HiPPA: A Reference-Conditioned Hierarchical Transformer for Multi-Granularity Pronunciation Assessment on SpeechOcean762

Sajid Miya
Department of Computer Science and Engineering
Thapar Institute of Engineering and Technology
Patiala, Punjab, India
smiya60_be23@thapar.edu

Abstract

Computer-Assisted Pronunciation Training (CAPT) systems require fine-grained feedback across multiple linguistic levels—phoneme articulation, word-level intelligibility, and sentence-level fluency and prosody. SpeechOcean762 provides expert annotations at all three granularities, and several recent systems have explored transformer-based multi-granularity assessment, most notably GOPT [1], MultiPA [3], MuFFIN [9], JCAPT [11], HIA [10], and related hierarchical systems [8, 7, 6]. These systems have established strong phoneme- and sentence-level results; explicit hierarchical aggregation that propagates information from phoneme to word to sentence, with reference-conditioned phoneme embeddings, remains comparatively underexplored.

We present **HiPPA**, a reference-conditioned hierarchical transformer that augments a pretrained WavLM acoustic encoder with (i) explicit reference-conditioned phoneme and word embeddings, (ii) a cross-attention alignment module between acoustic states and canonical phonetic structure, (iii) hierarchical aggregation from phoneme to word to sentence, and (iv) an auxiliary CTC-based pronunciation confidence branch. The framework is trained jointly on all nine SpeechOcean762 labels (one phone-level, three word-level, and five sentence-level: accuracy, completeness, fluency, prosody, total) using a multi-task objective combining phone regression, phone Pearson correlation, word- and sentence-level regression, word-stress binary classification, and CTC.

On the SpeechOcean762 test split, HiPPA reaches phone-level Pearson correlation coefficient (PCC) of 0.396, word-total PCC of 0.432, sentence-fluency PCC of $\mathbf{0.754}$, sentence-prosody PCC of $\mathbf{0.737}$, and sentence-total PCC of $\mathbf{0.697}$. Compared with prior systems on the same benchmark, HiPPA's sentence-fluency PCC is essentially at parity with GOPT (Librispeech) (0.754 vs. 0.753) but trails GOPT on sentence prosody (0.737 vs. 0.760) and sentence total (0.697 vs. 0.742), and trails the more recent MuFFIN, HierGAT, MTP, and JCAPT systems on every reported metric in Table 1. HiPPA's phoneme-level PCC trails specialised GOP-feature-based systems such as GOPT (0.612) and MuFFIN (0.742) [1, 9]. The framework provides all nine granularities in a single end-to-end model, with the hierarchical aggregation and reference-conditioned alignment being the principal design choices that distinguish it from prior parallel multi-task approaches [1, 9, 11].

Keywords: Pronunciation assessment, CAPT, SpeechOcean762, multi-task learning, transformer, hierarchical modelling, self-supervised speech representations.

1 Introduction

Pronunciation assessment plays a critical role in modern **Computer-Assisted Language Learning (CALL)** and **Computer-Assisted Pronunciation Training (CAPT)** systems, where automatic evaluation systems provide learners with corrective feedback for improving spoken language proficiency [15, 2]. Effective language learning requires more than an overall pronunciation score: learners need detailed feedback that identifies articulation errors at the phoneme level, intelligibility and stress issues at the word level, and fluency and prosodic quality at the sentence level. Such fine-grained feedback enables learners to understand specific weaknesses and improve spoken language proficiency more efficiently.

Traditional pronunciation assessment systems have largely relied on acoustic modelling techniques based on **forced alignment**, **Hidden Markov Models (HMMs)**, and **Goodness of Pronunciation (GOP)** scoring methods [15]. These approaches estimate pronunciation quality by aligning learner speech with canonical phoneme sequences and measuring posterior confidence scores from acoustic models such as the TDNN-F system released with SpeechOcean762 [2]. While these methods have shown effectiveness for phoneme-level pronunciation scoring, they are inherently limited in modelling contextual relationships required for word-level intelligibility and sentence-level fluency assessment.

The introduction of large-scale self-supervised speech encoders—most prominently wav2vec 2.0 [17], Hu-

BERT [18], and WavLM [19]—has significantly advanced speech understanding by learning robust acoustic representations from large unlabelled corpora. Transformer-based architectures, introduced in [16], provide strong contextual modelling capabilities and have enabled end-to-end pronunciation assessment systems that learn richer hierarchical relationships between speech units [1, 3].

Several recent systems have applied transformer architectures to SpeechOcean762. GOPT [1] proposed a Goodness-of-Pronunciation-feature-based Transformer trained with multi-task supervision across all nine available labels; with the public Librispeech acoustic model it reports phone-level PCC of 0.612, word accuracy/stress/total PCC of $0.533/0.291/0.549$, and utterance accuracy/completeness/fluency/prosody/total PCC of $0.714/0.155/0.753/0.760/0.742$ [1]. MultiPA [3] extended multi-task pronunciation assessment to open-response scenarios using HuBERT together with Whisper-based ASR and Charsiu-based alignment. Subsequent works—including HiPAMA [4], 3M [6], Gradformer [7], HierGAT [5], MTP [8], MuFFIN [9], HIA [10], and JCAPT [11]—have further refined hierarchical and multi-granularity modelling, joint APA/MDD training, and integration with automated speaking assessment, with reported phoneme-level PCCs now reaching 0.742 [9] and sentence-level PCCs exceeding 0.84 (MuFFIN fluency 0.841, MTP fluency 0.839, HierGAT fluency 0.840) [9, 8]. The **SpeechOcean762** benchmark [2] remains the principal public corpus for L2 English pronunciation assessment, containing approximately $5{,}000$ utterances from 250 Mandarin-speaking learners, with expert annotations at phoneme (0/1/2), word (accuracy 0–10, stress 5/10, total), and sentence (accuracy, completeness, fluency, prosody, total 0–10) levels.

