The 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)
Taipei City, Taiwan, October 20-21, 2023. The Association for Computational Linguistics and Chinese Language Processing

結合 BERT 與 Wav2vec 2.0 提升第二外語受試者之自動英語口說評測
Enhancing Automated English Speaking Assessment for L2 Speakers with BERT and Wav2vec2.0 Fusion

Wen-Hsuan Peng[1], Hsin-Wei Wang[1], Sally Chen[2], Berlin Chen[1]
[1]National Taiwan Normal University
[2]The Language Training & Testing Center
[1]{61147006s, hsinweiwang, berlin}@ntnu.edu.tw
[2]sallychen@lttc.ntu.edu.tw

摘要

英語逐漸作爲許多國家的第二語言 (English as a Second Language, ESL)，同時也帶動電腦輔助語言學習的發展，近年來又以發展自動口語評測較爲熱門。然而，英語口語能力評測的過程需要耗費許多人力，也相當費時。因此，建立出一套自動英語口語評分的方法不但能節省人力、時間，亦能提供更加一致的評估標準。本研究中我們使用公開資料集 ICNALE，建構一套融合 BERT 和 Wav2vec 2.0 模型來進行自動英語口說能力分級。研究結果顯示，整合文字與語音的模型表現優於以人工轉錄訓練的 BERT 模型和單獨的 Wav2vec 2.0 模型。

Abstract

Due to the increasing popularity of English as a second language, there has been a growing interest in developing Compute-assisted Language Learning (CALL) applications that focus on automated assessing of spoken language proficiency. In the past, evaluating English speaking proficiency has been a time-consuming and labor-intensive process. Therefore, developing an efficient method for automated grading can establish consistent evaluation standards in a more timely and cost-effective manner. In this study, we explore the fusion of BERT and Wave2vec2.0 modeling strategies to assess holistic English speaking proficiency scores, withe an extensive set of experiments conducted on the publicly available ICNALE dataset. The experimental results indicate the superiority of our approach in relation to the existing baselines.

關鍵字：自動發音檢測、英語能力分級、多模態系統

Keywords: Automatic assessment of spoken language proficiency, Compute-assisted language learning, Multi-modal system

1 緒論

隨著全球化的影響，參加標準化英語測驗的受試者與日俱增，再加上新冠肺炎的影響之下，對於線上教育的需求大幅提升，也逐步推動語言學習輔助工具 (Computer-Assisted Language Learning, CALL) 相關研究的增長，早期電腦語言學習工具，主要以輔助發音訓練 (Computer-Assisted Pronunciation Training, CAPT) 最爲熱門。在題目以朗讀爲主的口說練習中，受試者會先根據語言學習工具所提供的文本提示 (prompt) 進行朗讀，並透過自動語音辨識 (Automatic Speech Recognition, ASR) 檢視受試者的音素序列 (Phoneme Sequence)，再與系統中母語者的規範音素 (canonical phone) 進行比對，以提供語者發音正確與否的回饋。近年來，有許多學者投入到基於第二外語受試者的口說評測研究，相較於先前的發音評測，考慮重音、韻律、流暢程度，口說評測不僅需要涵蓋到發音，更需要考慮用字遣詞、文法以及內容等部分。

在過去對於整體面向或是單一面向的口說評測方法，多半使用由人工收集或製作的特徵，而人工所製作的特徵有很大的程度仰賴當時的基本假設，並且可能會遺漏部分的重要面向。針對整體面向的評測問題，目前已能透過端對端系統 (Chen et al., 2018)，或是多階段模型的方法 (Cheng et al., 2020)，以自動生成特徵來代替人工製作的特徵。在發展第二外語受試者的自動口說評測中，過去的模型會先使用 ASR 技術將受試者的回答轉換成文字，其中包含識別音素、音節、單詞和聲學特徵等元素，並將這些元素進行強制對齊，以提取出相關的特徵。識別後的詞序列接著會被輸入到自然語言處理模塊，以生成與詞彙、語法、內容和結構相關的特徵。上述特徵皆經由人工標記過後，用於訓練口說評測模型，以預測等級。

先前提到的整體口說評測研究中，模型除了考慮原本的輸入資料，還包括了特定面向的資訊，如：發音、韻律、文字等資訊，然而，即

便這些資訊有助於發展出特定面向的模型，但仍然局限於人工可以標記的資訊範圍。此外，使用 ASR 轉錄的資訊同樣存在風險，ASR 本身具有部分的詞錯率 (Word Error Rate)，因此無法完整捕捉受試者的回答內容。儘管 ASR 可以提供部分發音的訊息，但仍無法提供其他重要面向的資訊，例如：語調、節奏、情感等，而這些資訊對 CALL 是重要的依據。

