Oral Assessment Model: Assessing the Quality of Pronunciation in English Reading

Xiaoming Sun

Zhengzhou Shengda University, Zhengzhou, Henan 451191, China

(Received 8 January 2024; accepted 7 April 2024)

When learning English, it is very important to correct the wrong pronunciation. This paper briefly introduces the speech recognition-based mispronunciation recognition algorithm and the mispronunciation recognition algorithm based on prior text. An accent recognition loss function is added to the algorithm based on prior text to improve its performance. After that, the performance of the algorithm is tested by using LibriSpeech and L2-Arctic datasets in the simulation experiment. When the number of hidden layers was 5 and the number of neurons in each hidden layer is 256, the proposed recognition algorithm performs the best. The algorithm converged faster during training, and the loss function is the smallest when it is stable. The proposed algorithm had the best recognition performance on both datasets, and the accent in L2-Arctic has little influence. The average recognition time of the proposed algorithm for the datasets is the least.

NOMENCLATURE

$Y(k)$	the frequency domain signal processed by fast Fourier transform
$y^{(n)}$	the original time domain signal
k	the sampling point number
n	the time sampling point of the time domain signal
$P(\omega)$	the instantaneous energy of $Y(k)$
$H_m(k)$	the frequency response of the triangular filter
m	the serial number in a group of M triangular filters
$c(l)$	the L-order MFCC characteristic parameter
$S(m)$	the energy spectrum function of the signal in the frequency domain after filtering
f_t	the output result of the forget gate
s_t	the output result of the cycle gate
g_t	the output result of the input gate
q_t	the output result of the output gate
h_t	the hidden state in the calculation process
$\omega, \omega_f, \omega_g, \omega_q$	the weight of the corresponding unit to hidden state h_{t-1} at the previous moment
u, u_f, u_g, u_q	the weight of the corresponding unit to current input data x_t
b, b_f, b_g, b_q	the bias term of the corresponding unit
H	intermediate vector sequence
$loss$	the overall loss function of the recognition model
$loss, loss_a$ and $loss_{asr}$	are the error state recognition loss function, accent recognition loss function, and phoneme recognition loss function
α	the weight of accent recognition
β	the weight of phoneme recognition
$\hat{a}$	the predicted accent label
a	the actual accent label
$\hat{P}$	the predicted phoneme sequence
P	the actual phoneme sequence
$CrossEntropy()$	the cross-entropy calculation function
F_1	the comprehensive recognition performance of the recognition model for pronunciation errors
TR	the number of errors that are predicted as errors in reading
FR	the number of errors predicted as errors but actually not in reading
FA	the number of errors predicted as correct
e_1	the error label value of the i-th phoneme in the actual phoneme error label sequence
$\hat{e_i}$	the error label value of the i-th phoneme in the predicted phoneme error label sequence
k	the length of the phoneme sequence
P	the recognition precision
R	the recognition recall rate
$F1$	a composite measure of precision and recall rate

1. INTRODUCTION

As a global language, English pronunciation plays an important role in cross-cultural communication. However, many learners from non-English speaking countries have some difficulties in oral English expression, among which the quality of pronunciation is an important aspect.[1] Accurate pronunciation is essential for effective communication, while mispronunciation will not only lead to difficulties in understanding, but also affect the effect of communication.[2] However, in the learning process, many students will encounter difficulties in pronunciation. In the traditional learning method, pronunciation difficulties are usually raised in the classroom and guided by teachers. However, due to a limited number of teachers compared to the student population, achieving individualized correction becomes challenging. Therefore, an algorithm model that can accurately evaluate English pronunciation[3] is needed to replace the manual evaluation of teachers, enabling learners to autonomously train their own pronunciation. According to the study conducted by Gang,[4] a high-speed hybrid model and artificial emotion recognition were utilized to analyze and eliminate different types of interference that impact the quality of speech, with the aim of enhancing students' ability to recognize English speech. The findings indicated that the performance of the model was highly satisfactory. Zhou et al.[5] proposed a classification model-based algorithm for correcting English grammar errors and verified its effectiveness by simulation experiments. Experiments conducted by Watanabe et al.[6] demonstrated the effectiveness of a combined CTC/attention end-to-end algorithm in automatic speech recognition. This paper briefly introduces the speech recognition-based mispronunciation recognition algorithm and the prior text-based mispronunciation recognition algorithm. An accent recognition loss function was introduced to the prior text-based algorithm to improve its performance. In the simulation experiments, the LibriSpeech and L2-Arctic datasets were used to test the performance of the proposed algorithm. The contribution of this article lies in utilizing an a priori text-based recognition method to identify pronunciation errors and enhancing the performance of pronunciation recognition using accent recognition loss function during training. This paper provides an effective reference for accurate evaluation of reading pronunciation quality.

