

Multi-granularity Interactive Attention Framework for Residual Hierarchical Pronunciation Assessment

Hong Han, Hao-Chen Pei, Zhao-Zheng Nie, Xin Luo, Xin-Shun Xu*

School of Software, Shandong University, Jinan, China
 sparkinhan@163.com, {202235343, 202435350}@mail.sdu.edu.cn,
 luoxin.lxin@gmail.com, xuxinshun@sdu.edu.cn

Abstract

Automatic pronunciation assessment plays a crucial role in computer-assisted pronunciation training systems. Due to the ability to perform multiple pronunciation tasks simultaneously, multi-aspect multi-granularity pronunciation assessment methods are gradually receiving more attention and achieving better performance than single-level modeling tasks. However, existing methods only consider unidirectional dependencies between adjacent granularity levels, lacking bidirectional interaction among phoneme, word, and utterance levels and thus insufficiently capturing the acoustic structural correlations. To address this issue, we propose a novel residual hierarchical interactive method, HIA for short, that enables bidirectional modeling across granularities. As the core of HIA, the Interactive Attention Module leverages an attention mechanism to achieve dynamic bidirectional interaction, effectively capturing linguistic features at each granularity while integrating correlations between different granularity levels. We also propose a residual hierarchical structure to alleviate the feature forgetting problem when modeling acoustic hierarchies. In addition, we use 1-D convolutional layers to enhance the extraction of local contextual cues at each granularity. Extensive experiments on the speechocean762 dataset show that our model is comprehensively ahead of the existing state-of-the-art methods.

Introduction

In the field of language learning, computer-assisted pronunciation training system (CAPT) (Eskenazi 2009; Tejedor-García et al. 2020), utilizing computer technology to assist language learners in improving their pronunciation skills, provides interactive training methods with immediate feedback. As the core component of CAPT, automatic pronunciation assessment (APA) (Li, Wu, and Meng 2017; Kheir, Ali, and Chowdhury 2023) aims to rate the quality of a speaker’s pronunciation and provides detailed feedback to better assist foreign language learning. Early researches on APA tend to be centered around signal granularity of speech data, such as assessing pronunciation accuracy at phoneme level (Wang and Lee 2012) or detecting various aspect at word or utterance levels (Tepperman and Narayanan 2005; Arias,

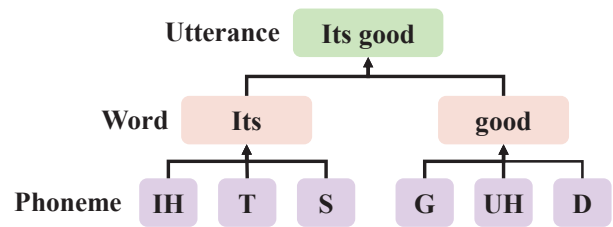


Figure 1: Schematic diagram of the acoustic hierarchical structure with a sample utterance “Its good”.

Yoma, and Vivanco 2010). These single-granularity assessment methods perform well in some specific tasks they are designed to address, but they have many limitations. In particular, they do not take the natural complexity and multi-granularity nature of speech into account (Lin et al. 2020).

The granularities among the pronunciation assessment tasks are not separated from each other (Cincarek et al. 2009), and they have some implicit correlations as shown in Fig. 1. Acoustic signals are typically characterized by their intricate hierarchical structure, with pronunciation results at lower granularity levels affecting higher granularity levels (Al-Barhamtoshy, Abdou, and Jambi 2014). However, modeling a single granularity level cannot fully reveal this implicit relations between different granularity levels.

Recently, to comprehensively study acoustic features at multiple levels of granularity in read-aloud scenario, research endeavors integrate multi-aspect multi-granular pronunciation assessment tasks into a single model to simultaneously evaluate multiple aspects of pronunciation including accuracy, fluency, prosody, and completeness within a unified model across different granularities (i.e., phoneme, word, and utterance).

However, existing methods have some limitations. GOPT (Gong et al. 2022) can effectively handle different granularity scoring tasks when modeling multi-granularity tasks in parallel, but lacks interaction between granularities, which may restrict the modeling of complex correlations between different granularities. HiPAMA (Do, Kim, and Lee 2023) uses a hierarchical structure to capture granularity dependencies, but its information flow is unidirectional, failing to consider bidirectional interaction. Gradformer (Pei et al.

*Corresponding Author

2024) focuses on utterance modeling and fails to capture the correlations between phoneme and word levels. HierGAT (Yan and Chen 2024) uses graph neural networks for hierarchical modeling, but its fixed graph structure limits the dynamic interaction between different granularity levels. As mentioned above, these methods only consider unidirectional relations between adjacent granularities, such as how phonemes form word pronunciations, and lack interactive modeling among phoneme, word, and utterance levels, failing to achieve bidirectional interaction. Additionally, for hierarchical modeling methods, as the granularity level increases, the corresponding model depth also increases, which may lead to the forgetting of initial encoded features.

Bidirectional interaction between different granularities is crucial (Gao et al. 2022). For example, the same word may be stressed differently depending on the utterances in English. The lack of modeling for this pronunciation pattern may be the reason why previous methods perform poorly on word stress.

