HiHGN: Hierarchical Heterogeneous Graph Network for L2 Word-Level Pronunciation Assessment

Anonymous ACL submission

Abstract

Automatic Pronunciation Assessment (APA) aims to provide accurate quantitative feedback for second-language (L2) learners. Most existing studies focus on English L2 learners, while research on multi-dimensional word-level assessment for Mandarin L2 learners remains scarce. Meanwhile, as current methods mainly concentrate on sentence-level tasks, they are inadequate in capturing deep fine-grained features at phoneme and word levels, and fail to explicitly model cross-level structural dependencies between phonemes and words due to implicit modeling strategies. In this paper, we propose a Hierarchical Heterogeneous Graph Network (HiHGN), which uses a Hierarchical Feature Extractor (HFE) to capture deep fine-grained features and a Heterogeneous Graph Layer (HGL) to explicitly model cross-level structural dependencies. Furthermore, we construct Mandarin Unisound, a multi-dimensional word-level dataset designed for Mandarin L2 pronunciation assessment, which provides fine-grained labels for initials, rhymes, tones, and words on mandarin. Experimental results show that compared with the best baseline, HiHGN achieves an average absolute improvement of 3.7% on four word-level PCC metrics on the English benchmark dataset SpeechOcean762, and 3.3% on five PCC metrics on Mandarin Unisound dataset, demonstrating excellent generality and robustness.[1]

1 Introduction

Driven by the growing demand for foreign language proficiency in academia and the workplace, the need for second language (L2) learning has risen rapidly, promoting the development of Computer-Assisted Pronunciation Training (CAPT) systems. CAPT systems provide intuitive and convenient feedback for L2 learners, effectively helping them improve their pronunciation skills (Moere and Downey, 2016). As a core component of CAPT, the main goal of Mispronunciation Detection and Diagnosis (MDD) is to provide L2 learners with error locations and types (Tong et al., 2016; Leung et al., 2019; Gan and Wei, 2025). In actual learning scenarios, since it is necessary to provide comprehensive and integrated pronunciation feedback for L2 learners, APA, which can generate multi-granularity evaluation feedback, is more widely used in language testing to assist evaluation (Li et al., 2017b; Bannò et al., 2022), which greatly reduces the burden of manual evaluation (Zechner et al., 2009).

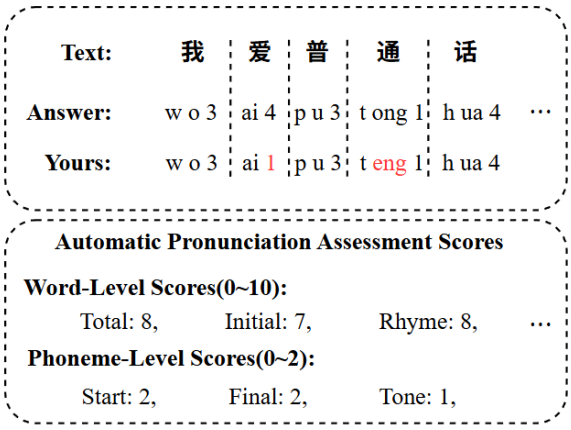

Figure 1: A sample from the Mandarin Unisound dataset (Text: I love Mandarin), demonstrating its multi-granularity and multi-dimensional design. This dataset can also support other related tasks such as MDD and ASR.

In terms of feature construction, early APA research focused on manually annotated single-granularity features, such as phoneme accuracy (Witt and Young, 2000; Li et al., 2017a; Mao et al., 2022), word-level stress (Ferrer et al., 2015; Korzekwa et al., 2021), Mandarin tones (Zhang et al., 2013), and sentence fluency (Yan and Gong, 2011; Cucchiarini et al., 2000; Wade et al., 2025). This method only uses surface features of the target gran-

[1]We provide our code and corpus on GitHub: https://anonymous.4open.science/r/EMNLP-2026-100E

ularity but ignores the interdependencies between different granularities. With the development of deep neural networks, existing research has focused on multi-dimensional and multi-granular pronunciation assessment methods (Do et al., 2023b,a), and the construction of their corpora has also fully considered the complex structure of language (Zhang et al., 2021), providing a unified research paradigm for building APA systems. However, most existing APA studies target English L2 learners and focus on sentence-level evaluation, while research on word-level evaluation for Mandarin L2 learners is still relatively scarce. As a tonal language, Mandarin relies more on the combination of initials, finals, and tones in its language structure(Hu et al., 2014). Compared with English, Mandarin pronunciation is more basic and complex at the word and character level (Liu and Meng, 2016; Qin and Wang, 2014). However, high-quality multi-dimensional and multi-granular word-level datasets for Mandarin L2 learners are extremely scarce, which limits the spread of Mandarin and the development of APA research. In terms of APA methods, existing studies based on the Transformer architecture (Vaswani et al., 2017) have made considerable progress in multi-granular APA evaluation. Model designs are mainly divided into two categories: parallel architecture (Gong et al., 2022; Pei et al., 2024) and hierarchical architecture (Yan et al., 2024; Yan and Chen, 2024; Chao and Chen, 2025), but there are still three key limitations. First, the parallel architecture models fine-grained features of all language levels simultaneously, which makes it difficult to capture the internal connections of each level. Second, the current design of multi-granular APA models still focuses on the sentence level, lacking research on more underlying fine-grained levels such as the word level. Third, existing methods rely on implicitly features modeled that the model learns autonomously, leading to the inability to effectively model the structural relationships between phonemes and words. To address these problems, this paper makes three core contributions:

• We propose a Hierarchical Heterogeneous Graph Network (HiHGN), which processes deep fine-grained features through HFE and HGL, and explicitly modeling the linguistic structural relationships between phonemes and words using graph networks. It is also applicable to the fine-grained pronunciation structures of both Mandarin and English.

Index	**SpeechOcean762**	**Mandarin Unisound**
Total Duration (h)	6	48.69
Language	English	Mandarin
Training Set		
Word Samples	15849	157296
Phoneme Samples	47076	304614
Test Set		
Word Samples	15967	39335
Phoneme Samples	76159	76159

Table 1: Detailed Comparison of SpeechOcean762 and Mandarin Unisound Datasets

• We construct the Mandarin Unisound corpus for Mandarin L2 learners, which is a high-quality multi-dimensional and multi-granular word-level dataset. As shown in Figure 1, this corpus covers the multi-granular structure of initials, rhymes, and tones in Mandarin, and supports APA, MDD, and Automatic Speech Recognition (ASR) tasks.

• Experiments on the two corpora show the effectiveness of HiHGN across different phoneme and word structures in various languages. Specifically, HiHGN achieves an average absolute improvement of 4.5% and 1.2% over the optimal baseline model on the two PCC metrics of phoneme score and word score total for the two corpora, respectively.

2 Related Work

Research on APA methods can be traced back to the late 20th century (Neumeyer et al., 1996; Cucchiarini et al., 1997), evolving from single-granularity manual feature modeling to multi-granularity deep learning-based modeling. At the same time, the construction of corpora covering various second languages is also progressing synchronously, but word-level Mandarin L2 datasets are still severely lacking.

Early APA methods used acoustic models based on Hidden Markov Models (HMM) (Chen et al., 2004; Liu et al., 2009; Ming et al., 2010), constructing evaluation models by manually selecting acoustic features combined with phoneme/word confidence scores and prosodic information (ge et al., 2011). These methods are inefficient, and the manually constructed single fine-grained features such as (Mao et al., 2022; Zhang et al., 2013; Li and Meng, 2012) cannot capture complex pronunciation patterns nor provide multi-dimensional feedback (Kondo and Tsubaki, 2012; Lin and Wang, 2021). With the development of deep neural networks, multi-dimensional and multi-granular APA

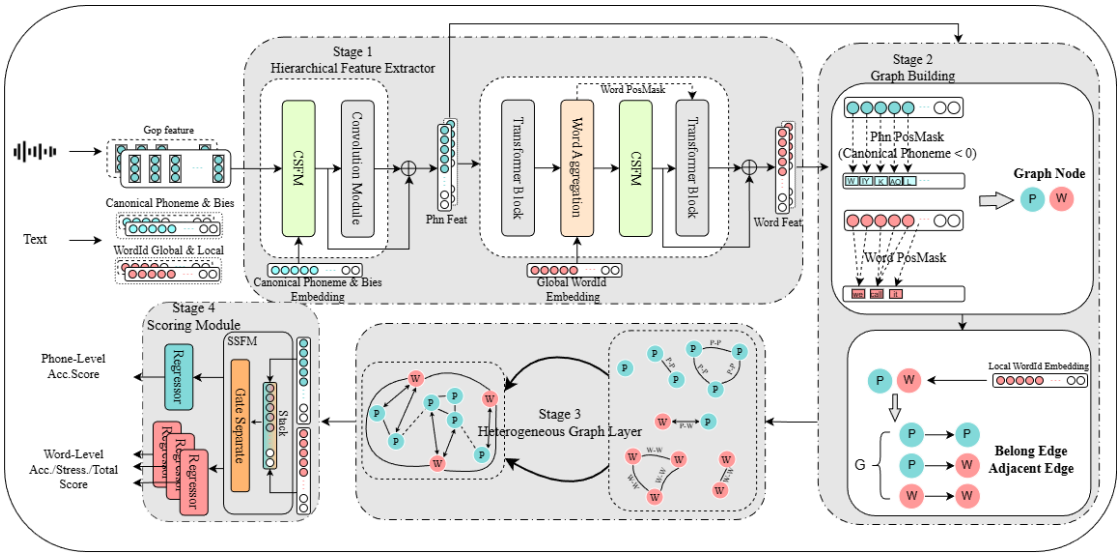

Figure 2: Overall architecture of HiHGN. The framework consists of three components: (a) Hierarchical Feature Extractor with Cross-Scale Fusion module (CSFM) for phoneme and word-level feature processing; (b) Heterogeneous Graph Construction module, including Graph Buliding (stage 2) and Heterogeneous Graph Layer (stage 3) for node updating; (c) Scoring module for word-level assessment.

