Hindawi
Advances in Multimedia
Volume 2021, Article ID 6783205, 12 pages
https://doi.org/10.1155/2021/6783205

Research Article

Research on Correction Method of Spoken Pronunciation Accuracy of AI Virtual English Reading

Shuli Wang and Xiuchuan Shi

Harbin University of Commerce, Harbin 150028, China

Correspondence should be addressed to Xiuchuan Shi; 102580@hrbcu.edu.cn

Received 20 October 2021; Revised 10 November 2021; Accepted 2 December 2021; Published 17 December 2021

Academic Editor: Qiangyi Li

In order to improve the pronunciation accuracy of spoken English reading, this paper combines artificial intelligence technology to construct a correction model of the spoken pronunciation accuracy of AI virtual English reading. Moreover, this paper analyzes the process of speech synthesis with intelligent speech technology, proposes a statistical parametric speech based on hidden Markov chains, and improves the system algorithm to make it an intelligent algorithm that meets the requirements of the correction system of spoken pronunciation accuracy of AI virtual English reading. Finally, this paper combines the simulation research to analyze the English reading, spoken pronunciation, and pronunciation correction of the intelligent system. From the experimental research results, the correction system of spoken pronunciation accuracy of AI virtual English reading proposed in this paper basically meets the basic needs of this paper to build a system.

1. Introduction

The virtual spoken English system has become an important English communication tool. With the continuous development of artificial intelligence technology, the AI virtual spoken English tool has gradually moved from theoretical research to real-world applications and the more widely used AI virtual English teaching pronunciation correction.

Most English speech synthesis models based on pronunciation mechanisms contain three main modules. Among them, the pronunciation movement model simulates the morphological structure of the pronunciation organ, the cooperative pronunciation model simulates the dynamic characteristics of the pronunciation organ, and the acoustic model simulates the aerodynamic process to generate the corresponding English speech signal. Any inappropriate approximation of these three main modules will affect the English speech quality. We try to build a more accurate pronunciation movement model to approximate the morphological characteristics of the articulation organs, so as to get a better pronunciation synthesis system. At present, there are two mainstream modeling strategies that are physiological models and geometric models. The physiological pronunciation model uses the finite element method to simulate the biomechanical properties of soft tissue and is embedded in the muscle structure to drive the model. However, the physiological pronunciation model faces the high computational load of the finite element module and the inappropriate distribution of the pronunciation organs and related muscles, which makes the control of the physiological pronunciation model extremely complicated. The geometric pronunciation model models the contours of the vocal organs, and the shapes of the vocal organs and vocal tract can be directly controlled by a predefined parametric set [1]. This parametric set is obtained through statistical analysis. Compared with the physiological pronunciation model, the geometric pronunciation model does not care about the biomechanical properties of soft tissues and the unplaning function of related muscles, so the calculation cost is greatly reduced, and the control of the vocal tract shape becomes simple. Therefore, the geometric model is more suitable for occasions where there is no need to understand and analyze

the internal structure of the pronunciation organs for English speech animation applications [2].

Although in most cases clear sound is sufficient to complete people's basic communication, what visual information provides is a more effective and vivid communication effect. In addition, when the voice is missing or unclear, visual information can help people guess and understand what the speaker wants to express. For example, for people with hearing impairment, effective lip reading or speculation and judgment based on changes in the speaker's facial expressions can help them understand the speaker's meaning accurately [3].

Based on the above analysis, this paper combines intelligent voice technology to construct the correction system of spoken pronunciation accuracy of AI virtual English reading, explore the effectiveness of the model, and improve the correction effect of spoken English reading.

2. Related Work

On the basis of speech visualization, speech-driven face modeling and animation technology are of great significance to improve the teaching effect of the multimodal Mandarin pronunciation teaching system [4]. In recent years, many 3D speaker simulation technologies have been proposed, which can be basically divided into the following six categories: based on vector graphics animation, based on raster graphics system for animation rendering, based on the data-driven synthesis, and based on anatomically modeling the head, based on deformation algorithm, and based on machine learning [5]. The technique 3D speaker modeling based on vector graphics animation uses a simple vector graphic animation to show the outline of the main facial articulation organs (mouth, tongue, teeth, and soft palate, etc.) [6]. The method 3D speaker modeling for animation rendering based on a raster image system uses complex polygons to form a human head model. The advantage is that the raster image system can provide a high rendering level and a more realistic head model. The disadvantage is that the time-varying motion parameters are difficult to calculate, and the raster image system is very expensive and animation renders a long time [7]. The method 3D speaker modeling based on data-driven synthesis uses digital image processing technology to extract features from digital images [8]. Literature [9] established a sound-to-speech reversal model based on generalized variable parameters-hidden Markov (GVP-HMM) to achieve 3D speaker modeling. Literature [10] modeled a 3D speaker from the anatomy of the head. Based on the physiological structure of the face, a muscle model is proposed, and the muscle vector is used to simulate the movement of the muscle to generate facial expression animation. The disadvantage of this method is that the muscle parameter derivation mechanism is very indirect, the measurement is also very complicated, and the control parameters of muscle characteristics are only partially visible. The 3D speaker modeling based on the deformation algorithm calculates the position of the deformation point of the entire face by capturing a small amount of facial control point displacement [11]. This method puts the face into a regular control grid, such as an $N \times N \times N$ cube, and establishes the corresponding relationship between the cubic control grid and the object to be deformed. Finally, the control grid can be moved to obtain the deformation of the deformed object to control the local deformation and global deformation of the object to be deformed according to the local movement and global movement of the control grid. First, it calculates the coordinates of the point to be deformed relative to the neighboring control point and gets the position of the point to be deformed from the displacement of the control point [12]. Concerning the 3D speaker modeling based on machine learning, using artificial intelligence techniques such as machine learning to learn the correspondence between speech or text and the movement of the articulator and expression movement [13], using any speech or text to drive the 3D head model, this method avoids immersive real-person data collection. This method is currently in the research stage. There have been many studies on 3D speakers abroad. Literature [14] developed a FAP-driven facial animation Italian speaker head model based on the MPEG-4 standard, which is automatically trained based on real data. The three-dimensional kinematics information is used to create a lip joint model and directly drive the speaker head model. The virtual speaker ARTUR developed in [15] shows the movement of the tongue and teeth, the pronunciation organs in the oral cavity. The visual speaker developed in [16] uses an electromagnetic pronunciation capture device to collect five control points of the tongue, two control points of the soft palate, and six control points of the mouth to simulate developmental pronunciation. The model of the pronunciation organ is obtained by three-dimensional reconstruction of nuclear magnetic resonance images. Literature [17] developed a visual pronunciation system based on a physiological model based on the deformation of the biological characteristics of each muscle in the face and vocal tract to simulate the movement of the articulation organs. Literature [18] developed a face animation system, which uses text/speech as the driving data of the system and uses the hidden Markov model to extract features of the speech signal. The speech is represented by the Meyer Frequency Cepstral Coefficient (MFCC). According to the information, the keyframe sequence of audio-viseme mapping is obtained through MFCC training, and a real-time synchronized face animation system is obtained according to the mapping relationship. Literature [19] developed a text-driven 3D Chinese pronunciation system, collected pronunciation corpus through EMA equipment, trained pronunciation model and acoustic model based on the hidden semi-Markov model (HSMM), and obtained a 3D network of pronunciation organs through MRI. The lattice model realizes the Chinese pronunciation system of simultaneous pronunciation. Literature [20] realized the correct pronunciation animation of the 3D human model and chose to