2. AN EVALUATION ALGORITHM FOR THE QUALITY OF ENGLISH PRONUNCIATION IN READING

In oral English training, students usually read along with the standard pronunciation to learn the characteristics of pronunciation. However, for non-native English students, the pronunciation rules of English are always different from their mother tongue. When reading, they will not only mispronounce terms because they are not familiar with the vocabulary,[7] but also cause inaccurate pronunciation because of the accent brought on by their mother tongue. In the traditional way of learning, teachers generally point out mistakes and correct them. However, given the larger number of students compared to teachers, it becomes impractical for teachers to correct each mistake individually. Moreover, if students are left to learn on their own, they may struggle in recognizing their own errors. With the emergence and development of intelligent algorithms, their scope of application is gradually expanding. The evaluation of oral English pronunciation required by this paper is the field where intelligent algorithms can be applied.

2.1. Oral Pronunciation Quality Assessment Based On Speech Recognition

Pronunciation evaluation of spoken English is a sequence-to-sequence speech recognition task, which means recognizing the audio feature sequence of speech as a phoneme sequence. In the face of sequence-to-sequence recognition tasks, long short-term memory (LSTM) is a commonly used algorithm.[8] Compared with other kinds of deep learning algorithms, LSTM extended from the recurrent neural network (RNN) can effectively use the context of sequence data, and its gate mechanism can effectively avoid the disappearance of gradients caused by long sequences. When converting from sequence to sequence, the length between the two sequences is not necessarily the same, and the direct conversion can fail to align. Therefore, an encoder-decoder structure can be used to solve this problem by constructing an intermediate sequence.

The basic principle of oral pronunciation quality assessment based on speech recognition is to first convert the spoken English audio into a phoneme sequence through the encoder-decoder[9] and then compare the recognized phoneme sequence with the actual phoneme sequence to find the pronunciation errors. However, the process of identifying pronunciation errors using this method is not continuous. When performing the pronunciation quality assessment, the default is that the phoneme recognition sequence of the actual pronunciation is correct, and then the actual pronunciation errors are found by comparison. This leads to the evaluation performance of the whole algorithm depending on the recognition accuracy of the pronunciation phonemes, rather than the recognition accuracy of the errors. The actual phoneme text of the target pronunciation is not needed in the training and optimization.

2.2. Prior Text-Based Oral Pronunciation Quality Assessment

The assessment method described above is divided into two parts. Firstly, phoneme recognition is performed on the user's actual pronunciation, and then the recognition results are compared with the accurate pronunciation phonemes. The two parts are independent of each other. In the training process, only the pronunciation phoneme recognition part is optimized, and the recognition result comparison part cannot be optimized.[10] Therefore, this paper chooses to align the actual pronunciation phonemes and the standard pronunciation phonemes in advance during the training process to obtain the labels of the pronunciation error sequence. Then, the spoken audio and the standard pronunciation phonemes are input into the recognition model to directly predict the pronunciation and error sequence labels, and the recognition model is inversely

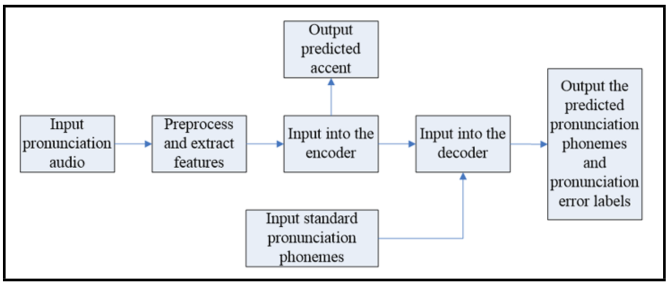

Figure 1. Oral pronunciation quality assessment process based on prior text.

optimized according to the error between the standard error label and the predicted error label.

