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

一種有效的發音評估方法：結合分層 Transformer 與預訓練策略

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

顏碧呈 [1*] 、李俊廷 [1] 、王亦澄 [1] 、王信威 [1] 、羅天宏 [1]
、許永昌 [2] 、趙偉成 [3] 、陳柏霖 [1*]

[1] National Taiwan Normal University, [2] EZAI

[1] 國立台灣師範大學、 [2] EZAI

[3] Advanced Technology Laboratory, Chunghwa Telecom Co., Ltd.

[3] 中華電信股份有限公司先進技術實驗室

{bicheng, berlin}@ntnu.edu.tw, weicheng@cht.com.tw

Abstract 摘要

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

自動發音評估（APA）能透過在不同語言層級提供多面向的發音分數與細緻回饋，量化第二語言（L2）學習者在目標語言上的發音能力。大多數現有的 APA 方法通常將建模過程平行化，也就是同時預測各語言層級的多個面向分數。這不可避免地使語言單位的階層結構與各發音面向之間的相關性被忽視。為了解決此一限制，本文首先提出 HierTFR，一種層級化的 APA 方法，能在共同建模語句內在結構的同時考量發音面向之間的相關性。我們並提出一種考慮相關性的正則化項，以強化估計分數與人工標註之間的連結。此外，也針對不同語言層級提出新穎的預訓練策略，以利更好的模型初始化。 在 speechocean762 基準資料集上進行的大量實證實驗顯示，本方法相較於若干具競爭力的基準方法具有可行性與有效性。

1 Introduction 1 引言

With the rising trend of globalization, more and more people are willing or being demanded to learn foreign languages. This surging need calls for developing computer-assisted pronunciation training (CAPT) systems, as they can offer tailored and informative feedback for L2 (second-language)

隨著全球化趨勢上升，越來越多人願意或被要求學習外語。這樣激增的需求促使發展電腦輔助發音訓練（CAPT）系統，因為它們能為 L2（第二語言）提供客製化且具資訊性的回饋

Reading-aloud Scenario 朗讀情境							
	We call it bear. 我們叫牠 bear。						
Automatic Pronunciation Assessment Results 自動發音評估結果							
Utterance level 句子層級		Word level 單字層級				Phone level 音素層級	
Aspects 面向	Scores 分數	Words 單字	Aspects 面向	Scores 分數		Phones 音素	Scores 分數
Accuracy 準確性	1.6	We 我們	Accuracy 準確性	2		W IY0	2.0
			Stress 重音	2			
			Total 總分	2			2.0
Fluency 流暢度	1.8	Call 呼叫	Accuracy 準確性	2		K AO0 L	2.0
			Stress 重音	2			1.8
Completeness 完整性	2		Total 總計	2			1.8
		It 它	Accuracy 準確性	2		IH0 T	2.0
Prosody 韻律	1.8		Stress 壓力	2			2.0 2.0
			Total 總計	2		B	
Total 總計	1.6	Bear 熊	Accuracy 準確性	1.2			2.0 1.0 1.0
			Stress 重音	2			
			Total 總計	1.2			

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

表 1：圖 1：來自 speechocean762 資料集 (Zhang et al., 2021) 的一個執行範例說明了在朗讀情境下 APA 系統的評估流程，該範例可為 L2 學習者提供深入的發音回饋。

learners to practice pronunciation skills in a stressfree and self-directed learning manner (Eskenazi 2009; Evanini and Wang, 2013; Evanini et al., 2017; Rogerson-Revell, 2021). As a crucial ingredient of CAPT, automatic pronunciation assessment (APA) aims to evaluate the extent of L2 learners' oral proficiency and then provide fine-grained feedback on specific pronunciation aspects in response to a target language (Bannò et al., 2022; Chen and Li, 2016; Kheir et al., 2023). A de-facto standard for APA systems is typically instantiated with a "reading-aloud" scenario, where an L2 learner is presented with a text prompt and instructed to pronounce it correctly. To offer in-depth feedback on learners' pronunciation quality, recent efforts have drawn attention to the notion of multi-aspect and multi-granular pronunciation assessments, which normally devises a unified scoring model to

讓學習者以無壓力且自我導向的方式練習發音技能 (Eskenazi 2009; Evanini and Wang, 2013; Evanini et al., 2017; Rogerson-Revell, 2021)。作為 CAPT 的重要組成部分，自動發音評估 (APA) 旨在評估 L2 學習者口語能力的程度，並針對特定目標語言提供細緻的發音面向回饋 (Bannò et al., 2022; Chen and Li, 2016; Kheir et al., 2023)。APA 系統的事實標準通常以「朗讀」情境實作，學習者會看到一段文字提示並被要求正確朗讀。為了就學習者的發音品質提供深入的回饋，近來研究關注於多面向與多層次的發音評估概念，通常會設計一個統一的評分模型來

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

(Chao et al., 2022; Do et al., 2023a; Gong et al., 2022).

如圖 1 所示的示例所描繪，該方法共同在不同語言層級（即音素、詞與語句層級）上評估發音能力，涵蓋多種面向（例如準確度、流暢度與完整性）。此類研究的方法通常遵循平行建模範式，採用 Transformer 網路及其變體作為主幹架構，將音素層級的發音特徵序列作為輸入，並透過多任務學習機制同時預測跨各語言層級的多項面向分數（Chao et al., 2022; Do et al., 2023a; Gong et al., 2022）。

Albeit models stemming from the parallel modeling paradigm have demonstrated promising results on a few APA tasks, they still suffer from at least two weaknesses. First, the language hierarchy of an utterance is nearly sidelined, which, for example, assumes that all phones within a word are of equal importance and might insufficiently capture the word-level structural traits. Second, most of these methods largely overlook the relatedness among the pronunciation aspects. As an illustration, we visualize the correlation matrix in Figure 2, which shows the Pearson Correlation Coefficients (PCCs) between any pair of expert annotated aspect scores on the training set. We can observe that except for the aspects of utterancecompleteness and word-stress, the rest pronunciation aspects exhibit strong correlations not only within the same linguistic level but also across different linguistic levels [2]. Building on these observations, we in this paper present a novel language hierarchy-aware APA model, dubbed HierTFR, which leverages a hierarchical Transformer-based architecture to jointly model the intrinsic multi-level linguistic structures of an utterance while considering relatedness among aspects within and across different linguistic levels. To explicitly capture the relatedness within and across different linguistic levels, an aspect attention mechanism and a selective fusion module are introduced. The proposed model is further optimized with an effective correlation-aware regularizer, which encourages the correlations of predicted aspect scores to match those of their counterparts provided by human annotations. Furthermore, distinct pre-training strategies tailored for three linguistic levels are put forward,

儘管源自平行建模範式的模型在少數自動發音評估（APA）任務上已展現出可喜的成果，但它們仍存在至少兩個弱點。首先，語句的語言階層幾乎被忽略，例如這類方法假設一個單字內的所有音素同等重要，可能不足以捕捉單字層級的結構特徵。其次，這些方法大多忽視發音各面向之間的相互關聯。舉例來說，我們在圖 2 中視覺化了相關矩陣，該矩陣顯示訓練集上任意一對專家註記面向分數之間的皮爾森相關係數（PCCs）。可以觀察到，除了語句完整度（utterancecompleteness）與單字重音（word-stress）這兩個面向外，其餘的發音面向不僅在相同語言層級內呈現強相關，也跨不同語言層級顯示出強相關 [2] 。 在上述觀察的基礎上，本文提出一種新穎的語言層級感知自動發音評估（APA）模型，稱為 HierTFR。該模型利用分層的 Transformer 架構，能夠在聯合建模語句的固有多層語言結構的同時，考量不同語言層級內與跨層級的面向相關性。為了明確捕捉不同語言層級內與跨層級的相關性，我們引入了面向注意力機制與選擇性融合模組。所提出的模型進一步以一個有效的關聯感知正則化項進行優化，該正則化項促使預測的面向分數之間的相關性與人類標註所提供的對應關聯性相匹配。此外，還提出了針對三個語言層級量身設計的不同預訓練策略，

