Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 6, pp. 4493–4506, DOI 10.12694/scpe.v25i6.3390

COMPUTER-ASSISTED ONLINE LEARNING OF ENGLISH ORAL PRONUNCIATION BASED ON DAE END-TO-END RECURRENT NEURAL NETWORKS

KANGSHENG LAI[*]AND LIUJUN MO[†]

Abstract. With the development of globalization, learning a second language has received increasing attention from people. To improve English oral proficiency, a computer-aided online learning system for English oral pronunciation is studied. A denoising autoencoder is integrated into the system to create a simplified end-to-end recurrent neural network for pronunciation detection and diagnosis based on deep learning. The study first collected and preprocessed oral pronunciation data of English learners, including enhancing speech signals and reducing noise. Next, an RNN model with Long Short-Term Memory (LSTM) as the core was constructed to capture time series characteristics in pronunciation. And use DAE to extract features and reduce the influence of background noise to enhance the recognition of pronunciation features. At the same time, the study utilized web crawler technology to collect a large amount of oral pronunciation data from non-native English learners, and constructed an English oral corpus containing pronunciation errors. And in order to simulate real situations, white noise and pink noise were artificially added to the corpus in the study, and they were divided into training and testing sets in a ratio of 60% to 40%. The results showed that the classification accuracy of the system in the training and testing sets under white noise environment was 78.97% and 94.01%, respectively, and the classification accuracy in the pink noise environment was 76.19% and 94.03%, respectively. The system's error detection accuracy in vowel and consonant pronunciation detection is 88.91% and 91.68%, respectively, and the error correction accuracy in vowel and consonant pronunciation detection is 90.67% and 91.96%, respectively. In summary, the research on computer-aided online learning of English oral pronunciation based on Denoising Auto Encoders end-to-end recurrent neural networks has effectively improved learning efficiency.

Key words: Denoising autoencoder; Spoken English pronunciation; Recurrent neural network; End-to-end

1. Introduction. Under the background of economic globalization, global cultural integration has become an inevitable development trend in the future [1]. English, as an international official language, is a universal language, and fluent spoken English is the basis and prerequisite for international cultural exchange [2]. China has long recognized the importance of English, and English learning has become one of the compulsory courses in schools, and English has always been a compulsory subject in all kinds of examinations for further studies. However, traditional English teaching, like other subjects, still adopts the traditional one-way teaching mode, neglecting students' oral application ability and independent learning ability [3]. As a result, many students' oral English proficiency is generally poor. Some English learners are afraid to speak up because they cannot understand or speak well, which will lead to a vicious circle and prevent them from improving their oral proficiency. Especially in the traditional one-to-many learning mode, teachers are unable to identify and correct students' oral pronunciation problems on a one-to-one basis, resulting in students not being able to get their oral pronunciation corrected, which leads to psychological fear of opening their mouths [4]. At the same time, failing to pronounce will also lead to students failing to speak and their listening will also be affected, thus affecting the whole foundation of English learning. With the development of computer technology, the development of computer-assisted English oral pronunciation online learning system provides students with an effective oral practice tool [5]. However, when students practise pronunciation on their own, they are often affected by the pronunciation of their mother tongue and have subtle pronunciation deviations. Current diagnostic techniques for automatic pronunciation detection are not as accurate as they should be due to the limitations of the corpus. Feng et al. designed an end-to-end pronunciation error detection algorithm that integrates attention mechanisms and is applied to an L2-ARCTIC corpus specifically labeled for pronunciation errors by non-native English speakers. However, this method is mainly limited to diagnosing and detecting

[*]Foreign Languages College, Pingxiang University, Pingxiang 337000, Jiangxi, China (Corresponding author, Kangsheng_Lai@outlook.com)

[†]Foreign Languages College, Pingxiang University, Pingxiang 337000, Jiangxi, China

pronunciation errors in L2 learners [6]. Zhang et al. proposed an end-to-end pronunciation error detection algorithm that combines connected temporal classification and attention mechanism. Due to the lack of expert annotated L2 pronunciation error data, this algorithm can only recognize pronunciation errors of L2 learners and cannot provide specific diagnostic information [7]. Against this background, this study innovatively adds a denoising autoencoder to the pronunciation detection and diagnosis module of the computer-assisted oral English pronunciation online learning system, and constructs a denoising autoencoder pronunciation detection and diagnosis system based on end-to-end recurrent neural networks using transfer learning. Therefore, in order to accurately detect errors in English pronunciation by learners in complex environments, and effectively conduct pronunciation training to improve the efficiency and quality of English oral pronunciation learning for English learners. The main contribution of the research is to integrate a denoising autoencoder into the pronunciation detection and diagnosis module of a computer-aided English oral pronunciation online learning system, thereby providing clearer feature information to enhance the detection and diagnosis capabilities of pronunciation errors. In order to effectively improve the accuracy of English oral pronunciation detection and diagnosis, and provide students with more accurate pronunciation correction and guidance. The research content mainly includes four parts. The second part is a review of the current research status of pronunciation detection diagnosis technology and recurrent neural networks both domestically and internationally; The third part discusses the design scheme of a computer-aided English oral English online learning model; The fourth part is to validate the online learning model proposed by the research institute and analyze its specific value in practical applications; The last part is a summary of the entire content and an outlook on future research directions.

2. Related works. Pronunciation detection and diagnostic techniques in second language learning have received a lot of attention from many researchers and have been studied extensively with fruitful results. A team of researchers from Algabri M used deep learning techniques to build a pronunciation detection and diagnostic system for Arabic in order to design a powerful computer-assisted pronunciation system with immediate feedback. The results show that the system has an error recognition rate of only 3.73% in the phoneme recognition process, which significantly improves the accuracy of pronunciation [8]. Wadud M A H and other scholars propose to combine non-self-recursive techniques with end-to-end neural modelling in order to improve the real-time detection of pronunciation errors and to design a new diagnostic model for the detection and diagnosis of mispronunciation. The results show that it exhibits significant advantages in improving detection efficiency [9]. Zhang and other researchers constructed an end-to-end automatic speech recognition system based on hybrid connectionism in order to design an automatic speech recognition system with high performance. The results show that the system can meet the requirements of automatic pronunciation error detection task and achieve high performance index [10].

