A Study on the Evaluation Model of In-depth Learning for Oral English Learning in Online Education

Yanli Ge
Office of Foreign Language, Basic Teaching Department,
Changchun University of Architecture and Civil Engineering, Changchun, 130607, China

***Abstract*—The trend of globalization in the world is becoming increasingly frequent, and people from different regions are communicating more closely. Therefore, the demand for a second language is constantly expanding, accelerating the development of the field of English oral evaluation and also accelerating the development of online education. The study proposes a text priori based oral evaluation model, which is based on the Transformer model and uses target phonemes as input to the Decoder. The model successfully predicts the relationship between actual pronunciation and error labels. At the same time, a self-supervised oral evaluation model with accent is constructed, which simulates the training process of misreading data by calculating semantic distance. The experimental results show that when the training set ratio reaches its maximum in the Speed Ocean dataset and the L2 Arctic dataset, the F1 values of the proposed method are 0.612 and 0.596, respectively; the length of the target phoneme has a smaller impact on this model compared to other models. Experiments have shown that the proposed deep learning method can alleviate deployment difficulties, directly optimize the effectiveness of oral evaluation, provide more accurate feedback, and also provide users with a better learning experience. This has practical significance for the development of the field of oral evaluation.**

Keywords—Spoken English; online education; transformer model; deep learning; evaluation model

I. Introduction

Deep learning technology has largely enhanced the efficiency of speech recognition. Deep speech recognition technology can recognize the phonemes of students' speech and compare them with the text they read. Compared with traditional evaluation methods, this method only needs to train a single recognition model, without complex modeling or providing additional comparative corpus. This speech recognition technology has become the main solution in spoken language testing [1-2]. However, the current spoken language testing methods based on deep learning mostly focus on speech recognition, mainly from improving the accuracy of speech recognition. These methods tend to use better acoustic models in speech testing to improve the effect of oral testing, thus ignoring the shortcomings of speech recognition in oral testing. The spoken language test algorithm based on speech recognition mainly aims at the phoneme and target phoneme in speech recognition to misread. Its optimization aims at improving the accuracy of speech recognition, rather than directly optimizing the effect of oral test. The misreading result generated by the algorithm is binary. It misreads the identified phoneme and target phoneme by aligning them, and judges whether to align the target phoneme or not, so it is unable to adjust the severity of the evaluation. At present, most of the mainstream recognition patterns need autoregressive recognition and decoding. This process is not real-time, which is a big defect for students who require fast feedback [3-4]. Therefore, to solve the above problems, a text priori oral evaluation model is proposed. This model uses the Transformer mode as the basis of the oral test, and appends a target text entered by the Decoder. By converting the non-differential calibration to the data preparation stage, the misreading of each target is improved, thus realizing the error recognition of each target. Furthermore, the study further discusses the role of phonemes in speech recognition models, demonstrating the key role of phonemes in English oral teaching. The innovation of this method lies in optimizing the speech recognition model from the perspective of oral phonemes, enabling learners from different regions and accents to learn from online oral teaching. The method proposed in the study can effectively recognize the phonemic features of spoken language, making oral learning more widely applicable and playing an important role in promoting online oral teaching.

II. Related Work

The gradual development of deep learning has become the research object of many international scholars and has achieved certain results. Wang et al. implemented a new fault location by using multiple feature groups for depth and breadth learning. They analyzed suspicious features based on spectrum and mutation by combining the combination features of invariants based on suspiciousness, static measurement, collapse stack tracking and invariants change features. Through testing a real software defect standard, Defects4J, higher early diagnosis performance than traditional methods were confirmed [5]. KotaV et al. adopted neural networks for emotion analysis. Methods CNN, double LSTM, attention mechanism and other methods were used for emotional analysis. CNN can reduce complexity, while dual LSTM can help handle long input text. This method uses the attention mechanism to determine the importance of each hidden state and weight it [6]. Seebeck and other scholars developed a DL method, which can automatically obtain comprehensive retinal sensitivity from OCT volume. The relative error of PWS and multiple sclerosis is 2.34 dB, and the minimum relative error is 5.70 and 3.07. Pearson correlation coefficient is 0.66 and 0.84, Spearman correlation coefficient is 0.68 and 0.83. Their research showed that predicting the retinal function of each measurement site based on OCT scanning can be used as a new visual function prediction method [7]. Kong and his team developed a second-order one-dimensional phase expansion method. The first step is to encode the phase of one dimension using quasi-Grami matrix. The second step is to use the deep convolution neural network to unwrap the phase. Both simulation and measurement results showed that the phase unwrapping quality of this algorithm is significantly improved when the SNR is less than 4 dB, and it can still maintain good performance under negative SNR [8].

There are many types of oral teaching models in the current research field. Chen S introduced an online oral English teaching platform based on the Internet of Things. This platform adopts the technology of Internet of Things to realize the design of the system structure of online oral English teaching platform and establish a virtual teaching environment. The platform corrects the user's mouth shape and pronunciation through the voice teaching system, and establishes a vocabulary tagging model based on long-term memory. He also introduced the attention mechanism into the long-term memory network. The test results showed that the network delay of the system is between 0.26 seconds and 0.37 seconds, which reduces the development time by 50% and increases the human-computer interface by 13.20% [9]. Liu used speech recognition technology to analyze and deal with differences in phoneme expression in spoken English, and made statistics on some errors and areas to be improved in English. The development and promotion of speech recognition technology can effectively reduce the cost of college oral English teaching and promote the improvement of college students' oral English ability [10]. Xu first proposed the concept of five dimensions of AR situational telepresence, i.e. the sense of scene, immersion, reality, interaction, and social telepresence. Then, combined with the actual situation of English teaching, he put forward a theoretical framework to strengthen the teaching of spoken English. Finally, he made a systematic analysis and discussion on the relationship between the proposed dimensions. He explored the application of augmented reality technology in classroom and online teaching from three levels of perception, acceptance and application [11].

To sum up, deep learning has been widely used in many fields and has shown strong performance. However, the field of oral evaluation is still in the development stage, so the research applies the deep learning technology to the oral evaluation model for the first time. The purpose of the study is to improve the oral evaluation model and further promote the development of oral evaluation.