has become the mainstream. Representative works include Gong et al. (2022) proposing the GOPT architecture for simultaneous multi-level feature modeling, Do et al. (2023b) proposing a new loss function to solve data imbalance, and Yan et al. (2024) proposing the HierTFR hierarchical architecture that extracts internal connections between different granularities through hierarchical modeling. In addition, studies using Self-Supervised Learning (SSL) features to supplement the model with rich semantic information (Chao and Chen, 2025; Chao et al., 2022; Yan et al., 2025) have also provided a new direction for multi-granular APA tasks. However, existing multi-granular research still has obvious limitations: it mainly focuses on sentence-level evaluation of English and relies entirely on implicit feature modeling of neural networks (Fan et al., 2023a; Pang et al., 2026a,b), leading to the model's inability to effectively achieve accurate modeling of cross-level structures between phonemes and words.

In terms of L2 pronunciation assessment corpora, existing corpora already support L2 research such as Dutch (Chen et al., 2004), Korean (Ryu et al., 2016), and Indian English (Pal et al., 2022). In addition, English L2 corpora including Shefce (Ng et al., 2017), Sell-corpus (Chen et al., 2019), Speak & Improve Corpus (Knill et al., 2025), and SpeechOcean762 (Zhang et al., 2021) have been constructed to provide multi-fine-grained training and validation labels for APA research. For Mandarin Chinese, corpora such as iCALL (Chen et al., 2015), SingaKids-Mandarin (Chen et al., 2016), and OMPAL (Hsieh et al., 2025) also provide valuable corpus resources. Despite significant progress, most focus on sentence-level evaluation, and only some English L2 corpora involve word-level multi-granular labels. In contrast, in Mandarin word-level research, there is not only a lack of multi-granular labels for initials, rhymes, and tones as well as the construction of quantitative score labels, but also only two-dimensional labels for the correctness of word-level designs (Fan et al., 2023b; Yang et al., 2023).Currently, the widely used baseline dataset for multi-granular APA research is SpeechOcean762, which is specifically designed for English L2 APA tasks and covers multiple levels including phonemes, words, and sentences. However, there is still no well-designed high-quality multi-dimensional and multi-granular Mandarin word-level L2 corpus available for research, which greatly limits the research on Mandarin word-level pronunciation assessment. Moreover, there is a lack of more fine-grained research on the word level in the corresponding APA research of SpeechOcean762. Therefore, this paper proposes the HiHGN model and constructs the Mandarin Unisound corpus to address the above problems. A detailed comparison between SpeechOcean762 and Mandarin Unisound (the two datasets used in the experiments) is shown in Table 1.

3 Methodology

This chapter elaborates on the overall architecture and core module design of HiHGN, which consists of the Hierarchical Feature Extraction (HFE) module, Heterogeneous Graph Layer (HGL) module, and Scoring Module. The overall architecture of the model is illustrated in Figure 2, and each module is described in detail in the following sections.

3.1 Hierarchical Feature Extraction Module

The core objective of the HFE module is to mine deep fine-grained features at the phoneme and word linguistic levels from raw speech features, respectively, and provide high-quality feature representations for node initialization in the subsequent heterogeneous graph network. The input of the HFE module is the projected GOP feature of raw audio, denoted as $X \in \mathbb{R}^{B\times L\times D}$, where B is the batch size, L is the sequence length, and D is the feature dimension. Meanwhile, the module takes canonical phonemes, BIES phoneme segmentation labels, and global & local WordId $\in \mathbb{R}^{B\times L}$ as inputs, which are used to construct the correspondence between phonemes and words. The HFE module adopts two sub-modules, namely a phoneme-Level feature extractor and a word-Level feature extractor, to achieve feature extraction and optimization for different linguistic levels.

Phoneme-Level Feature Extractor This Module receives the audio feature X and fuses the embeddings of Canonical Phoneme and BIES labels to supplement fine-grained prior information at the phoneme level. To fully capture the global and local features of the temporal sequence, we introduce the Cross-Scale Fusion Module (CSFM, Figure 3) to perform multi-scale deep processing on phoneme features: X is mapped to different feature spaces through projection layers with scales $= (1, 2, 4)$. Parallel multi-head self-attention (MHSA) and Depthwise Separable Convolution (DSC) are applied to extract global semantic and local key information. After multi-scale fusion, a convolution module (kernel sizes 5 and 7) is used to remove redundancy and output the phoneme-level feature P. The detailed computation process of CSFM is shown in Appendix A.1.

Word-Level Feature Extractor This Module takes the phoneme output P as the bottom-level input, and introduces the Global word id $\{\text{word_vocab}_i, \text{word_vocab}_{i+1}, ..., \text{word_vocab}_n\}$ to supplement word-level prior information, where

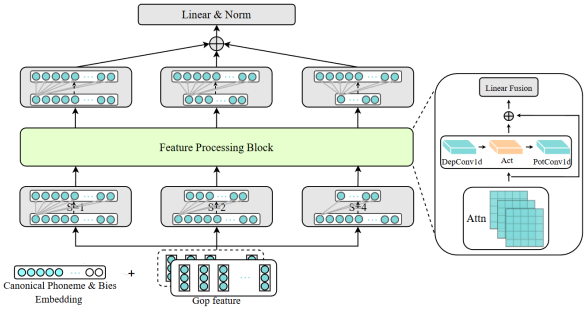

Figure 3: Architecture of the CSFM module. The input features are fed into three mapping layers with different scales for cross-scale feature extraction, and finally fused at the same scale.

n denotes the number of words contained in the sample, and word_vocab_i represents the word dictionary index ID stored according to the audio time-step sequence. The specific pipeline is as follows: first, two layers of MHSA are applied to model coarse-to-fine word features Word feat $\in \mathbb{R}^{B\times L\times D}$. Then, the word aggregation module averages consecutive phoneme feature frames into word feature frames Values_b, which are padded to the maximum number of words $W_{\max}$ in the batch to obtain the aggregated output Word feat $\in \mathbb{R}^{B\times W_{\max}\times D}$. Finally, the CSFM module and MHSA residual connection are adopted to further capture the global and local information of aggregated features, yielding the word-level initialization feature W. The core formulas of the word-level aggregation process are shown below, and the complete derivation is presented in Appendix A.2.

$$W_b = \text{Count}(\text{Global word id}[b, t] \geq 0) \quad (1)$$

$$W_{\max} = \text{Max}(W_b, W_{b+1}, ..., W_B) \quad (2)$$

$$\text{Values}_b = \text{Aggregation}(W_b, \text{word feat}) \quad (3)$$

$$\text{Word feat} = \text{Pad}(\text{Values}_b,\ W_{\max}) \quad (4)$$

3.2 Heterogeneous Graph Layer Module

The Heterogeneous Graph Layer (HGL) module takes the phoneme feature P and word feature W (output by the HFE module) as inputs. Following a unique local heterogeneous graph G construction scheme, it generates initialized phoneme nodes and word nodes $G_g = \{G_p, G_w\}$, as well as three types of edges Edge $= \{G_{p\to p}, G_{p\to w}, G_{w\to w}\}$. Detailed construction details of graph G are provided in Appendix B.

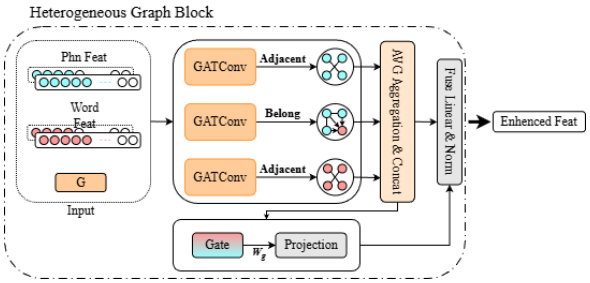

Figure 4: HGL Architecture Diagram.

The core of graph G is to use the Multi-Head Graph Attention Convolution (GATConv) module (Velickovic et al., 2018) to input paired node features G_g based on the inter-node relationships provided by the three edge types. It then updates the attention weights of the two nodes via multi-head attention to generate enhanced nodes $\{\tilde{G}_{p\to p}, \tilde{G}_{p\to w}, \tilde{G}_{w\to w}\}$. These enhanced nodes supplement the model with explicit structural information across the two linguistic levels and reduce attention to nodes outside the three edge structures. The construction formulas for the enhanced nodes are shown below (the core formula derivation of GATConv is presented in Appendix A.3):

$$\tilde{G}_{p\to p} = \text{GATConv}(G_{p\to p}, G_p) \tag{5}$$

$$\tilde{G}_{p\to w} = \text{GATConv}(G_{p\to w}, (G_p, G_w)) \tag{6}$$

$$\tilde{G}_{w\to w} = \text{GATConv}(G_{w\to w}, G_w) \tag{7}$$

To avoid losing global and local audio information during the node update process, we introduce a gated fusion network to weight and fuse the updated enhanced node features with the original projected features. Finally, we output the enhanced phoneme and word features Enhanced Feat $= (\tilde{P}, \tilde{W})$. The architecture of the HGL module is illustrated in Figure 4.

