Hindawi
Security and Communication Networks
Volume 2023, Article ID 9821752, 1 page
https://doi.org/10.1155/2023/9821752

*Retraction*

# Retracted: A Deep Learning-Based Assisted Teaching System for Oral English

**Security and Communication Networks**

Received 17 October 2023; Accepted 17 October 2023; Published 18 October 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This article has been retracted by Hindawi following an investigation undertaken by the publisher [1].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] H. Li and X. Liu, "A Deep Learning-Based Assisted Teaching System for Oral English," *Security and Communication Networks*, vol. 2022, Article ID 1882662, 10 pages, 2022.

Hindawi
Security and Communication Networks
Volume 2022, Article ID 1882662, 10 pages
https://doi.org/10.1155/2022/1882662



*Research Article*

# A Deep Learning-Based Assisted Teaching System for Oral English

**Haibo Li[1] and Xue Liu[2]**

[1]*Chuxiong Medical College, Chuxiong, Yunnan 675005, China*
[2]*Foreign Languages and Tourism Department, Hebei Petroleum University of Technology, Chengde 067000, China*

Correspondence should be addressed to Xue Liu; cdpc_lx1@cdpc.edu.cn

Received 5 July 2022; Revised 2 August 2022; Accepted 16 August 2022; Published 9 September 2022

Academic Editor: Tao Cui



The progress of global economic integration has forced English learners to have an urgent need to improve their oral English. College students' oral English ability is currently the worst of the four abilities of listening, speaking, reading, and writing. The main reasons are internal and external. The internal reason is that the pronunciation characteristics of Chinese students are different from those of English. The external cause is that the practice environment and tools of oral English are not ideal, which affects the improvement of learners' oral English. This study proposes using a deep learning algorithm (DLA) English in the evaluation of oral English quality to improve learners' oral English level. The quality of oral English can be comprehensively evaluated in terms of pitch, speed of sound, and rhythm. The standard of pronunciation is the foundation of oral English and is the most critical factor. In many DLAs, the input unit of DNN at a certain moment and its upper and lower moment input units have no relationship and are independent of each other, and the timing dependencies of adjacent units are not fully considered. The results are generally not very good on speech recognition tasks. This study proposes a time-delay neural network (TDNN) and a long short-term memory (LSTM) network to calculate the posterior probability of the model state to model context-dependent features in order to solve this problem. The fusion model TDNN-LSTM is applied in the English spoken pronunciation recognition task. To compare the accuracy of oral English pronunciation, several classic DLAs are introduced. The experimental results show that the method described in this study has a number of advantages. Although the performance improvement of this method in terms of recognition accuracy is not large, a certain degree of improvement is also very important for the oral English teaching assistant system.



## 1. Introduction

With the expansion of the global economy, economic and cultural exchanges between countries are becoming more common. At present, the international communication is still in English. Therefore, learning English is very important. Among the abilities of listening, speaking, reading, and writing in English, reading and writing have always been the strengths of Chinese students, but listening and speaking are relatively weak. Among them, speaking ability is the most difficult for the majority of students to overcome. The cultivation of oral expression ability plays a vital role in English teaching. If you learn any language, if you cannot speak it, you will not be able to achieve the expected communication purpose. However, the level of oral English of college students is generally low, and even many students cannot express in English, and they have a sense of fear of oral English. There are three main reasons why students' oral English is generally poor. First, there is no environment for practicing oral English. Many schools do not have professional and large-scale oral English communication institutions, and students have no place and object for practice. Second, Chinese people are generally introverted, and most students think that practicing oral English in public places will be unusual or feel ashamed to say it wrong. The third is the environment where English is not used. In China, English is not required in other places, except for some foreign companies that use English. Fourth, China has a large population studying English, and the number of professional oral English teaching staff is seriously insufficient. The situation of students' oral English learning cannot be quantitatively evaluated. The above is the fundamental reason

why the teaching effect of oral English has not been improved. In such a big environment, teachers need to improve students' oral English, which has become the focus and difficulty in English teaching. For the problems related to the general environment in oral English teaching, ordinary teachers have no ability to change, but for the fourth problem, teachers can use some auxiliary tools for oral English teaching to improve the teaching effect [1–3]. English speaking aids usually have the following functions. The first function is to score the oral audio uploaded by the learners. Students can clearly understand their speaking level according to the score. The second function is to correct mistakes in pronunciation. The system is able to point out the speaker's mistakes in the audio and tell you the correct pronunciation of a wrong word. The third function is to generate oral language learning reports for learners to understand and analyze their oral language learning status as a whole.

The birth of the English oral assistant system has improved the oral level of oral learners to a certain extent. Many scholars have also dedicated themselves to the design and development of an oral English-assisted teaching system. Reference [4] proposes a robot-assisted learner to practice oral English for Japanese adults. The characteristic of this study is that it is mainly aimed at the adult population. The experimental results show that the robot developed in this study can help adults improve their spoken English. Reference [5] proposes a human-machine dialogue system to increase learners' interest and accuracy in oral English practice. Reference [6] introduces a deep belief network to recognize the pronunciation of spoken language and judge whether it is correct or not. Reference [7] analyzes the difficulties existing in oral English teaching through data analysis tools and gives corresponding solutions. Reference [8] proposes an intelligent technology for students with visual impairment to guide their English learning. Reference [9] used a supervised machine learning method to evaluate the quality of students' oral English, including pronunciation accuracy and fluency. Reference [10] uses a hierarchical classification method to analyze the pronunciation categories of oral English, so as to judge whether the speaker's spoken language is standard. Reference [11] applies machine learning to the assessment of oral English proficiency levels. Experiments show that tone and intonation are key factors affecting its evaluation. The essence of oral English pronunciation recognition based on machine learning [12–14] is to classify and recognize each word in the input audio. The essence of oral English pronunciation recognition based on deep learning [15–17] is the same. The difference is that the process is different. DLAs incorporate both feature extraction of the data and final classification.

