[1,2*] Xue Wang

Research on Oral English Learning System Integrating AI Speech Data Recognition and Speech Quality Evaluation Algorithm

Journal of Electrical Systems

Abstract: - The application of AI speech data recognition (SDR) and speech quality assessment algorithms in spoken English learning systems can help to realize the intelligence and automation of the process related to spoken English learning. To better assist the people concerned with the learning of spoken English, it is necessary to integrate AI speech data recognition technology in the process of learning spoken English to bring learners a learning environment and learning conditions that are not bound by time and space. As an important way to realize the natural communication of information between machines and people, AI speech recognition technology has significant application advantages and application values in eliminating the influence of the native language of English speaking learners and innovating the learning mode in the local area. In addition, the assessment of the reliability of spoken English language learning can be carried out in several dimensions. However, it is still difficult to unify the criteria of many sizes, bringing new application scenarios for AI SDR. Based on this, this paper first analyzes the theoretical basis of AI SDR. Then it investigates the speech data recognition method of the fused AI English-speaking learning mechanism and the process of fused AI English-speaking speech quality assessment. Finally, it designs and verifies the functions of the spoken English learning mechanism incorporating AI speech data identification and quality assessment.

Keywords: Speech English Learning mechanism; AI Speech Data Recognition; Speech Quality; Assessment Algorithm.

Introduction

With the continuous expansion of the global economy, commerce, and integration, transnational exchanges are becoming increasingly close and frequent. In this context, English, as one of the most common and popular languages in the world, its importance and status cannot be underestimated [1]. On the other hand, with the continuous expansion of intelligence and info tech, intelligent algorithms and technologies represented by AI have been widely and deeply studied in many fields and have made remarkable achievements [2]. Specifically, in the area of oral English learning, the utilization of AI SDR and speech quality evaluation algorithm in oral English learning system has significantly released teachers' teaching pressure and students' learning flexibility, initiative, and enthusiasm [3]. On the one hand, the integration of AI speech data recognition and oral English learning system can significantly ameliorate the convenience of the learning procedure; On the contrary, the integration of a speech quality evaluation algorithm can effectively test the effectiveness of the whole system.

As an essential part of oral English learning, phonetic data learning is a meaningful way to master the key elements of language info transmission and communication. The English language is an important data element and carrier of the language. Its quality evaluation can be done from multiple dimensions, including clarity, naturalness, and intelligibility [4]. It is still difficult to unify the evaluation standards of these three dimensions, mainly because there are great difficulties in the naturalness evaluation of speech quality, and the emergence of intelligent algorithms provides a new opportunity for the unification of speech quality evaluation standards [5]. With the maturity of AI speech data recognition and speech signal processing tech, the links to oral English learning have gradually realized intelligence and automation and significantly accelerated learners' understanding and mastery of oral English and speech key points [6].

At present, speech data recognition tech has made significant progress, and the utilization of this tech in related fields has effectively facilitated the degree and effect of human-computer interaction [7]. However, in the area of oral English learning, the research and utilization of SDR and speech quality evaluation are still scarce, and it has been difficult to effectively support the practical requirements of oral English learning [8]. Under the background of more remote, terminal, and convenient oral English learning, to better assist relevant personnel in carrying out oral English learning, it is necessary to integrate AI speech data recognition tech into the process of oral English learning, to bring learners a learning environment and learning conditions free from time and area.

In addition, in the procedure of oral English learning, integrating speech quality evaluation algorithms can intuitively analyze and evaluate learners' oral pronunciation and provide learners with feedback info, such as data correction, to lay a foundation for improving learners' oral pronunciation quality [9]. The reliability of spoken English pronunciation is an important prerequisite for effective oral communication and is vital in establishing effective info transmission between the two sides of communication [10]. Speech quality

[1] Zhengzhou University of Economics and Business, Zhengzhou, Henan, 451191, China
[2] University of Gdańsk, Jana Ba_y_skiego 8, Gdansk, 80-309, Poland
e-mail: 1061670337@qq.com
Copyright © JES 2024 on-line : journal.esrgroups.org

assessment involves many specialties and disciplines, including many aspects such as speech recognition, linguistics, and psychology. Therefore, building a scientific English speech quality assessment algorithm requires the organic integration of multiple specialties and disciplines.

Integrating AI speech data recognition and speech quality evaluation algorithm can ameliorate the human-computer interaction of oral English learning systems, make the learning process more intelligent and automatic, and significantly reduce the cost of oral English learning [11]. The existing speech data recognition still has great potential and space for amelioration in intelligence, interactivity, and accuracy. The integration of AI speech data recognition and speech quality evaluation algorithm enhances the dynamics of speech evaluation [12]. The core of speech quality evaluation is to establish an evaluation model to compare the quality of the initial input signal and the signals processed by the system to evaluate the speech quality [13] objectively. In addition, because the performance, efficiency, and reliability of the traditional speech quality evaluation algorithm are still not enough to support the practical needs, the research on the Speech English learning mechanism integrating speech data recognition and speech quality evaluation algorithm has important practical value.

