Hindawi
Scientific Programming
Volume 2022, Article ID 6884637, 14 pages
https://doi.org/10.1155/2022/6884637

Research Article

Automatic Scoring of Spoken Language Based on Basic Deep Learning

Zhong Cheng[1,2] **and Zonghua Wang**[1]

[1]*School of Foreign Languages, Anhui University of Science and Technology, Huainan 232001, China*
[2]*School of English Studies, Shanghai International Studies University, Shanghai 201620, China*

Correspondence should be addressed to Zonghua Wang; 2011046@aust.edu.cn

Received 26 November 2021; Revised 13 December 2021; Accepted 3 January 2022; Published 25 January 2022

Academic Editor: Hangjun Che

The oral English test in domestic universities requires teachers to modify a large number of candidates' oral recordings. This is the work of using time repeatedly. Using the CALL system to realize the automation of conversation recording can reduce the burden of teachers' work. Therefore, it is of great practical significance to develop an automatic and accurate scoring system for oral English. With the development of artificial intelligence, deep learning technology has been gradually applied in various fields. Similarly, in the application of oral scoring, deep learning technology makes the implementation of such a system possible. Based on the deep learning technology, this paper proposes an automatic scoring algorithm for spoken language and implements a detailed design and evaluation system. The system consists of two modules. The pronunciation standard of spoken pronunciation and the content of spoken pronunciation are scored, and the sum of these two scores is the final score. Finally, this paper uses 650 oral English recordings from a college English test to train the artificial neural network. Experimental results show that if the training data set is small, the BP network model can obtain better comprehensive evaluation performance.

1. Introduction

In recent years, information technology has been widely used in the field of education. In language education, the popularity of English education in China is getting higher and higher, and the traditional language education methods are difficult to meet people's needs [1]. In this context, Computer Assisted Language Learning (CALL) has become a research hotspot [2]. CALL system not only is used in online education, but also includes English education platforms such as text, image, audio, and video, which also play an important role on the Internet. Instead of teachers automatically revising students' test questions and homework between classes, teachers are freed from taking time to revise. The automatic correction system like now has almost reached the completely correct level in the correction task facing objective problems. As for composition questions and oral questions, automatic revision is still the research focus that should be broken through. Oral problems can be divided into two types [3]. One is retelling, reading aloud, and reciting what is known. Another point is that candidates are free to play games around specific problems and topics. We are often called "open spoken English." With the development of speech recognition technology, the first question can be well evaluated by comparing and analyzing the examinee's pronunciation with the standard pronunciation [4], such as using the classical Goodness of Pronunciation (GOP) algorithm. In addition, it is necessary to comprehensively evaluate candidates' answers from multiple dimensions, for example, fluency, rhythm, intonation, richness of vocabulary, and meaning. For a long time, the research on open oral scoring technology has not made great breakthrough. With the development of machine learning technology, some scholars have studied how to apply it to automatic oral evaluation. Thus, the famous automatic scoring system of speed competition appeared [5].

In universities, the English skills training service system of universities is used for examination in the teaching of

situational English. At least mid-term and final evaluations are conducted every semester. In two exams, each teacher is usually responsible for the educational tasks of multiple classes. Because of the complicated manual grading method, teachers' burden is aggravated, and their educational energy is insufficient. If we study the intelligent correction system needed in the oral test of senior high schools in China [6], we can greatly reduce the pressure on teachers, and teachers can put more strength into practical teaching activities to improve their teaching ability.

There are extremely few ways to score speakers with speech disorders. We study an automatic speech score, which is a kind of assessment for speakers with language disorders [7]. With the development of society and the integration of global economy, people's demand for English learning is increasing day by day, so the research on automatic assessment of oral proficiency is particularly important. In the previous automatic evaluation system, recording conditions are a challenge for learners' pronunciation, noisy sounds, etc. In addition, it is necessary to deal with nonfluent, nongrammatical, and spontaneous sounds with unknown potential text. To solve these series of problems, we propose a method of combining speech recognition system based on deep learning with Gaussian process (GP) scorer, which is a measure to evaluate the performance of rejection scheme [8].

2. Related Technology Research

2.1. Research on Natural Language Processing Related Technologies

2.1.1. Latent Semantic Analysis. Latent Semantic Analysis, also known as Latent Meaning Index, is a document modeling method for natural language processing [9]. Like the previous vector space model, LSA method also uses vectors to represent words and documents and judges the relationship between words and documents according to the relationship between vectors, which leads to two shortcomings: (1) vector space model uses correct sentence matching. (2) You cannot ignore the meaning of a word and provide semantic search. LSA solves the above problems by statistically analyzing a large number of text libraries and mapping documents from sparse n-dimensional space to low-dimensional space. Vector space is called inclusion space. The document modeling process using LSA method is as follows, shown in Figure 1:

(1) Analyze the document set and create a word-document matrix

(2) Singular value decomposition of word-document matrix

(3) Dimension reduction of the matrix after singular value decomposition

The TF-IDF is calculated by the following formula:

$$A_{i,j} = \frac{N_{i,j}}{N_j} \log\left(\frac{D}{D_i}\right). \quad (1)$$

		Terms				
		T1	T2	T3	…	Tn
Documents	D1	0.2	0.1	0.5	…	0.1
	D2	0.1	0.3	0.4	…	0.3
	D3	0.3	0.1	0.1	…	0.5
	…	…	…	…	…	…
	Dm	0.2	0.1	0.2	…	0.1

Figure 1: Word-document matrix.

Matrix S is an $m \times m$ dimensional diagonal matrix, and each value on the diagonal represents the importance of various topics, and this value is also called a singular value. Then, in Step 3, the matrix after SVD decomposition only stores the largest K topics of U, and a dimension descent process is performed, in which only K topic vectors corresponding to S and V are maintained. As shown in Figure 2, the resulting Matrix A can be expressed by the following formula:

$$A_{m\times n} = U_{m\times k} S_{k\times k} V^T_{k\times n}. \quad (2)$$

If you use query text to calculate the similarity of all the text in the document set, you need to map the query text to the meaning space:

$$q_{1\times k} = q_{1\times n} V_{n\times k} S^{-1}_{k\times k}. \quad (3)$$

2.1.2. Word Embedding. To score the spoken language, considering the learning model, it is necessary to use the neural network model to score the spoken content of the examinee [10, 11]. The existing model scoring has the following main problems: (1) there is only one word in the number vector; so if there are N words in the text, it needs to use the N-dimensional vector for coding. Therefore, if the number of nonrepeated words in the text is large, the dimension of the vector becomes large. In addition, as the number of neurons increases, the computation becomes more complex. (2) Simple hot coding scheme cannot describe the meaning relationship between words. Words can be represented by low-dimensional vectors. For words with similar meanings, the vector displays are also close, as shown in Figure 3. More abundant information can be embedded into low-dimensional vectors, which are represented by single hot coding and term embedding, respectively.