The above studies used data analysis tools, machine learning, and DLAs to identify oral English, including the pronunciation of spoken language and whether the intonation is correct and fluent. By identifying the results and guiding oral learners on how to pronounce correctly, the quality of oral English teaching can be improved. This idea of improving oral English teaching is correct. The premise is that it is necessary to develop a method with high recognition accuracy and fast recognition speed. The problem with machine learning-based pronunciation recognition methods is that the final recognition results are generally not ideal, and the recognition results are easily affected by noise, feature extraction methods, and final classifiers. The problem with deep learning-based pronunciation recognition is that the model training event is too long, making it difficult to meet the needs of real-time recognition. Moreover, most deep models need to adjust multiple parameters, so the model is easily affected by parameters, resulting in unstable final results. Aiming at these problems, this study proposes an English speaking assisted teaching system based on DLA. The classic DLAs include convolutional neural network (CNN) [18], deep neural network (DNN) [19], and LSTM [20]. In this study, to improve the pronunciation accuracy of the model for oral English, a hybrid deep neural network is used. There is no relationship between the input unit of DNN at a certain time and its upper and lower time input units, and they are independent of each other. The timing dependencies of adjacent units are not fully considered, so the results of the fully connected feedforward DNN on speech recognition tasks are generally average. Neither was good. To solve this problem, this study first introduces TDNN and LSTM to calculate the posterior probability of the model state to model the context-dependent features. It will be applied to the English spoken pronunciation recognition task.

## 2. Relevant Knowledge

*2.1. The Teaching System Assists the Learning Process of Oral English.* Using machines to help people improve their oral English is the purpose of the auxiliary teaching system. The teaching system assists the oral English learning process as shown in Figure 1. A complete auxiliary teaching system mainly includes language input, language practice, and result output.

In the language input stage, both standard and test data are entered. Correct pronunciation, standard speaking rate, and rhythm are examples of standard data. Different standard data are put for different learners and situations. The test data are the user's oral English audio data. During the language practice phase, the assistance system acts as a learning buddy, conversing in English with the learner. The auxiliary system can be configured to reflect the learner's preferred English style and play a role in the task situation. This method can pique the learner's interest in communicating in English. The system can perform a variety of functions during the language practice phase. According to the user's identity, practice topics of different topics are imported. Exchanges on topics are practiced. Feedback adjustments are made based on the results of the communication. At this stage, the user can not only actively communicate with the system but also the system feedback and communicate the results. The system can also actively push the content that users need. The active push function is mainly determined according to the personal identity and hobbies and other tags set by the user. In the language output stage, the system provides evaluation feedback tools for

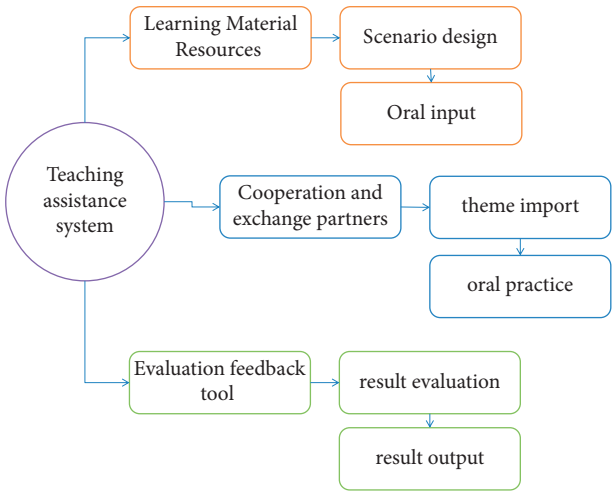


Figure 1: Teaching system to assist English learning process.

result evaluation. According to the evaluation results, users can strengthen the weak items. After a period of study, the learner can communicate with the auxiliary system again to assess whether their oral proficiency has reached the goal. If you do not reach your goal, keep learning. If the goal has been achieved, the assistance system also completes the task satisfactorily.

*2.2. Auxiliary Oral English Teaching Design.* Figure 2 depicts the design process for auxiliary oral English teaching. The design of auxiliary speaking instruction must consider three factors: teaching task analysis, teaching process, and teaching evaluation and reflection. Teaching task analysis mainly includes teaching objectives and teaching task analysis. Before the specific implementation of the teaching process, it is also necessary to understand the learner's situation, so there should also be participant analysis. The teaching process consists primarily of designing learning situations, resource tools, and learning strategies. After the learning process, it needs to be strengthened and consolidated to form a result evaluation. The evaluation link here also needs to use the assistance of the teaching system. The teaching process can be optimized through the assistance system. Finally, a summative evaluation is formed, and teaching reflection is carried out.

*2.3. Oral English Assessment Based on Speech Recognition.* Speech recognition is a technique for converting speech signals into words [21]. For a long time, the basic speech recognition framework depicted in Figure 3 has been the most traditional framework in the field of speech recognition, that is, a speech recognition system composed of acoustic models, pronunciation dictionaries, and language models. Audio input, feature extraction, an acoustic model, a pronunciation dictionary, and a language model are the main components of the traditional speech recognition framework. Mel-scale frequency cepstral coefficients (MFCCs) [22], perceptual linear predictive (PLP) [23], and other feature extraction methods for audio data are commonly used. After feature extraction, the audio data are fed into the trained admission model for scoring. The scoring result, pronunciation dictionary, and language model together form a decoding network and finally output the speech recognition result.

## 3. A Method for Identifying the Quality of Oral English

*3.1. Time-Delay Neural Network.* Many DLAs are appropriate for processing continuous language data. This study employs the time-delay neural network (TDNN) [24]. When TDNN is used instead of traditional feedforward neural networks, the input of each hidden layer is increased. The input of each hidden layer not only keeps up with the output of the previous layer at the current moment but also the output of several moments before and after is combined into the current input. TDNN is a context-sensitive model that is designed to retrieve more historical information from the previous layer at the same time. DNN cannot model longer context temporal information and TDNN. The network is multilayered, with each layer having the ability to abstract features. It has the ability to express the temporal relationship between speech features. The learning process does not necessitate exact temporal placement of the learned labels. The amount of computation is reduced by sharing weights. Figure 4 depicts the TDNN structure.

