# An Effective Hierarchical Graph Attention Network Modeling Approach for Pronunciation Assessment

Bi-Cheng Yan, *Student Member, IEEE*, and Berlin Chen, *Member, IEEE*

***Abstract*—Automatic pronunciation assessment (APA) manages to quantify second language (L2) learners' pronunciation proficiency in a target language by providing fine-grained feedback with multiple aspect scores (e.g., accuracy, fluency, and completeness) at various linguistic levels (i.e., phone, word, and utterance). Most of the existing efforts commonly follow a parallel modeling framework, which takes a sequence of phone-level pronunciation feature embeddings of a learner's utterance as input and then predicts multiple aspect scores across various linguistic levels. However, these approaches neither take the hierarchy of linguistic units into account nor consider the relatedness among the pronunciation aspects in an explicit manner. In light of this, we put forward an effective modeling approach for APA, termed HierGAT, which is grounded on a hierarchical graph attention network. Our approach facilitates hierarchical modeling of the input utterance as a heterogeneous graph that contains linguistic nodes at various levels of granularity. On top of the tactfully designed hierarchical graph message passing mechanism, intricate interdependencies within and across different linguistic levels are encapsulated and the language hierarchy of an utterance is factored in as well. Furthermore, we also design a novel aspect attention module to encode relatedness among aspects. To our knowledge, we are the first to introduce multiple types of linguistic nodes into graph-based neural networks for APA and perform a comprehensive qualitative analysis to investigate their merits. A series of experiments conducted on the speechocean762 benchmark dataset suggests the feasibility and effectiveness of our approach in relation to several competitive baselines.**

***Index Terms*—Automatic pronunciation assessment (APA), computer-assisted pronunciation training, deep regression models, pre-training mechanism.**

## I. Introduction

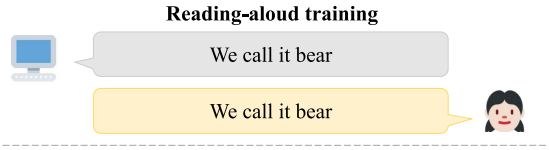
With the rising trend of globalization, an ever-growing number of people are willing or being asked to learn foreign languages. In response to this surging demand for foreign language learning, computer-assisted pronunciation training (CAPT) systems have garnered significant research attention, as they can offer L2 (second-language) learners a range of stress-free and self-directed scenarios to practicing pronunciation skills [1]. Among other things, CAPT systems have a broad spectrum of applications, which not only provide timely and informative feedback for L2 learners to improve their speaking skills [2], [3], but also serve as a handy reference for professionals (e.g., interviewers and examiners) on standardized tests to relieve their workload [4].As a crucial ingredient of CAPT, automatic pronunciation assessment (APA) aims to quantify oral proficiency and provide fine-grained feedback to learners by predicting multiple aspect scores at various linguistic levels [5], [6]. An APA system is typically instantiated in a read-aloud scenario, where an L2 learner is presented with a text prompt and instructed to pronounce it correctly. Early studies for APA mostly focused on single-aspect assessment, typically developed by extracting sets of hand-crafted features to construct scoring modules accordingly, such as phone-level accuracy [7], [8], [9], word-level lexical stress [10], [11], or various aspects of utterance-level proficiency scores [12], [13], [14]. Although these efforts possess the advantage of being easily interpretable, they rely solely on surface features and postulate implicitly that scoring aspects of different linguistic levels are independent of each other, often leading to suboptimal performance. More recently, with the synergistic breakthroughs in neural model architectures and optimization algorithms [15], [16], research endeavors have been advocated for the notion of multi-aspect and multi-granular pronunciation assessment, which creates a unified scoring model to jointly evaluate pronunciation proficiency at various linguistic levels (i.e., phone, word, and utterance) with diverse aspects (e.g., accuracy, fluency, and completeness), as the running example depicted in Fig. 1. Prior arts along this line of research usually follow a parallel modeling paradigm [17], [18], [19], wherein Transformer-based neural networks serve as the archetype to take as input a sequence of phone-level pronunciation feature embeddings of a learner's utterance while simultaneously predicting multiple aspect scores across different linguistic levels without accounting for their subtle dependency.

Albeit effective, such parallel modeling approaches suffer from at least two weaknesses. First, these approaches fall short in taking advantage of the hierarchical structure of an utterance, which assumes that all phones within a word are of equal importance and insufficiently capture the word-level structure cues that are prominent in the composition of an utterance-level representation when solely based on phone-level pronunciation features. Second, the relatedness among pronunciation aspects is mostly sidelined. As an illustration, we visualize the correlation matrix in Fig. 2, which shows the Pearson Correlation Coefficient (PCC) between any pair of expert annotated aspect scores on the training set. We can observe that except for the aspects of utterance-completeness and word-stress, the remaining aspects present strong correlations not only within the same linguistic

Received 1 February 2024; revised 29 June 2024; accepted 9 August 2024. Date of publication 26 August 2024; date of current version 9 September 2024. This work was supported by E.SUN Bank under Grant 202308-NTU-02. The associate editor coordinating the review of this article and approving it for publication was Dr. Samuel Thomas. *(Corresponding author: Berlin Chen.)*

The authors are with the Department of Computer Science and Information Engineering, National Taiwan Normal University, Taipei 11677, Taiwan (e-mail: 80847001s@ntnu.edu.tw; berlin@ntnu.edu.tw).

Digital Object Identifier 10.1109/TASLP.2024.3449111




**Automatic Pronunciation Assessment Results**

| Utterance level | | Word level | | | Phone level | |
|---|---|---|---|---|---|---|
| Aspects | Scores | Words | Aspects | Scores | Phones | Scores |
| Accuracy | 1.6 | We | Accuracy | 2 | W | 2.0 |
| | | | Stress | 2 | | |
| Fluency | 1.8 | | Total | 2 | IY0 | 2.0 |
| | | Call | Accuracy | 2 | K | 2.0 |
| Completeness | 2 | | Stress | 2 | AO0 | 1.8 |
| | | | Total | 2 | L | 1.8 |
| Prosody | 1.8 | It | Accuracy | 2 | IH0 | 2.0 |
| | | | Stress | 2 | T | 2.0 |
| | | | Total | 2 | B | 2.0 |
| Total | 1.6 | Bear | Accuracy | 1.2 | EH0 | 1.0 |
| | | | Stress | 2 | R | 1.0 |
| | | | Total | 1.2 | | |

Fig. 1. An example curated from the speechocean762 dataset [21] illustrates the evaluation flow of an APA system in the reading-aloud training scenario, which offers an L2 learner in-depth pronunciation feedback.


Fig. 2. Correlation matrix on training set of expert annotations. Each element in the matrix reveals the PCC score of any pair of measuring aspects.

level but also across different linguistic levels.[1] Building on these observations, we in this paper present a novel APA method, dubbed HierGAT, which leverages hierarchical graph attention architecture to jointly model the intrinsic structure of an utterance and meanwhile considers the interactions among disparate aspects at the same and across different linguistic levels. By representing an input utterance as a heterogeneous graph, HierGAT updates and learns meaningful node representations for various linguistic units through message passing, reformulating the APA task into a node estimation problem. More specifically, HierGAT first constructs a heterogeneous graph structured hierarchically with utterance, word, and phone levels, where each level contains its corresponding types of nodes. Subsequently, HierGAT learns hierarchical representations for various linguistic units by information aggregation with a dedicated designed mechanism of iterative intra-word, phone-word, inter-word, and word-utterance message passing. Furthermore, in order to capture the interactions among aspects, we also design an aspect attention module [20] to resolve the relatedness among aspects of the same linguistic level. Comprehensive experiments conducted on the speechocean762 benchmark dataset show that our proposed method achieves significant and consistent improvements over several cutting-edge baselines [21].

Our main contributions of this work can be summarized as follows:

1) To our knowledge, we are the first to construct a heterogeneous graph network for tackling the task of automatic pronunciation assessment, characterizing an input utterance by its constituent words and the corresponding phones. The proposed heterogenous graph is able to capture several types of relations, including utterance-word, word-word, phone-word, and phone-phone ones.
2) Our proposed modeling framework is highly flexible and can be easily extended to integrate other supra-segmental linguistic units, such as syllables and intonational phrases.
3) Without resort to any external self-supervised pre-trained feature extractors, our method is shown to outperform current cutting-edge methods [22], [23], [24]. Ablation studies and qualitative analysis confirm its effectiveness in capturing hierarchical structure of an utterance.

The remainder of this paper is organized as follows. In Section II, we review related work in the subfields of CAPT, including mispronunciation detection and diagnosis, as well as automatic pronunciation assessment. Next, we elaborate on the proposed modeling framework in Section III. Sections IV and V detail the experimental setup and results, respectively. Finally, in Section VI, we conclude the paper with a discussion of our findings and future work.

## II. Related Work

Research and development on CAPT date back to pioneering efforts conducted in the 60's of the last century [7], [25], which has attracted surging attention in recent years, showing good promise by leveraging many advanced deep learning technologies [26], [27], [28]. According to the types of diagnostic feedback being provided, research endeavors of CAPT fall into two broad categories: one is phone-level mispronunciation detection and diagnosis (MDD), and the other is automatic pronunciation assessment (APA).

