ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 1 - 2026

Semantic Prosodic Collaborative Modeling Method in the Intelligent Evaluation System of Spoken English Fluency in the Field of Green Energy Materials Engineering

Lili Li[1,*]

[1] College of Basic Courses, Shanxi Institute of Energy, Jinzhong 030600, Shanxi, China

SUMMARY: *In order to support the intelligent evaluation of Spoken English fluency in the field of green energy materials engineering, an automatic evaluation system based on semantic-prosodic collaborative modeling was proposed. The framework integrates speech transcription, term semantic representation, prosodic sequence encoding, grading score and error diagnosis feedback to jointly capture term usage, discourse advancement, pause, stress and speech rate changes. The corpus contains 4860 samples from 162 learners and 38 experts with synchronized audio, transcribed text, term labels, and prosodic annotations. Experiments show that SSPF-Net achieves 93.40% classification accuracy, 92.87% Macro-F1 and 0.918 QWK, which are better than the four baseline models, and gives consideration to scoring accuracy and deployment efficiency. At the same time, the average inference delay is 84 ms, which meets the requirements of edge device deployment. The system performance remains stable with only 3.6% scoring variance in the technical terms and noisy speech conditions. The feedback module can further locate semantic bias and prosodic imbalance, providing interpretable evaluation for computer-aided spoken English training in engineering scenarios.*

KEYWORDS: *Professional oral English; Fluency evaluation; Semantic prosody modeling; Intelligent evaluation system*

1 Introduction

The international communication in the field of green energy materials engineering has obvious professional spoken language characteristics. Battery materials, energy storage devices, interface reactions, cycle life and thermal stability are usually in the form of high-density terminology, compressed syntax and rapid information switching in academic reports, project defense and collaborative discussions. When evaluating such expressions, it is difficult to accurately reflect the true fluency only by looking at the segmental pronunciation or semantic integrity. An intelligent evaluation system for this scenario needs to process speech signals, transcribed text, term semantics and prosodic sequences simultaneously, so that word selection, pause organization, rhythm distribution and semantic coherence can be represented, scored and feedback in the same computational framework. Therefore, spoken English fluency evaluation in the field of green energy materials engineering is no longer a simple pronunciation detection task, but gradually turns to a system task of joint semantic and prosodic modeling, whose technical implementation directly relies on automatic speech recognition, sequence learning, multi-modal fusion and scoring calibration.

Focusing on the basic technology of automatic spoken language evaluation, Wei et al.

*lilyleehu@126.com
https://doi.org/10.65102/is2026084 Published: 30 April, 2026

studied Dutch non-native speech recognition and pronunciation error detection, and improved the usability of non-native spoken language modeling by accumulating multi-source speech resources [1]. Al-Ghezi et al. established an automatic evaluation framework for spontaneous spoken Finnish and Swedish, which made the machine scoring in the natural expression situation obtain a relatively stable structural support [2]. Kallio et al. focused on the relationship between prosody and fluency in spoken second language, and proposed that global prosody parameters could be used to describe rhythm features in automatic evaluation [3]. Inceoglu et al. compared the judgment of native listeners with the ASR results, indicating that the machine recognition output has been able to play an important reference role in intelligibility evaluation [4]. Farrus systematically collated ASR applications in second language learning from the review level, and showed the main paths for the transfer of recognition models to learning feedback systems [5]. Saito et al. discussed the design of training, validation and generalization around automatic intelligibility evaluation, which provided a reference research framework for the experimental organization of subsequent scoring models [6]. This phase of research shows that automatic evaluation has gradually expanded from simple score output to a computational link where recognition, representation, discrimination and feedback are interconnected.

In order to more clearly present the technical distribution of existing research on non-native spoken language recognition, prosodic modeling, automatic scoring and feedback generation and their reference to this paper, an overview of related work is shown in Table 1.

Table 1: Overview of related work

Reference	Research Object	Key Method	Inspiration for This Study
[1]	Non-native speech recognition	Resource accumulation and error correction	Expands domain-specific corpora
[2]	Spontaneous spoken language assessment	Automated scoring framework	Supports natural speaking scenarios
[3]	Prosody and fluency	Global parameter modeling	Supports prosodic encoding
[4]	Intelligibility assessment	Listener–ASR comparison	Supports dual evaluation
[5]	ASR-based learning applications	Systematic review	Supports module selection
[6]	Comprehensibility assessment	Training and generalization design	Supports experimental organization
[7]	Automatic prosody scoring	Deep learning algorithms	Supports the scoring network
[8]	Spoken proficiency measurement	Computer-assisted assessment tools	Supports system implementation
[9]	Analysis of prosodic effects	Correlation of phonological features	Supports robustness validation

Along the direction of deep modeling, Wang et al. applied deep learning to automatic prosody scoring in reading fluency, indicating that prosody indicators can be directly embedded into the scoring network [7]. Li et al. developed a machine-aided tool for automatic measurement of non-native Japanese spoken proficiency, and proved that the system-level implementation has been able to carry more fine-grained language ability discrimination [8].

Emara and Shaker analyzed the influence of phonological and prosodic features of non-native English on ASR accuracy, and showed that prosodic changes under complex acoustic conditions would directly interfere with machine judgment [9]. Zhu et al. proposed a pronunciation error detection model based on feature fusion, which makes the joint modeling of multi-source acoustic features an effective scheme in oral language evaluation [10]. These researches promote the transfer of automatic evaluation from single acoustic score to joint representation score, and also provide a transferable technical basis for semantic prosodic collaborative modeling in professional English scenes.

In terms of score interpretation and system feedback, Handley and Wang re-examined the explanatory power of common fluency indicators in automatic spoken language assessment, and pointed out that a clearer mapping relationship between fluency features and spoken language level was needed [11]. Ballier et al. introduced Whisper into spoken language scoring and demonstrated the usability of large model speech transcription in second language scoring [12]. Shi et al. examined the relationship between automatic scoring indicators and complexity, accuracy and fluency, which provided a more explicit basis for the linkage interpretation of multiple indicators [13]. Inceoglu et al. further incorporated learner behavior monitoring into ASR pronunciation practice, which promoted the linkage analysis between evaluation results and learning process [14]. Banno and Matassoni introduced the results of grammatical error correction into oral scoring, which expanded the analysis dimension of the scoring system for expression quality [15]. Johnson et al. used dictation tools to achieve pronunciation test scoring, indicating that lightweight tools can also form an automatic scoring link with practical value [16]. Nickolai et al. systematically summarized the evidence structure of ASR research conclusions in the field of CALL, which provided a more stable evidence base for method selection in system design [17]. Bashori et al. constructed an ASR-based English pronunciation learning system to form a closed loop between automatic recognition and feedback generation [18]. Kim et al. studied the automatic recognition of second language speech in noisy environment, indicating that robust recognition ability is still one of the core conditions for the implementation of the system [19]. Vidal et al. established an automatic pronunciation evaluation system for English learners to make the engineering realization path clearer [20]. The research in this stage makes the automatic oral scoring no longer stay at the level of "score giving", but gradually have the system attributes of behavior monitoring, indicator interpretation and feedback generation.

Compared with general learning scenarios, spoken English in the field of green energy materials engineering includes the characteristics of intensive professional terms, large semantic span, and obvious influence of pause position by concept organization. Term misreading, rhythm imbalance and semantic break often appear synchronously, and it is difficult to make stable judgments by only relying on local characteristics such as pitch, duration or speaking rate. It is easy to ignore the real impact of stress point, pause boundary and rhythm structure on fluency by only using text semantic scoring. Therefore, it is more suitable for intelligent evaluation in this field to adopt the idea of collaborative semantic and prosodic modeling: integrating speech and text information at the recognition level, jointly investigating the semantic integrity and prosodic organization ability of terms at the evaluation level, and outputing localized and interpretable diagnosis results at the feedback layer. Based on this goal, this paper constructs an English spoken corpus and a multi-modal feature representation method in the field of green energy materials engineering, designs a semantic prosody collaborative modeling and grading scoring network for fluency evaluation, and introduces error diagnosis and feedback generation mechanism in the scoring output layer to form an intelligent evaluation system that takes into account accuracy, stability and deployment efficiency.

