ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Research on Automatic Error Correction System for English Accent in Colleges and Universities Based on LSTM and Speech Recognition

Yan Zhang[1,*] *and Wanli Zhang*[2]

[1] Department of Foreign Languages, Zhongyuan Institute of Science and Technology, Zhengzhou, Henan, 451400, China
[2] Modern Education Technical Center, Zhongyuan Institute of Science and Technology, Zhengzhou, Henan, 451400, China

SUMMARY: *In this paper, based on the collection of English accent data, the Mel frequency cepstrum coefficient is used to obtain English accent signal characteristics. Then LSTM is combined with connected temporal classification network to establish the LSTM-CTC model applied to English accent recognition and error correction in colleges and universities, and the efficacy of the model is verified and scrutinized. Based on this, a DSP chip is incorporated as a foundation to develop an automated accent recognition and error rectification system for college English, and the accent recognition capabilities of the system are examined. The outcomes indicate that the phoneme error rate and word error rate of the LSTM - CTC model are 13.63% and 18.52% respectively and the system in this paper can recognize six different types of English accents. Combining deep learning with speech recognition technology can enhance the automatic recognition and error correction ability of English accent in colleges and universities, and help students better master different types of English accents.*

KEYWORDS: *mel frequency cepstrum coefficient; LSTM; connected temporal classification network; speech recognition; English accent*

1 Introduction

Pronunciation is the foundation of English learning in colleges and universities, and good pronunciation not only enables students to communicate with others more confidently, but also helps them better understand the culture and language habits of English-speaking countries [1, 2]. However, for non-native learners, there are serious accent problems in their pronunciation, which hinders students' sustainable English learning, and accent correction has become an important way for colleges and universities to improve the quality of English teaching [3-5]. Regarding the research on English accent problems of non-native learners, literature [6] examined the views of non-native English speakers on the accents of native speakers and non-native speakers, and the survey results showed that the respondents generally believed that the native speakers' accents were better than the non-native speakers' accents in terms of correctness, acceptability and pleasantness. Literature [7] discusses non-native English speakers' perceptions of accent speech and its relationship with identity, and questionnaires and interviews show that participants have greater uncertainty about the relationship between accent and socioeconomic and educational status, and that they are reluctant to demonstrate their native identity through speech with a native accent. Literature [8] explored the issue of accents of non-

*18339960512@163.com
https://doi.org/10.65102/is2026615
Published: 30 April, 2026

native English speakers and the results of the study showed that non-native speakers do not necessarily need to imitate the accents of their native speakers, rather the focus of English language teaching should be on communication between interlocutors in different contexts. Literature [9] examined the effect of non-native English speakers' accents on students in interactive listening and extensive speaking programs and the results showed that there are many challenges in what non-native English speakers say because of the difficulties in understanding the spoken language.

Currently, many computer-assisted English learning systems focus only on the textual learning of words and grammar, and relatively little on the learning of English pronunciation, and only give an approximate overall rating of the learner's pronunciation [10, 11]. In English learning in colleges and universities, an essential part is to have a timely and effective error correction feedback, which allows students to know their own deficiencies, a function that is not yet provided by various learning systems [12, 13]. In recent years, with the development of Long Short-Term memory networks (LSTM) and speech recognition technology, automatic English accent correction systems have become feasible to a certain extent. As a special recurrent neural network (RNN) structure, LSTM effectively solves the problems of vanishing and exploding gradients of traditional RNNS on long sequence tasks by introducing "memory units" and gating mechanisms [14, 15]. This makes LSTM better able to handle long sentences and complex contexts in English accent recognition, which in turn improves recognition accuracy and performance [16]. Regarding the study of LSTM and its related techniques for English accent recognition, the literature [17] describes the application of LSTM in automatic speech recognition, and reveals that the effectiveness of LSTM, with its accuracy exceeding that of other deep learning models, is revealed in experiments by applying it to spoken English. Literature [18] emphasized the practical value of recognizing and classifying English accents and discussed the LSTM-based approach for recognizing and classifying English accents, and the test results pointed out that the LSTM excelled in recognizing and classifying English accents. Literature [19] describes the optimization of LSTM in traditional English online learning platforms, which not only significantly improves the ability to convert spoken language into text and perception, but also significantly reduces the word error rate.

While speech recognition technology converts human speech into digital signals, it can automatically recognize and understand human language using computer technology to achieve human-computer interaction [20, 21]. Based on LSTM and speech recognition of college English accent automatic error correction system combines the advantages of the two technologies, in college English teaching, can be through the quality of the students' English pronunciation scoring and the existence of the accent of the effective feedback, which will help teachers to target the teaching measures, greatly improving the learning efficiency [22-24]. For the current exploration of automatic error correction systems for English accents, literature [25] proposes an ABO-LSTM-based system, which combines the ability of LSTM to capture long-term dependencies with the ability of the ABO algorithm to optimize model parameters to improve accuracy, and is able to identify and correct English speech samples with different accents. Literature [26] designs systems based on speech recognition technology and speech synthesis, which are able to assist English learners in improving their oral English development by assessing their pronunciation accuracy and providing targeted correction and training. Literature [27] designed an intelligent correction system for students' English pronunciation errors based on speech recognition technology in order to correct students' incorrect English pronunciation, it by combining the relevant hardware structure, the system is effectively applied, and based on comparative experiments reveals its practical application value.

In order to enhance the level of English accent recognition and error correction in colleges and universities, the article establishes an automatic recognition and error correction system for

English accent in colleges and universities based on LSTM and speech recognition, and explores the application effectiveness of the system. Through the effective integration of deep learning and speech recognition technology, this study realizes the accurate recognition and error correction of English accent, so as to better enhance the English accent expression level of college students.

2 Relevant theoretical and technical basis

In the context of the artificial intelligence age, the modes of English accent instruction in higher education institutions have become more diverse. Effectively leveraging technological methods to accurately identify and rectify English accents is of great significance for improving the quality of spoken English teaching in colleges and universities. As a result, comprehensively exploring the in - depth integration of technology and spoken English recognition has emerged as a vital approach to drive the high - quality development of spoken English in higher education.

2.1 Speech Recognition Framework and MFCC Algorithm

2.1.1 Speech recognition technology framework

Speech recognition, as one of the research centers in natural language technology, aims to enable machines to understand and respond correctly to natural speech as humans do. Its essence is to convert the input speech signal into corresponding computer commands after a series of decoding and recognition. Among them, the acoustic model plays the role of the core hub in speech recognition, how to quickly train a robust, high-precision acoustic model is the focus of this paper. Figure 1 depicts the compositional structure of speech recognition. This structure primarily consists of a feature extraction unit, an acoustic model unit, a pronunciation lexicon, a language model, and a decoding component.