Figure 1 shows the assessment process of oral pronunciation quality by the proposed recognition model, and the specific procedure is shown below.

1. The audio data of pronunciation is input.

2. The audio data is preprocessed by noise reduction and framing,[11] and then the audio features are extracted. In this paper, the Mel-frequency cepstral coefficient (MFCC) method is selected to extract the MFCC features of the audio.[12] The corresponding formula is:

$$\begin{cases} Y(k) = \sum_{n=0}^{N-1} y(n) \cdot e^{\frac{-2j\pi kn}{N}} \\ P(\omega) = |Y(k)|^2 \\ S(m) = \ln\left(\sum_{k=0}^{N-1} P(\omega) \cdot H_m(k)\right) \\ \sum_{m=0}^{M-1} H_m(k) = 1 \\ c(l) = \sum_{m=0}^{M-1} S(m) \cos\left(\frac{\pi l(2m+1)}{2M}\right) \end{cases} ; \quad l = 1, 2, 3, \cdots, L \quad (1)$$

where $Y(k)$ stands for the frequency domain signal processed by fast Fourier transform (FFT),[13] $y(n)$ indicates the original time domain signal, k indicates the sampling point number, n indicates the time sampling point of the time domain signal, $P(\omega)$ indicates the instantaneous energy of $Y(k)$, $H_m(k)$ is the frequency response of the triangular filter, m indicates the serial number in a group of M triangular filters, $c(l)$ indicates the L-order MFCC characteristic parameter, and $S(m)$ represents the energy spectrum function of the signal in the frequency domain after filtering.

3. The audio features are input into the encoder for forward calculation to obtain intermediate vector sequence . The LSTM serves as the encoder in this paper, and its related formula is:

$$\begin{cases} f_t = \sigma(b_f + u_f x_t + \omega_f h_{t-1}) \\ s_t = f_t s_{t-1} + g_t \sigma(b + u x_t + \omega h_{t-1}) \\ g_t = \sigma(b_g + u_g x_t + \omega_g h_{t-1}) \\ h_t = \tanh(s_t) g_t \\ g_t = \sigma(b_q + u_q x_t + \omega_q h_{t-1}) \end{cases} ; \quad (2)$$

where f_t, s_t, g_t, q_t are the output results of the forget, cycle, input, and output gates,[14] h_t is the hidden state in the calculation process, $\omega, \omega_f, \omega_g, \omega_q$ are the weight of the corresponding unit to hidden state h_{t-1} at the previous moment, u, u_f, u_g, u_q are the weight of the corresponding unit to current input data x_t, and b, b_f, b_g, b_q are the bias term of the corresponding unit. Meanwhile, an additional aggregation layer is incorporated at the output end of the encoder to identify the accent of the user, and the aggregation layer uses the mean-variance statistics to classify intermediate vector sequence H.

4. H is input to the decoder. At the same time, the corresponding standard pronunciation phoneme sequence is also input to the decoder, and the LSTM also serves as the decoder. Finally, the predicted sequence and the error state sequence of the pronunciation phoneme are calculated.

   If the recognition model is in the training stage, it is necessary to calculate the loss function of the recognition model,[15] and the relevant formula is:

$$\begin{cases} loss = loss_e + \alpha \cdot loss_a + \beta \cdot loss_{asr} \\ loss_a = CrossEntropy(\hat{a}, a) \\ loss_{asr} = CrossEntropy(\hat{P}, P) \\ loss_e = 1 - F_1 \end{cases} ; \quad (3)$$

   where $loss$ represents the overall loss function of the recognition model, $loss_e$, $loss_a$, and $loss_{asr}$ are the error state recognition loss function, accent recognition loss function, and phoneme recognition loss function, α is the weight of accent recognition, β is the weight of phoneme recognition, $\hat{a}$ and a are the predicted accent label and the actual accent label, $\hat{P}$ and P are the predicted phoneme sequence and the actual phoneme sequence, $CrossEntropy()$ is the cross-entropy calculation function, and F_1 is the comprehensive recognition performance of the recognition model for pronunciation errors, whose specific calculation is explained in the simulation experiment later. After the loss function of the recognition model is calculated, training is considered complete when either the loss function stabilizes, or the number of iterations reaches a predetermined threshold. Otherwise, the parameters in the model are inversely optimized according to the value of the loss function.