Recurrent neural networks play an important role in pronunciation detection and diagnostic techniques. Wang's research team, in order to predict future images from historical backgrounds, proposed to utilise the memory decoupling loss of recurrent neural networks for explicit decoupling of memory cells. The results show that this method also prevents the cells from learning redundant features and improves the efficiency and accuracy of the model [11]. Shang and other scholars, in order to be able to accurately predict the degree of haze pollution, proposed to use recurrent neural networks to construct a deep recurrent neural network haze prediction model with time series. The results show that the model can accurately and efficiently predict the degree of haze pollution [12]. Khan and other researchers designed a novel intrusion detection system in order to defend against cyber-attacks, which combines the neural recurrent network structure and machine learning techniques, aiming to effectively defend against cyber-attacks. The results show that the system has high intrusion detection performance [13]. In summary, Algabri M's research points out the effectiveness of deep learning techniques in pronunciation detection and diagnosis, which suggests that when designing English oral pronunciation learning models, this study can also use deep learning frameworks to improve the accuracy of the system. Meanwhile, Wang Y's team's research emphasizes the ability of recurrent neural networks to extract features from time-series data. Therefore, in speech detection, the model in this study can also use RNN to capture temporal information in speech, thereby more accurately identifying pronunciation errors. Therefore, the study aims to construct an end-to-end recurrent neural network model for computer-assisted online learning of spoken English pronunciation with a view to improving the accuracy of spoken English

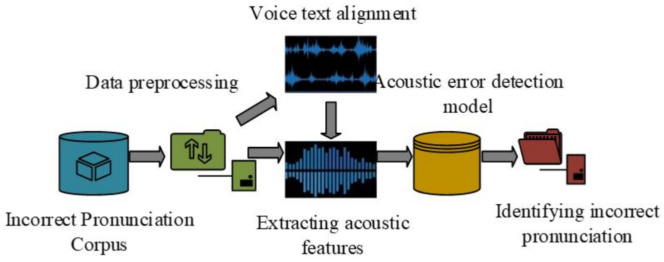

Voice text alignment
Data preprocessing
Acoustic error detection model
Incorrect Pronunciation Corpus
Extracting acoustic features
Identifying incorrect pronunciation

Fig. 3.1: Design Framework for English Spoken Pronunciation Detection and Diagnosis System

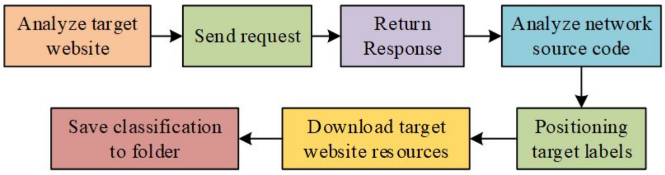

Analyze target website
Send request
Return Response
Analyze network source code
Positioning target labels
Download target website resources
Save classification to folder

Fig. 3.2: Data processing process

pronunciation detection.

3. Design of a Computer-Assisted English Spoken Pronunciation Online Learning Model Based on DAE End-to-End Recurrent Neural Networks. The computer-assisted learning system is the basis of this study, and the study builds a DAE-based framework for English spoken pronunciation detection based on the architecture of this system. Recurrent neural networks and end-to-end techniques are also introduced to design an acoustic detection system architecture based on DAE end-to-end recurrent neural networks, and finally the effectiveness of the system is verified using experiments.

3.1. Construction of DAE-based English Spoken Pronunciation Detection Framework. In recent years, with the continuous development of artificial intelligence and machine learning technology, the application of computer-aided learning systems has become more and more extensive [14]. Among them, deep learning technology has achieved remarkable results in the fields of speech recognition and natural language processing. This provides strong technical support for the design of English spoken pronunciation detection and diagnosis system. By using deep learning technology, automatic detection and diagnosis of spoken English pronunciation can be realized to help learners find pronunciation problems in time and take corresponding corrective measures [15]. The design framework of the spoken English pronunciation detection and diagnosis system constructed by using computer technology is shown in Fig. 3.1.

As shown in Fig. 3.1, the study built a database of English pronunciation errors using web scraping, then preprocessed the data with steps like cleaning, feature extraction, and standardization [16]. Then, by constructing an acoustic model for English spoken mispronunciation recognition, the system can identify and judge the correctness of pronunciation more accurately. Finally, the acoustic model is used for the recognition of spoken English mispronunciation to achieve real-time monitoring and correction of spoken English pronunciation [17]. The data processing flow is shown in Fig. 3.2.

As shown in Fig. 3.2, in the data processing process using web crawling technology, the target website is first analyzed, and then the server is accessed based on the website address to obtain a response. After parsing the webpage source code, the desired target label is located and the data is downloaded. Finally, the obtained data is processed and saved uniformly. The data is obtained from web pages, and research is conducted on using web crawler technology to automatically crawl audio data from web pages. When performing data scraping,

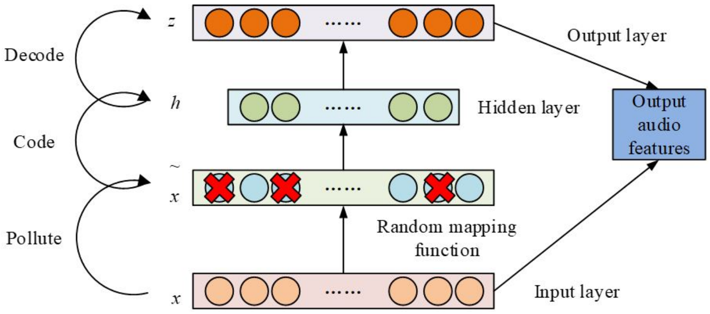


Fig. 3.3: Network structure diagram of DAE

first identify and collect the target webpage URLs containing audio, and place them in the processing queue. Next, using Python and the Requests library, the crawler initiates HTTP requests for URLs in the queue to retrieve webpage content. Then the crawler parses the HTML, identifying audio links and incorrect annotations. Finally, extract these audio download links and error information, crawl and download the audio, and save it locally. The main goal of the data preprocessing stage is to ensure that the audio data in the corpus has a high-quality and unified data format for subsequent model processing. Data cleaning is aimed at removing poor quality audio samples from the corpus, such as records that contain excessive background noise, recording errors, or unclear pronunciation. Then, in order to achieve uniformity in format, research was conducted to convert all audio files into WAV format, and pulse coding modulation was used as the encoding method for the audio. At the same time, during feature extraction, the sampling rate is unified to 16kHz and the data transmission rate of the audio signal is ensured to be 16 bits per second. Finally, in order to standardize the grouping of data, all audio files are set to mono format for saving. In real life, learners' learning environments may encompass a variety of complex and noisy environments, such as streets with noisy traffic, shopping malls with a lot of people, or indoor rooms with reverberating voices [18]. In order to remove or reduce the interference of such noise, the study innovatively adds Denoising Auto Encoders (DAE) to the English spoken pronunciation detection and diagnosis system. DAE achieves the function of removing noise and restoring data by introducing noise into the input data and trying to restore the original data from the noisy data [19]. The DAE's network structure is shown in Fig. 3.3.

