Construction of intelligent evaluation model for Topic Talk in the Mandarin proficiency test

Caihua Chen[a], Xinyuan Long[b*]
[a] Information Institute, Hunan Open University, Changsha 410007, Hunan, China; [b] Hunan Weina Space Information & Technology Ltd., Changsha 410205, Hunan, China

ABSTRACT

In order to solve the problem that the Topic Talk in the Mandarin proficiency test is still unable to achieve machine evaluation, an intelligent evaluation model for Topic Talk is constructed in this paper. First, according to the requirements of the intelligent evaluation model, a large vocabulary continuous speech recognition module is designed in the front of the model. Then the recognition module is optimized from the two aspects of the acoustic model and the language model. Finally, the experiment is performed on the dataset which is collected from the Mandarin Proficiency Test Center. The experimental results show that the speech recognition module constructed with the improved DBLSTM-HMM acoustic model and n-gram+RNN interpolation language model can better complete the recognition task of front part in the intelligent evaluation model. The improved acoustic model can reduce the word error rate and sentence error rate to 14.08% and 18.22%. The improved language model can increase the word recognition rate by 5.87%, and the correlation between the posterior probability and the manual score by 3.82%.

Keywords: Topic Talk, intelligent evaluation model, acoustic model, language model

1. INTRODUCTION

PSC (Putonghua Shuiping Ceshi) is the only authoritative test for the grading of Mandarin proficiency in China. The form of the test is an oral examination, and the test content is divided into four questions. The first three questions are reading the given characters, words, and articles respectively, which belong to the text-dependent oral evaluation. The first three questions have been successfully scored by machine, and the machine scoring is very close to the manual scoring. The fourth question belongs to Topic Talk. According to the topic determined by the lottery, candidates can express freely and improvise within a time limit of four minutes. Since the content expressed by the candidates cannot be accurately known in advance, the Topic Talk is a text-independent oral evaluation task. Limited to the existing research work on text-independent oral evaluation, manual scoring is still used for this task at present, which undoubtedly increases the manpower and workload of PSC.

Log posterior probability is recognized as the most important feature to measure pronunciation quality in oral evaluation tasks. It has been widely and maturely applied in the text-dependent oral evaluation task of PSC. Ge et al.[1] proposed a posterior probability algorithm based on phoneme confusion expansion network, which can significantly improve the calculation speed of posterior probability without changing the computational complexity of the system. Chen et al.[2] introduced the knowledge of Mandarin pronunciation linguistics into the posterior probability algorithm, and improved the pronunciation quality evaluation algorithm from the perspective of the phoneme scoring model, which can significantly improve the correlation between manual scoring and machine scoring. Chen et al.[3] optimized the probability space of the current computer-aided PSC system from the perspective of phonetics, which not only reduced the confusion caused by the probability space, but also significantly shortened the computing time of the system. Since there is no pre-set reference text in the Topic Talk evaluation of PSC, in order to realize the intelligent evaluation of this item, it is necessary to perform continuous speech flow recognition with a recognizer in advance, and take the optimal recognition result as the reference text of log posteriori probability. In this paper, we introduce a deep neural network to construct an intelligent evaluation model for the Topic Talk in PSC.

[*] 2322588796@qq.com

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 125061V
© 2022 SPIE · 0277-786X · doi: 10.1117/12.2662890

2. INTELLIGENT EVALUATION MODEL OF TOPIC TALK

The total score of Topic Talk item in PSC is 40 points. According to the requirements of the PSC official examination syllabus, the assessors should make a comprehensive evaluation from four aspects: the level of pronunciation standard (25 points), the degree of vocabulary and grammar standard (10 points), natural fluency (5 points), and the appropriate deduction of points if the effective time of expression is less than 3 minutes[4]. This item not only assesses the degree of standard and fluency of candidates' pronunciation, but also pays attention to the use of vocabulary and grammar. In the Topic Talk evaluation, since the machine cannot know the text expressed by the candidate in advance, the method of forcibly aligning the text with the model in the text-dependent evaluation task cannot be directly adopted. Therefore, it is necessary to add a complete large vocabulary continuous speech recognition module (LVCSR) at the front end of the intelligent evaluation model of Topic Talk item. The intelligent evaluation process of Topic Talk in Mandarin proficiency test is shown in Figure 1.

