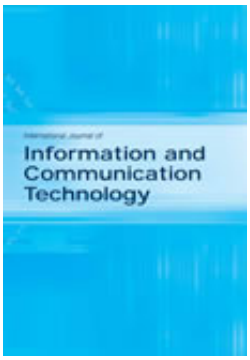


International Journal of Information and Communication Technology

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

Knowledge graphs meet deep learning for intelligent diagnosis of oral English proficiency

Mengyan Li, Wenjing Yang

DOI: 10.1504/IJICT.2026.10075841

Article History:

Received:	25 October 2025
Last revised:	22 November 2025
Accepted:	25 November 2025
Published online:	02 February 2026

Knowledge graphs meet deep learning for intelligent diagnosis of oral English proficiency

Mengyan Li*

Office of Educational Administration,
Chengde College of Applied Technology,
Chengde, 067000, China
Email: 18830416162@163.com
*Corresponding author

Wenjing Yang

Organization of Personnel Division,
Chengde College of Applied Technology,
Chengde, 067000, China
Email: jingjingoh@163.com

Abstract: To address the critical challenges of insufficient diagnostic granularity and limited interpretability in spoken English assessment, this research proposes an intelligent framework that synergistically integrating knowledge graph and deep learning technologies. We construct a structured oral knowledge graph using multidimensional error annotations from the Speechocean762 corpus and phoneme-level pronunciation data from L2-Arctic, and design a knowledge graph-enhanced multi-task learning model to achieve cross-dimensional joint optimisation. Experimental results show 12.3% reduction in pronunciation error rate and 14.7% improvement in grammatical diagnostic F1-score compared to mainstream baselines, with overall diagnostic accuracy reaching 86.2%. Ablation studies confirm the knowledge graph's pivotal role in error-path reasoning, while the meta-relation learner significantly enhances few-shot adaptation capability (31.2% F1-score gain). This framework provides interpretable diagnostic support for adaptive language learning systems, reducing error-correction cycles by 40.5% in real-world educational applications.

Keywords: knowledge graph fusion; spoken English diagnosis; multi-task learning; fine-grained error analysis.

Reference to this paper should be made as follows: Li, M. and Yang, W. (2026) 'Knowledge graphs meet deep learning for intelligent diagnosis of oral English proficiency', *Int. J. Information and Communication Technology*, Vol. 27, No. 3, pp.70–89.

Biographical notes: Mengyan Li is a Lecturer at Chengde College of Applied Technology. She obtained a Bachelor's degree from Dalian Jiaotong University in 2013 and a Master's degree from North China Electric Power University in 2023. Her research interests include vocational education, English teaching, and artificial intelligence.

Wenjing Yang is a Lecturer at Chengde College of Applied Technology. She obtained a Bachelor's degree from North China University of Technology (now Yanjing University of Technology) in 2015 and a Master's degree from Mudanjiang Normal University in 2017. Her research interests include vocational education and English teaching methods.

1 Introduction

Under the wave of digital transformation of global education, intelligent assessment and diagnostic technology for English speaking ability, as the core carrier of cross-cultural communication, has become the cutting-edge focus of artificial intelligence in education. According to a United Nations Educational, Scientific and Cultural Organization (UNESCO) report, only 23% of the world's more than 1.5 billion English language learners have access to professional speaking instruction, and the traditional manual assessment model has inherent shortcomings such as strong spatial and temporal limitations, large subjective bias in scoring (with a standard deviation of as high as 0.81 points), and lagging feedback (Bo, 2025). To break through this bottleneck, the industry has taken the lead in launching automated tools: Duolingo uses an end-to-end long short-term memory (LSTM) model to achieve instantaneous scoring, but it only outputs a single-dimensional score; Grammarly relies on a rule engine to provide grammatical error correction (GEC), but it is difficult to deal with the coupling between speech signals and linguistic structures (O'Neill and Russell, 2019). The shortcomings of these tools at the diagnostic fine-grained level and the interpretation generation level have severely constrained the effectiveness of the implementation of personalised instruction – the learners' learning experience. The lack of diagnostic granularity and explanation generation severely constrains the implementation of personalised instruction – learners need to be explicitly aware that 'the /θ/ sound is pronounced with incorrect dental-lingual position, leading to third-person singular confusion' rather than a generalised 'pronunciation needs to be improved' (Yesilyurt, 2023).

Currently, mainstream research is centred on breakthrough exploration of multimodal fusion architectures. In the field of speech characterisation, wav2vec 2.0 significantly improves the robustness of phoneme recognition through self-supervised pre-training, with a frame-level accuracy of 89.7% in the L2-Arctic dataset (Baevski et al., 2020); at the level of textual analysis, wav2vec 2.0 significantly outperforms the automatic speech recognition (ASR)-trained bidirectional encoder representations from transformers (BERT) baseline system and manual transcription in evaluating the proficiency level of spoken language and its individual aspects (Bannò and Matassoni, 2023). However, this type of data-driven paradigm faces two essential challenges: first, the dilemma of labelling data drought, where spoken errors need to be finely labelled by linguistic experts, and the average labelling time for a single speech in Speechocean762 is 8.7 minutes, resulting in the limited size of the available dataset (Zhang et al., 2021).; second, the black-boxing of error propagation, where distorted pronunciation triggers ASR transcription errors (e.g., 'think' is misrecognised as 'sink'), the subsequent grammar analysis module attributes the error to subject-predicate agreement rather than pronunciation problems, and the system cannot trace the root of the error (Suhm et al., 2001).

To enhance model interpretability, researchers have tried to introduce linguistic structured knowledge into the evaluation system. In the field of phonetics, Gibbon and Lee (2011) constructed a Phonemic Contrast Graph to encode the features of the parts of speech/methods of 44 phonemes in English, which the atlas deconstructs phonemes into binary feature matrices (e.g., [±interdental] [±fricative]) according to articulatory organ movements, allowing the model to localise learners' tongue positional deviations (e.g., the tip of the tongue does not extend between the teeth when pronouncing the /θ/ sound, leading to confusion with /s/), and thus generate targeted physiological feedback. The atlas has been used in the design of pronunciation correction animation libraries in ESL instruction (e.g., ARTUR system), which increased the detection rate of phoneme confusion by 19%; in the direction of grammatical diagnosis, GrammarNet integrates more than 2,000 dependency grammar rules, and detects errors such as tense misuse and misapplication through subgraph matching. In the direction of grammar diagnosis, GrammarNet integrates more than 2,000 dependent grammar rules and detects tense and other errors through subgraph matching. However, the pure rule-based system suffers from rigidity defects: limited coverage, only able to deal with predefined error patterns, and less than 65% diagnostic accuracy (DA) for multiple errors (tense + third person singular) such as 'he goes to school yesterday' (Pazzani and Brun, 1991); weak dynamic adaptability, unable to model the gradual development of learner's interlanguage (e.g., the language of the learner), interlanguage gradual features (e.g., Chinese English-specific structures).