use the EMA data as the support and the Dirichlet Free Deformation (DFFD) algorithm to drive the 3D human head talking model.

3. Statistical Parametric Speech Synthesis Based on Hidden Markov Chain

Generally speaking, unit splicing technology often does not involve the processing of speech signals, and the quality of synthesized speech often depends on the database produced. Since this paper only focuses on parametric speech synthesis technology, nonparametric speech synthesis methods are not in the scope of this paper. Parametric speech synthesis technology mainly uses data to train the model, so that the model can learn the mapping function from text to acoustic parametric from the data set. Compared with nonparametric speech synthesis technology, in the prediction stage, it no longer depends on the data set, and the model directly synthesizes the text into speech. Among parametric speech synthesis models, statistical parameter speech synthesis based on hidden Markov chains is the most popular technology, which is generally divided into text analysis module, acoustic module, and vocoder synthesizer. The process of the statistical parameter speech synthesis model based on the hidden Markov chain is shown in Figure 1.

Character to phoneme conversion converts words into phonetic representations, which are generally described by phonemes. The prosodic unit is composed of adjacent phonemes, which can generally reflect the speaker's mood in the speech and the mood of the sentence (declarative sentence, interrogative sentence, imperative sentence, etc.) and other pieces of information. The prosody feature embodies the pitch, length, and intensity of the speech. Adding prosodic information to the input helps to enhance the naturalness of the synthesized voice. Since the independent phonemes cannot model the context information, it is not conducive to the synthesized speech quality. Therefore, after the input is converted into phoneme, contextual information is often added to the phoneme information, which mainly includes phoneme, stress-related factors, and location-related factors. The common context information in English is shown in Table 1.

In the statistical parameter speech synthesis model based on the hidden Markov chain, the function of the acoustic module is to convert the phoneme-level context sequence output by the text analysis model into corresponding acoustic parameters, and the acoustic model often contains Mel cepstrum coefficients, basis frequency, and vocalization sign. Among them, the acoustic model is modeled by hidden Markov chains.

In the training stage, the acoustic parameters such as the cepstral coefficient sequence, the fundamental frequency sequence, and whether to pronounce the sequence in the audio are extracted through the corresponding signal processing algorithm. Each different context corresponds to a different state (hidden variable) in the hidden Marco chain and, at the same time, introduces the beginning and end substates in each state. The state is used to describe the context of prosody and linguistics. The acoustic parameter sequence corresponds to the observation value of the state in the hidden Markov chain, and the distribution of the observation value of each state is a multidimensional Gaussian mixture distribution. At the same time, the fundamental frequency information contains the fundamental frequency and whether to speak or not. Among them, the fundamental frequency is continuous, and the vocalization flag is discrete. Therefore, it needs to adopt multispace mixed distribution. It is worth noting that, according to the hidden Markov model, the probability of the duration of the state sequence is shown in

$$\log\ p(D|\lambda) = \sum_{k=1}^{K} lpgp_k(d_k). \tag{1}$$

Among them, K is the total number of states of the hidden Markov model obtained from the input context sequence, λ is the parameter in the hidden Markov model, and $p_k(d_k)$ is the probability that the state k continues for d_k basic time slots, and its probability expression is

$$p_k(d) = a_{kk}^{d} \cdot (1 - a_{kk}). \tag{2}$$

Among them, a_{kk} is the transition probability from state k to state k. This paper is based on the maximum conditional probability criterion. When calculating the maximum value of formula (2), the probability of the duration of each state will decrease proportionally as the number of continuous-time slots increases. Therefore, the duration of each state is 1 time slot. A hidden semi-Markov model is introduced to model the state duration, which makes the duration of each state obey a Gaussian distribution. And the mean and variance in the distribution of the duration of each state are the results after the most iteration in the forward-backward algorithm.

At the same time, the context sequence has different effects on acoustic parameters such as cepstral coefficients, fundamental frequency, and duration of the state. Therefore, it needs to establish a separate decision tree and corresponding problem set for each parameter. The training is based on the maximum conditional probability criterion. This will cause the output acoustic parameters of the model in a given state to be the mean value of the Gaussian mixture distribution, which will result in a large step between different states. As a result, the coherence of the entire synthesized speech deteriorates. In order to improve the coherence of synthesized speech, the first-order and second-order difference values of acoustic parameters are introduced into the observations of hidden Markov chains.

