自動口語評量中分數序數性與非均勻區間建模的有效策略

羅天宏*、陳思妤†、宋曜廷†、陳柏琳*
* 國立臺灣師範大學資訊工程學系
† 國立臺灣師範大學教育心理與輔導學系

摘要

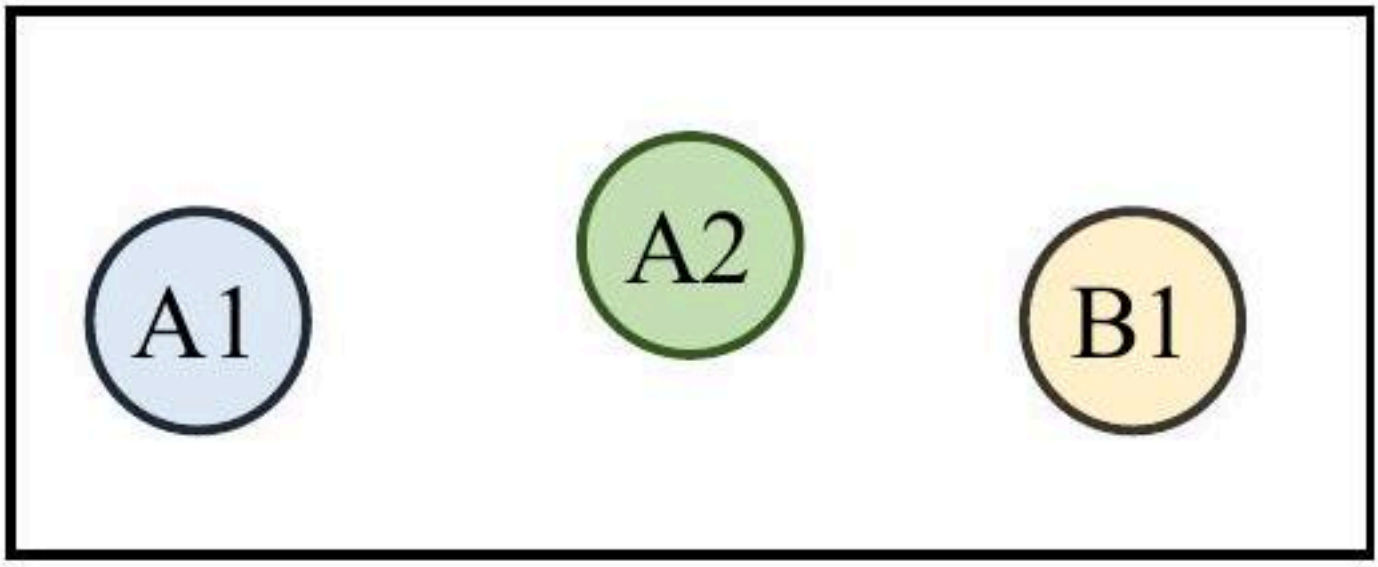
近期關於自動口語評估（ASA）的研究，已受益於自監督學習（SSL）表徵，這些表徵能在非母語語音中捕捉豐富的聲學和語言模式，且無需預設特徵篩選的假設。然而，基於語音的 SSL 模型捕捉的是與聲學相關的特徵，卻忽略了語言內容；而基於文本的 SSL 模型則依賴於 ASR 輸出，未能編碼語韻細微之處。此外，大多數現有技術將熟練度等級視為名目類別，忽略了其序數結構以及熟練度標籤之間不均勻的間隔。為了解決這些限制，我們提出了一種有效的 ASA 方法，透過新穎的建模範式，將 SSL 與手工製作的指標特徵相結合。我們進一步引入了一種多邊界序數損失，該損失能同時建模分數的序數性以及熟練度標籤的不均勻間隔。在 TEEMI 語料庫上進行的廣泛實驗表明，我們的方法始終優於強大的基準線，並且對未見過的提示具有良好的泛化能力。

索引詞彙—自動口語評估、自監督學習、多面向特徵、序數分類、不均勻分數間隔。

一、前言

隨著運算技術的快速進步以及全球第二語言（L2）學習者人數的增加，自動口語評估（ASA）受到了廣泛關注，並在電腦輔助語言學習（CALL）中扮演著日益重要的角色。ASA 系統旨在為學習者的口語表現提供即時回饋，促進口語能力的自主且低壓力的提升。此外，ASA 系統有助於減輕語言教師的負擔，同時為 L2 學習者的口語能力提供更一致且客觀的評估。鑑於這些技術發展，ASA 系統近年來已被廣泛採用，以在各種 CALL 使用案例中增強 L2 語言習得 [1]。

早期的 ASA 方法主要採用淺層分類器以及手工設計的特徵，這些特徵捕捉了語言能力的獨特面向，包括內容（例如，適當性和相關性）、表達（例如，流暢度和語調）和語言使用（例如，詞彙和語法）[2]-[10]。最近，自監督學習（SSL）建模範式的出現，例如 BERT 及其衍生模型 [11]，透過提供情境化嵌入為 ASA 帶來了新的機會。這些表示已在各種評估任務中得到有效利用。



