ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

An Intelligent Spoken English Evaluation System Based on a Corpus-driven SHO-SVM Model and Articulatory Feature Extraction

Cheng Huang[1,*]

[1] Foreign Language Department, Hainan Vocational University of Science and Technology, Haikou City, Hainan Province, China, 571100

SUMMARY: *This paper proposes a corpus-driven spoken English evaluation system, which combines pronunciation feature extraction, SHO search and SVM rank determination to form a computational link. The database is built based on 2400 spoken language corpus of 300 learners, and the expert ratings are coded into four levels of A, B, C, and D. The system extracted MFCC, fundamental frequency, formant, energy, speech rate, pause ratio, stress offset and phoneme duration, and formed a 126-dimensional feature vector after standardization. Spotted hyenas optimizer selected subsets according to classification fitness, feature scale and redundancy, so that the feature dimension was reduced to 43. SVM classifier completed the grade label and score interval determination under optimized parameters. The 7:2:1 partition results show that SHO-SVM achieves 94.6% Accuracy, 93.8% Precision, 92.9% Recall and 93.3% F1-score, which are better than SVM, RF and BP network. The results show that the system has stable automatic pronunciation scoring ability and repeatable oral assessment ability, and can be applied to different scoring levels.*

KEYWORDS: *Corpus driven; Pronunciation feature extraction; SHO-SVM; Evaluation of Intelligent Spoken Language*

1 Introduction

The core of intelligent spoken English assessment is not just to give a score, but to turn the pronunciation stability, rhythm control, stress distribution and phoneme bias in continuous speech into computable feature evidence. With the application of speech recognition, acoustic modeling and machine learning classification methods in language assessment scenarios, spoken language scoring has gradually shifted from manual subjective judgment to data-driven automatic evaluation. The computer system can extract multi-dimensional acoustic parameters from large-scale corpus, and put the scoring labels, pronunciation features and classification boundaries into a unified modeling link, so that the spoken language assessment has a repeatable, comparable and scalable technical foundation.

Al-Ghezi et al. constructed an automatic scoring framework for spontaneous second language spoken language, and used speech samples and scoring annotations to establish the model input relationship, indicating that the corpora driven method can support the recognition of spoken ability under the condition of cross-speaker [1]. Camara Arenas et al. compared the condition differences between automatic speech assessment and automatic speech recognition in L2 English tasks, and pointed out that recognition text accuracy is not equivalent to pronunciation quality assessment, and acoustic bias, phoneme substitution and

*hctotoo@outlook.com
https://doi.org/10.65102/is2026909 Published: 30 April, 2026

prosodic performance still need to be modeled independently [2]. Vidal et al. designed an automatic pronunciation evaluation system for Argentinian English learners, which calculated and compared the learner's speech with the target pronunciation, and provided a transferable engineering path for the English pronunciation scoring system [3]. Bashori et al. studied a language learning system based on automatic speech recognition and verified that speech input, feedback generation and scoring output can form a continuous human-computer interaction process [4]. Bashori et al. further analyzed the supporting effect of websites with automatic speech recognition technology on vocabulary and pronunciation training, indicating that online speech systems need to deal with word recognition, pronunciation deviation recording and user performance tracking at the same time [5].

Existing research has provided data basis and system experience for automatic spoken language assessment, but the computational model for English pronunciation quality still needs more fine-grained feature representation. Saito et al. conducted training, validation and generalization studies on the automatic assessment of L2 intelligibility, emphasizing the robustness judgment of the model under different sample distributions [6]. Kang et al. proposed an AI language tutoring system combining end-to-end automatic speech recognition and proficiency evaluation, which connected the recognition results with the output of ability level, reflecting the complete chain from speech input to level determination of intelligent evaluation system [7]. Lounis et al. reviewed the application of deep neural networks in mispronunciation detection and diagnosis, and pointed out that mispronunciation recognition should pay attention to acoustic patterns, phoneme boundaries and classification decisions at the same time [8]. Lounis et al. also used variational autoencoder for Arabic mispronunciation anomaly detection, which provided reference for unsupervised or weakly supervised pronunciation deviation identification [9]. Bahi et al. sorted out the automatic pronunciation evaluation and feedback generation method for Arabic learners, indicating that the pronunciation evaluation system not only needs to output the grade, but also needs to retain the interpretable feature basis [10].

Based on the above research vein, this paper constructs an intelligent spoken English evaluation system based on a corpus-driven SHO-SVM model and pronunciation feature extraction. Based on 2400 spoken English utterances from 300 learners, the system collects expert ratings, phoneme annotations and grade labels to form A training corpus containing four level results of A, B, C and D. The speech feature extraction module obtains parameters such as MFCC, fundamental frequency, formant, energy, speech rate, pause ratio, stress offset and phoneme duration from the speech signal. After framing, endpoint detection, noise suppression and standardization, the numerical vector suitable for machine learning model reading is generated. The SHO search module performs feature subset screening and SVM parameter optimization for high-dimensional pronunciation features to reduce the interference of redundant acoustic quantities on scoring boundaries. The SVM classifier learns the margin boundaries of different spoken language levels in the optimized feature space, and outputs rating prediction, rating decision and confidence state.

The focus of this paper is on the continuous computation link of corpus encoding, articulation feature modeling, SHO search optimization, and SVM score classification. Unlike systems that rely solely on automatic speech recognition text transcription, we map spoken language evaluation objectives directly between acoustic parameters and classification decisions, preserving discriminative evidence of speech quality itself. In the engineering implementation, the front end of the system is responsible for voice acquisition and format verification, the feature end completes batch calculation and caching, and the model end saves the optimal feature index, kernel function parameters and classification results. This structure can support offline training and online invocation, and provide a unified interface for

subsequent performance re-testing under different corpus sizes. The experimental part can be expanded around accuracy, precision, recall and F1 value, and combined with confusion matrix to observe the rank error distribution. This path is more in line with the writing requirements of technical journals for data, algorithms, parameters, and system verification. The remaining part of this paper is arranged as follows: Section II introduces related work; Section 3 gives the corpus construction, pronunciation feature extraction, SHO feature screening and SHO-SVM scoring model. Section IV explains the experimental setup and results. Section V summarizes the system performance and subsequent scalability directions.

2 Related work

The related research of intelligent spoken English assessment mainly focuses on speech signal representation, pronunciation deviation detection, automatic speech recognition aided scoring, and machine learning classification modeling. Compared with the traditional manual scoring, the computer method can segment continuous speech into computable segments, and record the phoneme duration, fundamental frequency trend, energy distribution, pause ratio, formant position and prosody changes from the acoustic level. The quality of spoken English is not exactly equivalent to the accuracy of the transcribed text. The scoring system also needs to identify articulation articulation, rhythm stability and stress matching. The resulting research path usually includes several computational steps such as corpus collection, signal preprocessing, feature extraction, feature selection, model training and rank output.

Calık et al. proposed an ensemble framework for Arabic phoneme mispronunciation detection, which fuses the results of multiple classifiers to enhance the stability of phoneme level error recognition [11]. This study illustrates that pronunciation evaluation cannot rely only on a single classification boundary, and the complementary results of multiple models on different phoneme classes can improve the reliability of mispronunciation judgments. Calık et al. further studied an audio-oriented Transformer mispronunciation detection framework, which maps speech segments into deep temporal representations and uses attention structures to capture local pronunciation deviations [12]. This method has strong modeling ability for high-dimensional acoustic sequences, but the model training usually requires a large labeled corpus and high computational cost.

Ahmed et al. constructed an Arabic mispronunciation recognition system based on LSTM network, and used the recurrent structure to deal with the temporal dependence in the speech sequence [13]. This study shows that pronunciation errors tend to be distributed between consecutive phonemes and adjacent acoustic states, and temporal modeling is able to retain more contextual information. Algabri et al. proposed a mispronunciation detection and diagnosis method for non-native Arabic speech, and introduced a feedback generation mechanism at the level of articulators [14]. This study connects acoustic bias with articulatory action interpretation, which provides a reference for interpretable output of articulatory assessment results.

