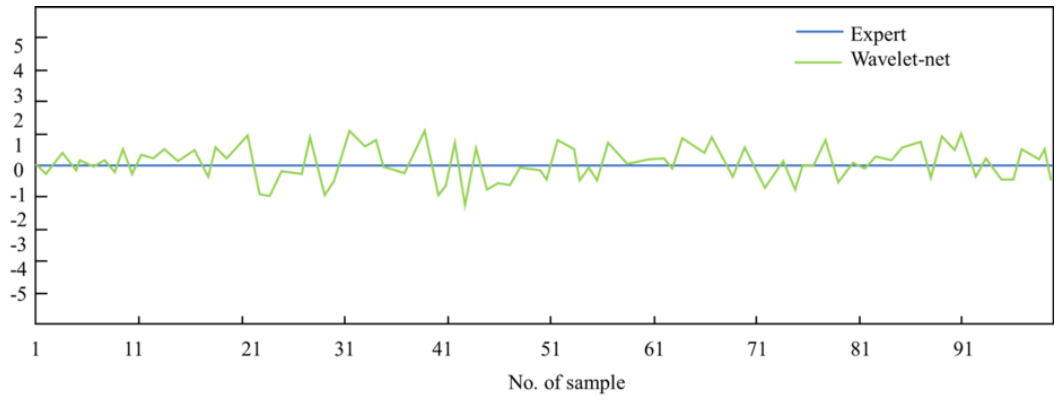


Figure 7. Scoring errors of students' voices by experts and wavelet neural networks



the output total error signal of the dynamic system to be studied is decomposed into independent frequency bands. The vector values in each frequency band represent different sub error signals. The wavelet multiresolution analysis method used as the input eigenvector of BP network, and the neural network is used for function approximation to obtain the final error tracing result. However, there is still a certain gap between the wavelet neural network and the expert scoring, and the stability is still not enough. Therefore, we intend to further improve the method.

Objective Evaluation Based on One-Dimensional Convolutional Neural Network

Comparing the experimental results with the results of expert scoring, as shown in figure 8.

It can be seen from Figure 8 that the scoring results of the one-dimensional convolutional neural network are similar to the experts' scoring trend, indicating that the method is feasible, and the error fluctuation is much smaller than that of the BP neural network and the wavelet neural network, and it is basically stable. In comparison, the one-dimensional convolutional neural network has the best stability, the smallest error, and the highest accuracy. Therefore, we intend to use a one-dimensional convolutional neural network to evaluate the pronunciation of vocal music students, because this method has the smallest error and is the closest to the actual scoring result, which is also the most effective method. Through this method, we can save human and other resources, while also giving students the best guidance.

In order to fully illustrate the scientific nature of the method in this paper, the following is a comprehensive comparison from the influence of time regularization, convergence speed, and the correct rate of recognition.

In order to compare the feasibility of the algorithm more fairly and effectively, before the samples are input into the trained network, the feature parameters of the input samples must first be unified in size, that is, set to the same number of frames. This process is called time-warping the number of frames. The relationship between the number of time rule frames and the recognition rate is shown in the Figure 9 below.

Because the characteristic parameters contained in different time-warped frames are different, their recognition rates are also different. When the number of time-warped frames increases, the recognition rate of both networks is improved, because the neural network originally has more feature parameters, and the recognition rate is higher. The higher the number of recognition frames, the better the sound quality can be analyzed to achieve the best evaluation effect.

In this paper, the influence of the structure of the neural network on the output value is analyzed when the learning rate, weight initialization criteria, and convergence criteria are all unchanged. The results show that the output value of the neural network is a valuable index to judge the network

Figure 8. Errors in scoring results of students' voices by experts and one-dimensional convolutional neural networks

