Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 228 (2023) 526–532

Procedia Computer Science

www.elsevier.com/locate/procedia

3rd International Conference on Machine Learning and Big Data Analytics for IoT Security and Privacy

Evaluation method of singing pronunciation quality based on artificial intelligence technology

Hanzhang Tang*

Southwest University of Science and Technology, Mianyang 621010, China

Abstract

Since people have limited ability to judge the quality of their own singing and pronunciation, it is difficult to find problems on their own after reaching a certain level, so they need external help. However, external help also comes from human beings in the past and has many limitations. Therefore, in order to provide better help, this paper will carry out research from the perspective of artificial intelligence. In this paper, the basic concepts of artificial intelligence technology are discussed, and then the evaluation ideas of singing pronunciation quality on the technical level are put forward. Finally, the method system is designed.

Keywords: Artificial intelligence technology; Singing pronunciation quality; Pronunciation evaluation

1. Introduction

Singing pronunciation quality can mainly be judged by intonation, tone strength, timbre and other indicators, and these indicators are special, there is no intuitive data display, so the past technical means can not judge these indicators, naturally can not evaluate the quality of singing pronunciation, also makes the evaluation work too artificial. Artificial itself but also have ability limit, on the one hand, the artificial work efficiency is limited, can't give accurate evaluation in a short time, on the other hand, artificial itself for the judgment easily loses the indicators, when the evaluation target of singing high levels, it is difficult to give evaluation (does not mean artificial unable to evaluate the high level, this phenomenon is mainly due to the inevitable interference of various factors in evaluation.

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .
E-mail address: 513342377@qq.com

10.1016/j.procs.2023.11.060

In view of this phenomenon, the emergence of artificial intelligence technology brings some ideas to deal with it, but the key lies in how to use this technology for evaluation, so specific methods need to be studied.

2. Basic concepts of artificial intelligence technology

Artificial intelligence technology is a technical system similar to human intelligence. Its main functions include data recognition, data processing, comprehensive data analysis, deep learning, etc. These functions cover the basic operation process of the technology, and the corresponding relationship is shown in Table 1.

Table 1 Corresponding relationship between artificial intelligence technical functions and basic operation processes

Basic operation process	Corresponding function
Data import	Data to identify
Data initialization	The data processing
Data modeling	Comprehensive data analysis
Data storage	Deep learning

Combined with table 1, the first artificial intelligence technology cannot operate independently, must first to import the data, and will start after the data import data recognition function, the function can be used to recognize to import data, identify the results generally show in mathematical form, concrete form depends on the data recognition, algorithm in common algorithm for LBP algorithm. The algorithm is essentially a feature recognition algorithm of image data, but in many cases, pure numbers can be regarded as a kind of image, so the algorithm can be used for pure digital data recognition. In principle, LBP algorithm uses a predefined operator in the image, and then obtains the texture parameters of the image by sliding through the serial port. The texture parameters are the eigenvalues of the image[1-3]. The operators of the classical LBP algorithm are mostly defined as 3*3 Windows, in which the existing center pixel value becomes the threshold value. The threshold value can be compared with its neighborhood, and then the next operation can be made according to the rules in Table 2.

Table 2 Operation rules after comparing the threshold value of classical LBP algorithm with the field

The serial number	The rules
1	If the field pixel value is greater than or equal to the threshold, it is marked as 1
2	When the domain pixel value is less than the threshold, it is marked as 0

Complete table 2 after operation, the operator of the left upper corner of the window pixels as a starting point, the window along the positive rotation, and then compared with the threshold value, can be a string of 8 bit binary number, converts it to a decimal after will find, decimal number with 8 bit image grey value, thus to save decimal number as the center pixel values store, traverse the entire image texture feature can be obtained.

Classical although LBP algorithm can be applied to pure digital data identification, but direct application has some defects, such as identification results easy to be controlled, so need to be in the application of two tasks: first, must to image conversion of pure digital data, such as converts it to a data graph, both the LBP algorithm can be used for processing; Second, in order to ensure the accuracy of the results, it is suggested to improve the classical LBP algorithm, that is, the classical LBP algorithm has a fixed window radius, which may not meet the texture frequency of different images[3-5]. Therefore, the window can be changed into a circle, and a circular LBP algorithm can be obtained. This algorithm can calculate the radius under any conditions, and set the interval and number between sampling points randomly based on the circular serial port. The circular window of this algorithm is shown in Figure 1.