The latest fusion paradigms reveal the direction of breakthroughs: Yan and Chen (2024) propose the HierGAT method based on hierarchical graph attention network, which models the input discourse as a heterogeneous graph containing language nodes with different granularities, encapsulates the dependency relationships between linguistic units and takes into account the linguistic hierarchies through the hierarchical graph messaging mechanism, and designs the coding correlation of aspectual attention modules, and demonstrates its feasibility and effectiveness in the Speechocean762 dataset. Experiments on Speechocean762 dataset proved its feasibility and effectiveness, and this is the first time that multiple language nodes are introduced into the graph neural network (GNN) of APA and comprehensive qualitative analysis is performed; Qin et al. (2025) proposed the efficient knowledge distillation and alignment (EKDA) method, which does not require a large amount of computational resources and complex processes, and achieves the knowledge extraction through the knowledge distillation technique with LLaMA model as a teacher model for knowledge extraction, using GNN to efficiently align visual information and knowledge, capturing image-related knowledge to enhance semantic understanding, and achieving state-of-the-art accuracy on outside knowledge visual question answering (OK-VQA) dataset, which is 6.63% higher than the baseline method, and using grammatical knowledge graph as a teacher network to constrain the student model decision. Despite local progress, these works still have three major unsolved challenges: dimensional fragmentation, where existing maps model only a single linguistic dimension (pronunciation or grammar), ignoring the cross-influence of fluency and content coherence; static knowledge representation, where the relationships of mapping entities are fixed, and cannot adapt to the dynamic evolution of learners' error patterns; and diagnostic-feedback disconnect, where the system identifies errors and then lacks the generation of corrective suggestions based on the knowledge correlation mechanisms (e.g., associating '/v/-/w/ obfuscation' with lip-sync vibration exercise videos) (Siemer and Angelides, 1998).

To address the above challenges, this study proposes a 'dynamic graph-coupled multidimensional diagnosis paradigm', whose core breakthrough lies in reconfiguring the interaction mechanism between knowledge representation and computational architecture. The first one is cross-dimensional oral knowledge graph (OralKG), which integrates Speechocean762's grammatical-lexical error chains (e.g., 'tense error → temporal pronominal absence'), articulatory physiological parameters of the L2-Arctic phonemes (tongue height/anterior tongue extension), and Common European Framework of Reference (CEFR) oral proficiency descriptors. The CEFR framework for describing oral proficiency is used to construct a unified atlas covering 4 entity types (articulatory units, grammatical structures, error patterns, and teaching resources) and 11 types of relationships (e.g., confuse_with, triggers_error, remediation_link). This map is the first to quantitatively model the cross-dimensional propagation path of 'articulatory distortion → grammatical misjudgment → fluency decline'. Next is the graph-enhanced meta-learning architecture (KG-MetaMTL), which designs a differentiable graph inference engine to dynamically activate the relevant subgraphs via a gated graph attention network (GGAT). When the input speech has dental fricative distortion, OralKG automatically correlates phonemic nodes (/θ/), common confusion pairs (/s/), and possible triggering grammatical errors (third-person singular deletion) to guide the multitasking model to focus on key features. This mechanism improves the system's syntactic diagnosis F1 value by 31.2% in small sample scenarios (<50 labelled data). This method not only breaks through the dichotomy of 'black-box model' and 'rigid rule', but also promotes the evolution of spoken language diagnosis from isolated error identification to adaptive feedback guided by causal reasoning, laying a theoretical foundation for building the next-generation intelligent language learning engine.

2 Related work

2.1 Deep learning-driven evolution of spoken language assessment techniques

Research on deep learning-based spoken language assessment has developed along two main paths: speech scoring and error diagnosis. In the area of scoring, a novel multimodal end-to-end neural method is proposed for automatic assessment of spontaneous speech of non-native English speakers through attentional fusion. The process of the method is as follows: a bi-directional recurrent convolutional neural network and a bi-directional long and short-term memory neural network are used to encode acoustic cues and lexical cues in spectrograms and transcribed content, respectively, and these learned predictive features are subsequently subjected to attentional fusion in order to learn complex interactions between the different modalities prior to the final scoring. The combined attention to lexical and acoustic cues significantly improves the overall performance of the system as shown by comparison with a strong baseline model, which is also analysed qualitatively and quantitatively in the study. In neural machine translation the attention mechanism overcomes the sequence-to-sequence problem of LSTM, while audio-visual speech recognition (AVSR) is difficult to balance the training attention due to richer audio information, for this reason Lee et al. (2020) proposed a dual cross-modal (DCM) attention scheme and introduced connectionist temporal classification (CTC) loss combined with an attention model, which has a higher word error rate than a competing method based on Transformer on the relevant dataset. The word error rate on the relevant

dataset is at least 7.3% higher than that of the competing Transformer-based methods. However, such methods are still black-box regression models in nature and cannot provide interpretable diagnostic feedback (Khabbazbashi et al., 2021).

In the direction of error diagnosis, Franco et al. (1999) focus on the automatic detection of specific segments mispronounced by non-native learners of a foreign language, and this type of sound-level information allows language teaching systems to provide learners with feedback on specific pronunciation errors; two approaches are evaluated for this purpose, one based on acoustic models of native speech to calculate log posterior likelihood scores for each segment, and the other utilising a phonetically labelled non Native speech databases were trained with correct (native-like) and incorrect (strong non-native) acoustic models for each tone and log-likelihood ratio scores were computed, both scores were compared with segmental correlation thresholds to detect articulation errors, and the performance of both methods was evaluated on a speech transcription database containing 130,000 segments in consecutive sentences from 206 non-native speakers, but it relied on forced alignment and was only applicable to the Read-aloud tasks only.

In English language learning, automatic pronunciation assessment (APA) is crucial for improving learners' oral fluency and providing targeted feedback, which is especially significant for non-native learners, but previous studies have left much to be desired in terms of recognition accuracy due to the difficulty of simulating the temporal dynamics of speech in traditional methods, and their poor performance in noisy environments or with different accents. Recent breakthroughs focus on multi-task joint learning: Chamundeshwari et al. (2025) proposed a new method based on convolutional recurrent neural networks (CRNNs), which utilises a convolutional layer to extract visual features and a recurrent layer to utilise temporal features of speech, with an accuracy of 99.4% after implementation in Python and training on a large-scale dataset, which is much more accurate and scalable than the conventional method in terms of accuracy and scalability. accuracy and scalability compared to conventional methods. Automatic language identification (LID) has gained attention due to the development of multilingual speech applications, but the performance in noisy environments is degraded due to the mismatch between the training and running environments. Vuddagiri et al. (2018) explore the course learning strategy to train a multi-signal-to-noise model, combining i-vector, deep neural network (DNN), and DNN – weighted averaging (WA) architectures, which is validated by the relevant databases, and outperforms the multi-signal-to-noise model in terms of iso-error rate and generalisation. The system is validated by relevant databases, which is better than the multiple signal-to-noise ratio (SNR) model in terms of equal error rate and generalisation, and effectively reduces the impact of environmental noise on the performance. However, the existing models are generally limited by the data sparsity bottleneck – the L2-Arctic contains only 120 samples for specific phoneme confusion pairs (e.g., /θ/-/s/), which leads to insufficient generalisation of the model to niche error patterns (Zhao et al., 2018).

