# Gradformer: A Framework for Multi-Aspect Multi-Granularity Pronunciation Assessment

Hao-Chen Pei, Hao Fang, Xin Luo, and Xin-Shun Xu, *Senior Member, IEEE*

***Abstract*—Automatic pronunciation assessment is an indispensable technology in computer-assisted pronunciation training systems. To further evaluate the quality of pronunciation, multi-task learning with simultaneous output of multi-granularity and multi-aspect has become a mainstream solution. Existing methods either predict scores at all granularity levels simultaneously through a parallel structure, or predict individual granularity scores layer by layer through a hierarchical structure. However, these methods do not fully understand and take advantage of the correlation between the three granularity levels of phoneme, word, and utterance. To address this issue, we propose a novel method, Granularity-decoupled Transformer (Gradformer), which is able to model the relationships between multiple granularity levels. Specifically, we first use a convolution-augmented transformer encoder to encode acoustic features, where the convolution module helps the model better capture local information. The model outputs both phoneme- and word-level granularity scores with high correlation by the encoder. Then, we use utterance queries to interact with the output of the encoder through the transformer decoder, ultimately obtaining the utterance scores. Through unique encoder and decoder architecture, we achieve decoupling at three granularity levels, and handling the relationship between each granularity. Experiments on the speachocean762 dataset show that our model has advantages over state-of-the-art methods in various metrics, especially in key metrics such as phoneme accuracy, word accuracy, and total score.**

***Index Terms*—Pronunciation assessment, transformer, computer-assisted pronunciation training, goodness of pronunciation, speech recognition.**

## I. Introduction

AUTOMATIC speech assessment or automatic pronunciation assessment is an essential technology in the field of Computer-Assisted Pronunciation Training (CAPT) [1], [2], [3], which aims to score the pronunciation quality of non-native language learners (called L2 learners) and provide feedback to help them better learn foreign languages [4], [5], [6]. Compared with manual or expert evaluation, CAPT is more convenient, efficient, and cost-effective [7].

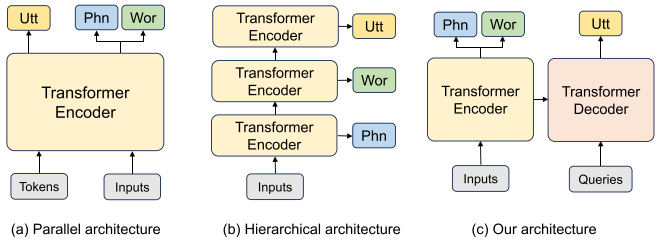


Fig. 1. Comparisons of parallel, hierarchical, and our architecture.

Automatic speech assessment has great practical significance and has been extensively researched. While much progress has been made, early works mainly focus on scoring phoneme-level pronunciation quality [8], [9]. A recent work achieves multi-granularity scoring through a hierarchical structure [10]. However, the speech assessment task encompasses multiple aspects, not just phoneme accuracy, but also word stress [11], overall fluency [12] and the prosody [11], [13]. There may be correlations between these aspects. For example, L2 learners who exhibit smooth intonation may have higher language proficiency and better pronunciation. Therefore, a unified joint modeling approach may yield better results than individual aspect evaluations.

Recently, a model called GOPT [14] incorporates multiple granularities and aspects of pronunciation into a unified model through multi-task learning. It uses a Transformer encoder [15] to simultaneously output scores for all granularities and aspects in parallel, as shown in Fig. 1(a). However, phonemes, words and utterances are not independent of each other, and there are strong linguistic dependencies between them [10]. The parallel architecture cannot capture this feature. A follow-up work, HiPAMA [16], proposes a hierarchical architecture with multiple modules of different granularities, which scores each granularity sequentially, as shown in Fig. 1(b). To some extent, it overcomes the problem of inability to capture linguistic hierarchy information between different granularities of speech in GOPT. However, like GOPT, HiPAMA also ignores the issue of correlation between different granularities. Furthermore, it uses a simple average operation on word-level aspect representations before obtaining utterance-level aspect representations, which may be one of the reasons for its poor performance on utterance fluency and prosody.

This paper conducts data correlation analysis and finds that there is a high correlation between phoneme-level scores and

Manuscript received 31 July 2023; revised 16 November 2023; accepted 18 November 2023. Date of publication 23 November 2023; date of current version 1 December 2023. This work was supported in part by the National Natural Science Foundation of China under Grants 62172256, 62202278, and 62202272, in part by Shandong Provincial Key Research and Development Program under Grant 2019JZZY010127, in part by Natural Science Foundation of Shandong Province under Grants ZR2019ZD06, ZR2020QF036, and ZR2021ZD15, and in part by the Major Program of the National Natural Science Foundation of China under Grant 61991411. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Juan Ignacio Godino-Llorente. *(Corresponding author: Xin-Shun Xu.)*

The authors are with the School of Software, Shandong University, Jinan 250101, China (e-mail: 202235343@mail.sdu.edu.cn; fanghaook@mail.sdu.edu.cn; luoxin.lxin@gmail.com; xuxinshun@sdu.edu.cn).

Digital Object Identifier 10.1109/TASLP.2023.3335807

word-level scores, which is demonstrated in Section V-C. Utterance-level scores have more indicators and are influenced by more factors. The correlation between phoneme/word-level scores and utterance-level scores is not very high, which also somewhat aligns with human intuition. In addition, in multi-task learning, some tasks that are difficult to evaluate or have extremely imbalanced data distribution may affect the performance of other tasks. Therefore, utterance-level representations require a separate module for processing. Previous models cannot address these two problems effectively. Therefore, we believe that the performance of a model with parallel or hierarchical structure, as shown in Fig. 1(a) and (b), may be limited.

Inspired by these analyses, we propose a novel multi-aspect multi-granularity pronunciation assessment framework with encoder-decoder structure, as shown in Fig. 1(c). Specifically, the encoder module encodes the acoustic features and outputs two fine-grained scores with high correlation simultaneously, i.e., phoneme-level and word-level. Thereafter, a small fixed number of learnable utterance representations called utterance queries are input to the decoder. The decoder interacts multiple times with the output of the encoder to assist in utterance-level scoring. This encoder-decoder architecture independently processes each granularity score with different correlations, effectively solving the problem of data correlation and the difficulty of evaluating utterance-level scores. In addition, a novel convolution module is introduced into the encoder to better model local information. Experiments show that our method can improve performance in various granularities and aspects of pronunciation assessment.

In summary, the main contributions of this paper are as follows:

- We propose a granularity-decoupled Transformer framework for multi-granularity and multi-aspect pronunciation assessment, Gradformer for short. Through a unique encoder-decoder architecture, Gradformer achieves the decoupling and merging of phoneme-, word-, and utterance-level scoring.
- We introduce a novel convolution module into the encoder, which helps the model better capture the local information of acoustic features.
- Experiments on the speechocean762 [17] dataset show that our model has advantages over state-of-the-art methods in almost all metrics, especially in key metrics such as phoneme accuracy, word accuracy, and total score.

