超越模態限制：一種統一的多模態大型語言模型方法，透過有效的課程學習實現自動口語評估

方禹軒、羅天宏、宋曜廷、陳柏琳
國立臺灣師範大學
{andyfang, teinhonglo, sungtc, berlin} @ntnu.edu.tw

摘要

傳統的自動口語評估（ASA）系統存在固有的模態限制：基於文本的方法缺乏語音資訊，而基於音訊的方法則遺漏語義上下文。多模態大型語言模型（MLLM）透過在統一框架內同時處理音訊和文本，為全面的 ASA 提供了前所未有的機會。本文首次系統性地研究了用於全面 ASA 的 MLLM，展示了 MLLM 在內容和語言使用方面的卓越性能。然而，對表達方面的評估揭示了獨特的挑戰，這被認為需要專門的訓練策略。因此，我們提出了「語音優先多模態訓練」（Speech-First Multimodal Training, SFMT），利用課程學習原則在跨模態協同融合之前建立更穩固的語音建模基礎。在一系列基準資料集上的實驗表明，基於 MLLM 的系統可以將整體評估性能從 0.783 的 PCC 值提升到 0.846。特別是，SFMT 在表達方面的評估中表現出色，相較於傳統訓練方法，實現了 $\mathbf{4\%}$ 的絕對準確度提升，這也為 ASA 開闢了一條新途徑。

索引詞彙—多模態大型語言模型（MLLM）、自動口語評估（ASA）、多模態訓練、第二語言能力、跨模態學習

一、前言

多模態大型語言模型（MLLM）的最新進展，開啟了前所未有的技術轉型時代，透過在多種模態之間共同整合資訊，從根本上重塑了人機互動的典範 [1]-[4]。諸如 GPT-4o [5] 等開創性努力，已展現出在統一框架內無縫處理文本、音訊和視覺輸入的卓越能力。特別值得注意的是，開源 MLLM（例如 Phi-4-multimodal [3]）的出現，在針對特定領域資料進行模型微調後 [6]-[10]，已在專業語言評估任務中展現出優於傳統單模態方法的性能。這種卓越的多模態能力也為解決以往傳統方法無法觸及的複雜現實世界應用開闢了新途徑。

在電腦輔助語言學習（CALL）領域中，自動口語評估（ASA）是其中最具挑戰性且多面向的任務之一 [11], [12]。評估第二語言（L2）口語能力的複雜性，源於需要同時評估口語能力的各個方面，包括
在傳遞（例如，發音準確性、流暢度、韻律特徵）、內容適當性（例如，主題相關性和連貫性）和語言使用（例如，詞彙豐富度和語法正確性）方面進行評估 [11], [13]。這些評估標準涵蓋了可量化的語言元素和細微的聲學特徵，例如重音模式、語調輪廓和語速 [14], [15]。口語評估的多維性質，加上第二語言口語產出固有的變異性，使得 ASA 系統成為現代語言學習環境中不可或缺的組成部分，提供客觀、一致且可擴展的評估能力，以補充人工評估 [12]。

然而，傳統的 ASA 方法存在根本性的模態特定限制，這限制了其有效性。以 BERT 為基礎的系統 [6], [8] 為例的文本分類器，在語義理解和上下文理解方面表現出色，但仍然嚴重依賴 ASR 轉錄品質，並且本質上缺乏對傳遞和韻律評估至關重要的聲學特徵的存取。相反，利用 wav2vec 2.0 [7], [9] 等自監督學習模型的音訊方法直接處理語音訊號，以捕捉豐富的聲學資訊進行傳遞評估，但卻犧牲了對評估語言使用複雜性和語法準確性至關重要的語義上下文和語言內容分析。儘管先前的研究已經探索了結合兩種模態的融合策略 [13]，但這些方法通常融合了獨立單模態系統的輸出，而不是實現統一架構中發現的真正跨模態資訊同步。這種根本性的限制促使我們研究 MLLM 是否能夠超越傳統模態界限，實現更有效的多模態整合，以進行全面的 ASA。

本文首次對 MLLM 進行了全面的 ASA 系統研究，探討了三個關鍵問題：1) 多模態大型語言模型能否有效解決傳統 ASA 系統中遇到的資訊融合挑戰，以及可以達到什麼樣的效能水準？2) 儘管 MLLM 有所進步，音訊模態對於傳遞評估任務是否仍然不可替代？3) 是否存在簡單而經濟高效的訓練策略，可以顯著提高 ASA 在口語能力評估不同方面的效能？為此，我們設計了使用 TEEMI 資料集進行了徹底的實驗，並提出了語音優先多模態訓練 (SFMT)，這是一種課程學習方法 [16]，它逐步從語音基礎過渡到跨模態整合，在傳遞方面的評估準確性方面實現了 4% 的絕對改進。