2 Methods and materials

2.1 Construction and Multi-modal Feature Representation of Spoken English Corpus in the field of Green Energy Materials Engineering

The goal of this section is not to describe the speech acquisition process separately, but to establish a unified input system that can directly serve the subsequent semantic prosodic collaborative modeling. The spoken language of green energy materials engineering contains a large number of high-frequency professional terms, compressed expressions and continuous declarative segments. The speech flow contains not only segments with high semantic density such as material names, performance indicators and technological processes, but also fluency cues such as pauses, extensions, stress shifts and speech rate changes. In order to make the system not only recognize professional content, but also maintain the sensitivity to the rhythmic structure of spoken language, the corpus construction stage adopts the processing method of "scene collection, transcription correction, term tagging, sequence alignment, and unified representation", and compresses the original speech, transcribed text and prosodic statistics into a computable multimodal input.

In order to form a continuous data processing link between the collection, labeling, screening and representation of spoken English samples in the field of green energy materials engineering, this paper constructs a domain corpus construction and unified input process, as shown in Fig. 1. The process takes the spoken language task unit as the basic organization granularity, and integrates the original recording, automatic transcription results, technical terminology tagging and prosodic statistical information into the same processing framework. The standardization is completed in five stages: sample segmentation, text correction, term positioning, prosodic extraction and input mapping. Through this structure, speech content information and fluent-related information are preserved synchronously, and a consistent data entry is provided for subsequent semantic prosodic collaborative modeling.

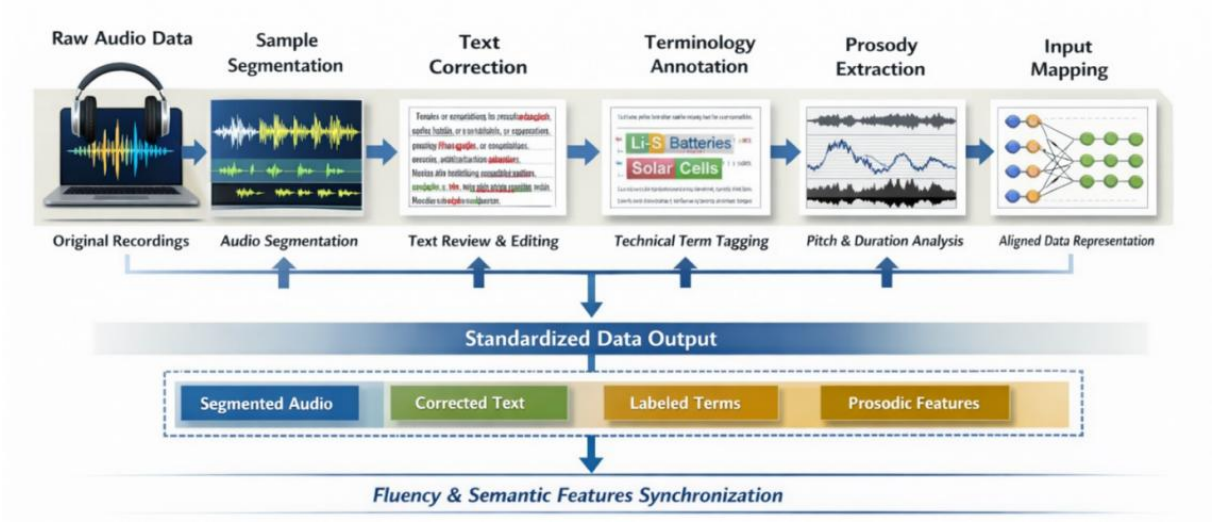

Figure 1: Construction and unified input process of spoken English corpus in the field of green energy materials engineering

In the speech channel, the raw recordings are first subjected to pre-emphasis, framing, and endpoint screening, followed by extraction of Mel-frequency cepstral coefficients, short-time energy, zero-crossing rates, and fundamental frequency profiles. In order to avoid the separation of local segment features and global prosodic cues in the representation stage, the spectral

vector and the prosodic vector are jointly mapped into a unified frame-level acoustic representation, which is calculated as shown in Equation (1).

$$a_t = \phi\left(W_m \begin{bmatrix} c_t \\ \Delta c_t \\ \Delta^2 c_t \end{bmatrix} + W_q \begin{bmatrix} F0_t \\ E_t \\ Z_t \end{bmatrix} + b_a\right) \tag{1}$$

Here, a_t denotes the unified acoustic representation of the t frame; c_t denotes the MFCC vector of the t frame, Δc_t and $\Delta^2 c_t$ denote the first and second difference, respectively. $F0_t$ represents the fundamental frequency of the frame, E_t represents the short-time energy, and Z_t represents the zero-crossing rate. W_m represents the spectral feature mapping matrix, W_q represents the prosodic feature mapping matrix, b_a represents the bias term, and $\phi(\cdot)$ represents the nonlinear activation function. The role of Eq. (1) is to map spectral details and prosodic statistics into the same frame-level representation space, so that subsequent models can preserve both the resolution of local articular features and the rhythm cues associated with fluency.

The text channel does not directly retain the automatic transcribing results, but adds a domain term awareness step after transcribing. According to the green energy materials engineering vocabulary, the system marks the material names, process terms, performance indicators and experimental action words, and then combines context embedding to construct term weighted semantic representation. The calculation process is shown in Formula (2):

$$s_i = \frac{\sum_{j=1}^{m_i} \omega_{ij}\, e_{ij}}{\sum_{j=1}^{m_i} \omega_{ij}}, \quad \omega_{ij} = \lambda_1 c_{ij} + \lambda_2 d_{ij} + \lambda_3 r_{ij} \tag{2}$$

Here, s_i represents the semantic vector of the i utterance unit, m_i represents the number of lemmas in the unit, e_{ij} represents the context embedding of the j lemma, and ω_{ij} represents the corresponding weight. c_{ij} denotes the context importance, d_{ij} denotes the term matching strength, r_{ij} denotes the paraphrase confidence, and λ_1, λ_2, and λ_3 are the normalization coefficients. The function of equation (2) is to improve the contribution ratio of specialized words in the semantic representation, so that "whether the expression is spoken" and "whether the term is accurate" can be perceived simultaneously in the same vector.

The acoustic and semantic vectors alone are still insufficient to fully describe fluency, because rhythm organization, pause boundaries, and stress distribution in spoken language evaluation tend to occur on longer time scales. Based on this consideration, this paper further constructs the sentence-level prosodic structure vector, and incorporates the average fundamental frequency, fundamental frequency fluctuation, average pause, pause dispersion, speech rate and stress density into a unified statistical framework, which is defined as Formula (3):

$$p_i = [\mu_{F0}, \sigma_{F0}, \mu_\tau, \sigma_\tau, \rho_i, \kappa_i], \quad \rho_i = \frac{N_i}{T_i}, \quad \kappa_i = \frac{1}{N_i}\sum_{k=1}^{N_i} \left|\frac{\Delta F0_k}{\Delta t_k}\right| \tag{3}$$

where p_i represents the prosodic vector of the i utterance unit, μ_{F0} and σ_{F0} represent the average fundamental frequency and its fluctuation, μ_τ and σ_τ represent the mean pause duration and its dispersion degree, ρ_i represents the number of effective syllables per unit duration, T_i represents the total duration, N_i represents the number of effective syllables, and κ_i represents the intensity of stress change. The function of Formula (3) is to organize the

scattered prosodic phenomena into comparable and quantifiable structural information, so that the speech speed, pause stability and intonation fluctuation no longer appear as isolated fragments, but form the fluency representation at the sentence level.

In order to ensure that frame-level acoustic features, unity-level semantic features and sentence-level prosodic features can establish stable correspondences on the same time axis, this paper further designs a multi-modal feature representation and temporal alignment structure, as shown in Fig. 2. With resampling, normalization and gated mapping as the core, the proposed structure uniformly projects inputs of different time scales and different dimensional forms into a shared representation space, which enables the formation of computable associations between local speech frames, term semantic units and global prosodic patterns. Such a processing method helps to reduce the timing offset and granularity inconsistency between cross-modal inputs, so that the subsequent scoring network can perform stable discrimination at the complete expression level.