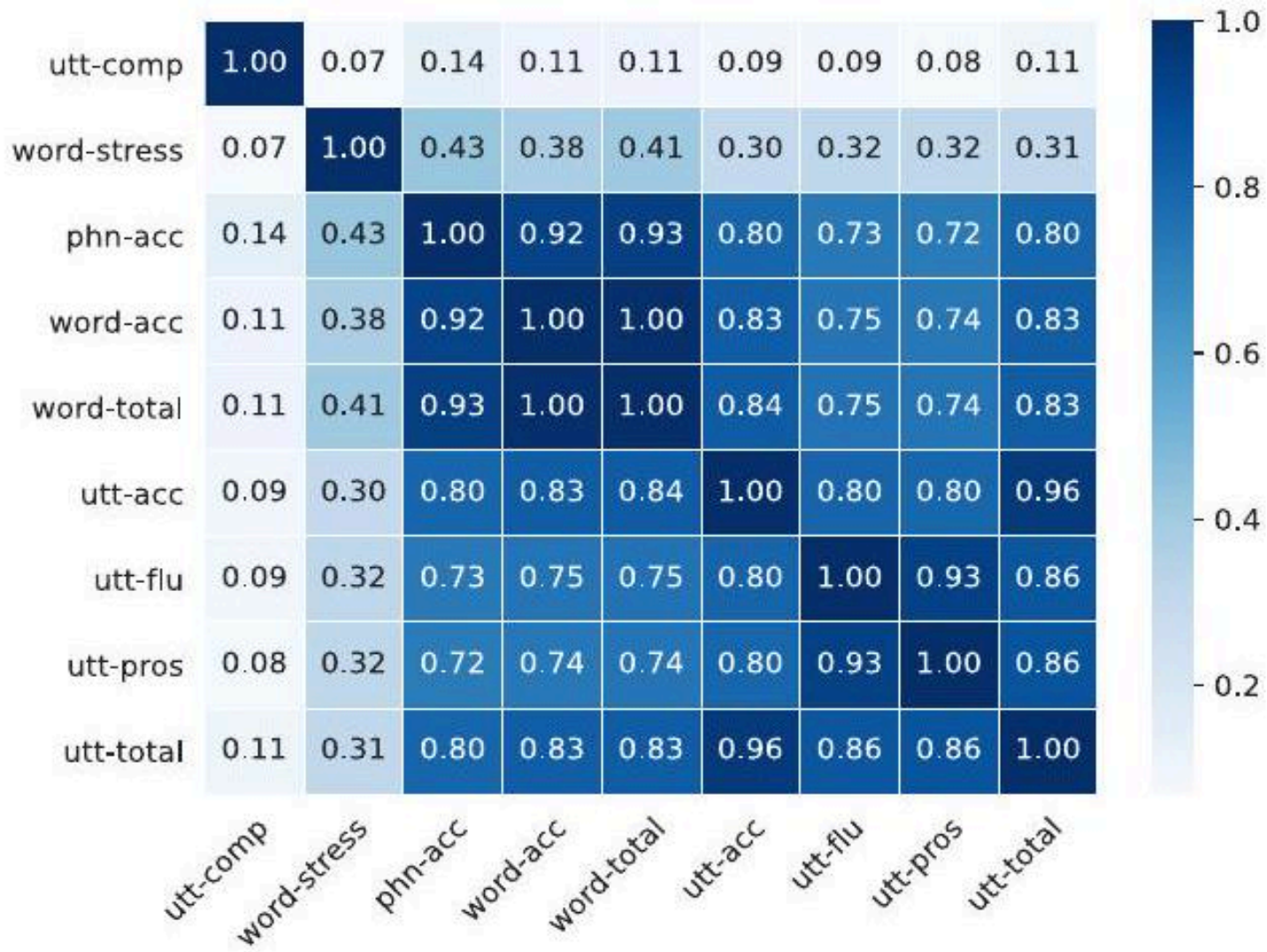

Figure 1: Figure 2: A correlation matrix derived from the expert annotations of the training set. Each element in the matrix corresponds to the PCC score of a pair of measured aspects.

圖 1：圖 2：從訓練集專家標註推導出的關聯矩陣。矩陣中的每個元素對應一對被測面向的皮爾森相關係數（PCC）分數。

so as to boost model initialization and hence reduce the reliance on large amounts of supervised training data. A comprehensive set of experimental results reveal that the proposed model achieves significant and consistent improvements over several strong baselines on the speechocean762 benchmark dataset (Zhang et al., 2021).

以增強模型初始化，從而減少對大量有監督訓練資料的依賴。一系列完整的實驗結果顯示，所提出的模型在 speechocean762 基準資料集 (Zhang et al., 2021) 上，相較於多項強大基線方法，達成了顯著且一致的改進。

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

綜合來說，我們工作的主要貢獻至少有三點：(1) 我們提出 HierTFR，一種用於自動發音評估（APA）的階層式神經模型，旨在階層化表示第二語言學習者的輸入語句，並有效捕捉不同語言層級內部及跨層級的關聯性；(2) 我們為模型訓練提出一種關聯感知正則化器，鼓勵預測分數考量各不同面向之間的關聯性；以及 (3) 在公開 APA 資料集上進行的大量實驗驗證了我們所提出的預訓練策略的有效性，這些策略大幅提升了各語言層級評估的成效。

2 Methodology

2.1 Problem Formulation

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

給定一個輸入語句 U ，由 L2 學習者發出的時間序列音訊信號 X 組成，以及一個含有 M 個單字與 N 個音素的參考文本提示 T，APA 模型被訓練以估計不同語言粒度下多個發音面向的能力分數。令 $\mathrm{G} = \{p, w, u\}$ 為一組語言粒度，其中 p, w, u 代表音素、單字與語句層級的語言單位，

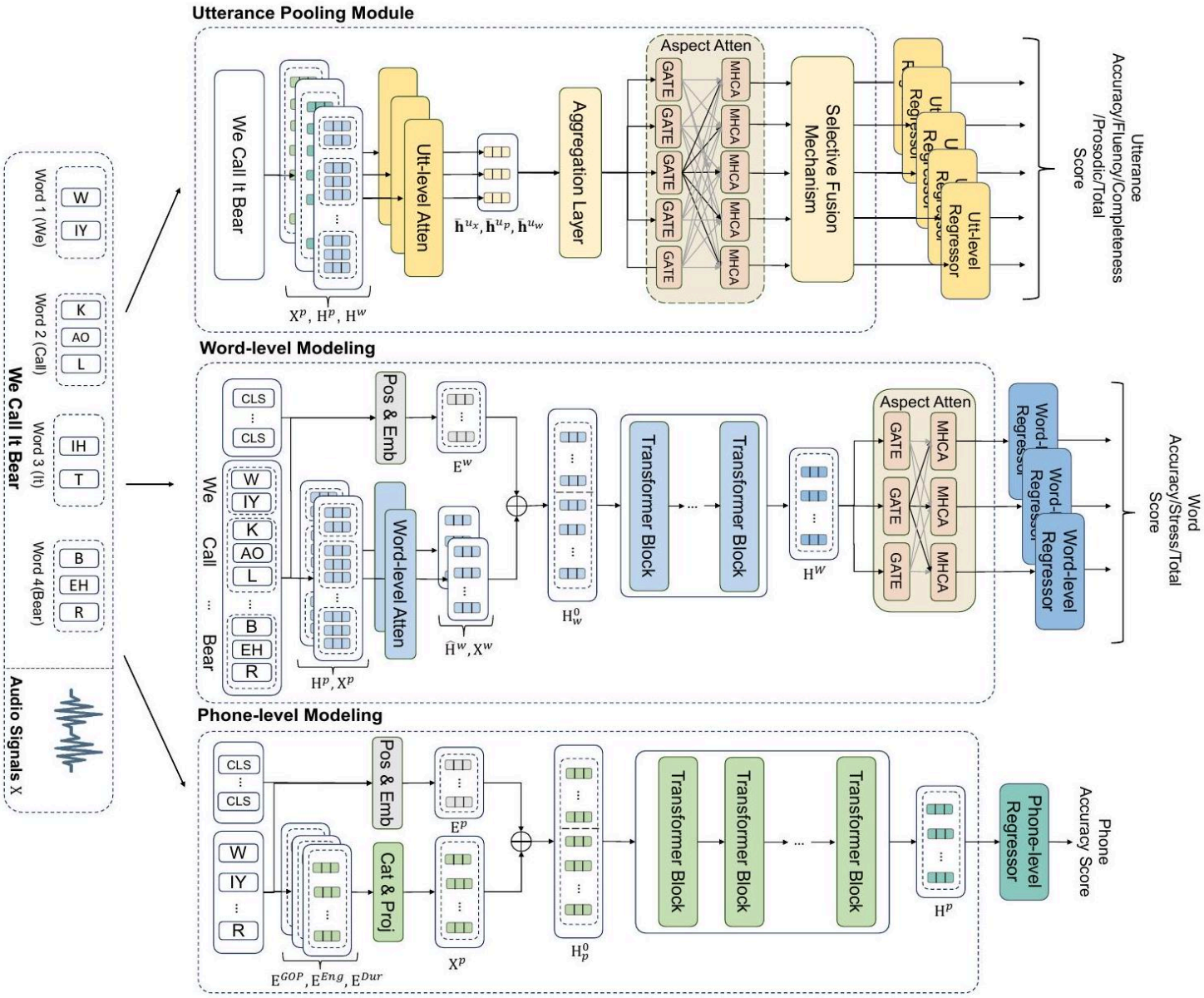

Figure 2: Figure 3: An architecture overview of the proposed model, which consists of a phone-level modeling component, a word-level modeling component, and an utterance pooling module.