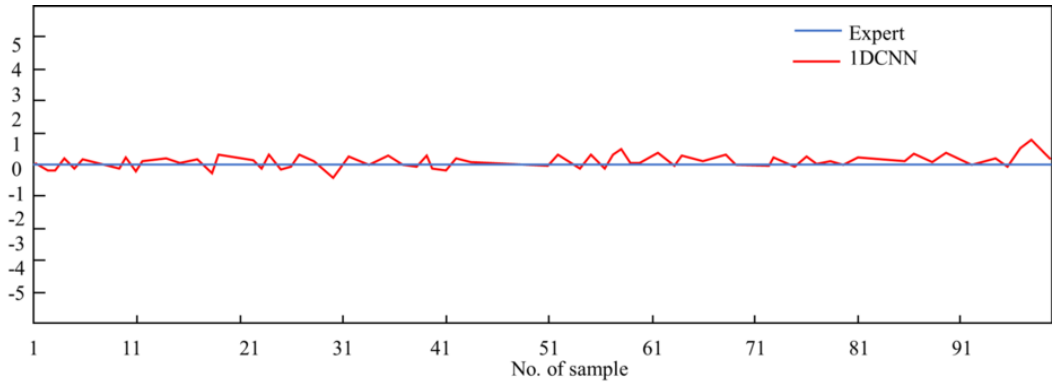
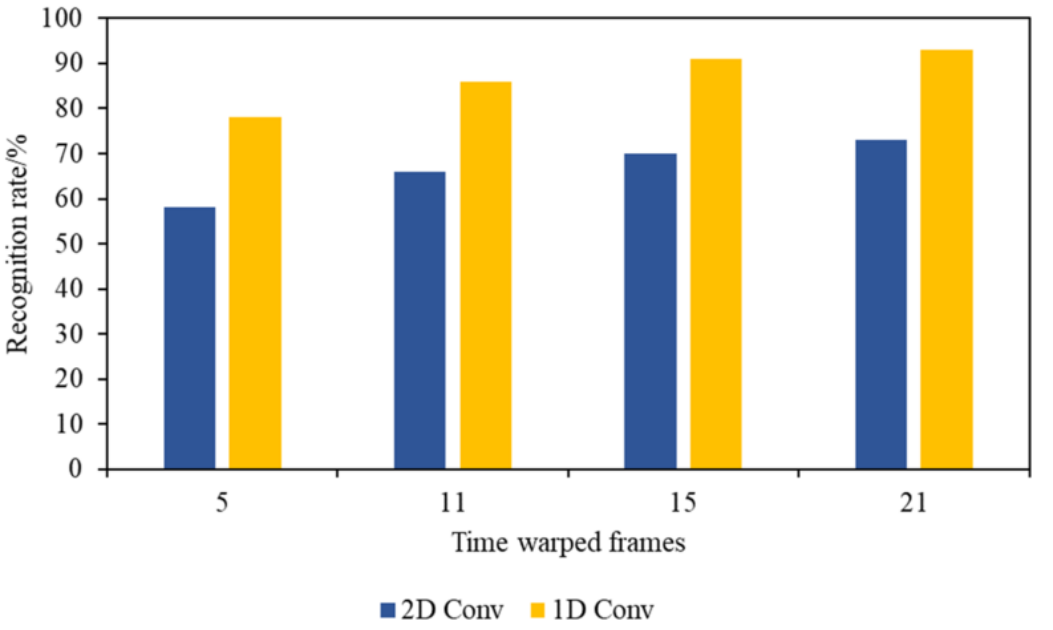


Figure 9. Influence of time warping frame number on recognition rate

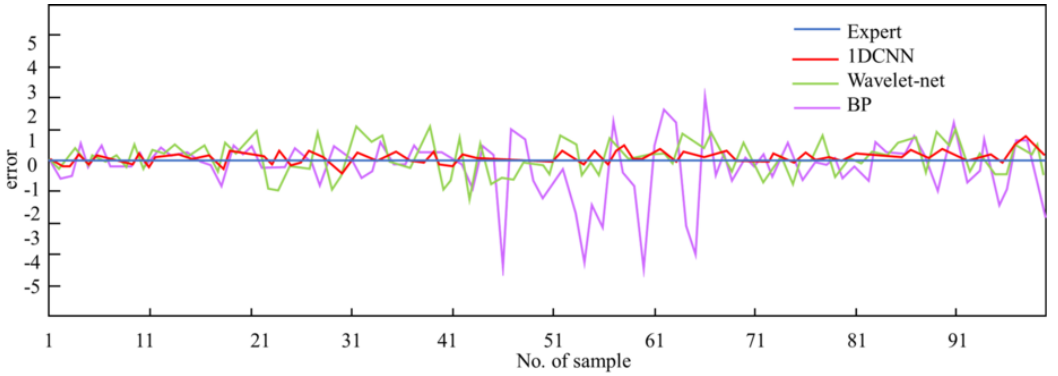


structure, and the change of the output value can well reflect the change of the network structure. Convergence speed indicates that the entire neural network can be stabilized, that is, the error rate is also stabilized. Because the structure of the entire neural network will continuously adjust the connection parameters through the relationship between the expected value and the actual output, it will further achieve the stability of the network. Different network structures have different effects on the error rate of the network output.