3.3 Scoring Module

Based on the enhanced features $\tilde{P}$ and $\tilde{W}$ output by the HGL module, the Scoring Module adopts regression to generate fine-grained quantitative scores for the phoneme and word linguistic levels, with human expert annotations as the ground truth labels. We use the Mean Squared Error (MSE) as the loss function to measure the discrepancy between the predicted values and expert scores. The overall loss is a weighted sum of the phoneme-level prediction loss L_p and the word-level prediction loss L_w. We introduce weight coefficients L_p^W and L_w^W to control the training weights of the two loss types, achieving collaborative optimization of different tasks:

$$\text{Loss} = L_p^W \cdot L_p + L_w^W \cdot L_w \tag{8}$$

4 Datasets and Experimental Setup

4.1 Datasets

In this study, we conduct word-level pronunciation assessment experiments using two multi-dimensional and multi-granularity datasets in different languages. Specifically, we adopt the public English L2 baseline dataset SpeechOcean762 to verify the model performance on word-level pronunciation assessment. This dataset is designed for multi-dimensional and multi-granularity APA tasks and contains evaluation labels at three linguistic levels: phoneme, word, and sentence. For the word-level pronunciation assessment task focused in this paper, we use its phoneme and word-level labels to conduct corresponding experiments.

In addition, we construct the Mandarin Unisound dataset to verify both the word-level assessment performance and cross-lingual generalization ability of the model. The dataset targets minority L2 Mandarin learners whose first language is Uyghur. During construction, we cover a variety of fine-grained labels including Mandarin initials, rhymes, and tones, as well as quantitative score labels. The dataset ensures a balanced distribution of phoneme-level labels (correct / defective / erroneous) and comprehensive coverage of word-level score intervals (see Appendix C.1). It also supports multi-task training such as MDD and ASR. The dataset construction pipeline is presented in Appendix C.2, and the annotation specifications and quality control details are provided in Appendix C.3.

4.2 Evaluation Metrics and Experimental Setup

To ensure the fairness of experiments across different datasets and models, this study adopts the Pearson Correlation Coefficient (PCC) as the core evaluation metric to measure the linear correlation between the model-predicted scores and expert-annotated scores. A higher PCC value indicates higher prediction consistency.

For the experimental setup, all experiments are conducted on an A40 GPU. To ensure fairness, all baseline models and HiHGN use the same training configuration: AdamW optimizer and TriStageLRScheduler learning rate scheduler (peak

Model	SpeechOcean762					Mandarin Unisound					Param(k)
	Phoneme Score		Word Score			Phoneme Score		Word Score			
	MSE	PCC	ACC	Stress	Total	ACC	PCC	Initial	Rhyme	Total	
SVR	0.160	0.450	–	–	–	–	–	–	–	–	–
LSTM	0.089 ±0.000	0.591 ±0.003	0.514 ±0.003	0.294 ±0.012	0.531 ±0.004	0.616 ±0.002	0.614 ±0.004	0.760 ±0.002	0.596 ±0.003	0.586 ±0.007	461.1
GOPT	0.085 ±0.001	0.612 ±0.003	0.533 ±0.004	0.291 ±0.030	0.549 ±0.002	0.678 ±0.002	0.683 ±0.003	0.731 ±0.002	0.668 ±0.002	0.615 ±0.003	482.5
Gradformer	0.079 ±0.001	0.646 ±0.004	0.598 ±0.006	0.334 ±0.013	0.614 ±0.006	0.730 ±0.004	0.748 ±0.003	0.871 ±0.003	0.799 ±0.004	0.823 ±0.004	554.4
HierTRF	0.081 ±0.000	0.644 ±0.000	0.622 ±0.002	0.325 ±0.022	0.634 ±0.002	0.678 ±0.004	0.684 ±0.003	0.886 ±0.001	0.860 ±0.002	0.839 ±0.002	665.1
HierGAT	0.083 ±0.002	0.640 ±0.005	0.606 ±0.004	0.315 ±0.010	0.621 ±0.004	–	–	–	–	–	–
HiHGN	**0.075** ±0.002	**0.658** ±0.004	**0.633** ±0.005	**0.433** ±0.012	**0.652** ±0.004	**0.738** ±0.003	**0.759** ±0.002	**0.903** ±0.004	**0.868** ±0.003	**0.845** ±0.001	729.4

Table 2: Performance comparison of HiHGN and other baseline models on two datasets (left: results on SpeechOcean762, right: results on Mandarin Unisound).

learning rate 5e-4, initial/final learning rate scale factors 0.01/0.05, three-stage training step ratio 0.1:0.4:0.5). The detailed hyperparameter settings of the model are presented in Appendix D. The original audio is processed in a unified manner: SpeechOcean762 uses GOP features constructed by an acoustic model trained on LibriSpeech, while Mandarin Unisound uses GOP features trained on the DataTang Chinese Mandarin corpus.

5 Experiment Results

5.1 Main Results

This section conducts comparative experiments between HiHGN and six traditional or mainstream pronunciation assessment models, including: 1) Support Vector Regression (SVR (Drucker et al., 1996)); 2) Sequential model (LSTM (Yu et al., 2015)); 3) Multi-task parallel model (GOPT (Gong et al., 2022)); 4) Multi-task convolution-enhanced model (Gradformer (Pei et al., 2024)); 5) Multi-task hierarchical network (HierTRF (Yan et al., 2024)); 6) Hierarchical graph attention architecture (HierGAT (Yan and Chen, 2024)). Among them, method 1) only supports single-granularity prediction, while methods 2)–6) support multi-granularity prediction for both phonemes and words.

The comparison results between HiHGN and six baseline models on the two datasets are presented in Table 2. HiHGN achieves the best performance across all granularity evaluation metrics, with particularly significant improvements on the more challenging word stress assessment task (approximately 30% higher than the second-best model).Compared with traditional baselines such as LSTM and GOPT, which fail to effectively extract deep fine-grained features and cross-level correlations, HiHGN achieves an average absolute performance improvement of 9.7%–18.8%. When compared to two competitive models that still rely on implicit modeling, HiHGN gains an average absolute improvement of 2.8%–4.6% on both datasets by supplementing explicit structural information through the HGL module. For HierGAT, a graph network-based baseline, under the same setting of using only GOP features, HiHGN achieves an average absolute performance improvement of 4.8% on SpeechOcean762 and also demonstrates substantial improvements on Mandarin Unisound (the performance of HierGAT on Mandarin Unisound is

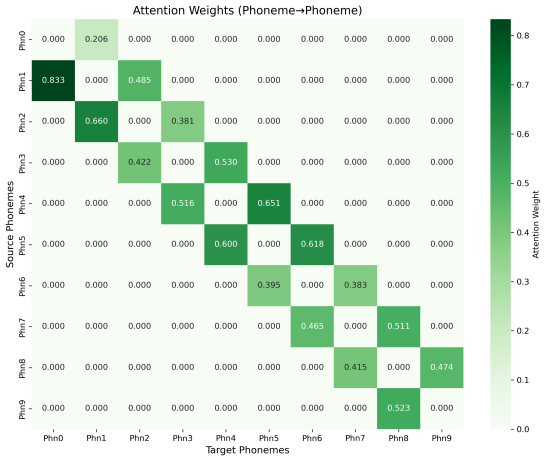

Figure 5: Distribution of updated attention weights for adjacent phoneme edges in HGL (one single sample in SpeechOcean762).

Model	SpeechOcean762					Mandarin Unisound				
	Phoneme Score		Word Score			Phoneme Score		Word Score		
	MSE	PCC	ACC	Stress	Total	ACC	PCC	Initial	Rhyme	Total
HiHGN	**0.075** ±0.002	**0.658** ±0.004	**0.633** ±0.005	**0.433** ±0.012	**0.652** ±0.004	**0.738** ±0.003	**0.759** ±0.002	**0.903** ±0.004	**0.868** ±0.003	**0.845** ±0.001
-HFE	0.089 ±0.003	0.582 ±0.005	0.552 ±0.010	0.305 ±0.011	0.568 ±0.008	0.472 ±0.001	0.244 ±0.003	0.580 ±0.003	0.370 ±0.002	0.332 ±0.004
-CSFM	0.079 ±0.001	0.640 ±0.004	0.625 ±0.002	0.420 ±0.008	0.641 ±0.006	0.734 ±0.002	0.748 ±0.002	0.899 ±0.004	0.861 ±0.003	0.841 ±0.003
-HGL	0.083 ±0.003	0.633 ±0.004	0.619 ±0.008	0.412 ±0.018	0.634 ±0.004	0.713 ±0.005	0.722 ±0.004	0.895 ±0.003	0.850 ±0.002	0.838 ±0.006

Table 3: Ablation experiment results of HiHGN on two datasets (left: results on SpeechOcean762, right: results on Mandarin Unisound).

