Hindawi
Scientific Programming
Volume 2023, Article ID 9769741, 1 page
https://doi.org/10.1155/2023/9769741

Retraction

Retracted: Construction of the Open Oral Evaluation Model Based on the Neural Network

Scientific Programming

Received 8 August 2023; Accepted 8 August 2023; Published 9 August 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] Z. Chen, X. Zhang, Z. Li, and A. Li, "Construction of the Open Oral Evaluation Model Based on the Neural Network," *Scientific Programming*, vol. 2021, Article ID 3928246, 11 pages, 2021.

Hindawi
Scientific Programming
Volume 2021, Article ID 3928246, 11 pages
https://doi.org/10.1155/2021/3928246

Research Article

Construction of the Open Oral Evaluation Model Based on the Neural Network

Zhixin Chen, Xu Zhang, Zhiyuan Li, and Anchu Li

Northeast Electric Power University, Jilin City, Jilin 132012, China

Correspondence should be addressed to Xu Zhang; 20132487@neepu.edu.cn

Received 20 July 2021; Accepted 7 September 2021; Published 22 September 2021

Academic Editor: Bai Yuan Ding

According to the problem of low efficiency and low scoring accuracy of the traditional oral language scoring system, this study builds an open oral language evaluation model based on the basic principles of deep learning technology. Firstly, the basic methods of the convolutional neural network (CNN) and long short-term memory (LSTM) neural network are introduced. Then, we combine the convolutional neural network (CNN) and long short-term memory (LSTM) neural network to design an open oral scoring model based on CNN + LSTM, which divides the oral evaluation model into the speech scoring model and text scoring model and makes a specific implementation of two scoring models, respectively. An experimental environment is then built to preprocess the data, and finally, the model built in this study is trained and simulated. The experimental results show that the CNN + LSTM network evaluation model has a better comprehensive scoring performance, higher scoring efficiency, and higher accuracy and has feasibility and practicability.

1. Introduction

The rapid development of the internet information level has promoted the continuous progress of education. In traditional oral English learning, teachers mostly adopt the manual correction method to students' oral test and training results, which is tedious, time-consuming, and inefficient and is not conducive to the long-term development of students' oral English learning. In recent years, computer-assisted learning system has been widely used in the field of education, and more and more students and teachers in China begin to use computer-assisted oral learning. However, the traditional computer-assisted oral learning system has some shortcomings in the open oral evaluation, such as the lack of intelligence and low performance. It is an inevitable trend for the development of education to design and implement an intelligent and open oral scoring system.

How to improve the performance and efficiency of the open oral evaluation model is still the focus of many scholars. According to the demand of diagnostic oral English proficiency assessment, this paper puts forward corresponding solutions and strategies for the demand of the oral English assessment model. Based on the GESE oral test method, a formative hierarchical evaluation model of oral English teaching in higher vocational colleges is constructed, and its positive "backwash effect" law is obtained. Many of the research tasks proposed in [1] focus on English, and it is difficult to explore a richer variety of languages within the framework of the SLU task. Minor languages need to be updated in the training process of the SLU model. This problem is explained from three aspects: the input of the neural network, the comparison between the French version and the best setting in different ways, and the comparison with the most advanced methods. Tsai et al. [2] aimed at the use of mobile phones, mailboxes, and other tools in daily life, which are realized by recurrent neural networks (RNNs). The character language model based on LSTM can not only capture the Boltzmann statistics of the system but also reproduce the dynamics. It can be applied to capture the time evolution of typical trajectories in chemistry and biophysics. An acoustic speech method for SLD proposed in [3] is a new method based on the original SLD. It uses an attention-based neural network model to capture the language and a Gaussian smoothing method to locate language changes. It is more effective in dealing with code-switching of monolingual

segments, but its performance will be affected by the duration of monolingual segments, which needs further study. Peng [4] proposed that learning assessment can let students know whether they have achieved the ideal academic achievements and goals through computer English teaching, so as to further improve their shortcomings. Based on the neural network and artificial intelligence technology, this model can realize this evaluation. It is based on hearing, introduces wavelet entropy features, and applies it to the adaptive model. Combined with the control experiment, the performance of the model is analyzed. Mathematical statistics can directly show the effect. Research proves that the model can meet the expectations. With the deepening development of higher education, English teaching has been paid more and more attention. While improving the quality of education [5], learning evaluation is the most fundamental measure. By studying the existing problems and characteristics of the learning evaluation system, analyzing the defects of the traditional system, then summing up the shortcomings, and developing the advantages, this paper puts forward a college English teaching quality evaluation system based on information fusion and optimized RBF neural network decision-making algorithm. Jiang et al. [6] proposed an improved fuzzy RBF neural network model based on back-propagation learning, which combines a large number of problems in English teaching before, such as large quantity and complexity, and then optimizes with the development limitations and existing shortcomings of neural networks. This model is an objective reference to a university teaching case and puts forward an improved quality evaluation method, which is subjective and random, so that the evaluation results are more in line with the actual situation. Mohammed et al. [7] used Arabic, English, and French to evaluate the recognition performance of linear predictive coding (LPC) and/or mel-frequency cepstrum coefficient (MFCC) and artificial neural network (ANN) and also tested LPC and MFCC, hidden layer, different neurons in the hidden layer, and different transfer functions to illustrate their usability. Li [8] proposed a teaching evaluation model based on the improved BP neural network, which is helpful to improve English teaching quality. It summarizes and analyzes several teaching elements, designs the evaluation system, and improves the shortcomings in calculation by improving nonmonotone linear search and adaptive step change. Finally, through practical verification, the model can better evaluate English teaching and improve the effect. Hermanto et al. [9] used two models, namely, recurrent neural network (RNN) and statistical network, using the n-gram model. Compared with the statistical language model, the neural language model achieves better results in the field of machine translation. The latter two are deficient in accuracy and calculation speed, while the RNN has improved its evaluation scores in these aspects. Esan et al. [10] proposed a machine translation method based on the recurrent neural network model. By testing the model with manual and automatic evaluation techniques, combined with real manual evaluation, it shows that the model is smooth and usable, basically consistent with human judgment, and relevant and has high translation effect and accuracy.