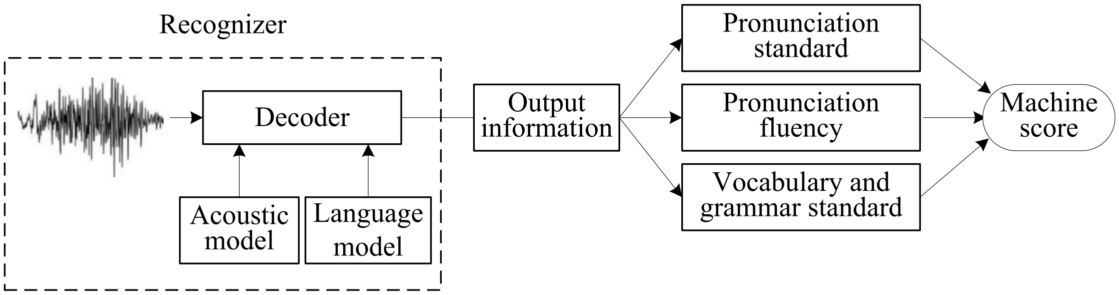

Figure 1. Intelligent evaluation process of Topic Talk items.

We input the speech of the candidate's Topic Talk item into the decoder after preprocessing and feature extraction. Based on the acoustic model and the language model, the decoder can obtain an optimal output sequence by using Viterbi algorithm. The output of the recognizer is used as the reference text for the subsequent text-independent evaluation tasks. First, the posterior probability of the obtained reference text relative to the pronunciation quality is calculated, and then the features of various evaluation indicators such as pronunciation standard degree, fluency degree, and vocabulary and grammar standard are calculated based on the posterior probability. Finally the total machine score is predicted based on all the obtained feature values.

For text-independent pronunciation quality evaluation, it mainly depends on the posterior probability of the recognition result relative to the pronunciation quality. Therefore, improving the recognition rate of the front-end recognizer is especially important for the posterior probability estimation of the Topic Talk item. The acoustic model and the language model has a great impact on the recognition performance of the front-end recognizer. Since the acoustic model and the language model have a great impact on the recognition performance of the front-end recognizer, in this paper, we mainly optimize the LVSCR module from the two aspects of the acoustic model and the language model to improve the speech recognition rate of the front-end, so as to calculate the posterior probability more accurately and better measure the pronunciation quality of candidates.

3. IMPROVED DBLSTM-HMM ACOUSTIC MODEL

Since Hidden Markov Model (HMM) can describe the relationship between the hidden state and feature sequence in speech information, it is widely used in acoustic modeling in speech recognition tasks. The traditional Gaussian Mixture Model-Hidden Markov Model (GMM-HMM)[5] has the advantages of fast training speed and easy transplantation, but it lacks the learning ability of deep nonlinear feature transformation. The leading Deep Neural Network-Hidden Markov Model (DNN-HMM)[6] can make better use of the relevant information between speech frames and has the ability to learn deep nonlinear feature transformations, but it lacks the ability to model the long-term correlation of speech. The Recurrent Neural Network-Hidden Markov Model (RNN-HMM)[7] can better solve the modeling problem of long-term correlation information of speech, but as the number of network layers increases, problems such as gradient disappearance or explosion are prone to occur. The Long Short-Term Memory-Hidden Markov Model (LSTM-HMM)[8] can effectively control the gradient disappearance or explosion problem in RNN, but like RNN, the memory of LSTM is unidirectional, when modeling the current moment, only historical information can be used, and future information cannot be introduced.