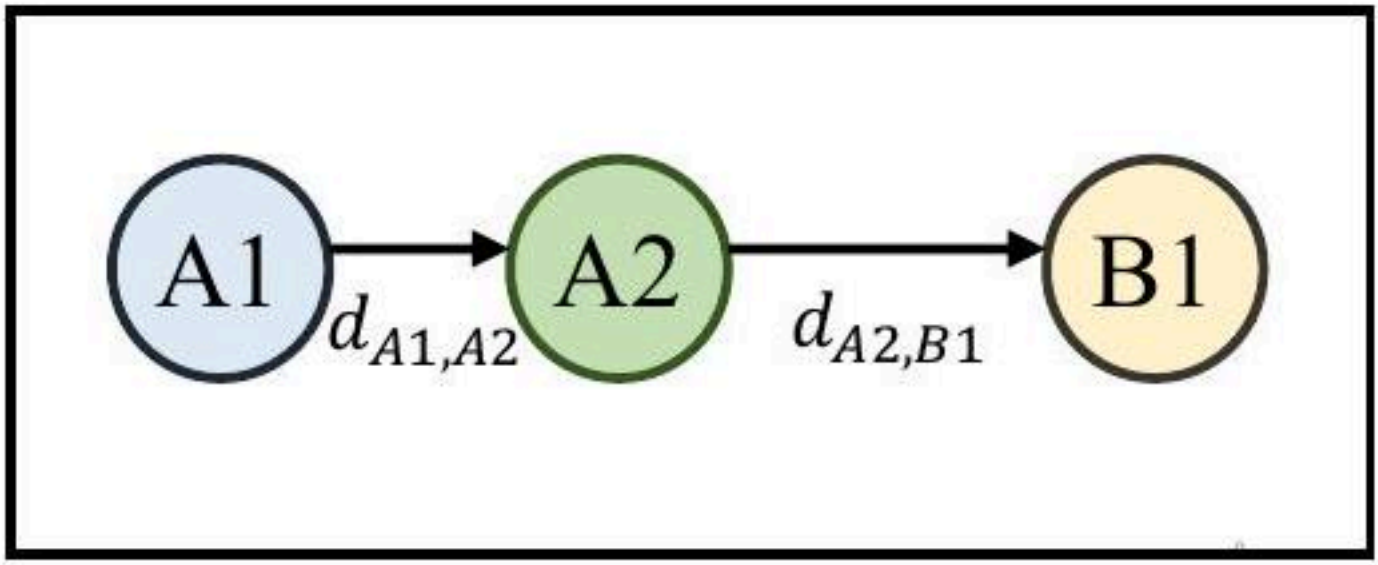


圖 1：圖 1. 概念圖說明了 CEFR 能力等級的名義和序數結構之間的區別。上圖將等級（例如，A1、A2、B1）視為名義類別，不反映它們之間任何固有的順序或距離。下圖透過表示能力等級之間的定向轉換（即 $\mathrm{A1} \rightarrow \mathrm{A2} \rightarrow \mathrm{B1}$），並以距離 d_{a_1,a_2} 和 d_{a_2,b_1} 表示非均勻分數區間，來建模序數結構。

任務，包括句子級評估 [12]、論文評分 [13]、[14]、口語獨白評估 [6]、[15]，以及其他 [16]。同時，基於語音的 SSL 模型的大力發展，透過提供豐富的聲學表示，支援更複雜的建模能力，進一步提升了 ASA 系統的有效性 [17]-[20]。

儘管持續努力，現有的基於 SSL 的 ASA 方法仍受限於模態特定的限制。基於語音的 SSL 編碼器擅長建模聲學特性，但通常無法捕捉學習者回應的語義內容。相較之下，基於文本的 SSL 模型依賴自動語音辨識 (ASR) 輸出，這使得它們容易受到轉錄錯誤的影響，並且無法表示對於評估表達方面至關重要的語調線索。此外，這兩種模態都傾向於缺乏可解釋性，並忽略了諸如音高變化和詞級準確性等明確指標。這些限制突顯了多面向建模策略的必要性，該策略將手工特徵與 SSL 衍生的嵌入相結合，以利用不同模態的互補優勢 [21][23]。除了與模態相關的限制之外，ASA 系統還必須解決國際語言標準（例如歐洲共同語言參考架構 (CEFR)）熟練度
等級（標籤）結構所帶來的挑戰。具體來說，CEFR 等級呈現出固有的序數關係（例如 $\mathrm{A1} < \mathrm{A2} < \mathrm{B1} < \mathrm{B2}$），但許多現有方法將它們視為名義類別，忽略了它們的有序性質。這種建模簡化可能導致次優的訓練目標和降低的校準性能。此外，CEFR 等級之間的得分間隔不均勻（例如，從 B1 到 B2 的進展與從 A1 到 A2 的進展不相等）[24]、[25]。如圖 1 所示，大多數先前的研究將所有等級差距視為等距，這在迴歸或分類設定中很常見，未能反映潛在的學習進程，這將阻礙可解釋性。

為了解決上述挑戰，我們探索了一種創新的 ASA 建模方法，該方法將手工特徵與自監督嵌入相結合，以共同捕捉學習者語音的聲學和語義特性。此外，我們提出了一種多邊界序數 (MMO) 損失函數，該函數明確地建模了 CEFR 熟練度等級中固有的序數結構和不均勻間隔。在 TEEMI 資料集上進行的實驗表明，所提出的框架始終優於強大的基準線，尤其是在識別代表性不足的熟練度等級（例如 Pre-A1 和 B2）方面。此外，該模型對未見過的提示和說話者表現出穩健的泛化能力，突顯了其在真實世界 ASA 場景中的適用性。這項工作的主要貢獻至少有兩方面：

我們提出了一個多面向的 ASA 框架，該框架將手工特徵與自監督嵌入相結合，以捕捉模態特定的特性，同時解決 CEFR 熟練度量表的序數和非均勻特性。在 TEEMI 資料集上進行的一系列實驗表明，與競爭性基準線相比，宏觀平均 F1 分數有顯著提高。
我們提出了一種新穎的 MMO 損失函數，據我們所知，這是第一個以完全資料驅動的方式，共同建模 CEFR 對齊分數的序數結構和非均勻等級間隔。這種設計提升了預測熟練度等級的效能和可解釋性。

二、相關工作

自動口語評估 (ASA) 旨在評估第二語言 (L2) 學習者的口語能力，通常透過反映整體熟練度的整體分數或評估特定表現面向的分析分數。

A. 手工特徵

早期的 ASA 系統主要採用淺層分類器，這些分類器經過訓練，能辨識人工設計的特徵，以捕捉口語的顯著面向，包括發音、流暢度、語調和文法等 [2]-[8], [10]。例如，[4] 採用元音空間度量來評估發音精確度，而 [5] 則納入從詞性分佈中得出的句法複雜度指標。語調和
節奏模式在 [7] 中進行了檢視，而句法分析準確性則在 [6] 中進行了探討。儘管這些人工設計的特徵具有可解釋性，但其提取往往建立在特定任務的假設之上，並且可能難以推廣到未曾見過的提示或不同的任務配置 [8]。

B. 基於文本的自監督特徵

自監督學習 (SSL) 的出現，使得上下文文本嵌入在 ASA 中被廣泛採用。BERT [11] 等模型在各種評估任務中取得了優異的表現，其中包括論文評分 [13]、[14]、可讀性預測 [12] 以及口語對話評估 [15]、[16]。這些模型能有效捕捉語義和句法特徵；然而，它們依賴於 ASR 生成的語音轉錄文本，而這些文本容易出現辨識錯誤，並且無法保留對於評估表達品質至關重要的語調和語音資訊。

C. 基於語音的自監督特徵

基於語音的自監督學習（SSL）模型，例如 wav2vec 2.0 [26]，有助於直接對原始聲學訊號進行建模，並且能夠編碼細緻的語音和語調表徵，而無需依賴自動語音辨識（ASR）轉錄。先前的研究已證明此類模型對於 CEFR 等級分類 [17]、[19] 和模態比較具有實用性。[18] 將這些模型擴展到對話情境，而 [20] 則將原型嵌入與損失重新加權策略結合，以緩解標籤不平衡相關的問題。儘管在建模與表達相關的特徵方面有效，但基於語音的 SSL 模型通常缺乏語義豐富性，而這對於評估內容而言可說是必要的。