### A. Mispronunciation Detection and Diagnosis

The goal of mispronunciation detection and diagnosis focuses primarily on pinpointing phone-level erroneous pronunciation segments and provide L2 learners with the corresponding diagnostic feedback [28], [29], [30]. Early work relies on pronunciation scoring based approaches, which make use of a well-trained acoustic model to derive various types of confidence measurements as indicators of mispronunciation. Commonly used indicators include, but is not limited to, phone durations [32], [33], likelihood ratios [29], [34], phone posterior probabilities [35], and their combinations [39]. Goodness of pronunciation (GOP) and its descendants are the most iconic instantiations [7]. The

[1]Both the aspects of utterance completeness and word stress suffer from label imbalance problems, with more than 90% of the assessments receiving the highest score [18].

principal idea behind GOP is to compute the ratio between the likelihoods of a canonical phone and the most likely pronounced phones predicted by an acoustic model via forced-alignment of the canonical phone sequence of a given text prompt to the speech signal uttered by a learner. A phone segment is identified as a mispronunciation if the corresponding likelihood ratio do not exceed a given phone-dependent threshold. However, pronunciation scoring based methods are untenable to provide specific diagnostic feedback for the mispronounced phone segments.

In order to better obtain informative diagnosis feedback, dictation-based methods alternatively frame MDD as a phone recognition task by employing a free-phone recognition process to dictate the most likely phone sequence uttered by an L2 learner. Consequently, the erroneous pronunciation portions can be easily identified by comparing the dictation result with the corresponding canonical phone sequence. To this end, for example, Leung et al. made attempts to employ a phone recognizer trained with the connectionist temporal classification (CTC) loss [36]. However, the conditional independence assumption of the CTC loss may hinder the fidelity of dictation results. As a workaround, Yan et al. [37] exploited the hybrid CTC-Attention ASR model as the dictation model and sought to capture deviant (non-categorical) phone productions by augmenting the canonical phone dictionary. To integrate historical mispronunciation patterns of L2 learners, Zhang et al. utilized a phonetic recurrent neural network Transducer (RNN-T) to transcribe learners' speech, which synergized RNN-T modeling with weakly supervised data augmentation and diversified beam search, so as to provide learners with comprehensive diagnostic feedback on erroneous pronunciation segments [38].

### B. Automatic Pronunciation Assessment

Automatic pronunciation assessment concentrates more on assessing and providing a suite of comprehensive pronunciation scores on a few specific aspects or traits of spoken language usage to reflect a learner's pronunciation quality [39], [40], [41]. Prior arts on APA focused exclusively on the single-aspect assessment, typically through constructing scoring modules individually to predict a holistic pronunciation proficiency score on a targeted linguistic level or some specific aspect with different sets of hand-crafted features. These hand-crafted features can be extracted directly from a learner's input speech signal or the associated transcription generated by automatic speech recognition (ASR), which may consist of acoustic features, confidence of recognized linguistic units (phones, syllables, or words) [43], time-alignment information [44], and other statistic measures such as fundamental frequency [45], speech rate, and filled pause [46]. As one of the pioneering attempts, Cucchiarini et al. utilized an ASR system to transcribe an input speech signal and then derived various statistic measures related to phonation quantity, like rate of speech, duration of pauses, and frequency of filled pauses, from phone-level alignment to assess the fluency of read speech [46]. Following a similar vein, Ferrer et al. quantified word-level stress according to the time-alignment information at the syllable nucleus, where Gaussian mixture models were employed to represent the distributions of prosody- and spectrum-related features, aiming to estimate possible manners of lexical stress (e.g., unstressed, primary, or secondary) for each syllable within a word [10]. Yet another approach is to frame lexical stress detection and pitch accent detection as a sequence labeling task [47], which created a synergy of canonical lexical stress patterns and syllable-based prosodic features to evaluate the learner's speaking proficiency and immediately provides diagnostic feedback on syllable boundaries. Despite that the aforementioned ASR-driven methods have the advantage of being easily interpretable, their performance is inevitably vulnerable to the errors made by ASR, potentially leading to an unfaithful rendering of the linguistic content inherent in a learner's utterance. This issue can be mitigated by replacing hand-crafted features with automatically derived features, through either an one-stage [48], [49], [50] or a multi-stage [51], [52] estimation process.

Due to the unprecedented breakthroughs brought about by deep learning, the notion of multi-aspect and multi-granular pronunciation assessment has made inroad into APA with good promise. Several neural scoring models have been proposed to jointly evaluate pronunciation proficiency at various linguistic levels with diverse aspects. For example, Lin et al. streamlined three linguistic-level scoring modules and introduced a single-aspect multi-granular hierarchical APA architecture, utilizing an attention mechanism to extract and aggregate linguistic representations from low to high linguistic levels for multi-granularity proficiency estimation [39]. Gong et al. proposed a GOP feature-based Transformer (GOPT) to jointly model multi-aspect pronunciation assessment at multiple granularities with a multi-task learning mechanism [19]. Since then, several subsequent extensions to the GOPT framework were developed. For example, Chao et al. integrated prosodic and self-supervised learning (SSL) based features into GOPT to achieve multi-view, multi-granularity, and multi-aspect (3M) pronunciation modeling [17]. Do et al. investigated the issue of data imbalance incurred by APA and proposed a score-balanced loss function that aims to nudge the prediction bias of a neural model towards the majority scores (i.e., high-performing proficiency scores) by assigning higher penalties when the predicted score belongs to a minority class and vice versa [18]. Departing from the aforementioned methods, we in this paper propose a novel hierarchical APA model based on a hierarchical graph attention Transformer architecture. By representing an input utterance as a hierarchical graph, the proposed method updates and learns the node representations across several linguistic levels by message passing, and aptly turns the pronunciation assessment task into a node regression problem.

## III. Methodology

### A. Problem Formulation

In this paper, we explore the task of multi-aspect and multi-granular automatic pronunciation assessment (APA), as illustrated in Fig. 1. Given an input utterance $\mathrm{U}$ which consists of a sequence of audio signals $\mathrm{X}$ uttered by an L2 learner and a text prompt $\mathrm{T}$ that the learner is expected to pronounce correctly, the objective of APA is to estimate proficiency scores for multiple aspects across various linguistic granularities. The proposed model (dubbed HierGAT) represents an input utterance as a hierarchical graph and formulates automatic pronunciation assessment as a node regression task.

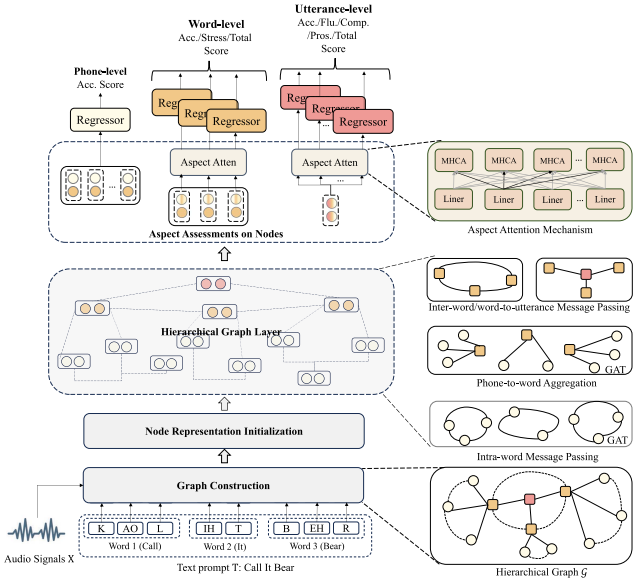



Fig. 3. The overall model architecture of HierGAT. We first construct a hierarchical graph for an input utterance and then learn hierarchy-aware representations for three linguistic level nodes (i.e., phone, word, and utterance nodes). The learned representations of different linguistic nodes are then fed to the corresponding regressors to access various aspect scores.

Formally, we denote a set of linguistic granularities as $\mathrm{G} = \{p, w, u\}$, where $p, w, u$ stands for the phone, word, and utterance levels, respectively. For a linguistic level $g \in \mathrm{G}$, our APA model targets to quantify pronunciation skill of an L2 learner with respect to multiple aspects, represented by $\mathrm{A}^g = \{a_1^g, a_2^g, \ldots, a_{N_g}^g\}$, where $N_g$ is the number of aspects, and each $a_j^g$ is framed as a regression task that estimates a sequence of aspect score $\mathbf{y}_{a_j}^g \in [0, 2]$. The overall model architecture of HierGAT is depicted in Fig. 3, which mainly consists of three parts: 1) node representation initialization, which is responsible for generating node features for phone-, word-, and utterance-level units; 2) hierarchical graph layer, which learns hierarchy-aware node representations with iteratively message passing; 3) aspect assessments on nodes, where regressors are built upon the learned node (aspect) representations to predict the corresponding proficiency score sequence.

