無需複雜聯合訓練的英語發音評估：以 LoRA 微調的語音多模態 LLM

Taekyung Ahn [1,2], Hosung Nam [2]
Enuma, Inc. [1], Korea University [2]
taekyung@enuma.com, hnam@korea.ac.kr

摘要

本研究展示了經由低秩調適（LoRA）調整的多模態大型語言模型（MLLM）能同時執行自動發音評估（APA）與誤讀檢測與診斷（MDD）。利用 Microsoft 的 Phi-4-multimodal-instruct，我們的微調方法省略了傳統上為這些不同任務所需的複雜架構變更或分別訓練程序。在 Speechocean762 資料集上微調後，模型預測的發音評估分數與人工評分之間呈現強烈的皮爾森相關係數（PCC > 0.7），同時達到低字錯誤率（WER）與音素錯誤率（PER）（兩者均 < 0.15）。值得注意的是，僅微調 LoRA 層即足以達到與微調所有音訊層相當的表現。本研究強調，可透過調適大型多模態模型而非完全微調，建立一個整合的發音評估系統，並採用相較於先前為同時處理 APA 與 MDD 而設計的聯合模型更為簡單的訓練方法。 這種高效的 LoRA 為基礎的方法為英語第二語言學習者提供更易取得、整合且具成效的電腦輔助發音訓練（CAPT）技術鋪路。

關鍵詞

ASR, LLM, APA, MDD, Multimodal, EFL, CAPT

1 介紹

隨著自動語音辨識（ASR）技術的進步，在電腦輔助發音訓練（CAPT）領域已開發出各種技術以協助非母語英語使用者改善英文發音[8]。CAPT 是一種可自動評估語音發音的系統。它主要可分為自動發音評估（APA）任務，對給定語句進行打分評估，以及誤發音偵測與診斷（MDD）任務，判斷講者朗讀給定句子時是否發生誤發音。

端到端預訓練語音辨識模型的出現，使得透過少量資料微調即可達到優異效能，並簡化訓練流程[3][2][5]。利用這些模型的 CAPT 系統主要朝兩個方向發展：透過微調技術提升對誤發音語音的辨識能力，以及基於辨識結果構建更為強健的發音評估框架[18][24][12]。這些模型固有的彈性也促進了同時執行 APA 與 MDD 任務之統一系統的研究。

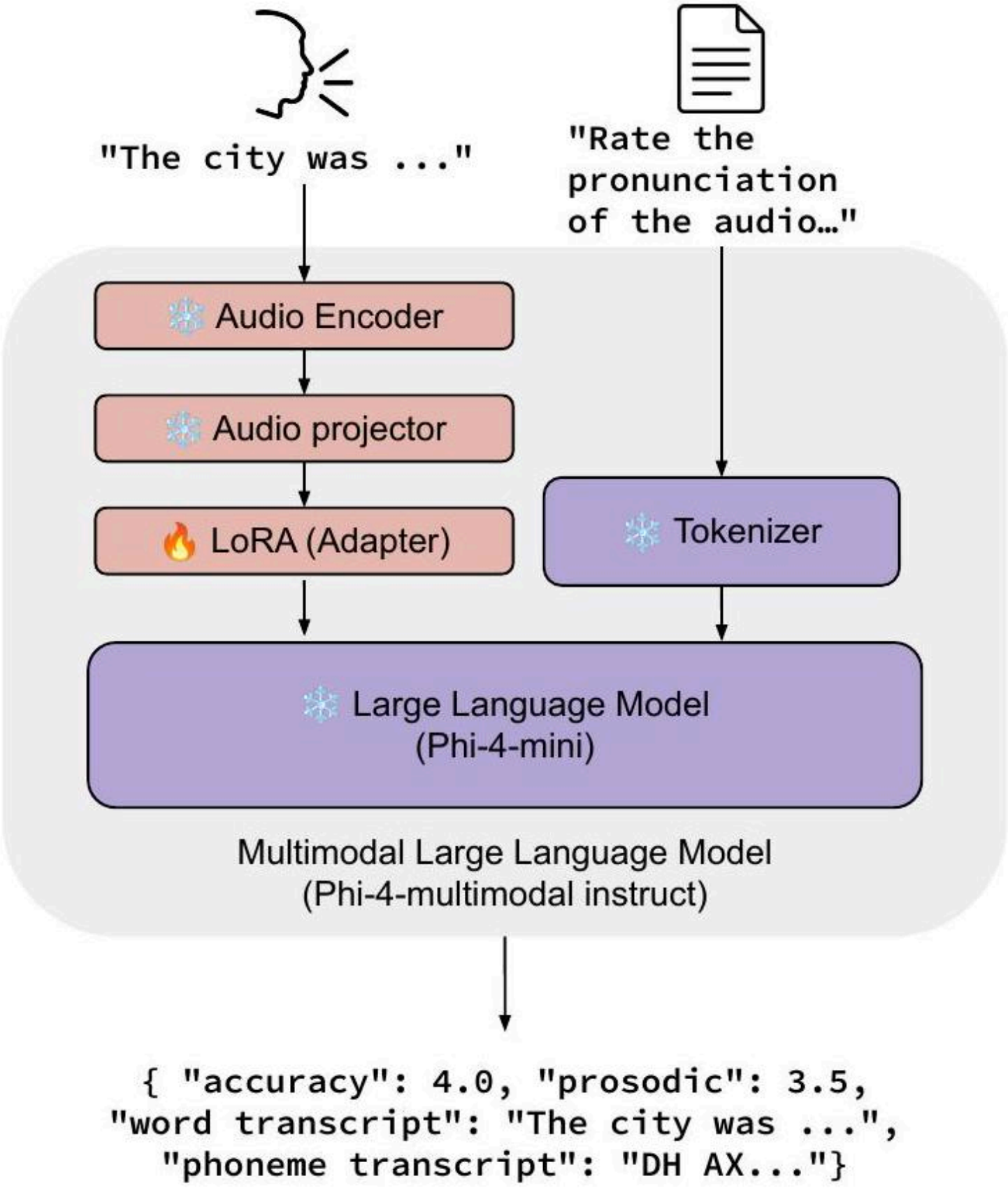

圖1：我們所提出方法的概覽。

[20]，揭示了這兩項功能之間存在顯著相關性。然而，這些方法仍然需要為每個任務使用獨立的資料集和獨立的模型架構，儘管它們在功能上相互依賴，但仍需不同的訓練程序和計算資源。

