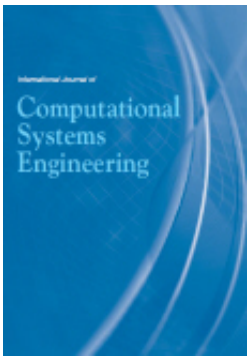
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Computational Systems Engineering

ISSN online: 2046-3405 - ISSN print: 2046-3391
https://www.inderscience.com/ijcsyse

A corpus-based study on the characteristics of the use of spoken English chunks

Rong Hu

DOI: 10.1504/IJCSYSE.2023.10054652

Article History:

Received:	08 August 2022
Last revised:	21 October 2022
Accepted:	30 January 2023
Published online:	19 March 2024

A corpus-based study on the characteristics of the use of spoken English chunks

Rong Hu

College of Humanities,
Ningbo University of Finance & Economics,
Ningbo, 315175, China
Email: 2016122577@jou.edu.cn

Abstract: This study constructs the English-speaking SELL corpus, proposes a CNN-LSTM-SA algorithm model-based English speaking recognition technique for the use of English-speaking blocks, and analyses the results of the SELL corpus and the speaking recognition model. The results show that the model's loss rate shows a trend of slow increase after a sharp decrease. When the number of iterations of the model is 300, the inflection points of the loss value and accuracy rate occur. At this point, the accuracy tends to converge, and its training accuracy is close to 88%, which is significantly higher than other algorithms. The CNN-LSTM network performs the best under the ReLu and tanh functions selected for the study, and the MAE and RMSE indexes 26.54 and 36.11, respectively. The model performance is higher than other algorithms under all six complexities, and its difference is about 4% at the lowest, with a very stable performance advantage.

Keywords: corpus; spoken English; chunk features; self-attentive mechanism; CNN-LSTM.

Reference to this paper should be made as follows: Hu, R. (2024) 'A corpus-based study on the characteristics of the use of spoken English chunks', *Int. J. Computational Systems Engineering*, Vol. 8, Nos. 1/2, pp.40–47.

Biographical notes: Rong Hu obtained her BA in English from Hunan Agricultural University in 2007. She obtained her MA in foreign linguistics and applied linguistics from Changsha University of Science & Technology in 2010. Presently, she is working in the College of Humanities, Ningbo University of Finance & Economics. Her areas of interest are applied linguistics, Corpus linguistics, English education and teaching.

1 Introduction

English is a universal language. About 1.75 billion people in the world use English, including 20% native speakers and 80% non-native speakers (Liu et al., 2021). At the same time, English is the international language for business, politics and diplomacy, as well as the first language for academic exchanges, especially in the fields of computers and the Internet. With China's increasing participation in globalisation, it has become the largest country with English as its second language. The number of English learners in China exceeds 300 million, including 210 million students in public schools and 90 million professionals. Students in public schools need to study English for at least nine years in compulsory education, and most of them will still continue to take English courses after the compulsory education. Traditional English teaching method in China adopts listening, grammar, reading and writing as main methods to test English proficiency of students, however, speaking is often ignored, so domestic students usually have strong reading and writing skills but poor speaking skills (He et al., 2020). As a result, English has lost its most important function of oral communication. This situation has also attracted the attention of the educational community who hopes to improve this situation through the reform of English teaching and the strengthening of oral tests. However, the slow progress of reform makes it difficult to achieve good results in a short time. In such teaching context, most students have poor pronunciation and are ReLuctant to open their mouths for practice (Zhang, 2021). Generally speaking, exposure to a pure English environment is the best choice to improve students' oral English ability. Therefore, in order to improve students' oral English ability, many parents will take their children to English speaking countries on holidays, or take their children to various commercial English training organisations to improve learners' oral English ability by learning and communicating with teachers in English speaking countries. These methods can indeed make a lot of progress, but they usually require a lot of money, time and energy. Not all families can afford these expenses. Therefore, the domestic education sector hopes to achieve a breakthrough in the teaching of spoken English by means of teaching reform. Corpora belong to a class of large structured texts, which are databases of a large number of natural language processing techniques, including syntactic analysis, text proofreading, and speech recognition (Han and Yin, 2021). Among them, the spoken language corpus contains audio and transcription files and is widely used in

acoustic model building, while the spoken language corpus is used to study dialects and phonemes in teaching linguistics courses (Rohdenburg, 2021). Some domestic scholars have used intelligent algorithms such as speech recognition architecture and sequence matching for English oral speech recognition design, but the selected feature parameters have a high number but a low degree of integration with the test construal, which means the system validity test and application are poor. Therefore, a corpus of spoken English chunks has been constructed by the study, and a convolutional neural network (CNN) fused with short and long-term memory networks (LSTM) model has been designed, which is optimised by using self-attention (SA) mechanism. The algorithm model of CNN-LSTM-SA for spoken English recognition is constructed through the above methods. This model improves the accuracy of spoken English recognition in the corpus system, can provide applied technical tools for oral English test activities, and reduce the errors and inefficiencies of manual scoring.