As can be seen in Fig. 3.3, the network structure of the DAE is very similar to that of an autoencoder, which can also be divided into three layers: the output layer, the input layer, and the hidden layer, and the overall structure consists of an encoder and a decoder [20-21]. The encoder maps the input data to a representation in the latent space, and unlike a normal autoencoder, the encoder is still responsible for mapping the noisy data to the latent representation after noise is introduced in the input data. The decoder maps the potential representation of the encoder output back to the original input space and tries to reduce the original data. The goal of the decoder is to minimize the interference of noise and restore a result similar to the original data. The training process of the denoising autoencoder can be divided into two steps: adding noise and reconstructing the data. When encoding with DAE, the ReLU activation function is used for calculation. In the calculation equation, the representation of the ReLU activation function is $f\left(t\right)$. The phonemic features of the input spoken English are encoded using DAE as shown in equation (3.1).

$$C_o = f\left(\omega x + b\right) \tag{3.1}$$

In equation (3.1),C_o denotes the coded output of the hidden layer,ω denotes the weight matrix from the input layer to the hidden layer,x denotes the input data,$f\left(t\right)$ denotes the ReLU activation function, and$b$ denotes the bias. The input data needs to be reconstructed with contamination and the mathematical expression for the reconstructed data is shown in equation (3.2).

$$Z = f_d\left(\omega C_o + b\right) \tag{3.2}$$

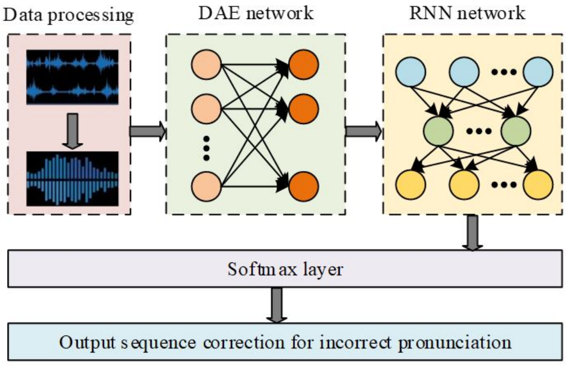


Fig. 3.4: Architecture of acoustic detection system based on DAE end-to-end recurrent neural network

In equation (3.2),Z denotes the reconstruction of the input data$x$ and$f_d\left(t\right)$ denotes the softplus function. Assuming the same number of neural nodes in the input and output layers, i.e., the data$t \in \left(0,1\right)$, then the mathematical expression of the softplus function is shown in Eq. (3.3).

$$f_d\left(t\right) = \begin{cases} \log\left(1+e^t\right), t \in \left(0,1\right) \\ t \qquad\qquad\quad , otherwise \end{cases} \tag{3.3}$$

The DAE network structure is learnt using a loss function as shown in equation (3.4).

$$L\left(x\right) = \sum_{i=1}^{d}\left[x_i \log\left(z_i\right) + \left(1-x_i\right)\log\left(1-z_i\right)\right] + \frac{\lambda}{2}\left\|W\right\|^2 \tag{3.4}$$

In Eq. (3.4),$L\left(x\right)$ denotes the loss function, and$\frac{\lambda}{2}\left\|W\right\|^2$ denotes the regularity term, whereλ denotes the parameter that controls the degree of regularisation. By stacking the network structure of DAE, a deep learning model can be formed, which is able to realise the cycle of the above process, continuously training from noisy speech data and obtaining a representation closer to the original data. In this way, i.e., the accuracy of spoken English recognition can be improved.

3.2. Architecture design of acoustic detection system based on DAE end-to-end recurrent neural network. Recurrent Neural Network (RNN) is a type of neural network specialized in processing sequential data, which recursively moves in the direction of sequence evolution and connects all nodes according to chaining. RNNs have an excellent memory capability, which enables them to combine the contents of previous memories with the current inputs, and thus apply the previous information to the current task. Unlike Convolutional Neural Network (CNN), RNN not only considers the current input, but also remembers the information from previous moments, and this memory capability gives RNN a great advantage in processing sequential data. In the process of English pronunciation detection, the mispronunciation of spoken English is often associated with its preceding and following phonemes. Therefore, in order to strengthen the connection between spoken English in terms of the preceding and following phonemes, it is necessary to process the features using RNN after extracting them using the DAE English Spoken Pronunciation Detection System. Since the input speech signals of spoken English are often very long and the dimension of the input is large, the duration of a phoneme is usually ten times the length of a frame [22]. In order to deal with such audio with excessive input length, the study introduces end-to-end processing in the construction of the system. The architecture of the acoustic detection system based on DAE end-to-end recurrent neural network is shown in Fig. 3.4.

As shown in Fig. 3.4, the acoustic detection system architecture combines DAE and RNN with the goal of recovering the original data from noisy speech data. Firstly, the spoken English audio data is captured and acoustic features are extracted, then input to DAE network for learning, after that the learned features are input to RNN network structure. The output of RNN is classified by softmax layer for output, and finally the student's mispronunciation is corrected based on the predicted output sequences. LSTM is a kind of RNN designed to deal with the long term dependency problem. RNN is difficult to handle long-term dependency problems, while LSTM can solve this problem by introducing several gating units. This unique structure allows the network to selectively remember or forget information, which is very effective for capturing long-term dependencies. Meanwhile, the forget gate of LSTM allows the model to discard unwanted information, which is very useful when dealing with continuous speech streams. And the unique gating mechanism of LSTM can protect gradients from damage during long sequence transmission, making the training process more stable. Therefore, choosing LSTM as a specific network architecture is more advantageous for research. When dealing with sequence data such as speech and text, if the sequence is too long, the RNN will suffer from the gradient explosion or vanishing problem, making the network difficult to train and optimize. To solve this problem, the study introduces LSTM instead of traditional RNN structure. The expression of LSTM network forgetting gate is shown in equation (3.5).

$$f_t = \sigma\left(\omega_f * [y_{t-1}, x_t] + b_f\right) \tag{3.5}$$

In equation (3.5),f_t denotes the forgetting gate,σ denotes the activation function,ω_f denotes the weight matrix,x_t denotes the current neuron input,t denotes the time,,b_f denotes the residual value, andy_{t-1} denotes the output of the previous neuron. The activation functionσ is generally used as a sigmoid function and its mathematical expression is shown in equation (3.6).

$$\sigma\left(x\right) = \frac{1}{1 + e^{-x}} \tag{3.6}$$

Candidate cells are utilised for updating and the formula for updating is shown in equation (3.7).

$$\left\{\begin{array}{l} i_t = \sigma\left(\omega_i * [y_{t-1}, x_t] + b_i\right) \\ C_t = f_t \bullet C_{t-1} + i_t \bullet C'_t \\ C'_t = \tanh\left(\omega_C * [y_{t-1}, x_t] + b_c\right) \end{array}\right. \tag{3.7}$$

In Equation (3.7),C denotes the memory cell,i_t denotes the input gate,C_t denotes the candidate value cell,$C_t{}'$ denotes the updated candidate value cell,ω_C denotes the weight value of the memory cell,ω_i denotes the weight value matrix of the input gate,b_c denotes the offset value of the memory cell, andb_i denotes the offset value of the input gate. The expression of the output gate is shown in equation (3.8).

$$\left\{\begin{array}{l} y_t = O_t \bullet \tanh\left(C_t\right) \\ O_t = \sigma\left(\omega_o \bullet [y_{t-1}, x_t] + b_o\right) \end{array}\right. \tag{3.8}$$

