HiPPA: A Hierarchical Multi-Task Transformer Framework for Multi-Granularity Pronunciation Assessment

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, including phoneme articulation, word-level intelligibility, and sentence-level fluency and prosody. Existing pronunciation assessment benchmarks such as SpeechOcean762 provide annotations across these hierarchical levels; however, prior baseline approaches primarily focus on phoneme-level scoring using traditional Goodness of Pronunciation (GOP)-based pipelines, leaving higher-level assessment tasks underexplored.

In this work, we propose **HiPPA**, a hierarchical multi-task transformer framework for multi-granularity pronunciation assessment. The proposed architecture leverages pretrained **WavLM** acoustic representations, canonical phoneme and word reference embeddings, cross-attention mechanisms for spoken-reference alignment, and auxiliary **CTC-based confidence modeling** to jointly predict pronunciation quality at phoneme, word, and sentence levels within a unified learning framework. Unlike conventional pipelines that optimize independently for low-level phoneme scoring, our approach models hierarchical dependencies between linguistic units and enables end-to-end multi-level assessment.

Experiments conducted on the SpeechOcean762 corpus demonstrate strong performance on sentence-level tasks, achieving Pearson correlation coefficients of **0.754 for fluency**, **0.737 for prosodic quality**, and **0.697 for sentence-level total scoring**, while also producing competitive phoneme- and word-level predictions. The proposed framework demonstrates that hierarchical transformer-based learning can significantly expand pronunciation assessment capability beyond traditional phoneme-only systems, establishing a strong foundation for next-generation intelligent pronunciation tutoring systems.

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 [5]. Effective language learning requires more than an overall pronunciation score; learners need detailed feedback that identifies articulation errors at the phoneme level, pronunciation quality 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 modeling techniques based on **forced alignment**, **Hidden Markov Models (HMMs)**, and **Goodness of Pronunciation (GOP)** scoring methods [5]. These approaches estimate pronunciation quality by aligning learner speech with canonical phoneme sequences and measuring posterior confidence scores from acoustic models. While these methods have shown effectiveness in phoneme-level pronunciation scoring, they are inherently limited in modeling broader contextual relationships required for word-level intelligibility and sentence-level fluency assessment.

The introduction of large-scale self-supervised speech models such as **WavLM** [2], **wav2vec 2.0** [6], and **HuBERT** [7] has significantly advanced speech understanding tasks by learning robust acoustic representations from large unlabeled speech corpora. Transformer-based architectures have demonstrated strong contextual modeling capabilities and provide new opportunities for building end-to-end pronunciation assessment systems capable of learning richer hierarchical relationships between speech units [3].

The **SpeechOcean762** benchmark is a valuable resource for pronunciation assessment research, containing annotated pronunciation scores across phoneme, word, and sentence levels [1]. However, existing baseline implementations released with the benchmark primarily focus on phoneme-level pronunciation scoring using classical GOP-based pipelines, leaving higher-level assessment tasks largely unexplored.

To address these limitations, we propose **HiPPA (Hierarchical Multi-Granularity Pronunciation Assessment)**, a unified transformer-based framework for hierarchical pronunciation scoring. Our architecture integrates pretrained speech representations, reference-conditioned phoneme and

word embeddings, cross-attention alignment mechanisms, hierarchical representation aggregation, and multi-task learning objectives to jointly predict pronunciation quality across multiple linguistic levels. Unlike prior approaches that optimize only for low-level phoneme scoring, the proposed system learns dependencies across phoneme, word, and sentence structures within a single end-to-end framework.

To the best of our knowledge, this is the first transformer-based framework developed for the SpeechOcean762 benchmark that jointly models pronunciation quality at phoneme, word, and sentence levels within a unified multi-task learning architecture.

The main contributions of this work are summarized as follows:

1. We propose a hierarchical transformer-based architecture for multi-granularity pronunciation assessment.
2. We jointly model phoneme-, word-, and sentence-level pronunciation scoring through a unified multi-task learning framework.
3. We introduce reference-conditioned cross-attention mechanisms for aligning expected pronunciation structure with spoken audio representations.
4. We evaluate the framework on SpeechOcean762 and demonstrate strong performance across sentence-level and word-level pronunciation tasks while extending benchmark coverage beyond existing phoneme-only baselines.

The proposed framework moves pronunciation assessment systems closer to real-world intelligent tutoring applications by enabling comprehensive hierarchical pronunciation feedback rather than isolated phoneme-level evaluation.

