An Effective Pronunciation Assessment Approach Leveraging Hierarchical Transformers and Pre-training Strategies

Bi-Cheng Yan[1]*, Jiun-Ting Li[1], Yi-Cheng Wang[1], Hsin-Wei Wang[1], Tien-Hong Lo[1],
Yung-Chang Hsu[2], Wei-Cheng Chao[3], Berlin Chen[1]*

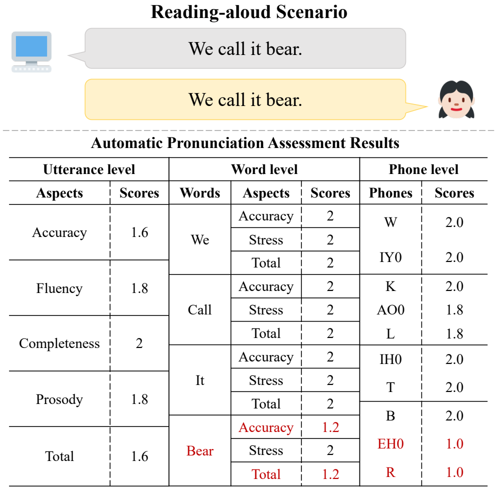
[1]National Taiwan Normal University, [2]EZAI
[3]Advanced Technology Laboratory, Chunghwa Telecom Co., Ltd.
{bicheng, berlin}@ntnu.edu.tw, weicheng@cht.com.tw

Abstract

Automatic pronunciation assessment (APA) manages to quantify a second language (L2) learner's pronunciation proficiency in a target language by providing fine-grained feedback with multiple pronunciation aspect scores at various linguistic levels. Most existing efforts on APA typically parallelize the modeling process, namely predicting multiple aspect scores across various linguistic levels simultaneously. This inevitably makes both the hierarchy of linguistic units and the relatedness among the pronunciation aspects sidelined. Recognizing such a limitation, we in this paper first introduce HierTFR[1], a hierarchal APA method that jointly models the intrinsic structures of an utterance while considering the relatedness among the pronunciation aspects. We also propose a correlation-aware regularizer to strengthen the connection between the estimated scores and the human annotations. Furthermore, novel pre-training strategies tailored for different linguistic levels are put forward so as to facilitate better model initialization. An extensive set of empirical experiments conducted on the speechocean762 benchmark dataset suggest the feasibility and effectiveness of our approach in relation to several competitive baselines.

1 Introduction

With the rising trend of globalization, more and more people are willing or being demanded to learn foreign languages. This surging need calls for developing computer-assisted pronunciation training (CAPT) systems, as they can offer tailored and informative feedback for L2 (second-language) learners to practice pronunciation skills in a stress-free and self-directed learning manner (Eskenazi 2009; Evanini and Wang, 2013; Evanini et al., 2017; Rogerson-Revell, 2021). As a crucial ingredient of CAPT, automatic pronunciation assessment (APA) aims to evaluate the extent of L2 learners' oral proficiency and then provide fine-grained feedback on specific pronunciation aspects in response to a target language (Bannò et al., 2022; Chen and Li, 2016; Kheir et al., 2023). A de-facto standard for APA systems is typically instantiated with a "reading-aloud" scenario, where an L2 learner is presented with a text prompt and instructed to pronounce it correctly. To offer in-depth feedback on learners' pronunciation quality, recent efforts have drawn attention to the notion of multi-aspect and multi-granular pronunciation assessments, which normally devises a unified scoring model to



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
			Stress	2	AO0	1.8
Completeness	2		Total	2	L	1.8
		It	Accuracy	2	IH0	2.0
Prosody	1.8		Stress	2	T	2.0
			Total	2		
Total	1.6	Bear	Accuracy	1.2	B	2.0
			Stress	2	EH0	1.0
			Total	1.2	R	1.0

Figure 1: A running example curated from the speechocean762 dataset (Zhang et al., 2021) illustrates the evaluation flow of an APA system in the reading-aloud scenario, which offers an L2 learner in-depth pronunciation feedback.

* Corresponding author.

[1] https://github.com/bicheng1225/HierTFR

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1737–1747
August 11-16, 2024 ©2024 Association for Computational Linguistics

jointly evaluate pronunciation proficiency at various linguistic levels (i.e., phone-, word-, and utterance-levels) with diverse aspects (e.g., accuracy, fluency, and completeness), as the running example depicted in Figure 1. Methods along this line of research usually follow a parallel modeling paradigm, wherein the Transformer network and its variants serve as the backbone architecture to take as input a sequence of phone-level pronunciation features and in turn predict multiple aspect scores across various linguistic levels simultaneously via a multi-task learning regime (Chao et al., 2022; Do et al., 2023a; Gong et al., 2022).

Albeit models stemming from the parallel modeling paradigm have demonstrated promising results on a few APA tasks, they still suffer from at least two weaknesses. First, the language hierarchy of an utterance is nearly sidelined, which, for example, assumes that all phones within a word are of equal importance and might insufficiently capture the word-level structural traits. Second, most of these methods largely overlook the relatedness among the pronunciation aspects. As an illustration, we visualize the correlation matrix in Figure 2, which shows the Pearson Correlation Coefficients (PCCs) 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 rest pronunciation aspects exhibit strong correlations not only within the same linguistic level but also across different linguistic levels[2]. Building on these observations, we in this paper present a novel language hierarchy-aware APA model, dubbed HierTFR, which leverages a hierarchical Transformer-based architecture to jointly model the intrinsic multi-level linguistic structures of an utterance while considering relatedness among aspects within and across different linguistic levels. To explicitly capture the relatedness within and across different linguistic levels, an aspect attention mechanism and a selective fusion module are introduced. The proposed model is further optimized with an effective correlation-aware regularizer, which encourages the correlations of predicted aspect scores to match those of their counterparts provided by human annotations. Furthermore, distinct pre-training strategies tailored for three linguistic levels are put forward, so as to boost model initialization and hence reduce the reliance on large amounts of supervised training data. A comprehensive set of experimental results reveal that the proposed model achieves significant and consistent improvements over several strong baselines on the speechocean762 benchmark dataset (Zhang et al., 2021).