In Eq. (3.8),O_t denotes the output gate,ω_o denotes the weight value of the output gate, andb_o denotes the deviation value of the output gate. In the acoustic detection system architecture, the DAE network is first used to learn and compute the loss function, then this loss function is used to train the LSTM network, and finally the output of the LSTM network is fed back to the DAE network through the sigmoid function in order to optimise the system performance. The expression of the loss function after training is shown in equation (3.9).

$$L'\left(x, z\right) = -\ln\left[P\left(z \left| x\right.\right)\right] \tag{3.9}$$

In Eq. (3.9),$L'\left(x, z\right)$ denotes the loss function after training and$P$ denotes the output probability. After performing forward and backward calculations through the LSTM network, the forward and backward variables will be initialized. In LSTM networks, before each processing of sequence data begins, it is necessary to initialize the forward variable. These variables include a set of loss function gradients, which will be updated during the training process. The mathematical expression of the initialized forward variable is shown in equation (3.10).

$$\alpha\left(1, u\right) = \left\{\begin{array}{l} C_o^b, u = 1 \\ C_o^Z, u = 2 \\ 0, Q \end{array}\right. \tag{3.10}$$

In Eq. (3.10),α denotes the initialized forward variable,u denotes the set of gradients of the$L'(x,z)$ loss function and$Q$ denotes the others. Backward variables also need to be initialized before processing begins, and these variables involve a set of all phoneme labels. The mathematical expression of the initialized backward variable is shown in equation (3.11).

$$\beta(T,u) = \left\{ \begin{array}{l} 1, u = Z \\ 0, Q \end{array} \right. \tag{3.11}$$

In Eq. (3.11),β denotes the initialised backward variable and$T$ denotes the set of all phoneme labels. Forward propagation refers to the process of data transmission from the input layer to the output layer in a network. The mathematical expression for forward propagation is shown in equation (3.12).

$$\alpha(t,u) = y_k^t \sum_{i=f(u)}^{u} \alpha(t-1,i) \tag{3.12}$$

In Eq. (3.12),$\alpha(t,u)$ denotes the forward propagation,y_k^t denotes the output of the softmax layer at the time$t$ and$k$ denotes the phoneme labels. Backpropagation is a crucial step in the learning process, which involves calculating the gradient of weights for each layer based on the loss function and updating weights according to these gradients. The mathematical expression for backward propagation is shown in equation (3.13).

$$\beta(t,u) = \sum_{i=u}^{t} \beta(t+1,i)\, y_k^t \tag{3.13}$$

In Eq. (3.13),$\beta(t,u)$ denotes backward propagation. The boundary conditions for the forward and backward variables are shown in equation (3.14).

$$\left\{ \begin{array}{l} \alpha(t,0) = 0, \forall t \\ \beta(t,|Z|+1) = 0, \forall t \end{array} \right. \tag{3.14}$$

As shown in equation (3.14), the boundary conditions of the forward and backward variables define the values of the forward and backward variables under specific conditions. The difference between the predicted and actual values of the output layer is the error gradient, which can adjust the weight of the network to reduce future prediction errors. The mathematical expression of the output layer output error gradient is shown in equation (3.15).

$$\tau_k^t = Z_k^t - \frac{1}{P(Z\,|x)} \sum_u \alpha(t,u)\,\beta(t,u) \tag{3.15}$$

In Eq. (3.15),τ_k^t denotes the gradient of the output error of the output layer. When training the LSTM model, the gradient of the output error can be used to update and learn the parameters of the network. Once the LSTM model is trained, the target pronunciation sequence can be recognised. However, during the initial phoneme training recognition process, the LSTM model is prone to overfitting in the training set. To avoid this, a Dropout strategy can be used, where a portion of neurons are randomly discarded during the training process.The Dropout process of the LSTM model is shown in Fig. 3.5.

As shown in Fig. 3.5, the Dropout process of the LSTM model is a neuron dropout on the feedforward network of the LSTM, with a probability of 0.5 to randomly drop some neurons at each layer. This process can effectively reduce the sequence modelling ability of the recurrent neural network, thus effectively mitigating the overfitting phenomenon. At the same time, since Dropout is performed on the feedforward neural network, it can prevent information from being lost during the looping process, thus ensuring the performance and accuracy of the model.

4. Validation of a DAE end-to-end recurrent neural network-based model for computer-assisted online learning of spoken English pronunciation. In this chapter, the configuration of the experimental environment and parameters is carried out, then the analysis of the model parameters of the LSTM network structure is analysed, followed by the performance validation of the model for learning spoken English pronunciation

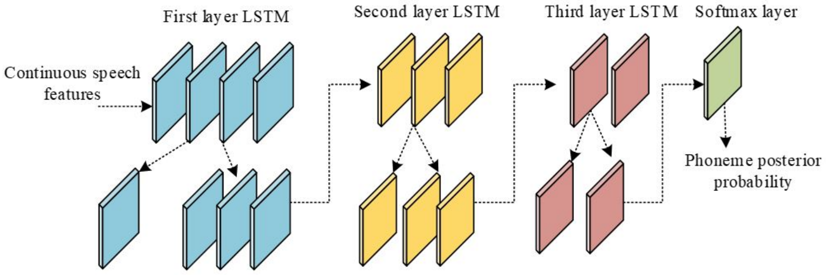


Fig. 3.5: Dropout process of LSTM model

Table 4.1: Experimental Environment and Parameter Configuration

Experimental environment	Configuration	Parameter	Value
Operating system	Windows 10	Optimizer	Adam
Memory	64GB	Iterations	500
GPUs	NVIDIA TITAN BLACK GPUs	Batchsize	64
Video storage	6G	Learning Rate	0.001
Programming Language	Python	Dropout	0.5

4.1. Experimental environment and parameter configuration. In order to verify the effectiveness of the computer-assisted English oral pronunciation online learning model based on DAE end-to-end recurrent neural network, the experimental environment is firstly constructed and the parameters are set. The experimental operating system is Windows 10, the programming language is Python, and the acoustic model for pronunciation detection is TensorFlow 1.4.0.The dataset used in the experiment is collected from the Internet by web crawler technology to construct a corpus of English spoken pronunciation errors. On this basis, in order to be closer to real scenarios, the study added white noise and pink noise to this corpus, and divided this corpus into a training set and a test set according to the ratio of 60%:40%. In the experiment, a neural network with three hidden layers was constructed, with a number of neurons of 39, 50, and 50, respectively. To enhance generalization and avoid overfitting, both the input layer and hidden layer adopt a 10% Dropout rate. The training parameters are set to 500 Epochs, with a batch size of 40 and a learning rate of 0.0001. The study selected Adam as the optimizer, with hyperparameters for first-order and second-order moment estimation set to 0.9 and 0.999, respectively. The batch size during training is set to 64 to save memory. To further prevent overfitting, the Dropout rate is set to 0.5. he specific experimental environment and parameter configuration are shown in Table 4.1.