2 Related Work

Automatic pronunciation assessment has been an active research area within speech processing and language learning systems, particularly in the domain of **Computer-Assisted Pronunciation Training (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 modeling approaches toward modern deep learning and transformer-based architectures.

2.1 Traditional Pronunciation Assessment Approaches

Early pronunciation assessment systems relied heavily on **Goodness of Pronunciation (GOP)**, a confidence-based scoring method introduced for evaluating phoneme-level pronunciation quality [5]. GOP-based systems typically perform forced alignment between learner speech and canonical phoneme sequences using acoustic models such as Hidden Markov Models (HMMs), followed by posterior probability estimation to quantify pronunciation correctness.

The original work by Witt and Young established GOP as a foundational method for automatic pronunciation scoring and demonstrated strong performance for phoneme-level articulation assessment [5]. However, these systems depend heavily on precise acoustic alignment and are limited in capturing higher-level linguistic structures such as word intelligibility, sentence fluency, and prosodic variation. As a result, their usefulness in practical tutoring systems remains constrained.

2.2 Deep Learning for Speech Representation Learning

Recent progress in self-supervised learning has significantly transformed speech processing research. Models such as **wav2vec 2.0** [6], **HuBERT** [7], and **WavLM** [2] learn contextual acoustic representations directly from large-scale unlabeled 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 automatic speech recognition, speaker verification, speech enhancement, and pronunciation assessment tasks. Their ability to model long-range dependencies makes them particularly suitable for evaluating pronunciation quality beyond isolated phoneme prediction.

2.3 Transformer-Based Pronunciation Assessment Systems

Transformer architectures have increasingly been adopted in speech modeling due to their ability to capture sequential dependencies using self-attention mechanisms introduced in the Transformer architecture [3]. Unlike recurrent neural networks and conventional acoustic pipelines, transformers can jointly model contextual relationships across multiple linguistic units.

Several recent pronunciation assessment systems have integrated transformer-based acoustic encoders for end-to-end scoring. However, most existing approaches focus on isolated evaluation tasks, primarily phoneme classification or sentence-level scoring. These architectures often fail to jointly model hierarchical dependencies between phonemes, words, and sentence-level pronunciation structures.

2.4 SpeechOcean762 Benchmark and Existing Limitations

The **SpeechOcean762** benchmark introduced a publicly available large-scale corpus for non-native English pronunciation assessment [1]. The dataset contains approximately 5,000 utterances from 250 speakers, with expert annotations provided 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 primarily relies on a traditional GOP-based scoring pipeline implemented using the Kaldi speech recognition toolkit [9] and combined with Support Vector Regression (SVR), evaluating only phoneme-level

Table 1: Comparative Analysis of Existing Approaches

Approach	Core Method	Limitations
GOP-Based Assessment	Forced alignment + posterior scoring	Limited to phoneme-level evaluation
GOP + SVR Baseline	Handcrafted features + regression	No word or sentence scoring
Transformer-Based Models	Self-supervised acoustic learning	Optimized for isolated tasks
SpeechOcean762 Baseline	Kaldi model + GOP pipeline	Underutilizes dataset hierarchy
Proposed HiPPA	Hierarchical transformer + multi-task learning	Supports phoneme, word, sentence assessment

pronunciation accuracy [1]. Consequently, word-level and sentence-level annotations remain largely underutilized despite being essential for real-world pronunciation tutoring applications.

2.5 Research Gap and Motivation

Existing pronunciation assessment systems suffer from three major limitations. First, traditional GOP-based pipelines focus exclusively on low-level phoneme scoring and fail to capture broader contextual pronunciation patterns [5]. Second, transformer-based pronunciation systems often target isolated scoring tasks rather than modeling hierarchical relationships between linguistic units [3]. Third, publicly available benchmarks such as SpeechOcean762 provide rich multi-level annotations, yet existing baseline implementations do not fully exploit this hierarchical supervision [1].

These limitations motivate the development of a unified framework capable of jointly modeling pronunciation quality across multiple granularities within a single architecture.

2.6 Comparative Analysis

The proposed HiPPA framework addresses these limitations by introducing a hierarchical multi-task architecture capable of jointly modeling pronunciation quality across multiple linguistic levels while leveraging modern transformer-based acoustic representations.