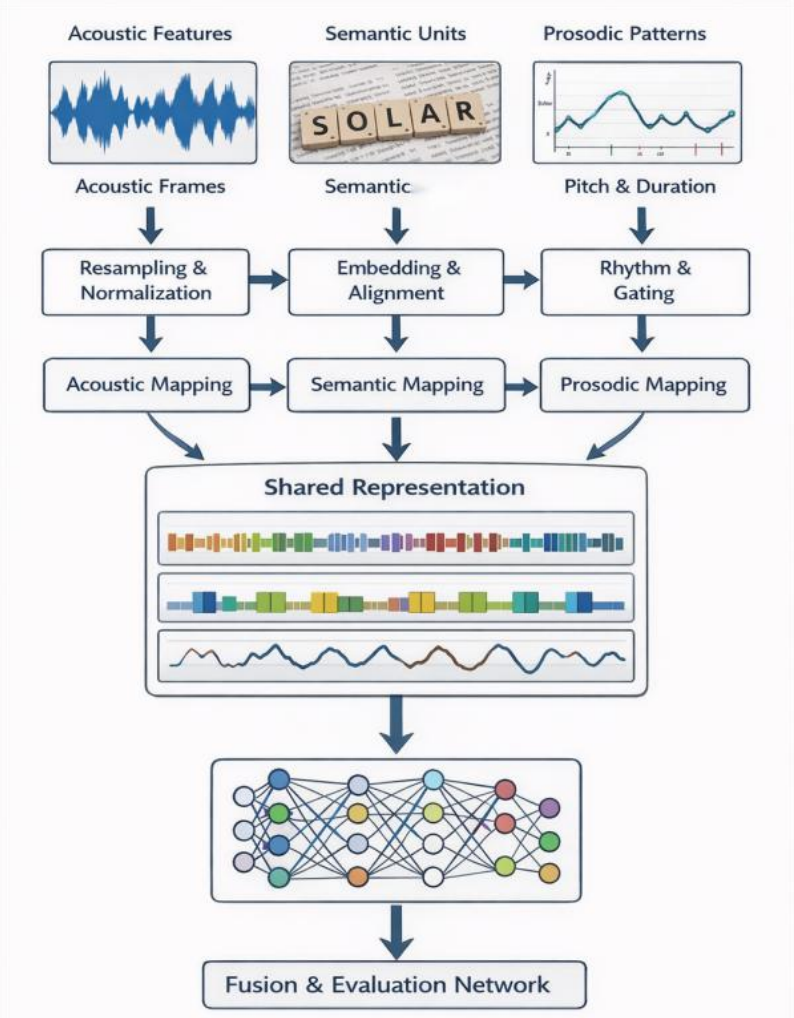

Figure 2: Speech-semantic-prosodic multimodal feature representation with time-aligned structure

In the unified representation stage, this paper uses the cross-modal mapping with gated weights to compress the three types of inputs into the same latent space, and the calculation form is shown in Formula (4).

$$h_i = \sum_{t=1}^{T_i} \alpha_{it} W_a a_t + W_s s_i + W_p p_i, \quad \alpha_{it} = \frac{\exp(u^\top \tanh(U_a a_t + U_s s_i))}{\sum_{t'=1}^{T_i} \exp(u^\top \tanh(U_a a_{t'} + U_s s_i))} \tag{4}$$

Here, h_i represents the unified representation of the i utterance unit, W_a, W_s and W_p represent the acoustic, semantic and prosodic mapping matrices, respectively, α_{it} represents the contribution weight of the t frame to the current unit, and U_a, U_s and u are the attention parameters. The function of equation (4) is to use semantic information to dynamically select more critical acoustic frames, while injecting sentence-level prosodic vectors into the unified representation, thus forming an input result that contains both content information and retains fluency cues.

After the above processing, the original corpus is organized into a three-layer input structure of "frame-level acoustic representation, unit-level semantic representation, sentence-level prosodic representation". The structure can not only cover the professional term information in green energy materials engineering spoken language, but also retain fluency signals such as pause, rhythm and stress, which provides a stable, continuous and interpretable data basis for subsequent semantic prosody collaborative modeling and grading scoring network.

2.2 Semantic Prosody Collaborative Modeling and Grading scoring Network for fluency evaluation

The fluency discrimination of spoken English in the field of green energy materials engineering is essentially a joint estimation of the semantic advancement of terms, the stability of rhythmic organization and the continuity of expression. Material names, performance indicators, process steps and experimental conclusions often appear in a high-density manner in oral expressions. The change of the semantic center of gravity will directly affect the stress position, pause boundary and speech rate distribution. If the judgment is only based on local acoustic segments, the model is easy to retain surface speech information and weaken the logical cohesion between professional content. Scoring only based on text content would weaken the true impact of rhythm imbalance, pause drift and prosody fluctuation on fluency. Based on this feature, this section constructs a semantic prosody collaborative modeling and grading scoring network, which enables the interaction, aggregation and discrimination of term semantics, temporal prosody and local acoustic state in the same computing path.

Fig. 3 shows the hierarchical composition and calculation path of the semantic prosody collaborative modeling and hierarchical scoring network. In the figure, the modal projection layer, interactive enhancement layer, temporal state update layer, salient segment aggregation layer and hierarchical output layer are given in turn. The first two layers are responsible for pulling inputs from different sources into a shared space and establishing semantic-prosodic coupling relationships, the middle layer is responsible for propagating expression states along the time axis, and the last two layers promote key segments as global scoring basis. Through this structure, local term pronunciation, intra-sentence rhythm variation and whole expression coherence are incorporated into the unified scoring link, and the scoring results are no longer dependent on a single segment, but are generated based on the complete expression process.

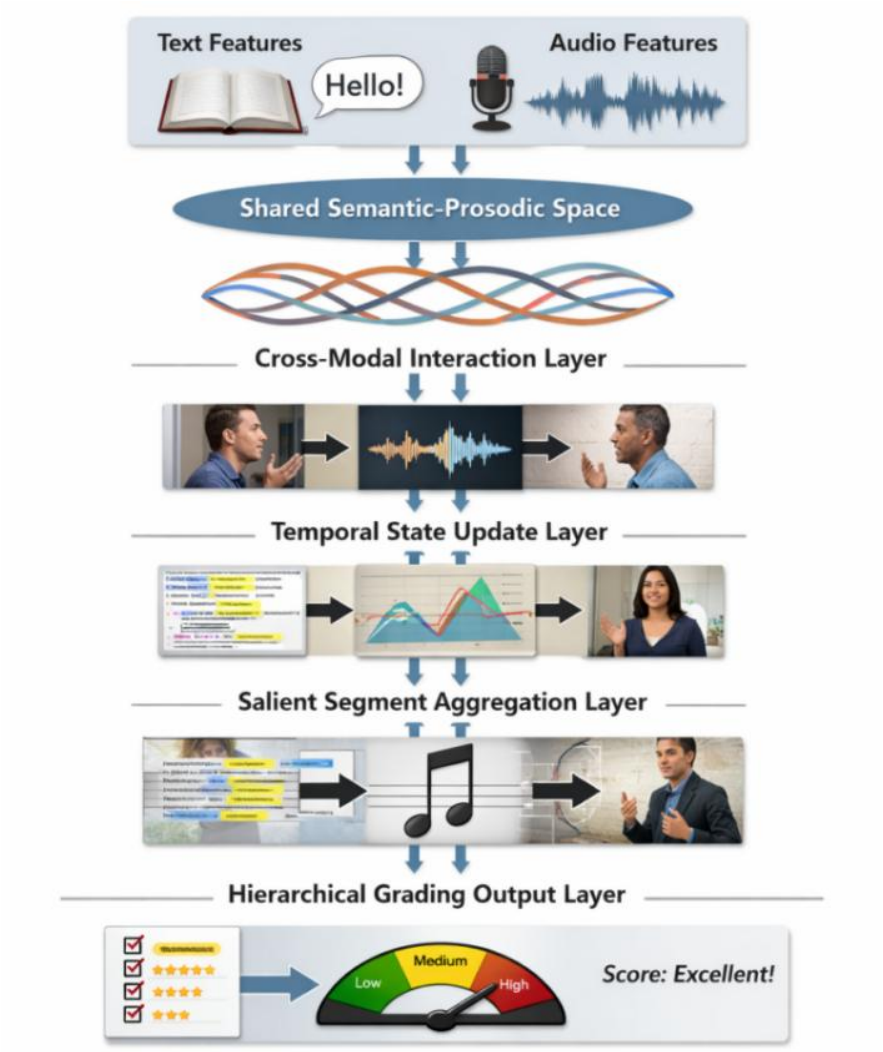

Figure 3: Semantic prosodic collaborative modeling and hierarchical scoring network structure

In the shared space construction stage, this paper first performs the normalized projection on the acoustic input, semantic input and prosodic input, and its calculation form is shown in Equation (5).

$$\tilde{a}_t = \mathrm{LN}(W_a a_t + b_a), \quad \tilde{s}_t = \mathrm{LN}(W_s s_t + b_s), \quad \tilde{p}_t = \mathrm{LN}(W_p p_t + b_p) \tag{5}$$

Here, $\tilde{a}_t$,$\tilde{s}_t$ and $\tilde{p}_t$ represent the acoustic vector, semantic vector and prosodic vector after projection at the t time step, respectively; W_a, W_s and W_p are modal mapping matrices; b_a, b_s and b_p are bias terms; and $\mathrm{LN}(\cdot)$ represents layer normalization. The function of formula (5) is to compress different modalities into a representation space with consistent dimensions and comparable scales, which provides a unified entrance for subsequent interactive calculations.