When using neural network to solve text problems, the network architecture shown in Figure 4 is usually used. The first layer of the network is the word embedding layer, which transforms the words in the input text into word vector representation. For example, the length of the word embedding vector is set to 50 for the text containing 20 words, and it becomes 20 through the word embedding layer. Two-dimensional matrix of ×50. The word embedding layer is interpreted as a dictionary model, and the word index and its corresponding word vector graph are stored in the dictionary. This model can be obtained through data training or

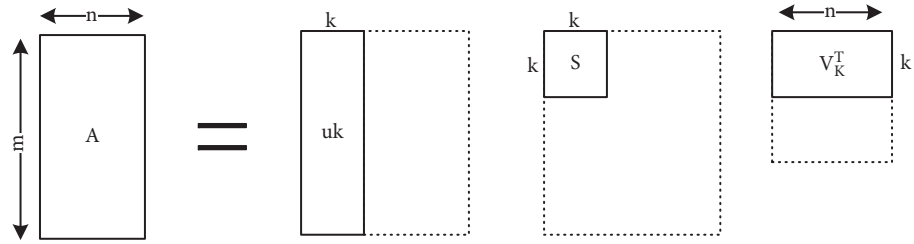

Figure 2: Svd matrix.

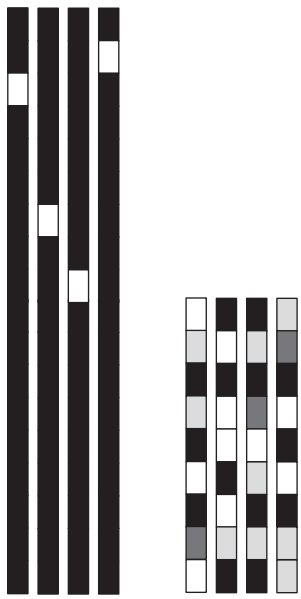

Figure 3: Word vectors represented by one-hot encoding and word embedding.

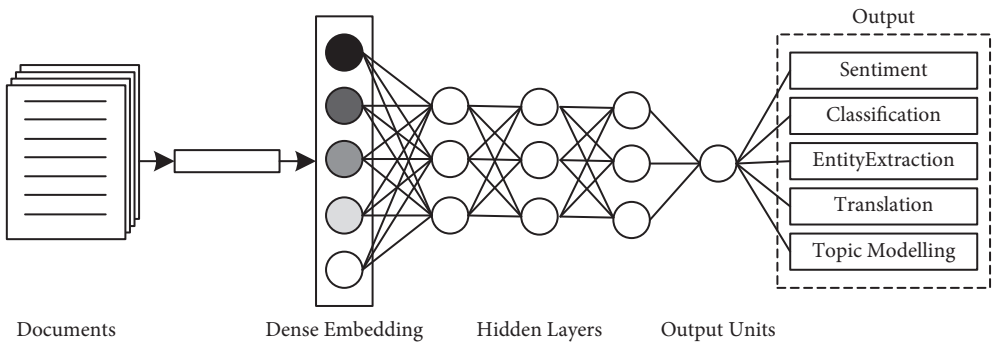

Figure 4: Neural network model architecture with word embedding layer.

loaded with trained models. Word 2 vector and GloVe are the commonly used models of preparation training language [12]. Based on the latter, this paper constructs a scoring model of oral content.

2.2. Research on Related Technologies of Scoring Model

2.2.1. Basic Concepts of Artificial Neural Networks. When an artificial neuron is stimulated, if the stimulus exceeds a certain threshold, the neuron will be activated and convey information to other neurons. The process of information transmission between neurons can be explained by Figure 5. Artificial neurons receive m input signals from other neurons. These input signals have a weight w during transmission. The weighted value can be abstractly understood as signal strength. After weighted addition, these input signals are processed by an "activation function" to generate an output signal Y.

The learning ability of neural network is strong because of its great activation function. If the activation function is not used, the network can only perform simple linear

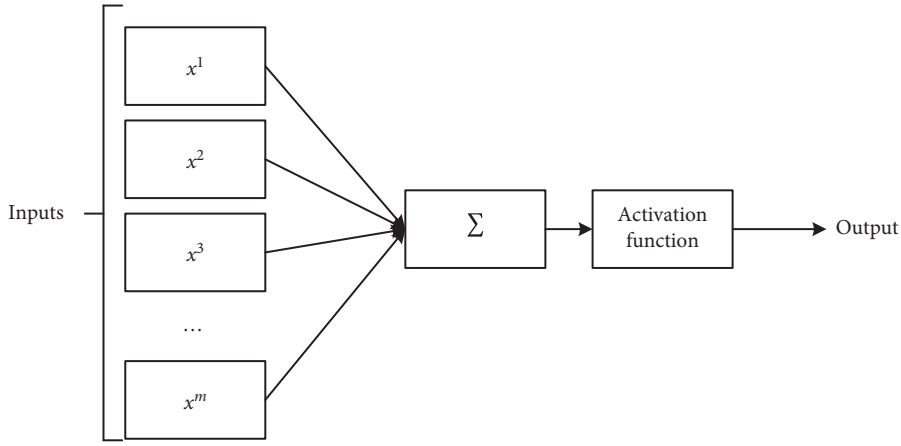

Figure 5: Mechanism of information transmission between neurons.

transformation, so the performance of this network is limited. On the other hand, the activation function introduces nonlinear elements into the network, which makes the neural network approximate to various nonlinear curves arbitrarily and makes the network have strong representation ability. The general active functions are Sigmoid, Hyperbolic Tangent, and Lireer, as shown in Figure 6, and there are function graphs of these three active functions.

The information transfer relationship between nodes is explained in the following formula:

$$y_j = \text{activation}\left(b_j + \sum_i x_i w_{ij}\right), \tag{4}$$

where x_i represents the output value of the i-th node of the previous layer (or the input value of the current node j), wij represents the weight value between the i-th node of the previous layer and the j-th node of the current layer, b_j represents the paranoid value of the j-th node of the current layer (the paranoid value is introduced to make the model converge better), and y_j represents the output value of the j-th node of the current layer.

2.2.2. Basic Concepts of Deep Learning. The differences between deep learning and traditional machine learning are as follows. Feature items are fully automated, so people do not have to go all out to find a more suitable initial input feature. Data becomes higher-level and more abstract display form through the network. This process is the core step in traditional machine learning. In Figure 7, the process can be described simply (or as close as possible to the expected result).

Briefly introduce some core terms contained in the above figure.

(1) Loss function: it is used to calculate the difference between the predicted data and the actual data of neural network.