Many standard voice augmentation techniques are accessible and have drawn much interest in automated speech identification. These improvement techniques, nevertheless, only function effectively with straightforward, constant audio sources. As a result, a “Hybrid Speech Enhancement Algorithm (HSEA)” is suggested in this work to improve speech detection efficiency. With the help of the Hidden Markov Approach, the well-known non-linear spectrum subtraction speech improvement technique is optimized and brought to the test on 1440 "Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) audio" recordings and 6660 clinical voice translation audio files. As the level of speech filtration is raised, it becomes more mathematically sophisticated, which continues to be a drawback of the suggested approach [22].To tackle concerns about voice recognition data protection, we provide a unique decentralized feature extraction strategy in federated training. In a proposed design, Mel-spectrogram is initially extracted from an input voice using a quantum computing server, and the related convolutional features are then encoded employing a quantum circuit approach with random variables. This increases modeling parameter security. More computation time is needed [23]. Using profound feature alteration, deep learning may provide high-level and high-quality characteristics, increasing categorization reliability. Specific samples are of poor quality for categorizing many causes, including data acquisition. Sample learning is thus required. So, in this study, a deep “dual-side learning ensemble framework” is developed. The extensive dual-side learning of Parkinson's voice data is made possible in this framework by the construction and combination of a deep sample learning technique with a deep network [24]. The lack of certified sign language teachers and the high cost of assistive equipment are the significant challenges confronting students who are deaf or hard of hearing. In this work, they have developed a visual speech identification technique using modern deep-learning models. Additionally, the current approaches are problematic. They thus suggest a unique method by fusing the outcomes from audio and visual speech to overcome the inadequacies found. To effectively read lips, this research proposes a novel deep learning-based audio-visual speech recognition model. The rate of errors is greater [25]. Currently, end-to-end (E2E) approaches founded on recurrent patterns have lost ground to transformer-based E2E automatic speech recognition (ASR) ideas, which have been proven to function better on various ASR applications. Transformer ASR, similar to other E2E designs, calls for the whole input series to calculate the attention on the encoder and decoder, which increases latency and presents a problem for continuous ASR. To overcome the delay problem and promote interactive ASR, the study suggests the Decoder end Adaptive Computation Steps (DACS) method. The power consumption is very high. [26]. Due to this existing literature drawback, we proposed an AI speech data recognition and quality evaluation algorithm.

Theoretical Basis

Speech data recognition involves many specialties and disciplines. Man-machine communication and related interactive operations are realized through the technical integration of interdisciplinary fields [14]. Speech data recognition at the level of oral English learning mainly includes word speech recognition, phrase speech recognition, and paragraph speech recognition. The speech data recognition system can also be further divided according to the word vocabulary and the dependence on the speech subject. For example, it can be divided into small, medium, many, and infinite according to the difference in word vocabulary. According to the dependence on the speech subject, it can be divided into the need to designate a specific person and the need not to mean a person. The workflow of the spoken English speech recognition mechanism is shown in Figure 1 below.

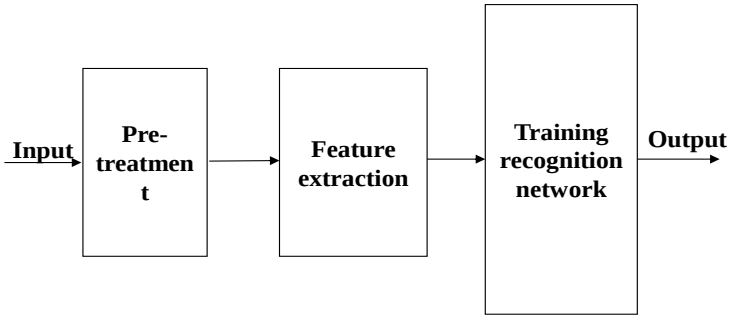

Figure1. Diagram of components of the speech recognition system

In addition, since spoken English will attenuate in the transmission process after a speech is sent out, it is necessary to enhance the speech signal before speech data signal processing. The enhancement calculation formula is shown in Formula 1 below, where λ the coefficient is located in the interval of (0,1).

$$f(x) = g(x) - \lambda g(x-1) \tag{1}$$

The recognition of spoken English speech data is inseparable from the discrimination of speech input point and termination point, so it is necessary to introduce a breakpoint detection mechanism in the recognition process [15]. Endpoint recognition of speech data includes short-term energy calculation and short-term mean zero combining ratio calculation. In the analysis of short-term average amplitude, the concept of the average amplitude of speech energy is introduced to avoid the amplitude difference between adjacent values. Secondly, in discrete speech signal sampling, there will be a zero crossing phenomenon, which can directly reflect the frequency and approximate spectral characteristics of the speech signal.

Oral English speech data includes several types, as shown in Figure 2 below. Through the combination of time and frequency field attributes, the time spectrum is formed to mobilize the time-series data of oral English speech.

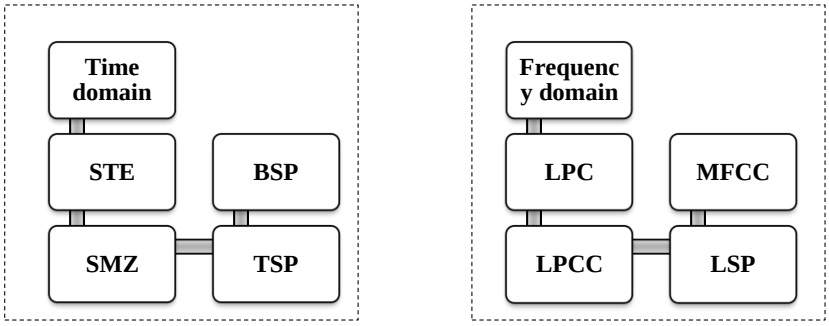

Figure2. Types of Spoken English speech data

The calculation of LPC is based on the actual speech of spoken English and the sampling of predicted values [16]. The generation model of spoken English speech is calculated as shown in equation 2 below, where a_i is the model parameter. Compared with LPC, LPCC has more significant advantages in describing standard peak features of spoken English speech and the dependence on incentive data.

$$Q(x) = \frac{1}{1 - \sum_{i=1}^{n} a_i x^{-i}} \tag{2}$$

The calculation process of MFCC is shown in equation 3 below, and L is the number of MFCC coefficients. Firstly, the voice data is converted into a spectrum; secondly, the range is transformed into an energy spectrum, and finally, through IFT.

$$M_n = \sum_{i=1}^{K} \log Z(i) \cos[\pi(i-0.5)n/K] \qquad n = 1, 2, \cdots L \tag{3}$$

ZCPA has a significant impact on the stability and anti-noise of oral English speech recognition system, which requires the speech data recognition system to have auditory solid perception ability and anti-environmental noise ability. The working principle of ZCPA is depicted in Figure 3 below. The voice data recognition system uses data filtering, detection, receiving, and other devices to collect and synthesize voice data.