3.1 Network structure of DBLSTM

In order to realize the simultaneous modeling of front and rear bidirectional information, make up for the defect of unidirectional modeling in LSTM-HMM, and further improve the recognition rate, in this paper, we build a Deep Bidirectional Long Short-Term Memory model (DBLSTM) based on LSTM, as shown in Figure 2.

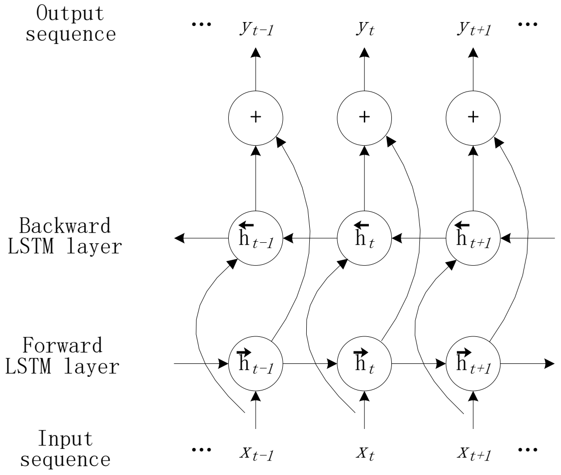

Figure 2. Intelligent evaluation process of Topic Talk items.

We input the input sequence x into the forward LSTM layer and the backward LSTM layer at the same time. The forward hidden layer vector $\overrightarrow{h}$ is iteratively calculated from front to back through the forward layer, and the backward hidden layer vector $\overleftarrow{h}$ is iteratively calculated from back to front through the backward layer. We update the output sequence $y=\{y_1,y_2,....,y_M\}$ based on two LSTM layers, where M is the number of output data. The iterative formula of the DBLSTM network is given as follows:

$$\overrightarrow{h_t} = H\left(W_{x\vec{h}} x_t + W_{\vec{h}\vec{h}} \vec{h}_{t-1} + b_{\vec{h}}\right) \tag{1}$$

$$\overleftarrow{h_t} = H\left(W_{x\overleftarrow{h}} x_t + W_{\overleftarrow{h}\overleftarrow{h}} \overleftarrow{h}_{t-1} + b_{\overleftarrow{h}}\right) \tag{2}$$

$$y_t = W_{\vec{h}y} \vec{h}_t + W_{\overleftarrow{h}y} \overleftarrow{h}_t + b_y \tag{3}$$

3.2 Improved DBLSTM acoustic model

With the deepening of Deep Neural Network, the gradient disappearance of DBLSTM will occur in both time domain and space domain[9]. We use the gate signal of linear cyclic connection in DBLSTM to deal with the gradient disappearance problem in the time domain. By introducing maxout neural network into DBLSTM to increase the depth of DBLSTM, the gradient disappearance problem in DBLSTM spatial domain can be solved by using maxout neuron to generate a constant gradient. In order to solve the problem of model over fitting in the process of DBLSTM training, we introduce dropout regularization algorithm. The improved DBLSTM depth mixing acoustic model proposed in this paper is shown in Figure 3.

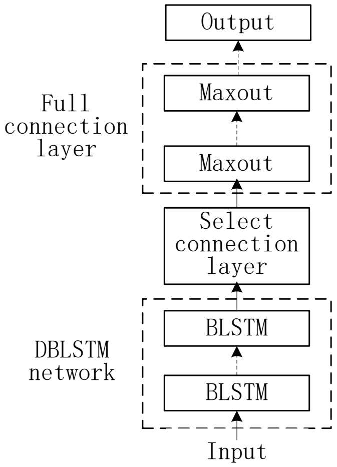

Figure 3. Depth mixed acoustic model of DBLSTM + Maxout.

The bottom layer of BLSTM in the figure mainly models the long-term related information of the input speech signal. The middle select connection layer transforms the output data of BLSTM network according to equation (3) and transmits it to the full connection layer. The maxout neurons in the full connection layer are regularized and trained according to the dropout algorithm, and finally enter the softmax output layer to output the results.