Table 1: Computational methods of related speech assessment studies and relationship to this paper

Reference	Research Method	Technical Feature	Relationship to This Paper
[11]	Ensemble classification framework	Integrates multi-model results	Supports pronunciation deviation classification
[12]	Audio Transformer	Captures deep temporal features	Supports high-dimensional acoustic modeling
[13]	LSTM network	Processes continuous speech dependencies	Supports sequential feature representation
[14]	Articulatory-level feedback	Connects acoustic deviation with explanation	Supports evaluation interpretability
[15]	ASR-based adaptive learning	Builds speech input and model feedback	Supports systematic spoken English evaluation
[16]	Prosody modeling	Uses rhythm and pause features to predict performance	Supports prosodic feature extraction
[17]	ASR review	Summarizes the development of recognition models	Supports the speech recognition background
[18]	End-to-end recognition review	Analyzes deep recognition structures	Supports speech front-end modeling
[19]	Speaker verification review	Emphasizes voiceprint and individual differences	Supports cross-speaker robustness
[20]	Feature extraction and classification review	Summarizes acoustic features and classifiers	Supports SHO-SVM method design

Wilschut et al. studied an adaptive vocabulary learning model based on automatic speech recognition, which connected speech input, model feedback and learning state update as a continuous process [15]. Although this study focuses on vocabulary learning as the application scenario, its speech acquisition, recognition output and model update mechanism have reference value for intelligent spoken language evaluation system. Sabu and Rao used prosodic modeling to predict children's reading proficiency, focusing on analyzing the role of pause, rhythm and intonation and other features on speech performance judgment [16]. This study shows that prosodic information is able to complement phoneme level features and is particularly suitable for distinguishing between articulatory fluency and speech naturalness.

O'Shaughnessy summarized the development trend of automatic speech recognition research, pointing out that acoustic models, language models, and end-to-end architectures are driving speech systems from feature engineering to deep representation learning [17]. Prabhavalkar et al. systematically summarized end-to-end speech recognition methods and analyzed the application of connection timing classification, attention model, and Transducer structure in speech recognition [18]. These studies provide the technical foundation in terms of recognition front-end and acoustic coding for spoken English assessment systems.

O'Shaughnessy has also studied automatic speaker verification methods, and summarized the computational processing of speaker differences around voiceprint embedding, similarity calculation, and discriminative models [19]. For spoken language assessment, speaker differences can affect the fundamental frequency range, speaking rate habits and energy distribution. Therefore, the model needs to reduce the score offset caused by individual

acoustic differences in feature standardization and classification training. Yadav et al. reviewed feature extraction and classification technology in speech recognition, and summarized MFCC, LPC, PLP, spectral features, SVM, KNN, neural network and other classification methods [20]. This research has a direct connection with the method design of this paper, which shows that articulational feature extraction and classifier selection are the basic links of speech evaluation system.

From the perspective of system implementation, relevant results also suggest that the spoken language evaluation model needs to maintain the consistency of input and output. The score labels in the corpus should correspond to the acoustic feature window, and the whole sentence score should not be retained while ignoring the segment-level pronunciation differences. The feature selection process should record the name and index of the selected feature, which is convenient for subsequent retest. In addition to the rank, the classification output should also save the confidence and error distribution to provide a basis for manual review and model iteration. Therefore, the related work of this paper is no longer a simple reference of speech recognition technology, but converts the existing research into executable modeling basis: corpus is used to form training samples, acoustic features are used to describe the quality of speech, search optimization is used to compress the feature space, SVM classifier is used to establish the level boundary, and system interface is used to output and record the results. This approach enables related work to form a continuous relationship with the proposed method, avoids the separation of speech processing, feature selection and scoring models into disjoint descriptions, and enhances the integrity of model reproduction experiments.

Combined with the above studies, it can be seen that the existing methods have formed a clear accumulation of technologies in depth sequence modeling, mispronunciation detection, automatic speech recognition and acoustic feature classification. On this basis, the corpus driven method is used to construct the spoken English scoring samples, and the pronunciation parameters such as MFCC, fundamental frequency, formant, energy, speaking rate, pause ratio and stress shift are uniformly encoded. Then the SHO algorithm is used to perform high-dimensional feature search and SVM parameter optimization, so that the scoring model can form a stable classification boundary on a small feature subset. This path emphasizes more on the linkage between interpretable acoustic features, feature subset selection and machine learning classification decision, which meets the requirements of intelligent spoken English assessment system for accuracy, reproducibility and engineering deployment.

3 Methods

3.1 English Corpus Construction and scoring label Coding for Intelligent Spoken Language Assessment

The input basis of the intelligent Oral English assessment system is the spoken English corpus that can be labeled, calculated and tracked. The corpus of this paper collects 2400 spoken and semi-open answer speech from 300 learners, the sampling rate is set to 16kHz, and it is saved as mono uncompressed format. The speaker number, question type number, text prompt, recording duration, expert rating and quality status were recorded synchronously for each speech, which was convenient for subsequent pronunciation feature extraction and SHO-SVM classification training. In the corpus construction stage, the original score is not directly used as the only basis, but the four scoring items of pronunciation accuracy, fluency, prosodic stability and integrity are encoded into a unified label to form the training samples for machine learning.

In order to uniformly describe the mapping relationship between spoken language samples and rating labels and support subsequent feature calculation and label tracking calls, the encoding of corpus samples is defined as follows:

$$D = \{x_i = (s_i, u_i, q_i, r_i, y_i) \mid i = 1,2, \dots, N\} \tag{1}$$

where D represents the set of spoken English corpus; x_i represents the i spoken sample; s_i represents the speaker number. u_i represents the path and duration information of the voice file. q_i represents the question type and prompt text. r_i represents a multidimensional rating vector. y_i denotes the final rank label; N denotes the total number of samples. The speech, title, rating and label were put into the same data unit, which could avoid the loss of sample sources after feature extraction, so that the sample index and corresponding level could be directly called in the model training stage.

Score labels were given independently by three English speech assessors on a scale of 0 to 100. The system divides the score into four dimensions: pronunciation accuracy, stress rhythm, pause control and speech integrity, and then performs numerical normalization to reduce the influence of the difference in scale given by different raters on label boundaries. The processed scores are entered into the grade mapping module to generate classification labels according to the four levels A, B, C, and D, and the continuous scores are retained for subsequent error analysis.

In order to weaken the label shift caused by the differences in scoring dimensions and maintain the cross-item scores in a stable and comparable numerical interval, the scoring normalization process is shown in the following equation:

$$\hat{r}_{i,k} = \frac{r_{i,k} - \min(R_k)}{\max(R_k) - \min(R_k) + \varepsilon} \tag{2}$$

where $r_{i,k}$ represents the original score of the i sample on the k scoring dimension. R_k represents the set of all ratings for that dimension; $\hat{r}_{i,k}$ represents the normalized score value; Let ε denote the minimal constant that prevents the denominator from being zero. The calculation makes the four types of rating dimensions in the same numerical interval, which facilitates the subsequent formation of a unified label and provides a stable input for the nonlinear mapping between pronunciation features and rating levels.

In order to test the consistency of labeling among multiple raters and reduce the influence of subjective score deviation on the formation of grade boundaries, the label reliability coefficient is calculated as follows:

$$\rho_i = 1 - \frac{\sum_{a=1}^{M} \sum_{b=a+1}^{M} | r_i^{(a)} - r_i^{(b)} |}{\left(\frac{M}{2}\right) \cdot R_{span}} \tag{3}$$

Here, ρ_i denotes the scoring consistency of the i sample. M denotes the number of raters; $r_i^{(a)}$ and $r_i^{(b)}$ represent the combined score given by the two raters, respectively; R_{span} represents the rating interval width. The closer this coefficient is to 1, the more stable the sample label is. If the consistency of samples is lower than the set threshold, the system will put them into the review queue instead of entering the first round of model training, so as to reduce the interference of noise labels on SVM classification boundaries.

Table 2 presents the rank distribution of the corpus and the statistics of the underlying recordings. This table is used to show whether the training samples are basically balanced,

and also provides data basis for the division of training set, validation set and test set in subsequent experiments. The four-level labels A, B, C, and D correspond to four categories of oral performance: excellent, good, qualified, and weak, respectively. The system synchronously records the average duration and average rating when saving the grade labels, so as to facilitate the observation of acoustic differences between different grades.