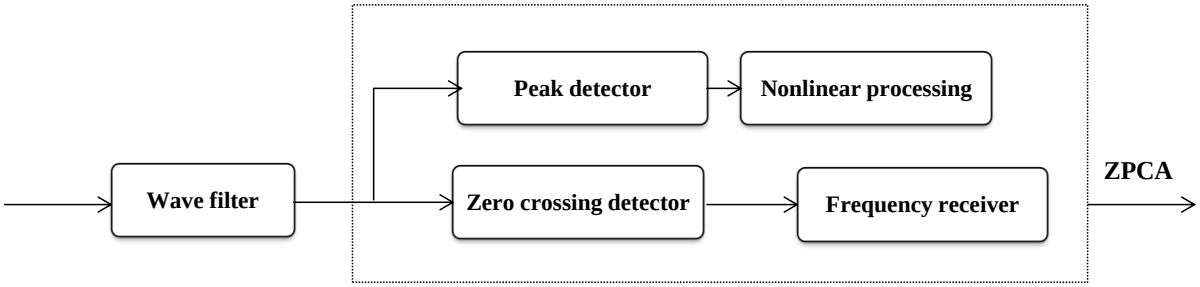

Figure3. Working principle of ZCPA

The extraction of spoken English speech data is to collect the internal attributes of speech data and distinguish the attribute info of speech, to judge the source of speech. Secondly, the calculation of spoken English speech feature parameters helps to ease the recognition effect of speech data and realize the practical analysis of feature parameter data in speech signals [17]. In addition, the feature extraction of spoken English speech data recognition should pay attention to the collection of typical said English signal feature parameters and data compression. Because there are significant differences in spoken English speech features among different groups, to attain precise speech recognition, it is important to collect more feature parameters containing semantic data to eliminate and reduce unnecessary personalized difference data.

Speech Data Recognition Method

The common types of speech data recognition in oral English learning systems include ANN, channel model, pattern matching, knowledge speech, and statistical model. With the iterative progress of intelligent tech, the practicability of traditional knowledge speech and channel models has gradually lagged behind the requirements of reality, and the intelligent speech recognition methods represented by ANN, BP, and RBF have become the focus of research and development.

Under the iterative progress in global integration, English, the primary language for communication among countries, has become a vital supporting force for cross regional and cross-industry communication. Firstly, the utilization of speech recognition in oral English learning can effectively capture and record learners' errors in the process of pronunciation to establish a premise for subsequent improvement. Secondly, there are many problems in traditional oral English teaching, such as uninteresting courses and challenges in realizing students' subject status. The AI oral learning system can effectively trigger students' learning initiatives. In addition, the learning system can establish interactive channels between students and between teachers and students to achieve accurate teaching assistance and interaction. Finally, because AI's oral English learning mechanism is not limited by teaching space and time, this system can make learners' oral English learning process more flexible and convenient.

AI speech data recognition system represented by ANN uses the principle of adaptive nonlinear dynamics to establish a self-organizing high error tolerance learning and prediction network to realize adaptive learning and recognition of speech data in multiple scenes and multiple environments [18]. With the support of ANN tech, the spoken English speech data recognition system can continuously carry out self-organizing training and learning, and constantly iterate and optimize its speech data recognition performance. The data parallel processing ability of the mechanism makes the data processing of the whole system more stable and fault-tolerant. Therefore, it has significant utilization advantages in the field of oral English speech data recognition. The principle architecture of AI integrated spoken English speech data recognition mechanism is depicted in Figure 4 below.

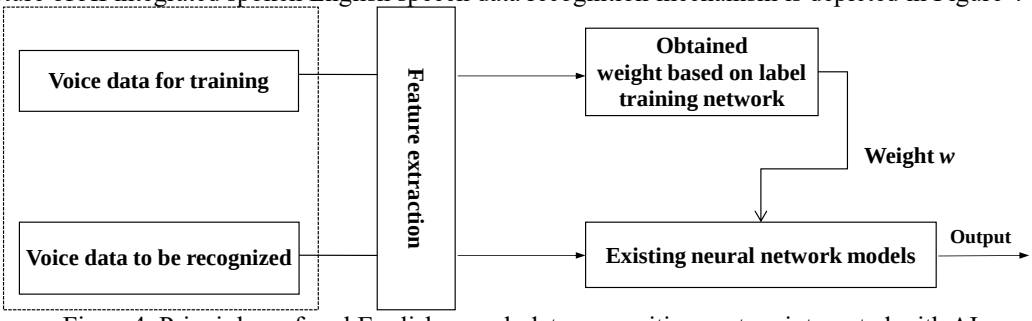

Figure4. Principle a of oral English speech data recognition system integrated with AI

After capturing different speech data, the spoken English speech data recognition system determines the connection weight of the neural network through nonlinear calculation [19]. The speech data recognition system integrated with AI can proofread and adjust the recognition results of speech data adaptively, so the system has robust and intelligent perception ability. In addition, the reference template of speech recognition is established through vector quantization to recognize the input vector sequence of spoken English speech data. In the process of vector sequence recognition of spoken English speech data, the distortion quantization error of sequence data is calculated and compared to judge whether the recognition result meets the minimization requirement of total

average distortion error. The vector quantization recognition procedure of the speech data recognition system is depicted in Figure 5 below.