3.2.1 Maxout Neural Network. The maxout neural network structure in the full connection layer is shown in Figure 4.

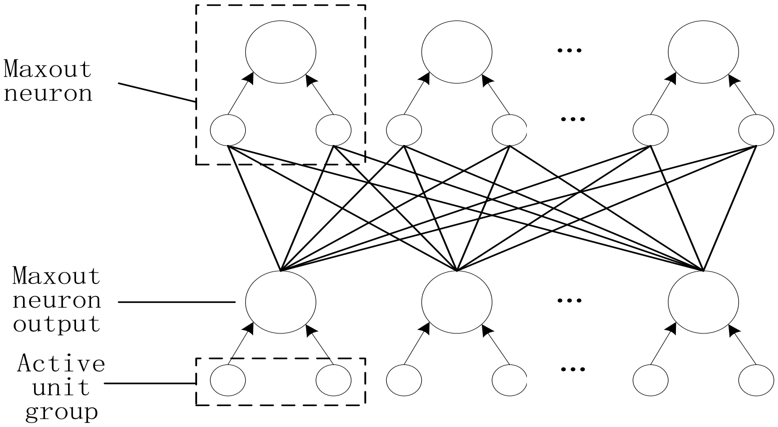

Figure 4. Neural network structure of Maxout.

The activation unit group of Maxout neuron contains multiple optional activation units. We select the maximum value from the activation unit group as the output of Maxout according to equation (4):

$$h_l^i = \max_{j \in 1,\cdots,k} z_l^{ij} \tag{4}$$

where h_l^i represents the output of the i-th Maxout neuron in the *l*-th layer, and k is the number of active units in the active unit group. z_l^{ij} represents the j-th activation unit of the *i*-th Maxout neuron in the *l*-th layer, which is obtained from the forward LSTM propagation layer of the previous layer:

$$z_l = W_l^T h_{l-1} + b_l \tag{5}$$

where W_l^T is the weight matrix from the neuron to the active unit z_l in the previous layer, and b_l is the bias vector. Maxout solves the gradient disappearance problem by generating a constant gradient during training. The gradient of maxout neuron is:

$$\frac{\partial h_l^i}{\partial z_l^{ij}} = \begin{cases} 1, & z_l^{ij} \geq z_l^{is}, \forall s \in 1,...,k \\ 0, & \text{others} \end{cases} \tag{6}$$

When the value of the active unit is the largest, the gradient of the Maxout neuron is 1, and the other is 0.

3.2.2 Dropout algorithm. In order to avoid the over-fitting problem of DBLSTM network when there are few training *samples*, we introduce Dropout regularization algorithm. In the iterative process of the neural network, the weight update of the hidden layer nodes is set to prevent the over-fitting phenomenon of the network, that is, the weights of some hidden layer nodes are not updated in one iteration, but *are* activated and updated in the next iteration. Different regularization methods are used in the training and testing stages of the Dropout algorithm[10].

(1) Training stage

We use a binary mask m_l in the original activation unit and calculate the output value of the Maxout neuron as follows:

$$h_l = m_l \cdot \theta\left(W_l^T h_{l-1} + b_l\right) \tag{7}$$

Among them, θ represents the nonlinear transformation operation of neurons, m_l obeys the $1-r$ Bernoulli distribution, and r is the Dropout rate. The selection of the r value is very important in the network training process. The smaller the r value is, the more useful information is retained. The larger the r value, the higher the degree of regularization.

(2) Testing stage

In the test stage of dropout, it is not necessary to omit the activated neurons, but in order to compensate for Dropout training, it is necessary to reduce the activation value of neurons by $1-r$. In order to prevent Dropout regularization from damaging the long short term information of DBLSTM network[11], the Maxout network is only applied to the full connection layer.

