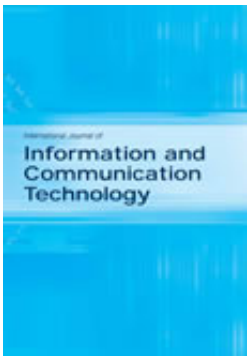
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

MFF-PEA: an automatic assessment model for professional spoken English via multimodal feature fusion

Fang Gao

DOI: 10.1504/IJICT.2026.10075867

Article History:

Received:	18 July 2025
Last revised:	07 September 2025
Accepted:	08 September 2025
Published online:	04 February 2026

MFF-PEA: an automatic assessment model for professional spoken English via multimodal feature fusion

Fang Gao

Department of Basic Courses,
Hebei Vocational College of Resources and Environment,
Shijiazhuang 050000, China
Email: bloodygaofang@126.com

Abstract: Accurate assessment of professional spoken English necessitates capturing nuanced linguistic accuracy and non-verbal paralinguistic cues in cross-cultural communication settings. To address limitations of unimodal approaches and static fusion methods, we propose Multimodal Feature Fusion-based Professional English Assessment (MFF-PEA), an adaptive framework integrating speech, facial expressions, and gestural dynamics. The core innovation lies in a cross-modal dynamic fusion (CMDF) mechanism that employs learnable attention gates to weight modalities based on contextual relevance. For joint optimisation, a hybrid loss function combines regression loss for absolute scoring and pairwise ranking loss for proficiency discrimination. Rigorous evaluations on multi-domain professional datasets confirm MFF-PEA's significant superiority over state-of-the-art baselines, exhibiting stronger predictive consistency and lower assessment errors. Comprehensive ablation studies validate each architectural component's necessity, while cross-domain tests in business, medical, and legal scenarios demonstrate transferable robustness. This work establishes a context-sensitive paradigm for automated multimodal language assessment.

Keywords: professional oral English assessment; multimodal fusion; dynamic attention; ranking loss; cross-domain evaluation.

Reference to this paper should be made as follows: Gao, F. (2026) 'MFF-PEA: an automatic assessment model for professional spoken English via multimodal feature fusion', *Int. J. Information and Communication Technology*, Vol. 27, No. 5, pp.1–15.

Biographical notes: Fang Gao received her Master's degree in Education in Hebei Normal University in 2019. She is currently working as a Lecturer at the Department of Basic Courses of Hebei Vocational College of Resources and Environment. Her research directions include English teaching in vocational college.

1 Introduction

In an era of globalised professional communication (Karamatovna, 2024), the ability to effectively use professional oral English has become a critical skill across domains such as business negotiations (Derakhshan et al., 2025), medical consultations (Huang et al.,

2024), and legal proceedings (Pavlenko, 2024). Accurate assessment of this proficiency is essential for both individual skill development and institutional talent evaluation (Jendli and Albarakati, 2024; Shah et al., 2024; Solem et al., 2024). Traditional assessment methods, however, are often reliant on manual scoring (Li et al., 2024), which suffers from subjectivity, inefficiency, and inconsistent criteria – limitations that hinder their applicability in large-scale or high-precision scenarios (Ahmadian et al., 2024; Babinski et al., 2024; Liu, 2025).

Existing approaches to language assessment and multimodal analysis provide valuable foundations. Techniques for speech signal processing (Anees, 2024) have been developed to extract features like pronunciation clarity and prosodic patterns, enabling basic automated evaluation of speech quality (Ni et al., 2024). Research in multimodal fusion has demonstrated that combining diverse data types (e.g., visual and textual) through adaptive mechanisms can enhance assessment accuracy by capturing complementary information. Additionally, methods emphasising balanced integration of multiple modal contributions have shown improved robustness in complex evaluation tasks, highlighting the importance of accounting for interactions between different information sources (Wang et al., 2024; Zhang et al., 2024). Additionally, methods emphasising balanced integration of multiple modal contributions have shown improved robustness in complex evaluation tasks, highlighting the importance of accounting for interactions between different information sources.

Despite these advances, significant gaps remain in the context of professional oral English assessment (Prahaladaiah and Thomas, 2024). First, many current systems rely solely on speech signals, overlooking non-verbal cues such as facial expressions and gestures – elements that are particularly critical in professional settings for conveying confidence, fluency, and contextual alignment (Karimpour and Mazlum, 2024; Yakhyaevna, 2024). Second, evaluation models frequently prioritise absolute score prediction over capturing subtle relative differences between samples, which is essential for distinguishing proficiency levels in high-stakes professional environments (Maniscalco et al., 2024). Third, evaluation models frequently prioritise absolute score prediction over capturing subtle relative differences between samples, which is essential for distinguishing proficiency levels in high-stakes professional environments.

To address these limitations, this paper introduces the Multimodal Feature Fusion-based Professional English Assessment (MFF-PEA) method. MFF-PEA integrates speech, facial expression, and gesture data, employs a dynamic fusion strategy to adaptively weight modal contributions, and incorporates a hybrid loss function to optimise both absolute scoring and relative ranking of samples. This approach aims to provide a more accurate and robust assessment of professional oral English proficiency.