## II. Related Work

CAPT technology has two application scenarios, namely read-aloud scenario and open-response scenario [18]. Currently, the majority of works, including ours, focus on the read-aloud scenario. The mainstream methods of pronunciation assessment can be divided into two types, GOP-based methods and non-GOP methods, respectively.

### A. GOP-Based Methods

GOP [8] (Goodness of Pronunciation) is a classic and effective method for pronunciation assessment that has been the mainstream approach in the past decades [19], [20], [21]. GOP is a confidence score derived from automatic speech recognition (ASR) systems in essence, which is computed as the normalized frame-level posterior probability of phonemes. Initially, GOP is calculated using acoustic models based on Gaussian Mixture Models-Hidden Markov Models (GMM-HMM). Thereafter, with the rapid development of deep learning, acoustic models based on DNN-HMM have stronger acoustic representation abilities, simplifying and improving the calculation and performance of GOP [22], [23]. Thereinto, context-aware GOP [9] is an improved GOP method based on TDNN-HMM. It overcomes the limitations of traditional GOP, which ignores transitions between phonemes within the segment and the contextual effects across phonemes, by introducing transition and duration factors.

There are generally two forms of using the GOP method for pronunciation assessment: GOP score based and GOP feature based. Specifically, a GOP score-based method directly uses the calculated GOP value as the score for pronunciation quality, or sets a threshold for the GOP score to achieve Mispronunciation Detection and Diagnosis (MDD) [24], [25], [26]. A GOP feature-based method extracts frame-level or segmental-level GOP features and trains a task-specific regressor or classifier to achieve speech assessment or MDD [23]. The latter often has better scoring performance than the former [27]; however, it relies on supervised pronunciation assessment data, which is typically costly.

Early works mainly focus on phoneme-level pronunciation scoring. To obtain scores at the word or utterance level, a common approach is to average the corresponding phoneme-level scores [20], [22], [28]. Recently, a method [10] is proposed for multi-granularity scoring using a hierarchical architecture, but it requires a relatively complex training scheme for optimization. Gong et al. [14] further proposes a parallel architecture using multi-task learning, which achieves multi-aspect and multi-granularity scoring for the first time. A follow-up work [16] uses a hierarchical architecture to capture hierarchical information between different granularities, further improving the performance of multi-aspect multi-granularity scoring.

### B. Non-GOP Methods

Since GOP is a kind of manual acoustic feature, it may not provide sufficient information for speech assessment [29]. Therefore, researchers are also exploring non-GOP methods.

Thereinto, a transfer learning based method [30] is proposed to extract deep features for pronunciation assessment. It replaces GOP features for utterance-level pronunciation assessment with deep features [31] generated by a model, which is pre-trained on a speech recognition dataset from native speakers (L1) and fine-tuned on a speech assessment dataset from L2 speakers. It partially alleviates the problem of insufficient data in the field of speech assessment. In addition, an end-to-end model consisting of an audio encoder and a text encoder for multi-granularity scoring is proposed in [32]. However, due to the limitation of modeling granularity, the end-to-end model can only provide scores at the word- and utterance-level, but not at the fine-grained level of phonemes.

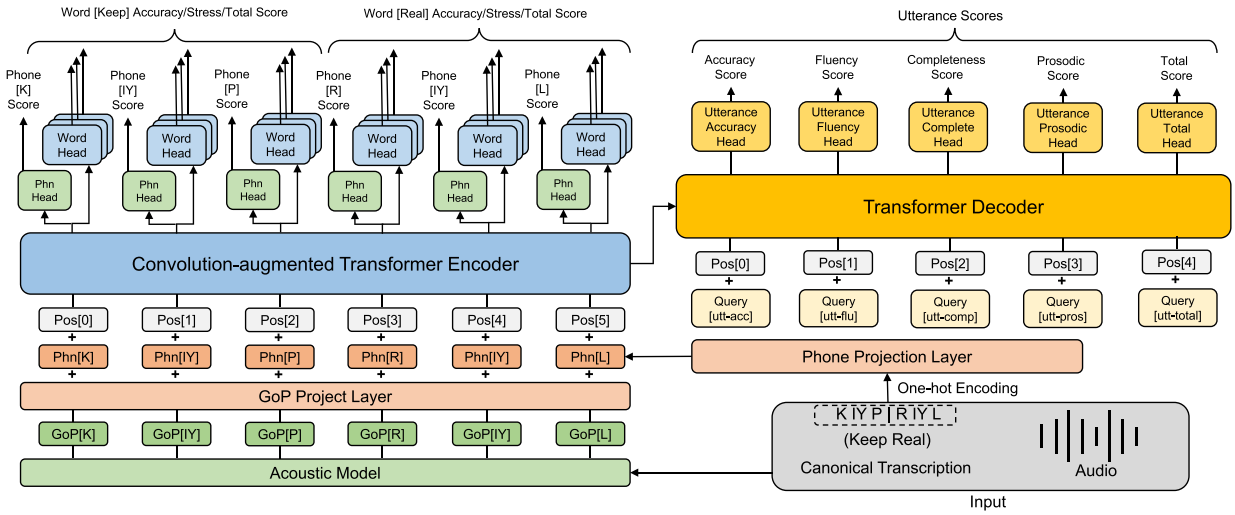


Fig. 2. Overall architecture of Gradformer. Gradformer takes the GOP features extracted from the acoustic model and the projected phoneme embedding as the input. Then, a convolution-augmented transformer encoder is applied to output the phoneme- and word-level scores. Finally, a decoder, which takes utterance queries as input, outputs the five utterance-level aspect scores.

Recently, self-supervised learning pre-trained models have performed well in downstream tasks in the field of speech [33], [34], [35], [36]. Many studies have demonstrated the feasibility of using self-supervised learning models for pronunciation assessment [37], [38] and MDD [39], [40], [41], indicating that self-supervised learning features can also be a good substitute for GOP features for pronunciation assessment task.

Nevertheless, the above-mentioned methods still rely on the acoustic features or the forced alignment between acoustic features and canonical transcription derived from ASR systems. Currently, certain ASR-free methods have shown progress and can be applied to open scenarios to some extent [29], [42].

## III. Method

### A. Overview

It is worth noting that the previous methods process all granularity levels in parallel or use a hierarchical architecture to handle each granularity level independently. By contrast, we design an encoder-decoder architecture for the multi-aspect multi-granularity pronunciation assessment task for the first time. As shown in Fig. 2, firstly, the phoneme embedding is added to the GOP features obtained from acoustic model as input to the model. Then, a convolution-augmented transformer encoder is used to encode the features and output phoneme-level and word-level scores. Finally, a transformer decoder is used to decode phoneme-level features into utterance queries, thereby outputting utterance-level scores. As mentioned previously, our model achieves the merging and decoupling of scoring at different granularities through the encoder-decoder architecture. It can reasonably model the relationships between each granularity level and help the model learn better utterance-level representations. Additionally, our model employs novel convolution modules and self-attention mechanisms to simultaneously focus on local and global acoustic features. In the following subsections, the model is described in details.

