The Application of Intelligent Speech Recognition Technology in the Tone Correction of College Piano Teaching

Jing Wang1,*

1Mianyang Teachers’ College, Mianyang, Sichuan, 621000

*Corresponding author e-mail: 754090077@mtc.edu.cn

Abstract. In recent years, with the continuous deepening of college teaching reform and the rapid development of artificial intelligence, the teaching methods used in college classrooms have been greatly different from traditional ones. The original teaching programs and teaching methods need to be changed with the development of the times. The main purpose of this article is to improve the correction rate of intelligent voice recognition technology for college piano timbre without violating the safety of artificial intelligence. This article mainly conducts experiments by consulting relevant domestic and foreign literature, using observation and comparison methods, and using some neural network algorithms (BP neural network algorithm and convolutional neural network algorithm) to obtain experimental results, and finally complete our experimental goal - research intelligence The application of speech recognition technology in the tone correction of piano teaching in colleges. The test results show that the use of multiple algorithms can improve the application of intelligent speech recognition technology in the tone correction of piano teaching in colleges and universities, mainly to improve the correction rate and enhance the ability of guiding teaching.

Keywords: Intelligent speech recognition, piano tone correction, BP neural network algorithm, convolutional neural network algorithm.

1. Introduction

The basic principle of the BP neural network is that the artificial neural network does not need to pre-determine the mathematical equations of the mapping relationship between the input and the output. It only needs to learn some rules through its own training, and obtain the closest expected output value when the input value is given [1-2]. Its basic idea is the gradient descent method. The gradient search technique is used to minimize the error mean square error between the actual output value and the expected output value of the network [3].Convolutional neural network (CNN) is basically the result of several layers of convolution with non-linear activation function, and the application of pooling layer in convolution [4]. Different filters (hundreds or thousands) are applied to
each layer.

The key to understanding is that the filter is not preset, but learned during the training phase to minimize the appropriate loss function. It has been observed that lower levels learn to detect basic features, while higher levels detect more complex features, such as shapes or faces [5]. Because of the convergence layer, neurons in the back layer can see more original images, so they can edit the basic features learned in the previous layer. So far, the basic concepts of convnet have been introduced [6]. CNN performs one-dimensional convolution and pooling operations on audio and text data in the time dimension. Images are processed in two-dimensional (height \( \times \) width) dimensions, and videos are processed in three-dimensional (height \( \times \) width \( \times \) time) dimensions. For images, sliding the filter on the input will generate a feature map that provides the filter response for each spatial location [7].

That is, the convolutional network consists of multiple filters stacked together to learn to recognize specific visual features independent of the location information in the image. These visual functions are very simple at the front end of the network, and then as the network deepens, they are combined into more complex global functions [8].

In recent years, deep learning has been widely used in various practical applications. The function of automatically extracting features of input data replaces the traditional function selector. Due to the limited number of samples, this article recommends using transfer learning technology to fine-tune the pre-trained model with a small amount of cloud data. Feature extraction is a research hotspot in cloud classification research [9]. Previous researchers usually design features manually or use feature selectors to select the best combination of features. One of the advantages of deep learning is that it can automatically extract and combine features of input data according to tasks. So we conducted experiments on this [10].

2. Algorithm establishment

2.1. BP neural network algorithm

Artificial Neural Networks (ANN) are widely used in many fields, and ANN (Artificial Neural Network) systems appeared after the 1840s. It is connected by many neurons with adjustable connection weights. It has the characteristics of large-scale multi-line processing, distributed information storage, good self-organization and self-learning ability. It has been widely used in information processing, pattern recognition, intelligent control and system modeling. Especially, the error back propagation training (BP network) can be close to any continuous function, and it has strong nonlinear mapping ability. Moreover, the network parameters can be set according to specific conditions, such as the number of intermediate layers, the number of processing units in each layer and the learning coefficient of the network, so it plays an important role in many application fields.

As mentioned above, because of the limited number of samples in the training set, this article only fine-tunes on the basis of the training set, so the correction rate cannot be too large. In this article, the initial correction rate of the model is 30%, and the correction rate increases by 80% every 2 cycles of training.

\[
l_r = l_{r0} \alpha^{\lfloor \frac{t}{2} \rfloor}
\]

(1)

In the formula: \( l_{r0}=30\% \); \( \alpha=0.2 \), \( t \) is the current number of training cycles; rounded down. The loss function is defined as cross entropy, and the formula is as follows:

\[
L = \sum_{i=1}^{K} y_i \log y'_i
\]

(2)

Where: \( y \) and \( y' \) are the true label and predicted input of the picture, respectively; \( K=6 \) is the number of categories. This paper uses the stochastic gradient descent algorithm with momentum as the optimizer, and each parameter update is:
\[ \omega_{l+1} = \omega_l - \alpha \nabla L(\omega_l) + \gamma (\omega_l - \omega_{l-1}) \]  

(3)

In the formula: \( \gamma = 0.9 \) is the momentum; \( \nabla L(\omega_l) \) is the loss function to derive the parameter \( \omega_l \). This is the result we finally got.