## B. Graph Construction

For an input text prompt T with $M$ words and $N$ phones, we first represent it as a hierarchical graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ stands for the node set and $\mathcal{E}$ are edges between nodes. In order to utilize the linguistic structures of the text prompt, the undirected hierarchical graph $\mathcal{G}$ contains phone nodes, word nodes, and an utterance node, defined by $\mathcal{V} = \mathcal{V}_p \cup \mathcal{V}_w \cup \mathcal{V}_u$, where each phone node $v_{p_n} \in \mathcal{V}_p$ corresponds to a phone $p_n$ in the canonical phone sequence of T, $v_{w_m} \in \mathcal{V}_w$ represents a word $w_m$ in T, and $v_u \in \mathcal{V}_u$ is a special supernode that signifies the whole utterance. The edge connection of $\mathcal{G}$ is defined as $\mathcal{E} = \mathcal{E}_p \cup \mathcal{E}_w \cup \mathcal{E}_{pw} \cup \mathcal{E}_{wu}$, where $\mathcal{E}_p$ denotes the connections between phone nodes within a particular word, $\mathcal{E}_w$ stands for the connections between word nodes within the text prompt, $\mathcal{E}_{pw}$ is the cross-linguistic connections between a word node and its constituent phone nodes, and $\mathcal{E}_{wu}$ describes cross-linguistic connections between an utterance node and its constituent word nodes. A schematic depiction of the hierarchical graph is illustrated in Fig. 4.

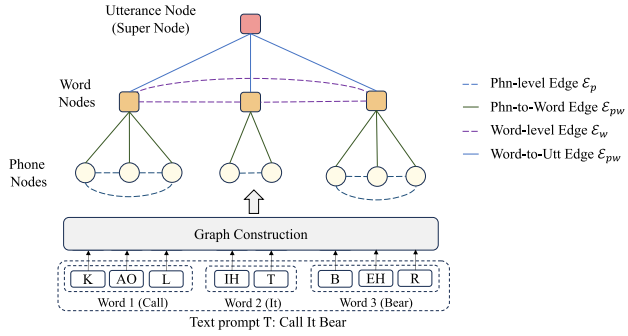



Fig. 4. An illustration of a hierarchical graph for an input text prompt comprising three types of nodes: an utterance node, word nodes, and phone nodes.

*Edge Connection:* This hierarchical graph $\mathcal{G}$ is an unweighted graph; namely, the connected node pairs have weight 1, and disconnected node pairs have weight 0 in the adjacency matrix $\mathcal{A}$. For the phone-level connections, an edge $e_{p_{i,j}}$ connects phone nodes $v_{p_i}$ and $v_{p_j}$ if they are within the same word, facilitating the aggregation of intra-word information. All word nodes are fully-connected by word-level edges which seeks to capture inter-word information. For the cross-linguistic relations, the phone-to-word edge $e_{pw_{i,k}}$ connects the phone node $v_{p_i}$ to its corresponding word node $v_{w_k}$, enabling message passing from the phone nodes to word nodes. All word nodes are linked to an utterance supernode $v_u$ with word-to-utterance connections, thereby gathering information from the word nodes to an utterance node. In the resulting hierarchical graph, each phone node can only interact with neighboring phone nodes within the same word, while interacting indirectly with the phone nodes of other words via word-level node connections.

## C. Node Representation Initialization

*Pronunciation Feature Extraction:* For an input utterance U, we start by converting the text prompt into a canonical phone sequence through looking up a pronunciation dictionary. Next, various pronunciation features are extracted to assess the L2 learner's pronunciation quality at the phone level, which are then concatenated and projected to obtain a sequence of dense pronunciation features $\mathrm{X}_p = (\mathbf{x}_{p_1}, \mathbf{x}_{p_2}, \ldots, \mathbf{x}_{p_N})$:

$$\mathrm{X}_p = \mathrm{W}_x \cdot \tilde{\mathrm{X}}_p + \mathbf{b}_x, \tag{1}$$

$$\tilde{\mathrm{X}}_p = [\mathrm{E}^{\mathrm{GOP}} || \mathrm{E}^{\mathrm{Eng}} || \mathrm{E}^{\mathrm{Dur}} || \mathrm{E}^{\mathrm{Fbank}}], \tag{2}$$

where $\mathrm{E}^{\mathrm{GOP}}$ is goodness of pronunciation-based (GOP) feature [7], [26], $\mathrm{E}^{\mathrm{Dur}}$ and energy $\mathrm{E}^{Eng}$ are prosodic features of duration and energy statistics, and spectral features $\mathrm{E}^{\mathrm{Fbank}}$ (viz. log Mel-filterbank features). $\mathrm{W}_x$ and $\mathbf{b}_{\boldsymbol{x}}$ are learnable parameters, and $||$ denotes concatenation operation. Notably, the extracted pronunciation features include both frame- and phone-level features. To align with the phone-level features, the frame-level features are averaged over time frames based on aligned phone boundaries. In addition, the word-level pronunciation features are denoted by

$\mathrm{X}_w = (\mathbf{x}_{w_1}, \mathbf{x}_{w_2}, \ldots, \mathbf{x}_{w_M})$, where $\mathbf{x}_{w_m}$ stands for the features of the $m$-th word, which is the sum of its constituent (connected) phone-level pronunciation features.

*Node Representation Initialization:* We explore to use a convolution-augmented Branchformer (ConvBFR) [56] to initialize node features at both the phone and word levels, with the aim of capturing contextualized pronunciation patterns at their respective granularities. Subsequently, the utterance-level node is initialized by summing the features of its connected words. More specifically, the proposed ConvBFR comprises two parallel branches to dynamically model various ranged contexts at different linguistic granularities, with one branch following the original Transformer network architecture employing self-attention to capture long-range dependencies [53] and the other branch utilizing a convolution module introduced in [55] to capture local dependencies. Specifically, for the phone-level nodes, we first map the canonical phone sequence into phone embeddings $\mathrm{E}_p$ via a phone and position embedding layer, which are then point-wisely added to $\mathrm{X}_p$ to provide a rendition of the positional information and phonetic characteristics. Next, a phone encoder is followed to initialize the phone-level node representations $\tilde{\mathrm{H}}_p$:

$$\tilde{\mathrm{H}}_p^0 = \mathrm{X}_p + \mathrm{E}_p, \tag{3}$$

$$\tilde{\mathrm{H}}_p = \mathrm{PhnEnc}\left(\tilde{\mathrm{H}}_p^0\right), \tag{4}$$

where $\mathrm{PhnEnc}(\cdot)$ consists of 3 stacked ConvBFR blocks. Afterward, for the word-level nodes, $\mathrm{X}_w$ are enriched with the textual information $\mathrm{E}_w$, and then a word encoder is employed to generate the initial node representations $\tilde{\mathrm{H}}_w$:

$$\tilde{\mathrm{H}}_w^0 = \mathrm{X}_w + \mathrm{E}_w, \tag{5}$$

$$\tilde{\mathrm{H}}_w = \mathrm{WordEnc}\left(\tilde{\mathrm{H}}_w^0\right), \tag{6}$$

where $\mathrm{E}_w$ is generated by passing the text prompt T into a word and position layer, and $\mathrm{WordEnc}(\cdot)$ encompasses a stack of 3 ConvBFR blocks. For the utterance node representation $\tilde{\mathbf{h}}_u$, it is initialized by summing the representations of its connected words $\tilde{\mathbf{h}}_u = \sum_{k \in \mathcal{N}_u} \mathrm{X}_{w_k}$, where $\mathcal{N}_u$ is the set of the neighboring word nodes of the utterance node $v_u$.

### D. Hierarchical Graph Layer

After constructing the hierarchical graph $\mathcal{G}$ with the adjacency matrix $\mathcal{A}$ and node representations at three linguistic levels $(\tilde{\mathrm{H}}_p \cup \tilde{\mathrm{H}}_w \cup \tilde{\mathbf{h}}_u)$, we use the graph attention network (GAT) [54] to update the node representations.

*Graph Attention Network:* Given a constructed graph $\mathcal{G}$ with the corresponding hidden representations of input nodes H, a GAT layer updates a node $v_i$ with the representation $\mathbf{h}_i$ as follows:

$$e_{ij} = \mathrm{LeakyReLU}(\mathrm{W}_a[\mathrm{W}_q\mathbf{h}_i || \mathrm{W}_k\mathbf{h}_j], \tag{7}$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{l \in \mathcal{N}_i} \exp(e_{il})}, \tag{8}$$

$$\mathbf{u}_i = \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \mathrm{W}_v \mathbf{h}_j\right), \tag{9}$$

where $\sigma$ is an activation function instantiated with rectified linear units (ReLU), $\mathcal{N}_i$ is the set of neighboring nodes of $v_i$, $\alpha_{ij}$ stands for the attention weight between $\mathbf{h}_i$ and $\mathbf{h}_j$, and $\mathrm{W}_a$, $\mathrm{W}_q$, $\mathrm{W}_k$, and $\mathrm{W}_v$ are trainable weight matrices. The multi-head attention can be expressed by

$$\mathbf{u}_i = ||_{t=1}^{T}\, \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij}^t \mathrm{W}_v^t \mathbf{h}_j\right), \tag{10}$$

where $T$ is the number of independent attention mechanisms, $\alpha_{ij}^t$ are normalized attention weights computed by the $t$-th attention mechanism, and $\mathrm{W}_v^t$ is the corresponding transformation matrix. Next, a residual connection is in turn employed to prevent gradient vanishing. The updated node representation $\mathbf{h}_i'$ can be denoted by

$$\mathbf{h}_i' = \mathbf{h}_i + \mathrm{W}_o\mathbf{u}_i, \tag{11}$$

where $\mathrm{W}_o$ is a linear projection adjusting the dimension of $\mathbf{u}_i$ to align with $\mathbf{h}_i$. Finally, stacking on each graph attention layer, we introduce a position-wise feed-forward (FFN) layer consisting of two linear transformations, in the same vein as Transformer [15].