(2) Optimizer: determine the algorithm to update the network weight by using the loss value, among which the commonly used optimization algorithms are Adam, SGD, and RMSProp.

(3) Backpropagation error: reverse transmission loss values are sent from the output layer to the input layer (the loss values obtained at each node are allocated by a weighted contribution ratio), and the weight values and polarization values in the network are updated using a gradient descent algorithm during transmission.

In the current field of deep learning, various deep learning models have been developed. In this paper, we pay attention to convolution neural network and cyclic neural network.

2.2.3. BP Neural Network. BP neural network has strong nonlinear mapping ability and can approximate any discontinuous function with high precision [13]. It is an extremely effective model to solve problems such as regression and classification. The training process of BP neural network is mainly divided into two stages [14]. The first stage is the forward propagation of signals, from the input layer to the hidden layer and finally to the output layer. In the second stage, backpropagation algorithm is used to backpropagate the error from the output layer to the hidden layer, and finally the weight and bias voltage are adjusted to the input layer in turn. When a network is trained using a large amount of data including a plurality of samples, the mean square error is exemplified as a loss function, and the mean square value of the error after forward propagation of the training data is as follows:

$$E = \frac{1}{N}\sum_{k=1}^{N}\left(y_k(i) - t_k(i)\right)^2, \tag{5}$$

where $y_k(i)$ represents the true output value of the i-th data sample, and $t_k(i)$ represents the predicted value obtained after the i-th data sample passes through the neural network. The BP neural network uses learning rate and gradient descent algorithm to update the connection weights and

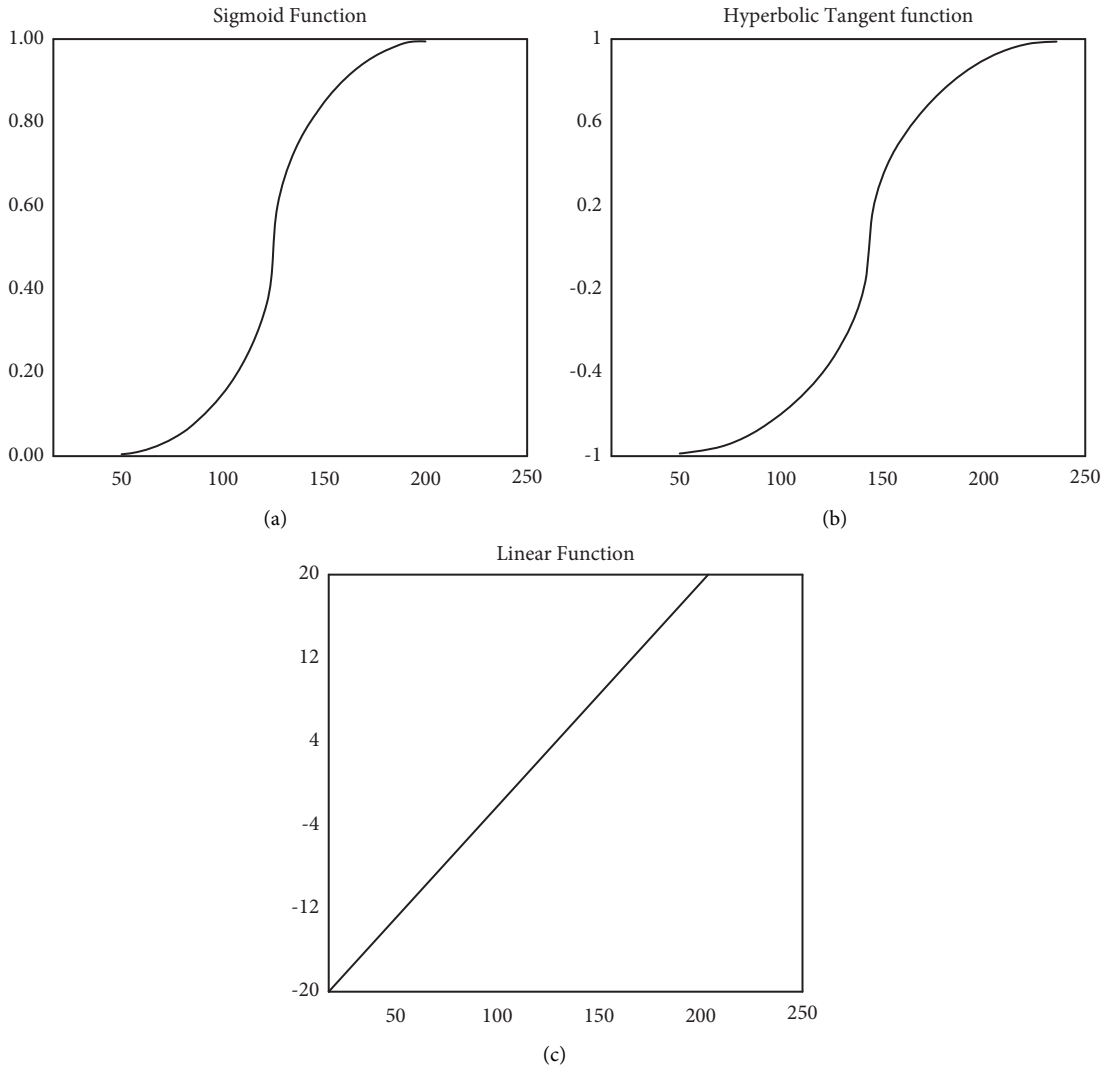

Figure 6: Activation function graph. (a) Sigmoid function image. (b) Image of hyperbolic tangent function. (c) Pure linear function image.

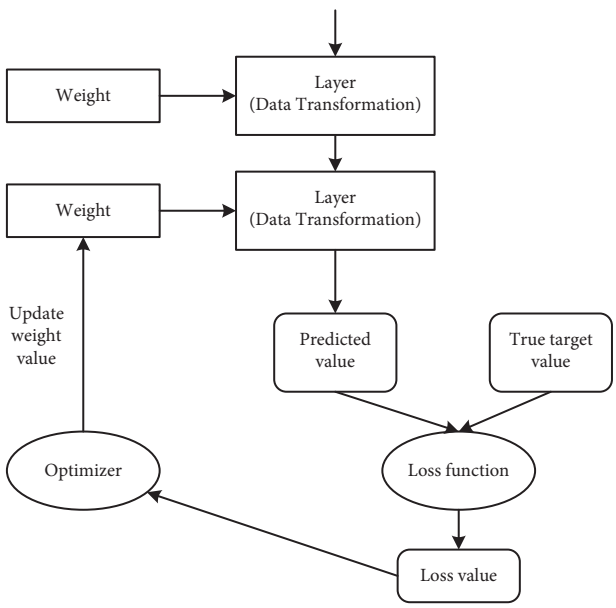

Figure 7: Working principle of neural network.