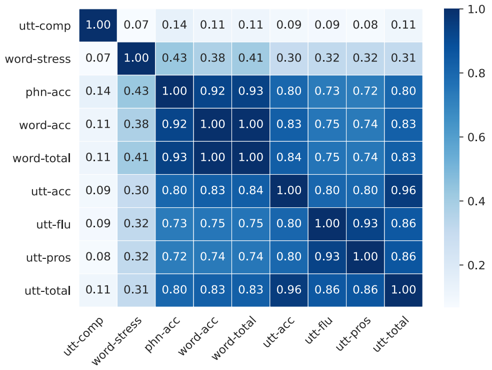


Figure 2: A correlation matrix derived from the expert annotations of the training set. Each element in the matrix corresponds to the PCC score of a pair of measured aspects.

In summary, the main contributions of our work are at least three-fold: (1) we introduce HierTFR, a hierarchical neural model for APA, which is designed to hierarchically represent an L2 learner's input utterance and effectively capture relatedness within and across different linguistic levels; (2) we propose a correlation-aware regularizer for model training, which encourages prediction scores to consider the relatedness among disparate aspects; and (3) extensive sets of experiments carried out on a public APA dataset confirm the utility of our proposed pre-training strategies, which considerably boosts the effectiveness of assessments across various linguistic levels.

2 Methodology

2.1 Problem Formulation

Given an input utterance U, consisting of a time sequence of audio signals X uttered by an L2 learner, and a reference text prompt T with M words and N phones, an APA model is trained to estimate the proficiency scores pertaining to multiple pronunciation aspects at various linguistic granularities. Let $\mathrm{G} = \{p, w, u\}$ be a set of linguistic granularities, where p, w, u stands for the phone-, word-, and utterance-level linguistic units,

[2] 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 (Do et al., 2023a).

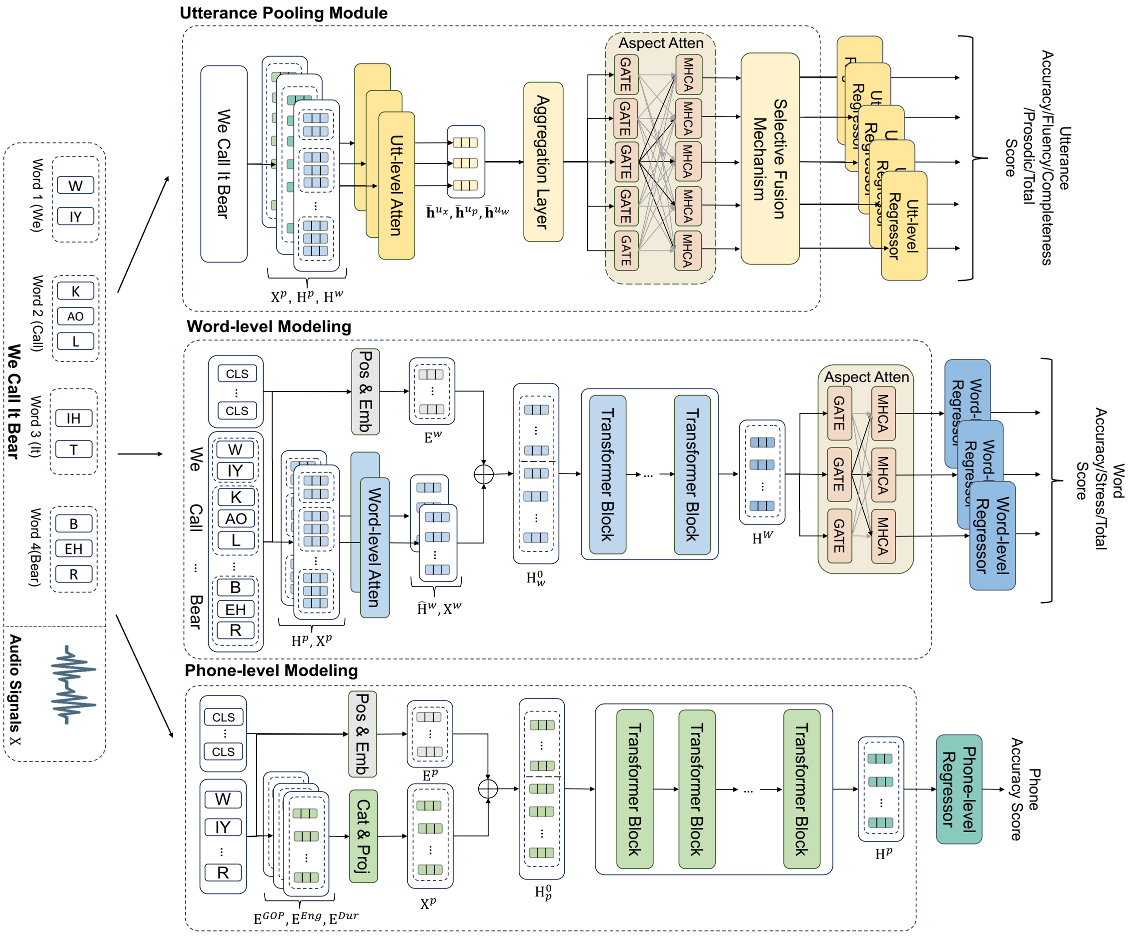


Figure 3: An architecture overview of the proposed model, which consists of a phone-level modeling component, a word-level modeling component, and an utterance pooling module.

respectively. For each linguistic unit $g \in G$, the APA model learns to predict a set of aspect scores $A^g = \{a_1^g, a_2^g, \ldots, a_{N_g}^g\}$, where N_g is the number of pronunciation aspects of the linguistic unit g.