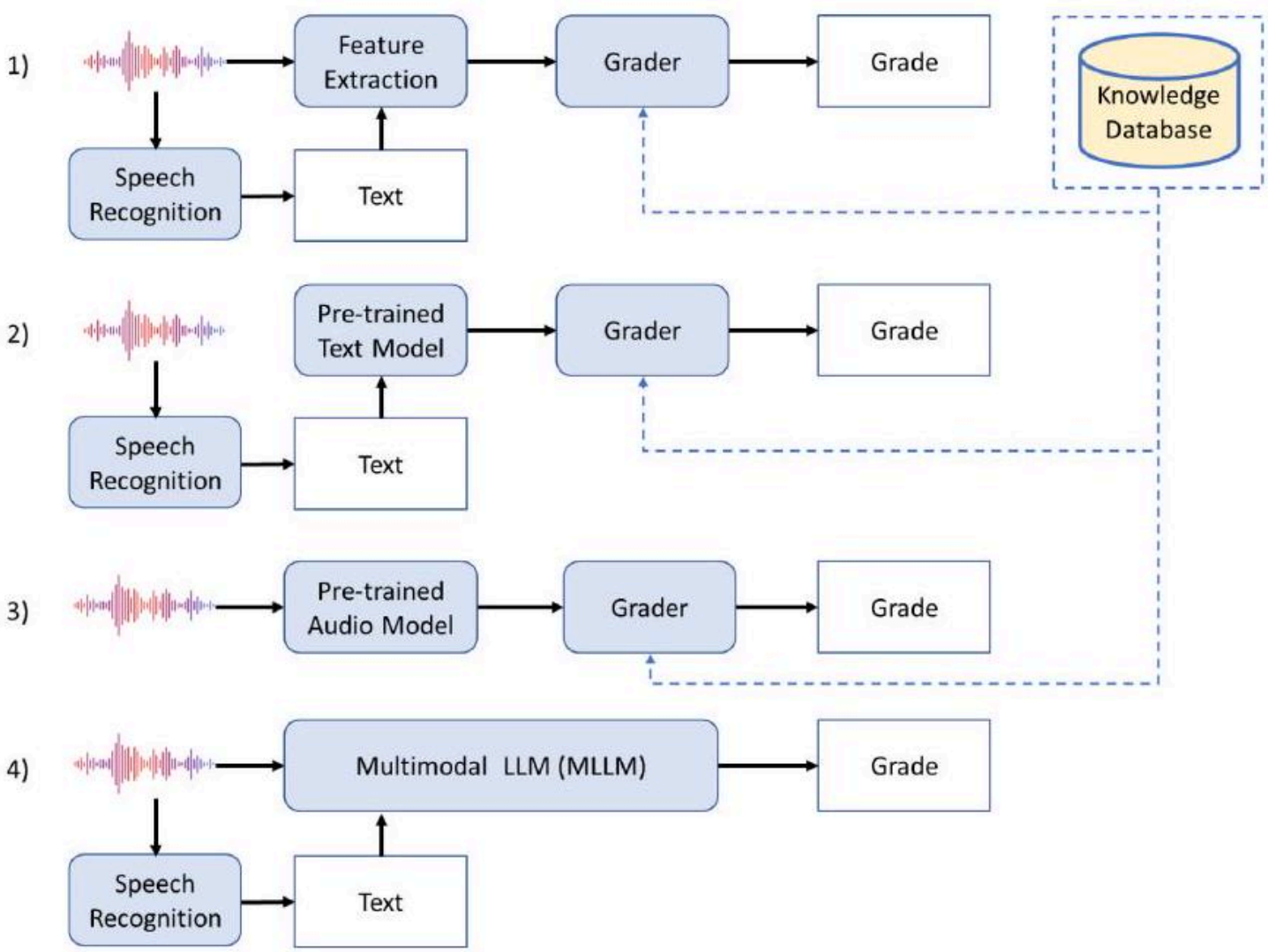


圖 1：圖 1. 自動口語評分系統已從手工特徵工程，透過自監督學習方法，演進到能夠進行全面評估和回饋生成的統一多模態框架（改編自 [14]）。

二、相關工作

A. 自動口語評分系統的演進

自動口語評分（ASA）已歷經三個截然不同的典範演進，每個典範都標誌著第二語言學習者口語能力評估自動化方面的根本性進展。圖 1 闡述了從手工特徵系統，透過自監督模型，到統一多模態框架的演進過程。

手工特徵系統：早期的 ASA 系統通常依賴明確的特徵工程流程（圖 1 (1)），從語音中提取手工聲學特徵（頻譜、韻律、時間），並從 ASR 轉錄中提取語言特徵 [17]。傳統機器學習演算法處理這些特徵以預測熟練度，其中美國教育測驗服務社（ETS）透過廣泛的特徵工程研究開創了基礎方法 [18][20]。最近，Wu 等人 [21] 表明，專家定義的知識線索（表達/語言使用標準）顯著提升了評估效能。儘管這些系統具有可解釋性，但其泛化能力有限，且需要大量的領域專業知識。
自監督學習範式：自監督學習透過基於文本或基於音訊的預訓練模型來處理 ASA（圖 1 (2) 和 (3)）。

基於文本的模型：基於 BERT 的模型能夠從 ASR 轉錄稿中進行複雜的語義評估（語法、語言使用、內容）[6]、[8]，但受限於 ASR 品質
並且缺乏用於評估表達方面的聲學資訊。

基於音訊的模型：自監督語音模型，例如 wav2vec 2.0，處理原始語音以捕捉聲學模式 [7]、[9]。Lo 等人 [11] 發現 wav2vec 2.0 本身就編碼了語法資訊，揭示了跨模態特徵提取的潛力。然而，它們缺乏用於全面評估的語義上下文。

這兩種方法在各種 ASA 任務上都取得了一些成功，但仍受限於模態限制。為了解決這個限制，先前的融合策略通常在模型層級操作，這將無法實現真正的跨模態同步 [13]。
3) 多模態大型語言模型：當代 MLLM 標誌著統一多模態處理的典範轉移（圖 1(4)）。像 Qwen-Audio [2]、SALMONN [1] 和 Phi-4-multimodal [3] 這樣的模型，在單一框架中同時處理語音和文本，透過跨模態注意力實現真正的多模態整合。

MLLM 透過提供超越分數的全面教育回饋，超越了傳統的評估限制。然而，如何設計最佳的多模態整合訓練策略，特別是針對需要細緻聲學分析的表達方面評估，仍然有待深入探索。

B. 多模態訓練的課程學習

課程學習（Curriculum learning）認為，從簡單到複雜任務的結構化進程能提升模型效能 [16]。近期多模態語音應用，例如 WavLLM [22] 和 SALMONN [1]，也證實了漸進式訓練在語音與文本聯合建模中的有效性。此外，Zhang 等人 [23] 透過策略性資料排序，將課程學習應用於口語評估，顯示在資料有限情境下的改進。然而，現有方法側重於資料層級的課程（依難度排序樣本），而非解決多模態整合中的根本挑戰。