### B. Model Architecture

*Model Inputs:* For fair comparison, we follow the baseline model [14] to use GOP features as input to the model. GOP features are derived from the acoustic features outputted by the acoustic model and the forced alignment between acoustic features and canonical transcription, which is the log phone posterior (LPP) and log posterior ratio (LPR) defined in [23]. Specifically, the LPP of a phone $p_i$ is defined as follows:

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

$$p(p_i|o_t) = \sum_{s \in p_i} p(s|o_t), \tag{2}$$

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

$$LPR\left(p_j|p_i\right) = \log p\left(p_j|o; t_s, t_e\right) - \log p\left(p_i|o; t_s, t_e\right). \tag{3}$$

The Librispeech acoustic model[1] we use to extract acoustic features and generate forced alignment has a total of 42 pure phones, including 39 phonemes according to CMU Pronouncing Dictionary as well as symbols of placeholder, silence and noise. Thus the GOP feature of phone $p_i$ can be defined as an 84-dimensional vector as follows:

$$\left[LPP(p_1) \ldots, LPP(p_{42}), LPR\left(p_1|p_i\right) \ldots, LPR\left(p_{42}|p_i\right)\right]. \tag{4}$$

Due to the fact that different phonemes possess distinct characteristics and canonical phoneme embeddings are crucial for the model performance, we also use one-hot canonical phoneme embeddings as the input of our model, same as the baseline model [14]. To better fuse the two types of features, we project

[1][Online]. Available: https://kaldi-asr.org/models/m13

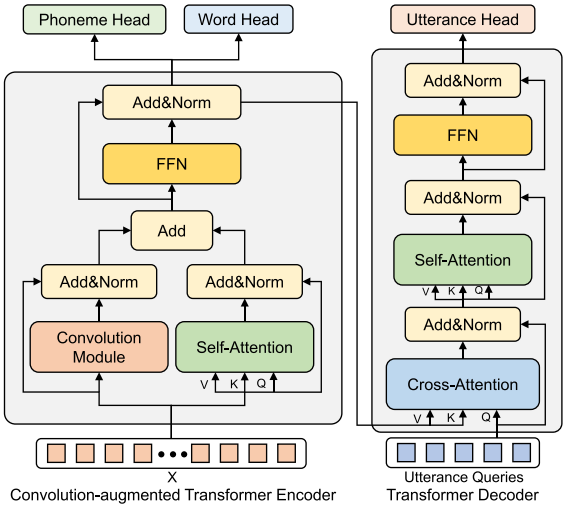


Fig. 3. Illustration of the Convolution-augmented Transformer Encoder and Transformer Decoder.

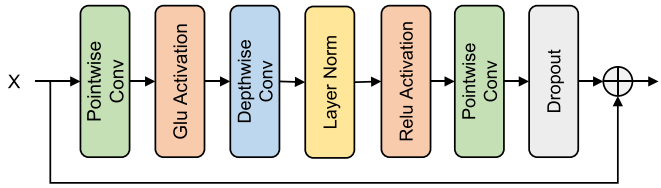


Fig. 4. Illustration of the Convolution Module. The convolution module contains a pointwise convolution with a GLU activation layer. Then, the depthwise convolution with a layer normalization and a Relu activation layer is followed by a same pointwise convolution and a dropout layer.

both of them into the same $C$-dimension and add them directly as the final input $x \in \mathbb{R}^{C \times N}$ of the model, where $N$ is the number of phonemes of the audio input.

*Convolution-augmented Transformer Encoder:* Because there is a significant correlation between phoneme-level and word-level scores, we use a transformer encoder to output these two granularities scores simultaneously. Specifically, we encode input features through an $L$-layer convolution-augmented transformer encoder, each layer consisting of a convolution module, a self-attention module, and a feed forward network (FFN) as shown in Fig. 3.

Firstly, we use the self-attention module to obtain global contextual information of features as follow:

$$x^l_{self} = SelfAttn\left(x^{l-1}\right), \tag{5}$$

where $x^{l-1}$ is the output of the previous layer, $x^l_{self}$ is the output of the self-attention module.

Multiple phonemes within the same word clearly have more significant interrelationships. However, the self-attention module extracts global features and introduces a large amount of redundant information, resulting in the loss of important local information. Alternatively, convolutions have been successful for ASR [43], [44], which can effectively capture local context. Therefore, we introduce the convolution module to better capture local information between phonemes, the operation is defined as follow:

$$x^l_{conv} = Conv\left(x^{l-1}\right), \tag{6}$$

where $x^l_{conv}$ is the output of the convolution module.

As shown in Fig. 4, firstly, pointwise convolution is a 1-D convolution with a kernel size of 1, which doubles the feature dimension. The GLU activation function is used after it and changes the feature dimension to C. Then, the depthwise convolution with a kernel size of 3 further extracts local information, followed by a layer normalization and a relu activation function. Finally, there is the same pointwise convolution, followed by a dropout layer. The entire module is a residual structure.

We further fuse the global information extracted by the self-attention module and the local information extracted by the convolution module through a feed forward network. The fusion operation is formulated as follows:

$$x^l = FFN\left(x^l_{conv} + x^l_{self}\right), \tag{7}$$

where $x^l$ is the output of the $l$-th layer of the transformer encoder.

As shown in Fig. 3, we add phoneme- and word-level regression heads above the output of each corresponding phoneme in the last layer of the encoder. Thereinto, each phoneme has a phoneme-level regression head that outputs phoneme-level accuracy scores, and has three word-level regression heads that output accuracy, stress, and total score. Each regression head is a $48 \times 1$ linear layer with layer normalization. During the training phase, the word-level score of each phoneme is supervised by the score of the word to which it belongs. In the inference stage, the average score of all phonemes belonging to a word is used as the final word-level score.

*Utterance Transformer Decoder:* Utterance-level scoring is a comprehensive assessment of the entire utterance, which requires an independent module to model the relationship between the features output by the encoder and this level. Inspired by DETR, we use a transformer decoder to generate the final utterance-level scores.

Firstly, we initialize a set of learnable vectors, we call utterance queries, with each query corresponding to an utterance-level aspect. Randomly initialized query features of the first self-attention layer are phoneme-independent and do not have information from the phoneme-level features, thus applying self-attention is unlikely to enrich information. Therefore, unlike the original transformer decoder design, we switch the order of self- and cross-attention to make computation more effective. The cross attention between utterance queries and encoder output is defined as follows:

$$q^d_{cross} = CrossAttn\left(q^{d-1}, x^L\right), \tag{8}$$

where $q^{d-1}$ is the output of the previous layer of decoder, $q^0$ is obtained through random initialization. $x^L$ is the phoneme level features output by the last layer of encoder, which contains rich contextual information. In the cross-attention module, $Q$ is $q^{d-1}$, and $K, V$ is $x^L$. $q^d_{cross}$ is the output of the cross-attention module.