2.2 Hierarchical Interactive Transformer Architecture

The overall architecture of our proposed APA model is schematically depicted in Figure 3, which consists of three ingredients: phone-level modeling, word-level modeling, and utterance pooling modules. After obtaining the representations of various pronunciation aspects, fully-connected neural layers is functioned as the regressors to collectively generate the corresponding aspect score sequence for an input utterance.

Phone-level Modeling. For an input utterance U, various pronunciation features are extracted to portray the L2 learner's pronunciation quality, which includes the goodness of pronunciation (GOP)-based features E^{GOP}, as well as prosodic features composed of duration E^{Dur} and energy E^{Eng} statistics (Witt and Young, 2000; Hu et al., 2015; Zhu et al., 2022; Shen et al., 2021)[3]. All these features are then concatenated and subsequently projected to from a sequence of acoustic features X^p. In the meantime, the phone-level text prompt is mapped into an embedding sequence E^p via a phone and position embedding layer and then point-wisely added to X^p for enriching the phonetic information of X^p. The resulting representations H_p^0 are prepend with five trainable "[CLS]" embeddings and in turn fed into a phone-level transformer to obtain the contextualized representations H^p (Vaswani et al., 2017):

$$X^p = W \cdot [E^{GOP}; E^{Dur}; E^{Eng}] + \mathbf{b}, \tag{1}$$

$$H_p^0 = X^p + E^p, \tag{2}$$

$$H^p = \text{Transformer}_{\text{phn}}(H_p^0), \tag{3}$$

[3] Further details on pronunciation feature extractions can be found in Appendix A.

where W and $\mathbf{b}$ are learnable parameters. To assess a sequence of phone-level aspect scores, H^p (excluding the first 5 embeddings) is forward propagated to the corresponding regressors. The excluded embeddings $\mathrm{H}^p_{1:5}$ are expected to convey the holistic pronunciation information and are further fed into the subsequent selective fusion mechanism for use in utterance-level assessments.

Word-level Modeling. For the word-level assessments, a word-level attention pooling is used to produce a word representation vector from its corresponding phones, which can be implemented as a multi-head attention layer followed by an average operation. The word-level input representations H^0_w can be obtained by applying the word-level attention to the phone-level representations X^p and H^p individually, followed by a linear combination with the word-level textual embeddings E^w. Next, H^0_w is prepend with five trainable “[CLS]” embeddings and fed into a transformer to calculate the contextualized representations H^w at word-level:

$$\mathrm{X}^w = \mathrm{Atten}_{\mathrm{word}}(\mathrm{X}^p), \tag{4}$$

$$\widehat{\mathrm{H}}^w = \mathrm{Atten}_{\mathrm{word}}(\mathrm{H}^p), \tag{5}$$

$$\mathrm{H}^0_w = \mathrm{X}^w + \widehat{\mathrm{H}}^w + \mathrm{E}^w, \tag{6}$$

$$\mathrm{H}^w = \mathrm{Transformer}_{\mathrm{word}}(\mathrm{H}^0_w). \tag{7}$$

Note here that H^w (excluding the first 5 embeddings) is utilized in the word-level assessments while the excluded embeddings $\mathrm{H}^w_{1:5}$ are fed into in subsequent selective fusion mechanism for use in the utterance-level assessments.

After that, an aspect attention mechanism is introduced to capture the relatedness among disparate aspects (Do et al., 2023b; Ridley et al., 2021). This mechanism consists of two sub-layers: a self-gating layer and a multi-head cross-attention layer. Specifically, for the j-th word-level aspect, the relation-aware representations $\widehat{\mathrm{H}}^{wr_j}$ are first derived from H^w via a self-gating layer which aims to abstract away from redundant information while considering the information gathered from other aspects. In addition, a multi-head cross-attention (MHCA) process alongside a masking strategy is employed to calculate aspect representations H^{w_j} from a collection of all relation-aware aspect representations $\mathrm{C}^{ra} = \left[\widehat{\mathrm{H}}^{wr_1}, \ldots, \widehat{\mathrm{H}}^{wr_{N_w}}\right]$. The following equations illustrate the operations of aspect attention:

$$\widehat{\mathrm{H}}^{w_j} = \mathrm{W}_j \cdot \mathrm{H}^w + \mathbf{b}_j, \tag{8}$$

$$\widehat{\mathrm{H}}^{wr_j} = \sigma\left(\mathrm{W}_{g_j} \cdot \mathrm{C}^w + \mathbf{b}_{g_j}\right) \otimes \widehat{\mathrm{H}}^{w_j}, \tag{9}$$

$$\mathrm{H}^{w_j} = \mathrm{MHCA}\left(\widehat{\mathrm{H}}^{wr_j}, \mathrm{C}^{ra}\right), \tag{10}$$

where $\widehat{\mathrm{H}}^{w_j}$ are aspect-specific representations, and $\mathrm{C}^w = [\widehat{\mathrm{H}}^{w_1}, \ldots, \widehat{\mathrm{H}}^{w_{N_w}}]$ includes all aspect-specific representations. In MHCA, $\widehat{\mathrm{H}}^{wr_j}$ is linearly projected to act as the query matrix, while C^{ra} is linearly projected to form the key and value matrixes. Additionally, the masking strategy ensures that the output representation at a specific position is only influenced by the other aspects of the word unit. Lastly, the aspect representations H^{w_j} are taken as input to the corresponding regressor to predict a score sequence for the j-th word-level pronunciation aspect.

