Modular Design of English Pronunciation Level Evaluation System Based on Deep Learning

Jing Cheng

School of International Culture and Communication, Jingdezhen Ceramic University, Jingdezhen, Jiangxi, China
cinbon2007@163.com

Abstract: *In today's world economic integration and internationalization process is accelerating, the demand for improving English teaching and English pronunciation is also increasing rapidly. In this paper, deep learning techniques were applied to the evaluation of English pronunciation level, and Deep Belief Networks (DBN) were used to build English pronunciation level models. Therefore, according to the characteristics of Chinese college students' English pronunciation, this paper improved the traditional computer English pronunciation quality evaluation method, and comprehensively evaluated various evaluation indicators such as intonation, speech speed, rhythm, and intonation. The tests showed that the English pronunciation quality evaluation model had high credibility as evidenced by the 88.33% total agreement rate between machine evaluation and manual evaluation and the 0.723 Pearson correlation coefficient, and it could evaluate and give feedback in a timely, accurate and objective manner. This helped learners find the difference between their pronunciation and standard pronunciation and correct their English pronunciation in time, thereby improving their English pronunciation.*

Keywords: *English Pronunciation Level; Deep Learning; Speech Recognition; Pronunciation Quality Evaluation; Multi-Parameter Evaluation Index*

1. Introduction

The measurement of English speech quality still has a lot of issues. For instance, some computer-assisted language learning programs for oral language acquisition concentrate on teaching vocabulary and grammar while using only one or two assessment indicators as evaluation criteria. This has a functional deficiency, and can only comprehensively score students' English pronunciation. In the oral English test, manual evaluation is still the main method, and its reproducibility is poor. Deep learning is a deep nonlinear network that can approximate a complex function and describe the distributional representation of the input data. It can learn the essential characteristics of a large amount of data from a small number of samples, and has better performance in simulating the analysis and learning of the human brain.

In view of these problems in the evaluation of English pronunciation quality, many scholars have studied it. The goal of Abbasinia S's research was to better understand English expressiveness by looking at how speech processed functions in video games. On this basis, a software that could correct English pronunciation for 8-10-year-old children in teaching centers, language schools, and even at home was designed [1]. In order to improve the effect of college English pronunciation quality assessment, Davila A M used artificial emotion recognition technology combined with high-speed mixing mode to analyze and screen various interference factors, thereby improving the English pronunciation level of college students [2]. Li X used computer technology combined with metacognition to simulate and test the effectiveness of oral English teaching. The research results showed that it had certain reference significance for improving the level of students' English pronunciation [3]. Chika used a variety of indicators to test and evaluate students' English. The results showed that it was very necessary to improve students' English pronunciation and syllable characteristics [4]. Piotrowska M proposed a special Convolutional Neural Network (CNN) detection method for English allophone variants [5]. It can be found that the relevant researchers have conducted in-depth research on English pronunciation and have discussed many aspects of English pronunciation improvement.

Numerous academics have also undertaken similar research on the Internet of Things-based deep learning English pronunciation level evaluation system's modular architecture. Among them, Lasi F aimed to develop a deep learning English pronunciation level evaluation model to analyze the shape of human lips to improve the English pronunciation level [6]. In view of the difference between English pronunciation and Chinese Pinyin, Wang X constructed a neural network-based model focusing on the requirements of oral English, and explained the concept of restricted Boltzmann machine. Combined with the Back Propagation (BP) algorithm, the differential of the multi-parameter evaluation was generated, which provided a reference frame for the comprehensive evaluation of English pronunciation level [7]. On the basis of transfer learning technology, Suparman U constructed a deep learning model that can automatically evaluate the lexical stress of non-native English speakers, and systematically evaluated the students' pronunciation with correct vocabulary [8]. With the development of modern educational technology based on computer technology and network technology, Gu X compared various algorithms and designed a teaching model for English learning [9]. According to the needs of English online teaching, Han Y used remote monitoring and deep learning algorithms to establish the architecture of the English online speech quality assessment system [10]. Although many scholars have used deep learning models on the English pronunciation evaluation system to improve the traditional computerized English pronunciation quality evaluation methods for the features of Chinese college students' English pronunciation, but the research on intonation evaluation of fundamental frequency is still blank.