polarization values of each layer. The whole backpropagation process can be explained by the following formula:

$$E^{(n)} = \left(w_c\right)^T \cdot E^{(n+1)}, \tag{6}$$

$$w_{ji}^{(n-1)} = w_{ji}^{(n-1)} + \text{learn}_{\text{rate}} * \frac{\partial E_i^{(n)}}{\partial w_{ji}^{(n-1)}}, \tag{7}$$

$$b_j^{(n)} = b_j^{(n)} + \text{learn}_{\text{rate}} * \frac{\partial E_i^{(n)}}{\partial b_j^{(n-1)}}. \tag{8}$$

2.2.4. Convolution Neural Network. One-dimensional convolution neural network is well applied to sequence data, such as audio signals and text data, and in some cases, the performance of this network can match that of cyclic neural network [15, 16]. The computational cost is usually quite small, and the model can achieve better performance. As shown in Figure 8, as the operation principle of one-dimensional convolution network, feature is the data length of each feature. The network output data format after convolution operation is samples. The new step is the length of the feature sequence after the convolution operation, and filters are the number of convolution kernels.

2.2.5. Cyclic Neural Network. Cyclic neural networks (RNN) can circulate information in the network, but unlike networks such as CNN, their output only considers the influence of the previous input and does not consider the influence of other time inputs. In RNN, the output of each moment is not only related to the input of the current moment, but also related to the input of the previous moment. The network has the function of "storage." Therefore, RNN is extremely suitable for processing sequence data, especially text data. h_t in terms of o_t can be calculated by the following formulas:

$$h_t = \text{activation}\left(Uh_{t-1} + Wx_t + b\right), \tag{9}$$

$$o_t = Vh_t. \tag{10}$$

Conventional RNN model is only applied to the processing of short sequence data. In order to solve the problem of insufficient "long-term storage" capacity in traditional RNN networks, some researchers improve the model, which is called Short Term Storage Network (LSTM). LSTM model selectively adds new information and selectively forgets previously accumulated information by introducing grid control mechanism. A new state c_t is introduced in LSTM network for circulating information transmission.

The states of the hidden layer and the memory cell are represented by the following equations:

$$c_t^* = \tanh\left(W_c x_t + U_c h_{t-1} + b_c\right), \tag{11}$$

$$c_t = f_t \tanh \odot \tanh\left(c_t\right). \tag{12}$$

The states of the three gate controllers can be calculated from the following equations:

$$f_t = \sigma\left(W_f x_t + U_f h_{t-1} + V_f c_{t-1} + b_f\right), \tag{13}$$

$$i_t = \sigma\left(W_i x_t + U_i h_{t-1} + V_i c_{t-1} + b_i\right), \tag{14}$$

$$o_t = \sigma\left(W_o x_t + U_o h_{t-1} + V_o c_{t-1} + b_o\right). \tag{15}$$

3. Design of Oral Scoring System

3.1. Overall System Design. Combining deep learning technology and object-oriented design idea, the oral scoring system designed in this paper includes six modules as shown in Figure 9.

(1) Oral scoring module: call the scoring mode module, load the training scoring mode, automatically correct the oral data, and save the scoring results in Excel file form.

(2) Sound noise reduction module: in order to make the results of speech recognition and feature extraction more accurate, the examinee's spoken language is noise reduced.

(3) Speech recognition module: convert the examinee's dialogue recording into the corresponding text content through the speech recognition engine.

(4) Data processing module: this module mainly extracts spoken speech recording and speech recognition text. The score of CNN + LSTM is converted into digital display form for spoken language recognition text.

(5) Systematic evaluation module: analyze the evaluation results of main evaluation and evaluation models.

(6) Scoring Model Module: define the scoring model based on BP + CNN + LSTM, respectively, and save the training model for loading directly.

Intelligent spoken language evaluation refers to the dynamic process from audio to total point output and can be described as the scoring system in Figure 10 [17]. The speech recognition engine first performs noise reduction processing through the sound noise reduction module and then transfers the beautiful recording to the corresponding text content. The general scoring system fits the characteristic value according to the scoring model. Two scoring models are used here. Speech Scoring Model and Text Scoring Model are designed to improve the accuracy of the scoring system. In addition, in the actual correcting environment, the teacher also evaluates the examinee's conversation from the level of sound and content. This design is consistent with the manual scoring method. The design of the core module of the system is described in detail.

3.2. Design of Speech Noise Reduction Module. Because of the problem of the recording device, the recording of spoken language is often mixed with current sound and noise. This

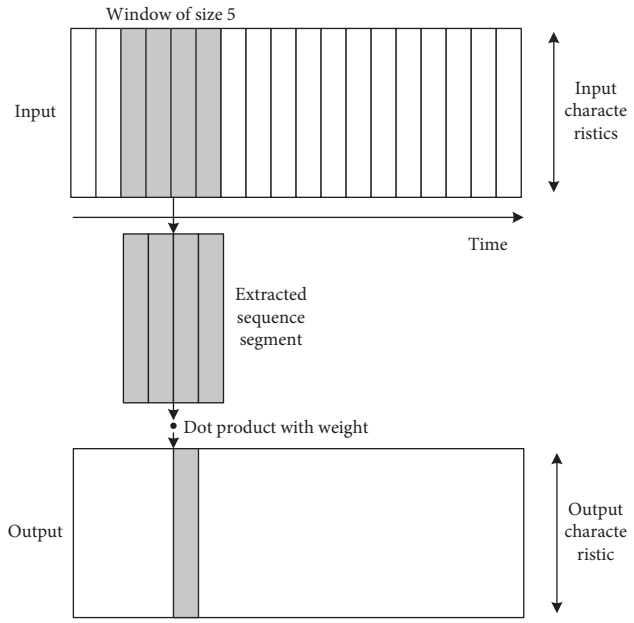

Figure 8: Working principle of one-dimensional convolution network.