最近，LLMs（Large Language Models）在理解自然語言語境方面展現了顯著進步[14]。LLMs 可透過少量的額外訓練或僅以提示對齊來達成所需目標，而無需蒐集大量訓練資料與計算資源。跟隨這些進展，結合視覺與音訊編碼器的多模態 LLMs（MLLMs）應運而生。它們採用將編碼器連接到現有 LLMs 的架構，能將原始影像或音訊檔案轉換為類似於文字嵌入的向量表示。透過訓練適配器去理解此向量，LLMs 就能在與文字相同的語境中理解影像或音訊輸入。

為了解決現有分開訓練方法的限制，我們的研究提出一種統一方法，在單一訓練框架與模型架構中同時執行 APA 和 MDD 任務。我們特別採用 Phi-4-multimodal-instruct [1]，
利用其 MLLM 能力以克服傳統上對獨立資料集與模型層的需求。透過使用只學習語音適配器的高效 LoRA（低秩適配）微調，我們的方法在維持效能的同時，避開了整體微調的計算負擔。我們的目標是展示這個統一的 MLLM 框架能有效同時處理兩種發音評估任務，提供一個比現有多階段 CAPT 系統更節省資源且更簡化的替代方案。

2 相關工作

2.1 結合 ASR 的 CAPT 系統

CAPT（電腦輔助發音訓練）於 1980 年代出現[8]。在 CAPT 中，透過結合 ASR（自動語音辨識）技術，發音教學被積極應用與研究。自 1990 年代以來，發音評估方法便成為研究主題[15][4][10]，此領域大致可分為自動發音評估（APA）與錯誤發音偵測與診斷（MDD）兩大任務。

APA 指的是一系列透過建立能自動為學習者語音給分的系統，來協助語言學習的方法。隨著 ASR 技術的發展，得以構建能辨識語音音素單位並透過與正確音素比較來自動評分發音正確性的系統[9][10]。這類研究在評估方法上有多種探討，能以稱為發音良好度（goodness of pronunciation, GOP）的指標，使用基於 HMM 的 ASR 系統來衡量發音準確性[15][23]。

MDD 研究透過將擷取出的音素層級聲學特徵與正確音素進行比較來進行。傳統的錯誤發音偵測與診斷技術通常倚賴專家制定的語音規則、隱馬可夫模型（HMM）與高斯混合模型（GMM）等統計方法，以及包含 MFCC 與共振峰追蹤在內的聲學特徵分析，這些方法需要大量的人工工程，且常常難以應對情境依存的變化、重音韻律特徵與說話者差異 [23] [6] [19]。

2.2 端到端語音辨識建模

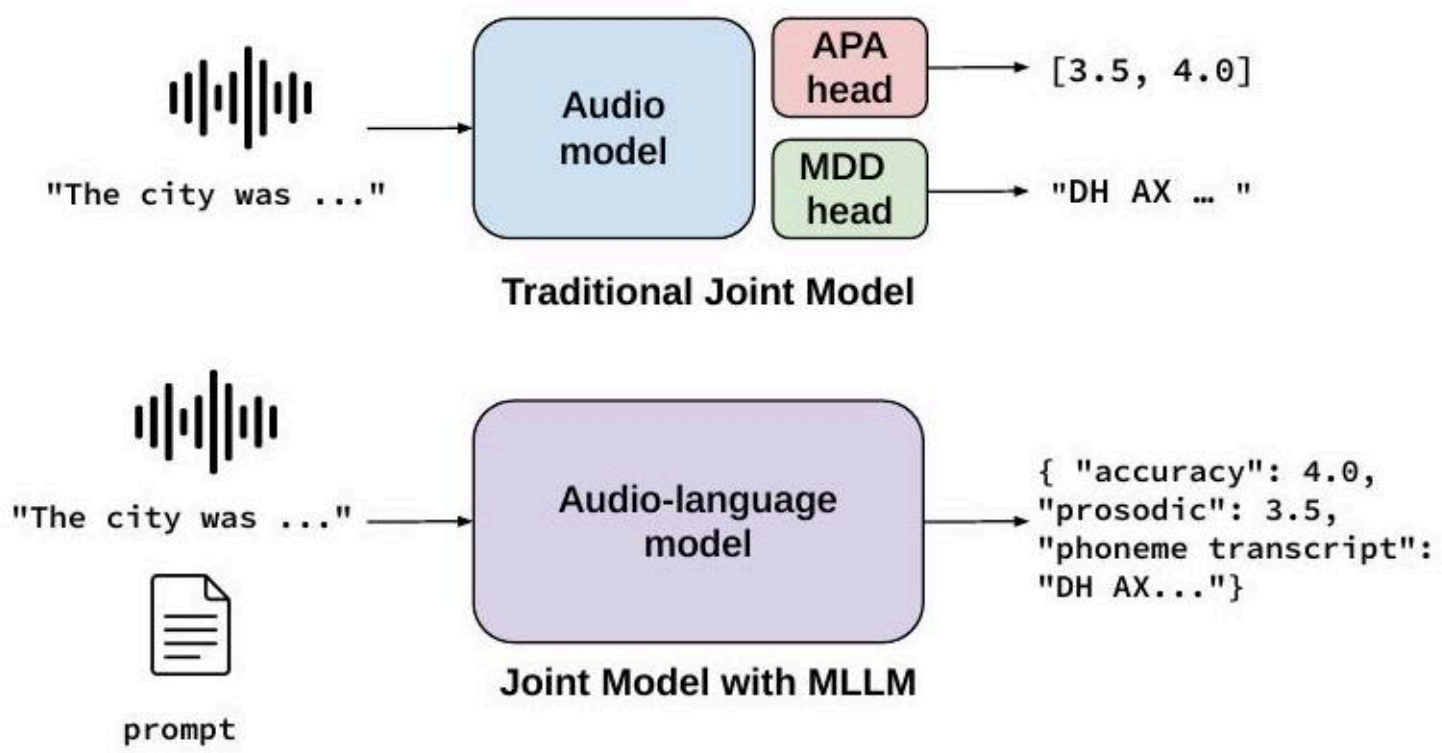

圖 2：圖 2：我們的模型與過去研究的比較。

隨著機器學習技術的進步，語音識別模型現在可以在沒有複雜前處理的情況下進行訓練
由於深度學習模型能自主學習語音與文字之間的關係，故引入了相關程序[7]。Gong[13]提出了 GOPT，一種基於 GOP 特徵的 Transformer 模型，用於評估非英語母語者的發音。該模型透過多任務學習，同時評估發音品質的多個面向（準確性、流暢度、完整性、韻律），涵蓋不同層級的特徵（音素、單字、語句），從而提升各評估任務的表現。在 Speechocean762[25]資料集上的實驗顯示，GOPT 在發音評估上顯著優於先前的傳統方法。