In this paper, the pronunciation level of college students were used as the object of study, and the deep learning technology was applied to English speech recognition and pronunciation level assessment to establish a relevant model. Through speech recognition experiments and speech evaluation experiments, factors such as intonation, speech speed, rhythm, and intonation were discussed in depth. Finally, the established English pronunciation level evaluation system model was tested to verify the reliability of the above evaluation indicators. At the same time, a more scientific and objective English phonetic quality assessment model was established by using regression analysis method.

2. Modular Design Method of Deep Learning English Pronunciation Level Evaluation System

2.1. Basic principles of speech recognition

Speech recognition overview:

The main function of speech recognition is to convert speech signals into corresponding text information [11].

Acoustic model:

The acoustic model is a very critical link, which can be used to describe the feature sequence of each unit. For a specific acoustic feature vector, the model can be used to obtain the probability of it belonging to different sound sources, and the maximum likelihood basis can be used to convert it into a corresponding state sequence.

Acoustic primitive selection:

In the acoustic model, the first problem to be solved is how to select the appropriate acoustic unit. Its selection involves three basic elements: trainability, generality, and accuracy [12].

Hidden Markov model:

The current speech recognition technology is mainly based on Hidden Markov Model (HMM), which is a statistical-based pattern recognition technology.

The algorithm extracts features from the original speech signal and converts them into corresponding feature vectors. When the given speech feature sequence is $O_1^T = \{o_1, o_2, ...o_T\}$, $\tilde{W}$ can be expressed by Formula (1):

$$\tilde{W} = \arg\max_{w} P(W \middle| O_1^T) = \arg\max_{w} \frac{P(W \middle| O_1^T)P(W)}{P(O_1^T)} \tag{1}$$

In Formula 1, P(W) represents the probability of W appearing; $P(O_1^T|W)$ represents the probability that the output acoustic feature is W when the given word sequence is W; $P(O_1^T)$ is removed to obtain Formula (2) [13]:

$$\tilde{W} = \arg\max_{w} P(O_1^T|W)P(W) \tag{2}$$

The right-hand part of Formula (2) is taken as the logarithm and further simplified to:

$$\tilde{W} = \arg\max_{w} P\{\log P(O_1^T|W) + \lambda * \log P(W)\} \tag{3}$$

In the formula, $\log P(O_1^T|W)$ represents the acoustic score, and $\log P(W)$ represents the language score. The correlated acoustic and speech models are used for the calculations. Since the acoustic and speech models are both trained from speech and text corpora, an adjustable parameter λ is added to this formula to measure the effect of the two modes on the selection of word sequences W.

HMM acoustic modeling:

HMM is essentially a doubly stochastic process, which is an implicit finite-state Markov chain that continuously transitions between different states. However, the state sequence cannot be directly observed, and can only be indirectly reflected by the observation vector [14]; in another case, the observation is a random process. The result is that the observation value is determined by the hidden state, and the corresponding observation vector is randomly output in any given state.

HMM can be described by the following five sets of parameters, namely:

$$M = \{S, O, A, B, \pi\} \tag{4}$$

In the formula, S represents the set of limited hidden states contained in the HMM model; O represents the set of possible observations corresponding to each hidden state; A represents the transition probability between states, which can be represented by a matrix; B represents the probability of taking the corresponding output observation value under the premise of a given state; π represents the set formed by the probability of the initial state of the system. When the HMM is used as an acoustic model, as shown in Figure 1. b_{ij} in the figure represents the transition probability from state i to state j.