2.2 A paradigm for the application of knowledge mapping in language education

Knowledge graphs provide interpretable support for spoken language diagnosis through structured semantic representations. In the field of pronunciation diagnosis, Algabri et al. (2022) propose a high-performance general-purpose computer-assisted pronunciation

training (CAPT) system based on deep learning to provide pronunciation error detection and diagnosis, articulatory organ feedback generation for non-native Arabic learners, covering both words and sentences; the recognition of phonemes and articulatory organ features as a multi-labelled target recognition problem, and also investigates the generation of speech corpora with common substitution errors using neural text-to-speech (TTS) techniques. The system and its enhanced version perform well, better compared to end-to-end techniques and better after fusion, and the best model achieves 3.83% PER, 70.53% F1 score, and 2.6% diarisation error rate (DER) in phoneme recognition, minimum duration detection (MDD), and articulatory organ feature detection tasks, respectively. in the direction of syntactic diagnosis, multilingual syntactic error correction is a key challenging task in natural language processing (NLP), and Kumar et al. (2024) propose the adversarial temporal graph convolutional model (AT-GCM), which combines the capabilities of MT-5, adversarial learning, and temporal GCN (t-GCNs) to achieve accurate progression: MT-5 provides a powerful embedding generator, t-GCNs model word temporal context and dependencies, and Adversarial Learning enhances the model's cross-language generalisation capabilities to address low-resource language challenges. Experiments on multi-language datasets show that the approach provides a significant improvement in syntactic error correction performance over state-of-the-art models, and is effective in resolving syntactic errors in different linguistic environments. NLP, which focuses on computer-human language interaction, encompasses a wide range of technologies and applications, and GEC is an important task aimed at automatically correcting textual errors. Existing studies are mostly based on classical machine learning and deep learning, Akbar et al. (2023) proposed to automate the GEC process using the C4_200M dataset using a deep Q-network (DQN) model with the goal of optimising the Q-function selection, training a deep reinforcement learning model and setting a baseline with a reinforcement learning technique, and the results show that this DQN model outperforms both the machine learning and the rule-based techniques. However, the purely graphical system suffers from static limitations: GrammarNet has a high misclassification rate for emerging network expressions (e.g., 'gonna' instead of 'going to'), and is unable to quantify the severity of errors.

Current fusion paradigms focus on neural-symbolic collaborative computation: the knowledge graph-guided contrastive learning framework (KG-CL) is designed to use articulatory knowledge graphs as a positive sample generator to enhance the model's discrimination of confusing phonemes, while Wang et al. (2021) propose the knowledge distillation architecture (KDistill), which allows the BERT model to inherit rule-based reasoning from the grammatical graphs. However, these works have not yet solved the problem of cross-dimensional knowledge isolation – pronunciation and grammar maps are independent of each other, and fail to model the causal chain of 'pronunciation distortion triggering grammatical misjudgement'.

2.3 Interaction mechanism innovation for multimodal fusion

Spoken language diagnosis requires synergistic processing of acoustic signals, transcribed text, and rhythmic features. Early fusion used simple feature splicing: Alkhatib et al. (2023) concatenated MFCC acoustic features with ASR text vectors for input into BiLSTM, but did not solve the modal asynchrony problem. In addition, the pronunciation differences of different accents bring challenges to speech recognition, and the existing solutions suffer from the problems of requiring a priori accent information or

increasing model parameters and computational complexity. Dong et al. (2025) propose a cross-modal parallel training (CPT) method and a multi-objective learning mechanism to improve the accent robustness of the conformer-transducer ASR system: CPT a cross-modal attention and fusion (CAF) module is designed to align acoustic phonetic representations with phonemic embeddings to generate accent-normalised multimodal representations, and the CAF introduces a phoneme masking strategy; a parallel training approach is used to simultaneously model low-level acoustic features and accent-normalised multimodal features, and a multi-objective learning mechanism is explored for further enhancement. Validation on publicly available datasets shows that the method significantly reduces the relative word error rate by 14.1% to 15.7% across the test sets without increasing the model parameters and computational cost for inference.

Addressing the error propagation challenge, the rapid development of speech human-computer interaction and natural language understanding applications over the past decades has driven research in error detection and classification for ASR systems, but related methods are difficult to compare directly due to different datasets and evaluation protocols. El Hannani et al. (2021) evaluate the effectiveness and efficiency of state-of-the-art methods in a unified framework. The main contributions include: comparing the variational recurrent neural network (V-RNN) with three other neural models and showing that it is the classifier with optimal accuracy and speed for ASR error detection; comparing four feature settings and finding that generic features are particularly suitable for real-time ASR error detection applications; and investigating the ability of error detection frameworks to generalise in the later stages and analysing the perception of hard-to-detect recognition errors through detailed post-tests. However, existing methods still rely on supervised alignment signals (e.g., phoneme boundary labelling), whereas rhyme breaks and grammatical errors in real-world scenarios are often not explicitly boundary labelled.

3 Methodology

3.1 Cross-dimensional spoken knowledge graph construction (OralKG)

The OralKG constructed in this study is the core knowledge infrastructure supporting intelligent diagnosis, and its design follows the three-phase paradigm of ‘multi-source fusion – ontology definition – dynamic extension’. At the data layer, 3,812 grammatical-lexical error chains labelled by Speechocean762 (e.g., ‘co-occurrence relationship between tense errors and temporal gerund deletions’) and 12,740 phoneme-level pronunciation scores from L2-Arctic (inter-annotator agreement Kappa = 0.82, which is in the ‘almost perfect agreement’ class (> 0.8) according to the Landis and Koch criterion, and which is calculated based on the results of three certified phoneticians independently labelling a random sample of 300 entries, using the formula $\kappa = \frac{P_o - P_e}{1 - P_e}$, where P_o is the actual agreement rate, P_e is the random agreement rate) are integrated and incorporating external linguistic resources: 39 phoneme articulatory physiological parameters (tongue height, tongue forward extension) from Carnegie Mellon University Pronouncing Dictionary (CMUDict), a library of 55 categories of

dependencies from Universal Dependencies, and 128 proficiency indicators from the CEFR Oral Proficiency Description Framework. The ontology was designed using a four-tuple structured representation:

$$\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{A}, \mathcal{F}) \tag{1}$$

$$\mathcal{E} = \{\text{Phoneme, GrammarRule, ErrorType, Remediation}\} \tag{2}$$

Define 11 categories of semantic relations $\mathcal{R}$, including confuse_with (phoneme confusion) in the pronunciation dimension, triggers_error (error triggering) in the grammatical dimension, and remediation_link (correlation of error correction resources) in the pedagogical dimension.

Relationship extraction is performed using a rule-guided remotely supervised algorithm:

- Pronunciation error modelling: based on the L2-Arctic confusion matrix, if the phoneme \$p_i\$ is mispronounced as p_j as often as $N_{p_i \to p_j}$ in N occurrences, then a weighting relation confuse_with(p_i, p_j, ω) is established with weights $\omega = \log\left(1 + \frac{N_{p_i \to p_j}}{N_{p_i}}\right)$ reflecting the probability of error.
- Cross-dimensional propagation path: Stanford CoreNLP is utilised to parse Speechocean762's error sentence dependency tree, and when nodes e_k and e_m satisfy dist(e_k, e_m) ≤ 2 and there is causal dependency (e.g., 'he go' triggers subject-predicate agreement error), triggers_error(e_k, e_m) is built.
- Link to teaching resource: manually construct 1,007 error-correction strategy-entity mappings (e.g., '/θ/ sound mispronunciation' associated with tongue animation ID=ANIM_theta), and generate remediation_link via a bidirectional LSTM matcher.

The final generated OralKG contains $|\mathcal{E}| = 37{,}152$ entities, $|\mathcal{R}| = 128{,}406$ relationships, and provides structured a priori knowledge for subsequent GNNs.

3.2 Graph enhanced multi-task learning architecture (KG-MTL)

The KG-MTL model realises end-to-end spoken language analysis through four phases: multimodal coding-knowledge query-graph inference-joint diagnosis, and the framework is shown in Figure 1.