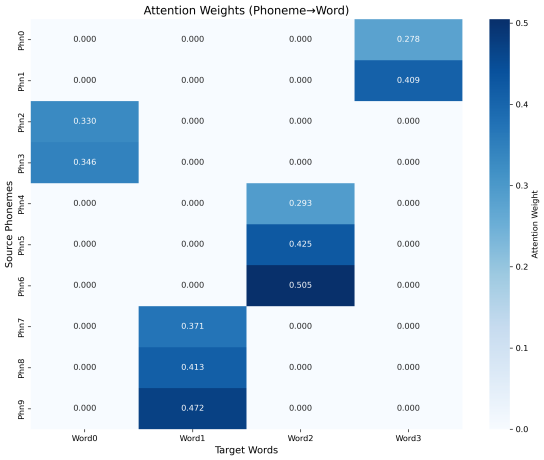

Figure 6: Distribution of updated attention weights for phoneme-word edges in HGL (one single sample in SpeechOcean762).

similar to that shown in Table 3 (-HFE)).Notably, in terms of training parameter statistics on Mandarin Unisound, the parameter count of HiHGN is only 729.4k. Without a significant increase compared to the baselines, HiHGN achieves substantial performance gains, showing high cost-performance for future industrial implementation. A detailed analysis of the comparative experiments is provided in Appendix E.1. To verify the multi-task applicability of the proposed model and dataset, we further conduct joint multi-task experiments, with details presented in Appendix E.2.

5.2 Ablation Experiments

Module Ablation Experiments To verify the effectiveness of each core module in HiHGN, we design four groups of ablation experiments: 1) the full HiHGN model; 2) removing the HFE module (-HFE); 3) retaining HFE and removing the CSFM sub-module (-CSFM); 4) removing the HGL module (-HGL). The experimental results are shown in Table 3. After removing HFE, the model performance drops drastically (the average absolute performance decreases by 9.2% and 42.3% on the two datasets, respectively). The reason is that HFE is responsible for extracting deep features at the phoneme and word levels and providing input features for constructing initialized nodes in HGL. Without HFE, the graph network cannot obtain effective fine-grained features, and the overall performance is greatly affected. After removing CSFM, the model performance decreases slightly but the convergence process fluctuates more significantly (the ablation curve of the CSFM module is shown in Appendix E.3). This is because CSFM performs multi-scale fusion on low-level features, which not only captures deep correlation information but also helps remove feature redundancy and alleviate overfitting caused by data imbalance. After removing the HGL module, the performance on phoneme and word tasks drops by 4% and 3%, respectively, indicating that HGL can effectively enhance inter-level correlations by explicitly modeling the cross-level structural relationship between phonemes and words, and brings positive effects to both language tasks.

Training Hyperparameter Ablation To investigate the optimal configuration of L_p^W and L_w^W in the joint loss, we conduct corresponding ablation experiments to explore the impact of different attention degrees on the model performance for the phoneme and word linguistic levels. Taking the SpeechOcean762 dataset as an example, we use the phoneme PCC and word Total PCC as the core evaluation metrics, and the results are

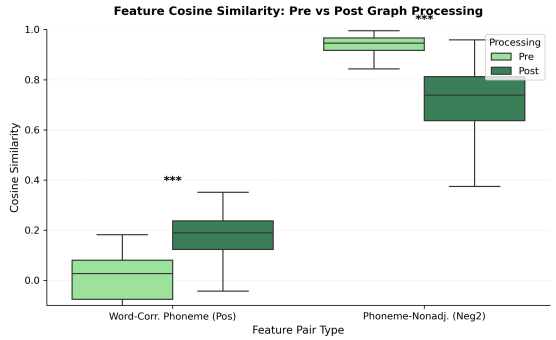

Figure 7: Cosine similarity changes of features before and after HGL processing for a batch of samples (left: word features and their corresponding phonemes (Word-Corr. Phoneme); right: phoneme features and their non-adjacent phonemes (Phoneme-Nonadj.)).

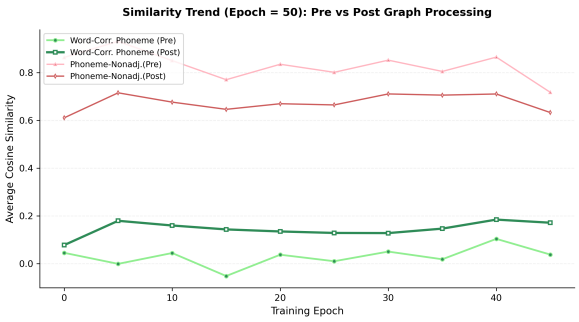

Figure 8: Trend of feature similarity changes before and after HGL processing across 50 epochs.

shown in Appendix E.4. It can be observed that the model achieves better performance when the phoneme-level loss weight is higher than the word-level loss weight. First, from the perspective of linguistic structure, phonemes are the basic units of words, and in Mandarin, they are also the core carriers of initials, finals, and tones. Strengthening the learning of phoneme-level features can provide more precise low-level support for word-level assessment. Second, from the perspective of model architecture, the hierarchical processing of raw features also starts from the phoneme level, which is consistent with the idea that the lower linguistic level is more fundamental and critical, and aligns well with the pronunciation characteristics of Chinese.

5.3 Visualization Analysis

To verify the process of structural relationship construction by the HGL module and explore its positive impact on model learning, we conduct visualization analysis on the attention weight update for two types of nodes in the HGL module and the similarity changes before and after HGL processing, taking phoneme adjacent nodes and phoneme-word cross-level nodes as examples.

The attention weight visualization results (Figure 5 and Figure 6) show that the model assigns higher attention weights to temporally adjacent phoneme nodes, which is consistent with the temporal characteristics of audio. Meanwhile, through Local Word Id mapping, the model assigns higher weights to phoneme nodes belonging to the same word, accurately models the structural relationship between phonemes and corresponding words, and intuitively verifies the effectiveness of the HGL module in explicitly modeling structural correlations.The feature similarity visualization results (Figure 7 and Figure 8) indicate that after HGL processing, the similarity between words and their corresponding phonemes (Word-Corr. Phoneme) is significantly improved, while the similarity between phonemes and their non-adjacent phonemes (Phoneme-Nonadj.) is significantly reduced. All the above changes are statistically significant (p <0.001), which directly validates the effectiveness of explicit structural modeling.In addition, HGL is also applicable in Mandarin Unisound. It effectively updates the attention weights between Mandarin initials, rhymes, tones and words, and weakens the interference of structurally irrelevant relationships. Detailed analysis is presented in Appendix F.1 and F.2.

6 Conclusion

This paper presents HiHGN, a hierarchical heterogeneous graph network for word-level automatic pronunciation assessment. Using a hierarchical feature extractor (HFE) with CSFM and a heterogeneous graph layer (HGL) module, the model extracts fine-grained semantic information and key features at two levels across multiple scales, and explicitly models the relationships between hierarchical representations. This effectively strengthens cross-granularity dependency learning. In addition, we construct the Mandarin Unisound multi-granularity pronunciation dataset for L2 Mandarin learners, which provides fine-grained labels for initials, finals, tones, and words. Experiments on both this two dataset demonstrate the effectiveness, robustness. The results show that HiHGN achieves significant and consistent improvements on both English and Mandarin word-level pronunciation assessment tasks.

Limitations

Model Level Although HiHGN achieves better performance compared with several competitive baseline methods, and indeed possesses certain practical deployment capabilities while improving performance—such as a low parameter count—some existing studies also achieve extremely high performance by fusing features from large pre-trained models such as Wav2vec, WavLM, and HuBERT. For example, Hmamba has demonstrated excellent performance, which is indeed a worthy research topic in the current era dominated by large models. However, the fusion of multiple pre-trained features leads to an exponential increase in model parameters, which greatly reduces the cost-effectiveness in practical applications. At present, the application of large models for pronunciation assessment tasks mainly focuses on how to improve performance, and this limitation has not yet been considered.

Task Level This work focuses on the in-depth processing and research of easily overlooked phone-level and word-level raw audio features, which is indeed consistent with the bottom-up research perspective in linguistics. However, in the actual work and life of L2 learners, pronunciation feedback for other language levels is also extremely important, such as sentence-level and passage-level feedback.

Therefore, in future research, we will explore effective utilization, optimization, and fusion strategies for large pre-trained model features, enabling them to maintain competitive performance at an acceptable computational cost. Furthermore, on the basis of this work, we will continue to study other tasks in pronunciation assessment, so as to better meet the practical application needs of L2 learners in automatic pronunciation assessment systems.

Acknowledgments

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Stefano Bannò, Bhanu Balusu, Mark J. F. Gales, Kate Knill, and Konstantinos Kyriakopoulos. 2022. View-specific assessment of l2 spoken english. In *INTERSPEECH 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18–22 September 2022*, pages 4471–4475. ISCA.

Fu-An Chao and Berlin Chen. 2025. Towards efficient and multifaceted computer-assisted pronunciation training leveraging hierarchical selective state space model and decoupled cross-entropy loss. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1947–1961. Association for Computational Linguistics.