2 Related work

At present, the research work based on the use of spoken English chunks has become an important direction in the field of education, and the teaching of spoken English has received attention from many scholars at home and abroad. Cao and Guo (2020) constructed an evaluation model of spoken English based on fuzzy metrics and speech recognition technology by using fuzzy metrics to evaluate spoken English. The study showed that the speaking evaluation algorithm based on fuzzy metrics and speech recognition technology is superior compared with the traditional algorithm. Yu (2020) proposed a fault-tolerant alignment and search filtering algorithm based on syllable unit WFST network to solve the difficulties of repeated spoken English error detection and correction in computer-assisted language teaching. Experiments have shown that multivariate hybrid search filtering with syllable as the modelling unit achieves relatively optimal results without using secondary fault-tolerant alignment. There are also more studies on CNN algorithms fusing LSTM algorithms by domestic and foreign scholars. Xie et al. (2020) proposed an enhanced grey wolf optimiser (GWO) to design evolutionary CNN-LSTM networks for time series analysis. The results show that the proposed CNN-LSTM optimisation model produces results that significantly outperform the optimisation results of classical search methods and the GWO fusion particle swarm algorithm, yielding that the CNN-LSTM network provides better representational capabilities that not only capture the interaction of important features but also encapsulate complex dependencies in complex temporal contexts to perform time-series tasks. Liu and Ren (2019) proposed an enhanced emotion-triggered system to achieve smooth interactions with the robot, where the emotion triggering was replaced by a deep neural network of CNN-LSTM. The results showed that the CNN-LSTM-based model required only ten ms or less to complete the classification without degrading accuracy, significantly outperforming other algorithms. Song et al. (2021) proposed a spatio-temporal hybrid CNN-LSTM-based prediction for more reasonable prediction of the heating load of SDHS model. The experimental results showed that the prediction performance of the CNN-LSTM-based algorithm has obvious accuracy advantages, and the MAPE evaluation indexes of the four heat exchange stations were distributed between 3.1% and 4.1%. The algorithm has good adaptability to heat load data with different value ranges and can better meet the needs of field engineering applications. Guan et al. (2020) used CNN networks to extract and reorganise Chinese character features in whole sentences. The CNN network combined the reorganised features into a dual LSTM network, and finally weighted each word in the whole sentence to classify the corresponding classification of each word is the result output. The results showed that the learning ability of the dual LSTM network was improved and the output accuracy was increased to 98%. Jia et al. (2020) proposed an LSTM-CNN-based recognition model to build a driving behaviour recognition dataset by detecting extreme acceleration and deceleration points through statistical analysis of real car driving data. Training results showed that LSTM-CNN can achieve better results. Huang et al. (2022) proposed a CNN-LSTM network-based damage detection method for laser ultrasound guided wave scanning detection, which can detect each scanning point signal without relying on the surrounding detection point signals. The results showed that the accuracy of the method was 99.9%, 99.9%, 99.8% and 99.8% for the detection of 0.5 mm deep crack damage, penetration crack damage, corrosion damage and internal crack damage of copper tubes, respectively, and the location and size of the damage could be accurately detected. Yu and Zhang (2021) fused feature parsimony (FFR) and GACLN for model construction to obtain an FFR-GACLN model and combined it with a CNN-LSTM network approach for recognition task. The results showed that the method outperformed previous techniques. Huang and Kuo (2018) combined CNN and LSTM and applied them to a PM2.5 prediction system in order to monitor and estimate PM2.5 concentrations. The experimental results showed that the proposed CNN-LSTM model had the highest prediction accuracy among the models.

In the field of education, there is still less research on the use of spoken English chunks based on corpora, and the research on the speech recognition of the teaching system is not comprehensive. Therefore, the research will focus on the construction of spoken English chunks corpus and the spoken language recognition technology, and optimise spoken English recognition through improving CNN algorithm to improve the accuracy of spoken English recognition of the corpus system. It is expected to provide practical technical tools for oral English test activities, and reduce manual scoring errors and inefficiencies.

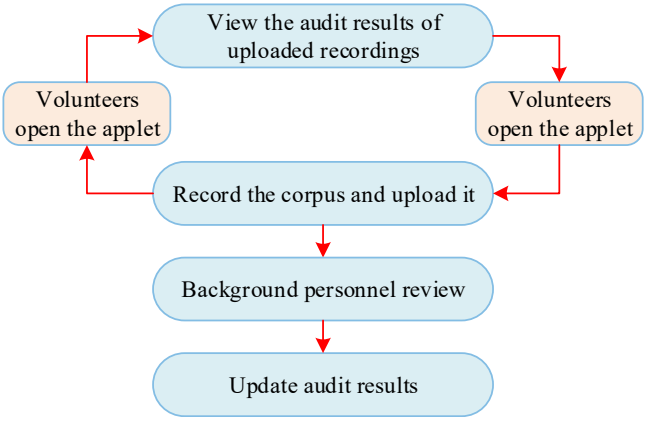
3 Research on the construction of English spoken language block corpus and spoken language recognition technology