3. SIMULATION EXPERIMENTS

3.1. Experimental Data

Both the LibriSpeech (`www.openslr.org/12/`) and L2-Arctic datasets (`psi.engr.tamu.edu/l2-arctic-corpus/`) were selected for simulation experiments. The LibriSpeech dataset is a large public speech dataset, consisting of speech data from more than 3,000 individuals and covering various speech styles such as everyday conversations, speeches, and news broadcasts. it is mainly used for speech recognition and natural language processing research. The native languages of the speakers in the L2-Arctic dataset include Hindi, Chinese, Spanish, Arabic, and Vietnamese. This dataset is used for researching recognition for speeches with accents and natural language

Table 1. Basic parameters of the recognition model.

	Encoder	Decoder
Number of hidden layers	5	5
Number of neurons in the hidden layer	256	256
Activation function in the hidden layer	Sigmoid	Sigmoid
Learning rate	0.1	0.1
Maximum number of training sessions	1,000	

processing. The required speech data were selected from the above datasets to form the training set and the test set. The final training set had 5,000 speech data from the Fisher dataset and 2,500 speech data from the L2-Arctic dataset. In the test set, there was 1,500 speech data from the Fisher dataset and 500 speech data from the L2-Arctic dataset. The Montreal Forced Aligner program was used for the phonetic labeling of speech data in the dataset. It is a forced alignment tool that can provide time-aligned annotations of audio files.

3.2. Experimental Setup

The quality evaluation algorithm for English oral pronunciation also adopted the encoder-decoder structure, and its basic parameters are shown in Tab. 1. In addition, the MFCC feature was selected as audio feature, and its dimension was set to 39. Weight α was used to measure accent recognition in the loss function used for reversely optimizing the parameters during training was set to 0.1. Weight β was used to measure phoneme recognition was also set to 0.1. Meanwhile, the performance of recognition models with different encoder and decoder parameters were also tested. The parameters for comparison were a hidden layer count and several neurons in each hidden layer. The hidden layer counts were set as 3, 4, 5, 6, and 7 respectively, while the number of neurons in each hidden layer were set as 64, 128, 256, 512, and 1,024.

To further validate the effectiveness of the proposed algorithm, a comparison was conducted with two alternative algorithms. The other two algorithms were based on speech recognition and based on prior text without considering accent, respectively. The former recognizes the phoneme sequence of the speech first and then compares the standard phoneme sequence of the spoken language. The algorithm also adopts the encoder-decoder structure, so its relevant parameters were the same as the prior text-based algorithm. The prior text-based oral pronunciation quality assessment algorithm without considering accent was obtained by setting weight α in the prior text-based algorithm to 0, and the other parameters remained unchanged.

3.3. Evaluation Criteria

The main goal of the oral pronunciation quality evaluation algorithm based on prior text is to identify the errors of the user's pronunciation and assist the user to pronounce accurately. Therefore, the evaluation criteria was used to evaluate the identification performance of the algorithm for correct and incorrect pronunciation, and the corresponding calculation formula is:

Table 2. The $F1$ value of the proposed algorithm under different hidden layer counts and different number of neurons in the hidden layer.

Number of hidden layers	3 layers	4 layers	5 layers	6 layers	7 layers
64	0.831	0.873	0.889	0.878	0.839
128	0.849	0.898	0.907	0.887	0.839
256	0.867	0.916	0.936	0.919	0.868
512	0.847	0.897	0.906	0.885	0.846
1,024	0.829	0.871	0.886	0.871	0.826

$$\begin{cases} TR = \sum_{i=1}^{k}(\hat{e}_i \cdot e_i) \\ FR = \sum_{i=1}^{k}(\hat{e}_i \cdot (1 - e_i)) \\ FA = \sum_{i=1}^{k}((1 - \hat{e}_i) \cdot e_i) \\ P = \frac{TR}{TR+FR} \\ R = \frac{TR}{TR+FA} \\ F1 = \frac{2 \cdot P \cdot R}{P+R} \end{cases}; \quad (4)$$

where TR was the number of errors that are predicted as errors in reading, FR was the number of errors predicted as errors but actually not in reading, FA was the number of errors predicted as correct, e_i represented the error label value of the i-th phoneme in the actual phoneme error label sequence (0 or 1, 0 means correct, 1 means wrong), $\hat{e}_i$ represented the error label value of the i-th phoneme in the predicted phoneme error label sequence, k indicated the length of the phoneme sequence, P indicated the recognition precision, R represent the recognition recall rate, and $F1$ was a composite measure of precision and recall rate and was also the value used to calculate the loss function in the previous text.