### 3. Modeling method

3.1. **Combined modeling of security risk rating and correction rate**

The model implementation steps are as follows:

1) First, preprocess the piano tone signal. Use wavelet transform to normalize and denoise the piano tone. The wavelet mother function is sym5, and the number of decomposition layers is 5.

2) Perform feature extraction and dimensionality reduction of the piano tone signal. When the piano tone signal is not dimensionized, the sample dimension is 4900. This paper uses the FastPCA method to reduce the dimensionality of the signal. First, the sample is de-averaged, and then the covariance function is calculated. The variance function solves the eigenvalues and eigenvectors, and selects the dimensions of the principal components according to the cumulative contribution rate. The contribution rate calculation formula is:

\[
1 - \frac{\sum_{i=1}^{k} D_{ii}}{\sum_{i=1}^{n} D_{ii}} \leq t
\]

(4)

Among them: \( D \) represents the \( 112 \times 1 \) dimensional feature vector; \( n \) represents the number of rows of \( D \); \( k \) is the dimension of the final selected principal component; \( t \) is the error that can be set by yourself. The results of many experiments show that when \( t = 0.1 \), the cumulative contribution rate is the highest. Therefore, this paper finally selects the principal component dimension as 36.

3) Normalize the piano tone signal after dimensionality reduction, and the normalization method is:

\[
X = \frac{D - D_{\min}}{D_{\max} - D_{\min}}
\]

(5)

4) Set the number of neurons in the input layer to 30, the number of neurons in the hidden layer to 40, and the number of neurons in the output layer to 3.

5) Input a total of 810 piano sound samples from the training set after dimensionality reduction and normalization into the BP neural network optimized by the SSSE algorithm. The weights and biases of the BP neural network are optimized, and the weight update process is as follows: ① According to uniform distribution According to the rule of, \( m \) particles are randomly initialized and placed in the domain of the objective function; ② Find a new position on each particle in the search space according to the simplex neighborhood and multi-role search strategy; Evaluate the advantages and disadvantages of the particles according to the error function, and determine the three action states of the particles: central state, mining state and action state; ③ Record the best position of the particle, when the convergence position of the particle is in the required Stop the search when the accuracy is within the range; ④ Set the weight of the BP neural network to the best position of the particle, and update the deviation and weight in parallel. Piano tone cycles of 198 test samples are input into the training network. The recognition rate and error curve of piano tone are obtained through statistical analysis.

### 4. Evaluation results and research

4.1. **Evaluation results**
The data shown in Figure 1 and Figure 2 are intelligently correcting piano timbre, and processing them with artificial intelligence will lead to data discontinuities and abnormal analysis. The consequences of this situation are more serious. On the other hand, in a large number of normal operating conditions and abnormal operating conditions, managers are mainly concerned about abnormal operating conditions, and the rest of the normal operating conditions are actually out of consideration. They are usually included in the data set for risk assessment, analysis and processing; and in semi-supervised learning, this kind of sample describes the data distribution.

**Table 1.** The output value of the model and two types of error analysis table

|                | Precise | Error-times-1 | Error-times-2 |
|----------------|---------|---------------|---------------|
| **Training set data** | 76.21%  | 24.21%        | 11.97%        |
| **Test set data**     | 83.27%  | 35.21%        | 13.21%        |

**Table 2.** Accuracy table of convolutional neural network

| sample  | Observed | Predicted | Correct rate |
|---------|----------|-----------|--------------|
| training| 0        | 80        | 20           | 82.50%       |
From the data in Table 1 and Table 2, it can be seen that the classification accuracy of the accuracy model of the convolutional neural network algorithm is 72.09%, and the prediction accuracy is 85.50%. Among them, the first type of error in the training set and test set (normal data is regarded as abnormal data) are 24.21% and 35.30%, respectively, the second type of error (normal data is regarded as abnormal data) are 11.97% and 13.21%, and the possibility of making up the numbers with fuzzy timbre is low.

This data set contains a large number of piano rhythms and pitches disclosed by users, a total of 12 variables describing piano timbre characteristics. First, the data is preprocessed, and the method of fewer features and more selection is used to remove the variables that have little influence on tone recognition and correction, and retain 8 feature variables.

4.2. The meaning and connection of intelligent speech recognition technology and artificial intelligence

Speech recognition technology, also known as automatic speech recognition (ASR), aims to convert the vocabulary content of human speech into computer-readable input, for example, into binary codes or character sequences. Different from speaker recognition and speaker verification, the latter tries to identify or confirm the speaking speaker, not the vocabulary content contained in the speech. Speech recognition is an interdisciplinary subject that includes multiple disciplines. In recent years. Speech recognition technology has made significant progress and began to move from the laboratory to the market. With the development of the times, speech recognition technology has entered various consumer fields in the market.