Then, through self-attention modules and FFN, the query features that refine and integrate phoneme-level information are

further enhanced:

$$q^d = FFN\left(SelfAttn\left(q^d_{cross}\right)\right), \tag{9}$$

where $q^d$ is the output of the $d$-th layer of the transformer decoder. Our query based method uses compact and learnable embedding vectors to represent utterance level granularity, and uses them as queries to decode scores from phoneme level features.

Similar to the phoneme level and word level, we add regression heads corresponding to the utterance aspect for each query output by the last layer of decoder to predict the final utterance-level score.

### C. Loss

We consider each sub-task of the pronunciation assessment as a regression problem. Therefore, we adopt mean squared error (MSE) loss for each task as follows:

$$\mathcal{L}_{MSE} = \frac{1}{M}\sum_{i=1}^{M}(f-y)^2, \tag{10}$$

where $M$ is the number of samples, $f$ is the prediction of the model, and $y$ is the true score.

We normalize all output scores to the same scale. Because each granularity has multiple aspects of loss, we consider the average loss of multiple aspects as the loss at each granularity level. The total loss is calculated as the sum of each granularity-level loss:

$$\mathcal{L}_{total} = \mathcal{L}_{phn} + \mathcal{L}_{word} + \mathcal{L}_{utt}, \tag{11}$$

where $\mathcal{L}_{phn}$ is the phoneme-level loss, $\mathcal{L}_{word}$ is the average of the three word-level aspect losses, and $\mathcal{L}_{utt}$ is the average of the five utterance-level aspect losses.

## IV. Experiments

### A. Dataset and Metrics

Speechocean762 is a free and open-source speech assessment dataset available on openslr.[2] It should be noted that speechocean762 is currently the only available dataset for multi-aspect and multi-granularity pronunciation assessment task. It contains 5000 English utterances from 250 non-native speakers whose mother tongue is Mandarin, half of whom are children and the other half are adults. The dataset is divided into a training set and a test set, each containing 2500 utterances. The training set includes 15,849 words and 47,076 phonemes, while the test set includes 15,967 words and 47,369 phonemes.

Moreover, it is a multi-granularity and multi-aspect speech assessment dataset that provides rich label information. Specifically, it includes three granularity levels: phoneme-, word-, and utterance-level. The phoneme-level score is accuracy. The word-level has three aspects of scores: accuracy, stress, and total. The utterance-level has five aspects of scores: accuracy, completeness, fluency, prosodic, and total. Each score is independently annotated by five experts using the same scoring criteria, and the final score is the average score of the five experts. The scoring range for word-level and utterance-level is [0, 10], while for phoneme-level it is [0, 2]. In the experiments, the scores for the three granularities are uniformly scaled to a range of [0, 2].

Pearson Correlation Coefficient (PCC) is a statistical measure of linear correlation between two data sequences. We use PCC as the evaluation metric to measure the correlation between the model's scores and expert scores. Specifically, the PCC between model scoring and expert scoring can be calculated as (12):

$$PCC(f,y) = \frac{\sum_{i=1}^{M}\left(f_i-\bar{f}\right)\left(y_i-\bar{y}\right)}{\sqrt{\sum_{i=1}^{M}\left(f_i-\bar{f}\right)^2}\sqrt{\sum_{i=1}^{M}\left(y_i-\bar{y}\right)^2}}, \tag{12}$$

where $f_i$ is the $i$-th score predicted by the model for one aspect, $y_i$ is the corresponding $i$-th score given by the expert, $M$ is the total number of samples, which is 2500.

Pronunciation assessment is a relatively subjective task. Even experts, when scoring under the same criteria, are inevitably influenced by subjective factors. We calculate the mean PCC values between each pair of the five experts for each metric, as shown in Table I. It is evident that the consistency among the experts is not very high. Therefore, we use the average scores given by the five experts as the final scores in the experiments, aiming to minimize the impact of subjective factors. In addition, the consistency of individual scoring on different metrics is also noteworthy, and this will be discussed in detail in Section V-C.

### B. Experimental Setup

*Acoustic Model:* Due to the scarcity of L2 data, acoustic models often use L1 data which is abundant and easy to access for pre-training and fine-tune on L2 data to generate better forced alignment and output better GOP features [45]. For fair comparison, in the experiments, we use the same acoustic model as in [14] to extract GOP features. The acoustic model is based on the factorized time-delay neural network (TDNN-F) and trained on the 960 h L1 speech recognition dataset Librispeech [46] using Kaldi recipe.

*Training Configuration:* During the model training phase, we use the Adam optimizer to train the Gradformer model with an initial learning rate of 1e-3 and a batch size of 25. The maximum number of epoch is set to 60, and the learning rate is halved every 5 epochs after the 20th epoch. We save the model with the minimum phoneme-level MSE loss as the optimal model.

*Model Configuration:* In Gradformer, both encoder and decoder have 3 layers. The embedding dimensions of the encoder and decoder are both 48. The dimension of FFN is 256. Considering the problem that the dataset and feature dimension are not large enough, we only set 1 head for self-attention and cross-attention. The dropout ratio is set to 0.1 to suppress overfitting. All transformer weights are randomly initialized with Xavier init [47]. Experiments show that 5 utterance queries are sufficient to interact with the acoustic features from the encoder output and represent the utterance-level features well. Therefore, we use 5 utterance queries as the input of the decoder.

[2][Online]. Available: http://www.openslr.org/101/

TABLE I
PERFORMANCE COMPARISON OF GRADFORMER AND VARIOUS METHODS ON SPEECHOCEAN762

| Model | Phoneme score | | Word score (PCC) | | | Utterace score (PCC) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MSE↓ | PCC↑ | Accuracy↑ | Stress↑ | Total↑ | Accuracy↑ | Completeness↑ | Fluency↑ | Prosodic↑ | Total↑ |
| Human | - | - | 0.589 | 0.212 | 0.602 | 0.618 | 0.658 | 0.665 | 0.651 | 0.675 |
| SVR [17] | 0.160 | 0.450 | - | - | - | - | - | - | - | - |
| UOR [48] | 0.120 | 0.520 | - | - | - | - | - | - | - | - |
| Mixup-pretrain [28] | - | - | - | - | 0.610 | - | - | - | - | - |
| Deep feature [30] | - | - | - | - | - | - | - | - | - | 0.720 |
| Wav2vec2-based [49] | - | - | - | - | - | - | - | - | - | 0.725 |
| LAS [50] | - | - | - | - | - | - | - | - | - | **0.766** |
| LSTM [14] | 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 | 0.745 ±0.002 | 0.747 ±0.005 | 0.741 ±0.002 |
| GOPT [14] | 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 [16] | 0.084 ±0.001 | 0.616 ±0.004 | 0.575 ±0.004 | 0.32 ±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** | **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 |

## V. RESULTS AND DISCUSSIONS

### A. Main Results

We compare Gradformer with traditional single-granularity scoring models and multi-aspect multi-granularity scoring models. Specifically, we compare Gradformer with the following 9 models: 1) Support vector regression (SVR) based model presented in [17]; 2) A universal ordinal regression (UOR) model proposed in [48]; 3) A phoneme-level mixup data augmentation and multi-source information method for improving word-level scoring proposed in [28]; 4) Deep feature based model presented in [30]; 5) A model [49] based on self-supervised learning model wav2vec2; 6) A linguistic-acoustic similarity (LAS) measure and phone-level GOP pre-training method proposed in [50]; 7) A parallel architecture model using LSTM as the encoder proposed in [14]; 8) A parallel architecture model named GOPT using transformer encoder proposed in [14]; 9) A hierarchical architecture model named HiPAMA [16]. Models 1-6 can only score specific granularities of speech, while models 7-9 use a multi-task learning approach for multi-aspect multi-granularity scoring.