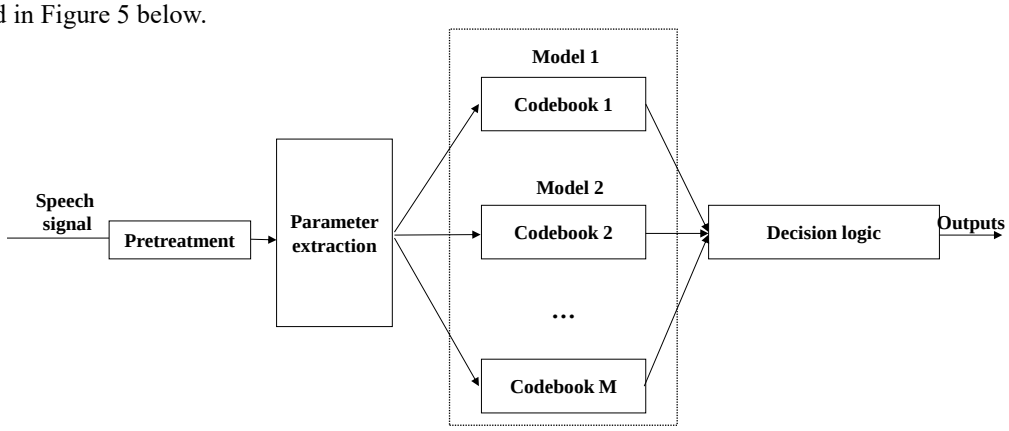

Figure5. Vector quantization recognition process of speech data recognition system

The speech quality evaluation algorithm needs to systematically, comprehensively, and scientifically analyze and evaluate the fused speech data, and verify the effectiveness of the data. Secondly, the speech quality evaluation algorithm needs to have high robustness and real-time. The algorithm can meet the utilization needs of multiple scenes so that learners can apply the system more conveniently and quickly. In addition, to further ensure the efficiency of the whole evaluation system, the speech data evaluation algorithm should also be able to adapt to the diversified and personalized characteristics of learners. It should not affect the reliable performance of the whole system due to learners' personalized acoustic features.

Algorithm Process

Firstly, for the spoken English speech data sequence $E=\{E_1,E_2,...E_T\}$ and its model $\lambda=(X,Y,\pi)$ to be recognized, the probability of recognition sequence under specific conditions must be calculated. The definition of the forward variable is shown in equation 4 below. The variable is expressed as the probability that the spoken English speech data sequence to be recognized before time *t* is in a particular state *A* at time *t*.

$$F_t(i)=P(E_1,E_2,...,E_t;\ f_t=A_i\mid\lambda) \tag{4}$$

First, initialize FF, as shown in equation 5 below:

$$F_1(i)=\pi_i y_i(E_1)\quad 1\le i\le N \tag{5}$$

Then, the iterative calculation is shown in equation 6 below, in which x_{ij} $y_j(E_{t+1})$ state transition and observation symbol elements, respectively.

$$F_{t+1}(j)=[\sum_{i=1}^{N}F_t(i)x_{ij}]y_j(E_{t+1}),1\le t\le T-1;1\le j\le N \tag{6}$$

Similarly, the definition of the backward variable of the probability that the spoken English speech data sequence to be recognized after time t is in a particular state *A* at time *t* is as shown in equation 7:

$$B_t(i)=P(E_{t+1},E_{t+2},...,E_T;\ f_t=A_i\mid\lambda) \tag{7}$$

The iterative calculation $B_t(i)$ is shown in the following equation 8:

$$B_t(i)=\sum_{i=1}^{N}x_{ij}y_j(E_{t+1})B_{t+1}(j),t=T-1,T-2,...,1,1\le i\le N \tag{8}$$

The probability that the whole spoken English speech data sequence to be recognized is in state A_i at time *t* is calculated and depicted in below equation 9:

$$P(E\mid\lambda)=\sum_{i=1}^{N}F_t(i)B_t(i)=\sum_{i=1}^{N}P(E,\ f_t=A_i\mid\lambda),\ \ t=1,2,\cdots,T;\ \ i=1,2,...N \tag{9}$$

Secondly, for the spoken English speech data sequence $E=\{E_1,E_2,...E_T\}$ and its model $\lambda=(X,Y,\pi)$ to be recognized, it is necessary to determine the best state of the series to be identified. The probability that the speech data sequence *E* and the model to be recognized are in the form A_i at time *t* is shown in equation 10 below.

$$\Psi_t(i) = P(f_t = A_i \mid E, \lambda) = \frac{F_t(i)B_t(i)}{\sum_{i=1}^{N} F_t(i)B_t(i)} = \frac{F_t(i)B_t(i)}{P(E \mid \lambda)} \tag{10}$$

The calculation of the most likely state of the speech data to be recognized at any time is shown in the following equation 11:

$$f_t = \arg\max_{1 \le i \le N}[\Psi_t(i)], 1 \le t \le T \tag{11}$$

In addition, for a given spoken English speech data sequence to be recognized, the parameters must be adjusted to $P(E \mid \lambda)$ reach the maximum value. The adjustment of the parameter model involves the training and re-evaluation of the model. The variables shown in equation 12 below are defined to calculate the probability of oral English speech data sequence in state A_i at time *t* and state A_j at time *t + 1*.

$$\Gamma_t(i, j) = \frac{F_t(i)x_{ij}y_j(E_{t+1})B_{t+1}(j)}{P(E \mid \lambda)} = \frac{F_t(i)x_{ij}y_j(E_{t+1})B_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} x_t(i)x_{ij}y_j(E_{t+1})B_{t+1}(j)} \tag{12}$$

Spoken English Speech Quality Evaluation

With the iterative progress of smart tech represented by a neural network, its utilization in oral English learning has effectively accelerated the maturity and perfection of the oral English learning mechanism. In the natural language field represented by oral English, grammar and semantics are independent. AI and other intelligent neural networks create a deep-level model by describing phonetic grammar and semantic data to characterize the effective measurement of multiple-interactive info. In the decoupling process of spoken English speech data recognition, it is necessary to change the model of nonverbal factors and establish image coding.

The decoupling of each level of speech is realized through the interaction of many groups. The speech is divided into many discrete elements sets through speech understanding, and the representation of grammatical structures of different language structures under neural oscillation is analyzed [20]. The band acquisition distribution of spoken English speech data is shown in Figure 6 below. These band activities at different levels are an inevitable process to realize spoken English speech processing to divide the time sequence of speech data and establish an interconnected internal organizational structure. In addition, in the complex speech interaction environment, it is necessary to systematically encode the critical info in the speech stream to facilitate the recognition accuracy of oral English speech data.