To address the aforementioned issues, we propose a new residual hierarchical interactive multi-aspect multi-granular pronunciation assessment framework, HIA. Specifically, we design an interactive attention module that enables bidirectional interaction at each granularity level. This module processes the features of each granularity in the acoustic embeddings through the attention mechanism and generates interactive attention heads for each granularity to effectively capture the correlations between different granularities, thereby achieving the bidirectional interaction between granularity levels. Additionally, HIA optimizes the hierarchical structure using a residual connection (He et al. 2016), i.e., introducing acoustic embeddings from the Transformer encoder when modeling the target granularity. By adopting the residual structure, we alleviate the forgetting and processing limitations of the original embedding features caused by the increased depth of the model.

Contributions of this paper are summarized as follows:

- We first note that prior methods perform poorly on word stress, as the same word can be stressed differently across utterances in English, and then introduce the HIA framework to address this limitation.
- To address the issue of insufficient inter-granularity interaction, we design an interactive attention module to enable bidirectional interaction across phoneme, word, and utterance levels, thereby capturing their correlations more effectively and overcoming prior interaction limitations.
- To alleviate the feature forgetting in hierarchical modeling, we propose a residual hierarchical structure, which allows HIA to effectively leverage the hierarchical structure characteristics of speech signals while mitigating the forgetting of initial encoding features by the hierarchical structure, thereby improving the overall performance of the model
- We conduct extensive experiments and analyses on the speechocean762 dataset, experimental results show that our model achieves state-of-the-art performance on all metrics.

Related Work

As the core technology of CAPT research, pronunciation assessment can be simply divided into two categories according to task scenarios: open-response pronunciation assessment and read-aloud pronunciation assessment.

Open-response Pronunciation Assessment

Open-response pronunciation assessment demands the system to handle learners' spontaneous pronunciation without pre-specified texts, making it particularly critical in open-response scenarios, such as IELTS. In these scenarios, learners must accomplish free or semi-free pronunciation tasks through oral expression, which poses higher demands on speech assessment technology.

In this field, the MultiPA (Chen, Yu, and Hirschberg 2024) represents a significant advancement. In concrete terms, the model leverages pre-trained self-supervised learning models and Automatic Speech Recognition (ASR) models to identify potential words. In addition, researchers are also exploring methods for scoring that do not rely on ASR. Cheng et al. (2020) investigated an ASR-free scoring approach that is derived from the marginal distribution of raw speech signals. Cheng et al. (2023a) proposed a novel ASR-free approach for automatic fluency assessment using self-supervised learning.

Read-aloud Pronunciation Assessment

Unlike open-response pronunciation assessment, in read-aloud pronunciation assessment tasks, learners are required to read pre-specified text in a read-aloud scenario.

In early researches on APA, Witt et al. (2000) proposed a phoneme-level pronunciation scoring and evaluation method to derive the posterior probabilities of phonemes, thus assessing the "Goodness of Pronunciation" (GOP). Hu et al. (2015) enhanced mispronunciation detection and diagnosis (MDD) (Strik et al. 2009; Li, Qian, and Meng 2017) by employing an acoustic model trained with deep neural networks and a transfer-learning-based logistic regression classifier. Although such approaches are interpretable, they implicitly assumed that different granularity levels are independent which leads to suboptimal performance.

With breakthroughs in neural network architectures and optimization algorithms (Vaswani et al. 2017; Gao et al. 2017), the research has shifted towards multi-aspect Multi-granularity pronunciation assessment. A notable research of this transition is the GOPT (Gong et al. 2022), which introduces an innovative Transformer-based multi-task learning framework, achieving better results than a single-task-specific assessment task. Building on GOPT, Do et al. (2023) proposed the HiPAMA, which adopts hierarchical structure to sequentially assess pronunciation at various granularity levels. Furthermore, Pei et al. (2024) introduced the Gradformer with granularity-decoupled structure, which incorporates a convolution-enhanced Transformer encoder to encode acoustic features. In addition to GOP based methods, several studies have used non-GOP methods such as transfer learning and self-supervised learning (Kim et al. 2022; Chao et al. 2023; Lin and Wang 2023) to cope with limited L2 training data.