We train the model five times with different random seeds on the training set and report the mean and standard deviation of the results on the test set, as shown in Table I. From this table, we have the following observations:

- Our model achieves state-of-the-art performance on various granularities and aspects, especially on phoneme-level and word-level scoring.
- Gradformer achieves 0.079 MSE and 0.646 PCC for phoneme-level accuracy, which is a significant improvement over previous methods. None of the previous methods use convolution except HiPAMA. HiPAMA only employs simple convolutional layers. This demonstrates the effectiveness of the convolution module in Gradformer, which is further demonstrated in subsequent ablation experiments.
- For word-level scoring, Gradformer also achieves markedly better performance than GOPT and is also notably better than HiPAMA, which also uses convolution. It is worth mentioning that Gradformer achieves 0.614 PCC on the word-level total score, which is even better than Mixup-pretrain method that uses additional internal L2 data for pre-training. This verifies the effectiveness of both the convolution module in Gradformer and the architecture of Gradformer.
- In terms of the utterance-level scoring, Gradformer outperforms various previous models on all other metrics except for the total score. Gradformer achieves 0.756 PCC, which is better than previous models except for LAS method. One possible reason is that LAS uses additional 4000 hours of internal L2 speech data for the pre-training of the acoustic model, which is far more than the 960 hours of L1 data we use to train our acoustic model. Gradformer overcomes the problem of hierarchical models being unable to effectively model utterance-level representations, resulting in a comprehensive improvement in utterance-level performance. LAS only achieves 0.702 PCC without using additional internal L2 data. Our acoustic model is not even fine-tuned on speechocean762. Even so, we are able to achieve very competitive results, further demonstrating the effectiveness of Gradformer.
- The previous models do not perform well on the stress and completeness metrics, mainly due to the extremely imbalanced data distribution. Even so, our model outperforms other models on these two metrics, indicating its robustness.

### B. Ablation Studies

To explore the factors affecting the performance of Gradformer, we conduct comprehensive ablation studies. Specifically, we carry out four sets of ablation experiments to explore the effects of the model architecture, convolution modules, model depth, and embedding dimensions on model performance.

*Architecture Ablation:* We first verified the effectiveness of the proposed model architecture. Thereinto, the baselines are defined as follows:

TABLE II
RESULTS OF THE ABLATION EXPERIMENTS ON THE EFFECTIVENESS OF ARCHITECTURE

| Model | Phoneme score | | Word score (PCC) | | | Utterace score (PCC) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MSE↓ | PCC↑ | Accuracy↑ | Stress↑ | Total↑ | Accuracy↑ | Completeness↑ | Fluency↑ | Prosodic↑ | Total↑ |
| Baseline | 0.087 ±0.001 | 0.603 ±0.003 | 0.526 ±0.004 | 0.333 ±0.030 | 0.543 ±0.002 | 0.707 ±0.004 | 0.205 ±0.039 | 0.761 ±0.008 | 0.756 ±0.006 | 0.733 ±0.005 |
| Baseline+Decoder | 0.084 ±0.001 | 0.618 ±0.003 | 0.542 ±0.006 | 0.338 ±0.024 | 0.558 ±0.007 | 0.726 ±0.003 | 0.126 ±0.071 | 0.762 ±0.001 | 0.760 ±0.001 | 0.750 ±0.001 |
| Baseline+Decoder+C | 0.081 ±0.002 | 0.636 ±0.006 | 0.584 ±0.009 | **0.340** **±0.036** | 0.600 ±0.009 | 0.723 ±0.006 | 0.061 ±0.126 | 0.756 ±0.005 | 0.756 ±0.006 | 0.746 ±0.005 |
| **Gradformer** | **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** |

Baseline represents the parallel architecture model, which is the GOPT. Baseline+Decoder represents the encoder-decoder architecture without convolution enhancement. Baseline+Decoder+C represents the encoder-decoder architecture using simple convolutions with kernel size of 3.

*Baseline:* Our implemented parallel architecture model, which is GOPT. It consists of a three-layer transformer encoder with the embedding dimension of 48.

*Baseline+Decoder:* An encoder-decoder architecture model that achieves decoupling between phoneme-level, word-level and utterance-level scoring.

*Baseline+Decoder+C:* A convolutional variation of Gradformer by introducing a simple 1D convolution layer with kernel size of 3 into the transformer encoder to help capture local information.

It should be noted that the embedding dimensions of the model above are all 48, and the number of layers for both the encoder and decoder (if exists) is 3.

The ablation results are summarized in Table II. From which, we have the following observations:

- Due to the limited size of speechocean762, Baseline does not perform well when embedding dimension is 48, as the number of parameters becomes excessively large. This is consistent with the observation in [14].
- Baseline+Decoder outperforms Baseline on all other metrics except for a slight decrease on completeness. This implies that the decoupling can correctly handle the data correlation issue between each granularity and partially overcome the problem of low-quality indicators affecting high-quality indicators in multi-task learning.
- Compared with Baseline+Decoder, Baseline+Decoder+C shows significant improvement on phoneme-level and word-level performance, demonstrating the importance of convolution.
- Gradformer, with novel convolution modules, improves the overall performance on various aspects of three granularities except for stress, demonstrating that our convolution module not only can better capture fine-grained information but also help the model obtain more comprehensive and meaningful representations.

*Convolution Module Ablation:* The convolution module is crucial to improve the performance of Gradformer. To further investigate how the topology of convolution module affects the final performance of the model, we set up ablation experiments for the topology of the convolution module and other modules. Thereinto, different topology of the convolution module and other modules are defined as follows:

*Before encoder:* We only apply a convolution module once to the GOP features before they enter the transformer encoder, without adding any convolution modules in the transformer encoder.

TABLE III
ABLATION EXPERIMENT RESULTS OF THE EFFECT OF DIFFERENT TOPOLOGICAL STRUCTURES BETWEEN CONVOLUTION MODULE AND OTHER MODULES ON MODEL PERFORMANCE

| Topology | MSE↓ | Phone↑ | Word↑ | Utterace↑ |
|---|---|---|---|---|
| Before encoder | 0.083 ±0.001 | 0.626 ±0.002 | 0.593 ±0.004 | 0.749 ±0.003 |
| Serial | 0.082 ±0.001 | 0.630 ±0.004 | 0.600 ±0.006 | 0.749 ±0.003 |
| Parallel* | **0.079** **±0.001** | **0.646** **±0.004** | **0.614** **±0.006** | **0.756** **±0.003** |

Due to space limitation, we only present three of the most representative metrics, namely phoneme accuracy, word-level and utterance-level total score. * denotes the setting used in gradformer.