隨著如 wav2vec2.0 [3] 等自監督學習（SSL）方法的出現，語音辨識（ASR）模型能以比音素更小的單位學習語音特徵。研究人員可以透過少量資料的微調，輕易建構發音評估（APA）與錯誤檢測與診斷（MDD）引擎 [18][24][12]。儘管有這些進展，多數方法仍然需要針對 APA 與 MDD 任務分別進行訓練程序。[20] 建立了一個可同時執行 APA 與 MDD 的聯合模型：先以音素層級轉錄的語音資料微調 wav2vec2.0 模型，然後以分離的 APA 與 MDD 頭進行多任務學習。然而，此法仍需為每個任務編碼器準備獨立的資料集與不同的訓練流程，限制了可擴展性與資源效率。此外，僅以語音輸入訓練的這些模型無法整合像文字提示等其他模態以提高靈活性。

2.3 多模態大型語言模型（MLLM）

隨後，多模態大型語言模型（MLLMs）出現，能夠處理影像和語音以及文字。這些模型透過專門的編碼器將視覺或音頻輸入轉換為向量維度，然後將這些向量表示與 LLM 嵌入對齊，以實現全面的多模態理解。Wang [21] 探討了 GPT-4o [16] 在發音評估上的零樣本能力，評估其在多層級（音素、單字、句子）打分的表現。研究顯示 GPT-4o 的零樣本打分準確度顯著低於其他評估方法。該研究得出結論，像 GPT-4o 這類的 MLLMs 在未經微調的情況下，可能尚未能取代專門用於發音評分的工具。

Fu [11] 開發了一個基於 MLLMs 的發音評估系統。在他們的研究中，他們訓練了一個專用的語音編碼器並將其與 LLM 整合，達成系統預測的發音分數與人工評估者之間的高皮爾森相關係數（PCC）值。這項研究成功展示了基於 MLLM 的發音評估系統的可行性。然而，該研究使用了約 1,000 小時的語音資料來訓練與 LLM 整合的語音編碼器。透過適配器（adapters）將預訓練語音編碼器與 LLM 連接以建立 MLLMs，呈現出顯著的挑戰。這種方法涉及複雜的整合流程，導致訓練與部署的成本相當高昂。

2.4 Phi-4-multimodal 與 LoRA 微調

Phi-4-multimodal [1] 是微軟發佈的一個視覺-語音語言模型，是一個將視覺與語音編碼器連接到 LLM 的多模態模型。Phi-4 採用了 Mixture-of-LoRAs（低秩

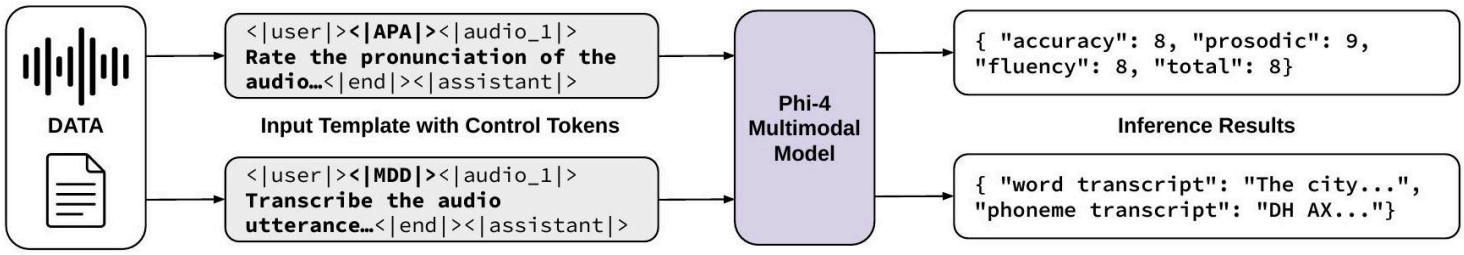

圖 3：圖 3：使用提示詞（prompts）與控制標記（control tokens）的方法。

在網路中採用適配器（LoRA）結構，設計架構以在僅訓練 LoRA 層而非整體微調的情況下，達到良好效能並將不同模態間的干擾降到最低。憑藉此優勢，相較於模型規模，可在有限的計算資源下進行各種微調。

利用 Phi-4-multimodal 的這些能力，本研究解決了現有 CAPT 系統的基本侷限：以往需為 APA 與 MDD 任務分別進行不同的訓練程序與資料集。透過採用統一的多模態大語言模型（MLLM）框架並以高效的 LoRA 微調，我們證明兩項任務可以在單一模型架構與訓練流程中同時完成。此方法不僅消除了以往方法所需的大量資源，還透過提示工程實現系統的靈活自訂，代表朝向更易獲取且更高效的 CAPT 技術邁出重要一步。我們的工作確立了構建真正統一發音評估系統的可行性，克服了歷來將這些功能相關任務分隔開來的計算與架構限制。

3 Method

3.1 Control Tokens

為了在不修改模型架構的情況下區分任務，我們使用了控制標記：<|APA|> 用於 APA 提示，<|MDD|> 用於 MDD 提示（見圖 3）。在進行監督式微調（SFT）前，這些標記被置於每個提示的開頭。在 SFT 過程中，模型學會將每個控制標記與對應任務的具體需求相關聯。接著我們實驗性地評估在推理階段明確使用這些任務專屬控制標記時的效能提升。作為基線比較，我們在不使用控制標記或額外訓練的情況下，採用了 Phi-4-multimodal-instruct 模型 [1] 進行推理。