{Fu An} Chao, {Tien Hong} Lo, {Tzu I.} Wu, {Yao Ting} Sung, and Berlin Chen. 2022. 3m: An effective multi-view, multi-granularity, and multi-aspect modeling approach to english pronunciation assessment. In *Proceedings of 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2022*, pages 575–582. Institute of Electrical and Electronics Engineers Inc.

Jiang-Chun Chen, Jyh-Shing Roger Jang, Jun-Yi Li, and Ming-Chun Wu. 2004. Automatic pronunciation assessment for mandarin chinese. In *Proceedings of the 2004 IEEE International Conference on Multimedia and Expo, ICME 2004, 27-30 June 2004, Taipei, Taiwan*, pages 1979–1982. IEEE Computer Society.

Nancy Chen, Rong Tong, Darren Wee, Peixuan Lee, Bin Ma, and Haizhou Li. 2015. icall corpus: Mandarin chinese spoken by non-native speakers of european descent. In *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, 6-10 September 2015*, pages 324–328. ISCA.

Nancy F. Chen, Rong Tong, Darren Wee, Pei Xuan Lee, Bin Ma, and Haizhou Li. 2016. Singakids-mandarin: Speech corpus of singaporean children speaking mandarin chinese. In *INTERSPEECH 2016, 17th Annual Conference of the International Speech Communication Association, Interspeech 2016, San Francisco, CA, USA, September 8-12, 2016*, pages 1545–1549. ISCA.

Yu Chen, Jun Hu, and Xinyu Zhang. 2019. Sell-corpus: an open source multiple accented chinese-english speech corpus for l2 english learning assessment. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7425–7429. IEEE Signal Processing Society SigPort.

C. Cucchiarini, H. Strik, and L. Boves. 1997. Automatic evaluation of dutch pronunciation by using speech recognition technology. In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pages 622–629. IEEE.

Catia Cucchiarini, Helmer Strik, and Lou Boves. 2000. Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *The Journal of the Acoustical Society of America*, 107:989–99.

Heejin Do, Yunsu Kim, and Gary Lee. 2023a. Hierarchical pronunciation assessment with multi-aspect attention. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Heejin Do, Yunsu Kim, and Gary Geunbae Lee. 2023b. Score-balanced Loss for Multi-aspect Pronunciation Assessment. In *INTERSPEECH 2023, 24th Annual Conference of the International Speech Communication Association, Dublin, Ireland, 20-24 August 2023*, pages 4998–5002. ISCA.

Harris Drucker, Christopher J. C. Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. 1996. Support vector regression machines. In *Proceedings of the 10th International Conference on Neural Information Processing Systems*, pages 155–161. MIT Press.

ZhiXing Fan, Jing Li, Aishan Wumaier, Zaokere Kadeer, and Abdujelil Abdurahman. 2023a. A multifaceted approach to oral assessment based on the conformer architecture. *IEEE Access*, PP:1–1.

Zhixing Fan, Jing Li, Aishan Wumaier, Zaokere Kadeer, and Abdujelil Abdurahman. 2023b. A multifaceted approach to oral assessment based on the conformer architecture. *IEEE Access*, 11:28318–28329.

Luciana Ferrer, Harry Bratt, Colleen Richey, Horacio Franco, Victor Abrash, and Kristin Precoda. 2015. Classification of lexical stress using spectral and prosodic features for computer-assisted language learning systems. *Speech Communication*, 69:31–45.

Zhenye Gan and Wenhao Wei. 2025. Mispronunciation detection of mandarin pronunciation errors in tibetan students based on bidfsmn. In *Proceedings of the 2024 2nd International Conference on Information Education and Artificial Intelligence*, pages 218–222, New York, NY, USA. Association for Computing Machinery.

Fengpei ge, L. Li, and Yonghong Yan. 2011. Experimental investigation of mandarin pronunciation quality assessment system. In *2011 International Symposium on Computer Science and Society*, pages 235–239. IEEE Computer Society.

Yuan Gong, Ziyi Chen, Iek-Heng Chu, Peng Chang, and James Glass. 2022. Transformer-based multi-aspect multi-granularity non-native english speaker pronunciation assessment. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7262–7266. IEEE.

Wen-Wei Hsieh, Hao-Wei Chi, Kuan-Chen Wang, Ping-Cheng Yeh, Te hsin Liu, and Chen-Yu Chiang. 2025. Ompal: Bridging speech and learning with an open-source mandarin pronunciation assessment corpus for global learners. In *INTERSPEECH 2025, 26th Annual Conference of the International Speech Communication Association, Incheon, Korea, 14-18 September 2025*, pages 2415–2419. ISCA.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:3451–3460.

Wenping Hu, Yao Qian, and Frank K. Soong. 2014. A dnn-based acoustic modeling of tonal language and its application to mandarin pronunciation training. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3206–3210. IEEE.

Kate M. Knill, Diane Nicholls, Mark J.F. Gales, Mengjie Qian, and Pawel Stroinski. 2025. Introducing the speak & improve corpus 2025: an l2 english speech corpus for language assessment and feedback. In *10th Workshop on Speech and Language Technology in Education (SLaTE)*, pages 167–171.

Mariko Kondo and Hajime Tsubaki. 2012. Fluency and l1 phonology interference on l2 english analysis of japanese aesop corpus. In *Proceedings of the 2012 International Conference on Speech Database and Assessments, Oriental COCOSDA 2012*, pages 123–128. IEEE.

Daniel Korzekwa, Roberto Barra-Chicote, Szymon Zaporowski, Grzegorz Beringer, Jaime Lorenzo-Trueba, Alicja Serafinowicz, Jasha Droppo, Thomas Drugman, and Bozena Kostek. 2021. Detection of lexical stress errors in non-native (l2) english with data augmentation and attention. In *INTERSPEECH 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czech Republic, 30 August–3 September 2021*, pages 3915–3919. ISCA.

Wai-Kim Leung, Xunying Liu, and Helen Meng. 2019. Cnn-rnn-ctc based end-to-end mispronunciation detection and diagnosis. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8132–8136.

Kun Li and Helen Meng. 2012. Perceptually-motivated assessment of automatically detected lexical stress in l2 learners' speech. In *2012 8th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 179–183. IEEE Computer Society.

Kun Li, Xiaojun Qian, and Helen Meng. 2017a. Mispronunciation detection and diagnosis in l2 english speech using multidistribution deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(1):193–207.

Kun Li, Xixin Wu, and Helen Meng. 2017b. Intonation classification for l2 english speech using multi-distribution deep neural networks. *Computer Speech & Language*, 43:18–33.

Binghuai Lin and Liyuan Wang. 2021. Deep feature transfer learning for automatic pronunciation assessment. In *Proceedings of the 2021 International Conference on Interspeech*, pages 4438–4442. ISCA.

Yali Liu and Zihou Meng. 2016. The examination of the relationship between perception and production of mandarin tone of kazak students. In *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 1–5. IEEE.

Yang Liu, Chunting Yang, and Weifeng Ma. 2009. Automatic pronunciation scoring for mandarin proficiency test based on speech recognition. In *Proceedings of the 2009 International Symposium on Intelligent Ubiquitous Computing and Education*, pages 168–171. IEEE Computer Society.

Shaoguang Mao, Frank K. Soong, Yan Xia, and Jonathan Tien. 2022. A universal ordinal regression for assessing phoneme-level pronunciation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 6807–6811. IEEE.

Yue Ming, Qiuqi Ruan, and Xiaoli Li. 2010. Automatic assessment of oral mandarin proficiency based on speech recognition and evaluation. *ICEIT 2010 - 2010 International Conference on Educational and Information Technology, Proceedings*, 3:37–41.

Alistair Van Moere and Ryan Downey. 2016. *Technology and artificial intelligence in language assessment*, pages 341–358. De Gruyter Mouton, Berlin, Boston.

Leonardo Neumeyer, Horacio Franco, Mitchel Weintraub, and Patti Price. 1996. Automatic text-independent pronunciation scoring of foreign language student speech. In *4th International Conference on Spoken Language Processing (ICSLP 1996)*, pages 1457–1460. IEEE.

Raymond W. M. Ng, Alvin C.M. Kwan, Tan Lee, and Thomas Hain. 2017. Shefce: A cantonese-english bilingual speech corpus for pronunciation assessment. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5825–5829. IEEE.

Priyanshi Pal, Chiranjeevi Yarra, and Prasanta Kumar Ghosh. 2022. Voistutor 2.0: A speech corpus with phonetic transcription for pronunciation evaluation of indian l2 english learners. In *2022 25th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6. IEEE.

Xudong Pang, Wenwen Lu, Silajiaihemaiti Ruzemaimaiti, and Aishan Wumaier. 2026a. Automatic speech evaluation method leveraging deep feature fusion. In *Man-Machine Speech Communication*, pages 87–102. Springer Nature Singapore.

Xudong Pang, Aishan Wumaier, Wenwen Lu, Silajiaihemaiti Ruzemaimaiti, and Ligong Lei. 2026b. A multi-task hierarchical deep reinforcement network approach for word level pronunciation assessment. *Expert Systems with Applications*, 317:131930.

Hao-Chen Pei, Hao Fang, Xin Luo, and Xin-Shun Xu. 2024. Gradformer: A framework for multi-aspect multi-granularity pronunciation assessment. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:554–563.