*Serial:* We add a convolution module in each transformer encoder layer, but the convolution module and self-attention module are in series, with the features passing through the convolution module before the self-attention module.

*Parallel:* Similar to the second one, the convolution module and self-attention module are parallel. The outputs of the two modules are fused by addition and then enter the FFN module.

The experimental results are shown in Table III. From the table, we can draw the following conclusions:

- Incorporating convolution module into transformer encoder for multiple feature extractions is beneficial, rather than using it only once before the encoder.
- Self-attention is generally good at modeling global information, while convolution focuses more on local information. Serial combination disrupts the independence of these two types of information extraction, so parallel processing is a better choice.

*Transformer Layer Ablation:* We also explore the effect of network depth on the model performance. We keep the number of encoder and decoder layers consistent and test the performance of Gradformer with 1, 3, and 6 layers, respectively. It should be noted that the GOPT uses a 3-layer transformer encoder. The HiPAMA, due to its fixed network architecture, cannot be compared in terms of network depth, but its number of parameters is similar to that of GOPT.

The results are shown in Table IV. From which, we can find that our model achieves significantly better performance than

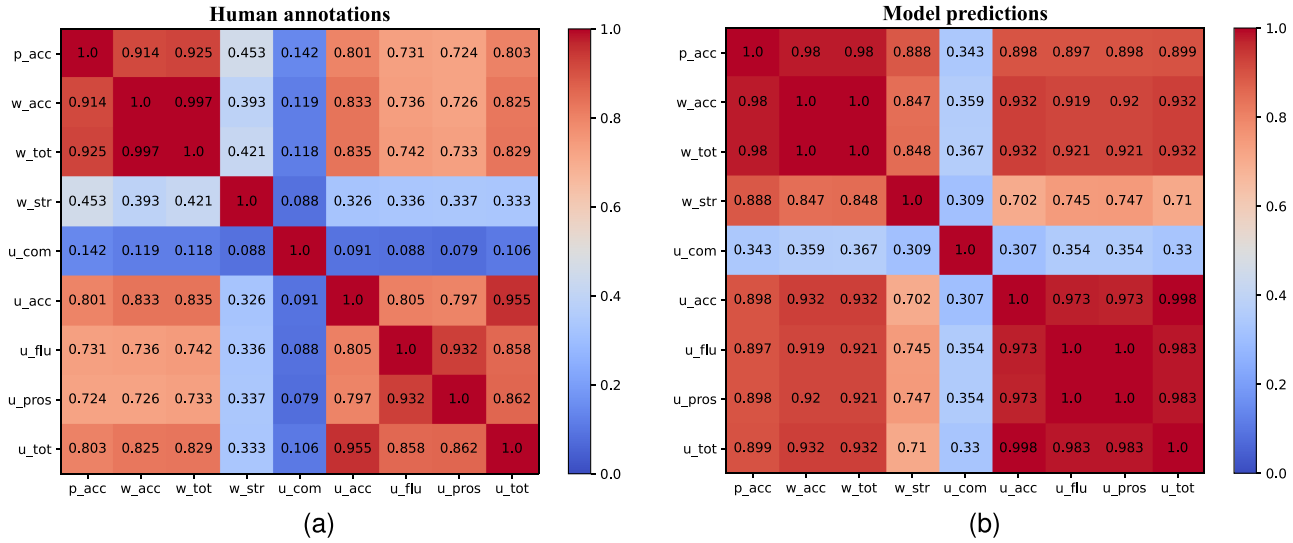


Fig. 5. Correlation matrix of different metrics at three granularities. Thereinto, p_acc stands for phoneme-level accuracy; w_acc, w_tot and w_str stand for word-level accuracy, total and stress, respectively; u_com, u_acc, u_flu, u_pros and u_tot stand for utterance-level completeness, accuracy, fluency, prosodic and total score, respectively.

TABLE IV
EFFECT OF THE NUMBER OF LAYERS ON MODEL PERFORMANCE

| Model | MSE↓ | Phone↑ | Word↑ | Utterace↑ |
|---|---|---|---|---|
| 1layer | 0.081 ±0.001 | 0.635 ±0.002 | 0.602 ±0.003 | 0.753 ±0.002 |
| 3layers* | **0.079 ±0.001** | **0.646 ±0.004** | **0.614 ±0.006** | **0.756 ±0.003** |
| 6layers | 0.081 ±0.001 | 0.637 ±0.006 | 0.606 ±0.007 | 0.748 ±0.006 |

TABLE V
EFFECT OF THE SIZE OF EMBEDDING DIMENSION ON MODEL PERFORMANCE

| Embedding size | MSE↓ | Phone↑ | Word↑ | Utterace↑ |
|---|---|---|---|---|
| 24 | 0.080 ±0.001 | 0.641 ±0.004 | 0.605 ±0.005 | 0.746 ±0.004 |
| 48* | **0.079 ±0.001** | **0.646 ±0.004** | **0.614 ±0.006** | **0.756 ±0.003** |
| 84 no projection | 0.081 ±0.000 | 0.640 ±0.002 | 0.606 ±0.004 | 0.753 ±0.003 |
| 128 | 0.081 ±0.000 | 0.640 ±0.006 | 0.611 ±0.005 | 0.747 ±0.004 |

No projection layer is needed when the embedding dimension is 84.

GOPT with 3 layers of encoder, even with only 1 layer of encoder and 1 layer of decoder. It also outperforms HiPAMA in terms of phoneme- and word-level scoring. In addition, the performance of Gradformer reaches its best with a 3-layer encoder and a 3-layer decoder. However, it begins to decline with 6 layers due to the excessively large number of parameters that are difficult to optimize.

*Embedding Dimension Ablation:* Finally, we investigate the effect of different embedding dimensions on the performance of the model. Keeping the number of layers of encoder and decoder as 3, we test the performance of Gradformer with embedding dimensions of 24, 48, 84, and 128. The results are summarized in Table V.

From the table, we can find that even when the embedding dimension is only 24, the model can still achieve good phoneme-level and word-level performance except for utterance level. We suppose that the information carried by the 24-dimensional encoder output is too limited, making it difficult for utterance queries to obtain enough useful information to help with utterance-level scoring. Furthermore, when the embedding dimension is 48, the model performs the best, while the model's performance decreases when the embedding dimension is 84 or 128 due to the excessive number of model parameters.

### C. Data Correlation Analysis

As mentioned in section I, we find that there is a high correlation between phoneme-level scores and word-level scores. To demonstrate this, we conduct experiments to analyze the correlation between scores at various granularities and aspects. We average the phoneme-level scores and word-level scores of each utterance of the training set, and then compute the correlation between each pair of aspects. The visualization of the results is shown in Fig. 5.