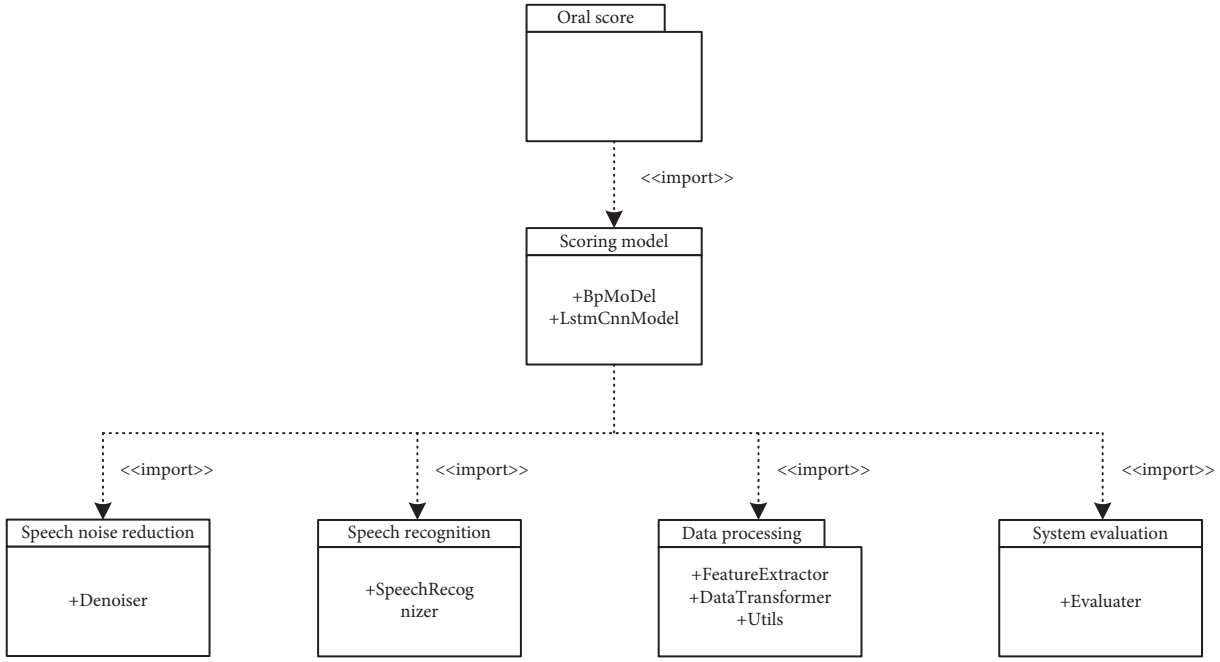

Figure 9: System module design.

affects the correctness of subsequent feature extraction and speech recognition. Traditional noise reduction methods use spectrum subtraction or adaptive filtering. In recent years, due to the successful application of learning in the field of sound processing, the use of deep learning technology in reducing sound noise has been improved and is popular. In this paper, RnNoise, an open source noise suppression library, is used to realize the header noise reduction module, in which RNNOIS uses grid control loop unit to realize noise reduction neural network, and GRU is a variant of LSTM. By introducing grid control mechanism, GRU network can store information for a long time. RnNoise uses beautiful sound data (English conversation recordings) and noise data (computer fan sounds, office noises, street people noises, etc.) to train the model. Therefore, a wider range of signal-to-noise ratio is obtained, and the noise reduction effect becomes more remarkable. In addition, RnNoise is made in C language. In the speech noise reduction module, the RnNoise wrapper is made by using Python language, and RnNoise is integrated into the system.

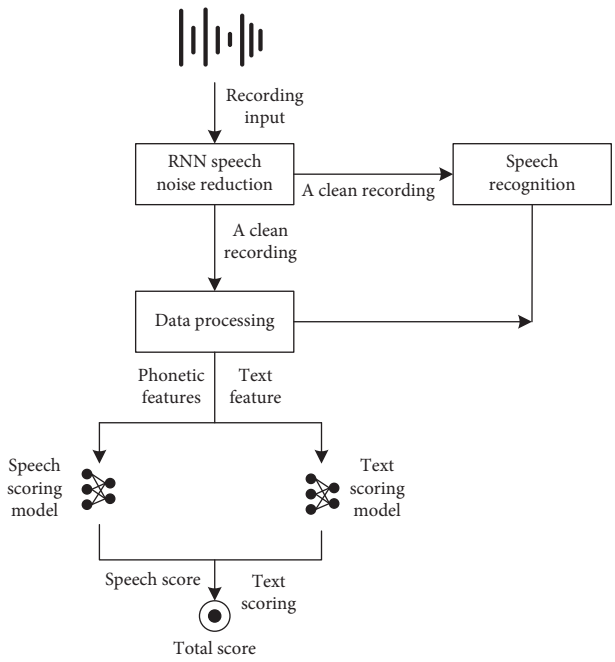

Figure 10: System scoring process.

3.3. Design of Speech Recognition Module. On the basis of evaluating the accuracy of speech recognition engine, a unified standard of Word Error Rate (WER) is reached. WER can be calculated using the following formula:

$$\text{WER} = \frac{\text{Number of words replaced + Number of words deleted + Number of words inserted}}{\text{The total number of words in the correct recognition sequence}}. \tag{16}$$

Microsoft uses a local recognition engine. The recognition speed is the fastest, but the ambiguity is extremely high.

3.4. Design of Data Processing Module

3.4.1. Data Cleaning. Because of the oral fluency of the examinee and the recognition error of the speech recognition engine itself, there are often recognition results that affect the accuracy of the text scoring model in the speech recognition text. For example, this video is about the Chinese and China great wall, um; the great wall is built by the king in dynasty. These features include the number of syntax errors and the depth of syntax tree. In addition, there are also onomatopoeia words like uh and um. In addition to these onomatopoeia words, you can be more specific about the grammar of the text without affecting the entire text content. To build the topic model of LSA, "stop words" such as "the," "is," and "at" must be removed [18]. These stop words have little substantive meaning for the topic model. In addition, the generated model can be more efficient.

3.4.2. Feature Extraction. Feature extraction is an extremely important step before machine learning, which determines the reliability and accuracy of the evaluation model. In this paper, in feature screening, the importance of each feature can be measured by calculating the Pearson correlation coefficient with manual scoring, and the feature with correlation coefficient below 0.2 should not be selected [19]. In this paper, there is generally no fixed reference answer for open oral scoring, so when choosing features, besides the features of similarity in meaning, we mainly choose the features of common type. As shown in Table 1, each feature finally selected and used here will be briefly described.

In this paper, the characteristics of four scales are extracted to evaluate the oral scoring model. The characteristic of speed is often called Rate of Speech (ROS), which is mainly used to explain the fluency of spoken language and calculated by the following formula:

$$\text{ROS} = \frac{N_{\text{words}}}{t - t_s}, \tag{17}$$

where N_{words} represents the total number of words contained in the examinee's spoken language, t represents the total duration of oral recording, and t_s represents the mute duration of recording.

Besides the characteristics of sound speed, the number of quiet sounds during recording can also reflect the fluency of oral English of the tester to a certain extent. In the evaluation of pronunciation quality, the probability characteristic after

Table 1: Summary of features.