我們的研究將課程學習的概念擴展到模態層級的進程，探討聲學資訊與文本資訊對於基於 MLLM 的 ASA 任務的相對重要性。我們提出了 SFMT，這是一種從簡單到複雜的學習方法，它首先建立穩固的聲學基礎，然後再處理跨模態整合。這種模態層級的課程方法專門解決了優化 MLLM 效能的問題，適用於需要有效整合聲學和語義資訊的細粒度評估任務，同時保留了對於準確能力評估至關重要的區辨能力。

三、方法論

A. 用於 ASA 的多模態大型語言模型架構

我們利用 Phi-4-multimodal [3] 進行全面的自動口語評估。此模型採用混合 LoRA 架構，可實現高效的多模態微調，同時保留基礎語言能力。如圖所示

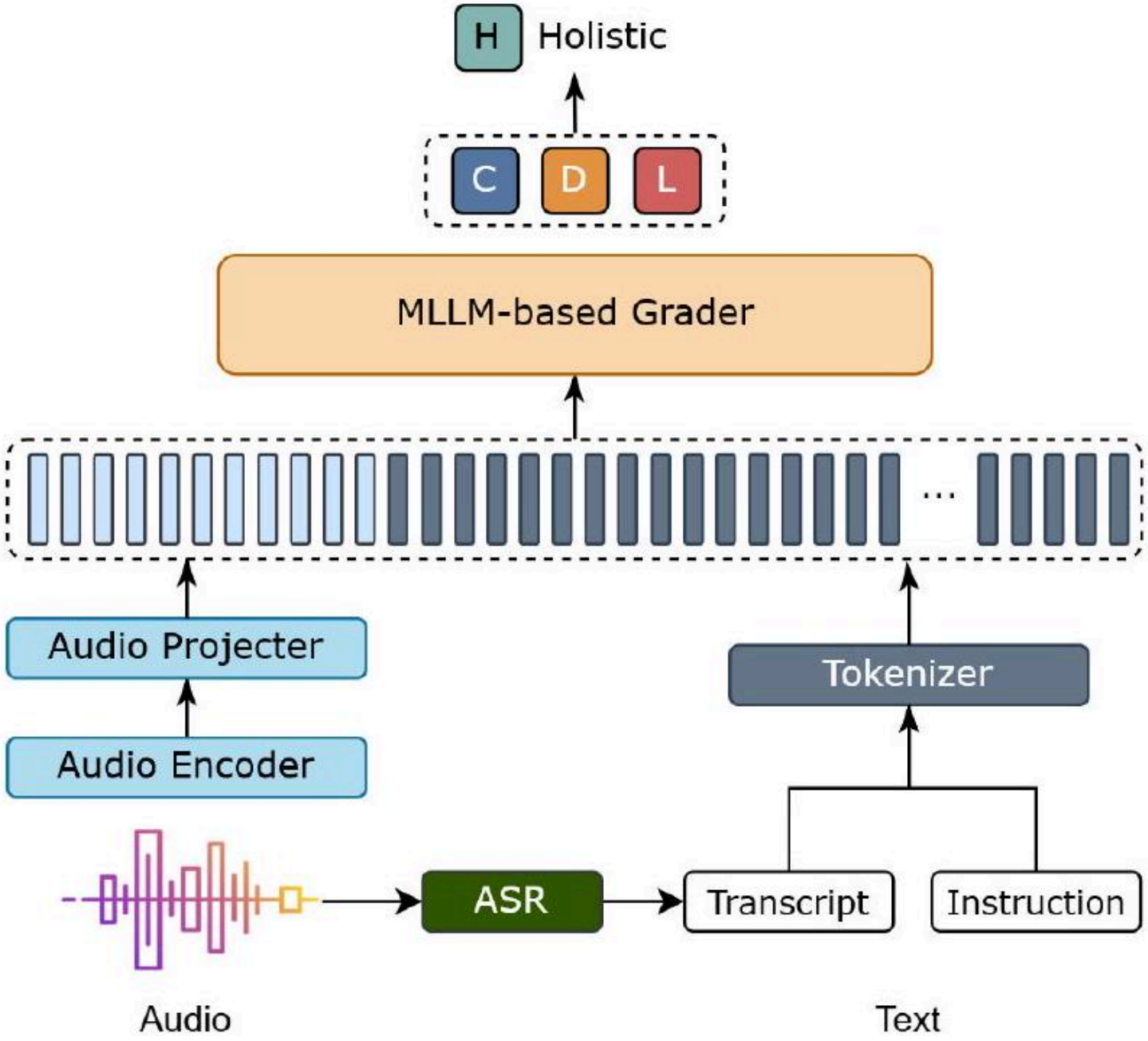


圖 2：圖 2。所提出的 MLLM 架構透過專門的途徑處理音訊和文字輸入，以在內容、表達、語言使用和整體評估方面產生多面向的熟練度分數。

圖 2 所示，該系統在整合之前，透過特定模態的途徑處理原始音訊和 ASR 生成的文字記錄，其中包括：(1) 一個 3.8 B 參數的僅解碼器 Transformer 作為推理骨幹，(2) 一個音訊處理管道，其中包含 460 M 參數的編碼器，使用 conformer 區塊和音訊投影器進行共享嵌入空間映射，以及 (3) 一個特定模態的音訊適配器（LoRA $_{\text{audio}}$，460M 參數），可學習目標聲學特徵，而不會干擾語言能力。

為了對 TEEMI 資料集的口語回答進行全面評估，我們訓練了三個專門模型，分別針對內容 (C)、表達 (D) 和語言使用 (L) 等方面。每個模型在訓練期間都會收到特定方面的指令，從而實現重點優化。整體 (H) 分數整合了從所有三個方面收集的評估結果，提供與 CEFR 標準一致的整體熟練度指標。

B. 語音優先多模態訓練 (SFMT) 策略

標準的多模態訓練方法在 ASA 中遇到一個根本性的挑戰：模態不平衡。這些方法的模型傾向於對文本特徵表現出系統性的偏好，因為文本特徵具有結構化的表示和計算效率，因此未能充分利用對於語音傳遞評估至關重要的聲學資訊 [24]。這種不平衡損害了模型學習細緻聲學模式的能力，包括發音準確性、流暢度變化和韻律特徵，而這些是文本表示本身無法編碼的。