圖 2：圖 3：所提出模型的架構概覽，包含一個音素層建模元件、一個單字層建模元件，以及一個語句池化模組。

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

分別。對於每個語言單位 $g \in G$ ，APA 模型學習預測一組面向分數 $\mathrm{A}^g = \left\{a_1^g, a_2^g, \ldots, a_{N_g}^g\right\}$ ，其中 N_g 為語言單位 g 的發音面向數量。

2.2 Hierarchical Interactive Transformer Architecture

2.2 分層互動 Transformer 架構

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

如圖 3 所示，本研究所提出的 APA 模型整體架構示意圖包含三個構成要素：音素層級建模、字詞層級建模與語句池化模組。在取得各種發音面向的表示後，多層全連接神經層作為回歸器，共同為輸入語句產生相對應的面向分數序列。

Phone-level Modeling. For an input utterance U, various pronunciation features are extracted to portray the L2 learner's pronunciation quality, which includes the goodness of pronunciation

音素層級建模。對於輸入語句 U，會擷取各種發音特徵以刻畫第二語言學習者的發音品質，這些特徵包含發音準確度（goodness of pronunciation）

(GOP)-based features E^{GOP}, as well as prosodic features composed of duration $\mathrm{E}^{\mathrm{Dur}}$ and energy $\mathrm{E}^{\mathrm{Eng}}$ statistics (Witt and Young, 2000; Hu et al., 2015; Zhu et al., 2022; Shen et al., 2021) [3]. All these features are then concatenated and subsequently projected to from a sequence of acoustic features X^{p}. In the meantime, the phone-level text prompt is mapped into an embedding sequence E^{p} via a phone and position embedding layer and then point-wisely added to X^{p} for enriching the phonetic information of X^{p}. The resulting representations H_{p}^{0} are prepend with five trainable "[CLS]" embeddings and in turn fed into a phonelevel transformer to obtain the contextualized representations H^{p} (Vaswani et al., 2017):

基於 (GOP) 的特徵 E^{GOP} ，以及由時長 $\mathrm{E}^{\mathrm{Dur}}$ 和能量 $\mathrm{E}^{\mathrm{Eng}}$ 統計組成的韻律特徵（Witt and Young, 2000；Hu et al., 2015；Zhu et al., 2022；Shen et al., 2021） [3] 。所有這些特徵隨後被串接並投影，形成一個聲學特徵序列 X^{p} 。同時，音素層級的文字提示經由音素與位置嵌入層映射為嵌入序列 E^{p} ，然後逐點加到 X^{p} 上，以豐富 X^{p} 的語音資訊。所得的表示 H_{p}^{0} 前置五個可訓練的 "[CLS]" 嵌入，接著送入音素層級的 Transformer 以獲得情境化表示 H^{p} （Vaswani et al., 2017）：

$$
\begin{aligned}
\mathrm{X}^{p} &= \mathrm{W} \cdot \left[\mathrm{E}^{GOP}; \mathrm{E}^{Dur}; \mathrm{E}^{Eng}\right] + \mathbf{b} \\
\mathrm{H}_{p}^{0} &= \mathrm{X}^{p} + \mathrm{E}^{p} \\
\mathrm{H}^{p} &= \mathrm{Transformer}_{\mathrm{phn}}\left(\mathrm{H}_{p}^{0}\right)
\end{aligned}
$$

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

其中 W 與 $\mathbf{b}$ 為可學習參數。為了評估一序列音素層級的分數， H^{p} （不包含最前面的 5 個嵌入）被前向傳播至相應的迴歸器。被排除的嵌入 $\mathrm{H}_{1:5}^{p}$ 預期傳遞整體的發音資訊，並進一步輸入後續的選擇性融合機制，用於語句層級的評估。

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

詞級建模。對於詞級評估，使用詞級注意力池化從其對應的音素產生詞表示向量，該操作可實作為一個多頭注意力層，後接一個平均化運算。詞級輸入表示 H_{w}^{0} 可透過對音素層表示 X^{p} 和 H^{p} 個別應用詞級注意力得到，之後再與詞級文字嵌入 E^{w} 進行線性組合。接著，將 H_{w}^{0} 在前端加上五個可訓練的 "[CLS]" 嵌入，並輸入至 transformer，以計算詞級的語境化表示 H^{w} ：

$$
\begin{aligned}
\mathrm{X}^{w} &= \mathrm{Atten}_{\mathrm{word}}\left(\mathrm{X}^{p}\right) \\
\widehat{\mathrm{H}}^{w} &= \mathrm{Atten}_{\mathrm{word}}\left(\mathrm{H}^{p}\right) \\
\mathrm{H}_{w}^{0} &= \mathrm{X}^{w} + \widehat{\mathrm{H}}^{w} + \mathrm{E}^{w} \\
\mathrm{H}^{w} &= \mathrm{Transformer}_{\mathrm{word}}\left(\mathrm{H}_{w}^{0}\right)
\end{aligned}
$$

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

此處注意 H^{w} （不含前五個嵌入）被用於詞級評估，而被排除的嵌入 $\mathrm{H}_{1:5}^{W}$ 則在後續的選擇性融合機制中被用於句子級評估。

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

接著，引入了一種面向注意力機制以捕捉不同面向間的相關性 (Do et al., 2023b; Ridley et al., 2021)。該機制由兩個子層組成：自我門控層（self-gating layer）與多頭交叉注意力層（multi-head cross-attention layer）。具體而言，對於第 j 個詞層面向，關係感知表示 $\widehat{\mathrm{H}}^{wr_j}$ 先經由自我門控層從 H^{w} 推得，該層旨在在考量從其他面向所蒐集資訊的同時，抽除冗餘資訊。此外，採用多頭交叉注意力（MHCA）過程以及遮罩策略，從所有關係感知面向表示的集合 $\mathrm{C}^{ra} = \left[\widehat{\mathrm{H}}^{wr_1}, \ldots, \widehat{\mathrm{H}}^{wr_{N_w}}\right]$ 中計算出面向表示 H^{w_j} 。下列方程式說明了面向注意力的運算：

$$
\begin{aligned}
\widehat{\mathrm{H}}^{w_j} &= \mathrm{W}_j \cdot \mathrm{H}^{w} + \mathbf{b}_j \\
\widehat{\mathrm{H}}^{wr_j} &= \sigma\left(\mathrm{W}_{g_j} \cdot \mathrm{C}^{w} + \mathbf{b}_{g_j}\right) \otimes \widehat{\mathrm{H}}^{w_j} \\
\mathrm{H}^{w_j} &= \mathrm{MHCA}\left(\widehat{\mathrm{H}}^{wr_j}, \mathrm{C}^{ra}\right)
\end{aligned}
$$

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

其中 $\widehat{\mathrm{H}}^{w_j}$ 為面向特定層面的表示，而 $\mathrm{C}^{w} = \left[\widehat{\mathrm{H}}^{w_1}, \ldots, \widehat{\mathrm{H}}^{w_{N_w}}\right]$ 包含所有面向特定層面的表示。在 MHCA 中， $\widehat{\mathrm{H}}^{wr_j}$ 被線性投影作為查詢矩陣，而 C^{ra} 被線性投影以形成鍵和值矩陣。此外，遮罩策略確保特定位置的輸出表示僅受該單字單位其他層面的影響。最後，面向特定層面的表示 H^{w_j} 作為對應迴歸器的輸入，以預測第 j 個單字層級發音層面的分數序列。

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

語句層匯聚模組。對於語句層評估，引入語句層注意力匯聚（utterance-level attention pooling）以從相應的輸入表示中生成語句層的整體表示，該方法可透過注意力匯聚有效實現（Peng et al., 2022）。更詳細地，語句層表示 $\mathbf{h}^{u}$ 可通過分別將向量序列 $\mathrm{X}^{p}, \mathrm{H}^{p}$ 與 H^{w} 輸入到語句層注意力匯聚模組，然後進行彙整操作來獲得：

$$
\begin{aligned}
\overline{\mathbf{h}}^{u_x} &= \mathrm{Atten}_{\mathrm{utt}}\left(\mathrm{X}^{p}\right) \\
\overline{\mathbf{h}}^{u_p} &= \mathrm{Atten}_{\mathrm{utt}}\left(\mathrm{H}^{p}\right) \\
\overline{\mathbf{h}}^{u_w} &= \mathrm{Atten}_{\mathrm{utt}}\left(\mathrm{H}^{w}\right) \\
\mathbf{h}^{u} &= \mathrm{W}_u\left(\overline{\mathbf{h}}^{u_x} + \overline{\mathbf{h}}^{u_p} + \overline{\mathbf{h}}^{u_w}\right) + \mathbf{b}_u
\end{aligned}
$$