Utterance Pooling Module. For the utterance-level assessments, utterance-level attention pooling is introduced to generate an utterance-level holistic representation from the corresponding input representations, which can be effectively implemented by attention pooling (Peng et al., 2022). In more detail, the utterance-level representation $\mathbf{h}^u$ can be obtained by feeding the vector sequences X^p , H^p , and H^w into an utterance-level attention pooling module individually, followed by an aggregation operation:

$$\bar{\mathbf{h}}^{u_x} = \mathrm{Atten}_{\mathrm{utt}}(\mathrm{X}^p), \tag{11}$$

$$\bar{\mathbf{h}}^{u_p} = \mathrm{Atten}_{\mathrm{utt}}(\mathrm{H}^p), \tag{12}$$

$$\bar{\mathbf{h}}^{u_w} = \mathrm{Atten}_{\mathrm{utt}}(\mathrm{H}^w), \tag{13}$$

$$\mathbf{h}^u = \mathrm{W}_u\left(\bar{\mathbf{h}}^{u_x} + \bar{\mathbf{h}}^{u_p} + \bar{\mathbf{h}}^{u_w}\right) + \mathbf{b}_u, \tag{14}$$

where W_u, $\mathbf{b}_u$ are trainable parameters.

Next, a selective fusion mechanism is proposed to integrate contextualized representations across multiple linguistic levels for the utterance-level pronunciation assessments (Xu et al., 2021). Specifically, for the estimation of j-th utterance-level aspect score, an aspect attention operation is first performed on $\mathbf{h}^u$ to a produce intermediate representation $\hat{\mathbf{h}}^{u_j}$. Note also that the gate values for the phone ($g_p^{u_j}$), word ($g_w^{u_j}$) and utterance ($g_u^{u_j}$) granularities are used to control the extent to which these contextualized representations can flow into the fused representation $\mathbf{h}^{u_j}$:

$$g_p^{u_j} = \sigma\left(\mathbf{w}_{p_j} \cdot [\mathbf{h}_j^p; \mathbf{h}_j^w; \hat{\mathbf{h}}^{u_j}] + b_{p_j}\right), \tag{15}$$

$$g_w^{u_j} = \sigma\left(\mathbf{w}_{w_j} \cdot [\mathbf{h}_j^p; \mathbf{h}_j^w; \hat{\mathbf{h}}^{u_j}] + b_{w_j}\right), \tag{16}$$

$$g_u^{u_j} = \sigma\left(\mathbf{w}_{u_j} \cdot [\mathbf{h}_j^p; \mathbf{h}_j^w; \hat{\mathbf{h}}^{u_j}] + b_{u_j}\right), \tag{17}$$

$$\mathbf{h}^{u_j} = g_p^{u_j} \cdot \mathbf{h}_j^p + g_w^{u_j} \cdot \mathbf{h}_j^w + g_u^{u_j} \cdot \hat{\mathbf{h}}^{u_j}, \qquad (18)$$

where $\mathbf{h}_j^p$ and $\mathbf{h}_j^w$ are j-th representation vectors of H^p and H^w; and $\mathbf{w}_{p_j}$, $\mathbf{w}_{w_j}$, $\mathbf{w}_{u_j}$, b_{p_j}, b_{w_j}, and b_{u_j} are trainable parameters. The fused representation $\mathbf{h}^{u_j}$ is then passed to the corresponding regressor to assess the proficiency score for a given utterance-level aspect.

2.3 Optimization

Automatic Pronunciation Assessment Loss. The loss for multi-aspect and multi-granular pronunciation assessment, $\mathcal{L}_{APA}$, is calculated as a weighted sum of the mean square error (MSE) losses corresponding to different linguistic levels.

$$\mathcal{L}_{APA} = \frac{\lambda_p}{N_p} \sum_{j_p} \mathcal{L}_{p^{j_p}} + \frac{\lambda_w}{N_w} \sum_{j_w} \mathcal{L}_{w^{j_w}} + \frac{\lambda_u}{N_u} \sum_{j_u} \mathcal{L}_{u^{j_u}}, \qquad (19)$$

where $\mathcal{L}_{p^{j_p}}$, $\mathcal{L}_{w^{j_w}}$, and $\mathcal{L}_{u^{j_u}}$ are phone-level, word-level, and utterance-level losses for disparate aspects, respectively. The parameters λ_p, λ_w, and λ_u are adjustable parameters which control the influence of different granularities, and N_p, N_w, and N_u mark the numbers of aspects at the phone-, word-, and utterance-levels, respectively.

Correlation-aware Regularization Loss. The correlation-aware regularization loss is defined as the difference between the correlation matrix of the predicted aspect scores $\hat{\Sigma}$ and the correlation matrix of the corresponding target labels Σ:

$$\mathcal{L}_{cor} = \ell(\hat{\Sigma}, \Sigma), \qquad (20)$$

where ℓ is the regularization loss function, and each element in $\hat{\Sigma}_{ij}$ (or Σ_{ij}) is defined as a Pearson correlation coefficient between i-th aspect score and j-th aspect score[4]. We adopt the MSE criterion for computing ℓ; the overall loss thus can be expressed by:

$$\mathcal{L} = \mathcal{L}_{APA} + \lambda \mathcal{L}_{cor}, \qquad (21)$$

where $\lambda \in [0, 1]$ is a tunable parameter, which is experimentally set to 0.01 based on the development set.

2.4 Pre-training Strategies

It is without doubt that a proper initialization is vital for the estimation of a neural model, due mainly to the highly nonconvex nature of the training loss function (Tamborrino et al., 2020; Lakhotia et al., 2021). At lower linguistic levels, we leverage the mask-predict objective (Ghazvininejad et al., 2019) in the pre-training stage. To this end, we first mask a portion of input text prompt at phone- and word-levels. The corresponding Transformers are then tasked on recovering the masked tokens conditioning on the unmasked prompt sequence and the associated pronunciation representations (i.e., H_p^0, and H_w^0). For the utterance level, we base the proposed pre-training strategy on predicting the relatively high or low accuracy scores for a pair of utterances. Namely, given any two utterances, the objective is to predict whether the former has a higher, lower, or the same accuracy score as the latter. Note here that, utterance pairs are randomly selected from a training batch, and this mechanism is employed to pretrain their utterance-level representations, denoted as $\mathbf{h}_{out_1}^u$, and $\mathbf{h}_{out_2}^u$. Next, we feed the concatenation of these vector representations $\mathbf{h}_{out}^u = [\mathbf{h}_{out_1}^u; \mathbf{h}_{out_2}^u]$ into a three-way classifier, using the cross-entropy loss as the training objective.