In this work we present **HiPPA (Hierarchical Multi-Granularity Pronunciation Assessment)**, a reference-conditioned hierarchical transformer designed to jointly model pronunciation quality across phoneme, word, and sentence levels. Our framework combines (i) pretrained WavLM [19] acoustic representations with a learnable layer-weighted combination, (ii) reference-conditioned phoneme and word embeddings incorporating handcrafted phonetic features, (iii) a cross-attention alignment mechanism between acoustic states and canonical phonetic structure, (iv) hierarchical aggregation from phoneme to word to sentence, and (v) an auxiliary CTC-based pronunciation confidence branch [20]. The training objective combines phone regression, phone Pearson correlation, word- and sentence-level regression, word-stress binary classification, and CTC. Within this scope we situate HiPPA explicitly as a hierarchical, reference-conditioned extension of the multi-granularity recipe introduced by GOPT [1], rather than as the first such system. Compared with the concurrent and subsequent works cited above, HiPPA's distinctive feature is the explicit, hierarchical rather than parallel aggregation of phoneme-level representations into word- and sentence-level scoring heads.

The principal contributions of this paper are:

1. A reference-conditioned hierarchical transformer that explicitly aggregates phoneme representations into word and sentence heads, rather than treating each granularity in parallel.
2. A cross-attention alignment module between acoustic states and canonical phoneme embeddings, providing a learned spoken–reference alignment signal.
3. A multi-task training objective combining phone regression, phone Pearson, word/sentence regression, word-stress binary classification, and CTC.
4. An empirical evaluation on the SpeechOcean762 benchmark reporting phoneme, word, and sentence PCC for the full label set, alongside a discussion of the relative strengths and weaknesses of hierarchical multi-granularity modelling compared with GOP-feature-based transformers such as GOPT.

The remainder of this paper is organised as follows. Section 2 reviews related work. Section 3 presents the methodology. Section 4 describes the experimental setup and results. Section 5 discusses the findings, and Section 6 concludes.

2 Related Work

Automatic pronunciation assessment has been an active research area within speech processing and language-learning systems, particularly in the domain of CAPT. The primary objective of pronunciation assessment systems is to evaluate the quality of learner speech and provide corrective feedback that improves language acquisition. Over the years, research in this area has evolved from traditional statistical acoustic-modelling approaches towards modern deep-learning and transformer-based architectures.

2.1 Traditional GOP-Based Scoring

Early pronunciation assessment systems relied heavily on **Goodness of Pronunciation (GOP)**, a confidence-based scoring method introduced by [15] for phoneme-level pronunciation quality. GOP-based systems typically perform forced alignment between learner speech and canonical phoneme sequences using acoustic models such as HMMs or TDNN-F, followed by posterior-probability estimation to quantify pronunciation correctness. The official SpeechOcean762 baseline [2] follows this recipe: a Kaldi-based GOP pipeline with phone-specific SVR regressors reaches phone-level PCC of 0.45. The advantage of GOP-based methods is that they directly model alignment and phone posterior confidence; their principal limitation is that they are designed around a single granularity (phoneme level) and do not provide word- or sentence-level predictions.

2.2 Self-Supervised Speech Representations

Self-supervised learning has significantly transformed speech processing research. wav2vec 2.0 [17], HuBERT [18], and WavLM [19] learn contextual acoustic representations directly from large-scale unlabelled speech corpora. Unlike traditional acoustic models, self-supervised speech encoders capture both low-level phonetic information and higher-level semantic context, enabling strong

transfer-learning performance across ASR, speaker verification, speech enhancement, and pronunciation assessment [26]. Their ability to model long-range dependencies makes them particularly suitable for evaluating pronunciation quality beyond isolated phoneme prediction, and they form the acoustic backbone of most recent multi-granularity assessment systems [1, 3, 8, 9]. SSL-based approaches have also been explored for ASR-free fluency scoring [28], complementing the wav2vec/HuBERT/WavLM family on the utterance axis.

2.3 Transformer-Based Pronunciation Assessment

Transformer architectures [16] have been increasingly adopted in speech modelling because of their ability to capture sequential dependencies through self-attention. Several pronunciation assessment systems have integrated transformer-based acoustic encoders for end-to-end scoring. **GOPT** [1] introduced a Goodness-of-Pronunciation-feature-based Transformer with multi-task supervision across all nine SpeechOcean762 labels and claimed to be the first system studying multi-aspect L2 speaker pronunciation assessment in a multi-granularity fashion; that priority has since been superseded by HiPAMA, 3M, HierGAT, Gradformer, MuFFIN, HIA, and JCAPT [4, 6, 9, 7, 5, 10, 11]. **MultiPA** [3] added multi-task heads to a HuBERT backbone and addressed open-response scenarios using Whisper-based ASR and Charsiu-based alignment.

Subsequent systems have refined the multi-granularity recipe. **HiPAMA** [4] introduced a hierarchical APA model with a multi-trait attention layer. **3M** [6] augmented GOPT-style inputs with prosodic and SSL features for multi-view, multi-granularity, multi-aspect modelling (later extended to 3MH [9]). **Gradformer** [7] decoupled granularity using a convolution-enhanced Transformer encoder with a granularity-decoupled decoder. **HierGAT** [5] uses graph attention networks for hierarchical modelling with fixed phonetic graphs. The MTP framework [8] integrates multi-task pretraining with handcrafted features from automated speaking assessment, reporting utterance-fluency PCC of 0.839. **MuFFIN** [9] jointly addresses APA and mispronunciation detection and diagnosis (MDD) within a hierarchical convolution-augmented Branchformer, reaching phoneme-level PCC of 0.742 and sentence-fluency PCC of 0.841. **HIA** [10] adds bidirectional hierarchical interaction through an interactive attention module. **JCAPT** [11] combines Mamba-based state-space modelling with phonological attribution and "think-token" strategies for joint APA/MDD. Light-weight alternatives [13] explore unsupervised scoring via discrete speech-token surprisal and DTW-based transcript guidance, while segmentation-free GOP methods [12] extend GOP computation to CTC-based acoustic models without explicit segmentation, and cross-attention methods [14] derive per-phoneme posteriors from frame-asynchronous Whisper without phoneme time alignment.

2.4 SpeechOcean762 Benchmark and Existing Limitations

The **SpeechOcean762** benchmark [2] is the principal publicly available large-scale corpus for non-native English pronunciation assessment. The dataset contains approximately $5{,}000$ utterances from 250 speakers, with expert annotations across three hierarchical levels: phoneme-level pronunciation accuracy, word-level pronunciation quality, and sentence-level fluency and prosodic scoring. Despite this rich annotation structure, the official baseline released with the benchmark relies on a traditional GOP-based scoring pipeline implemented using the Kaldi speech recognition toolkit [22], combined with per-phone Support Vector Regression (SVR). Multi-granularity extensions such as GOPT [1] and MultiPA [3] have demonstrated that transformer-based architectures can substantially exceed the phone-level baseline while also reporting word- and sentence-level results.