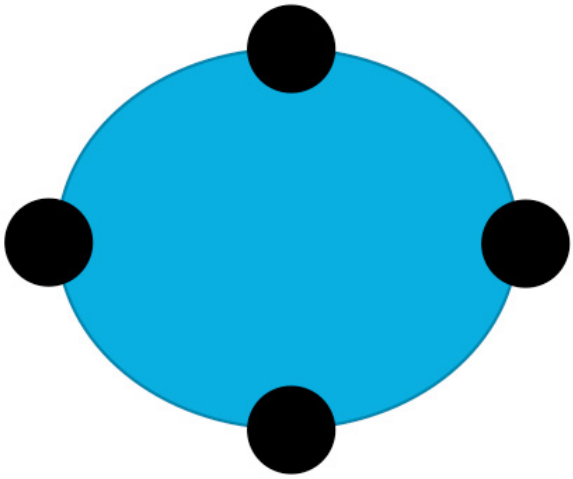

FIG. 1 Circular window example diagram of circular LBP algorithm

(Clockwise rotation, blue circle is the window, black circle is the sampling point on the window)

On the basis of Figure 1, the setting of sampling points of circular Windows can be realized by coordinate calculation, and the expression is shown in Formula (1).

$$x_p = x_c + R\cos\frac{2\pi p}{P} \tag{1}$$

Where, xp is the coordinate of the PTH sampling point, xc is the center coordinate of the window, P is the PTH sampling point, p is the total number of sampling points, and R is the radius of the neighborhood.

3. Thoughts on evaluation of singing pronunciation quality

Combining artificial intelligence technology and circular LBP algorithm, the evaluation ideas of singing pronunciation quality are as follows: first, taking computer equipment as the input end, the evaluation target sings through computer peripherals, forming audio in computer equipment; Second, the audio is fed into an ARTIFICIAL intelligence technology system (which is also mounted on a computer device and can be input via a syndomain transfer) and converted into a graph using the system's data conversion tool. The data curve is mainly formed according to the sound pattern in the audio. Taking the sound intensity indicator as an example, if the sound intensity of the audio reaches 1HZ in the 30s, 1HZ data nodes will appear in the data curve, and the data curve can be obtained after these nodes are arranged in chronological order. Thirdly, standard voicing data curves are input into the system data in advance, and artificial intelligence technology and circular LBP algorithm are used to extract the characteristic values of the curves, and the extracted texture parameter features will be saved in the knowledge base. Fourthly, the same method is adopted to extract the features of the input audio data curve, and the index is compared with the standard audio data curve features in the knowledge base (each index has a corresponding curve), so as to obtain accurate evaluation results. Figure 2 is a schematic diagram of the idea.

