Multi-task Pretraining for Enhancing Interpretable L2 Pronunciation Assessment

Jiun-Ting Li*, Bi-Cheng Yan†, Yi-Cheng Wang‡, Berlin Chen†,
*Advanced Technology Laboratory, Chunghwa Telecom Co., Ltd., Taiwan
†National Taiwan Normal University, Taiwan, ‡National Taiwan University, Taiwan
E-mail: jtli@cht.com.tw

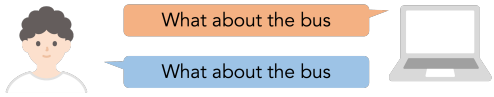
***Abstract*—Automatic pronunciation assessment (APA) analyzes second-language (L2) learners' speech by providing fine-grained pronunciation feedback at various linguistic levels. Most existing efforts on APA typically adopt segmental-level features as inputs and predict pronunciation scores at different granularities via hierarchical (or parallel) pronunciation modeling. This, however, inevitably causes assessments across linguistic levels (e.g., phone, word, and utterance) to rely solely on phoneme-level pronunciation features, nearly sidelining supra-segmental pronunciation cues. To address this limitation, we introduce multi-task pretraining (MTP) for APA, a simple yet effective strategy that attempts to capture long-term temporal pronunciation cues while strengthening the intrinsic structures within an utterance via the objective of reconstructing input features. Specifically, for a phoneme-level encoder of an APA model, the proposed MTP strategy randomly masks segmental-level pronunciation features and reconstructs the masked ones based on their surrounding pronunciation context. Furthermore, current APA systems lack integration with automated speaking assessment (ASA), limiting holistic proficiency evaluation. Drawing on empirical studies and prior knowledge in ASA, our framework bridges this gap by incorporating handcrafted features (HCFs), such as fluency (speech rate, silence duration) and stress (pitch accent strength), derived from human-designed formulas via regressors to generate interpretable proficiency scores. Experiments on speechocean762 show improved pronunciation scoring and ASA proficiency correlation, enabling targeted training and comprehensive proficiency assessment.**

***Index Terms*—computer-assisted language learning, automatic pronunciation assessment, automated speaking assessment, multi-task learning.**

I. Introduction

Fueled by frequent global economic and cultural exchanges, foreign language acquisition, such as English, has become increasingly vital [1]. However, the insufficient supply of language instructors struggles to keep pace with the growing needs of the second language (L2) learners. In response, computer-assisted language learning (CALL) emerges as a promising solution to bridge this gap, which provides effective self-directed learning environments through automatic scoring systems that ensure greater consistency and speed at a lower cost. [2], [3]. As a core component of CALL, automatic pronunciation assessment (APA) aims to evaluate the oral skills of L2 learners by analyzing their pronunciation quality in terms of the presented text prompts (or reference texts) in a target language [4].

As depicted in Figure 1, an L2 learner is presented with a reference text, and in turn, the APA system analyzes their



Automatic Pronunciation Assessment Results

Granularities						
Utterance level		Word level			Phone level*	
Aspects	Scores	Words	Aspects	Scores	Phones	Scores
Accuracy	9	What	Accuracy	10	W	2.0
			Stress	10	AH0	1.8
			Total	10	T	2.0
Fluency	9	About	Accuracy	10	AH0	2.0
			Stress	10	B	2.0
Completeness	10		Total	10	AW1	2.0
					T	2.0
Prosody	8	The	Accuracy	10	DH	2.0
			Stress	10		
			Total	10	AH0	2.0
Total	8	Bus	Accuracy	10	B	2.0
			Stress	10	AH0	2.0
			Total	10	S	2.0

Fig. 1. An illustration of a partial pronunciation assessment sample in the speechocean762 [5] corpus, demonstrating multi-granularity evaluation at the phone, word, and utterance levels, with multi-aspect pronunciation metrics applied across each granularity to support comprehensive L2 feedback. * indicates the single aspect called accuracy at the phoneme-level (granularity).

speech alongside the reference text by assessing their pronunciation proficiency with fine-grained pronunciation aspects (e.g., accuracy, stress, and fluency) across multiple linguistic levels (viz., phoneme, word, and utterance levels) [6], [7].

