Hindawi
Scientific Programming
Volume 2022, Article ID 4774677, 10 pages
https://doi.org/10.1155/2022/4774677

Research Article

Automatic Assessment Method of Oral English Based on Multimodality

Xiaoyan Chen

School of Education, Shanghai Aurora College, Shanghai 201908, China

Correspondence should be addressed to Xiaoyan Chen; xy.chen@aurora-college.cn

Received 8 November 2021; Revised 20 December 2021; Accepted 27 December 2021; Published 19 January 2022

Academic Editor: Fazli Wahid

With the rapid development of Internet technology and educational informatization, there are more and more oral materials available on the Internet. Therefore, how to adapt to learners' dynamic abilities and provide them with personalized learning materials has become a very important issue in educational technology. Aiming at the inefficiency of the existing automatic assessment of spoken English, a multimodal-based automatic assessment method of spoken English is proposed. The Word2Vec model is used to extract text features, and then, speech and text are input into GRU temporal structure, and encoder coding is used for multimodal fusion to realize automatic evaluation of multimodal spoken English. Simulation results show that the multimodal model proposed in this paper is superior to the traditional oral English automatic assessment model in terms of fluency, emotional expression, and sense of rhythm and can better improve learners' oral English level.

1. Introduction

With the importance of international English communication, how to enhance learners' oral English level has become the focus of current thinking. However, in traditional oral English assessment, manual scoring is usually adopted, which not only inefficient and impartial is but also cannot objectively evaluate learners' oral English level. For this reason, many scholars propose to combine modern information technology to improve learners' oral level, such as Constantinides G A et al. [1] build an automatic speech assessment model according to speech signals, Kovacs G et al. [2] use deep learning methods to optimize the automatic speech evaluation model, and Zhang and Qin [3] evaluate speech and use DTW template matching to establish an automatic evaluation model. Yi et al. [4] use multimodal neural network to predict the evaluation of press conference professionally by using collected text and audio data. It is composed of three parts: language model, audio model, and feature fusion network. There are two kinds of feature fusion methods in the audio model of multimodal neural network, which are shared attention network, text feature generation, and audio feature generation. Compared with the former, the performance of the latter is much better, and the accuracy rate is generally 60%. In [5], the automatic evaluation method of speech quality based on speech signal detection and dynamic synchronous recognition can be used for automatic evaluation of spoken English. Using DSP chip, the design is divided into two parts, which are speech signal processing algorithm design and hardware circuit design. Sound sensor is used to collect sound signals and can decompose and analyze the features of spoken speech and extract the features of speech signals, so as to recognize the quality of spoken speech. In today's universities, the evaluation system of college English teaching writing plays an important role in promoting online courses [6]. The system uses ASP.NET, SQL Server, and other technologies to achieve, for the need to upload courseware, documents and students' homework, after-school communication, and so on to provide a platform. In [7], the Internet of Things used to obtain intelligent data is the current process evaluation of college English based on the Internet of Things, which can not only increase the flexibility of learning but also analyze and judge the generated data, so as to judge whether the process evaluation can be satisfied in English teaching. Su et al. [8] use Sugeno integral technology to solve the problems existing in oral English speech evaluation. It is a technology to ensure the reliability of speech processing system and integrate learners' personality into the evaluation system. The first step is to collect English phonemes (HDP) and classify them into different HDP sets. The second step is to recognize English phonemes in the set and then integrate them under the Sugeno integral framework. Bauman [9] points out that more and more Asians are learning English, which leads to the stereotype of Asian accent. Referring to Americans' evaluation of Asian accents, several different characteristics were rated in three dimensions: attractiveness, status, and vitality. Finally, it shows that most people's evaluation of English Asian accent is negative. Compared with mainstream American English, Asian accent English, and Brazilian Portuguese accent English, RM variance analysis ($N = 69$) showed that they were all low. Suzuki et al. [10] focus on the rhythm of speech in the process of communication between people. In many auxiliary language learning systems, the accuracy of prosody evaluation is still not enough. A new evaluation method is used to evaluate the rhythm and prosody of spoken English. It uses the least square method to automatically estimate the importance factor of words in the calculation of intonation score. Through experiments, the correlation coefficient between rhythm score and system is -0.55, while the conventional one is only-0.11, but 1.0 is the best. Other applications [11–15] in spoken English are based on related algorithms, and the application of intelligent methods speeds up the application of automatic English translation.

However, the abovementioned automated oral English assessment method only considers the phonetic level, does not involve the text content, and cannot accurately reflect all the information of oral English. To solve the above problems, this paper proposes a multimodal approach for automatic assessment of oral English, which combines pronunciation and text, and constructs an automatic assessment model of oral English by means of joint learning, so as to provide a more effective auxiliary tool for oral English scoring.

2. Word2Vec Text Feature Extraction

The appearance of the language model promotes the ability of natural language processing. The Word2Vec language model [16, 17] has strong text representation ability and can learn deep text relations, which can solve the problem of polysemy. Therefore, this paper chooses Word2Vec to extract text features. The parameter settings for the Word2Vec model are shown in Table 1.

The Word2Vec model is characterized by supporting CBOW and Skip-Gram [18]. Because the CBOW model is a single hidden layer fully connected neural network, its structure is very simple and it is easy to train the model, so this paper chooses the CBOW method to train the Word2Vec model. CBOW algorithm calculates the probability of a word according to the C words before and after the word. The specific principle is shown in Figure 1.