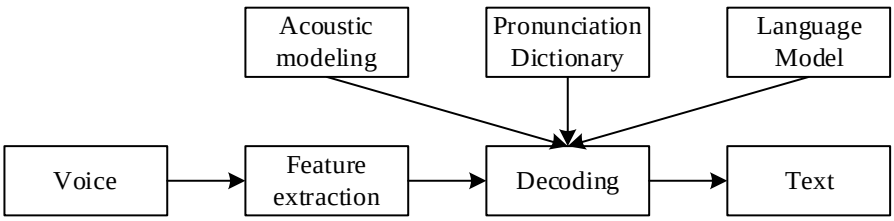


Figure 1: The overall framework of speech recognition

(1) Feature extraction. Extract features from the input speech and turn the speech into vectors for processing by the acoustic model.

(2) Acoustic modeling. Used to recognize speech vectors.

(3) Pronunciation Dictionary. Includes the pronunciation of words and can map words to a sequence of phonemes (or words). The pronunciation lexicon provides support for the link between the acoustic model and the language model.

(4) Language Model. The language model models the language targeted by the system.

(5) Decoding. For the input signal, the maximum probability word string is found based on the acoustic and language models and the lexicon.

If X is a feature vector (MFCC 40 dimensional) and W represents a word sequence, find the most probable word sequence, i.e.:

$$W^* = \arg\max_W P(W \mid X) \tag{1}$$

According to Bayes' formula:

$$P(W \mid X) = \frac{P(X \mid W)P(W)}{P(X)} \tag{2}$$

can be approximated:

$$W^* = \arg\max_W P(X \mid W)P(W) \tag{3}$$

where X and W are the observation vectors of the training samples and their corresponding word sequences, respectively, and $P(W \mid X)$ denotes the posterior probability used in Bayesian, which can be viewed as $P(W \mid X)$ maximal W^* as the output of speech recognition. The $P(X \mid W)$ can be obtained from the acoustic model for probability, which is used to indicate the degree of match between the acoustic features and the word sequence W. The $P(W)$ is computed from the language model, and indicates the probability of occurrence of the word sequence W.

2.1.2 Mel frequency cepstrum coefficients

Mel cepstrum coefficient (MFCC) is a cepstrum parameter extracted in the frequency domain of Mel scale, which describes the nonlinear characteristics of the human ear frequency, i.e., to analyze the audio spectrum based on the results of human hearing experiments, and to obtain Mel frequency cepstrum coefficients which are easy to be analyzed through a series of data conversions of the audio signals.

(1) Audio signals show great differences between macroscopic and microscopic aspects, which are reflected in macroscopic instability and microscopic smoothness. In this case, partially consecutive n samples $\left\{x_j, x_{j+1}, \cdots, x_{j+n-1}\right\}$ are merged into a single frame $Chunk_i$, with the time covered by each frame being $T_i \in [20,30]$ms. In order to avoid too large a difference in values between neighboring frames, a suitable frame shift k is set to solve this problem with a certain number of coincident sampling points, and the distribution of sampling points for two neighboring frames can be expressed as respectively:

$$Chunk_i = \left\{x_j, x_{j+1}, \cdots, x_{j+n-1}\right\} \tag{4}$$

$$Chunk_{i+1} = \left\{x_{j+n-1-k}, x_{j+n-k}, \cdots, x_{j+2n-1}\right\} \tag{5}$$

(2) Audio windowing. For each frame signal $Chunk_i$ after frame splitting, windowing is performed through the windowing function $\delta(n)$ to avoid the influence of high-frequency components by increasing the attenuation of high-frequency components in the audio signal, and reduce the risk of spectral energy leakage, so as to get the time-domain signal of the music $w_i(n)$, which is made possible through windowing that each frame of audio is mapped in a segment of the spectrum, i.e:

$$w_i(n) = \delta(n) * Chunk_i \tag{6}$$

(3) Fast Fourier Transform (FFT) is used to transform the signal after adding windows and splitting frames to obtain the Fourier spectrum, $w_i(n)$ is the ith frame of the input signal $w(n)$, $Chunk_i$ is the number of the ith frame of the sampling points, the FFT can be expressed as follows:

$$P = \frac{\left| FFT\left(w_i(n)\right)^2 \right|}{Chunk_i} \tag{7}$$

After processing by FFT, The signal in the time domain is transformed into a signal in the frequency domain, filtering the influence of higher than the highest frequency in the sampled signal, and at the same time realizing the dimensionality reduction.

(4) The energy spectrum is defined in a filter bank with M filters by means of a triangular filter bank, and the center frequency is set to be $f(m), m = 1, 2, 3\cdots, M$, and the size of each region varies with the value of m, and the definition of the triangular filter is denoted as:

$$H_m(k) = \begin{cases} 0, k < f(m-1) \\ \dfrac{2(k - f(m-1))}{(f(m+1) - f(m-1))(f(m) - f(m-1))}, k(m-1) \leqslant k \leqslant f(m) \\ \dfrac{2(f(m+1) - k)}{(f(m+1) - f(m-1))(f(m) - f(m-1))}, k(m) \leqslant k \leqslant f(m+1) \\ 0, k \geqslant f(m+1) \end{cases} \tag{8}$$

By means of a triangular filter, the spectrum is smoothed, the resonance peaks in the original signal are strengthened, and harmonic elimination is achieved with fewer operations.

(5) The actual frequency f is mapped into the Mel frequency $Mel(f)$ by the filter to achieve a unification shift in frequency. Logarithmic operations are performed on all filter outputs to obtain the logarithmic spectrum, and then the values of the spectral line energies are obtained. I.e:

$$Mel(f) = 2595 * \log_{10}(1 + f / 700) \tag{9}$$

(6) Discrete Cosine Transform (DCT) is performed on the results of each logarithmic operation. Since filters are usually overlapped, the filter energies are correlated with each other and the DCT requires a de-correlation operation on the energies. The Mel inverted spectral coefficient characteristic MFCC coefficients of the audio are finally obtained, then:

$$MFCC(i, n) = \sqrt{\frac{2}{M}} \sum_{m=0}^{M-1} \log[S(i, m)] \cos \frac{\pi n(2m-1)}{2M} \tag{10}$$

where i is the audio signal, m denotes the number of filters, and n is the DOC (discrete cosine first conversion) spectrum.

2.2 Deep Learning Related Technologies for Speech Recognition

2.2.1 Long and short-term memory neural networks

The fundamental architecture of a Long Short - Term Memory (LSTM) neural network features

distinct components within the hidden layer known as memory blocks. A typical LSTM configuration comprises cells, input gates, and output gates. To overcome its constraints, forgetting gates are incorporated. These forgetting gates are responsible for modifying the state of the LSTM. Forgetting gates eliminate the stored inputs by resetting the cell variables. Meanwhile, input and output gates control the inflow and outflow of data. The gating mechanism governs the functioning of the memory blocks. The forgetting gate assigns weights to the information inside the cell, enabling the state to be reset when the information is of little significance. Moreover, the forgetting gate facilitates continuous prediction and curtails the bias in the prediction process.