D. 多面向特徵

為了捕捉口語表達的多面向性質，最近的研究致力於多面向建模方案，這些方案使用平行或階層式架構 [27]-[29] 共同預測表達、內容和語言使用等面向的分數。儘管大多數先前的研究都集中在分析性評分，但有些研究已將多面向建模擴展到整體口語評估。例如，[21] 將多面向特徵納入整體評分，而 [23] 則在口語評估的背景下引入了軟標籤建模。[22] 系統地比較了 wav2vec 2.0 和 BERT 在不同評分面向上的表現，結果顯示前者在建模表達方面表現出色，而後者在與內容相關的任務上略勝一籌。研究發現，兩種模態的融合能產生最佳的整體表現。

III. 資料集

為了評估所提出的 ASA 框架的有效性，本研究採用了 TEst for English-Medium Instruction (TEEMI) 語料庫 [30]，這是一個專為英語授課 (EMI) 和 ASA 研究而策劃的專有資料集。該語料庫包含大學部和研究所程度的第二語言學習者所產生的自發性英語口語回應。每位第二語言學習者的口語回應都經過標註

表 1：表一
teemi 資料集中每個 CEFR 能力等級的說話者人數。

任務	用法	前 A	A1	A1+	A2	A2+	B1	B1+	B2
	訓練	34	61	76	156	150	169	79	65
A01	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
B02	未見	15	14	21	41	48	62	31	16
C01	未見	10	12	9	17	16	21	18	16
總計	-	87	130	160	320	315	369	206	139

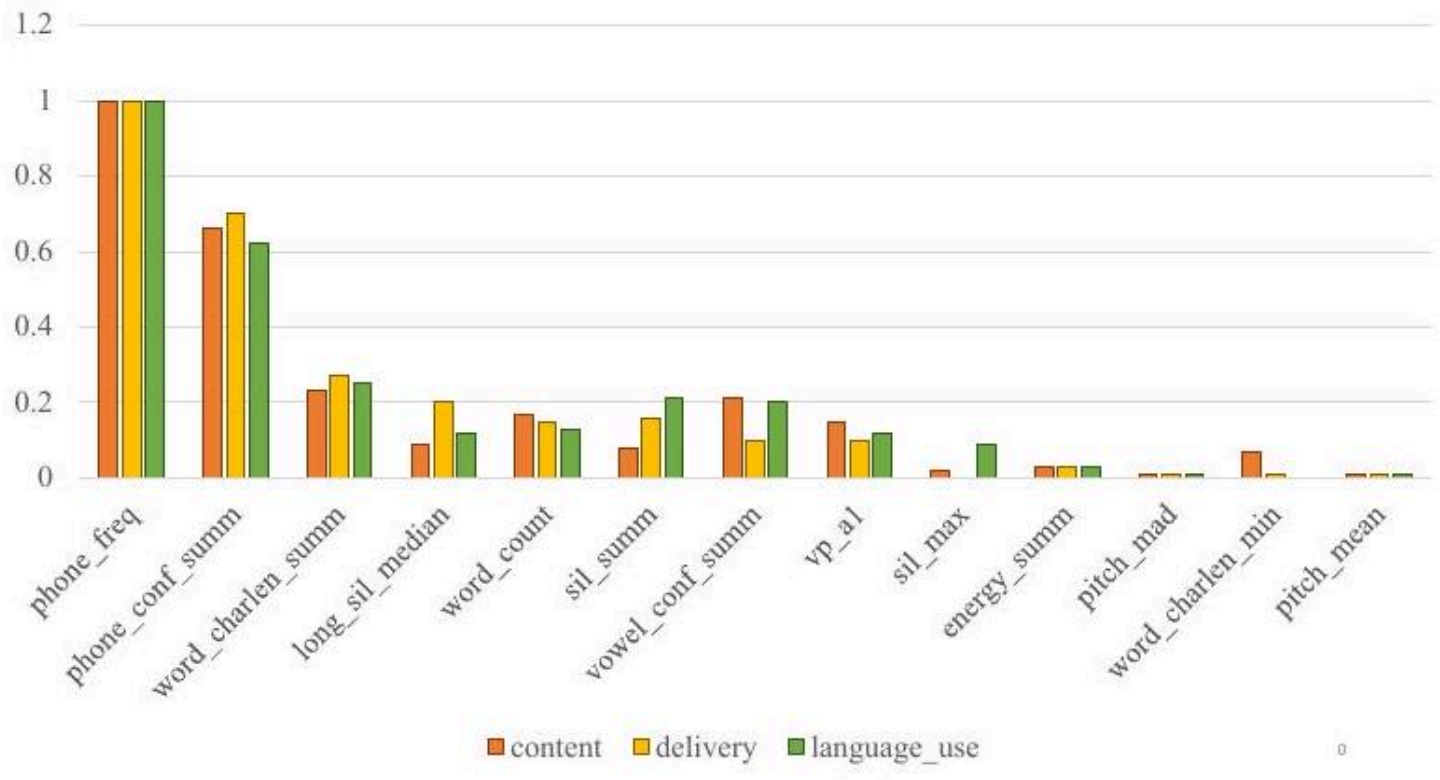


圖 2：圖 2。在 TEEMI 語料庫的 A01 測試集上，針對內容、表達和語言使用預測，手工特徵的相對重要性，使用針對每個評分面向單獨訓練的隨機森林迴歸器計算。

根據符合 CEFR 標準的評分標準，給出一個整體分數和三個分析分數（即內容、語言使用和表達）。每個回答至少由三名受過訓練的評分員進行註釋，最終標籤由多數投票決定，以確保可靠性。

TEEMI 語料庫的組成部分包括三種任務形式：一般聽力與回答 (A)、情境問答 (B) 和主題問答 (C)。在本研究中，我們專注於包含任務 A01、A02、B02 和 C01 的子集，共產生 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。其餘任務（A02、B02 和 C01）被保留，以評估 ASA 模型推廣到以前未見過提示的能力。每個任務的詳細 CEFR 等級分佈和相應的分區如表 I 所示。

另一方面，為了驗證傳統手工製作的特徵在不同提示和任務類型中，其泛化能力是好是壞，我們根據 [9] 和 [10] 中提供的方法和特徵定義，透過訓練獨立的隨機森林迴歸器來分析每個特徵對於內容、表達和語言使用的重要性。如圖 2 所示，個別特徵的效用在不同評分面向中有所差異。例如，音素頻率統計 (phone_freq) 和總靜音持續時間 (sil_summ) 對於表達評分非常有用，而 CEFR-A1 級詞彙項的頻率 (vp_a1) 和字數 (word_count) 則更能反映內容評估。這項觀察促使我們納入 SSL 表示，預期能捕捉更全面的
資訊以用於 ASA。