The CBOW algorithm is defined as follows.

Let NEG(w) represent a negative sample set; then, define positive and negative sample labels as

$$L^{w}(\tilde{w}) = \begin{cases} 1, & \tilde{w} = w, \\ 0, & \tilde{w} \neq w. \end{cases} \tag{1}$$

Maximize the positive sample to obtain

$$g(w) = \prod_{N\in\{w\}\cup NEG(w)} p(u \mid \text{Context}(w)). \tag{2}$$

Among them,

$$p(u \mid \text{Context}(w)) = \left[\sigma\left(\overline{x}_{w}^{T}\theta^{u}\right)\right]^{L^{w}(u)}\left[1 - \sigma\left(\overline{x}_{w}^{T}\theta_{j-1}^{w}\right)\right]^{1-L^{w}(u)}. \tag{3}$$

The objective function is obtained by taking logarithm as follows:

$$\begin{aligned} l &= \log \prod_{w\in C} g(w) = \sum_{w\in C} \log\, g(w) \\ &= \sum_{w\in C} \log \prod_{w\in\{w\}\cup NEG(w)}^{l^{w}} \left\{\left[\sigma\left(\overline{x}_{w}^{T}\theta^{u}\right)\right]^{L^{w}(u)} \bullet \left[1 - \sigma\left(\overline{x}_{w}^{T}\theta^{u}\right)\right]^{1-L^{w}(u)}\right\} \\ &= \sum_{w\in C} \sum_{w\in\{w\}\cup NEG(w)} \left\{L^{w}(u) \bullet \log\left[\sigma\left(\overline{x}_{w}^{T}\theta^{u}\right)\right] + \left[1 - L^{w}(u)\right] \bullet \log\left[1 - \sigma\left(\overline{x}_{w}^{T}\theta^{u}\right)\right]\right\}. \end{aligned} \tag{4}$$

The gradient is

$$\frac{\partial l(wu)}{\partial \theta^{u}} = \left[L^{w}(u) - \sigma\left(\overline{x}_{w}^{T}\theta^{u}\right)\right]\overline{x}_{w}. \tag{5}$$

Therefore, the update method of θ^{u} is

$$\theta^{u} = \theta^{u} + \eta\left[L^{w}(u) - \sigma\left(\overline{x}_{w}^{T}\theta^{u}\right)\right]\overline{x}_{w}. \tag{6}$$

Using symmetry to get the gradient of $\overline{x}_{w}$,

$$\frac{\partial l(w.u)}{\partial \overline{x}_{w}} = \left[L^{w}(u) - \sigma\left(\overline{x}_{w}^{T}\theta^{u}\right)\right]\theta^{u}. \tag{7}$$

Update the above gradient and apply it to each word vector results in

$$v(\tilde{w}) := v(\tilde{w}) + \eta \sum_{u\in\{w\}\cup NEG(w)} \frac{\partial l(w.u)}{\partial \overline{x}_{w}}, \quad \tilde{w} \in \text{Context}(w). \tag{8}$$

The input to the Word2Vec model is the one-hot vector representation of each English word. Let V denote the size of corpus, C denote the size of context window, N denote the dimension size of word vector finally trained, and w denote the shared weight matrix of training model learning; then, $w = V \times N$. The $1 \times V$ output vector of the model can be

Table 1: Training parameters of the Word2Vec model.

Parameters	Historical experience value
Generator vector maintenance size N	415
Window size C	6
Minimum word frequency	1
Training model	CBOW
Optimization model	Hierarchical softmax

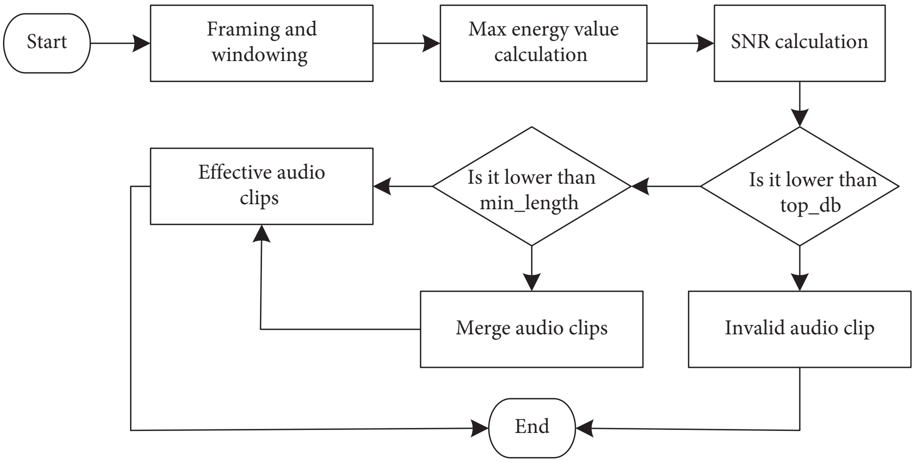

Figure 1: The CBOW principle.

obtained by softmax activation function, and the loss value can be obtained by subtracting this vector with the real one-hot vector.

If the above methods are directly used for training, there are problems such as high requirements for hardware performance and long training time. Therefore, it is necessary to optimize the training process of the Word2Vec model. Because of the fast training speed and good performance of Hierarchical Softmax, this paper chooses the Hierarchical Softmax method [19, 20] based on adjusted structure to optimize the training process. By optimizing the output layer to build a tree structure, the multiclassification problem is transformed into $\log(V)$ binary classification problem, which greatly improves the training efficiency of the model.