3.2 Encoder Layer Unfreeze

本研究探討解凍特定音頻相關層的影響。使用相同資料集，我們比較了兩種不同的微調策略：

僅 LoRA 微調：僅更新 LoRA 適配器層的權重，同時保持其他層（包括音頻編碼器和音頻投影器）凍結（見圖 1）。
解凍微調：除了更新 LoRA 適配器層的權重外，還解凍並更新音訊編碼器（Audio Encoder）和音訊投影器（Audio Projector）層的權重（見圖 4）。

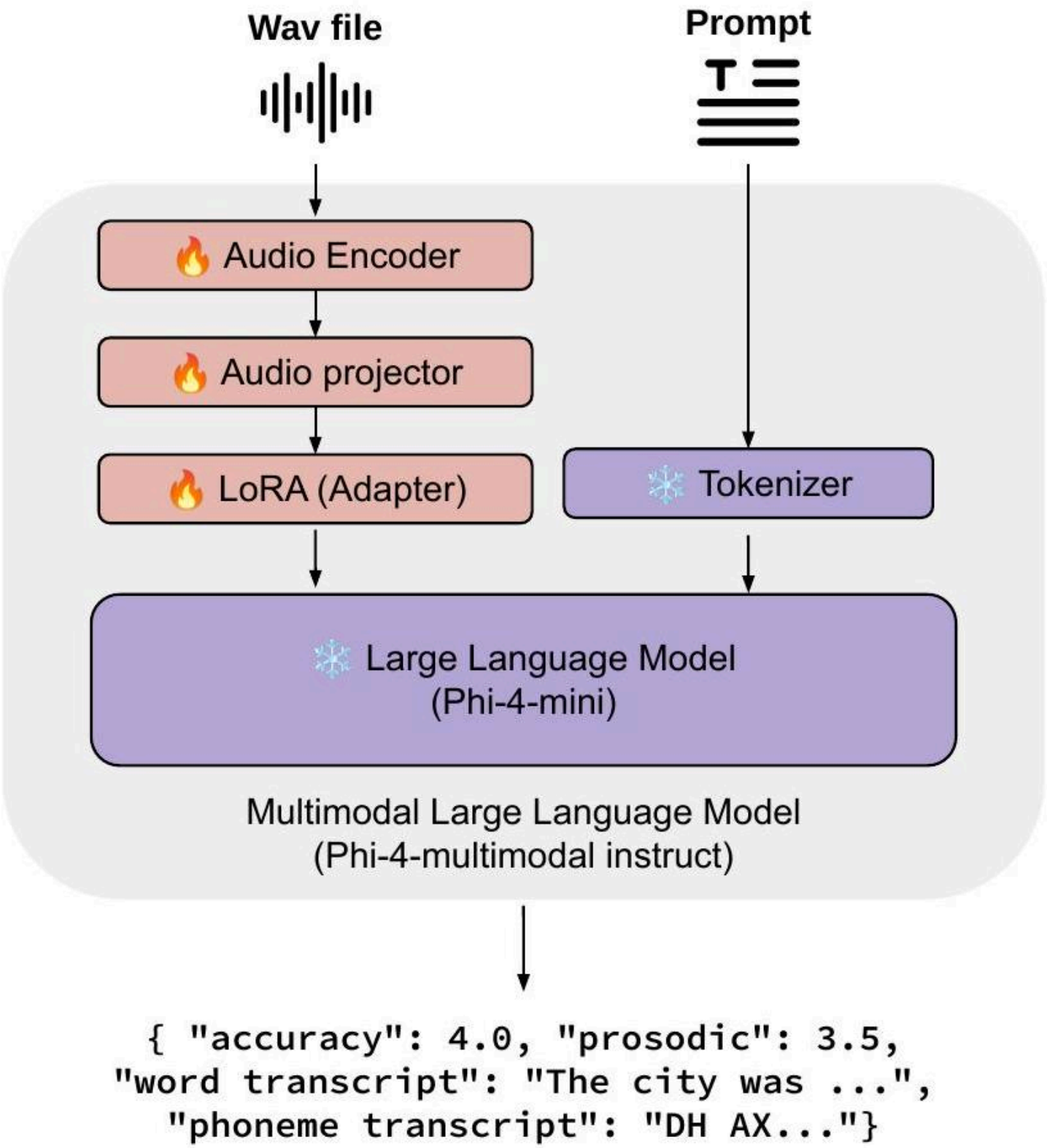

圖4：圖4：解凍層的方法。

雖然傳統做法通常將微調限制在 LoRA 適配器層，但官方文件[17]指出解凍所有音訊層可能帶來好處。基於此，我們的實驗直接比較這些方法，以確定在音訊相關任務中，與 LoRA 適配器同時解凍音訊編碼器和投影器是否能提供性能上的優勢。

3.3 找出 APA 與 MDD 任務之間的相關性

基於 Ryu 等人的研究 [20]，我們分析了發音準確度分數與發音辨識表現之間的相關性。我們檢驗了來自單一模型（在相同語句資料集上共同訓練 APA 與 MDD 任務）所產生的分數，是否真正反映發音品質，而非僅僅將資料投射為分數。我們具體調查了人類與模型的準確度分數與來自 MDD 任務的音素錯誤率（PER）之間的關係。我們假設

表 1：Speechocean762 資料集資訊

類型（大小）		訓練（2.5k）	測試（2.5k）
檔案總數		2500	2500
講者		125	125
wav 總時長		2 小時 52 分鐘	2 小時 41 分鐘
年齡（n）	20 歲以下	49.6% (62)	56.0% (70)
	20 秒	44.0% (55)	36.0% (45)
	30 秒	6.4% (8)	7.2% (9)
	40 秒	-	0.8% (1)
	超過 50 秒	-	-
性別（n）	男性	46.4% (67)	53.6% (58)
	女性	53.6% (58)	46.4% (67)

負相關，即較高的準確度分數對應較低的發音錯誤率（PER），這表明發音評估確實受到錯誤發音檢測的影響。

4 實驗

4.1 資料集

為了驗證以我們提出的學習方法構建之模型的效能，並便於與先前研究進行比較，我們使用 Speechocean762 資料集 [25] 進行實驗。我們使用了整個資料，包含訓練集 2.5k 樣本與測試集 2.5k 樣本（見表 1）。