Feature category	Feature name	Brief description of characteristics
Phonetic class	articulationRate	Speed of speech
	numSilence	Number of voice pauses
	posteriorScore	Number of pronunciation pauses
	speakingRatio	Posterior probability score of pronunciation
Text class	eassyLength	Total number of words in text
	uniqueWords	Number of nonrepeating words in the text
	parseTreeDepth	Sum of all syntactic tree depths in text
	semanticSimilarity	Semantic similarity between text and theme
	goodGrammerRatio	Correct rate of text grammar

pronunciation is adopted by many oral scoring systems. This paper uses this characteristic to explain the correctness of the examinee's pronunciation. In addition, when extracting effective spoken language, the proportion of long-term recording can also reflect a certain degree of rich spoken content. In the oral evaluation of traditional reading problems, the standard oral sequence corresponding to the benchmark text is usually displayed, the test speech is forced to be configured, and the postprobability average of each phoneme is calculated by the classical GOP algorithm. However, there is no reference text in the open oral score, so it is necessary to combine the speech recognition engine with the speech model of standard English pronunciation training and calculate the average postprobability as the feature of pronunciation quality.

Chapter structure and other features are not suitable for text scoring model of text design. For such short text, sentence structure is a very good alternative, and the depth of grammar tree is used to describe the structural features of sentences. Candidates who are not used to dialogue will have a lower depth of grammar tree than usual. There are algorithms to calculate the similarity of the meanings of commonly used articles. Vector Space Model (VSM) [20], Latent Meaning Analysis (LSA), and Latent Directory Distribution (LDA) are three methods that are based on the word back model, but the degree of meaning varies depending on the method. As a result of the actual test, it is found that the topic model of LSA is more effective in the data set used here. As shown in Figure 11, it is the process of building the topic model of LSA.

Some common part-of-speech tags are shown in Table 2.

There are no grammatical errors in famous English original novels. This paper refers to the method in EASE, an open source composition scoring system. After the part-of-speech tags of Sherlock Holmes' novel collections are displayed, the combination of 3 Yuan tags and 4 Yuan tags is taken out, and the extracted results are saved as a retrieval library of local tag combinations. If you cannot find it, the grammar is wrong. We use the following formula to calculate the correct rate of text syntax:

$$\text{correct}_{\text{ratio}} = \frac{N_g}{N_s}. \tag{18}$$

3.4.3. Data Conversion. Deep learning can automatically extract features, so feature engineering is not needed. As shown in Figure 12, the quantization flow of the entire text removes onomatopoeia words by first performing data cleansing on the speech recognition text and eliminates duplicate words in the text due to recognition errors.

Mel Frequency Cepstrum Coefficients (MFCC) are extracted from spoken recording data as input to the sound scoring model. MFCC contains integrated voice information. Figure 13 is a schematic flowchart showing converting spoken speech recording data into MFCC feature vectors:

3.5. Scoring Model Module Design. Using Keras deep learning framework, all neural networks in this study are constructed. Keras is a highly neural network framework made by *Python* and can run on TensorFlow, CNTK, or Thano.

3.5.1. Scoring Model Based on BP Neural Network. Through repeated experiments, the number of hidden layers and the number of neurons are determined. When the training results do not converge, the number of hidden layers or layer nodes is increased. After the results converge, reduce the number of nodes appropriately and observe whether better results will be obtained. Taking the text scoring model as an example, the sound scoring model with the number of input segments other than 4 has the same structure.

3.5.2. Scoring Model. If the manually extracted features are always invalid, and the correlation between manually extracted features and manually evaluated features is low, it is difficult for the trained model to fit the data accurately. Deep learning technology can automatically mine features, and the learning data can be displayed deeper, which can break through the boundaries of artificial design features. Combine these two networks to construct speech scoring mode and text scoring mode. The computational cost of cyclic neural network is very high when dealing with very long sequence data, so one-dimensional convolution neural network is used as preprocessor step before LSTM network, and shortening sequence can take out higher-level feature display to deal with LSTM layer. As shown in Figure 14, the design of the speech scoring model includes two consecutive convolution blocks. Finally, all connection layers pseudo-combine the one-dimensional vectors to output corresponding speech evaluation results.

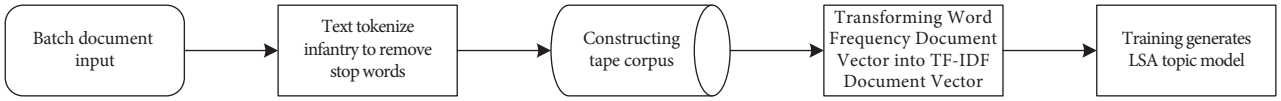

Figure 11: LSA topic model building process.

Table 2: Part-of-speech standard effect.

Part-of-speech tags	Describe
NN	Noun (singular)
NNS	Noun (plural)
VB	Verb (prototype)
VBD	Verb (past tense)
VBNJJ	Verb (past participle)
RB	Adjectives
IN	Adverb
CC	Subordinate conjunctions
PRP	Conjunctions
	Personal pronoun

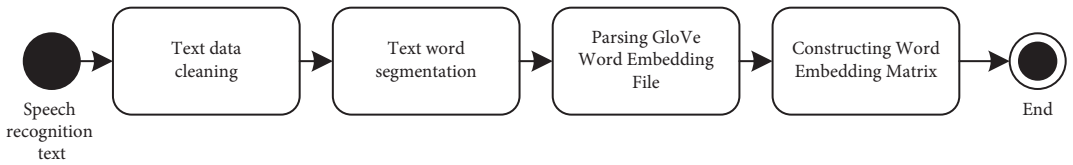

Figure 12: Text vectorization process.

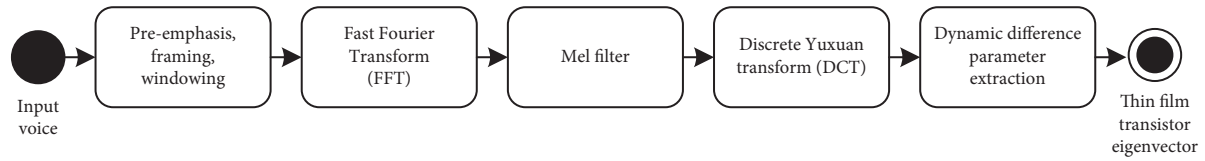

Figure 13: MFCC feature extraction flow.

The design of the text scoring model is shown in Figure 15, and the neural network model shares five layers of networks. The first layer is the word embedding layer, which is defined by GloVe model. The second layer is a one-dimensional flip layer for reducing the length of the network input sequence and extracting more effective features. The third layer network is the LSTM layer, and the LSTM network can select "stored" and "forgotten" information. And it is a one-dimensional vector after pseudooutput MeanOverTime processing and outputs the evaluation result of spoken content.