我們透過系統性消融研究（第 $\mathrm{V}-\mathrm{B}$ 節）進行的實證調查揭示了一個反直覺的發現：與文本相比，音訊模態對於基於 MLLM 的評分器表現出卓越的學習效率，特別是在
語音傳遞方面的評估。在相同的訓練條件下，音訊表現出更強的初始性能和更快的收斂速度，這促使我們採用語音優先策略。

聲學學習的這種實證優勢源於三個基本因素：
(1) 資訊完整性：原始音訊訊號保留了語音資訊的完整頻譜——從語音細節到韻律輪廓——為 MLLM 提供了未經濾波的管道，以獲取熟練度評估所需的所有聲學證據。相較之下，ASR 轉錄的文字是一種有損轉換，會丟棄對表達評估至關重要的副語言特徵。
(2) 直接訊號存取：音訊輸入繞過了基於文字方法固有的錯誤傳播，提供了對真實聲學模式的直接存取。這消除了 ASR 轉錄錯誤的連鎖效應，以及主要針對母語語音訓練的 ASR 系統所產生的系統性偏差。
(3) 優先學習模式：當同時接觸兩種模態時，模型會表現出對基於文字特徵的優先最佳化，將其視為計算效率高的途徑 [25]，特別是對於內容和語言使用評估。這種偏好抑制了聲學辨別能力的發展，因為模型會收斂於未充分利用聲學資訊的解決方案。

基於這些洞察，我們提出了語音優先多模態訓練（Speech-First Multimodal Training, SFMT），這是一種兩階段課程學習策略，利用了所發現的學習層次結構。透過在引入文字資訊之前建立強大的聲學特徵提取能力，SFMT 確保模型能發展出強大的表達評估能力，並在後續的多模態整合中持續發揮作用（圖 3）：

階段 1 – 聲學基礎（圖 3(a)）：給定訓練資料 $\mathcal{D}_{\text{audio}} = \{(\mathbf{a}_i, I_i, y_i)\}_{i=1}^{N}$，其中 $\mathbf{a}_i$ 是音訊輸入向量，$I_i \in \{I_C, I_D, I_L\}$ 是特定面向的指令，而 y_i 是目標分數，我們最佳化：

$$\theta_{\text{LoRA}}^{1} = \arg\min_{\theta_{\text{LoRA}}} \sum_{(\mathbf{a},I,y)\in\mathcal{D}_{\text{audio}}} \mathcal{L}\left(f_{\text{Phi-4}}\left(\mathbf{a}, I; \theta_{\text{LoRA}}\right), y\right),$$

其中 $f_{\text{Phi-4}}$ 表示 MLLM，而 $\mathcal{L}$ 是損失函數。僅更新 LoRA 音訊轉接器參數 θ_{LoRA}。

階段 2 – 跨模態整合（圖 3(b)）：使用包含額外轉錄向量 $\mathbf{t}_i$ 的多模態資料 $\mathcal{D}_{\text{multi}} = \{(\mathbf{a}_i, \mathbf{t}_i, I_i, y_i)\}_{i=1}^{N}$，我們從階段 1 繼續最佳化：

$$\theta_{\text{LoRA}}^{2} = \arg\min_{\theta_{\text{LoRA}}^{1}} \sum_{(\mathbf{a},\mathbf{t},I,y)\in\mathcal{D}_{\text{multi}}} \mathcal{L}\left(f_{\text{Phi-4}}\left(\mathbf{a}, \mathbf{t}, I; \theta_{\text{LoRA}}^{1}\right), y\right),$$