Table 2: Rank labels and sample statistics for spoken English corpus

Grade Label	Number of Samples	Number of Speakers	Average Duration/s	Mean Overall Score
A	486	92	18.7	88.4
B	728	146	20.1	76.9
C	814	171	22.6	64.2
D	372	89	25.4	51.7

Mute segment location, abnormal audio elimination and text prompt matching are needed before corpus is put into the database. Each utterance is segmented into sentence-level segments and phoneme level time Windows, and the segmentation results are aligned with the phoneme sequence of the cue text. This process can extend the whole sentence score to the segment position, which provides the location basis for the subsequent extraction of pause ratio, stress shift, phoneme duration and energy change.

In order to map the sentence-level speech segmentation results to the phoneme time axis, and keep the pause, stress boundary and segment position information synchronously, the segment alignment vector is constructed as follows:

$$A_i = [(p_{i,j}, t^s_{i,j}, t^e_{i,j}, d_{i,j}, b_{i,j})]_{j=1}^{L_i} \tag{4}$$

where A_i represents the segment-aligned sequence of the i speech; $p_{i,j}$ denotes the j phoneme or syllable unit; $t^s_{i,j}$ and $t^e_{i,j}$ denote the start and end times; $d_{i,j}$ denote the duration; $b_{i,j}$ denote pause, accent, or boundary states; L_i denotes the number of segments. This vector binds the speech timeline to the text structure, provides an explicit computational location for the feature extraction module, and avoids global statistics to mask local pronunciation bias.

Since the number of samples in different grades is not completely consistent, the sample weights need to be calculated after label coding. The weights do not change the original labels, but only adjust the contribution of samples of different grades during model training, preventing A large number of C or B level samples from compressing the classification space of both levels of A and D. This processing enables SHO-SVM to consider both boundary samples when searching feature subset and learning classification margin.

In order to reduce the influence of the difference in the number of grade samples on model training, and keep the classification boundary stable and the sample contribution relatively balanced, the sample weight is defined as follows:

$$\omega_i = \frac{N}{K \cdot n_{y_i}} \cdot (1 + \lambda(1 - \rho_i)) \tag{5}$$

Here, ω_i denotes the training weight of the i sample; K represents the number of grade categories; n_{y_i} denotes the number of classes the sample belongs to; Let λ denote the consistency adjustment coefficient. Let ρ_i denote the scoring consistency. The formula considers both the number of categories and label reliability, so that the minority level

samples can obtain reasonable training contributions, and the low consistency samples will not be simply amplified. After the above processing, the corpus forms five kinds of data fields: speech file, score vector, rank label, alignment sequence and sample weight, which can be directly entered into the pronunciation acoustic feature extraction and standardization module.

After the above processing, the corpus is no longer just a collection of raw audio files, but forms a structured data unit containing speech files, question type information, score vectors, rank labels, alignment sequences, sample weights and review status. Each sample can be traced in the system to the recording source, scoring basis and label generation process, which provides a stable data entry for subsequent acoustic feature extraction. This data structure also facilitates the SHO algorithm to call the sample weight and label reliability information in the feature search phase, so that the feature selection is not separated from the scoring task itself. The SVM classifier can directly read the rank labels, the optimized feature matrix and the sample weights during training, thus forming a continuous calculation link from corpus collection to score output. Therefore, the construction of spoken English corpus and the coding of scoring labels not only assume the role of data preparation, but also lay a reproducible data foundation for subsequent acoustic parameter modeling, feature standardization processing and SHO-SVM grade determination.

3.2 Acoustic parameter modeling and standardization processing based on articulatory feature extraction

Spoken English assessment systems need to convert continuous speech signals into trainable numerical features that enable articulation accuracy, prosodic variation, pause control, and speech fluency to enter machine learning models. The original speech waveform itself cannot be directly used as the stable input of SHO-SVM classifier, so it needs endpoint detection, framing and windowing, spectrum conversion, acoustic parameter extraction and standardization to form a feature matrix with clear structure and consistent dimension. This paper focuses on the task of oral English scoring to extract parameters such as MFCC, fundamental frequency, formant, short-time energy, zero-cross rate, speaking rate, pause ratio, stress offset and phoneme duration, and encodes these parameters into sampled-level utterance feature vectors. This processing can preserve the timbre, rhythm and duration information in speech, and also provide computable input for subsequent SHO feature search and SVM rank determination.

Fig. 1 shows the complete calculation process of spoken English speech from the original file into the pronunciation feature matrix. The left input end includes the learner's voice, the topic text, the scoring label and the speaker number. The system first checks the format of the voice file, unifies the sampling rate and marks the silent segment, and then enters the endpoint detection and framing and windowing module. The middle computing layer extracts parameters such as MFCC, fundamental frequency, formant, short-time energy, zero-crossing rate and phoneme duration according to frame-level processing logic, and combines the phoneme sequence in the title text to complete the alignment of pronunciation segments. The right output aggregates the frame-level parameters into sentence-level statistical features, which are written into the feature matrix together with the score labels and sample weights, so that the subsequent SHO search module can directly read the trainable data structure.

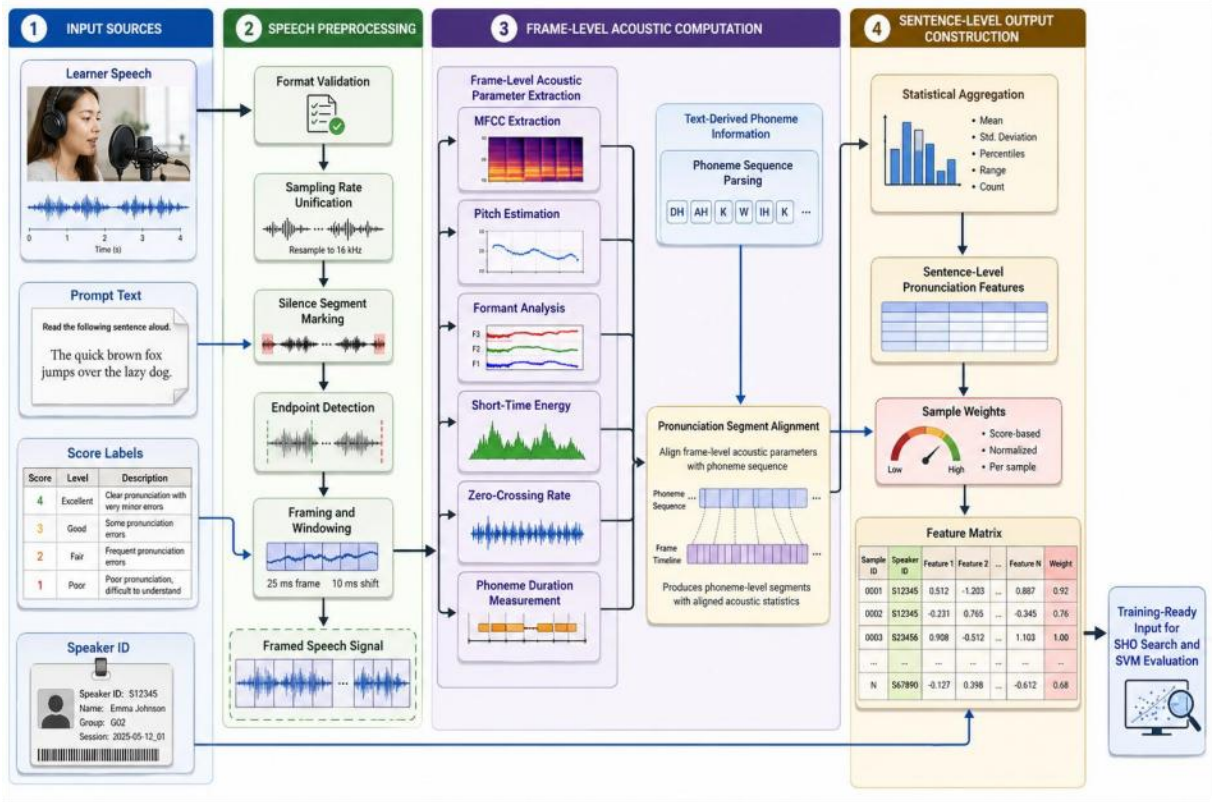

Figure 1: Computational flow of articulatory acoustic feature extraction

In order to convert the original speech into a computable frame sequence within a fixed window, the system pre-emphasizes and Windows the sampled waveform and saves the frame number, as shown in the following equation:

$$z_{i,m}(n) = [u_i(n + mH) - \alpha u_i(n + mH - 1)] \cdot w(n) \tag{6}$$

where $z_{i,m}(n)$ represents the windowed signal of the m frame of the i speech; $u_i(n)$ represents the original sampling sequence. H represents frame shift; Let α denote the pre-weighting coefficient; w(n) is the window function. This formulation converts continuous speech into a locally stable frame-level signal, which enables articulational fluctuations, energy variations, and local phoneme boundaries to enter subsequent spectrum calculations.