The existing studies on APA typically employ a unified pronunciation assessment model with a parallel [6], [8] or hierarchical [7], [9], [10] neural architecture to assess pronunciation aspects across linguistic granularities by leveraging segmental-level pronunciation features (e.g., the statistics of energy and duration, the features of phoneme-level pronunciation quality and textual information) in conjunction with a feature aggregation mechanism. A building block for these models, goodness of pronunciation (GOP) feature measures the pronunciation deviation between L2 learners and native speakers at the phoneme-level [6]. It first utilized GOP features as the input for APA systems, derived from an automatic speech recognizer (ASR). While GOP features remain in use, recent APA models increasingly leverage deep learning paradigms, such as self-supervised learning (SSL) models [7]–[9], to just name a few, Wav2vec2.0 [11], HuBERT [12] and WavLM [13], to automatically extract contextual features. Pretrained on large-scale unlabeled speech data, these SSL models capture phonological patterns in English, mitigating

979-8-3315-7206-8/25/$31.00

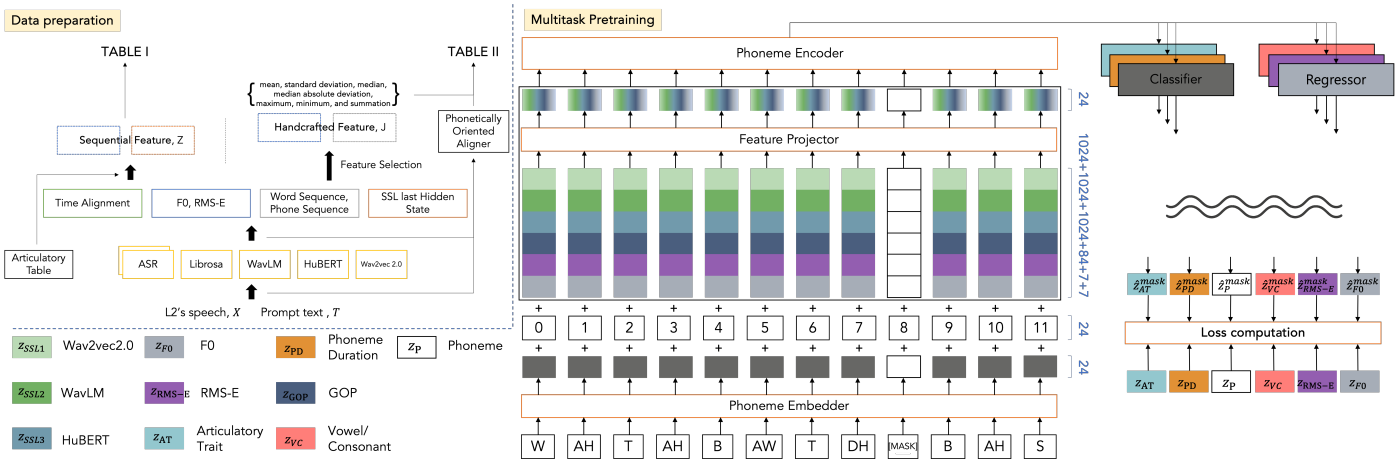


Fig. 2. An illustration of the data preparation process and the MTP framework. In the data preparation process, L2 speech and prompt text are processed by pre-trained acoustic models to extract time alignment information, deep features, and acoustic features (e.g., F0 contours, RMS energy). These are transformed into sequential features and aggregated into handcrafted features using human-designed formulas, supporting MTP and APA training.

TABLE I
TASK DETAILS WITH VIEWS, DENOTATIONS, AND DIMENSION SIZE

Tasks	Views	Denotations	Dimension size
Phonetic	Articulation Trait	S_A	1
	Phoneme Sequence	S_P	1
	Vowel/Consonant	S_V	1
Prosodic	RMS-E Statics	S_R	7
	F0 Statics	S_F	7
	Phoneme Duration	S_D	1

phonetic variability and enhancing scalability and phoneme-level precision. This evolution opens new avenues for advancing pronunciation assessment.

Albeit models stemming from the effective modeling paradigm have demonstrated promising results on a few APA tasks, they still suffer from at least two weaknesses. First, the assessments across linguistic levels (e.g., phone, word, and utterance) rely solely on segmental-level pronunciation features, which inevitably sideline the supra-segmental pronunciation cues. While recent APA models, while leveraging several features for overall precision, such as phonetic features (e.g., phoneme sequence), prosodic features (e.g., energy, pitch and phoneme duration), to create multi-view APA [7], [8], [14], they do not jointly model phonetic and prosodic content and rely heavily on predictors, with encoders failing to intrinsically encode phonetic and prosody-aware representations. To address the limitations of accumulating features without optimization, and treating features solely as inputs, we propose multi-task pretraining (MTP) for APA, which optimizes the phoneme encoder to integrate sequential phonetic and prosodic features as inputs and outputs within hierarchical architectures [7], [14], reconstructing them to enhance APA. Our framework incorporates phonetic subtasks, including phoneme prediction [15]–[17], articulatory traits prediction, and vowel/consonant prediction, alongside prosodic subtasks like phoneme duration prediction [17], pitch prediction and energy prediction aggregated from phoneme alignment information, representing fundamental frequency (F0), and root-mean-square energy (RMS-E), respectively.