4.2. Analysis of model parameters for LSTM network structure. In order to verify the effect of the number of layers of the LSTM network on the performance of the DAE end-to-end recurrent neural network-based computer-assisted oral English pronunciation online learning model, the number of LSTM layers was set from 1 to 4, and the phoneme error recognition accuracies were compared with different numbers of LSTM network layers as shown in Fig. 4.1. From Fig.‘4.1, it can be seen that the values of the parameters such as insertion, deletion, substitution and phoneme error recognition accuracy gradually increase with the increase in the number of LSTM network layers. When the number of LSTM network layers reaches 4, the recognition accuracy reaches the highest value of 84.04%. This is an improvement of 20.07%, 12.48% and 7.81% as compared to when the number of layers is 1, 2 and 3 respectively. Thus it can be seen that the performance of phoneme recognition is better when the number of layers of LSTM network is 4.

In order to investigate the effect of the number of LSTM network nodes on the performance of the computer-assisted oral English pronunciation online learning model with DAE end-to-end recurrent neural network, the study experimented with the number of network nodes in the hidden layer of 100, 150, 200, and 300, and the

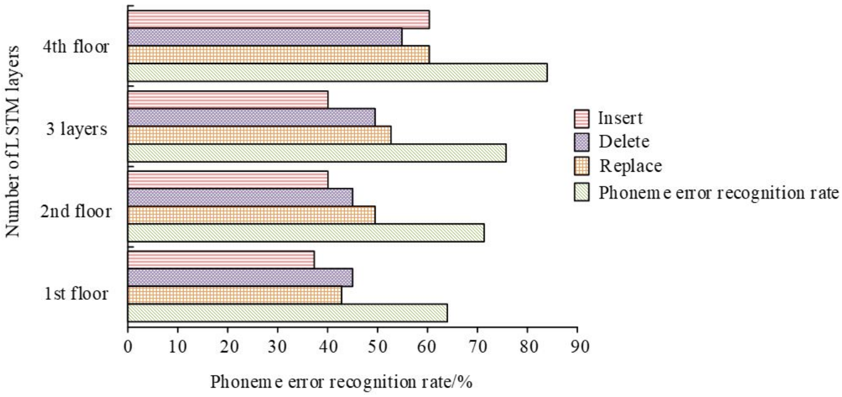


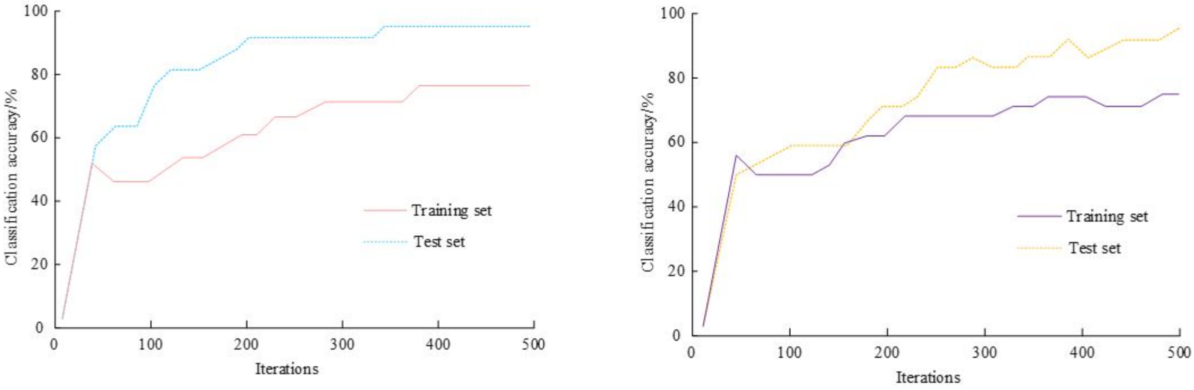
Fig. 4.1: The accuracy of phoneme error recognition for different LSTM network layers

Table 4.2: The impact of different numbers of LSTM network nodes on model performance

Number of network nodes	Insert/%	Delete/%	Replace/%	Phoneme sequence recognition rate/%	Accuracy/percent	Recall/%	F1 value/%	Recognition rate of mispronounced phonemes
100	2.65	7.33	2.65	85.65	89.01	81.35	88.18	84.16
150	2.65	5.33	2.65	87.69	87.98	87.48	87.15	87.06
200	2.65	13.00	6.33	75.97	82.16	76.15	78.61	77.94
300	6.00	15.33	1.33	78.12	82.16	86.70	86.19	83.76

effect of the different number of LSTM network nodes on the performance of the model is shown in Table 4.2. From Table 4.2, it can be seen that when the number of LSTM network nodes is 150, the phoneme sequence recognition rate is 87.69%, which is an improvement of 2.04%, 11.72% and 9.57% compared to the cases of number of nodes 100, 200 and 300, respectively. When the number of nodes in the LSTM network is 150, the mispronunciation phoneme recognition rate reaches a maximum value of 87.06%, which is an improvement of 2.9%, 9.12% and 3.3% compared to the cases of 100, 200 and 300 nodes, respectively. In summary, considering various factors such as phoneme sequence recognition rate and mispronunciation phoneme recognition rate, it can be seen that the number of nodes of LSTM network set to 150 is the most appropriate. When the number of nodes in the LSTM network is set to 150, the model performs the best in phoneme sequence and mispronunciation recognition tasks. This finding emphasizes the importance of selecting an appropriate network size, that is, a depth and width that is neither excessive nor insufficient, for the model to capture key information and generalize new data.

4.3. Performance Validation of English Spoken Pronunciation Learning Model. In order to verify the classification performance of the English spoken pronunciation learning model in a noisy environment, a comparison of the classification performance of the model in different noise environments is plotted as shown in Fig. 7. As can be seen from Fig. 7(a), before 50 iterations, the classification accuracy of the training and testing sets was the same. However, after 50 iterations, the classification accuracy of the testing set gradually widened compared to the training set, indicating overfitting in the training set. In the white noise environment, the model gradually becomes stable after 400 iterations, and the classification accuracy in the training set and test set reaches 78.97% and 94.01%, respectively. This indicates that the model has a better classification performance under white noise environment. As can be seen in Fig. 7(b), after 50 iterations, the