As the number of context factors in Table 1 increases, the number of factors that can be combined will increase exponentially as the number of factors increases. This leads to an exponential increase in the number of states in the hidden Markov chain, and training such a model often requires a larger training data set. At the same time, the increase in the number of states will more easily lead to the problem of uneven data distribution. For example, some combinations have a large number of training data, while some combinations have a small number of training data, which

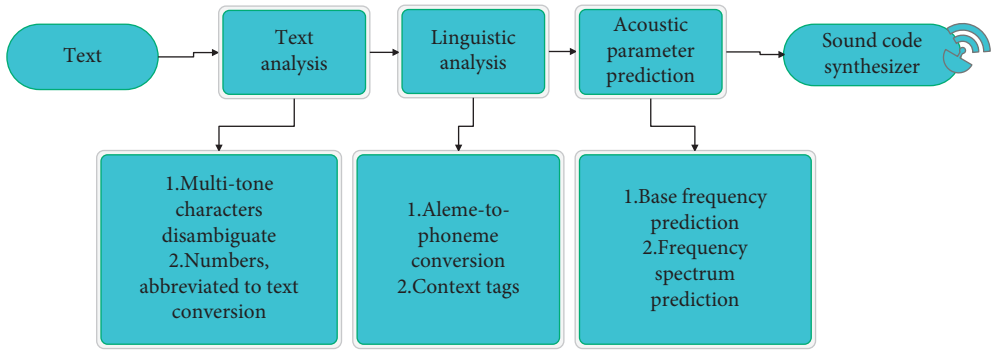

Figure 1: Statistical parametric speech synthesis model based on hidden Markov chain.

Table 1: Context information.

Current phoneme, first two phonemes, last two phonemes
The position of the current phoneme in the current syllable
The number of phonemes in the current syllable, the previous syllable, and the next syllable
The type of accent in the current syllable, the previous syllable, and the next syllable
Whether the current syllable, the previous syllable, and the next syllable are stressed
The position of the current syllable in the current word and current phrase
The number of syllables in the current phrase, the previous phrase, and the next phrase
The number of accented syllables in the current phrase, the previous phrase, and the next phrase
The number of syllables from the current position to the previous and next stressed syllable
Part of speech of the current word, the previous word, and the next word
The number of syllables in the current word, previous word, and next word
The position of the current word in the current phrase
The number of words before and after the current position in the current phrase
The number of words in the previous phrase and the next phrase
The number of syllables in the current, previous, and next phrase
The position of the current phrase in the main sentence
The distance from the current position to the stressed syllable
The number of phonemes, syllables, words, and phrases in the current sentence

ultimately leads to insufficient training of the model. When there is no training in the test phase, the incorrect acoustic parameters predicted by the model will affect the quality of the synthesized speech. In order to improve the generalization performance of the model and solve the problem of data sparseness, decision trees are often introduced into the model. Each leaf in the decision tree corresponds to a state in the hidden Markov chain. In the process of training the decision tree, the pruning strategy is adopted, and some leaves in the decision tree will correspond to multiple contexts so that the final state number of the model is reduced, the data distribution space is reduced, and the model training is more adequate. In the prediction phase, when encountering a context without training, the model can also determine the state corresponding to the context according to the decision tree. According to the training data set to train the parameters in the hidden Markov chain model, the mathematical expression of the maximum conditional probability of the observation is shown in formula (3). Among them, O, w, λ, q, a_{ij}, and b_q are, respectively, the acoustic feature (Meier cepstrum coefficient, fundamental frequency) sequence, context feature sequence, parameters in the hidden Markov chain, and the state in the hidden Markov chain, state transition probability, and observation state probability.

$$p_k(O|\lambda, W) = \sum_{\forall q} \pi_{qo} \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(O_t), \tag{3}$$

$$\lambda_{\max} = \arg\max_{\lambda} p_k(O|\lambda, W). \tag{4}$$

During training, the number of frames T of the observation value and the state sequence of the hidden Markov chain are known, and the forward-backward algorithm and the EM algorithm are used to obtain the parameters λ. In the testing phase, this paper first analyzes the text input required, extracts the context sequence, and obtains the hidden Markov model sequence according to the context sequence and the duration of the corresponding state of each context. Subsequently, the state duration is used to adjust the hidden Markov model sequence into a frame-level sequence, combined with the speech parameter generation algorithm I col to obtain smooth acoustic parameters, Finally, a vocoder is used to synthesize speech.

Given a text input, the context sequence obtained from the input text is w, and the probability that the model generates observations (the observations correspond to the acoustic parameters) is shown in

$$\begin{aligned} o_{\max} &= \arg\max_{0} p(o|\lambda_{\max}, w) \\ &= \arg\max_{0} \sum_{\forall q} p(o, q|\lambda_{\max}, w) \\ &\approx \arg\max_{0,q} p(o, q|\lambda_{\max}, w) \\ &= \arg\max_{0} p(o|q, \lambda_{\max}) p(q|\lambda_{\max}, w) \\ &\approx \arg\max_{0} p(q|q_{\max}, \lambda_{\max}, w) \\ &= \arg\max_{0} \prod_{t=1}^{T'} N'\left(o_t; u_{q_{\max}}, \Sigma_{q_{\max},t}\right). \end{aligned} \tag{5}$$

Among them, $q_{\max}$ is the state sequence, T' is preset, which determines the duration of the synthesized speech, and $\lambda_{\max}$ is the acoustic model based on the hidden Markov chain.

The vocoding synthesizer restores the frame-level acoustic features (fundamental frequency, Mel cepstrum coefficients) predicted in the acoustic model to the time domain signal through a digital filter, that is, the final speech. Its mathematical expression is as follows:

$$x(n) = h(n) * e(n). \tag{6}$$

Among them, $x(n)$ is the synthesized speech signal, $h(n)$ is the formant filter, whose parameters are determined by acoustic parameters such as the Mel cepstrum coefficient, and $e(n)$ is the excitation signal, which corresponds to the output of the acoustic model.