However, the above research has not solved the intelligent problem of open oral learning, and the research direction needs to be further studied. The above literature research also puts forward many oral evaluation methods and has achieved good results. However, many solutions are based on machine learning or algorithms, which have low effect, do not realize intelligent processing, and cannot meet the requirements of oral evaluation. Although there are some neural network solutions, the translation efficiency and accuracy are relatively low, and this paper combined with the CNN + LSTM model can solve such problems. Based on this, this study combines the learning characteristics of the neural network, according to the basic principles and structural characteristics of the convolution neural network (CNN) and long short-term memory (LSTM) neural network, and designs and implements an open oral scoring model based on CNN + LSTM. The final results show that this model can improve the intelligent level of open oral learning, and the evaluation model has relatively high scoring efficiency and accuracy, which further shows that the evaluation model is feasible and practical.

This paper proposes that the neural network can solve the translation problem from the perspective of intelligent analysis of oral English, which can improve the accuracy of oral English and build a standard evaluation model. The CNN-LSTM model proposed in this paper has the highest evaluation accuracy, which can reach 82%, and the effect is the best. The key points will be to score the oral pronunciation and oral content separately, and finally, the total score after summation will be taken as the final score result of the examinee. When constructing the scoring model, this paper uses different neural networks to implement and test. Finally, the model with the highest relevance of the man-machine score will be applied to the actual correction task to help teachers reduce work pressure. At the same time, it can also be used as a tool for students' oral self-evaluation.

2. Basic Method of the Neural Network

2.1. Convolution Neural Network (CNN)

2.1.1. Fundamentals of the CNN. Convolutional neural network (CNN) is a representative algorithm of deep learning [11]. It is a feedforward neural network with convolution calculation and depth structure, which is very suitable for processing one-dimensional time-domain sequence data and image data. Compared with other neural networks, CNN can learn the original data efficiently and quickly, so as to extract the specific features of the data; that is, it has the ability of representation learning. CNN is widely used in computer vision, natural language processing, and other fields. The network has four remarkable characteristics, which are local perception, weight sharing (convolution operation), pooling processing, and multiconvolution kernel operation [12]. Among them, convolution operation is an operation that defines two integrable functions for convolution operation. It mainly includes two operation modes, namely, continuous convolution operation and discrete convolution operation.

Let both $f(x)$ and $g(x)$ belong to integrable functions in the real number field, and the new function

$J(x) = (f * g)(x)$ is the convolution of functions $f(x)$ and $g(x)$, in which $(f * g)(x) = (g * f)(x)$ and $(f * g)(x)$ belong to integrable functions.

The continuous convolution operation is expressed as

$$J(x) = (f * g)(x) = \int_{-} f(a)g(x-a)\mathrm{d}a. \tag{1}$$

The expression of the discrete convolution operation is

$$J(t) = (f * g)(t) = \sum_{n=-\infty}^{\infty} f(n)g(t-n). \tag{2}$$

Among them, the convolution operation mode in the CNN is discrete convolution, which can easily deal with discrete data problems in actual measurement.

2.1.2. Central Neural Network Structure. As shown in Figure 1, the network structure of the convolution neural network (CNN) is mainly divided into five network layers, namely, input layer, convolution layer, pooling layer, full connection layer, and output layer, belonging to multilayer perceptron (MLP) [13]. The following five layers are described in detail.

(1) Input Layer. The input layer of the CNN can process multidimensional data. Like other neural network algorithms for deep learning, CNN's learning method is calculated by the gradient descent algorithm, and the input data features of the CNN need to be standardized. Standardizing the features of the input data can greatly improve the learning efficiency of the CNN.

(2) Convolution Layer. Convolution layer, pooling layer, and full connection layer are all hidden layers in the CNN. The convolution layer has the function of feature extraction from the input data. There are many convolution kernels in the convolution layer, and the composition of convolution kernels is similar to that of neurons in the feedforward neural network [14], such as

$$\begin{aligned} Z^{l+1}(i,j) &= \left[Z^l \otimes \omega^{l+1}\right](i,j) + b \\ &= \sum_{k=1}^{K_l} \sum_{x=1}^{f} \sum_{y=1}^{f} \left[Z_k^l(s_0 i + x, s_0 j + y)\omega_k^{l+1}, \quad (i,j) \in \{0,1,\ldots,L_{l+1}\} L_{l+1} = \frac{L_l + 2p - f}{s_0} + 1. \right. \end{aligned} \tag{3}$$

In formula (3), the summation part is equivalent to solving a cross-correlation, B represents the deviation, Z^l and Z^{l+1} represent the input and output of the first layer, respectively, which can also be called the feature map, L_{l+1} represents the size of Z_{l+1}, if the length and width of the feature map are the same, $Z(i, j)$ represents the pixels of the feature map, f, s_0, p represents the number of channels of the feature map, and GG is the parameters of the convolution layer, which are convolution kernel size, convolution step size, and filling layer, respectively.