4. N-GRAM+RNN INTERPOLATION LANGUAGE MODEL

The intelligent evaluation of text-independent Topic Talk item depends greatly on the recognition rate of the front-end recognizer. If the decoding result of the front-end recognizer is wrong, the posterior probability calculated based on the recognition result will hardly provide useful information for the evaluation of pronunciation quality. A good language model can effectively improve the decoding efficiency of the recognizer, thereby improving the speech recognition rate.

The existing LVCSR recognizers generally use the n-gram language model based on statistics[12], which can only memorize the historical information of the previous two to three words, and the further historical information has no effect on the score of the current word. It is obviously that the lack of historical information reduces the reliability of n-gram language model scores. The language model based on the RNN[13] can introduce further sentence history information in the training process, but the decoding efficiency of whose is not high. In order to improve the front-end speech recognition rate, we integrate the advantages of the two language models and perform interpolation operations on the n-gram language model and the RNN language model[14, 15]. Firstly, we use the n-gram language model to obtain the one-pass decoding result of the decoder, and then use the RNN language model to re-estimate the score of the N-best candidate result[16] decoded in one pass, and finally take the sentence with the highest re-estimated score as the new recognition result. The process is shown in Figure 5.

For the N-best candidate result decoded by the decoder in one pass with the n-gram language model, we keep its acoustic model score (AC Score) unchanged, and apply RNN to re-estimate the language model score (LM Score). Based on the original acoustic model score and the re-estimated language model score, we obtain a new score for each candidate sentence, and select the candidate sentence with the highest score as the new speech recognition result.

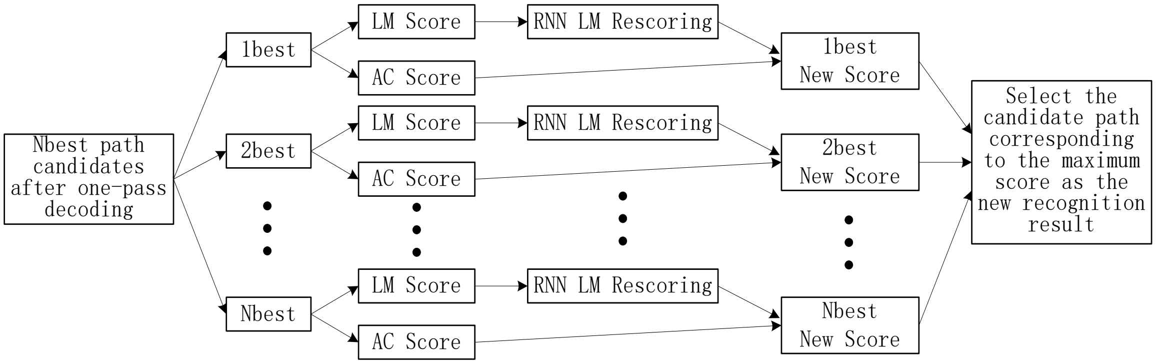

Figure 5. Process of RNN rescoring identification result.

The language model score calculation formula for re-estimating the *i*th candidate sentence in the RNN language model is as follows:

$$Score_i = AcScore_i + W_i \otimes C + \left[\lambda \cdot Lm^i_{ngram} + (1-\lambda) \cdot Lm^i_{RNN}\right] \cdot LmScale \tag{8}$$

Among them, $AcScore_i$ is the acoustic model score of the *i*-th candidate sentence, which remains unchanged during the re-estimation of the language model score by using n-gram and RNN interpolation, W_i is the number of words in the *i*-th candidate sentence, C is the word penalty, λ is the interpolation coefficient, Lm^i_{ngram} is the n-gram language model score, Lm^i_{RNN} is the RNN language model score, and $LmScale$ is the scaling factor of the language model score during decoding.

In the following text-dependent evaluation tasks, the reference text updated by the RNN can not only reduce the recognition errors caused by the language model, but also make the calculated posterior probability more conducive to the evaluation of pronunciation quality.