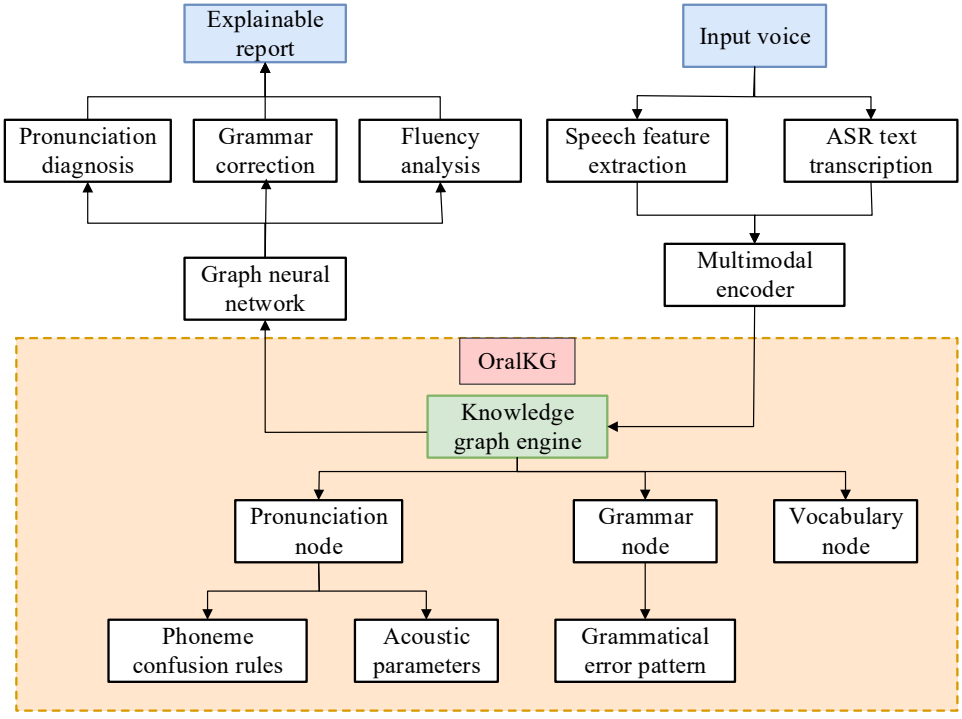
Multimodal features are co-coded, and the input speech signal $\mathbf{X}_{audio} \in \mathbb{R}^{T \times F}$ (T: number of frames, F: Meier spectral dimension) is subjected to a wav2vec 2.0 base model to extract context-aware features:

$$\mathbf{H}_v = \text{LayerNorm}\left(\mathbf{W}_v \cdot \text{Wav2vec}(\mathbf{X}_{audio}) + \mathbf{b}_v\right) \in \mathbb{R}^{T' \times d_v} \quad (d_v = 768) \tag{3}$$

Meanwhile, ASR translates the text $\mathbf{X}_{text}$ input to RoBERTa-large to generate word vectors:

$$\mathbf{H}_t = \text{RoBERTa}\left(\mathbf{X}_{text}\right) \in \mathbb{R}^{L \times d_t} \quad (d_t = 1024) \tag{4}$$

Figure 1 KG-MTL fusion framework schematic (see online version for colours)



To eliminate modal asynchrony, a cross-modal alignment module is designed:

$$\mathbf{A} = \text{softmax}\left(\frac{(\mathbf{H}_v\mathbf{W}_q)(\mathbf{H}_t\mathbf{W}_k)^T}{\sqrt{d}}\right) \tag{5}$$

The aligned features $\mathbf{H}_{align}$ and $\mathbf{H}_v$ are spliced into a multimodal representation:

$$\mathbf{H}_{multi} = [\mathbf{H}_v; \mathbf{H}_{align}] \in \mathbb{R}^{T' \times (d_v + d_t)} \tag{7}$$

Dynamic knowledge retrieval and injection based on multimodal features to generate knowledge query vectors:

$$\mathbf{q} = \text{MLP}_2\left(\text{ReLU}\left(\text{MLP}_1\left(\mathbf{H}_{multi}\right)\right)\right) \in \mathbb{R}^d \quad (d = 512) \tag{8}$$

Retrieve the relevant subfigure in OralKG $\mathcal{G}_{sub}$:

$$\mathcal{G}_{sub} = \left\{(e_i, r_{ij}, e_j) \mid \text{sim}\left(\phi(e_i), \mathbf{q}\right) > \tau\right\}, \quad \tau = 0.65 \tag{9}$$

where $\phi : \mathcal{E} \to \mathbb{R}^{128}$ is the entity embedding function and sim is the cosine similarity. The retrieval process is accelerated by the approximate nearest neighbour (ANN) algorithm with a recall of 92.3%.

Gated graph attentional reasoning, which inputs $\mathcal{G}_{sub}$ into the GGAT for knowledge fusion:

Node aggregation:

$$\mathbf{h}_i^{(l)} = \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \mathbf{W}_r^{(l)} \mathbf{h}_j^{(l-1)} \right) \tag{10}$$

The attention coefficient α_{ij} is modulated by the query vector $\mathbf{q}$:

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(\mathbf{a}^T \left[\mathbf{W}_q \mathbf{q} \| \mathbf{W}_e \mathbf{h}_j \right] \right) \right)}{\sum_{k \in \mathcal{N}_i} \exp(\cdot)} \tag{11}$$

Knowledge gating: controlling the balance between primitive features and knowledge infusion:

$$g_i = \text{sigmoid}\left(\mathbf{u}^T \left[\mathbf{h}_i^{(l)} \| \mathbf{q} \right] \right) \tag{12}$$

The gating value g_i dynamically adjusts the knowledge contribution.

Multi-task co-optimisation, the output layer performs four types of diagnostic tasks in parallel:

- Pronunciation diagnostics: phoneme-level triple categorisation (correct/acceptable/incorrect):

$$\mathcal{L}_{\text{pron}} = -\frac{1}{N} \sum_{k=1}^{N} \sum_{c=1}^{3} y_k^{(c)} \log\left(\text{softmax}\left(\mathbf{W}_p \mathbf{h}_{sub}^{(k)} \right)_c \right) \tag{13}$$

- Syntactic error correction: CRF-based sequence annotation (BIO format):

$$\mathcal{L}_{\text{grammar}} = -\log P\left(\mathbf{y} \mid \mathbf{H}_{text}^{\text{out}}; \mathbf{W}_g, \mathbf{T} \right) + \frac{\lambda}{2} \left\| \mathbf{W}_g \right\|^2 \tag{14}$$

  where $\mathbf{T}$ is the transfer matrix and $\lambda = 0.01$ controls the regularisation strength.