After the shared space is established, the model needs to further characterize the coupling relationship between term semantics and prosodic structure. Instead of simple concatenation, this paper introduces interactive enhanced representation, which is calculated as shown in Equation (6):

$$c_t = \tanh(W_c[\tilde{a}_t; \tilde{s}_t; \tilde{p}_t; \tilde{s}_t \odot \tilde{p}_t] + b_c) \tag{6}$$

Here, c_t represents the interaction enhancement vector at the t time step, $[\cdot;\cdot]$ represents the vector sequence, $\odot$ represents element-wise multiplication, and W_c and b_c are interaction mapping parameters. The function of Eq. (6) is to explicitly preserve the multiplicative interaction terms between semantic vectors and prosodic vectors, so that semantic changes in termdense segments can directly affect prosodic modeling results.

On this basis, this paper further constructs a gated temporal state update mechanism to continuously propagate the collaborative information along the expression process, whose expression is shown in Formula (7):

$$g_t = \sigma(W_g[\tilde{s}_t; \tilde{p}_t] + b_g), \quad h_t = \mathrm{GRU}(c_t, h_{t-1}) + g_t \odot \tilde{s}_t + (1 - g_t) \odot \tilde{p}_t \tag{7}$$

Here, g_t represents the semantic prosodic gating vector, $\sigma(\cdot)$ is Sigmoid function, h_t represents the timing state at the t time step, and $\mathrm{GRU}(\cdot)$ represents the gated recurrent unit. The function of formula (7) is not simply to do sequence encoding, but to dynamically adjust the proportion of semantic contribution and prosodic contribution in the time recursion, so that the concept advancement and rhythm change in expression can synchronize into the state update process.

In order to avoid dilutes the scoring results of non-critical segments in long sequences, this paper introduces a temporal attention aggregation mechanism under term significance constraint, and the definition of attention weight is shown in Formula (8).

$$\alpha_t = \frac{\exp(u^\top \tanh(W_h h_t + W_d \beta_t + b_h))}{\sum_{\tau=1}^{T} \exp(u^\top \tanh(W_h h_\tau + W_d \beta_\tau + b_h))} \tag{8}$$

Here, α_t denotes the attention weight at the t time step, T denotes the total number of time steps, β_t denotes the term saliency at that moment, and u, W_h, W_d, and b_h are the attention parameters. The effect of equation (8) is to make the scoring network pay more attention to the segments that are more indicative of fluency quality, such as term interpretation, concept turning and pause boundaries, rather than looking at the whole speech flow on average.

In the output layer, this paper uses ordinal hierarchical scoring instead of ordinary Softmax classification to maintain order constraints between the levels. The discriminant form is shown in Equation (9).

$$P(y \le k \mid r) = \sigma\big(\theta_k - w^\top [r; \bar{p}; \bar{\beta}]\big), \quad k = 1,2,3 \tag{9}$$

Here, $P(y \le k \mid r)$ represents the cumulative probability that the sample belongs to the k level and below, $r = \sum_{t=1}^{T} \alpha_t h_t$ represents the global expression state, $\bar{p}$ represents the sentential prosodic statistical summary, $\bar{\beta}$ represents the mean significance of the whole term, θ_k is the classification threshold, and w is the discriminant parameter. The function of equation (9) is to put the "weak-so-good-excellent" four-level results into an ordered probability space, so that the scoring network can not only give the grade judgment, but also maintain the continuous boundary between adjacent grades.

Through the above structure, the semantic-prosodic collaborative modeling and grading scoring network no longer only judges whether the spoken language is close to the surface form of English, but further judges whether the professional terms are natural cohesion, whether the information is stable, and whether the rhythm organization is consistent with the semantic boundary. In this way, the scoring results obtained are closer to the actual needs of the spoken expression of green energy materials engineering, and also provide a stable input for the error diagnosis and feedback generation mechanism in the next section.

2.3 Error diagnosis and feedback generation mechanism based on scoring output

The automatic assessment of spoken English in the field of green energy materials engineering does not end at the grade output. For professional expressions containing high-density terms, continuous instructions and conceptual turning points, the grading results can only reflect the overall state, and it is difficult to reveal the specific sources of term offset, rhythm imbalance, stress drift and pause misalignment. In order for the scoring results to be truly transformed into executable training information, it is necessary to continue to complete the error decomposition, principal cause discrimination and feedback organization after the classification output. Based on this requirement, this section constructs an error diagnosis and feedback generation mechanism based on scoring output, which integrates grade probability, semantic deviation, prosodic deviation and task context into the diagnosis link, and generates feedback results with priority on this basis, so that the system can be further extended from "scoring" to "explaining the score" and "guiding the correction".

At the entrance of diagnosis, the system first calculates the score reliability by combining the classification probability and distribution entropy to determine whether the current sample needs to enter the depth diagnosis. Its calculation form is shown in Equation (10):

$$r_i = \lambda_1\left(1 - \hat{y}_{i,y_i^*}\right) + \lambda_2\left(-\sum_{k=1}^{K} \hat{y}_{ik} \log \hat{y}_{ik}\right) \tag{10}$$

Here, r_i represents the diagnostic trigger value of the i sample, $\hat{y}_{i,y_i^*}$ represents the prediction probability corresponding to the target level, $\hat{y}_{ik}$ represents the prediction probability that the sample belongs to the k level, K represents the number of scoring levels, and λ_1 and λ_2 are the weight coefficients. The function of formula (10) is to compress the "degree of deviation from the target level" and "uncertainty of prediction distribution" into a single trigger indicator. When r_i is high, it means that although the sample has been scored, its classification boundary is unstable or deviates from the target greatly, and it needs to enter the subsequent error decomposition process.

In the stage of semantic diagnosis, the system does not directly compare the original text, but calculates the semantic distance between the learner expression vector and the target template vector to identify the phenomenon of technical term omission, term substitution and loose logical connection. Its calculation form is shown in Equation (11):

$$d_i^{sem} = 1 - \frac{r_i^{\top} t_i}{\|r_i\|_2 \|t_i\|_2} \tag{11}$$

Here, d_i^{sem} represents the semantic deviation score of the i sample, r_i represents the global semantic representation expressed by the learner, t_i represents the semantic representation of the target template, and $\|\cdot\|_2$ represents the two-norm. The function of formula (11) is to measure the directional consistency between the learner expression and the standard expression from the representation space. When the value rises, it indicates that the term content and semantic advancement are significantly deviated.

Fig. 4 shows the overall computational flow of the error diagnosis and feedback generation mechanism. The process takes the grading results as the starting point, and organizes the confidence assessment, semantic deviation localization, prosodic alignment analysis, error type discrimination and feedback ranking into a continuous data processing path. The nodes in the

diagram are not isolated functional modules, but follow the interpretation requirements of the scoring results: the front-end nodes are responsible for determining whether the current output is stable enough, the middle nodes are responsible for decomposing the sources of grade deviation, and the back-end nodes are responsible for condensating the diagnosis results into readable feedback, visual tips, and training suggestions. Through this structure, the scoring output originally stuck in the probability space is transformed into the specific correction information for learners, so that the evaluation system has stronger interpretation ability and interactive value.