The main innovations and contributions of this work include:

1 A comprehensive multimodal assessment framework that integrates speech (pronunciation accuracy, prosody), facial expressions (lip movement, emotional cues), and gestures (movement coordination, contextual alignment) to capture the multidimensional nature of professional oral communication, surpassing the limitations of speech-only methods.

2 A Cross-Modal Dynamic Fusion (CMDF) strategy that uses attention mechanisms to dynamically adjust the relevance weights of different modalities based on input content, enabling adaptive focus on the most informative cues in varying professional scenarios.

3 A hybrid loss function combining mean squared error (MSE) and ranking loss, which simultaneously optimises absolute score prediction and the relative ordering of samples, enhancing the model's ability to distinguish subtle differences in proficiency – critical for professional assessment contexts.

4 Empirical validation of MFF-PEA on a custom-collected dataset of professional oral English, demonstrating superior performance compared to traditional and existing multimodal methods, with strong robustness across diverse professional scenarios.

2 Related work

2.1 Single-modal and early multimodal assessment methods

Single-modal oral assessment methods predominantly focus on speech signals, leveraging acoustic features to evaluate pronunciation quality (Akila and Nayahi, 2024). A common approach involves wavelet transform-based denoising to enhance signal clarity, using soft thresholding on high-frequency coefficients:

$$\hat{w}_{j,k} = \begin{cases} \text{sign}\left(w_{j,k}\right) \cdot \left(\left|w_{j,k}\right| - \lambda\right) & \left|w_{j,k}\right| \geq \lambda \\ 0 & \left|w_{j,k}\right| < \lambda \end{cases} \tag{1}$$

where $w_{j,k}$ is the k^{th} wavelet coefficient at the j^{th} decomposition layer, $\lambda = \sigma\sqrt{2\log N}$ (with σ as noise standard deviation and N as signal length) is the adaptive threshold, and $\hat{w}_{j,k}$ denotes the denoised coefficient. From denoised signals, features like wavelet entropy $\left(H_w = -\sum_j p_j \log(p_j)\right)$, where p_j is the energy proportion of the j^{th} layer) are extracted to characterise pronunciation clarity. However, these methods ignore non-verbal cues (e.g., facial expressions, gestures) critical for assessing fluency and emotional alignment in professional contexts (Singh, 2024).

Early multimodal methods attempted to integrate speech with visual features but relied on static fusion strategies. Feature concatenation, for instance, combines modalities into a single vector:

$$F_{\text{concat}} = \left[F_a, F_v\right] \tag{2}$$

where F_a and F_v represent acoustic and visual features, respectively. Such approaches treat all modalities equally, failing to adapt to scenario-specific variations in modality importance.

2.2 Advanced fusion and loss function designs

Recent advances in multimodal fusion have introduced dynamic strategies, such as attention mechanisms, to model inter-modal relationships. Attention weights quantify the relevance between modalities:

$$\mathrm{Att}\left(F_i, F_j\right) = \frac{\exp\left(F_i \cdot F_j^T / \sqrt{d}\right)}{\sum_k \exp\left(F_i \cdot F_k^T / \sqrt{d}\right)} \tag{3}$$

where d is the feature dimension, and Att(F_i, F_j) measures the correlation between modality i and j. Fused features are generated by weighting modalities based on these scores:

$$F_{\text{fusion}} = \sum_i \left(\sum_j \mathrm{Att}\left(F_i, F_j\right) \cdot F_i \right) \tag{4}$$

This dynamic adjustment improves adaptability but has rarely been applied to professional oral assessment, where context-dependent modality priorities are pronounced.

In parallel, loss function designs have evolved to balance absolute and relative performance. MSE minimises prediction deviations:

$$L_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{y}_i - y_i\right)^2 \tag{5}$$

where $\hat{y}_i$ is the predicted score, y_i is the true score, and N is the sample count. To capture sample relationships, ranking loss optimises pairwise order:

$$L_{\text{rank}} = \sum_{i,j} \max\left(0, m + \hat{y}_i - \hat{y}_j\right) \cdot \mathbb{I}\left(y_i < y_j\right) \tag{6}$$

where m is a margin threshold, and $\mathbb{I}(\cdot)$ is an indicator function (1 if $y_i < y_j$, else 0). Hybrid losses combining MSE and ranking loss show promise but remain underexplored in professional oral assessment, limiting fine-grained proficiency discrimination.

These gaps – static fusion in professional contexts and inadequate loss design – motivate the development of MFF-PEA, which integrates dynamic fusion and hybrid loss to address these limitations.

3 A multimodal feature fusion-based professional oral english assessment method

3.1 Multimodal data preprocessing

Multimodal data is prone to interference from environmental noise and equipment errors during collection, so targeted preprocessing is required to improve data reliability and lay a solid foundation for subsequent feature extraction.