3 Experimental Settings

3.1 Evaluation Dataset and Metrics

We conducted APA experiments on the speechocean762 dataset, which is a publicly available open-source dataset specifically designed for research on APA (Zhang et al., 2021). This dataset contains 5,000 English-speaking recordings spoken by 250 Mandarin L2 learners. The training and test sets are of equal size, and each of them has 2,500 utterances, where pronunciation proficiency scores were evaluated at multiple linguistic granularities with various pronunciation aspects. Each score was independently assigned by five experienced experts using the same rubrics, and the final score was determined by selecting the median value from the five scores. The evaluation metrics include Pearson Correlation Coefficient (PCC) and Mean Square Error (MSE). PCC is the primary evaluation metric, quantifying the linear correlation between predicted and ground-truth scores. A higher PCC score reflects a stronger correlation between the predictions and human annotations. In the following experiments, we report the MSE value in order to evaluate the

[4] To calculate PCC scores between aspects across different granularities, we duplicate the aspect scores of higher granularities to match the aspect scores at the lower granularities.

Models	Phone Score		Word Score (PCC)			Utterance Score (PCC)				
	MSE↓	PCC↑	Accuracy↑	Stress↑	Total↑	Accuracy↑	Completeness↑	Fluency↑	Prosody↑	Total↑
Lin2021	-	-	-	-	-	-	-	-	-	0.720
Kim2022	-	-	-	-	-	-	-	0.780	0.770	-
Ruy2023	-	-	-	-	-	0.719	-	0.775	0.773	0.743
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
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
HierTFR	0.081 ±0.000	**0.644** **±0.000**	**0.622** **±0.002**	**0.325** **±0.022**	**0.634** **±0.002**	**0.735** **±0.008**	**0.513** **±0.204**	**0.801** **±0.004**	**0.795** **±0.002**	**0.764** **±0.002**

Table 1: The performance evaluations of our model and all compared methods on speechocean762 test set.

phoneme-level APA accuracy in comparison with prior arts.

3.2 Implementation Details

For the input feature extraction of the phone-level energy and the duration statistics, we follow the processing flow suggested by Zhu et al. (2022) and Shen et al. (2021), where a phone-level feature is constructed from time-aggregated frame-level features according to the forced alignment. Both the phone- and word-level Transformers for contextual representation modeling consist of 3 processing blocks utilizing multi-head attention with 3 heads and 24 hidden units, respectively. In addition, for the word- and utterance-level attention pooling, we use a single-layer multi-head attention with 3 heads and 24 hidden units. The combination weights used in Eq. (19) for the APA loss (λ_p , λ_w , λ_u) are assigned as $(1, 1, 1)$, respectively. To ensure the reliability of our experimental results, we repeated 5 independent trials, each of which consisted of 100 epochs with different random seeds. The test set results are reported by averaging those achieved by the top 100 best-performing models which are determined based on their PCC scores on the development set.