3. Automatic Oral English Assessment Model Based on Multimodality

3.1. Overall Frame Design of Multimodal Model. The overall structure of the multimodal-based oral English automatic assessment model includes text module, multimodal fusion module, and speech module, as shown in Figure 2. The whole structure of the model connects the modules through a gated loop unit (GRU). The speech module is a speech data input module and the text module is a text data input module. These data are input into the GRU timing structure, and the last output timing feature is the encoder coding representation of the speech module and the text module. Then, the encoder encodes into the multimodal fusion module. In this module, the speech modal features and text modal features are effectively fused by interactive calculation, and the fused features are taken as the final coding of multimodal processing. Finally, the prediction scores are output through softmax function and full connection layer.

3.1.1. Input Structure Design of the Model. According to the overall framework of the multimodal oral assessment model, the input structure of the model includes two parts.

The first part is the voice module. This part includes prosody features and MFCC cepstrum features. Prosody feature vectors reflect the whole situation of speech data, do not participate in the training of sequence models, and perform mosaic operation when the speech sequence models are output at the last moment. MFCC features exist in the form of a matrix in the speech module. Each frame of MFCC can generate 39-dimensional feature vectors, and different data have different frame lengths. Table 2 shows the distribution of data numbers in different frame lengths. As can be seen from the table, there is a big difference in the number of data in different frame length ranges. Therefore, it is necessary to unify the voice data form. Set the fixed frame length to 2377 frames, which can cover 99% of the data; (2377, 99) is the size of the MFCC feature matrix for each speech, and the prosody feature vector generated for each speech is 35 dimensions.

The speech module inputs the MFCC frame sequence into the sequence model according to the time sequence, and its structure is shown in Figure 3. Each MFCC frame vector is 39 dimensions, and the prosody speech feature vector is 35 dimensions.

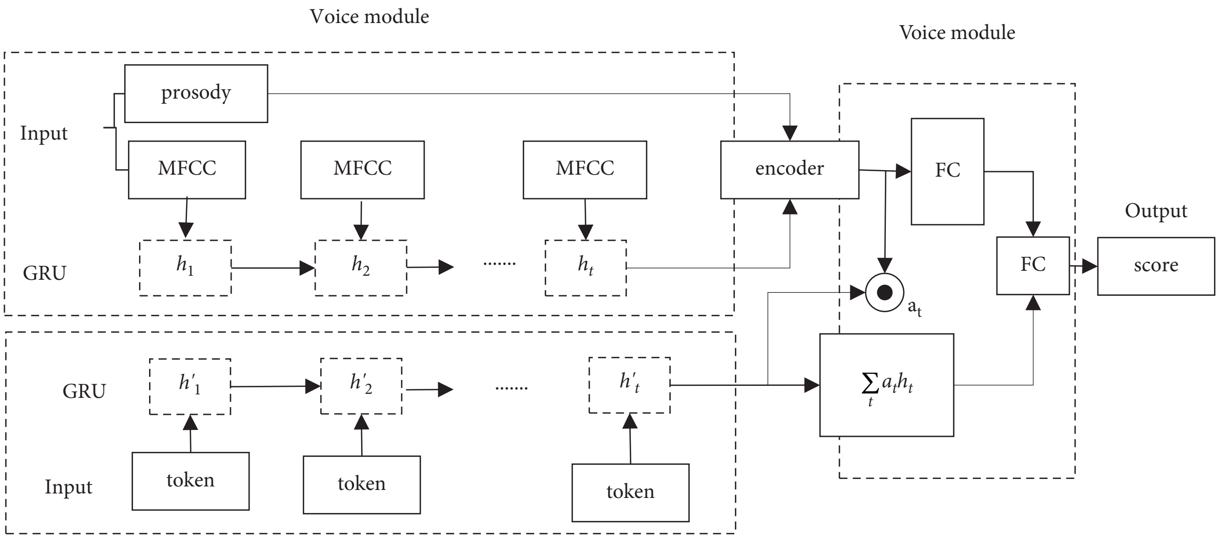

Figure 2: The framework of a multimodal spoken English assessment model.

Table 2: Frame length distribution of voice data.

Frame length	0–1500	1500–1700	1700–1900	1900–2100	2100–2300	2300–2500	2500-
Number	792	525	549	523	388	261	389

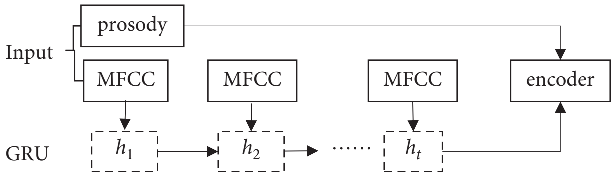

Figure 3: Input structure of speech module in the multimodal model.

The second part is the text module, and its structure is shown in Figure 4. The data in the text module are all English data, and a token sequence can be formed by segmenting the data and removing stop words because the length of token in different text data is different, and the length difference is large. Therefore, it is necessary to unify the text data form. Let the fixed token length be 256, covering 99% of the text data. Then, the token words are transformed into vectors that can be recognized by computer and encoded to generate 400-dimensional word vectors, that is, the text module input is completed. Thereby, the feature matrix size of each text data can be determined to be (256,400).