Under special circumstances, if the convolution kernel size $f = 1$ and convolution step size $S_0 = 1$ do not include filling, the cross-correlation calculation in the convolution layer is similar to matrix multiplication, thus constructing a fully connected network between convolution layers, as shown in the following:

$$\begin{aligned} Z^{l+1} &= \sum_{k=1}^{K_l} \sum_{i=1}^{L} \sum_{j=1}^{L} \left(Z_{i,j,k}^l \omega_k^{l+1}\right) + b = \omega_{l+1}^T Z_{l+1} + b, \\ L^{l+1} &= L. \end{aligned} \tag{4}$$

(3) Pooling Layer. After the convolution layer extracts the features, the pooling layer selects the features and filters the information. The pooled area selected by the pooled layer is similar to the step of scanning the feature map by the convolution kernel and is also determined by the pooled size, the pooled step size, and the number of filled layers [15]. Among them, the pooling layer is divided into Lp pooling and random pooling or mixed pooling. Lp pooling is mainly inspired by the hierarchical structure in the visual cortex, and its expression is as follows:

$$A_k^l(i,j) = \left[\sum_{x=1}^{f} \sum_{y=1}^{f} A_k^l(S_0 i + x, S_0 j + y)^p\right]^{1/p}. \tag{5}$$

In equation (5), S_0 is the pooling step size, (i, j) is the pixel, and p is the prespecified parameter. If $p = 1$ and Lp pooling takes the average value in the pooling area, it is mean pooling; if $p \longrightarrow \infty$ and Lp pooling takes the maximum value in the region, it is maximum pooling. These two pooling methods are the most commonly used methods in the CNN.

Hybrid pooling and random pooling are mainly extended from the concept of Lp pooling. Random pooling is randomly selected according to a specific probability in the pooled area, which can ensure that a part of nonmaximum excitation signals enter the next construction. Hybrid pooling can represent a linear combination of mean pooling and maximum pooling, as shown in the following:

$$A_k^l = \lambda L_1\left(A_k^l\right) + L_\infty\left(A_k^l\right); \quad \lambda \in [0,1]. \tag{6}$$

It is found that hybrid pooling and random pooling have a regularization function compared with mean pooling and maximum pooling, which can avoid overfitting of the CNN.

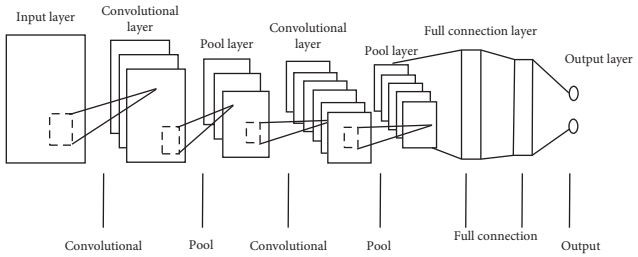

Figure 1: Central neural network structure.

(4) Fully Connected Layer. The fully connected layer belongs to the hidden layer in the traditional feedforward neural network. It is mainly located at the end of the hidden layer of the CNN, which only transmits signals to other fully connected layers. The main function of the full connection layer is to combine the extracted data features nonlinearly and output them.

(5) Output Layer. The upper part of the output layer is mostly a fully connected layer, and its structure and working principle are the same as those of the output layer in the traditional feedforward neural network. The output layer only outputs the image classification label through the logic function or normalized exponential function. The output layer recognizes objects according to the center coordinates, size, and classification of output objects. The output layer directly outputs the classification results of each pixel in the image to complete the semantic segmentation of the image.

2.2. Long Short-Term Memory (LSTM) Neural Network

2.2.1. Fundamentals of LSTM. Long short-term memory (LSTM) is a kind of time cycle neural network, which is designed to solve the long-term dependence of the RNN of the cycle neural network [16]. The main working mechanisms are mainly divided into the gating mechanism, forgetting mechanism, and circular memory mechanism. The three working mechanisms can effectively guarantee the normal operation of LSTM and make it play its long-term and circular memory functions, thus avoiding gradient disappearance and explosion. The following three mechanisms are analyzed in detail. LSTM output dimension setting is 300, return_sequences parameter is set to true, and the activation parameter is set to sigmoid.

(1) Gating Mechanism. In the LSTM network model, the key element of the gating mechanism is the activation function, which is usually sigmoid function and tan*h* function. *S* function is a commonly used activation function in neural networks. The characteristic of this function is that it can control the output real value within the range of [0, 1], as shown in the following:

$$\sigma(x) = \frac{1}{\left(1 + e^{-x}\right)}. \tag{7}$$

Tan*h* function, that is, hyperbolic tangent function: this function can control the output real value in the range of [−1, 1], with 0 bit as the center, which is just complementary to the *S* function without 0 as the center. The function is shown in the following:

$$\tan h(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}. \tag{8}$$

The gating mechanism can control the storage, memory, and update of information, and the basic principle is shown in formulas (9) and (10):

$$f(x) = \sigma(Wx + b), \tag{9}$$

$$\sigma(x) = \frac{1}{\left(1 + e^{-x}\right)}. \tag{10}$$

(2) Forgetting Mechanism. In the LSTM network model, forgetting mechanism must be realized based on the gating mechanism. By analyzing the influence of historical information on the memory unit, it is determined whether to retain or forget some information [17].