*Hierarchical Massage Passing:* The proposed hierarchical graph layer begins by updating representations of phone nodes using their locally-neighboring phones within a word via the intra-word message passing. Then, the intermediate representations of a word node $\mathrm{H}_w'$ are derived by gathering information from its constituent phone nodes:

$$\mathrm{H}_{p \leftarrow p} = \mathrm{GAT}\left(\tilde{\mathrm{H}}_p, \tilde{\mathrm{H}}_p\right), \tag{12}$$

$$\mathrm{H}_w' = \mathrm{GAT}\left(\tilde{\mathrm{H}}_w, \mathrm{H}_{p \leftarrow p}\right), \tag{13}$$

where $\mathrm{H}_{p \leftarrow p}$ is updated representations of phone nodes. $\mathrm{GAT}(\tilde{\mathrm{H}}_p, \tilde{\mathrm{H}}_p)$ denotes that $\tilde{\mathrm{H}}_p$ is linear projected to form query, key, and value matrices, respectively, while $\mathrm{GAT}(\tilde{\mathrm{H}}_w, \mathrm{H}_{p \leftarrow p})$ means that $\tilde{\mathrm{H}}_w$ is used as query matrix, and $\mathrm{H}_{p \leftarrow p}$ serves as the key and value matrices, respectively. To propagate information from word nodes to the utterance node, we first perform inter-word message passing to update the representations of word nodes for capturing the interactions among words. The representation of the utterance node is then refined by aggregating information from its connected word nodes:

$$\mathrm{H}_{w \leftarrow w} = \mathrm{GAT}\left(\mathrm{H}_w', \mathrm{H}_w'\right), \tag{14}$$

$$\mathbf{h}_{u \leftarrow w} = \mathrm{GAT}\left(\tilde{\mathbf{h}}_u, \mathrm{H}_{w \leftarrow w}\right), \tag{15}$$

where $\mathrm{H}_{w \leftarrow w}$ and $\mathbf{h}_{u \leftarrow w}$ are updated representations of word and utterance nodes, respectively. In $\mathrm{GAT}(\tilde{\mathbf{h}}_u, \mathrm{H}_{w \leftarrow w})$, $\tilde{\mathbf{h}}_u$ acts as a query vector, and $\mathrm{H}_{w \leftarrow w}$ is projected to construct the key and value matrices. In this way, HierGAT updates and learns hierarchy-aware node representations through the hierarchical graph layer at three linguistic levels.[2]

### E. Aspect Assessments on Nodes

The proposed HierGAT model is a unified architecture that can be optimized in an end-to-end manner using the mean square

[2]The proposed hierarchical message passing can be easily generalized to k-hop message passing by iterating (11) to (15) $k$ times. However, since the proposed hierarchical graph $\mathcal{G}$ is undirected, it may include redundant information in the $k$-hop neighbors.

error (MSE) loss for each aspect at different linguistic levels. Once the aspect representations are obtained, a fully-connected layer acting as the regressor is in turn employed to calculate the corresponding aspect score sequence.

*Aspect Assessments via Phone-level Nodes:* For the phone-level node aspect assessment, we first concatenate phone node representations with their corresponding word node representations, which are then activated by the ReLU function to derive the aspect representations $\mathrm{H}_p$ for phone nodes:

$$\mathrm{H}_p = \sigma(\mathrm{W}_p\,[\mathrm{H}_{p\leftarrow p}||\mathrm{H'}_{w\leftarrow w}])\,, \tag{16}$$

$\mathrm{H'}_{w\leftarrow w}$ is a sequence of augmented word-level node representations, repeated for each phone node based on the phone-to-word connections. Next, the regression head is built on top of $\mathrm{H}_p$ to access phone accuracy scores.

*Aspect Assessments via Word-level Nodes:* For the word-level node aspect assessments, the word node representations are first concatenating with the average representations of their constituent phone nodes:

$$\mathrm{H}_w = \sigma(\mathrm{W}_w\left[\mathrm{H}_{w\leftarrow w}||\bar{\mathrm{H}}_w\right])\,, \tag{17}$$

where $\bar{\mathrm{H}}_w = (\bar{\mathbf{h}}_{w_1}, \bar{\mathbf{h}}_{w_w}, \ldots, \bar{\mathbf{h}}_{w_M})$ with $\bar{\mathbf{h}}_{w_m}$ being the average vector of constituent phone-level representations derived from $\mathrm{H}_{p\leftarrow p}$ for the $m$-th word. Afterward, an aspect attention mechanism is introduced to capture the relatedness among the aspects [20], [41]. Specifically, for the $j$-th word-level aspect, the intermediate aspect representations $\tilde{\mathrm{H}}^{w_j}$ are linearly projected from $\mathrm{H}_w$, and a multi-head cross-attention (MHCA) with a masking strategy is followed to derive word-level aspect representations $\mathrm{H}^{w_j}$ from a collection of all intermediate representations $\mathrm{C}^w = [\tilde{\mathrm{H}}^{w_1}, \tilde{\mathrm{H}}^{w_2}, \ldots, \tilde{\mathrm{H}}^{w_{N_w}}]$. The following equations illustrate the operations of aspect attention:

$$\tilde{\mathrm{H}}^{w_j} = \mathrm{W}_{w_j} \cdot \mathrm{H}_w + \mathbf{b}_{w_j}\,, \tag{18}$$

$$\mathrm{H}^{w_j} = \mathrm{MHCA}\left(\tilde{\mathrm{H}}^{w_j}, \mathrm{C}^w\right), \tag{19}$$

where $\mathrm{W}_{w_j}$ and $\mathbf{b}_{w_j}$ are aspect specific projection weights. In the operation of MHCA, $\tilde{\mathrm{H}}^{w_j}$ is linearly projected as query matrix, while $\mathrm{C}^w$ serves as key and value matrices. The masking strategy ensures that the output representation at a specific position is only influenced by the other aspects of the word. Lastly, the aspect representations $\mathrm{H}^{w_j}$ are taken as the input to the corresponding regressor for evaluating the $j$-th word-level pronunciation aspect.

*Utterance-level Node Aspect Estimations:* For the utterance-level node aspect assessments, the node representations $\mathrm{H}_{p\leftarrow p}$ and $\mathrm{H}_{w\leftarrow w}$ are individually fed into an attention pooling mechanism to obtain holistic vector representations $\bar{\mathbf{h}}_{p\leftarrow p}$ and $\bar{\mathbf{h}}_{w\leftarrow w}$ at the phone and word levels, respectively. The utterance node representation $\mathbf{h}_u$ is then generated by packing these vectors together via concatenation and projection:

$$\bar{\mathbf{h}}_{p\leftarrow p} = \mathrm{AttPool}_p\ (\mathrm{H}_{p\leftarrow p})\,, \tag{20}$$

$$\bar{\mathbf{h}}_{w\leftarrow w} = \mathrm{AttPool}_w\ (\mathrm{H}_{w\leftarrow w})\,, \tag{21}$$

$$\mathbf{h}_u = \sigma\left(\mathrm{W}_u\left[\mathbf{h}_{u\leftarrow w}\,\middle|\middle|\,\bar{\mathbf{h}}_{p\leftarrow p}\,\middle|\middle|\,\bar{\mathbf{h}}_{w\leftarrow w}\right] + \mathbf{b}_u\right), \tag{22}$$

where $\sigma$ is the ReLU function, and $\mathrm{W}_u$ and $\mathbf{b}_u$ are trainable parameters. After that, an aspect attention mechanism is performed on $\mathbf{h}_u$ to derive various aspect representations $\mathrm{h}^{u_j}$, which are then passed through regression heads to derive the utterance-level proficiency scores.