Second is the lack of connection from APA to automated speaking assessment (ASA), hindering comprehensive proficiency evaluation. We propose to integrate the handcrafted features (HCFs) from automated speaking assessment (ASA) into APA, which exploit alignment information and follow human-designed formulas to compute, including fluency [18], [19], pronunciation [20], [21], and rhythm [22], aggregated across utterance-level for offering interpretable L2 scores [23], [24], e.g., speech rate too slow. Our approaches pave the way for more holistic and precise APA systems. To summarize the contributions in this work: (1) a MTP framework that infuses phonetic and prosodic features into the phoneme encoder for promoting the overall APA model's performance; and (2) the HCFs in ASA are integrated into the APA, paving the way to a comprehensive system.

II. METHODOLOGY

A. Task Formulation and Model Architecture

We tackle the APA task with the following flow. Given X is the L2 learner's speech input and T is the reference text prompt, the goal is to produce a proficiency score vector $\mathbf{s}$, encompassing aggregate and multi-aspect scores at phone, word, and utterance levels. Formally, APA is defined as a mapping $\mathbf{s} = f(X, T; \theta)$, where f is a model parameterized by θ, evaluating pronunciation by comparing X to T [4] as depicted in Figure 2, the preprocessing extracts phonetic and prosodic sequential features Z (e.g., phoneme alignments, duration, pitch) from X and T, and the fine-grained HCFs J, which are used to generate multi-aspect proficiency scores.

The APA model comprises encoders and scoring modules. The phoneme encoder, denoted as θ_p, processes Z to produce a sequence of latent representations $\mathrm{H} = \{\mathbf{h_1}, \mathbf{h_2}, \ldots, \mathbf{h_n}\}$, where each $\mathbf{h_i}$ encodes phoneme-level features. We adopt a hierarchical architecture, inspired by [7], to encode multi-granularity (phone, word, utterance) representations, using an E-Branchformer as the backbone [25], integrating convolutional and Branchformer layers to capture local and global

TABLE II
DELIVERY IS AN IMPORTANT ASPECT OF PRONUNCIATION ASSESSMENT, FOCUSING ON THE ANALYSIS AND USE OF ACOUSTIC INFORMATION. IT IS DIVIDED INTO PRONUNCIATION AND FLUENCY, ALONG WITH THE FEATURE ELEMENTS OF GRANULARITY LISTED BELOW [24], [26].

Evaluation	Aspects	Features	Items
Delivery	Pronunciation	Phone	Confidence
		Word	Confidence
		Pitch	F0
		Energy	RMS-E
		Rhythm	CCI
			rPVI
			nPVI
			Varco
			Number of Nucleus
			Number of Consonants
			Vowel-to-consonant Ratio
		Error Rate	Error Rate (ER)
			Match Error Rate (MER)
	Fluency	Silence	Silence
			Long Silence
		phone	Duration
			Count
			Frequency
		Word	Duration
			Character Length
			Count
			Frequency
			Repeat Count
			Distinct Count
			Number of Filled Pauses

dependencies, enhancing each level's precision. The scoring modules, denoted as θ_s, receive the projected J and map H to proficiency scores $\mathbf{s}$, including aggregate scores (e.g. total) and multi-aspect metrics (e.g., pronunciation accuracy, prosody quality), via regression heads for each granularity level. To optimize θ_p, we employ the MTP method to promote overall performance.

B. MTP Framework with Phonetic and Prosodic Subtasks

To enhance APA accuracy, we develop a multi-task learning (MTL) framework that pretrains a phoneme encoder to jointly model phonetic and prosodic features, as shown in Figure 2. The phoneme encoder, a 3-layer transformer, processes phonetic and prosodic sequential features as inputs and outputs, as detailed in Table I. Inspired by BERT [27], we apply a masking strategy, randomly masking 15% of phonetic features, with 90% replaced by a mask token and 10% unchanged, to enhance robustness. Three masking methods are used: 1) replacing phonemes with a mask token, 2) zeroing one-dimensional features (e.g., F0 values), and 3) zeroing multi-dimensional prosodic vectors (e.g., energy profiles). A curriculum learning approach progresses from non-masking teacher-forcing training to masking-based training, optimizing multi-granularity feature integration.

Phoneme Prediction. Predicts phones in the input sequence. This technique has been applied in text-to-speech [15], [16] and has achieved success, [17] also applies this task in their framework. It has 42 categories (39 phones, one `[PAD]` padding token, one `[MASK]` masking token, and one `[UNK]` special token).