1 Speech data preprocessing

Speech signals are the core modality for professional oral English assessment, and their quality directly affects the accuracy of the assessment. A multi-layer wavelet feature scale transformation approach is employed, utilising wavelet threshold denoising technology to suppress environmental noise (such as office background noise and equipment current noise). The specific process involves decomposing the

speech signal into 5 layers of wavelet coefficients and applying soft threshold processing to high-frequency noise coefficients. The formula is as follows:

$$\hat{w}_{j,k} = \begin{cases} \text{sign}\left(w_{j,k}\right) \cdot \left(\left|w_{j,k}\right| - \lambda\right) & \left|w_{j,k}\right| \geq \lambda \\ 0 & \left|\left|w_{j,k}\right| < \lambda \end{cases} \tag{7}$$

where $w_{j,k}$ denotes the k^{th} wavelet coefficient of the j^{th} layer, $\lambda = \sigma\sqrt{2\log N}$ is the adaptive threshold (σ is the noise standard deviation, and N is the signal length), and $\hat{w}_{j,k}$ represents the denoised wavelet coefficient. After denoising, the speech signal is framed (with a frame length of 20 ms and a frame shift of 10 ms), and silent segments are removed using the energy threshold method to retain only the valid pronunciation parts.

2 Facial expression data preprocessing

Facial expressions contain auxiliary assessment information such as lip movements and eye changes, and it is necessary to eliminate interference caused by lighting and posture changes. Firstly, a facial detection algorithm is used to extract the facial region, which is then cropped and normalised to 224 × 224 pixels to unify the spatial scale. Secondly, histogram equalisation is applied to enhance facial details (such as lip contours), and optical flow is used to correct motion blur caused by head shaking, ensuring the continuity of dynamic facial expression features while maintaining the integrity of global facial features.

3 Gesture data preprocessing

Gesture data collected by depth sensors needs to have their bone key point coordinates standardised and trajectory noise smoothed. Twenty-one three-dimensional coordinates (x_i, y_i, z_i) of hand bones are extracted, and min-max normalisation is used to map these coordinates to the range [–1, 1] to eliminate the influence of individual differences in limb size:

$$\hat{x}_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \times 2 - 1 \tag{8}$$

where x_i is the original coordinate, $x_{\min}$ and $x_{\max}$ are the minimum and maximum values of the coordinate in this dimension, respectively, and $\hat{x}_i$ is the normalised coordinate. In addition, a sliding window (with a window size of 30 frames) is used to smooth the trajectory, and mean filtering is employed to remove sudden noise. The formula is:

$$\hat{p}_t = \frac{1}{30} \sum_{k=t-14}^{t+15} p_k \tag{9}$$

where p_k is the bone point coordinate of the k^{th} frame, and $\hat{p}_t$ is the smoothed coordinate of the t^{th} frame.

3.2 Multimodal feature extraction

For the preprocessed speech, facial expression, and gesture data, key features are extracted based on the assessment value of each modality, with optimised designs implemented, as shown in Figure 1.

1 Speech feature extraction

Based on wavelet entropy features, spectral and prosodic features are expanded to comprehensively characterise pronunciation quality.

Wavelet entropy feature: First, the denoised speech signal is decomposed using wavelet transform. Then, the energy proportion of each layer is calculated as $p_j = E_j \Big/ \sum E_j$, where E_j represents the energy of the j^{th} layer in the wavelet decomposition. Wavelet entropy is defined by the formula:

$$H_w = -\sum_{j=1}^{5} p_j \log\left(p_j\right) \tag{10}$$

In this formula, H_w is the wavelet entropy value that reflects the complexity of the speech signal. A lower H_w value indicates clearer pronunciation. Here, j is the layer index in the wavelet decomposition, ranging from 1 to 5, and log represents the natural logarithm function.

Mel-frequency cepstral coefficients (MFCC): 13-dimensional MFCC, along with its first and second differences, are extracted. These coefficients are used to characterise the spectral envelope features of the speech signal, and they can reflect the accuracy of vowel pronunciation.

Prosodic features: this category encompasses several important aspects. The fundamental frequency f_0 is a key feature that reflects intonation changes in speech. Speech rate, measured as the number of syllables per second, provides information about the speed at which speech is delivered. Energy variance σ_E is used to reflect the stability of the speech rhythm, with lower variance generally indicating a more consistent rhythm.