Speechocean762 資料集由母語為中文普通話的非英語母語者所產生的英語語音錄音組成。對於 APA 任務，該資料集提供五項句子層級的評估指標：accuracy、fluency、prosody、completeness 和 total score，每項指標的分數範圍為 0 到 10。我們在訓練與評估中排除了「completeness」指標，因為所有測試集的值均為 10，無法與模型預測進行有意義的相關性分析。

此外，Speechocean762 提供了補充的逐字稿，標示出發音錯誤的音素，可用於 MDD 模型訓練。這些註釋使用來自 CMUDict 標準 [22] 的 46 單位音素集合。該音素集合包含以 ARPABET 格式表示的 39 個音素、一個表示未知音素的 <unk> 標記，以及六個專門用於捕捉第二語言（L2）發音模式的額外音素。

4.2 模型配置

Phi-4-multimodal-instruct model [1] 的結構包含一個語音編碼器、投影器，以及連接到大型語言模型 (LLM) 的 LoRA 適配器。根據該研究，語音編碼器和投影器已在大量多語種語音資料上充分訓練，無需進行額外訓練 [1]。研究主張僅需微調 LoRA 適配器即可執行特定任務。因此，在本實驗中，我們計劃透過對模型進行 LoRA 微調，建立一個統一的 APA 與 MDD 系統。微調時我們使用了一張 NVIDIA A100 SXM（80GB VRAM）GPU。在批次大小為 8 的情況下，我們將梯度累積步驟設為 8，使用 Adam 優化器，初始學習率為 $2 \times 10 - 5$ 。

Table 2: Table 2: Training Values

類型	鍵	資料類型	值
APA	準確度	整數	發音的準確性。子音與母音的準確性。
	韻律	整數	韻律流暢度。語調、重音、節奏。語速。說話中的停頓。
	流暢度	整數	講話者在語流中是否自然流暢，沒有明顯停頓、重複或結巴的情況。
	總計	整數	對發音品質的整體評估，考量語音的所有面向。
MDD	音素稿本	字串	模型辨識輸入語音後的音素層級文字記錄（CMUDict 語料庫）。
	單字稿本	字串	模型識別輸入語音後逐字轉錄的文本。

如表 2 所示，模型被設定為預測多項評估指標。對於 APA 任務，模型被訓練以測量每句話的音素準確度、韻律品質、流暢度及總分。此外，在 MDD 任務中，模型對相同的語句執行了音素及正字法的轉錄。傳統的 MDD 方法通常僅著重於音素層級的辨識以找出發音錯誤，但我們的方法有所不同。由於我們的預訓練模型已具備強大的自動語音辨識能力，我們設計它同時識別正確句子（正字法轉錄）以及實際聽到的音素（音素轉錄）。

根據表 2 所述的配置，我們為 APA 和 MDD 任務構建了專用的監督式微調（SFT）提示（參見附錄 7.1 和 7.2）。依照 Phi-4 訓練腳本格式，包含控制標記與音訊標記的提示構成了每個訓練對的「user」部分（見圖 3），而訓練值（表 2）則使用提示中指定的相同 JSON 格式轉換為「assistant」部分。這些 user-assistant 配對，連同其對應的向量化音訊檔，構成了完整的 SFT 訓練實例。

4.3 評估

為評估模型的 APA 表現，我們使用皮爾森相關係數（PCC）來衡量人類與模型預測發音分數之間的相關性。對於 MDD 評估，我們以字錯率（WER）評估模型的語音辨識表現。此外，針對音素辨識結果，我們透過將模型輸出與參考音素轉錄比對，計算音素錯誤率（PER）與 F1 分數以進行評估。

表 3：表 3：訓練結果。所有 APA 任務的 **p** 值均小於 0.05，除非為底線標示的值

訓練語音層 epoch		APA 任務 - PCC 值				MDD 任務 指標				
		準確率	流暢度	韻律	總分	WER	PER	F1 分數	精準度	召回率
未訓練	0	-0.041	$\underline{-0.017}$	-0.0493	-0.104	0.97	0.792	0.143	0.154	0.134
LoRA	1	0.547	0.585	0.567	0.544	0.15	0.137	0.686	0.689	0.682
LoRA	2	0.637	0.726	0.709	0.662	0.139	0.118	0.722	0.725	0.719
LoRA	3	0.656	0.727	0.711	0.675	0.14	0.114	0.724	0.728	0.721
LoRA	4	0.645	0.733	0.714	0.668	0.148	0.121	0.721	0.723	0.72
解凍	1	0.535	0.559	0.564	0.55	0.159	0.199	0.575	0.579	0.57
解凍	2	0.624	0.668	0.653	0.634	0.145	0.154	0.651	0.651	0.651
解凍	3	0.621	0.669	0.655	0.637	0.148	0.15	0.663	0.665	0.662
解凍	4	0.743	0.717	0.704	0.666	0.142	0.142	0.667	0.671	0.663

衡量誤讀偵測的準確性。

$$\text{WER} = \frac{I + D + S}{N}$$
$$\text{PER} = \frac{I + D + S}{N}$$

WER 指標量化語音辨識模型的效能。在 WER 公式 (1) 中，I 代表插入錯誤的數量，D 代表刪除錯誤的數量，S 代表替換錯誤的數量，N 代表參考文字中的總詞數。將這些錯誤數相加後除以總詞數即可得到錯誤率。類似地，PER (2) 使用與 WER 相同的方法計算，但在音素層級而非詞層級進行。

$$\text{Precision} = \frac{TP}{TP + FP}$$
$$\text{Recall} = \frac{TP}{TP + FN}$$
$$\text{F1 score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

在精確率與召回率的公式 (3 與 4) 中，TP 代表真正陽性（正確識別的發音錯誤），FP 代表偽陽性（錯誤標記為發音錯誤的正確發音），FN 代表偽陰性（遺漏的發音錯誤）。接著使用這些指標計算 F1 分數 (5)，以提供模型發音錯誤偵測能力的平衡性衡量。