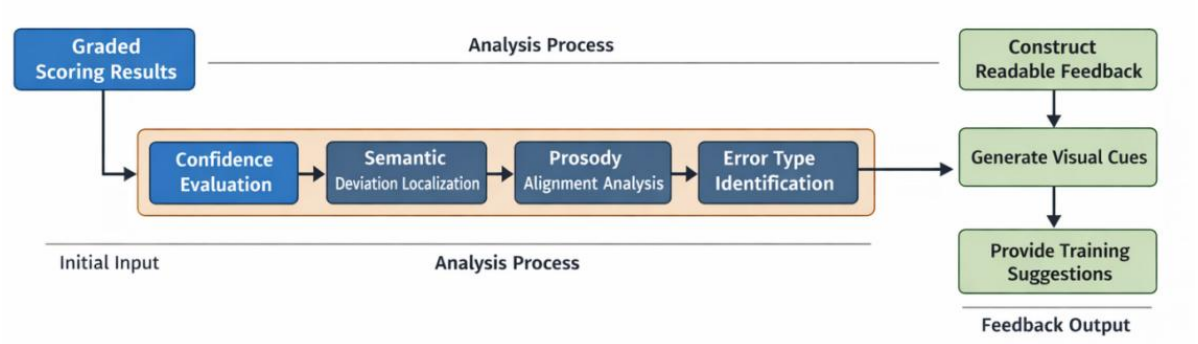

Figure 4: Error diagnosis and feedback generation mechanism based on scoring output

Semantic distance alone is still not sufficient to explain fluency loss, as pauses, stress, and rhythm changes in spoken speech have independent discriminant value. To this end, the system further performs temporal alignment between the learner prosodic sequence and the template prosodic sequence, and constructs the cumulative cost matrix, whose recurrence relationship is shown in Equation (12) :

$$D(m,n) = c(m,n) + \min\{D(m-1,n), D(m-1,n-1), D(m,n-1)\} \quad (12)$$

Here, $D(m,n)$ represents the minimum cumulative cost between the m position of the learner's prosodic sequence and the n position of the template sequence, and $c(m,n)$ represents the local distance between the two positions. The function of equation (12) is to find the optimal alignment path through dynamic programming, so that prosodic patterns at different speaking rates and durations can be meaningfully compared.

After obtaining the cumulative path, the system normalizes it to a comparable prosodic deviation score, as defined in Equation (13) :

$$d_i^{pro} = \frac{D(M_i, N_i)}{M_i + N_i} \quad (13)$$

Here, d_i^{pro} represents the prosodic deviation score of the i sample, M_i and N_i represent the length of the learner sequence and template sequence respectively, and $D(M_i, N_i)$ represents the cumulative cost at the end point. The function of formula (13) is to normalize the timing deviations under different length conditions to the same scale, so that the speech rate differences will not directly amplify the diagnosis results, and the real rhythm imbalance and stress drift can be stably identified.

In the error type discrimination stage, the system inputs the semantic deviation, prosodic deviation and scoring trigger value into the diagnosis layer together to distinguish four types of results: "term semantic deviation", "rhythm imbalance", "accent abnormality" and "compound deviation". The probability expression is given in Equation (14).

$$p_i^{err} = \mathrm{softmax}\left(W_e[d_i^{sem}; d_i^{pro}; r_i; \bar{p}_i] + b_e\right) \tag{14}$$

Here, p_i^{err} represents the error type probability vector, $\bar{p}_i$ represents the sentence-level prosodic statistical summary, and W_e and b_e are diagnostic layer parameters. The function of equation (14) is to compress the originally scattered deviation indicators into interpretable error categories, so that the subsequent feedback will no longer remain in general hints, but can point to specific sources of imbalance.

In the feedback generation phase, the system also needs to judge which prompts should be presented preferentially. Simply juxtaposing the output often reduces the readability of the feedback when multiple biases occur at the same time. Therefore, the feedback priority score is constructed in this paper, and its calculation form is shown in Formula (15).

$$q_{ij} = \eta_1 p_{ij}^{err} + \eta_2 d_{ij}^{sev} + \eta_3 c_{ij}^{ctx} \tag{15}$$

Here, q_{ij} represents the priority score of the j feedback in the i sample, p_{ij}^{err} represents the error class probability corresponding to this feedback, d_{ij}^{sev} represents the deviation severity, c_{ij}^{ctx} represents its relevance to the current spoken task context, and η_1, η_2, and η_3 are the weight coefficients. The function of equation (15) is to rank the candidate feedback so that the system preferentially outputs the prompt content that most affects the expression quality and is most worthy of immediate correction.

In the interface output phase, the system generates two types of results based on the feedback items with the highest priority. One is structured text feedback, which is used to point out specific phenomena such as "incomplete term description", "pause falling on non-concept boundaries" or "stress deviation from performance index words". The other category is visual feedback, which displays semantic propulsion trajectories, prosodic alignment curves, and bias position hot spots. Instead of stacking all the metrics directly, the interface displays the current level, type of error, and main feedback before moving on to detailed explanations and training recommendations. The real-time feedback link formed in this way enables learners to see the grade judgment, the source of deviation and the correction direction after one expression, so as to transform the output of the scoring module into action information for further training.

Through the above mechanism, the output of the hierarchical scoring network is further transformed into a structured result with diagnostic significance. The system can not only indicate the grade of the current expression, but also indicate whether the grade deviation comes from term semantics, rhythm organization or stress control. The feedback content with clear priority can be generated according to the deviation strength and task context. For spoken English in the field of green energy materials engineering, this closed loop from scoring, diagnosis to feedback is closer to the actual needs of professional expression training, and also makes the evaluation system transform from a static scoring tool to an intelligent auxiliary module that can continuously intervene in the process of expression correction. The output results thus formed not only retained the stability of the algorithm discrimination, but also enhanced the readability and executability of the feedback content, which provided direct support for subsequent interface interaction and training iterations.

3 Results

3.1 Comparison between evaluation performance and baseline of semantic prosodic collaboration model

The experiments were run on a server equipped with dual Intel Xeon Gold 6338 CPU, 512 GB DDR4 memory and four NVIDIA A100 80 GB Gpus. The operating system was Ubuntu 22.04. The deep learning framework is PyTorch 2.1, the programming environment is Python 3.10, and CUDA 12.1 is enabled synchronously with cuDNN 8.9. The model was trained using AdamW optimizer with batch size set to 48, maximum number of training rounds set to 120, initial learning rate set to 2×10−4, and with cosine annealing scheduler to control late convergence. The scoring task uses a four-level label system, and the joint index of weighted Kappa and Macro-F1 is used to select the optimal model in the validation stage. The corpus contains 4860 entries from 162 learners and 38 experts, with 3888, 486 and 486 entries for training, validation and test sets, respectively. Each sample contains speech, transcribed text, term hit labels, sentence-level prosodic statistics, and manual ratings, which can cover common scenarios such as term explanation, experiment description, mechanism explanation, and result reporting.

To make the experimental setup clearer, the main training configuration is shown in Table II. This table does not directly serve performance comparison, but gives the basic parameters for the experiments in this section to be reproducible.

Table 2: Training and evaluation environment configuration

Configuration Item	Parameter Setting
Training Framework	PyTorch 2.1
Programming Environment	Python 3.10
GPU Configuration	4 × NVIDIA A100 80 GB
Optimizer	AdamW
Batch Size	48
Number of Epochs	120
Initial Learning Rate	(2×10^{-4})
Learning Rate Schedule	Cosine Annealing
Evaluation Metrics	Accuracy, Macro-F1, QWK, MAE, RTF
Label System	Four-level Fluency Classification

The baseline model is set around the input form and scoring method, and a total of four types of methods are selected for comparison: Audio-BiLSTM only uses acoustic input, Text-BERT only uses transcribed semantic input, AudioText-Transformer uses bimodal joint modeling of acoustic and Text, and Whister-Score mainly uses large model transcribed results with additional lightweight scoring heads. The core feature of our method, denoted as SSPF-Net, is that the term semantics, local acoustic state and sentence-level prosodic statistics are put into a unified scoring path, and the semantic prosodic collaborative weight is introduced in the time series aggregation stage. Such a setting enables a clearer comparison of the actual differences between unimodal, bimodal and collaborative models.