4. Experimental Results and Analysis

4.1. Means for Evaluating System Performance. In this paper, Pearson correlation coefficient is used to evaluate the performance index of oral evaluation, which is used to evaluate the correlation of different vectors. Its mathematical expression is as follows:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} . \tag{19}$$

The second evaluation index is the difference of man-machine scoring, which is mainly used to describe the difference between manual and machine scoring. Its calculation formula is as follows:

$$d = E\left|S_{\text{Machine}} - S_{\text{Human}}\right| . \tag{20}$$

The third evaluation index is accuracy. This paper establishes the maximum value of man-machine evaluation error to determine whether the evaluation result is correct or not.

4.2. Analysis and Evaluation of Scoring Model

4.2.1. Effectiveness Analysis of Feature Extraction. There are Pearson correlation coefficients for different features in the speaking score, and the results are shown in Table 3 and Table 4. As can be seen from the following two tables, the characteristics of speech types are numbSilence and speakingRatio. This shows that when grading oral English, teachers are most concerned about the fluency of oral English and the long effective time of oral English. In particular, fluency is characterized by recording the more stops, and the lower it is, the lower the score is. This shows that, for oral content, teachers are more interested in candidates' vocabulary grasp and rich conversation content. ParsetreedDepth and goodGrammerRatio features are affected by the recognition accuracy of the speech recognition engine.

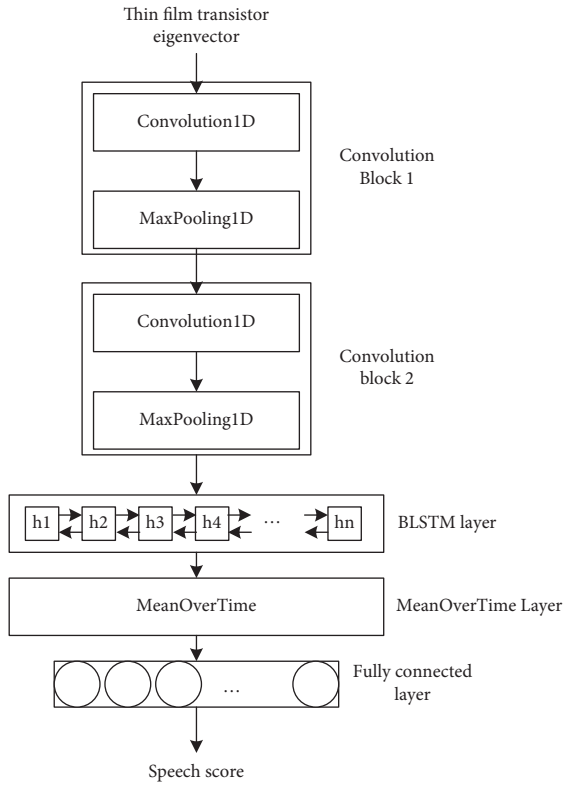

Figure 14: Speech scoring model.

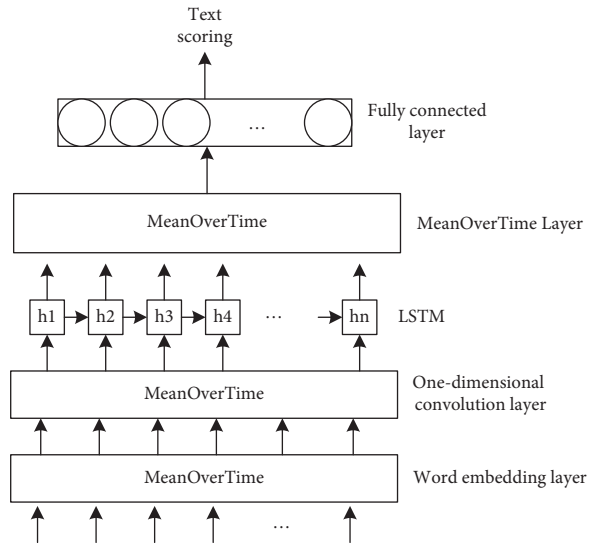

Figure 15: Text scoring model.

On the other hand, teachers also make great efforts to analyze the grammatical errors and sentence structures of candidates' dialogues when scoring manually, and the relationship between these two characteristics and manual scoring is low.

4.2.2. Analysis of Model Scoring Results. We use 150 pieces of test data to test two different neural network scoring models and calculate three evaluation indexes introduced in the first section of this paper to comprehensively evaluate the performance of the two scoring models.

Figures 16 and 17 are the prediction results of oral comprehensive evaluation of BP scoring model and CNN + LSTM scoring model, respectively. It is found from the figure that BP model shows better fitting effect than CNN + LSTM model. In addition, from the lowest students' scores, BP neural network shows better adaptability in the face of extreme values (minimal and maximal). As shown in

Table 3: Correlation between phonetic features and manual scores.

Phonetic features	Pearson correlation coefficient
articulationRate	0.38
numSilence	0.45
posteriorScore	0.32
speakingRatio	0.43

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

Text class feature	Pearson correlation coefficient
contentLength	0.58
uniqueWords	0.60
parseTreeDepth	0.28
semanticSimilarity	0.34
goodGrammerRatio	0.25

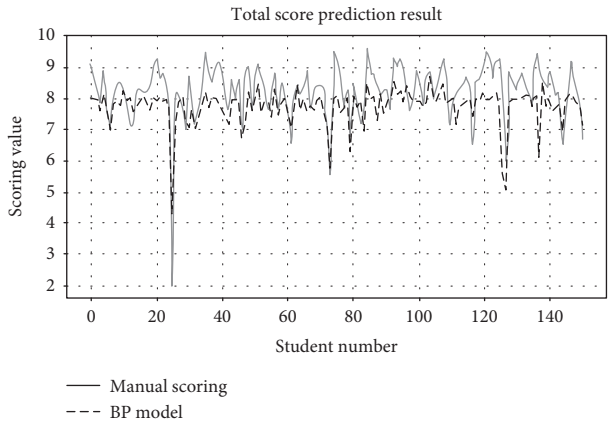

Figure 16: Scoring results of BP model.

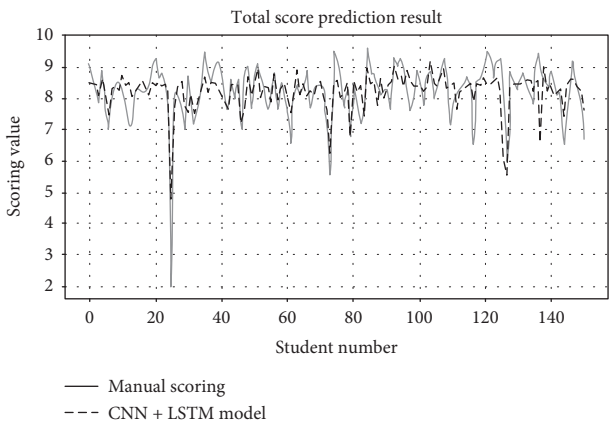

Figure 17: Scoring results.