Vowel/Consonant Prediction. Classifies phones as vowels or consonants. This idea has been used as input in [8]. It has vowel, consonant and a trash can category.

Articulation Trait Prediction. Predicts articulation traits, extending a binary vowel/consonant prediction, containing vowel, stop, affricate, fricative, aspirate, liquid, nasal, semivowel, and a trash can category. These features have been incorporated in the mispronunciation detection and diagnosis field [28], to the best of our knowledge, it is a novel use in the APA domain.

Phoneme Duration Prediction. Classifies phoneme-level durations, categorizing frame counts from 1 to 100 (capped at 100 for longer durations). [17] also uses this task.

Pitch Prediction and Energy Prediction. Predicts phoneme-level fundamental frequency (F0) and root-mean-square energy (RMS-E) vectors. Using `librosa.pyin` [29], [30] to extract F0 and `librosa.feature.rms` to extract energy from raw speech data, we aggregate by phoneme duration to compute seven metrics: mean, standard deviation, median, median absolute deviation, maximum, minimum, and sum, yielding seven-dimensional feature vectors.

The MTP loss is a weighted sum of subtask losses, where their weights are hyperparameters tuned to balance subtask contributions, ensuring the encoder captures phonetic and prosodic features for multi-view APA.

C. HCFs in Automated Speaking Assessment

In ASA, we categorize the HCFs related to pronunciation assessment into two main aspects: pronunciation and fluency. According to the definition in phonetics [31], the features used for these aspects include elements such as duration, speed, volume, and pitch, which can be obtained through duration, energy, and F0, respectively. The detailed HCF items used are shown in Table II. Some of the listed HCFs include confidence, energy, silence, long silence, duration, and phoneme length, which are further processed statistically to obtain values such as mean, standard deviation, median, median absolute deviation, maximum, minimum, and summation. Additionally, the total duration of the audio file and the duration of the actual spoken content are included. These are iconic components in ASA for evaluating the overall speaking proficiency. These features are conveyed to the projection linear layers, fusing into the final decisions.

III. EXPERIMENTS

A. Experimental Setup

We adhere to the hyperparameters from [7] for MTP and training of our APA model. To quantify epistemic uncertainty, we train with five different random seeds. To evaluate the APA model's performance, the Pearson correlation coefficient (PCC) is adopted, while the mean squared error (MSE) is only used for the phoneme accuracy.

To address the skewed word distribution and limited vocabulary of word inputs for the APA model, we employ modern-BERT [32], a BERT variant optimized for robust tokenization, to generate token-level embeddings, replacing the inherent word embeddings in current APA models. To ensure accurate phone-word alignment, we use the Myers diff algorithm [33],

TABLE III

PERFORMANCE METRICS FOR DIFFERENT METHODS. HIERCB-IMP IS THE RESULT OF OUR IMPLEMENTED VERSION OF HIERCB. HIERCB-BPE-F0 DENOTED HCBBF. COMP., ACC. INDICATES COMPLETENESS AND ACCURACY, RESPECTIVELY. **BOLD** AND UNDERLINE DENOTE THE BEST AND THE SECOND-BEST PERFORMANCE IN EACH ASPECT, RESPECTIVELY. ⊗ SYMBOL INDICATES THE ACCUMULATION OF SUBTASKS IN MTP.

