International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18 (2026) pp. 1-12

https://doi.org/10.70917/ijcisim-2026-0029
Article

Optimization Design of Intelligent Pronunciation Assessment System Based on Speech Recognition Technology in Higher Vocational English Classroom Teaching under the Background of Industry-Teaching Integration

Tingting He *, Jiguo Yao, Yonghui Lai and Yuhan Zou

School of International Logistics, Hunan Modern Logistics College, Changsha, Hunan, 410131, China;
hebe679@163.com

Abstract: The in-depth development of the integration of industry and education requires that English teaching in higher vocational colleges and universities be equipped with more objective and convenient methods and means for students' speech recognition, detection and assessment. In this paper, a computer-aided English pronunciation assessment system is established by utilizing speech recognition technology and combining the characteristics of English pronunciation with the verification of speech segments, the cutting of language signals and the assessment of pronunciation. MFCC is selected as the voice feature parameter of the proposed system to meet the needs of pronunciation model training. For accent assessment, the neural network-based GOP method is introduced as a deep learning-based accent assessment method to construct a theoretical modeling method for the English pronunciation assessment system. Compared with the traditional methods, the modeling method proposed in this paper consistently maintains an average error between 0.001-0.1 on the automatic assessment of English spoken pronunciation quality, which demonstrates a superior performance of automatic assessment of spoken pronunciation quality.

Keywords: English pronunciation assessment system; speech feature parameters; deep learning; industry-education integration; speech recognition

1. Introduction

Under the social background of the intelligent era, the application of artificial intelligence in the field of education is gradually deepened and normalized, and the application of intelligent technology will bring about a thorough change in the education mode, education environment, education evaluation system and talent training mode [1]. With the changes in the teaching mode, educational environment and talent cultivation mode, as well as the continuous development of artificial intelligence technology, the way of integration of artificial intelligence and education will gradually deepen from the relatively broad integration of the whole discipline to the detailed integration of specialized discipline-oriented [2-4]. With the application of artificial intelligence in English teaching assessment, teaching collaborative assistance and other aspects, the evaluation and teaching methods of the English language subject are also gradually transforming [5]. Artificial intelligence technologies have begun to gradually enter the English classroom teaching and performance assessment, and the application of these technologies improves teaching efficiency while optimizing students' learning experience [6-8].

The automatic assessment method of English spoken pronunciation quality is built on the basis of speech signal detection and feature extraction, using artificial intelligence technology, combined with

time-frequency feature analysis and spectral analysis methods of English spoken pronunciation signals, aiming to improve the automation level and intelligence level of the assessment of English spoken pronunciation quality [9-11]. In the automatic English spoken pronunciation quality assessment system, the English spoken pronunciation signal is affected by the perturbation and distortion of the spoken pronunciation channel, which leads to the poor accuracy of the assessment of the quality of English spoken pronunciation, and the changes in the characteristics of the spoken language produce speech attenuation and distortion, which leads to a decline in the accuracy of the detection performance of the automatic system of the assessment of the quality of English spoken pronunciation, and the optimization of the design of the signal processing algorithm is needed [12-13]. Among the traditional methods, the research on automatic English spoken pronunciation quality assessment methods mainly includes multi-resolution feature detection methods, wavelet analysis methods, scale decomposition methods, time-frequency analysis methods and fractional-order Fourier feature extraction methods [14-15].

Certain research results have been achieved by combining artificial intelligence control and feature extraction to improve the performance of an automatic assessment system for English spoken pronunciation quality. Among them, Jing, W designed a system combining speech recognition and speech synthesis in their research to improve students' spoken English pronunciation level by automatically assessing and correcting learners' pronunciation, thus effectively improving students' pronunciation ability [16]. Fan, Z proposed a spoken pronunciation quality assessment method based on Conformer modeling and multi-task learning with multi-width band technique and achieved good results in pronunciation error detection and pronunciation quality assessment on English dataset [17]. Xiao, W explored the effectiveness of automatic speech recognition technology in diagnosing English pronunciation errors, and the study found that there was an overlap between human ratings and machine ratings, and that they met the different needs of learners, thus providing insights into facilitating English pronunciation assessment and learning [18]. Duan, J and He, Z proposed a model for assessing the pronunciation quality of spoken English based on the Dynamic Time Warping (DTW) algorithm, which comprehensively evaluates multiple dimensions, such as pronunciation standard, fluency and intonation, thus improving the overall effectiveness of pronunciation quality assessment and contributing to the enhancement of the students' learning of spoken English [19]. Shi, X et al. explored the relationship between automated speech scoring (ASS) metrics and the complexity, accuracy and fluency (CAF) dimensions, and found that “oral pronunciation” and “completeness” had a significant effect on Chinese university students' spoken English performance [20]. The above evaluation systems have basically realized the assessment of English pronunciation quality, but there are still the problems of too few evaluation indexes and low stability.

This paper firstly briefly summarizes the module composition and corresponding functions of the computer-assisted English pronunciation evaluation system supported by speech recognition technology. It also focuses on the basic principle of vowel cutting and the operation steps of the verification mechanism in the segmental verification module. Secondly, it describes in detail the pre-processing process of speech signal and the extraction process of MFCC feature parameters. The model training method of GOP method based on deep neural network is also described to build the theoretical model of English pronunciation evaluation system. The theoretical model of the English pronunciation evaluation system is used again to analyze the distribution of resonance peaks on common vowels of different speakers at different levels. Finally, the speech performance is evaluated by combining the results of speech parameter extraction.