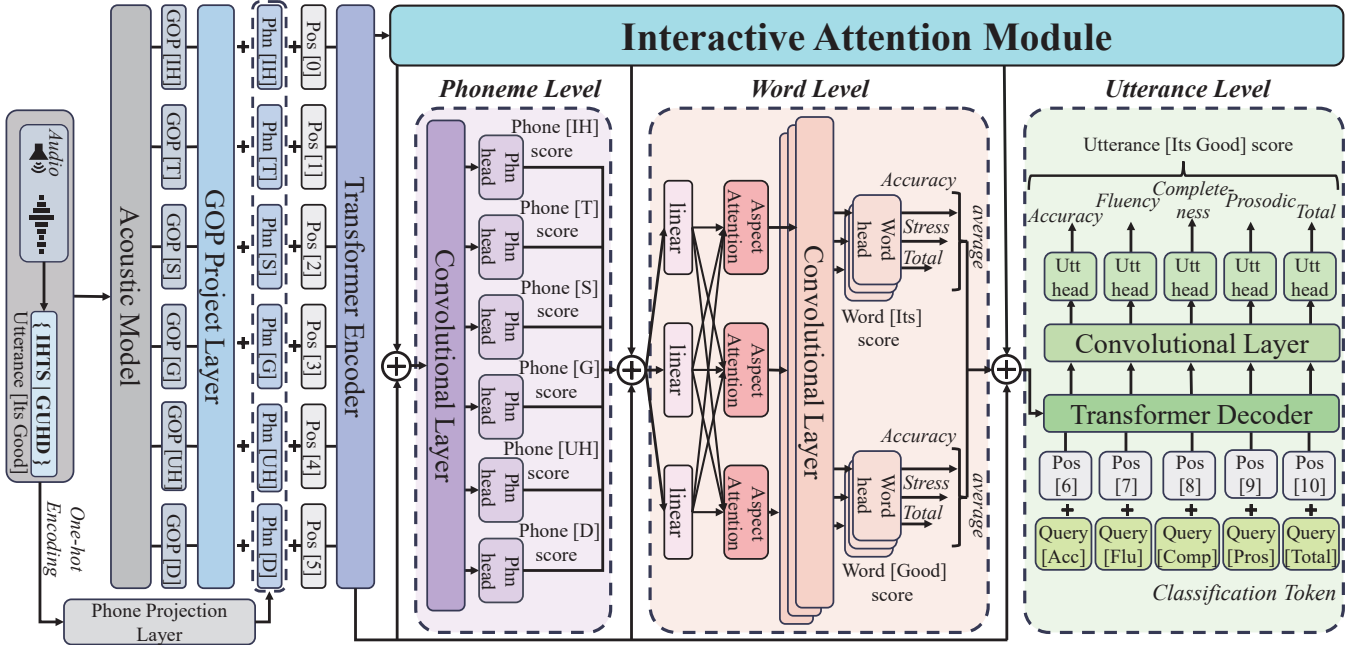


Figure 2: Main architecture of HIA. HIA takes the GOP features extracted from the acoustic model and the projected canonical phoneme embeddings as input. Then, Transformer encoder is applied to encode the input to obtain the acoustic embeddings. Finally, the integrated residual hierarchical structure is used to obtain the scores at each granularity in turn.

Methodology

Overview

As shown in Fig. 2, our model adopts residual structure, namely utilizes acoustic feature embeddings initially encoded by the Transformer encoder for each granularity. These embeddings are combined with the output of the interactive attention heads for each granularity. For word- and utterance-level granularity, we also incorporate scoring results from the phoneme- and word-level, respectively, to model the hierarchical structure. Each component is detailed in the following subsections.

Acoustic Feature Processing

For fair comparison, we follow the baseline model (Gong et al. 2022) to use GOP features (Tu et al. 2018; Shi, Huo, and Jin 2020) as input to the model. In our experiments, ASR acoustic model is used to extract GOP feature which is the log phone posterior (LPP) and log posterior ratio (LPR) defined in (Hu et al. 2015). Specifically, the LPP of a phone p is defined as follows:

$$P(p|o_t) = \sum_{s \in p} P(s|o_t), \quad (1)$$

$$LPP(p) \approx \frac{1}{t_e - t_s + 1} \sum_{t=t_s}^{t_e} \log P(p|o_t), \quad (2)$$

where o_t is the input observation of the frame t , s is the state belonging to the phone p ; t_s and t_e are the start and end frame indexes, respectively. LPR of a phone p_j versus p_i is

defined as:

$$LPR(p_j|p_i) = \log P(p_j|\mathbf{o}; t_s, t_e) - \log P(p_i|\mathbf{o}; t_s, t_e). \quad (3)$$

The Librispeech (Panayotov et al. 2015) acoustic model we use to process audio and generate forced alignment has a total of 42 pure phones, thus the GOP feature of phone p can be defined as an 84-dimensional vector as follows:

$$[LPP(p_1), \dots, LPP(p_{42}), LPR(p_1|p), \dots, LPR(p_{42}|p)]. \quad (4)$$

Considering that different phonemes exhibit distinct characteristics, we use the canonical phoneme embedding to provide useful information same as the baseline model (Gong et al. 2022). Then, we add the projected GOP feature, canonical phoneme embedding, and a trainable positional embedding together and input them to the Transformer encoder.

Interactive Attention Module

In the field of multi-aspect multi-granularity pronunciation assessment, effectively leveraging correlations among granularities is critical for accurately predicting pronunciation scores. Previous studies have only considered unidirectional relations between adjacent granularities (i.e., phoneme \rightarrow word \rightarrow utterance), thereby neglecting the bidirectional correlations between multiple granularities. For the first time, we introduce an Interactive Attention Module that jointly encodes all pairwise bidirectional interaction within a single self-attention operation, thereby enabling simultaneous bottom-up and top-down information exchange across phoneme, word, and utterance levels as shown in Fig. 3.

First, we initialize a set of query vectors for each granularity by projecting acoustic feature embeddings, referred

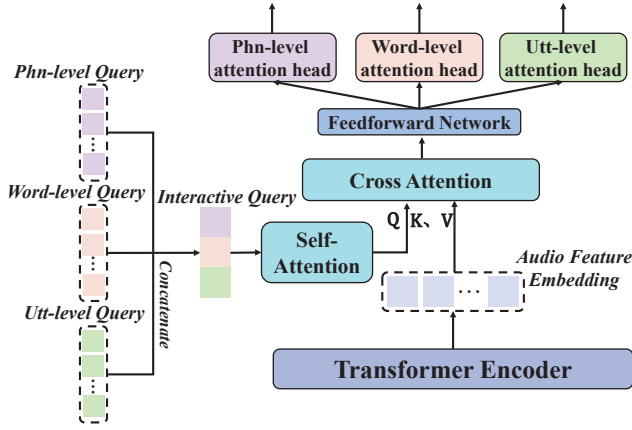


Figure 3: Network structure of interactive attention module. For simplicity, the residual connection and norm layers are omitted. Phn is Phoneme, Utt is Utterance.