3.1 *Approaches to the construction of a corpus of spoken English chunks*

The teaching of English mainly focuses on listening, reading and writing, while the teaching and testing of oral English are less, so most Chinese students have weak oral English ability. In order to realise the localisation of spoken English teaching, the influence of dialects in various regions of China on English pronunciation is mainly considered, and a second language learning (SELL) corpus is constructed. The corpus includes seven major dialect regions from China, which represent the main types of dialects in China. The corpus is built by inviting volunteers of different dialect including Mandarin, Cantonese, Wu, Hunan, Min, Hakka, and Gan to read prepared materials. The recording sampling rate is 16 kHz, the total duration is 31.6 h, and 389 people participate in the recording, including 186 men and 203 women. Because the volunteers who participated in the recording have lived in dialect environment without professional oral English pronunciation learning, and overseas study experience, the influence of local dialect on oral English pronunciation retains to a large extent. In addition, the corpus also adds standard documents of artificial phonemes and wrong pronunciation to evaluate the pronunciation accuracy of spoken English on the basis of dialect. With the development of technology, various network communication tools have become popular. WeChat has become one of the most used communication tools that many people use every day in China. In 2017, WeChat launched WeChat applet, which is an application that can be used on the WeChat platform without downloading and installing. Because of its portability, it is favoured by the majority of users. At present, the number of users of WeChat applet has exceeded 400 million. In this context, this paper uses WeChat apple to collect corpus data. It does not require volunteers to spend time and energy. In any place, the corpus can be recorded through small programs and uploaded to the server for the operators to review. After receiving the feedback, the volunteers can re record according to their own conditions without additional communication with the corpus builder. It effectively reduces the costs of building a spoken language corpus.

The SELL corpus constructed in this study consists of five main steps: first, text screening, in which English texts from online books are screened to obtain the initial material for the recordings; second, audio acquisition, in which audio recordings of various regional dialects are collected through a low-cost mobile applet; third, data review, in which the uploaded audio files are reviewed for validity and feedback; fourth, audio processing, in which the audio is noise reduced; fifth, Phoneme tagging, in which the uploaded audio of each dialect are manually annotated (Liao et al., 2019). The specific acquisition steps of the SELL corpus, as shown in Figure 1.

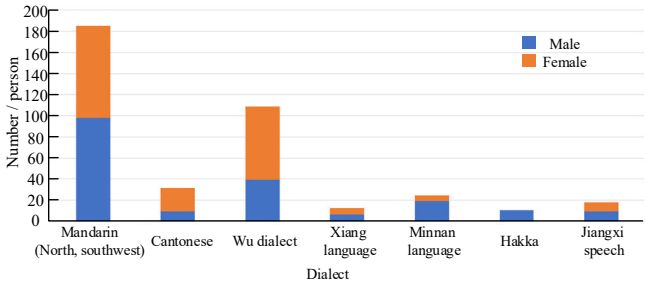
Figure 1 Specific collection steps of SSS corpus (see online version for colours)



Throughout the collection process, uploading volunteers can record the uploaded materials without the management of system personnel, which effectively reduces the cost of corpus construction. In this study, by collecting free recorded materials from the Gutenberg project, 11,000 utterances were finally selected, of which the percentages of monophthongs, diphthongs, and triphthongs were 100%, 88.90%, and 31.21%, respectively. Since the corpus of this study includes various types of dialects from a wide range of regions in China, the speech materials were recorded and collected through an applet, so that the recording of speech materials could not be unrestricted by time and place.

The SELL corpus proposed in this study consists of three main parts. First is the audio file; the second is the transcription file, in which the collected audio files are transcribed and then stored as text files after proofreading by manual means; the third is the phoneme calibration file, which includes about 1,600 manual phoneme calibration files stored in TextGrid format. Figure 2 shows the dialect acquisition of all volunteers in the SELL corpus, with a total duration of 31.6 h, including 16.7 h for males and 14.9 h for females. Each recorded audio exists has a unique number, which contains detailed information about the audio, such as the gender and dialect region of the volunteer (Yin, 2018).

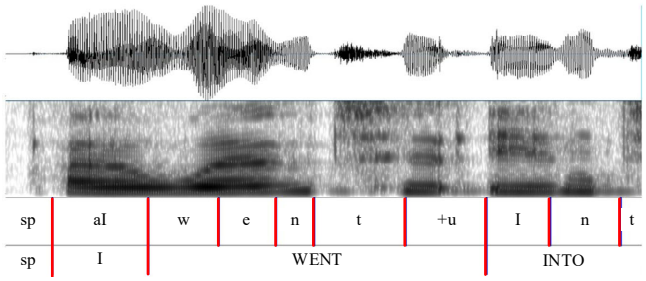
Figure 2 Statistics of corpus collection in various regions (see online version for colours)



In order to be able to evaluate the pronunciation, eight datasets are selected from seven dialect regions in this study, and each dataset has around 200 recordings, with a total of 1,600 recording files. Then, all the recordings are manually phoneme calibrated. In this study, all audio files are first transcribed and phoneme alignment is carried out

by the P2FA project. Then the audio is manually adjusted and labelled by the PRAAT project, as shown in Figure 3. In Figure 3, the red line indicates the time boundary, and the whole figure from top to bottom indicates the audio waveform, spectrogram, phoneme marker, and text content respectively.