5. POSTERIOR PROBABILITY ESTIMATION

After the one-pass decoding result of the candidate's Topic Talk item is obtained through the front-end recognizer, the posterior probability of the decoded phoneme t is calculated as follows:

$$pp(t \mid O) = \frac{1}{N}\ln\frac{p(O \mid t)\,p(t)}{\sum_{q \in Q_t} p(O \mid q)\,p(q)} \approx \frac{1}{N}\ln\frac{p(O \mid t)}{\sum_{q \in Q_t} p(O \mid q)} \tag{9}$$

Among them, $O = [o_1, o_2, ..., o_N]$ is the acoustic observation vector corresponding to the decoded phoneme t. We assume that the prior probability $p(q)$ of all phonemes is equal. In order to make the calculation of the posterior probability more targeted, the posterior probability denominator space Q_t is composed of the error prone phonemes of the decoded phoneme t [2.

For the decoded phoneme t, if the optimal path obtained by Viterbi decoding is $\Theta = \{s_1, s_2, ..., s_N\}$, then $\ln p(O \mid t)$ can be approximately calculated as follows:

$$\ln p(O \mid t) = \sum_{i=1}^{N} \ln p(o_i \mid s_i) \tag{10}$$

It is assumed that as long as the transition probability a_{ij} of the HMM is greater than zero, the jump from state i to state j can be completed, that is, the likelihood score calculation can ignore a_{ij}. For the DBLSTM-HMM acoustic model, $p(o_i \mid s_i)$ can be calculated as follows:

$$p(o_i \mid s_i) = \frac{p(s_i \mid o_i) p(o_i)}{p(s_i)} \tag{11}$$

Among them, $p(s_i)$ is the prior probability of each HMM state obtained from the training set, $p(o_i)$ can be regarded as a constant during the decoding process, and $p(s_i \mid o_i)$ is the score of the neural network output corresponding to the state s_i after the softmax activation operation. The likelihood score of the DBLSTM-HMM acoustic model is given by using the following formula:

$$\ln p(O \mid t) = \sum_{i=1}^{N} \ln \frac{p(s_i \mid o_i)}{p(s_i)} \tag{12}$$

First, we estimate the posterior probability of each phoneme in the decoded phoneme according to equation (9), then calculate the average value of the posterior probability of all phonemes in a sentence, and finally calculate the average value of the posterior probability of all sentences in a speech, which is the final estimate of the posterior probability of the speech.

6. EXPERIMENTAL VERIFICATIONS

6.1 Experimental dataset

(1) Acoustic model dataset

We collected the recorded speech data of the Topic Talk item from the PSC Centers across the country, and selected about 800 hours of speech data as the training dataset for the experimental acoustic model. The pronunciation level of all selected speech data is good, with scores above 80. For the candidates who passed the PSC and scored between 60 and 100, we randomly selected about 20 hours of speech data of the Topic Talk item as the testing dataset to verify the recognition rate. All the data in the dataset is the real test data of the candidates, without artificial noise reduction, the sampling frequency is 16 KHz, and the quantization is 16 bit.

(2) Language model dataset

The language model corpus required for the experiment is obtained by manually converting the speech of the candidate’s Topic Talk item into text. We collected about 480,000 sentences in total, with about 3.65MB words after word segmentation. The subsequent training of n-gram + RNN interpolation language model is based on this corpus.

(3) Pronunciation evaluation dataset

From the PSC data over the years, we selected 4,000 speech data of Topic Talk item with precise manual annotation as the pronunciation evaluation dataset, and each speech data has an independent score by two scoring experts. In order to reflect the reliability of the score, the difference between the expert scores of the selected speech data is required to be within 3 points, and the correlation is about 0.8. In the experiment, we take the average value of the scores of the two experts as the final reference manual standard score. First, we extract different evaluation features from the evaluation dataset, then calculate their posterior probability values, and finally calculate the correlation between the extracted features and the manual standard score to measure the performance of the features.The machine scoring performance can also be measured by the correlation.