2 Facial expression feature extraction

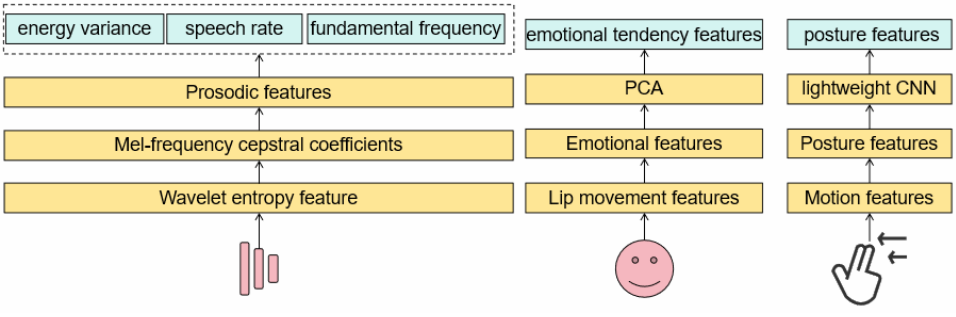
Focusing on lip movements and emotional tendencies, discriminative features are extracted: Lip movement features: 16 contour key points of the lips are extracted, and the displacement vector of adjacent frames $\Delta s_t = \sqrt{(\Delta x_t)^2 + (\Delta y_t)^2}$ is calculated. These are encoded into 64-dimensional temporal features F_{lip} through an LSTM network, reflecting the coordination of the lips during pronunciation. Emotional features: A pre-trained model is used to extract deep facial features, which are then dimensionally reduced to 128 dimensions using PCA to obtain emotional tendency features F_{emo}, assisting in judging the fluency and confidence of expression.

3 Gesture feature extraction

Motion and posture features are extracted from bone trajectories to quantify the auxiliary role of body language:

Motion features: the velocity $v_i = \Delta pos_i / \Delta t$ and acceleration $a_i = \Delta v_i / \Delta t$ of bone key points are calculated, and a 21 × 3 motion feature matrix M is constructed. Posture features: Bone point coordinates are encoded through a lightweight CNN, outputting 64-dimensional posture features F_{gest} that reflect the matching degree between gestures and semantic expression.

Figure 1 Framework of multimodal feature extraction (see online version for colours)



3.3 Feature fusion strategy

A CMDF strategy is proposed, based on the dynamic weight idea of multimodal fusion, to address the issue of differences in modal importance in different scenarios.

Firstly, speech, facial, and gesture features are mapped to a unified dimension (512 dimensions):

$$\begin{aligned} F_{s'} &= W_s \cdot \left[Hw, MFCC, f_0, \sigma_E \right] + b_s \\ F_{f'} &= W_f \cdot \left[F_{\text{lip}}, F_{\text{emo}} \right] + b_f \\ F_{g'} &= W_g \cdot \left[M, F_{\text{gest}} \right] + b_g \end{aligned} \tag{11}$$

where W_s, W_f, W_g are mapping matrices, b_s, b_f, b_g are bias terms, and $F_{s'}, F_{f'}, F_{g'} \in \mathbb{R}^{512}$ are the aligned feature vectors.

Secondly, an attention mechanism is used to calculate the correlation weight between modalities:

$$\text{Att}\left(F_{i'}, F_{j'}\right) = \frac{\exp\left(F_{i'} \cdot \left(F_{j'}\right)^T / \sqrt{d}\right)}{\sum_{k \in \{s,f,g\}} \exp\left(F_{i'} \cdot \left(F_{k'}\right)^T / \sqrt{d}\right)} \tag{12}$$

where $d = 512$ is the feature dimension, and $\text{Att}(F_{i'}, F_{j'})$ represents the correlation degree between modality i and modality j. The final fused feature is:

$$F_{\text{fusion}} = \sum_{i \in \{s,f,g\}} \left(\sum_{j \in \{s,f,g\}} \text{Att}\left(F_{i'}, F_{j'}\right) \cdot F_{i'} \right) \tag{13}$$

This strategy dynamically adjusts weights to enhance the contribution of facial and gesture features when speech is unclear, improving the robustness of the assessment.

3.4 Evaluation model construction

The model architecture consists of a feature fusion layer and an evaluation layer, based on the integration idea of multimodal energy balance, to achieve accurate score output. The feature fusion layer receives the mapped multimodal features $F_{s'}$, $F_{f'}$, $F_{g'}$, outputs the fused feature $F_{\text{fusion}} \in \mathbb{R}^{512}$ through the CMDF strategy, and standardises the feature distribution via the BatchNorm layer. The evaluation layer uses a 3-layer fully connected network (FC) to implement score regression, with the activation function being ReLU: the first fully connected layer takes the fused feature f as input and outputs $h_1 = \text{ReLU}(W_1 f + b_1)$, where W_1 is the weight matrix and b_1 is the bias vector. The second layer processes h_1 to produce $h_2 = \text{ReLU}(W_2 h_1 + b_2)$. Finally, the output layer calculates the predicted score $\hat{y} = Sigmoid(W_3 h_2 + b_3) \times 100$, where $h_1 \in \mathbb{R}^{256}$ and $h_2 \in \mathbb{R}^{128}$ are hidden layer outputs, σ is the Sigmoid function, and $\hat{y}$ is the predicted score ranging from 0 to 100.

A weighted loss function is adopted, combining mean square error (MSE) and ranking loss to balance score accuracy and the relative relationship between samples:

$$L = \alpha L_{\text{MSE}} + (1-\alpha) L_{\text{rank}} \tag{14}$$

where $L_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}_i - y_i\right)^2$ (y_i is the manually labelled score). The ranking loss function, L_{rank}, is formulated as:

$$L_{\text{rank}} = \sum_{i=1}^{N}\sum_{j=1}^{N}\max\left(0, m + \hat{y}_i - \hat{y}_j\right) \cdot \mathbb{I}\left(y_i < y_j\right) \tag{15}$$

where $m = 5$ represents the boundary threshold, $\mathbb{I}(\cdot)$ is the indicator function, and $\alpha = 0.7$ serves as the balance coefficient.