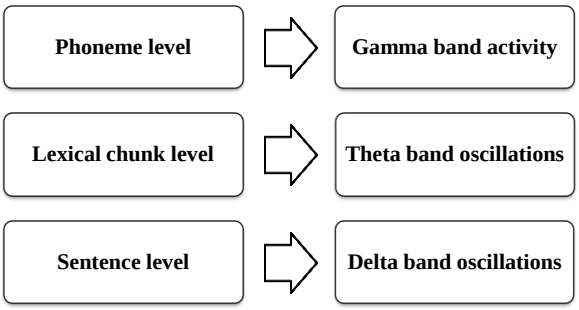

Figure6. Band acquisition and distribution of spoken English speech data

The Evaluation Mechanism

Establishing the assessment mechanism for English speech accuracy evaluation needs to be evaluated according to the similarity between the estimated speech and the comparative reference speech in the speech database. Firstly, it is necessary to preprocess the spoken English speech data and extract high-value parameter features. Secondly, the evaluation of oral English speech quality needs to be carried out on speech perception, clarity, and content rhythm. In addition, the intelligent algorithm is used to evaluate and grade the oral English speech data, and the final assessment findings are given. The assessment mechanism of spoken English speech quality must also introduce acoustic and perceptual models to obtain the evaluation results after data fusion through correlation calculation.

As a fundamental part of oral English pronunciation quality assessment, evaluating pronunciation quality can directly reflect the standard degree of pronunciation. The speech quality evaluation scheme will have a direct and significant impact on the effectiveness of the whole learning system. At present, there are two types of evaluation schemes for oral English speech quality: the speech feature evaluation method and the speech acoustic model evaluation method. Among them, the former mainly uses the comparative analysis between the reference data in the standard speech database and the evaluated data to obtain the evaluation results, while the latter does not need to establish a traditional speech reference database.

In addition, the speech feature evaluation method extracts the feature parameters of the evaluated speech signal data. It compares them with the standard speech data to obtain their similarity to output more objective evaluation results. The acoustic model evaluation method classifies the speech data signals to be evaluated, establishes the corresponding classification model, trains and aligns the extracted feature parameters, and obtains the evaluation results through the scoring mechanism. It can be seen that acoustic.

The subject of oral English speech quality evaluation is mainly composed of objective subject and subjective issues. Among them, the objective evaluation subject primarily uses computer software and hardware media system to record and evaluate the oral English data. According to the differences in the evaluation methods adopted, the subjective evaluation subjects include many kinds, mainly including a dam, MOS, DRT, sharpness measurement, and so on. These two dimensions of speech quality evaluation methods can effectively judge the quality of spoken English speech. Objective evaluation still has a large optimization space at the level of natural perception. Therefore, it is crucial to further study the construction of AI Integrated English speech data recognition at the level of eliminating the interference of perception factors.

As a virtual extension of the task of speech data recognition in oral English learning, the evaluation of oral speech quality can comprehensively and intelligently evaluate its pronunciation quality. The assessment of oral English pronunciation accuracy from the perspective of learning has local explanatory characteristics [21]. By establishing an intelligent analysis and evaluation model, it could compare, locate and solve oral pronunciation, defects, and problems. Selecting a systematic analysis model is the premise of human-computer interaction in the oral English learning system. Through the training and learning of the model, it could realize the understanding of pronunciation, and then output the follow-up action plan to learners and system users. The intelligent algorithm, on account of the local interpretation framework, can achieve a good balance in local prediction, transparency, and accuracy, achieve good prediction transparency, and give intuitive analysis and interpretation.

Because the complexity of the objective function in the spoken English speech quality evaluation algorithm is too high, the evaluation process requires too much computational power. Therefore, it is necessary to establish an alternative function to analyze the objective attributes. By inputting training data, Lasso, decision tree, and LRM models are selected to replace the sample instance data points to obtain the disturbed data set. The calculation of the sample case and the corresponding prediction model is shown in equation 13, where y is the sample case, m is the prediction model, $\Phi(m)$ is the function of model m, and ξ_y is the disturbance measure.

$$E(y) = \arg\min_{m \in M} L(f, m, \xi_y) + \Phi(m) \tag{13}$$

As a data enhancement tech under disturbed data sets, it includes a variety of transformation forms and realizes the combined evaluation of speech data through specific transformation. Secondly, the speech segment of oral English speech data is replaced, and its quality is evaluated. In addition, the AI neural network program is used to proofread and optimize the computer learning program, and the local interpretable alternative model is used to complete the self-construction.

System Design and Results Verification

The database design of oral English learning system integrating AI speech data recognition and quality evaluation includes the design and implementation of the concept and logical structure, and the storage of speech data is realized through the design and construction of these two dimensions. Among them, at the level of conceptual structure design, it mainly includes entity analysis and the construction of an entity relationship diagram. Secondly, at the design level of logical architecture, its design and implementation have many data tables, such as users, voice, evaluation, and text. In addition, the creation of the user login module of oral English learning system includes the design of input and output items, the operation process of the user login module, and the design of the user login interface. The login process of oral English learning system integrating AI speech data recognition and quality evaluation is shown in Figure 7 below.