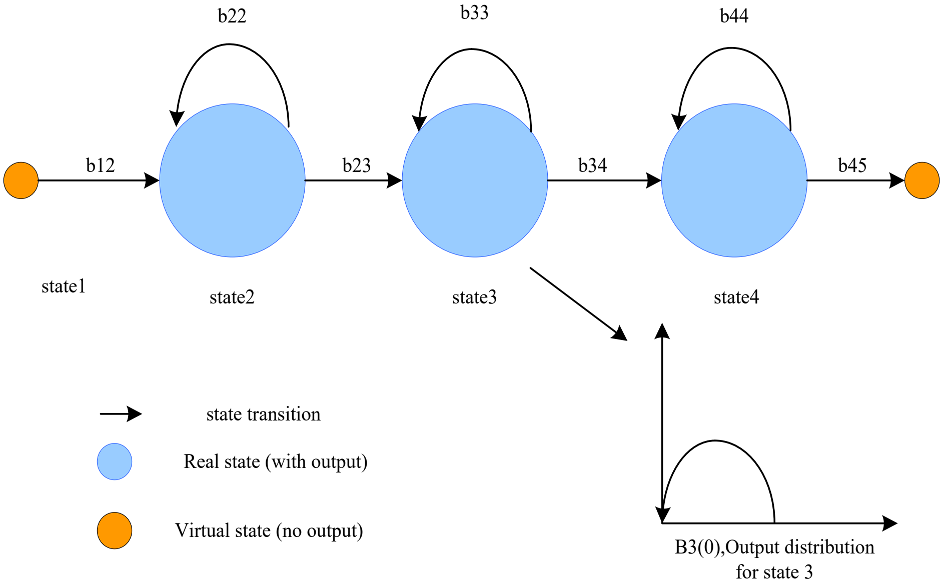

Figure 1. Structure diagram of acoustic model based on HMM

The characteristic distribution of speech signals is difficult to describe with a simple probability distribution model. The Gaussian mixture model is a commonly used method in speech recognition, which mainly uses the Gaussian mixture model to represent the output probability B.

According to different output probabilities, HMM models can be divided into three categories: Discrete Hidden Markov Model (DHMM), Semi-Continuous Hidden Markov Model (SCHMM) and Continuous Hidden Markov Model (CHMM) [15]. In terms of model accuracy, CHMM has the highest accuracy; DHMM is the worst; SCHMM is in the middle; in terms of training complexity, it's the exact opposite.

Language model:

Language patterns are a way of reflecting people's language habits, which reflect the internal rules of word formation. In this paper, the language model used is the probability calculation from the basic sequence to the word, and its precise description would directly affect the performance of the system. Statistical language model is essentially a probabilistic model. It cannot simply determine whether to follow grammatical rules, but uses probabilistic methods to express all the words in the language, that is, the probability of occurrence of sentences in real situations [16].

A sentence S consisting of a sequence of words $W_1^n = \{W_1, W_2, ..., W_n\}$ is given, and its probability of occurrence can be expressed by probabilistic methods:

$$P(S) = P(W_1^{\ n}) = P(W_1)P(W_2|W_1)P(W_3|W_1W_2) \cdots P(W_n|W_1W_2 \cdots W_{n-1}) \qquad (5)$$

In the formula, P() represents the probability function, and $P(W_2|W_1)$ represents the probability of W_2 occurrences in the case of W_1 occurrences. In Formula (5), the close relationship between different words can be determined by P(S).

2.2. Speech feature extraction

Overview of voice characteristics:

Its time-domain properties include the speech pitch period, short-term zero-crossing rate, and short-term energy. Fast Fourier transform spectrum coefficients, linear prediction coefficients, the Linear Prediction Cepstrum Coefficient (LPCC), the Mel Frequency Cepstrum Coefficient (MFCC), and others are among the frequency characteristics of speech [17].

MFCC has the following advantages over LPCC:

(1) According to the research on the human hearing mechanism, the human ear has different hearing sensitivity to sounds of various frequencies, and the sound of 200-5 kHz has a great influence on the clarity of the sound. MFCC converts linear signals into Mel signals to achieve noise suppression and highlight important information. LPCC is based on linear frequency and does not have this advantage.

(2) MFCC is not affected by the characteristics of the input signal. It has no presuppositions and constraints and is suitable for different occasions. LPCC assumes that the processed signal is an autoregressive one, and this assumption cannot be fully applied to consonants with high dynamic characteristics. In addition, when there is noise, the autoregressive signal would become an autoregressive moving average signal, and MFCC has better anti-noise performance than LPCC. A large number of experimental results show that the parameters of MFCC are more robust than those of LPCC, and can better improve the processing effect of speech signals. In view of this, this paper uses the feature parameters of MFCC as speech features.

Speech feature parameter extraction:

The original speech signal not only has a large amount of data, but also causes a lot of information that interferes with semantics due to factors such as the loudness and length of the sound, so it cannot be directly applied to speech processing. The quality of the feature parameters would have a great impact on the processing effect of speech, and appropriate feature extraction technology can achieve better results. The following are guidelines for characteristic parameters:

(1) Each order parameter has good independence.

(2) These characteristic parameters can not only reflect the characteristics of the sound, but also

show the characteristics of hearing and vocal tract, and at the same time have good regional division ability.