3.1.2. Timing Structure Design. In this paper, GRU temporal structure is selected as the temporal structure of the multimodal oral English automatic assessment model. Because the model includes two modules, speech module and text module, the temporal structure is also divided into speech module temporal structure and text module temporal structure. The input structure of voice module and text module are closely related, and the length of time sequence structure is the length of input data, so the research sets the number of layers and the number of cells of the two time sequence structures as 1 and 200, respectively. Specific parameter settings are shown in Table 3.

3.1.3. Multimodal Fusion Structure Design. The multimodal fusion structure in this paper is a structure based on attention mechanism, which can realize multimodal fusion by fusing speech coding and text coding. Let *XF* represent the original speech feature, and the attention module takes it as input to obtain the formula:

$$\begin{aligned} z_{i,j} &= F\left(w^T x_{i,j} + b\right), \\ a_{i,j} &= \frac{z_{i,j}}{\sum_{i,j} z_{i,j}}, \end{aligned} \tag{9}$$

where *F* represents the nonlinear activation function, and the enhanced speech features are obtained by combining the above formula:

$$\widehat{x}_{i,j} = \left(1 + a_{i,j}\right) x_{i,j}. \tag{10}$$

By expanding the attention mechanism space, we obtain

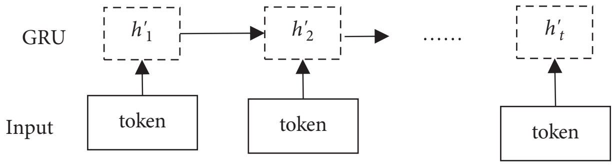

Figure 4: Input structure of this module in the multimodal model.

Table 3: Parameter setting of timing structure of the multimode model.

	Parameter name	Parameter value	Parameter description
Voice module	encoder_size_audio	2369	GRU length
	num_layer_audio	1	GRU number of structural layers
	hidden_dim_audio	202	GRU number of units
Text module	encoder_size_text	249	GRU length
	num_layer_text	1	GRU number of structural layers
	hidden_dim_text	202	GRU number of units

$$
\begin{aligned}
z_{i,j}^{c} &= F\left(w_c^T x_{i,j}^{c} + b^c\right),\\
a_{i,j}^{c} &= \frac{z_{i,j}^{c}}{1+\exp\left(-z_{i,j}^{c}\right)},
\end{aligned} \tag{11}
$$

where C represents the Cth channel; the enhanced spatial characteristics are expressed by the following formula:

$$\widehat{x}_{i,j}^{c} = \left(1 + a_{i,j}^{c}\right)x_{i,j}^{c}. \tag{12}$$

The specific steps of multimodal fusion are as follows. The first step is to multiply the coding of text vector and speech vector and obtain the attention weight of text vector coding by activating function softmax, as shown in formula (15). The second step is to obtain the sum of the attention weight of the text vector encoding and the weight of the text vector encoding, thereby obtaining the final text attention vector encoding, as shown in formula (16). The third step is to splice the text attention vector coding and the speech vector coding to obtain the multimodal fusion vector coding, as shown in Formula (15), and use it as the final output of the multimodal fusion module:

$$a = \operatorname{soft\,max}(te * ae), \tag{13}$$

$$\mathrm{ate} = \sum_{i=1}^{T} a_i te, \tag{14}$$

$$fe = \operatorname{concat}(ae, \mathrm{ate}). \tag{15}$$

3.2. Model Hyperparameter Tuning. The multimodal model includes hyperparameters such as loss function, activation function, optimizer, learning rate, and batch size. These parameters have a certain impact on the model performance, so this paper selects and optimizes these hyperparameters. In deep neural network, different activation functions should be selected for different application scenarios. Softmax activation function and sigmoid activation function are often used in multiclassification tasks, and sigmoid function is the most used activation function in two-classification tasks, which mainly realizes the application of multiclassification tasks in multiple two-classification tasks, and these two-classification tasks are interrelated. The application scenario in this paper belongs to multiple categories, and multiple categories exist independently and are not related, so sigmoid activation function is not suitable for this paper. Softmax activation function is used to deal with multiclassification tasks, and multiple categories are independent of each other, so the independent categories can be normalized by softmax activation function. Therefore, this paper chooses softmax activation function to deal with multimodal models.

Besides selecting activation function, other parameters in model training need to be set, such as loss function, learning rate, and optimizer. The task of oral English automatic assessment is a classification problem. When calculating the loss value, the cross-entropy loss function (such as equation (16)) is selected to calculate it. The update mode of network parameters is determined by the optimizer. There are many types of optimizers, and different optimizers need to be used in different scenarios. Because Adam optimizer is the most widely used optimizer with the best application effect, Adam optimizer is selected to update network parameters. Learning rate has an influence on model training efficiency and model performance. In order to prevent overfitting, the dropout operation is used to encode and set the speech and text, and the network parameters are adjusted according to the model performance:

$$\mathrm{Loss} = -\sum_{i} y_i \ln a_i. \tag{16}$$

4. Experiment and Analysis

4.1. Model Validation. In order to verify the performance of the multimodal oral English automatic assessment model, this paper evaluates the model from the modules of emotional expression, fluency, and sense of rhythm, and

Table 4: Scores the sense of rhythm.