(a) Classification accuracy under White noise

(b) Classification accuracy under White noise

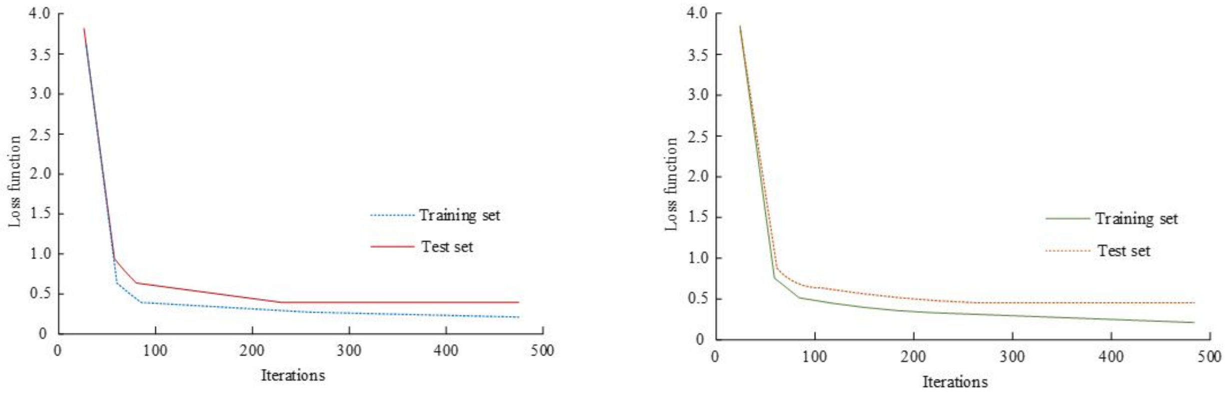
Fig. 4.2: Comparison of Model Classification Performance in Different Noise Environments

classification accuracy of the training set showed a downward trend, which remained stable for a period of time before returning to a slow upward trend. This indicates that the model can better generalize to unseen data after further learning and adjustment, indicating its robustness. The classification accuracy of the model in pink noise environment reaches 76.19% and 94.03%, respectively. This indicates that the model also has good classification performance in pink noise environment. In summary, it can be seen that the model still maintains a high classification accuracy in white and pink noise environments, indicating that the model still has good adaptability in the presence of background noise.

In order to further verify the recognition performance of the English spoken pronunciation learning model in the noise environment, the loss function profiles of the model in different noise environments are plotted as shown in Fig. 4.3. As can be seen from Fig. 4.3, before the number of iterations approaches 60, the value of the loss function shows a rapid decline trend. After 60 iterations, the decline rate gradually slows down and gradually stabilizes around 100 iterations. This indicates that the model learns quickly in the initial iteration process, effectively reducing errors. However, as learning progresses, the model's fitting of the data gradually approaches its potential optimal state, and the decline rate of the loss function slows down until it reaches a relatively stable state, reflecting that the learning process of the model is beginning to saturate. The loss function curves in white noise and pink noise environments have very similar trends, both of which begin to converge around 100 iterations, and the values of the loss function eventually converge and stabilize at 0.5 or below. It shows that the model has good adaptability in both white noise and pink noise environments, and can effectively reduce the value of the loss function and improve the classification accuracy of the model.

In order to validate the effectiveness of the spoken English pronunciation learning model, a comparison of the audio data pairs of spoken English before and after DAE network processing is shown in Fig. 9. As can be seen from Fig. 4.5(a) and Fig. 4.5(b), the vowel signals after DAE network processing are sparser compared with the original vowel signals. At the same time, the vowel signal processed by DAE network has some enhancement in the high frequency part of the signal compared to the original vowel signal. As can be seen in Fig. 9(c) and Fig. 9(d), the spectral images of the consonant signals processed by the DAE network become smoother and smoother compared with the original consonant signals. This indicates that the DAE network can effectively reduce the noise and interference in the audio signal and improve the quality and clarity of the signal. In summary, it can be seen that the audio signal processed by the DAE network becomes sparser in the high-frequency part, indicating that the network can effectively filter out unnecessary noise and retain useful speech information.

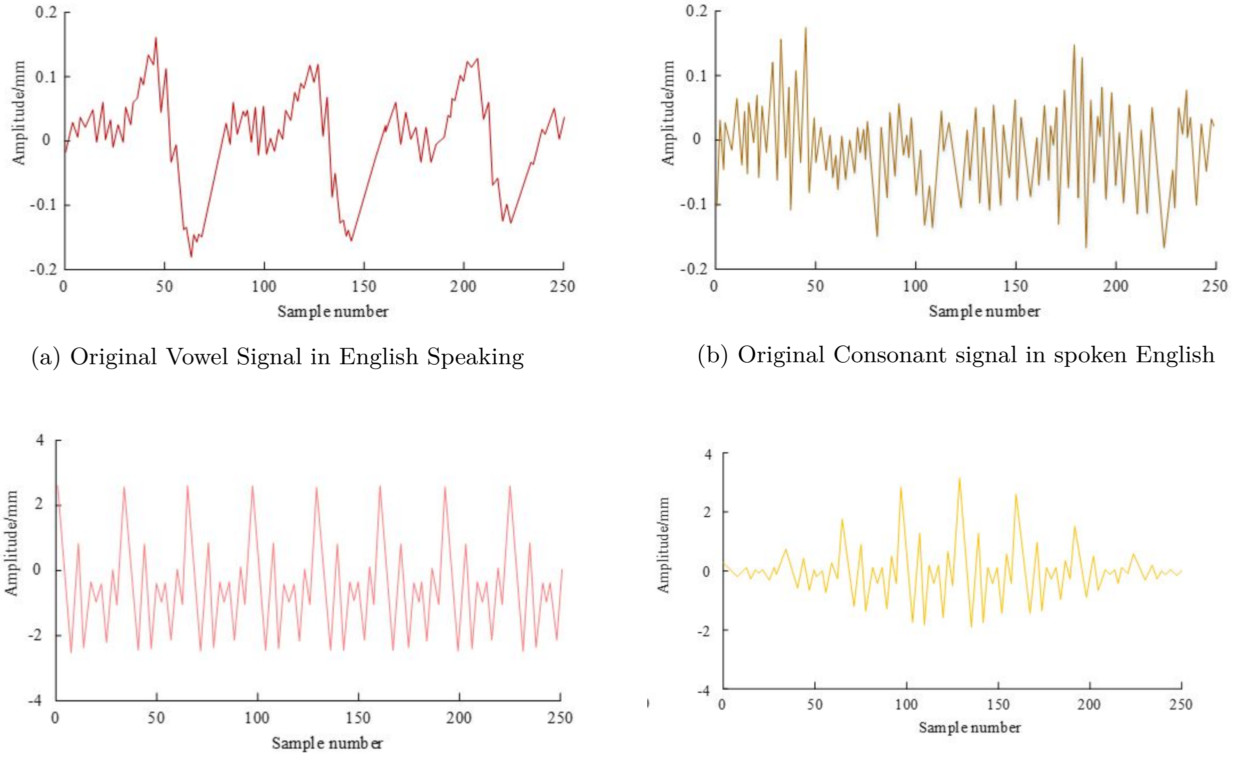
In order to verify the effectiveness of the English spoken pronunciation learning model in practical applications, the acoustic detection system based on DAE end-to-end recurrent neural network is compared with the



(a) Classification accuracy under White noise

(b) Classification accuracy under White noise

Fig. 4.3: Comparison of loss functions of models under different noise environments