(3) In order to facilitate the calculation of feature parameters, effective calculation methods are used to reduce the storage requirements under the premise of ensuring the processing quality of the voice signal, so as to achieve real-time processing of the voice signal.

The speech recognition process:

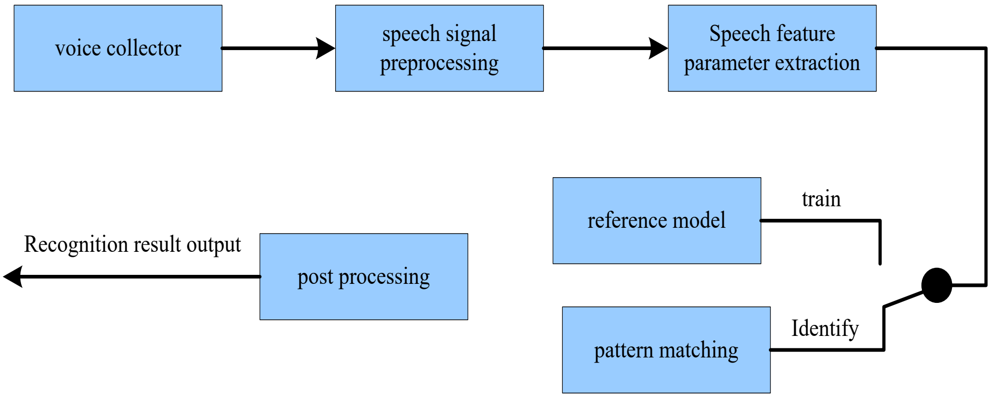
voice collector
speech signal preprocessing
Speech feature parameter extraction
reference model
train
Recognition result output
post processing
pattern matching
Identify

Figure 2. Speech recognition flowchart

A general flow of voice recognition is described in Figure 2. The Nyquist sampling principle states that if the sampling frequency fs max during the analog to digital conversion is greater than twice the maximum frequency Fmax in the signal, it is as shown in Formula (6):

$$f_{s\_\max} \geq 2 * F_{\max} \tag{6}$$

2.3. Deep learning algorithms

In essence, deep learning is a technique for information extraction that makes use of multi-level nonlinear transformations. It can train a series of network parameters with rich content without supervision or supervision, and apply them to feature extraction, transformation and classification.

Key features of deep learning:

(1) The deep structure of the model is highlighted;

(2) The important role of big data in optimizing complex models is emphasized. When the training data is rich and complex, the model can truly demonstrate the ability to model a large amount of data;

(3) The concept of "feature learning" is emphasized.

The composition of a deep learning model:

The network node function:

In deep learning networks, nonlinear transformations are realized by using the nonlinear representation of the input and output of hidden layer units. In fact, the network node and the neuron in the neural network are the same concept here. Combined with the neuron concept, the following model can be obtained:

$$u_i = \Sigma_{j=0}^{j<N_{(l-1)}} w_{ij} \cdot x_j + \theta_i \tag{7}$$

In Formula 7, $N_{(l-1)}$ can be regarded as the number of nodes in the $l-1$ th layer.

For each node in the network, such as the network node model in Figure 3, if the input value is x, the output is an activation probability value with an activation function of $y = f(x)$. There are several different excitation functions:

Sigmoid activation function:

$$f(x_i) = \frac{1}{1+\exp(-x_i)} \tag{8}$$

Hyperbolic tangent activation function:

$$f(x_i) = \tanh(x_i) \tag{9}$$

Modified linear activation function:

$$f(x_i) = \max(x_i, 0) \tag{10}$$

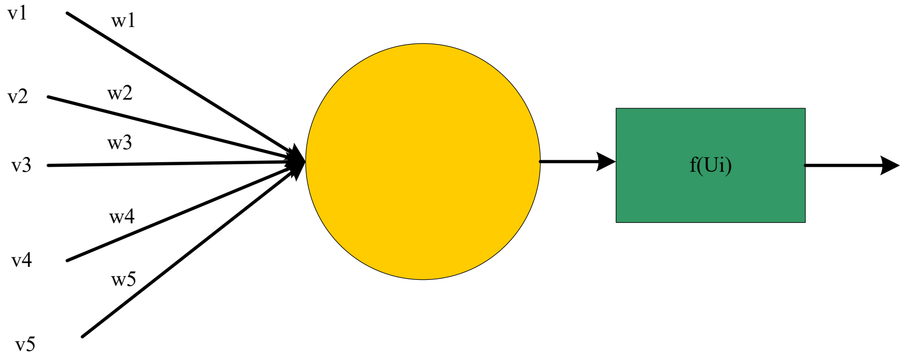

Figure 3. Network node model diagram

RBM model:

RBM models are a special type of Markov random fields. Its upper layer is a random hidden unit, and its lower layer is a random layer of visible or observable units, as shown in Figure 4. Here, for the convenience of calculation, all visible units and recessive units are set as binary variables, namely: $\forall i, j, v \in \{0,1\}, h_j \in \{0,1\}$.

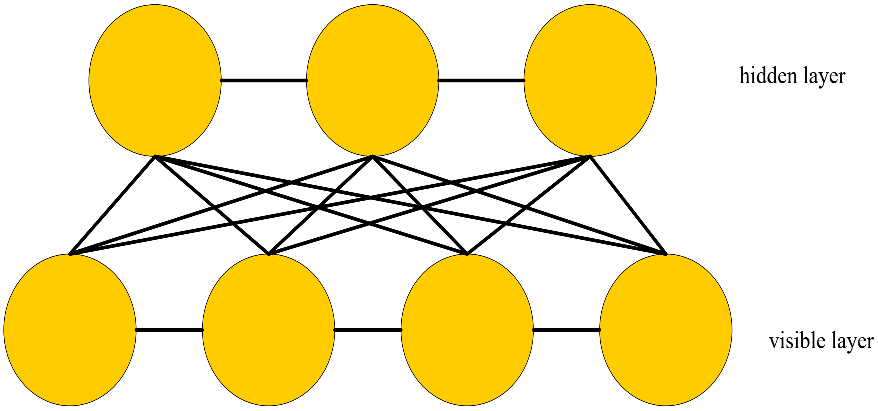

Figure 4. Structure diagram of restricted Boltzmann machine

Since the RBM structure is symmetric, the following formulas can be derived:

$$P(v_j = 1|h) = sigm(a_j + h) \tag{11}$$

Among them, v and h stand for the states of the model's visible unit and hidden unit, which are each represented by a vector.

DBN model:

A deep belief network is a probabilistically generated model that does not belong to a discriminative model like a neural network. Its production mode is mainly to establish and evaluate joint distributions between observed data and markers. The discriminative model only evaluates the subsequent cases.

The bottom layer of DBN is one-way diffusion, and the upper layer is a fully connected network.

The process is as follows:

(1) First, according to the Contrastive Divergence (CD) algorithm, the training is repeated to obtain the RBM network;

(2) The weighting and bias of the first RBM remain unchanged;

(3) Using the Contrastive Divergence (CD) algorithm to repeatedly train the RBM for many times, the following RBM structure is obtained: as shown in Figure 5:

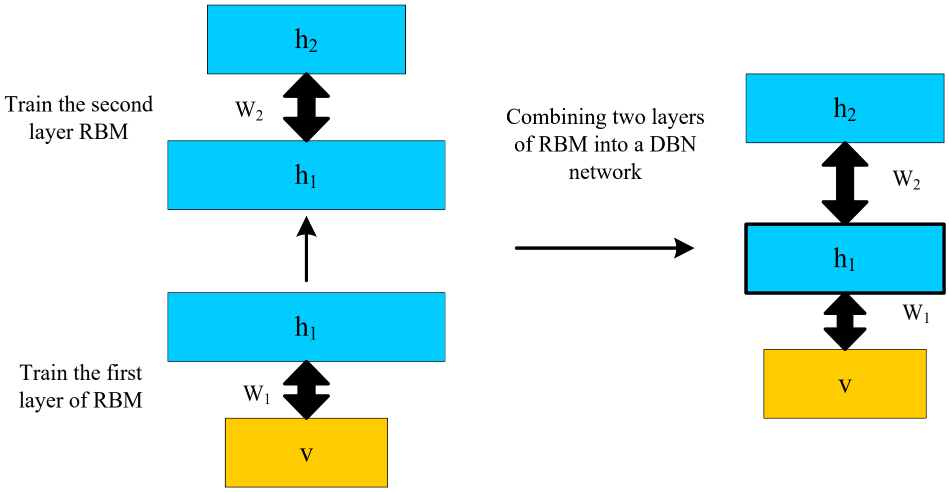
h2
Train the second layer RBM
W2
h1
Combining two layers of RBM into a DBN network
h2
W2
h1
h1
W1
Train the first layer of RBM
W1
v
v