	Consistency rate	Adjacent consistency rate	Root mean square error
Naive Bayes	0.2940	0.9350	0.9489
SVM	0.5712	0.9719	0.7159
MA_SE + Audio + Text + LSTM	0.6095	0.9210	0.9809
MA_SE + Audio + LSTM	0.6986	0.9922	0.5703
MA_SE + Audio + Text + GRU	0.7349	0.9491	0.6881
MA_SE + Audio + GRU	0.7479	0.9919	0.5252

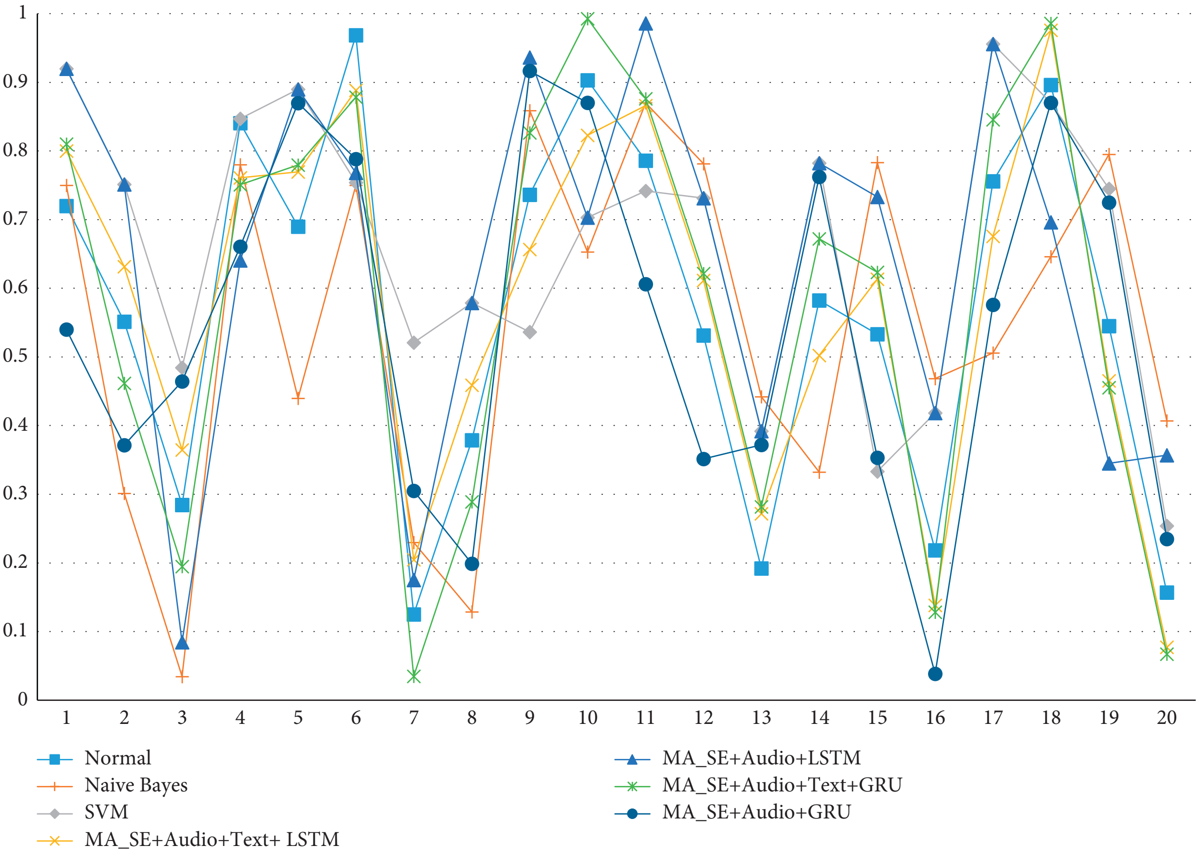

Figure 5: Comparison of the predictive value of rhythm under different models.

compares the proposed model (MA_SE + Audio + Text) with the traditional machine learning model. In order to make the comparative experiment rigorous, before the experiment, the data are resampled and denoised, and the input features of the two traditional machine learning and deep learning models are also set as MFCC features and prosody features. The features are transformed into the required feature vectors through normalization processing, and the input features of the pure speech model and the multimodal model are consistent. Then, the parameters of the traditional machine learning model are optimized. Finally, the scoring results of different models on oral English rhythm, as shown in Table 4, are obtained.

It can be seen from Table 4 that compared with the two traditional machine learning models, the multimodal model and pure speech model of GRU structure are higher than the traditional model in terms of consistency rate and adjacent consistency rate. Compared with the multimodal model using LSTM structure, the experimental results using the GRU temporal structure model are higher. The experimental results of pure speech are higher than those of multimodal experiments. Therefore, deep learning has great influence on rhythm, and GRU structure has better performance than LSTM structure, while text has little effect on rhythm score.

Through the analysis of the rhythm prediction results of the six models, we can see that MA_SE + Audio + Text has the best performance. The results of rhythm scoring were evaluated, and 20 samples were randomly selected for prediction. The prediction effect is shown in Figure 5.

Table 5 shows the results of fluency scores of different models. It can be seen from the table that the multimodal model using GUR time-series structure has the best scoring result. The multimodal deep learning model using GRU is higher than the pure speech model with GRU structure in

Table 5: Scores the fluency.