The computational operations within the LSTM block are such that the input values can be saved in the cell state only if the input gate allows it. The input value i_t of its input gate and the expected value $\tilde{C}_t$ of the memory cell at time step t are computed as follows:

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right) \tag{11}$$

$$C_t = \tanh\left(W_c\left[h_{t-1}, x_t\right] + b_c\right) \tag{12}$$

where $W\left[h_{t-1}, x_t\right]$ and b denote the weight matrix and bias, respectively. The forgetting gate controls the weights of the state cell cells and the value of the forgetting gate is calculated as follows:

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t + b_f\right)\right. \tag{13}$$

With this procedure, the new state of the storage unit is updated to:

$$\tilde{C}_t = i_t.\tilde{C}_t + f_t \cdot \tilde{C}_{t-1} \tag{14}$$

Given a new state storage cell, the output value of the gate is computed as:

$$o_t = \sigma\left(W_x \cdot \left[h_{t-1} \times X_t\right] + b_o\right) \tag{15}$$

The final output value of the unit can be interpreted as:

$$h_t = O_t * \tanh\left(c_t\right) \tag{16}$$

where σ, g and h are point-by-point nonlinear activation functions, and i, f, o and c are the input gates, forgetting gates, output gates and cellular activation vectors, respectively. All features of the LSTM network architecture can be trained using the $sig-moid(\varphi)$ and tanh activation functions. With this architecture, the network can store inputs for long periods of time, thus utilizing the trained series of extended time cases.

2.2.2 Connecting temporal classification networks

Connected Timing Classification (CTC) is a temporal classification method.CTC differs from traditional DNN-HMM based acoustic models in that it does not require frame-level alignment of labels in the temporal dimension, and the input of speech features is sufficient for predicting the results.The process of reducing the lossy values of the CTC through training and thus

reducing the differences between the predicted values and the true labels greatly simplifies the training process of acoustic models. It is important to note that CTC additionally introduces blank labels for modeling silence, inter-word overlap, etc., which simplifies the modeling process, and thus CTC is particularly suitable for sequence modeling.

Let the given sequence $X=\left(x_1,x_2,\cdots,x_T\right)$ denote the input T frames of speech features, and the prediction of each frame outputted by the neural network is $Y=\left(y_1,y_2,\cdots,y_{T'}\right)$. Due to the pooling function in the CNN, the length of the sequence is made exponentially shorter $T=nT'$, and n is the number of times that the feature map is reduced after the pooling calculation, where:

$$y_i=\left(y_I^1,y_i^2,\cdots,y_i^k,\cdots,y_i^m\right) \tag{17}$$

where m is the total number of modeling units and y_i^k is the k th modeling unit position in the i th frame. Then the posterior probability that the k th modeling unit at the moment of a given input sequence X,t is output by the neural network SoftMax function is:

$$P(k\mid t,X)=\frac{\exp\left(y_t^k\right)}{\sum_{k'}\exp\left(y_t^{k'}\right)} \tag{18}$$

From the above equation, the probability distribution of the corresponding modeling units in T' frames is obtained sequentially as:

$$P(\pi\mid X)=\prod_{t=1}^{T'}P\left(\pi_t\mid t,X\right) \tag{19}$$

where π is the path that generates the sequence of predicted T', and the probability of the corresponding path π is obtained by accumulating it; since π and y is a many-to-one relationship, and ψ is the conversion function of the path to the prediction, the probability of obtaining the sequence of the path's corresponding prediction is expressed as:

$$P(y\mid X)=\sum_{\pi\in\psi^{-1}(y)}P(\pi\mid X) \tag{20}$$

Eq. (21) obtains the final CTC loss function given the real label y^*, and the CTC loss value is continuously reduced through training so that the prediction results are gradually approximated towards the real label. That is:

$$CTC(X)=-\lg\left(P\left(y^*\mid X\right)\right) \tag{21}$$

Currently, there are three main types of CTC decoding, i.e., maximal path decoding, prefix bundle decoding, and bundle decoding. Maximum path decoding aims to find the label corresponding to the first $z(z\leqslant m,m$ is the number of modeling units) path with the largest probability for each path, without the need for a priori knowledge such as dictionaries, language models, etc., and the decoding process is extremely simple, and its computational process is as follows:

$$\pi^* = \operatorname{Arg}\max_{\pi}(P(\pi \mid X)), z \leqslant m \tag{22}$$

$$y' \approx \psi\left(\pi^*\right) \tag{23}$$

y' is the final decoding result.

3 Intelligent Error Correction Model for English Accents

Speech recognition, a newly emerging technology that combines multiple disciplines, has found extensive applications across numerous sectors. Regarding English accent recognition in higher education institutions, this paper develops an intelligent English accent recognition and error rectification model founded on LSTM - CTC. The primary objective of this model is to significantly enhance the English accent proficiency of college and university students.

3.1 Accent Data Collection and Feature Extraction

3.1.1 English Accent Collection

The model undergoes training and testing with LibriSpeech, a dataset that encompasses 900 hours of spoken English. The training, development, and testing subsets are utilized. These subsets serve two main purposes: first, to fine - tune the model's hyper - parameters, and second, to assess the model's efficacy. LibriSpeech contains voiced speech, which has more standardized pronunciation rules, and has reference value for testing the recognition ability of the model, but it cannot fully represent spoken speech. In order to make up for the insufficiency of the dataset and enhance the model's recognition ability of spoken language, we collected English spoken speech data by ourselves. We also recorded natural conversations with different English accents via Skype and Discriminator software to obtain 90 hours of spoken English speech data, covering a variety of spoken expressions and topics. The recording sampling rate is 16kHz, the quantization bit is 16bit, and it is saved in .wav format, and the speech data is labeled with the purpose of enabling the computer to understand the content of the speech data. In the English spoken speech recognition system, text conversion is generally used to annotate the speech data.

The English accent dataset obtained in this study encompasses six main categories, namely AM (American accent), AU (Australian accent), BR (British accent), CA (Canadian accent), EU (European accent), and IN (Indian accent). In total, there are 25,192 speech data samples. To comprehensively analyze the experimental outcomes, the aforementioned dataset is partitioned into a training set and a test set at a ratio of 7:3. This division aims to validate the efficacy of error correction in English accent recognition.

3.1.2 Signal Feature Extraction

The Mel Frequency Cepstrum Coefficient (MFCC), a correlation coefficient employed in speech recognition and speaker identification, is commonly utilized to depict speech signals. Consequently, it was adopted for the feature extraction of English accents. First, the English accent signal data from the test subjects were gathered. Then, this data underwent pre - emphasis, frame splitting, and windowing processes. After that, the processed speech signal was examined in the time domain.