III. Oral English Learning Evaluation Model for Online Education based on Improved Deep Learning Algorithm

A. *The Construction of Transformer Oral Evaluation Model based on Text Priori*

With the continuous development of computer deep learning technology, the task of speech recognition has been greatly improved. The evaluation model based on in-depth learning of spoken English has been applied in online oral English education [12-14]. The research first proposes a text priori-based oral evaluation model, whose structure is shown in Fig. 1.

The model built in Fig. 1 obtains the error status label of the phoneme by comparing the actual phoneme with the target phoneme. The obtained wrong label can avoid the model from introducing an improbable alignment operation during training, and at the same time, the alignment operation will be transferred to the data preparation stage. This model is significantly different from the spoken language evaluation model using speech recognition. The model used in the research does not use the actual pronunciation phoneme as the input of Decoder, but uses the target phoneme. The prediction expression of actual pronunciation and error label is shown in formula (1).

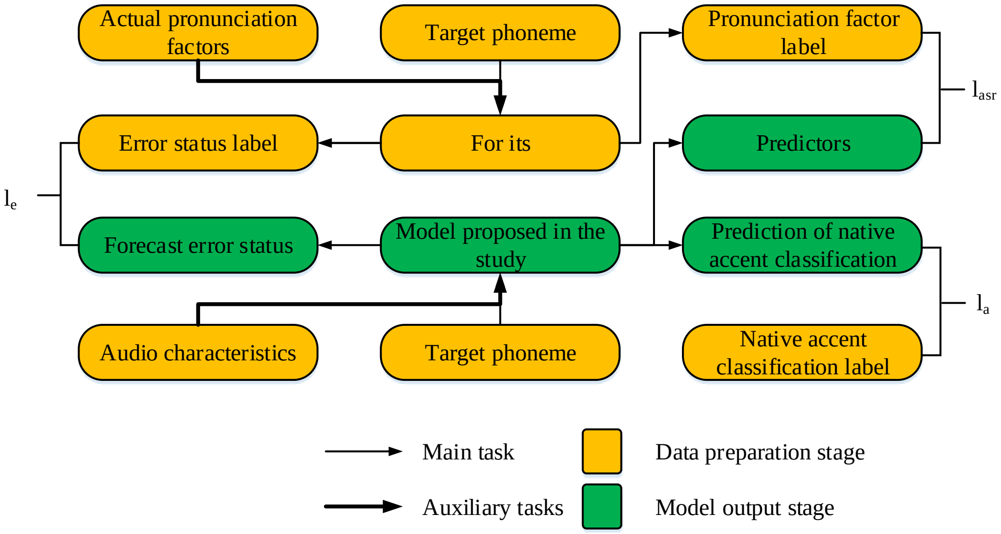


Fig. 1. Work flow of oral evaluation based on text priori.

$$\hat{P},\hat{E} = Decoder(H, P^{tgt}) \quad (1)$$

In formula (1), $\hat{P}$ represents the actual pronunciation. $\hat{E}$ indicates an error label. P^{tgt} is the target phoneme. H represents the characteristics of audio. When outputting the model, the output length should be paid attention. In the oral language evaluation model of speech recognition, when to output <EOS> determines the output length of speech sequence. Moreover, when the length of speech sequence is aligned with the target text, the model will output an error status sequence, which is the same length as the target text. The model used in the study has used phoneme tagging to align the speech sequence with the target text in the training process, which can also make the length of the error state equal to the target text. The research takes “hi” as an example, and aligns the labeled actual phoneme “HH EY” with the target text phoneme “HH AY” in the data preparation stage, and then obtains the phoneme error status target. For audio features and target phonemes, the output of the model reflects the matching between the two. A two-layer convolutional network is superimposed on the back end of Decoder, and the convolution core size of the network is 3 * 3. ReLU function is used as the activation function, and the output value is mapped to [0,1] interval through linear layer and sigmoid function. The oral language evaluation model based on text priori is shown in Fig. 2.

Since the evaluation can be differentiated in the output of the evaluation error state, the loss function between the predicted state and the real label can be directly calculated and optimizes the whole model using back-propagation. The research uses Binary Cross-Entropy (BCE) to train the predicted value and label, and its expression is shown in formula (2) [15-17].

$$l_e^{BCE} = BCE(\hat{E}, E) \quad (2)$$

In formula (2), E represents the real label. In the evaluation task, BCE loss does not represent a loss function. When the model extracts the relevant acoustic features of the speaker's mother tongue as an auxiliary task, the research adds a module about mother tongue prediction to the output of Encoder. This module predicts the speaker's mother tongue through the input audio features. The input sequence audio features need to be converted into a single classification output, so the research uses the Statistics pooling layer based on mean variance statistics to achieve the above purpose. Its expression is shown in formula (3).

$$\hat{a} = StatisticsPooling(H) \quad (3)$$

In formula (3), $\hat{a}$ represents the single classification output of the sequence audio feature transformation. Cross Entropy (CE) is also used for the training of $\hat{a}$ and a. Its expression is shown in formula (4).

$$l_a = CrossEntropy(\hat{a}, a) \quad (4)$$

The criterion for classifying the error state is the misreading of the target phoneme. The model needs more samples for training, otherwise the model cannot combine the output target with the corresponding audio and target phonemes, and then over-fitting occurs. At present, there are few data sets that can be used in oral evaluation, so the research needs to obtain an acoustic model first. The model can be trained by standard speech recognition dataset. After completing the pre-training task, the research will still require the model to complete the auxiliary task of speech recognition, so as to mine more relations between audio features and phonemes. Finally, the loss function is integrated to obtain the formula (5).

$$\begin{cases} l = l_e + \alpha l_a + \beta l_{asr} \\ l_{asr} = CrossEntropy(\hat{P}, P^{out}) \end{cases} \quad (5)$$

In formula (5), α represents the weight of the auxiliary task of native language recognition. β represents the weight of loss function of speech recognition auxiliary task.