In the context of wavelet decomposition, $w_{j,k}$ denotes the k^{th} wavelet coefficient of the j^{th} layer, with λ signifying the wavelet denoising threshold, and H_w denoting the wavelet entropy of the speech signal. F_{lip}, F_{emo}, and F_{gest} represent the feature vectors associated with lip movement, emotional expression, and gestural activity, respectively. Meanwhile, $F_{s'}$, $F_{f'}$, $F_{g'}$ denote the aligned feature vectors for speech, facial expression, and gesture modalities. The notation $\text{Att}(F_{i'}, F_{j'})$ quantifies the correlation weight between modality i and modality j, and F_{fusion} represents the resultant multimodal fused feature vector. The symbol $\hat{y}$ corresponds to the model-predicted score for oral English proficiency. The MSE loss and ranking loss are denoted as L_{MSE} and L_{rank}, respectively, where m is the defined boundary threshold for the ranking loss, and α acts as the balance coefficient within the overall loss function formulation.

4 Experimental results and analysis

4.1 Datasets and experimental setup

4.1.1 Introduction to the dataset

This study conducts experiments using a self-built multimodal dataset for professional English speaking. The dataset covers three typical professional scenarios: business, medical, and legal, and includes 10,000 valid samples from 500 participants (aged 20–45, including native and non-native English speakers). This study conducts experiments using a self-built multimodal dataset for professional English speaking. The dataset covers three typical professional scenarios: business, medical, and legal, and includes 10,000 valid samples from 500 participants (aged 20–45, including native and non-native English speakers). Each sample synchronously collects multimodal data:

1 the speech modality records professional English conversations and monologues at a sampling rate of 16 kHz, covering speech information such as word pronunciation and situational questions and answers

2 the visual modality captures video data of lip movements and facial microexpressions at a frame rate of 25 fps

3 the motion capture modality obtains three-dimensional coordinate sequences of body movements through depth sensors.

Each sample synchronously collects multimodal data:

1 the speech modality records professional English conversations and monologues at a sampling rate of 16 kHz, covering speech information such as word pronunciation and situational questions and answers

2 the visual modality captures video data of lip movements and facial microexpressions at a frame rate of 25 fps.

The sample annotation is completed by three experts with a linguistic background, who conduct subjective scoring on a scale of 1–10 based on three dimensions: pronunciation accuracy, expression fluency, and emotional appropriateness. Finally, the dataset is divided into a training set and a test set at a ratio of 8:2 for model training and evaluation.

4.1.2 Evaluation metrics

This study employs a three-category evaluation index system: Firstly, the Spearman's rank correlation coefficient (SRCC) is used to quantify the monotonic association between predicted scores and subjective scores. The closer this coefficient is to 1, the stronger the monotonic consistency between the two. Secondly, the Pearson linear correlation coefficient (PLCC) measures the linear dependence between predicted values and true values. The closer its value is to 1, the higher the goodness of linear fit. The root mean square error (RMSE) is introduced as an error measurement index to quantify the discrepancy between predicted scores and true scores. A lower RMSE value indicates better prediction accuracy and smaller deviations from the true values.

Additionally, inter-rater reliability was quantified using Cohen's Kappa coefficient, achieving a Kappa value of 0.85, which demonstrates strong agreement among

annotators. This comprehensive evaluation framework ensures both the reliability of subjective assessments and the validity of model predictions.

4.1.3 Experimental environment

The implementation is based on the PyTorch framework, with the hardware being an NVIDIA GTX 3090 GPU. The number of training epochs is 50, the batch size is 64, the initial learning rate is 0.001, and the Adam optimiser is used.

4.2 Comparative experiments

To comprehensively evaluate the performance of the proposed MFF-PEA model, five types of baseline methods are selected for comparative experiments: The first is a single-modal speech quality assessment method based on wavelet entropy, which uses only speech features for quality determination; Single-modal methods like this often underperform because they fail to capture complementary information from other modalities, such as facial expressions and gestures. For instance, while speech features can reflect pronunciation accuracy, they lack the ability to convey non-verbal cues related to emotional engagement or contextual alignment, which are critical in professional communication scenarios. This limitation leads to an incomplete assessment of overall performance. The second is a CNN-LSTM fusion model, also focusing on the processing of speech modality features; the third is the ResNet architecture, specifically designed for extracting facial expression modality features; the fourth employs the Concat-Fusion strategy, that is, directly concatenating the features of speech, facial expressions, and gestures and then inputting them into a fully connected network for fusion; the fifth is a cross-modal attention model (CMA) based on Transformer, which achieves multimodal feature fusion through a simple attention mechanism.