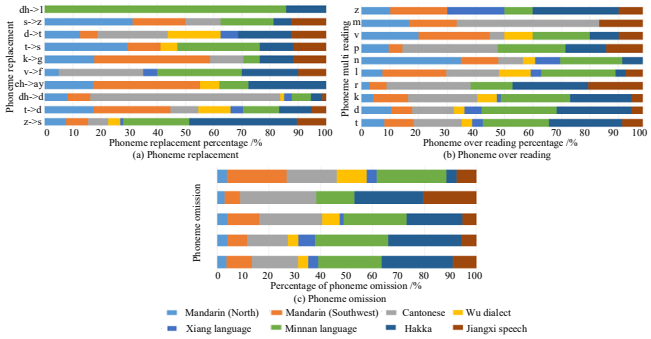
Figure 3 Screenshot of PRAAT phoneme annotation interface (see online version for colours)



When calibrating phonemes, it is not only necessary to mark phonemes that are correctly pronounced, but also to mark phonemes with Omission, overpronunciation, or mispronunciation. Omission is indicated by '–', overpronunciation is indicated by '+', and mispronunciation is indicated by '?'. The calibration content and audio are processed in a time-aligned manner and stored in TextGrid format, where the calibration file includes the specific start and end times of each phoneme, and the time unit is S.

A total of 1,600 phonetic sentences were marked with obvious phoneme errors in this study, which included 2,158 omissions, 230 overpronunciations, and 2,018 substitutions. The statistics of phoneme mispronunciation for each regional dialect are shown in Figure 4.

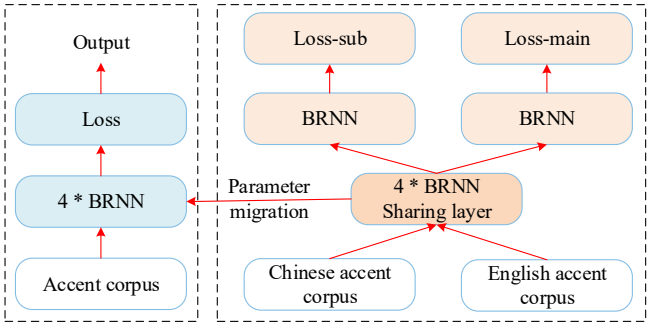
Figure 4 Statistics of phoneme error markers in various dialect regions (see online version for colours)



The common phoneme insertion errors, phoneme substitution errors, and phoneme multiple reading errors are counted in Figure 4 and ranked by the frequency of occurrence. The statistical results fully demonstrate that the dialect of English learners can have an impact on English learning. The SELL corpus includes dialect samples from many regions, which can help Chinese English learners practice oral pronunciation.

Compared with the traditional pronunciation corpus, the Chinese and English spoken data is added for multi task recognition training to identify the similarities and differences between the two languages through online autonomous learning. The structural model of multi-task learning by multi-recognition task network is shown in Figure 5.

Figure 5 Accent speech recognition network and multi task recognition network structure (see online version for colours)



(a) Accent speech recognition network (b) Multi identification task network

English spoken language recognition is the main task of the multi-recognition task network structure, followed by learning Chinese spoken language recognition. The spoken Chinese and English data are imported through the network, followed by a four-layer shared double recurrent recurrent neural network (BRNN) to obtain common features of both tasks (Zhang, 2018). Then, the two tasks are imported into the corresponding Loss layer, so that the loss function can be computed and the information parameters in the corpus can be updated. In this case, two learning tasks will get two loss layers to update the parameters respectively, and the total task of $Loss_{sum}$ is computed as shown in equation (1).

$$Loss_{sum} = \lambda Loss_{main} + (1-\lambda) Loss_{sub} \tag{1}$$

In equation (1), $Loss_{main}$ and $Loss_{sub}$ denote the hyperparameters of the primary task and secondary task, respectively, and the task weights are λ and $1 - \lambda$, respectively. In this study, the multi-recognition task network is regarded as a multi-task recognition model, and when the task recognition is finished, the obtained four layers of parameters will be imported into the accent speech recognition network, and finally the accent speech recognition network will be further trained to obtain the final recognition model.

3.2 CNN-LSTM-SA model construction for spoken English recognition

In order to improve the accuracy of spoken English recognition in the English corpus, this study introduces a CNN-LSTM-SA model. After incorporating CNN and LSTM algorithms for the recognition model, the SA mechanism is then introduced for optimisation, which can effectively improve the model signal processing and feature extraction capabilities. The attention mechanism is a model that can mimic the way of human brain attention activity, which can input and calculate the attention of a key point and calculate the impact size of the key point output model (Zheng et al., 2019). The state of a RNN neuron is derived

from the previous neuron state and the input information X of the current neuron together as the computational input, as shown in equation (2).

$$o_t = f\left(o_{t-1}, x_t\right) \tag{2}$$

In the encoding process, the hidden layer state of the last input o_t is considered as the semantic vector, and all the hidden layer states of the input sequence can be obtained from the nonlinear variation of the semantic vector C, while setting t as the number of words in the input sequence, see equation (3).

$$\begin{cases} C = o_{T_x} \\ C = q\left(o_1, o_2, o_3, \ldots, o_{T_x}\right) \end{cases} \tag{3}$$

During the decoding process, the semantic vector is converted into a sequence of specified length. The next output word y_t is identified by the semantic vector C and the output sequence o_1, o_2, o_3, …, o_t see equation (4).