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

其中 $W_u, \mathbf{b}_u$ 為可訓練參數。

Next, a selective fusion mechanism is proposed to integrate contextualized representations across multiple linguistic levels for the utterance-level pronunciation assessments (Xu et al., 2021). Specifically, for the estimation of j-th

utterancelevel aspect score, an aspect attention operation is first performed on $\mathbf{h}^u$ to a produce intermediate representation $\hat{\mathbf{h}}^{u_j}$. Note also that the gate values for the phone $\left(g_p^{u_j}\right)$, word $\left(g_w^{u_j}\right)$ and utterance $\left(g_u^{u_j}\right)$ granularities are used to control the extent to which these contextualized representations can flow into the fused representation $\mathbf{h}^{u_j}$:

接著，提出一種選擇性融合機制以整合跨多種語言層級的情境化表示，用於句子層級的發音評估 (Xu et al., 2021)。具體來說，為了估計第 j 個句子層級面向分數，先對 $\mathbf{h}^u$ 執行一個面向注意力操作以產生中間表示 $\hat{\mathbf{h}}^{u_j}$ 。同時注意，手機 $\left(g_p^{u_j}\right)$ 、詞 $\left(g_w^{u_j}\right)$ 與句子 $\left(g_u^{u_j}\right)$ 粒度的閘值用來控制這些情境化表示流入融合表示 $\mathbf{h}^{u_j}$ 的程度：

$$
\begin{aligned}
g_p^{u_j} &= \sigma\left(\mathbf{w}_{p_j} \cdot \left[\mathbf{h}_j^p ; \mathbf{h}_j^w ; \hat{\mathbf{h}}^{u_j}\right] + b_{p_j}\right) \\
g_w^{u_j} &= \sigma\left(\mathbf{w}_{w_j} \cdot \left[\mathbf{h}_j^p ; \mathbf{h}_j^w ; \hat{\mathbf{h}}^{u_j}\right] + b_{w_j}\right) \\
g_u^{u_j} &= \sigma\left(\mathbf{w}_{u_j} \cdot \left[\mathbf{h}_j^p ; \mathbf{h}_j^w ; \hat{\mathbf{h}}^{u_j}\right] + b_{u_j}\right)
\end{aligned}
$$

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

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

其中 $\mathbf{h}_j^p$ 與 $\mathbf{h}_j^w$ 為 j 個表示向量，分別來自 H^p 與 H^w ；而 $\mathbf{w}_{p_j}, \mathbf{w}_{w_j}, \mathbf{w}_{u_j}, b_{p_j}, b_{w_j}$ 與 b_{u_j} 為可訓練參數。融合後的表示 $\mathbf{h}^{u_j}$ 隨後被送入對應的回歸器以評估給定句子層級面向的熟練度分數。

2.3 Optimization 2.3 最適化

Automatic Pronunciation Assessment Loss.

自動發音評估損失。

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

多面向、多粒度發音評估的損失 $\mathcal{L}_{APA}$ 被計算為對應於不同語言層級的均方誤差 (MSE) 損失之加權總和。

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

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

其中 $\mathcal{L}_{p^{j_p}}, \mathcal{L}_{w^{j_w}}$ 、 $\mathcal{L}_{u^{j_u}}$ 、 λ_p, λ_w 分別為針對不同面向在音素層、詞彙層與句子層的損失。參數 λ_u 、 N_p, N_w 為可調整參數，用以控制不同粒度的影響力，而 N_u 、{6} 則標示音素層、詞彙層與句子層的面向數量。

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

相關性感知正則化損失。相關性感知正則化損失定義為預測面向分數 $\widehat{\Sigma}$ 的相關矩陣與對應目標標籤 $\sum$ 的相關矩陣之差：

$$
\mathcal{L}_{cor} = \ell(\widehat{\Sigma}, \Sigma),
$$

where ℓ is the regularization loss function, and each element in $\widehat{\sum}_{ij}$ (or $\sum_{ij}$) is defined as a Pearson correlation coefficient between i-th aspect score and j-th aspect score [4]. We adopt the MSE criterion for computing ℓ; the overall

loss thus can be expressed by:

其中 ℓ 為正則化損失函數，而 $\widehat{\sum}_{ij}$ （或 $\sum_{ij}$ ）中的每個元素被定義為第 i 個面向分數與第 j 個面向分數之皮爾森相關係數 [4] 。我們採用 MSE 準則來計算 ℓ ；因此整體損失可表示為：

$$\mathcal{L} = \mathcal{L}_{APA} + \lambda \mathcal{L}_{cor},$$

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

其中 $\lambda \in [0, 1]$ 是一個可調參數，根據開發集的實驗結果將其設定為 0.01。

2.4 Pre-training Strategies

2.4 預訓練策略

It is without doubt that a proper initialization is vital for the estimation of a neural model, due mainly to the highly nonconvex nature of the training loss function (Tamborrino et al., 2020;

毫無疑問，適當的初始化對於神經模型的估計至關重要，這主要是由於訓練損失函數高度非凸的性質（Tamborrino et al., 2020;

Lakhotia et al., 2021). At lower linguistic levels, we leverage the mask-predict objective (Ghazvininejad et al., 2019) in the pre-training stage. To this end, we first mask a portion of input text prompt at phone- and word-levels. The corresponding Transformers are then tasked on recovering the masked tokens conditioning on the unmasked prompt sequence and the associated pronunciation representations (i.e., H_p^0, and H_w^0). For the utterance level, we base the proposed pretraining strategy on predicting the relatively high or low accuracy scores for a pair of utterances. Namely, given any two utterances, the objective is to predict whether the former has a higher, lower, or the same accuracy score as the latter. Note here that, utterance pairs are randomly selected from a training batch, and this mechanism is employed to pretrain their utterance-level representations, denoted as $\mathbf{h}^u_{\text{out}\,t_1}$, and $\mathbf{h}^u_{\text{out}\,t_2}$. Next, we feed the concatenation of these vector representations $\mathbf{h}^u_{\text{out}} = \left[\mathbf{h}^u_{\text{out}\,1}; \mathbf{h}^u_{\text{out}\,2}\right]$ into a three-way classifier, using the cross-entropy loss as the training objective.

Lakhotia 等人，2021）。在較低的語言層級，我們在預訓練階段採用 mask-predict 目標（Ghazvininejad 等人，2019）。為此，我們首先在音素與詞級別遮蔽部分輸入文字提示。相應的 Transformer 模型隨後被要求在已解鎖的提示序列與相關發音表示（即 H_p^0 與 H_w^0 ）的條件下，還原被遮蔽的標記。至於句子層級，我們將所提出的預訓練策略建立在預測一對語句相對較高或較低的正確度分數上。也就是說，給定任兩個語句，目標是預測前者的正確度分數是否比後者高、低或相同。此處需注意，語句對是從訓練批次中隨機選取，該機制用於預訓練它們的句子層級表示，分別記為 $\mathbf{h}^u_{\text{out}\,t_1}$ 與 $\mathbf{h}^u_{\text{out}\,t_2}$ 。接著，我們將這些向量表示的串接結果 $\mathbf{h}^u_{\text{out}} = \left[\mathbf{h}^u_{\text{out}\,1}; \mathbf{h}^u_{\text{out}\,2}\right]$ 輸入到一個三分類器，並以交叉熵損失作為訓練目標。

3 Experimental Settings

3 實驗設定

3.1 Evaluation Dataset and Metrics

3.1 評估資料集與指標

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

in order to evaluate the

我們在 speechocean762 資料集上進行了自動發音評估（APA）實驗。該資料集為公開可得的開源資料集，專為 APA 研究設計（Zhang et al., 2021）。此資料集包含由 250 名以中文為母語的第二語言學習者所錄製的 5,000 則英語語音。訓練集與測試集大小相同，兩者各有 2,500 則語句，且在多種語言層級上針對不同發音面向進行了發音能力評分。每一個分數皆由五位經驗豐富的專家根據相同評分規準獨立評定，最終分數則以五個分數的中位數作為結果。評估指標包括皮爾森相關係數（PCC）與均方誤差（MSE）。PCC 為主要評估指標，用以量化預測分數與實際分數之間的線性相關程度。PCC 值越高代表模型預測與人工標註之間的相關性越強。在下列實驗中，我們同時報告 MSE 值以評估