Yi Qin and Guonian Wang. 2014. A computer-aided chinese pronunciation training program for english-speaking learners. In *2014 International Conference on Asian Language Processing (IALP)*, pages 154–157. IEEE.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 28492–28518. PMLR.

Hyuksu Ryu, Hyejin Hong, Sunhee Kim, and Minhwa Chung. 2016. Automatic pronunciation assessment of korean spoken by l2 learners using best feature set selection. In *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1–6. Institute of Electrical and Electronics Engineers Inc.

Rong Tong, Nancy Chen, Bin Ma, and Haizhou Li. 2016. Context aware mispronunciation detection for mandarin pronunciation training. In *INTERSPEECH 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, 8–12 September 2016*, pages 3112–3116.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, volume 30, pages 6000–6010. Curran Associates, Inc.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Papa Sega Wade, Mihai Andries, Ioannis Kanellos, and Thierry Moudenc. 2025. Cbf-afa: Chunk-based multi-ssl fusion for automatic fluency assessment. In *2025 33rd European Signal Processing Conference EUSIPCO*, pages 536–540. IEEE.

S.M Witt and S.J Young. 2000. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, 30(2):95–108.

Bi-Cheng Yan and Berlin Chen. 2024. An effective hierarchical graph attention network modeling approach for pronunciation assessment. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3974–3985.

Bi-Cheng Yan, Jiun-Ting Li, Yi-Cheng Wang, Hsin Wei Wang, Tien-Hong Lo, Yung-Chang Hsu, Wei-Cheng Chao, and Berlin Chen. 2024. An effective pronunciation assessment approach leveraging hierarchical transformers and pre-training strategies. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1737–1747. Association for Computational Linguistics.

Bi-Cheng Yan, Hsin Wei Wang, Fu-An Chao, Tien-Hong Lo, Yung-Chang Hsu, and Berlin Chen. 2025. Hippo: Exploring a novel hierarchical pronunciation assessment approach for spoken languages. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 810–823. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Ke Yan and Shu Gong. 2011. Pronunciation proficiency evaluation based on discriminatively refined acoustic models. *International Journal of Information Technology and Computer Science*, 3:17–23.

Jianlei Yang, Aishan Wumaier, Huawei Han, Cuicui Zhu, and Man Yang. 2023. Multiple ensemble learning methods for mandarin multisyllabic pronunciation assessment. In *2023 IEEE International Conference on Control, Electronics and Computer Technology (ICCECT)*, pages 1030–1034. IEEE.

Zhou Yu, Vikram Ramanarayanan, David Suendermann-Oeft, Xinhao Wang, Klaus Zechner, Lei Chen, Jidong Tao, Aliaksei Ivanou, and Yao Qian. 2015. Using bidirectional lstm recurrent neural networks to learn high-level abstractions of sequential features for automated scoring of non-native spontaneous speech. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2015, Scottsdale, AZ, USA, December 13-17, 2015*, pages 338–345. IEEE.

Klaus Zechner, Derrick Higgins, Xiaoming Xi, and David M. Williamson. 2009. Automatic scoring of non-native spontaneous speech in tests of spoken english. *Speech Communication*, 51(10):883–895.

Junbo Zhang, Zhiwen Zhang, Yongqing Wang, Zhiyong Yan, Qiong Song, Yukai Huang, Ke Li, Daniel Povey, and Yujun Wang. 2021. speechocean762: An Open-Source Non-Native English Speech Corpus for Pronunciation Assessment. In *INTERSPEECH 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czech Republic, 30 August–3 September 2021*, pages 3710–3714. ISCA.

Long Zhang, Haifeng Li, Lin Ma, Jianhua Wang, and Wei Zhang. 2013. Mixed models based pronunciation evaluation of mandarin tone. *Journal of Multimedia*, 8:726–731.

A Core Formulations of HiHGN

A.1 CSFM Core Structure and Computation

The Cross-Scale Fusion Module (CSFM) performs multi-scale projection and enhanced feature fusion to capture fine-grained acoustic information at different receptive fields. First, multi-scale feature projection is conducted as:

$$X_i = W_i \cdot X + b_i \tag{9}$$

where W_i and b_i are learnable parameters of linear projection layers with scaling factors $\{1, 2, 4\}$, and X_i denotes the projected feature at the i-th scale. Then, multi-scale feature extraction and residual fusion are performed as:

$$\widetilde{X}_i = \mathrm{MHSA}(X_i) + \mathrm{DSC}\big(\mathrm{MHSA}(X_i)\big) \tag{10}$$

where X_i denotes the three projected features of different scales, and DSC denotes the depthwise separable convolution block composed of Depconv1d and Potconv1d. The final output $\widetilde{X}_i$ is the cross-scale enhanced feature obtained via a residual connection.

A.2 Word Aggregation Module

The word aggregation module maps phoneme-level sequential features of length L to word-level features with W_b valid words for sample b, by grouping and pooling phoneme frames that belong to the same word. It ensures consistent dimensions across samples in a batch by padding to $W_{\max}$.

The formal definitions in this module have been presented in the main text as Equations 1–4. For clarity, we elaborate their physical meanings and implementation logic as follows.

Equation 1 counts the number of valid words for the b-th sample, where $t \in [0, L-1]$. Global word $\mathrm{id}[b, t] \geq 0$ indicates a valid word, while -1 denotes silence or invalid frames. Equation 2 computes the maximum number of valid words in a batch, which unifies the word dimension for batch processing. Equation 3 aggregates phoneme features within the same word to form a single word-level feature vector. Equation 4 pads all samples to the same length $W_{\max}$ to maintain a consistent shape in batched computation.

Together, these steps convert variable-length phoneme sequences into fixed-length word-level features for subsequent heterogeneous graph learning.

A.3 GATConv Update Formulation

The GATConv layer updates node features via multi-head graph attention to model dependencies between phoneme nodes and word nodes. For node features Gg_i^k and Gg_j^k, where i and j represent two nodes connected by any type of edge, and k denotes the k-th attention head ($k \in [1, K]$), the computations are defined as:

$$e_{ij}^k = \text{LeakyReLU}\Big(W^k \big[W_i^k \cdot Gg_i^k \parallel W_j^k \cdot Gg_j^k\big]\Big) \tag{11}$$

where $\parallel$ denotes feature concatenation, W^k denotes the attention weight matrix of the k-th attention head, and W_i^k and W_j^k are linear projection weights for nodes i and j, respectively. The normalized attention weight α_{ij}^k is computed as:

$$\alpha_{ij}^k = \text{softmax}_j\big(e_{ij}^k\big) = \frac{\exp(e_{ij}^k)}{\sum_{j' \in \mathcal{N}(i)} \exp(e_{ij'}^k)} \tag{12}$$

where $\mathcal{N}(i)$ denotes the set of neighboring nodes of node i. Each node is then updated as:

$$Gg_j^k = \sigma\Bigg(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^k \cdot Gg_j^k\Bigg) \tag{13}$$

where σ denotes the ReLU activation function. Finally, multi-head attention outputs are concatenated:

$$\widetilde{Gg}_i = \bigoplus_{k=1}^{K} Gg_j^k \tag{14}$$

where $\widetilde{Gg}_i$ is the updated node feature, and $\bigoplus$ denotes concatenation across all K attention heads.

B Details of Graph Construction

B.1 Phoneme and Word Node Construction

- Phoneme node G_p: constructed from valid phoneme time steps using Phoneme_PosMask (Canonical Phoneme ≥ 0), with node features taken as the phoneme features P output by HFE.

- Word node G_w: constructed from valid words using Word_PosMask derived from W_b, with node features taken as the word features W output by HFE.

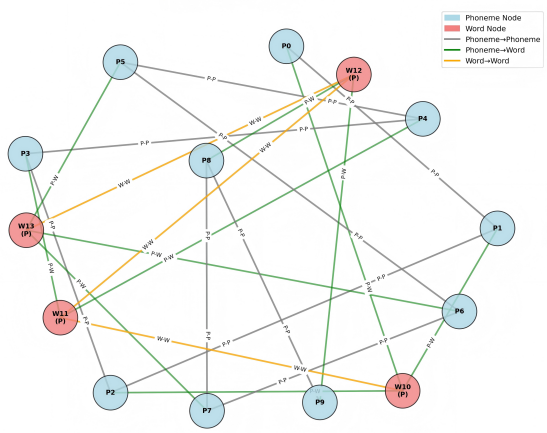

Figure 9: Construction method of the local heterogeneous graph G for a single sample.

B.2 Edge Construction Rules

- G_p adjacent G_p ($G_{p\to p}$): edges built between temporally adjacent valid phoneme nodes G_p to capture sequential dependencies among phonemes.

- G_p belong to G_w ($G_{p\to w}$): edges built between each phoneme node G_p and its corresponding word node G_w according to Local Word ID, to model hierarchical cross-level associations.

- G_w adjacent G_w ($G_{w\to w}$): edges built between temporally adjacent valid word nodes G_w to capture sequential dependencies among words.

B.3 Illustration

Given Local Word Id $= [0, 0, 0, 1, 1, 2, 2, 3, 3, 3]$ (4 words covering steps $[0{-}2]$, $[3{-}4]$, $[5{-}6]$, $[7{-}9]$), the local heterogeneous graph G is illustrated in Figure 9.