$$\begin{aligned} o_t &= \arg\max P(o_t) = \prod_{t=1}^{T} p\left(o_t \left| \{o_1, o_2, \ldots, o_{t-1}\}, C\right.\right) \\ &= g\left(\{o_1, o_2, \ldots, o_{t-1}\}, C\right) \end{aligned} \tag{4}$$

Equation (4) can be transformed into equation (5) in the RNN model.

$$o_t = g\left(o_{t-1}, s_t, C\right) \tag{5}$$

where o_{t-1} denotes the input at the current moment and also the output at the previous moment; g denotes the multilayer nonlinear neural network. Attention mechanisms are usually combined with seq2seq and are able to be applied in both encoding and decoding modules. When Attention is applied in the decoding module of the seq2seq model, the conditional probability of the decoding module to identify the output by X is calculated as shown in equation (6).

$$p\left(y_t \left| \{y_1, y_2, \ldots, y_{t-1}\}, X\right.\right) = g\left(y_{t-1}, s_i, c_i\right) \tag{6}$$

In equation (6), s_i denotes the state of the hidden layer at moment t when decoding, which is calculated as shown in equation (7).

$$s_i = f\left(s_{i-1}, y_{i-1}, c_i\right) \tag{7}$$

The linguistic vector c_i corresponding to each target output value o_t has an impact on the conditional probability, while c_i is obtained from the sequence of hidden layer vectors (o_1, o_2, o_3, …, o_T) in the decoding module after summing them according to the weights, as shown in equation (8).

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} o_j \tag{8}$$

where α_{ij} denotes the attention allocation coefficient of the i^{th} output sequence for the j^{th} input sequence, and its value is determined by the hidden layer state of each input and the $i-1^{\text{th}}$ output. When the value is smaller, the influence is smaller, as shown in equation (9).

$$\alpha_{ij} = \frac{\exp\left(e_{ij}\right)}{\sum_{k=1}^{T_x} \exp\left(e_{jk}\right)} \tag{9}$$

After calculating α_{ij}, the attention allocation vector of the output at the moment i in the T_x input hidden layer states is obtained by the softmax function, which is used to calculate the weights of c_i. As shown in equation (10).

$$e_{ij} = \alpha\left(s_{i-1}, o_j\right) = v^T \tanh\left(w_s s_{i-1} + w_o o_j\right) \tag{10}$$

The RNN-based attention mechanism suffers from the long-range dependence problem and the disadvantages of recurrent neural networks. In order to extract features efficiently, reduce the computational difficulty of each layer and optimise the performance of the model, the SA mechanism is proposed. If V denotes value, K denotes key, and Q denotes query, all three vectors represent the sentence itself in the algorithm, which is obtained in the encoding by multiplying the input vector X with the weight matrix, as shown in equation (11).

$$\begin{cases} Q = W^Q X \\ K = W^K X \\ V = W^V X \end{cases} \tag{11}$$

The value of the three vectors is obtained by the above formula, then calculate scaled dot-product attention by equation (12) for

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{12}$$

In equation (12), d_k denotes the word vector dimension of K and Q. $\frac{1}{\sqrt{d_k}}$ plays a moderating role to avoid over large inner product of K and Q. The probability distribution is normalised by softmax to obtain the weights relative to V, and the weighted sum is the result of multiplying by V. V, K and Q are each linearly projected h times. Then equation (12) is calculated h times, and the results of h times calculating are used to obtain the multi-head attention mechanism as shown in equation (13).

$$\begin{cases} Multi\text{-}head = Concat\left(head_1, head_2, \ldots, head_n\right)W^O \\ Head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \end{cases} \tag{13}$$

In equation (13), the model word vector dimensions are $W_i^Q \in R^{d_m \times d_k}$, $W_i^K \in R^{d_m \times d_k}$, $W_i^V \in R^{d_m \times d_v}$, $W^O \in R^{hd_v \times d_m}$, d_m and the model word vector dimensions of V are d_v. The input mapping layer in the CNN algorithm can effectively implement the vector transformation and representation of sentence sequence information and build up the matrix, as shown in equation (14).

$$\begin{cases} S = \left(w_1, w_2, w_3, \ldots, w_n\right) \\ r^w = W^{word} * V^w \end{cases} \tag{14}$$

In equation (14), S is the input sentence sequence and n is the number of words in the sentence, r^w is the word vector, W^{word} is the word vector matrix of all word vectors, and V^w is the word encoding form. The pooling layer is mainly used to reduce the dimensionality of the extracted information to improve the fault tolerance of the model, and the maximum value of the vector is selected for feature extraction, see equation (15).

$$p_i = \max(C) = \max\left(c_1, c_2, \ldots, c_{n-Hi+1}\right) \tag{15}$$

In equation (15), p^i is a vector of fixed dimensions and C is the final computation result of the convolutional layer. In order to ensure the strong correlation of the spoken information data, LSTM network is used to improve the problem of 'insufficient long-term memory' of the traditional recurrent network. In the LSTM network recurrent cell structure, h_{t-1} and C_{t-1} are the output and structure information of the previous layer, h_t is the output, C_t is the LSTM structure information, σ is the regression function, tanh is the activation function, x_t is the input at the time of t. The input gate i, the forgetting gate f and the output gate o are calculated as shown in equation (16).