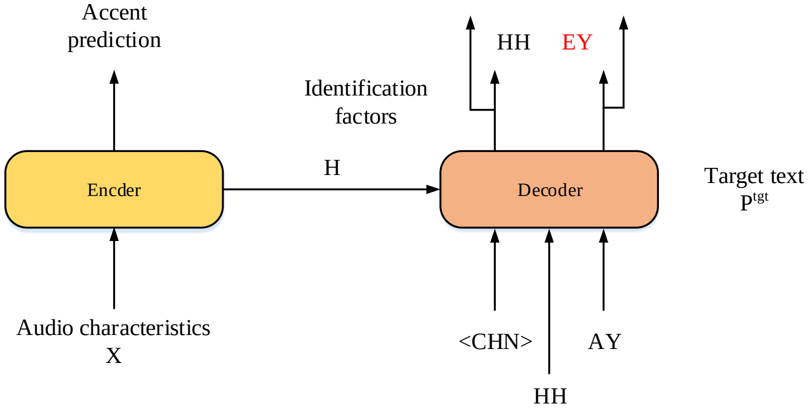


Fig. 2. Structure of spoken language evaluation model based on text priori.

B. Self-Supervised Acoustic Model Construction

In the practice of second language learning, learners are extremely vulnerable to the objective influence of their mother tongue pronunciation, which leads to the deviation of their second language pronunciation from the standard pronunciation. There is unavoidable misreading in oral English assessment. There is a large error between the text annotation of the Second Language (L2) phonetic feature and the actual pronunciation. Therefore, in the absence of actual pronunciation marking, it is difficult to accurately model L2 speech features based on the text priori oral evaluation model. Self-supervised learning (SSL) is a common machine learning method. It can use auxiliary tasks to accurately mine the supervision information related to itself from the massive unsupervised data without external tag data, and then realize effective network training [18-20]. In recent years, the research on SSL in the voice field has attracted more and more attention. The most classic structure in SSL is the Noise Reduction Auto Encoder (DAE). Its main structure is divided into three parts, namely encoder, bottleneck layer and decoder, as shown in Fig. 3.

According to Fig. 3, the DAE will first learn a compressed feature vector from the original features. Then the decoder will process the feature vector to recover the corresponding original data. The feature vectors in the original feature will be compressed once in the encoder and bottleneck layer respectively. Therefore, the original data recovered by the decoder is no longer accompanied by noise and other influence elements, and will be more representative and typical. From Fig. 1, the DAE will modify the original features before importing them. According to the different types of input features, there are also some differences in the modification methods of features, mainly including transformation, masking, and comparative learning. Transformation refers to converting the original voice input information into spectrum information, and then requiring the network model to recover the original waveform from it. Masking refers to treating the input speech feature as 0 randomly, and the most widely used is the Bidirectional Encoder Representations from Transformer (BERT) model. Comparative learning can ensure that the model can screen out more typical and distinctive speech features under specified conditions. This modification method no longer only destroys and modifies the original input information, but helps the network model learn valuable representative features by adding interference items. The most typical model is Wav2Vec model. The BERT model and Wav2Vec model are organically combined, and a discrete process is added after the encoder completes the encoding of the original features. This realizes the effective integration of discretization process and comparative learning, and obtains the Wav2Vec2 network model, whose structure is shown in Fig. 4.

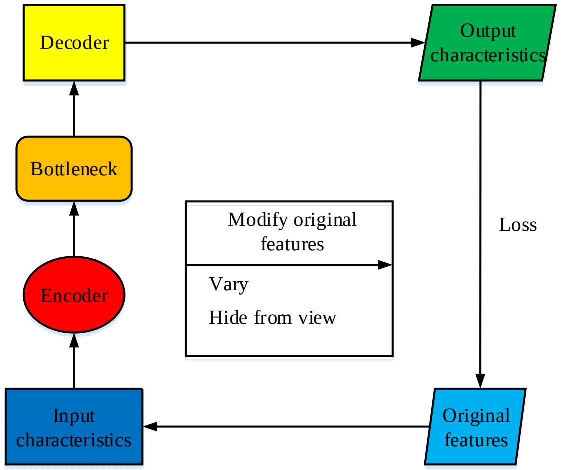


Fig. 3. Network structure diagram of DAE.

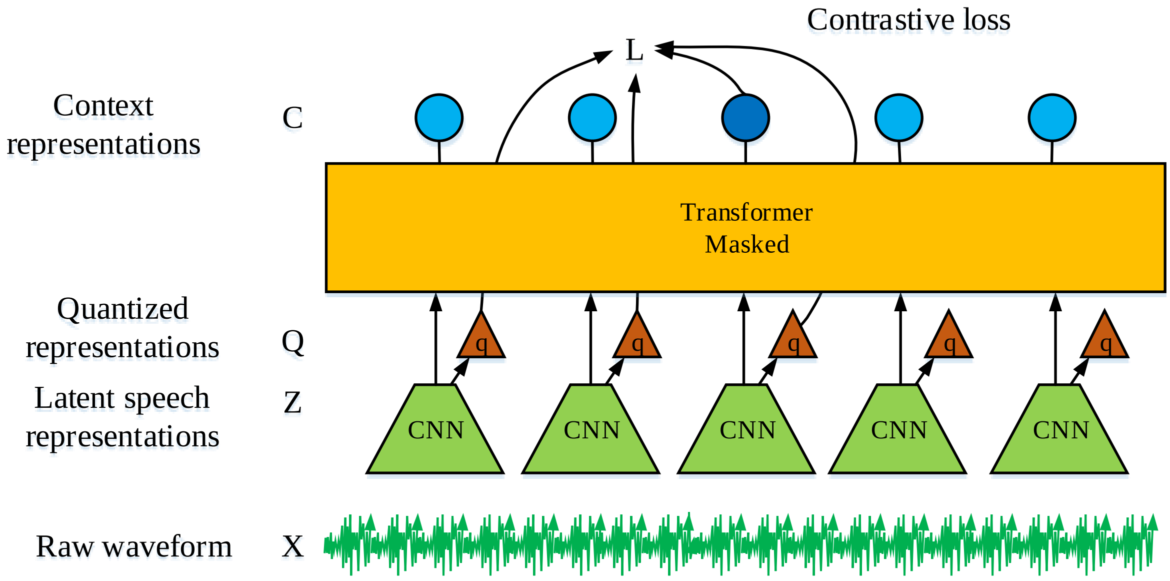


Fig. 4. Structure and training diagram of Wav2Vec2 network model.