(3) Cyclic Memory Mechanism. Among them, in the LSTM network model, a new state *ct*, namely, memory unit, is added in the LSTM network model, and the main function of this memory unit is to transmit information circularly.

2.2.2. Remote Access Service Network Architecture. As shown in Figure 2, LSTM network structure is mainly divided into four parts: input gate, forgetting gate, memory unit, and output gate. The basic principles and methods of the four parts are analyzed in detail.

(1) Input Gate. The input gate mainly controls the filtering input data through the sigmoid function, wherein the input is mainly divided into two inputs, namely, the data directly input are controlled as shown in equation (11), and the other control input is stored in the candidate memory unit as shown in equation (12):

$$Z_i = \sigma\left(\sum_j K_{i,j} x_j^i + \sum_j W_{i,j} h_j^{(i-1)} + b_i, \right. \tag{11}$$

$$Z = \tan h\left(\sum_j K_{i,j} x_j^i + \sum_j W_{i,j} h_j^{(i-1)} + b. \right. \tag{12}$$

In the above formula, x_j^i represents the input data at time *t*, $K_{i,j}$ represents the input weight, $W_{i,j}$ represents the cycle weight of the forgetting gate, $h_j^{(i-1)}$ represents the output value of LSTM cells at the previous time, b_i represents the bias of the input gate, and *b* represents the bias of candidate memory cells.

(2) Forgetting the Door. Forgetting gate Z_f mainly uses the sigmoid function to map the input data x_j^i at time *t* in the range of [0, 1]. When the value is 0, it means that all information cannot pass this gate; when the value is 1, it means that all information is allowed to pass, as in the following:

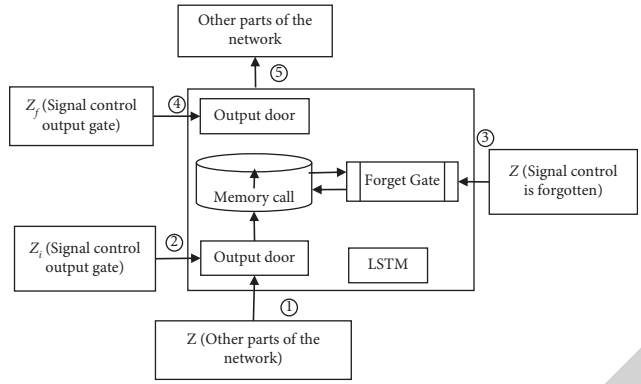

Figure 2: LSTM network architecture.

$$Z_f = \sigma\left(\sum_j K_{i,j}^f x_f^j + \sum_j W_{i,j}^f h_j^{(i-1)} + b_f.\right. \tag{13}$$

In equation (13), x_j^i represents the data input at time t; $K_{i,j}^f$ represents the input weight; $W_{i,j}^f$ represents the cyclic weight of the forgetting gate; $h_j^{(i-1)}$ represents the cell output value of LSTM at last time; $_b^f$ stands for the offset of the forgetting gate [18].

(3) Memory Unit. The memory cell C_i is cyclically updated mainly through the candidate memory cell Z and the memory cell C_{i-1} at the previous time and is adjusted by the input gate Z_i and the forgetting gate Z_f, as shown in the following:

$$C_i = Z_f \cdot C_{i-1} + Z_i \cdot Z. \tag{14}$$

(4) Output Gate. The output gate Z_o mainly controls whether the memory unit C_i outputs through the sigmoid function. When the value is 0, it means that all information cannot pass this gate [19]; when the value is 1, it means that all information is allowed to pass, as shown in equation (15). The final output value H_i of LSTM cells is controlled by the $\tan h$ activation function, as shown in equation (16):

$$Z_o = \sigma\left(\sum_i K_{i,j}^o x_j^i + \sum_j W_{i,j}^o h_j^{(i-1)} + b_o,\right. \tag{15}$$

$$h_i = Z_o \cdot \tan h\left(C_i\right). \tag{16}$$

In the above formula, x_j^i represents the data input at time t; $K_{i,j}^o$ represents the input weight; $W_{i,j}^o$ represents the cyclic weight of the output gate; $h_j^{(i-1)}$ represents the cell output value of LSTM at last time; $_b^o$ represents the offset of the output gate.

3. Construction of the Open Oral Scoring Model Based on CNN + LSTM

Combined with the basic principles and network structure of the CNN and LSTM neural network models, this study proposes an open oral scoring model based on the CNN + LSTM neural network. Because the model data trained by the traditional machine learning model and BP neural network model have low accurate fitting degree, the correlation between the manual feature extraction and manual score is low. This study breaks through the limitations of artificial extraction through the one-dimensional convolution neural network and long short-term memory network to integrate and process data more effectively, combined with the CNN + LSTM neural network to build the speech scoring model and text scoring model. Among them, the cyclic neural model has a large computational workload when dealing with lengthy sequence data. In this study, before the LSTM neural network, the one-dimensional convolution neural network is used to preprocess the data to shorten the sequence data, so as to improve the computational efficiency, extract more accurate features, and input them to the LSTM layer for processing.