Models based on deep learning have gradually emerged in many fields such as image classification and segmentation, video understanding, machine translation and understanding, and speech recognition and synthesis, and their performance indicators have been constantly refreshed. It is worth noting that after the size of the data set increases to a certain extent, the performance of traditional machine learning algorithms no longer increases significantly as the size of the data set becomes larger, as shown in Figure 2.

Due to the great similarity between adjacent points of the speech signal, this will bring great redundancy to the model implicitly learning the alignment of text and audio. Therefore, 80 power values in the specified frequency range under the Mel scale are used to represent 1024 points in each frame. This article assumes that T is used to represent the number of slots of the decoder; then, the decoder will eventually generate a prediction value of $80 * T$ dimension. The output of the decoder is processed by the postprocessing module to synthesize the final speech. The rest of this section will be expanded with the encoder, decoder, and postprocessing network in the Tacotron model. The overall model structure of Tacotron is shown in Figure 3.

The decoder of the Tacotron model is composed of a cyclic neural network and a preprocessing module. It mainly

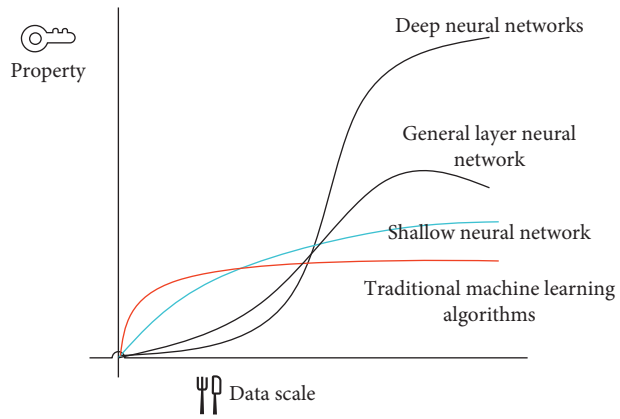

Figure 2: The relationship between model performance and data scale.

uses the output of the encoder in the Tacotron model as its input to predict the acoustic parameters at the next moment. The cyclic neural network is composed of a cyclic neural network combined with the attention mechanism and a two-layer cyclic neural network, as shown in Figure 4.

The Tacotron model splices the context obtained by the attention mechanism with the decoder output value at the previous moment and uses it as the decoder's next moment input (the spliced value may need to be projected to the specified dimension using a fully connected neural network). The structure of the attention mechanism in Tacotron's decoder is shown in Figure 5, and the specific mathematical operation process in the attention mechanism is shown in formulae (7) to (11).

$$e_{i,j} = \text{score}\left(s_{i-1}, h_j\right). \tag{7}$$

$$\alpha_{i,j} = \frac{\exp\left(e_{i,j}\right)}{\sum_{k=1}^{T} \exp\left(e_{i,k}\right)}, \tag{8}$$

$$c_i = \sum_{j=1}^{T} \alpha_{i,j} h_j, \tag{9}$$

$$s_i = f\left(s_{i-1}, \left[y_{i-1}; c_i\right]\right), \tag{10}$$

$$y_i = g\left(s_i, c_i\right). \tag{11}$$

In formula (7), $\text{score}\left(s_{i-1}, h_j\right)$ is a score function, which is used to calculate the similarity between the state value s at each time in the decoder and the output value h at each time in the encoder. The common form of the score is shown in

$$\text{score}\left(s_{i-1}, h_j\right) = \begin{cases} s_{i-1}^{T} h_j \\ s_{i-1}^{T} W_a h_j \\ v_a^{T} \tanh\left(W_a \left[s_{i-1}; h_j\right]\right) \end{cases} \tag{12}$$

In formula (9), c_i is the context information at time i. In formula (10), $\left[y_{i-1}; c_i\right]$ is the splicing of the output value y_{i-1} and c_i of the decoder at time $i-1$, which is used as the input

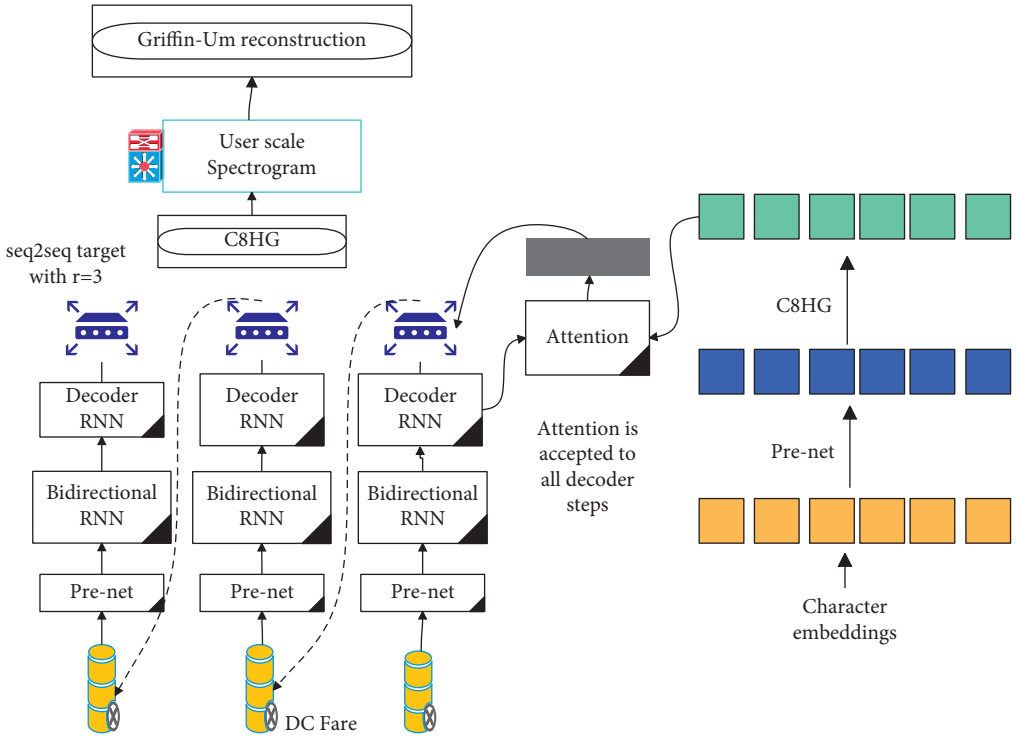

Figure 3: The overall structure of the Tacotron model.