	Consistency rate	Adjacent consistency rate	Root mean square error
Naive Bayes	0.2941	0.9159	0.9825
SVM	0.4633	0.9656	0.7996
MA_SE + Audio + GRU	0.6109	0.9703	0.6968
MA_SE + Audio + LSTM	0.6456	0.9874	0.6259
MA_SE + Audio + Text + LSTM	0.6488	0.9582	0.7172
MA_SE + Audio + Text + GRU	0.6669	0.9779	0.6441

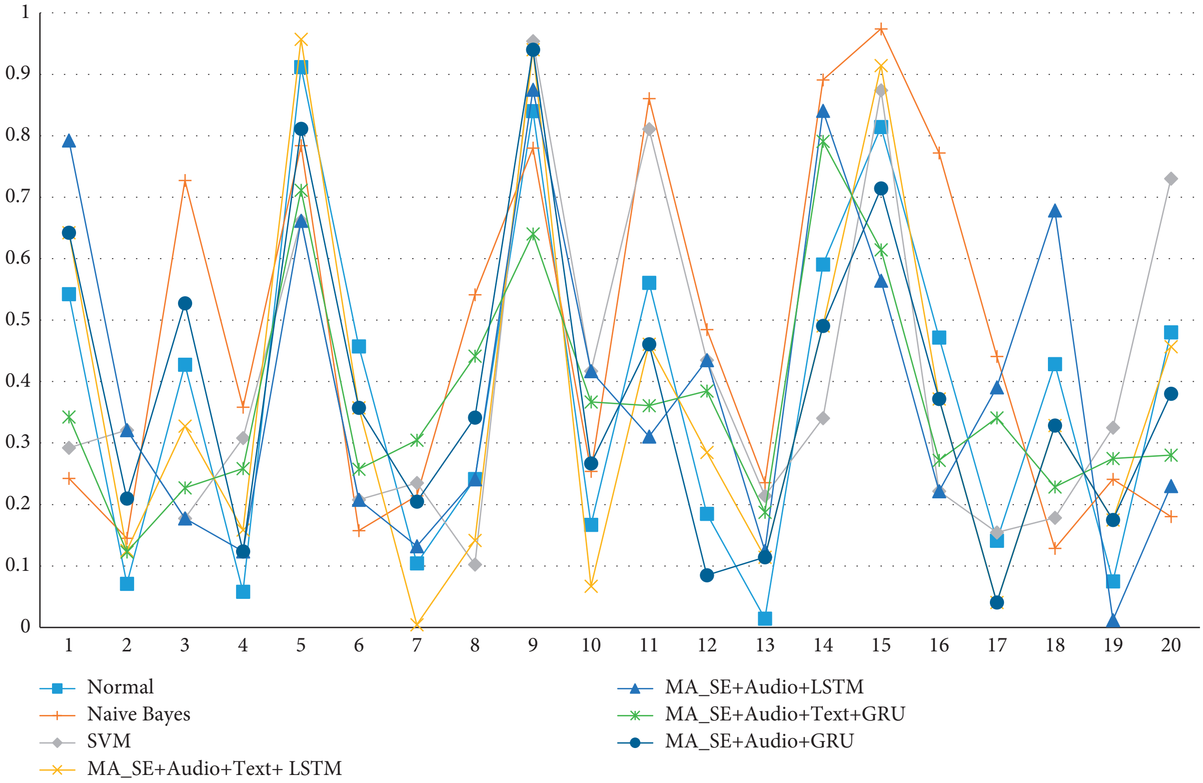

Figure 6: Comparison of predicted values of profitability under different models.

terms of consistency index. Compared with GRU structure and LSTM structure, the experimental results of the multimodal model with GRU structure are higher, while the experimental results of the pure speech model with LSTM structure are higher than those of the pure speech model with GRU structure. This shows that, in the fluency scoring module, the deep learning method plays a great role in it. Text information also has a great influence on fluency.

Through the analysis of the fluency prediction results of the six models, we can see that MA_SE + Audio + Text shows the best performance. Evaluate the scoring results of profitability, and randomly select 20 samples for prediction. The prediction effect is shown in Figure 6.

In Table 6, among the results of emotion scoring module, the consistency evaluation index of the multimodal model combining GRU, speech, and text is the best. The model using LSTM and pure speech has the best experimental results in terms of adjacent coincidence rate and root mean square error. On the whole, in the consistency index, compared with the pure speech model and the multimodal model, the multimodal model has better performance. This shows that the introduction of deep learning methods and texts plays a great role in the emotional performance scoring module.

Through the analysis of the emotion prediction results of the six models, we can see that MA_SE + Audio + Text has the best performance. The results of emotional scoring were evaluated, and 20 samples were randomly selected for prediction. The prediction effect is shown in Figure 7.

In Figures 5–7, 20 random samples are selected to compare and analyze different algorithms and combined algorithms. The application scenarios of data samples are different, and many application scenarios in the same figure are different. Compared with the model in the same scene, the experimental effect is better in different scenes.

4.2. Example Verification. In order to explore the practical application effect of the multimodal method of combining speech and text, the multimodal model is applied

Table 6: Scores the emotional performance.