2. Establishment of English Pronunciation Assessment System

2.1. Computer-Aided English Pronunciation Assessment System Architecture

The architectural flow of the English pronunciation assessment system in this paper is shown in Fig. 1. Firstly, the pre-processed learners' English pronunciation is verified by speech segmentation, including vowel segmentation cutting, the establishment of the verification system and the reliability of the verification system. Then we train the acoustic model on a large number of standard pronunciation databases through the HMM model, use the Viterbi algorithm to cut and decode the speech segments, and then send this information, including the extraction of evaluation parameters, evaluation parameter regularization, parameter association process and evaluation mechanism, to the core of the English pronunciation evaluation system, i.e., the pronunciation evaluation module, through which we find out the weights of each evaluation parameter in the English pronunciation evaluation. The feedback results provided to the learners contain comprehensive scores, corrective opinions after comparing through the expert knowledge base, and the feedback results provided to the learners.

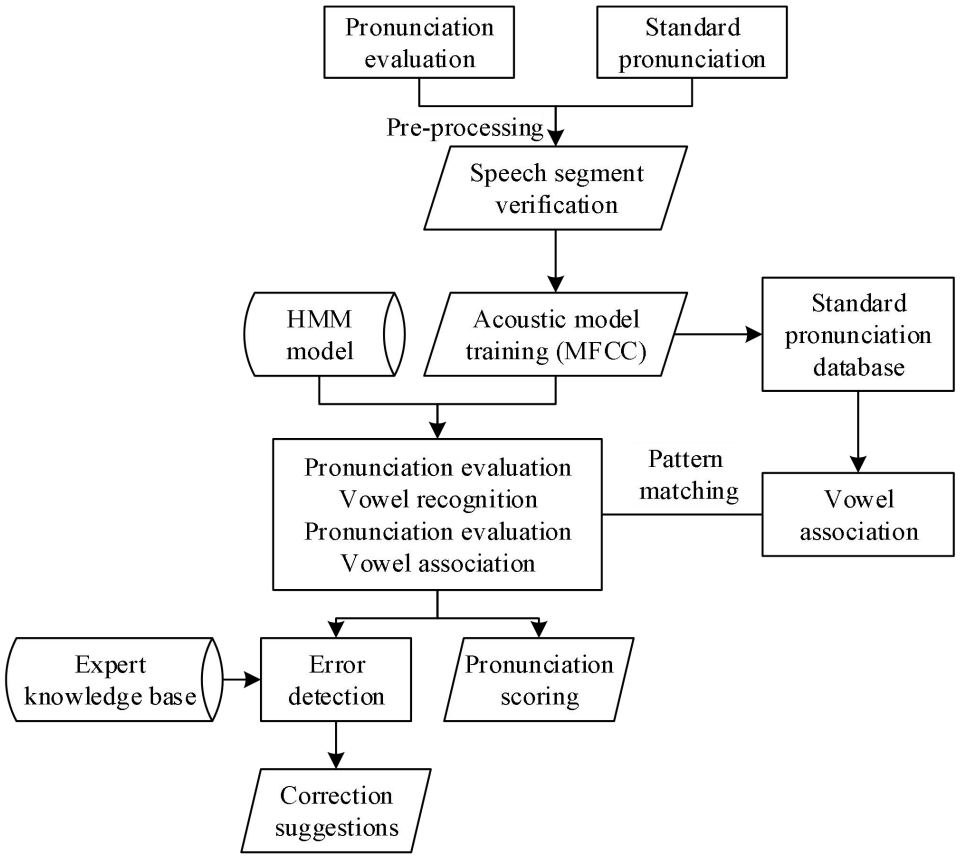

Figure 1. English pronunciation evaluation system diagram.

2.2. Voice Segment Validation

The so-called voice segment verification is to generate judgment thresholds for different evaluation voices and make judgment on the correctness of the evaluation voice content based on the thresholds. The process of speech segment verification is shown in Fig. 2. When the verification system receives the evaluation speech segment, it performs pattern matching on each vowel separately, and then gives the final confidence threshold value according to the distance size of the matching result combined with the verification mechanism.

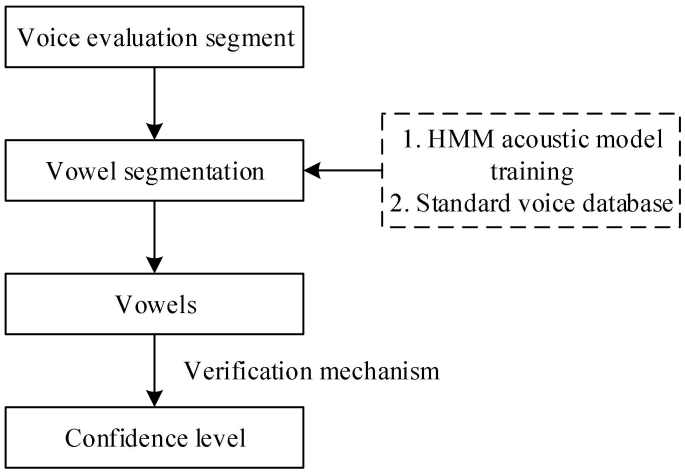

Figure 2. Speech verification process.

2.2.1. Vowel Cutting

In this paper, we use the forced alignment approach commonly used in the Viterbi algorithm to cut speech segments into the smallest vowel pronunciation segments possible. In this case, after the evaluation speech is cut, if the similarity between the evaluation content and the standard pronunciation content is high, the number of vowel segments produced by the cut will be close to or equal to the number of vowel segments of the standard pronunciation. If the first n vowel-consonant segments of the evaluation content are the same as the standard pronunciation, and the following vowel-consonant segments are speech segments outside the standard pronunciation pool, the number of vowel-consonant segments after forced alignment will be approximately n. For example, if the standard pronunciation is "She had your dark suit in greasy wash water all year", and the evaluation pronunciation is "She had your dark suit", the number of meta-consonant segments that can be cut out of the system after vowel segmental cutting is 15. For the part of the segments that are not cut out, the confidence level of the segments is set to 0 in this paper in order to minimize the number of segments that can be cut out. For the part of speech segments that are not cut out, this paper sets the confidence level to 0 to enhance the reliability of the verification of speech segments and make the recognition rate of the system higher.