In order to obtain the cepstrum parameters that can reflect the differences between timbre and phoneme, the system calculates the Mel-band energy for each frame of speech and compresses the spectral shape for classification, as shown in the following equation:

$$c_{i,m,l} = \sum_{p=1}^{P} \log\left(\sum_{\omega=1}^{\Omega} |F_{i,m}(\omega)|^2 B_p(\omega)\right) \cos\frac{\pi l(p - 0.5)}{P} \tag{7}$$

where $c_{i,m,l}$ denote the l dimension MFCC coefficient; $F_{i,m}(\omega)$ denotes the spectrum of the m frame; Let $B_p(\omega)$ denote the p Mel-filter; P is the number of filters; Ω represents the number of frequency sampling points. The formula compresses the resonance structure in pronunciation into a low-dimensional cepstral representation, which is able to characterize vowel opening, consonant articulation and phoneme transition morphology.

To preserve accuracy, prosody, and fluency information simultaneously, the system concatenates multiple acoustic parameters into a unified vector representation for the model to train, as shown in the following equation:

$$v_i = [\mu(C_i), \sigma(C_i), \mu(F0_i), \sigma(F0_i), E_i, Z_i, R_i, P_i, S_i, T_i] \quad (8)$$

where v_i represents the integrated acoustic vector of the i speech. C_i stands for MFCC matrix; $F0_i$ represents the fundamental frequency sequence; E_i represents short-term energy statistics; Z_i is the zero-crossing rate; R_i is the speaking rate; P_i is the proportion of pauses. S_i stands for stress offset; T_i denotes the phoneme duration. This vector unifies the frequency spectrum, prosody and time structure into the same feature space, which can support SVM to establish classification boundaries for different spoken language levels.

Fig. 2 illustrates the transformation of acoustic parameters from original feature vectors to normalized training matrices. The left side shows the extracted multi-dimensional acoustic parameters, including spectral features, prosodic features and temporal structure features. In the middle part, the mean value, standard deviation and outlier range are calculated according to the training set samples, and then the same statistical aperture is applied to the training set, validation set and test set to ensure that there is no information leakage between different data partitions. The right output holds the normalized feature matrix, feature indices, training set statistics, and rank labels. This structure can ensure that SHO algorithm can select features under the same numerical scale, and also make SVM classifier not be interfered by dimensional differences when calculating kernel function distance.

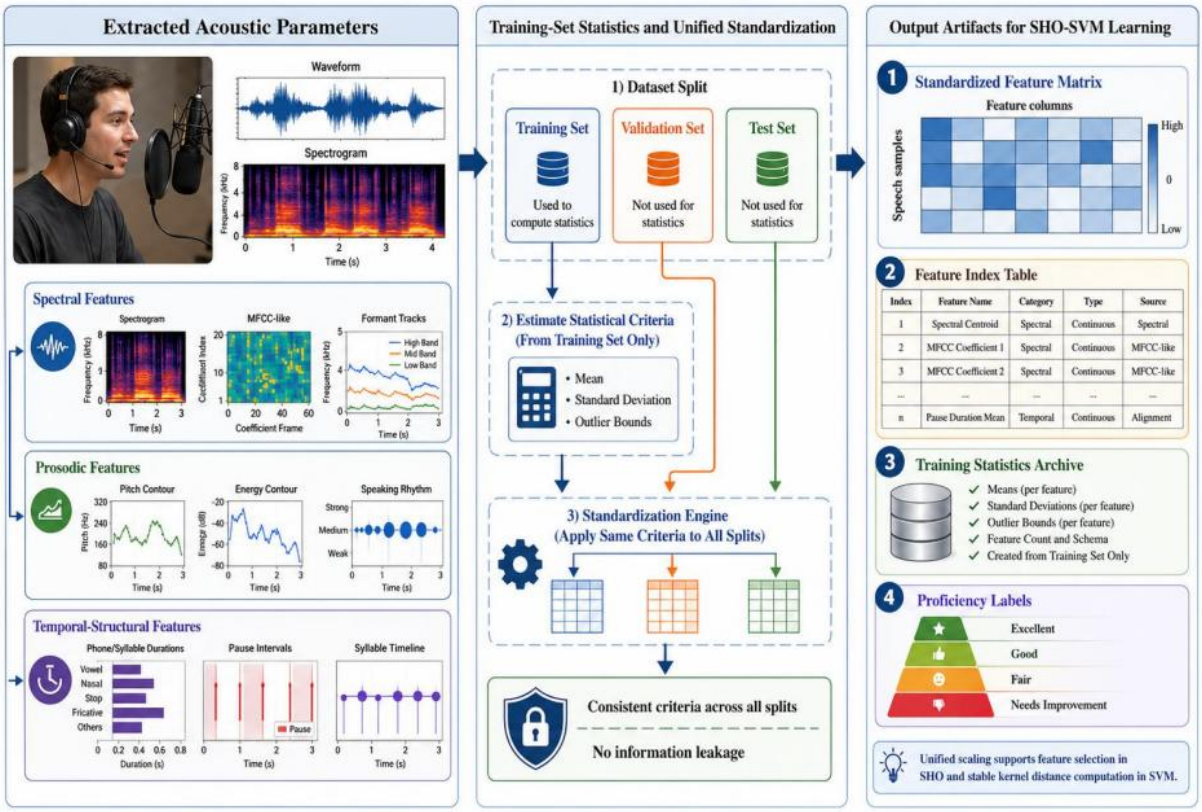

Figure 2: Acoustic parameter normalization and feature matrix generation structure

In order to depict the dynamic change amplitude between adjacent speech frames, the system calculates the first-order difference feature of acoustic parameters and retains the

change direction for training, as shown in the following equation:

$$\Delta g_{i,m} = \frac{\sum_{\tau=1}^{Q} \tau \left(g_{i,m+\tau} - g_{i,m-\tau} \right)}{2 \sum_{\tau=1}^{Q} \tau^2} \tag{9}$$

Here, $\Delta g_{i,m}$ denotes the first-order difference of acoustic parameters in the m frame. $g_{i,m}$ denote the original frame-level parameters; Q denotes the difference window radius. This formula can describe the changing speed of energy, fundamental frequency and cepstrum parameters during pronunciation, so that the model can not only read the static pronunciation state, but also capture the dynamic features in phoneme transition and intonation transition.

In order to eliminate the influence of different parameter scale differences on SHO search and SVM training, the system performs the training set statistical normalization writing matrix, as shown in the following equation:

$$\tilde{v}_{i,h} = \frac{v_{i,h} - \mu_h^{train}}{\sigma_h^{train} + \delta} \tag{10}$$

where $\tilde{v}_{i,h}$ represents the h feature after normalization; $v_{i,h}$ denote the original acoustic parameters; μ_h^{train} and σ_h^{train} are the mean and standard deviation of the h feature in the training set, respectively. Let δ denote the smoothness constant. This processing ensures that MFCC, fundamental frequency, energy, speech rate and pause ratio will not change the model weight distribution due to the difference of numerical range, so that SHO feature search and SVM kernel function calculation remain stable.

After the above processing, the system obtains a unified articulatory feature matrix, rank labels and sample weights. Each row corresponds to a sample of spoken English, and each column corresponds to an interpretable acoustic parameter. This matrix not only preserves the information of pronunciation accuracy, prosodic rhythm and fluency, but also provides an optional high-dimensional feature space for subsequent SHO search optimization. Compared with the direct input of original speech, this processing method is more suitable for reproducible experimental design in technical journal papers, and can clearly explain the source of model input, the method of parameter calculation and the standardization caliber. The normalized feature matrix will enter the SHO search optimization module in Section 3.3 and be used as the base input for SVM score classification.