	Consistency rate	Adjacent consistency rate	Root mean square error
Naive Bayes	0.2318	0.7039	0.3522
SVM	0.4189	0.9382	0.8852
MA_SE + Audio + GRU	0.6436	0.9632	0.7031
MA_SE + Audio + LSTM	0.6132	0.9745	0.6845
MA_SE + Audio + Text + LSTM	0.6559	0.9636	0.7122
MA_SE + Audio + Text + GRU	0.6889	0.9388	0.7254

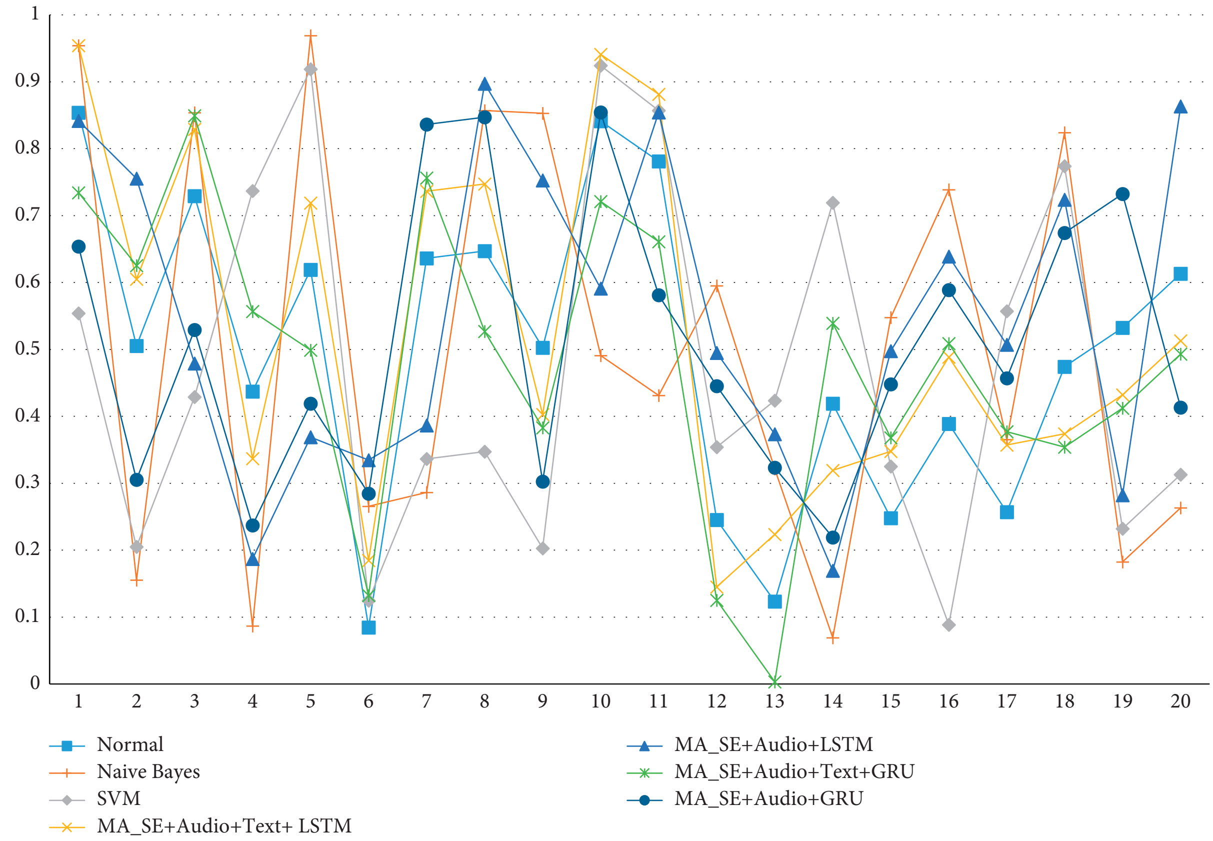

Figure 7: Comparison of emotion prediction values under different models.

> "Garbage sorting is to separate waste into different categories, mainly wet or dry. It's commonly applied in major cities and can be seen as a critical part of the whole process of garbage recycling. To realize environmental sustainability, garbage sorting is a necessity. First of all, waste sorting regulations are needed and communities should be provided with necessary instructions and facilities. Secondly, citizens should be fully informed of garbage sorting knowledge so as to improve the efficiency in the whole process. Finally, only when the whole society is mobilized can we truly bring some positive changes and build an environment-friendly society. So education will be very important."

Figure 8: Text example of unsmooth data.

to specific data examples, and the results are shown in Figure 8. In Figure 8, red indicates the unfluent part, reflecting the intermittent and tense performance of the speaker. There are many unsmooth parts of the data, so it is difficult to judge the data correctly only by voice data, but the model proposed in this study can directly and accurately judge the data as unsmooth. Therefore, the multimodal method combining speech and text can fully judge speech information and has good application effect. The comparison in the experiment is based on the combined application of various algorithms. The time complexity of the algorithm will increase, but in order to improve the efficiency of the algorithm, the time complexity is acceptable. Therefore, in consideration of the time complexity under the condition, it is advantageous to improve the accuracy of the algorithm.

"Garbage sorting is to separate waste into different categories, mainly wet or dry. It is commonly applied in major cities and can be seen as a critical part of the whole process of garbage recycling. To realize environmental sustainability, garbage sorting is a necessity. First of all, waste sorting regulations are needed and communities should be provided with necessary instructions and facilities. Secondly, citizens should be fully informed of garbage sorting knowledge so as to improve the efficiency in the whole process. Finally, only when the whole society is mobilized can we truly bring some positive changes and build an environment-friendly society. So, education will be very important."