to as $Q^l \in \mathbb{R}^{B \times D}$, here l represents different granularity level, B is the batch size, and D is the embedding dimensions for each granularity respectively. These queries represent the initial states of different granularities. Subsequently, we concatenate the multi-granularity queries as $Q = \{Q^{phn}, Q^{word}, Q^{utt}\}$, $Q \in \mathbb{R}^{B \times 3 \times D}$. Then, the self-attention mechanism is applied to Q not only enables bidirectional interaction between different granularity levels but also effectively captures the correlations within each granularity level, generating self-attention heads for concatenated multi-granularity query as follows:

$$Q_{self} = SelfAttn(Q). \quad (5)$$

By concatenating multi-granularity queries and introducing the self-attention mechanism, the model achieves bidirectional interaction between different granularity levels. The generated self-attention heads not only contain multi-granularity bidirectional interaction features but also preserve each level's native cues, further improving the model's performance in multi-granularity tasks.

Subsequently, we input self-attention heads together with the pre-processed acoustic feature embeddings, into the cross-attention mechanism. The self-attention heads Q_{self} serve as queries, while the acoustic feature embeddings X act as keys and values, thereby mapping multi-granularity interaction features to the acoustic feature space. The operation is formulated as follows:

$$Q_{cross} = CrossAttn(Q_{self}, X). \quad (6)$$

Finally, the output of the cross-attention mechanism is fed into a feed forward network and the output H is projected to obtain interactive attention heads for each granularity, denoted as H^{phn} , H^{word} and H^{utt} . Subsequently, the interactive attention heads for each granularity are used in the corresponding granularity's modeling process, enabling more precise scoring in multi-granularity pronunciation assessment tasks.

Residual Hierarchical Multi-granularity Modeling

Phoneme-level Modeling The output X of the Transformer encoder is added to the phoneme-level attention heads in the interactive attention module. Finally, the fused features are input into the convolutional layer, which further extracts and refines the phoneme-level features by learning the characteristic patterns of local regions (Abdel-Hamid et al. 2014).

As shown in Fig. 2, we add phoneme-level regression heads after the output of each corresponding phoneme in the convolutional layer. Thereinto, each phoneme has a phoneme-level regression head, a 48×1 linear layer with layer normalization, that outputs phoneme-level accuracy scores. The model outputs phoneme-level scores that reflect the learner's pronunciation quality in terms of phoneme accuracy. The formula modeling process is as follows:

$$S^{phn} = Conv(X + H^{phn}). \quad (7)$$

Word-level Modeling There is a high correlation between phoneme level and word level, so we leverage phoneme-level scores to calculate word-level scores. Specifically, we first sum the output X of the Transformer encoder, the phoneme-level scoring results S^{phn} , and the word-level attention head H^{word} as word-level inputs:

$$X^{word} = X + S^{phn} + H^{word}. \quad (8)$$

There are many different aspects of word-level granularity, and scores of multiple aspects are related to each other and affect each other. In Word-level Modeling, we use the aspect attention mechanism (Do, Kim, and Lee 2023; Ridley et al. 2021) to capture the correlations between different aspects of the same granularity as well as the difference between the different aspects of scoring:

$$S^{word} = AspectAttn(X^{word}). \quad (9)$$

Similar to the phoneme level, we add convolutional layer and regression heads to output the final accuracy, stress, and total score.

Utterance-level Modeling The Transformer decoder is only used at the utterance level because the utterance level involves complex contextual information, and the decoder can capture long-range dependencies and global features (Pei et al. 2024). Therefore, we use the decoupling method to model the utterance-level scoring task. First, we initialize a set of learnable vectors as queries, $Q^{utt} = \{q_k^{utt}\}_{k=1}^N$, N is the number of utterance-level aspects. Then, the word-level scoring results S^{word} , the output X of the Transformer encoder, and the utterance-level attention head H^{utt} in the interactive attention module are summed up as the key and value into the Transformer decoder. The formula is defined as follows:

$$X^{utt} = X + S^{word} + H^{utt}, \quad (10)$$

$$S^{utt} = TransDecoder(Q^{utt}, X^{utt}). \quad (11)$$

Finally, the output of Transformer decoder is first processed by convolutional layer and regression heads are added to predict the final utterance-level scores.

Model	Phoneme score		Word score (PCC)			Utterance score (PCC)				
	MSE↓	PCC↑	Acc↑	Stress↑	Total↑	Acc↑	Comp↑	Fluency↑	Prosodic↑	Total↑
Human	-	0.555	0.589	0.212	0.602	0.618	0.658	0.665	0.651	0.675
RF (Zhang et al. 2021)	0.130	0.440	-	-	-	-	-	-	-	-
SVR (Zhang et al. 2021)	0.160	0.450	-	-	-	-	-	-	-	-
UOR (Mao et al. 2022)	0.120	0.520	-	-	-	-	-	-	-	-
Mixup-pretrain (Fu et al. 2022)	-	-	-	-	0.610	-	-	-	-	-
Deep feature (Lin and Wang 2021)	-	-	-	-	-	-	-	-	-	0.720
Wav2vec2-based (Lin and Wang 2023)	-	-	-	-	-	-	-	-	-	0.725
LAS (Liu et al. 2023b)	-	-	-	-	-	-	-	-	-	0.766
LSTM (Gong et al. 2022)	0.089 ±0.000	0.591 ±0.003	0.514 ±0.003	0.294 ±0.012	0.531 ±0.004	0.720 ±0.002	0.076 ±0.086	0.045 ±0.002	0.747 ±0.005	0.741 ±0.002
GOPT (Gong et al. 2022)	0.085 ±0.001	0.612 ±0.003	0.533 ±0.004	0.291 ±0.030	0.549 ±0.002	0.714 ±0.004	0.155 ±0.039	0.753 ±0.008	0.760 ±0.006	0.742 ±0.005
HiPAMA (Do, Kim, and Lee 2023)	0.084 ±0.001	0.616 ±0.004	0.575 ±0.004	0.320 ±0.021	0.591 ±0.004	0.730 ±0.002	0.276 ±0.177	0.749 ±0.001	0.751 ±0.002	0.754 ±0.002
Gradformer (Pei et al. 2024)	0.079 ±0.001	0.646 ±0.004	0.598 ±0.006	0.334 ±0.013	0.614 ±0.006	0.732 ±0.005	0.318 ±0.139	0.769 ±0.006	0.767 ±0.004	0.756 ±0.003
HIA (Ours)	0.076 ±0.001	0.657 ±0.004	0.613 ±0.003	0.436 ±0.043	0.628 ±0.005	0.743 ±0.002	0.354 ±0.131	0.778 ±0.006	0.784 ±0.003	0.764 ±0.002