From Fig. 5, we have the following observations and conclusions:

- From the fourth and fifth rows of the Fig. 5(a), it can be found that the correlation between the two indicators, word-level stress and utterance-level completeness, and other indicators is very low, due to the extremely imbalanced data distribution of these two indicators.
- It can be found from the top-left area of the Fig. 5(a) that there is a high correlation between fine-grained scores such as phoneme-level accuracy, word-level accuracy, and total score.

- It can be found from the bottom-right and top-right areas of the Fig. 5(a) that the various aspect scores at the larger granularity of the utterance level are highly correlated with each other. However, the correlation between phoneme-level, word-level and utterance-level scores is not very high, indicating that utterance-level scores are more influenced by the overall impression rather than being determined by a specific fine-grained indicator.
- Jointly considering Figs. 5(a) and (b), we observe that the correlation between different granularities of model scoring is higher than that of experts. In addition, the correlation within various granularities of model scoring is higher than the correlation between different granularities, which is consistent with human scoring. In other words, our model is capable of providing scores similar to those of experts.

The above analysis demonstrate that it is reasonable to use the proposed encoder-decoder architecture to independently process two sets of scores with different correlations.

## VI. Limitations and Future Works

Despite achieving optimal performance, our model does not perfectly address the issue of imbalanced data distribution in the dataset, resulting in poor performance in terms of utterance completeness and word stress. Moreover, the handcrafted GOP features extracted by the acoustic model are not updated during the model training. Consequently, the entire system is not end-to-end. In future work, we will explore an end-to-end system based on self-supervised learning pre-trained models, aiming to overcome the aforementioned limitations.

## VII. Conclusion

In this paper, we propose a novel multi-aspect multi-granularity pronunciation assessment model named Gradformer. The model uses a transformer encoder to encode the acoustic features and output fine-grained scores. Then, a fixed number of utterance queries interact with the encoded acoustic features through cross-attention mechanisms multiple times for larger granularity scoring. Through the encoder-decoder architecture, Gradformer effectively solves the data correlation issue among various granularity and aspect scores and the problem that a model cannot obtain effective utterance-level representations in pronunciation assessment task. In addition, we introduce a novel convolution module in parallel with the self-attention module into the encoder to form a convolution-augmented transformer encoder, which can effectively capture local information between acoustic features and obtain more comprehensive and meaningful feature representations, significantly improving the model's performance on various granularities and aspects. The experimental results demonstrate that the proposed method outperforms the previous state-of-the-art multi-aspect multi-granularity method in all 9 metrics. Specifically, there is an average increase of 2.3% in phoneme-level and word-level PCC. The most significant improvement, observed at the utterance-level completeness, amounts to 4.2%. These results demonstrate the effectiveness of our method. We hope that Gradformer can serve as a strong baseline to facilitate future research on pronunciation assessment.

## Acknowledgment

The authors would like to thank the Associate Editor and the anonymous reviewers for their deep and careful work, which is much helpful in improving this paper.

## References

[1] M. Eskenazi, "An overview of spoken language technology for education," *Speech Commun.*, vol. 51, no. 10, pp. 832–844, 2009.
[2] K. Li, X. Qian, and H. M. Meng, "Mispronunciation detection and diagnosis in L2 english speech using multidistribution deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 1, pp. 193–207, Jan. 2017.
[3] C. Agarwal and P. Chakraborty, "A review of tools and techniques for computer aided pronunciation training (CAPT) in english," *Educ. Inf. Technol.*, vol. 24, no. 6, pp. 3731–3743, 2019.
[4] C. T. García, V. Cardeñoso-Payo, M. J. Machuca, D. E. Mancebo, A. Ríos, and T. Kimura, "Improving pronunciation of spanish as a foreign language for L1 japanese speakers with japañol CAPT tool," in *Proc. IberSPEECH*, 2018, pp. 97–101.
[5] C. T. García, "Design and evaluation of mobile computer-assisted pronunciation training tools for second language learning," Ph.D. dissertation, Universidad de Valladolid, Valladolid, Spain, 2020.
[6] C. T. García, D. E. Mancebo, E. C. Arenas, C. G. Ferreras, and V. Cardeñoso-Payo, "Assessing pronunciation improvement in students of english using a controlled computer-assisted pronunciation tool," *IEEE Trans. Learn. Technol.*, vol. 13, no. 2, pp. 269–282, Apr.–Jun. 2020.
[7] M. P. Black, J. Tepperman, and S. S. Narayanan, "Automatic prediction of children's reading ability for high-level literacy assessment," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 1015–1028, May 2011.
[8] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Commun.*, vol. 30, no. 2-3, pp. 95–108, 2000.
[9] J. Shi, N. Huo, and Q. Jin, "Context-aware goodness of pronunciation for computer-assisted pronunciation training," in *Proc. Interspeech*, 2020, pp. 3057–3061.
[10] B. Lin, L. Wang, X. Feng, and J. Zhang, "Automatic scoring at multi-granularity for L2 pronunciation," in *Proc. Interspeech*, 2020, pp. 3022–3026.
[11] J. Tepperman and S. S. Narayanan, "Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2005, pp. 937–940.
[12] C. Cucchiarini, H. Strik, and L. Boves, "Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology," *J. Acoust. Soc. Amer.*, vol. 107, no. 2, pp. 989–999, 2000.
[13] J. P. Arias, N. B. Yoma, and H. Vivanco, "Automatic intonation assessment for computer aided language learning," *Speech Commun.*, vol. 52, no. 3, pp. 254–267, 2010.
[14] Y. Gong, Z. Chen, I. Chu, P. Chang, and J. R. Glass, "Transformer-based multi-aspect multi-granularity non-native english speaker pronunciation assessment," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 7262–7266.
[15] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[16] H. Do, Y. Kim, and G. G. Lee, "Hierarchical pronunciation assessment with multi-aspect attention," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5, doi: 10.1109/ICASSP49357.2023.10095733.
[17] J. Zhang et al., "speechocean762: An open-source non-native english speech corpus for pronunciation assessment," in *Proc. Interspeech*, 2021, pp. 3710–3714.
[18] B. Lin and L. Wang, "Fast task-specific adaptation in spoken language assessment with meta-learning," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 7257–7261.
[19] H. Strik, K. P. Truong, F. de Wet, and C. Cucchiarini, "Comparing different approaches for automatic pronunciation error detection," *Speech Commun.*, vol. 51, no. 10, pp. 845–852, 2009.
[20] S. Sudhakara, M. K. Ramanathi, C. Yarra, and P. K. Ghosh, "An improved goodness of pronunciation (GOP) measure for pronunciation evaluation with DNN-HMM system considering HMM transition probabilities," in *Proc. Interspeech*, 2019, pp. 954–958.
[21] V. Laborde, T. Pellegrini, L. Fontan, J. Mauclair, H. Sahraoui, and J. Farinas, "Pronunciation assessment of japanese learners of french with GOP scores and phonetic information," in *Proc. Interspeech*, 2016, pp. 2686–2690.