Models 模型	Phone Score 音素分數		Word Score (PCC) 詞彙分數 (PCC)			Utterance Score (PCC) 語句分數 (PCC)				
	MSE ↓	PCC ↑	Accuracy ↑ 準確率 ↑	Stress ↑ 重音 ↑	Total ↑ 總計 ↑	Accuracy 準確性	Completeness 完整度	Fluency ↑ 流暢度 ↑	Prosody ↑ 韻律 ↑	Total ↑ 總計 ↑
Lin2021	-	-	-	-	-	-	-	-	-	0.720
Kim2022	-	-	-	-	-	-	-	0.780	0.770	-
Ruy2023	-	-	-	-	-	0.719	-	0.775	0.773	0.743
LSTM	0.089	0.591	0.514	0.294	0.531	0.720	0.076	0.745	0.747	0.741
	±0.000	±0.003	±0.003	±0.012	±0.004	±0.002	±0.086	±0.002	±0.005	±0.002
GOPT	0.085	0.612	0.533	0.291	0.549	0.714	0.155	0.753	0.760	0.742
	±0.001	±0.003	±0.004	±0.030	±0.002	±0.004	±0.039	±0.008	±0.006	±0.005
HiPAMA	0.084	0.616	0.575	0.320	0.591	0.730	0.276	0.749	0.751	0.754
	±0.001	±0.004	±0.004	±0.021	±0.004	±0.002	±0.177	±0.001	±0.002	±0.002
HierTFR	0.081	0.644	0.622	0.325	0.634	0.735	0.513	0.801	0.795	0.764
	±0.000	±0.000	±0.002	±0.022	±0.002	±0.008	±0.204	±0.004	±0.002	±0.002

Table 2: Table 1: The performance evaluations of our model and all compared methods on speechocean762 test set.

表 2：表 1：我們的模型與所有比較方法在 speechocean762 測試集上的效能評估。

phoneme-level APA accuracy in comparison with prior arts.

與先前技術相比的音素層級 APA 準確性。

3.2 Implementation Details

3.2 實作細節

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

在音素層級能量與時長統計量的輸入特徵擷取方面，我們遵循 Zhu et al. (2022) 與 Shen et al. (2021) 建議的處理流程，根據強制對齊將時間聚合的逐幀特徵構造成音素層級特徵。用於語境表示建模的音素與詞層級 Transformer 各由 3 個處理區塊組成，分別使用具有 3 個 head 與 24 個隱藏單元的多頭注意力。此外，詞與語句層級的注意力池化使用單層多頭注意力，具有 3 個 head 與 24 個隱藏單元。用於 Eq. (19) 中 APA 損失 ($\lambda_p, \lambda_w, \lambda_u$) 的組合權重分別設為 ($1, 1, 1$)。為確保實驗結果的可靠性，我們重複進行 5 次獨立試驗，每次試驗包含以不同隨機種子訓練的 100 個 epoch。測試集結果以在開發集上根據 PCC 分數選出的前 100 名最佳模型所達成結果的平均值呈報。

3.3 Compared Methods 3.3 比較方法

We compare our proposed model (viz. HierTFR) with several families of top-of-the-line methods. Lin et al. (2021) and Kim et al. (2022) are singleaspect assessment models. The former develops a bottom-up hierarchical scorer evaluating the accuracy scores at the utterance level. The latter leverages self-supervised features (Baevski et al., 2020) to describe the learner's pronunciation traits

我們將所提模型（即 HierTFR）與數種頂尖方法家族進行比較。Lin et al. (2021) 與 Kim et al. (2022) 屬於單一面向評估模型。前者提出一個自下而上的分層評分器，在語句層級評估準確度分數；後者則利用自監督特徵 (Baevski et al., 2020) 來描述學習者的發音特徵。

and then separately models the corresponding utterance-level aspects with recurrent neural models. In addition, LSTM, GOPT (Gong et al., 2022; Ruy et al. 2023), and HiPAMA (Do et al., 2023b) are multi-aspect and multi-granular pronunciation assessments. First, LSTM and GOPT follow a parallel modeling regime, both of which treat the phone-level input features as a flattened sequence and assess higher level pronunciation scores through stacking LSTM layers or Transformer blocks. Second, Ruy et al. (2023) introduces a unified model architecture that jointly optimizes phone recognition and APA tasks. Lastly, HiPAMA is a hierarchical APA model that more resembles our model than the other methods compared in this paper. Different from our method, HiPAMA extracts high-level pronunciation features from low-level features based on a simple average pooling mechanism. Furthermore, the aspect attention mechanism used in HiPAMA performs on the logistics, whereas our model operates on the intermediate representations.

然後分別以遞歸神經模型對應地建模發話層級的相關面向。此外，LSTM、GOPT (Gong et al., 2022；Ruy et al. 2023) 與 HiPAMA (Do et al., 2023b) 都屬於多面向與多粒度的發音評估方法。首先，LSTM 與 GOPT 採用平行建模模式，兩者皆將音素層級的輸入特徵視為展平的序列，並透過堆疊 LSTM 層或 Transformer 區塊來評估更高層級的發音分數。其次，Ruy et al. (2023) 提出一個統一的模型架構，可共同優化音素識別與自動發音評估（APA）任務。最後，HiPAMA 是一個分層的 APA 模型，其結構與本文提出的模型較為相似。與我們的方法不同，HiPAMA 從低階特徵提取高階發音特徵時，基於簡單的平均池化機制。此外，HiPAMA 使用的面向注意力機制是在後勤層（logistics）上運作，而我們的模型則是在中間表示（intermediate representations）上運作。

4 Experimental Results 4 實驗結果

4.1 Main Results 4.1 主要結果

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

表 1 報告了在 speechocean762 資料集上的結果，該資料集分為三個部分：第一部分顯示單一面向評估模型的結果，第二部分呈現多面向與多層級發音方法的結果，第三部分則報告我們模型的結果。我們在第三部分進一步與另一個分層 APA 模型（即 HiPAMA）進行比較。

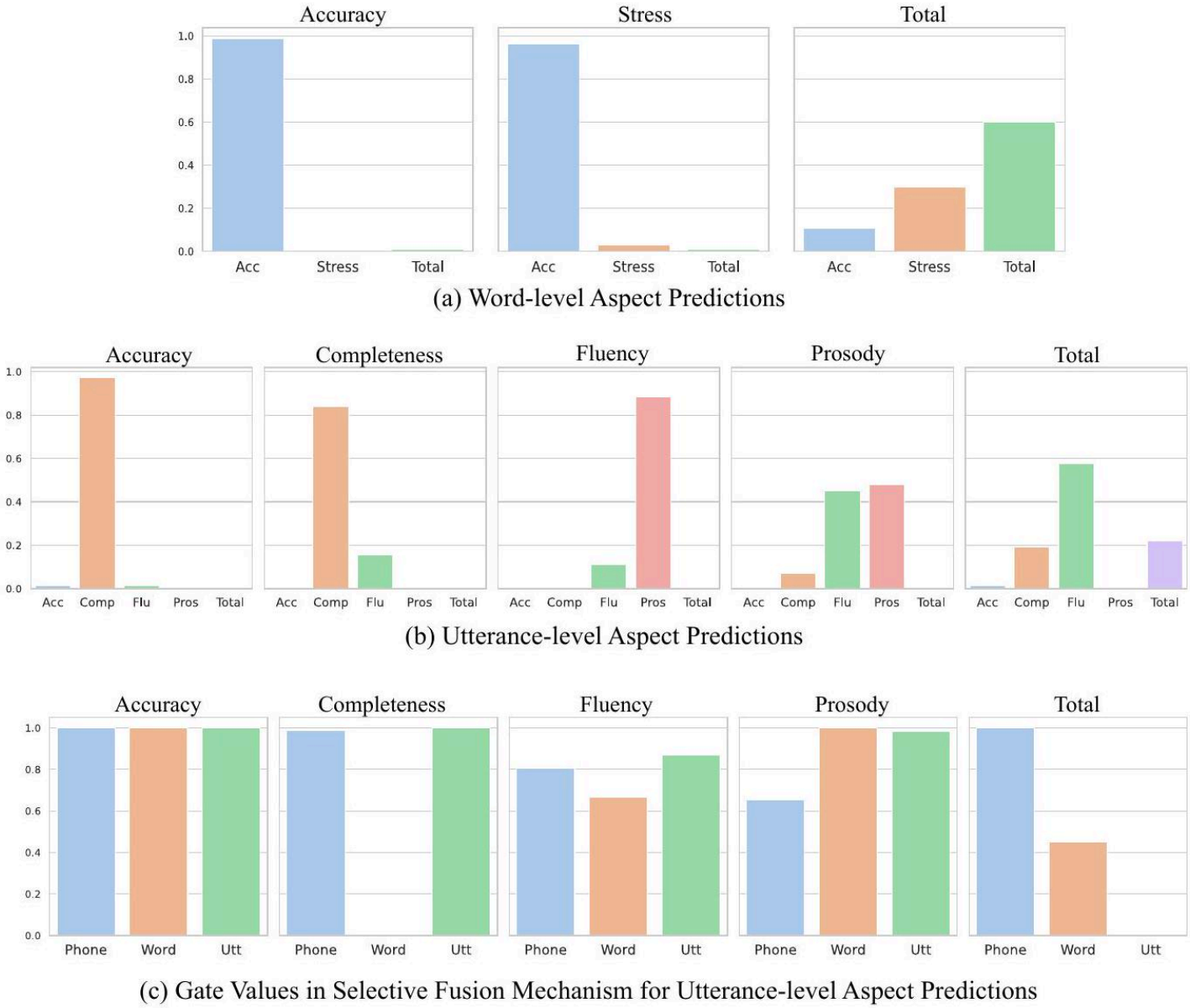

Figure 3: Figure 4: Qualitative visualization of model parameters when predicting each aspect score. We show (a) the averaged attention values for word-level aspects, (b) the averaged attention weights for utterance-level aspects, and (c) the averaged gate values for three linguistic levels.