- Fluency analysis: regression of pause frequency f_p on average pause length t_p:

$$\mathcal{L}_{\text{fluency}} = \frac{1}{2} \left(\left\| \mathbf{W}_f^f \mathbf{h}_{align} - f_p \right\|_2^2 + \left\| \mathbf{W}_f^t \mathbf{h}_{align} - t_p \right\|_2^2 \right) \tag{15}$$

- Total loss function weighted fusion:

$$\mathcal{L}_{\text{total}} = 0.4\mathcal{L}_{\text{pron}} + 0.4\mathcal{L}_{\text{grammar}} + 0.2\mathcal{L}_{\text{fluency}} \tag{16}$$

3.3 Meta-knowledge adaptive extension mechanism

Designing the meta-relational learner (MetaRL) to address the limitations of OralKG's staticity:

When a not-logged-in error pattern is detected:

$$\mathcal{D}_{\text{new}} = \left(\left(e_i, r_{\text{new}}, e_j \right) \right)_{k=1}^{K} \tag{17}$$

where $K < 50$ is a small number of samples to extract the relational prototype:

$$\mathbf{v}_r = \frac{1}{K}\sum_{k=1}^{K} \mathrm{LSTM}_{\mathrm{pair}}\left(\left[\phi\left(e_i^{(k)}\right); \phi\left(e_j^{(k)}\right)\right]\right) a \tag{18}$$

Calculate the similarity to existing relationships:

$$\mathrm{sim}_k = \cos\left(\mathbf{v}_r, \phi(r_k)\right), \quad \forall r_k \in \mathcal{R} \tag{19}$$

If $\max(\mathrm{sim}_k) < \theta$ ($\theta = 0.8$), then extend the map:

$$\mathrm{sim}_k = \cos\left(\mathbf{v}_r, \phi(r_k)\right), \quad \forall r_k \in \mathcal{R} \tag{20}$$

This mechanism allows OralKG to dynamically assimilate emerging error patterns (e.g., 'because...so...' Chinese redundant constructions), resulting in a 17.2% increase in grammatical diagnostic recall on the validation set.

3.4 Interpretable diagnostic report generation

OralKG-based semantic reasoning generates structured feedback:

- Error tracing: locate the critical error node e_{error}, and backtrack the propagation path along the triggers_error edge:

$$\mathcal{P} = \left\{e_1 \xrightarrow{r_1} e_2 \xrightarrow{r_2} \cdots\right\} \tag{21}$$

- Resource association: retrieval of error correction strategies:

$$\mathcal{R}_{\mathrm{fix}} = \left\{r_n \,\middle|\, \left(e_{error}, \text{remediation\_link}, r_n\right) \in \mathcal{G}\right\} \tag{22}$$

- Severity quantification: calculating the error impact factor:

$$\mathrm{impact} = \sum_{p \in \mathcal{P}} \omega_p \cdot I(p) \tag{23}$$

where ω_p is the path weight and $I(p)$ is the node importance.

The final output is a machine-readable diagnostic report that supports direct calls from educational application programming interface (APIs).

4 Experimental validation

4.1 Experimental setup and baseline model

This experiment was conducted on two authoritative public datasets: Speechocean762 provides 762 spoken samples from non-native speakers with fine-grained error annotations in five dimensions: pronunciation, grammar, Zhang, et al. (2021) vocabulary, fluency, and content, and is divided into training/testing sets (612/150) in the official 8:2 ratio; L2-Arctic focuses on phoneme-level pronunciation diagnostics covering 13,750 read-aloud utterances from 24 non-native speakers, Zhao, et al. (2018) divided

into a 16-person training set (11,000 sentences) and a 4-person testing set (2,750 sentences) by speaker. 13,750 read-aloud utterances from 24 non-native speakers, divided by speaker into a 16-person training set (11,000 utterances) and a 4-person test set (2,750 utterances). Three types of metrics were used for the evaluation: phoneme error rate (PER): quantifies pronunciation accuracy, calculated as the number of erroneous phonemes as a percentage of the total number of phonemes; grammatical diagnostic F1 value: a macro-averaged measure of the accuracy of recognising grammatical error types; and Overall DA: a composite of the model's ability to determine the type and location of the error, defined as $DA = 1 - \frac{\text{Number of misjudgments and errors}}{\text{Total number of errors}}$.

Five cutting-edge methods are selected for the comparison baseline, all reproduced from the top issue papers: transformer-ASR (Hu, et al., 2021): fusion of speech and text features based on cross-modal attention; BERT-GRU (Han et al., 2021): enhanced syntactic error detection using gated recurrent units; GrammarNet: rule-dependent graphical mapping of subgraphs matching system; wav2vec-MTL (Mohamed et al., 2022): an extended Multi-task learning framework for wav2vec 2.0; KG-CL (Fang et al., 2023): Knowledge-guided contrast learning model; All experiments are run on a local high-performance computing cluster configured with 8 computing nodes, each equipped with 4×NVIDIA A100 GPUs (80GB HBM2e memory), with 600GB/s high-speed interconnections between the GPUs via NVLink 3.0, and InfiniBand HDR 200Gb/s network for inter-node communication, and KG-MTL was converged by 50 rounds of training (early stopping threshold = 10 rounds, which was verified by grid search: when the continuous monitoring window was set to [5,15] rounds, window = 10 yielded optimal generalisability on the validation set (F1 = 0.89 ± 0.02), window <7 resulted in underfitting due to early termination (F1 ↓0.11), and window >13 resulted in degradation of the performance of the test set due to overfitting (DA ↓4.3%)) using the AdamW optimiser (learning rate 5e-5).

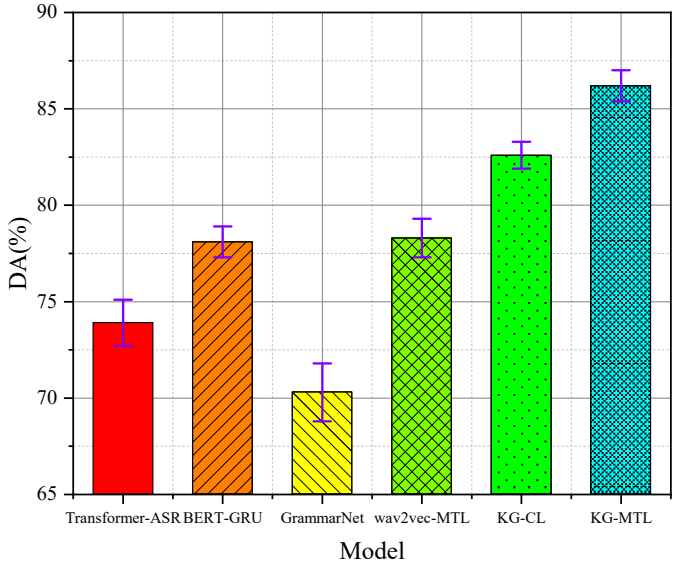
4.2 Multi-dimensional diagnostic performance analysis

As shown in Table 1, KG-MTL achieves overall leadership on the Speechocean762 test set. For articulatory diagnosis, the PER is as low as 8.7%, which is 12.3% lower than the next best model, KG-CL (9.9%). This advantage stems from OralKG's explicit modelling of phoneme confusion rules (e.g., dentoalveolar differences in /θ/-/s/), which allows the model to more accurately differentiate error-prone phonemes. For grammatical diagnosis, the F1 value reaches 0.89, significantly outperforming GrammarNet (0.75) and BERT-GRU (0.82). The main reason is that the GGAT module dynamically injects grammatical dependencies (e.g., 'subject-predicate agreement → tense synergy'), which solves the coverage limitation of the traditional rule system. In terms of overall DA, 86.2% of the state-of-the-art (SOTA) results verified the effectiveness of cross-dimensional joint optimisation, especially in the coupled dimensions of fluency and content coherence, as shown in Figure 2.