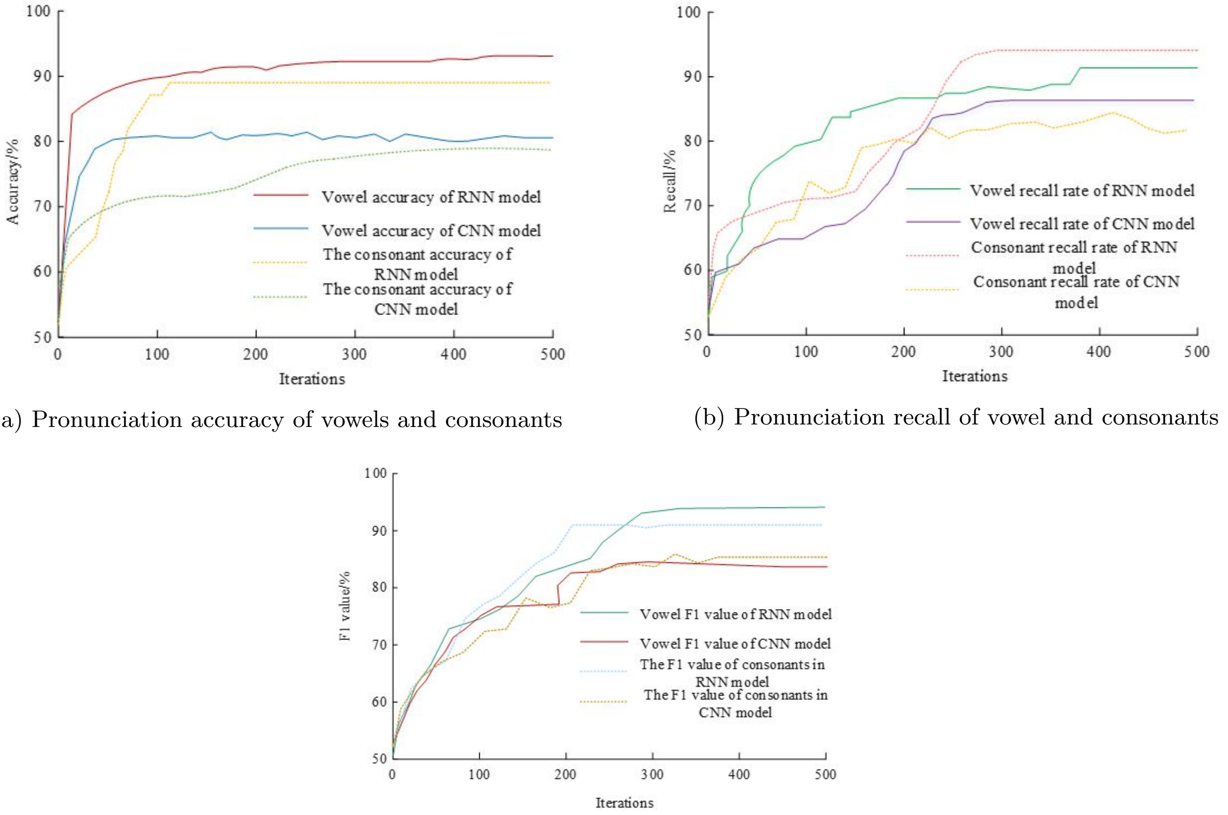
(a) Original Vowel Signal in English Speaking

(b) Original Consonant signal in spoken English

(c) English Vowel signal processed by DAE network

(d) English Consonantsignal processed by DAE network

Fig. 4.4: Comparison of English-Speaking Audio Data before and after DAE Network Processing



(a) Pronunciation accuracy of vowels and consonants

(b) Pronunciation recall of vowel and consonants

(c) The F1 value of vowel and consonant Pronunciation

Fig. 4.5: The pronunciation detection performance of the model

acoustic detection system based on convolutional neural network. The pronunciation detection performance of the validation model is shown in Fig. 4.5a. As can be seen in Fig. 4.5a(a), in vowel articulation detection, the accuracy of the recurrent neural network-based model reaches 94.07%, which is 13.43% higher than the accuracy of the convolutional neural network-based model. In consonant articulation detection, the accuracy of the recurrent neural network-based model is 89.12%, which is 10.15% higher than the accuracy of the convolutional neural network-based model. As can be seen in Fig. 4.5a(b), in vowel pronunciation detection, the recall based on the recurrent neural network model is 91.89%, which is 5.55% higher than the recall based on the convolutional neural network model. In consonant articulation detection, the recall based on the recurrent neural network model is 94.27%, which is 12.24% higher than the recall based on the convolutional neural network model of 82.03%. As can be seen in Fig. 10(c), in vowel articulation detection, the F1 value based on the recurrent neural network model is 93.76%, which is 10.49% higher than the F1 value based on the convolutional neural network model. In consonant articulation detection, the F1 value based on recurrent neural network model is 91.52%, which is 6.29% higher than the F1 value based on convolutional neural network model. Thus it can be seen that the DAE end-to-end recurrent neural network based acoustic detection system has better articulation detection performance compared to the convolutional neural network based acoustic detection system.

In order to more intuitively verify the effectiveness of the English spoken pronunciation learning model in practical applications, the acoustic detection system based on DAE end-to-end recurrent neural network was inductively compared with the acoustic detection system based on convolutional neural network, and the

Table 4.3: Performance of Two Models in English Oral Pronunciation

Performance	CNN based system		RNN based system	
	Vowel	Consonant	Vowel	Consonant
Accuracy/per cent	80.64	78.97	94.07	89.12
Recall/%	86.34	82.03	91.89	94.27
F1 value/%	83.27	85.23	93.76	91.52
Error detection accuracy/%	75.32	74.06	88.91	91.68
Error correction accuracy/%	73.16	60.16	90.67	91.96

performance of the two models on English spoken pronunciation is shown in Table 4.3. As can be seen in Table 4.3, in terms of error detection correctness, the correctness of the recurrent neural network-based model for vowels and consonants is 88.91% and 91.68%, which is an improvement of 13.59% and 17.62%, respectively, compared to the convolutional neural network-based model. In terms of the correct rate of error correction, the correct rates of vowels and consonants based on the recurrent neural network model are 90.67% and 91.96%, which are improved by 17.51% and 31.8%, respectively, compared with the model based on convolutional neural network. In summary, the effectiveness of the computer-assisted oral English pronunciation online learning model in practical applications is verified through comparative experiments.