3.3 SHO Search Optimization and feature subset Selection for high-dimensional pronunciation features

High-dimensional features include MFCC, fundamental frequency, formant, energy, speech rate, pause ratio, stress offset and phoneme duration. The contribution of different features to spoken language scoring is not consistent, and there is strong correlation between some parameters. If all the features directly enter the SVM training, the distance of the kernel function will be interfered by redundant dimensions, and the classification boundary is easy to be drawn by low-contribution features. In this paper, the SHO algorithm is used to establish the feature search process, and each candidate individual is represented as a feature selection scheme, and the cross-validated classification performance, number of features and redundancy together constitute the fitness evaluation.

In order to transform the continuous feature space into a searchable subset, it is necessary to establish a mapping relationship between the candidate mask and the acoustic dimension and keep the index stable and reliable, as shown in the following equation:

$$m_h = \begin{cases} 1, & \sigma(X_h) > \theta \\ 0, & \sigma(X_h) \le \theta \end{cases}, \quad \sigma(X_h) = \frac{1}{1 + e^{-X_h}} \tag{11}$$

where m_h represents the selection state of the h dimension acoustic feature. X_h denotes the continuous position of the SHO individual in this dimension. Let θ denote the selection threshold. The mask converts the continuous search position into a binary feature switch, so that the parameters such as MFCC, fundamental frequency and pause ratio can enter the candidate subset in a unified way. The mask index is kept consistent during the training, validation and testing phases to ensure that the feature columns read by the subsequent SVM are not offset.

Fig. 3 shows the internal process of SHO feature search and subset filtering. On the left, the inputs are normalized pronunciation feature matrix, rank labels, and sample weights. In the middle part, a cyclic search structure was composed of candidate individual initialization, binary mask transformation, SVM cross validation and fitness calculation. The right side outputs the optimal feature index, kernel function parameter candidates, and the compressed articulation feature matrix. This process binds feature filtering to the goal of rating classification and avoids retaining features only by a single relevance.

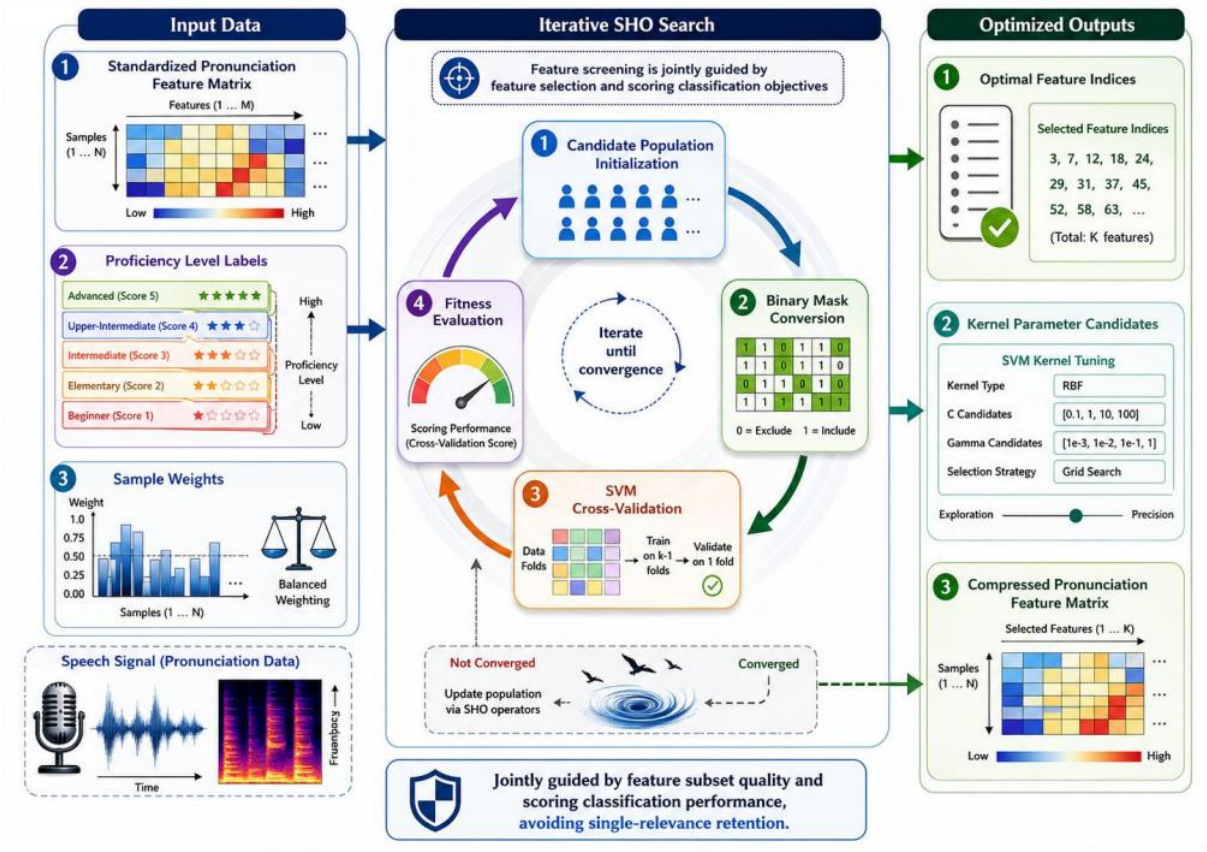

Figure 3: SHO search optimization and articulational feature subset screening process

In order to measure the classification value of the candidate feature subset, the three constraint weights of recognition error, dimension scale and redundancy should be considered simultaneously, as shown in the following equation:

$$J(m) = \eta[1 - Acc_{cv}(m)] + \beta\frac{|m|}{H} + \gamma\frac{2}{|m|(|m|-1)}\sum_{a<b}|\,corr(f_a, f_b)| \qquad (12)$$

where $J(m)$ represents the fitness value of the candidate subset; $Acc_{cv}(m)$ is the classification accuracy of this subset in cross-validation. $|m|$ denotes the number of selected features; H represents the total number of features; $corr(f_a, f_b)$ is the correlation between two features. Let η, β, and γ denote the three-term weights. A smaller fitness indicates that the subset has lower dimensionality and less redundancy while maintaining the classification ability. This design is suitable for the task of oral English scoring, because the pronunciation quality is determined by multiple types of acoustic cues, and simply compressing the number of features will weaken the scoring explanatory power.

In order to simulate the search migration of SHO population in the feature space, the candidate positions are iteratively updated according to the optimal individual and the shrinkage coefficient, as shown in the following equation:

$$X_h^{t+1} = X_{best,h}^t - A_t\left|C_t X_{best,h}^t - X_h^t\right| + \kappa\xi_h^t \qquad (13)$$

Here, X_h^{t+1} denotes the h dimension position in the $t+1$ iteration. $X_{best,h}^t$ denotes the current best individual position;A_t and C_t denote the control coefficients that change with iteration. κ is the magnitude of the disturbance; Let ξ_h^t denote the random disturbance term. This update method takes into account both global search and local convergence, so that the algorithm can find effective combinations from a wide range of acoustic features, and refine around the better subset in the later stage.

In order to apply the searched binary mask to the normalized pronunciation feature matrix, we need to generate a subset of input features that can be read by the SVM, as shown in the following equation:

$$Z_m = Z\Pi_m, \quad \Pi_m = [e_h]_{m_h=1,h=1}^H \qquad (14)$$

Here, Z represents the complete pronunciation feature matrix after normalization; Z_m represents the input matrix after the mask filter; Let Π_m denote the projection matrix consisting of the selected feature columns; e_h denotes the h dimensional unit column vector. After this step, the system obtains the compressed training input and retains the acoustic parameters that contribute to the rank determination. The filtering results not only reduce the computational overhead of SVM, but also make it easier to backtrack the model output to specific articulatory factors, such as too long pauses, insufficient fundamental frequency fluctuations, or phoneme duration shifts.

After the feature subset screening is completed, the system synchronously saves the fitness curve, the selected feature frequency and the final index table for each iteration. This record can support experimental review and facilitate the analysis of the contribution of different pronunciation parameters in the grade determination. Compared with manually setting fixed features, the search process of SHO can automatically adjust the input space according to the corpus distribution, so that the subsequent SVM classifier can maintain a stable interval in a small dimension. Finally, the optimal subset will enter the SHO-SVM scoring classification stage together with sample labels and weights, which provides a direct input basis for the online scoring module to reduce the calculation delay, keep the results available for review, and support the deployment and operation of the system.