TABLE I
STATISTICS OF THE SPEECHOCEAN762 DATASET

| Linguistic Granularities | Pronunciation Aspects | Score Interval | Number of Counts | |
|---|---|---|---|---|
| | | | Train | Test |
| Phone | Accuracy | [0, 2] | 47,076 | 47,369 |
| Word | Accuracy | [0, 10] | 15,849 | 15,967 |
| | Stress | | | |
| | Total | | | |
| Utterance | Accuracy | [0, 10] | 2500 | 2500 |
| | Completeness | | | |
| | Fluency | | | |
| | Prosody | | | |
| | Total | | | |

*Model Optimization:* The total loss is computed as a weighted sum of the MSE losses from different levels, where the loss at each linguistic level is calculated as an average of multiple aspects:

$$\mathcal{L}_{APA} = \frac{\lambda_p}{N_p}\sum_{i_p}\mathcal{L}_{p^{i_p}} + \frac{\lambda_w}{N_w}\sum_{i_w}\mathcal{L}_{w^{i_w}} + \frac{\lambda_u}{N_u}\sum_{i_u}\mathcal{L}_{u^{i_u}}\,, \tag{23}$$

where $\mathcal{L}_{p^{i_p}}$, $\mathcal{L}_{w^{i_w}}$, and $\mathcal{L}_{u^{i_u}}$ are phone-level, word-level, and utterance-level losses at disparate aspects, respectively; $\lambda_p$, $\lambda_w$, and $\lambda_u$ are adjustable parameters controlling the influence of different granularities; and $N_p$, $N_w$, and $N_u$ refer to the numbers of aspects at phone, word, and utterance levels.

## IV. Experimental Setups

### A. Experimental Data

We conducted APA experiments on the speechocean762 dataset, a publicly available open-source dataset specifically designed for multi-aspect and multi-granular pronunciation assessment [21]. This dataset contains 5000 English-speaking recordings spoken by 250 Mandarin L2 learners. The training and test sets are of equal size, each of which has 2500 utterances. This corpus contains comprehensive annotation information, and the pronunciation proficiency scores were evaluated at multiple linguistic granularities alongside disparate aspects. Table I summarizes the detailed statistics of used speech corpus. Each score was independently assigned by five experts using the same rubrics, and the final score was determined by selecting the median value from the five scores.

### B. Pronunciation Feature Extractions

*GOP Feature:* To extract the GOP feature, we first aligned audio signals X with the text prompt T by using an ASR model[3] to obtain the timestamps for each phone in the canonical phone sequence. Next, frame-level phonetic posterior probabilities were produced by the ASR model and then averaged

[3]A public-assessable ASR model trained with English speech corpus: https://kaldi-asr.org/models/m13.

over time based on the phone-level timestamps. The resulting phone-level posterior probabilities are converted into a GOP feature vector as a combination of log phone posterior (LPP) and log posterior ratio (LPR). Owing to the used ASR model containing 42 phones, the GOP feature of a canonical phone $p$ was thus represented by an 84-dimensional vector:

$$[\mathrm{LPP}\left(p_1\right), \ldots, \mathrm{LPP}\left(p_{42}\right),\ \mathrm{LPR}\left(p_1|p\right) \ldots,\ \mathrm{LPR}\left(p_{42}|p\right)], \tag{24}$$

$$LPP\left(p_i\right) = \log p\left(p_i|\mathbf{o}; \mathrm{t}_s, \mathrm{t}_e\right),$$

$$= \frac{1}{t_e - t_s + 1} \sum_{t=t_s}^{t_e} \log p\left(p_i|\mathrm{o}_t\right), \tag{25}$$

$$\mathrm{LPR}\ \left(p_i|p\right) =\ \log p\left(p_i|\mathbf{o}; \mathrm{t}_s, \mathrm{t}_e\right) - \log p\left(p|\mathbf{o}; \mathrm{t}_s, \mathrm{t}_e\right), \tag{26}$$

where LPR is the log posterior ratio between phones $p_i$ and $p$; $t_s$ and $t_e$ are the start and end timestamps of phone $p$, and $o_t$ is the input acoustic observation of the time frame $t$.

*Energy Feature:* The energy feature is a 7-dimensional vector comprised of statistics (viz. [mean, std, median, mad, sum, max, min]) over phone segments, where the root-mean-square energy (RMSE) is employed to compute energy value for each time frame, with 25-millisecond windows and a stride of 10 milliseconds.

*Duration Feature:* The duration feature is a 1-dimensional vector indicating the length of each phone segment in seconds.

*Log Mel-filterbank Feature:* The log Mel-filterbank feature is an 80-dimensional vector computed over 25-millisecond windows with 10-millisecond strides, which are then averaged over each phone segment to from the corresponding phone-level feature.

### C. Implementation Details

*Model Configurations:* In accordance with [19], we normalized utterance-level and word-level scores to the same scale as the phone-level score [0, 2] for training APA models. Both the feature encoders at the phone and word levels consist of three blocks, each with a single-head attention mechanism and 24 hidden units. The proposed hierarchical graph layer consists of 3 stack graph attention layers, each with a single attention head and a hidden size of 24.

*Training Configurations:* In the training phase, we use a batch size of 25 and apply Adam optimizer with a learning rate 1e-3. To ensure the reliability of our experimental results, we repeated 5 independent trials, each consisting of 100 epochs using different random seeds with a learning rate scheduler that warms up at the beginning and cuts in half every five epochs after the 20-th epoch. The experimental results are reported by averaging 100 experiments with the minimum phone-level MSE values, where the mean and standard deviation values for different evaluation metrics, as described below, are reported.

*Evaluation Metrics:* The primary evaluation metric is PCC, which measures the linear correlation between predicted scores and ground-truth scores. In addition, mean squared error value (MSE) is used to assess phone-level accuracy.

### D. Compared Methods

We first report the inter-annotator agreement for the five annotators (**Human-agreement**), and compare the proposed model with the following top-of-the-line methods:

*Lin2021 [14]:* This method uses a single-aspect multi-granular pronunciation scorer with a hierarchical architecture which takes phone-level surface features as inputs and assesses the learner's utterance-level accuracy score.

*Kim2022 [28]:* This approach employs a single-aspect pronunciation assessment model designed to separately measure oral skills on the utterance level. Each aspect-specific scorer is implemented as a Bi-LSTM network, with the input features extracted from a self-supervised learning model (HuBERT Large [23]).

*LSTM [19]:* This method frames multi-aspect and multi-granular pronunciation assessment as sequential labeling tasks, deriving a sequence of phone-level features and utilizing a 3-layer LSTM to generate the representations across different linguistic units based on distinct timestamps.

*GOPT [19]:* This model extends the sequential modeling strategy by replacing the backbone model of LSTM with a 3-stacked Transformer block and performs pronunciation assessment at various granularities with diverse aspects.

*Ryu2023 [40]:* This method leverages is a unified model architecture that adopts a self-supervised model as the backbone model, which is optimized with phone recognition and utterance-level pronunciation assessment tasks jointly.

*Gradformer (GFR) [42]:* This model approaches multi-aspect and multi-granular pronunciation assessment tasks with a granularity-decoupled Transformer network, which decouples the linguistic units of an utterance into two sub-groups: phone and word levels, and utterance level. A Conformer encoder is employed to jointly model pronunciation aspects at phone and word levels, while a Transformer decoder takes a sequence of aspect vectors as the input and interacts with the encoder outputs for utterance-level pronunciation scoring.

*HiPAMA [41]:* This model is built on top of a hierarchical architecture for multi-aspect and multi-granular pronunciation assessment, which more resembles our model in relation to all the other methods. In contrast to our model, HiPAMA extracts high-level pronunciation features from low-level features based on a simple average pooling mechanism. In addition, the aspect attention mechanism used in HiPAMA performs on the internal logistics, while our model operates on the intermediate representations.

*3M [17]:* This approach is a state-of-the-art APA model with a paralle pronunciation modeling technique, which enhances the input features of GOPT with three types of SSL-based features to capture supra-segmental pronunciation cues, while also integrating vowel and consonant features to enhance phone-level textual embeddings.

*HierCB [56]:* This approach is aslo a cutting-edge APA model with a hierarchical neural structure, stacking multiple ConvBFR blocks at three linguistic granularities (phone, word, and utterance) for pronunciation modeling and a combination of mean pooling and attention pooling mechanisms is further employed to capture cross-granularity relationships.

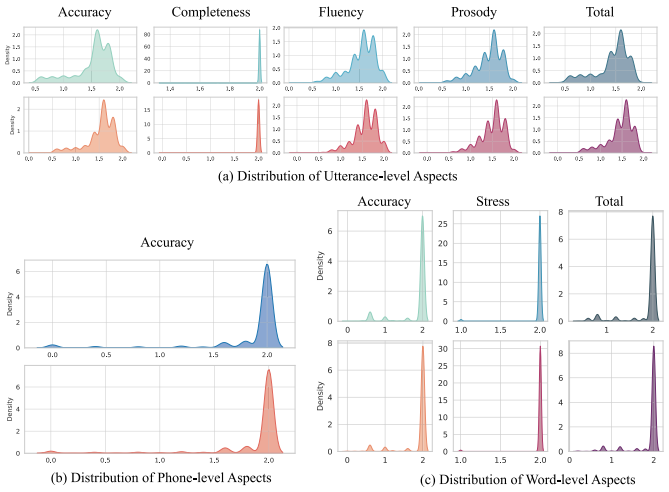



Fig. 5. Score distributions for the aspects across different linguistic granularities on the training and test sets: (a) utterance-level aspects (accuracy, completeness, fluency, prosody, and total score), (b) word-level aspects (accuracy, stress, and total score), and (c) phone-level accuracy score. The first row of each figure shows the distribution on the training set, while the second row shows the distribution on the test set.