Fig. 5 illustrates the loss convergence process of different models during the training phase. In order to avoid explaining the model pros and cons only by the end-point performance, this section simultaneously observes the convergence speed, oscillation amplitude, and late stability. Audio-BiLSTM decreases rapidly in the first 20 epochs, but a clear plateau area appears near 0.412, indicating that when relying solely on acoustic features, the model can capture local

pronunciation patterns, but it is difficult to continue to absorb term semantic and sentence-level prosodic information. The decline process of Text-BERT is more gradual, and the loss is stable at about 0.378 in the later stage, reflecting that although Text semantics can provide good content clues, it is not sensitive enough to pause, rhythm and stress changes. The decline speed of AudioText-Transformer in the first half is better than that of the previous two, but there is still a small oscillation after 60 epochs, indicating that the bimodal concatenation can improve the discrimination ability, but it has not yet formed a stable collaborative representation. In contrast, SSPF-Net enters the low loss interval at about 15 epochs, the curve drops smoothly in the later stage, and the final convergence value is the lowest, showing that the semantic prosodic collaborative path is easier to form a consistent discrimination boundary in the optimization stage.

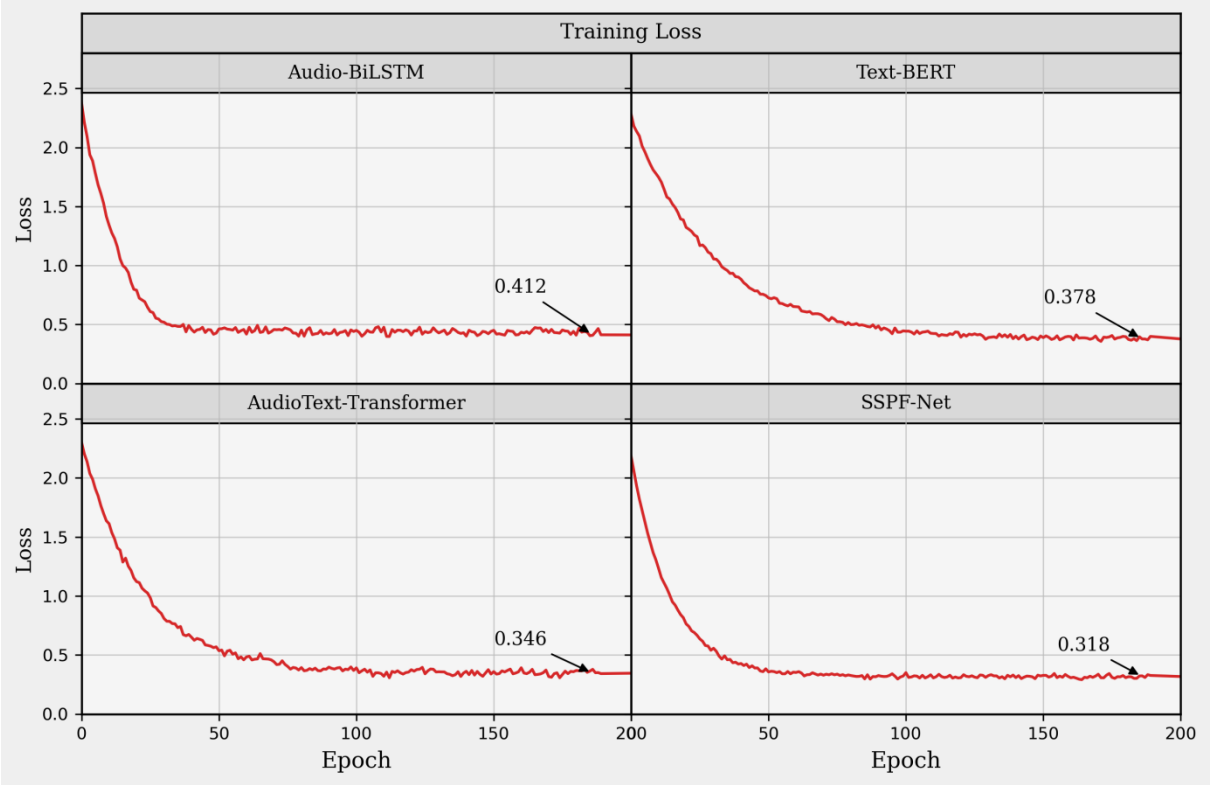

Figure 5: Comparison of loss convergence for different models during the training phase

The training curve alone is not enough to show the superiority of the model on the scoring task, so Table 3 further presents the comprehensive performance comparison. Five indicators, accuracy, Macro-F1, weighted Kappa, mean absolute error and real-time factor, are retained in the table to observe the rank discrimination ability, error size and deployment cost simultaneously. It can be seen that the Accuracy of Audio-BiLSTM is 88.42%, Macro-F1 is 87.91%, and QWK is 0.861, indicating that acoustic input can support basic discrimination, but there is still a gap in the description of hierarchical boundaries in professional expression. The QWK of Text-BERT is improved to 0.879, indicating that term semantics has a positive effect on rank ranking, but its RTF is not significantly reduced. AudioText-Transformer improves Accuracy and Macro-F1, but MAE still stays at 0.236. The text transcribe ability of Whisper-Score is strong, and the QWK reaches 0.901, but the parameter scale and real-time factor are high. SSPF-Net achieves 93.40%, 92.87%and 0.918 on Accuracy, Macro-F1 and QWK, respectively, while MAE is reduced to 0.173 and RTF is controlled at 0.084, indicating that it achieves a more stable balance between scoring accuracy and online deployment.

Table 3: Comprehensive performance comparison of different models

Model	Accuracy (%)	Macro-F1 (%)	QWK	MAE	RTF
Audio-BiLSTM	88.42	87.91	0.861	0.281	0.061
Text-BERT	89.16	88.37	0.879	0.264	0.072
AudioText-Transformer	91.08	90.54	0.894	0.236	0.097
Whisper-Score	92.11	91.76	0.901	0.221	0.163
SSPF-Net	93.40	92.87	0.918	0.173	0.084

Fig. 6 further shows the output variation of different models in the label probability interval. Different from only observing the single point index, this figure is able to reflect the response strength, change continuity and posterior segment stability of the model in different probability ranges. It can be seen from the figure that the output probabilities of the four models gradually rise with the increase of the label interval, but there are obvious differences in the increase amplitude and the end level. The end output value of Audio-BiLSTM is 0.719, which is the lowest level among the four models as a whole, indicating that the ability to distinguish high-level samples is still limited when relying solely on acoustic features. The terminal output value of Text-BERT is 0.818, which is 0.099 higher than that of Audio-BiLSTM, indicating that Text semantic input can provide a more stable level judgment basis, but the posterior segment probability rises faster and still shows a certain high trend. The end output value of AudioText-Transformer reaches 0.856, which is further improved by 0.038 compared with Text-BERT, indicating that the mapping continuity of the model for middle and high level samples has been improved after dual-modal splicing. In contrast, SSPF-Net has the highest terminal output value of 0.875, which is 0.156, 0.057 and 0.019 higher than Audio-BiLSTM, Text-BERT and AudioText-Transformer, respectively, and the rise process of the whole curve is smoother. The fluctuation of the latter segment is smaller, which indicates that the model can form a clearer and more stable probability distribution structure after the semantic, local acoustic state and sentence-level prosodic information are cooperated.