爲了解決先前提到的問題，本研究使用基於 BERT(Baevski et al., 2018) 和 Wav2vec2.0(Baevski et al., 2020) 的自我監督式學習 (Self-supervised Learning, SSL) 表示法來進行實驗。最近的研究指出，自我監督式學習能夠有效的處理語音的下游任務，例如 ASR、關鍵詞偵測、語者識別等領域。在這些研究中，多半使用預訓練模型的上下文表示法 (contextual representation)，並已證實這些預訓練模型能夠從不同語言水平的語者（如：L1, L2）中，提取出流暢度、發音、句法，甚至語意特徵。(Tsai et al., 2022) 在電腦輔助語言學習領域中，自我監督學習的表示法目前已成功被應用於發音錯誤檢測和發音診斷 (Peng et al., 2021)，以及自動發音評估 (Kim et al., 2022)。此外，BERT 的預訓練模型在自動文章評分或文章可讀性 (Deutsch et al., 2020)(Martinc et al., 2021) 等研究中展現出卓越的效果，特別是在捕捉文章的語言特徵方面，例如語義、詞彙和文章上下文的一致性等。

在 (Stefano Bannò, 2022) 中，作者運用 BERT 和 Wav2vec2.0 兩個預訓練模型進行英語口說評測的實驗，研究發現 Wav2vec2.0 在公開語料集 ICNALE 上的表現達到 77.8% 的準確率。受此研究之啓發，我們進一步比較與延伸此方法。爲了方便進行分析與比較，本研究選擇在相同的公開語料集 ICNALE 上進行實驗，該數據集包含五個不同等級的標籤資訊，並以歐洲通用語言參考框架 (Common European Framework of Reference for Language, CEFR) 作爲評估標準，基於上述背景，我們進一步提出了融合 BERT 和 Wav2vec2.0 的自動口說評測架構，整合文字與語音的資訊，有效將英語口說表現分級。

2 預訓練模型 (Pre-trained Model)

2.1 Wav2vec2.0 模型

Wav2vec2.0 爲 Facebook AI 於 2020 所開發的自我監督預訓練模型。(Baevski et al., 2020)。其中包含三個模塊，特徵編碼器 (feature encoder)$f : X \mapsto Z$，上下文變換器 (contextual block transformer block)，$g : Z \mapsto C$ 還有量化模塊 (quantization block)$Z \mapsto Q$。目標是將語音數據轉換成有意義的向量或表示法。如圖 1 所示。