其中 θ_{LoRA} 是預訓練的轉接器。這種進程確保在多模態整合之前，能有穩固的聲學專業化，特別是增強了評估在傳遞方面的表現。

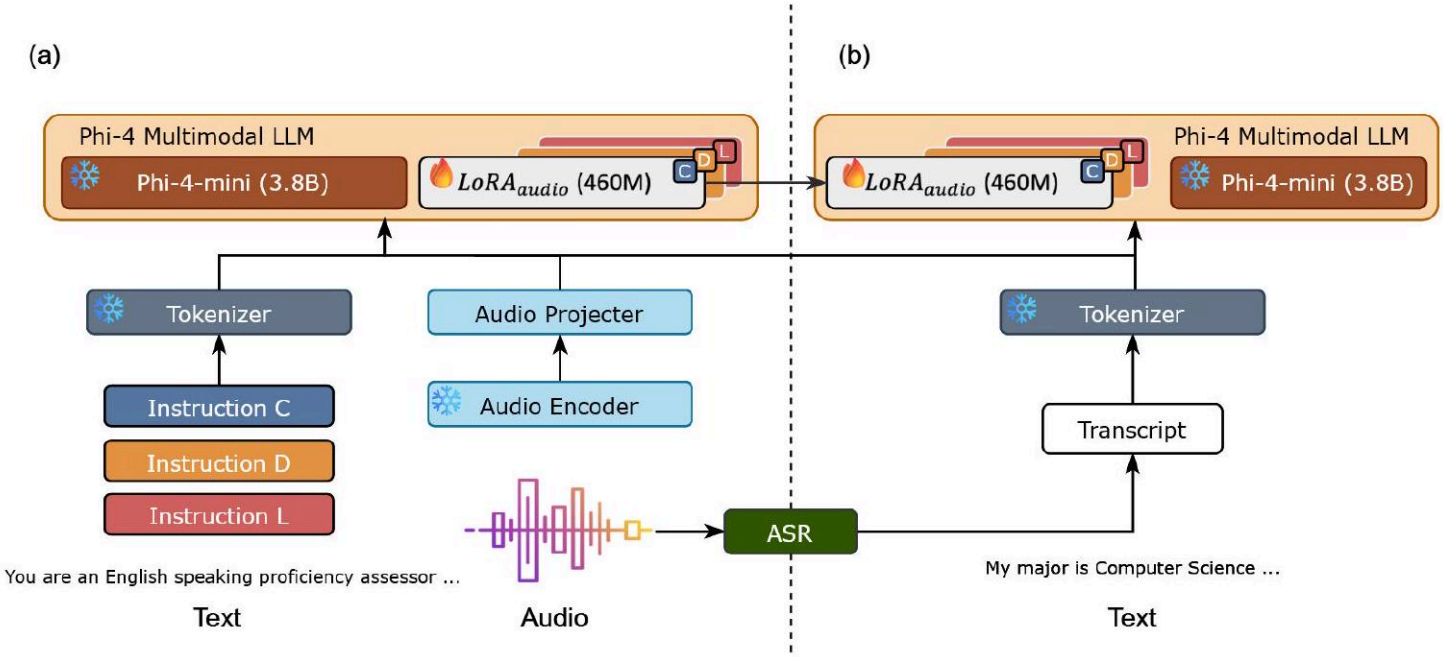


圖 3：圖 3. SFMT 採用兩階段課程學習方法，首先透過純音訊訓練建立聲學基礎，然後再引入與文字資訊的跨模態整合。

肆、實驗

A. 資料集

我們在兩個不同的資料集上評估我們提出的模型：專有的 TEEMI 語料庫和公開的 Speak & Improve 語料庫。

TEEMI 語料庫：TEEMI 語料庫（Test for English Medium Instruction）[26] 是一個綜合性的第二語言能力資料集，專為高等教育背景下的 EMI 研究而設計。該語料庫收錄了大學部和研究所第二語言學習者的自發性英語口語，每個回答都使用八級 CEFR 對齊量表（PreA1 到 B2）從四個方面進行評估：整體、內容、語言使用和表達。TEEMI 配備了三位評分者的註釋，並採用多數決以確保評分可靠性。

TEEMI 的口語評估包括三種任務形式：一般聽力與回答 (A)、情境問答 (B) 和主題問答 (C)。在本文中，我們專注於包含任務 A01、A02 的子集，總計 8,214 個回答。模型訓練和驗證僅在 A01 上進行，其中包含來自 1,231 位說話者的 6,152 個回答。A02 任務被保留下來，用於評估模型對先前未見過提示的泛化能力。本研究中使用的 A01 和 A02 任務的詳細 CEFR 等級分佈如表一所示。

表 1：表一
TEEMI 資料集中選定 CEFR 能力等級（A01、A02）的統計資訊。

任務	用法	前 A	A1	A1+	A2	A2+	B1	B1+	B2
A01	訓練	34	61	76	156	150	169	79	65
	驗證	8	16	19	38	39	43	23	12
	測試	11	20	23	49	50	48	32	15
A02	未見	9	7	12	19	12	26	23	15
總計	-	62	104	130	262	251	286	157	107

SLaTE 2025 Speak & Improve 語料庫：我們使用 Speak & Improve 語料庫 2025 [27]，其中包含 315 小時的第二語言英語語音，CEFR 能力等級從 A2 到 C1+。該語料庫包含四種任務類型：訪談、意見、簡報和溝通活動，並配有不同面向的平均整體分數。我們遵循官方資料分割來建構相應的訓練、開發和測試集。

B. 實作細節

模型配置是使用 Phi-4-multimodal-instruct [1] 進行初始化，並將 LoRA 適應 [29]（rank=320）應用於音訊編碼器。訓練採用 AdamW 優化器（$\mathrm{lr} = 4\mathrm{e} - 5$）進行 3 個 epoch，批次大小為 32（梯度累積步驟：16），並在單一 NVIDIA RTX 3090 上使用 bfloat16 混合精度。Flash attention [30] 用於提高記憶體效率。

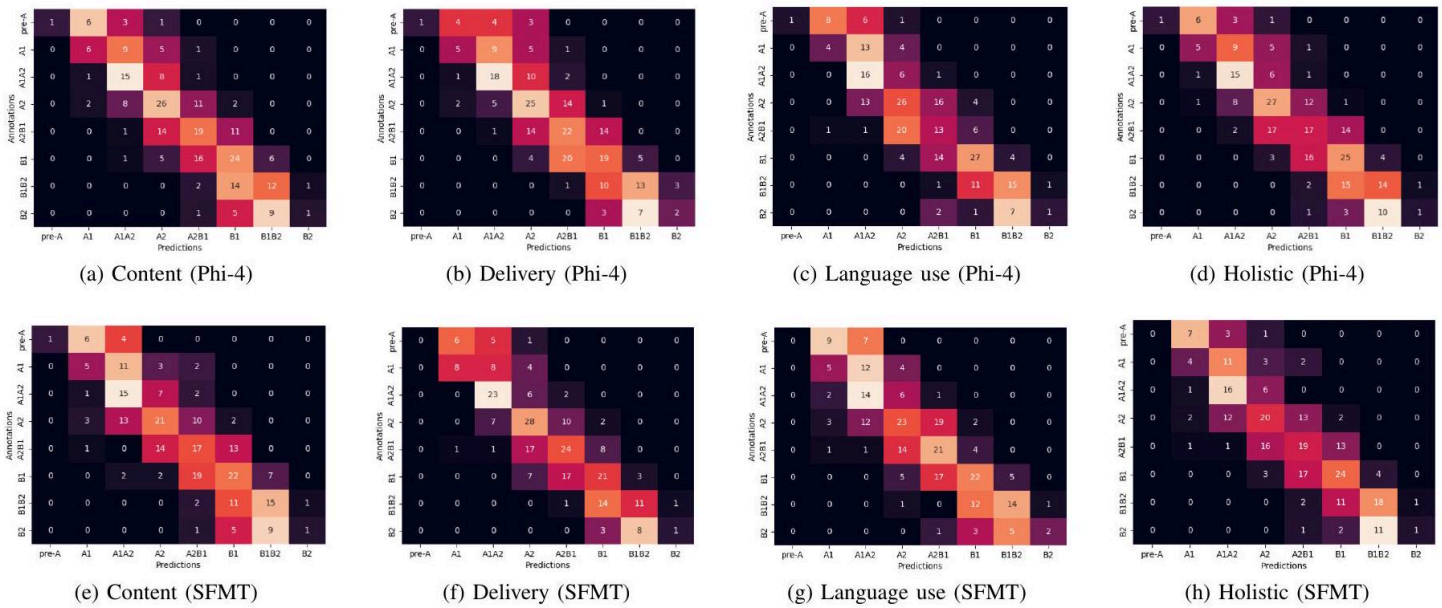
在語音辨識方面，我們將 Whisper large v2（14.75% WER）與 Phi4 的整合式 ASR 模組（18.25% WER）在 TEEMI 語料庫上進行比較。輸出生成限制為 10 個 token，以防止幻覺。SFMT 訓練遵循規定的兩階段課程：階段 1 處理僅音訊輸入，帶