3.4 Oral English score prediction and grade determination based on SHO-SVM model

After normalization and SHO subset screening, articulatory features form a database of feature vectors that can be read by supervised classification models. Each sample in the database corresponds to A compressed pronunciation feature vector, while retaining the four-level rating labels of A, B, C and D and the sample weights. In this paper, SVM is used as the spoken language scoring classifier, and the feature subset, penalty factor and kernel function parameters obtained by SHO search are jointly written into the model configuration. This structure makes the score prediction not rely on a single acoustic index, but establishes a stable classification boundary in the multi-dimensional pronunciation feature space.

In order to make the sample weight, class label and classification margin enter the training objective at the same time, the weighted optimization form of SHO-SVM model is as follows:

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \omega_i \xi_i, \quad y_i(w^T\phi(z_i) + b) \geq 1 - \xi_i, \xi_i \geq 0 \tag{15}$$

where z_i represents the i utterance feature vector after SHO filtering. y_i denotes the rank label; Let ω_i denote the sample weights; C is the penalty factor; Let ξ_i, denote the slack variable; Let $\phi(\cdot)$ denote the kernel mapping function. The formula maximizes the margin while incorporating the sample weights, so that the minority grades and high reliable labels can obtain reasonable contributions in training, and reduce the influence of class boundaries by sample distribution skew.

The SHO-SVM system structure consists of input layer, kernel mapping layer, classification judgment layer and output recording layer. The input layer reads the compressed feature matrix, the kernel mapping layer completes the nonlinear space conversion, the classification and judgment layer outputs the grade results, and the record layer saves the confidence, parameters and sample index.To clarify the functional connections between these modules and their corresponding input and output, Table 3 summarizes the classification model structure.

Table 3: SHO-SVM spoken language score classification model structure

Module	Input Content	Computational Task	Output Content
Feature Input Layer	Compressed pronunciation feature vectors	Read samples and weights	Training matrix
Kernel Mapping Layer	Training matrix and kernel parameters	Compute sample similarity	Kernel matrix
Classification Decision Layer	Kernel matrix and grade labels	Learn classification boundaries	Grade results
Output Recording Layer	Prediction results and parameters	Save confidence status	Scoring records

In order to depict the nonlinear similarity relationship between different pronunciation feature samples and adapt to the complex boundary distribution between spoken English grades, the following formula is shown:

$$K(z_i, z_j) = \exp\left(-\gamma\|z_i - z_j\|_2^2\right) \tag{16}$$

where $K(z_i, z_j)$ represents the kernel function value between two articulatory feature vectors; Let γ denote the kernel width parameter; $\|z_i - z_j\|_2^2$ denotes the square Euclidean distance. The kernel function can map the original acoustic features into a high-dimensional space, so that the nonlinear relationship among pronunciation accuracy, pause ratio and stress offset can be expressed. Compared with linear division, this processing is more suitable for scoring scenarios where the spoken language levels have cross boundaries.

The SHO algorithm not only screens the feature subset, but also participates in the SVM parameter search. Each candidate individual contains the feature mask, penalty factor, and kernel width, and the system updates the optimal configuration based on the validation set results. The search objective is not only to pursue a single accuracy, but also to consider macro-average F1 value, classification error and model complexity.

In order to uniformly measure the classification performance, model size and rank balance performance of parameter combinations on the validation set, the model configuration objective function is as follows:

$$\Theta^* = \arg\min_{\Theta} \left[1 - F1_{macro}(\Theta) + \lambda_1 \frac{|m|}{H} + \lambda_2 Err_{grade}(\Theta)\right] \tag{17}$$

Here, Θ denotes the candidate configuration consisting of the feature mask, C, and γ. $F1_{macro}$ represents the macro average F1 value. $|\frac{|m|}{H}$ represents the feature compression ratio; Err_{grade} is the grade error classification rate. Let λ_1 and λ_2 denote the constraint weights. This function makes SHO search process give consideration to both prediction accuracy and model compactness, and avoids getting a classifier that is only suitable for the majority class samples.

In the scoring output stage, the multi-classification decision method is used. The system calculates the decision score of the four levels of the sample respectively, and takes the level corresponding to the maximum score as the final prediction result. At the same time, the output saves the score gap as the confidence basis to facilitate the subsequent review of boundary samples. This mechanism enables the intelligent spoken language evaluation system not only to give the grade, but also to explain whether the prediction result is stable.

In order to convert the multi-class decision score into the spoken language rating and retain the classification confidence for the system to review and call, the prediction rule is as follows:

$$\hat{y}_i = \arg\max_{c \in \{A,B,C,D\}} s_c(z_i), \quad Conf_i = \frac{\exp(s_{\hat{y}_i})}{\sum_c \exp(s_c(z_i))} \tag{18}$$

where $\hat{y}_i$ denotes the prediction level of the i sample; $s_c(z_i)$ is the decision score of the sample on class c. $Conf_i$ denotes the prediction confidence. The rule transforms the SVM classification results into grade labels and confidence states that can be written into the system, so that the scoring output has engineering attributes that can be recorded, reviewed and counted.

After model training, the system generates SHO-SVM score classifier, optimal parameter configuration, feature index table, and rank output interface. In the online evaluation, the new input speech is converted into the model input according to the same process of feature extraction, normalization and subset projection, and then the classifier outputs the rank and confidence. This structure keeps the consistency between the training end and the application end, and also enables the intelligent evaluation system of spoken English to complete stable prediction at a low feature dimension. Finally, the SHO-SVM model connects pronunciation

features, search optimization and rank determination as a continuous computing link, which provides a clear model basis for subsequent experimental verification.

4 Experiments

4.1 Experimental Setup

The experiment focuses on the training, validation and testing of SHO-SVM intelligent spoken English evaluation system. The corpus is 2400 spoken English samples from 300 learners, including three types of tasks: reading sentences, short question answering and semi-open expression. The recordings are unified in 16kHz, 16bit, monaural WAV format. Each sample contains an expert rating, a rating label, a speaker number, and a question type number. The training set, validation set and test set were divided according to the ratio of 7:2:1, and the four-level label distribution of A, B, C and D was kept consistent in the three parts to avoid the influence of class sample skew on model comparison.

The experimental environment was completed with Python 3.10, Scikit-learn, LibSVM and self-written SHO search program. The hardware platform is configured with Intel i7 processor, 32GB memory and RTX 3060 graphics card. Speech preprocessing includes endpoint detection, pre-emphasis, framing and windowing, silent segment marking and abnormal audio removal. The feature extraction module calculates MFCC, fundamental frequency, formant, short-time energy, zero-crossing rate, speech rate, pause ratio, stress offset and phoneme duration, and uses training set statistics to complete normalization. The population size of the SHO algorithm is set to 30, the maximum number of iterations is 80, and the search objects include feature masks, SVM penalty factors, and radial basis kernel width.

In order to verify the performance of the model, SVM, random forest, BP neural network and SVM without SHO screening were set as comparison methods. All models use the same training, validation, and test sets, and the input features are of the same standardized caliber. The evaluation metrics include Accuracy, Precision, Recall, and F1-score, while the number of features, training time, and grade confusion are recorded. The model selection is based on the macro average F1 value of the validation set, and the final results are reported on the test set. To reduce the fluctuation caused by random partition, the experiment was repeated 10 times, and the average value was taken as the final performance, and the paired t-test was used to test the difference between SHO-SVM and the comparison model, and the significance level was set at 0.05. The experimental log is saved synchronously, which facilitates the reproduction of experimental process, parameter source, feature index and result tracking.

4.2 Experimental Results

The experimental results are analyzed from the aspects of the overall performance of the model, the scoring performance of different task types, the level confusion, the feature contribution distribution and the module ablation effect. All test results are obtained based on independent test sets, the model input uses the same set of standardized pronunciation feature matrices, and the scoring labels are coded with four levels A, B, C, and D. The SHO-SVM model retains 43-dimensional pronunciation features after feature selection, covering parameters such as MFCC statistics, fundamental frequency fluctuation, pause ratio, stress offset and phoneme duration. Compared with the full 126-dimensional feature input, the filtered feature matrix retains the main pronunciation discriminant information, and reduces the impact of redundant dimensions on the classification boundary. Subsequent charts focus

on the test set results to illustrate the specific performance of the model in terms of task type, hierarchical division, and feature contribution, respectively.