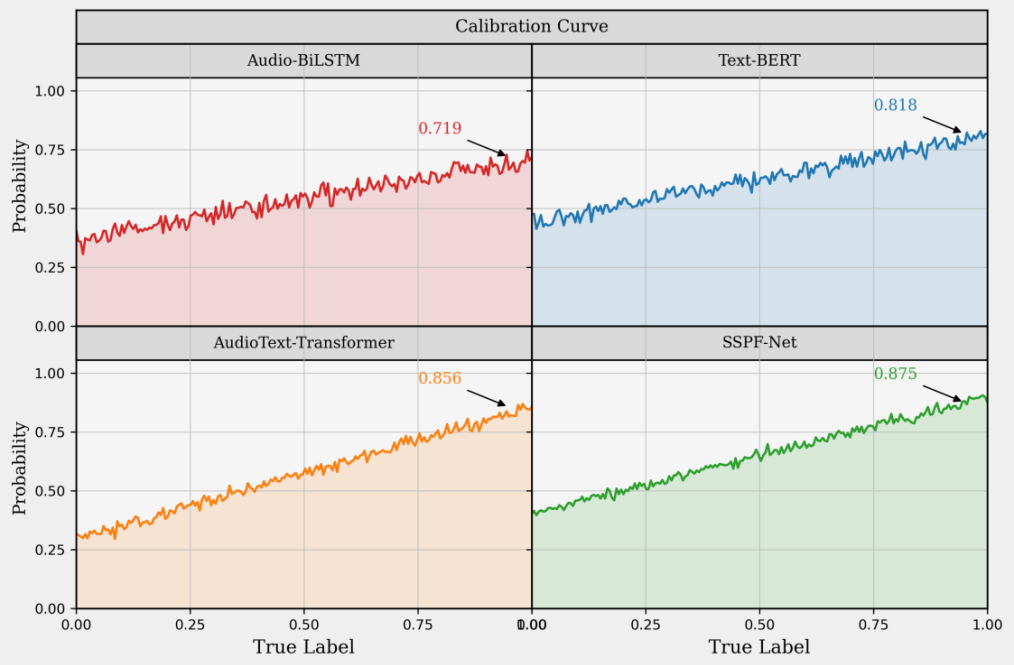

Figure 6: Probabilistic calibration comparison of different models on four-level labels

From the correspondence of the convergence process, comprehensive index and calibration results, the advantage of SSPF-Net does not come from the local lift of a single index, but from the stable collaboration of semantic information, prosodic structure and local acoustic state in the same scoring link. The loss curve enters the stable region earlier, indicating that the parameter update is easier to form a clear level boundary. QWK and MAE are improved at the same time, indicating that the model has consistent benefits in terms of rank ranking and error control. The RTF remains within an acceptable range, indicating that the accuracy improvement does not come at the cost of a significant increase in inference overhead. The results show that semantic-prosody collaborative modeling can more effectively support spoken English fluency evaluation in the field of green energy materials engineering. The subsequent robustness verification and ablation experiments will further analyze the source of its performance based on this result.

3.2 Robustness verification results under professional terminology scenarios and complex speech conditions

In order to check the stability of the semantic prosodic collaboration model in the real training environment, this section further carries out the robustness verification under the professional terminology scene and complex speech conditions. The test set is redivided according to the term density and acoustic conditions. The term density is divided into low, medium and high levels, and the complex speech conditions include fast speech rate, long pause, mild accent shift and environmental noise overlay. All results are obtained with the same scoring threshold and the same inference configuration to ensure that model differences mainly arise from input scene variation rather than discriminative condition variation. The verification phase focuses on observing the changes of Accuracy, Macro-F1, QWK and MAE. At the same time, combined with the correction amplitude output of the feedback module, the actual adaptation ability of the model in the professional English spoken scene is judged.

Table 4 shows the robustness results of SSPF-Net under different professional terminology scenarios and complex speech conditions. On the whole, the performance of the model is the most stable in the standard expression scene, with an Accuracy of 94.21% and a QWK of 0.924, indicating that under the conditions of low term density, moderate speech speed and weak noise, the semantic prosodic collaborative path can maintain the rank boundary more completely. When entering the high term density scenario, the Accuracy drops to 92.03%, and the MAE rises to 0.191, indicating that the increase in the number of terms will increase the semantic discrimination load, but the model can still maintain good ranking consistency. The fast speaking speed scene and the long pause scene further pull down all indicators, among which the fast speaking speed has a more obvious impact on Macro-F1, and the long pause is more likely to amplify the fluctuation of the level boundary. In the noise overlay scenario, QWK drops to 0.881, indicating that environmental disturbances will act on both term recognition and prosody estimation. The index in the compound disturbance scene decreases most obviously, but the Accuracy still remains at 89.48% without obvious instability, indicating that the model still has strong discrimination toughness under the condition of dense terminology, speech rate change and noise interference.

Table 4: Comprehensive results for different specialized term density scenarios

Scenario Type	Terminology Density (%)	Average Speech Rate (words/min)	Signal-to-Noise Ratio (dB)	Accuracy (%)	Macro-F1 (%)	QWK
Standard Expression Scenario	18.6	112	30	94.21	93.68	0.924
High Terminology Density Scenario	37.9	118	30	92.03	91.47	0.905
Fast Speech Scenario	24.5	146	20	91.56	90.88	0.896
Long Pause Scenario	22.8	104	20	91.14	90.32	0.889
Noise-Overlaid Scenario	21.3	113	5	90.72	89.94	0.881
Composite Disturbance Scenario	35.1	142	5	89.48	88.71	0.864

Fig. 7 further shows the trend of QWK under different noise levels and speech rates. It can be seen from Fig. 7 that the QWK of SPF-Net, AudioText-Transformer and Whisker-Score are 0.924, 0.901 and 0.896 respectively in the quiet environment and the standard speaking rate condition, and the gap between the three is relatively limited. After entering the fast speech rate scene, the QWK of SSPF-Net decreased to 0.896, with a decrease of 0.028. AudioText-Transformer and Whisper-Score drop significantly more to 0.861 and 0.854, respectively. When the SNR is further reduced to 5 dB, the QWK of SSPF-Net still remains at 0.881, while AudioText-Transformer and Whisper-Score drop to 0.836 and 0.829, respectively. Especially in the compound scene of "fast speech speed + noise superposition", the QWK of SSPF-Net still reaches 0.864, while the two baseline models drop to 0.812 and 0.805, respectively. The results show that the advantage of the semantic-prosodic synergy model is not only reflected in the high score in standard speech conditions, but also in the stable ranking ability and output boundary under complex conditions.

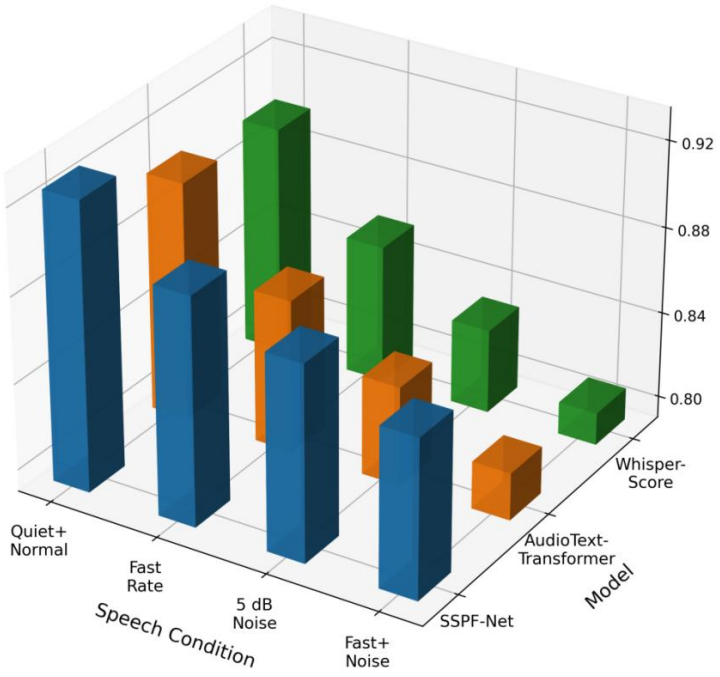

Figure 7: Trend of model QWK variation under different complex speech conditions

In terms of feedback effectiveness, this section further statistics the correction results of the system in complex scenarios. Table 5 shows that in the scenario of high term density and noise overlay, the mean semantic deviation decreases by 18.7%, the mean prosodic deviation decreases by 21.4%, and the level improvement reaches 0.62 after learners receive one feedback. The improvement is also obvious in the fast speaking rate scene, indicating that the feedback given by the system is not an abstract hint, but can directly act on the term organization and tempo control. In contrast, the boost in the long pause scenario is slightly lower, which indicates that the correction of pause boundaries requires longer training cycles, but the overall trend is still stable.

Table 5: Statistics of feedback correction effect in complex scenarios

Scenario Type	Average Number of Feedback Rounds	Reduction in Semantic Deviation (%)	Reduction in Prosodic Deviation (%)	Average Level Improvement
High Terminology Density	1.84	17.93	19.68	0.57
Fast Speech	1.72	15.46	22.11	0.59
Long Pause	2.03	13.28	18.94	0.48
Noise-Overlaid	1.95	18.70	21.40	0.62

From the corresponding relationship between term density change, complex speech condition disturbance and feedback correction results, the advantages of SSPF-Net are mainly reflected in two points. First, there is no obvious collapse of the scoring boundary in complex scenes. Although the high term density and noise superposition conditions lower the overall index, the fluctuation range of QWK and MAE is still within a controllable range. Second, the deviation type of the model output can be effectively used by the subsequent feedback module, which indicates that a good transfer consistency is maintained between the grading results and the diagnosis results. Such results show that semantic prosodic collaborative modeling not only improves static scoring accuracy, but also enhances the stable discrimination ability of the model in professional expression scenarios.