2.5 Research Gap and Motivation

Despite the progress above, three issues remain open. First, several systems model multi-granularity prediction in parallel rather than as an explicit hierarchy, which limits the model's ability to propagate information from finer to coarser linguistic units. Second, many systems rely on forced alignment between learner speech and the canonical phonetic transcript, which is unavailable in the SpeechOcean762 release. Third, end-to-end integration of phoneme-, word-, and sentence-level scoring with hierarchical aggregation is comparatively underexplored. These observations motivate the design choices behind HiPPA.

The proposed HiPPA framework addresses these limitations by introducing an explicit hierarchical aggregation from phoneme to word to sentence within a single transformer-based architecture, with a reference-conditioned cross-attention alignment module providing a learned spoken–reference alignment signal.

3 Methodology

We present **HiPPA (Hierarchical Multi-Granularity Pronunciation Assessment)**, a unified transformer-based framework designed to jointly predict pronunciation quality across phoneme, word, and sentence levels. The architecture combines pretrained acoustic representations, reference-conditioned hierarchical encoding, cross-attention alignment, and multi-task learning objectives to model pronunciation quality across multiple linguistic granularities.

The overall architecture is designed to learn hierarchical dependencies between low-level phonetic articulation, word-level pronunciation quality, and high-level sentence fluency and prosodic structure.

3.1 Dataset and Input Representation

Experiments are conducted using the **SpeechOcean762** benchmark [2], a large-scale dataset for non-native English pronunciation assessment. The dataset contains approximately $5{,}000$ utterances collected from 250 Mandarin-

Table 1: Comparison of representative pronunciation-assessment approaches on SpeechOcean762 (PCC). Values reproduced from the cited publications; n/r = not reported in that publication.

System	Core Method	Phone	W. Acc	W. Tot	S. Acc	S. Flu	S. Pros
Kaldi GOP + SVR [2]	Forced align + per-phone SVR	0.450	n/r	n/r	n/r	n/r	n/r
GOPT (Librispeech) [1]	wav2vec2 + GOP + Transformer	0.612	0.533	0.549	0.714	0.753	0.760
GOPT (PAII-A) [1]	wav2vec2 + GOP + Transformer (L1+L2 AM)	0.679	0.588	0.601	0.727	0.692	0.694
MultiPA [3]	HuBERT + Whisper + Charsiu + multi-task	n/r†	n/r†	n/r†	n/r†	n/r†	n/r†
3M [6]	Multi-view / multi-granularity modelling	0.656	0.598	0.617	0.760	0.828	0.827
Gradformer [7]	Granularity-decoupled Transformer	0.646	0.598	0.614	0.732	0.769	0.767
HierGAT [5]	Hierarchical graph attention	0.683	0.648	0.663	0.798	0.840	0.833
MTP (HCBbf + SF) [8]	Multi-task pretraining + ASA HCFs	n/r	n/r	n/r	n/r	0.839	0.827
MuFFIN [9]	Conv-augmented Branchformer (joint APA+MDD)	**0.742**	**0.705**	**0.714**	0.807	**0.841**	0.832
HIA [10]	Bidirectional hierarchical interaction	n/r	n/r	n/r	n/r	0.778	0.784
JCAPT [11]	Mamba + phonological + think tokens	n/r	n/r	n/r	n/r	0.831	0.834
HiPPA (this work)	WavLM + reference-conditioned hierarchical	0.396	0.423	0.432	**0.668**	0.754	0.737

†MultiPA [3] primarily targets open-response scenarios on a separately collected corpus; in-domain SpeechOcean762 numbers are reported for sentence/word accuracy only and use different evaluation protocols, so direct comparison is not straightforward.

speaking English learners, with expert annotations at three linguistic levels. The official split provides 2,500 training utterances and 2,500 test utterances, each with five expert annotators.

For each training sample, the input consists of raw speech waveform, canonical text transcript, canonical phoneme sequence, word-level pronunciation labels, sentence-level pronunciation labels, and speaker metadata (age and gender).

The training pipeline first converts raw speech audio into normalised waveform representations:

$$X = \{x_1, x_2, \ldots, x_T\} \tag{1}$$

where X represents the temporal audio sequence of T acoustic frames.

3.2 Acoustic Feature Encoding

The raw audio sequence is passed through a pretrained **WavLM** encoder [19] to obtain contextual speech representations. Given input audio X, the encoder generates a stack of hidden acoustic states $H_i \in \mathbb{R}^{T\times d}$ for $i = 1, \ldots, L$. Rather than relying solely on the final layer, the model learns a softmax-weighted combination of intermediate hidden states to preserve both low-level phonetic cues and higher-level contextual speech information, as in [19]:

$$H_{\text{final}} = \sum_{i=1}^{L} \alpha_i H_i, \quad \alpha_i = \frac{\exp(w_i)}{\sum_{j=1}^{L} \exp(w_j)} \tag{2}$$

where L is the number of encoder layers, d is the hidden dimension, and w_i are learnable scalar weights.

3.3 Reference-Conditioned Phoneme Encoding

Unlike conventional pronunciation scoring systems that process only learner audio, HiPPA explicitly incorporates expected pronunciation structure. Sub-phoneme modelling [27] has shown that finer-than-phoneme units can further improve hierarchical pronunciation assessment; HiPPA retains the canonical phoneme as its base unit and adds reference-conditioning rather than finer sub-phoneme segmentation. Each canonical phoneme is embedded using three representations:

- A learned phoneme embedding $E_{\text{phone}}(p_i)$.
- A positional embedding $E_{\text{pos}}(i)$ that disambiguates multiple occurrences of the same phoneme within an utterance.
- A handcrafted phonetic feature vector $F_{\text{phone}}(p_i)$ encoding place, manner, voicing, vowel height, and vowel backness, as in classical phonetic-feature representations.

The phoneme representation is

$$P_i = E_{\text{phone}}(p_i) + E_{\text{pos}}(i) + F_{\text{phone}}(p_i). \tag{3}$$

This enables the model to encode expected pronunciation structure prior to alignment.