As shown in Figure 3, the designed speech scoring model is mainly composed of MFCC feature vectors; 2 convolution blocks; bidirectional LSTM layer, namely, BLSTM, MeanOverTime layer, and full connection layer. The specific steps are as follows:

(1) Input the MFCC eigenvector firstly

(2) The filters in convolution block 1 and convolution block 2 convolve the input MFCC feature vectors in a translation way, and after the calculation is completed, the data are pooled by using the maximum pooling layer in the convolution block [20]

(3) After the data are processed by the convolution layer, they are input to the bidirectional LSTM network in the lower layer, that is, BLSTM, and the speech features are better extracted and fitted by the bidirectional LSTM network

(4) After feature extraction and fitting, the MeanOverTime layer is used to sum and average the corresponding positions of N vectors with equal lengths output by the upper LSTM network, and finally, a one-dimensional vector is output

(5) Finally, the one-dimensional vector output by the MeanOverTime layer is fitted through the full connection layer, so as to output the corresponding speech score results

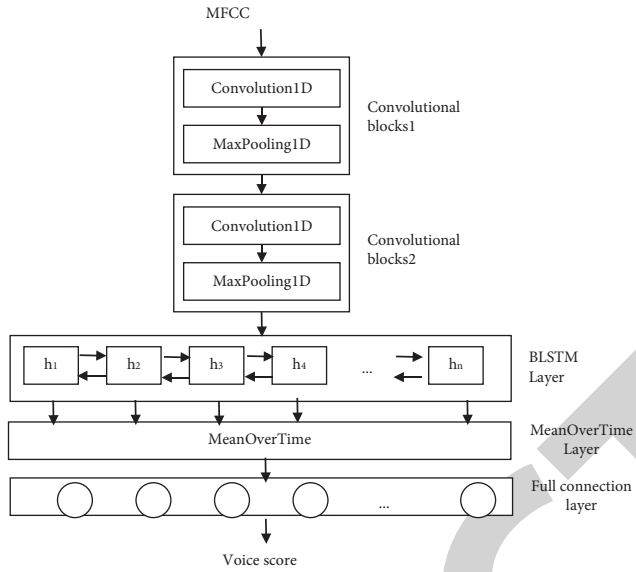

Figure 3: Speech scoring model based on CNN + LSTM.

As shown in Figure 4, the text scoring model based on CNN + LSTM is mainly composed of five layers: word embedding layer, one-dimensional convolution layer, LSTM layer, MeanOverTime layer, and full connection layer. The main function of the MeanOverTime layer is to sum and average the corresponding positions of N equal length vectors (including all intermediate state outputs of the LSTM network) output by the LSTM layer and finally output a one-dimensional vector. Finally, the full connection layer combines the one-dimensional vector and outputs the corresponding speech scoring results.

The specific steps of text scoring design are as follows:

(1) Firstly, the word embedding matrix based on the GloVe model is designed, and the dimension size of the word embedding layer is set to 50. If words cannot be found in the matrix, the word embedding layer vector is set to 0.

(2) The main function of the one-dimensional convolution layer is to shorten the length of the network input sequence and better extract speech features.

(3) The convolution layer is followed by the LSTM layer, which uses the function of selecting "memory" and "forget" information of the LSTM network to better extract and fit text features.

(4) MeanOverTime layer in the text scoring model has the same function as the MeanOverTime layer in the speech scoring model.

(5) Finally, the one-dimensional vector output by the MeanOverTime layer is fitted by the full connection layer, so as to output the corresponding text scoring results.

4. Experimental Verification

4.1. Environmental Construction. To verify the feasibility of the speech scoring model and text scoring model based on the CNN + LSTM network, the experimental data are the recordings of 600 students' open oral test in a university in Hebei province. The recording files are all in the MP3 form, and the audio attributes are 16 bits and 16 kHz sampling rates. The spoken pronunciation and content are scored by manual scoring. Before training and testing, the FFmpeg tool is used to convert the MP3 format of the recording to the PCM format. Finally, the dataset is divided into two parts: one is the training dataset and the other is the test dataset. Among them, the training set is 480 pieces of data, and the test set is 120 pieces of data.

In the above, this study constructs a speech scoring model and a text scoring model based on the CNN + LSTM network. For these two models, this study uses the RMSProp optimizer and mean square error loss function to train. The training iteration times are set to epoch 50, the batch value is set to 10, and the learning rates are set to 0.001 and 0.01, respectively. On the contrary, because the activation function is used in the output layer of the two scoring models, only the scoring results in the range of 0–1 can be obtained. Therefore, this study needs to normalize the manual scoring results before training the two models. The normalization process is shown in the following:

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}}. \tag{17}$$

4.2. Data Preprocessing. In the scoring model based on CNN + LSTM constructed in this study, the working form of the scoring model is to transform spoken recording and speech recognition text into a numerical vector representation. In this study, too many datasets are selected, which will lead to jumbled experimental process and increase the difficulty of the experiment. Therefore, it is necessary to preprocess the data in order to achieve better experimental results.

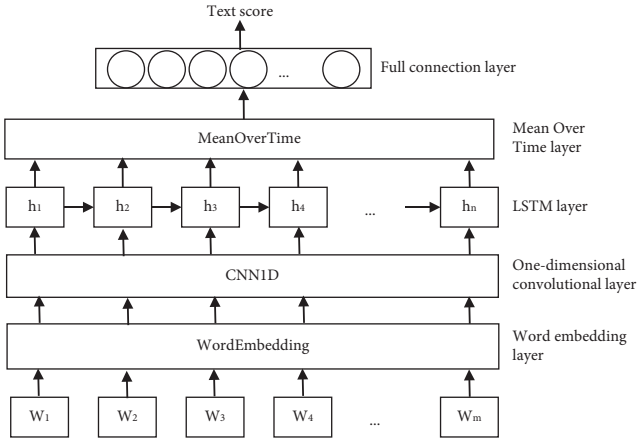

Figure 4: Text scoring model based on CNN + LSTM.