Let the speech signal be $x(n)$, then the linear spectrum $X(k)$ can be expressed as:

$$X(k)=\sum_{n=0}^{N-1}x(n)e^{-j2\pi nk/N} \tag{24}$$

Let $H_p(k)$ be the p th Mel-frequency filter, use the Mel-frequency filter for the linear spectrum $X(k)$, and logarithmize the resulting Mel-frequency spectrum:

$$S(p)=\ln\left(\sum_{k=0}^{N-1}|X(k)|^2\ H_p(k)\right) \tag{25}$$

The logarithmic energy output from each filter bank is processed by the discrete cosine transform, and the final, Mel-frequency cepstrum coefficient eigenparameters $c(n)$ are:

$$c(n)=\sum_{p=1}^{p-1}S(p)\cos\left(\frac{\pi n(p+1/2)}{P}\right) \tag{26}$$

3.2 Recognition and Error Correction Model for English Accents

3.2.1 LSTM-CTC modeling

In this paper, LSTM network and CTC model are combined together to construct an end-to-end English accent recognition error correction model. The LSTM-CTC based speech recognition model is given in Figure 2. The language data used in this paper are without text, and the model input timings are per-frame speech features, and the output timings are International Phonetic Alphabet (IPA). A series of pre-processing is first performed on the experimental audio data, then the spectrum of speech is analyzed and relevant features are extracted, followed by modeling the long sequences using LSTM network to fully exploit the contextual information. During the decoding phase, as Connectionist Temporal Classification (CTC) can be considered an objective function capable of directly optimizing the probability of the input sequence and the output target sequence, under this objective function, CTC autonomously learns and refines the relationship between the input and output sequences throughout the training process.The output layer of the CTC network is a SoftMax layer, and the quantity of nodes is equivalent to that of the labeled sequences. Moreover, the blank nodes are crucial in addressing the issue of overlapping words in the labeling.

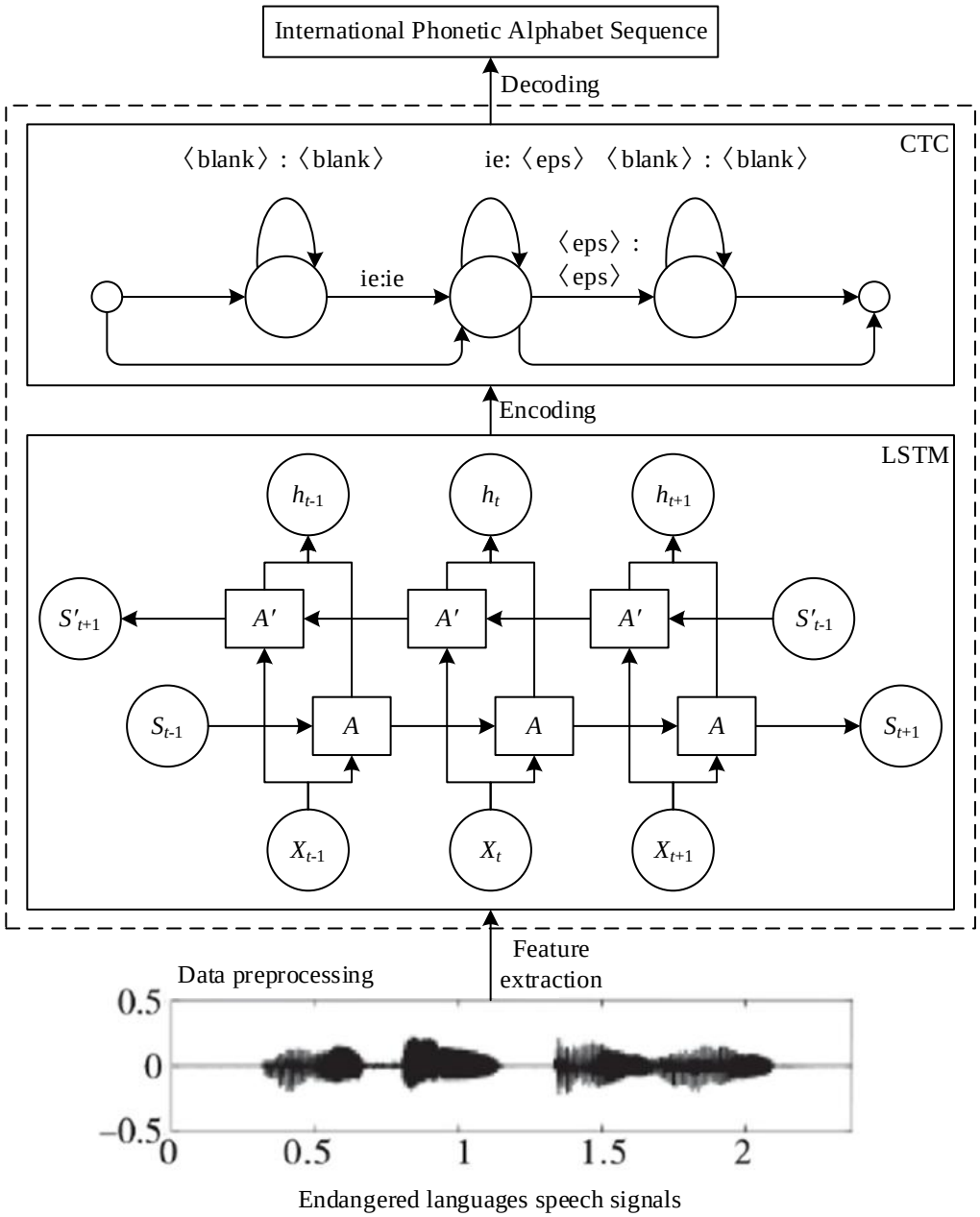


Figure 2: A recognition and error correction model for English accents

Assigning a label to a single frame of data in speech recognition poses a challenge. However, when dealing with tens of frames, it becomes straightforward to identify the corresponding pronunciation label. In CTC networks, the existence of the blank node enables the frame - skipping method. The output of the CTC and the label adhere to the following equivalence, namely:

$$F(i-ie-) = F(-ii-ie) = iie \tag{27}$$

where "i" and "e" are international phonetic symbols "-" is blank.

Evidently, numerous output sequences can be mapped to a single output. As a result, CTC not only accelerates the decoding rate but also autonomously refines the correspondence between input and output sequences during the training phase.

3.2.2 Model Identification Process

In this project, Python language is used for programming, and the environment of TensorFlow is built on Ubuntu, and then the LSTM-CTC acoustic model is constructed for the recognition of English accent signal. Based on the continuous audio LSTM-CTC of English accent signal process, in this paper, the acquired English accent original sound is pre-processed, followed by the extraction of MFCC feature parameters, firstly, the MFCC feature parameters are input into the pre-trained LSTM layer with 128 LSTM units, and then the features of the English accent signals are learnt by using the long term memory units of the LSTM and the gating mechanism, and then the MFCC feature parameters are input into the pre-trained LSTM layer, and the MFCC feature parameters are output into the pre-trained LSTM layer. two hidden layers, which correspond to the probability distribution of the output possible labels, are trained using the Adam optimizer. The probability matrix is used as an input to the CTC model, and the labeling of the training set is used to obtain the loss of the model. The training of the LSTM-CTC acoustic model for the whole continuous audio is completed after the above steps.