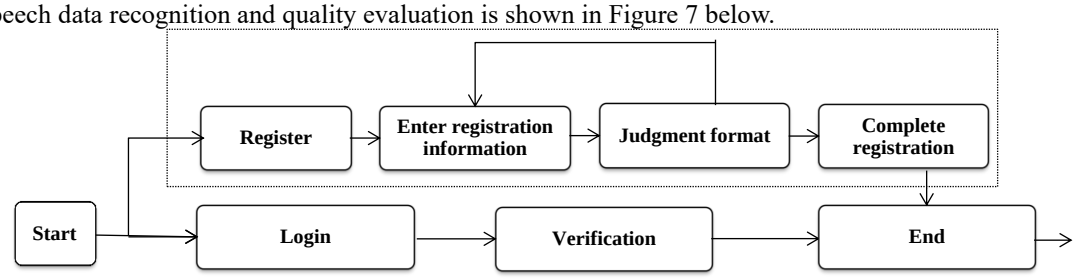

Figure7. Login process of oral English learning system

Speech Recognition Module

The speech recognition module in the oral English learning system integrating AI speech data recognition and quality evaluation includes hardware interaction, data acquisition, function reading, and interface provision. From the perspective of acquisition parameters of spoken English speech data, it is necessary to set the trigger

and end thresholds for the acquisition of these parameters. Secondly, the training process of spoken English speech recognition is shown in Figure 8.

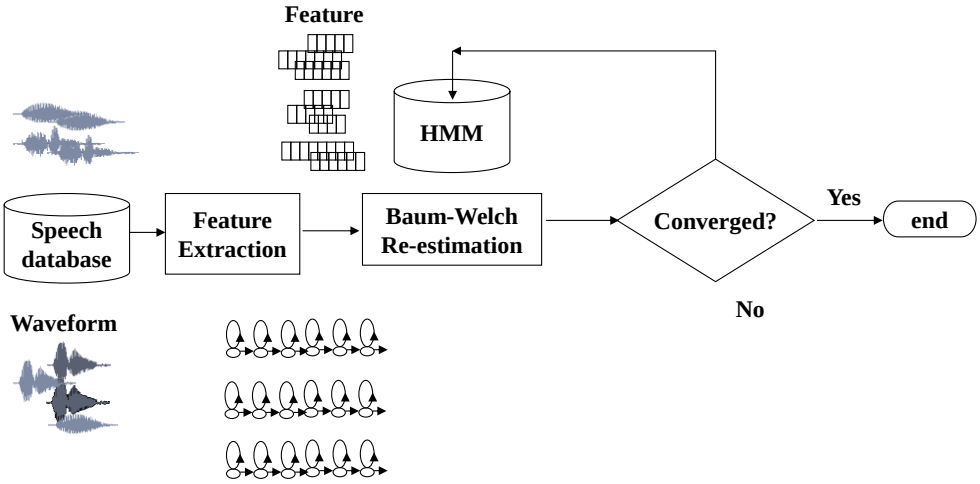

Figure8. The training process of spoken English speech recognition

Through reading the codebook and template file, collect spoken English speech data, preprocess the data, extract feature parameters, vector quantization, parameter probability calculation, and template file matching, and finally obtain and output correct text info. In addition, the acquisition and recognition of spoken English speech are mainly controlled by three signals: trigger, pause, and termination. The spoken English speech recognition library loads the local speech recognition library by calling the speech recognition interface to realize the framing of speech data, endpoint detection, and dynamic reading of characteristic parameters and templates.

The speech quality evaluation process of the oral English learning system is to load the template file through speech recognition, and because the system has the iterative function of self-training and learning, it can realize systematic speech quality evaluation and analyze the evaluation results. First, the voice quality evaluation module of the system will collect voice input data after triggering wake-up. Secondly, after beginning, the system will compare the input voice with the voice and semantic database in the database to objectively evaluate the quality of users' oral English voice, and send the evaluation results to the display terminal system.

The fusion of spoken English speech data is mainly the accurate and complete integration of multivariate and multi-source data to ensure the objectivity, scientificity, and accuracy of the data evaluation results. The speech data to be fused mainly comprises acoustics, perception, energy, and pitch. The similarity, perception, and phonological differences between speech and standard library are evaluated, respectively. The evaluation data and results of these different dimensions need further data fusion processing to ensure the integrity and objectivity of the output evaluation results. In addition, the voice data fusion process needs to establish the corresponding data mapping channel and then realize the attribute mapping of the evaluation results by selecting the proper mapping function and data fusion algorithm to obtain the final result.

To verify the performance and effect of the oral English learning system integrating AI speech data recognition and quality evaluation, it is necessary to further test each module of the system. By testing and confirming the actual functions of the whole system, the specific utilization scenarios applicable to the entire system can be verified. Firstly, at the functional verification level of the oral English learning system, it mainly tests the preconditions, uses steps, and actual effects of the whole system, including the login module, speech recognition module, and quality evaluation module. The speech data recognition performance of the learning system is verified by setting different experimental control groups. Table 1 below shows the recognition effect of the system under different amounts of oral English vocabulary data. It can be seen from Table 1 that the speech data recognition rate of the learning system is different under different background environments and different vocabulary data, but it can meet most oral English learning scenarios. Correct identification and recognition accuracy are more important to speech recognition technology. It shows the efficiency of the method. Figure 9, 10, and 11 shows the correct identification with a different environment. Figure 12, 13, and 14 depicts the recognition accuracy of different environments.

Table 1: Recognition effects under different amounts of spoken English vocabulary data.

Environment	Vocabulary range/quantity	Number of tests	Correct identification	Recognition accuracy
Normal classroom	100	500	497	99.4%
	200	500	494	98.8%
	500	500	486	97.2%
Normal outdoor	100	500	495	99.0%

	200	500	487	97.4%
	500	500	483	96.6%
Noisy outdoor	100	500	493	98.6%
	200	500	486	97.2%
	500	500	478	95.6%