Table 1: The results of HIA and compared baselines on various pronunciation assessment tasks with average MSE (phoneme level) and PCC (phoneme, word, and utterance level) scores and standard deviations of five different runs.

Loss Function

In this work, we use mean squared error (MSE) loss as loss function, which is widely used for pronunciation assessment. The formula is as follows:

$$L_{MSE} = \frac{1}{N} \sum_{i=1}^N (s_i - y_i)^2, \quad (12)$$

where N is the number of samples, s_i is the i -th prediction score of the model, and y_i is the i -th ground truth.

As the reason of multi-aspect and multi-granularity pronunciation assessment task, we consider the total loss is calculated as the sum of each granularity level loss, and the loss at each granularity level is an average sum of corresponding multiple aspects:

$$L_{total} = \sum_{i=1}^M \frac{1}{N} \sum_{j=1}^N L_{ij}, \quad (13)$$

here M and N refer to the total number of granularity levels and corresponding aspect levels, respectively.

Experiments

Dataset

Speechocean762 (Zhang et al. 2021), currently the only open-source standard dataset designed specially for pronunciation assessment in read-aloud scenario, is used for our experiments. It consists of 5000 English sentences and the recorders are 250 non-native English speakers, half of whom are children.

In addition, this dataset has a rich variety of data annotation types, independently annotated by five experts at the phoneme, word, and utterance levels. Specifically, for each utterance, it provides five utterance-level aspect scores: accuracy, fluency, completeness, prosody, and total score

(ranging from 0-10). For each word, it provides three word-level aspect scores: accuracy, stress, and total score (ranging from 0-10). For each phoneme, it also provides an accuracy score (ranging from 0-2). In the experiments, the scores for word and utterance are uniformly rescaled to (0-2), making them on the same scale as the phoneme scores.

Evaluation Metrics

We use MSE to measure the difference between predicted scores and truth scores for phoneme level, the formula is shown in Eq. (12).

Pearson Correlation Coefficient (PCC) is also used as evaluation metric to measure the correlation of predicted values and labeled values of different aspects at each granularity level, it can be calculated as follows:

$$PCC(S, Y) = \frac{\sum_{i=1}^N (s_i - \bar{s})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (s_i - \bar{s})^2} \sqrt{\sum_{i=1}^N (y_i - \bar{y})^2}}, \quad (14)$$

where s_i , y_i are the i -th predicted score given by our proposed model and corresponding true score given by the experts, respectively, N is the total number of sentences.

Experimental Setup

Training Configuration For the model training phase, we use Adam optimizer to train the HIA and initialize the learning rate to 1e-3, the learning rate is halved every 5 epochs after the 20th epoch. The maximum number of epoch is set to 100 and we save the model with the minimum phoneme-level MSE loss as the optimal model. For all experiments, we perform five times with different random seeds for all models, whose mean and standard deviation are reported.

Model Configuration In HIA, the layers of Transformer encoder and decoder are set to 3 and their embedding dimensions are 48. For dimension matching, the embedding

Model	Phoneme score		Word score (PCC)			Utterance score (PCC)				
	MSE↓	PCC↑	Acc↑	Stress↑	Total↑	Acc↑	Comp↑	Fluency↑	Prosodic↑	Total↑
w/o P w/o W w/o U	0.085 ±0.000	0.626 ±0.006	0.551 ±0.004	0.335 ±0.020	0.605 ±0.006	0.717 ±0.003	0.272 ±0.159	0.751 ±0.003	0.754 ±0.003	0.748 ±0.004
w/o P w/ W w/ U	0.083 ±0.001	0.621 ±0.005	0.606 ±0.006	0.429 ±0.024	0.617 ±0.007	0.737 ±0.005	0.344 ±0.118	0.765 ±0.004	0.765 ±0.005	0.758 ±0.003
w/ P w/o W w/ U	0.079 ±0.000	0.661 ± 0.005	0.569 ±0.004	0.328 ±0.023	0.604 ±0.006	0.734 ±0.002	0.322 ±0.105	0.765 ±0.005	0.771 ±0.004	0.759 ±0.004
w/ P w/ W w/o U	0.080 ±0.001	0.653 ±0.004	0.615 ± 0.003	0.421 ±0.011	0.621 ±0.006	0.723 ±0.001	0.302 ±0.074	0.754 ±0.003	0.753 ±0.003	0.754 ±0.002
w/ P w/ W w/ U (HIA)	0.076 ± 0.001	0.657 ±0.004	0.613 ±0.006	0.436 ± 0.043	0.628 ± 0.007	0.743 ± 0.002	0.354 ± 0.151	0.778 ± 0.003	0.784 ± 0.004	0.764 ± 0.002

Table 2: Ablation results on the effectiveness of Interactive Attention Module. P, W and U denote the interactive attention heads generated by the interaction attention module at the phoneme-, word-, and utterance-level granularity, respectively.