3 Methodology

We propose **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, a large-scale dataset for non-native English pronunciation assessment [1]. The dataset contains approximately **5,000 utterances collected from 250 Mandarin-speaking English learners**, with expert annotations available at three linguistic levels.

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
- Speaker metadata (age and gender)

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

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

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

3.2 Acoustic Feature Encoding

The raw audio sequence is passed through a pretrained **WavLM** encoder to obtain contextual speech representations [2]. Given input audio (X), the encoder generates hidden acoustic states:

$$H_a = Encoder(X) \tag{2}$$

where:

$$H_a \in \mathbb{R}^{T \times d} \tag{3}$$

and (d) denotes the hidden representation dimension.

Rather than relying solely on the final transformer layer, the model learns a weighted combination of intermediate hidden states to preserve both low-level phonetic cues and higher-level contextual speech information, following transformer-based contextual representation learning principles [3]:

$$H_{final} = \sum_{i=1}^{L} \alpha_i H_i \tag{4}$$

where:

- (L) represents encoder layers
- (α_i) are learnable layer weights

This allows the model to combine acoustic information from multiple representation levels.

3.3 Reference-Conditioned Phoneme Encoding

Unlike conventional pronunciation scoring systems that process only learner audio, the proposed framework explicitly incorporates expected pronunciation structure.

Each canonical phoneme is embedded using three representations:

- Phoneme embedding
- Positional embedding
- Linguistic feature embedding

The phoneme representation is defined as:

$$P_i = E_{phone}(p_i) + E_{pos}(i) + F_{phone}(p_i) \tag{5}$$

where:

- (p_i) represents canonical phoneme
- (E_{phone}) is phoneme embedding
- (E_{pos}) is position embedding
- (F_{phone}) represents handcrafted phonetic features

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 Vaswani et al. [3]:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6}$$

where:

- Query = acoustic speech representation
- Key = canonical phoneme representation
- Value = canonical phoneme representation

The aligned representation becomes:

$$H_{align} = Attention(H_a, P, P) \tag{7}$$

This mechanism enables the model to learn deviations between expected and observed pronunciation.

3.5 Hierarchical Representation Learning

Pronunciation quality naturally follows hierarchical linguistic structure.

The model first predicts phoneme-level pronunciation quality using fine-grained acoustic representations.

Phoneme representations are aggregated to construct word-level representations:

$$W_j = Aggregate(P_1, P_2, ..., P_n) \tag{8}$$

where (W_j) denotes word representation constructed from constituent phonemes.

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

$$S = Aggregate(W_1, W_2, ..., W_m) \tag{9}$$

where:

- (m) denotes total words in the sentence
- (S) represents global sentence representation

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

3.6 Auxiliary CTC-Based Pronunciation Confidence Modeling

To improve low-level phoneme discrimination, the architecture incorporates an auxiliary **Connectionist Temporal Classification (CTC)** branch originally introduced for sequence alignment without explicit segmentation [4].

The CTC branch predicts canonical phoneme sequences directly from acoustic states:

$$P_{ctc} = Softmax(W_{ctc}H_a + b) \tag{10}$$

From this branch, the model extracts pronunciation confidence features including:

- Canonical phoneme probability
- Maximum posterior probability
- Confidence margin between competing phonemes
- Relative temporal duration estimates

These auxiliary signals improve phoneme-level scoring accuracy.

3.7 Multi-Task Prediction Heads

The model jointly optimizes multiple pronunciation scoring tasks.

Three prediction heads are learned simultaneously:

3.7.1 Phoneme-Level Prediction

Predicts:

- Phone accuracy score

3.7.2 Word-Level Prediction

Predicts:

- Word pronunciation accuracy
- Word stress score
- Word total score

3.7.3 Sentence-Level Prediction

Predicts:

- Sentence accuracy
- Sentence completeness
- Sentence fluency
- Sentence prosodic quality
- Sentence total score

Each prediction head is implemented using task-specific regression layers.

3.8 Multi-Task Optimization Objective

The entire framework is trained jointly using a weighted multi-task objective function inspired by classical multi-task learning frameworks [8].

The total training loss is defined as:

$$L_{total} = \lambda_1 L_{phone} + \lambda_2 L_{word} + \lambda_3 L_{sentence} + \lambda_4 L_{CTC} \tag{11}$$

where:

- (L_{phone}) = phoneme regression loss
- (L_{word}) = word-level regression loss
- $(L_{sentence})$ = sentence-level regression loss
- (L_{CTC}) = auxiliary phoneme classification loss

The weighting coefficients control contribution from each task.