## V. Experimental Results

### A. Qualitative Analysis

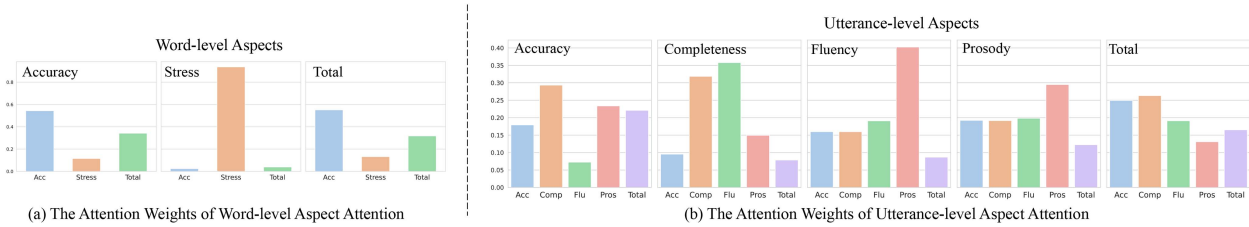
*Distributions of Aspect Scores:* Before launching into a series of experiments on the APA tasks, we perform quantitative analysis on the score distributions of aspects across different linguistic granularities on both the training and test sets. As shown in Fig. 5, the speechocean762 is a well-curated dataset, where though the majority of aspect scores skew towards high proficiency scores, the distributional trends are consistent between the training and test sets. Furthermore, both the distributions of utterance-completeness and word-stress demonstrate a notable high-score-biased phenomenon. The scores for these two aspects are densely distributed on high-performing labels, and a plurality of the data instances belong to a small number of labels. This label imbalance problem poses a challenge for regression models to determine proficiency scores accurately.

*Qualitative Visualization of Attention Weights in the Aspect Attention Mechanisms:* In the second set of experiments, we examine the relatedness among disparate aspects at both word and utterance levels on the training set by analyzing the attention weights of the aspect attention mechanisms when assessing a specific aspect score. Fig. 6(a) presents attention weights among the word-level aspects, which reveals the attention weights for the assessments on the accuracy and total aspects are influenced by various other aspects. In contrast, the aspect of stress is a specific evaluation task concerned with identifying emphasis on particular syllables within a word, resulting in attention weights being focused on itself [52]. We then move to analyzing the relatedness among the utterance-level pronunciation aspects. As shown in Fig. 6(b), the attention weights for the prosody and the total aspects are more uniformly distributed, whereas the fluency aspect is primarily complemented by the prosody aspect. This could be attributed to the fact that the total and prosody scores measure holistic oral skills, including speaking style, rhythm, and intonation. Consequently, our model considers multiple aspects when evaluating the total and prosody aspects. Interestingly, the aspect of completeness is primarily influenced by fluency and its own. Interestingly, the aspect of completeness is primarily influenced by fluency and its own. We postulate that our model reflects the halo effect present in the human annotations on the training set. Specifically, when assessing the completeness score, it seems that the decisions of human annotators might be influenced by the score pertaining to the prosody, making word-level pronunciation clarity maybe psychometrically redundant.

### B. Main Results

Table II presents the APA results on the speechocean762 dataset, organized into two groups, where the first group includes the results of models built upon the GOP-based features, while the second group for other models utilizing the SSL-based features. Furthermore, for fair comparisons, we report on the performance of GOPT and HierGAT variants, where the input features of these models are enhanced by concatenating GOP features with three types of SSL-based features (i.e., Wav2vec2.0, HuBERT, WavLM), following the processing flow suggested in [17].

With respect to the models built on the GOP-based features (the first group of Table II), we can make the following observations. First, on the whole, our model (HierGAT) consistently outperforms human-human agreement on all assessment tasks, expect for the aspect of utterance-completeness. Second, Lin2020, a single-aspect assessment method, fails to harness the dependency between aspects through the multi-task learning scheme, resulting in inferior performance compared to other multi-aspect and multi-granular pronunciation assessment models. Third, compared to the baseline methods with the parallel modeling techniques, HierGAT excels on most assessment tasks, particularly for assessments of higher linguistic granularities (utterance and word levels), achieving average improvements of 9.94%, and 8.28% over LSTM, and GOPT, respectively. This performance gain underscores the significance of capturing the hierarchical structure of an utterance when modeling cross-linguistic relationships with the proposed hierarchical graph layer. In terms of the hierarchical modeling architecture, our model outperforms HiPAMA across various pronunciation assessment tasks with an average improvement of up to 5.47%, while maintaining a top-performing phone-level accuracy score. HiPAMA generates high-level pronunciation features from phone-level pronunciation features with a simple average operation, which can be seen as fully-connected relationships with uniform weights. In contrast, our graph structures effectively prune unnecessary connections between phones or words when modeling cross-linguistic relationships. Finally, our model outperforms the state-of-the-art APA model, GFR, in most word- and utterance-level assessment tasks, achieving an average PCC score improvement of 3.56%, while maintaining comparable results at the phone-level pronunciation aspect. As opposed to GFR, our model streamlines the pronunciation assessment tasks at three linguistic levels with tactfully designed hierarchical architecture and a newly proposed aspect attention mechanism to capture the relatedness among aspects.




(a) The Attention Weights of Word-level Aspect Attention

(b) The Attention Weights of Utterance-level Aspect Attention

Fig. 6. Qualitative visualization of model parameters when predicting each aspect score on the speechcoean762 training data set. We show (a) the averaged attention values for word-level aspects, and (b) the averaged attention weights for utterance-level aspects.

TABLE II
PERFORMANCE EVALUATIONS OF OUR MODEL AND ALL COMPARED METHODS ON SPEECHOCEAN762, WHERE ACC., AND COMP. REFER TO THE ASPECTS OF ACCURACY AND COMPLETENESS, RESPECTIVELY

| Input Feat. | Model | Phone Score | | Word Score (PCC) | | | Utterance Score (PCC) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | MSE↓ | PCC↑ | Acc.↑ | Stress↑ | Total↑ | Acc.↑ | Comp.↑ | Fluency↑ | Prosody↑ | Total↑ |
| Human-agreement | | - | 0.555 | 0.589 | 0.212 | 0.602 | 0.618 | 0.658 | 0.665 | 0.651 | 0.675 |
| GOP | Lin2021 | - | - | - | - | - | - | - | - | - | 0.720 |
| | LSTM | 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 | 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 |
| | GFR | **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 |
| | HiPAMA | 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 |
| | HierGAT | 0.083 ±0.002 | 0.640 ±0.005 | **0.606** ±0.004 | 0.315 ±0.010 | **0.621 ±0.004** | 0.726 ±0.004 | **0.606 ±0.042** | **0.795 ±0.005** | **0.785 ±0.003** | **0.760 ±0.004** |
| SSL | Kim2022 | - | - | - | - | - | - | - | 0.780 | 0.770 | - |
| | Ryu2023 | - | - | - | - | - | 0.719 | - | 0.775 | 0.773 | 0.743 |
| | GOPT | 0.081 ±0.003 | 0.640 ±0.013 | 0.584 ±0.016 | 0.352 ±0.046 | 0.603 ±0.152 | 0.748 ±0.006 | 0.290 ±0.123 | 0.817 ±0.011 | 0.807 ±0.009 | 0.778 ±0.005 |
| | 3M | 0.078 ±0.001 | 0.656 ±0.005 | 0.598 ±0.005 | 0.289 ±0.033 | 0.617 ±0.005 | 0.760 ±0.004 | 0.325 ±0.141 | 0.828 ±0.006 | 0.827 ±0.008 | 0.796 ±0.004 |
| | HierCB | 0.076 ±0.000 | 0.680 ±0.000 | 0.630 ±0.003 | **0.355 ±0.033** | 0.645 ±0.003 | 0.772 ±0.002 | **0.677 ±0.121** | 0.827 ±0.003 | 0.823 ±0.004 | 0.796 ±0.001 |
| | HierGAT | **0.073 ±0.001** | **0.683 ±0.004** | **0.648 ±0.003** | 0.327 ±0.011 | **0.663 ±0.002** | **0.798 ±0.002** | 0.531 ±0.047 | **0.840 ±0.002** | **0.833 ±0.002** | **0.821 ±0.002** |

The best results for each group (GOP, SSL) in various pronunciation assessment tasks are highlighted in bold.

When we pair the GOP-based features with the SSL-based features, the assessment results are consistently improved across all pronunciation aspects. By combining the SSL-based features, HierGAT obtains average performance gains of 4.30%, 3.20% and 3.26% on phone, word, and utterance levels, respectively, compared to its base form. Second, as the focus is shifted to the single-aspect pronunciation scorer, we find that Kim2022 additionally with the SSL-based features can boost the assessment results for the utterance-level pronunciation assessments, reaching almost the same performance level with other multi-aspect and multi-granularity pronunciation scorers developed based on the GOP-features.