Methods	Phoneme Score		Word Score (PCC)				Utterance Score (PCC)					
	MSE ↓	PCC ↑	Acc. ↑	Stress ↑	Total ↑	Avg. PCC ↑	Acc. ↑	Comp. ↑	Fluency ↑	Prosodic ↑	Total ↑	Avg. PCC ↑
HierCB [7]	0.076	0.680	0.630	0.355	0.645	-	0.772	0.677	0.827	0.822	0.796	-
HierCB-imp	0.078 (0.001)	0.660 (0.002)	0.608 (0.011)	0.385 (0.045)	0.625 (0.011)	0.500 (0.020)	0.757 (0.006)	0.685 (0.153)	0.835 (0.004)	0.826 (0.003)	0.787 (0.004)	0.723 (0.033)
HierCB-f0	0.077 (0.000)	0.665 (0.001)	0.616 (0.006)	0.375 (0.025)	0.635 (0.006)	0.542 (0.012)	0.758 (0.008)	0.836 (0.129)	0.836 (0.007)	0.826 (0.005)	0.787 (0.007)	0.758 (0.031)
HCBbf	0.080 (0.000)	0.651 (0.003)	0.602 (0.004)	0.397 (0.033)	0.618 (0.004)	0.539 (0.014)	0.753 (0.007)	0.705 (0.106)	0.833 (0.005)	0.824 (0.004)	0.782 (0.004)	0.779 (0.025)
$+S_P$	**0.077** (0.001)	0.670 (0.004)	**0.631** (0.004)	0.402 (0.037)	**0.647** (0.004)	**0.560** (0.015)	**0.772** (0.004)	**0.791** (0.101)	0.836 (0.004)	0.826 (0.003)	**0.796** (0.004)	**0.804** (0.024)
$\otimes S_D$ [17]	0.078 (0.000)	0.659 (0.002)	0.622 (0.002)	0.361 (0.026)	0.637 (0.022)	0.540 (0.010)	0.762 (0.002)	0.755 (0.099)	0.835 (0.004)	0.824 (0.003)	0.788 (0.002)	0.793 (0.022)
$\otimes S_V$	0.078 (0.001)	0.659 (0.002)	0.620 (0.008)	0.316 (0.022)	0.637 (0.009)	0.524 (0.013)	0.769 (0.002)	0.694 (0.124)	0.833 (0.008)	0.826 (0.005)	0.794 (0.002)	0.783 (0.028)
$\otimes S_A$	0.079 (0.001)	0.659 (0.002)	0.617 (0.005)	**0.425** (0.039)	0.634 (0.004)	0.559 (0.016)	0.764 (0.003)	0.542 (0.191)	0.834 (0.003)	0.824 (0.003)	0.789 (0.002)	0.751 (0.041)
$\otimes S_F$	0.078 (0.001)	0.660 (0.003)	0.623 (0.004)	0.370 (0.017)	0.639 (0.003)	0.544 (0.008)	0.766 (0.002)	0.709 (0.116)	**0.839** (0.001)	**0.827** (0.002)	0.790 (0.002)	0.786 (0.025)
$\otimes S_R$	0.080 (0.001)	0.649 (0.003)	0.598 (0.013)	0.400 (0.041)	0.617 (0.011)	0.538 (0.021)	0.755 (0.010)	0.504 (0.234)	0.832 (0.007)	0.822 (0.006)	0.780 (0.008)	0.739 (0.053)

which efficiently aligns token sequences to their corresponding words. Multiple token embeddings are then aggregated into a single word embedding using attention pooling, where a clipped softmax of the attention matrix weights the contributions of each token. For preparing the HCFs in Table II, we adopt the pretrained ASR model with TDNN-F framework to retrieve the alignment information [5]. But for the error rates (ERs), we adopt an ASR under the framework of a SSL acoustic model with differentiable weighted finite-state transducers [34], pretrained on Librispeech [35] corpus, a 960-hour training set, to produce the transcriptions. The transcribed results may deviate from the prompts due to the listener's perceptions of L2 pronunciation errors. Thus, we adopt a phonetically oriented aligner [36], which retains the phonetic and linguistic relationships, to compute the ERs in reference and hypothesis transcripts, including the case of homophonic errors. After extracting HCFs, we eliminate those with skewed distributions, binary features exceeding a threshold (0.8), and duplicates. We then apply feature selection using a multivariate regressor model [37] to identify the most relevant HCFs for building an optimized utterance-level proficiency assessment model. Finally, we obtained 56, 66, 51, 50, and 53 dimensions of HCFs for each aspect at the utterance level, totalling 112 unique dimensions.

B. Corpus

We conducted APA experiments on the speechocean762 corpus [5], which is a publicly available dataset specifically designed for research on APA [27]. 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 annotates utterance-level scores (accuracy, stress, completeness, fluency, prosodic, total), word-level scores (accuracy, stress, total) and phoneme accuracy score.

C. Experimental Results

Table III presents the performance of our APA models compared to the original HierCB model [7]. Our reproduced baseline, HierCB-imp replicates HierCB by rerunning the training without altering the implementation, resulting in minor performance differences. We introduce HierCB-f0, which incorporates fundamental frequency (F0) as an additional input feature, and HierCB-bpe-f0 (HCBbf), which further replaces word embeddings with token-level embeddings from modern-BERT [32]. Additional MTP subtasks are evaluated, including phoneme prediction (S_P), duration prediction (S_D) [17], vowel/consonant classification (S_V), articulation prediction (S_A), F0 prediction (S_F), and RMS-E prediction (S_R).

The results show that HierCB-f0 outperforms HierCB-imp in many aspects, while HCBbf underperforms slightly compared to HierCB-f0 but HCBbf addresses the issue of word embedding; we prefer to use this one in our further experiments.