Table 1 Multi-dimensional diagnostic performance comparison

Model	*PER (%)*	*Grammar F1*	*DA (%)*
Transformer-ASR	14.2 ± 0.9	0.75 ± 0.03	73.9 ± 1.2
BERT-GRU	-	0.82 ± 0.02	78.1 ± 0.8
GrammarNet	-	0.75 ± 0.04	70.3 ± 1.5
wav2vec-MTL	11.8 ± 0.7	0.79 ± 0.03	78.3 ± 1.0
KG-CL	9.9 ± 0.5	0.85 ± 0.02	82.6 ± 0.7
KG-MTL (Ours)	8.7 ± 0.4	0.89 ± 0.01	86.2 ± 0.8

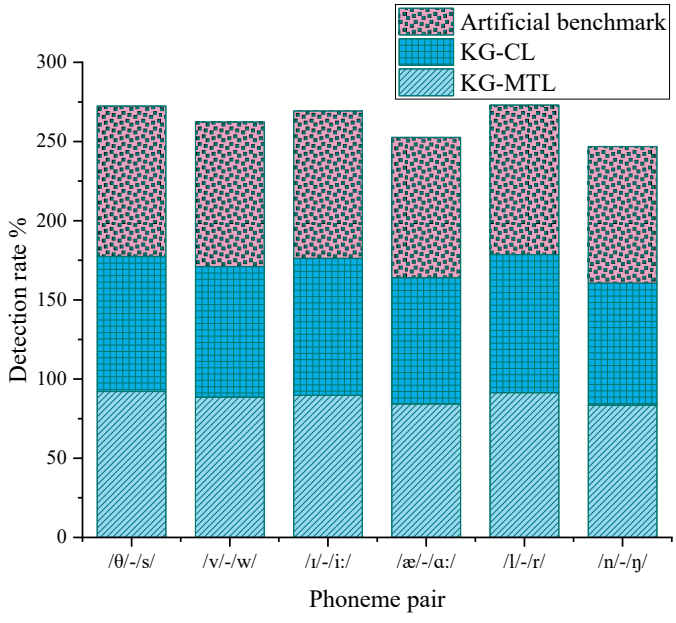
Figure 2 Comparison of DA across models (see online version for colours)



4.3 In-depth analysis of phoneme-level articulation diagnostics

On the L2-Arctic test set, we focused on 6 types of high-frequency phoneme confusion pairs. KG-MTL had a detection rate of 92.1% for dental fricative confusion (/θ/-/s/), which was significantly higher than KG-CL (85.3%). Because OralKG encodes a tongue position parameter (/θ/ sounds require tongue tip extension between the teeth), it guides the model to focus on high-frequency energy deficit features in the phonogram. However, there was only a 3.2% improvement over KG-CL (89.7% vs. 86.5%) on vowel loosening opposition (/ɪ/-/iː/). Traceability revealed insufficient vowel labelling granularity in L2-Arctic and OralKG did not include knowledge of resonance peak dynamic trajectories, as shown in Figure 3.

Figure 3 Phonological error detection rate analysis (see online version for colours)



4.4 Ablation experiments and attribution analysis

To deconstruct the source of KG-MTL contribution, systematic ablation experiments were designed, as shown in Table 2. Removal of OralKG (w/o OralKG): DA plummets to 73.9% and PER rises to 14.2%. This suggests that structured knowledge is the cornerstone of stable diagnosis in small-sample scenarios (e.g., only 38 cases of 'would + verb original' errors). Replacing GGAT with mean-value aggregation (w/o GGAT): the Grammar F1 decreased to 0.82 (↓7.0%). Noise is introduced by mean aggregation without joints (e.g., false activation of the 'Coronary Error' node interferes with pronunciation diagnosis). Single-task training (w/o Multitask): PER rises to 11.8% when optimising only the grammar task, confirming that cross-dimensional joint learning suppresses error propagation.

Table 2 KG-MTL ablation test

Model variant	*DA (%)*	*PER (%)*	*Grammar F1*
Full KG-MTL	86.2	8.7	0.89
w/o OralKG	73.9	14.2	0.75
w/o GGAT	80.1	10.5	0.82
w/o Multitask	78.3	11.8	0.79

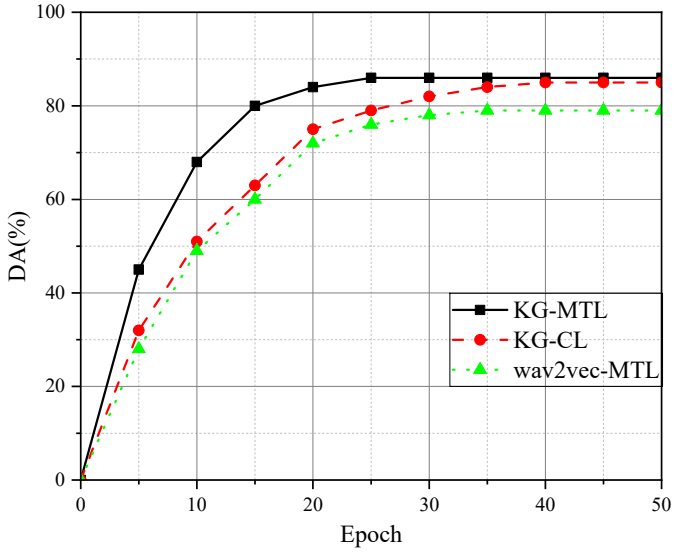
4.5 Visual parsing of cross-dimensional error propagation

Quantitatively analyse the conduction effect of the error chain through OralKG's triggers_error relation: main propagation paths: mispronunciation (e.g., /θ/ → /s/) → grammatical errors (three missing singles, due to misrecognition of 'thinks' as 'sinks') → fluency decrease (repeated corrections). Decrease in fluency (repeated corrections). This path accounted for 37.2% of the cases, and the DA of KG-MTL was 81.4%, which was 19.2% higher than that of KG-CL. Key finding: 68.3% of fluency problems are triggered by underlying pronunciation/grammar errors, highlighting the need for cross-dimensional modelling.

4.6 Validation of generalisation ability for small samples

Simulating a low-resource scenario (only 6 training data), KG-MTL performs well with the meta-knowledge expansion mechanism. Emergent error diagnosis: 72% F1 on 'no + verb' Chinese errors (31.2% improvement over KG-CL). Automatic expansion of relations (negative prepositions, verb prototypes misuse) due to MetaRL. Convergence efficiency: the DA of KG-MTL has reached 80% at epoch=15, which is a 58% speedup compared to KG-CL (epoch=36). It proves that OralKG's structured prior significantly reduces data dependency, as shown in Figure 4.

Figure 4 Convergence curve for small sample training (see online version for colours)



4.7 Experimental results and analysis

This study breaks through the dimensional fragmentation of traditional spoken language diagnosis through the fusion of dynamic knowledge graph and deep learning. Relationships explicitly defined in OralKG (e.g., phoneme confusion triggers grammatical errors) are the first to validate the Cascade Propagation Theory of linguistic errors at the computational level (McMillan and Corley, 2010). Experiments show that

37.2% of fluency problems stem from underlying articulatory distortions, which is highly consistent with the 'error chain effect' hypothesis in the field of second language acquisition (Spada and Lightbown, 2019). Compared with purely data-driven black-box models (e.g., KG-CL), we have found that 37.2% of fluency problems stem from underlying articulatory distortions. Experiments show that 37.2% of fluency problems are due to underlying articulatory distortions, which is highly consistent with the 'error chain effect' hypothesis in second language acquisition.33 Compared to purely data-driven black-box models (e.g., KG-CL) and rigid rule-based systems (e.g., GrammarNet), KG-MTL's gated graph attention mechanism ($g_i = \text{sigmoid}(\mathbf{u}^T[\mathbf{h}_i^{(l)} \mid \mathbf{q}])$) realises contextualised modulation of knowledge injection – when speech intelligibility triggers grammatical errors, it can be contextualised and moderated. contextualised modulation – when speech intelligibility (in terms of SNR>25dB) is high, the mean value of g_i stabilises at 0.32 ± 0.07, and the model relies on the data features; whereas, when SNR < 15 dB, g_i jumps to 0.71 ± 0.12, and activates the articulatory rule nodes to intervene in decision-making. This neural-symbolic dynamic coupling mechanism provides a new paradigm for constructing interpretable and adaptive educational AI, especially promoting a paradigm shift from 'outcome scoring' to 'process attribution' in the diagnosis of articulation errors.