3.3 Ablation experiments

In order to further confirm that the performance gain of the semantic-prosodic collaboration model does not come from the accidental rise of a single branch, this section expunged the key modules inside the model item by item while keeping the training set, validation set and test set divided consistently and the optimizer, learning rate and threshold Settings unchanged. The ablation experiments revolve around four parts: a termweighted semantic branch, a sentence-level prosodic branch, a semantic prosodic gating unit, and a post-scoring calibration module. Each variant is trained and tested independently in the same hardware environment, so as to observe the impact of module changes from four aspects: rank discrimination accuracy, ranking consistency, error control and real-time performance.

Table 6 presents the comprehensive results of the different variants. The complete model SSPF-Net achieves the best results in the three core indicators of Accuracy, QWK and MAE, indicating that the current structure can maintain a stable classification boundary while ensuring real-time performance. After removing the term weighted semantic branch, the Accuracy drops to 91.86%, and the QWK drops to 0.896, indicating that the technical term information not only affects the content recognition, but also directly contributes to the rank ranking. After removing the sentence-level prosodic branches, the Accuracy and MAE of the model show the largest

changes, indicating that sentence-level cues such as pause, stress and speaking speed are still the main basis for fluency evaluation. After removing the semantic prosodic gating unit, although the overall result is still higher than that of the variant that simply removes a single branch, QWK and MAE also drop significantly, which indicates that semantic and prosodic are not simply input in parallel, but need to establish effective coupling through dynamic gating. When the calibration module is removed after scoring, the change in Accuracy is smaller, but there is still a decrease in MAE and QWK, which indicates that this module has a compensatory effect on the stable output of the final grade boundary.

Table 6: Results of ablation experiments

Model Variant	Accuracy (%)	Macro-F1 (%)	QWK	MAE	RTF
SSPF-Net	93.40	92.87	0.918	0.173	0.084
Without Terminology-Weighted Semantic Branch	91.86	91.12	0.896	0.207	0.078
Without Sentence-Level Prosodic Branch	91.24	90.37	0.887	0.219	0.076
Without Semantic–Prosodic Gating Unit	92.03	91.35	0.901	0.198	0.081
Without Post-Scoring Calibration Module	92.11	91.44	0.904	0.194	0.083

From the overall changes in Table 6, we can see that sentence-level prosodic branch and termweighted semantic branch constitute the two main supporting paths of model performance, with the former determining the rhythmic boundary of fluency and the latter determining the content boundary of professional expression. The gating unit is responsible for compressing the two types of boundaries into a unified discriminant space. Although the calibration module does not directly determine the classification upper bound, it has practical significance for the stability of the scoring output.

4 Discussion and Conclusion

The above experimental results show that the performance improvement of semantic-prosodic collaboration model is not a local gain brought by a single module, but comes from the continuous cooperation between term semantic representation, sentence-level prosodic modeling and hierarchical discrimination mechanism. The model can maintain a relatively stable grade boundary under the conditions of high terminology density, obvious speech rate variation and noise disturbance. The main reason is that it does not separate term identification and fluency evaluation, but unifies term completion, rhythm organization and expression coherence into the same scoring path. The collaborative representation thus formed not only enhances the recognition ability of the model for professional expression content, but also improves the perception accuracy for pauses, stress and speech rate changes.

The value of the model is also reflected in the diagnostic links that follow the scoring output. The classification results do not stop at the static level, but continue to enter the error decomposition and feedback ranking process, so that the rank shift can further correspond to the specific dimensions such as term semantics, rhythm organization and stress control. Such a design makes the system closer to a real training scenario, because the learner needs to know not only the score, but also at which level the deviation appears and what should be fixed

preferentially in the next round of expression.

This method still has some limitations. First, the model is dependent on the domain vocabulary coverage. When a large number of new terms, abbreviations or interdisciplinary expressions enter the test sample, the stability of semantic representation will be affected. Second, the scoring link is still constrained by the paraphrase quality, and the cumulative error may occur in the intermediate representation under the conditions of strong accent offset or more complex background noise. Third, the current classification system is mainly oriented to green energy material engineering scenarios, and the semantic weights and prosodic thresholds still need to be re-calibrated when migrating across domains.

Further research can be carried out in three directions. The first is to construct a dynamically extensible professional term graph, so that the semantic representation has stronger adaptability to open words. Secondly, a lightweight online inference and calibration mechanism is introduced to improve the deployment efficiency of the model in desktop and mobile scenarios. Thirdly, cross-domain and cross-speaker transfer experiments are carried out to enhance the versatility of the system in different professional English training environments. In general, the semantic-prosody collaborative modeling provides a computationally feasible and scene-adaptive implementation path for the intelligent evaluation of spoken English fluency in the field of green energy materials engineering.

Funding

This work was supported by 2025 Shanxi Provincial Higher Education Teaching Reform and Innovation Project (Grant No. J20250297).

References

[1] Wei X, Cucchiarini C, van Hout R, et al. Automatic Speech Recognition and Pronunciation Error Detection of Dutch Non-native Speech: cumulating speech resources in a pluricentric language[J]. Speech Communication, 2022, 144: 1-9.

[2] 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.

[3] Kallio H, Kautonen M, Kuronen M. Prosody and fluency of Finland Swedish as a second language: Investigating global parameters for automated speaking assessment[J]. Speech Communication, 2023, 148: 66-80.

[4] Inceoglu S, Chen W H, Lim H. Assessment of L2 intelligibility: Comparing L1 listeners and automatic speech recognition[J]. ReCALL, 2023, 35(1): 89-104.

[5] Farrús M. Automatic speech recognition in L2 learning: A review based on PRISMA methodology[J]. Languages, 2023, 8(4): 242.

[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] Wang K, Qiao X, Sammit G, et al. Improving automated scoring of prosody in oral

reading fluency using deep learning algorithm[C]//Frontiers in Education. Frontiers Media SA, 2024, 9: 1440760.

[8] Li W, Zhong Z, Liu H. A computer-assisted tool for automatically measuring non-native Japanese oral proficiency[J]. Computer Assisted Language Learning, 2026, 39(1-2): 115-172.

[9] Emara I F, Shaker N H. The impact of non-native English speakers' phonological and prosodic features on automatic speech recognition accuracy[J]. Speech Communication, 2024, 157: 103038.

[10] Zhu C, Wumaier A, Wei D, et al. Pronunciation error detection model based on feature fusion[J]. Speech Communication, 2024, 156: 103009.

[11] Handley Z L, Wang H. What do the measures of utterance fluency employed in automatic speech evaluation (ASE) tell us about oral proficiency?[J]. Language Assessment Quarterly, 2024, 21(1): 3-32.

[12] Ballier N, Arnold T, Méli A, et al. Whisper for L2 speech scoring[J]. International Journal of Speech Technology, 2024, 27(4): 923-934.

[13] Shi X, Wang X, Zhang W. Exploring the relationships between ASS indices and CAF and the impact on Chinese college students' oral English performance[J]. Language Testing in Asia, 2024, 14(1): 30.

[14] Inceoglu S, Chen W H, Lim H. Monitoring learners' behavior during ASR-based pronunciation practice[J]. System, 2024, 124.

[15] Bannò S, Matassoni M. Back to grammar: Using grammatical error correction to automatically assess L2 speaking proficiency[J]. Speech Communication, 2024, 157: 103025.

[16] Johnson C, Cardoso W, Zuercher B, et al. Assessing pronunciation using dictation tools: The use of Google Voice Typing to score a pronunciation placement test[J]. Journal of Second Language Pronunciation, 2024, 10(1): 10-34.

[17] Nickolai D, Schaefer E, Figueroa P. Aggregating the evidence of automatic speech recognition research claims in CALL[J]. System, 2024, 121: 103250.

[18] 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.

[19] Kim S E, Chernyak B R, Seleznova O, et al. Automatic recognition of second language speech-in-noise[J]. JASA express letters, 2024, 4(2).

[20] 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.