The node corresponding to a specific moment in the hidden layer, as well as all the nodes corresponding to the time span before and after it, forms the basic unit of TDNN, which is known as time-delay neural network (TDN). Assume that TDN has a time span of $T$ and that the node has $N$ inputs $(x_1(t), x_2(t), \ldots, x_N(t))$ at time $t$. The first $T$ moments of each input $x_i(t)$ are input $x_i(t-n)$, $n = 1, 2, \ldots, T$. The weight is $(w_{i1}, w_{i2}, \ldots, w_{iT})$, and the calculation formula of the neuron output value $h(t)$ is as follows:

$$h(t) = f\left(\sum_{i=1}^{N}\left[\sum_{n=1}^{T} w_{in} \cdot x_i(t-n) + b_i\right]\right), \quad (1)$$

where $f(\cdot)$ is the activation function and $b_i$ is the bias coefficient.

*3.2. Long Short-Term Memory Networks.* On time-dependent problems, recurrent neural network (RNN) [25] performs well. RNN remembers the current output data and adds it to the next input data. It is to add self-loop feedback information to each neuron in the hidden layer in the time domain; that is, the hidden layer's input contains information from the input layer and information from the hidden layer at the previous moment. As a result, RNN can highlight the strong modeling ability of time series-related information tasks more effectively. Figure 5 depicts the RNN structure.

Timing is represented by $t$-1, $t$, and $t + 1$ in the diagram. $x$ represents the input data. The memory at time $t$ is represented by $s_t$. $W$ represents the input's weight. $U$ represents the current weight of the input parameter. $V$ denotes the

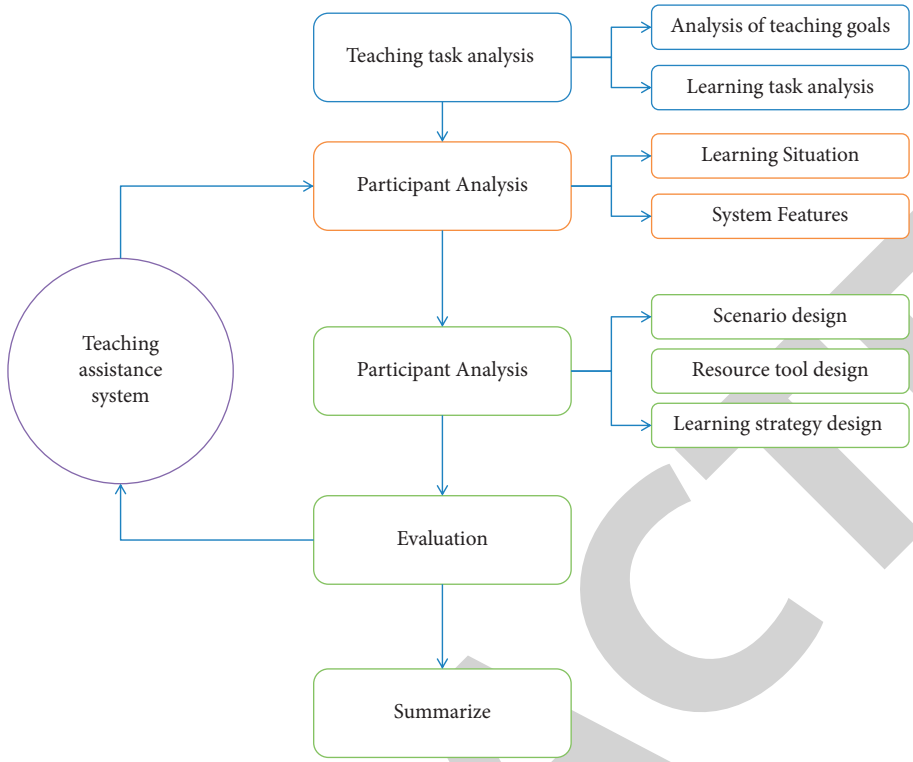


Figure 2: Design process for auxiliary oral English instruction.

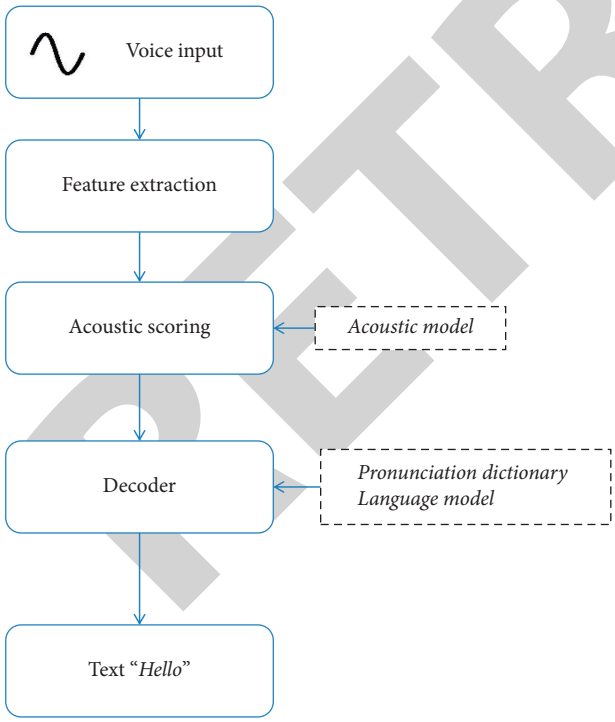


Figure 3: Classic framework of speech recognition.

output parameter's weight. At time $t = 1$, input $s_0$ is set to 0. $W$, $U$, and $V$ values are initialized at random. Equation (2) calculates the $h_1$ value, $s_1$ value, and $o_1$ value.

$$h_1 = Ux_1 + Ws_0, \tag{2}$$

$$s_1 = f(h_1), \tag{3}$$

$$o_1 = g(Vs_1). \tag{4}$$

$f(\cdot)$ and $g(\cdot)$ are both activation functions. $f(\cdot)$ can be an activation function such as tanh, ReLU, and sigmoid, while $g(\cdot)$ is typically a softmax function. The state $s_1$ at this time is used as the memory state of the current moment, according to the sequence. It will take part in the next moment's activities.