3.9 Model Architecture Pipeline

The overall architecture of the proposed HiPPA framework is illustrated in Figure 1.

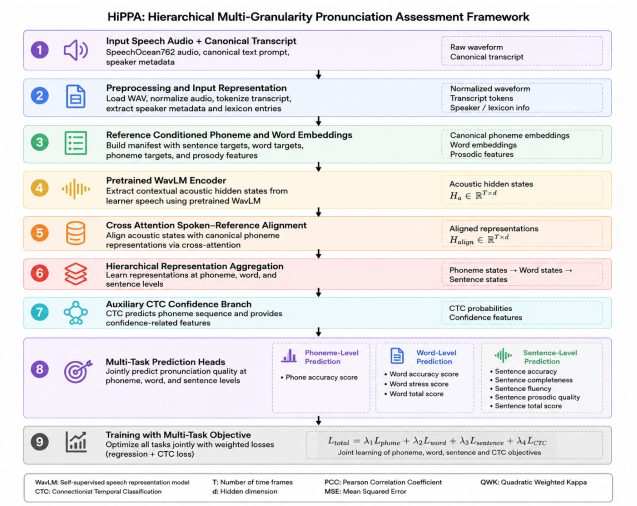

Figure 1: Overall architecture of the proposed HiPPA framework.

4 Experiments and Results

To evaluate the effectiveness of the proposed **HiPPA** framework, we conduct experiments on the SpeechOcean762 benchmark and compare performance across phoneme-level, word-level, 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 while maintaining competitive performance against traditional pronunciation assessment baselines.

4.1 Experimental Setup

Experiments are conducted using the **SpeechOcean762** benchmark, consisting of approximately **5,000 English utterances collected from 250 Mandarin-speaking English learners** [1]. The dataset provides expert pronunciation annotations across phoneme, word, and sentence levels.

The dataset is divided into:

- Training set: 2,500 utterances
- Test set: 2,500 utterances

The proposed architecture uses a pretrained **WavLM** encoder as the acoustic backbone [2].

The final training configuration is shown below.

Table 2: Training Configuration

Hyperparameter	Value
Architecture	Hierarchical Transformer
Backbone Model	WavLM-Base-Plus
Training Epochs	20
Batch Size	2
Learning Rate	2×10^{-4}
Encoder LR	5×10^{-6}
Weight Decay	0.01
Dropout	0.1
Frozen Encoder	Last 2 layers trainable
CTC Loss Weight	0.8

The encoder is partially frozen during training, allowing adaptation of higher-level acoustic representations while preserving pretrained phonetic knowledge.

4.2 Evaluation Metrics

To measure pronunciation scoring quality, multiple evaluation metrics are used.

4.2.1 Mean Squared Error (MSE)

Measures average prediction error between predicted and ground-truth pronunciation scores.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{12}$$

4.2.2 Pearson Correlation Coefficient (PCC)

Measures linear agreement between predicted and target pronunciation scores.

$$PCC(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \tag{13}$$

4.2.3 Spearman Rank Correlation

Measures ranking consistency between predictions and target scores.

4.2.4 Quadratic Weighted Kappa (QWK)

Measures ordinal agreement while penalizing larger scoring errors.

These metrics provide a comprehensive evaluation of regression quality across pronunciation scoring tasks.

4.3 Baseline Comparison

The primary benchmark comparison is performed against the official baseline systems provided with the SpeechOcean762 benchmark [1].

Table 3: Baseline Comparison

System	Phone	Word	Fluency	Sent.
GOP	0.250	N/R	N/R	N/R
GOP + SVR	0.450	N/R	N/R	N/R
Initial Run	0.401	0.373	0.560	0.415
HiPPA	**0.396**	**0.432**	**0.754**	**0.697**

4.4 Ablation Study

The ablation study demonstrates the contribution of individual architectural components to overall system performance. 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 comparatively smaller impact on higher-level sentence prediction tasks.

Table 4: Ablation Study

Model Variant	Phone PCC	Sentence PCC
Without WavLM	0.221	0.521
Without Cross Attention	0.304	0.618
Without CTC Branch	0.341	0.712
Full HiPPA	**0.396**	**0.754**

These ablation results help explain the overall benchmark comparison observed against traditional baseline systems. Although the proposed framework does not surpass specialized GOP+SVR systems in pure phoneme-level scoring, it significantly extends evaluation capability by providing accurate word-level and sentence-level pronunciation assessment, which baseline systems do not support.