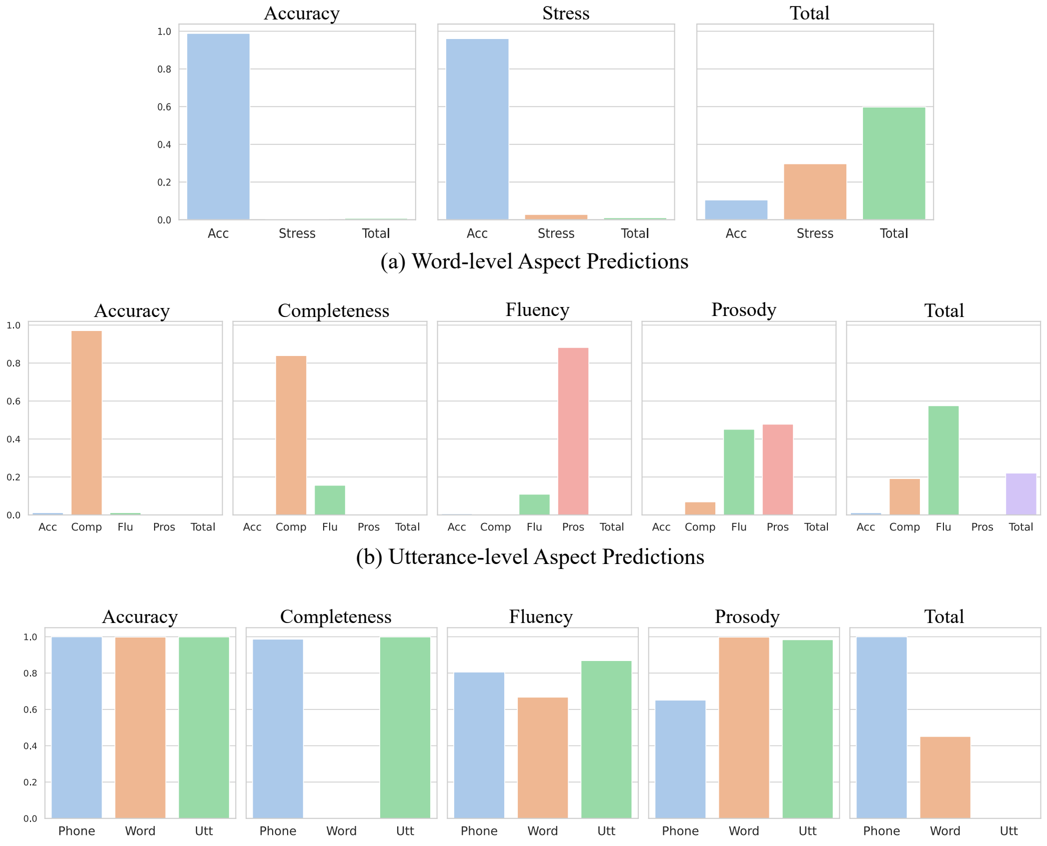
3.3 Compared Methods

We compare our proposed model (viz. HierTFR) with several families of top-of-the-line methods. Lin et al. (2021) and Kim et al. (2022) are single-aspect assessment models. The former develops a bottom-up hierarchical scorer evaluating the accuracy scores at the utterance level. The latter leverages self-supervised features (Baevski et al., 2020) to describe the learner's pronunciation traits and then separately models the corresponding utterance-level aspects with recurrent neural models. In addition, LSTM, GOPT (Gong et al., 2022; Ruy et al. 2023), and HiPAMA (Do et al., 2023b) are multi-aspect and multi-granular pronunciation assessments. First, LSTM and GOPT follow a parallel modeling regime, both of which treat the phone-level input features as a flattened sequence and assess higher level pronunciation scores through stacking LSTM layers or Transformer blocks. Second, Ruy et al. (2023) introduces a unified model architecture that jointly optimizes phone recognition and APA tasks. Lastly, HiPAMA is a hierarchical APA model that more resembles our model than the other methods compared in this paper. Different from our method, HiPAMA extracts high-level pronunciation features from low-level features based on a simple average pooling mechanism. Furthermore, the aspect attention mechanism used in HiPAMA performs on the logistics, whereas our model operates on the intermediate representations.

4 Experimental Results

4.1 Main Results

Table 1 reports the results on the speechocean762 dataset, which is divided into three parts: the first part shows the results of single-aspect assessment models, the second part presents the results of multi-aspect and multi-granular pronunciation methods, and the third part reports the results of our model. We further provide a comparison with another hierarchical APA model (viz. HiPAMA) in the third part.



(a) Word-level Aspect Predictions

(b) Utterance-level Aspect Predictions

(c) Gate Values in Selective Fusion Mechanism for Utterance-level Aspect Predictions

Figure 4: Qualitative visualization of model parameters when predicting each aspect score. We show (a) the averaged attention values for word-level aspects, (b) the averaged attention weights for utterance-level aspects, and (c) the averaged gate values for three linguistic levels.

First, a general observation is that our approach, HierTFR, excels in all assessment tasks, especially at the linguistic levels of utterance and word. This performance gain confirms that our model works comparably better for capturing the relationships between linguistic units than the other competitive methods. In terms of the utterance-level total score, the single-aspect assessment method (viz. Lin2021) largely falls behind the other multi-aspect and multi-granular pronunciation assessment models, which we attribute to the fact that the single-aspect assessment method is unable to harness the dependency relationships between aspects through the multi-task learning paradigm. By leveraging self-supervised learning features, Kim2022 achieves significant improvements over most APA methods in terms of the utterance-level assessments. Next, we scrutinize the performance of multi-aspect and multi-granular pronunciation assessment methods. Ruy2023 demonstrates significant advancements in the utterance-level fluency and prosody assessments due probably to the joint training of the APA model on the phone recognition task simultaneously. In comparison with the parallel modeling approaches (i.e., GOPT and LSTM), we can observe that HierTFR substantially improves the performance across all tasks, where its performance gains reveal the importance of capturing the hierarchical linguistic structures of an input utterance. Notably, compared to the HiPAMA, our model consistently achieves superior performance on a variety of pronunciation assessment tasks. This superiority stems from our tactfully designed selective fusion mechanism and the correlation-aware loss. The former allows our model to assess utterance-level aspect scores by leveraging information from diverse linguistic levels, while the latter explicitly models the relatedness among different aspects during the optimization.

4.2 Qualitative Analysis

Qualitative Visualization of Relatedness Among Aspects. In the second set of experiments, we examine the relatedness among disparate aspects at both word- and utterance-levels, where the

Models	Phone Score	Word Score			Utterance Score				
	Accuracy	Accuracy	Stress	Total	Accuracy	Completeness	Fluency	Prosody	Total
HierTFR	**0.644**	**0.622**	0.325	**0.634**	**0.735**	0.513	**0.801**	**0.795**	**0.764**
w/o CorrLoss	0.639	0.605	**0.348**	0.620	0.728	**0.520**	0.796	0.789	0.758
w/o Pretrain	0.621	0.545	0.318	0.559	0.716	0.215	0.770	0.772	0.739
w/o SFusion	0.630	0.608	0.328	0.622	0.728	0.378	0.784	0.782	0.756
w/o AspAtt	0.636	0.584	0.290	0.596	0.724	0.383	0.784	0.775	0.746

Table 2: Ablation study on HierTFR, reporting PCC scores on three linguistic levels.

attention weights of the aspect attention mechanisms were determined based on the development set when assessing a specific aspect score. For the word-level assessments, the distributions of attention weights are in close accordance with the manual scoring rubrics of the speechocean762 dataset. In Figure 4(a), the total aspect serves as a comprehensive assessment and the corresponding weights are contributed from various pronunciation aspects. In contrast, the accuracy aspect measures the percentage of mispronounced phones within a word, leading to the attention weights being more concentrated on a word-level unit itself. Furthermore, the stress score also highly attends to the accuracy aspect, reflecting the strong relation between lexical stress and word-level pronunciation accuracy (Korzekwa et al., 2022). In regard to the relatedness within the utterance-level aspects, inspecting Figure 4(b) we find that the attention weights of the prosody and total aspects scatter across various pronunciation aspects, whereas the attention weights of the accuracy and completeness center primarily on the completeness aspect. One possible reason is that the prosody and total scores both measure high-level oral skills, and when the human annotators judge the proficiency scores, they also take multiple pronunciation aspects into account simultaneously. Next, the completeness aspect measures the percentage of words with good pronunciation quality in an utterance. This implicitly reflects the intelligibility of a learner's pronunciation and is vital to the accuracy assessment.