2.2.2. Validation Mechanisms

To summarize, there are two cases when speech segments are cut: one is that the number of vowel segments obtained after cutting is basically the same as the number of vowel segments in the standard pronunciation, and the other is that part of the speech content is not cut out of the vowel segments. This section focuses on the first case, where a suitable confidence threshold is obtained by regularizing the evaluation parameters. The ranking of vowel segments directly affects the confidence value of the verification system, and in this paper, the formula for finding the confidence value is improved and optimized to obtain equation (1):

$$C_{Vow} = \frac{2}{1+e^{\left(k \cdot R \cdot \frac{\log_{39}^{P} R}{\log_{39}^{P} Ro}\right)}} \tag{1}$$

($\frac{\log_{39}^{P} R}{\log_{39}^{P} Ro}$ denotes the degree of difference in the log probability value of the vowel)

C_{vow} denotes the confidence value, k is the adjusted parameter value, and $R(0 \sim 38)$ denotes the ranking of this vowel among the 39 models, with 0 denoting the 1st place. From Equation (1), the confidence value of a vowel is 1 when the vowel is ranked 1st relative to the 39 models.

Because of the similarity in vowel pronunciation, the comparison cannot rely only on the ranking when making the confidence value judgment. For example, the ranking of [a] and [æ] after comparing 39 models is 2nd place, and its 1st place is [ei]. It can be seen that the log odds difference degree between [a] and the 1st place is small, so the log odds difference degree factor is added to the above formula to ensure the accuracy of the verification system. When the confidence value of all vowels in the speech segment is calculated, the confidence value of the speech segment (0~100) is deduced using the ratio of the time length $(T\_len(Vowel_n))$ of each vowel and the length of the speech time period $(T\_len(Segment))$ as the weights as in Eq. (2):

$$C_{seg} = 100 \cdot \sum_{n=1}^{N} \frac{T\_len(Vowel_n)}{T\_len(Segment)} \cdot C_{Vow_n} \tag{2}$$

N is the number of vowels in the word.

3. Theoretical Model of English Pronunciation Assessment System

3.1. Speech Signal Preprocessing

Speech signal preprocessing is the preparatory work before the extraction of speech characteristics, mainly for the characteristics of the speech signal frequency domain processing. After the analog speech signal is sampled and quantized into a digital signal, it needs to be pre-emphasized to make the high and low frequency amplitudes comparable, and then frame-splitting and windowing are performed to obtain the speech frame. If the voice data is read directly from an audio file (e.g., a wav file), there is no need for

sample quantization.

Because the human voice from the lips, the high-frequency part of the attenuation, so that the low-frequency part of the energy is always higher than the high-frequency part of the energy, which leads to the high-frequency part of the spectral value of the smaller, not easy to analyze and process. Pre-emphasis is to let the speech pass through a high-pass filter, the high-frequency part of the enhancement, so that the high and low-frequency amplitudes are comparable, and its equation is shown in equation (3).

$$S_{n'} = S_n - \alpha S_{n-1} \tag{3}$$

where S_n is the n th speech data and $S_{n'}$ is the n th speech data after pre-emphasis processing. α is the pre-emphasis coefficient, the value range is: $0 \le \alpha < 1$, here take 0.97.

Speech signal is a slow time-varying signal with short-time smoothness. For a speech signal, if a short enough time (about 6-30ms) is taken, the characteristics of the signal are found to be basically unchanged, but from a longer period of time (0.6s or more), the characteristics of the speech signal are constantly changing, and thus reflecting on what the speech is trying to express. Because of this characteristic of speech, it is necessary to analyze the speech by dividing it into several short time segments, and this process is called “frame splitting”. There is a certain overlapping area between two adjacent frames, which makes the frames smoother and maintains the continuity of speech characteristics. Usually the overlapping part is half or one-third of the frame length, and the size of the frame length is between 20 and 30ms, because the characteristics of the speech signal are more stable in this time period. Assuming that the signal sampling frequency is 16kHz, the frame length is 25ms, and the frame rate is 100 frames/second, there are 400 sampling points in each frame, and there are 40,000 sampling points per second in the split-frame speech signal.

It is difficult to analyze the changes of the signal in the time domain to see the characteristics of the signal, which is usually converted to the frequency domain for analysis, and the common method is to perform the Fast Fourier Transform (FFT). In order to avoid more signals in the FFT operation, resulting in errors or mistakes, the need for multiple signals to add window processing. The most widely used windows in speech signal processing are Hamming window, Hanning window and rectangular window. In this system Hamming window is used and its expression is shown in equation (4). Multiplying the speech frames by Hamming window increases the left-right continuity between frames and removes the boundary effect.

$$S_{n'} = \left\{ 0.54 - 0.46\cos\left(\frac{2\pi(n-1)}{N-1} \right) \right\} S_n \tag{4}$$

3.2. MFCC Extraction Process

The extraction process of MFCC is shown in Fig. 3: Firstly, the fast Fourier transform is performed on the speech signal that has been processed by the front-end, which is converted from the time domain to the frequency domain, and then mapped to the Mel frequency domain, and then passes through a triangular band-pass filter bank, logarithmic to the resultant energy value, and finally performs the discrete cosine transform (DCT).