On a separate front, Ruy2023 further boots the performance of utterance-level assessments by jointly training the APA model with the MDD task. This highlights the potential of combining the APA and MDD tasks, as it encourages the APA model to produce more phonetic-aware representations. Finally, compared to the existing APA models with parallel neural architectures, our model demonstrates remarkable performance improvements across various aspects at three granularities, outperforming GOPT and 3M with lifts of 5.97% and 5.11%, respectively, in the PCC score. Furthermore, compared to the cutting-edge hierarchical APA model, HierCB, our model significantly excels for most assessments at word and utterance levels. This superiority stems from the efficacy of the proposed hierarchical graph attention layer in modeling cross-linguistic relationships.

### C. Ablation Studies

To better understand the contributions of different modules to the performance of HierGAT, we conduct here a series of ablation studies for in-depth analysis. First, we compare the prediction distributions of HierGAT with different input features thorugh boxplots for various aspects, as shown in Fig. 7. Next, we remove different model components of HierGAT and report the corresponding PCC scores for the accuracy evaluations at three granularities in Table III. Finally, we examine the learned weights in both the phone and word encoders when using the weighted combination mechanism in ConvBFR.

*Comparison of Input Features:* Fig. 7 shows the distribution of predicted scores estimated by HierGAT with different input features for various pronunciation aspects in the training set.

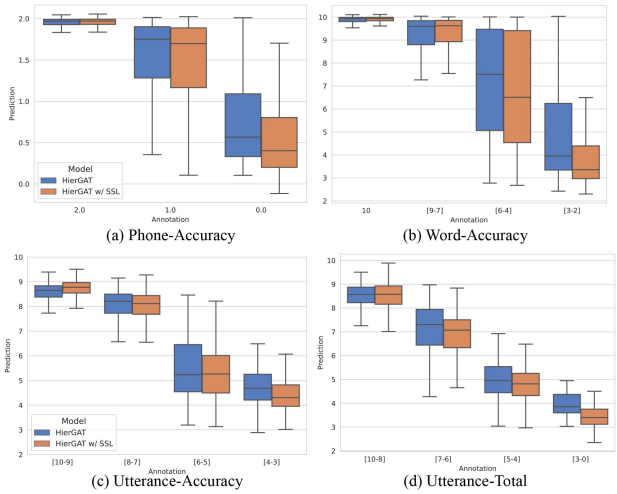



Fig. 7. Boxplots of the HierGAT's predictions for various pronunciation aspects in the development set using different input features.

TABLE III
THE ABLATION STUDIES ON SPEECHOCEAN762

| Models | Phone | Word | Utterance |
|---|---|---|---|
| *Model Components* | | | |
| w/ weighted merge | 0.637 | 0.612 | 0.724 |
| w/o aspect attention | 0.626 | 0.602 | 0.714 |
| w/o GAT | 0.615 | 0.532 | 0.714 |
| *Number of GAT Layers (Each layer with 1-head)* | | | |
| 3* | 0.640 | 0.606 | 0.726 |
| 2 | 0.639 | 0. 98 | 0.722 |
| 1 | 0.618 | 0.587 | 0.717 |
| *Number of Heads for GAT (In a 3-layer GAT setting)* | | | |
| 3 | 0.630 | 0.596 | 0.721 |
| 2 | 0.632 | 0.599 | 0.712 |
| 1* | 0.640 | 0.606 | 0.726 |

* Denotes the setting used in the HierGAT model.

First, we focus on the predictions at lower-level granularities (viz. phone and word levels). In Fig. 7(a) and (b), we observe that the predicted scores of these two models (HierGAT and HierGAT with SSL-based features) are more accurate for instances of high and low annotation scores, which are tightly concentrated around the high and low score intervals but are more scattered for the intermediate score interval. A possible reason is that the annotated scores of training instances for phone-accuracy and word-accuracy are densely located at high scores, leading to the model's predictions for instances with intermediate scores to be biased towards higher scores. Next, as we look at Fig. 7(c) and (d) for the utterance-level aspect assessments, we notice that the predicted scores for accuracy and total aspects have the tendency of consistently decreasing from high to low scores for both models, which closely aligns with the score intervals provided by human experts. This is evident that when the model additionally resorts to the SSL-based features, the corresponding predictions are more accurate for most aspects, resulting in smaller interquartile ranges compared to the base form of the model.

| Phone-level ConvBFR (3 Layers) | | | Word-level ConvBFR (3 Layers) | | |
|---|---|---|---|---|---|
| Layer | Attention | Convolution | Layer | Attention | Convolution |
| 1 | 0.539 | 0.461 | 1 | 0.611 | 0.389 |
| 2 | 0.892 | 0.108 | 2 | 0.802 | 0.198 |
| 3 | 0.244 | 0.756 | 3 | 0.772 | 0.228 |

Fig. 8. Visualization of branch weights with respect to various layers of ConvBFR at different linguistic levels (phone level and word level).

*Comparison of Model Components:* The first part of Table III presents an ablation study with the following settings: 1) replacing the concatenation operator with a weight average mechanism for merging two branches in both phone and word feature encoders [53], 2) removing the aspect attention mechanism, and 3) replacing the hierarchical graph layer with a simple attention pooling. First, we can observe that the weighted average mechanism is slightly worse than the concatenation operator, where we see performance drops at phone and utterance levels and a modest improvement at the word level. Next, we notice the performance significantly declines at the utterance level and slightly drops at the word-level when the aspect attention mechanisms are removed from the proposed hierarchical architecture. The proposed aspect attention mechanism can effectively leverage the relatedness among aspects, as evident by the proportional decrease in performance corresponding to the number of aspects at different linguistic granularities. Finally, the employ of the hierarchical graph layer is indispensable for HierGAT, as the removal of such a layer leads to performance degrades for all linguistic granularities.

*Depth and Width of GAT Layers:* In the second and third parts of Table III, we investigate the impact on the performance of HierGAT when varying the width or depth of the GAT layer in the proposed hierarchical graph layer. We observe that the model performance is gradually improved as the number of layers increases; however, this improvement is limited. Meanwhile, there is a tendency of performance degradation when with an increase in the number of heads. One possible reason is that Speechocean762 by itself is not a large-scale dataset, and our model is capable of sufficiently learning the data characteristics when equipped with a single GAT head. As such, to strike a balance between performance and computational efficiency, the proposed model is configured as a 3-layer GAT with a single attention head.

*Learned Weights for Merging Operation in the Phone and Word Encoders:* To examine the learned weights for merging two branches at the phone and word encoders, we visualize the average weights on the training set while accessing pronunciation aspects in the variant of HierGAT which replaces the concatenation operator with a weighted average mechanism in the ConvBFR blocks. As shown in Fig. 8, we can observe several certain patterns in the learned weights. For example, in the initial layers of the phone- and word-level ConvBFR modules, the two types of branches are utilized in an interleaving fashion, with the learned weights being distributed almost uniformly between the two branches. This indicates that both models use local

and global relationships to learn hidden representations. In the following layers of the phone encoder, the attention block and the convolution block are utilized, showing that global context and local relationships are equally important in the phone-level modeling. On the other hand, the word encoder leveraging consecutive attention blocks is observed, highlighting the importance of global dependencies in word-level pronunciation modeling.

## VI. Conclusion

In this paper, we have proposed HierGAT, a hierarchical graph-based architecture for automatic pronunciation assessment. Notably, we are the first to explore constructing a heterogeneous graph network to streamline the three linguistic units for the pronunciation assessment. Evaluation on the speechocean762 benchmark datasets proves the effectiveness of HierGAT and demonstrates capturing the language hierarchy and interactions between pronunciation aspects are beneficial to the assessments.

*Limitations and Future Work:* In this research, the proposed model focus on the "reading-aloud" pronunciation training scenario, where the assumption is that the L2 learner pronounces a predetermined target sentence correctly. This assumption restricts the applicability of our models to other learning scenarios, such as freely-speaking or open-ended conversations. We leave this for a future extension. We further plan to delve into resolving the data imbalance issue for the proposed model to enhance its generalizability to unseen learners.

## References


[1] A. Van Moere and R. Downey, "Technology and artificial intelligence in language assessment," in *Handbook of Second Language Assessment*. Boston, MA, USA: De Gruyter Mouton2016, pp. 341–357.

[2] M. Eskenazi, "An overview of spoken language technology for education," *Speech Commun.*, vol. 51, no. 10, pp. 832–844, 2009.

[3] K. Evanini and X. Wang, "Automated speech scoring for nonnative middle school students with multiple task types," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2013, pp. 2435–2439.

[4] K. Evanini, M. C. Hauck, and K. Hakuta, "Approaches to automated scoring of speaking for K–12 English language proficiency assessments," *ETS Res. Rep. Ser.*, vol. 2017, pp. 1–11, 2017.

[5] K. Li, X. Wu, and H. Meng, "Intonation classification for L2 English speech using multi-distribution deep neural networks," *Comput. Speech Lang.*, vol. 43, pp. 18–33, 2017.