Consequently, we adopt HCBbf using MTP. The phoneme-, word-, and utterance-level averaged performance peaks with S_P, achieving MSE 0.077, PCCs 0.560 and 0.804, respectively. Utterance-level fluency and prosodic metrics peak with S_F, reaching PCCs of 0.839 (fluency) and 0.827 (prosodic), outperforming baselines. Adding S_A and S_R subtasks further enhances word-level stress (PCC 0.425, 0.400). These findings suggest that phoneme- and word-level metrics primarily benefit from the phonetic subtask, whereas utterance-level metrics depend on both phonetic and prosodic subtasks, with prosodic features aiding stress and prosody. Besides, unlike [17]'s phoneme duration and phoneme subtasks, which underperform, our S_V, S_A, and S_F subtasks yield superior results.

We also pretrained on only-phonetic and only-prosodic groups to observe the efficacy of each task group. Table

TABLE IV
THE EFFICACY OF ONLY-PHONETIC (PHN.) AND ONLY-PROSODIC (PROS.) GROUPS OVER PHONEME AND THE AVERAGED (AVG.) PCC IN WORD AND UTTERANCE.

Methods		Phoneme		Word	Utterance
		MSE ↓	PCC ↑	Avg. PCC ↑	Avg. PCC ↑
HCBbf		0.080 (0.000)	0.651 (0.003)	0.539 (0.014)	0.779 (0.025)
phn.	+S_P	0.077 (0.001)	**0.670** (0.004)	**0.560** (0.015)	**0.804** (0.024)
	+$S_P \otimes S_V$	**0.076** (0.000)	**0.670** (0.001)	0.545 (0.011)	0.767 (0.012)
	+$S_P \otimes S_V \otimes S_A$	0.077 (0.001)	0.664 (0.001)	0.527 (0.019)	0.786 (0.030)
pros.	+S_D	0.081 (0.001)	0.647 (0.005)	0.525 (0.016)	0.768 (0.027)
	+$S_D \otimes S_F$	0.080 (0.001)	0.648 (0.003)	0.525 (0.012)	0.773 (0.031)
	+$S_D \otimes S_F \otimes S_R$	0.081 (0.001)	0.641 (0.007)	0.531 (0.035)	0.757 (0.037)

TABLE V
COMPARISON WITH ER [38] AND HCF. PHN. ONLY, WRD. ONLY AND UTT. ONLY INDICATES HCFS APPENDING ONLY ON THE APA PHONE, WORD OR UTTERANCE REGRESSORS, RESPECTIVELY.

Methods	Phoneme		Word	Utterance
	MSE ↓	PCC ↑	Avg. PCC ↑	Avg. PCC ↑
HCBbf	0.080 (0.000)	0.651 (0.003)	0.539 (0.014)	0.779 (0.025)
+ER (All Heads) [38]	0.081 (0.001)	0.645 (0.003)	0.519 (0.037)	0.781 (0.046)
+ER (Phn. Only)	0.080 (0.001)	0.649 (0.002)	0.530 (0.011)	0.759 (0.015)
+ER (Wrd. Only)	0.078 (0.001)	0.663 (0.002)	0.559 (0.015)	0.757 (0.028)
+ER (Utt. Only)	0.078 (0.001)	0.660 (0.003)	0.528 (0.025)	0.770 (0.039)
(1) +ER (Utt. Only)	**0.076** (0.000)	**0.674** (0.001)	**0.548** (0.015)	0.738 (0.042)
(2) +ER (Utt. Only)	0.078 (0.001)	0.659 (0.002)	0.526 (0.012)	0.752 (0.058)
+HCF (All Heads)	0.084 (0.001)	0.629 (0.001)	0.538 (0.016)	0.783 (0.014)
+HCF (Phn. Only)	0.080 (0.001)	0.650 (0.004)	0.537 (0.013)	0.768 (0.030)
+HCF (Wrd. Only)	0.080 (0.001)	0.655 (0.002)	0.529 (0.026)	0.769 (0.029)
+HCF (Utt. Only)	0.079 (0.001)	0.658 (0.003)	0.527 (0.024)	0.793 (0.010)
(1) +HCF (Utt. Only)	0.077 (0.003)	0.670 (0.009)	0.540 (0.016)	**0.795** (0.015)
(2) +HCF (Utt. Only)	0.079 (0.002)	0.656 (0.005)	0.537 (0.015)	0.784 (0.023)

IV shows that the phonetic group contributes more than the prosodic group does. Despite this, fluency and prosodic in the utterance-level get their peak performance when adding the S_F subtask. This indicates the importance of both groups of subtasks in MTP.