Qualitative Visualization of Interactions Across Linguistic Levels. In Figure 4(c), we report on the average gate values of utterances for three linguistic granularities by estimating the utterance-level pronunciation aspect scores based on the development set. We can observe that the phone-level representations bear high impacts on the utterance-level aspect assessments, in comparison to the other linguistic levels. Next, the word-level and utterance-level representations exhibit minimal impact on the completeness and total aspects, respectively. One possible reason is that the completeness aspect somehow reflects pronunciation intelligibility, and our model learns to distill the information from the phone- and utterance-level representations. On the other hand, the total aspect evaluates an overall speaking skill. Our model thus tends to capture the subtle information by distilling the fine-grained traits inherent in the phone- and word-levels.

4.3 Ablation Study

To gain insight into the effectiveness of each model component of HierTFR, we conduct an ablation study to investigate their impacts. These variations include excluding the correlation-aware regularizer (w/o CorrLoss), removing the proposed pre-training strategies (w/o Pretrain), omitting the selective fusion mechanism (w/o SFusion), and eliminating the aspect attention mechanism at both word and utterance levels (w/o AspAtt). From Table 2, we can observe that the proposed correlation-aware regularization loss is beneficial for most pronunciation assessment tasks. Next, the proposed pre-training strategies are crucial to obtaining better performance as the model trained without them tends to perform relatively worse for all pronunciation assessment tasks. This highlights the efficacy of the pre-training strategies for hierarchical APA models, thereby alleviating the requirement for large amounts of supervised training data. Third, removing the selective fusion mechanism leads to degradations in the utterance-level aspect assessments, while removing the aspect attention mechanism deteriorates the performance on word-level aspect assessments.

5 Related Work

Early studies on APA focused primarily on single-aspect assessments, typically through individually constructing scoring modules to predict a holistic pronunciation proficiency score on a targeted linguistic level or some specific aspect with different sets of hand-crafted features, such as the phone-level posterior probability (Witt and Young, 2000), word-level lexical stress (Ferrer et al., 2015), or various utterance-level pronunciation aspects (Coutinho et al., 2016). More recently, with the rapid progress of deep learning (Vaswani et al., 2017; Raffel et al., 2020; Hsu et al., 2021), several neural scoring models have been successfully developed for multi-aspect and multi-granular pronunciation assessment. Gong et al. (2022) proposed a GOP feature-based Transformer (GOPT) architecture to model pronunciation aspects at multiple granularities with a multi-task learning scheme. Do et al. (2023b) employed a neural scorer with a hierarchical structure to mimic the language hierarchy of an utterance to deliver state-of-the-art performance for APA.

6 Conclusion

In this paper, we have put forward a novel hierarchical modeling method (dubbed HierTFR) for multi-aspect and multigranular APA. To explicitly capture the relatedness between pronunciation aspects, a correlation-aware regularizer loss has been devised. We have further developed model pre-training strategies for our HierTFR model. Extensive experimental results confirm the feasibility and effectiveness of the proposed method in relation to several top-of-the-line methods. In future work, we plan to examine the proposed HierTFR model on open-response scenarios, where learners speak freely or respond to a given task or question (Wang et. al., 2018; Park and Choi, 2023). In addition, the issues of explainable pronunciation feedback are also left as a future extension.

Limitations

Limited Applicability. 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 text prompt correctly, which restricts the applicability of our models to other learning scenarios, such as freely speaking or open-ended conversations.

Lack of Accent Diversity. The used dataset merely contains Mandarin L2 learners, hindering the generalizability of the proposed model and could be untenable when assessing the L2 learners with diverse accents.

The lack of Interpretability. The model of the proposed method simply trains to mimic expert’s annotations without resorting to manual assessment rubrics or other external knowledge, making it not straightforward to provide reasonable explanations for the assessment results.

Ethics Statement

We hereby acknowledge that all of the co-authors of this work compile with the provided ACL Code of Ethics and honor the code of conduct. Our experimental corpus, speechocean762, is widely used and publicly available. We think there are no potential risks for this work.

References

Stefano Bannò, Bhanu Balusu, Mark Gales, Kate Knill,and Konstantinos Kyriakopoulos. 2022. View-specific assessment of L2 spoken English. In *Proceedings of Interspeech (INTERSPEECH)*, pages 4471–4475.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. Wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS)*, pages 12449–12460.

Fu An Chao, Tien Hong Lo, Tzu I. Wu, Yao Ting Sung, Berlin Chen. 2022. 3M: An effective multi-view, multigranularity, and multi-aspect modeling approach to English pronunciation assessment. In *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 575–582.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, Furu Wei. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, volume 16, pages1505–1518.

Nancy F. Chen, and Haizhou Li. 2016. Computer-assisted pronunciation training: From pronunciation scoring towards spoken language learning. In *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1–7.

Eduardo Coutinho, Florian Hönig, Yue Zhang, Simone Hantke, Anton Batliner, Elmar Nöth, and Björn Schuller. 2016. Assessing the prosody of non-native speakers of English: Measures and feature sets. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 1328–1332.

Heejin Do, Yunsu Kim, and Gary Geunbae Lee. 2023a. Score-balanced Loss for Multi-aspect Pronunciation Assessment. In *Proceedings of Interspeech (INTERSPEECH)*, pages 4998–5002.

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

Maxine Eskenazi. 2009. An overview of spoken language technology for education. *Speech communication*, volume 51, pages 832–844.

Keelan Evanini, and Xinhao Wang. 2013. Automated speech scoring for Nonnative middle school students with multiple task types. In *Proceedings of Interspeech (INTERSPEECH)*, pages 2435–2439.