Figure 3. The extraction of MFCC

The conversion relationship between Hertz frequency and Mel frequency is shown in equation (5), f is the linear frequency in Hertz, and the converted Mel frequency has a logarithmic curve.

$$Mel(f) = 2595\lg\left(1 + \frac{f}{700} \right) \tag{5}$$

Let the number of filters in the triangular filter bank be M $(m = 1, 2, \ldots, M)$, then the equation for the m th filter can be expressed by equation (6).

$$H_m(k) = \begin{cases} 0 & k < f[m-1] \\ \dfrac{2(k - f[m-1])}{\big(f[m+1] - f[m-1]\big)\big(f[m] - f[m-1]\big)} & f[m-1] \le k \le f[m] \\ \dfrac{2(f[m+1] - k)}{\big(f[m+1] - f[m-1]\big)\big(f[m+1] - f[m]\big)} & f[m] \le k \le f[m+1] \\ 0 & k > f[m+1] \end{cases} \tag{6}$$

where $f[i]$ is the center frequency of the delta filter, which satisfies Eq. (7), i.e., equally spaced distribution in the Mel frequency domain.

$$Mel(f[i+1]) - Mel(f[i]) = Mel(f[i]) - Mel(f[i-1]) \tag{7}$$

When M is taken to be 20, the filters are denser in the low-frequency portion and relatively sparse in the high-frequency portion, which is in line with the characteristic of the Mel frequency, i.e., human hearing is more sensitive to the low-frequency signals, and tries to take as much as possible of the filter bank while taking as little as possible of the high-frequency portion. From the horizontal axis, the span of the filters is getting larger and larger as the frequency increases, but after mapping to the Mel frequency, the spans of the filters are equal, which is the difference between the human perceptual system and the linear frequency.

The output of the filter in Eq. (8) can be logarithmic and then summed, or it can be summed and then logarithmic, but the latter is more robust to spectral estimation errors and noise.

$$S[m] = \ln\left[\sum_{k=0}^{N-1} |X_a[k]|^2 H_m[k]\right], \quad 0 \le m < M \tag{8}$$

Finally, the DCT operation is performed, as shown in equation (9), to decorrelate the individual frequency bands and map the output of the filter to the cepstrum.

$$c[n] = \sum_{m=0}^{M-1} S[m] \cos\left(\frac{\pi n}{M}\left(m + \frac{1}{2}\right)\right), \quad 0 \le n < M \tag{9}$$

3.3. Deep Neural Network Based GOP Approach

The GOP algorithm, as one of the most classical methods in the field of articulatory assessment, also has an important use in the era of deep learning. in 2015 scholars first proposed to extend the GOP method based on GMM-HMM to DNN-HMM. In the DNN model training, a multilayer neural network is trained to represent speech as a nonlinear basis function, and the last layer of the network outputs a posteriori probability of senone, which is a smaller state of articulation than a triphone. A senone is a smaller articulatory state than a triphone. Unlike previous GOP algorithms that use the HMM decoding lattice to approximate the denominator, a frame-level a posteriori probability approach is proposed to approximate the GOP based on the DNN-HMM system.

Assuming that the articulatory segment to be evaluated is $O_{t_s}^{t_e}$ and the corresponding reference phoneme is p, the corresponding sequence of states is obtained by forced alignment as in Eq. (10):

$$s^* = s_{t_s}, s_{t_s+1}, \ldots, s_{t_e} \tag{10}$$

t_s and t_e are the start and end times of the pronunciation segment, the likelihood of $O_{t_s}^{t_e}$ is equation (11):

$$
\begin{aligned}
P(o \mid p;t_s,t_e) &\approx \arg\max_s P(o,s \mid p;t_s,t_e) \\
&= \pi_{s_{ts}} \prod_{t=t_s+1}^{t_e} a(s_t \mid s_{t-1}) \prod_{t=t_s}^{t_e} P(o_t \mid s_t) \\
&\approx \prod_{t=t_s}^{t_e} P(o_t \mid s_t) \\
&= \prod_{t=t_s}^{t_e} P(s_t \mid o_t) P(o_t) / P(s_t)
\end{aligned}
\tag{11}
$$

where π is the initial state sequence distribution of the HMM, $a(s_t \mid s_{t-1})$ is the inter-state transfer probability, $P(s_t \mid o_t)$ is the DNN model Softmax output, $P(s_t)$ is obtained from the training corpus of the DNN model, and the computation of the sene transfer probability is omitted for the sake of computational simplicity, and the calculation of the likelihood probability can further be omitted The calculation of the firing probability $P(o_t)$, then the likelihood probability is further abbreviated as equation (12):

$$
\log P(o \mid p;t_s,t_e) \approx \sum_{t=t_s}^{t_e} \log \frac{P(s_t \mid o_t)}{P(s_t)} \tag{12}
$$

Compared with the definition of GOP proposed in the GMM-HMM system, the DNN-based GOP estimation does not require decoding lattice and its corresponding backward and forward computations, and thus is suitable for supporting fast, online, multi-channel applications.

Considering that each senone may be shared by multiple states, the senone transfer probability is introduced into the GOP computation, and the GOP formula introducing the senone transfer probability is redefined as Eq. (13):

$$
GOP(p) = \frac{1}{T}\left[\sum_{t=1}^{T} \log P(s_t \mid o_t) + \sum_{t=2}^{T} \log P(s_t \mid s_{t-1}) + (T-1)\log n\right] \tag{13}
$$

where T is the duration of phoneme p, $P(s_t \mid o_t)$ has the same meaning as above, and n is the number of senone in the acoustic model.

Two DNN-HMM speech recognizers were first trained with two native pronunciation corpora, and then three previous DNN-HMM-based GOP methods were compared, which achieved up to 14.89% improvement in phoneme-level score relevance on the Indian English language learners' pronunciation dataset as compared to the previous method that does not consider the probability of senone transfer.