The training process of LSTM-CTC model is as follows:

(1) Convert the sampling rate of the labeled English accent signal dataset in wav format to 15000Hz.

(2) Produce the location text and labeled text for each wav audio in the English accent signal dataset and put them in the same folder.

(3) Download the pre-training model, change the English accent categories in the category file in the model to five categories from 0 to 5, and CTC as the loss function of LSTM.

(4) Set the maximum number of iterations in training to 500, the number of samples in one training is 64, and the initial learning rate is 0.0005.

(5) Start training.

3.3 Validation of the English accent recognition model

3.3.1 Model Loss Curve

For comparison with the LSTM - CTC model, CNN - CTC and RNN - CTC are selected as baseline models. For all three models, the number of network layers is configured to 3. To avoid overfitting, Dropout is employed. The weights are initialized using random orthogonal initialization, and the learning rate is set at 0.005.To make it easier to observe the decrease in the loss value, this paper arranges for the loss value to be saved once for every 500 speech data samples during training. The resulting decline curve of the model's loss value is presented in Fig. 3.From the figure, when comparing the three models, the CNN - CTC model exhibits the fastest convergence rate. The LSTM - CTC model also shows a rapid decrease in the loss value at the start. However, after a certain period, it tends to decline steadily, and the loss value at the final convergence is lower than that of the CNN - CTC model.In the initial stage, the RNN - CTC model has a quicker decline in the loss value compared to the other two. Nevertheless, after some time, the decline becomes extremely slow. Moreover, the loss value of the LSTM - CTC model decreases more slowly than that of the CNN - CTC model. At this point, the LSTM - CTC and CNN - CTC models have reached convergence, while the loss value of the RNN - CTC model remains relatively high, indicating that the training is not yet complete.Overall, the LSTM - CTC model designed in this paper demonstrates better overall performance, and the model training is more stable.

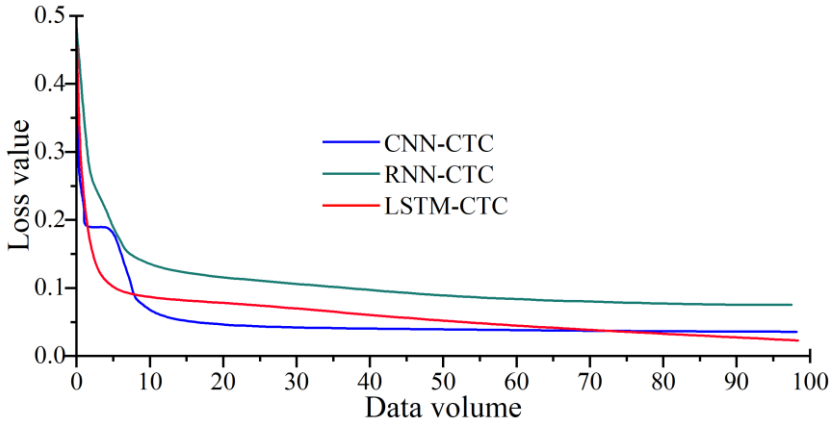
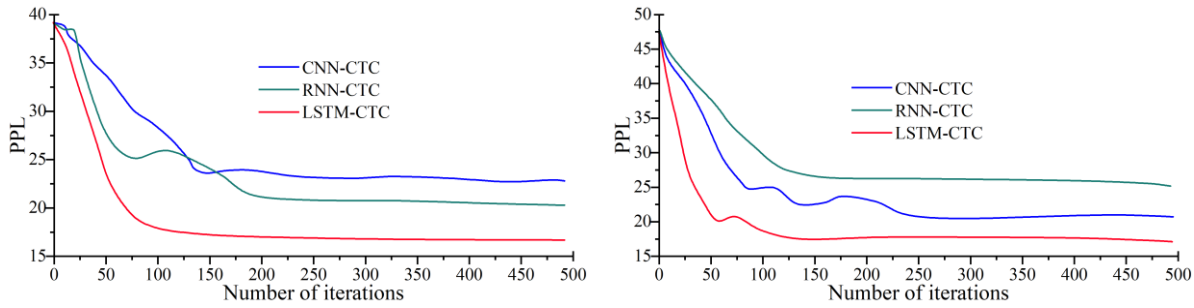


Figure 3: Loss value decline curve

3.3.2 Accent Recognition Performance

In the process of model testing and training, the LibriSpeech dataset and the self-built Oral English dataset are selected for the training of the English accent recognition model, and 70% of both datasets are randomly selected as the training set and 30% as the test set, and the language model perplexity (PPL) and accuracy are used as the evaluation criteria. CNN-CTC model and RNN-CTC model are added as experimental comparisons to compare and analyze the performance of the LSTM-CTC model designed in this paper. The PPL comparisons under different models are shown in Fig. 4, where Fig. 4(a)~(b) shows the PPL comparison results on different datasets.

As can be seen from the figure, with the increase of iteration number, the PPL values of the three models are decreased, in which the training effect of LSTM-CTC model in the two datasets is similar, and it tends to be stabilized around 150 iterations to achieve the effect that the PPL value is as low as 17.5, and the fluctuation of the change is smaller and more stable.The training effect of CNN-CTC model in the Oral English dataset is is better, reaching the lowest PPL value after 250 iterations (22), and the training effect of the RNN-CTC model in the two datasets is more different, with the lowest iteration and PPL values of about 150 and 27, respectively. It can be seen that the LSTM-CTC model improves the data adaptation ability, enhances the applicability of the language model and optimizes its performance.

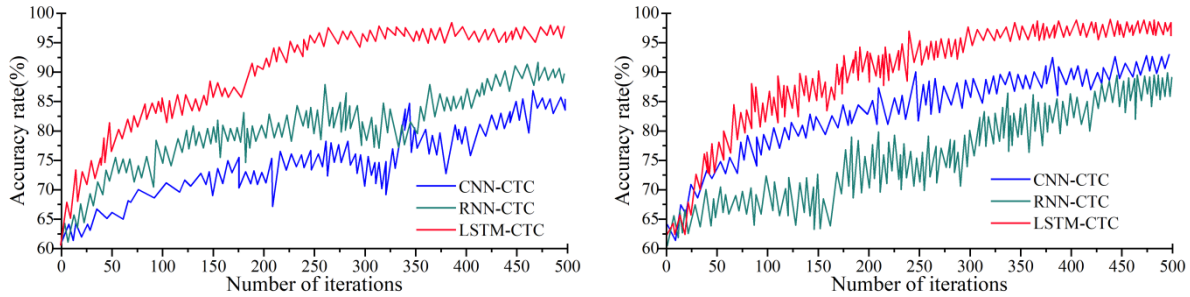


(a) LibriSpeech Dataset (b) Oral English Dataset

Figure 4: Comparison results of PPL on different datasets

The accuracy comparison results of English accent recognition under different models are shown in Fig. 5, where Fig. 5(a)~(b) shows the accuracy comparison results on different datasets. As can be seen from the figure, the accuracy rate of LSTM-CTC model is the highest during training and testing, in which the effect is better in the training and testing of Oral