[22] W. Hu, Y. Qian, and F. K. Soong, "A new dnn-based high quality pronunciation evaluation for computer-aided language learning (CALL)," in *Proc. Interspeech*, 2013, pp. 1886–1890.
[23] W. Hu, Y. Qian, F. K. Soong, and Y. Wang, "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," *Speech Commun.*, vol. 67, pp. 154–166, 2015.
[24] Y. Wang and L. Lee, "Improved approaches of modeling and detecting error patterns with empirical analysis for computer-aided pronunciation training," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2012, pp. 5049–5052.
[25] H. Huang, H. Xu, X. Wang, and W. Silamu, "Maximum f1-score discriminative training criterion for automatic mispronunciation detection," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 23, no. 4, pp. 787–797, Apr. 2015.
[26] R. Duan, T. Kawahara, M. Dantsuji, and H. Nanjo, "Cross-lingual transfer learning of non-native acoustic modeling for pronunciation error detection and diagnosis," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 391–401, 2020.
[27] A. Lee, "Language-independent methods for computer-assisted pronunciation training," Ph.D. dissertation, Massachusetts Inst. Technol., Cambridge, USA, 2016.
[28] K. Fu, S. Gao, K. Wang, W. Li, X. Tian, and Z. Ma, "Improving non-native word-level pronunciation scoring with phone-level mixup data augmentation and multi-source information," 2022, *arXiv:2203.01826*.
[29] S. Cheng, Z. Liu, L. Li, Z. Tang, D. Wang, and T. F. Zheng, "Asr-free pronunciation assessment," in *Proc. Interspeech*, 2020, pp. 3047–3051.
[30] B. Lin and L. Wang, "Deep feature transfer learning for automatic pronunciation assessment," in *Proc. Interspeech*, 2021, pp. 4438–4442.
[31] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 30–42, Jan. 2012.
[32] B. Lin and L. Wang, "Attention-based multi-encoder automatic pronunciation assessment," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 7743–7747.
[33] A. T. Liu, S. Yang, P. Chi, P. Hsu, and H. Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 6419–6423.
[34] A. T. Liu, S. Li, and H. Lee, "TERA: Self-supervised learning of transformer encoder representation for speech," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2351–2366, 2021.
[35] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 12449–12460.
[36] W. Hsu et al., "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3451–3460, 2021.
[37] S. Bannò and M. Matassoni, "Proficiency assessment of L2 spoken english using wav2vec 2.0," in *Proc. Spoken Lang. Technol. Workshop*, 2022, pp. 1088–1095.
[38] E. Kim, J. Jeon, H. Seo, and H. Kim, "Automatic pronunciation assessment using self-supervised speech representation learning," in *Proc. Interspeech*, 2022, pp. 1411–1415.
[39] X. Xu, Y. Kang, S. Cao, B. Lin, and L. Ma, "Explore WAV2VEC 2.0 for mispronunciation detection," in *Proc. Interspeech*, 2021, pp. 4428–4432.
[40] L. Peng, K. Fu, B. Lin, D. Ke, and J. Zhang, "A study on fine-tuning WAV2VEC2.0 model for the task of mispronunciation detection and diagnosis," in *Proc. Interspeech*, 2021, pp. 4448–4452.
[41] M. Wu, K. Li, W. Leung, and H. Meng, "Transformer based end-to-end mispronunciation detection and diagnosis," in *Proc. Interspeech*, 2021, pp. 3954–3958.
[42] W. Liu et al., "An ASR-free fluency scoring approach with self-supervised learning," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5, doi: 10.1109/ICASSP49357.2023.10095311.
[43] S. Kriman et al., "Quartznet: Deep automatic speech recognition with 1D time-channel separable convolutions," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, 2020, pp. 6124–6128.
[44] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 10, pp. 1533–1545, Oct. 2014.
[45] M. Tu, A. Grabek, J. Liss, and V. Berisha, "Investigating the role of L1 in automatic pronunciation evaluation of L2 speech," in *Proc. Interspeech*, 2018, pp. 1636–1640.
[46] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 5206–5210.
[47] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, vol. 9, pp. 249–256.
[48] S. Mao, F. K. Soong, Y. Xia, and J. Tien, "A universal ordinal regression for assessing phoneme-level pronunciation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 6807–6811.
[49] B. Lin and L. Wang, "Exploiting information from native data for non-native automatic pronunciation assessment," in *Proc. Spoken Lang. Technol. Workshop*, 2022, pp. 708–714.
[50] W. Liu et al., "Leveraging phone-level linguistic-acoustic similarity for utterance-level pronunciation scoring," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5, doi: 10.1109/ICASSP49357.2023.10096699.

**Hao-Chen Pei** received the bachelor's degree in computer science and technology from Soochow University, Suzhou, China, in 2022. Currently, He is a graduate student at the School of Software, Shandong University, Jinan, China. His current research interests include machine learning, automatic speech recognition, and pronunciation assessment.

**Hao Fang** received the B.S. degree in intelligence science and technology from Hangzhou Dianzi University, Hangzhou, China, in 2022. He is currently pursuing the M.S. degree with the School of software, Shandong University, Jinan, China. His research interests include computer vision, saliency detection, and video segmentation.

**Xin Luo** received the Ph.D. degree in computer science from Shandong University, Jinan, China, in 2019. He is currently an Assistant Professor with the School of Software, Shandong University, Jinan, China. His research interests mainly include machine learning, multimedia retrieval and computer vision. He has published over 20 papers on TIP, TKDE, ACM MM, SIGIR, WWW, IJCAI, et al. He serves as a Reviewer for ACM International Conference on Multimedia, International Joint Conference on Artificial Intelligence, AAAI Conference on Artificial Intelligence, the IEEE Transactions on Cybernetics, Pattern Recognition, and other prestigious conferences and journals.

**Xin-Shun Xu** (Senior Member, IEEE) received the M.S. and Ph.D. degrees in computer science from Shandong University, China, in 2002, and Toyama University, Japan, in 2005, respectively. He is currently a Professor with the School of Software, Shandong University. He joined the School of Computer Science and Technology at Shandong University as an Associate Professor in 2005, and joined the LAMDA group of Nanjing University, China, as a postdoctoral fellow in 2009. From 2010 to 2017, he was a Professor at the School of Computer Science and Technology, Shandong University. He is the founder and the leader of MIMA (Machine Intelligence and Media Analysis) group of Shandong University. His research interests include machine learning, information retrieval, data mining and image/video analysis and retrieval. He has published in TIP, TKDE, TMM, TCSVT, AAAI, CIKM, IJCAI, MM, SIGIR, WWW and other venues. He also serves as a SPC/PC member or a reviewer for various international conferences and journals, e.g. AAAI, CIKM, CVPR, ICCV, IJCAI, MM, TCSVT, TIP, TKDE, TMM and TPAMI.