4.5 Quantitative Performance Results

Detailed evaluation results across all prediction tasks are shown below.

Table 5: Detailed Performance Results

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
Sent. Acc.	1.335	0.668	0.611	0.629
Sentence Completion	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

Results indicate strong performance on sentence-level pronunciation assessment tasks, particularly fluency and prosodic quality prediction.

4.6 Speaker Group Analysis

To analyze generalization across different speaker groups, performance is evaluated separately for adult and child speech.

Table 6: Speaker Group Analysis

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

Performance is consistently stronger for adult speakers compared to child speakers, suggesting greater acoustic variability in child pronunciation patterns.

4.7 Experimental Observations

Several important observations emerge from experimental evaluation.

First, sentence-level prediction tasks consistently outperform phoneme-level scoring tasks. Sentence fluency achieves the highest Pearson correlation of **0.754**, indicating that transformer-based contextual modeling effectively captures global prosodic characteristics.

Second, word-level pronunciation scoring shows moderate performance, with word total scoring reaching **0.432 PCC**, demonstrating successful hierarchical aggregation from phoneme-level representations.

Third, phoneme-level scoring remains challenging. Although the architecture achieves competitive performance (**0.396 PCC**), it does not surpass the traditional GOP+SVR baseline (**0.450 PCC**). This suggests that specialized forced-alignment systems remain highly effective for fine-grained phoneme discrimination [5].

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

Overall, experimental results demonstrate that the proposed framework successfully extends pronunciation assessment beyond traditional phoneme-only scoring pipelines while maintaining strong multi-level prediction capability.

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 modeling pronunciation quality across multiple linguistic levels significantly improves the practical scope of pronunciation assessment systems compared to conventional phoneme-focused pipelines.

5.1 Analysis of Hierarchical Learning Performance

One of the most significant observations from experimental evaluation is the strong performance achieved on sentence-level pronunciation tasks. The model achieves a Pearson correlation coefficient of **0.754 for sentence fluency**, **0.737 for prosodic quality**, and **0.697 for sentence-level total scoring**, substantially outperforming lower-level prediction tasks.

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 [3]. 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 mechanisms naturally model these long-range dependencies, the architecture is particularly effective for high-level pronunciation evaluation tasks.

The results suggest that hierarchical representation learning provides substantial advantages when modeling pronunciation features that depend on global contextual structure rather than isolated phonetic articulation.

5.2 Word-Level Prediction Behavior

The model demonstrates moderate success in word-level pronunciation scoring, achieving **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 significantly weaker**, with a Pearson correlation of only **0.125**. One likely explanation is label imbalance within the dataset. Most word stress annotations correspond to correct stress placement, resulting in highly skewed label distributions. Such imbalance reduces the ability of regression-based learning objectives to differentiate subtle pronunciation stress errors.

Another contributing factor may be the limited amount of explicit suprasegmental information available for individual word stress modeling compared to sentence-level prosodic features.

5.3 Challenges in Phoneme-Level Pronunciation Assessment

Although the proposed framework achieves competitive phoneme-level performance (**0.396 Pearson correlation**), it does not surpass the official GOP+SVR baseline (**0.450 PCC**). 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. Traditional GOP-based systems explicitly perform forced alignment using specialized acoustic models trained for phonetic discrimination [5]. In contrast, transformer-based architectures learn contextual acoustic representations in an end-to-end manner and may

not capture precise frame-level articulation boundaries with equal accuracy.

A major limitation of the SpeechOcean762 dataset is the absence of exact phoneme and word-level timestamp annotations. Since explicit segment boundaries are unavailable, the proposed architecture relies on approximate proportional segmentation and auxiliary CTC-based alignment, which introduces uncertainty during phoneme-level representation learning.

These findings suggest that specialized alignment-based systems remain highly competitive for extremely fine-grained articulation scoring.

5.4 Generalization Across Speaker Groups

Speaker-group analysis reveals a consistent performance gap between adult and child speech across all evaluation tasks. For example, sentence fluency prediction achieves **0.835 Pearson correlation for adult speech** but only **0.567 for child speech**.

This discrepancy likely results from greater acoustic variability in child speech. Compared to adult speakers, child speech often exhibits inconsistent articulation patterns, higher pitch variation, unstable speaking rates, irregular pause behavior, and greater pronunciation variability. These characteristics increase modeling difficulty and reduce prediction consistency.