Looking at Fig. 4, the original speech feature X can be encoded as Z under the action of the encoder, and Z has a higher degree of abstraction. Then the Wav2Ve++c2 network model will enter the discretization process, from which Z can be converted into speech feature Q. Finally, Transformer can integrate it into a new context feature C. The Wav2Vec2 network model also includes the contrast loss function training model, which can generate a richer discrete code table. On this basis, the discrete acoustic units can be obtained by clustering them with k-Means algorithm, as shown in Fig. 5.

Fig. 5 shows the complete acoustic unit construction process. The circle represents the cluster, the red label and the blue label represent the original acoustic unit and the replaced acoustic unit respectively. Black dots represent each voice data in the training process of the clustering model network. The semantic vector is extracted from the target speech standard and L2 speech feature, and then Class-K speech information is obtained under the clustering effect of k-Means algorithm. Finally, it is converted into discrete acoustic unit sequence-U, as shown in Eq. (6).

$$U = \{u_1, u_2, \cdots, u_T\} \tag{6}$$

In formula (6), T is the corresponding length of the discrete acoustic unit sequence, which can replace the original voice features as the input of the network model. For any U, ranking the other $K-1$ acoustic units according to their distance from U, and their corresponding distance is obtained as shown in Formula (7).

$$S^d = \{s_1^d, s_2^d, \cdots, s_{K-1}^d\} \tag{7}$$

In formula (7), d is the single semantic distance, that is, the difference between the vector distance of the acoustic unit before and after the replacement in the semantic subspace. It can accurately reflect the matching degree between the replaced acoustic unit sequence and the original speech features. The research selects the unit closest to k from the dataset as a substitute, and uses normal distribution to select it, as shown in formula (8).

$$k = \min(K-1, \mathrm{int}(abs(\frac{r(K-1)}{3}))) \tag{8}$$

The above method can obtain the vector distance difference of the acoustic model in the semantic subspace before and after the replacement. Euclidean distance is selected as the calculation method of distance difference, and the expression is shown in formula (9).

$$d(u, u^r) = MSE(v, v^r) \tag{9}$$

In formula (9), d refers to the distance difference of the vector. The mute part of the original audio will have a huge distance from other acoustic units after clustering. These outliers will interfere with the model, so the replacement method used in the study should be used under the condition of $d(u, u^r) < H$. If the condition is not met, it will not be replaced. The pre-training process based on acoustic unit replacement is shown in Fig. 6.

In Fig. 6, the original audio is converted into a discrete audio sequence U. Then the new sequence U^r is obtained by replacing the acoustic unit. The distance difference between the two in the quantum space is shown in formula (10).

$$D = \{d_1, \cdots, d_T\} \tag{10}$$

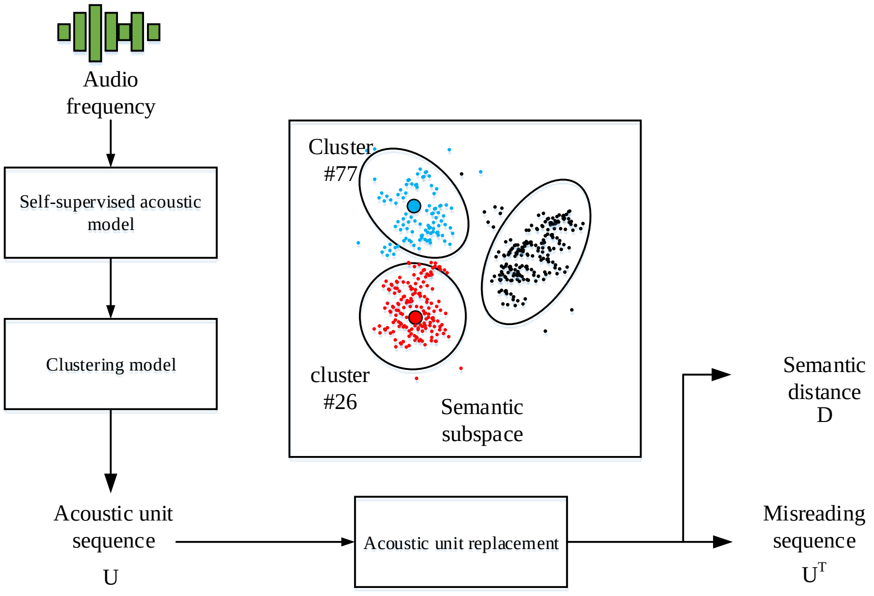


Fig. 5. Self-supervised clustering of original speech features and its replacement.

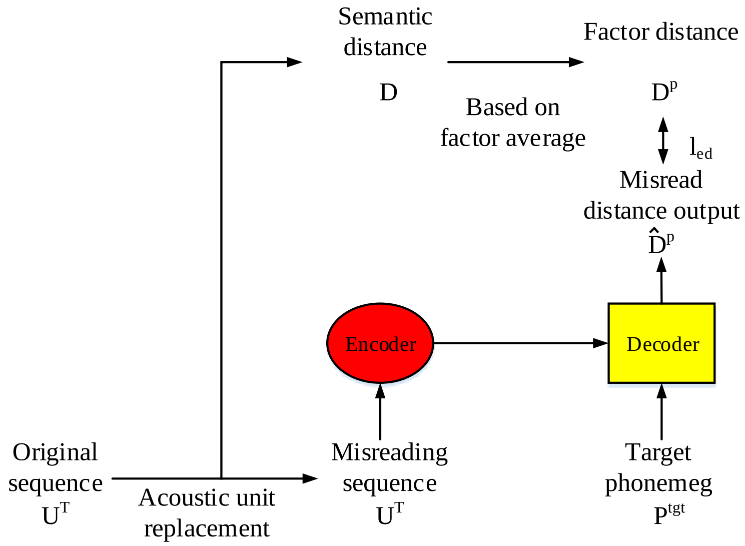


Fig. 6. Pre-training process based on acoustic unit replacement.