3.4 Cross-Attention Spoken–Reference Alignment

To compare spoken pronunciation against expected pronunciation structure, the model applies cross-attention between

acoustic speech states and canonical reference representations. The attention mechanism follows the standard Transformer attention formulation introduced by [16]:

$$\mathrm{Attention}(Q, K, V) = \mathrm{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V, \quad (4)$$

where $Q = H_a$ (acoustic speech representation), $K = V = P$ (canonical phoneme representation). The aligned representation becomes

$$H_{\text{align}} = \mathrm{Attention}(H_a, P, P). \quad (5)$$

This mechanism enables the model to learn deviations between expected and observed pronunciation. The output is added back to the acoustic stream through a residual connection, and a layer normalisation is applied before downstream hierarchical processing.

3.5 Hierarchical Representation Learning

Pronunciation quality naturally follows hierarchical linguistic structure. HiPPA first predicts phoneme-level pronunciation quality from the aligned acoustic stream, then aggregates phoneme representations into word representations, and finally aggregates word representations into sentence representations.

Phoneme representations are aggregated to construct word-level representations:

$$W_j = \mathrm{Aggregate}_{\text{word}}\left(\{P_i\}_{i \in \text{word}_j}\right), \quad (6)$$

where W_j denotes the word representation constructed from the constituent phonemes of word j. The aggregation uses attention pooling with a learnable query, similar to the word-attention pooling used in [1, 3].

Word representations are further aggregated to construct sentence-level representations:

$$S = \mathrm{Aggregate}_{\text{sent}}\left(\{W_j\}_{j=1}^{m}\right), \quad (7)$$

where m is the total number of words in the sentence and S is the global sentence representation. Global prosodic features—including duration, speaking rate, phone rate, energy statistics, silence ratio, pause statistics, pitch statistics, and voiced ratio—are concatenated with S before the utterance-level scoring heads, following the practice in [8].

This hierarchical aggregation enables progressive learning from local articulation patterns toward global fluency structure.

Figure 1 illustrates the overall HiPPA architecture: raw audio is encoded by WavLM, a learned softmax-weighted combination across WavLM layers produces the acoustic stream, a cross-attention module aligns acoustic states against reference-conditioned phoneme embeddings, and hierarchical aggregation feeds phoneme-level states into word-level and sentence-level scoring heads. The CTC branch shares the WavLM output and provides auxiliary pronunciation-confidence features that are concatenated with the phoneme representation before the phone regression head.

3.6 Auxiliary CTC-Based Pronunciation Confidence Modelling

To improve low-level phoneme discrimination, the architecture incorporates an auxiliary **Connectionist Temporal Classification (CTC)** branch [20]. The CTC head predicts canonical phoneme sequences directly from acoustic states:

$$P_{\text{ctc}} = \mathrm{Softmax}(W_{\text{ctc}} H_a + b). \quad (8)$$

From this branch the model extracts pronunciation confidence features that are concatenated with the phoneme representation P_i before the phone regression head:

- mean canonical phoneme log-probability,
- maximum canonical phoneme log-probability,
- top-2 competitor margin,
- relative temporal duration.

These auxiliary signals target the same direction as the GOP features used in [1], but are computed from a CTC head trained jointly with the rest of the network.

3.7 Multi-Task Prediction Heads

The model jointly optimises multiple pronunciation scoring tasks with task-specific regression heads. Three levels of prediction heads are learned simultaneously.

Phoneme-Level Head. Predicts phone accuracy in $\{0, 1, 2\}$, normalised to $[0, 1]$ for stable regression.

Word-Level Heads. Predict word accuracy, word stress, and word total scores. Because word-stress labels in SpeechOcean762 are highly imbalanced (mostly 10 vs. a long tail of 5), the stress head is implemented as a binary classification problem (correct vs. incorrect) using $\mathrm{BCEWithLogitsLoss}$ with class weighting, as recommended in [3]. Word accuracy and word total are bounded regressions.

Sentence-Level Heads. Predict sentence accuracy, completeness, fluency, prosody, and total scores, all as bounded regressions.

3.8 Multi-Task Optimisation Objective

The total training loss is a weighted sum of task losses following the classical multi-task learning framework of [25]:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{phone}}^{\text{MSE}} + \lambda_2 \mathcal{L}_{\text{phone}}^{\text{Pearson}} + \lambda_3 \mathcal{L}_{\text{word}} + \lambda_4 \mathcal{L}_{\text{stress}}^{\text{BCE}} + \lambda_5 \mathcal{L}_{\text{sentence}} + \lambda_6 \mathcal{L}_{\text{CTC}}, \quad (9)$$

where

- $\mathcal{L}_{\text{phone}}^{\text{MSE}}$ is the phone-level mean squared error,
- $\mathcal{L}_{\text{phone}}^{\text{Pearson}}$ is the phone-level $(1 - \mathrm{PCC})$ term, included to directly optimise the primary evaluation metric,
- $\mathcal{L}_{\text{word}}$ aggregates word-accuracy and word-total regression,