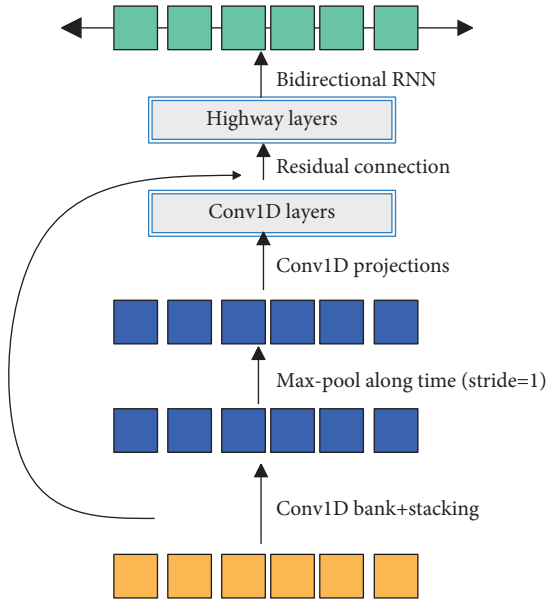

Figure 4: CBHG structure in Tacotron model.

of the decoder at time *à*, *c* is the input of the decoder at time *i*, and *f* is the cyclic neural network. In formula (11), *g* is a fully connected neural network, which takes the main state value *g* and splicing value of the encoder as input to obtain the output of the decoder at the current time.

The Seq2Seq model combined with the attention mechanism makes it have longer sequence modeling capabilities. But with different attention mechanisms, there will be different performances under different tasks. Formulae (7) to (11) correspond to the attention mechanism in the Tacotron model, that is, the output value gy of the decoder at each moment. And its solution process is as follows. First, according to the state value s_{i-1} of the decoder at the previous time $i-1$ and the output h of the encoder, the attention mechanism is used to obtain the context information c_i, and then c_i and n are used as the input of the recurrent neural network to obtain s_i. Finally, s_i and c_i are used as the input of the fully connected neural network to

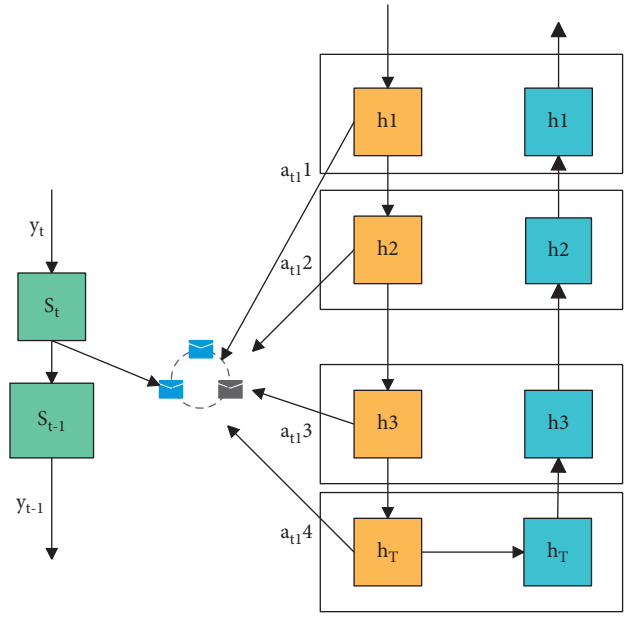

FIGURE 5: Schematic diagram of attention mechanism in Tacotron model decoder.

predict the output value of the decoder at time i. In addition to the attention mechanism used in Tacotron, researchers have also designed other forms of attention mechanism. The solving process of the output value of the decoder at each moment is as follows. First, it clarifies the state value s of the decoder at the current moment and the output h of the encoder, then obtains the context variable c_i according to the attention mechanism, and then uses s_i and c_i to obtain y_i. In addition, a location-based attention mechanism is proposed, as shown in

$$\alpha_i = \text{soft max}\left(W_a h_t\right). \tag{13}$$

A local attention mechanism is proposed; that is, the decoder only pays attention to the local information within a specific window size of the input sequence at each moment. The width of this window is generally much smaller than the length of the input, which can save a lot of calculations. This article presents three strategies for calculating local attention mechanisms. (1) It empirically specifies the size of the window $2 * D + 1$. The model calculates the position p of the center point of interest of the decoder in the encoder output sequence at each moment. Then, the area of the calculation context is $[\max(p_i - D), \min(p_i + D, L_{\text{encoder}})]$, where $a = L_{\text{encoder}}/L_{\text{decoder}}$ is the length of the sequence output by the encoder. (2) It empirically specifies the size of the window $2 * D + 1$. We assume that the correspondence between input and output satisfies a monotonic increase; then, each output moment of the decoder is at the center point position $p_i = a * i$ in the encoder. Among them, $a = L_{\text{encoder}}/L_{\text{decoder}}$. L_{decoder} is the length of the time sequence of the decoder, and then, the area of the calculation context is $([p_i - D, p_i + D]; 3)$. (3) It empirically specifies the size of the window $2 * D + 1$ and predicts the center point position p_i in the model, and the solving process of p_i is shown in

$$p_i = L_{\text{input}} \cdot \text{sigmoid}\left(v_p^T \tanh\left(W_p h_t\right)\right). \tag{14}$$

At the same time, in order to reflect the relationship that the closer the point to the center point, the greater the impact on the output; that is, the response is on the weight α. We assume that the weight α at each position obeys a Gaussian distribution with mean p_i and variance σ. This relationship is shown in

$$\alpha_i(s) = \alpha_{i,s} \exp\left(-\frac{(s - p_i)^2}{2\sigma^2}\right). \tag{15}$$

Among them, s is the position index of the encoder, and the variance $\sigma = D/2$ and $\alpha_{i,s}$ are obtained using formula (8).