To obtain the corresponding distance between phonemes, the initial position of phonemes is searched through the forced alignment tool. Therefore, the corresponding distance of phonemes is shown in formula (11).

$$D^p = \{d_1^p, \cdots, d_L^p\} \quad (11)$$

In formula (11), L represents the sequence length of the target phoneme. Finally, the research attempts to predict the distance model of each phoneme with the model, and then make the model learn the degree of deviation from the target text. The loss of the model is expressed by the mean square error function, as shown in formula (12).

$$l_{ed} = MSE(\hat{D}^p, D^p) \quad (12)$$

Formula (13) is the loss function when using the substitution method based on acoustic units for migration prediction.

$$l = l_{ed} + \beta l_{asr} \quad (13)$$

The clustering model is trained in L1 and L2 speech, and the characteristics of standard and non-standard pronunciation of the target speech are obtained. These features can obtain more fine-grained audio replacement by changing the number of clusters, which greatly increases the authenticity of misread samples.

IV. Performance Verification of Oral English Learning Evaluation Model for Online Education

A. *Performance Analysis of Oral Language Evaluation Model based on Text Priori*

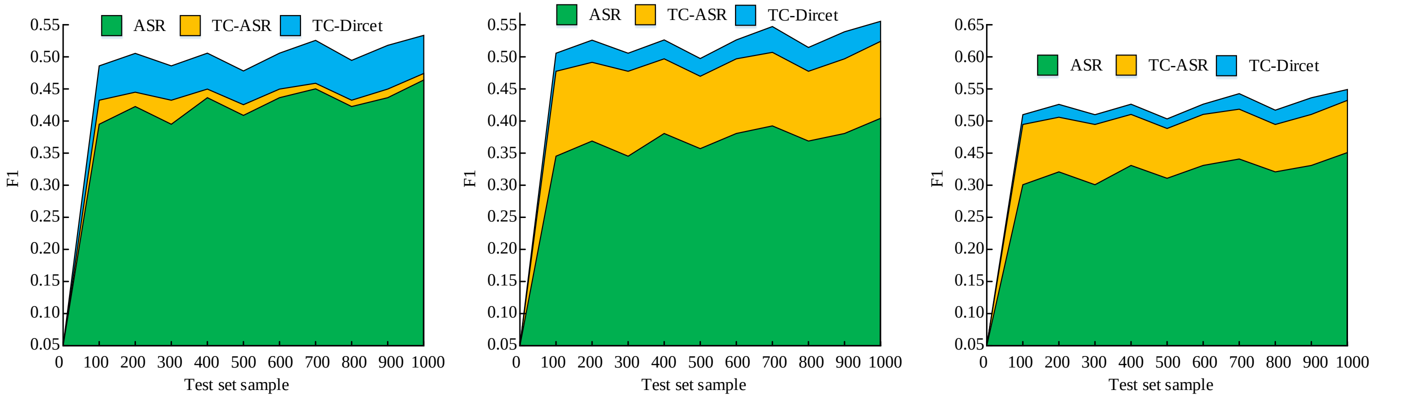
Before the experimental analysis, the weight of the auxiliary task of mother tongue recognition was set to 0.1. If no native language information is added, the weight is set to 0. The weight of loss function of speech recognition auxiliary task is set to 0.1. The ASR pre-training uses the Librispeech data set. The ratio of training set, verification set and test set is approximately 10:1:1 in this data set. The L2-Arctic data set is used in the oral evaluation task. The data set is divided into training set, verification set and test set according to the ratio of 10:1:4. The test of model performance is mainly evaluated by seven indicators: Phoneme Error Rate (PER), Precision, Accuracy, Recall, F1, False Rejection Rate (FRR) and False Acceptance Rate (FAR). The effect comparison of different models in oral evaluation is Table I.

TABLE I. Oral Evaluation of Different Methods

Model		PER	PRE	ACC	REC
Primitive phoneme	ASR	0.224	0.426	0.787	0.524
	TC-ASR	0.120	0.452	0.833	0.396
	TC-Direct	0.129	0.506	0.824	0.474
Extended phoneme	ASR	0.286	0.403	0.789	0.401
	TC-ASR	0.155	0.569	0.842	0.502
	TC-Direct	0.173	0.500	0.823	0.521
Extended phoneme+	ASR	0.293	0.398	0.786	0.413
	TC-ASR	0.172	0.552	0.838	0.519
	TC-Direct	0.181	0.488	0.818	0.629

Table I shows the performance comparison of different models under different conditions. In Table I, there is a certain gap between the PER, RRE, ACC and REC indicators of ASR model and the other two models. The ACC index of TC-ASR model is higher than TC-Direct, while the remaining three indexes are lower than TC-Direct model. The PER value of TC-Direct model in “extended phoneme+” is 0.181. The ACC value in “extended phoneme” is 0.823. The REC value in “extended phoneme+” is 0.629. The experimental results show that the TC-Direct model has better performance in oral evaluation.

Fig. 7 shows the F1 value result of the translation model. Fig. 7 (a) shows the F1 value of the model in the original phoneme. The F1 of ASR is 0.464. TC-ASR model is a text priori phoneme level speech recognition model, its F1 value in the original phoneme score is 0.462. The TC-Direct model is a text priori model proposed by the study, and its F1 score in the original phoneme is 0.538. Fig. 7 (b) shows the F1 value of the model in the extended phoneme. The F1 score of ASR model is 0.402. The F1 score of TC-ASR model is 0.533. The F1 of TC-Direct is 0.554. Fig. 7 (c) shows the F1 value of the model in “extended phoneme+”. The F1 score of ASR model is 0.405. The F1 score of TC-ASR model is 0.535. The F1 score of TC-Direct model is 0.549. The experimental results show that the F1 value of TC-Direct model is the highest in the three environments, and the validity of extended phoneme is also shown.