Model	Phoneme	Stress	Word	Utterance
HIA	0.657 ± 0.004	0.436 ± 0.043	0.628 ± 0.007	0.764 ± 0.002
-Res	0.647 ±0.007	0.382 ±0.021	0.603 ±0.009	0.748 ±0.003
-Hi	0.645 ±0.001	0.374 ±0.016	0.593 ±0.001	0.753 ±0.003

Table 3: Ablation results on the effectiveness of Residual Hierarchical structure. Res denotes residual structure, Hi denotes hierarchical structure. Because of the space limitation, only the PCC of phoneme accuracy, word stress, word total and utterance total scores are reported.

dimension of interactive attention module query is also set to 48. Due to the dataset and feature dimension are not large enough, we set the number of heads for self-attention and cross-attention in interactive attention module and Transformer to 1. The dropout ratio is set to 0.1 to suppress overfitting. In addition, the kernel size of convolutional layers is set to 5 for each granularity and stride is set to 1.

Results and Discussions

Main Results

In this section, we compare our proposed HIA with traditional single-granularity scoring models and state-of-the-art multi-aspect multi-granularity scoring baseline models, all baseline results are quoted from their original papers and summarized in Table 1. According to the results, we have the following observations:

- Our model outperforms the evaluation results of human experts in all but the utterance-level completeness metric. This gap is mainly attributed to the distributional bias in the dataset, in which 4975 out of 5000 sentences in the dataset have completeness scores of 10.
- Compared with the single-granularity scoring methods, HIA demonstrates significant performance advantages in

Layer	Phoneme	Stress	Word	Utterance
0 layer	0.638 ±0.007	0.415 ±0.011	0.601 ±0.012	0.754 ±0.003
1 layer*	0.657 ± 0.004	0.436 ± 0.043	0.628 ± 0.007	0.764 ± 0.002
2 layers	0.646 ±0.002	0.427 ±0.023	0.618 ±0.004	0.759 ±0.004
3 layers	0.645 ±0.008	0.421 ±0.007	0.617 ±0.013	0.755 ±0.005

Table 4: Ablation results on the effectiveness of the number of convolutional layers. * denotes the setting used in HIA model.

all metrics except the total score at utterance level. This suggests that the multi-aspect multi-granularity scoring approach can better utilize the different inter-granularity correlations and dependencies in audio data.

- Compared with multi-aspect multi-granularity scoring baseline models, our model consistently achieved the state-of-the-art results. Its highest PCC scores highlights the ability of HIA to handle complex pronunciation features, and demonstrates our proposed model is capable of processing and evaluating articulatory features of different granularities more effectively.

Ablation Studies

In order to delve deeper into the key factors that enhance the effectiveness of HIA, we conduct ablation experiments to study the effects of the interactive attention module, the residual hierarchical structure, the number of convolutional layers and the model Configuration on model performance.

Interactive Attention Module Ablation To validate the effectiveness of the interactive attention module, we conduct ablation study on interactive attention heads at each granularity level, and the results are shown in Table 2. The first row represents the ablation of all granularity interactive at-

Setting	Phoneme	Stress	Word	Utterance
<i>Embedding Size</i>				
24	0.649	0.420	0.613	0.752
48*	0.657	0.436	0.628	0.764
96	0.654	0.432	0.611	0.762
<i>Number of Heads</i>				
1*	0.657	0.436	0.628	0.764
2	0.652	0.431	0.618	0.759
4	0.648	0.433	0.623	0.751

Table 5: Ablation results on different model configuration. * denotes the setting used in HIA model.

tention heads, with only the residual hierarchical structure used to score each granularity.

It can be seen that using the corresponding interactive attention heads at each granularity level (rows 2 to 4) improves the performance of metrics at each granularity level, demonstrating the interactive attention module benefits each granularity level. In particular, using word-level attention heads significantly improves performance on word stress, validating the correctness of using the interactive attention module for bidirectional interaction modeling. Using interactive attention heads at all granularity levels (row 5) achieves the best performance, further confirming that the interactive attention module can effectively capture the interdependencies between different granularity levels.

Residual Hierarchical Structure Ablation As shown in Table 3, after ablating the residual connection, all the metrics decline to some extent, especially for the word stress scores, validating the residual structure affords overall performance improvements. The hierarchical structure is embodied in score passes between adjacent granularities, so we ablate the hierarchical structure by removing score passes. It can be seen that the removal of the hierarchical structure also resulted in decreases on all metrics, which further confirms the importance of the hierarchical structure in improving model performance.

Convolutional Layer Ablation: To address the neglect of local context clues that may result from the feature extraction process, we introduce convolutional layer to enhance the model’s ability to capture local features. Table 4 shows that compared with not using convolutional layers (0 layer), HIA achieves performance improvements on all pronunciation assessment metrics with the introduction of convolutional layers. It begins to decline with 2 and 3 convolutional layers, because the the dataset is not large enough and the increased parameters are difficult to optimize.