The main focus position at the current moment is p_i and the focus point p_{i-1} at the previous moment $i - 1$; then, $p_i \geq p_{i-1}$. The specific process is as follows: we assume that the focus point of the last time $i - 1$ is p_{i-1}; then, the range of the focus position at the current time à is $p_{i-1}, \ldots, L_{\text{encoder}}$. We assume that the position of the attention point of i at the current time satisfies the Bernoulli distribution, and a Bernoulli distribution experiment is carried out from p_{i-1}. If the output of p_j is 1, where $p_j \in \{p_{i-1}, \ldots, L_{\text{decoder}}\}$, then, the position of the attention point of i at the current time is considered to be p_j, The context information at the current moment is h_{pj}, as shown in Figure 6.

The historical alignment information is taken into account in calculating the context information at the current moment. The newly added historical alignment information helps to further strengthen the model's ability to model long sequences. That is, score (s_{i-1}, h_j) in formula (7) becomes as shown in

$$\begin{aligned} e_{i,j} &= \text{score}\left(s_{i-1}, h_j, f_i\right) \\ &= w^T \tanh\left(W s_{i-1} + V h_j + U f_{i,j} + b\right). \end{aligned} \tag{16}$$

Among them, $f_i = F * \alpha_{i-1}$. F adjusts the α dimension to the specified dimension.

4. Research on Correction Method of Spoken Pronunciation Accuracy of AI Virtual English Reading

The correction system of spoken pronunciation accuracy of AI virtual English reading is constructed on the basis of the previous algorithm improvement. The core framework of the AI virtual English reading system is shown in Figure 7.

We can directly create a subprocess to call and realize data transfer through an anonymous pipe. The process of AI virtual English reading is shown in Figure 8.

When the main process needs to call other subroutines according to the system logic, the system first creates an anonymous pipe, sets its read handle to A and write handle to B, and sets the startup information of the subprocess according to these handle pointers. After the preparatory work is completed, the child process can be created. After the child process starts, it first sets the read handle of the process to B and the write handle to A according to the startup information, which is just the

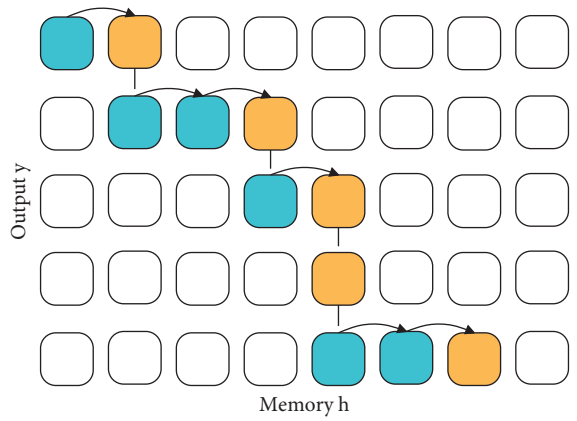

Figure 6: Monotonically increasing attention mechanism.

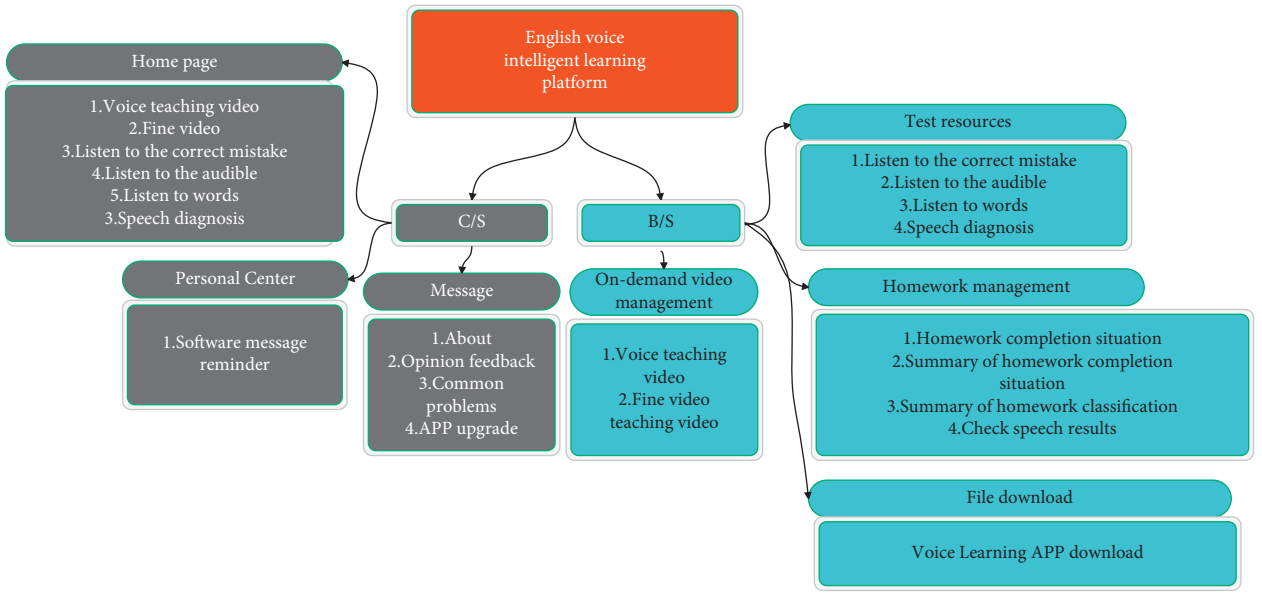

Figure 7: Core framework of AI virtual English reading system.

opposite of the main process. After the interface setting is completed, the system can run the main program and wait for the input command. The main process waits for the child process to start, then writes the command through the B handle, and reads the output result through the A handle. When the child process gets the command, it calls the corresponding function according to the logic and writes the output result to the A handle. If it is the end command, it exits the process. In this way, the subroutine call is completed.