IV. 方法論

在本節中，我們將自動口語評估 (ASA) 定義為針對第二語言學習者口語回應的分類任務。每個訓練、驗證或測試實例都表示為輸入-輸出對 ($\mathbf{x}_i, y_i$)，其中 $\mathbf{x}_i$ 表示從學習者回應的多種模態中提取的輸入特徵，包括原始音訊訊號 $\mathbf{a}$ 、ASR 生成的轉錄 $\mathbf{w}$ 和給定的提示 $\mathbf{p}$ ，以及 y_i 對應的（或預測的）CEFR 標籤。每個口語回應都記錄為原始音訊序列 $\mathbf{a} = \{a_1, a_2, \ldots, a_t\}$ ，其中 t 表示聲學幀的數量。為了提取語義內容，應用 ASR 系統生成詞級轉錄 $\mathbf{w} = \{w_1, w_2, \ldots, w_m\}$ ，其中 m 表示識別出的單詞數量。此外，每個回應都與提示 $\mathbf{p} = \{p_1, p_2, \ldots, p_k\}$ 相關聯，提示為評估任務提供上下文資訊（例如，問題）。先前的基於 SSL 的 ASA 方法通常依賴於語音模態 a (參見圖 3b) 或文本模態 $\mathbf{w}$ (參見圖 3a) 進行熟練度預測。對應的（或預測的）標籤 $y_i \in \mathcal{Y} = \{1.0, 1.5, 2.0, \ldots, 5.0\}$ 表示 CEFR 對齊的熟練度分數，隨後將其數位化為 $\{1, 2, \ldots, 8\}$ ，用於在涵蓋 Pre-A1 到 B2 的 8 級量表上進行分類。

A. 多面向能力建模

如圖 3c 所示，所提出的架構使用獨立的 Transformer [31] 編碼器，並進行了微小的調整 [32]，以分別對內容、表達和語言使用等面向進行建模，從而生成特定面向的表示，這些表示將有效地整合以進行評分。

內容模組：內容模組旨在捕捉學習者口語回答與給定提示之間的語義對齊。提示 $\mathbf{p}$ 使用預訓練的 BERT 模型進行編碼，以從 [CLS] 標記中導出句子級嵌入。同時，從音訊原始波形 a 中使用 wav2vec 2.0 (W2V) 提取幀級語音特徵，從而產生一系列上

下文表示：

$$\mathbf{e}^{[\mathrm{CLS}]} = \mathrm{BERT}([[CLS]; \mathbf{p}]),$$

$$\mathbf{h}^{a}_{1:t} = \mathrm{W2\ V}(\mathbf{a}).$$

提示嵌入 $\mathbf{e}^{[\mathrm{CLS}]}$ 被複製並與每個幀級語音特徵連接，然後輸入到 Transformer 編碼器層：

$$\mathbf{h}^{c}_{1:t} = \mathrm{Transformer}_{\mathrm{content}}\left(\left[\mathbf{e}^{[\mathrm{CLS}]}; \mathbf{h}^{a}_{1}\right], \ldots, \left[\mathbf{e}^{[\mathrm{CLS}]}; \mathbf{h}^{a}_{t}\right]\right),$$
$$\mathbf{v}^{c} = \mathrm{Pooler}\left(\mathbf{h}^{c}_{1:t}\right),$$

其中注意力池化器 (attention pooling Pooler) 應用於 $\mathbf{h}^{c}_{i:t}$ 以獲得內容表示 $\mathbf{v}^{c}$ ，其封裝了口語回應與提示之間與內容相關的關聯性。

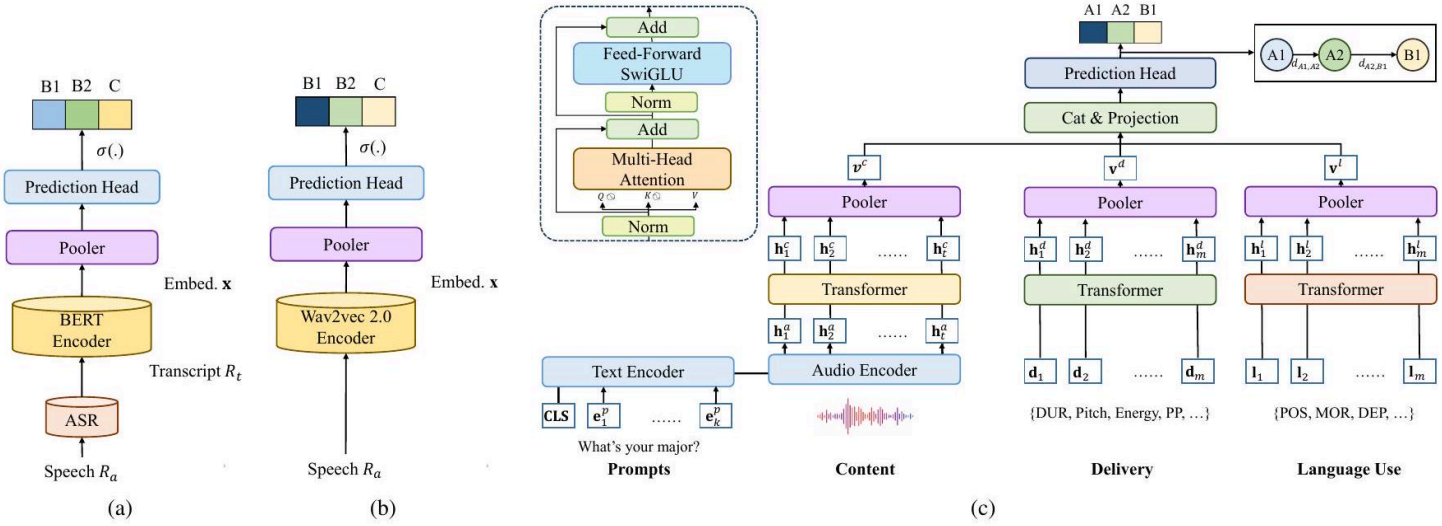


圖 3：圖 3. 呈現了自動口語評估的模型架構概述。(a) 描繪了一個基於文本的評分器，它利用 ASR 生成的轉錄稿和 BERT 嵌入來執行 CEFR 等級分類。(b) 說明了一個基於語音的評分器，它使用 wav2vec 2.0 直接編碼原始音訊輸入。(c) 呈現了我們的多面向框架，它使用獨立的 Transformer 編碼器對內容、表達和語言使用進行建模。將產生的表示 ($\mathbf{v}^{*}$) 串聯、池化，然後傳遞給使用交叉熵和 MMO 損失訓練的預測頭。