5. Conclusion

To sum up, compared with the traditional models, the multimodal model proposed in this paper has better performance. In this paper, the multimodal deep learning method combining GRU, speech, and text has auxiliary effect on fluency and emotional expression, which depend on text, but has little auxiliary effect on rhythm. Therefore, the multimodal automated oral English assessment model in this paper has great influence on the assessment tasks and is helpful to the assessment of oral English proficiency. At the same time, there are still some shortcomings in this study, such as the data quality and data scale are different from the actual situation, so it is necessary to further improve the data quality and expand the data scale. In addition, more input features should be added to the model to improve the accuracy of the model. In the next step, the research will further study the above shortcomings in order to improve the accuracy and practicability of the model. In the future research work, it is necessary to further integrate the performance comparison of different algorithms. In terms of parameter optimization, intelligent calculation methods can be considered, and the prediction effect will be further improved. Whether it is suitable for the application of other languages in the application scenario needs further verification.

Data Availability

The experimental data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declared that they have no conflicts of interest regarding this work.

References

[1] G. A. Constantinides, P. Y. K. Cheung, and W. Luk, "Wordlength optimization for linear digital signal processing," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 22, no. 10, pp. 1432–1442, 2003.

[2] G. Kovács, L. Tóth, and D. Van Compernolle, "Selection and enhancement of Gabor filters for automatic speech recognition," *International Journal of Speech Technology*, vol. 18, no. 1, pp. 1–16, 2015.

[3] J. Zhang and B. Qin, "DTW speech recognition algorithm of optimization template matching," in *Proceedings of the World Automation Congress*, pp. 1–4, Puerto Vallarta, Mexico, June 2012.

[4] S. Yi, K. Mochitomi, I. Suzuki, X. Wang, and T. Yamasaki, "Attention-based multimodal neural network for automatic evaluation of press conferences," *International Journal of Multimedia Data Engineering & Management*, vol. 11, no. 3, pp. 1–19, 2020.

[5] L. Deng, "Design of automatic evaluation system for spoken English pronunciation quality," *Automation & Instrumentation*, vol. 6, pp. 175–179, 2019.

[6] J. Guo, "Design and implementation of automatic evaluation system in college English writing teaching based on ASP.net. Big data analytics for cyber-physical system in smart city. BDCPS 2020," *Advances in Intelligent Systems and Computing*, vol. 1303, pp. 1622–1626, 2021.

[7] L. H. Li, "Design of college English process evaluation system based on data mining technology and internet of things," *International Journal of Data Warehousing and Mining*, vol. 16, no. 2, pp. 18–33, 2020.

[8] P. F. Su, Q. C. Chen, and X. L. Wang, *A Fuzzy Pronunciation Evaluation Model for English Learning. International Conference on Machine Learning & Cybernetics*, IEEE, Dalian, China, 2006.

[9] C. Bauman, "Social evaluation of asian accented English," *university of Pennsylvania Working Papers In Linguistics*, vol. 19, no. 2, pp. 11–20, 2013.

[10] M. Suzuki, T. Konno, A. Ito, and S. Makino, "Automatic evaluation system of English prosody based on word importance factor," *Journal of Systemics Cybernetics & Informatics*, vol. 6, no. 4, pp. 83–90, 2008.

[11] M. Tremmel, "What to make of the five-paragraph theme: history of the genre and implications," *Teaching English in the Two-Year College*, vol. 39, no. 1, pp. 29–42, 2011.

[12] G. Chen and S. Li, "Research on location fusion of spatial geological disaster based on fuzzy SVM," *Computer Communications*, vol. 153, pp. 538–544, 2020.

[13] D. Bolaos, A. R. ColeH, W. Ward, A. G. Tindal, and J. P. Schwanenflugel, "Automatic assessment of expressive oral reading," *Speech Communication*, vol. 55, no. 2, pp. 221–236, 2013.

[14] L. Fathi and S. Sidgi, "The usefulness of automatic speech recognition (ASR) eyespeak software in improving Iraqi EFL students' pronunciation," *Advances in Language and Literary Studies*, vol. 8, no. 1, pp. 1–6, 2017.

[15] J. Nakamoto, K. A. Lindsey, and F. R. Manis, "A longitudinal analysis of English language learners' word decoding and reading comprehension," *Reading and Writing*, vol. 20, no. 7, pp. 691–719, 2007.

[16] N. Tu, H. Thu, and V. A. Nguyen, "language model combined with Word2Vec for product's aspect based extraction," *ICIC Express Letters*, vol. 14, no. 11, pp. 1033–1040, 2020.

[17] K.-H. Kim, D. Lee, M. Lim, and J.-H. Kim, "Input dimension reduction based on continuous word vector for deep neural network language model," *Phonetics and Speech Sciences*, vol. 7, no. 4, pp. 3–8, 2015.

[18] Z. Xiong, Q. Shen, Y. Xiong, Y. Wang, and W. Li, "New generation model of word vector representation based on CBOW or skip-gram," *Computers, Materials & Continua*, vol. 58, no. 2, pp. 259–273, 2019.

[19] F. Wang, F. Xie, S. Shen, L. Huang, R. Sun, and J. Le Yang, "A novel multiface recognition method with short training time and lightweight based on ABASNet and H-softmax," *IEEE Access*, vol. 8, pp. 175370–175384, 2020.

[20] F. Wang, S. Yang, Q. Li, and C. Wang, "An internet of things malware classification method based on mixture of experts neural network," *Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 5, pp. 1–12, 2020.