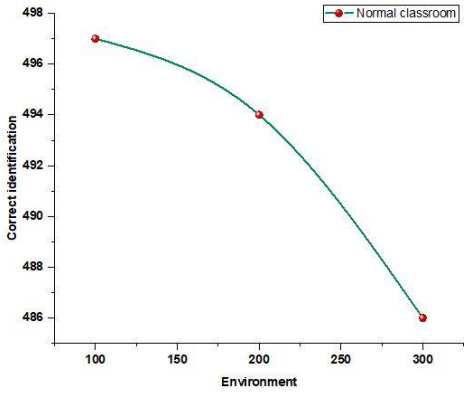

Figure 9. Correct identification with normal classroom

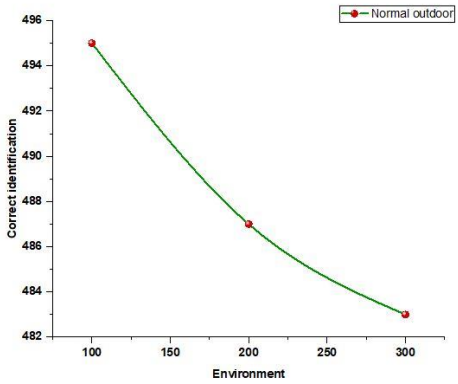

Figure 10. Correct identification with normal outdoor

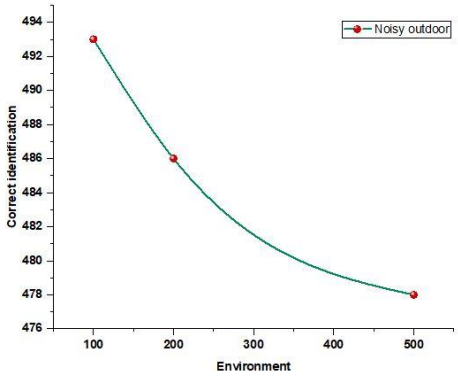

Figure 11. Correct identification with noisy outdoor

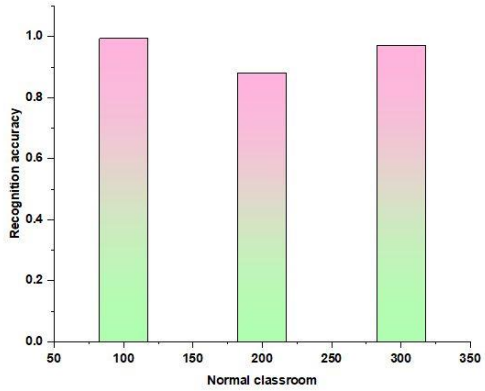

Figure 12. Recognition accuracy in a normal classroom

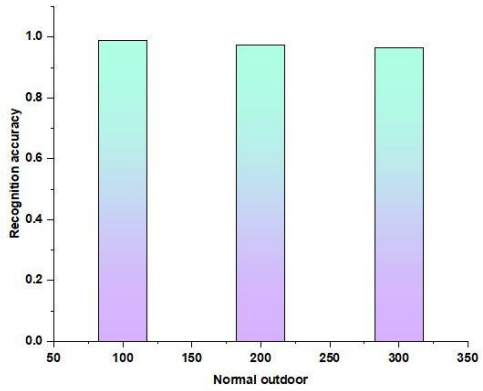

Figure 13. Recognition accuracy with normal outdoor

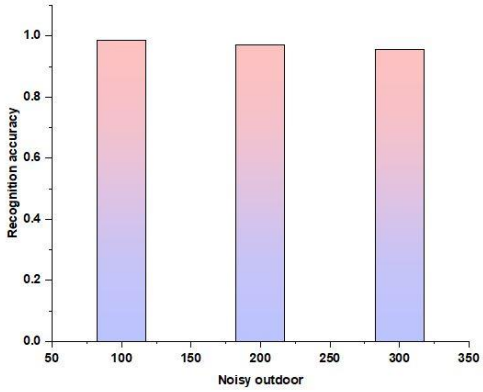

Figure 14. Recognition accuracy with noisy outdoor

In addition, to further verify the efficiency of speech quality evaluation, the real-time performance of recognition and assessment of the system is tested, and the results are shown in Table 2 below. Response time and mean time are necessary to analyze and effectively predict the computation time of the suggested method. As seen from table 2, the learning system can complete the recognition and evaluation of data in a short time after capturing voice data input so that it can serve most utilization scenarios. Figure 15 shows the response time (s) of the vocabulary range.

Table 2: Test results of system recognition rate and real-time evaluation.

Group	Vocabulary range/quantity	Response time (s)	Mean time
1	897	0.347	0.307
2	798	0.259	
3	856	0.323	

4	791	0.251	
5	912	0.355	

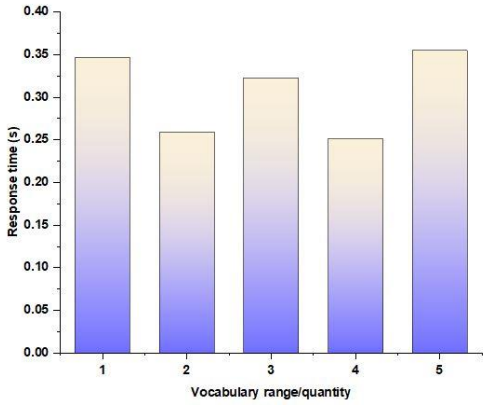

Figure 15. Vocabulary range with response time (s)

Conclusion

Integrating AI speech data recognition and speech quality assessment algorithms enhances the dynamics of speech assessment and releases teachers' teaching pressure and students' learning flexibility, initiative, and motivation. Integrating AI speech data recognition with speech quality assessment algorithms enhances the dynamics of speech assessment, releasing teachers' teaching pressure and students' flexibility, initiative, and motivation to learn. In addition, the integration of AI speech data recognition with English speaking learning systems can significantly enhance the convenience of the learning process and enable the testing of the effectiveness of the whole system. In this article, we analyze the theoretical basis of AI speech data recognition and study the feature extraction of spoken English speech data recognition. The algorithm operating principles and procedures systems are examined by analyzing the speech data recognition method of the English conveying learning system incorporating AI. Finally, each functional module of the spoken English learning system for speech data recognition and quality assessment is designed, and the performance of the whole system is analyzed and verified. In conclusion, this paper provides a diversified learning platform and channel for English speaking learners by integrating AI speech data recognition and speech quality assessment algorithms in English speaking learning. However, the research content of this paper still needs further research and improvement in the improvement of speech model, optimization of speech recognition algorithm, and real-time spoken English speech recognition function.