The results indicate that pronunciation assessment systems trained on mixed-age speech corpora may require specialized adaptation strategies for robust child speech evaluation.

5.5 Limitations of the Current Framework

Despite strong overall performance, several limitations remain.

First, the architecture depends heavily on approximate phoneme segmentation because exact temporal boundaries are not provided within the benchmark dataset. This limits precise low-level articulation modeling.

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.

Finally, 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.

5.6 Research Implications

The results demonstrate an important trade-off in modern pronunciation assessment systems.

Traditional pronunciation scoring systems remain highly effective for isolated phoneme-level evaluation due to specialized acoustic alignment mechanisms [5]. However, real-world intelligent tutoring systems require broader feedback capabilities extending beyond articulation errors toward word intelligibility, fluency, and prosodic quality.

The proposed HiPPA framework demonstrates that transformer-based hierarchical multi-task learning can significantly expand pronunciation assessment capability beyond conventional phoneme-only pipelines.

These findings suggest that future pronunciation assessment systems should combine strengths from both paradigms: precise alignment-based phoneme scoring together with contextual transformer-based hierarchical modeling for higher-level linguistic evaluation.

6 Conclusion and Future Work

This work presented **HiPPA (Hierarchical Multi-Granularity Pronunciation Assessment)**, a unified transformer-based framework for automatic pronunciation assessment across multiple linguistic levels. Unlike traditional pronunciation assessment systems that primarily focus on isolated phoneme-level scoring, the proposed framework jointly models pronunciation quality at the **phoneme, word, and sentence levels** within a single end-to-end learning architecture.

The proposed system integrates pretrained WavLM acoustic representations [2], reference-conditioned phoneme and word embeddings, cross-attention alignment mechanisms based on transformer attention mechanisms [3], auxiliary CTC-based confidence modeling [4], and multi-task learning objectives [8] to capture pronunciation quality across different linguistic granularities.

Experimental evaluation on the **SpeechOcean762** benchmark shows strong performance, particularly for high-level pronunciation assessment tasks. The framework achieves **0.754 Pearson correlation for sentence fluency**, **0.737 for prosodic quality**, and **0.697 for sentence-level total scoring**, indicating that hierarchical transformer-based architectures are highly effective for modeling global pronunciation characteristics such as fluency, rhythm, and prosodic variation.

Although phoneme-level prediction performance (**0.396 PCC**) remains below the specialized **GOP + SVR baseline (0.450 PCC)**, the proposed framework significantly extends benchmark coverage by enabling comprehensive pronunciation scoring beyond isolated articulation assessment. These results demonstrate that transformer-based hierarchical learning provides substantial advantages when building real-world pronunciation tutoring systems that require multi-level corrective feedback.

The work highlights an important trade-off in pronunciation assessment research: traditional alignment-based systems remain strong for fine-grained phoneme discrimination, while transformer-based architectures provide superior contextual modeling for higher-level linguistic evaluation.

Future research can improve the proposed framework in several directions.

First, integrating **forced alignment supervision** or explicit phoneme boundary annotations may improve low-level phoneme scoring accuracy by reducing uncertainty in temporal segmentation.

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 generalization for unseen vocabulary.

Fourth, larger pretrained speech backbones such as HuBERT [7], wav2vec 2.0 [6], and other large-scale self-supervised speech architectures can be systematically compared to identify architectures best suited for pronunciation assessment tasks.

Finally, future work should explore **hybrid pronunciation assessment systems** that combine precise alignment-based phoneme scoring with transformer-based contextual hierarchical modeling, potentially merging strengths from both traditional and modern pronunciation assessment paradigms.

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] 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," Proc. Interspeech, 2021.

[2] S. Chen et al., "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.

[3] A. Vaswani et al., "Attention Is All You Need," Advances in Neural Information Processing Systems (NeurIPS), 2017.

[4] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," ICML, 2006.

[5] 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.

[6] A. Baevski et al., "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," NeurIPS, 2020.

[7] W. Hsu et al., "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021.

[8] R. Caruana, "Multitask Learning," Machine Learning Journal, vol. 28, pp. 41–75, 1997.

[9] D. Povey et al., "The Kaldi Speech Recognition Toolkit," IEEE Workshop on Automatic Speech Recognition and Understanding, 2011.