After the above model system is constructed, the performance of the AI virtual English reading system is verified with experiments, and English reading and spoken pronunciation are analyzed and evaluated through simulation research. The scoring results are shown in Table 2 and Figure 9.

On the basis of the above research, the effect of pronunciation correction in English reading is evaluated, and the results are shown in Table 3 and Figure 10.

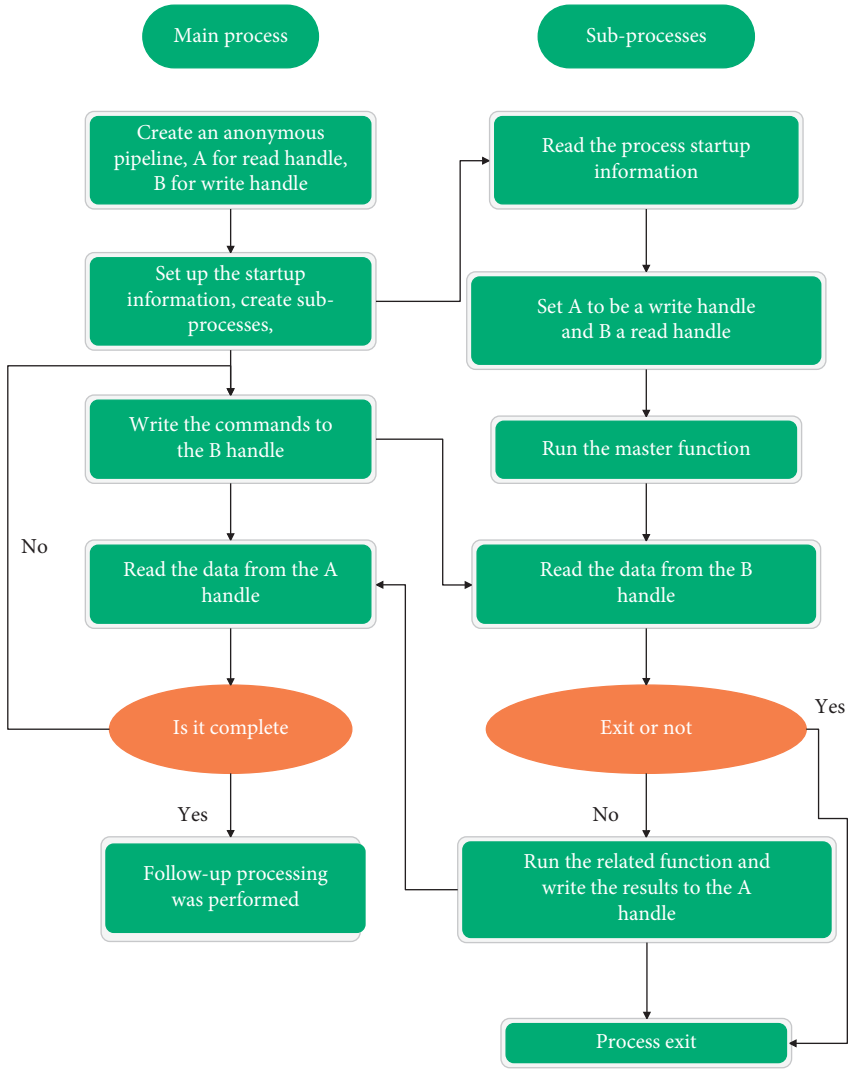

Figure 8: Calling of subroutines of the spoken pronunciation evaluation engine for AI virtual English reading.

Table 2: Performance evaluation of the correction system of spoken pronunciation accuracy of AI virtual English reading.

Number	Smart reading	Oral English
1	91.9	89.5
2	83.8	86.3
3	85.5	81.9
4	83.2	85.3
5	91.2	88.7
6	83.8	76.5
7	81.5	78.3
8	83.8	75.1
9	81.5	80.6
10	82.9	82.1
11	87.2	89.2
12	87.9	75.4
13	90.2	79.3
14	82.0	89.9
15	79.1	84.7
16	83.0	79.0
17	84.5	80.0
18	81.4	75.3
19	90.5	79.4

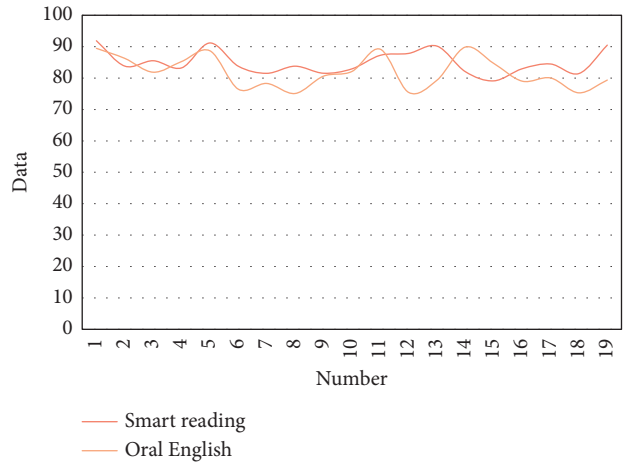

Figure 9: Statistical diagram of the evaluation of English reading and spoken pronunciation.

Table 3: Evaluation data of pronunciation correction effect of spoken English reading.