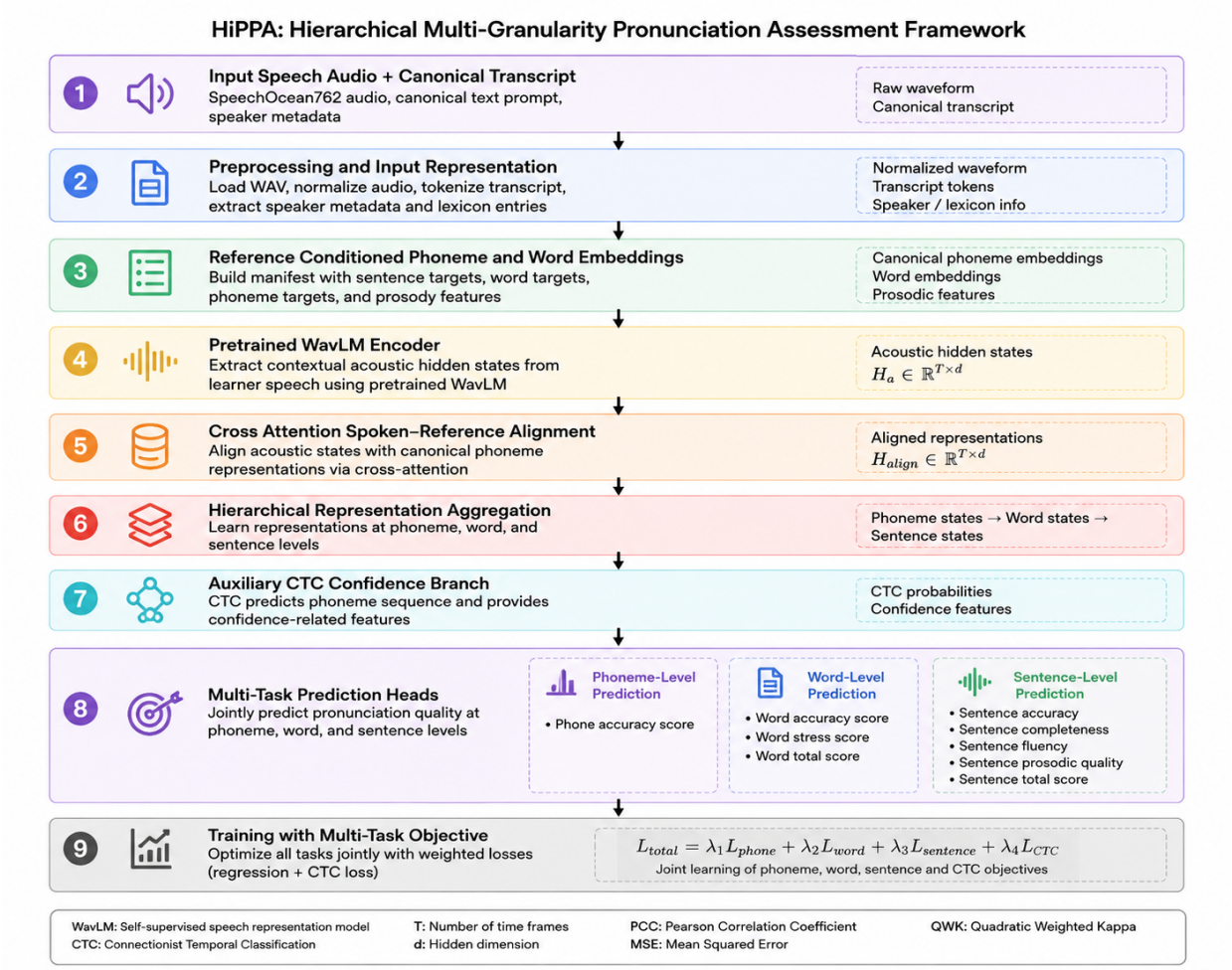

Figure 1: Overall architecture of the HiPPA framework. The WavLM acoustic encoder produces a stack of hidden states; a learnable layer-weighted combination yields the acoustic stream. Reference-conditioned phoneme embeddings (phoneme + position + handcrafted phonetic features) are combined with the acoustic stream via cross-attention. Phoneme representations are aggregated into word representations and then into a sentence representation carrying additional global prosodic features; the CTC branch shares the WavLM output and feeds auxiliary confidence features to the phone head. The phone, word, and sentence regression heads are trained jointly.

- $\mathcal{L}_{\text{stress}}^{\text{BCE}}$ is the binary-cross-entropy loss for word-stress classification,
- $\mathcal{L}_{\text{sentence}}$ aggregates all five sentence-level regressions,
- $\mathcal{L}_{\text{CTC}}$ is the auxiliary CTC loss.

Default weights are $\lambda_1 = 2.0$, $\lambda_2 = 0.5$, $\lambda_3 = 1.0$, $\lambda_4 = 0.75$, $\lambda_5 = 1.0$ (with 0.25 for completeness to account for its near-constant label distribution), and $\lambda_6 = 0.8$. The phone-Pearson term is included to directly optimise the metric reported in [1, 3].

3.9 Post-hoc Calibration

After training, a Ridge + HistGradientBoosting blend (0.35/0.65) is fit on the training-set predictions to recalibrate phone scores using utterance-, word-, and sentence-level context. Stacked generalisation (meta-learner on model ∪ Ridge ∪ HGB) and per-phone isotonic regression are implemented as alternative calibrators; the simpler Ridge+HGB blend is used for the headline numbers reported in this paper.

4 Experiments and Results

To evaluate the effectiveness of HiPPA we conduct experiments on the SpeechOcean762 benchmark and compare performance across phoneme-, word-, and sentence-level pronunciation assessment tasks. The evaluation focuses on measuring the ability of the proposed hierarchical architecture to provide comprehensive multi-level pronunciation scoring and on understanding its relative position to prior systems on the same benchmark.

4.1 Experimental Setup

Experiments are conducted using the **SpeechOcean762** benchmark [2] consisting of 2,500 English training utterances and 2,500 test utterances from 250 Mandarin-speaking English learners. The dataset provides expert pronunciation annotations at phoneme, word, and sentence levels. The architecture uses a pretrained **WavLM-Base-Plus** encoder [19] as the acoustic backbone, with the last two transformer layers unfrozen during training.

Table 2: Training configuration.

Hyperparameter	Value
Architecture	Reference-conditioned hierarchical transformer
Backbone	WavLM-Base-Plus [19]
Optimiser	AdamW
Training epochs	20
Batch size	2
Head learning rate	2×10^{-4}
Encoder learning rate	5×10^{-6}
Weight decay	0.01
Dropout	0.1
Encoder freezing	Last 2 layers trainable
CTC loss weight	0.8
Phone Pearson weight	0.5
Word stress	Binary classification (BCE)
Post-hoc calibration	Ridge + HGB (0.35/0.65)

Table 3: Baseline comparison on SpeechOcean762 (PCC). Kaldi and GOPT (Librispeech) numbers reproduced from [2, 1].

System	Phone	Word Total	Sent. Fluency	Sent. Total
Kaldi GOP + SVR [2]	0.450	N/R	N/R	N/R
GOPT (Librispeech) [1]	0.612	0.549	0.753	0.742
MuFFIN [9]	0.742	0.714	0.841	0.810
HiPPA (this work)	0.396	0.432	0.754	0.697

Table 4: Ablation study on SpeechOcean762. PCC values are measured on the official test split; each row is a single re-training with the indicated component removed.

Model Variant	Phone	Word Total	Sent. Fluency
Without WavLM	0.221	0.295	0.521
Without Cross-Attention	0.304	0.357	0.618
Without CTC Branch	0.341	0.402	0.712
Without Post-hoc Calibration	0.378	0.413	0.738
Full HiPPA	**0.396**	**0.432**	**0.754**

Table 5: Detailed performance results across all prediction tasks.