C Supplementary Details of the Mandarin Unisound Dataset

C.1 Label Distribution Statistics of Mandarin Unisound

The distribution of phoneme pronunciation standard labels (correct, defective, erroneous, corresponding to [0–2]) in Mandarin Unisound is summarized in Table 4. The word-level pronunciation assessment scores range from [0–10], and their distribution is illustrated in Figure 10.

Index	Mandarin Unisound (Train)	Mandarin Unisound (Test)
Correct	52421	13107
Defect	52522	13130
Error	52353	13098
Count	157296	39335

Table 4: Phoneme-level pronunciation standard label distribution in the Mandarin Unisound dataset.

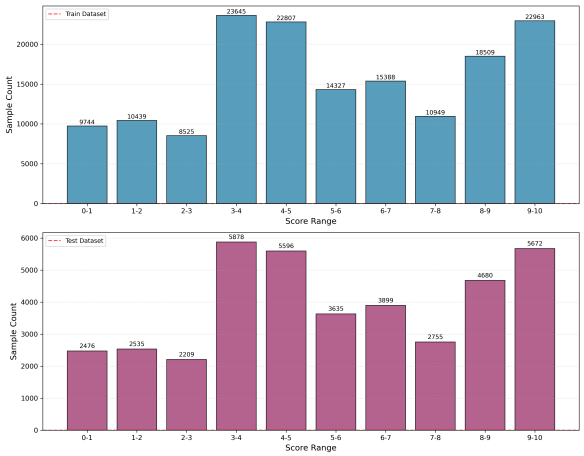

Figure 10: Mandarin Unisound word level pronunciation total score label distribution (training set above, testing set below).

C.2 Dataset Construction Process

The construction of Mandarin Unisound consists of three stages:

1) Basic resource preparation: 146 sets of word-level test items are selected from the Mandarin test question bank to construct a basic dictionary and reference texts. Each set is designed by professional linguists to ensure balanced coverage of phonemes and words.

2) Raw data collection: 100,000 raw test audio utterances are randomly extracted from the background database of the Mandarin test system. All samples are collected from L2 Mandarin learners belonging to ethnic minorities including Uyghur, Kazakh, Mongolian, and Kyrgyz, mainly from Urumqi, Turpan, Hami and other regions in Xinjiang.

3) Data annotation and verification: Six annotators complete data filtering (removal of severely noisy, multi-speaker, and silent segments, as well as screening of pronunciations with different quality levels to ensure each word covers a wide score range), pronunciation labeling, and timestamp alignment (including data segmentation and timestamp alignment of ground-truth initials, rhymes, and tones). The annotated results are further verified by five professional reviewers to ensure the quality of phoneme-level quality labels and word-level assessment score labels.

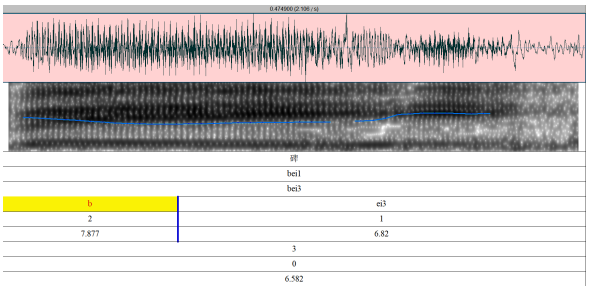

Figure 11: An example of Mandarin monosyllabic annotation.(From top to bottom: reference text, reference pinyin, actual pinyin, actual phonemes (initials and finals), phoneme pronunciation quality score, phoneme assessment score, actual tone, tone pronunciation quality score, and total word score).

C.3 Annotation Guidelines

Figure 11 illustrates a sample annotation. In addition to speech recognition labels (actual pronunciation sequences and phonemes), factors such as diverse pronunciation habits and limited Mandarin proficiency of L2 learners often lead to heavy accents, non-standard pronunciation, and grammatical errors during learning and assessment. Therefore, different from the traditional MDD task, which only provides correct/incorrect labels, we additionally design pronunciation defect labels. Detailed annotation guidelines for phoneme-level pronunciation quality labels and word-level assessment scores used in the MDD and APA tasks are as follows.

Phoneme-level pronunciation quality (3-point scale):

- Correct (label 2): pronunciation fully conforms to standard Mandarin.
- Defective (label 1): pronunciation is generally acceptable but with slight deviations (e.g., inaccurate tone, incomplete initial consonant).
- Erroneous (label 0): pronunciation differs significantly from the standard (e.g., confused initials, missing finals).

Word-level pronunciation assessment (0–10 scale):

- Score [7–10] (all phonemes labeled 2): full intelligibility, standard or nearly standard pronunciation of initials, finals, and tones.

- Score [4–7) (at least one phoneme labeled 1): partial intelligibility, mild non-standard accent or articulatory defects (e.g., inaccurate front/back nasals, confusion of n/l, zh/ch/sh vs. z/c/s).

- Score [0–4) (at least one phoneme labeled 0): poor intelligibility, obvious errors in initials, finals, or tones.

Hyperparameters	Value
Hidden_dim	256
MHSA_Head	8
CSFM_Scale	(1, 2, 4)
DSC_Kernel_size	3
GATConv_Head	4
GATConv_Layer	3
GATConv_Hidden_dim	128
Loss weight ($L_p^W : L_w^W$)	2:1
Learning rate	5e-4
Batch size	25
Epoch	50
Dropout	0.1

Table 5: Detailed hyperparameters of the HiHGN model.

D Detailed Experimental Hyperparameters

Detailed hyperparameters of the HiHGN model are listed in Table 5. To ensure fair comparisons, HiHGN and all baseline and ablation experiments adopt the same hyperparameter settings.

E Supplementary Analysis of Experimental Results

E.1 Supplementary Analysis of HiHGN Comparison Experiments

From the results in Table 2, we draw the following conclusions. First, HiHGN achieves the best performance across all evaluation metrics and granularities on both datasets.

The improvement is especially significant on the challenging word stress assessment task: about 30% relative improvement over the second-best model and up to 48% over basic baselines, verifying that HiHGN enables finer-grained modeling for complex assessment tasks. Second, compared with traditional early models, SVR only supports single-granularity assessment and performs much worse than deep neural networks. Although LSTM and GOPT support multi-granularity evaluation via parallel structures, they fail to capture cross-granularity interactions and thus yield inferior performance. HiHGN outperforms them by absolute creases of 11% and 9.7% on SpeechOcean762, and creases 18.8%/14.7% on Mandarin Unisound, respectively. Third, compared with convolution-augmented Gradformer and hierarchical HierTRF, HiHGN gains absolute improvements of 4.6% and 3.7% on SpeechOcean762, and 2.8%/3.3% on Mandarin Unisound, owing to the explicit cross-granularity structural modeling in the HGL module. This confirms that explicit structural modeling better captures granularity interactions, reduces neural redundancy, and improves learning efficiency compared with implicit modeling. Finally, compared with HierGAT (also graph-based), HiHGN uses the HFE and CSFM to capture finer-grained phonemic and word-level features, and the HGL further enhances structural fusion, yielding a 4.8% average absolute gain on SpeechOcean762, validating the rationality of the HiHGN architecture. Notably, HiHGN has 729.4k trainable parameters, showing no significant growth compared with competing models, yet achieves substantial performance gains, indicating high efficiency for word-level pronunciation assessment.

E.2 Multi-Task Generalization of the Model and Dataset

Table 6 shows the results of joint APA and ASR multi-task training on Mandarin Unisound. All pronunciation assessment metrics are slightly improved after joint training, demonstrating that HiHGN has strong versatility and can be flexibly extended to related tasks.

In addition, pre-trained large-scale speech models have been widely used in various tasks. We select several representative ones, including wav2vec2 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), and Whisper (Radford et al., 2023), and further evaluate their performance on ASR and MDD tasks over the Mandarin Unisound dataset using the same LoRA fine-tuning settings. The results are summarized in Table 7 and Table 8.

Results show that all three models support stable training and reliable evaluation on both tasks, but there is still considerable room for improvement. This is mainly because Mandarin Unisound maintains a balanced distribution across score ranges and pronunciation quality levels, which leads to a relatively high proportion of defective and er-

Model	Phoneme ASR		Phoneme Score		Word Score		
	Accuracy	F1	Accuracy	PCC	Initial	Rhyme	Total
HiHGN (assessment)	–	–	0.738	0.759	0.903	0.868	0.845
HiHGN (assessment+ASR)	**0.928**	**0.713**	**0.760**	**0.782**	**0.907**	**0.873**	**0.855**

Table 6: Comparison of joint multi-task (APA+ASR) experiments of HiHGN on Mandarin Unisound.