$$\begin{cases} f_t = \sigma\left(M_f \cdot \left[h_{t-1}, x_t\right] + b_f\right) \\ i_t = \sigma\left(M_i \cdot \left[h_{t-1}, x_t\right] + b_i\right) \\ o_t = \sigma\left(M_o \cdot \left[h_{t-1}, x_t\right] + b_o\right) \end{cases} \tag{16}$$

In equation (17), M, b are the weight matrix and deviation values. By means of the three kinds of gate controller states, the problem of information removal can be solved. The cell structure C_t and the implicit state h_t are calculated as shown in equation (17).

$$\begin{cases} C_t = f_t * C_{t-1} + (1 - f_t) * \left[\tanh\left(M_c \cdot \left[h_{t-1}, x_t\right] + b_c\right)\right] \\ h_t = o_t * \sigma\left(C_t\right) \end{cases} \tag{17}$$

The LSTM network counts the hidden moments and input values to determine the state of the three gates, and the forgetting and input gates are used to realise the update and information transfer to the memory unit. Then the information is hand over to the output gate to transfer the hidden information state to the next loop unit. The LSTM network can realise two-way memory information transmission and feedback, connecting up the data information and realising the integrity and selectability of data information.

Self-attentive mechanism layer is introduced into the recognition model, first calculate the key element and a certain metric correlation and derive the similarity, then normalise the similarity between the two, then calculate the feature weight coefficients, and finally weight the summed weight vector as shown in equation (18).

$$\begin{aligned} &Q = W^Q X, K = W^K X, V = W^V X \\ &Q = K = V \\ &Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \end{aligned} \tag{18}$$

In equation (18), V denotes the weight value, K denotes the metric value, and Q denotes the key element. The output layer outputs the dimensionality reduced data with full connection layer features, and the output $Y = [y_1, y_2, \ldots, y_m]^T$ of prediction step m is calculated, and the output value y_t at moment t is calculated as shown in equation (19).

$$y_t = f\left(w_o s_t + b_o\right) \tag{19}$$

In equation (19), b_o denotes the deviation vector and w_o denotes the weight matrix.

4 Performance analysis of SELL corpus and application analysis

4.1 Algorithm performance test analysis

In order to evaluate the training efficiency of the proposed algorithm, we studied and tested the changes of recognition accuracy and model loss rate under different iterations. The traditional CNN algorithm, LSTM algorithm, and the CNN LSTM algorithm in Mustaqeem (2021) are used as a comparison. The comparison results are shown in Figure 6.

According to Figure 6, the model shows a trend of slow increase after a sharp decrease in the loss rate, and 300 iterations is the inflection point. The accuracy also tends to converge around 300 iterations, when its training accuracy is close to 88%, which is significantly higher than the other two algorithms. Next, the original activation function of the algorithm was replaced with other common functions to test the performance of the activation function selected in the study, and the results were evaluated in terms of root mean square error (RMSE) and mean absolute error (MAE), as shown in Table 1, and some representative data are listed here due to the large number of experimental combinations.

Figure 6 Model accuracy and loss rate change (see online version for colours)

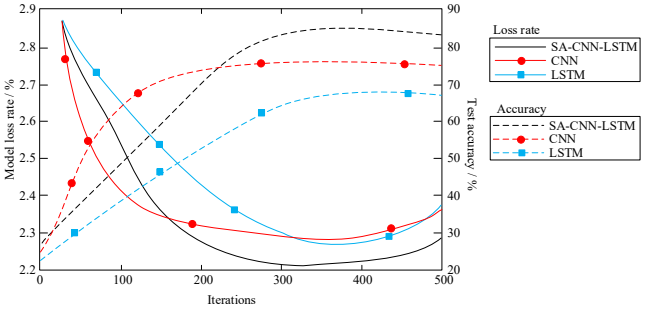
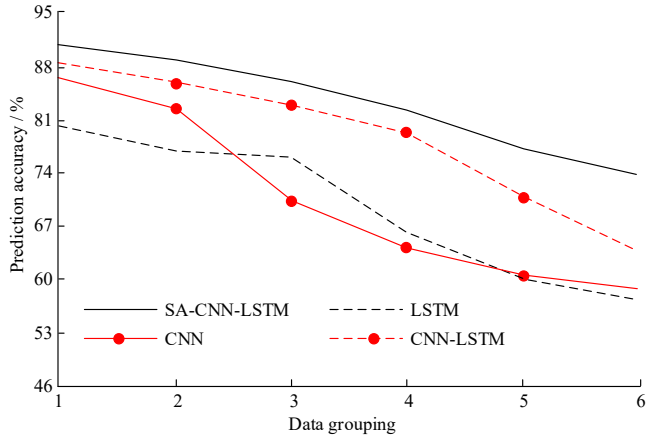


Table 1 Algorithm training results under different activation functions

Neural network activation function	*RMSE (MW)*	*MAE (MW)*
Convolution-ReLu, LSTM-tanh	35.11	25.54
Convolution-tanh, LSTM-ReLu	35.23	26.01
Convolution-ReLu, LSTM-ReLu	36.21	26.49
Convolution-sigmoid, LSTM-ReLu	40.95	31.65
Convolution-sigmoid, LSTM-tanh	41.13	31.96

Analysis of Table 1 reveals that the CNN-LSTM network performs best with the study's selected ReLu and tanh functions, with MAE and RMSE metrics of 26.54 and 36.11, respectively. The performance of the recognition algorithm in the case of the convolutional network layer with tanh, the LSTM network with ReLu and both networks with ReLu is not far from the study's original combination the minimum difference is about 0.5 for MAP and only about 0.1 for RMSE.