$$h_2 = Ux_2 + Ws_1, \tag{5}$$

$$s_2 = f(h_2), \tag{6}$$

$$o_2 = g(Vs_2). \tag{7}$$

By analogy, the final output is as follows:

$$h_t = Ux_t + Ws_{t-1}, \tag{8}$$

$$s_t = f(h_t), \tag{9}$$

$$o_t = g(Vs_t). \tag{10}$$

RNN has a good effect on time-series problems. However, when the parameters are unchanged during the training process, the gradient will be continuously multiplied during the backpropagation process, and the value will

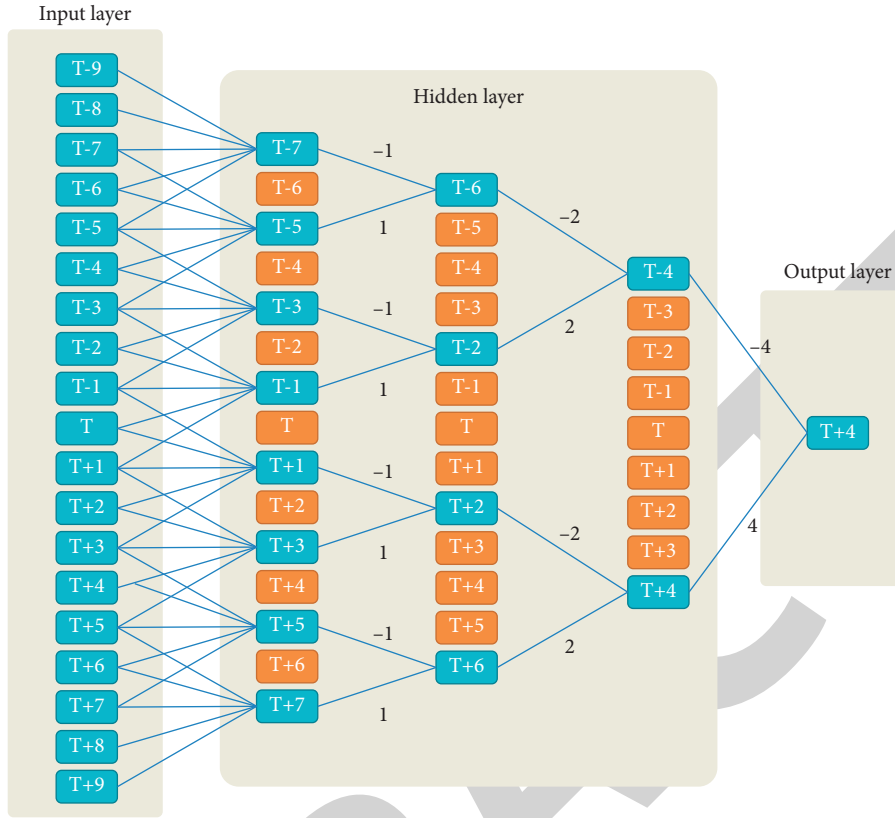


Figure 4: TDNN structure.

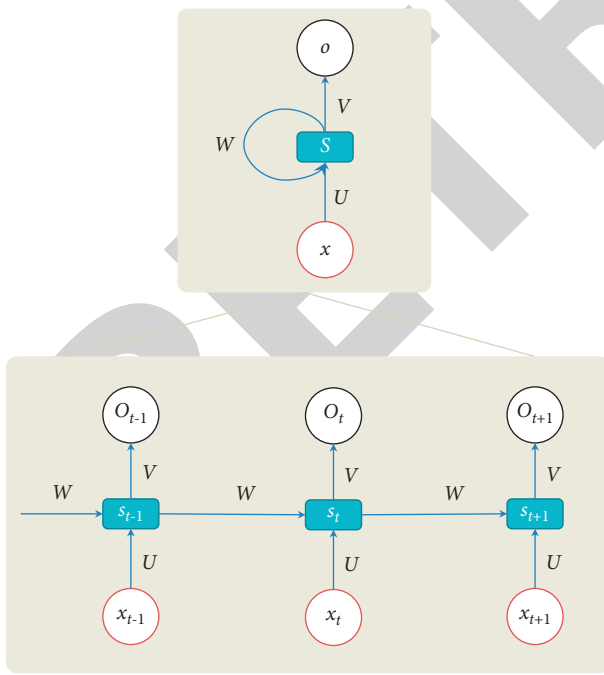


Figure 5: RNN structure.

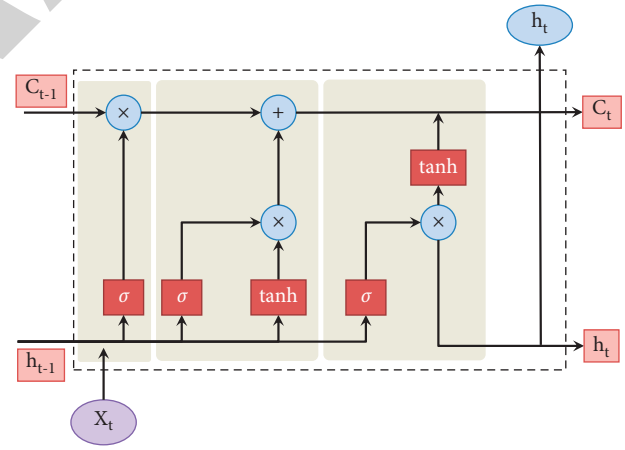


Figure 6: LSTM structure.

become larger or smaller. This can lead to problems such as exploding gradients or vanishing gradients. LSTM is a special kind of RNN that can solve the time dependence problem very well. Therefore, LSTM is introduced to solve the above problems. Compared with RNN, LSTM is calculated based on the input and the output of the hidden layer at the previous moment, but it changes the internal structure of the RNN hidden layer. The neurons of LSTM include input gate $i$, forget gate $f$, output gate $o$, and internal memory unit $C$. Figure 6 shows the LSTM structure.