3.4. Experimental Results

The $F1$ values of the proposed algorithm under different numbers of hidden layers and hidden layer neurons are shown in Tab. 2. It can be observed that with the same number of hidden layer nodes, the performance of this recognition algorithm initially improves and then declines as the number of hidden layers increased. Similarly, with the same number of hidden layers, the performance also follows an upward-downward trend as the number of hidden layer neurons increases. When there are five layers and 256 neurons in each hidden layer, this recognition algorithm performs the best.

Figure 2 illustrates the convergence curves of the three oral English pronunciation quality recognition algorithms during training. It can be observed from Fig. 2 that as training times rise, the loss functions of the three algorithms gradually decreases and remains stable after a certain number of iterations. Among them, the loss function of the prior text-based algorithm decreases the fastest, the loss function of the prior text-based algorithm without considering the accent decreases the second, and the loss function of the algorithm decreases the fastest. Moreover, the loss function of the prior text-based algorithm added with the accent recognition loss function is also the lowest after the loss function converges to be stable.

The phoneme recognition performance of the three algorithms on the LibriSpeech dataset, which consists of standard speech, and the L2-Arctic dataset with accents, are shown in Fig. 3. It can be observed from Fig. 3 that for both the LibriSpeech dataset and the L2-Arctic dataset, the $F1$ value of

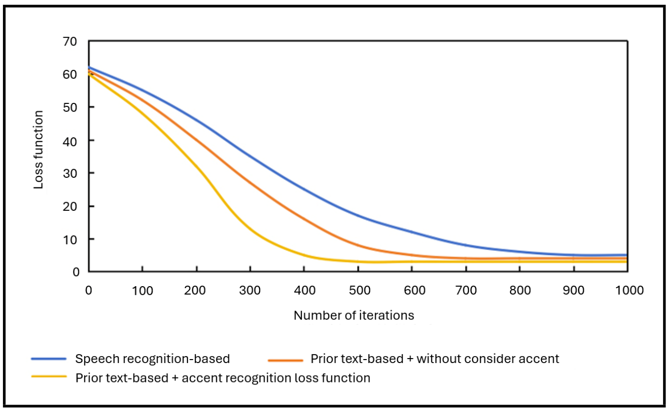

Figure 2. Convergence curves of the three algorithms.

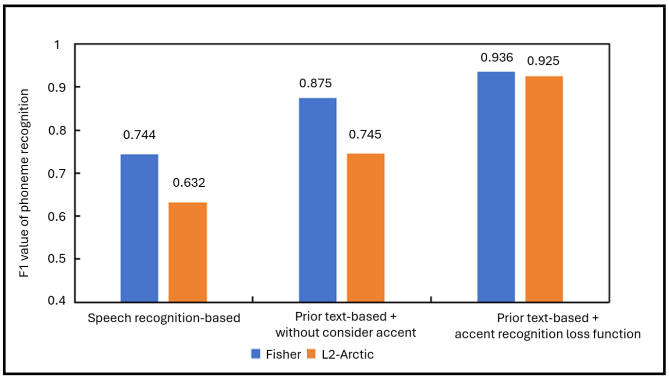

Figure 3. Phoneme recognition performance of the three algorithms.

the prior text-based algorithm is the highest. The prior text-based algorithm without considering accent is the second, and the speech recognition algorithm is the lowest. The horizontal comparison of the phoneme recognition performance of the same algorithm on different datasets showed that the phoneme recognition performance of the speech recognition-based algorithm and the prior text-based algorithm without considering the accent are significantly reduced when facing the L2-Arctic dataset with different accents, and the prior text-based algorithm added with the accent recognition loss function also experiences a decrease, but it is relatively insignificant.