圖 3：圖 4：在預測各面向分數時對模型參數的定性視覺化。我們展示 (a) 詞級面向的平均注意力值、(b) 句子級面向的平均注意力權重，以及 (c) 三個語言層級的平均閘值。

First, a general observation is that our approach, HierTFR, excels in all assessment tasks, especially at the linguistic levels of utterance and word. This performance gain confirms that our model works comparably better for capturing the relationships between linguistic units than the other competitive methods. In terms of the utterance-level total score, the single-aspect assessment method (viz. Lin2021) largely falls behind the other multi-aspect and multi-granular pronunciation assessment models, which we attribute to the fact that the single-aspect assessment method is unable to harness the dependency relationships between aspects through the multi-task learning paradigm. By leveraging self-supervised learning features, Kim2022 achieves significant improvements over most APA methods in terms of the utterance-level assessments. Next, we scrutinize the performance of multi-aspect and multi-granular pronunciation assessment methods. Ruy2023 demonstrates significant advancements in the utterance-level fluency and prosody assessments due probably to the joint training of the APA model on the phone

首先，一般性的觀察是，我們的方法 HierTFR 在所有評量任務上表現出色，特別是在句子層級與詞語層級的語言學評量上。這項效能提升證實我們的模型在捕捉語言單位之間關係方面，表現明顯優於其他具有競爭力的方法。就句子層級的總分而言，單一面向評量方法（即 Lin2021）大幅落後於其他多面向且多粒度的發音評量模型，我們認為這是因為單一面向評量方法無法透過多任務學習範式來利用面向之間的相依關係。透過利用自我監督學習特徵，Kim2022 在句子層級評量上相較大多數 APA 方法取得了顯著的改進。接著，我們檢視多面向與多粒度發音評量方法的表現。Ruy2023 在句子層級的流暢度與韻律評量上展現了明顯進步，這很可能是因為在音素層（phone）上的共同訓練。

recognition task simultaneously. In comparison with the parallel modeling approaches (i.e., GOPT and LSTM), we

can observe that HierTFR substantially improves the performance across all tasks, where its performance gains reveal the importance of capturing the hierarchical linguistic structures of an input utterance. Notably, compared to the HiPAMA, our model consistently achieves superior performance on a variety of pronunciation assessment tasks. This superiority stems from our tactfully designed selective fusion mechanism and the correlation-aware loss. The former allows our model to assess utterance-level aspect scores by leveraging information from diverse linguistic levels, while the latter explicitly models the relatedness among different aspects during the optimization.

同時進行識別任務。與平行建模方法（例如 GOPT 和 LSTM）相比，我們可以觀察到 HierTFR 在所有任務上大幅提升了性能，其性能增益揭示了捕捉輸入語句層次語言結構的重要性。值得注意的是，與 HiPAMA 相比，我們的模型在各種發音評估任務上持續達到更優異的表現。這種優勢來自我們巧妙設計的選擇性融合機制與關聯感知損失。前者使模型能夠利用來自不同語言層級的資訊來評估語句層級的面向分數，後者則在優化過程中明確建模不同面向之間的相關性。

4.2 Qualitative Analysis

4.2 質性分析

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

面向之間相關性的質性視覺化。在第二組實驗中，我們檢視了詞級與語句級上不同面向之間的相關性，其中

Models	Phone Score 音素分數	Word Score 詞彙分數			Utterance Score 語句分數				
	Accuracy 準確性	Accuracy 準確性	Stress 重音	Total 總分	Accuracy 準確性	Completeness 完整性	Fluency 流暢度	Prosody 韻律	Total 總計
HierTFR	0.644	0.622	0.325	0.634	0.735	0.513	0.801	0.795	0.764
w/o CorrLoss 不含 CorrLoss	0.639	0.605	0.348	0.620	0.728	0.520	0.796	0.789	0.758
w/o Pretrain 不含預訓練	0.621	0.545	0.318	0.559	0.716	0.215	0.770	0.772	0.739
w/o SFusion 不含 SFusion	0.630	0.608	0.328	0.622	0.728	0.378	0.784	0.782	0.756
w/o AspAtt 不含 AspAtt	0.636	0.584	0.290	0.596	0.724	0.383	0.784	0.775	0.746

Table 3: Table 2: Ablation study on HierTFR, reporting PCC scores on three linguistic levels.

表 3：表 2：關於 HierTFR 的消融研究，報告三個語言層級的 PCC 分數。

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

annotators judge the proficiency scores, they also take multiple pronunciation aspects into account simultaneously. Next, the completeness aspect measures the percentage of words with good pronunciation quality in an utterance. This implicitly reflects the intelligibility of a learner's pronunciation and is vital to the accuracy assessment.

在評估特定面向分數時，面向注意力機制的注意力權重是根據開發集決定的。對於詞層級評估，注意力權重的分佈與 speechocean762 資料集的人工評分規則高度一致。在圖 4(a) 中，total 面向作為綜合評估，其相對應的權重來自各種發音面向。相較之下，accuracy 面向衡量詞內發音錯誤音素的比例，導致注意力權重更集中在詞層級單位本身。此外，stress 分數亦高度關注 accuracy 面向，反映詞彙重音與詞層級發音準確性之間的緊密關聯（Korzekwa et al., 2022）。關於句子層級面向間的相關性，檢視圖 4(b) 我們發現 prosody 與 total 面向的注意力權重分佈在多個發音面向上，而 accuracy 與 completeness 的注意力權重則主要集中在 completeness 面向。 一個可能的原因是韻律（prosody）分數與總分都衡量高階口語能力，當人工評註者判定能力分數時，他們也會同時考量多個發音面向。接著，完整性（completeness）面向衡量一個語句中具有良好發音品質的字詞比例。這隱含反映出學習者發音的可懂度（intelligibility），對準確性評估至關重要。

Qualitative Visualization of Interactions Across Linguistic Levels. In Figure 4©, we report on the average gate values of utterances for three linguistic granularities by estimating the utterancelevel pronunciation aspect scores based on the development set. We can observe that the phonelevel representations bear high impacts on the utterance-level aspect assessments, in comparison to the other linguistic levels. Next, the word-level

跨語言層級互動的定性視覺化。在圖 4© 中，我們報告了針對三種語言粒度之語句的平均門控值（average gate values），方法是基於開發集估計語句層級的發音面向分數。我們可以觀察到，與其他語言層級相比，音素（phone）層級的表示對語句層級面向評估有較高的影響力。接著，詞（word）層級...

and utterance-level representations exhibit minimal impact on the completeness and total aspects, respectively. One possible reason is that the completeness aspect somehow reflects pronunciation intelligibility, and our model learns to distill the information from the phone- and utterance-level representations. On the other hand, the total aspect evaluates an overall speaking skill. Our model thus tends to capture the subtle information by distilling the fine-grained traits inherent in the phone- and word-levels.