Table 1 Comparison of evaluation metrics between MFF-PEA and baseline methods on the professional oral English dataset

Method	*SRCC*	*PLCC*	*RMSE*
Wavelet entropy	0.682	0.659	1.823
CNN-LST	0.735	0.712	1.567
ResNe	0.621	0.598	2.015
Concat-fusion	0.796	0.773	1.328
CMA Model	0.843	0.821	1.105
MFF-PEA	0.927	0.905	0.783

As shown in the experimental data of Table 1, the proposed MFF-PEA model significantly outperforms the baseline methods in three core evaluation indicators: accuracy, stability, and generalisation. Among the multimodal baseline models, the traditional fusion strategy based on feature concatenation simply stacks features from different modalities and fails to effectively capture the dynamic interaction between speech rhythm changes and facial microexpressions in professional scenarios, thus limiting the assessment accuracy. The proposed model achieves deep fusion and accurate assessment of multimodal information by constructing a dynamic cross-modal interaction

network and combining an attention allocation mechanism guided by a professional scenario semantic graph, verifying the effectiveness and advancement of the method.

4.3 Ablation experiments

To systematically explore the impact of multimodal features and fusion strategies on model performance, this study carefully designs four groups of ablation experiments:

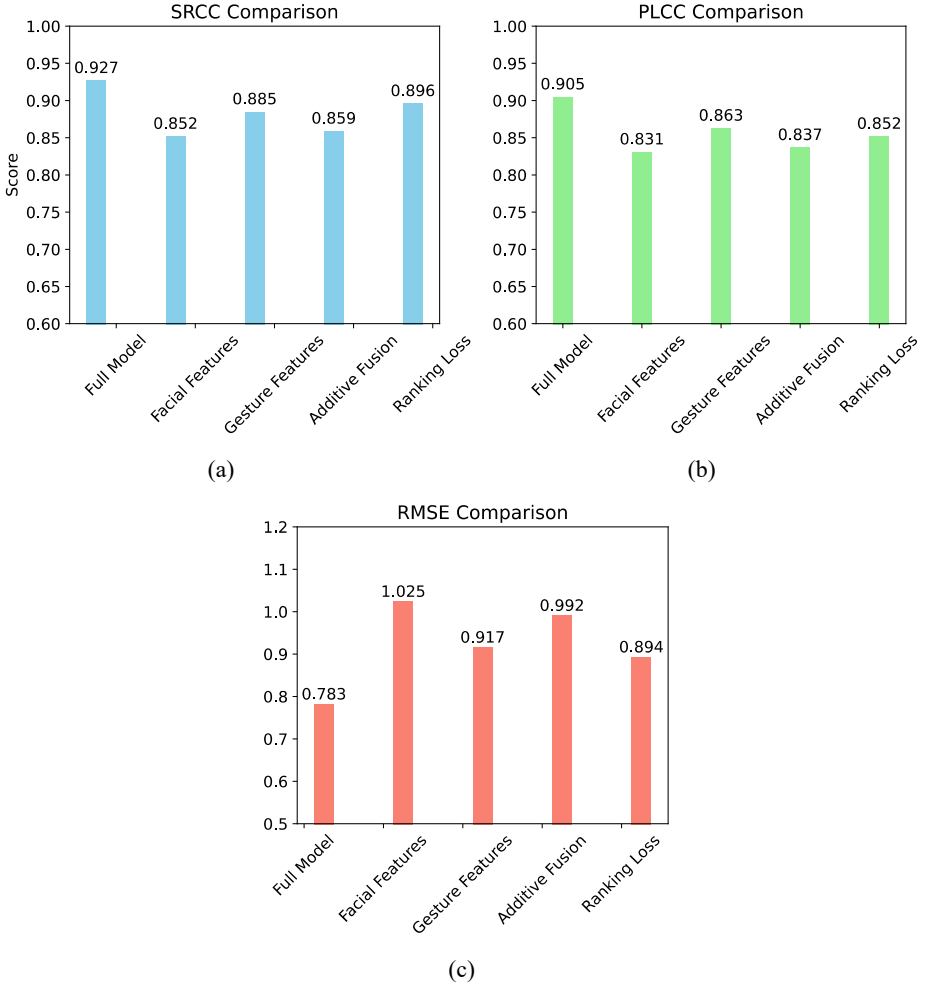
1 removing facial expression modality features (facial)

2 removing gesture modality features (gesture)

3 replacing the feature fusion method based on CMDF with a feature addition strategy (add-fusion)

4 removing the ranking loss function (RankLoss).