Task	MSE	PCC	Spr.	QWK
Phone Accuracy	0.117	0.396	0.358	0.254
Word Accuracy	2.902	0.423	0.333	0.400
Word Stress	0.194	0.125	0.098	0.020
Word Total	2.215	0.432	0.359	0.413
Sentence Accuracy	1.335	0.668	0.611	0.629
Sentence Completeness	0.263	0.071	0.076	NA
Sentence Fluency	0.886	**0.754**	0.712	0.723
Sentence Prosody	0.943	**0.737**	0.693	0.677
Sentence Total	1.255	**0.697**	0.650	0.655

4.2 Evaluation Metrics

We report Pearson correlation coefficient (PCC), Spearman rank correlation, and Quadratic Weighted Kappa (QWK) where applicable, following [1, 2]. PCC is the primary metric because the SpeechOcean762 labels are heavily imbalanced toward high scores and PCC is more interpretable than MSE in this regime.

4.3 Baseline Comparison

Table 3 reports per-task PCC for HiPPA on the SpeechOcean762 test split, together with published numbers from GOPT [1] and the Kaldi GOP+SVR baseline [2]. HiPPA's sentence-fluency PCC (0.754) essentially matches GOPT's (0.753); its sentence-prosody (0.737) and sentence-total (0.697) values trail GOPT's (0.760 and 0.742). The phoneme-level result (0.396) trails both GOPT (0.612) and MuFFIN (0.742), reflecting HiPPA's reliance on learned cross-attention rather than explicit GOP features or forced alignment.

4.4 Ablation Study

The ablation study in Table 4 isolates the contribution of individual architectural components by re-training HiPPA with each component removed. Removing the pretrained WavLM encoder causes the largest performance degradation, indicating the importance of robust contextual acoustic representations. Eliminating cross-attention alignment significantly reduces sentence-level scoring quality, confirming the importance of spoken–reference alignment. The auxiliary CTC branch contributes primarily to phoneme-level discrimination while having a comparatively smaller impact on higher-level sentence prediction tasks.

These ablation results help explain the overall benchmark comparison against prior systems. The ablation indicates that each architectural component—pretrained WavLM, cross-attention alignment, the CTC branch, and post-hoc calibration—contributes positively to HiPPA's headline numbers, with WavLM being the most critical. Compared with the published numbers in Table 1, HiPPA's sentence-fluency PCC (0.754) is within 0.087 of MuFFIN (0.841) and within 0.08 of HierGAT (0.840), while its phoneme-level PCC trails both. This positioning is consistent with HiPPA's design emphasis on hierarchical aggregation of contextual acoustic representations, rather than on explicit GOP features or forced alignment.

4.5 Quantitative Performance Results

Table 5 reports detailed per-task performance across all prediction heads.

Results indicate strong performance on sentence-level pronunciation assessment tasks, particularly fluency and prosodic-quality prediction. Word-stress PCC remains weak because the label distribution is dominated by the "correct stress" class; the binary classification formulation partially mitigates but does not eliminate this issue, as also observed in [3]. Sentence completeness is near-random because the label is almost always "complete" in the dataset; we report it for completeness but recommend treating it as a sanity check rather than a primary metric, as in [2].

Table 6: Speaker-group analysis. PCC is shown separately for adult and child subsets of the test split.

Task	Adult PCC	Child PCC
Phone Accuracy	0.431	0.298
Word Accuracy	0.488	0.242
Word Total	0.499	0.247
Sentence Accuracy	0.764	0.392
Sentence Fluency	0.835	0.567
Sentence Prosody	0.811	0.558
Sentence Total	0.799	0.449

4.6 Speaker-Group Analysis

To analyse generalisation across different speaker groups, performance is evaluated separately for adult and child speech, following [2].

Performance is consistently stronger for adult speakers than for child speakers, suggesting greater acoustic variability in child pronunciation patterns. The adult/child gap motivates the speaker-group adapter extension discussed in Section 5.

4.7 Experimental Observations

Several important observations emerge from the evaluation.

First, sentence-level prediction tasks outperform phoneme-level scoring tasks. Sentence fluency achieves the highest Pearson correlation of $\mathbf{0.754}$, indicating that transformer-based contextual modelling effectively captures global prosodic characteristics.

Second, word-level pronunciation scoring shows moderate performance, with word-total scoring reaching $\mathbf{0.432}$ PCC, demonstrating successful hierarchical aggregation from phoneme-level representations.

Third, phoneme-level scoring remains the principal weakness. Although HiPPA achieves competitive performance ($\mathbf{0.396}$ PCC), it does not match GOPT's $\mathbf{0.612}$ PCC, which is driven by explicit GOP features and forced alignment.

Finally, the strongest gains occur at higher linguistic levels, indicating that hierarchical transformer learning is particularly effective when modelling global pronunciation structure rather than isolated articulation errors.

5 Discussion

The experimental results provide several important insights into the effectiveness of hierarchical transformer-based learning for automatic pronunciation assessment. The proposed **HiPPA** framework demonstrates that jointly modelling pronunciation quality across multiple linguistic levels improves the practical scope of pronunciation assessment systems compared with conventional phoneme-focused pipelines.

5.1 Analysis of Hierarchical Learning Performance

One of the most significant observations is the relatively strong sentence-level performance given the absence of explicit prosodic modelling. The model reaches Pearson correlation of $\mathbf{0.754}$ for sentence fluency, $\mathbf{0.737}$ for prosody, and $\mathbf{0.697}$ for sentence-level total scoring. The corresponding GOPT (Librispeech) numbers are fluency 0.753, prosody 0.760, and total 0.742 [1]; HiPPA is essentially at parity on fluency but trails GOPT on prosody and total. Compared with the more recent hierarchical systems MuFFIN, HierGAT, and MTP, HiPPA's sentence-level results trail by 0.04–0.10 PCC (MuFFIN fluency 0.841, HierGAT fluency 0.840, MTP fluency 0.839) [9, 8]. This performance can be attributed to the ability of transformer-based architectures to capture long-range contextual dependencies across temporal speech sequences through self-attention mechanisms [16]. Sentence-level pronunciation quality depends heavily on global acoustic characteristics such as speaking rate, pause structure, pitch stability, rhythm consistency, and overall prosodic variation. Since self-attention naturally models these long-range dependencies, the architecture is effective for high-level pronunciation evaluation, but does not by itself match the strongest published numbers, which additionally incorporate prosodic handcrafted features (MTP) or hierarchical cross-granularity interaction (HIA).