4.2.1. Data Cleaning. In the above dataset, due to the nonstandard spoken language of most students and the recognition defects of the speech recognition system itself, there are large errors in feature extraction of the tested speech recognition text, and the accuracy rate is relatively low, thus affecting the final recognition results. Therefore, this experiment cleans the experimental data, filters and removes some useless spoken words, which make the speech text structure clearer, and increases the effectiveness of the generated model. In this paper, adaptive filter (LMS) is used for data cleaning. Voice data formats include MP3 format, VOC format, and Au format. In the system, it is expressed as binary format coding. The system samples audio data in KB/s.

4.2.2. Feature Extraction. Feature extraction is very important in the scoring model, and the reliability and accuracy of the scoring model are determined by feature extraction. In the neural network scoring model based on CNN + LSTM constructed in this study, two kinds of features, speech class and text class, are mainly extracted. Specific feature extraction is shown in Table 1. Speech class feature extraction is directly based on the speech signal, and text class feature extraction is based on the output of the speech recognition engine.

According to the summary of feature extraction categories in Table 1, this study selects four phonetic features to evaluate students' oral pronunciation quality, fluency, and content richness. Among them, the specific function of the rate of speech (ROS) is to describe oral fluency, and the expression is

$$\mathrm{ROS} = \frac{N_{\mathrm{words}}}{t - t_s}. \tag{18}$$

In equation (18), N_{words} is the total number of words in students' spoken English, T is the total duration of oral recording, and t_s is the mute duration in recording.

In this study, five text features are selected to evaluate the oral content of candidates. Text feature extraction can better reflect the richness of the oral content and effectively deal with special data. In text feature extraction, grammar is regarded as the standard to judge students' oral English. Through part-of-speech tags, we can judge whether there are grammatical problems in the spoken content in vocabulary. The row label is shown in Table 2.

In the text extraction of speech recognition, grammatical errors exist in sentences and are difficult to be found. In this study, the correct rate of text grammar is calculated by

$$\mathrm{correct\_ratio} = \frac{N_g}{N_s}. \tag{19}$$

In equation (19), N_S represents the total number of ternary or quaternary tag combinations in the text, and N_g represents the total number of correct tag combinations.

4.2.3. Data Conversion. In this study, a pretrained word embedding model is used to transform the speech recognition text into a vector representation, as shown in Figure 5.

For spoken speech recording data, the extracted mel-frequency cepstrum coefficient (MFCC) is used as the input of the speech scoring model, as shown in Figure 6.

4.3. Empirical Results

4.3.1. Indicators for Evaluating System Performance. In the scoring system constructed in this study, the speech and text scoring model is based on the CNN + LSTM neural network. For the comprehensive evaluation and comparison of the scoring performance of the two models, this study sets up three evaluation indicators to quantitatively analyze the experimental results.

(1) Pearson's correlation coefficient is the most widely used performance test index in the field of automatic scoring. The main function of this coefficient is to reflect the linear correlation between two sequences, as shown in the following:

$$\rho_{X,Y} = \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y}. \tag{20}$$

In formula (20), $\mathrm{cov}(X, Y)$ is denoted as the oblique variance of X and Y, and σ_X and σ_Y are both standard deviations. X is the system score, Y is the teacher score, and the range of $\rho_{X,Y}$ is [−1, 1]. If the result is positive, it indicates that X and Y are positively correlated, and if it is negative, it indicates that X and Y are negatively correlated. The closer the absolute value of this coefficient is to 1, the higher the correlation between X and Y.

(2) Average difference of the man-machine score: it mainly describes the deviation degree between the machine score and manual score. The indicator is represented by the letter d, as shown in the following:

$$d = E\left|S_{\mathrm{Machine}} - S_{\mathrm{Human}}\right|. \tag{21}$$

(3) Accuracy: this experiment uses the method of the maximum error of man-machine scoring to judge

Table 1: Summary of the feature extraction of the speech class and text class.

Feature category	Name to be signed	Brief description of characteristics
Phonetic class	Articulation rate	Speed of speech
	Num silence	Number of voice pauses
	Posterior score	Postpronunciation verification probability score
	Speaking ratio	Ratio of effective speaking time to total recording time
Text class	Content length	Total number of words in the text
	Unique words	Number of nonrepeating words in the text
	Syntactic tree depth	Sum of all syntactic tree depths in the text
	Semantic similarity	Semantic similarity between the text and topic
	Good grammar ratio	Correct rate of text grammar

Table 2: Part-of-speech tags.

Part-of-speech tags	Description
NN	Noun (singular)
NNS	Noun (plural)
VB	By bar (prototype)
VBD	Verb (past tense)
VUN	Moving river (past participle)
JJ	Adjectives
RB	Adverb
IN	Subordinate conjunctions
CC	Conjunctions
PRP	Personal pronoun

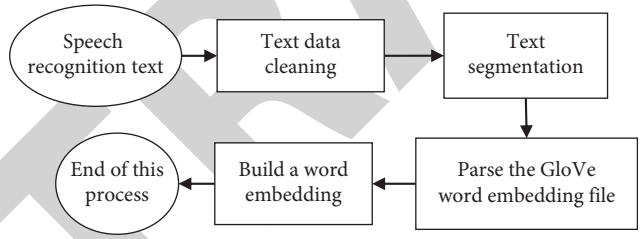

Figure 5: Text vector flow.