[6] S. Banno, B. Balusu, M. J. F. Gales, K. M. Knill, and K. Kyriakopoulos, "View-specific assessment of L2 spoken English," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2022, pp. 4471–4475.

[7] 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.

[8] K. Li, X. Qian, and H. Meng, "Mispronunciation detection and diagnosis in L2 English speech using multi-distribution deep neural networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 1, pp. 193–207, Jan. 2017.

[9] S. Mao, F. Soong, Y. Xia, and J. Tien, "A universal ordinal regression for assessing phone-level pronunciation," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, 2022, pp. 6807–6811.

[10] L. Ferrer, H. Bratt, C. Richey, H. Franco, V. Abrash, and K. Precoda, "Classification of lexical stress using spectral and prosodic features for computer-assisted language learning systems," *Speech Commun.*, vol. 69, pp. 31–45, 2015.

[11] D. Korzekwa et al., "Detection of lexical stress errors in non-native (L2) English with data augmentation and attention," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2021, pp. 3915–3919.

[12] E. Coutinho et al., "Assessing the prosody of non-native speakers of English: Measures and feature sets," in *Proc. Lang. Resour. Eval. Conf.*, 2016, pp. 1328–1332.

[13] C. Cucchiarini et al., "Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology," *J. Acoustical Soc. Amer.*, vol. 107, no. 2, pp. 989–999, 2000.

[14] B. Lin and L. Wang, "Deep feature transfer learning for automatic pronunciation assessment," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2021. pp. 4438–4442.

[15] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pretraining of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2019, pp. 4171–4186.

[17] F.-A. Chao, T.-H. Lo, T.-I. Wu, Y.-T. Sung, and B. Chen, "3M: An effective multi-view, multigranularity, and multi-aspect modeling approach to English pronunciation assessment," in *Proc. IEEE Asia-Pac. Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2022, pp. 575–582.

[18] H. Do, Y. Kim, and G. G. Lee, "Score-balanced loss for multi-aspect pronunciation assessment," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2023, pp. 4998–5002.

[19] Y. Gong, Z. Chen, I.-H. Chu, P. Chang, and J. Glass, "Transformer-based multi-aspect multigranularity non-native English speaker pronunciation assessment," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 7262–7266.

[20] R. Ridley, L. He, X.-Y. Dai, S. Huang, and J. Chen, "Automated cross-prompt scoring of essay traits," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, pp. 13745–13753.

[21] J. Zhang et al., "Speechocean762: An open-source non-native English speech corpus for pronunciation assessment," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2021, pp. 3710–3714.

[22] 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, pp. 12449–12460.

[23] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE Trans. Audio Speech Lang. Process.*, vol. 29, pp. 3451–3460, 2021.

[24] S. Chen et al., "WavLM: Large-scale self-supervised pre- training for full stack speech processing," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1505–1518, Oct. 2022.

[25] E. B. Page, "Statistical and linguistic strategies in the computer grading of essays," in *Proc. Conf. Comput. Linguistics*, 1967, pp. 1–13.

[26] 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.

[27] Y. Qian et al., "Neural approaches to automated speech scoring of monologue and dialogue responses," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 8112–8116.

[28] E. Kim, J.-J. Jeon, H. Seo, and H. Kim, "Automatic pronunciation assessment using self-supervised speech representation learning," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2022, pp. 1411–1415.

[29] J. Shi, N. Huo, and Q. Jin, "Context-aware goodness of pronunciation for computer-assisted pronunciation training," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 3057–3061.

[30] B.-C. Yan, H.-W. Wang, Y.-C. Wang, and B. Chen, "Effective graph-based modeling of articulation traits for mispronunciation detection and diagnosis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5.

[31] C. Richter and J. Guðnason, "Relative dynamic time warping comparison for pronunciation errors," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5.

[32] Q.-T. Truong, T. Kato, and S. Yamamoto, "Automatic assessment of L2 English word prosody using weighted distances of F0 and intensity contours," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2018, pp. 2186–2190.

[33] C. Graham and F. Nolan, "Articulation rate as a metric in spoken language assessment," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 3564–3568.

[34] 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. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 954–958.

[35] S. Mao, Z. Wu, R. Li, X. Li, H. Meng, and L. Cai, "Applying multitask learning to acoustic-phonemic model for mispronunciation detection and diagnosis in L2 English speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 6254–6258.
[36] W.-K. Leung, X. Liu, and H. Meng, "CNN-RNN-CTC based end-to-end mispronunciation detection and diagnosis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 8132–8136.
[37] B.-C. Yan, M.-C. Wu, H.-T. Hung, and B. Chen, "An end-to-end mispronunciation detection system for L2 English speech leveraging novel anti-phone modeling," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 3032–3036.
[38] D. Y. Zhang, S. Saha, and S. Campbell, "Phonetic RNN-transducer for mispronunciation diagnosis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5.
[39] B. Lin, L. Wang, X. Feng, and J. Zhang, "Automatic scoring at multi-granularity for L2 pronunciation," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 3022–3026.
[40] H. Ryu, S. Kim, and M. Chung, "A joint model for pronunciation assessment and mispronunciation detection and diagnosis with multi-task learning," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2023, pp. 959–963.
[41] H. Do, Y. Kim, and G. G. Lee, "Hierarchical pronunciation assessment with multi-aspect attention," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5.
[42] H.-C. Pei, H. Fang, X. Luo, and X.-S. Xu, "Gradformer: A framework for multi-aspect multi-granularity pronunciation assessment," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 554–563, 2024.
[43] P. Muller, F. De Wet, C. Van Der Walt, and T. Niesler, "Automatically assessing the oral proficiency of proficient L2 speakers," in *Proc. Workshop Speech Lang. Technol. Educ.*, 2009, pp. 29–32.
[44] H. Franco et al., "EduSpeak: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications," *Lang. Testing*, vol. 27, no. 3, pp. 401–418, 2010.
[45] K. Laskowski, J. Edlund, and M. Heldner, "An instantaneous vector representation of delta pitch for speaker-change prediction in conversation dialogue system," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2008, pp. 5041–5044.
[46] 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.
[47] K. Li, S. Mao, X. Li, Z. Wu, and H. Meng, "Automatic lexical stress and pitch accent detection for L2 English speech using multi-distribution deep neural networks," *Speech Commun.*, vol. 96, pp. 28–36, 2018.
[48] L. Chen, J. Tao, S. Ghaffarzadegan, and Y. Qian, "End-to-end neural network based automated speech scoring," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 6234–6238.
[49] W. Liu et al., "An ASR-free fluency scoring approach with self-supervised learning," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5.
[50] K. Fu, S. Gao, S. Shi, X. Tian, W. Li, and Z. Ma, "Phonetic and prosody-aware self-supervised learning approach for non-native fluency scoring," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2023, pp. 949–953.
[51] S. Cheng, Z. Liu, L. Li, Z. Tang, D. Wang, and T. F. Zheng, "ASR-free pronunciation assessment," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 3047–3051.
[52] D. Korzekwa, J. Lorenzo-Trueba, T. Drugman, and B. Kostek, "Computer-assisted pronunciation training—Speech synthesis is almost all you need," *Speech Commun.*, vol. 142, pp. 22–33, 2022.
[53] Y. Peng, S. Dalmia, I. Lane, and S. Watanabe, "Branchformer: Parallel MLP-attention architectures to capture local and global context for speech recognition and understanding," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 17627–17643.
[54] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," in *Proc. Int. Conf. Learn. Representations*, 2018.
[55] A. Gulati et al., "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 5036–5040.
[56] B.-C. Yan, Y.-C. Wang, J.-T. Li, H.-W. Wang, W.-C. Chao, and B. Chen, "ConPCO: Preserving phoneme characteristics for automatic pronunciation assessment leveraging contrastive ordinal regularization," 2024, *arXiv:2406.02859*.

**Bi-Cheng Yan** (Student Member, IEEE) received the M.S. degree in computer science and information engineering in 2017 from National Taiwan Normal University, Taipei, Taiwan, where he is currently working toward the Ph.D. degree in computer science and information engineering. He was with ASUSTeK Computer Inc., Beitou, Taiwan, from 2017 to 2020. He is the author/coauthor of more than 20 academic publications. His research interests include computer-assisted language learning, speech recognition, and speech enhancement.

**Berlin Chen** (Member, IEEE) received the B.S. and M.S. degrees in computer science and information engineering from National Chiao Tung University, Hsinchu, Taiwan, in 1994 and 1996, respectively, and the Ph.D. degree in computer science and information engineering from National Taiwan University, Taipei, Taiwan, in 2001. He was with the Institute of Information Science, Academia Sinica, Taipei, from 1996 to 2001, and then with the Graduate Institute of Communication Engineering, National Taiwan University, from 2001 to 2002. In 2002, he joined the Graduate Institute of Computer Science and Information Engineering, National Taiwan Normal University, Taipei. He is currently a Professor with the Department of Computer Science and Information Engineering of the same university. He is the author/coauthor of more than 200 academic publications. His research interests include speech recognition and natural language processing, multimedia information retrieval, computer-assisted language learning, and artificial intelligence.