Figure 4 illustrates the average duration taken by the three algorithms in recognizing pronunciation errors in speech data. It is evident from Fig. 4 that, compared with the algorithm based on speech recognition, the average time of the prior text-based algorithm without considering accent and the proposed algorithm is significantly smaller, and the average recognition time of the proposed algorithm is the least.

4. DISCUSSION

Automated evaluation of English pronunciation not only reduces the workload for teachers but also provides effective reference for students to practice speaking independently. This paper first briefly introduces a speech pronunciation quality evaluation algorithm based on speech recognition and a prior text-based algorithm for evaluating the quality of spoken pronunciation. To enhance its evaluation performance, an accent recognition loss function is introduced during the training process. Subsequently, simulation tests are conducted on the pro-

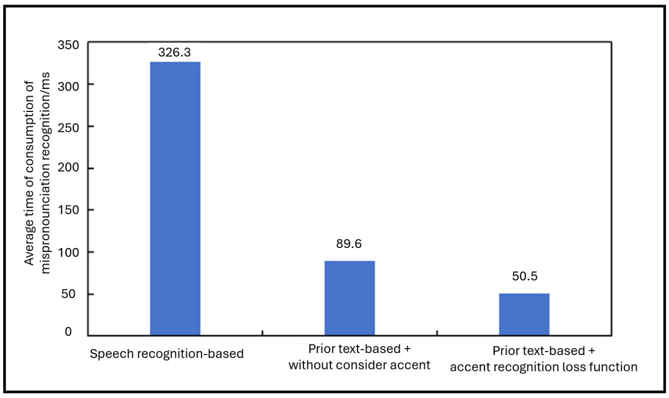

Figure 4. The average time consumption of mispronunciation recognition by the three algorithms.

nunciation quality evaluation algorithm to examine the impact of different numbers of hidden layers and neurons in LSTM on algorithm performance. The pronunciation evaluation algorithms based on speech recognition, prior text-based without considering accents, and prior text-based with consideration of accents are compared. The results are shown in the previous section.

When the number of hidden layers and neurons in the LSTM used in both the encoder and decoder increases, it becomes more capable of capturing intricate hidden patterns, resulting in improved recognition performance. However, once it reaches a certain point, the discovered patterns become overly detailed, which not only reduces computational efficiency but also decreases the proportion of effective information provided by these intricate patterns. Moreover, the excessively detailed patterns may also impact the model's generalization ability. Among the three comparison algorithms, the proposed pronunciation evaluation algorithm based on prior text and considering accents performed the best. The reason behind this is that the pronunciation evaluation algorithm based on speech recognition first identifies phonemes in the pronunciation and then compares them with a standard phoneme sequence. These two parts are independent of each other and have their own recognition errors, which leads to error accumulation. On the other hand, the pronunciation evaluation algorithm based on prior text aligned actual pronunciation phonemes with standard ones during training, obtaining the label for pronunciation error sequences. This label is then used to train the algorithm, combining phoneme recognition and comparison into one step to reduce error accumulation. Additionally, the proposed algorithm introduces a loss function for accent recognition, allowing it to consider accent influence during training. Hence, the proposed algorithm achieves superior performance.

5. CONCLUSIONS

This paper briefly introduces the speech recognition-based mispronunciation recognition algorithm and the prior text-based algorithm. An accent recognition loss function is introduced to the algorithm based on prior text to improve its performance. After that, the performance of the algorithm is tested by using LibriSpeech and L2-Arctic datasets in the simulation experiment. The outcomes are presented below.

1. When the number of hidden layers is 5 and the number of neurons in each hidden layer is 256, the proposed recognition algorithm performs the best.

2. The loss function of the prior text-based algorithm added with the accent recognition loss function decreases fastest, that of the recognition algorithm based on prior text without considering accent decreases second, and that of the algorithm based on speech recognition decreases fastest.

3. For both LibriSpeech dataset and L2-Arctic dataset, the prior text-based algorithm added with the accent recognition loss function has the highest $F1$ value, followed by the prior text-based algorithm without considering accent, and the speech recognition algorithm has the lowest $F1$ value.