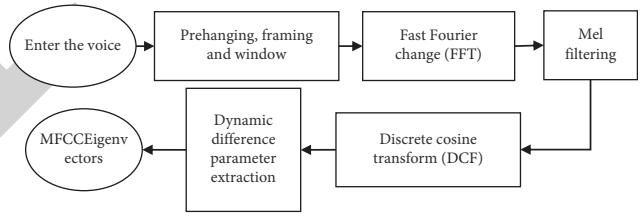

Figure 6: MFCC feature extraction flow.

whether the scoring result is accurate or not. If the difference between the manual score and machine score is less than the maximum error, it indicates that the score result is accurate.

Finally, the ratio of the number of accurate scoring results in the test set to the total number of samples is calculated, which is used as the accuracy index of the scoring model. In practical applications, the error of man-machine scoring is less than 1 point, which shows that the machine prediction result is accurate.

4.3.2. Analysis and Evaluation of the Scoring Model. In the CNN + LSTM neural network model, 9 features are extracted as the input of speech and text scoring models, and Pearson's correlation coefficients between each feature and manual scoring are calculated, respectively. The calculated results are shown in Tables 3 and 4.

It can be seen from Tables 3 and 4 that the Pearson correlation coefficients of num silence and speaking ratio are 0.44 and 0.42, respectively, which are highly correlated with manual scores. It shows that teachers pay more attention to

Table 3: Correlation between phonetic features and manual scores.

Phonetic features	Pearson's correlation coefficient
Articulation rate	0.37
Num silence	0.44
Posterior score	0.31
Speaking ratio	0.42

Table 4: Correlation between text class features and manual scoring.

Text class feature	Pearson's correlation coefficient
Content length	0.57
Unique words	0.59
Syntactic tree depth	0.27
Semantic similarity	0.33
Good grammar ratio	0.24

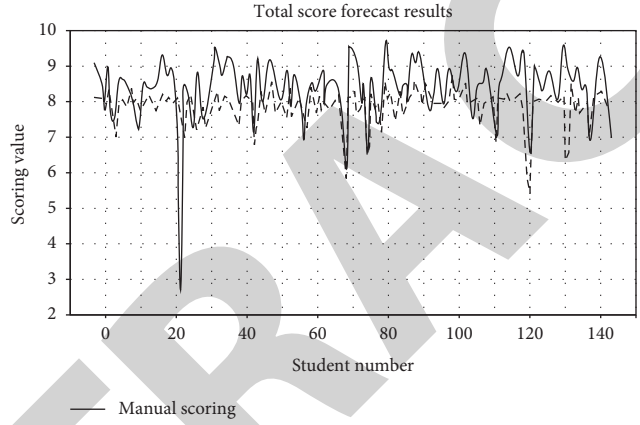

Figure 7: Scoring results of the BP scoring model.

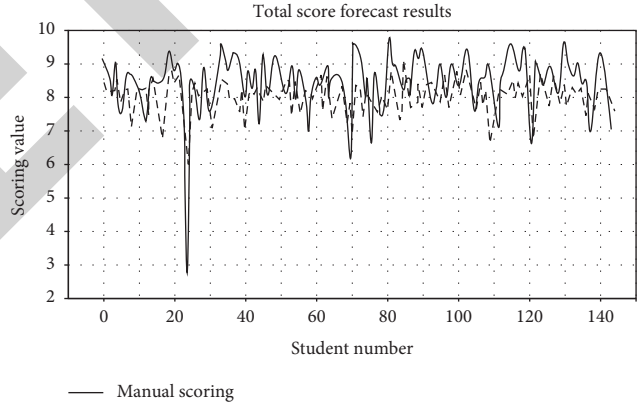

Figure 8: Scoring results of the CNN + LSTM neural network scoring model.

Table 5: Performance evaluation of the scoring model.

	Pearson's correlation	Average difference	Accuracy (%)
Manual scoring model	0.764	0.485	—
CNN + LSTM model	0.694	0.639	82.3
BP model	0.533	0.601	80.0

students' oral fluency and effective oral duration when grading oral English. The Pearson correlation coefficients of content length and unique words are 0.57 and 0.59, respectively, which have high correlation with manual scoring. It shows that teachers pay more attention to students' mastery of vocabulary and the richness of the oral content.

4.3.3. Test Results. As shown in Figures 7 and 8, 120 test data were selected to test the scoring models of the BP neural network and CNN + LSTM neural network, and the performance of these two scoring models was comprehensively evaluated through the three evaluation indexes set above.

It can be seen from Figures 7 and 8 that the CNN + LSTM neural network scoring model has a better fitting effect than the BP neural network scoring model, and the CNN + LSTM neural network scoring model has stronger adaptability. In Figures 7 and 8, the dotted line represents the standard actual test sample data. In the BP scoring model in Figure 7, it is quite different from the standard sample data, indicating that the evaluation effect is poor. In Figure 8, the difference between the CNN + LSTM model and the standard sample data is small, indicating that the evaluation effect is good.

It can be seen from Table 5 that the CNN + LSTM neural network model is better than the BP neural network model in the Pearson correlation coefficient and accuracy, and the difference between them in man-machine score correlation is obvious. CNN + LSTM is slightly better than the BP model in the average difference index.