4. Application of the Model and Performance Evaluation

4.1. Resonance Peak Frequency Analysis

The designed model was used to extract the resonance peaks of vowels from different speakers at different levels in spoken English. The first resonance peaks of the extracted vowels are labeled as F1 and the second resonance peaks are labeled as F2. Acoustic-phonetic studies have shown that the first and second resonance peaks are the main factors that affect the tongue position of vowel articulation. When the tongue position is high, F1 is small, and when the tongue position is low, F1 is large. In front of the tongue, F2 is large, and after the tongue, F2 is small. In Mandarin, /ι/ and /τ/ are lingual vowels, and they are more specific, /ι/ is a tongue front vowel and /ι/ is a tongue back vowel. With the resonance peak acoustic features extracted from the designed model for the standard speakers and the two groups of learners, the first and second resonance peak distribution patterns of the standard speakers (male CM, female CF), high level learners (male AUM, female AUF) and beginner level learners (male BUM, female BUF) can be plotted separately, and the selected reasons for the Mandarin are a, o, γ, u, y, i, ι, τ.

The pattern of resonance peaks for high-level male speakers is shown in Figure 4, and the pattern of resonance peaks for beginner-level male speakers is shown in Figure 5.

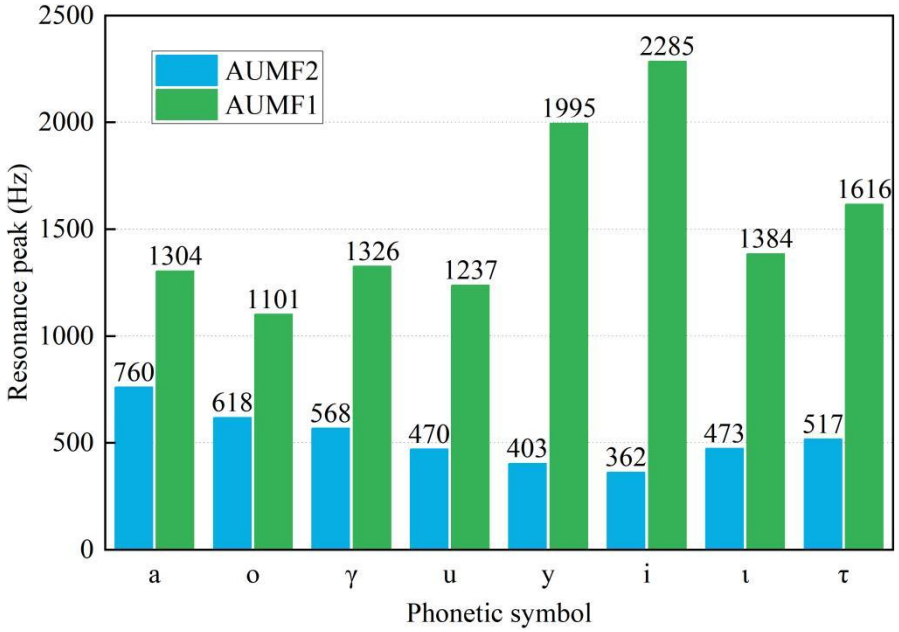

Figure 4. High-level male vocalist formant pattern.

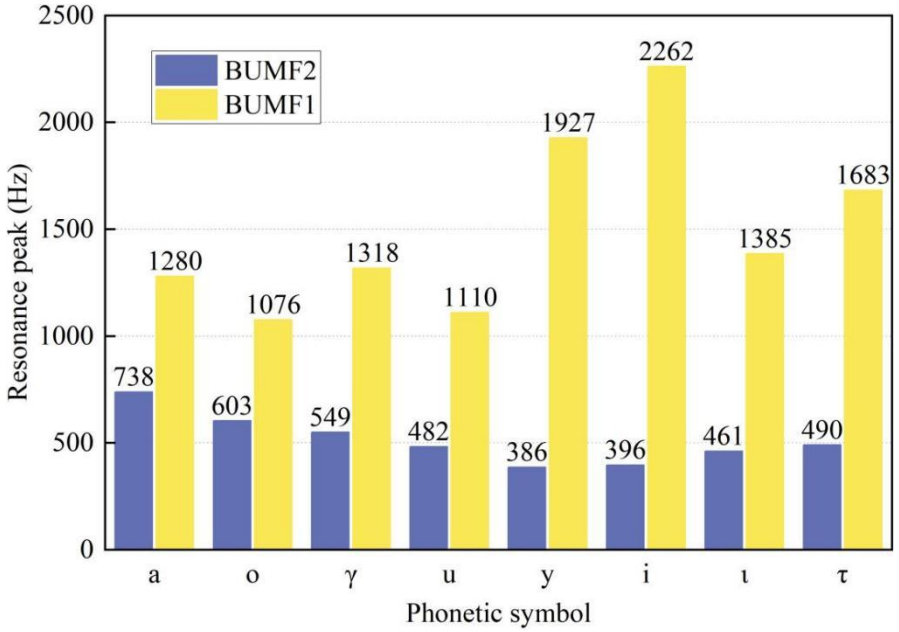

Figure 5. The formant mode of the male speaker at the initial level.

By comparing the distribution patterns of the resonance peak modes of the vowels of the two groups of male learners, it can be seen that the distribution of the resonance peaks when pronouncing the vowels of the beginner-level and high-level male learners are relatively consistent, with the overall first resonance peaks being greater than 1,000 Hz, and the second resonance peaks being in the range of 300-800 Hz.

The pattern of resonance peaks for standard female articulators is shown in Figure 6, and the pattern of resonance peaks for high-level female articulators is shown in Figure 7.