6.2 Performance of acoustic model

Considering that Mandarin is a tonal language, we add 4 dimensional pitch features to the 39 dimensional MFCC input features of the acoustic model[17]. The number of input layer units of the neural network is $43 \times 11 = 473$ (the feature of the current frame is 43 dimensions, and 5 frames are spliced in front and back of the current frame, totaling 473 dimensions). The DNN network contains five hidden layers, each with 1024 nodes. It is activated by sigmoid function, and the output is transformed by softmax. In the training process, the Stochastic Gradient Descent (SGD) method is used to optimize the network parameters. There are 150 nodes in each recurrent layer of RNN network, and the BPTT algorithm is used to optimize the network parameters[18]. The DBLSTM network contains 6 hidden layers (2 BLSTM hidden layers, 1 select connection layer, 3 full connection hidden layers), each BLSTM hidden layer contains 256 forward and backward LSTM storage units respectively. The number of nodes in the connection layer is 256, and each full connection hidden layer contains 1024 nodes. The CSC-BPTT algorithm is used to optimize the network parameters.

In order to verify the performance of the proposed acoustic model, it is necessary to compare the recognition results of different acoustic models, as shown in Table 1.

Table 1. Speech recognition performance of different acoustic models.

Acoustic model	Word error rate	Sentence error rate
DNN-HMM	18.83%	23.95%
RNN-HMM	16.52%	20.53%
LSTM-HMM	15.03%	19.78%
DBLSTM-HMM	14.08%	18.22%

Through the analysis of table 1, it can be seen that the recognition performance of the proposed acoustic model is higher than that of the literature model. The word error rate of the proposed acoustic model is 14.08%, which is 4.75%, 2.44%, and 0.95% lower than that of Dahl et al. (2012), Hasim et al. (2015), and Sak et al. (2014), respectively. It shows that it is possible to improve the speech recognition rate by making good use of the context information of LSTM network structure.

6.3 Performance of language model

In the experiment, we use the Srilm tool to train the 3-gram model, and the training text is 480,000 sentences in the prepared language model dataset. After obtaining 50 candidate N-bests reserved by one-pass decoding, the n-gram+RNN interpolation language model is used to re-estimate the score of each best. We set the vector dimension of the input words in the RNN network to be the same as the dictionary size, set the number of hidden layer nodes to 500, and the number of output categories to 100. The BPTT algorithm is used for training, the interpolation coefficient is set to 0.5, and the acoustic model score remains unchanged. In order to verify the performance of the proposed language model, it is necessary to compare the recognition results of different acoustic models, as shown in Table 2.

Table 2. Word recognition performance of different language models.

Language model	Word recognition rate
N-gram	84.65%
N-gram+RNN interpolation	90.52%

Through the analysis of Table 2, we can see that the word recognition rate of the n-gram+RNN interpolation language model is better than that of the n-gram language model. The n-gram+RNN interpolation language model only retrains an RNN network on the basis of the n-gram language model, and the word recognition performance has been significantly improved, which also shows that most of the corrected recognition errors come from the lack of historical information in the language model. In order to verify the correlation between the posterior probability and the manual standard score, we apply the n-gram+RNN interpolation language model to re-estimate the N-best candidate of each sentence in the 4000 data in the pronunciation evaluation dataset, and re estimated the posterior probability of the sentence with the highest score. The experimental results are shown in Table 3.

Table 3. Correlation between a posterior probability and manual score of different language models.

Language model	Correlation between a posterior probability and manual score
N-gram	53.15%
N-gram+RNN interpolation	56.97%

From Table 3, we can see that n-gram+RNN interpolation can improve the correlation between the posterior probability and the manual score, but in terms of recognition performance, the improvement of the correlation is still insufficient, because the RNN language model retains a longer historical information, which makes those sentences with non-standard pronunciation but strong logic become the largest candidates after the RNN re-estimates the score, resulting in a large improvement in the recognition rate.