4.4 結果

表 3 彙整了 APA 與 MDD 任務的訓練結果。兩種微調策略相較於未訓練的基準都有顯著改善，基準顯示出高初始錯誤率（WER 0.97、PER 0.792）以及低偵測表現（F1 分數 0.143）。在所有指標上，LoRA 方法持續優於 Unfreeze 策略。LoRA 方法在第 2 個 epoch 達到最低的詞錯誤率（WER 0.139），並取得最佳表現

表 4：表 4：與其他研究之 PCC 結果比較。

Model	準確性	流暢度	韻律的	總計	PER
LoRA（四個世代）	0.645	0.733	0.714	0.668	0.121
解凍（四個世代）	0.743	0.717	0.704	0.666	0.142
GOPT [13]	0.714	0.753	0.760	0.742	-
Data2vec2 [11]	0.713	0.777	-	-	-
Joint-CAPT-L1 [20]	0.719	0.775	0.773	0.743	0.099
Azure PA [21]	0.7	0.715	0.842	0.782	-

在第 3 個訓練週期取得音素錯誤率（PER）0.114、F1 分數 0.724、精確度 0.728 和召回率 0.721。雖然 Unfreeze 策略在第 4 個週期前表現持續提升，其最佳成績（字錯率 WER 0.142、PER 0.142、F1 分數 0.667）仍不及 LoRA 方法所達到的水準。這些結果顯示，

在本實驗中，LoRA 微調（特別是在第 2 和第 3 個週期附近）比對整個音訊層進行完全微調更適合用於 MDD 任務。

表 4 比較了我們在 Speechocean762 資料集上經微調模型（特別是以 4 個訓練週期評估的 LoRA 與 Unfreeze 策略）與文獻中數種既有方法的表現。一個顯著的成就是我們的 Unfreeze 模型在準確度上達到最新的皮爾森相關係數

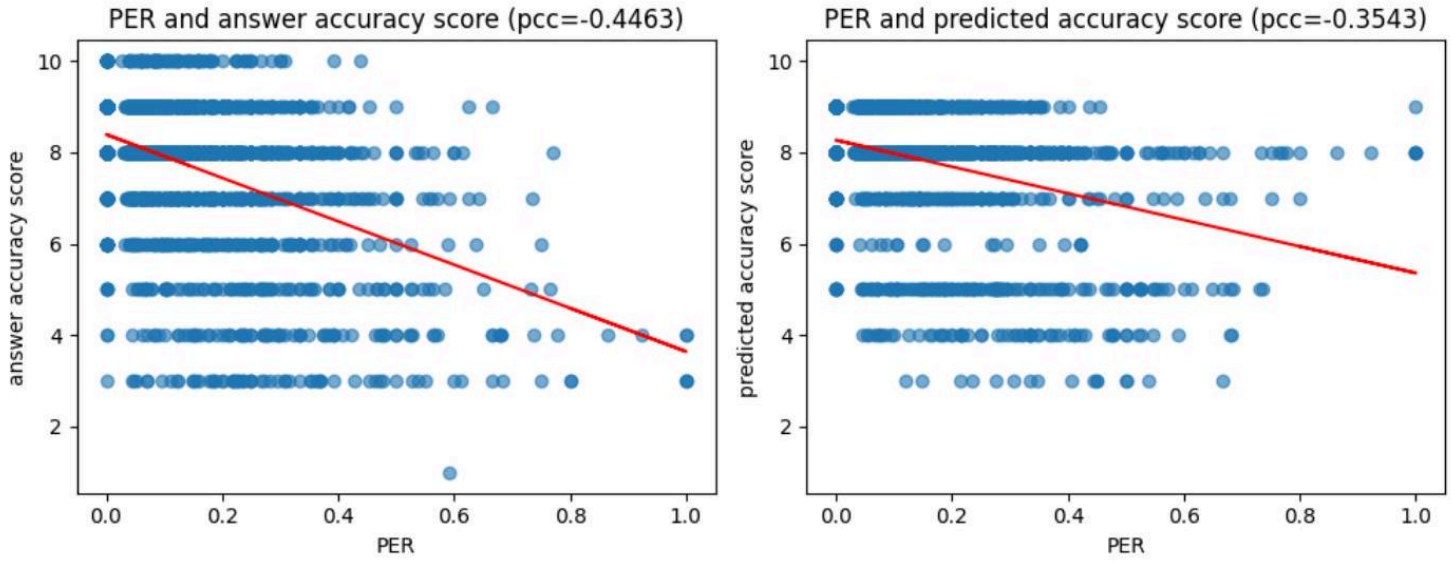

圖5：圖5：準確度分數與音素錯誤率之間的皮爾森相關係數。

（PCC）為 0.743，超越了所有其他列出的基準，包括 GOPT [13]、Data2vec2 [11]、Joint-CAPT-L1 [20] 與 Azure PA [21] 在此特定維度上的表現。

然而，此一在準確性評估上的優勢並未在其他發音評估面向上普遍顯現。在流暢度方面，Data2vec2 與 Joint-CAPT-L1 報告的皮爾森相關係數（PCC）分別為 0.777 與 0.775，均高於我們的模型（LoRA：0.733、Unfreeze：0.717）。類似地，在韻律與整體評分（總分）方面，Azure PA 服務展現了更高的相關性（韻律 PCC 0.842、總分 PCC 0.782）。GOPT（0.742）與 Joint-CAPT-L1（0.743）的總分 PCC 也都高於我們的 LoRA（0.668）與 Unfreeze（0.666）模型。此外，針對以音素錯誤率（PER）衡量的 MDD 面向，[20] 研究達成了較低（較佳）的 PER 0.099，顯示其音素層級辨識準確度高於我們的 LoRA 模型（0.121）與 Unfreeze 模型（0.142）。

總結而言，如表 4 所示，我們提出的微調方法——尤其是 Unfreeze 策略——在 Speechocean762 基準上的音素層級準確度評分（Accuracy PCC）達到最先進的表現。然而，在評估流暢度與韻律等更廣泛面向時，現有的專業系統與商業服務仍優於我們的模型。這些成熟的解決方案在全面發音評估上展現更強的能力，導致更高的整體相關性分數與更低的音素辨識錯誤率。