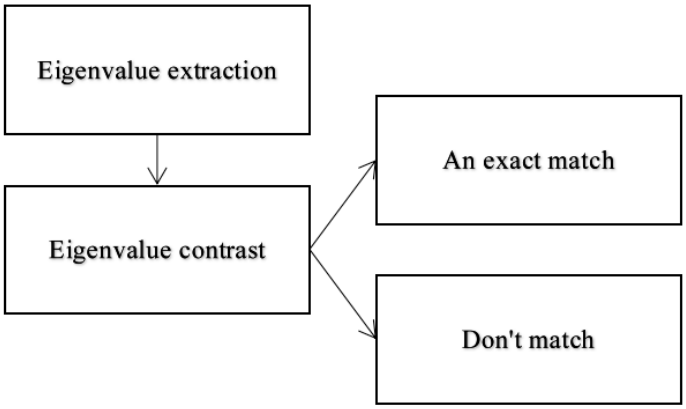

Figure 2 Schematic diagram of ideas

4. Design and application of artificial intelligence technology method system

4.1. Overview of the system

Combining with the evaluation needs of singing pronunciation quality, the system is mainly composed of six parts: one is the data input layer, which mainly realizes the evaluation audio and standard audio input; The second is the data storage layer, which is mainly responsible for storing input evaluation audio, standard audio and data characteristics and other knowledge data. A third is neural network layer, or the formation of the data graph in addition to the need to have audio support, also need to be logical, and artificial intelligence technology is responsible for providing logical support in neural networks, which is a mimic human nervous system to produce the data model, composed of several neural network node and neural link, the main function is to support data transmission, make connections between the data, which helps curve the data; Fourth layer is the data pretreatment, mainly be responsible for the evaluation of input audio with standard audio preprocessing, namely whether evaluation or standard audio, audio input is likely to be affected by external interference factors and appear quality problem, if direct to evaluate such audio,, we must lead to a preprocessing, so need to in front of the evaluation of audio to solve the problem, Ensure audio quality, and then both can be evaluated; The fifth is the algorithm layer, which is mainly equipped with circular LBP algorithm[6-7]. It relies on the algorithm to extract features and then make comparisons to obtain evaluation results[8]. The sixth is the visualization layer, which is mainly used to display the evaluation results and realize the output of the results.

4.2. Design method

According to the overall structure of the system, the design method of each component is as follows.

4.2.1 Data entry Layer

The Internet, ethernet and other networks are used to connect the artificial end equipment and the computer input end equipment. Under the support of the network, the evaluation target and standard audio input personnel can record audio and send the audio to the computer input end equipment to achieve data input.

4.2.2 Data storage layer

Because the data storage layer is mainly responsible for data storage, so must have enough capacity, and for the sake of long-term use, singing sound quality evaluation is a continuous work, so will continue to output audio data, it's brought capacity challenge to traditional database, namely traditional database capacity regardless of size, must exist limit, In this case, the problem can only be solved by expansion. However, expansion will bring high cost, so unlimited use cannot be allowed. In order to solve this problem, this paper chose the cloud database, it has unlimited capacity, thus meet the demand of long-term use of the audio data storage, at the same time because of the cloud and the security of the database itself is insufficient, so in order to improve the level of security, need to take a firewall

technology protection, make cloud database capacity is limited, also need to increase in the actual application, however, the cloud database expansion operation is very simple and does not have high cost, so the cloud database can be used as the data storage layer of the system[9-10].

4.2.3 Neural network layer

Neural network has many forms, such as feedforward type, feedback type, etc., but only feedforward type neural network needs to be used in the evaluation of singing pronunciation quality. The neural network model is characterized by logical forward and suitable for solving unidirectional problems, so this paper chooses feedforward type neural network. Mainly can be divided into the neural network input layer, hidden layer and output layer three parts, including the number of input layer and output layer itself can only be 1, unlimited number of hidden layer, unlimited number of nodes in the input layer and hidden layer, and the number of nodes in the hidden layer is generally greater than the input layer and output layer of the node number is 1. The function of input layer and output layer is to realize data input and output, while the hidden layer is mainly responsible for the integration of each node in the input layer to establish data relationship and obtain data relationship model, which is the final evaluation result and will be output through the output layer.

4.2.4 Data preprocessing layer

There are two ways to realize the data preprocessing layer: First, according to the practical problems, select the corresponding data preprocessing tool Chambers in the resource can be, this paper chose this method, the choice of tools have data to heavy tools, data, noise reduction, the former can remove repetitive meaningless data, to reduce the data scale, improve the efficiency of processing, which can remove the noise of the audio data in the common interference, ensure accuracy; Secondly, if there are no ready-made tools to choose from, Java language programming technology can be considered to achieve this, which can develop various personalized pre-processing tools, but the operation is difficult, so there is no need to adopt various methods in this paper.

4.2.5 algorithm layer

Algorithm layer is mainly through a program database and Java language programming technology, according to the basic concept of circular LBP algorithm algorithm program development, the development of the program saved in the program database.

4.2.6 Visualization layer

Artificial intelligence technology itself the result doesn't directly by artificial to consult, so through the visualization technology to solve this problem, this paper chose the multimedia technology, this technology can convert technological results into artificial can refer to the form, such as words, pictures, and then showed the results in related forms, artificial language as well as tea.