The forget gate is controlled by sigmoid. A $f_t$ value from 0 to 1 is generated based on the output $h_{t-1}$ at the previous moment and the current input $x_t$. This value is used to decide whether to forget or forget part of the information $C_{t-1}$ learned at the last moment., where $w$ represents the weight matrix, $b$ represents the bias vector, and $\sigma$ represents the

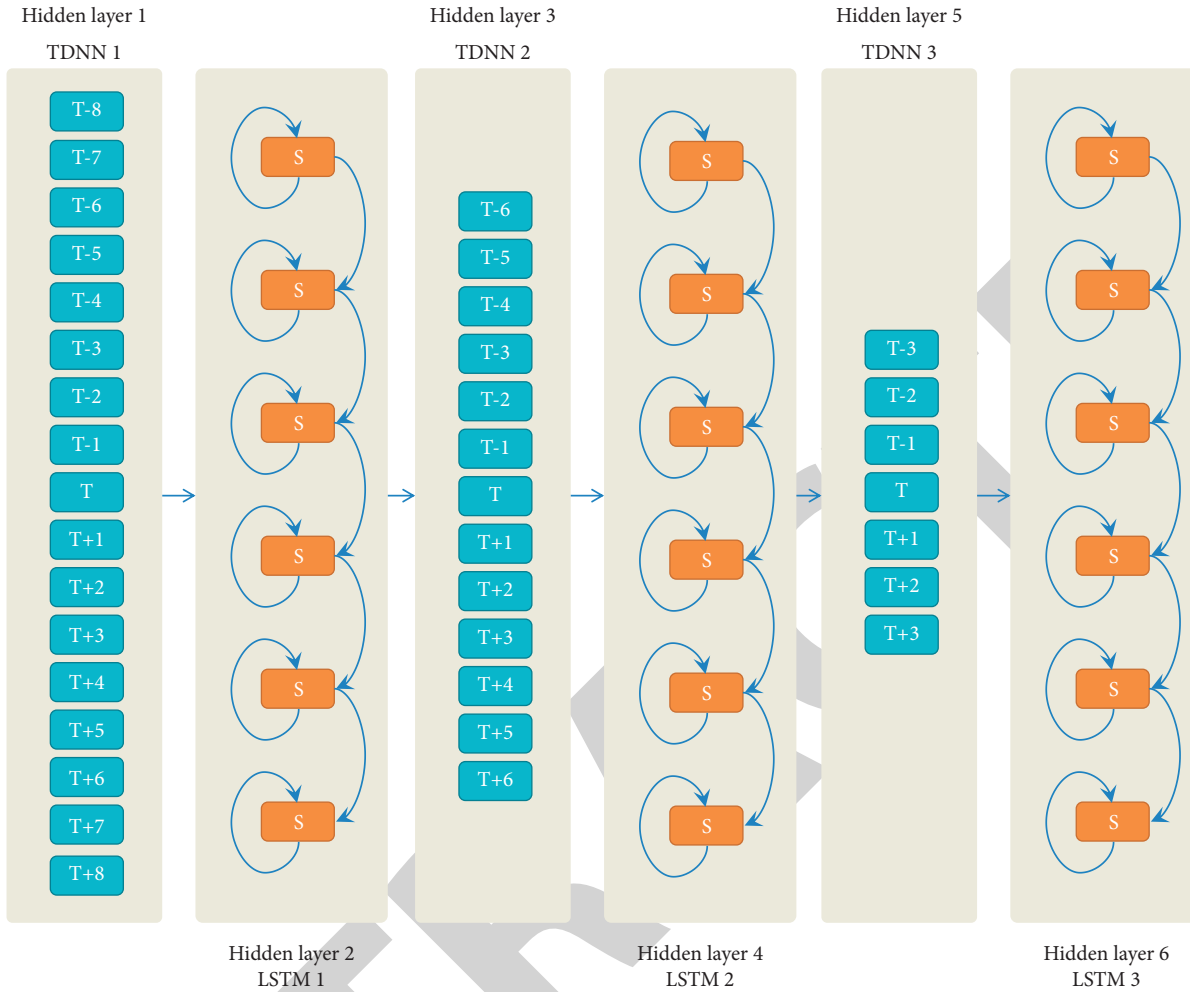


Figure 7: TDNN-LSTM structure.

nonlinear activation function. The formula for calculating the $f_t$ value is shown as follows:

$$f_t = \sigma\left(w_f \cdot [h_{t-1}, x_t] + b_f\right). \tag{11}$$

The input gate uses sigmoid to decide which information needs to be updated. The tanh layer is to generate a new candidate value $C_t$, which may be added to the internal memory unit as a candidate value generated by the current layer. Combining the values generated by the above two parts, the model is updated as follows. First, the product information of the internal memory unit of the previous layer and ft is used to forget the unnecessary information and then added with $i_i \times \widetilde{C}_t$ to obtain the candidate value $C_t$. The calculation formula is as follows:

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right), \tag{12}$$

$$\widetilde{C}_t = \tanh\left(W_C \cdot [\mu_{t-1}, x_t] + b_C\right), \tag{13}$$

$$C_t = f_t \times C_{t-1} + i_t \times \widetilde{C}_t. \tag{14}$$

The model's output is obtained by multiplying an initial output through the sigmoid layer and scaling the $C_t$ value to a value between -1 and 1.

$$o_t = \sigma\left(W_o \cdot [h_{t-1}, x_t] + b_o\right), \tag{15}$$

$$h_t = o_t \times \tanh\left(C_t\right). \tag{16}$$

*3.3. TDNN-LSTM Model.* For tasks with strong timing information correlation, both TDNN and LSTM have superior modeling capabilities. LSTM training is more difficult than TDNN training. As a result, the TDNN-LSTM fusion model of TDNN and LSTM is used in oral English speech recognition. The model captures enough contextual information while reducing computational complexity. The TDNN-LSTM structure is depicted in Figure 7. As shown in the diagram, the network has six hidden layers. The single layer is known as TDNN, and the double layer is known as LSTM. The two models are arranged alternately. A unit module consists of a TDNN and an LSTM.