D. Fusing HCFs and Discussions

Fusing HCFs into utterance-level decisions preserves APA model interpretability while enhancing holistic proficiency

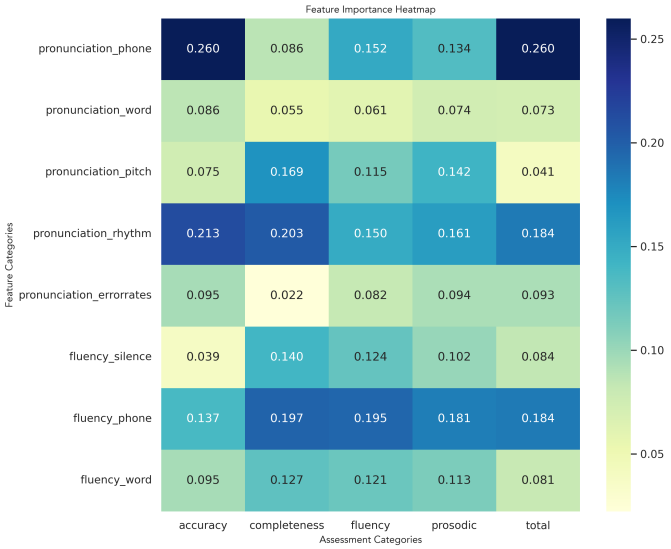


Fig. 3. The illustration of the relation in assessment categories (the aspects in the utterance-level) and feature categories (the features column in Table II).

assessment. We compare two configurations from Table III: HCBbf+S_P (denoted (1)) and HCBbf+$S_P \otimes S_D \otimes S_V \otimes S_A \otimes S_F$ (denoted (2)), representing for the phone-word and utterance groups, respectively. In Table V, we first compare our approach with [38], adding ER metrics, which introduce the character error rate (CER) and phoneme match error rate (MER-P), improves the utterance-level PCC from 0.779 to 0.781. However, applying ER fusion solely at utterance-level regressors (Utt. Only) outperforms fusion across all regressors (All Heads), as well as phone-only (Phn. Only) and word-only (Wrd. Only) configurations, making it our preferred strategy for further comparisons.

Then, we integrate highly relevant HCFs into the final decisions. Figure 3 illustrates HCF importance. Surprisingly, ER contributes less than expected, while with `fluency_phone` and `pronunciation_phone` scoring above 0.150, highlighting the significance of phonetic features, `pronunciation_rhythm` contributes across multiple aspects. +HFC Phn. Only and Wrd. Only do not outperform Utt. Only at the utterance level. And (1) +HCF (Utt. Only) achieves a PCC of 0.795. This suggests that HCFs enhance utterance-level ASA more than phoneme/word-level APA, ideal for holistic evaluation in CALL systems, while phoneme- and word-level evaluations benefit more from fine-grained features.

To bridge APA and ASA, we further discuss how APA's phonetic and prosodic outputs contribute to holistic speaking proficiency. Correlation analysis reveals that APA's phoneme accuracy and stress alignment strongly influence ASA's fluency scores (e.g., speech rate, silence duration), as well as ASA's rhythm scores, which represent prosodic variations, making it more vivid and natural through elements. Regular prosodic changes help speakers maintain stable speech rates while making it easier for listeners to understand meaning, validating the integration of HCFs. These HCFs, derived from human-

designed formulas, produce transparent feedback, such as slow speech rate or excessive pauses, enabling learners to address specific issues and educators to design targeted exercises.

CONCLUSIONS

This work proposes a joint modelling approach for phonetic and prosodic subtasks during the pretraining of the APA model's encoder, enhancing its efficacy. Additionally, we explore HCFs, bridging the gap toward a comprehensive system for holistic assessment, offering personalized feedback for L2 learners and data-driven insights for educators.

REFERENCES

[1] P. Howson, *The English effect*. London: British Council, 2013.

[2] M. Zhang, "Contrasting automated and human scoring of essays," *R & D Connections*, vol. 21, no. 2, pp. 1–11, 2013.

[3] P. M. Rogerson-Revell, "Computer-assisted pronunciation training (CAPT): Current issues and future directions," *RELC Journal*, vol. 52, no. 1, pp. 189–205, 2021.

[4] Y. El Kheir, A. Ali, and S. A. Chowdhury, "Automatic pronunciation assessment - a review," in *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023, pp. 8304–8324.

[5] J. Zhang, Z. Zhang, Y. Wang, Z. Yan, Q. Song, Y. Huang, K. Li, D. Povey, and Y. Wang, "speechocean762: An open-source non-native English speech corpus for pronunciation assessment," in *Proc. of Interspeech*, 2021, pp. 3710–3714.

[6] Y. Gong, Z. Chen, I.-H. Chu, P. Chang, and J. Glass, "Transformer-based multi-aspect multi-granularity non-native English speaker pronunciation assessment," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7262–7266.