5. Conclusion. As a globally used communication language, fluent spoken English can significantly enhance one's social competitiveness. In order to improve the pronunciation and error correction ability of online learning of spoken English, the study is based on a computer-assisted online learning system for spoken English pronunciation, which extracts the acoustic features of the audio by introducing a DAE module and processes the audio data using a recurrent neural network structure. The results show that under the white noise environment, the classification accuracies of the training set and test set are as high as 78.97% and 94.01%, respectively. In pink noise environment, the classification accuracy is 76.19% and 94.03%, respectively. In the vowel pronunciation detection task, the accuracy, recall and F1 value of the recurrent neural network-based model reached 94.07%, 91.89% and 93.76%, respectively. In the consonant pronunciation detection task, the values of these three metrics were 89.12%, 94.27% and 91.52%, respectively. In terms of the correct rate of error detection, the model was 88.91% and 91.68% correct in vowel and consonant pronunciation detection, respectively. In terms of the correct rate of error correction, the correct rates of the recurrent neural network-based model in vowel and consonant articulation detection are 90.67% and 91.96%, respectively. In summary, the DAE end-to-end recurrent neural network-based acoustic detection system has significant advantages in terms of error detection, error correction, and overall classification performance for spoken English pronunciation. However, only vowels and consonants in spoken pronunciation were analyzed in this study, and the study can be further improved in the future to include other types of phonemes in the analysis in order to evaluate the performance of the model more comprehensively. Faced with existing challenges such as the lack of immediate feedback in online learning environments, this system provides learners with a powerful self-learning platform through its high accuracy error correction function. The potential application of this technology in research is not limited to personalized tutoring for language learners, but may also extend to the diagnosis of acoustic barriers, promotion of cross-cultural communication, and optimization of distance education resources. In the future, research can be expanded to include analysis of more phonemes to comprehensively evaluate and enhance the model's universality and adaptability in multilingual environments. In addition, researchers should also consider how to integrate this technology into existing digital learning platforms to achieve a wider educational impact and contribute to language education in the era of globalization.

REFERENCES

[1] Garcia-Perez, D., Pérez-López, D., Diaz-Blanco, I., Gonzalez-Muniz, A., Dominguez-Gonzalez, M. & Vega, A. Fully-convolutional denoising auto-encoders for NILM in large non-residential buildings. *IEEE Transactions On Smart Grid.* **12**, 2722-2731 (2020)

[2] Gheller, C. & Vazza, F. Convolutional deep denoising autoencoders for radio astronomical images. *Monthly Notices Of The Royal Astronomical Society.* **509**, 990-1009 (2022)

[3] Larrazabal, A., Martínez, C., Glocker, B. & Ferrante, E. Post-DAE: anatomically plausible segmentation via post-processing with denoising autoencoders. *IEEE Transactions On Medical Imaging.* **39**, 3813-3820 (2020)

[4] Li, X., Liu, Z. & Huang, Z. Deinterleaving of pulse streams with denoising autoencoders. *IEEE Transactions On Aerospace And Electronic Systems.* **56**, 4767-4778 (2020)

[5] Liu, P., Zheng, P. & Chen, Z. Deep learning with stacked denoising auto-encoder for short-term electric load forecasting. *Energies.* **12**, 2445-2447 (2019)

[6] Feng, Y., Fu, G., Chen, Q. & Chen, K. SED-MDD: Towards sentence dependent end-to-end mispronunciation detection and diagnosis. *ICASSP 2020-2020 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP).* **21**, 3492-3496 (2020)

[7] Zhang, L., Zhao, Z. & Ma, C. End-to-End Automatic Pronunciation Error Detection Based on Improved Hybrid CTC/Attention Architecture. *Sensors.* **20**, 1809-1811 (2020)

[8] Algabri, M., Mathkour, H., Alsulaiman, M. & Bencherif, M. Mispronunciation detection and diagnosis with articulatory-level feedback generation for non-native arabic speech. *Mathematics.* **10**, 2727-2728 (2022)

[9] Wadud, M., Alatiyyah, M. & Mridha, M. Non-autoregressive end-to-end neural modeling for automatic pronunciation error detection. *Applied Sciences.* **13**, 109-112 (2022)

[10] Wang, Y., Wu, H., Zhang, J., Gao, Z., Wang, J., Philip, S. & Long, M. End-to-end automatic pronunciation error detection based on improved hybrid ctc/attention architecture. *Sensors.* **20**, 1809 (2022)

[11] Wang, Y., Wu, H., Zhang, J., Gao, Z., Wang, J., Philip, S. & Long, M. Predrnn: A recurrent neural network for spatiotemporal predictive learning. *IEEE Transactions On Pattern Analysis And Machine Intelligence.* **45**, 2208-2225 (2022)

[12] Shang, K., Chen, Z., Liu, Z., Song, L., Zheng, W., Yang, B. & Yin, L. Haze prediction model using deep recurrent neural network. *Atmosphere.* **12**, 1625 (2021)

[13] Khan, M. HCRNNIDS: Hybrid convolutional recurrent neural network-based network intrusion detection system. *Processes.* **9**, 834-836 (2021)

[14] Raj, D. & Ananthi, J. Recurrent neural networks and nonlinear prediction in support vector machines. *Journal Of Soft Computing Paradigm.* **1**, 33-40 (2019)

[15] Guo, K., Hu, Y., Qian, Z., Liu, H., Zhang, K., Sun, Y. & Yin, B. Optimized graph convolution recurrent neural network for traffic prediction. *IEEE Transactions On Intelligent Transportation Systems.* **22**, 1138-1149 (2020)

[16] Onan, A. Bidirectional convolutional recurrent neural network architecture with group-wise enhancement mechanism for text sentiment classification. *Journal Of King Saud University-Computer And Information Sciences.* **34**, 2098-2117 (2022)

[17] Apaydin, H., Feizi, H., Sattari, M., Colak, M., Shamshirband, S. & Chau, K. Comparative analysis of recurrent neural network architectures for reservoir inflow forecasting. *Water.* **12**, 1500-1503 (2020)

[18] Hewamalage, H., Bergmeir, C. & Bandara, K. Recurrent neural networks for time series forecasting: Current status and future directions. *International Journal Of Forecasting.* **37**, 388-427 (2021)

[19] Yu, Y., Si, X., Hu, C. & Zhang, J. A review of recurrent neural networks: LSTM cells and network architectures. *Neural Computation.* **31**, 1235-1270 (2019)

[20] Mokayed, H., Quan, T., Alkhaled, L. & Sivakumar, V. Real-time human detection and counting system using deep learning computer vision techniques. *Artificial Intelligence And Applications.* **1**, 221-229 (2023)

[21] Kumar, V., Arulselvi, M. & Sastry, K. Comparative Assessment of Colon Cancer Classification Using Diverse Deep Learning Approaches. *Journal Of Data Science And Intelligent Systems.* **1**, 128-135 (2023)

[22] Garai, S., Paul, R., Kumar, M. & Choudhury, A. Intra-Annual National Statistical Accounts Based on Machine Learning Algorithm. *Journal Of Data Science And Intelligent Systems.* **2**, 12-15 (2023)

Edited by: Mudasir Mohd
Special issue on: Scalable Computing in Online and Blended Learning Environments:
Challenges and Solutions
Received: Feb 21, 2024
Accepted: May 7, 2024