而子音（phone）和語句（utterance）層級的表示分別對完整性（completeness）與總分（total）方面影響甚微。一個可能的原因是，完整性在某種程度上反映了發音的可懂度，而我們的模型學會從子音與語句層級的表示中蒸餾相關資訊。另一方面，總分評估的是整體口說能力，因此模型傾向於透過從子音與詞（word）層級中蒸餾出的細緻特徵來擷取微妙資訊。

4.3 Ablation Study 4.3 消融研究

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

為了深入了解 HierTFR 各個模型組件的效用，我們進行了消融研究以調查它們的影響。這些變體包括排除相關性感知正則化器（w/o CorrLoss）、移除提出的預訓練策略（w/o Pretrain）、省略選擇性融合機制（w/o SFusion），以及在單字與語句層級同時刪除面向注意機制（w/o AspAtt）。從表 2 中可以觀察到，所提出的相關性感知正則化損失對大多數發音評估任務是有利的。接著，所提出的預訓練策略對取得較佳表現至關重要，因為未使用這些策略訓練的模型在所有發音評估任務上表現相對較差。這突顯了預訓練策略對分層自動發音評估（APA）模型的效用，從而降低了對大量有標註訓練數據的需求。第三，移除選擇性融合機制會導致語句層級面向評估的退化，而移除面向注意機制則會使單字層級面向評估的表現惡化。

5 Related Work

Early studies on APA focused primarily on singleaspect assessments, typically through individually constructing scoring modules to predict a holistic

早期關於自動發音評估（APA）的研究主要集中在單一面向的評估，通常透過單獨構建評分模組來預測整體表現

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

在目標語言層級或某些特定面向上，使用不同組手工設計的特徵來評分發音熟練度，例如音素層級的後驗機率 (Witt and Young, 2000)、單字層級的詞彙重音 (Ferrer et al., 2015)，或各種語句層級的發音面向 (Coutinho et al., 2016)。近年來，隨著深度學習的快速進展 (Vaswani et al., 2017; Raffel et al., 2020; Hsu et al., 2021)，已成功開發出數個用於多面向與多粒度發音評估的神經評分模型。Gong et al. (2022) 提出一種基於 GOP 特徵的 Transformer (GOPT) 架構，透過多任務學習方案來建模多種粒度的發音面向。Do et al. (2023b) 採用具有階層結構的神經評分器，模擬語句的語言階層，以在自動發音評估 (APA) 上達到最先進的表現。

6 Conclusion 6 結論

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

在本文中，我們提出一種新穎的分層建模方法（稱為 HierTFR），用以處理多面向與多粒度的自動發音評估（APA）。為了明確捕捉發音面向之間的相關性，本研究設計了一種具相關性感知的正則化損失。我們也為 HierTFR 模型開發了模型預訓練策略。大量實驗結果證實所提出方法相較於多項頂尖方法的可行性與有效性。未來工作中，我們計畫在開放回應情境下檢驗所提出的 HierTFR 模型，在此情境中學習者可自由講話或回應指定任務或問題（Wang et. al., 2018；Park and Choi, 2023）。此外，可解釋的發音回饋問題亦將作為未來的擴充方向。

Limitations 限制

Limited Applicability. In this research, the proposed model focus on the “reading-aloud” pronunciation training scenario, where the assumption is that the L2 learner pronounces a predetermined text prompt correctly, which restricts the applicability of our models to other learning scenarios, such as freely speaking or openended conversations.

適用範圍有限。本研究所提出的模型著重於「朗讀」發音訓練情境，該情境假設第二語言學習者會照預先給定的文本提示發音正確，這限制了我們模型在其他學習情境（例如自由說話或開放式對話）中的適用性。

Lack of Accent Diversity. The used dataset merely contains Mandarin L2 learners, hindering the generalizability of the proposed model and could

缺乏口音多樣性。所使用的資料集僅包含普通話第二語言學習者，限制了所提出模型的泛化能力，且可能

be untenable when assessing the L2 learners with diverse accents.

在評估具有多樣口音的第二語言學習者時，這將變得無法成立。

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

缺乏可解釋性。所提出方法的模型僅透過訓練來模仿專家的標註，未依賴人工評分規準或其他外部知識，因此無法直接為評分結果提供合理的解釋。

Ethics Statement 倫理聲明

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

茲此聲明，本研究之所有共同作者皆遵守所提供之 ACL 倫理守則並尊重行為準則。我們的實驗語料庫 speechocean762 為廣泛使用且公開取得之資料。我們認為本研究不存在潛在風險。

References 參考文獻

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

Stefano Bannò、Bhanu Balusu、Mark Gales、Kate Knill 與 Konstantinos Kyriakopoulos。2022。L2 口語英語的觀點特定評估。刊於 Interspeech (INTERSPEECH) 會議論文集，頁 4471-4475。

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

Alexei Baevski、Henry Zhou、Abdelrahman Mohamed 與 Michael Auli。2020。Wav2vec 2.0：用於語音表徵自監督學習之框架。刊於國際神經資訊處理系統會議 (NIPS) 論文集，頁 12449-12460。

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

Fu An Chao、Tien Hong Lo、Tzu I. Wu、Yao Ting Sung、Berlin Chen。2022。3M：一種有效的多視角、多粒度與多面向英語發音評估建模方法。刊於亞太訊號與資訊處理協會年會與研討會 (APSIPA ASC) 論文集，頁 575-582。

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

Sanyuan Chen、Chengyi Wang、Zhengyang Chen、Yu Wu、Shujie Liu、Zhuo Chen、Jinyu Li、Naoyuki Kanda、Takuya Yoshioka、Xiong Xiao、Jian Wu、Long Zhou、Shuo Ren、Yanmin Qian、Yao Qian、Jian Wu、Michael Zeng、Xiangzhan Yu、Furu Wei。2022。Wavlm：用於全棧語音處理的大規模自監督預訓練。IEEE Journal of Selected Topics in Signal Processing，卷 16，頁 1505–1518。

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

Nancy F. Chen 與 Haizhou Li。2016。電腦輔助發音訓練：從發音評分走向口語語言學習。收錄於 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) 會議論文集，頁 1–7。

Eduardo Coutinho, Florian Hönig, Yue Zhang, Simone Hantke, Anton Batliner, Elmar Nöth, and Björn Schuller. 2016. Assessing the prosody of non-native

Eduardo Coutinho、Florian Hönig、Yue Zhang、Simone Hantke、Anton Batliner、Elmar Nöth 與 Björn Schuller。2016。評估非母語英語使用者的韻律

speakers of English: Measures and feature sets. In Proceedings of the International Conference on Language Resources and Evaluation (LREC), pages 1328-1332.

英語使用者：度量與特徵集。收錄於 International Conference on Language Resources and Evaluation (LREC) 會議論文集，頁 1328–1332。

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

Heejin Do、Yunsu Kim、Gary Geunbae Lee。2023a。《Score-balanced Loss for Multi-aspect Pronunciation Assessment》。收錄於 Interspeech (INTERSPEECH) 會議論文集，第 4998–5002 頁。

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

Heejin Do、Yunsu Kim、Gary Geunbae Lee。2023b。《Hierarchical pronunciation assessment with multiaspect attention》。收錄於 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 會議論文集，第 1–5 頁。

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

Maxine Eskenazi。2009。《An overview of spoken language technology for education》。發表於 Speech Communication，卷 51，第 832–844 頁。

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

Keelan Evanini、Xinhao Wang。2013。《Automated speech scoring for Nonnative middle school students with multiple task types》。收錄於 Interspeech (INTERSPEECH) 會議論文集，第 2435–2439 頁。

Keelan Evanini, Maurice Cogan Hauck, and Kenji Hakuta. 2017. Approaches to automated scoring of speaking for K-12 English language proficiency assessments. ETS Research Report Series, pages 111.

Keelan Evanini、Maurice Cogan Hauck 與 Kenji Hakuta。2017。K-12 英語能力評量之口語自動評分方法。ETS Research Report Series，頁數 111。

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

Luciana Ferrer、Harry Bratt、Colleen Richey、Horacio Franco、Victor Abrash 與 Kristin Precoda。2015。使用頻譜與韻律特徵對電腦輔助語言學習系統進行詞彙重音分類。Speech Communication，卷 69，頁 31–45。

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

Marjan Ghazvininejad、Omer Levy、Yinhan Liu 與 Luke Zettlemoyer。2019。Mask-Predict：條件遮蔽語言模型的平行解碼。收錄於《自然語言處理實證方法會議與國際聯合自然語言處理會議論文集（EMNLP-IJCNLP）》，頁 6112–6121。

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

Yuan Gong、Ziyi Chen、Iek-Heng Chu、Peng Chang 與 James Glass。2022。基於 Transformer 的多面向多粒度非母語英語說話者發音評量。收錄於 IEEE 國際聲學、語音與訊號處理會議（ICASSP）論文集，頁 7262–7266。

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

Wei-Ning Hsu、Benjamin Bolte、Yao-Hung Hubert Tsai、Kushal Lakhotia、Ruslan Salakhutdinov、Abdelrahman Mohamed。2021。HuBERT：透過對隱藏單位的遮蔽預測進行自監督語音表徵學習。IEEE/ACM Transactions on Audio, Speech and Language Processing，卷 29，頁 3451–3460。

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

Wenping Hu、Yao Qian、Frank K. Soong、Yong Wang。2015。結合以深度神經網路訓練的聲學模型與基於轉移學習的邏輯回歸分類器之改良發音錯誤偵測。Speech Communication，卷 67，頁 154–166。

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

Eesung Kim、Jae-Jin Jeon、Hyeji Seo、Hoon Kim。2022。利用自監督語音表徵學習之自動發音評估。收錄於 Proceedings of Interspeech (INTERSPEECH)，頁 1411–1415。

Yassine Kheir, Ahmed Ali, and Shammur Chowdhury. 2023. Automatic Pronunciation Assessment - A Review. In Findings of the Association for Computational Linguistics: EMNLP, pages 83048324.