Keelan Evanini, Maurice Cogan Hauck, and Kenji Hakuta. 2017. Approaches to automated scoring of speaking for K–12 English language proficiency assessments. *ETS Research Report Series*, pages 1–11.

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

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-Predict: Parallel Decoding of Conditional Masked Language Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6112–6121.

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

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, volume 29, pages 3451–3460.

Wenping Hu, Yao Qian, Frank K. Soong, and Yong Wang. 2015. Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers. *Speech Communication*, volume 67, pages 154–166.

Eesung Kim, Jae-Jin Jeon, Hyeji Seo, Hoon Kim. 2022. Automatic pronunciation assessment using self-supervised speech representation learning. In *Proceedings of Interspeech (INTERSPEECH)*, pages 1411–1415.

Yassine Kheir, Ahmed Ali, and Shammur Chowdhury. 2023. Automatic Pronunciation Assessment - A Review. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 8304–8324.

Daniel Korzekwa, Jaime Lorenzo-Trueba, Thomas Drugman, and Bozena Kostek. 2022. Computer-assisted pronunciation training—Speech synthesis is almost all you need. *Speech Communication*, volume 142, pages 22–33.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. On Generative Spoken Language Modeling from Raw Audio. *Transactions of the Association for Computational Linguistics*, volume 9, pages 1336–1354.

Binghuai Lin and Liyuan Wang. 2021. Deep feature transfer learning for automatic pronunciation assessment. In *Proceedings of Interspeech (INTERSPEECH)*, pages 4438–4442.

Silke M. Witt and S. J. Young. 2000. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, volume 30, pages 95–108.

Jungbae Park and Seungtaek Choi. 2023. Addressing cold start problem for end-to-end automatic speech scoring. In *Proceedings of Interspeech (INTERSPEECH)*, pages 994–998.

Yifan Peng, Siddharth Dalmia, Ian Lane, and Shinji Watanabe. 2022. Branchformer: Parallel mlp-attention architectures to capture local and global context for speech recognition and understanding. In *International Conference on Machine Learning* (PMLR), pages 17627–17643

Yu Wang, M.J.F. Gales, Kate M Knill, Konstantinos Kyriakopoulos, Andrey Malinin, Rogier C van Dalen, Mohammad Rashid. 2018. Towards automatic assessment of spontaneous spoken English. *Speech Communication*, volume 104, pages 47–56.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, volume 21, pages 5485–5551.

Robert Ridley, Liang He, Xin-yu Dai, Shujian Huang, and Jiajun Chen. 2021. Automated cross-prompt scoring of essay traits. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, volume 35, pages 13745–13753.

Pamela M Rogerson-Revell. 2021. Computer-assisted pronunciation training (CAPT): Current issues and future directions. *RELC Journal*, volume 52, pages 189–205.

Hyungshin Ryu and Sunhee Kim and Minhwa Chung. 2023. A joint model for pronunciation assessment and mispronunciation detection and diagnosis with multi-task learning. In *Proceedings of Interspeech (INTERSPEECH)*, pages 959–963.

Yang Shen, Ayano Yasukagawa, Daisuke Saito, Nobuaki Minematsu, and Kazuya Saito. 2021. Optimized prediction of fluency of L2 English based on interpretable network using quantity of phonation and quality of pronunciation. In *Proceedings of IEEE Spoken Language Technology Workshop (SLT)*, pages 698–704.

Alexandre Tamborrino, Nicola Pellicanò, Baptiste Pannier, Pascal Voitot, and Louise Naudin. 2020. Pretraining is (almost) all you need: An application to commonsense reasoning. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 3878–3887.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pages 5998–6008.

Heng-Da Xu, Zhongli Li, Qingyu Zhou, Chao Li, Zizhen Wang, Yunbo Cao, Heyan Huang, and Xian-Ling Mao. 2021. Read, listen, and see: Leveraging multimodal information helps Chinese spell checking. In *Findings of the Association for Computational Linguistics (ACL-IJCNLP Findings)*, pages 716–728.

Junbo Zhang, Zhiwen Zhang, Yongqing Wang, Zhiyong Yan, Qiong Song, Yukai Huang, Ke Li, Daniel Povey, and Yujun Wang. 2021. Speechocean762: An open-source non-native English speech corpus for pronunciation assessment. In *Proceedings of Interspeech (INTERSPEECH)*, pages 3710 –3714.

Chuanbo Zhu, Takuya Kunihara, Daisuke Saito, Nobuaki Minematsu, Noriko Nakanishi. 2022. Automatic prediction of intelligibility of words and phonemes produced orally by japanese learners of English. In *IEEE Spoken Language Technology Workshop (SLT)*, pages. 1029–1036.

A Pronunciation Feature Extractions

GOP Feature. To extract the GOP feature, we first align audio signals X with the text prompt T by using an ASR model[5] to obtain the timestamps for each phone in the canonical phone sequence. Next, frame-level phonetic posterior probabilities are produced by the ASR model and then averaged over the time dimension 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 can be represented as an 84-dimensional vector:

$$\begin{aligned}&[\mathrm{LPP}(p_1), \ldots, \mathrm{LPP}(p_{42}),\\ &\qquad \mathrm{LPR}(p_1|p), \ldots, \mathrm{LPR}(p_{42}|p)]\end{aligned} \tag{22}$$

$$\begin{aligned}\mathrm{LPP}(p_i) &= \log p(p_i|\mathbf{o}; \mathrm{t}_s, \mathrm{t}_e)\\ &= \frac{1}{t_e - t_s + 1}\sum_{t=t_s}^{t_e} \log p(p_i|\mathrm{o}_t),\end{aligned} \tag{23}$$

$$\begin{aligned}\mathrm{LPR}(p_i|p) = \log p(p_i|\mathbf{o}; \mathrm{t}_s, \mathrm{t}_e)&\\ - \log p(p|\mathbf{o}; \mathrm{t}_s, \mathrm{t}_e)&,\end{aligned} \tag{24}$$

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 (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.

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