7. CONCLUSIONS

In order to realize the intelligent evaluation of Topic Talk item in PSC, we studied the text-independent oral evaluation task from the perspective of recognition, and constructed an intelligent evaluation model of Topic Talk item. Based on the recorded speech data of candidates collected from the PSC Centers across the country, we analyzed the speech recognition performance and word recognition performance of the proposed intelligent evaluation model from the perspectives of acoustic models and language models. Experiments show that the intelligent evaluation model proposed in this paper can better identify the text content of the candidate's expression, and the posterior probability calculated based on the recognized text has a high correlation with the manual standard score. During the experiment, it was found that the contextual information of the recognition results was helpful for the improvement of the recognition performance, which also pointed out a direction for the follow-up research.

ACKNOWLEDGMENTS

This work is supported by both the Special Project of Applied Linguistice in China's Hunan Province (Grant No. XYJ2021GB09), and the Scientific Research Project of Hunan Open University (Grant No. XDK2020-C-26).

REFERENCES

[1] Ge, F. P., Lu, L. and Yan, Y. H., "Experimental investigation of mandarin pronunciation quality assessment system," International Symposium Computer Science and Society (ISCCS), 235-239(2011).

[2] Chen, C. H., "Improvement in automatic Putonghua pronunciation quality assessment algorithm," Journal of Guizhou Normal University (Natural Sciences), 31(06), 95-99(2013).

[3] Chen, C. H., "Research of speech recognition network in Putonghua level test system," Journal of Xihua University (Natural Science Edition), 33(02), 17-21(2014).

[4] Mandarin Training and Testing Center of State Language and Writing Commission, [Implementation Outline of Mandarin Proficiency Test], Commercial Press, 7, 461-462(2017).

[5] Rabiner, L. and Juang, B. H., [Fundamentals of Speech Recognition], Prentice-Hall, Inc., 353-356(2009).

[6] Dahl, G. E. and Dong, Y., "Member S. context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," IEEE Transactions on Audio Speech & Language Processing, 20(1), 30-42(2012).

[7] Sak, H., Senior, A., Rao, K., et al., "Fast and accurate recurrent neural network acoustic models for speech recognition," Computer Science, 2(1), 10-15(2015).

[8] Sak, H., Senior, A. and Beaufays, F., "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," Computer Science, 6(1), 338-342(2014).

[9] Zaremba, W., Sutskever, I. and Vinyals, O., "Recurrent neural network regularization," Computer Science, 4(1), 1-8(2014).

[10] Li, J., Wang, X. R. and Xu, B., "Understanding the dropout strategy and analyzing its effectiveness on LVCSR," IEEE Inter. Conf. on Acoustics, Speech and Signal Processing, 7614-7618(2013).

[11] Glorot, X. and Bengio, Y., "Understanding the difficulty of training deep feedforward neural networks," Journal of Machine Learning Research, 9(1), 249-256(2010).

[12] Manning, C. D., [Foundations of Statistical Natural Language Processing], MIT Press, vol 5, (1999).

[13] Mikolov, T., [Statistical Language Models Based on Neural Networks], Brno University of Technology, Doctor's Thesis, 120-122(2012).
[14] Mikolov, T., Kombrink, S., Burget, L., et al., "Extensions of recurrent neural network language model," Proc. of 2011 IEEE Inter. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 5528-5531(2011).
[15] Mikolov, T., Deoras, A., Kombrink, S., et al., "Empirical evaluation and combination of advanced language modeling technique," Proceedings of Interspeech, 605-608(2011).
[16] Young, S., Evermann, G., Gales, M., et al., [The HTK Book (for HTK version 3.4)], Cambridge University Press, 2-3(2006).
[17] Boersma, P., "Accurate short term analysis of the fun amental frequency and the harmonics-to-noise ratio of a sampled sound," Proc. of the Institute of Phonetic Sciences, 97-110(1993).
[18] Stolcke, A., "SRILM-an extensible language modeling toolkit," Proc. of the Interspeech, 901-904(2002).