表達編碼器：表達編碼器捕捉口語表達的時間特徵，著重於語調變化和語音清晰度。ASR 輸出中每個與單詞對齊的片段都由一個表達特徵向量 $\mathbf{d}_i \in \mathbb{R}^{16}$ 表示，其中包含音高和能量統計數據（平均值、標準差、中位數、中位數絕對偏差、總和、最大值、最小值）、片段持續時間和置信度分數。表達向量序列 $\mathbf{d}_{1:m}$ 經由 Transformer 層處理以獲得上下文嵌入。然後應用基於注意力的池化層以產生代表表達相關資訊的固定維度向量 $\mathbf{v}^{d}$ ：

$$\mathbf{v}^{d} = \mathrm{Pooler}\left(\mathrm{Transformer}_{\mathrm{delivery}}\ \left(\mathbf{d}_{1:m}\right)\right).$$

語言使用模組：此模組旨在捕捉學習者口語回應中存在的詞彙選擇和句法結構。給定 ASR 系統產生的轉錄稿，使用 Stanza NLP 工具包 [33] 獲得詞級語言註釋，包括詞性 (POS) 標籤、依存關係 (DEP) 和形態特徵 (MOS)。這些特徵被編碼成語言特徵向量序列 $\mathbf{l}_{1:m} \in \mathbb{R}^{m \times 263}$ ，然後通過 Transformer 編碼器以捕捉詞語之間的上下文依賴關係。將基於注意力的池化機制應用於上下文序列以產生代表語言使用的固定長度嵌入：

$$\mathbf{v}^{l} = \mathrm{Pooler}\left(\mathrm{Transformer}_{\mathrm{lang}}\ \left(\mathbf{l}_{1:m}\right)\right).$$

為了建構口語回應的整體表示，三個特定面向的嵌入，即 $\mathbf{v}^{c}$ (內容)、 $\mathbf{v}^{d}$ (表達) 和 $\mathbf{v}^{l}$ (語言使用)，被串聯起來並通過一個線性投影層。融合後的向量隨後被送入預測頭，以計算 CEFR 能力等級的 logits，表示為 $\mathbf{z} \in \mathbb{R}^{C}$ ：

$$\mathbf{z} = \mathrm{PredictionHead}\left(\mathrm{Projection}\ \left(\left[\mathbf{v}^{c}; \mathbf{v}^{d}; \mathbf{v}^{l}\right]\right)\right).$$

B. 多邊界序數損失

為了同時考慮 CEFR 評分中固有的序數結構和非均勻間隔，本研究引入了一種基於 logit 的多邊界序數 (MMO) 損失。雖然先前的研究 [34] 已在隱藏表示層實施多邊界約束以模擬圖像分類的序數性，但所提出的方法與之不同之處在於直接在 logit 層施加這些約束。這種設計選擇有助於對預測輸出進行更明確的監督，並提供與 CEFR 等級不對稱進展的更好對齊。據我們所知，我們是第一個將此概念擴展並概念化用於 ASA 的研究。

對於每個輸入實例，MMO 損失被定義為應用於正向 $(\mathbf{z}, \mathbf{z}_j) \in \mathcal{S}^+$ 和負向 $(\mathbf{z}, \mathbf{z}_k) \in \mathcal{S}^-$ logit 對集合的成對約束：

$$\mathcal{L}_{\mathrm{MMO}}(\mathbf{z}, y) = \max\left(0, d_{y,y_k} + \phi\left(\mathbf{z}, \mathbf{z}_k\right) - \phi\left(\mathbf{z}, \mathbf{z}_j\right)\right)$$
$$d_{y,y_k} = d_{y,y+1} + \cdots + d_{y_{k-1},y_k}$$

其中 $\phi(\cdot)$ 表示餘弦相似度函數。累積邊界 d_{y,y_k} 代表真實標籤 y 與負面標籤 y_k 之間的序數距離，從而強制在對數空間中，對於在 CEFR 量表上距離較遠的標籤對，產生更大的分離。

最終的損失函數將 MMO 損失與傳統的交叉熵目標結合，以共同優化分類準確性和序數一致性：

$$\mathcal{L} = \lambda \cdot \mathcal{L}_{\mathrm{CE}} + (1 - \lambda) \cdot \mathcal{L}_{\mathrm{MMO}},$$

其中超參數 $\lambda \in [0, 1]$ 控制著標準分類監督與序數感知學習之間的平衡。

表 2：表二
模型在 teemi 測試集上的表現。

模型	內容		交付		語言使用		整體	
	ACC ↑	F1 ↑	ACC ↑	F1 ↑	ACC ↑	F1 ↑	ACC ↑	F1 ↑
W2V [17]	35.08	29.55	39.92	37.11	36.29	31.53	34.67	30.17
BERT [17]	33.47	28.31	37.90	31.19	36.29	31.66	35.48	31.19
W2V-BERT [22]	35.08	27.83	38.31	31.46	41.13	35.15	38.71	30.35
W2V-PT [20]	30.24	24.23	38.71	34.33	42.74	36.00	34.68	29.87
BERT-PT [20]	29.44	27.25	40.73	37.22	35.08	33.90	33.87	32.49
MA	35.89	31.60	41.53	39.04	38.31	31.83	33.87	26.28
MA + MMO	37.10	34.77	42.34	40.87	42.34	40.22	36.29	35.55

表 3：表三
模型在未見測試資料集上的表現。

模型	任務	內容		表達		語言運用		整體性	
		ACC ↑	F1 ↑	ACC ↑	F1 ↑	ACC ↑	F1 ↑	ACC ↑	F1 ↑
MA	A02	30.08	31.91	39.84	39.12	34.15	36.02	32.52	35.30
	B02	32.66	27.76	36.29	34.30	34.27	28.89	33.47	27.63
	C01	20.17	14.30	26.05	21.85	37.82	34.61	25.21	21.27
+MMO	A02	35.77	35.77	43.09	44.84	39.02	40.44	30.89	32.32
	B02	31.85	29.54	38.71	37.74	30.65	31.24	32.66	31.40
	C01	30.25	29.62	44.54	43.04	39.50	35.27	31.93	31.14

V. 實驗設置

A. 實作細節

模型配置是使用來自 HuggingFace Transformers 函式庫 [35] 的預訓練模型進行初始化。兩個基於 SSL 的模型，bert-base-uncased [1] 和 wav2vec2-base [2] ，分別用作文字和語音編碼器。對於所有 Transformer 編碼器，注意力頭的數量設定為 1，以鼓勵輕量級建模並減少過度擬合。目標函數中的加權係數 λ 透過在驗證集上進行網格搜尋來調整，其中 $\lambda = 0.5$ 是根據評估指標的經驗性能選擇的。所有模型都在 NVIDIA 3090 GPU 上使用 AdamW 優化器進行訓練，批次大小為 32，初始學習率為 $1\mathrm{e}-4$ 。所有分類器的訓練過程都根據驗證集的平均宏觀平均分數，在 30 個耐心週期後提前停止。