Figure 5. RBM stacking process diagram

(4) The above three steps are repeated;

(5) When training a high-level RBM, in addition to the visible units, the visible layer of the RBM also contains data representing category labeling information. These two data are used to comprehensively train a supervised DBN network;

The above assumes that there are 150 visible units in the upper layer of the RBM. In the input layer, the training data can be divided into 10 categories; the input model network of any sampled data is used for training. When judging the labeled data, the relative label value of this sampling type is 1, while the values of other labels are 0.

The trained DBN structure is shown in Figure 6:

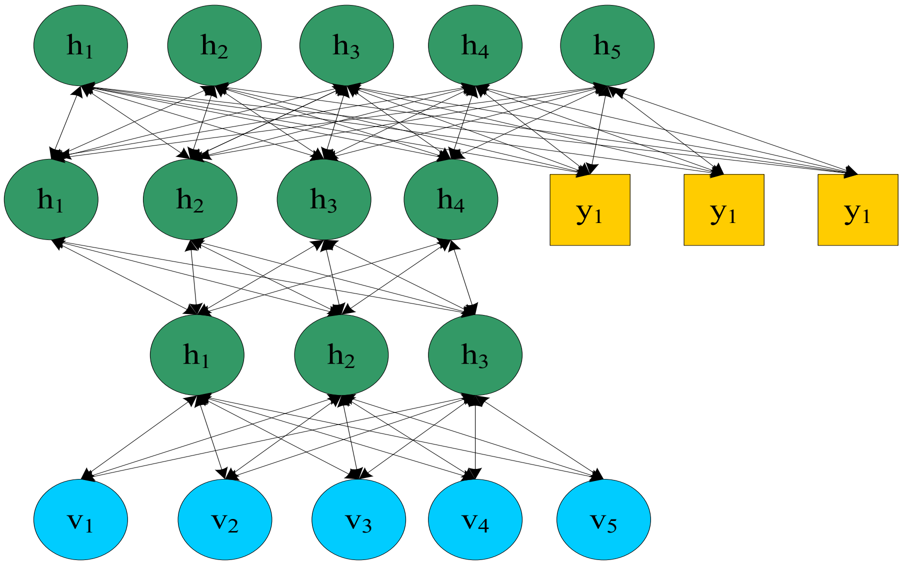

Figure 6. Trained DBN

3. Experiment Simulation and Result Evaluation of English Pronunciation Level of Deep

3.1. Speech recognition experiment

In this paper, a total of 9900 Arabic numerals are included in the speech recognition experiment (99 people's pronunciation of 10 Arabic numerals, and 10 numerals are repeated 10 times), 5500 words from the first 55 people are used as practice, and 4400 words from 44 people are used as tests. This paper uses the MatlabR2013a as the test platform.

Experimental results and analysis:

It has been shown via several trials that the DBN model has four levels: an input layer, two hidden layers, and an output layer. Among them, the quantity of nodes in the input layer and the quantity of nodes at the output end correspond to the distinctive properties of the input speech, respectively. The models above are contrasted with this one, the results shown in Table 1.

Table 1. Comparison of different recognition rates

Model abbreviation	Recognition rate
Dhmm	90.78%
Cdhmm	94.06%
Tda-mwst	93.13%
Tda-gts	93.07%
Bp_adaboost	89.27%
Kaswt	92.65%
Model of this paper	96.56%

3.2. Speech evaluation experiment

The phonetic evaluation experiment aims to test the effect of the model and method of the English pronunciation level evaluation system proposed in this paper.

The linear correlation between two variables is described by the Pearson correlation coefficient. r is a measure of how strongly the connection is linear. From -1 to +1, the larger the absolute value, the stronger the correlation. It is generally considered that r is between 0 and 0.2 for very weak, between 0.2 and 0.4 for weak, between 0.4 and 0.6 for moderate, between 0.6 and 0.8 for strong, and between 0.8 and 1 for extremely strong. The formula is as follows:

$$r=\frac{N\sum x_i y_i-\sum x_i\sum y_i}{\sqrt{N\sum x_i^2-(\sum x_i)^2}\sqrt{N\sum y_i^2-(\sum y_i)^2} \tag{12}$$

Experimental results and analysis:

Inspection of evaluation indicators:

The four components of intonation, speed, rhythm, and intonation of 240 phrases in 10 sentences of 24 students can be collected and compared with human judgment using the approach covered in this work, as shoen in Table 2 and Figure 7.