Number	Pronunciation correction
1	85.9
2	78.8
3	79.2
4	78.0
5	78.2
6	85.7
7	70.8
8	75.9
9	83.1
10	75.0
11	75.4
12	77.0
13	85.9
14	73.2
15	85.3
16	78.2
17	86.7
18	78.5
19	81.5

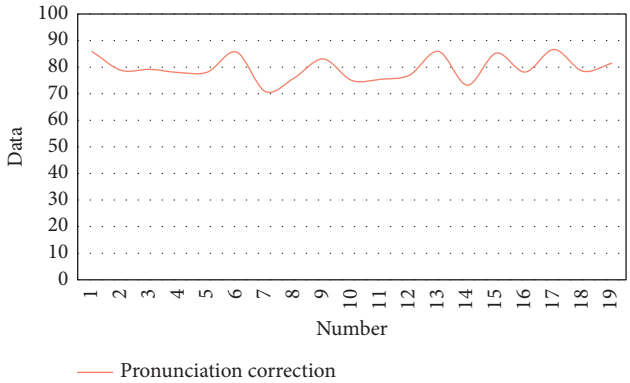

Figure 10: Statistical diagram of evaluation of pronunciation correction effect of spoken English reading.

From the above research, we can see that the correction system of spoken pronunciation accuracy of AI virtual English reading has good practical effects.

5. Conclusion

Traditional speech synthesis technology is often divided into nonparametric speech synthesis technology and parametric speech synthesis technology. Nonparametric speech synthesis technology is mainly based on unit selection. The main idea is that speech is spliced from speech unit fragments, and the speech unit database is made with sufficient coverage. In the prediction stage, the text is transformed into a phoneme sequence marked with prosodic features (fundamental frequency, duration, etc.). Using the set loss function as the evaluation criterion, the optimal speech unit is selected from the database, and the selected speech unit sequence is spliced into the final speech. This paper combines the intelligent voice technology to carry out the AI virtual English reading oral pronunciation accuracy correction model, to verify the performance of the AI virtual English reading system, and to analyze the English reading and oral pronunciation and pronunciation correction. From the experimental research point of view, the correction system of spoken pronunciation accuracy of AI virtual English reading proposed in this paper basically meets the basic needs of this paper to build a system.

Data Availability

The labeled dataset used to support the findings of this study is available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This work was supported by the Heilongjiang Philosophy and Social Sciences Program, Research on Innovation and Practice of Project-based Business English Teaching Strategies in the Post-mooc Era. (No. 22YYB020), and Key Project of Heilongjiang Province Economic and Social (The Special Project for Foreign Language Discipline), Research and Practice of Business English Teaching Based on "Introducing Enterprises into Education" under the Concept of Project Teaching (No. WY20211083-C).

References

[1] M. Woźniak and D. Połap, "Voice recognition through the use of Gabor transform and heuristic algorithm," *Nephron Clinical Practice*, vol. 63, no. 2, pp. 159–164, 2017.

[2] T. Haderlein, M. Döllinger, V. Matoušek, and E. Nöth, "Objective voice and speech analysis of persons with chronic hoarseness by prosodic analysis of speech samples," *Logopedics Phoniatrics Vocology*, vol. 41, no. 3, pp. 106–116, 2015.

[3] S. S. Nidhyananthan, K. Muthugeetha, and V. Vallimayil, "Human recognition using voice print in LabVIEW," *International Journal of Applied Engineering Research*, vol. 13, no. 10, pp. 8126–8130, 2018.

[4] F. L. Malallah, K. N. Y. M. G. Saeed, S. D. Abdulameer, and A. W. Altuhafi, "Vision-based control by hand-directional gestures converting to voice," *International Journal of Scientific & Technology Research*, vol. 7, no. 7, pp. 185–190, 2018.

[5] S. Morgan, "Contact effects on voice-onset time in Patagonian Welsh[J]," *Acoustical Society of america Journal*, vol. 140, no. 4, p. 3111, 2016.

[6] G. Mohan, K. Hamilton, A. Grasberger, A. C. Lammert, and J. Waterman, "Realtime voice activity and pitch modulation for laryngectomy transducers using head and facial gestures," *The Journal of the Acoustical Society of America*, vol. 137, no. 4, p. 2302, 2015.

[7] T. G. Kang and N. S. Kim, "DNN-based voice activity detection with multi-task learning," *IEICE-Transactions on Info and Systems*, vol. E99.D, no. 2, pp. 550–553, 2016.

[8] H.-N. Choi, S.-W. Byun, and S.-P. Lee, "Discriminative feature vector selection for emotion classification based on speech," *The Transactions of the Korean Institute of Electrical Engineers*, vol. 64, no. 9, pp. 1363–1368, 2015.

[9] C. T. Herbst, S. Hertegard, D. Zangger-Borch, and P. A. Lindestad, "Freddie mercury—acoustic analysis of speaking fundamental frequency, vibrato, and subharmonics," *Logopedics Phoniatrics Vocology*, vol. 42, no. 1, pp. 1–10, 2016.

[10] J. Al-Tamimi, "Revisiting acoustic correlates of pharyngealization in Jordanian and Moroccan Arabic: implications for formal representations," *Laboratory Phonology*, vol. 8, no. 1, pp. 1–40, 2017.

[11] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[12] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1315–1329, 2016.

[13] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Audio-visual speech recognition using deep learning," *Applied Intelligence*, vol. 42, no. 4, pp. 722–737, 2015.

[14] Y. Qian, M. Bi, T. Tan, and K. Yu, "Very deep convolutional neural networks for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2263–2276, 2016.

[15] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, 2014.

[16] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: a survey," *Speech Communication*, vol. 56, no. 3, pp. 85–100, 2014.

[17] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[18] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, vol. 46, no. 3, pp. 535–557, 2017.

[19] P. Swietojanski, A. Ghoshal, and S. Renals, "Convolutional neural networks for distant speech recognition," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1120–1124, 2014.

[20] K. Angell and E. Tewell, "Teaching and un-teaching source evaluation: questioning authority in information literacy instruction," *Comminfolit*, vol. 11, no. 1, pp. 95–121, 2017.