4. When the L2-Arctic dataset is tested, compared with the Fisher dataset, the performance of the three algorithms is reduced, but only the prior text-based algorithm added with the accent recognition loss function is not significantly reduced.

5. Compared to the algorithm based on speech recognition, the average identification time taken by the prior text-based algorithm without considering accent and the prior text-based algorithm added with the accent recognition loss function is significantly smaller, and the time of the latter is the least.

REFERENCES

[1] Mann, V. and Venkatasubramanian, V. Retrosynthesis Prediction using Grammar-based Neural Machine Translation: An Information-Theoretic Approach, *Comput. Chem. Eng.*, **155**, 107533-1 - 107533-15, (2021). https://doi.org/10.1016/j.compchemeng.2021.107533

[2] Tripathi, S. and Kansal, V. Machine Translation Evaluation: Unveiling the Role of Dense Sentence Vector Embedding for Morphologically Rich Language, *Int. J. Pattern Recogn.*, **34** (1), 2059001.1 - 2059001.18, (2020). https://doi.org/10.1142/S0218001420590016

[3] Eddine, M. A New Concept of Electronic Text Based on Semantic Coding System for Machine Translation, *ACM T. Asian Low-Reso.*, **21** (1), 1-16, (2022). https://doi.org/10.1145/3469655

[4] Gang, Z. Quality evaluation of English pronunciation based on artificial emotion recognition and gaussian mixture model, *J. Intell. Fuzzy Syst.*, **40** (4), 7085-7095, (2021). https://doi.org/10.3233/JIFS-189538

[5] Zhou, S. and Liu, W. English Grammar Error Correction Algorithm Based on Classification Model, *Complexity, 2021 (Pt.2)*, 6687337-1-6687337-11, (2021). https://doi.org/10.1155/2021/6687337

[6] Watanabe, S., Hori, T., Kim, S., Hershey, J. R. and Hayashi, T. Hybrid CTC/Attention Architecture for End-to-End Speech Recognition, *IEEE J. STSP*, **11** (8), 1240-1253, (2017). https://doi.org/10.1109/JSTSP.2017.2763455

[7] Lee, J. W. A comparison study on EFL learner and teacher perception of grammar instruction and error correction, *English Teach.*, **73** (2), 139-159, (2018). https://doi.org/10.5539/elt.v4n4p36

[8] Barzang, E. Fostering EFL Learners' Grammar Achievement Using Recasts and Meta-linguistic Awareness Error Correction Feedbacks, *J. Stud. English Linguist.*, **8** (1), 35-46, (2019). https://doi.org/10.5296/jsel.v8i1.16065

[9] Zhu, J., Shi, X. and Zhang, S. Machine Learning-Based Grammar Error Detection Method in English Composition, *Sci. Programming, 2021 (Pt.13)*, 1-10, (2021). https://doi.org/10.1155/2021/4213791

[10] O'Brien, J. Consciousness-raising, Error Correction and Proofreading, *J. Scholarship Teach. Learn.*, **15** (3), 85-103, (2015). https://doi.org/10.14434/josotl.v15i3.13284

[11] Eckstein, G. Grammar Correction in the Writing Centre: Expectations and Experiences of Monolingual and Multilingual Writers, *Can. Mod. Lang. Rev.*, **72** (3), 1-23, (2016). https://doi.org/10.3138/cmlr.3605

[12] Evers, K. and Chen, S. Effects of Automatic Speech Recognition Software on Pronunciation for Adults With Different Learning Styles, *J. Educ. Comput. Res.*, **59** (4), 669-685, (2021). 10.1177/0735633120972011

[13] Cao, Q. and Hao, H. Optimization of Intelligent English Pronunciation Training System Based on Android Platform, *Complexity, 2021 (Pt.10)*, 1-11, (2021). https://doi.org/10.1155/2021/5537101

[14] Zhan, W. and Chen, Y. Application of machine learning and image target recognition in English learning task, *J. Intell. Fuzzy Syst.*, **39** (4), 5499-5510, (2020). https://doi.org/10.3233/JIFS-189032

[15] Sui, P. Research on Interactive English Speech Recognition Algorithm in Multimedia Cooperative Teaching, *Int. English Educ. Res.*, (1), 79-82, (2018). https://doi.org/10.1109/ICITBS.2018.00095