圖 5 呈現了 PER 與發音準確度分數之間的相關性分析。左側圖顯示 PER 與答案（人工）準確度分數之間的關係，表現出中度負相關（Pearson 相關係數為 -0.4463）。這表示較低的 PER 值通常對應較高的答案準確度分數。右側圖顯示的預測準確度分數也觀察到類似趨勢。雖然相關性略微較弱（-0.3543），但仍確認了 PER 與預測準確度分數之間明顯的負向關係。因此，兩項分析一致地表明，較佳的音素辨識（較低的 PER）與較高的實際答案準確度及模型預測之準確度評估相關。

5 討論

5.1 貢獻

本研究的意義在於展示能夠建構一個以多模態大型語言模型（MLLM）為基礎的發音評估系統，並僅透過微調 MLLM 的部分層就同時執行 APA 與 MDD。透過為訓練與推論設計提示（prompt）而非採用複雜流程，即可用少量訓練資料建立發音評估系統。利用模型即使在體積龐大下僅訓練與語音相關的部分層仍能達到足夠表現的特性，我們得以降低不必要的訓練成本。此外，拼字與音素單位不需分別訓練，也減少了不必要的訓練步驟。由於 ARPABET 的發音符號由字母構成，我們能夠透過將其作為提示，在不需訓練或新增音素標記的情況下，有效率地建立同時執行拼字與音素層級識別的系統。

雖然本實驗使用僅由以中文為母語的受試者構成的英語語料來建立模型，但要為具有各種母語的英語學習者建立發音評估系統並不困難。由於 Phi-4 的 LLM 在 23 種語言的文字資料上學習，而其語音層則在 8 種語言的語音資料上學習，因此相比從頭訓練以英語為外語的學習者語音資料，使用較小的資料集與簡單的訓練方法便有可能建立適合不同目標語者的發音評估模型。

5.2 限制

儘管這些結果令人振奮，我們的方法仍有幾項重要的限制。主要限制來自運算資源的限制，阻礙了我們進行完整的微調實驗。完整微調涉及更新整個模型架構中的參數，包括語音元件與核心 LLM，這可能比僅修改特定層或使用 LoRA 提供更佳的調適能力。儘管既有研究（例如 [1]）與我們的結果顯示 LoRA 微調能達到足夠的效能，但要進行全面的評估與比較，仍需實施完整微調。
這些實驗在我們有限的運算資源下是不可行的。

另一個重要的限制在於我們的發音評估方法缺乏超音段（韻律）資訊。我們無法為模型建立從資料集中學習重音與語調等韻律特徵的指標。因此，我們無法驗證韻律與流暢度評分方法的有效性，例如檢驗準確度分數與音素錯誤率（PER）之間的相關性。然而，若使用專門設計以捕捉超音段資訊的資料集，模型便有可能對韻律與流暢度分數的計算提供解釋。

5.3 進一步研究

本研究衍生出數個未來研究方向。首先，本模型的運算需求在實際部署上帶來挑戰。模型約有 58 億個參數，推理速度較慢且需仰賴高效能 GPU。在對回應速度與成本效益有嚴格要求的教育場域中，要實務化並商業化，透過量化來優化模型變得至關重要。雖然官方文件中提供了量化腳本，但仍需進行完整實驗，比較量化前後的效能，以評估模型效率與準確度之間的取捨。

其次，完整的發音評估不僅需檢視音素準確度，還要評估重音、語調與節奏等韻律要素。我們目前的方法僅用單一的韻律分數來評估韻律特性，這只能有限地反映特定韻律特徵（如重音、語調與節奏）。未來研究應著重於從訓練資料中擷取更詳細的韻律資訊，使模型能同時理解並評估音素準確度與韻律要素。此一強化將使多模態語言模型得以提供更進階的語音分析，並生成針對韻律模式的詳細說明。

第三，Phi-4-multimodal-instruct 模型[1] 的多模態能力為整合式分析提供了有前景的機會。該模型支援視覺與語音編碼器的聯合訓練，能同時處理視覺與聽覺資訊。透過微調此架構以同時識別發音模式與唇動，我們可以達成對語音產生之更高維度的理解與分析。此類方法將在統一框架中同時利用聲學與視覺線索，可能提升發音評估的準確性，並為學習者提供更全面的回饋。

6 結論

本研究建立了一個統一的電腦輔助發音訓練（CAPT）框架，將自動發音評估（APA）與誤讀偵測與診斷（MDD）整合於單一的語音多模態大型語言模型（MLLM）中。透過對預訓練 Phi-4-multimodal-instruct 模型進行高效的低秩適配（LoRA）微調，我們達成了與人類評估高度相關的發音評估表現，同時維持高準確度
在語音與音素識別任務中的應用。此項工作提供了一個實用且統一的框架，降低了開發先進發音訓練技術的技術門檻。

7 附錄

7.1 APA 提示

```
Rate the pronunciation of the audio.
**Accuracy**
Score range: 0-10
* 9-10: The overall pronunciation of the sentence is
    excellent, with accurate phonology and no obvious
    pronunciation mistakes
* 7-8: The overall pronunciation of the sentence is good,
    with a few pronunciation mistakes
* 5-6: The overall pronunciation of the sentence is
    understandable, with many pronunciation mistakes and
    accent, but it does not affect the understanding of
    basic meanings
* 3-4: Poor, clumsy and rigid pronunciation of the sentence
    as a whole, with serious pronunciation mistakes
* 0-2: Extremely poor pronunciation and only one or two
    words are recognizable
**Fluency**
Score range: 0-10
* 8-10: Fluent without noticeable pauses or stammering
* 6-7: Fluent in general, with a few pauses, repetition,
    and stammering
* 4-5: The speech is a little disfluent, with many pauses,
    repetition, and stammering
* 0-3: Intermittent, very disfluent speech, with lots of
    pauses, repetition, and stammering
```

```
**Prosodic**
Score range: 0-10
* 9-10: Correct intonation at a stable speaking speed,
    speak with cadence, and can speak like a native
* 7-8: Nearly correct intonation at a stable speaking
    speed, nearly smooth and coherent, but with little
    stammering and few pauses
* 5-6: Unstable speech speed, many stammering and pauses
    with a poor sense of rhythm
* 3-4: Unstable speech speed, speak too fast or too slow,
    without the sense of rhythm
* 0-2: Poor intonation and lots of stammering and pauses,
    unable to read a complete sentence
**Total**
Score range: 0-10
Provide an overall assessment of the pronunciation quality
    considering all aspects of the speech. This should
    reflect your holistic evaluation of the speaker's
    pronunciation abilities based on the entire audio
    sample.
* 9-10: Excellent overall pronunciation that sounds nearly
    native-like
* 7-8: Good pronunciation with minor issues that don't
    affect comprehension
```

```
* 5-6: Fair pronunciation with noticeable non-native
    features but generally understandable
* 3-4: Poor pronunciation that requires effort to understand
* 0-2: Very poor pronunciation that is largely
    incomprehensible
Provide the results in the following JSON format:
{'accuracy': ACCURACY_SCORE, 'fluency': FLUENCY_SCORE,
    'prosodic': PROSODIC_SCORE, 'total': TOTAL_SCORE}
```