Table 2. Experimental results of evaluation indicators-number of samples

Index Number of samples (pieces)	Intonation	Speed of speech	Rhythm	Intonation
Consistent	206	194	203	193
One level difference	33	46	33	42
A difference of two	1	0	4	5
Difference of three	0	0	0	0

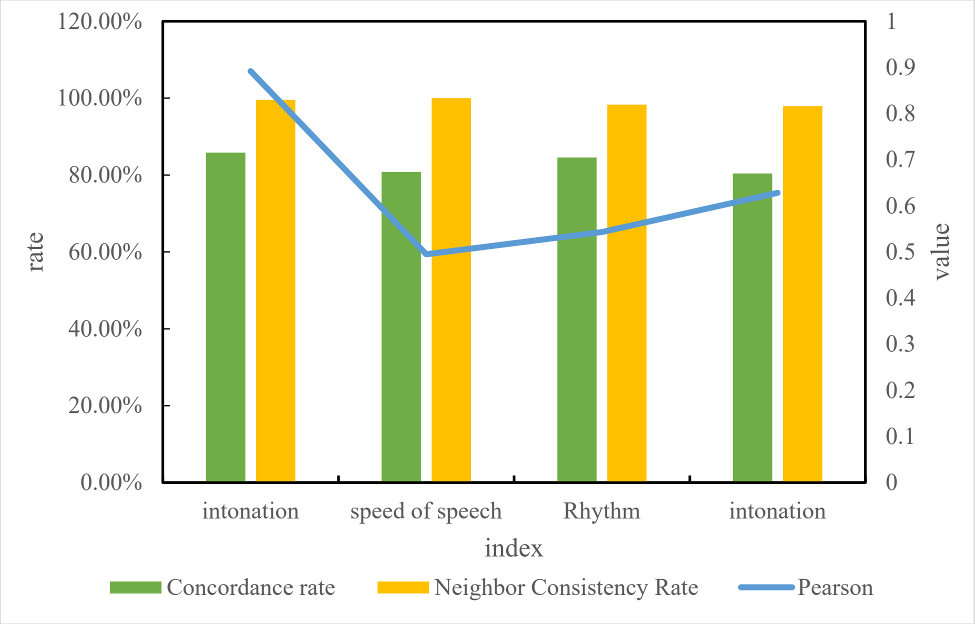

Figure 7. Experimental results of evaluation indicators-statistical indicators

From the test data in Table 2 and Figure 7, the speech quality assessment method proposed in this paper has high credibility.

3.3. Test of the English pronunciation level evaluation system model

This research uses the regression analysis method to establish a relatively scientific and objective English phonetic quality evaluation and analysis model.

Regression analysis is the process of creating a statistical model using mathematical statistical tools and analyzing the statistical relationship between several objective elements. The statistical principles underlying several seemingly ambiguous occurrences have been discovered for model prediction in a vast number of experiments and observations.

In this paper, the comprehensive score of manual evaluation is used as the dependent variable, and the above four factors are used as independent variables. The English sentences with the same mechanical and manual evaluation in English are selected, and the weight of each index is calculated through the Statistical Product Service Solutions (SPSS) software by using the multiple regression analysis method, and the formula is as follows:

$$\begin{aligned}\text{Score} &= \text{AccuracyScore}\times 0.45+\text{AccuracyScore}\times 0.107\\ &+\text{RhythmScore}\times 0.351+\text{IntonationScore}\times 0.313-0.398\end{aligned} \tag{13}$$

In Table 3, there are 212 samples that are evaluated mechanically and manually, and there are 28 samples with a one-level difference. The third level and the second level are identical. It demonstrates that there is a significant link between automated and manual evaluation.