5. Conclusion

To sum up, the open oral English scoring model based on CNN + LSTM designed in this study has better fitting effect and adaptability and can improve the intelligent level of open oral English learning. Compared with the traditional artificial scoring and BP neural network scoring model, the CNN + LSTM neural network scoring model has higher scoring accuracy and efficiency. Experimental results show that the CNN + LSTM neural network model has better scoring performance and has certain practicability when the training dataset is small.

Data Availability

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

Conflicts of Interest

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

References

[1] S. Ghannay, C. Servan, and S. Rosset, "Neural networks approaches focused on french spoken language understanding: application to the MEDIA evaluation task," in *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 2722–2727, Barcelona, Spain, 2020.

[2] S. T. Tsai, E. J. Kuo, and P. Tiwary, "Learning molecular dynamics with simple language model built upon long short-term memory neural network," *Nature Communications*, vol. 11, no. 1, p. 5115, 2020.

[3] S. E. Chazan, S. Gannot, and J. Goldberger, "Attention-based neural network for joint diarization and speaker extraction," in *Proceedings of the 2018 16th International Workshop on Acoustic Signal Enhancement*, pp. 301–305, Tokyo, Japan, 2018.

[4] N. Peng, "Performance evaluation of English learning through computer mode using neural network and AI techniques," *Journal of Intelligent and Fuzzy Systems*, vol. 40, no. 3, pp. 1–11, 2020.

[5] Y. Chen, "College English teaching quality evaluation system based on information fusion and optimized RBF neural network decision algorithm," *Journal of Sensors*, vol. 2021, Article ID 6178569, 9 pages, 2021.

[6] Y. Jiang, J. Zhang, and C. Chen, "Research on a new teaching quality evaluation method based on improved fuzzy neural network for college English," *International Journal of Continuing Engineering Education and Life-Long Learning*, vol. 28, no. 3-4, p. 293, 2018.

[7] E. M. Mohammed, M. S. Sayed, A. M. Moselhy, and A. A. Abdelnaiem, "LPC and MFCC performance evaluation with artificial neural network for spoken language identification," *International Journal of Signal Processing Image Processing & Pattern Recognition*, vol. 6, no. 3, pp. 55–66, 2013.

[8] H. Li, "Application research of BP neural network in English teaching evaluation," *Telkomnika Indonesian Journal of Electrical Engineering*, vol. 11, no. 8, pp. 4602–4608, 2013.

[9] A. Hermanto, T. B. Adji, and N. A. Setiawan, "Recurrent neural network language model for English-Indonesian machine translation: experimental study," in *Proceedings of the International Conference on Science in Information Technology*, pp. 132–136, Yogyakarta, Indonesia, 2016.

[10] A. Esan, J. Oladosu, C. Oyeleye et al., "Development of a recurrent neural network model for English to Yorùbá machine translation," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 5, pp. 602–609, 2020.

[11] T. Maximilian, L. J. Lennart, and E. Nicole, "Deep learning-based detection and classification of geographic atrophy using a deep convolutional neural network classifier," *Graefes Archive for Clinical & Experimental Ophthalmology*, vol. 256, pp. 2053–2060, 2018.

[12] T. Liang, G. Yang, F. Lv, J. Zhang, Z. Cao, and Q. Li, "Convolutional neural networks for text classification with multi-size convolution and multi-type pooling," *Database Systems for Advanced Applications*, Springer, vol. 10829, pp. 3–12, Berlin, Germany, 2018.

[13] A. H. Fath, F. Madanifar, and M. Abbasi, "Implementation of multilayer perceptron (MLP) and radial basis function (RBF) neural networks to predict solution gas-oil ratio of crude oil systems," *Petroleum*, vol. 6, no. 1, pp. 80–91, 2020.

[14] Y. Gao, G. Cai, H. Li et al., "Diagnosis and prognosis prediction of ovarian cancer with feedforward neural network by mining real-world laboratory tests," *Gynecologic Oncology*, vol. 159, pp. 338-339, 2020.

[15] L. Liu, C. Shen, and A. Hengel, "Cross-convolutional-layer pooling for image recognition," *IEEE Transactions on Pattern*

Analysis & Machine Intelligence, vol. 39, no. 11, pp. 2305–2313, 2015.

[16] S. Basumallik, R. Ma, and S. Eftekharnejad, "Packet-data anomaly detection in PMU-based state estimator using convolutional neural network," *International Journal of Electrical Power & Energy Systems*, vol. 107, pp. 690–702, 2018.

[17] L. Pullagura, "An effective LSTM network model for accurate prediction of delays in Indian railway networks," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 4, pp. 5111–5116, 2020.

[18] H. Wang, M. Li, and X. Yue, "IncLSTM: incremental ensemble LSTM model towards time series data," *Computers & Electrical Engineering*, vol. 92, no. 8, Article ID 107156, 2021.

[19] G. Cheng, P. Zhang, and J. Xu, "Automatic speech recognition system with output-gate projected gated recurrent unit," *IEICE Transactions on Information and Systems*, vol. E102.D, no. 2, pp. 355–363, 2019.

[20] H. Yamaguchi, Y. Hashimoto, G. Sugihara et al., "Three-dimensional convolutional autoencoder extracts features of structural brain images with a "diagnostic label-free" approach: application to schizophrenia datasets," *Frontiers in Neuroscience*, vol. 15, Article ID 652987, 2021.