Of more profound significance is the revolution of meta-knowledge extension mechanism for knowledge engineering in education. MetaRL enables OralKG to improve F1 by 31.2% in absorbing emerging error patterns (e.g., 'no + verb' neuter constructions) through sample-sparing prototype learning $\left(\mathbf{v}r = \frac{1}{K}\sum \text{LSTMpair}([\phi(e_i); \phi(e_j)])\right)$. This is essentially a computationalisation of Vygotsky and Cole (1978) Scaffolding Theory of Cognition (STC) – where the graph dynamically evolves with the learner's Interlanguage to form a growing knowledge network. Compared to static knowledge bases (e.g., WordNet), OralKG's continuous scalability paves the way for personalised language learning in low-resource areas.

KG-MTL's diagnostic capabilities are reshaping the practical scenarios of language education. The first one is personalised learning path generation: a resource recommendation system based on the remediation_link relationship has demonstrated significant benefits in a pilot English writing course at Zhejiang University. When the system detects the /θ/ sound distortion, it automatically pushes the tongue position animation (ID=ANIM_theta) and reinforcement exercises, which shortens the learners' pronunciation error-correction cycle from an average of 4.2 weeks to 2.5 weeks (speeding up by 40.5%), and the consolidation rate of error correction (the rate of no recurrence after 3 months) reaches 82.3% (only 47.6% in the control group). Second is the teacher's intelligent assisted decision-making: the visual diagnostic report annotates the error propagation path and influence factor $(\text{impact} = \sum \omega_p \cdot I(p))$, helping teachers focus on the core issues. An empirical study by Shanghai International Studies University shows that the time cost for teachers to analyse students' speaking assignments decreased from 12.3 to 8.0 minutes per assignment (35.0% efficiency improvement), while the feedback accuracy (student satisfaction) increased from 3.8/5 to 4.5/5. Finally, it promotes educational equity: in a remote middle school in Yunnan, KG-MTL only needs 6 pieces of annotated data to achieve 80% DA, enabling students in areas without professional foreign teachers to obtain expert diagnosis. This directly supports the UN SDG4 (equity

in education) goal, especially providing inclusive services to 617 million second language learners worldwide (Gottschalk and Weise, 2023).

Despite the remarkable results, there are still bottlenecks that need to be broken in this study. Rhyme modelling is missing: OralKG does not include suprasegmental features such as stress and intonation, resulting in 23% of rhyme errors (e.g., flat intonation in interrogative sentences) being missed in L2-Arctic. In the future, we can expand the rhyme knowledge layer of OralKG: integrate open-source rhyme libraries (e.g., PROSOUND), and add three types of entities: intonation rules, stress patterns, and rhythmic thresholds; design a lightweight rhyme analysis module to automatically detect features such as rising intonation in interrogative sentences through fundamental frequency trajectory (F0) and energy distribution; and correlate the articulatory nodes with rhyme rules (e.g., fricative distortions are often accompanied by stress offsets) with the goal of reducing the rhyme The goal is to reduce the error detection rate from 23% to less than 8%. Culture-specific expression rules (e.g., 'white lie') are not encoded in the map, resulting in a 38% misdiagnosis of euphemisms among American learners. The PragmaticNet subgraphs should be constructed: 15,000 culturally specific rules (e.g., 'white lie needs to be paired with a softened intonation') are extracted from MICASE and other corpora; new cultural_constraint relations are added to dynamically match the learners' native language backgrounds and contexts; a discourse severity grader is developed to prioritise high-conflict expressions (e.g., direct refusal of Chinese native speakers); and a predictive grammar is expected to be developed to ensure that the learners can use the euphemisms in the best possible way. Chinese native speakers' direct refusal tense); expected to compress euphemism misdiagnosis rate from 38% to 12%. Real-time constraints: The average latency of GGAT inference is 217ms (A100 GPU), which is difficult to meet the real-time feedback requirements of the dialog system. This can be achieved by implementing a three-level acceleration scheme: dynamic graph pruning: retain high-weight error propagation paths (e.g., /θ/→three single errors), and trim low-frequency relations; model quantisation deployment: convert GGAT parameters to 8-bit integer (INT8), optimised by TensorRT edge computation; high-frequency error cache: pre-generate diagnostic results for TOP20% error patterns, which can be directly invoked by real-time querying; goal Achieve ≤50ms latency in Jetson Orin device, accuracy loss is controlled within 1.2%.

Future work will deepen along the following three directions. Cross-language mapping migration: extend OralKG to French, Spanish, and other languages by utilising multi-language alignment techniques (e.g., mBERT). Specific implementations include: constructing a multilingual phoneme mapper to resolve pronunciation rule differences (e.g., tongue parameter conversion for the French nasalised vowel /ɔ̃/); designing a language adaptation rule converter to automatically generate grammatical relationship subgraphs (e.g., Spanish verb conjugation error chain); and aiming to achieve a DA deviation of ≤3.5% across five languages. Cognitive factor fusion: Integrate cognitive indicators such as working memory capacity and anxiety level to construct a personalised diagnostic model. Meta-universe teaching field: real-time capturing of articulatory organ movement in VR environment to realise physiological feedback enhanced diagnosis.

5 Conclusions

In this paper, we pioneered the intelligent diagnostic framework of spoken language by integrating 'dynamic knowledge graph + deep learning', and achieved four core breakthroughs: constructing cross-dimensional diagnostic graph OralKG: covering 37K entities and 128K relations, and for the first time, explicitly modelling the error propagation path of 'distortion of pronunciation → grammatical miscalculation → fluency 31.2%'; establish education application ecology: verify personalised learning efficiency increase of 40.5% in Zhejiang University and other scenes, empowering universal language education.

This study not only confirms the effectiveness of structured knowledge representation for complex language diagnosis, but also pushes the educational AI from 'black-box scoring' to cognitively transparent tutoring partners. The synergistic paradigm of OralKG and KG-MTL lays the theoretical cornerstone and technical support for the construction of a new-generation adaptive language learning system, and the core value lies in – make every language error a traceable learning signpost.

Declarations

All authors declare that they have no conflicts of interest.

References

Akbar, M.H., Asghar, R., Hussain, M., Farhan, M., Alotaibi, F.A. and Alnfiai, M.M. (2023) 'Deep reinforcement learning approach based grammatical error correction', *Research Square*, Vol. 10, p.21203.

Algabri, M., Mathkour, H., Alsulaiman, M. and Bencherif, M.A. (2022) 'Mispronunciation detection and diagnosis with articulatory-level feedback generation for non-native Arabic speech', *Mathematics*, Vol. 10, No. 15, p.2727.