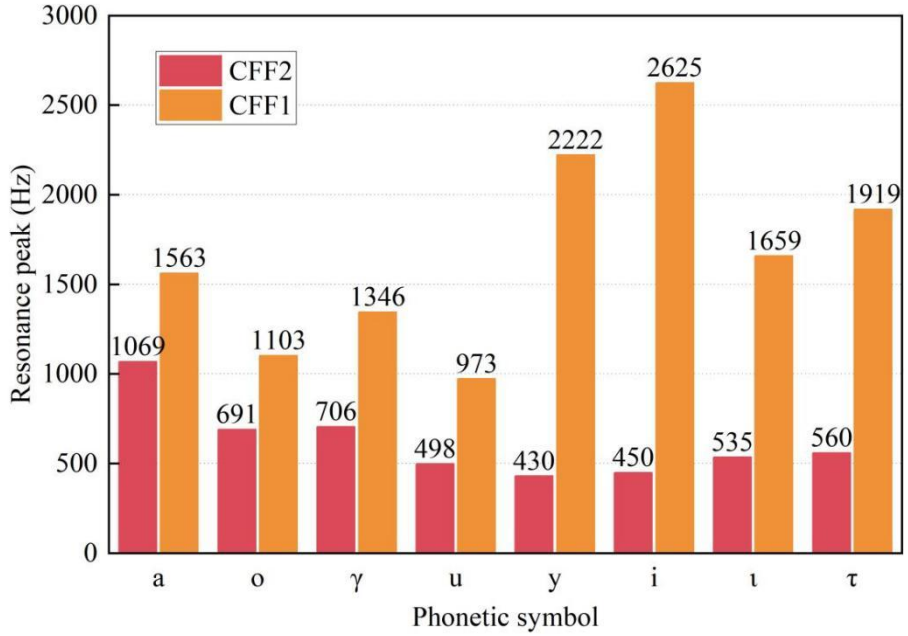

Figure 6. Standard female speaker formant mode.

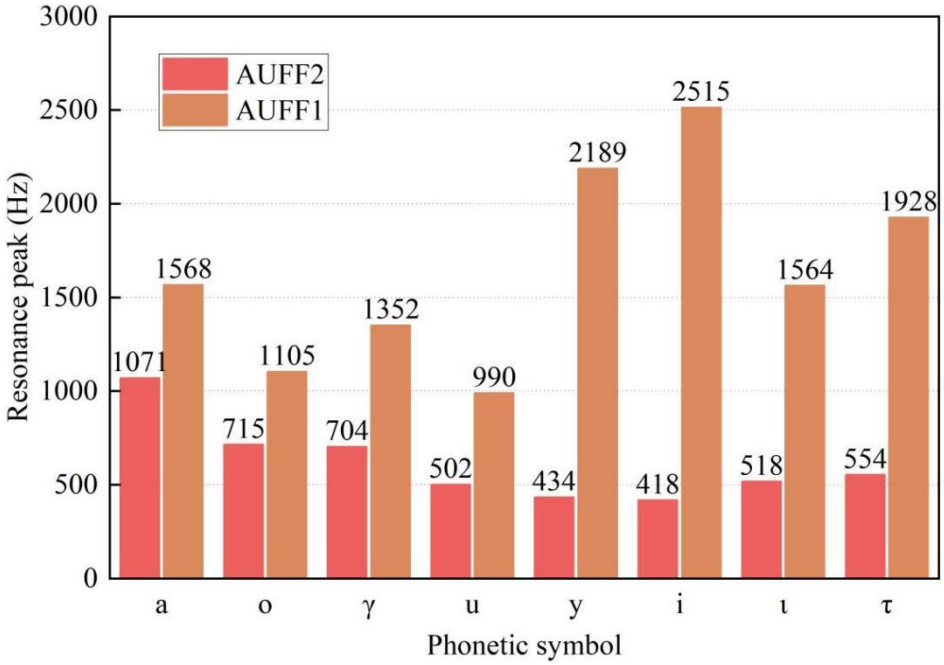

Figure 7. High-level female vocalist formant pattern

Through the resonance peak mode distribution patterns of standard female speakers and high-level female learners, it can be seen that the resonance peak distribution of vowels pronounced by female learners is more stable, and the patterns of the first resonance peaks and the second resonance peaks of some vowels are basically similar to the resonance peak distribution patterns of the vowels pronounced by the standard female speakers, but there are still differences in some vowels. The overall resonance peaks of female speakers were higher, up to 2625 Hz.

4.2. Quality Assessment Testing Performance

Randomly select a pronunciation signal as the experimental object, using the model method of this paper for the feature decomposition of the spoken English pronunciation signal, according to the feature decomposition results of the spoken English pronunciation signal adaptive filtering detection and

spectral analysis, to achieve the signal detection and recognition, to obtain the detection results are shown in Figure 8.