有空文字佔位符，而階段 2 則包含完整的多模態輸入。特定面向的提示在推論期間引導目標評估。

模型效能評估使用皮爾遜相關係數 (PCC) 來衡量預測一致性，絕對準確度 (Absolute Accuracy) 用於精確的 CEFR 等級分類，鄰近準確度 (Adjacent Accuracy) 用於 ± 0.5 等級內的預測，以及巨集準確度 (Macro Accuracy) 用於平衡的跨等級效能測量，同時考慮資料集類別不平衡。此外，對於基於迴歸的評分任務評估，則使用均方根誤差 (RMSE) 來評估預測與實際連續分數之間誤差的平均大小。

表 2：表二
模型在 teemi 測試集上的表現。

模型	內容 (C)			傳遞 (D)			語言使用 (L)			整體 (H)		
	PCC ↑	ABS ↑	ADJ ↑	PCC ↑	ABS ↑	調整 ↑	PCC ↑	ABS ↑	調整 ↑	PCC ↑	ABS ↑	ADJ ↑
基準模型												
W2V [7]	0.755	35.08	81.85	0.768	39.92	83.06	0.740	36.29	79.03	0.771	34.67	83.87
BERT [6]	0.774	33.47	84.68	0.794	38.31	84.68	0.759	36.29	80.24	0.781	35.48	82.66
W2V-BERT [13]	0.735	35.08	81.45	0.794	38.71	87.10	0.798	41.13	82.66	0.771	38.71	84.68
W2V-PT [11]	0.733	30.65	79.84	0.796	39.11	83.06	0.779	42.74	81.45	0.785	34.68	83.07
BERT-PT [11]	0.756	29.44	79.84	0.783	40.73	83.06	0.788	35.08	81.85	0.777	33.87	81.85
多面向 [28]	0.760	37.10	80.24	0.810	41.94	85.48	0.785	39.92	81.45	0.783	38.31	84.27
我們的方法												
Phi-4	0.826	41.93	87.90	0.831	42.34	89.11	0.840	41.53	89.52	0.846	42.34	90.32
Phi-4 (SFMT)	0.821	39.11	88.31	0.848	46.77	89.11	0.835	40.73	88.31	0.838	41.13	90.73



(a) Content (Phi-4) (b) Delivery (Phi-4) (c) Language use (Phi-4) (d) Holistic (Phi-4)

(e) Content (SFMT) (f) Delivery (SFMT) (g) Language use (SFMT) (h) Holistic (SFMT)

圖 4：圖 4. 比較標準 Phi-4 和 SFMT 在 CEFR 量表上的表現的混淆矩陣，顯示對交付評估的對角線集中度有所提高。

為了促進多模態 ASA 研究的重現性並推動社群進步，我們將在發表後公開所有原始碼和微調實作 [2]

V. 結果

A. 整體 MLLM 效能

表二顯示，在所有評估面向中，MLLM（即 Phi-4）相較於目前最先進的模型，展現出顯著的優勢。標準的 Phi-4 在 PCC 分數上持續保持在 0.82 以上，這相較於所有 PCC 結果低於 0.80 的對照模型，代表著顯著的改進。這似乎驗證了 MLLM 多模態整合能力對於全面性 ASA 的有效性。

圖 4 中的混淆矩陣提供了視覺確認，證實了在執行分類時，分類精確度有所提升

ASA 與基於 MLLM 的模型；基於 MLLM 的模型與傳統模型相比，展現出卓越的對角線集中度，這表明 MLLM 是一種能夠超越固有模態限制的全能型主力。

B. 模態分析與 SFMT 有效性

消融研究以皮爾森相關係數 (PCC) 和巨集準確度 (Macro Acc) 作為關鍵績效指標（表三），揭示了多模態大型語言模型 (MLLM) 輔助語音情感評分器 (ASA) 中模態貢獻的基本見解，並驗證了我們 SFMT 策略的有效性。

模態貢獻：表三報告了基於不同模態及其組合運作的 MLLM 模型之效能水準。僅音訊的配置展現出強大的整體效能，尤其在評估語音傳遞方面表現出色。相較之下，僅文字的模型則普遍效能下降，

表 3：表三

消融研究比較模態貢獻對 MLLM 輔助語音情感評分器 (ASA) 效能的影響。

訓練配置	音訊	文字	內容 (C)		傳遞 (D)		語言使用 (L)		整體 (H)	
			PCC ↑	巨集 Acc ↑	PCC ↑	巨集 Acc ↑	PCC ↑	巨集 Acc ↑	PCC ↑	巨集 Acc ↑
Phi-4	✓	✓	0.826	82.00	0.831	82.48	0.840	85.27	0.841	84.27
Phi-4 (僅限文字)	×	✓	0.784	74.76	0.776	75.83	0.768	72.80	0.776	73.35
Phi-4 (僅限音訊)	✓	×	0.811	82.33	0.835	82.94	0.830	86.16	0.836	86.83
Phi-4 (SFMT)	✓	✓	0.821	83.41	0.848	84.01	0.835	83.67	0.838	86.75