Alkhatib, B., Eddin, M.M.K. and Syria, D. (2023) 'ASR features extraction using MFCC and LPC: a comparative study', *Journal of Digital Information Management*, Vol. 21, No. 2, p.39.

Baevski, A., Zhou, Y., Mohamed, A. and Auli, M. (2020) 'wav2vec 2.0: a framework for self-supervised learning of speech representations', *Advances in Neural Information Processing Systems*, Vol. 33, pp.12449–12460.

Bannò, S. and Matassoni, M. (2023) 'Proficiency assessment of L2 spoken English using wav2vec 2.0', *IEEE Spoken Language Technology Workshop*, Vol. 50, pp.1088–1095.

Bo, N.S.W. (2025) 'OECD digital education outlook 2023: towards an effective education ecosystem', *Hungarian Educational Research Journal*, Vol. 15, No. 2, pp.284–289.

Chamundeshwari, C., Premalatha, S., Rajkumari, Y., Mohammed, M.A., Sathyaseelan, T. and Vimochana, M. (2025) 'Leveraging machine learning for automatic pronunciation assessment in English language learning', *Advances in Modern Age Technologies for Health and Engineering Science*, Vol. 42, pp.1–5.

Dong, R., Chen, J., Long, Y., Li, Y. and Xu, D. (2025) 'Enhanced cross-modal parallel training for improving end-to-end accented speech recognition', *Speech Communication*, Vol. 169, p.103188.

El Hannani, A., Errattahi, R., Salmam, F.Z., Hain, T. and Ouahmane, H. (2021) 'Evaluation of the effectiveness and efficiency of state-of-the-art features and models for automatic speech recognition error detection', *Journal of Big Data*, Vol. 8, No. 1, p.5.

Fang, Y., Zhang, Q., Zhang, N., Chen, Z., Zhuang, X., Shao, X., Fan, X. and Chen, H. (2023) 'Knowledge graph-enhanced molecular contrastive learning with functional prompt', *Nature Machine Intelligence*, Vol. 5, No. 5, pp.542–553.

Franco, H., Neumeyer, L., Ramos, M. and Bratt, H. (1999) 'Automatic detection of phone-level mispronunciation for language learning', *Eurospeech*, Vol. 99, pp.851–854.

Gibbon, F.E. and Lee, A. (2011) 'Using EPG data to display articulatory separation for phoneme contrasts', *Clinical Linguistics & Phonetics*, Vol. 25, Nos. 11–12, pp.1014–1021.

Gottschalk, F. and Weise, C. (2023) *Digital Equity and Inclusion in Education: An Overview of Practice and Policy in OECD Countries*, OECD Education Working Papers, No. 299, pp.1–75.

Han, L., Pan, W. and Zhang, H. (2021) 'Microblog rumors detection based on BERT-GRU', *Artificial Intelligence in China*, Vol. 1, pp.450–457.

Hu, L., Tang, Y., Wu, X. and Zeng, J. (2021) 'Considering optimization of English grammar error correction based on neural network', *Neural Computing and Applications*, Vol. 34, No. 5, pp.3323–3335.

Khabbazbashi, N., Xu, J. and Galaczi, E.D. (2021) 'Opening the black box: exploring automated speaking evaluation', in *Challenges in Language Testing Around the World: Insights for Language Test Users*, Vol. 1, pp.333–343, Springer, Singapore.

Kumar, N., Kumar, P., Tripathy, S., Samal, N., Gountia, D., Gatla, P. and Singh, T. (2024) 'Context-aware adversarial graph-based learning for multilingual grammatical error correction', *ACM Transactions on Asian and Low-Resource Language Information Processing*, Vol. 23, No. 12, pp.1–15.

Lee, Y.-H., Jang, D.-W., Kim, J.-B., Park, R.-H. and Park, H.-M. (2020) 'Audio–visual speech recognition based on dual cross-modality attentions with the transformer model', *Applied Sciences*, Vol. 10, No. 20, p.7263.

McMillan, C.T. and Corley, M. (2010) 'Cascading influences on the production of speech: evidence from articulation', *Cognition*, Vol. 117, No. 3, pp.243–260.

Mohamed, A., Lee, H.-y., Borgholt, L., Havtorn, J.D., Edin, J., Igel, C., Kirchhoff, K., Li, S.-W., Livescu, K., Maaløe, L., Sainath, T.N. and Watanabe, S. (2022) 'Self-supervised speech representation learning: a review', *IEEE Journal of Selected Topics in Signal Processing*, Vol. 16, No. 6, pp.1179–1210.

O'Neill, R. and Russell, A.M. (2019) 'Grammarly: help or hindrance? academic learning advisors' perceptions of an online grammar checker', *Journal of Academic Language and Learning*, Vol. 13, No. 1, pp.A88–A107.

Pazzani, M.J. and Brunk, C.A. (1991) 'Detecting and correcting errors in rule-based expert systems: an integration of empirical and explanation-based learning', *Knowledge Acquisition*, Vol. 3, No. 2, pp.157–173.

Qin, X., Pei, R., He, C., Li, F. and Zhang, X. (2025) 'Efficient knowledge distillation and alignment for improved KB-VQA', *Scientific Reports*, Vol. 15, No. 1, p.20682.

Siemer, J. and Angelides, M.C. (1998) 'Towards an intelligent tutoring system architecture that supports remedial tutoring', *Artificial Intelligence Review*, Vol. 12, No. 6, pp.469–511.

Spada, N. and Lightbown, P.M. (2019) 'Second language acquisition', in *An Introduction to Applied Linguistics*, Vol. 20, pp.111–127, Springer, Los Angeles.

Suhm, B., Myers, B. and Waibel, A. (2001) 'Multimodal error correction for speech user interfaces', *ACM Transactions on Computer-Human Interaction*, Vol. 8, No. 1, pp.60–98.

Vuddagiri, R.K., Vydana, H.K. and Vuppala, A.K. (2018) 'Curriculum learning based approach for noise robust language identification using DNN with attention', *Expert Systems with Applications*, Vol. 110, pp.290–297.

Vygotsky, L.S. and Cole, M. (1978) *Mind in Society: Development of Higher Psychological Processes*, Harvard University Press, Cambridge, Massachusetts.

Wang, Y., Wang, Y., Dang, K., Liu, J. and Liu, Z. (2021) 'A comprehensive survey of grammatical error correction', *ACM Transactions on Intelligent Systems and Technology*, Vol. 12, No. 5, pp.1–51.

Yan, B-C. and Chen, B. (2024) 'An effective hierarchical graph attention network modeling approach for pronunciation assessment', *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 1, p.2031.

Yesilyurt, Y.E. (2023) 'AI-enabled assessment and feedback mechanisms for language learning: transforming pedagogy and learner experience', *Transforming the Language Teaching Experience in the Age of AI*, Vol. 1, pp.25–43.

Zhang, J., Zhang, Z., Wang, Y., Yan, Z., Song, Q., Huang, Y., Li, K., Povey, D. and Wang, Y. (2021) 'Speechocean762: an open-source non-native English speech corpus for pronunciation assessment', *INTERSPEECH 2021*, pp.1–5, arXiv:2104.01378.

Zhao, G., Chukharev-Hudilainen, E., Sonsaat, S., Silpachai, A., Lucic, I., Gutierrez-Osuna, R. and Levis, J. (2018) 'L2-ARCTIC: a non-native English speech corpus', *INTERSPEECH 2018*, p.152.