English dataset, and the accuracy rate can reach the highest (about 98%) at about 300 iterations, and the accuracy rate grows faster.The training of the RNN-CTC model in LibriSpeech dataset is better, and the accuracy rate reaches the highest (about 98%) at about 4,000 iterations. The RNN-CTC model trained on the LibriSpeech dataset has a better effect, and the accuracy rate stabilizes after 450 iterations and stays at a high level of 90%, while the CNN-CTC model has a maximum accuracy rate of about 88%, which is about 10% lower compared to the LSTM-CTC model. In terms of overall stability, the accuracy of the LSTM-CTC model fluctuates less and the model is more solid during the model run. It can be seen that the LSTM-CTC model reduces the error of English accent recognition, improves the stability of the English accent recognition model, and helps to guarantee the accuracy of English accent recognition.



(a) LibriSpeech Dataset (b) Oral English Dataset

Figure 5: Comparison results of Accuracy rate on different datasets

3.3.3 Model performance analysis

To further demonstrate the error - rectifying capacity of the LSTM - CTC model in English accent recognition, this paper selects the phoneme error rate (PER) and word error rate (WER) as assessment indicators. These are then used to make comparisons with the GMM - HMM, DNN - HMM, CNN - CTC, and RNN - CTC models.Moreover, to examine the decoding speed of these models for English accents, this paper also conducts tests on the decoding time of the models, specifically the disparity in real - time rate (RTF).The outcomes of the comparison regarding the error - correction performance of different models in English accent recognition are presented in Figure 6.

As can be seen from the figure, the phoneme error rate and word error rate of the English accent recognition error correction model based on LSTM-CTC are 13.63% and 18.52%, respectively, which are 65.84% and 43.99% lower in the phoneme error rate and 66.66% and 56.79% lower in the word error rate, respectively, when compared with the traditional GMM-HMM and DNN-HMM models. And compared with the deep learning neural network CNN-CTC and RNN-CTC models, this paper's model obtains superior results for both phoneme error rate and word error rate. The experimental results show that the LSTM-based acoustic model, after applying the CTC training criterion, fully exploits the modeling ability of recurrent neural networks on sequence data, which is significantly better than the GMM and DNN in terms of model representation ability, and also makes the English accent recognition error correction model free from unreasonable HMM conditional assumptions. In addition, the real-time rate of the LSTM-CTC model reaches 9.88s, which possesses a higher English accent decoding rate compared to the individual comparison models. This indicates that the end-to-end acoustic recognition model effectively reduces the English accent recognition error rate while also effectively optimizes the model structure and obtains huge savings in storage space and decoding time.

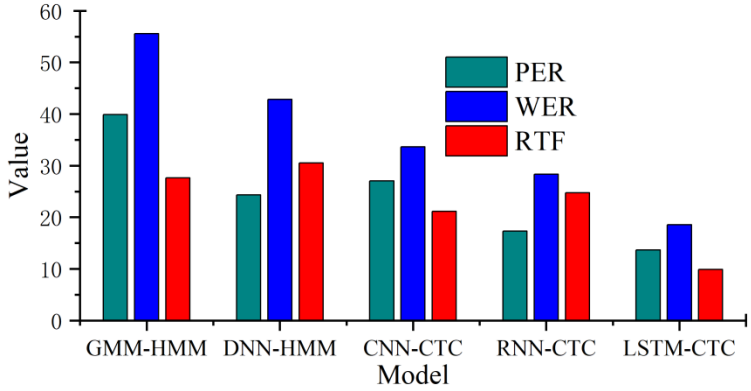


Figure 6: Performance comparison results of different models

4 English Accent Automatic Error Correction System

The improvement of speech recognition technology can realize the high accuracy of microphone speech input to text, thus enhancing its application value in various fields, and the technical research of English accent recognition can improve the accuracy of speech recognition within a certain range. This chapter mainly designs an automatic English accent recognition system based on DSP chip, and applies the LSTM-CTC model designed in the previous section to this system, so as to help college English speaking learners master a more standard English accent.

4.1 System hardware and software design and development

4.1.1 System hardware design

The hardware structure of the entire English accent automatic identification and error correction system is shown in Figure 7, including TMS320 DSP, crystal, power supply, JTAG interface, keyboard, LCD display module, data memory, program memory, audio codec chip, wireless transceiver module. The core processing chip adopts Y company TMS320VC5509 chip, the chip has the advantages of low power consumption, high speed rate, high cost-effective and so on, it is widely used in the field of voice processing, portable equipment terminal signal processing and so on. Voice codec chip TLV320AIC23B, using advanced Sigma-Delta oversampling technology, can provide 16-bit, 32-bit and 64-bit sampling within the range of 10 ~ 100kHz sampling rate, ADC and DAC signal-to-noise ratio of up to 80dB and 120dB, respectively, the chip has been widely used in a variety of audio signal processing field. The wireless part adopts the PTR2000 module designed based on the nRF401 wireless communication chip, which has low transmit power consumption, high sensitivity and simple peripheral interface, and is the ideal choice for low-power radio transmission at present.

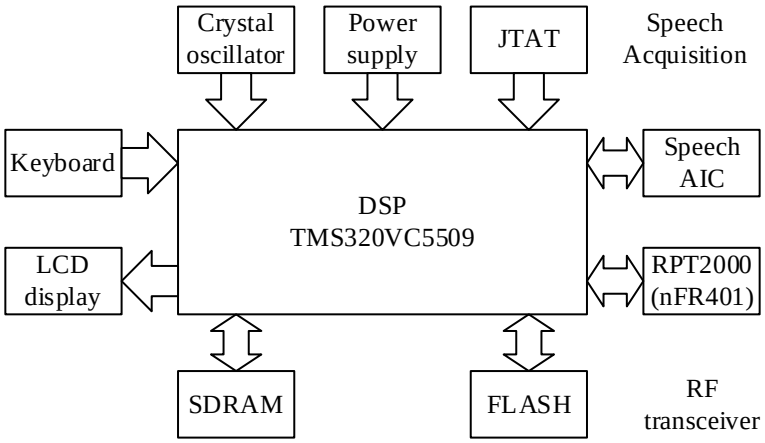


Figure 7: Hardware structure of the system

After the system is powered on, firstly, the tester passes the voice signal to the audio codec chip TLV320AIC23B by means of microphone or linear input, and transmits it to the system core processing chip (DSP) through the MCBSP serial port. Then, the DSP analyzes the voice control signal after receiving it using the LSTM-CTC model to recognize the voice control command. Finally, the system passes part of the recognized results to the LCD display module for observation and the other part to the wireless transceiver module to realize remote interaction with the smart device. In addition, the keyboard can query the working status of the system and control the display circuit to display it, in order to monitor the working of the system in real time.