Because the parameters adjusted for various iterations are different, the error rates are also different. It can be seen from Figure 10 that when the number of iterations increases, the error rates of the two networks are reduced, because the neural network is originally a stable parameter adjustment, and the error rate is lower. In addition, the slope of the convergence curve can represent the convergence speed, and it is obvious that the one-dimensional convolutional neural network is better regardless of the number of iterations from lower to higher.

Although the performance of wavelet neural network is superior to that of BP neural network, its connection mode is fully connected, which makes the entire network more complex. Whether this is due to training time or training difficulty, there are certain drawbacks. The most prominent feature of one-dimensional convolutional neural networks is weight sharing, which reduces the complexity of the network. In addition, an important feature of one-dimensional convolutional neural networks is that they are top-heavy (the smaller the input weight value, the more the output weight), reflecting an inverted triangle shape. This prevents the back propagation loss of gradient wavelet neural networks from being too fast. The maximum evaluation error value of the one-dimensional convolutional neural network method is 0.34, which is smaller than the maximum error value of the wavelet neural network method, which is 0.82; the minimum error value of the evaluation error of the one-dimensional convolutional neural network method is 0.01; the average error of the wavelet neural network method is 0.56, which is larger than the average error of 0.23 of the one-dimensional convolutional neural network method.

Figure 10. Comparison of errors of various methods



In this chapter, the corresponding indicators and evaluation results of the subjective evaluation are given first; then the objective evaluation of voice based on the BP neural network, wavelet neural network, and improved one-dimensional convolutional neural network are each presented. Compare the recognition results of one-dimensional convolutional neural networks and two-dimensional convolutional neural networks, as well as the evaluation results of BP neural networks and wavelet neural networks, and the evaluation results of wavelet neural networks and one-dimensional convolutional neural networks. Comparison of the results of the different models is needed to determine which method can more accurately assess artistic sounds. Finally, we chose a one-dimensional convolutional neural network to score the students' voices when singing and compared them with the actual experts' scores at the same time This model can give students the best guidance in vocal music training, effectively improve the quality of students in the process of vocal music learning, and provide more high-quality talents in the field of vocal music in China.

A detailed description and discussion were conducted on the methods of improving the pronunciation quality of vocal students based on big data technology. Three methods – the BP neural network, wavelet neural network, and one-dimensional convolutional neural network – were used for objective evaluation of artistic voice, and their effects were compared and analyzed. By comparing experimental data, it can be seen that the one-dimensional convolutional neural network method has the best stability and accuracy, the smallest error, and the highest accuracy. Although the evaluation results of wavelet neural networks are also similar to expert scores, the error still needs to be improved. It is recommended to further explore other possibilities and advantages of one-dimensional convolutional neural networks in future research. Compared to traditional two-dimensional convolutional neural networks and BP neural networks, one-dimensional convolutional neural networks have fewer parameters, resulting in higher training efficiency. Additionally, one-dimensional convolutional neural networks use weight sharing to better process the temporal information of speech signals, thereby improving the ability to extract and model sound features. In addition, one-dimensional convolutional neural networks can perform convolution operations across multiple time steps when processing speech signals, thereby better capturing the long-term dependence of sound features. Finally, one-dimensional convolutional neural networks can also utilize techniques such as residual connections and batch standardization to improve the convergence speed and generalization ability of the network. Overall, one-dimensional convolutional neural networks perform well in objective evaluation of voice, with high scoring results and stability, and they can be used as an effective evaluation tool.

CONCLUSION

This paper uses convolutional neural network (CNN) as the basic network, improves the traditional two-dimensional CNN (2DCNN) network into a one-dimensional CNN (1DCNN) network that is more suitable for one-dimensional sound signals through correlation preprocessing and parameter optimization and structural adjustment of the CNN network, and proposes a method for evaluating the quality of singing voice based on the 1DCNN network. This method can not only be used for the evaluation of singing voice, but also for the diagnosis of voice diseases. In addition, this article analyzes the factors that need to be considered in the training process of neural networks, such as time warped frames, convergence speed, and errors, so that they can be more comprehensively applied in professional environments. The acoustic evaluation method based on one-dimensional convolutional neural networks proposed in this article can provide strong support and assistance for vocal music students in scientific selection, teaching, training, and voice disease diagnosis. By using this method, teachers can objectively evaluate students' pronunciation quality and provide corresponding guidance, thereby improving students' singing level. At the same time, for people with voice diseases, this method can quickly and accurately diagnose the type and degree of the disease and provide corresponding treatment suggestions. The simulation experiment completed on the MatlabR2016a platform compares the predicted evaluation results with the subjective evaluation results of professionals to obtain error statistical results. A comparative analysis of the BP neural network, wavelet neural network, and traditional 2D CNN network shows that the average error of the proposed method is 0.23, 0.50 lower than the BP neural network, and 0.33 lower than wavelet neural network.