4.3. Evaluation methods

Overall, as a two-dimensional image, a spectrogram can present the changes in the speech spectrum that occur over time. Among them, time is represented by the horizontal axis in the spectrogram, frequency is represented by the vertical axis in the spectrogram, and the values of different coordinate points represent the energy generated by a certain sound frequency component at a specific time point. So, in the spectrogram, the visual image texture represented by various resonance spectra formed by speech during time changes can be well displayed. Nowadays, spectrograms have been applied in many scientific research projects, and in this study, spectrograms were used to evaluate the quality of singing pronunciation. Compared with traditional machine learning, Convolutional neural network has a very complex network structure and more hidden layers. Therefore, this kind of neural network has a very prominent ability to express and learn features, and has been more applied in large-scale classification and recognition learning tasks. Because of the characteristics of two-dimensional image data, combined with the characteristics of Convolutional neural network, these two methods are used to evaluate the quality of singing pronunciation in this study.

4.3.1 Extracting Mel spectrogram

Sound waves themselves belong to one-dimensional existence, and people cannot directly observe the frequency changes of sound waves through their eyes. In this case, using spectrograms can effectively help people understand sound waves. Then, the sound wave is transformed into a Mel spectrogram using a Mel scale filter bank, which effectively presents the energy information, frequency domain information, and time domain information of the sound.

After obtaining the singing audio samples, the next step is to preprocess the samples, which includes pre

emphasis processing, windowing processing, and framing processing. Usually, a FIR high-pass digital filter is used to first perform pre emphasis processing, using a first-order transfer function, as shown in formula (2):

$$H(Z) = 1 - \mu Z^{-1} \quad (2)$$

In this formula, the pre emphasis coefficient is calculated using μ To represent, in general, its value is set to 0.98.

Short term stable audio file signals can be obtained through framing processing. Compared to Haining window, rectangular window, etc., the Hamming window has the smallest spectral leakage, and the so-called windowing mainly refers to the use of each frame signal multiplied by the Hamming window. The specific processing process is shown in formula (3).

$$S\omega(n) = S(n)\,\omega(n) \quad (3)$$

In this formula, the original signal is represented by S (n), and the Window function is represented by ω (n) To represent.

After observing the Mel spectrogram, it can be found that the samples of singing audio present significant differences in the spectrogram images. By analyzing the relevant research content in the field of machine vision, it can be found that in recognizing Mel spectrograms, the problem of singing evaluation can be handled through image classification.

4.3.2 Parameter optimization of Convolutional neural network

As a deep neural network, CNN itself has a convolutional structure, which not only reduces the network parameters but also effectively solves the overfitting problem of the model. In this process, considering the energy information, frequency domain information, time domain information and other characteristics contained in the Mayer spectrogram, after repeated testing, the parameters of the Convolutional neural network were optimized in this study, and then the Dropout layer and data enhancement module were added, so that the loss value of the network could be minimized, and the training time could also be shortened, In this way, while ensuring the Receptive field, the fine-grained features can be extracted.

①Convolutional layer: The main function of this convolutional layer is to accurately recognize spatial patterns in the image, including local objects and lines. The application of convolution operation plays a very important role, as it not only extracts image features and enhances them, but also reduces noise. In this study, the CNN network constructed mainly consists of three convolutional layers, with 32 cores, 32 cores, and 64 connotations, all of which have 3 convolutional cores × The size of 3. In order to effectively solve the problem of gradient disappearance, Relu Activation function is used in both the hidden layer and the input layer.

②Pooling layer: In the pooling layer, downsampling is mainly used, which can effectively reduce the number of parameters and also have robustness to translation and deformation. Compared with average pooling, maximum pooling can maximize the control of convolutional layer parameter errors, further ensuring that the mean does not shift, which can preserve the most texture information. Therefore, in this study, maximum pooling was chosen. In order to sample twice the feature map, 2 was used in this study × A pooling layer of 2 sizes is set with 2 as the stride.

③ Fully connected layer: In this study, in order to prevent the loss of feature information, a fully connected layer was used for processing, which can effectively enhance the features. In this study, two fully connected layers were mainly used. The output of the first connection layer was 64, and the output of the second connection layer was 1. Among these two parameters, 1 represents the number of output categories. Compared with the ordinary Convolutional neural network, the data enhancement module is also used in this study. Through this module, the diversity of data features can be significantly enhanced. During specific operation, trusted images can be randomly transformed through color change and Geometric transformation, and further sample expansion can be carried out on this basis. This way can be used to dig deeper features when training models, so that higher generalization ability can be obtained. By using the Image Data Generator to obtain images, multiple random transformations can be executed in Keras, thereby achieving the goal of data enhancement.