4.1.2 System software design

The software of this paper is developed using the DSP integrated development environment code debugger (CCS) with Windows 10 operating system provided by Y Company. The system software based on LSTM-CTC model has four main modules such as recognition, training, learning and USB. At startup, the system plays a voice prompt after initializing the settings through the system, and then detects whether there is a key press through keyboard scanning. When a key is pressed, the system will prompt for mode selection, and after the selection is confirmed, the system will enter the cycle of training mode, learning mode and USB mode, and the system will enter the corresponding mode according to the different keys. When there is no button pressed, it will enter the recognition mode by default.

4.1.3 System development and deployment

The server side of the platform is developed in Java, which is used as the main web development language. The system uses Spring Boot framework for rapid iterative development, which inherits the excellent genes of Spring, and at the same time can simplify coding, simplify configuration, and simplify deployment. The components in Spring Cloud and Spring Cloud Alibaba are used in the system architecture to provide a distributed solution for the system. Data access and storage, using the mainstream relational database MySQL, as well as non-relational database Redis to support data caching, while through the Redis cluster to improve the high availability of the cache server and the concurrency of the system, through the ElasticSearch to support full-text search. Table 1 shows the main development environment information of the system.

Table 1: The system mainly develops environment information

Type	Entry	Development tools and versions
Programming language	Front-end language	NPM, Promise, ES6, TypeScript
	Back-end language	Java2.0, SQL, Shell
Development environment	Operating system	Windows10, CentOS7.8
	Server	Docker, Tomcat
	JDK	JDK2.0
	Database	MySQL9.0
	Cache	Redis5.4
	Message middleware	RabbitMQ3.8
Tools	Back-end development tools	IntelliJ IDEA 2.5
	Front-end development tools	VC Code
	Project management tools	GIT, Maven3.8

4.2 Testing of the English Accent Correction System

4.2.1 Accent Recognition Performance

The English accent recognition and error correction system designed in this paper is based on the LSTM-CTC model designed in the previous paper, and its purpose is to enhance the recognition effect of English accent, and the recognition rate is an effective index to characterize the performance of the system. Based on this, the article selects three kinds of fixed-length speech for testing according to the English accent usage scenarios, which are 8s, 16s and 32s, and for each kind of fixed-length speech, multiple words are chosen as the word length. Figure 8 shows the time consumption of English accent fixed-length speech. Based on the results in the figure, it can be seen that in most of the English accent usage scenarios, the accent recognition speed of the English accent recognition and error correction system designed in this paper has not exceeded 900ms, and the overall English accent recognition speed of the system is better, and it can complete the recognition and error correction of English accent beyond the user's perception. Overall, thanks to the LSTM-CTC model, it can be close to real-time in short-time English accent recognition and error correction, provide a seamless and smooth interactive experience of English accent recognition, and provide reliable recognition results for timely correction of possible English accent pronunciation problems.

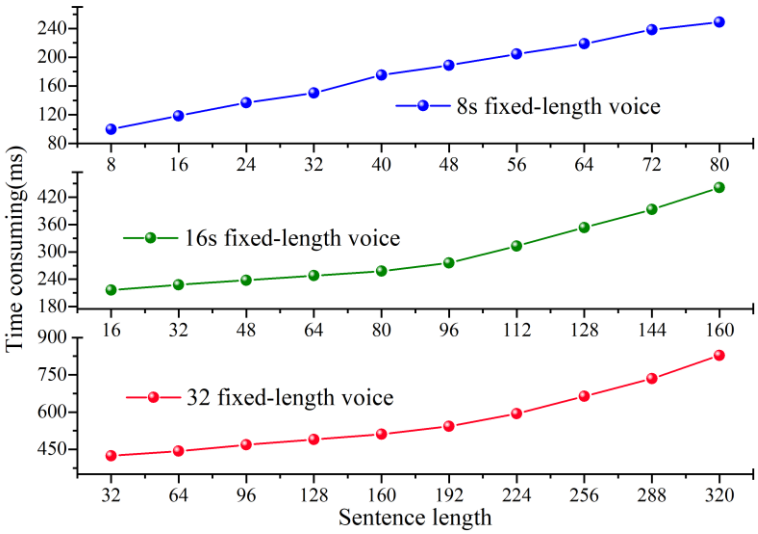


Figure 8: The English accent is long and time-consuming

4.2.2 Examples of system applications

Based on the English accent recognition error correction system established in this paper, the homemade English accent dataset established in the previous paper is utilized to verify the English accent recognition results of the system through the confusion matrix. Figure 9 shows the English accent recognition confusion matrix of the system. From the recognition results of the confusion matrix, the English accent recognition and error correction system is quite accurate in recognizing and correcting the prediction results for each type of English accent, and the checking rate for each type of English accent exceeds about 85%, and the checking rate for American accent (AM) exceeds 90%. Combined with the comprehensive F1 calculation results, the experimental results for each type of English accent do not have serious deviations, and there is no situation in which the F1 value of a certain type of English accent is very low, which indicates that the system in this paper has high robustness. Therefore, the combination of LSTM model and CTC network to establish LSTM-CTC model applied to English accent recognition and error correction has high applicability, and the established English accent recognition and error correction system can effectively recognize different types of English accents. The application of this system to the learning of English accent correction in colleges and universities can effectively help students to distinguish the pronunciation of different types of English accents, so as to better correct the accent pronunciation problems of students in the learning process.

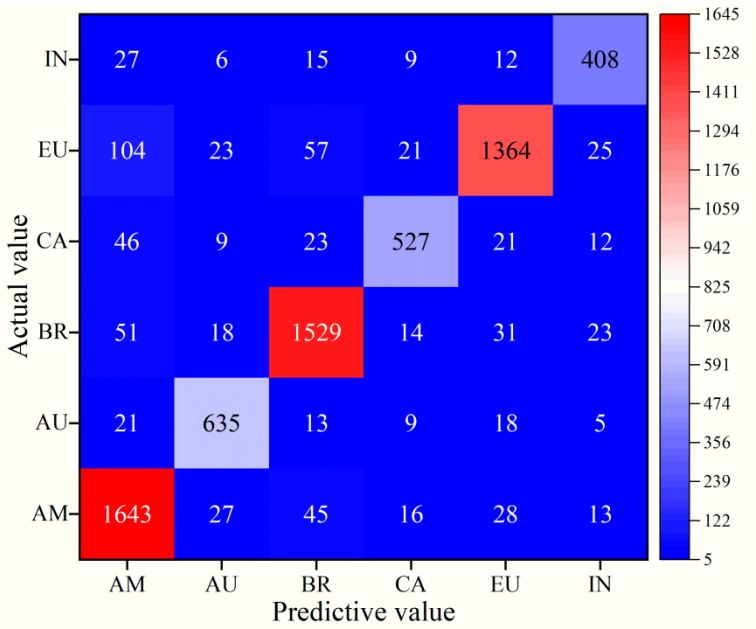


Figure 9: English accent recognition confusion matrix

5 Conclusion

The article establishes an automatic English accent recognition and error correction model for colleges and universities based on LSTM, develops an English accent recognition system based on DSP, and verifies and analyzes its performance. The results show that the LSTM-CTC model has better overall English accent recognition performance, with lower phoneme error rate, word error rate and decoding time than the comparison model. When the system is applied to different types of English accent recognition, it has a higher checking rate of English accent recognition, which can provide help to ensure the correction of errors in English accent recognition in