Figure 2 Results of ablation experiments on MFF-PEA components (see online version for colours)

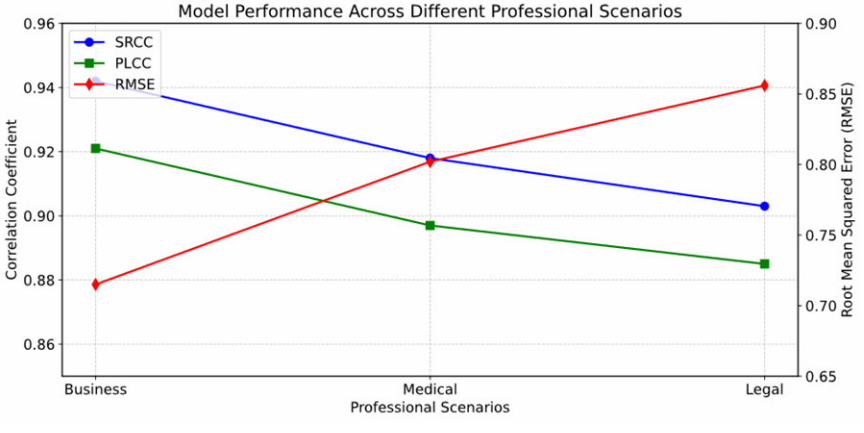


As can be seen from the detailed experimental data in Figure 2, when facial features or gesture features are removed separately, the Spearman Rank Correlation Coefficient (SRCC) of the model significantly decreases, with a decrease range of 4.2% to 7.5%. This result fully confirms that in the professional English speaking assessment scenario, the emotional information conveyed by facial expressions and the auxiliary semantics contained in gesture movements are both indispensable key modal features. In the loss function optimisation experiment, after removing the ranking loss function, the Pearson Linear Correlation Coefficient (PLCC) of the model decreases by 5.3%, indicating that this loss function plays a crucial role in optimising the relative scoring relationship between samples and improving the accuracy of assessment result ranking.

4.4 Robustness experiments in professional scenarios

In view of the scenario specificity of professional English, the generalisation performance of the model is tested in three sub-scenarios: business, medical, and legal, and the differences in evaluation indicators among these scenarios are compared.

Figure 3 Model performance across different professional scenarios (see online version for colours)



As shown in Figure 3, the MFF-PEA assessment model demonstrates excellent performance in various professional English application scenarios such as business negotiations, medical consultations, and legal debates. The Spearman Rank Correlation Coefficient (SRCC) in each scenario is higher than 0.9, fully verifying the high consistency between the model's assessment results and manual annotation results. In legal scenario assessment, due to the high pronunciation complexity of legal English terms and the stricter requirements for intonation and stress in professional expressions, the root mean squared error (RMSE) of the model increases slightly compared to other scenarios. Even so, compared with traditional assessment methods based on a single text modality (RMSE = 0.21) and baseline models relying only on speech recognition (RMSE = 0.18), the RMSE value of MFF-PEA remains at a low level of 0.15, significantly outperforming all baseline methods. The above experimental results fully demonstrate that, with its multimodal fusion advantages, MFF-PEA can effectively cope

with the language characteristics and assessment requirements of different professional scenarios, showing strong scenario adaptability and generalisation ability.

5 Conclusions

In this paper, a Multimodal Feature Fusion-based Professional Oral English Assessment method (MFF-PEA) is proposed, which effectively solves the limitations of traditional methods in dealing with professional oral English assessment scenarios, such as ignoring non-verbal cues and lacking adaptive fusion mechanisms. By integrating speech, facial expression, and gesture features, the comprehensiveness of assessment is significantly improved. The CMDF strategy is introduced to ensure the adaptability to different professional scenarios. In addition, a hybrid loss function combining MSE and ranking loss is designed to enhance the model's ability to distinguish subtle differences in pronunciation quality. The optimised feature extraction process for each modality further improves the model performance. The following conclusions can be drawn from the experiments on self-collected professional oral English datasets:

1 integrating speech, facial expression, and gesture features significantly enriches the assessment information and enhances the comprehensiveness of the system in evaluating professional oral English

2 the introduction of the CMDF strategy improves the adaptability of the assessment model, especially performing well in different professional scenarios where the importance of modalities varies

3 the hybrid loss function combining MSE and ranking loss enhances the model's ability to distinguish subtle differences in pronunciation quality, improving both absolute score accuracy and relative relationship discrimination

4 the optimised feature extraction process for each modality, such as extracting wavelet entropy from speech and lip movement features from facial expressions, further improves the effectiveness of feature representation and the overall performance of the model.

The experimental data in this paper, mainly from self-collected professional oral English datasets covering business, medical, and legal scenarios, validates the effectiveness and usefulness of the proposed method. However, the limitations of the dataset scale may affect the generalisation ability of the model in more diverse professional environments. Future work should consider expanding the dataset with more samples from various professional domains to validate the effectiveness of the model in a wider range of application scenarios and further optimise the fusion strategy to adapt to more complex cross-domain assessment tasks.

Declarations

All authors declare that they have no conflicts of interest.

References

Ahmadian, S., Brevik, L.M. and Öhrn, E. (2024) 'Adventures with anxiety: gender bias in using a digital game for teaching vocational English', *Journal of Computer Assisted Learning*, Vol. 40, No. 6, pp.2715–2734.

Akila, B. and Nayahi, J.J.V. (2024) 'Parkinson classification neural network with mass algorithm for processing speech signals', *Neural Computing and Applications*, Vol. 36, No. 17, pp.10165–10181.