The training of neural networks takes a long time, and this article only conducts simulation in the Matlab R2016a environment. If conditions permit, the research still hopes to try the method of distributed cluster computing, using CUDA for training, which will greatly accelerate the training process of the network and improve research efficiency. The method presented in this article has some limitations, such as the need to perform a time rule on the number of frames for the input samples in order to unify the size of the feature parameters of the samples before inputting them into the trained network – that is, to set them to the same number of frames. In the future, we can explore the performance of models at different frame rates to seek more efficient and accurate methods for analyzing sound signals.

AUTHOR NOTE

The authors declare that they have no conflicts of interest.

This work was supported by Heilongjiang Provincial Education Science Planning Key project in 2023, Project name: Research on the Ideological and Political Construction of College Vocal Music Course of Chinese Ancient Poetry Art Songs from the Perspective of Cultural Confidence (project No. GJB1423389).

REFERENCES

Atmowardoyo, H., Weda, S., & Sakkir, G. (2020). Information technology used by millennial good English language learners in an Indonesian university to improve their English skills. *Solid State Technology*, *63*(5), 9532–9547.

Foote, J. A., & McDonough, K. (2017). Using shadowing with mobile technology to improve L2 pronunciation. *Journal of Second Language Pronunciation*, *3*(1), 34–56. doi:10.1075/jslp.3.1.02foo

Fouz-González, J. (2020). Using apps for pronunciation training: An empirical evaluation of the English File Pronunciation app. *24*(1), 62-85.

Geng, L. (2021). Evaluation model of college English multimedia teaching effect based on deep convolutional neural networks. *Mobile Information Systems*, *2021*, 1–8. doi:10.1155/2021/1874584

Hao, W. (2021). Pronunciation correction of students in music classroom based on computer voice simulation. *Journal of Intelligent & Fuzzy Systems, (Preprint)*, 1-10.

Hsieh, Y. Z., Lin, S. S., Luo, Y. C., Jeng, Y. L., Tan, S. W., Chen, C. R., & Chiang, P. Y. (2020). ARCS-assisted teaching robots based on anticipatory computing and emotional big data for improving sustainable learning efficiency and motivation. *Sustainability (Basel)*, *12*(14), 5605. doi:10.3390/su12145605

Huang, C. (2022). Vocal music teaching pharyngeal training method based on audio extraction by big data analysis. *Wireless Communications and Mobile Computing*, *2022*, 1–11. doi:10.1155/2022/4572904

Jiang, H. (2019). Computer-assisted interactive platform design for online music teaching based on cloud computing. *International Journal of Engineering Intelligent Systems*, *27*(2), 47–54.

Jiang, M. Y. C., Jong, M. S. Y., Lau, W. W. F., Chai, C. S., & Wu, N. (2021). Using automatic speech recognition technology to enhance EFL learners' oral language complexity in a flipped classroom. *Australasian Journal of Educational Technology*, *37*(2), 110–131. doi:10.14742/ajet.6798

Jing, J. (2022). Deep learning-based music quality analysis model. *Applied Bionics and Biomechanics*, *2022*, 1–6. doi:10.1155/2022/6213115 PMID:35733449

Kessler, G. (2018). Technology and the future of language teaching. *Foreign Language Annals*, *51*(1), 205–218. doi:10.1111/flan.12318

Lange, C., & Costley, J. (2020). Improving online video lectures: Learning challenges created by media. *International Journal of Educational Technology in Higher Education*, *17*(1), 1–18. doi:10.1186/s41239-020-00190-6

Mason, W. P. (Ed.). (2013). *Physical acoustics: principles and methods*. Academic press.

O'Brien, M. G., Derwing, T. M., Cucchiarini, C., Hardison, D. M., Mixdorff, H., Thomson, R. I., & Levis, G. M. (2018). Directions for the future of technology in pronunciation research and teaching. *Journal of Second Language Pronunciation*, *4*(2), 182–207.