Table 5: Performance evaluation of scoring model.

	Pearson correlation	Average difference	Accuracy
Between manual scoring	0.764	0.485	—
BP model	0.695	0.645	82.6%
CNN + LSTM model	0.545	0.602	80.1%

Table 5, the performance of Pearson correlation coefficient and accuracy of BP model is better than that of CNN + LSTM model, and the evaluation of the two machines is highly correlated. In the average difference index, CNN + LSTM is slightly better than BP model.

5. Conclusion

Firstly, this paper introduces the overall design and scoring process of the scoring system. After that, the detailed designs of voice noise reduction module, speech recognition module, data processing module, and scoring model module are explained, respectively.

Then, we analyze the experimental results of the oral scoring system and evaluate the performance of the scoring model. This paper introduces three evaluation indexes to evaluate the performance of the model. There are Pearson correlation coefficient, average score difference of man-machine evaluation, correctness of scoring model, and so on. After using these evaluation indexes to analyze the training and evaluation results of the evaluation model, it is found that the comprehensive evaluation performance of BP model is higher than that of CNN + LSTM scoring model when the data set is small. The spoken language scoring model is based on deep learning or other algorithm models, and there are different scoring effects under different algorithms, which lead to different scoring differences. Therefore, the later work to solve this problem needs to combine the advantages of different algorithms for fusion research.

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.

Acknowledgments

This work was sponsored in part by the Anhui University of Science and Technology School of Foreign Languages Curriculum Ideological and Political Project and Anhui University of Science and Technology University Student Innovation and Entrepreneurship Project.

References

[1] A. Bartolomeo, S. Shukla, H. A. Westra, N. S. Ghashghaei, and A. Olson, "Rolling with resistance: a client language analysis of deliberate practice in continuing education for psychotherapists," *Counselling and Psychotherapy Research*, vol. 21, no. 2, pp. 1–9, 2020.

[2] J. Buendgens-Kosten, "The monolingual problem of computer-Assisted Language learning," *ReCALL*, vol. 32, no. 3, pp. 307–322, 2020.

[3] T. Yamaguchi, W. Endo, and Y. Shinoda, "Interrogation system with automatic recognition and delay correction functions of fiber bragg gratings by pulse modulation with wavelength-swept laser," *IEEE Sensors Journal*, vol. 19, no. 22, Article ID 10519, 2019.

[4] J. Joseph, Z. E. H. Moore, D. Patton, T. O'Connor, and L. E. Nugent, "The impact of implementing speech recognition technology on the accuracy and efficiency (time to complete) clinical documentation by nurses: a systematic review," *Journal of Clinical Nursing*, vol. 29, no. 13-14, pp. 2125–2137, 2020.

[5] M. D. Kiselev and O. E. Pudovikov, "Optimization of parameters of automatic speed control system of a freight train with distributed traction," *Russian Electrical Engineering*, vol. 91, no. 9, pp. 568–576, 2020.

[6] T. Lin and X. Liu, "An intelligent recognition system for insulator string defects based on dimension correction and optimized faster R-CNN," *Electrical Engineering*, vol. 103, no. 6, pp. 1–9, 2021.

[7] A. Loukina and H. Buzick, "Use of automated scoring in spoken language assessments for test takers with speech impairments," *ETS Research Report Series*, vol. 2017, no. 1, pp. 1–10, 2017.

[8] Y. Wang, M. J. F. Gales, K. M. Knill et al., "Towards automatic assessment of spontaneous spoken English," *Speech Communication*, vol. 104, pp. 47–56, 2018.

[9] J. Zhang, D. Fang, W. Zhao et al., "An Improved Biomedical Event Trigger Identification Framework via Modeling Document with Hierarchical Attention," in *Proceedings of the 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 583–589, IEEE, San Diego, CA, USA, November 2019.

[10] H. Mohammadzadeh Jamalian, M. Tamjidi Eskandar, A. Chamanara, R. Karimzadeh, and R. Yousefian, "An artificial neural network model for multi-pass tool pin varying FSW of AA5086-H34 plates reinforced with Al 2 O 3 nanoparticles and optimization for tool design insight," *CIRP Journal of Manufacturing Science and Technology*, vol. 35, pp. 69–79, 2021.

[11] M. Shirmohammadi, S. J. Goushchi, and P. M. Keshtiban, "Optimization of 3D printing process parameters to minimize surface roughness with hybrid artificial neural network model and particle swarm algorithm," *Progress in Additive Manufacturing*, vol. 6, pp. 1–17, 2021.

[12] S. Bairoliya, A. Goel, W. Zhang, and B. Cao, "Laboratory preparation of monochloramine for environmental research: a comparison of four commonly used protocols," *Environmental Research*, vol. 197, no. 7, Article ID 111009, 2021.

[13] B. Shao, C. Ni, J. Wang, and Y. Wang, "Research on venture capital based on information entropy, BP neural network and

CVaR model of digital currency in Yangtze River Delta," *Procedia Computer Science*, vol. 187, pp. 278–283, 2021.

[14] N. Yang, H. Tang, J. Yue, X. Yang, and Z. Xu, "Accelerating the training process of convolutional neural networks for image classification by dropping training samples out," *IEEE Access*, vol. 8, Article ID 142393, 2020.

[15] G. Li, H. Tang, Y. Sun, J. Kong et al., Hand gesture recognition based on convolution neural network," *Cluster Computing*, vol. 22, no. 2, pp. 2719–2729, 2019.

[16] J. Feng, S. Cai, and X. Ma, "Enhanced sentiment labeling and implicit aspect identification by integration of deep convolution neural network and sequential algorithm," *Cluster Computing*, vol. 22, no. 6, pp. 1–19, 2019.

[17] H. S. Das and P. Roy, "A deep dive into deep learning techniques for solving spoken language identification problems," *Intelligent Speech Signal Processing*, vol. 2019, pp. 81–100, 2019.

[18] G. Jorge-Botana, R. Olmos, and J. M. Luzón, "Bridging the theoretical gap between semantic representation models without the pressure of a ranking: some lessons learnt from LSA," *Cognitive Processing*, vol. 21, no. 1, pp. 1–21, 2020.

[19] D. Wu, R. Yang, and C. Shen, "Sentiment word co-occurrence and knowledge pair feature extraction based LDA short text clustering algorithm," *Journal of Intelligent Information Systems*, vol. 56, no. 20, pp. 1–23, 2020.

[20] C. Ke, Z. Jiang, H. Zhang, Y. Wang, and S. Zhu, "An intelligent design for remanufacturing method based on vector space model and case-based reasoning," *Journal of Cleaner Production*, vol. 277, Article ID 123269, 2020.