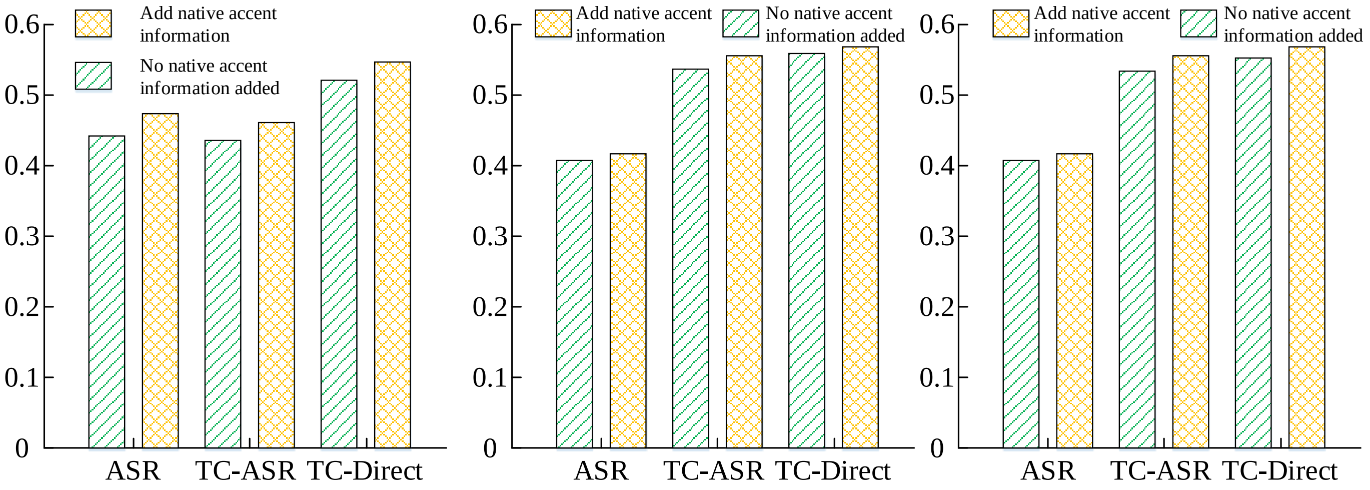


(a) Model F1 value in the original phoneme

(b) Model F1 value in extended phoneme

(c) Model F1 value in extended phoneme

Fig. 7. F1 values of machine translation models in different models.



(a) The model adds the F1 score of mother tongue feature information to the original phoneme

(b) The model adds the F1 score of mother tongue accent information to the expanded phoneme

(c) The model adds the F1 score of mother tongue accent information to the expanded phoneme

Fig. 8. F1 scores of the model before and after adding mother tongue accent information.

Fig. 8 shows F1 scores of each model after adding native language feature information. Fig. 8(a) shows the F1 score of the model adding native language feature information to the original phoneme. The F1 value of ASR is 0.464 without adding native language accent label; The F1 score of the ASR model is 0.473 when the mother tongue accent label is added. Similarly, the F1 scores of TC-ASR model before and after adding mother tongue information were 0.462 and 0.469 respectively; The F1 scores of TC-Direct model before and after adding mother tongue information were 0.538 and 0.550 respectively. Fig. 8(b) shows the F1 score of the model in which the mother tongue accent information is added to the expanded phoneme. The F1 scores of ASR model were 0.402 and 0.404 before and after the addition of mother tongue accent information. The F1 scores of TC-ASR model before and after adding mother tongue information were 0.533 and 0.551 respectively. The F1 of TC-Direct before and after adding mother tongue information were 0.554 and 0.562 respectively. Fig. 8(c) shows the F1 score of the model in which the mother tongue accent information is added to “extended phoneme+”. The F1 scores of ASR model were 0.405 and 0.408 respectively before and after adding native accent information. The F1 scores of TC-ASR model before and after adding mother tongue information were 0.535 and 0.552 respectively. The F1 scores of TC-Direct model before and after adding mother tongue information were 0.549 and 0.555 respectively. From the comparative analysis of model results, TC-Direct model has a higher F1 score compared with other models, indicating that the model has better performance. From the analysis of the results before and after adding the mother tongue accent information, adding the mother tongue accent information can improve the F1 score of the model, indicating that the mother tongue accent information can help the model obtain better recognition performance.

B. *Analysis of Influence of Parameters on Model Performance*

Adjusting the weight θ between FAR and FRR can make the oral evaluation model different in difficulty, as shown in Fig. 9. Fig. 9(a) shows the change of the Recall-Recision curve of the loss function when adjusting θ. The larger the value of θ, the effective adjustment range of Focal function is between 0.05 and 0.95. The effective adjustment range of BCE function is about 0.38 to 0.78. The effective adjustment range of F1 function is 0.59 to 0.62. Fig. 9(b) shows the change of the FAR-FRR curve of the loss function when adjusting θ. The effective adjustment range of Focal function is also between 0.05 and 0.95. The effective adjustment range of BCE function is about 0.22 to 0.42. The effective adjustment range of F1 function is 0.37 to 0.39. The experimental results show that Focal loss function has a wider adjustment range, which is a better choice for practical application.

Fig. 10 shows the effect of target phonemes of different lengths on the reasoning duration. Fig. 10(a) shows the reasoning time results of ASR model for different length phonemes. The longer the length of the target phoneme is, the longer the reasoning time of ASR model is, which is in positive proportion. Fig. 10(b) shows the reasoning time results of TC-Direct model for different length target phonemes. The reasoning time of the model is also positively correlated with the phoneme length, but the influence is low. The model has the best performance when the target phoneme length is 25-40.

Fig. 11 shows the effect of the scale of the training set on the performance of the model. Fig. 11(a) shows the F1 value of the model in the L2-Arctic dataset. The F1 of ASRis 0.372 without training; The F1 value is 0.473 when the training set proportion reaches the highest. The F1 value of TC-Direct is 0.596 when the training set proportion is the highest. Fig. 11(b) shows the F1 value of the model in the Speed Ocean dataset. The F1 of ASR is 0.356 without training, and the ratio of training set is 0.446 when it reaches the maximum. The F1 of TC-Direct is 0.612 when the training set proportion is the highest. The F1 value is higher than that of ASR model in both data sets regardless of the proportion of the training set. Comparing the two data sets, the F1 value of ASR in the Speed Ocean data set is slightly lower than that in the L2-Arctic data set. The F1 value of the TC-Direct model in the Speed Ocean dataset is higher than that in the L2-Arctic dataset. Thus, the TC-Direct has wider applicability.