## 4. Experimental Analysis

*4.1. Evaluation Indicators of Oral English Pronunciation.* Many factors influence the quality of oral English output, including intonation, pitch, rhythm, and duration. In most

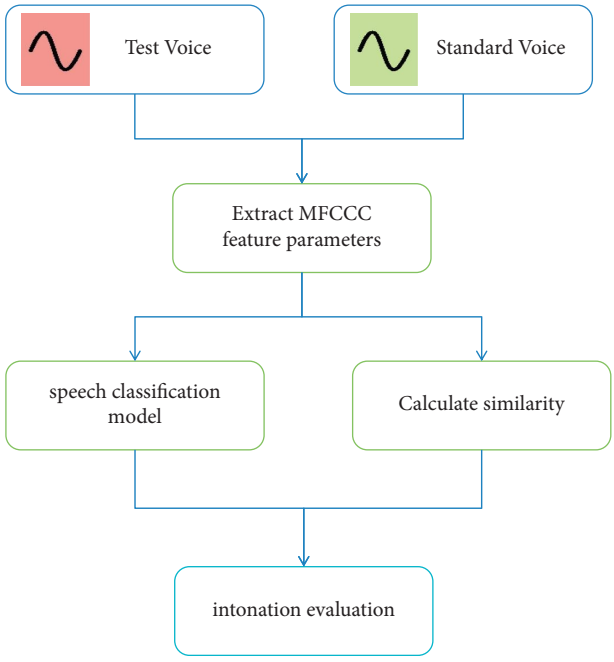


Figure 8: Principle of pitch evaluation.

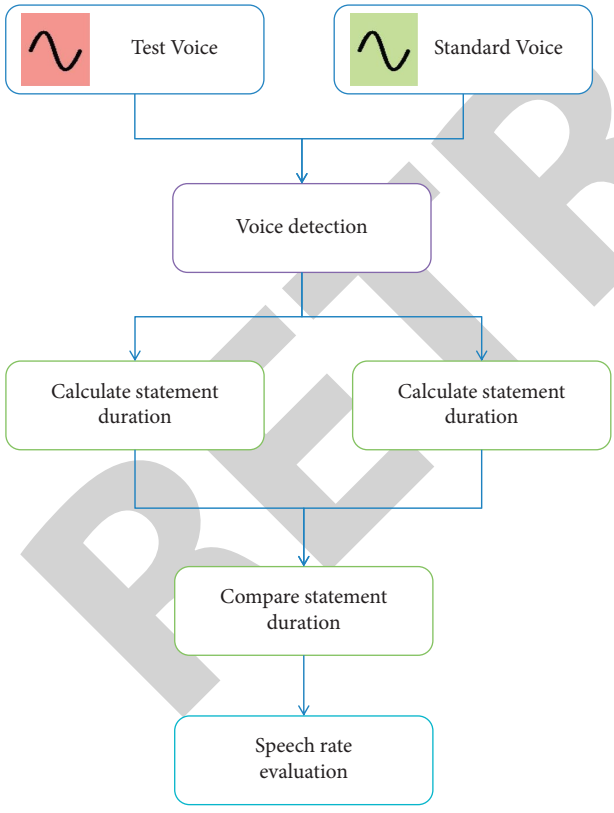


Figure 9: Principle of sound velocity evaluation.

cases, intonation is used as the primary indicator of the quality of oral English output, with other factors serving as auxiliary indicators. Pitch is primarily used to determine whether each word in the output sentence is correct and understandable. Figure 8 depicts the pitch evaluation principle.

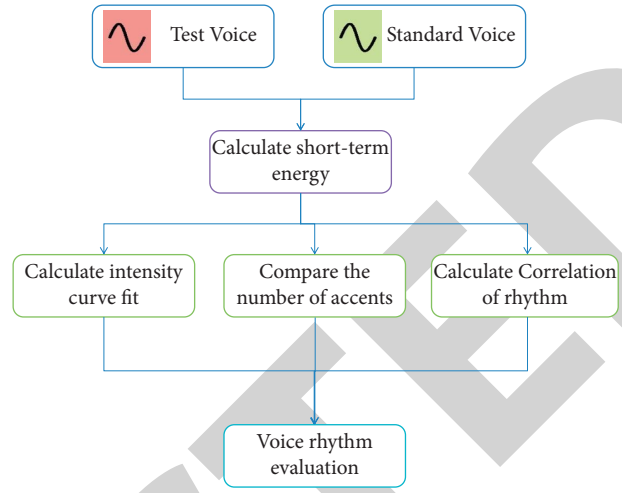


Figure 10: Principle of rhythm evaluation.

Speaking rate is also a key indicator of oral evaluation. The number of words produced by the speaker per unit time is referred to as the speech rate. If a student speaks 120 words in one minute, his speech rate is 2 words per second. Regardless of whether such a speech rate is fast or slow, it must be compared to the standard speech rate. The standard speaking rate for the same 120 words is 80 seconds, but the student only took 60 seconds, indicating that the student's speaking rate is too fast. Figure 9 depicts the sound speed evaluation principle.

Speaking rhythm is also very important. The difference between heavy, light, long, and short sounds when speaking a sentence is referred to as rhythm. The content of spoken language output differs, as does its rhythm. Many stressed syllables characterize the rhythm of oral English. Unstressed syllables between stressed syllables sound a little hazy. Figure 10 depicts the oral rhythm evaluation principle.

Many of the above aspects should be able to be evaluated by a good speaking assistance system. Furthermore, the weight of each factor is not the same. This study first validates the method's performance using the most fundamental and critical pitch indicator. The main purpose of pitch is to calculate the number of words recognized by the speech recognition module. The more correctly recognized words there are, the better the intonation.

*4.2. Experimental Data.* Thirty college students, 18 boys and 12 girls, were chosen for this study. In a quiet and comfortable classroom, students read prepared English sentences. Some examples of English sentences are shown in Table 1, with a total of 50 English sentences. At the same time, Sonar recording software is used. The sampling frequency is set to 16 KHz, and the encoding is 16 bit encoding. 30 students read out 50 sentences and recorded 1500 pieces of audio data. 1050 sentences of 1500 sentences are selected as training data and 450 sentences as test data.