Intelligent speech recognition technology is born out of artificial intelligence, and artificial intelligence is born out of human intelligence. Human intelligence has two obvious levels, namely sensitivity and rationality, reaction and thinking, emergency processing and analysis and judgment, and knowledge and understanding, which represent the two levels of human perceptual intelligence and cognitive intelligence, respectively. Perceptual intelligence and cognitive intelligence are both the performance of knowledge-based abilities. Perceptual intelligence is the emergency processing of the brain's experience knowledge (knowledge base), and the coping ability based on feelings. Cognitive intelligence is brain reasoning based on rational knowledge (knowledge map). For example, if you drive to work every day, the tool location and route are already fixed. You don’t need to think about it. You can only rely on experience to deal with it. If you are on a business trip to a new place, you need to check the map, understand the situation of public transportation, and decide to take an airplane, high-speed rail or autopilot according to the time requirements, and finally choose a feasible solution, which requires cognitive intelligence. Any knowledge point in the knowledge base will have infinite development of associated knowledge, which determines the infinite development prospect of artificial intelligence. So with the passage of time, intelligent speech recognition technology is getting stronger and stronger.

4.3. Development of Intelligent Speech Recognition Technology

Since the beginning of research on speech recognition technology, the development of speech recognition technology has a history of more than half a century. The Audry system developed by
Davis et al. It was the beginning of speech recognition research. This is the first system that can obtain multiple English letters. Dynamic programming and linear predictive analysis technology solve the most important problem in speech recognition—the model problem of speech signal generation; in the 1970s, speech recognition technology has made major breakthroughs, and dynamic time warping technology (DTW) is basically mature, so the length of speech can be equal. In addition, vector quantization (VQ) and hidden Markov model theory (HMM) speech recognition functions are constantly improving, which paved the way for the development of speech recognition. In the 1980s, the research on speech recognition was more in-depth, and various speech recognition algorithms were proposed, among which outstanding achievements include the HMM model human neural network (ANN). After entering the 1990s, speech recognition technology began to be applied to the global market. Many well-known technology Internet companies, such as IBM, Apple, etc., have invested a lot of money in the development and research of speech recognition technology. In the 21st century, the research focus of speech recognition technology has shifted to impromptu oral and natural dialogue and simultaneous translation of multiple languages. With the advancement of data processing technology and the rapid popularization of mobile Internet, computer technology has been widely used in various fields of society, followed by the generation of massive data. Among them, voice data has received more and more attention.

Domestic research and exploration on speech recognition technology started in the 1980s, and achieved many results and developed rapidly. For example, the speech recognition technology developed by Tsinghua University uses 1183 monosyllables as the recognition primitives, and decomposes the syllables, and finally recognizes them, so that the accuracy of three-character and four-character words is as high as 98%; the Chinese Academy of Sciences uses continuous density HMM, the recognition rate of the whole system reaches 89.5%, and the recognition rates of tones and words are 99.5% and 95%, respectively. At present, our country's speech recognition technology has the strength of the international superpowers, and its comprehensive error rate can be controlled within 10%, and finally mature.

5. Conclusion
In short, computer technology and communication technology, and is the product of the continuous development of science and technology. As the most advanced science and technology in today's social development, artificial intelligence technology has brought many new changes to people's learning, life and production, and can better meet people's different needs for life and work. But we cannot be satisfied with the status quo. At present, there are still some problems in the application of intelligent speech recognition technology and tone correction. Therefore, we should find better algorithms to solve the problem and improve the correction rate. We believe that this will be solved perfectly in the future.

References
[1] Cai J, Liu Y. Research on English pronunciation training based on intelligent speech recognition[J]. International Journal of Speech Technology, 2018, 21(3):633-640.
[2] Verma A. INTELLIGENT STORE FOR SMART TECHHIES[J]. Pc Quest, 2019, 32(7):48-49.
[3] Park J S, Kim S H. Noise Cancellation Based on Voice Activity Detection Using Spectral Variation for Speech Recognition in Smart Home Devices[J]. Intelligent automation and soft computing, 2020, 26(1):149-159.
[4] Isa M A, Sims-Waterhouse D, Piano S, et al. Volumetric error modelling of a stereo vision system for error correction in photogrammetric three-dimensional coordinate metrology[J]. Precision Engineering, 2020, 58:188-199.
[5] Zhang Y, Grice-Patil Z, Burton P, et al. Neural mechanisms of selective listening to pitch and timbre in musical sounds[J]. The Journal of the Acoustical Society of America, 2018, 143(3):1962-1962.
[6] Tiedong C, Weiyang X. Post-evaluation on financial support highway traffic project based on
[7] Zhang D, Li W, Wu X, et al. Application of simulated annealing genetic algorithm optimized back propagation (BP) neural network in fault diagnosis [J]. International Journal of Modeling Simulation & Scientific Computing, 2019, 10(04):46-49.

[8] Ding W, Sun Y, Ren L, et al. Multiple Lesions Detection of Fundus Images based on Convolution Neural Network Algorithm with Improved SFLA [J]. IEEE Access, 2020, PP(99):1-1.

[9] Feng J, Cai S, Ma X. Enhanced sentiment labeling and implicit aspect identification by integration of deep convolution neural network and sequential algorithm [J]. Cluster Computing, 2019, 22(6):1-19.

[10] Guo Y, Budak, ümit, Vespa L J, et al. A retinal vessel detection approach using convolution neural network with reinforcement sample learning strategy [J]. Measurement, 2018, 125:586-591.