B. 評估指標

對 ASA 模型有效性的具體評估對於評分應用至關重要，因為在所有層級上都需要準確的預測。然而，由於 CEFR 等級的分佈不平衡，傳統的評估指標（例如準確度 (ACC)）可能會低估 ASA 模型的效能。因此，我們使用宏觀平均 F1 分數來懲罰那些對次要類別處理不佳的模型。

VI. 結果與討論

A. 整體表現

首先，我們將報告基準系統和所提出模型在 TEEMI 語料庫上的表現，評估標準涵蓋四個基於 CEFR 的評分面向：內容、表達、

語言運用和整體熟練度。此處比較的基準模型包括基於語音的 SSL (W2V) [17]、基於文本的 SSL (BERT) [17]、兩者的多模態融合 (W2VBERT) [22]，以及它們各自的原型變體 (W2VPT 和 BERT-PT) [20]。如表二所示，W2V-BERT 在整體熟練度方面表現最強，達到最高的絕對準確度 (38.71%)，這突顯了整合語音和語言線索的優勢。雖然原型模型在表達和語言運用方面的評估上顯示出具競爭力的 F1 分數，但它們在內容和整體熟練度方面的表現顯著下降，這表明其在不同面向上的泛化能力有限。所提出的多面向 (MA) 框架，將手工特徵與 SSL 衍生的嵌入相結合，在宏觀平均 F1 方面，始終優於所有基準系統，涵蓋所有四個評分面向。在表達方面觀察到特別強勁的表現，同時在內容和語言運用方面也具有競爭力的準確度。引入多邊界序數 (MMO) 損失後，進一步提升了性能。MA+MMO 變體實現了最高的整體 F1 分數 (35.55%)，與基礎 MA 模型相比，在內容和語言運用方面的宏觀平均 F1 分別絕對增加了 3.17 和 8.39。性能的顯著提升證實了我們所提出的 ASA 建模策略具有廣闊的潛力。

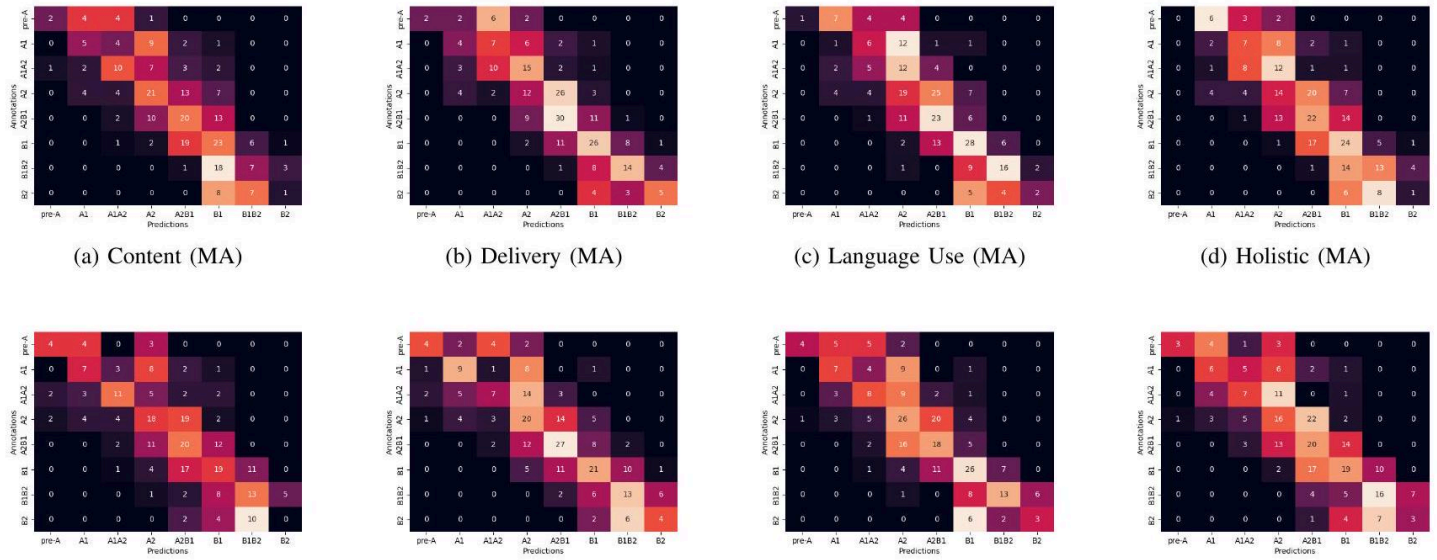
B. 在未見提示和任務上的評估

為了評估所提出模型的泛化能力，我們在 TEEMI 語料庫中三個未見任務上進行了第二組實驗：A02、B02 和 C01，如表一所述。A02 與訓練任務 A01 具有相同的任務類型，但在提示內容上有所不同，而 B02 和 C01 則涉及提示和任務類型的轉變，分別對應情境式和主題式問答任務。

如表三所示，MA+MMO 模型在所有 CEFR 等級上都持續優於 MA 模型。對於 A02 等級，變異主要來自提示，MA+MMO 將整體 F1 分數提高了 1.02% ，語言使用方面的 F1 分數提高了 4.42% ，這證明了其對提示級別差異的穩健性。對於引入更廣泛任務變異的 B02 和 C01 等級，效能提升更為顯著。MA+MMO 在 B02 上實現了整體 F1 分數 3.77% 的增長，而在 C01 上，在表達（ $+21.19\%$ ）和整體 F1（ $+9.87\%$ ）方面觀察到顯著改進。這些發現表明，納入 MMO 損失增強了模型泛化到未見提示和任務配置的能力。

C. 混淆矩陣的可視化

圖 4 呈現了兩種配置下每個評分面向的混淆矩陣：所提出的多面向 (MA) 模型（頂部行）及其透過多邊際序數 (MMO) 損失增強的變體（底部行）。這些混淆矩陣揭示了預測的 CEFR 之間的對應關係



(a) Content (MA) (b) Delivery (MA) (c) Language Use (MA) (d) Holistic (MA)

(e) Content (MA+MMO) (f) Delivery (MA+MMO) (g) Language Use (MA+MMO) (h) Holistic (MA+MMO)

圖 4：圖 4. 比較多面向 ASA 分類器在有無所提出的多邊際序數 (MMO) 損失情況下，在四個評分面向（內容 (a,e)、表達 (b,f)、語言使用 (c,g) 和整體熟練度 (d,h)）上的效能的混淆矩陣。上方行 (a-d) 顯示了 MA 模型的結果，下方行 (e-h) 顯示了 MA+MMO 的結果。每個矩陣都將真實 CEFR 等級（行）與預測等級（列）繪製出來，說明了準確性和序數對齊。