在交付方面的評估中觀察到最顯著的下降。這突顯了純文字模型所面臨的挑戰，部分原因是它們僅依賴 ASR 轉錄（在 TEEMI 上使用 Whisper large v2 實現 14.75% WER），以及在交付方面評估時固有的缺乏直接聲學線索。

SFMT 驗證：SFMT 在引入文本資訊之前，策略性地強調建立穩固的語音處理基礎，這帶來了顯著的提升，尤其是在評估「表達」方面——其成功最關鍵地取決於細緻的聲學辨識。這在與 Phi-4 基準線的比較中得到了明確的證明（表三）：在「表達」方面的評估顯示出顯著的 PCC 優勢（SFMT 的值為 0.848，而 Phi-4 基準線為 0.831）。此外，SFMT 將此方面的巨集觀準確度從 82.48% 提高到 84.01% 。這些結果驗證了 SFMT 作為一種有效的課程學習方法，突顯了在跨模態整合之前建立穩固聲學表徵的益處。

C. 泛化至未見任務

對 TEEMI 未見任務的評估（參見表四）證實了微調後的 Phi4 在各方面的強大泛化能力。在傳遞方面的評估展現出最強的遷移性能，表明可遷移聲學特徵的有效學習。儘管任務提示中存在語義差異，內容和語言使用方面的結果也顯示出很強的相關性。這再次驗證了 MLLM 開發可泛化多模態表示以用於跨任務 ASA 應用的能力。

表 4：表四

在未見過的 teemi 資料集上的模型效能。

面向	PCC ↑	ABS Acc ↑	ADJ Acc ↑
內容 (C)	0.851	32.52	78.86
傳遞 (D)	0.863	44.72	86.18
語言使用 (L)	0.855	33.33	78.86
整體 (H)	0.846	32.52	78.86

D. 跨語料庫評估

在 Speak & Improve 語料庫上的跨語料庫評估（表 V）進一步證實了我們的模型在不同第二語言（L2）人群和評估任務中的有效性。SFMT 策略在所有評估指標上始終優於傳統
基準模型和標準 Phi-4 實作，展現出卓越的預測準確性和與人類判斷的相關性。這種跨語料庫的成功驗證了所提出的模型和訓練方案不僅適用於 TEEMI 語料庫的特定特性，還能推廣到更廣泛的國際評估情境。在不同資料集和學習者群體中持續的性能提升，確立了 SFMT 在實際自動口語評估（ASA）部署場景中的實用性。

表 5：表 V

在 Speak & Improve 語料庫上的表現。

方法	均方根誤差 ↓	PCC ↑	準確率 ± 0.5 ↑	準確率 ± 1.0 ↑
BERT [6]	0.445	0.727	76.0	96.3
W2V [7]	0.394	0.790	**81.3**	**99.3**
Phi-4	0.412	0.796	74.7	98.0
Phi-4 (SFMT)	**0.387**	**0.800**	79.7	99.2

第六節 結論與未來工作

本文首次系統性地研究了多模態大型語言模型（MLLM）在綜合自動口語評估（ASA）中的應用，並解決了三個基本研究問題。我們的研究結果表明，MLLM 有效地解決了傳統資訊融合所面臨的挑戰，與單模態模型相比，在所有評估面向都取得了卓越的表現。消融研究證實了音訊模態對於表達評估的不可替代性，而我們提出的 SFMT 策略透過語音優先的課程學習顯著提升了效能，特別有助於細粒度的聲學辨識。在 TEEMI 和 Speak & Improve 語料庫上進行的一系列實驗驗證，證實了我們模型在不同第二語言學習者群體和評估情境中具有強大的泛化能力。這些結果也表明，基於 MLLM 的模型將成為 ASA 的變革性骨幹，實現更準確、更全面、更具泛化性的評估系統。未來的研究將探索用於多面向評估的多任務學習框架，並將綜合回饋生成整合到 ASA 中，以期實現創建智慧、適應性語言學習環境的更廣泛目標，為第二語言學習者在各種電腦輔助語言學習情境中提供個人化、即時的指導。

參考文獻

[1] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang, “SALMONN: Towards generic hearing abilities for large language models,” in The Twelfth International Conference on Learning Representations, 2024. [Online]. Available: https://openreview.net/forum?id=Vti_B5p116

[2] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin et al., “Qwen2-audio technical report,” arXiv preprint arXiv:2407.10759, 2024.

[3] Microsoft and Others, “Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras,” arXiv preprint arXiv:2503.01743, 2025.

[4] A. Rouditchenko, S. Bhati, E. Araujo, S. Thomas, H. Kuehne, R. Feris, and J. Glass, “Omni-r1: Do you really need audio to fine-tune your audio llm?” arXiv preprint arXiv:2505.09439, 2025.

[5] OpenAI, J. Achiam, S. Adler, and …, “Gpt-4 technical report,” 2024. [Online]. Available: https://arxiv.org/abs/2303.08774

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, 「BERT: Pre-training of deep bidirectional transformers for language understanding，」收錄於《2019 年北美計算語言學會年會論文集：人類語言技術，第一卷（長篇與短篇論文）》。明尼亞波利斯，明尼蘇達州：計算語言學會，2019 年 6 月，頁 4171-4186。[線上]。網址：https://aclanthology.org/N19-1423

[7] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, 「wav2vec 2.0: A framework for self-supervised learning of speech representations，」收錄於《神經資訊處理系統進展 33》，2020 年，頁 12449-12460。

[8] X. Wang, K. Evanini, Y. Qian, and M. Mulholland, 「Automated scoring of spontaneous speech from young learners of english using transformers，」收錄於《2021 IEEE 語音技術研討會，SLT 2021》，中國深圳，2021 年 1 月 19-22 日。IEEE，2021 年，頁 705-712。[線上]。網址：https://doi.org/10.1109/SLT48900.2021. 9383501