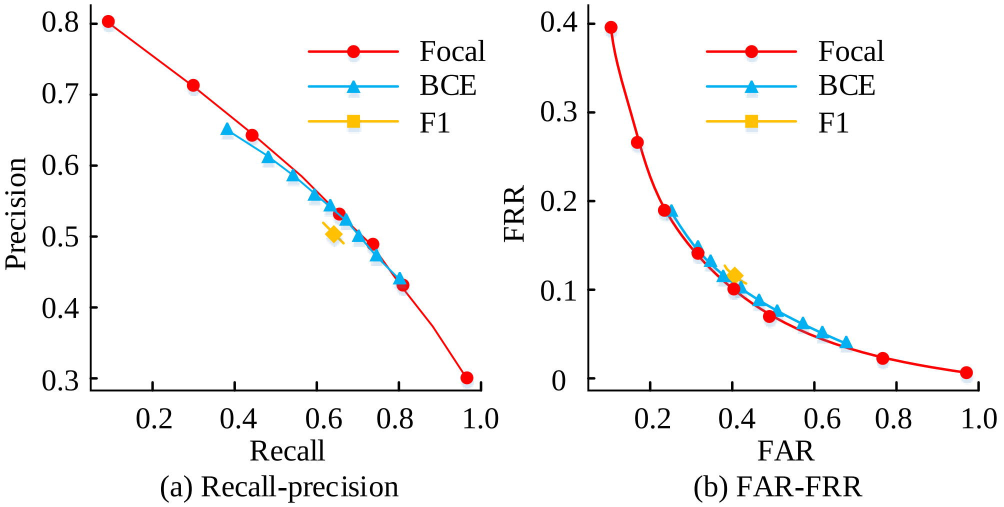
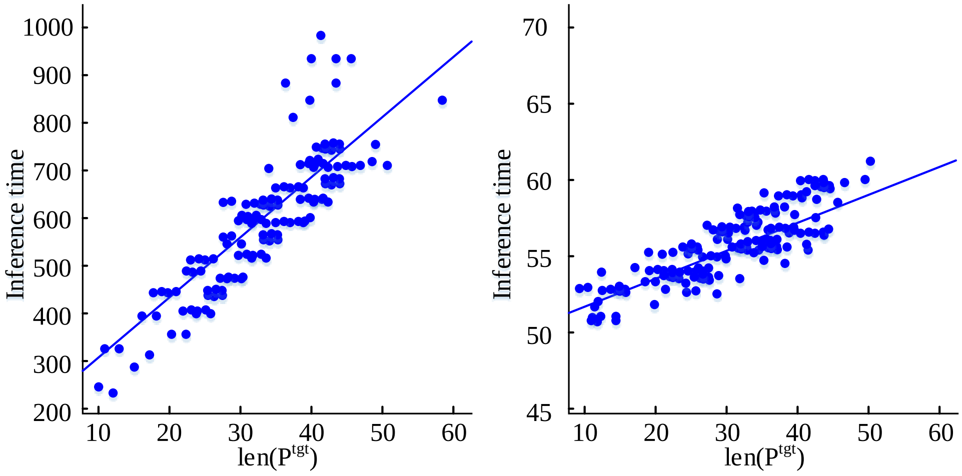


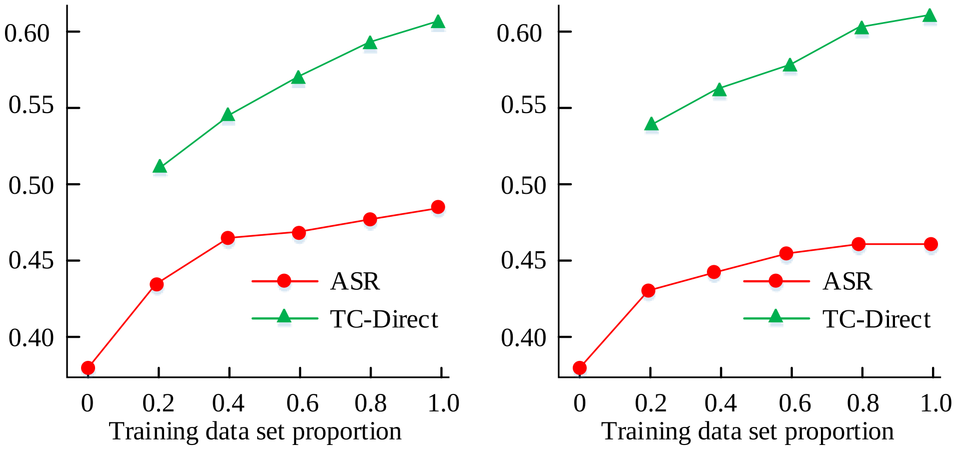
Fig. 9. Influence of the weight between FAR and FRR on the loss function.



(a) F1 of model in L2-Arctic data set

(b) F1 value of the model in the SpeedOcean dataset

Fig. 10. Reasoning duration results of target phonemes with different lengths.



(a) L2-Arctic

(a) Speech Ocean

Fig. 11. F1 value of model in different training set proportions.

V. Discussion

The proposed model was experimentally validated through comparative experiments. Analyzing the performance of the model from three datasets, among which the TC-ASR and TC-Direct models have better performance. In the original phonemes, the PER and ACC indicators of TC-ASR are slightly better than those of TC-Direct, but both exceed 0.1; The PER and REC indicators of the TC-Direct model are significantly higher than those of the TC-ASR model. In the "extended phoneme" and "extended phoneme+" models, the TC-ASR model showed lower REC indicators than the TC-Direct model, while the PER, PRE, and ACC indicators were higher than the TC-Direct model. Through the above four indicators, it is difficult to distinguish the performance gap between models. The study continued to use the F1 value to measure the superiority of the model's performance. After introducing the F1 value, the TC-Direct model showed the highest F1 value in all three phoneme conditions, indicating that the model had good performance. Not only that, the study also added native accent information, which can enhance the F1 value of the model, indicating that native accent information helps the model to perform more accurate recognition. To verify the rigor of the experiment, several parameters were validated. Adjust the weight of parameters, adjust the length of phonemes, and adjust the size of the training set. In the experiment, the weight of FAR-FRR has a direct impact on the performance of the model. If its weight ratio is about, the worse the model performance; The length of phonemes does not directly affect the performance of the model, indicating that the model has a wide range of applications; The performance of the model also depends on the size of the training set, and with sufficient training sets, the model can perform better.