References

[1] PAGLIERANI P, PETRI D. Uncertainty evaluation of objective speech quality measurement in VoIP Systems [J]. IEEE Transient instrumentation and Measurement, 2019, 58 (3):46-51.

[2] Yu.K.G, Ze.G.S. Improving AMDF for Pitch Period Detection [C]. The Ninth International Conference on Electronic Measurement & Instruments, 2019: 283-286.

[3] Zechner K Bejar I. Towards automatic scoring of non-native spontaneous speech[C], In Proceedings of the Human Language Tech Conference of the North American Chapter of the ACL,2016,: 146-150.

[4] Heiga, Z. Takashi, N, Yamagishi. The HMM-based speech synthesis system (HTS) version 2.0 [C] .Proc of ISCA Bonn Germany. Germany. 2017: 22-24.

[5] YU K, Mairesse F,Young S. Word-level emphasis modeling in HMM-based speech synthesis[C] . ICASSP 2010. NJ: IEEE Press, 2010 : 4238-4241.

[6] Olivier C, Vladimir V, Olivier B, Sayan M. Choosing Multiple Parameters for Support Vector Machines [J]. Machine Learning, 2012, 46 (1-3)..

[7] Jeih-weih Hung, A. Acero, H, Hon, Spoken Language Processing: A Guide to Theory, Algorithm, and System Development, Prentice Hall, 2011:356-408.

[8] Mezura-Montes E, Coello C A. A simple multi-membered evolution strategy to solve constrained optimization problems [J]. IEEE Transactions on Evolutionary Computation, 2005, 9(1):1-17.

[9] Witt,Silke M., and Steve J. Young. Phone-level pronunciation scoring and assessment for interactive language learning [J]. Speech communication, 2010, 30(2) :95-108..

[10] Li J Y, Jiang X L. A new spectral clustering method for large-scale complex image segmentation [J]. Computer utilization research, 2011, 28 (5): 1994-1997.

[11] Bengio,Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives[J].IEEE transactions on pattern analysis and machine intelligence,2013,35(8),1798-1828.

[12] Chen, T. Q., Li, X., Grosse, R. B. et al. Isolating sources of disentanglement in variational auto encoders [C]. Samy Bengio,Hanna M. Wallach, Hugo Larochelle, editor. Advances in Neural Info Processing Systems (Neur IPS), Montréal CANADA.2018, pp.2615-2625.

[13] Ribeiro, Marco Tulio, Singh, Sameer, et al. “why should I trust you?” Explaining the predictions of any classifier [EB/OL]. In Knowledge Discovery and Data Mining (KDD), 2016. https://arxiv.org/abs/1602.04938.

[14] Fukuda N. Orthogonalized distinctive phonetic feature extraction for noise robust automatic speech recognition [C]. IEICE Transactions on Info and Systems. 2014. 1110-1118.

[15] Yao K S, Paliwal K K, Nakamura S. Noise adaptive speech recognition on account of sequential noise parameter estimation [J]. Speech Communication, 2014. 42(1) : 5-23.

[16] Liang W Q, Wang G L, Liu J et al. Phoneme based pronunciation quality evaluation algorithm[J]. Journal of Tsinghua University (Natural Science Edition), 2015,45(1):5-8.

[17] Qi J Y, Zhao H, He S. A study of HMM-based method to evaluate the accuracy of Mandarin single character pronunciation. HMM-based method for evaluating the accuracy of Mandarin pronunciation [J]. Computer Engineering and Applications, 2017, 43(7):224-229..

[18] Xu M, Huang C W, Yang L. A comprehensive evaluation method of multi-level phoneme template for Mandarin pronunciation training[J]. A comprehensive evaluation method of multi-level phoneme template for Mandarin pronunciation training [J]. Computational Engineering and Applications, 2017, 43(28):237-239. .

[19] Zhao Zongbiao, Li Wenxin, Gao R. Neural network-based vector quantization algorithm in speech recognition system in speech recognition system[J]. Henan Science, 2018, (07). : 15-23.

[20] Zen Heiga,Takashi,Nose,Yamagishi. The HMM-based speech synthesis system (HTS) version 2.0[C] .Proc of ISCABonn Germany. Germany. 2017 : 22-24.

[21] Badino L, Andersson J S, Yamagishi J, et al. Identificationof contrast and its emphatic realization in HMM-based speechsynthesis [C] .INTERSPEECH 2009. Grenoble: ISCA, 2019 :520-523.

[22] Gnanamanickam, J., Natarajan, Y. and KR, S.P., 2021. A hybrid speech enhancement algorithm for voice assistance application. Sensors, 21(21), p.7025.

[23] Yang, C.H.H., Qi, J., Chen, S.Y.C., Chen, P.Y., Siniscalchi, S.M., Ma, X. and Lee, C.H., 2021, June. Decentralizing feature extraction with quantum convolutional neural network for automatic speech recognition. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6523-6527). IEEE.

[24] Ma, J., Zhang, Y., Li, Y., Zhou, L., Qin, L., Zeng, Y., Wang, P. and Lei, Y., 2021. Deep dual-side learning ensemble model for Parkinson speech recognition. Biomedical Signal Processing and Control, 69, p.102849.

[25] Li, M., Zorilă, C. and Doddipatla, R., 2021, January. Transformer-based online speech recognition with decoder-end adaptive computation steps. In 2021 IEEE spoken language technology workshop (SLT) (pp. 1-7). IEEE.

[26] Kumar, L.A., Renuka, D.K., Rose, S.L. and Wartana, I.M., 2022. Deep learning based assistive technology on audio visual speech recognition for hearing impaired. International Journal of Cognitive Computing in Engineering, 3, pp.24-30.