Figure 7 Algorithm performance under different data complexity (see online version for colours)



Considering that the recognition performance of the algorithm may fluctuate greatly with the change of complexity in the actual speech recognition process, the study selected six sets of data according to the complexity from low to high and evaluated the performance of the algorithm recognition under the condition of different data complexity, as shown in Figure 7.

From Figure 7, the prediction accuracy of all the algorithms involved in the experiment is decreasing with the increase of the complexity of the sample data, and the difference between the correct rate of the same algorithm in group 6 and group 1 is close to 15%. The performance of the recognition algorithm proposed in this study is higher than other algorithms at all six complexities, and its difference is about 4% at the lowest, with a very stable performance advantage. In order to make a visual analysis of the error situation of the algorithm proposed in the study, the model was tested in this study using the speech data recorded during one week, respectively, and multiple algorithms were used for comparison, the results of which are shown in Figure 8.

Figure 8 Comparison of model prediction errors (see online version for colours)

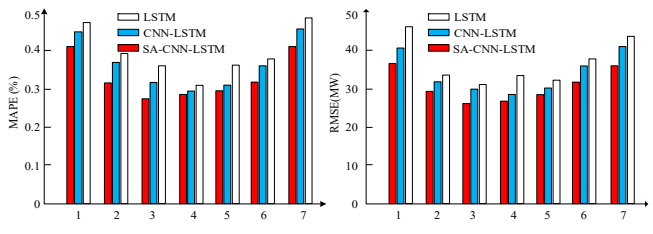


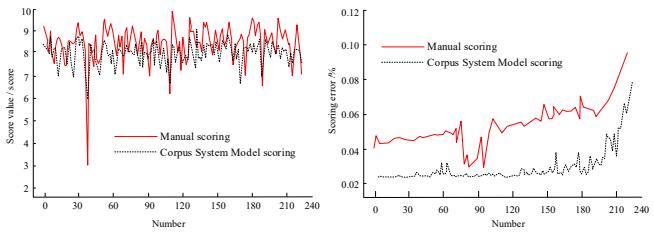
Figure 8(a) shows the MAPE analysis and Figure 8(b) shows the RMSE analysis, it can be seen that the error profile of the algorithm proposed in the study is steadily better than other algorithms, and the fluctuation of the daily prediction error is smaller, which is a considerable advantage over both the LSTM algorithm as a control and the CNN-LSTM algorithm without the self-attentive mechanism.

4.2 The effect of applying spoken blocks of English corpus

In order to further verify the application performance of the spoken English corpus proposed in this study, the model was tested and the results were compared and analysed with the original manual recognition data. The score is 1–10, and the results are shown in Figure 9.

The results in Figure 9 show that under the manual recognition scoring, the score value fluctuates widely, with the lowest value being 3 and the maximum value being close to 10, which indicates that the scoring is subjective. While the spoken recognition scoring of the corpus recognition model has less fluctuation and the score is basically between 7 and 10, and the overall recognition result is more objective and stable. The scoring error refers to the scoring error between each scoring data, and a small error indicates that the overall scoring tends to be smooth and the error rate is small. In Figure 9, the scoring errors of both the manual recognition scores and the corpus recognition model are below 10%, but the lowest scoring error of 3.01% for the manual scores is still higher than the lowest error of 2.14% for the model scoring system, which is due to the scoring errors caused by phantom listening or missing listening due to the long-time overload work intensity when facing a large number of speaking recognition tasks. The results show that the corpus speech recognition model proposed in the study can effectively reduce the error rate and improve the objectivity and accuracy of scoring. Further applying the corpus-based English-speaking blocks proposed in this study to practical teaching, the score satisfaction of 79.84%, 81.05%, 76.49% and 90.11% for word correctness, grammar usage, utterance coherence and speaking fluency were significantly better than the score without the proposed algorithm.

Figure 9 Comparison between speech recognition and artificial recognition of spoken chunks in English corpus, (a) score value (b) scoring error (see online version for colours)



5 Conclusions

In this study, a SELL corpus based on English spoken language is constructed for the study of English spoken language chunk usage features, and a CNN-LSTM-SA model-based English spoken language recognition method is proposed. The results show that the loss rate of the model proposed in this study decreases sharply and then increases slowly. The accuracy also tends to converge around 300 iterations. Its training accuracy is close to 88%, which is significantly higher than the other two algorithms. CNN-LSTM network performs best under the ReLu and tanh functions selected in the study. The MAE and RMSE indexes are 26.54 and 36.11 respectively. When the convolution network layer adopts tanh, and the LSTM network adopts ReLu, the performance of the recognition algorithm is not far from the original combination of the study. The minimum map gap is about 0.5, and the RMSE is only about 0.1. The recognition algorithm proposed in this study is higher than other algorithms in six kinds of complexity, and the lowest difference is about 4%. Compared with other algorithms, the advantage of the proposed algorithm is very stable. The error of the proposed algorithm is stable and better than other algorithms, and the fluctuation of prediction error is small. The corpus speech recognition model can effectively reduce the error rate and improve the objectivity and accuracy of scoring. This study mainly focuses on the accent pronunciation problem of Chinese students, and the corpus recognition system is designed. The recognition rate of the system for native English pronunciation is not high, so in future work, more corpus contents will be added to make the system recognise more speech sounds.