Yassine Kheir、Ahmed Ali、Shammur Chowdhury。2023。自動發音評估——綜述。收錄於 Findings of the Association for Computational Linguistics: EMNLP，頁 8304–8324。

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

Daniel Korzekwa、Jaime Lorenzo-Trueba、Thomas Drugman 與 Bozena Kostek。2022。Computer-assisted pronunciation training—Speech synthesis is almost all you need。Speech Communication，卷 142，頁 22–33。

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. On Generative Spoken Language Modeling from Raw Audio. Transactions of the Association for Computational Linguistics, volume 9, pages 13361354.

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

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

Binghuai Lin 與 Liyuan Wang。2021。Deep feature transfer learning for automatic pronunciation assessment。收錄於 Interspeech (INTERSPEECH) 會議論文集，頁 4438–4442。

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

Silke M. Witt 與 S. J. Young。2000。Phone-level pronunciation scoring and assessment for interactive language learning。Speech Communication，卷 30，頁 95–108。

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

Jungbae Park 和 Seungtaek Choi。2023。針對端到端自動語音評分的冷啟動問題。收錄於 Interspeech (INTERSPEECH) 會議論文集，第 994–998 頁。

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

Yifan Peng、Siddharth Dalmia、Ian Lane、和 Shinji Watanabe。2022。Branchformer：並行 mlpattention 架構以捕捉語音識別與理解的局部與全域上下文。收錄於國際機器學習會議 (PMLR)，第 17627–17643 頁。

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

pages 47-56.

Yu Wang、M.J.F. Gales、Kate M Knill、Konstantinos Kyriakopoulos、Andrey Malinin、Rogier C van Dalen、Mohammad Rashid。2018。朝向自發口語英語的自動評估。Speech Communication，卷 104，第 47–56 頁。

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

Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li、和 Peter J. Liu。2020。以統一的文字對文字 Transformer 探索遷移學習的極限。The Journal of Machine Learning Research，卷 21，第 5485–5551 頁。

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

Robert Ridley、梁賀、戴昕宇、黃書堅、陳家駿。2021。自動化跨題目作文特徵評分。收錄於 AAAI 人工智慧會議論文集（AAAI），第 35 卷，頁 13745–13753。

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

Pamela M Rogerson-Revell。2021。電腦輔助發音訓練（CAPT）：當前議題與未來方向。RELC Journal，第 52 卷，頁 189–205。

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

Hyungshin Ryu、Sunhee Kim、Minhwa Chung。2023。使用多任務學習之發音評估與誤發音偵測及診斷的聯合模型。收錄於 Interspeech（INTERSPEECH）論文集，頁 959–963。

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

Yang Shen、Ayano Yasukagawa、Daisuke Saito、Nobuaki Minematsu、Kazuya Saito。2021。基於發聲量與發音品質之可解釋網路的 L2 英語流暢度最佳化預測。收錄於 IEEE Spoken Language Technology Workshop（SLT）論文集，頁 698–704。

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

Alexandre Tamborrino、Nicola Pellicanò、Baptiste Pannier、Pascal Voitot 以及 Louise Naudin。2020。Pretraining is (almost) all you need: An application to commonsense reasoning。收錄於 Proceedings of the Association for Computational Linguistics (ACL)，頁 3878–3887。

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

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 以及 Illia Polosukhin。2017。Attention is all you need。收錄於 Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)，頁 5998–6008。

Heng-Da Xu, Zhongli Li, Qingyu Zhou, Chao Li, Zizhen Wang, Yunbo Cao, Heyan Huang, and XianLing Mao. 2021. Read, listen, and see: Leveraging multimodal information helps Chinese spell checking. In Findings of the Association

for Computational Linguistics (ACL-IJCNLP Findings), pages 716-728.

Heng-Da Xu、Zhongli Li、Qingyu Zhou、Chao Li、Zizhen Wang、Yunbo Cao、Heyan Huang 以及 XianLing Mao。2021。Read, listen, and see: Leveraging multimodal information helps Chinese spell checking。收錄於 Findings of the Association for Computational Linguistics (ACL-IJCNLP Findings)，頁 716–728。

Junbo Zhang, Zhiwen Zhang, Yongqing Wang, Zhiyong Yan, Qiong Song, Yukai Huang, Ke Li, Daniel Povey, and Yujun Wang. 2021. Speechocean762: An open-source non-native English speech corpus for pronunciation assessment.

Junbo Zhang、Zhiwen Zhang、Yongqing Wang、Zhiyong Yan、Qiong Song、Yukai Huang、Ke Li、Daniel Povey 以及 Yujun Wang。2021。Speechocean762: An open-source non-native English speech corpus for pronunciation assessment。

In Proceedings of Interspeech (INTERSPEECH), pages 3710-3714.

於 Interspeech (INTERSPEECH) 會議論文集中，頁 3710-3714。

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

Chuanbo Zhu、Takuya Kunihara、Daisuke Saito、Nobuaki Minematsu、Noriko Nakanishi。2022。自動預測日本英語學習者口說單字與音素可懂度。收錄於 IEEE Spoken Language Technology Workshop (SLT)，頁 1029-1036。

A Pronunciation Feature Extractions

發音特徵擷取

GOP Feature. To extract the GOP feature, we first align audio signals X with the text prompt T by using an ASR model [5] to obtain the timestamps for each phone in the canonical phone sequence. Next, framelevel phonetic posterior probabilities are produced by the ASR model and then averaged over the time dimension based on the phone-level timestamps. The resulting phone-level posterior probabilities are converted into a GOP feature vector as a combination of log phone posterior (LPP) and log posterior ratio (LPR). Owing to the used ASR model containing 42 phones, the GOP feature of a canonical phone p can be represented as an 84 -dimensional vector:

GOP 特徵。為擷取 GOP 特徵，我們首先使用 ASR 模型 [5] 將音訊訊號 X 與文本提示 T 對齊，以取得典型音素序列中每個音素的時間戳。接著，由 ASR 模型產生逐幀的語音音素後驗機率，並根據音素層級的時間戳在時間維度上進行平均。所得的音素層級後驗機率被轉換為 GOP 特徵向量，該向量由對數音素後驗 (LPP) 與對數後驗比 (LPR) 組合而成。由於所使用的 ASR 模型包含 42 個音素，典型音素 p 的 GOP 特徵可表示為一個 84 維向量：

$$
\begin{aligned}
&[\mathrm{LPP}\left(p_1\right), \ldots, \mathrm{LPP}\left(p_{42}\right), \\
&\mathrm{LPR}\left(p_1 \mid p\right), \ldots, \mathrm{LPR}\left(p_{42} \mid p\right)] \\
&\mathrm{LPP}\left(p_i\right)=\log p\left(p_i \mid \mathbf{o} ; \mathrm{t}_s, \mathrm{t}_e\right) \\
&=\frac{1}{t_e-t_s+1} \sum_{t=t_s}^{t_e} \log p\left(p_i \mid \mathrm{o}_t\right)
\end{aligned}
$$

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

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

其中 LPR 為音素 p_i 與 p; t_s 的對數後驗比（log posterior ratio），t_e 為音素 p 的起始與結束時間戳，而 o_t 為時間框架 t 的輸入聲學觀測值。

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

能量特徵。能量特徵為一個 7 維向量，包含在音素片段上的 [mean, std, median, mad, sum, max, min]，其中對每個時間框架計算能量值時使用的是均方根能量（RMSE），視窗長度為 25 毫秒，步幅為 10 毫秒。

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

時長特徵。時長特徵為一個 1 維向量，表示每個音素片段的長度（秒）。

Corresponding author. 通訊作者。

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

[2] Both the aspects of utterance completeness and word stress suffer from label imbalance problems, with more than 90%

[2] 句子完整性與重音兩個面向都存在標籤不平衡問題，超過 90%

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

[3] 有關發音特徵擷取的更多細節，請參見附錄 A。

[4] To calculate PCC scores between aspects across different granularities, we duplicate the aspect scores of higher

[4] 為了計算不同粒度之間面向的 PCC 分數，我們複製較高層級的面向分數

granularities to match the aspect scores at the lower granularities.

細緻度以匹配較低層級的面向分數。

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

[5] 一個可公開評估的 ASR 模型，使用英語語音語料訓練： https://kaldi-asr.org/models/m13.