[7] B.-C. Yan, Y.-C. Wang, J.-T. Li, M.-S. Lin, H.-W. Wang, W.-C. Chao, and B. Chen, "ConPCO: Preserving phoneme characteristics for automatic pronunciation assessment leveraging contrastive ordinal regularization," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[8] F.-A. Chao, T.-H. Lo, T.-I. Wu, Y.-T. Sung, and B. Chen, "3M: An effective multi-view, multi-granularity, and multi-aspect modeling approach to English pronunciation assessment," in *Proc. of IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, pp. 575–582.

[9] ——, "A hierarchical context-aware modeling approach for multi-aspect and multi-granular pronunciation assessment," in *Proc. of Interspeech*, 2023, pp. 974–978.

[10] H. Do, Y. Kim, and G. G. Lee, "Hierarchical pronunciation assessment with multi-aspect attention," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[11] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2Vec2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems (NeurIPS)*, vol. 33, pp. 12 449–12 460, 2020.

[12] 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/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.

[13] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[14] B.-C. Yan, H.-W. Wang, Y.-C. Wang, J.-T. Li, C.-H. Lin, and B. Chen, "Preserving phonemic distinctions for ordinal regression: A novel loss function for automatic pronunciation assessment," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–7.

[15] Y. Jia, H. Zen, J. Shen, Y. Zhang, and Y. Wu, "PnG BERT: Augmented BERT on phonemes and graphemes for neural TTS," in *Proc. of Interspeech*, 2021, pp. 151–155.

[16] L. The Nguyen, T. Pham, and D. Q. Nguyen, "XPhoneBERT: A pre-trained multilingual model for phoneme representations for text-to-speech," in *Proc. of Interspeech*, 2023, pp. 5506–5510.

[17] 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. of Interspeech*, 2023, pp. 949–953.

[18] H. Strik and C. Cucchiarini, "Automatic assessment of second language learners' fluency," 1999.

[19] A. Preciado-Grijalva and R. F. Brena, "Speaker fluency level classification using machine learning techniques," *arXiv preprint arXiv:1808.10556*, 2018.

[20] K. Takai, P. Heracleous, K. Yasuda, and A. Yoneyama, "Deep learning-based automatic pronunciation assessment for second language learners," in *HCI International 2020 - Posters*. Springer International Publishing, 2020, pp. 338–342.

[21] L. Chen, K. Evanini, and X. Sun, "Assessment of non-native speech using vowel space characteristics," in *Proc. of IEEE Spoken Language Technology Workshop (SLT)*, 2010, pp. 139–144.

[22] K. Kyriakopoulos, K. M. Knill, and M. J. Gales, "A deep learning approach to automatic characterisation of rhythm in non-native English speech," in *Proc. of Interspeech*, 2019, pp. 1836–1840.

[23] N. F. Chen and H. Li, "Computer-assisted pronunciation training: From pronunciation scoring towards spoken language learning," in *Proc. of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2016, pp. 1–7.

[24] K. Zechner and K. Evanini, *Automated speaking assessment: Using language technologies to score spontaneous speech*. Routledge, 2019.

[25] K. Kim, F. Wu, Y. Peng, J. Pan, P. Sridhar, K. J. Han, and S. Watanabe, "E-branchformer: Branchformer with enhanced merging for speech recognition," in *Proc. of IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 84–91.

[26] K. Kyriakopoulos, "Deep learning for automatic assessment and feedback of spoken English," Ph.D. dissertation, Queens' College, 2021.

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional Transformers for language understanding," in *Proc. of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019, pp. 4171–4186.

[28] 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 *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[29] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in Python." in *SciPy*, 2015, pp. 18–24.

[30] M. Mauch and S. Dixon, "PYIN: A fundamental frequency estimator using probabilistic threshold distributions," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 659–663.

[31] R. Wayland, *Phonetics: A practical introduction*. Cambridge University Press, 2018.

[32] B. Warner, A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, A. Gallagher, R. Biswas, F. Ladhak, T. Aarsen, G. T. Adams, J. Howard, and I. Poli, "Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference," in *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025, pp. 2526–2547.

[33] E. W. Myers, "An o (nd) difference algorithm and its variations," *Algorithmica*, vol. 1, no. 1-4, pp. 251–266, 1986.

[34] Z. Zhao and P. Bell, "Advancing CTC models for better speech alignment: A topological approach," in *Proc. of IEEE Spoken Language Technology Workshop (SLT)*, 2024, pp. 279–285.

[35] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[36] N. Ruiz and M. Federico, "Phonetically-oriented word error alignment for speech recognition error analysis in speech translation," in *Proc. of IEEE Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 296–302.

[37] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 58, no. 1, pp. 267–288, 1996.

[38] H. Do, W. Lee, and G. G. Lee, "Acoustic feature mixup for balanced multi-aspect pronunciation assessment," in *Proc. of Interspeech*, 2024, pp. 312–316.