Model Size Ablation: To investigate the impact of model capacity on performance and assess the scalability of HIA, we conduct ablation studies on two configuration parameters: embedding size and number of attention heads. As shown in Table 5, increasing the embedding size from 24 to 48 leads to consistent improvements across all metrics.

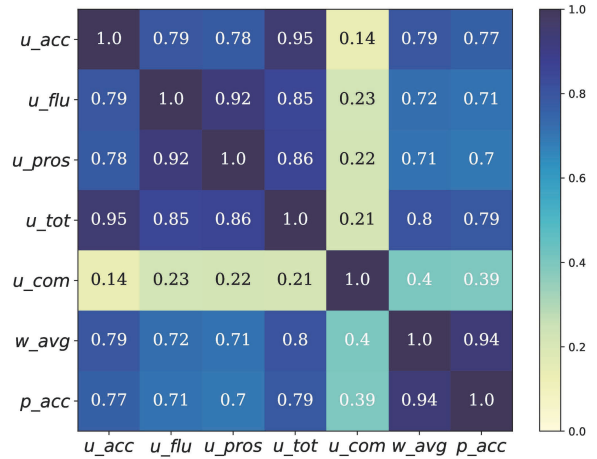


Figure 4: Correlation matrix of different metrics at three granularities. Thereinto, p_acc stands for phoneme-level accuracy; w_avg stands the mean value for word-level accuracy, total and stress; u_com, u_acc, u.flu, u_pros and u_tot stand for utterance-level completeness, accuracy, fluency, prosodic and total score, respectively.

However, further increasing the embedding size results in minor performance drops.

Similarly, we can observe slightly lower results with multiple heads, which we attribute to the limited data size and increased model complexity making optimization more difficult. These findings indicate that the selected configuration strikes a balance between expressiveness and efficiency.

Data Correlation Analysis

To validate the high correlation between phoneme-, word- and utterance-level scores, we calculate the correlation between each pair of aspects and visualize it.

As shown in Fig. 4, the relatively high correlations among phoneme accuracy, word mean scores, and utterance-level scores suggests that these scores are interdependent. This observation supports the use of bidirectional interaction mechanisms and hierarchical structures for modeling linguistic structures and accomplishing multi-granularity scoring tasks.

Conclusion

In this paper, we propose a novel multi-aspect multi-granular pronunciation assessment model named HIA. To achieve bidirectional interaction between different granularity levels, we design a novel Interactive Attention Module that generates interactive attention heads corresponding to each granularity, significantly improves the model performance, particularly on word stress. In addition, we propose a residual hierarchical structure through residual connections to mitigate feature forgetting, further improving the model performance. Experimental results on speechocean762 dataset show that our proposed model achieves the state-of-the-art of the multi-aspect multi-granularity pronunciation assessment on all granularities and aspects metrics.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62172256, 62202272, 62202278, in part by Natural Science Foundation of Shandong Province under Grant ZR2024LZH002 and Taishan Scholar Project of Shandong Province under Grant tstp20250704.