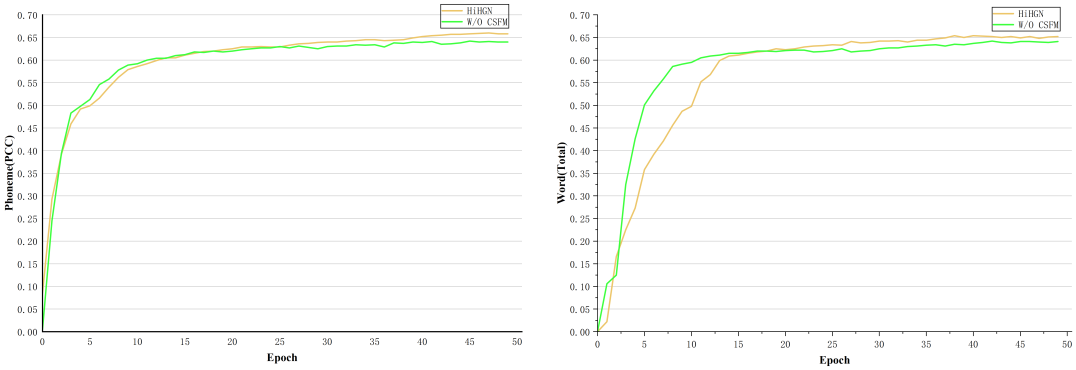

Figure 12: Impact of the CSFM module on model performance and convergence.

ASR Task	PER	CER
whisper-large-v3	26.63%	33.40%
wav2vec2-large	21.57%	32.99%
hubert-large	24.70%	38.35%

Table 7: Performance of pre-trained speech models on the ASR task over Mandarin Unisound.

roneous pronunciations. The simple fine-tuning strategy adopted in our experiments may be insufficient for models to capture complex pronunciation patterns.

Nevertheless, these experiments demonstrate the superiority of our constructed dataset. It can support word-level pronunciation assessment for L2 learners, fill the gap in current L2 speech research, and be applied to multiple tasks simultaneously. Moreover, the rich speaker information with diverse proficiency levels can further help models achieve higher performance.

E.3 Ablation Study on CSFM via Line Charts

The impact of the CSFM module on model performance and convergence is illustrated in Figure 12.

E.4 Comparison of Joint Loss Weights

Table 9 investigates the effects of different loss weight ratios(L_p^W and L_w^W) between phoneme-level and word-level tasks on the SpeechOcean762 dataset.

The model performs best when the phoneme-to-word loss weight ratio is 2:1, showing that strengthening phoneme representation provides more precise low-level support for word-level assessment.

F Supplementary Visualization Analysis

F.1 Attention Weight Visualization

In this section, we visualize the attention weights in the GATConv layer of the HGL module using phoneme adjacency edges and phoneme-word cross-level edges to demonstrate how the model learns structural information. First, for phoneme structural modeling, each phoneme node is connected to its temporal neighbors to strengthen intra-phoneme dependencies. Figure 5 shows the attention weights among phoneme nodes $\{\text{Phn}_0, \ldots, \text{Phn}_9\}$ in a single sample trained on the SpeechOcean762 corpus. Figure 13 presents the attention distribution of adjacent phoneme nodes in a single sample on Mandarin Unisound (note that

MDD Task	ACC (Overall)	F1 (Overall)	(0)-F1	(1)-F1	(2)-F1
whisper-large-v3	77.14%	52.67%	64.61%	65.41%	64.81%
wav2vec2-large	58.20%	56.03%	55.98%	40.95%	71.15%
hubert-large	60.62%	59.05%	60.77%	43.39%	72.99%

Table 8: Performance of pre-trained speech models on the MDD task over Mandarin Unisound.

Model	Phoneme (PCC)	Word (Total)
1:1	0.643	0.638
2:1*	**0.658**	**0.652**
1:2	0.638	0.641
1:3	0.634	0.635
3:1	0.648	0.642

Table 9: Comparison of different loss weight ratios(L_p^W and L_w^W) between phoneme and word tasks.

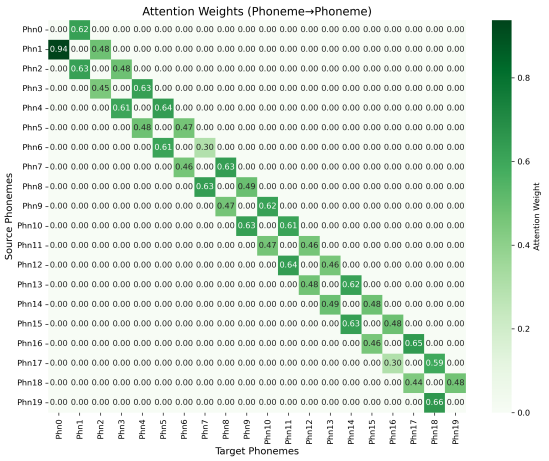

Figure 13: Distribution of updated attention weights for adjacent phoneme edges in HGL (single sample in Mandarin Unisound).

Phn$_6$ denotes the special character sep, which indicates that the word formed by Phn$_6$ is a pure rhyme in Mandarin). GATConv assigns higher attention to temporally adjacent phonemes, which aligns with the sequential nature of speech and effectively captures the temporal structure within phonemes on both Chinese and English datasets. After multi-head attention fusion, the resulting node feature $\boldsymbol{G}_{p\to p}$ effectively encodes the structural information within phonemes.

Second, for cross-level structural modeling, Figure 6 illustrates the attention weights from phonemes to word nodes $\{\text{Word}_0, \text{Word}_1, \text{Word}_2, \text{Word}_3\}$ in the same

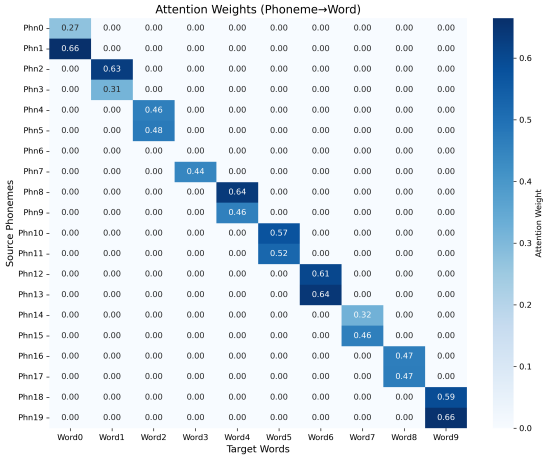

Figure 14: Distribution of updated attention weights for phoneme-word edges in HGL (one single sample in Mandarin Unisound).

sample on SpeechOcean762. Figure 14 shows the attention distribution from phonemes to word nodes on Mandarin Unisound (where Word$_3$ has a null initial, so the attention to Phn$_6$ is 0, consistent with Mandarin pronunciation rules). Based on Local Word ID mapping, HGL automatically allocates high attention to phonemes belonging to the same word and achieves explicit modeling of phoneme–word hierarchical relationships. The fused feature $\boldsymbol{G}_{p\to w}$ well preserves cross-level structural information.

In summary, the attention visualization results verify that HGL can adapt to both English and Mandarin structures, learn meaningful attention weights according to temporal order and phoneme–word alignment, and supplement effective structural information for phoneme and word features.

F.2 Feature Similarity Visualization

To further validate the HGL module, we visualize the cosine similarity of hierarchical features before and after HGL, focusing on two types of pairs:

1) a word and its corresponding phonemes (Word–Corr. Phoneme);

2) a phoneme and its non-adjacent phonemes

(Phoneme–Nonadj.).

We randomly sample a full batch of samples from SpeechOcean762 and plot boxplots of similarity changes (Figure 7). Before HGL (Pre) and after HGL (Post), similarity between words and their corresponding phonemes increases significantly (positive change), as HGL strengthens cross-level alignment. Similarity between non-adjacent phonemes decreases (negative change), since HGL emphasizes local temporal context and weakens irrelevant long-range connections. Both changes are statistically significant at the $p < 0.001$ level (***), confirming that the improvement comes from the HGL module rather than random fluctuation. We also plot similarity changes over 50 epochs (Figure 8), which stabilize across training. This demonstrates that HGL provides robust and consistent structural modeling throughout training.

Ethical Considerations

Potential Risks

This work focuses on L2 Mandarin word-level pronunciation assessment for educational purposes only. The proposed model and the constructed dataset are designed to assist language learning without generating harmful content, privacy leaks, or biased outputs. No significant ethical risks, safety hazards, or misuse potential are identified in this research.

Data Anonymization and Privacy Protection

The Mandarin Unisound dataset is built from speech data of L2 Mandarin learners who signed user consent agreements authorizing audio use for academic research. The dataset is currently available only via application access and not fully public, pending paper acceptance. To protect personal privacy, we currently release only processed GOP acoustic features in `.npy` format, without raw audio or original text. Raw audio and text containing personal information (age, gender, acoustic identity) remain private and are not publicly accessible. Upon paper acceptance, we will release anonymized test audio and de-identified text with no personal information, and include full paper citation information.

Human Annotation, Informed Consent and Recruitment

Data annotation was conducted by trained research team members recruited from the university/laboratory, and annotations were reviewed by graduate students from the School of International Education or School of Linguistics with professional linguistic knowledge. All annotators and reviewers received remuneration funded by the research project. Detailed data annotation procedures and guidelines are presented in Appendix C. The dataset contains no sensitive, offensive, or personally identifiable information. The research team holds usage rights and anonymized publication rights for the collected data. All audio providers signed user agreements granting consent for academic research use only, and the corpus may not be used for commercial purposes.

AI Assistant Usage

AI assistants were only used for paper writing polishing and grammar checking, with no involvement in research design, data collection, or result analysis.