Anees, M. (2024) 'Speech coding techniques and challenges: a comprehensive literature survey', *Multimedia Tools and Applications*, Vol. 83, No. 10, pp.29859–29879.

Babinski, L.M., Amendum, S.J., Carrig, M.M., Knotek, S.E., Mann, J.C. and Sánchez, M. (2024) 'Professional learning for ESL teachers: a randomized controlled trial to examine the impact on instruction, collaboration, and cultural wealth', *Education Sciences*, Vol. 14, No. 7, p.690.

Derakhshan, A., Teo, T. and Khazaie, S. (2025) 'Investigating the usefulness of artificial intelligence-driven robots in developing empathy for English for medical purposes communication: the role-play of Asian and African students', *Computers in Human Behavior*, Vol. 162, p.108416.

Huang, Y-p., Lin, L-C. and Tsou, W. (2024) 'Leveraging ESP teachers' roles: EMI university teachers' professional development in medical and healthcare fields', *English for Specific Purposes*, Vol. 74, pp.103–116.

Jendli, A. and Albarakati, M. (2024) 'Exploring motivational dynamics: the role of oral activities in improving Arab students' learning of English', *International Journal of Learning, Teaching and Educational Research*, Vol. 23, No. 3, pp.131–149.

Karamatovna, M.A. (2024) 'Fostering advanced professional communicative competence in university students via innovative technologies in the course' culture of speech and communication', *Central Asian Journal of Multidisciplinary Research and Management Studies*, Vol. 1, No. 2, pp.35–41.

Karimpour, P. and Mazlum, F. (2024) 'EAP practitioners' assessment behavior: bringing the hidden-away to light', *Journal of English for Academic Purposes*, Vol. 67, p.101321.

Li, J., Zong, H., Wu, E., Wu, R., Peng, Z., Zhao, J., Yang, L., Xie, H. and Shen, B. (2024) 'Exploring the potential of artificial intelligence to enhance the writing of English academic papers by non-native English-speaking medical students-the educational application of ChatGPT', *BMC Medical Education*, Vol. 24, No. 1, p.736.

Liu, J. (2025) 'Development of interactive English e-learning video entertainment teaching environment based on virtual reality and game teaching emotion analysis', *Entertainment Computing*, Vol. 52, p.100884.

Maniscalco, L., Veronese, N., Ragusa, F.S., Vernuccio, L., Dominguez, L.J., Smith, L., Matranga, D. and Barbagallo, M. (2024) 'Sarcopenia using muscle mass prediction model and cognitive impairment: a longitudinal analysis from the English longitudinal study on ageing', *Archives of Gerontology and Geriatrics*, Vol. 117, p.105160.

Ni, R., Boehlert, C.J., Zeng, Y., Chen, B., Huang, S., Zheng, J., Zhou, H., Wang, Q. and Yin, D. (2024) 'Automated analysis framework of strain partitioning and deformation mechanisms via multimodal fusion and computer vision', *International Journal of Plasticity*, Vol. 182, p.104119.

Pavlenko, A. (2024) 'Language proficiency as a matter of law: judicial reasoning on Miranda waivers by speakers with limited English proficiency (LEP)', *International Journal for the Semiotics of Law-Revue Internationale De Sémiotique juridique*, Vol. 37, No. 2, pp.329–357.

Prahaladaiah, D. and Thomas, K.A. (2024) 'Effect of phonological and phonetic interventions on proficiency in English pronunciation and oral reading', *Education Research International*, Vol. 2024, No. 1, p.9087087.

Shah, D.S.M., Othman, S., Salim, M., Salim, M., Khalil, M.I.M. and Kusmawan, U. (2024) 'The impact of immersive 360-degree video learning on enhancing oral communication skills', *Journal of Advanced Research in Applied Sciences and Engineering Technology*, Vol. 58, No. 1, pp.55–71.

Singh, M.K. (2024) 'Feature extraction and classification efficiency analysis using machine learning approach for speech signal', *Multimedia Tools and Applications*, Vol. 83, No. 16, pp.47069–47084.

Solem, M.S., Landmark, A.M.D., Stokoe, E. and Skovholt, K. (2024) 'Assessment in practice: achieving joint decisions in oral examination grading conversations', *Scandinavian Journal of Educational Research*, Vol. 68, No. 7, pp.1522–1539.

Wang, R., Zhu, J., Wang, S., Wang, T., Huang, J. and Zhu, X. (2024) 'Multi-modal emotion recognition using tensor decomposition fusion and self-supervised multi-tasking', *International Journal of Multimedia Information Retrieval*, Vol. 13, No. 4, p.39.

Yakhyaevna, N.S. (2024) 'Online teaching of English for medical purposes', *Central Asian Journal of Multidisciplinary Research and Management Studies*, Vol. 1, No. 8, pp.84–87.

Zhang, Z., Yin, W., Wang, S., Zheng, X. and Dong, S. (2024) 'MBFusion: multi-modal balanced fusion and multi-task learning for cancer diagnosis and prognosis', *Computers in Biology and Medicine*, Vol. 181, p.109042.