References

- Abdel-Hamid, O.; Mohamed, A.-r.; Jiang, H.; Deng, L.; Penn, G.; and Yu, D. 2014. Convolutional Neural Networks for Speech Recognition. *IEEE/ACM Trans Audio, Speech, Language Process.*, 22(10): 1533–1545.
- Al-Barhamtoshy, H.; Abdou, S.; and Jambi, K. 2014. Pronunciation Evaluation Model for None Native English Speakers. *Life Science Journal*, 11(9): 216–226.
- Arias, J. P.; Yoma, N. B.; and Vivanco, H. 2010. Automatic Intonation Assessment for Computer Aided Language Learning. *Speech Communication*, 52(3): 254–267.
- Chao, F.-A.; Lo, T.-H.; Wu, T.-I.; Sung, Y.-T.; and Chen, B. 2023. A Hierarchical Context-Aware Modeling Approach for Multi-aspect and Multi-granular Pronunciation Assessment. In *ISCA Interspeech*, 974–978.
- Chen, Y.-W.; Yu, Z.; and Hirschberg, J. 2024. MultiPA: A Multi-Task Speech Pronunciation Assessment Model for Open Response Scenarios. In *ISCA Interspeech*, 297–301.
- Cheng, S.; Liu, Z.; Li, L.; Tang, Z.; Wang, D.; and Zheng, T. F. 2020. ASR-Free Pronunciation Assessment. In *ISCA Interspeech*, 3047–3051.
- Cincarek, T.; Gruhn, R.; Hacker, C.; and Nakamura, S. 2009. Automatic Pronunciation Scoring of Words and Sentences Independent from the Non-Native’s First Language. *Computer Speech Language.*, 23(1): 65–88.
- Do, H.; Kim, Y.; and Lee, G. G. 2023. Hierarchical Pronunciation Assessment with Multi-Aspect Attention. In *ICASSP*, 1–5.
- Eskenazi, M. 2009. An Overview of Spoken Language Technology for Education. *Speech Communication*, 51(10): 832–844.
- Fu, K.; et al. 2022. Improving Non-Native Word-level Pronunciation Scoring with Phone-Level Mixup Data Augmentation and Multi-source Information. *arXiv:2203.01826*.
- Gao, L.; Guo, Z.; Zhang, H.; Xu, X.; and Shen, H. T. 2017. Video Captioning with Attention-Based LSTM and Semantic Consistency. *IEEE Trans Multimedia*, 19(9): 2045–2055.
- Gao, Z.; Zhang, S.; McLoughlin, I.; and Yan, Z. 2022. Paraformer: Fast and Accurate Parallel Transformer for Non-Autoregressive End-to-End Speech Recognition. *arXiv:2206.08317*.
- Gong, Y.; Chen, Z.; Chu, I.-H.; Chang, P.; and Glass, J. 2022. Transformer-Based Multi-Aspect Multi-Granularity Non-Native English Speaker Pronunciation Assessment. In *ICASSP*, 7262–7266.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*, 770–778.
- Hu, W.; Qian, Y.; Soong, F. K.; and Wang, Y. 2015. Improved Mispronunciation Detection with Deep Neural Network Trained Acoustic Models and Transfer Learning Based Logistic Regression Classifiers. *Speech Communication*, 67: 154–166.
- Kheir, Y.; Ali, A.; and Chowdhury, S. 2023. Automatic Pronunciation Assessment - A Review. In *EMNLP*, 8304–8324.
- Kim, E.; Jeon, J.-J.; Seo, H.; and Kim, H. 2022. Automatic Pronunciation Assessment Using Self-Supervised Speech Representation Learning. *arXiv:2204.03863*.
- Li, K.; Qian, X.; and Meng, H. 2017. Mispronunciation Detection and Diagnosis in L2 English Speech Using Multidistribution Deep Neural Networks. *IEEE/ACM Trans Audio, Speech, Language Process.*, 25(1): 193–207.
- Li, K.; Wu, X.; and Meng, H. 2017. Intonation Classification for L2 English Speech Using Multi-Distribution Deep Neural Networks. *Computer Speech Language.*, 43: 18–33.
- Lin, B.; and Wang, L. 2021. Deep Feature Transfer Learning for Automatic Pronunciation Assessment. In *ISCA Interspeech*, 4438–4442.
- Lin, B.; and Wang, L. 2023. Exploiting Information From Native Data for Non-Native Automatic Pronunciation Assessment. In *Spoken Language Technology*, 708–714.
- Lin, B.; Wang, L.; Feng, X.; and Zhang, J. 2020. Automatic Scoring at Multi-Granularity for L2 Pronunciation. In *ISCA Interspeech*, 3022–3026.
- Liu, W.; et al. 2023a. An ASR-Free Fluency Scoring Approach with Self-Supervised Learning. In *ICASSP*, 1–5.
- Liu, W.; et al. 2023b. Leveraging Phone-Level Linguistic-Acoustic Similarity for Utterance-Level Pronunciation Scoring. In *ICASSP*, 1–5.
- Mao, S.; Soong, F.; Xia, Y.; and Tien, J. 2022. A Universal Ordinal Regression for Assessing Phoneme-Level Pronunciation. In *ICASSP*, 6807–6811.
- Panayotov, V.; Chen, G.; Povey, D.; and Khudanpur, S. 2015. LibriSpeech: An ASR Corpus Based on Public Domain Audio Books. In *ICASSP*, 5206–5210.
- Pei, H.-C.; Fang, H.; Luo, X.; and Xu, X.-S. 2024. Gradformer: A Framework for Multi-Aspect Multi-Granularity Pronunciation Assessment. *IEEE/ACM Trans Audio, Speech, Language Process.*, 32: 554–563.
- Ridley, R.; He, L.; Dai, X.-y.; Huang, S.; and Chen, J. 2021. Automated Cross-Prompt Scoring of Essay Traits. In *AAAI*, 13745–13753.
- Shi, J.; Huo, N.; and Jin, Q. 2020. Context-Aware Goodness of Pronunciation for Computer-Assisted Pronunciation Training. *arXiv:2008.08647*.
- Strik, H.; Truong, K.; De Wet, F.; and Cucchiari, C. 2009. Comparing Different Approaches for Automatic Pronunciation Error Detection. *Speech Communication*, 51(10): 845–852.
- Tejedor-García, C.; Escudero-Mancebo, D.; Cámara-Arenas, E.; González-Ferreras, C.; and Cardeñoso-Payo, V. 2020. Assessing Pronunciation Improvement in Students of

- English Using a Controlled Computer-Assisted Pronunciation Tool. *IEEE/ACM Trans Learning Technologies*, 13(2): 269–282.
- Tepperman, J.; and Narayanan, S. 2005. Automatic Syllable Stress Detection Using Prosodic Features for Pronunciation Evaluation of Language Learners. In *ICASSP*, 937–940.
- Tu, M.; Grabek, A.; Liss, J.; and Berisha, V. 2018. Investigating the Role of L1 in Automatic Pronunciation Evaluation of L2 Speech. In *ISCA Interspeech*, 1636–1640.
- Vaswani, A.; et al. 2017. Attention Is All You Need. *NIPS*, 5998–6008.
- Wang, Y.-B.; and Lee, L.-S. 2012. Improved Approaches of Modeling and Detecting Error Patterns with Empirical Analysis for Computer-Aided Pronunciation Training. In *ICASSP*, 5049–5052.
- Witt, S.; and Young, S. 2000. Phone-Level Pronunciation Scoring and Assessment for Interactive Language Learning. *Speech Communication*, 30(2-3): 95–108.
- Yan, B.-C.; and Chen, B. 2024. An Effective Hierarchical Graph Attention Network Modeling Approach for Pronunciation Assessment. *IEEE/ACM Trans Audio, Speech, Language Process.*, 32: 3974–3985.
- Zhang, J.; et al. 2021. Speechocean762: An Open-Source Non-Native English Speech Corpus for Pronunciation Assessment. In *ISCA Interspeech*, 3710–3714.