In order to effectively reduce overfitting, in this study, the main purpose of adding the Dropout layer to the classifier is weight attenuation, which can effectively avoid parameter redundancy caused by the classifier.

4.4. System Applications

First input standard audio, and extract its characteristic parameters, it can be saved in the database for backup. Secondly, input the evaluation audio, extract the characteristic parameters and save the parameters in the knowledge base. At this time, the artificial intelligence technology system can identify the evaluation audio through the deep

learning function, and find the corresponding audio data set from the standby standard audio data based on the recognition results to establish the evaluation foundation. Again, according to the characteristic parameters of standard audio and evaluation audio, data curves are drawn on the tone strength, pitch accuracy, timbre and other indicators, and the evaluation results can be obtained by comparison. For example, the tone strength data feature of evaluation audio in the 30s is 1HZ, while that of standard audio in the 30s is 1.3Hz. This indicates that the pronunciation intensity of the evaluation target in the first 30s is insufficient, with a difference of 0.3, which can be strengthened in the follow-up. Finally, this evaluation result can be displayed to the artificial by means of visualization technology.

5. Conclusion

To sum up, the evaluation of singing pronunciation quality could not be achieved by technical means in the past, but the emergence of artificial intelligence technology has solved the relevant problems, and the use of this technology can accurately and comprehensively evaluate singing pronunciation quality. The evaluation result given by artificial intelligence technology is in the form of data, so it is more intuitive. At the same time, as long as there is no problem in the input sample, the accuracy of the result is higher than that of manual. In addition, as a tool, artificial intelligence technology does not have the limitation of human ability and can be evaluated efficiently, indicating that it has advantages and is worth promoting.

References

[1] Wang Z,Wu Q.Research on automatic evaluation method of Mandarin Chinese pronunciation based on 5G network and FPGA[J].Microprocessors and Microsystems,2021,80(3):103534.

[2] Zhang N,Jiang T,Deng F,et al.Automatic Singing Evaluation without Reference Melody Using Bi-dense Neural Network[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP).IEEE,2019.

[3] Tan T.Singing Evaluation based on Deep Metric Learning[C]//ISCSIC 2019:2019 3rd International Symposium on Computer Science and Intelligent Control.2019.

[4] Lu P,Wu J,Luan J,et al.XiaoiceSing:A High-Quality and Integrated Singing Voice Synthesis System[C]//2020.

[5] Pertiwi A L,Sari L H,Munir A,et al.Evaluation of air quality and thermal comfort in classroom[J].IOP Conference Series:Earth and Environmental Science,2021,881(1):012028(8pp).

[6] Konane D,Tiemounou S,Ouedraogo W Y S B.Impact of Languages and Accent on Perceived Speech Quality Predicted by Perceptual Evaluation of Speech Quality(PESQ)and Perceptual Objective Listening Quality Assessment(POLQA):Case of Moore,Dioula,French and English[J].Open Journal of Applied Sciences,2021,11(12):9.

[7] Bs R K,Ramanna P K,Shilpa M,et al.Evaluation of Anticaries Efficacy of Various Fluoride Varnishes on Artificial Enamel Lesion:An In Vitro Study[J].The journal of contemporary dental practice,2021,2021,22(1)7:774-777.

[8] Zhou Y.Research of artificial intelligence in computer network technology[J].Journal of Physics:Conference Series,2021,2083(4):042082-.

[9] Xu L,Zheng Y,Xu D,et al.Predicting the Preference for Sad Music:The Role of Gender,Personality,and Audio Features[J].IEEE Access,2021,PP(99):1.

[10] ACMD Silva,Silva D F,Marcacini R M.4MuLA:A Multitask,Multimodal,and Multilingual Dataset of Music Lyrics and Audio Features[C]//WebMedia'20:Brazillian Symposium on Multimedia and the Web.2020.