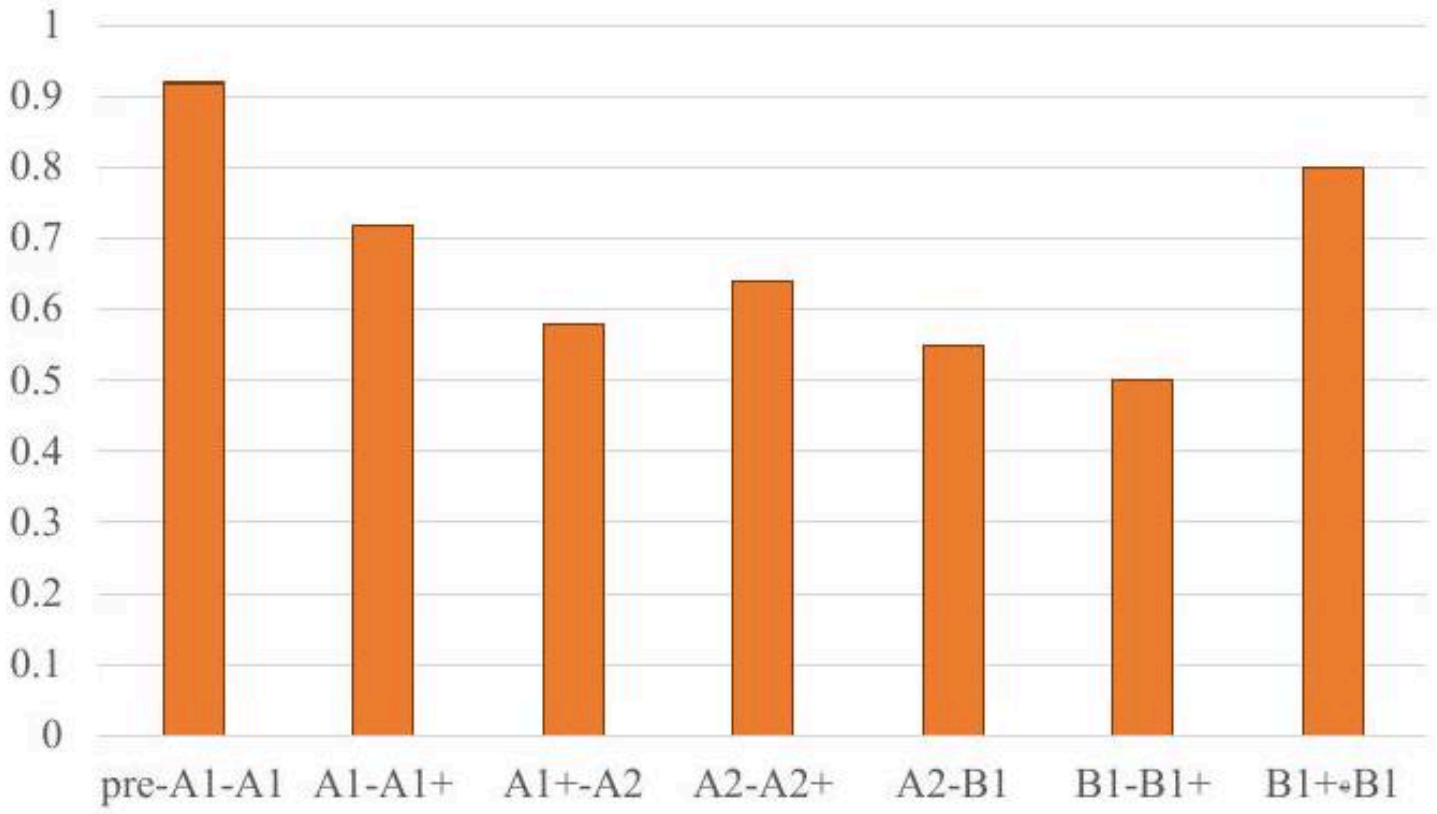


圖 5：圖 5. 從訓練實例計算出的相鄰 CEFR 等級之間的成對距離。

等級和參考標籤，提供類別行為和序數一致性的詳細視圖。

從 MA 模型獲得的結果顯示，相鄰等級之間經常發生混淆，例如 A2 與 A2+ 和 B1 與 B1+，特別是在內容和語言使用方面。還觀察到對中等程度（例如 A2+、B1）的過度估計偏差，這可能是由於訓練集中標籤分佈不平衡所致。加入 MMO 損失後，對角線集中度更清晰，並且遠距離等級之間的錯誤顯著減少。這些結果表明，MMO 損失引入了多個邊界，這些邊界考慮了等級間距離，從而提高了模型預測的性能。

D. CEFR 等級的距離分析

為了進一步探討 CEFR 能力等級的序數結構和潛在幾何，我們對相鄰類別標籤之間的語義距離進行了分析。如圖 5 所示，我們根據公式 9 計算了 CEFR 等級之間的成對距離。觀察到的距離顯示 CEFR 量表存在顯著的不均勻性。例如，Pre-A1 和 A1 之間的語義差距最大（0.9236），而 B1 到 B1+ 的過渡最小（0.4868）。這些發現經驗性地驗證了先前的觀察結果 [24]，證實了能力等級在標籤嵌入空間中並非等距的說法。這種不對稱過渡的存在挑戰了傳統序數分類方法中依賴固定邊界的假設，突顯了採用靈活、數據驅動策略（例如所提出的 MMO 損失）的必要性，以捕捉 ASA 中固有的非均勻進展。

VII. 結論與未來工作

本文提出了兩種用於自動口語評估（ASA）的創新建模策略：多面向分類和多邊界序數（MMO）損失。第一種策略旨在減輕 SSL 模型應用於 ASA 時的模態限制。第二種策略與多面向建模策略相容，同時解決了分數序數性和非均勻等級間隔的挑戰。在 TEEMI 資料集上的實驗證明了我們方法相對於先前方法的有效性。此外，對不同未見提示的評估證實了我們模型在不同 ASA 任務中的泛化能力。對於未來的工作，我們將計劃探索跨提示和評分面向的聯合訓練策略，以增強 ASA 在不同任務上的穩健性。此外，我們設想整合融合聲學和文本資訊的多模態大型語言模型（MLLM），從而提升 ASA 的全面性和可解釋性。

參考文獻

[1] A. Van Moere 和 R. Downey，「21. 語言評估中的科技與人工智慧」，第二語言評估手冊，第 12 卷，2016 年。
[2] C. Cucchiarini、H. Strik 和 L. Boves，「第二語言學習者流利度的量化評估：一種自動化方法」，收錄於 ICSLP 論文集，1998 年，第 2619-2622 頁。
[3] H. Strik 和 C. Cucchiarini，第二語言學習者流利度的自動評估。美國舊金山：sn，1999 年。
[4] L. Chen, K. Evanini, and X. Sun, “Assessment of non-native speech using vowel space characteristics,” in Proceedings of SLT, 2010, pp. 139-144.
[5] S. Bhat and S. Youn, “Automatic assessment of syntactic complexity for spontaneous speech scoring,” Speech