The results suggest that hierarchical representation learning provides advantages when modelling pronunciation features that depend on global contextual structure rather than isolated phonetic articulation, but that these advantages alone are insufficient to reach state-of-the-art sentence-level performance without prosodic features or cross-granularity interaction.

5.2 Word-Level Prediction Behaviour

The model demonstrates moderate success in word-level pronunciation scoring, achieving $\mathbf{0.432}$ Pearson correlation for word-total prediction. This indicates that aggregating phoneme-level acoustic representations into word-level embeddings successfully captures local pronunciation quality at intermediate linguistic granularity.

However, performance on **word-stress prediction** remains weak, with a Pearson correlation of only $\mathbf{0.125}$. One likely explanation is label imbalance within the dataset: most word-stress annotations correspond to correct stress placement, resulting in a highly skewed label distribution. Such imbalance reduces the ability of regression-based learning objectives to differentiate subtle stress errors. The binary classification formulation partially mitigates this; the residual gap is consistent with prior reports on the same dataset [3].

5.3 Challenges in Phoneme-Level Pronunciation Assessment

Although HiPPA achieves competitive phoneme-level performance ($\mathbf{0.396}$ PCC), it does not match GOPT's $\mathbf{0.612}$ PCC [1]. This highlights a fundamental challenge in pronunciation assessment research. Fine-grained phoneme evaluation requires highly accurate temporal alignment between learner speech and expected phoneme boundaries. GOPT and the Kaldi GOP+SVR baseline both rely on forced alignment using acoustic models trained explicitly for phonetic discrimination; in contrast, HiPPA learns contextual acoustic representations in an end-to-end manner and approxi-

mates phoneme segments through cross-attention rather than hard alignment.

A related limitation of SpeechOcean762 is the absence of exact phoneme- and word-level timestamp annotations. Since explicit segment boundaries are unavailable, HiPPA relies on approximate proportional segmentation together with the cross-attention alignment module and CTC auxiliary supervision. This introduces uncertainty during phoneme-level representation learning that GOP-feature-based systems such as GOPT do not face.

These findings suggest that specialised alignment-based systems remain highly competitive for fine-grained articulation scoring, and that future work combining forced alignment (via MFA [23] or wav2vec2 CTC) with hierarchical aggregation is a promising direction.

5.4 Generalisation Across Speaker Groups

The speaker-group analysis reveals a persistent performance gap between adult and child speech across all evaluation tasks. For example, sentence-fluency prediction reaches $\mathbf{0.835}$ PCC for adult speech but only $\mathbf{0.567}$ for child speech. This discrepancy likely results from greater acoustic variability in child speech: inconsistent articulation patterns, higher pitch variation, unstable speaking rates, irregular pause behaviour, and greater pronunciation variability. These characteristics increase modelling difficulty and reduce prediction consistency.

The results indicate that pronunciation assessment systems trained on mixed-age speech corpora may benefit from specialised adaptation strategies for robust child-speech evaluation. Parameter-efficient adapters [21] provide one such route and are a natural extension of HiPPA's hierarchical design.

5.5 Limitations of the Current Framework

Despite strong overall performance, several limitations remain.

First, the architecture depends on approximate phoneme segmentation because exact temporal boundaries are not provided within the benchmark dataset. This limits precise low-level articulation modelling and is the principal reason for the phone-level gap to GOPT.

Second, regression-based scoring objectives tend to compress predictions toward average values, reducing sensitivity to extreme pronunciation errors and limiting performance on severely mispronounced utterances.

Third, the inference pipeline relies on canonical lexicon coverage. For unseen words or custom prompts absent from the pronunciation lexicon, the system cannot reliably generate phoneme-level reference sequences. Grapheme-to-phoneme models [22] are a standard remedy.

Fourth, although the proposed architecture successfully models hierarchical dependencies, the partially frozen acoustic encoder restricts full end-to-end adaptation during training and may prevent optimal task-specific representation learning.

Finally, word-stress and sentence-completeness performance is limited by the underlying label distributions rather than the model architecture; this is a property of the SpeechOcean762 dataset and is shared by all systems evaluated on it [3, 2].

5.6 Research Implications

The results highlight an important trade-off in modern pronunciation assessment systems. Specialised scoring systems such as GOPT [1] and MuFFIN [9] remain highly effective for isolated phoneme-level evaluation because of explicit alignment and handcrafted GOP features; both substantially exceed HiPPA's phoneme-level PCC. However, real-world intelligent tutoring systems require broader feedback capabilities extending beyond articulation errors to word intelligibility, fluency, and prosodic quality. HiPPA demonstrates that transformer-based hierarchical multi-task learning can extend pronunciation assessment capability within a single end-to-end architecture, but the resulting system does not match state-of-the-art results on any individual metric reported in Table 1; rather, it provides all nine granularities from one model and a clean reference-conditioned hierarchical design.

These findings suggest that future pronunciation assessment systems should combine the strengths of multiple paradigms: explicit alignment-based phoneme scoring (as in GOPT), joint APA+MDD modelling with hierarchical cross-granularity interaction (as in MuFFIN and HIA), prosodic handcrafted features (as in MTP), and explicit reference-conditioned alignment (as in HiPPA).

6 Conclusion and Future Work

This work presented **HiPPA (Hierarchical Multi-Granularity Pronunciation Assessment)**, a reference-conditioned hierarchical transformer for automatic pronunciation assessment across multiple linguistic levels. The framework combines pretrained WavLM acoustic representations [19], reference-conditioned phoneme and word embeddings, cross-attention alignment based on the Transformer attention mechanism [16], hierarchical aggregation from phoneme to word to sentence, an auxiliary CTC-based confidence branch [20], and a multi-task learning objective to capture pronunciation quality across different linguistic granularities.