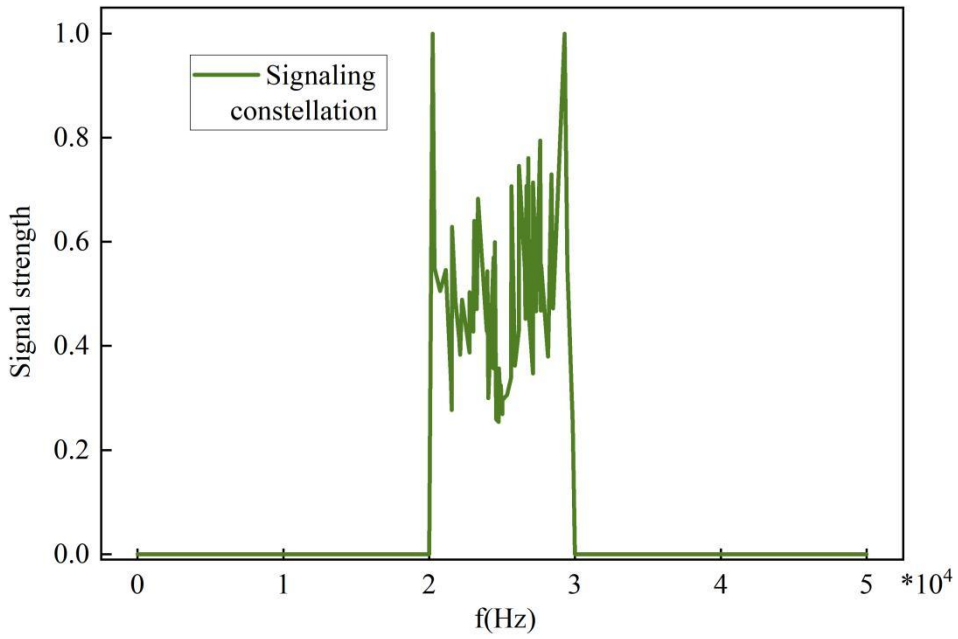

Figure 8. Quality assessment and detection of oral English pronunciation signals.

Analyzing Fig. 8, it can be seen that the discriminative power of the spoken English pronunciation signal detection and evaluation using this paper's method is strong, and it can accurately detect, capture and evaluate the pronunciation signals, and the quality assessment detection results provided are consistent with the original acquisition results.

The accuracy of this paper's method and the traditional method for automatic assessment of spoken English pronunciation quality is tested, and the comparison results are shown in Fig. 9. Overall, the quality assessment error of this paper's method is controlled between 0.001 and 0.1, and begins to converge at 70 iterations. In contrast, although the quality assessment error of the traditional method is 0.002 at 100 iterations, the overall error is higher, 0.53 at the 10th iteration.

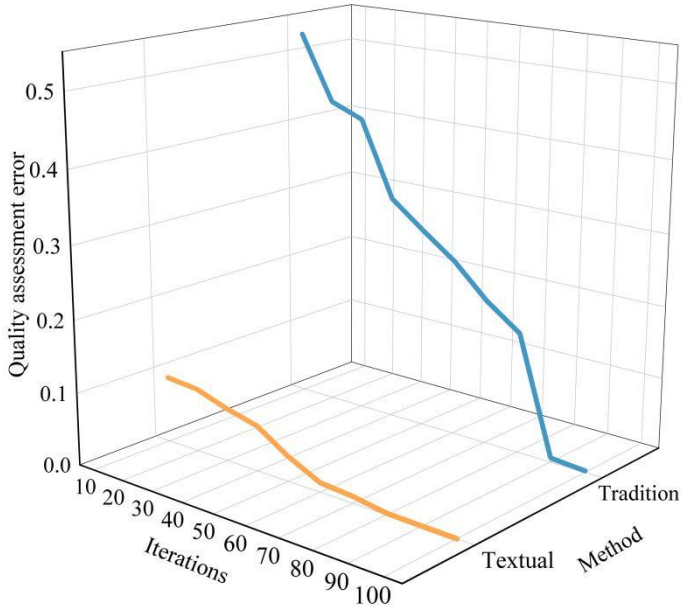

Figure 9. Automatic assessment of oral English pronunciation quality.

4.3. Voice Assessment Performance

In this section, the overall validation experiment of the designed model is conducted, T1 (containing 1050 native natural spoken language recordings and their scripts), T2 (containing 150 native natural spoken language recordings), and T3 (containing 200 spoken language recordings that intentionally ignored the contrapuntal voicing within the contrapuntal group) are used as the developmental test set to set up the evaluation results: (Excellent) Excellent, (Good) Good, (Medium) Pass, (NI) To be improved. Considering that different people recording these corpus sets have significantly different speaking levels, the system is set to have an average factor recognition rate of 85%, and is expected to be able to give significantly different evaluation results for T1, T2, and T3.

Figure 10 shows a comparison of the results of speech assessment for T1, T2, and T3. These results show that the system gives a 52% "Excellent" for Native-pronounced T1, compared to only 7% for T3. At the same time, T3 gets 30% "Medium" and 10% "NI", while T1 only gets 10% "Medium" and 0% "NI". The ratio of the different assessment levels to T2 was between T1 and T3. This is consistent with the expected results of the experiment, which demonstrates that reliable and robust evaluation results can be obtained at an average phoneme recognition rate of 85% of the system.

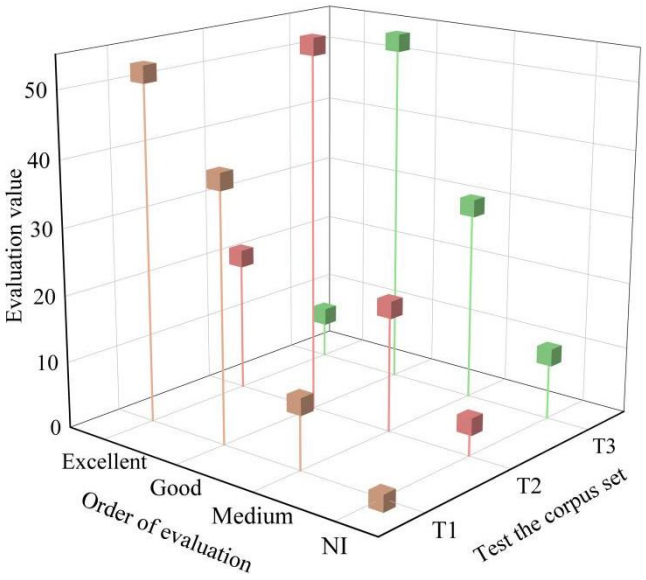

Figure 10. The comparison of opening test results for test corpora T1-T3.