Table 4 presents the numerical results of different classification models on the same test set. All four models use the same standardized feature entry, RF and BP networks do not join the SHO search, and SVM only uses the full feature vector. In particular, the F1-score of SHO-SVM is 3.9 percentage points higher than that of the basic SVM, which indicates that the optimized feature subset can improve the misclassification caused by the imbalance of samples between grades.

Table 4: Test set scoring results for different classification models

Model	Accuracy/%	Precision/%	Recall/%	F1-score/%	Feature Dimension
SVM	90.7	89.6	88.8	89.4	126
RF	91.5	90.8	90.1	90.4	126
BP Network	90.9	89.9	89.2	89.5	126
SHO-SVM	94.6	93.8	92.9	93.3	43

Fig. 4 is used to compare the scoring stability of the SHO-SVM model in the three classes of spoken English tasks. In the task of reading aloud sentences, the Accuracy, Precision, Recall and F1-score are 95.8%, 95.0%, 94.7% and 94.8%, respectively. The distribution of the four indicators is relatively balanced, indicating that phoneme boundaries and stress positions are easier to be captured by the model under the condition of fixed text. The four indexes of the short question answering task were 94.3%, 93.5%, 92.6% and 93.0%, respectively, which decreased slightly compared with the reading sentence, mainly reflecting the effect of the change of answer length on the pause proportion and speaking speed characteristics. The four indicators of the semi-open expression task are 93.5%, 92.9%, 91.7% and 92.3%, and the Recall is lower than the other two types of tasks, indicating that the fluctuation of speech rate and the change of pause position in free expression increase the difficulty of level recognition. The radar distribution is able to present the degree of contraction of the four indicators at the same time, which is more suitable than a single bar chart to show the differences in comprehensive scores under different task types.

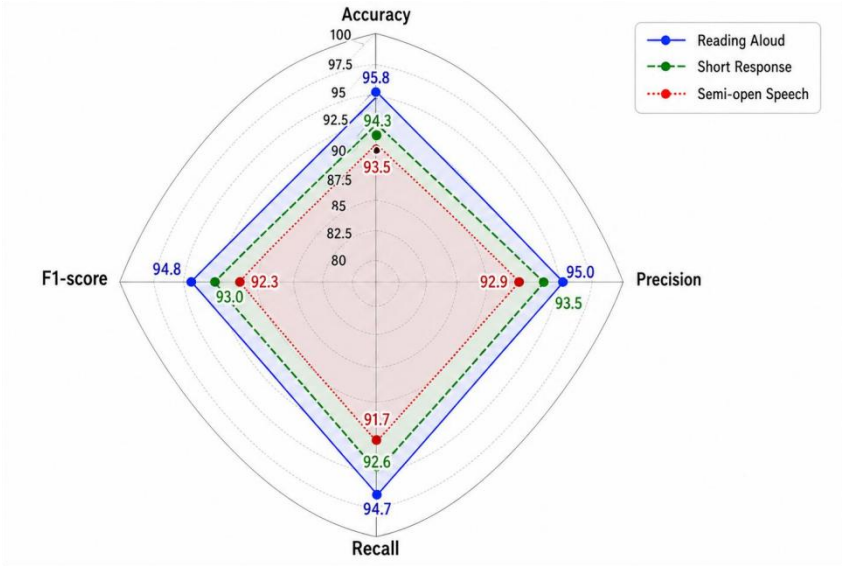

Figure 4: Radar distribution of scoring metrics for different spoken language task types

Table 5 shows the results of ablation experiments. After removing SHO search, the model can still complete the classification, but the F1-score decreases to 90.4%. After removing pause and accent features, the decrease of Recall is more obvious. After removing the sample weights, the misclassification of level D samples increases. The full model maintains the highest performance, which indicates that feature search, prosodic parameters and weight encoding jointly affect spoken language grade determination.

Table 5: Ablation experimental results of SHO-SVM module

Model Setting	Accuracy/%	Precision/%	Recall/%	F1-score/%
Without SHO search	91.2	90.9	89.8	90.4
Without pause and stress features	92.0	91.5	89.7	90.5
Without sample weights	93.1	92.4	91.1	91.7
Without SVM parameter optimization	92.4	91.8	90.9	91.3
Complete SHO-SVM	94.6	93.8	92.9	93.3

Fig. 5 shows the confusion matrix results of the SHO-SVM model in the four-level scores of A, B, C, and D. There were 49 class A samples in the test set, 47 of which were correctly identified as class A and 2 as class B. There were 73 samples of class B, of which 70 were correctly identified, 2 were judged as class A, and 1 was judged as class C. There were 81 samples of class C, of which 76 were correctly identified, 3 were judged as class B and 2 were judged as class D. Among the 37 samples of class D, 34 were correctly identified and 3 were judged as class C. The overall number of correctly identified samples was 227, and the Accuracy of the test set was 94.6%. Misclassification is mainly concentrated between adjacent grades, and there is no cross-level misclassification where class A is classified as class C or D, and no obvious deviation where class D is classified as class A or B. This result shows that the model has a strong ability to distinguish samples with large differences in pronunciation quality, and the misjudgment of boundary samples mainly comes from the overlap of acoustic features between adjacent scoring intervals.

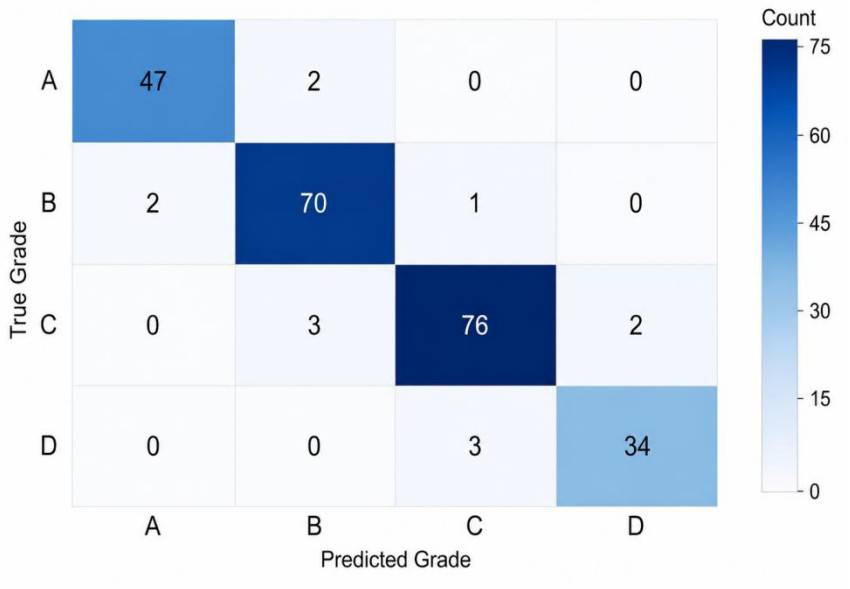

Figure 5: Heat map of the confusion matrix for the SHO-SVM four-level score

Table 6 further lists the classification details of each grade. The Precision and Recall of class A are both 95.9%, indicating that the acoustic features of the samples with high scores have a high degree of aggregation, and the model recognizes the samples with stable pronunciation and clear stress position more accurately. The number of class B samples is the largest, and the Recall reaches 95.9%, indicating that the model can better cover the middle and high level samples. The Precision of class C is 95.0%, and the Recall is 93.8%. There are still a few misclassifications of adjacent grades, which are mainly related to the pause proportion, speech speed fluctuation and close accent shift. The Recall of class D is 91.9%, which is lower than other grades, but still remains above 90%, indicating that the weak grade samples can be recognized by the model more stably. The comprehensive chart results show that SHO-SVM is superior to the comparison methods in feature compression, grade recognition and result stability, and can support the automatic scoring and grade output of the intelligent oral English assessment system.