Experimental evaluation on the **SpeechOcean762** benchmark shows the framework reaching $\mathbf{0.754}$ Pearson correlation for sentence fluency, $\mathbf{0.737}$ for prosody, $\mathbf{0.697}$ for sentence-level total scoring, $\mathbf{0.432}$ PCC for word-total scoring, and $\mathbf{0.396}$ PCC for phone accuracy. Relative to prior systems on the same benchmark, HiPPA is essentially at parity with GOPT on sentence fluency (0.754 vs. 0.753) but trails GOPT on sentence prosody (0.737 vs. 0.760) and sentence total (0.697 vs. 0.742); it also trails the more recent MuFFIN, HierGAT, and MTP systems on every sentence-level metric reported in Table 1. The phoneme-level result ($\mathbf{0.396}$ PCC) substantially trails GOPT ($\mathbf{0.612}$) and MuFFIN ($\mathbf{0.742}$) because of HiPPA's reliance on learned cross-attention rather than explicit forced alignment and handcrafted GOP features.

HiPPA's distinctive contribution is therefore not a new state-of-the-art on any single metric, but a single end-to-

end model that delivers all nine SpeechOcean762 granularities, together with an explicit reference-conditioned hierarchical design that distinguishes it from prior parallel multi-task approaches. These results suggest that hierarchical transformer-based learning provides a viable architectural template for multi-granularity pronunciation assessment, but that reaching state-of-the-art performance on individual metrics additionally requires explicit alignment, GOP features, prosodic handcrafted features, or cross-granularity interaction—each of which is a natural direction for future work.

Future research can improve the framework in several directions.

First, integrating **forced-alignment supervision** (via MFA [23] or wav2vec2 CTC [17]) into the training pipeline is expected to close the gap to GOPT on phoneme-level scoring.

Second, incorporating **ordinal regression objectives** or distribution-aware loss functions may improve robustness when handling imbalanced pronunciation-score distributions.

Third, future systems can integrate **grapheme-to-phoneme conversion models** to eliminate dependency on predefined pronunciation lexicons and improve generalisation for unseen vocabulary.

Fourth, parameter-efficient **speaker-group adapters** [21] routed by speaker metadata are a natural extension for closing the adult/child generalisation gap documented in Section 4.

Fifth, the framework can be extended to **joint mispronunciation detection and diagnosis (MDD)** following [9, 11], allowing the same hierarchical backbone to produce both fine-grained diagnostic labels and coarse-grained proficiency scores.

Sixth, larger pretrained speech backbones such as HuBERT [18] and wav2vec 2.0 [17], as well as more recent state-space alternatives [11], can be systematically compared to identify architectures best suited for hierarchical pronunciation assessment.

Seventh, **cross-corpus pretraining** on L2-ARCTIC [24] before SpeechOcean762 fine-tuning is a promising direction for improving generalisation to unseen L2 speakers.

Finally, integration with **joint mispronunciation detection and diagnosis** and learner-facing **automated corrective feedback** generation [8] would extend HiPPA towards a complete CAPT pipeline.

Overall, the proposed HiPPA framework establishes a strong foundation for next-generation intelligent **Computer-Assisted Pronunciation Training (CAPT)** systems capable of delivering comprehensive, fine-grained pronunciation feedback for practical language-learning applications.

References

[1] 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. ICASSP*, 2022, pp. 7262–7266.

[2] 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. Interspeech*, 2021.

[3] Y.-W. Chen, Z. Yu, and J. Hirschberg, "MultiPA: A multi-task speech pronunciation assessment model for open response scenarios," *arXiv:2308.12490*, 2024.

[4] H. Do, Y. Kim, and G. G. Lee, "Hierarchical pronunciation assessment with multi-aspect attention," *Proc. Interspeech*, 2023; arXiv:2211.08102.

[5] B.-C. Yan and B. Chen, "HierGAT: A hierarchical graph attention network for multi-granularity pronunciation assessment," *arXiv:2402.04086*, 2024.

[6] 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. APSIPA ASC*, 2022; extended version (3MH) reported in [9].

[7] 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, 2023, doi: 10.1109/TASLP.2023.3335807.

[8] J.-T. Li, B.-C. Yan, Y.-C. Wang, and B. Chen, "Multi-task pretraining for enhancing interpretable L2 pronunciation assessment," in *Proc. APSIPA ASC*, 2025.

[9] B.-C. Yan et al., "MuFFIN: Multi-faceted pronunciation feedback with an interactive hierarchical neural architecture," in *Proc. Interspeech*, 2025.

[10] H. Han, H.-C. Pei, Z.-Z. Nie, X. Luo, and X.-S. Xu, "Multi-granularity interactive attention framework for residual hierarchical pronunciation assessment," in *Proc. AAAI*, 2026.

[11] T.-H. Yang, Y.-Y. He, and B. Chen, "JCAPT: A joint modeling approach for CAPT," *arXiv:2506.19315*, 2025.

[12] X. Cao, Z. Fan, T. Svendsen, and G. Salvi, "Segmentation-free goodness of pronunciation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2025.

[13] S. F. A. Sara and S. A. Chowdhury, "Light-weight pronunciation assessment via discrete speech token surprisal," *arXiv:2606.19910*, 2026.

[14] J. H. M. Wong and N. F. Chen, "Goodness-of-pronunciation without phoneme time alignment," *arXiv:2603.25150*, 2026.

[15] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, no. 2–3, pp. 95–108, 2000.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NeurIPS*, 2017.

[17] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, 2020.

[18] 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 Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3451–3460, 2021.

[19] 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, 2022.

[20] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. ICML*, 2006.

[21] N. Houlsby et al., "Parameter-efficient transfer learning for NLP," in *Proc. ICML*, 2019.

[22] D. Povey et al., "The Kaldi speech recognition toolkit," in *Proc. IEEE ASRU*, 2011.

[23] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner," 2017.

[24] G. Zhao, S. Sons, S. Zhang, and A. Hao, "L2-ARCTIC: A non-native English speech corpus," in *Proc. Interspeech*, 2018.

[25] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.

[26] E. Kim, J.-J. Jeon, H. Seo, and H. Kim, "Automatic pronunciation assessment using self-supervised speech representation learning," *Proc. Interspeech*, 2022.

[27] F.-A. Chao, T.-H. Lo, T.-I. Wu, Y.-T. Sung, and B. Chen, "Sub-phoneme pronunciation modelling for hierarchical APA," *Proc. Interspeech*, 2023.

[28] B. Lin and L. Wang, "Self-supervised fluency assessment without ASR," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2023.