Oh, E. Y., & Song, D. (2021). Developmental research on an interactive application for language speaking practice using speech recognition technology. *Educational Technology Research and Development*, *69*(2), 861–884. doi:10.1007/s11423-020-09910-1

Saito, K., & Akiyama, Y. (2017). Video-based interaction, negotiation for comprehensibility, and second language speech learning: A longitudinal study. *Language Learning*, *67*(1), 43–74. doi:10.1111/lang.12184

Sun, J. (2019). Research on vocal sounding based on spectrum image analysis. *EURASIP Journal on Image and Video Processing*, *2019*(1), 1–10. doi:10.1186/s13640-018-0397-0

Sun, Z., Lin, C. H., You, J., Shen, H. J., Qi, S., & Luo, L. (2017). Improving the English-speaking skills of young learners through mobile social networking. *Computer Assisted Language Learning*, *30*(3-4), 304–324. doi:10.1080/09588221.2017.1308384

Syafitri, A., Asib, A., & Sumardi, S. (2018). An application of Powtoon as a digital medium: Enhancing students' pronunciation in speaking. *International Journal of Multicultural and Multireligious Understanding*, *5*(2), 295–317. doi:10.18415/ijmmu.v5i2.359

Tejedor-García, C., Escudero-Mancebo, D., Cámara-Arenas, E., González-Ferreras, C., & Cardeñoso-Payo, V. (2020). Assessing pronunciation improvement in students of English using a controlled computer-assisted pronunciation tool. *IEEE Transactions on Learning Technologies*, *13*(2), 269–282. doi:10.1109/TLT.2020.2980261

Vijayan, K., Li, H., & Toda, T. (2018). Speech-to-singing voice conversion: The challenges and strategies for improving vocal conversion processes. *IEEE Signal Processing Magazine*, *36*(1), 95–102. doi:10.1109/MSP.2018.2875195

Wang, X., & Liu, X. (2022). Interactive teaching system for remote vocal singing based on decision tree algorithm. *Mathematical Problems in Engineering*, *2022*, 1–10. doi:10.1155/2022/4957353

Wang, Z., & Wu, Q. (2021). Research on automatic evaluation method of Mandarin Chinese pronunciation based on 5G network and FPGA. *Microprocessors and Microsystems*, *80*, 103534. doi:10.1016/j.micpro.2020.103534

Wu, Y., Zheng, C., Hao, M., & Wang, L. (2022). Implementation of a system for assessing the quality of spoken English pronunciation based on cognitive heuristic computing. *Computational Intelligence and Neuroscience*, *2022*, 1–12. doi:10.1155/2022/5239375 PMID:35845915

Yang, Y., & Yue, Y. (2020). English speech sound improvement system based on deep learning from signal processing to semantic recognition. *International Journal of Speech Technology*, *23*(3), 505–515. doi:10.1007/s10772-020-09733-8

Yu, P. J., & Xiong, M. Z. (2022). Remote vocal singing course design based on embedded system and internet of things. *Mobile Information Systems*, *2022*, 1–10. doi:10.1155/2022/8712081

Zhang, H., & Tsai, S. B. (2021). An empirical study on big data model and visualization of internet+ teaching. *Mathematical Problems in Engineering*, *2021*, 1–10. doi:10.1155/2021/9974891

Zhang, Y., & Liu, L. (2018). Using computer speech recognition technology to evaluate spoken English. *Educational Sciences: Theory & Practice*, *18*(5).

Zhao, X., & Jin, X. (2022). Standardized evaluation method of pronunciation teaching based on deep learning. *Security and Communication Networks*, *2022*, 1–11. doi:10.1155/2022/8961836

Zhou, N. (2020). Database design of regional music characteristic culture resources based on improved neural network in data mining. *Personal and Ubiquitous Computing*, *24*(1), 103–114. doi:10.1007/s00779-019-01335-9

Zou, D., Huang, Y., & Xie, H. (2021). Digital game-based vocabulary learning: Where are we and where are we going? *Computer Assisted Language Learning*, *34*(5-6), 751–777. doi:10.1080/09588221.2019.1640745

Dan Shen was born in Heilongjiang, China in 1972. From 1991 to 1995, she studied at Harbin Normal University and obtained a bachelor's degree in 1995. At present, she works in Harbin University. She has published a total of sixteen papers. Her research interests include vocal music teaching and music education theory.

Wenjia Zhao was born in 1983 in Liaoning, China. From 2005 to 2007, she studied at Shenyang Conservatory of Music and received her bachelor's degree in 2007. From 2007 to 2010, she studied in Harbin Normal University and obtained a master's degree in 2010. At present, she works in Harbin University. She has published two papers in total. Her research interests include vocal music teaching and music education theory.