[9] S. Banno and M. Matassoni, 「Proficiency assessment of 12 spoken english using wav2vec 2.0，」收錄於《2022 IEEE 語音技術研討會 (SLT)》。卡達杜哈：IEEE，2023 年，頁 1088-1095。

[10] H. Nguyen 和 S. Park，「針對形成性科學評估提供自動化回饋：多模態大型語言模型的應用」，收錄於第 15 屆國際學習分析與知識會議論文集，系列 LAK '25。美國紐約州紐約市：Association for Computing Machinery，2025 年，第 803-809 頁。[線上]。網址：https://doi.org/10.1145/3706468.3706480

[11] T.-H. Lo、F.-A. Chao、T.-I. Wu、Y.-T. Sung 和 B. Chen，「一種有效緩解資料稀缺和分佈不平衡的自動化口語評估方法」，收錄於計算語言學協會研究成果：NAACL 2024。墨西哥墨西哥城：Association for Computational Linguistics，2024 年，第 1352-1362 頁。[線上]。網址：https://aclanthology.org/2024.findings-naacl。86

[12] N. H. de Jong，「評估第二語言口語能力」，Annual Review of Linguistics，第 9 卷，第 541-560 頁，2023 年。

[13] S. Park 和 R. Ubale，「用於細粒度語音評分的文本和語音表示多任務學習模型」，收錄於 2023 IEEE 自動語音辨識與理解研討會 (ASRU)。台灣台北：IEEE，2023 年，第 1-7 頁。

[14] S. Bannò, K. M. Knill, M. Matassoni, V. Raina, and M. Gales, 「使用自監督語音表徵學習評估 12 種口語能力，」收錄於第 9 屆教育語音與語言技術研討會 (SLaTE)。ISCA，2023 年，頁 126-130。[線上]。網址：https://www.isca-speech.org/archive/slate_2023/banno23_slate.html

[15] E. Kim, J.-J. Jeon, H. Seo, and H. Kim, 「使用自監督語音表徵學習進行自動發音評估，」收錄於 Proc. Interspeech 2022，2022 年，頁 1411-1415。

[16] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, 「課程學習，」收錄於國際機器學習會議，2009 年。[線上]。網址：https://api.semanticscholar.org/CorpusID:873046

[17] S. B. Davis and P. Mermelstein, 「連續語句中單音節詞彙辨識的參數表示比較，」IEEE Transactions on Acoustics, Speech and Signal Processing，第 28 卷，第 4 期，頁 357-366，1980 年 8 月。

[18] A. Loukina, K. Zechner, L. Chen, and M. Heilman, “Feature selection for automated speech scoring,” in Proceedings of the Tenth Workshop
on Innovative Use of NLP for Building Educational Applications (BEA). Association for Computational Linguistics, 2015, pp. 12-19.

[19] X. Xi, D. Higgins, K. Zechner, and D. M. Williamson, "Automated scoring of spontaneous speech using speechratersm v1.0,"ETS Research Report Series, vol. 2008, no. 2, pp. i-47, 2008.

[20] S. Xie, K. Evanini, and K. Zechner, “Exploring content features for automated speech scoring,” in Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2012, pp. 103-111.

[21] T.-I. Wu, T.-H. Lo, F.-A. Chao, Y.-T. Sung, and B. Chen, “A preliminary study on automated speaking assessment of

English as a second language (ESL) students," in Proceedings of the 34th Conference on Computational Linguistics and Speech Processing (ROCLING 2022), Y.-C. Chang and Y.-C. Huang, Eds. Taipei, Taiwan: The Association for Computational Linguistics and Chinese Language Processing (ACLCLP), Nov. 2022, pp. 174-183. [Online]. Available: https://aclanthology.org/2022.rocling-1.22/
[22] W. Chen, H. Liu, and X. Wang, "wavllm: Hierarchical curriculum learning for multimodal speaking assessment," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 32, pp. 1024-1036, 2024.
[23] C. Zhang, Y. Wang, Y. Zhang, B. Li, Y. B. Zhao, Y. Lu, Y. Li, and Z. Liu, "Oversampling, augmentation and curriculum learning for speaking assessment with limited training data," in Proc. INTERSPEECH 2024, Kos Island, Greece, September 2024, pp. 506-510.
[24] Y. Fan, W. Xu, H. Wang, J. Wang, and S. Guo, "Pmr: Prototypical modal rebalance for multimodal learning," in 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2002920038.
[25] T. Yu, X. Liu, Z. Hou, L. Ding, D. Tao, and M. Zhang, "Selfpowered 11 m modality expansion for large speech-text models," in Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP). Miami, Florida, USA: Association for Computational Linguistics, November 2024, pp. 12401-12417. [Online]. Available: https://aclanthology.org/2024.emnlp-main.690/
[26] S.-Y. Chen, T.-H. Lo, Y.-T. Sung, C.-Y. Tseng, and B. Chen, "A speaking practice tool on teemi for automated english-speaking assessment of chinese learners," in Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Kos, Greece, September 2024, pp. 2048-2049. [Online]. Available: https: //www.isca-archive.org/interspeech_2024/chen24aa_interspeech.pdf
[27] K. Knill, D. Nicholls, M. Gales, M. Qian, and P. Stroinski, "Speak & improve corpus 2025: an 12 english speech corpus for language assessment and feedback," ArXiv, vol. abs/2412.11986, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:274789386
[28] W.-H. Peng, S. Chen, and B. Chen, "Enhancing automatic speech assessment leveraging heterogeneous features and soft labels for ordinal classification," in 2024 IEEE Spoken Language Technology Workshop (SLT). Macao: IEEE, 2024, pp. 945-952.
[29] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in International Conference on Learning Representations, 2022. [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9
[30] T. Dao, "Flashattention-2: Faster attention with better parallelism and work partitioning," in The Twelfth International Conference on Learning Representations, 2024. [Online]. Available: https: //openreview.net/forum?id=mZn2Xyh9Ec

[1] https://huggingface.co/microsoft/Phi-4-multimodal-instruct
[2] https://github.com/ntnuYuhsuan/asa-grader.git