Table 3. Differences between mechanical assessment and manual assessment

Level of differences	**Tonation**
Concordance rate	88.33%
Neighbor consistency rate	100%
Pearson	0.723

4. Conclusions

This paper used deep learning technology to conduct in-depth discussions on English speech recognition and quality evaluation models, and used deep learning English speech recognition models, multi-parameter English speech quality evaluation models and methods to conduct in-depth research, so as to obtain better methods for English speech recognition and quality assessment and facilitate further research in this field. By comparing different algorithms and models, the better algorithms and models can be found.

Acknowledgement

This study is the culminating achievement of the 2023 General Teaching Reform Project "Teaching Practice and Curriculum Reform: Implementing Ideological-Political Education in SPOC English Listening and Speaking Classes" (Project No. TDJG-23-Y48) at Jingdezhen Ceramic University.

References

[1] Abbasinia S and Pouralvar K. Evaluation of Speech Processing Capability in Computer Games to Improve the Process of Learning English Pronunciation: An Action Research. Journal of Rehabilitation Sciences, 2019, 15(5):280-285.

[2] Davila A M. Using the Shadowing Technique to Improve Ecuadorian English Learners' Speaking Intelligibility 1, Angel M. Dávila and International Journal of Current Research, 2018, 10(12):76770-76772.

[3] Li X. Characteristics and rules of college English education based on cognitive process simulation. Cognitive Systems Research, 2019, 57(OCT.):11-19. https://doi.org/10.1016/j.cogsys.2018.09.014

[4] Chika, Fujiyuki, Sayoko. An Evaluation of English Pronunciation of Japanese EFL Learners Using Multiple Metrics. Journal of the Phonetic Society of Japan, 2018, 22(2):39-43.

[5] Piotrowska M, Czyewski A, Ciszewski T. Evaluation of aspiration problems in L2 English pronunciation employing machine learning. The Journal of the Acoustical Society of America, 2021, 150(1):120-132. https://doi.org/10.1121/10.0005480

[6] Lasi F. A Study on the Ability of Supra-Segmental and Segmental Aspects in English Pronunciation. Ethical Lingua Journal of Language Teaching and Literature, 2020, 7(2):426-437. https://doi.org/10.30605/25409190.222

[7] Wang X. The framework of the multi-parameter evaluation index system for college spoken english based on deep learning theory. Revista de la Facultad de Ingenieria, 2017, 32(15):583-590.

[8] Suparman U, Ridwan R and Hariri H. Overcoming Students' English Pronunciation in Remote Area, Indonesia. Asian EFL Journal, 2020, 27(4):213-229.

[9] Gu X. A study on college students' english level evaluation model based on cloud services platform. Boletin Tecnico/Technical Bulletin, 2017, 55(4):299-305.

[10] Han Y. Evaluation of English online teaching based on remote supervision algorithms and deep learning. Journal of Intelligent and Fuzzy Systems, 2020, 01(5):1-12.

[11] Rani S, Tina A A. The Impact of Bangla Regional Dialect on the Pronunciation of English at Tertiary Level. Humanities & Social Sciences Reviews, 2020, 8(2):513-522. https://doi.org/10.18510/hssr.2020.8259

[12] Zhou W, Deterding D and Nolan F . Intelligibility in Chinese English Spoken in Central China. Chinese Journal of Applied Linguistics, 2019, 42(4):449-465. https://doi.org/10.1515/CJAL-2019-0027

[13] Maslova A and Kolesnikova A. Through the eyes of high school students: Which English pronunciation norm to study in Russia? Vestnik - Moskvoskogo Universiteta, 2019, 19(2):115-123.

[14] Suciati S and Diyanti Y. Suprasegmental Features of Indonesian Students' English Pronunciation and the Pedagogical Implication. SAGA Journal of English Language Teaching and Applied Linguistics, 2021, 2(1):9-18. https://doi.org/10.21460/saga.2020.21.62

[15] Phuong T. Who Should Teach English Pronunciation? -Voices of Vietnamese EFL Learners and Teachers. Journal of Asia TEFL, 2021, 18(1):125-141. https://doi.org/10.18823/asiatefl.2021.18.1.8.125

[16] Thanh T. Vietnamese EFL Learners' Perspectives of Pronunciation Pedagogy. Asian EFL Journal, 2019, 23(6):180-201.

[17] Wu X. AHP-BP-Based Algorithms for Teaching Quality Evaluation of Flipped English Classrooms in the Context of New Media Communication. International Journal of Information Technologies and Systems Approach, 2023, 16(2), 1-12.