Table 6: Classification details results for different rating levels

Grade	Test Samples	Precision/%	Recall/%	F1-score/%
A	49	95.9	95.9	95.9
B	73	93.3	95.9	94.6
C	81	95.0	93.8	94.4
D	37	94.4	91.9	93.2

Fig. 6 presents the inclusion stability of the main pronunciation features after SHO search over ten repetitions of the experiment. The stability of the mean MFCC is 0.91, the standard deviation of the fundamental frequency is 0.88, the pause ratio is 0.84, the stress offset is 0.82, and the phoneme duration is 0.79. It shows that the spectral structure, pitch fluctuation and time control are the main basis for the model to judge the spoken language level. The stability of the short-time energy is 0.63, the zero-crossing rate is 0.58, and the formant shift is 0.55, which belongs to the medium contribution feature and can supplement the articulation articulation and local timbre differences. The low-contribution features are mainly concentrated on the energy peak and local spectrum disturbance, and the stability is lower than 0.45, so the inclusion frequency in different data partitions is low. This thermodynamic ranking shows that the SHO algorithm does not simply retain all acoustic quantities, but gives priority to the pronunciation parameters with more stable correlation with the rating levels.

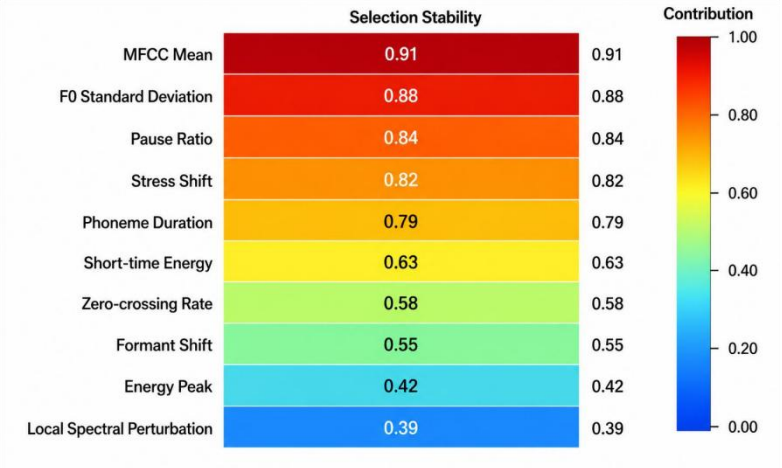

Figure 6: Heat ranking plot of articulatory feature contributions after SHO screening

From the overall results, the advantage of the model does not come from the improvement of a single index, but is formed by feature compression, parameter optimization and grade boundary stability. Forty-three dimensional features can still retain the discriminative information of pronunciation quality, indicating that the search and screening have a strong filtering effect on redundant acoustic parameters. The confusion matrix shows that the misclassification is mainly concentrated in the adjacent grades, which is consistent with the continuous change of oral score. The ablation results further prove that pause, accent, sample weight and parameter optimization all affect the final decision. The results show that the corpus-driven pronunciation feature modeling can provide a stable and recheckable calculation basis for intelligent spoken English assessment, and support the subsequent model deployment and corpus extension verification work.

5 Conclusion

Focusing on the task of corpus-driven spoken English assessment, this paper constructs an intelligent scoring system with corpus encoding, articulation feature extraction, SHO search and SVM rank determination. The experiment is based on 2400 speech samples from 300 learners, and the system forms a computable feature space from MFCC, fundamental frequency, formant, pause ratio, stress offset and phoneme duration. The redundant dimensions are reduced from 126 to 43 by SHO. The test results show that SHO-SVM achieves 94.6% Accuracy, 93.8% Precision, 92.9% Recall and 93.3% F1-score, which are better than basic SVM, random forest and BP network. The confusion matrix shows that the model misclassification is concentrated between adjacent grades, indicating that the classification boundary is consistent with the continuity of spoken score. Ablation experiments proved that SHO search, pause and accent features, sample weights, and SVM parameter optimization all have a practical impact on the scoring. The limitation is that the corpus is still dominated by a single English assessment scene, and the accent differences, speaking rate fluctuations and background noise in semi-open expressions have not been fully covered. SHO increases the offline training time, and the deployment efficiency of the model on mobile terminals or low computing power devices needs to be verified. In the future research, we will extend the multi-accent, multi-topic and multi-scene English corpus, introduce a lightweight acoustic encoder and an incremental learning mechanism, modify the feature weights by combining the online scoring log, and establish an interpretable feedback module, so that the system can output fine-grained pronunciation diagnosis evidence in addition to automatic scoring.

References

[1] Al-Ghezi R, Voskoboinik K, Getman Y, et al. Automatic speaking assessment of spontaneous L2 Finnish and Swedish[J]. Language Assessment Quarterly, 2023, 20(4-5): 421-444.

[2] Cámara Arenas E, Tejedor García C, Tomas Vázquez C J, et al. Automatic pronunciation assessment vs. automatic speech recognition: A study of conflicting conditions for L2-English[J]. 2024.

[3] Vidal J, Bonomi C, Riera P, et al. Automatic pronunciation assessment systems for English students from Argentina[J]. Communications of the ACM, 2024, 67(8): 63-67.

[4] Bashori M, van Hout R, Strik H, et al. I Can Speak: improving English pronunciation through automatic speech recognition-based language learning systems[J]. Innovation in Language Learning and Teaching, 2024, 18(5): 443-461.

[5] Bashori M, van Hout R, Strik H, et al. ‘Look, I can speak correctly’: learning vocabulary and pronunciation through websites equipped with automatic speech recognition technology[J]. Computer Assisted Language Learning, 2024, 37(5-6): 1335-1363.

[6] Saito K, Macmillan K, Kachlicka M, et al. Automated assessment of second language comprehensibility: Review, training, validation, and generalization studies[J]. Studies in second language acquisition, 2023, 45(1): 234-263.

[7] Kang B O, Jeon H B, Lee Y K. AI-based language tutoring systems with end-to-end automatic speech recognition and proficiency evaluation[J]. ETRI Journal, 2024, 46(1): 48-58.

[8] Lounis M, Dendani B, Bahi H. Mispronunciation detection and diagnosis using deep neural networks: a systematic review[J]. Multimedia Tools and Applications, 2024, 83(23): 62793-62827.

[9] Lounis M, Dendani B, Bahi H. Anomaly detection with a variational autoencoder for Arabic mispronunciation detection[J]. International Journal of Speech Technology, 2024, 27(2): 413-424.

[10] Bahi H, Dendani B, Lounis M. Automatic Pronunciation Assessment and Feedback for Arabic Learners: A Review[J]. International Journal of Asian Language Processing, 2024, 34(03n04): 2430001.

[11] Calık S S, Kucukmanisa A, Kilimci Z H. An ensemble-based framework for mispronunciation detection of Arabic phonemes[J]. Applied Acoustics, 2023, 212: 109593.

[12] Çalık Ş S, Küçükmanisa A, Kilimci Z H. A novel framework for mispronunciation detection of Arabic phonemes using audio-oriented transformer models[J]. Applied Acoustics, 2024, 215: 109711.

[13] Ahmed A, Bader M, Shahin I, et al. Arabic mispronunciation recognition system using LSTM network[J]. Information, 2023, 14(7): 413.

[14] Algabri M, Mathkour H, Alsulaiman M, et al. Mispronunciation detection and diagnosis with articulatory-level feedback generation for non-native arabic speech[J]. Mathematics, 2022, 10(15): 2727.

[15] Wilschut T, Sense F, van Rijn H. Speaking to remember: Model-based adaptive vocabulary learning using automatic speech recognition[J]. Computer Speech & Language, 2024, 84: 101578.

[16] Sabu K, Rao P. Predicting children’s perceived reading proficiency with prosody modeling[J]. Computer Speech & Language, 2024, 84: 101557.

[17] O'Shaughnessy D. Trends and developments in automatic speech recognition research[J]. Computer Speech & Language, 2024, 83: 101538.

[18] Prabhavalkar R, Hori T, Sainath T N, et al. End-to-end speech recognition: A survey[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 32: 325-351.

[19] O'Shaughnessy D. Review of methods for automatic speaker verification[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 32: 1776-1789.

[20] Yadav S, Kumar A, Yaduvanshi A, et al. A review of feature extraction and classification techniques in speech recognition[J]. SN Computer Science, 2023, 4(6): 777.