colleges and universities, so that students can better grasp the pronunciation defects in English accent recognition and improve their English pronunciation level. This study in the achievement of research results at the same time, there is a practical application of insufficient defects, in the follow-up research, the system in this paper will be applied to college students in English accent teaching in order to test its effectiveness.

Funding

Project Leader 1: Henan Provincial Federation of Social Sciences' 2022 Research Project (No: SKL-2022-1489) (Completed).

Project Leader 2: Brand Project of Industry-Education Integration in 2023 of Undergraduate Colleges.

References

[1] Djurayeva, Y. A. (2021). Enhancing English pronunciation in learning process. Academic research in educational sciences, 2(CSPI conference 2), 302-306.

[2] Sardegna, V. G., Lee, J., & Kusey, C. (2018). Self-efficacy, attitudes, and choice of strategies for English pronunciation learning. Language Learning, 68(1), 83-114.

[3] Bloem, J., Wieling, M., & Nerbonne, J. (2016). Automatically identifying characteristic features of non-native English accents. Future Dialects, 155-73.

[4] Dirham, U. R. (2022). English as a lingua franca: Perceptions of Indonesian non-native English-speaking teachers (NNESTs) on English pronunciation and accents identity. Scope: Journal of English Language Teaching, 7(1), 105-114.

[5] Vančová, H. (2019). Current issues in pronunciation teaching to non-native learners of English. Journal of Language and Cultural Education, 7(2), 140-155.

[6] Kaur, P., & Raman, A. (2014). Exploring native speaker and non-native speaker accents: The English as a Lingua Franca perspective. Procedia-Social and Behavioral Sciences, 155, 253-259.

[7] Tamimi Sa'd, S. H. (2018). Learners' views of (non) native speaker status, accent, and identity: an English as an international language perspective. Journal of World Languages, 5(1), 1-22.

[8] Kong, M. L., & Kang, H. I. (2022). Identity and accents: Do students really want to speak like native speakers of English?. RELC journal, 53(3), 505-518.

[9] Saipullah, H. M., Syahri, I., & Susanti, R. (2021). STUDENTS'PERCEPTIONS ON THE ACCENTS OF NON-NATIVE ENGLISH SPEAKERS IN INTERACTIVE LISTENING AND EXTENSIVE SPEAKING CLASS. English Community Journal, 5(1), 1-9.

[10] Choi, S. K., Kwon, O. W., & Kim, Y. K. (2017). Computer-Assisted English Learning System Based on Free Conversation by Topic. Research-publishing. net.

[11] Wang, C., Zhu, S., & Zhang, H. (2023). Computer-assisted English learning: Uncovering

the relationship between motivation and self-regulation. Journal of Computer Assisted Learning, 39(6), 1860-1873.

[12] Khansir, A. A., & Pakdel, F. (2018). Place of error correction in English language teaching. Educational Process: International Journal, 7(3), 189-199.

[13] Jing, H., Xiaodong, H., & Yu, L. (2016). Error Correction in Oral Classroom English Teaching. English Language Teaching, 9(12), 98-103.

[14] Van Houdt, G., Mosquera, C., & Nápoles, G. (2020). A review on the long short-term memory model. Artificial intelligence review, 53(8), 5929-5955.

[15] Palangi, H., Deng, L., Shen, Y., Gao, J., He, X., Chen, J., ... & Ward, R. (2016). Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. IEEE/ACM transactions on audio, speech, and language processing, 24(4), 694-707.

[16] Jiao, Y., Tu, M., Berisha, V., & Liss, J. M. (2016, September). Accent Identification by Combining Deep Neural Networks and Recurrent Neural Networks Trained on Long and Short Term Features. In Interspeech (pp. 2388-2392).

[17] Oruh, J., Viriri, S., & Adegun, A. (2022). Long short-term memory recurrent neural network for automatic speech recognition. IEEE Access, 10, 30069-30079.

[18] Ke, W. (2023). Study on recognition and classification of English accents using deep learning algorithms. Journal of Intelligent Systems, 32(1), 20230174.

[19] Orosoo, M., Raash, N., Treve, M., Lahza, H. F. M., Alshammry, N., Ramesh, J. V. N., & Rengarajan, M. (2025). Transforming English language learning: Advanced speech recognition with MLP-LSTM for personalized education. Alexandria Engineering Journal, 111, 21-32.

[20] Jiang, J., & Wang, H. H. (2021). Application intelligent search and recommendation system based on speech recognition technology. International Journal of Speech Technology, 24(1), 23-30.

[21] Shadiev, R., & Liu, J. (2023). Review of research on applications of speech recognition technology to assist language learning. ReCALL, 35(1), 74-88.

[22] Zhou, J., Su, S., & Li, R. (2024, October). Foreign language speech recognition and correction system aided by machine learning algorithm. In 2024 3rd International Conference on Data Analytics, Computing and Artificial Intelligence (ICDACAI) (pp. 851-856). IEEE.

[23] Zhang, F., & Sun, J. (2022). Multi-Feature Intelligent Oral English Error Correction Based on Few-Shot Learning Technology. Computational Intelligence and Neuroscience, 2022(1), 2501693.

[24] Bashori, M., van Hout, R., Strik, H., & Cucchiarini, C. (2024). I Can Speak: improving English pronunciation through automatic speech recognition-based language learning systems. Innovation in Language Learning and Teaching, 18(5), 443-461.

[25] Mayormente, M. D. (2024, December). Employing Deep Learning to Create Speech Recognition Systems for Accented English. In 2024 International BIT Conference (BITCON) (pp. 1-5). IEEE.

[26] Jing, W. (2024). Speech recognition sensors and artificial intelligence automatic evaluation application in English oral correction system. Measurement: Sensors, 32, 101070.

[27] Dai, M. (2022). Intelligent Correction System of Students' English Pronunciation Errors Based on Speech Recognition Technology. Journal of Information & Knowledge Management, 21(Supp02), 2240013.