*4.3. Accuracy Recognition of Oral English.* The speech signal must be preprocessed before it can be recognized as oral English. This preprocessing includes frame division

Table 1: Examples of English sentences.

| |
|---|
| (1) What's happening? |
| (2) Nice to see you again |
| (3) I would like to talk to you for a minute |
| (4) The meeting was scheduled for two hours, but it is now over yet |
| (5) It took years of hard work to speak good English |
| (6) Mama used to say that life was like a box of chocolates. You never know what you'll get |
| (7) Do you think you'd be surprised if I changed into something more comfortable? |
| (8) Everything you see exists together in a delicate balance |
| (9) You know some birds are not meant to be caged; their feathers are just too bright |
| (10) Hope is a good thing, perhaps the best thing. And nothing good ever dies |

Table 2: Recognition results of oral English by each model.

| Index | Reference [26] | Reference [27] | Reference [28] | Reference [29] | Proposed |
|---|---|---|---|---|---|
| Accuracy | 90.84 | 93.35 | 92.18 | 94.24 | 95.87 |
| Precision | 89.17 | 92.81 | 90.23 | 93.92 | 94.75 |
| Recall | 87.35 | 91.92 | 90.86 | 93.26 | 94.92 |

windowing, fast Fourier transform (FFT), Mel cepstral coefficient feature extraction, and other processes. The frame-by-frame windowing technique is used to compress the speech signal. When the length of the speech is between 10 and 30 milliseconds, it is considered a quasi-stationary signal. To divide the speech length into short segments, a window function must be added to the speech signal. Table 2 displays the results of each model's oral English recognition.

The comparison algorithms used are all DLAs, and the recognition rates of oral English pronunciation are all above 90%. This fully demonstrates the superior performance of DLAs. Compared with references [26–29], the proposed algorithm improves the accuracy by 5.5%, 2.7%, 4%, and 1.7%, respectively. The accuracy is improved by 6.3%, 2.1%, 5%, and 0.9%, respectively. The recall rates are improved by 8.7%, 3.3%, 4.5%, and 1.8%, respectively. The core model used in [26] is DNN. The core model used in [27] is LSTM. The core model used in [28] is BiLSTM. The core model used in [29] is CNN-BiRNN. The experimental results obtained by CNN-BiRNN in several models are relatively good. This method has a certain improvement on the basis of reference [29], although the improvement is not very obvious. The method in this study is based on the fusion of TDNN and LSTM model. In the classification of speech sequence data, LSTM can play a very good role. For oral English, since only the hidden state of the last time step is used in the LSTM classification task, it is not enough to fully express the spoken information. The complete accent information not only needs to obtain the pronunciation information of the past time but also needs to obtain the pronunciation information of the future time. There is a lot of prosody-related information in the accent. The prosody information is mostly information that echoes before and after; that is to say, it is necessary to determine the rhythm of the past moment. The rhythm of the future moment is also particularly important. TDNN-LSTM can simultaneously extract the prosody information of the past moment and the prosody information of the future moment to jointly determine the prosody information of the current moment.

## 5. Conclusion

The improvement of oral English learning demand has promoted the birth of various English-assisted teaching systems. The more classic software packages mainly include "talk to me" and "PhonePass Set". These auxiliary software packages for improving oral English are not completely suitable for Chinese students to learn oral English. To study an oral English-assisted teaching system suitable for Chinese students, this study introduces a DLA and applies it to the quality assessment and error correction of oral English. The improvement of oral English is mainly carried out from two levels. One is to identify the overall quality of the learner's pronunciation and give a certain score. If the full score is 100, 90 to 100 is excellent, 80 to 90 is good, 70 to 80 is fair, 60 to 70 is pass, and below 60 is failed. In this way, learners have an overall understanding of their oral English. Only when learners clearly recognize their current level they can formulate goals that can be achieved. The second is to find out which words have problems with pronunciation and tell them the correct pronunciation. Words with completely wrong pronunciation are marked in red, inaccurate words are marked in yellow, and completely correct words are marked in green. To achieve the above two points, it is necessary to correctly recognize the input speech. The speech recognition model used in this study is a fusion model of TDNN and LSTM. Firstly, a time-delay neural network and a long short-term memory network are introduced successively to calculate the posterior probability of the model state, so as to model the context-dependent features. Finally, according to the structural characteristics of the long short-term memory network, the TDNN-LSTM hybrid network structure is introduced, and it will be applied to the English spoken pronunciation recognition task. The simulation results show that the method in this study has advantages compared with other deep learning methods and has certain reference value. There are still some shortcomings in this study. For example, the evaluation indicators used only use the core pitch indicators, and the important indicators such

as sound speed and rhythm have not been experimentally studied. In addition, the training of the used models is relatively complex, and the process needs to be further simplified. This study is mainly devoted to optimizing the English speaking assistant system from the technical level. However, for teachers, it is also necessary to use these auxiliary tools reasonably in the teaching process. The auxiliary system should be able to be applied to each link before, during, and after class. Therefore, enriching the functions of the auxiliary system is also the follow-up research plan of this study.

## Data Availability

The labeled data set used to support the findings of this study is available from the corresponding author upon request.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

This work was supported by the Chuxiong Medical College.

## References

[1] J. J. Martins and PF. do Amaral, "English language teaching in BRAZILIAN public schools nowadays," *REVISTA INCLUSIONES*, vol. 6, pp. 127–142, 2019.

[2] I. Dzulkifli, "Teaching and learning aids to support the Deaf students studying Islamic Education," *PERTANIKA JOURNAL OF SOCIAL SCIENCE AND HUMANITIES*, vol. 29, no. 4, pp. 2263–2279, 2021.

[3] S. Sugiman, H. Suyitno, I. Junaedi, and D. Dwijanto, "The Creation of teaching aids for Disabled students as Mathematical-Thinking-Imaginative product," *International Journal of Instruction*, vol. 13, no. 3, pp. 777–788, 2020.

[4] T. Iio, R. Maeda, K. Ogawa et al., "Improvement of Japanese adults' English speaking skills via experiences speaking to a robot," *Journal of Computer Assisted Learning*, vol. 35, no. 2, pp. 228–245, 2019.