7.2 MDD 提示

將語音語句逐字轉寫，並同時提供字詞層級的轉錄與音素層級的解析。

對於音素解析，請使用 CMU 發音詞典格式（例如：AA、IH）。
若某個單字或音素不清楚，請標註為 '<unk>'。

請以以下 JSON 格式提供結果：
{'word_transcript': 'That's an interesting observation.', 'phoneme_transcript': 'DH EH S AX N IH N T AX R EH S T IH NG AA B Z AX R V EY IH SH AX N'}

參考文獻

[1] A. Abouelenin, A. Ashfaq, A. Atkinson, H. Awadalla, N. Bach, J. Bao, et al. 2025. Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs. arXiv preprint arXiv:2503.01743 (2025).
[2] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli. 2022. XLS-R: Self-supervised cross-lingual speech representation learning at scale. In Proc. Interspeech 2022. 2278-2282. doi:10.21437/Interspeech.2022-143
[3] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. 2020. Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In Advances in Neural Information Processing Systems, Vol. 33. 12449-12460.
[4] J. Bernstein, M. Cohen, H. Murveit, D. Rtischev, and M. Weintraub. 1990. Automatic evaluation and training in English pronunciation. In Proceedings of the ICSLP-90: 1990 International Conference on Spoken Language Processing. 1185-1188.
[5] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. In Proc. Interspeech 2021. 2426-2430. doi:10.21437/Interspeech.2021-329
[6] M. Eskenazi. 2009. An overview of spoken language technology for education. Speech Communication 51, 10 (2009), 832-844.
[7] Y. Feng, G. Fu, Q. Chen, and K. Chen. 2020. SED-MDD: Towards sentence dependent end-to-end mispronunciation detection and diagnosis. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 3492-3496.
[8] J. Fouz-González. 2015. Trends and directions in computer-assisted pronunciation training. In Investigating English pronunciation: Trends and directions. 314-342.
[9] H. Franco, V. Abrash, K. Precoda, H. Bratt, R. Rao, and J. Butzberger. 2000. The SRI EduSpeak system: recognition and pronunciation scoring for language learning. In Proceedings of Intelligent Speech Technology in Language Learning, InSTiLL2000.
[10] H. Franco, L. Neumeyer, V. Digalakis, and O. Ronen. 2000. Combination of machine scores for automatic grading of pronunciation quality. Speech Communication 30 (2000), 121-130.
[11] K. Fu, L. Peng, N. Yang, and S. Zhou. 2024. Pronunciation Assessment with Multi-modal Large Language Models. arXiv preprint arXiv:2407.09209 (2024).
[12] Y. Getman, R. Al-Ghezi, K. Voskoboinik, T. Grósz, M. Kurimo, G. Salvi, T. Svendsen, and S. Strömbergsson. 2022. Wav2vec2-Based Speech Rating System for Children with Speech Sound Disorder. In Interspeech 2022. ISCA.
[13] Y. Gong, Z. Chen, I. H. Chu, P. Chang, and J. Glass. 2022. Transformer-based multiaspect multi-granularity non-native english speaker pronunciation assessment. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 7262-7266.
[14] J. Heo. 2024. Practical AI application development using LLMs. onlybook.
[15] Y. Hong. 2021. 為母語非英語者自動化評估英語發音。博士論文。高麗大學研究生院。
[16] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card。arXiv preprint arXiv:2410.21276 (2024)。
[17] Microsoft. 2025. Phi-4-multimodal-instruct。Hugging Face。https://huggingface. co/microsoft/Phi-4-multimodal-instruct
[18] L. Peng, K. Fu, B. Lin, D. Ke, and J. Zhang. 2021. 對 wav2vec2.0 模型在誤讀檢測與診斷任務上微調的研究。發表於 Interspeech。ISCA。
[19] X. Qian, H. Meng, and F. Soong. 2010. 使用 DBN-HMM 於第二語言英語的發音錯誤偵測與診斷，以支援電腦輔助發音訓練。收錄於 INTERSPEECH 會議論文。775-778。
[20] H. Ryu, S. Kim, and M. Chung. 2023. 一個結合發音評估與發音錯誤偵測與診斷的多任務學習聯合模型。收錄於 Interspeech 2023 會議論文。959-963。doi:10.21437/Interspeech.2023-337
[21] K. Wang, L. He, K. Liu, Y. Deng, W. Wei, and S. Zhao. 2025. 探索大型多模態模型作為發音評估有效替代方案的潛力。arXiv 預印本 arXiv:2503.11229 (2025)。
[22] R. Weide et al. 1998. 卡內基美隆發音詞典（版本 0.6）。取自 www.cs.cmu.edu。
[23] S. M. Witt 與 S. J. Young。2000。用於互動式語言學習的音素層級發音評分與評估。Speech Communication 30, 2-3 (2000), 95-108。
[24] M. Yang、K. Hirschi、S. D. Looney、O. Kang 與 J. H. L. Hansen。2022。使用基於 Wav2vec2 的動量偽標籤提升語音口音與可懂度評估的誤讀檢測。在 Interspeech 2022。ISCA。
[25] J. Zhang、Z. Zhang、Y. Wang、Z. Yan、Q. Song、Y. Huang 等。2021。speechocean762：一個用於發音評估的開源非母語英語語音語料庫。arXiv preprint arXiv:2104.01378 (2021)。