References

Cao, D. and Guo, Y. (2020) 'Algorithm research of spoken English assessment based on fuzzy measure and speech recognition technology', *International Journal of Biometrics*, Vol. 12, No. 1, pp.120–129.

Guan, X., Liu, X. and Liu, Z. (2020) 'An algorithm based on simple CNN and BI_LSTM network for Chinese word segmentation', *Journal of Physics: Conference Series*, Vol. 1621, No. 1, p.12001.

Han, R. and Yin, Y. (2021) 'Application of web embedded system and machine learning in English corpus vocabulary recognition', *Microprocessors and Microsystems*, Vol. 80, No. 2, p.103634.

He, H., Yi, S. and Liu, W. (2020) 'Intelligent English learning model based on BPTT algorithm and LSTM network', *Journal of Intelligent and Fuzzy Systems*, Vol. 39, No. 153, pp.1–12.

Huang, C.J. and Kuo, P.H. (2018) 'A deep CNN-LSTM model for particulate matter (PM2.5) forecasting in smart cities', *Sensors*, Vol. 18, No. 7, pp.2220–2241.

Huang, L., Hong, X., Yang, Z. et al. (2022) 'CNN-LSTM network-based damage detection approach for copper pipeline using laser ultrasonic scanning', *Ultrasonics*, Vol. 121, p.106685.

Jia, S., Fei, H., Li, S. et al. (2019) 'LSTM-CNN for abnormal driving behavior recognition', *IET Intelligent Transport Systems*, Vol. 14, No. 7, pp.306–312.

Liao, W.L., Lee, T.T., Jiang, W.W. et al. (2019) 'Augmented reality teaching system based on cognitive theory of multimedia learning – an example system on four-agent soup', *Applied Science and Management Research*, Vol. 6, No. 1, pp.54–69.

Liu, N. and Ren, F. (2019) 'Emotion classification using a CNN_LSTM-based model for smooth emotional synchronization of the humanoid robot REN-XIN', *PLoS One*, Vol. 14, No. 5, p.e0215216.

Liu, X., Singh, P.K. and Pavlovich, P.A. (2021) 'Accent labeling algorithm based on morphological rules and machine learning in English conversion system', *Journal of Intelligent Systems*, Vol. 30, No. 1, pp.881–892.

Mustaqeem, K.S. (2021) '1D-CNN: speech emotion recognition system using a stacked network with dilated CNN features', *Computers, Materials and Continua*, Vol. 67, No. 3, pp.4039–4059.

Rohdenburg, G. (2021) 'Ambiguity avoidance by means of function words in English? Providing additional corpus-based counterevidence', *Zeitschrift für Anglistik und Amerikanistik*, Vol. 69, No. 3, pp.207–236.

Song, J., Zhang, L., Xue, G. et al. (2021) 'Predicting hourly heating load in a district heating system based on a hybrid CNN-LSTM model', *Energy and Buildings*, Vol. 243, No. 3, p.110998.

Xie, H., Zhang, L. and Lim, C.P. (2020) 'Evolving CNN-LSTM models for time series prediction using enhanced grey wolf optimizer', *IEEE Access*, Vol. 8, pp.161519–161541.

Yin, B. (2018) 'Practical teaching reform of art design based on virtual reality technology', *IPPTA: Quarterly Journal of Indian Pulp and Paper Technical Association*, Vol. 30, No. 7, pp.703–709.

Yu, S.S. (2020) 'Spoken English repetitive correction retrieval based on syllable unit WFST web search filter', *Journal of Physics Conference Series*, Vol. 1533, No. 3, p.32010.

Yu, Y. and Zhang, M. (2021) 'Control chart recognition based on the parallel model of CNN and LSTM with GA optimization', *Expert Systems with Applications*, Vol. 185, No. 15, p.115689.

Zhang, X. (2018) 'Design and analysis of music teaching system based on virtual reality technology', *IPPTA: Quarterly Journal of Indian Pulp and Paper Technical Association*, Vol. 30, No. 5, pp.196–202.

Zhang, Y. (2021) 'Interactive intelligent teaching and automatic composition scoring system based on linear regression machine learning algorithm', *Journal of Intelligent and Fuzzy Systems*, Vol. 40, No. 2, pp.2069–2081.

Zheng, S., Ouyang, P., Song, D. et al. (2019) 'An ultra-low power binarized convolutional neural network-based speech recognition processor with on-chip self-learning', *IEEE Transactions on Circuits and Systems I: Regular Papers*, Vol. 66, No. 12, pp.4648–4661.