VI. Conclusion

For the low performance of the conventional speech recognition oral evaluation model, a text priori-based oral evaluation model is proposed. By using the self-supervised learning method to build the acoustic model, the speech recognition and misreading detection are combined to achieve the purpose of error state prediction. The research verifies the performance of the proposed model through Librispeech dataset, L2-Arctic dataset and Speed Ocean dataset. The experimental results show that the F1 value of the model in the "original phoneme" is 0.538, the F1 value in the "extended phoneme" is 0.554, and the F1 value in the "extended phoneme+" is 0.549; After adding native accent features, the F1 value of the model is 0.550 in the "original phoneme", 0.562 in the "extended phoneme", and 0.555 in the "extended phoneme+". The proposed model has good F1 values in different phoneme datasets, indicating that the model has good performance in phoneme recognition and can improve the recognition performance of English spoken language. In the experiment, the study also verified the impact of the length of the target phoneme on the model performance, and the results showed that the change in model performance was less affected by the change in the length of the target phoneme. The experiment has verified that the performance of the model has a direct impact on the size of the training set. If the model is trained in sufficient training sets, it cannot continuously increase the F1 value, further improving the model's performance, and indicating that the model has a wider adaptability. However, there are still deficiencies in the study. The research did not explore the speech style when it was used for oral evaluation, so the follow-up research needs to preserve the speech style without accent.

REFERENCES

[1] N. Wang, X. Zhang, A. Sharma. “A Research on HMM based Speech Recognition in Spoken English”. Recent Advances in Electrical & Electronic Engineering (Formerly Recent Patents on Electrical & Electronic Engineering), 2021, 14(6):617-626.

[2] Q. Zhang, “Recognition of English spoken stressed syllables based on natural language processing and endpoint detection algorithm”. Journal of Intelligent and Fuzzy Systems, 2020, 39(4):5713-5724.

[3] F. Jiang, Y. Chiba, T. Nose, A. Ito. “Language modeling in speech recognition for grammatical error detection based on neural machine translation”. Acoustical Science and Technology, 2020, 41(5):788-791.

[4] P. M. Cuenca-Jimenez, J. Fernandez-Conde, J. M. Canas-Plaza. “FilterNet: Self-Supervised Learning for High-Resolution Photo Enhancement”. IEEE Access, 2022, 10:2669-2685.

[5] T. T. Wang, H. L. Yu, K. C. Wang, X. H. Su, “Fault localization based on wide & deep learning model by mining software behavior”. Future Generation Computer Systems, 2022, 127:309-319.

[6] V. R. Kota, S. D. Munisamy, “High accuracy offering attention mechanisms based deep learning approach using CNN/bi-LSTM for sentiment analysis”. International Journal of Intelligent Computing and Cybernetics, 2022, 15(1):61-74.

[7] P. Seebeck, W. D. Vogl, S. M. Waldstein, J. I. Orlando, M. Baratsits, T. Alten, T. Ankan, G. Mylonas, Bogunovic H, Schmidt-Erfurth U. “Linking Function and Structure with ReSensNet Predicting Retinal Sensitivity from OCT using Deep Learning”. Ophthalmology retina. 2022, 6(6):501-511.

[8] L. Kong, K. Cui, J. Shi, M. Zhu, S. Li. “1D Phase Unwrapping Based on the Quasi-Gramian Matrix and Deep Learning for Interferometric Optical Fiber Sensing Applications”. Journal of Lightwave Technology: A Joint IEEE/OSA Publication, 2022, 40(1):252-261.

[9] S. Chen. “Design of internet of things online oral English teaching platform based on long-term and short-term memory network”. International Journal of Continuing Engineering Education and Life-Long Learning, 2021, 31(1):104-118.

[10] C. Liu. “Application of speech recognition technology in pronunciation correction of college oral English teaching”, Application of Intelligent Systems in Multi-modal Information Analytics: Proceedings of the 2020 International Conference on Multi-model Information Analytics (MMIA2020), Volume 2. Springer International Publishing, 2021: 525-530.

[11] D. Xu. “Research on the Construction Strategy of the Theoretical Framework of Presence in Oral English Teaching Based on Augmented Reality Technology”. Creative Education, 2022, 13(10): 3162-3173.

[12] S. Kummin, S. Surat, F. M. Kutty, Z. Othman, J. Thompson., “The Use of Multimodal Texts in Teaching English Language Oral Skills”. Universal Journal of Educational Research, 2020, 8(12):7015-7021.

[13] Y. Hai. “Computer-aided teaching mode of oral English intelligent learning based on speech recognition and network assistance”. Journal of Intelligent and Fuzzy Systems, 2020, 39(4):5749-5760.

[14] M. Vojnovic, M. Mijic, D. S. Pavlovic, N. Vojnovic, “Influence of Overpressure Breathing on Vowel Formant Frequencies”. Archives of acoustics: journal of Polish Academy of Sciences, 2021,46(1):177-181.

[15] M. F. Biller, C. J. Johnson, “Examining Useful Spoken Language in a Minimally Verbal Child with Autism Spectrum Disorder: A Descriptive Clinical Single-Case Study”. American Journal of Speech-Language Pathology, 2020, 29(3):1-15.

[16] X. Liu, H. Zhang, Q. Liu, S. Dong, C. Xiao. “A cross-entropy algorithm based on Quasi-Monte Carlo estimation and its application in hull form optimization”. International Journal of Naval Architecture and Ocean Engineering, 2021, 13(4):115-125.

[17] Y. Zhao, Y. Han, Y. Liu, K. Xie, W. Li, J. Yu. “Cross-Entropy-Based Composite System Reliability Evaluation Using Subset Simulation and Minimum Computational Burden Criterion.” IEEE Transactions on Power Systems, 2021, 36(6):5198-5209.

[18] C. Munoz, H. Qi, G. Cruz, T. Küstner, R. M. Botnar, C. Prieto. “Self-supervised learning-based diffeomorphic non-rigid motion estimation for fast motion-compensated coronary MR angiography.” Magnetic Resonance Imaging, 2022, 85:10-18.

[19] C. Y. Liu, X. Chen, Z. Li, R. Proietti, et al., “SL-Hyper-FleX: a cognitive and flexible-bandwidth optical datacom network by self-supervised learning”. Journal of optical communications and networking, 2022, 14(2): A113-A121.

[20] A. A. Baffour, Z. Qin, J. Geng, J. Ding, et al., “Generic network for domain adaptation based on self-supervised learning and deep clustering”. Neurocomputing, 2022, 476:126-136.