Communication, vol. 67, pp. 42-57, 2015.
[6] R. Moore, A. Caines, C. Graham, and P. Buttery, “Incremental dependency parsing and disfluency detection in spoken learner english,” in Proceedings of TSD. Pilsen, Czech Republic: Springer, 2015, pp. 470479.
[7] E. Coutinho, F. Hönig, Y. Zhang, S. Hantke, A. Batliner, E. Nöth, and B. Schuller, “Assessing the prosody of non-native speakers of english: Measures and feature sets,” in Proceedings of LREC, 2016, pp. 13281332.
[8] P. Müller, F. de Wet, C. van der Walt, and T. Niesler, “Automatically assessing the oral proficiency of proficient 12 speakers,” in Proceedings of SLaTE, 2009, pp. 29-32.
[9] 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 ROCLING, 2022, pp. 174-183.
[10] L. Chen, K. Zechner, S.-Y. Yoon, K. Evanini, X. Wang, A. Loukina, J. Tao, L. Davis, C. M. Lee, M. Ma et al., “Automated scoring of nonnative speech using the speechrater sm v. 5.0 engine,” ETS Research Report Series, vol. 2018, no. 1, pp. 1-31, 2018.
[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pretraining of deep bidirectional transformers for language understanding,” in Proceedings of NAACL, 2019, pp. 4171-4186.
[12] Y. Arase, S. Uchida, and T. Kajiwara, 「CEFR-based sentence difficulty annotation and assessment」，收錄於 EMNLP 論文集，2022 年，第 6206-6219 頁。
[13] F. Nadeem, H. Nguyen, Y. Liu, and M. Ostendorf, 「Automated essay scoring with discourse-aware neural models」，收錄於 BEA 論文集，2019 年，第 484-493 頁。
[14] T.-I. Wu, T.-H. Lo, F.-A. Chao, Y.-T. Sung, and B. Chen, 「Effective neural modeling leveraging readability features for automated essay scoring」，收錄於 SLaTE 論文集，2023 年，第 81-85 頁。
[15] H. Craighead, A. Caines, P. Buttery, and H. Yannakoudakis, 「Investigating the effect of auxiliary objectives for the automated grading of learner English speech transcriptions」，收錄於 ACL 論文集，2020 年，第 2258-2269 頁。
[16] J. T. Li, T.-H. Lo, B.-C. Yan, Y.-C. Hsu, and B. Chen, 「Graph-enhanced transformer architecture with novel use of cefr vocabulary profile and filled pauses in automated speaking assessment，」 載於 SLaTE 論文集，2023 年，頁 109-113。
[17] S. Bannò and M. Matassoni, 「Proficiency assessment of 12 spoken english using wav2vec 2.0，」 載於 SLT 論文集，2023 年，頁 10881095。
[18] S. W. McKnight, A. Civelekoglu, M. Gales, S. Bannò, A. Liusie, and K. M. Knill, 「Automatic assessment of conversational speaking tests，」 載於 SLaTE 論文集，2023 年，頁 99-103。
[19] S. Bannò, K. M. Knill, M. Matassoni, V. Raina, and M. Gales, 「Assessment of 12 oral proficiency using self-supervised speech representation learning，」 載於 SLaTE 論文集，2023 年，頁 126-130。
[20] T.-H. Lo, F.-A. Chao, T.-i. Wu, Y.-T. Sung, and B. Chen, “An effective automated speaking assessment approach to mitigating data scarcity and imbalanced distribution,” in Proceedings of NAACL, 2024, pp. 13521362.
[21] Y. Qian, P. Lange, K. Evanini, R. Pugh, R. Ubale, M. Mulholland, and X. Wang, “Neural approaches to automated speech scoring of monologue and dialogue responses,” in Proceedings of ICASSP, 2019, pp. 81128116.
[22] S. Park and R. Ubale, “Multitask learning model with text and speech representation for fine-grained speech scoring,” in Proceedings of ASRU, 2023, pp. 1-7.
[23] W.-H. Peng, S. Chen, and B. Chen, “Enhancing automatic speech assessment leveraging heterogeneous features and soft labels for ordinal classification,” in Proceedings of SLT, 2024, pp. 945-952.
[24] M. Heilman, K. Collins-Thompson, and M. Eskenazi, 「An analysis of statistical models and features for reading difficulty prediction，」 in Proceedings of BEA, 2008, pp. 71-79.
[25] F. M. Lord, Applications of item response theory to practical testing problems. Routledge, 2012.
[26] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, 「wav2vec 2.0: A framework for self-supervised learning of speech representations，」 in Proceedings of NeurIPS, vol. 33, 2020, pp. 12449-12460.
[27] F.-A. Chao, T.-H. Lo, T.-I. Wu, Y.-T. Sung, and B. Chen, 「3m: An effective multi-view, multi-granularity, and multi-aspect modeling approach to english pronunciation assessment，」 in Proceedings of APSIPA, 2022, pp. 575-582.
[28] Y.-Y. He, B.-C. Yan, T.-H. Lo, M.-S. Lin, Y.-C. Hsu, and B. Chen, “Jam: A unified neural architecture for joint multi-granularity pronunciation assessment and phone-level mispronunciation detection and diagnosis towards a comprehensive capt system,” in Proceedings of APSIPA, 2024, pp. 1-6.
[29] F.-A. Chao, T.-H. Lo, T.-I. Wu, Y.-T. Sung, and B. Chen, “A hierarchical context-aware modeling approach for multi-aspect and multi-granular pronunciation assessment,” in Proceedings of Interspeech, 2023, pp. 974-978.
[30] S.-Y. Chen, T.-H. Lo, Y.-T. Sung, C.-Y. Tseng, and B. Chen, “Teemi: a speaking practice tool for 12 english learners,” in Proceedings of Interspeech, 2024, pp. 2048-2049.
[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, “Attention is all you need,” in Proceedings of NeurIPS, vol. 30, 2017.
[32] A. D. et al., 「The llama 3 herd of models」，載於 arXiv 預印本 arXiv:2407.21783，2024。
[33] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning，「Stanza: A python natural language processing toolkit for many human languages」，載於 Proceedings of ACL，2020，頁 101-108。
[34] D. Pitawela, G. Carneiro, and H.-T. Chen，「Cloc: Contrastive learning for ordinal classification with multi-margin n-pair loss」，載於 Proceedings of CVPR，2025，頁 15 538-15 548。
[35] T. W. et al., 「Transformers: State-of-the-art natural language processing」，載於 Proceedings of EMNLP，2020，頁 38-45。

[1] https://huggingface.co/bert-base-uncased
[2] https://huggingface.co/facebook/wav2vec2-base