5. Conclusion

This paper establishes an English pronunciation evaluation system with the three modules of segmental verification, language signal cutting and pronunciation evaluation as the main architecture. At the same time, MFCC is chosen as the feature parameter of English speech, and the theoretical model of the proposed English pronunciation evaluation system is constructed by combining the GOP method based on deep neural network.

The theoretical model of the English pronunciation evaluation system is used to extract the resonance peaks of English spoken vowels, and it is found that the first resonance peaks of male speakers are all larger than 1000 Hz, and the second resonance peaks are all between 300-800 Hz. The overall resonance peak distribution of female speakers is slightly higher than that of male speakers. In the feature decomposition of English pronunciation signals, the modeling method in this paper not only provides quality assessment detection results that are consistent with the original acquisition results, but also controls the quality assessment error to be between 0.001-0.1, which shows superior assessment performance and stability.

Funding

This research was supported by the: Fund Project 1: Excellent Youth Project of Science Research Project of Education Department of Hunan Province in 2022: Research on the Cultivation Path of “Cultural Confidence” in Public English Teaching in Higher Vocational Colleges under the Background of Digitalized Education (Project Number: 22B1022); Fund Project 2: Special Research Project of Hunan

Modern Logistics College in 2024: Application of Artificial Intelligence Technology in Higher Vocational English Teaching (Project Number: JYZ202401).

References

1. Hu, J. (2021). Teaching evaluation system by use of machine learning and artificial intelligence methods. International Journal of Emerging Technologies in Learning (iJET), 16(5), 87-101.
2. Lv, Z., & Shen, H. (2021). Artificial intelligence with fuzzy logic system for learning management evaluation in higher educational systems. Journal of Intelligent & Fuzzy Systems, 40(2), 3501-3511.
3. Owan, V. J., Abang, K. B., Idika, D. O., Etta, E. O., & Bassey, B. A. (2023). Exploring the potential of artificial intelligence tools in educational measurement and assessment. Eurasia journal of mathematics, science and technology education, 19(8), em2307.
4. Dong, J., Mohd Rum, S. N., Kasmiran, K. A., Mohd Aris, T. N., & Mohamed, R. (2022). Artificial intelligence in adaptive and intelligent educational system: a review. Future Internet, 14(9), 245.
5. Hang, Y., Khan, S., Alharbi, A., & Nazir, S. (2024). Assessing English teaching linguistic and artificial intelligence for efficient learning using analytical hierarchy process and Technique for Order of Preference by Similarity to Ideal Solution. Journal of Software: Evolution and Process, 36(2), e2462.
6. Zhang, X., & Chen, L. (2021). College English smart classroom teaching model based on artificial intelligence technology in mobile information systems. Mobile information systems, 2021(1), 5644604.
7. Fitria, T. N. (2021). The use technology based on artificial intelligence in English teaching and learning. ELT Echo: The Journal of English Language Teaching in Foreign Language Context, 6(2), 213-223.
8. Zhang, G. (2023). The Evaluation and Development of University English Teaching Quality Based on Wireless Network Artificial Intelligence. Journal of Combinatorial Mathematics and Combinatorial Computing, 117, 77-86.
9. Zhu, L., Yan, X., & Wang, J. (2022). A Recognition Method Based on Speech Feature Parameters-English Teaching Practice. Mathematical Problems in Engineering, 2022(1), 2287468.
10. Liu, L., Li, W., Morris, S., & Zhuang, M. (2023). Knowledge-Based Features for Speech Analysis and Classification: Pronunciation Diagnoses. Electronics, 12(9), 2055.
11. Wang, Y. (2021). Detecting pronunciation errors in spoken English tests based on multifeature fusion algorithm. Complexity, 2021(1), 6623885.
12. Sun, X. (2024). Oral Assessment Model: Assessing the Quality of Pronunciation in English Reading. International Journal of Acoustics & Vibration, 29(2), 187.
13. Sheng, Y., & Yang, K. (2021). Automatic Correction System Design for English Pronunciation Errors Assisted by High-Sensitivity Acoustic Wave Sensors. Journal of Sensors, 2021(1), 2853056.
14. Li, Q. (2023, July). Design of An Automatic Quality Evaluation System for English Pronunciation Machines Based on SVM. In 2023 International Conference on Data Science and Network Security (ICDSNS) (pp. 1-6). IEEE.
15. Arjmandi, M. K., & Pooyan, M. (2012). An optimum algorithm in pathological voice quality assessment using wavelet-packet-based features, linear discriminant analysis and support vector machine. Biomedical signal processing and control, 7(1), 3-19.
16. Jing, W. (2024). Speech recognition sensors and artificial intelligence automatic evaluation application in English oral correction system. Measurement: Sensors, 32, 101070.
17. Fan, Z., Li, J., Wumaier, A., Kadeer, Z., & Abdurahman, A. (2023). A multifaceted approach to oral assessment based on the conformer architecture. IEEE Access, 11, 28318-28329.
18. Xiao, W., & Park, M. (2021). Using automatic speech recognition to facilitate English pronunciation assessment and learning in an EFL context: pronunciation error diagnosis and pedagogical implications. International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT), 11(3), 74-91.
19. Duan, J., & He, Z. (2024). RETRACTED ARTICLE: An English pronunciation and intonation evaluation method based on the DTW algorithm. Soft Computing, 28(Suppl 2), 491-491.
20. Shi, X., Wang, X., & Zhang, W. (2024). Exploring the relationships between ASS indices and CAF and the impact on Chinese college students' oral English performance. Language Testing in Asia, 14(1), 30.