[5] M. Gasic, D. Hakkani-Tur, and A. Celikyilmaz, "Spoken language understanding and interaction: machine learning for human-like conversational systems," *Computer Speech & Language*, vol. 46, pp. 249–251, 2017.

[6] F. M. Jiao, J. Song, X. Zhao, P. Zhao, and R. Wang, "A spoken English teaching system based on speech recognition and machine learning," *INTERNATIONAL JOURNAL OF EMERGING TECHNOLOGIES IN LEARNING*, vol. 16, no. 14, pp. 68–82, 2021.

[7] P. M. I. Seraj, H. Habil, and M. K. Hasan, "Investigating the problems of teaching oral English communication skills in an EFL context at the Tertiary level," *International Journal of Instruction*, vol. 14, no. 2, pp. 501–516, 2021.

[8] RF. Brena, E. Zuvirie, A. Preciado, A. Valdiviezo, M. Gonzalez-Mendoza, and C. Zozaya-Gorostiza, "Automated evaluation of foreign language speaking performance with machine learning," *International Journal on Interactive Design and Manufacturing*, vol. 15, no. 2-3, pp. 317–331, 2021.

[9] S. Weigelt, V. Steurer, T. Hey, and W. F. Tichy, "Towards Programming in Natural language: learning new functions from spoken Utterances," *International Journal of Semantic Computing*, vol. 14, no. 02, pp. 249–272, 2020.

[10] O. Kang and D. Johnson, "The roles of suprasegmental features in predicting English oral proficiency with an automated system," *Language Assessment Quarterly*, vol. 15, no. 2, pp. 150–168, 2018.

[11] TI. Monastyrskaya, TB. Ganicheva, GV. Toropchin, and A. V. Katsura, "Need for a Differentiated approach to teaching English in higher school: a Sociological study," *Modern Journal of Language Teaching Methods*, vol. 8, no. 9, pp. 24–36, 2018.

[12] T. Hagendorff, "Linking human and machine Behavior: a new approach to evaluate training data quality for Beneficial machine learning," *Minds and Machines*, vol. 31, no. 4, pp. 563–593, 2021.

[13] JM. Bone, CM. Childs, A. Menon et al., "Hierarchical machine learning for high-Fidelity 3D Printed Biopolymers," *ACS Biomaterials Science & Engineering*, vol. 6, no. 12, pp. 7021–7031, 2020.

[14] G. Giantamidis, S. Tripakis, and S. Basagiannis, "Learning Moore machines from input-output traces," *International Journal on Software Tools for Technology Transfer*, vol. 23, no. 1, pp. 1–29, 2019.

[15] SA. Seshia, S. Jha, and T. Dreossi, "Semantic Adversarial deep learning," *IEEE DESIGN & TEST*.vol. 37, no. 2, pp. 8–18, 2020.

[16] J. S. Park and J. H. Park, "Enhanced machine learning algorithms: deep learning, reinforcement learning, ana Q-learning," *JOURNAL OF INFORMATION PROCESSING SYSTEMS*, vol. 16, no. 5, pp. 1001–1007, 2020.

[17] Y. Matsuo, Y. LeCun, M. Sahani et al., "Deep learning, reinforcement learning, and world models," *Neural Networks*, vol. 152, pp. 267–275, 2022.

[18] LA. Zavala-Mondragon, B. Lamichhane, L. Zhang, and G. d. Haan, "CNN-SkelPose: a CNN-based skeleton estimation algorithm for clinical applications," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 6, pp. 2369–2380, 2020.

[19] H. Kwon, M. Pellauer, A. Parashar, and T. Krishna, "Flexion: a quantitative Metric for Flexibility in DNN Accelerators," *IEEE Computer Architecture Letters*, vol. 20, no. 1, pp. 1–4, 2021.

[20] S. Dutta, JK. Mandal, TH. Kim, and S. K. Bandyopadhyay, "Breast Cancer prediction using Stacked GRU-LSTM-BRNN," *Applied Computer Systems*, vol. 25, no. 2, pp. 163–171, 2020.

[21] OZ. Mamyrbayev, K. Alimhan, B. Amirgaliyev, B. Zhumazhanov, D. Mussayeva, and F. Gusmanova, "Multimodal systems for speech recognition," *International Journal of Mobile Communications*, vol. 18, no. 3, pp. 314–326, 2020.

[22] K. Rangra and M. Kapoor, "Exploring the mel scale features using supervised learning classifiers for emotion classification," *INTERNATIONAL JOURNAL OF APPLIED PATTERN RECOGNITION*, vol. 6, no. 3, pp. 232–253, 2021.

[23] VV. Yerigeri and L. K. Ragha, "Speech stress recognition using semi-eager learning," *Cognitive Systems Research*, vol. 65, pp. 79–97, 2021.

[24] J. Monteiro, J. Alam, and T. H. Falk, "Multi-level self-attentive TDNN: a general and efficient approach to summarize speech into discriminative utterance-level representations," *Speech Communication*, vol. 140, pp. 42–49, 2022.

[25] J. Kim, Y. Lee, and E. Kim, "Accelerating RNN Transducer Inference via Adaptive expansion Search," *IEEE Signal Processing Letters*, vol. 27, pp. 2019–2023, 2020.

[26] J. Guglani and A. N. Mishra, "DNN based continuous speech recognition system of Punjabi language on Kaldi toolkit," *International Journal of Speech Technology*, vol. 24, no. 1, pp. 41–45, 2020.

[27] B. Fernandes and K. Mannepalli, "Speech emotion recognition using deep learning LSTM for Tamil language," *PERTANIKA JOURNAL OF SCIENCE AND TECHNOLOGY*, vol. 29, no. 3, pp. 1915–1936, 2021.

[28] V. Sornlertlamvanich and S. Yuenyong, "Thai Named Entity recognition using BiLSTM-CNN-CRF Enhanced by TCC," *IEEE Access*, vol. 10, pp. 53043–53052, 2022.

[29] JW. Wan, B. Chen, YQ. Liu, Y. J. Yuan, HW. Liu, and L. Jin, "Recognizing the HRRP by combining CNN and BiRNN with Attention Mechanism," *IEEE Access*, vol. 8, pp. 20828–20837, 2020.