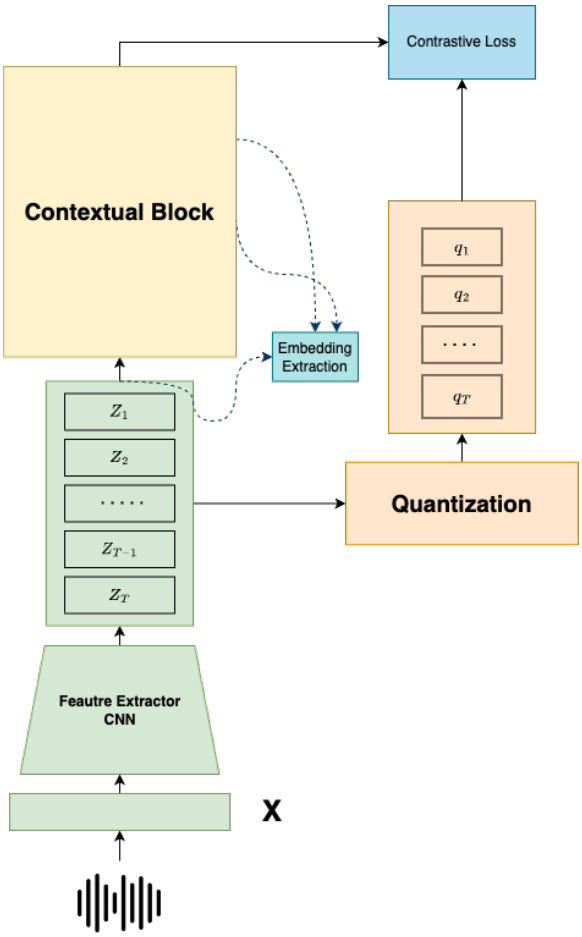

圖 1. Wav2vec2.0 架構圖

特徵編碼器是由多層一維卷積塊組成，原始輸入 X 經過批量正則化（Batch Normalization）和 GELU 激活函數標準化後，將其編碼成局部特徵表示，$Z = f(x)$。接著，這些大小爲 $Z^{T\times 768}$ 的特徵表示，將被送入到 contextual transformer 模組，以學習上下文的語音表示，$C = g(Z)$。同時，特徵表示 Z 也會傳入到由兩個編碼書 (codebook) 所組成的量化模塊。由於每個編碼書共有 320 種可能項目，對於每個 Z 中的向量表示，$z_i \in Z$，經由公式 (1) 後，會形成一個大小爲 $R^{2\times 320}$ 的 logit，並透過連接每個編碼書的相應項目，經過線性轉換後，生成出局部特徵編碼器表示 $z_i \in Z$ 的量化向量 q_i。

$$p_{g,v} = \frac{exp(l_g, v + \eta_v/\tau)}{\Sigma_{k=1}^{V} exp(l_g, v + \eta_v/\tau)} \quad (1)$$

其中，l 代表 logit，v 代表第 v 個編碼書項目，g 是編碼書群組，$\eta = -log(-log(u))$ 其

中 u 是從 $U(0,1)$ 均勻抽樣的樣本，而 τ 則是控制隨機性的參數。

模型的訓練方法以自我監督的方式進行預訓練。此方法和遮罩語言模型類似，透過公式 (2) 隨機遮蔽某些時間點的特徵表示向量。訓練目標是從一組 $K+1$ 個干擾項 (distractors)，重新生成量化的 $\tilde{q}_t$，其中候選向量包含 q^t 和 K 個 $\in Q$ 的干擾項，而這些干擾項是由相同語音片段的遮罩中均勻取樣而得到。

$$L_{cont} = -log\frac{exp(sim(c_t, q_t)/\tau}{\Sigma_{\tilde{q}\in Q} exp(sim(c_t, q_t)/\tau} \quad (2)$$

Wav2vec2.0 的自我監督學習模型先在 960 小時的 LibriSpeech 資料集上進行預訓練。作爲上游模型，Wav2vec2.0 預訓練模型在語音處理方面表現卓越，在 (Fan et al., 2021) 中，作者將 Wav2vec 應用於多任務學習，利用 Wav2vec2.0 作爲音頻編碼器，以提取語者和語言的特徵，其研究結果證明 Wav2vec 在語者識別和語音辨識相關任務的有效性。同時，在 (Pepino et al., 2021) 中，作者採用預訓練的 Wav2vec2.0 模型來實現語音情感識別任務，從而展示了預訓練的 Wav2vec2.0 模型也能夠有效捕捉豐富的語音信息。

2.2 BERT 模型

BERT，全名爲 Bidirectional Encoder Representations from Transformers，是由 Google 於 2018 年所開發的預訓練語言模型，其架構是由多層 Transformer 所建構，每層包含多頭自注意力 (Multi-head self-attention) 和殘差連接 (Residual connection) 的全連接子層。BERT 預訓練過程使用遮罩語言模型（Masked Language Model，MLM）和下一句預測 (Next Sentence Prediction, NSP) 進行實驗。該模型在 BooksCorpus（800M 單詞）(Zhu et al., 2015) 和英文維基百科上 (2,500M 單詞) 進行了預訓練，資料集涵蓋各種主題和領域，提供廣泛的語言模式和上下文信息。經由上述方法的訓練後，BERT 預訓練模型得以學習到豐富的表示法，而經過微調 (fine-tuning) 後的預訓練模型，在其他自然語言處理任務上均能獲得不錯的效果，例如：文本分類、命名實體識別、影像生成 (Niki Parmar, 2018)、機器翻譯、問答 (Dehghani et al., 2019)、語言理解。

相比於傳統的模型，BERT 有四項特點，一、遮罩語言模型（Masked Language Model，MLM）：BERT 使用遮罩語言模型來進行預訓練。在預訓練過程中，它隨機地將輸入文本中的一些單詞遮蔽，並使用特殊符號 [MASK] 替換，目標是讓模型能夠預測被遮蔽的單詞，參見表 1，使得 BERT 能夠學習到詞彙之間的上下文關係，並具有更好的語義理解能力。二、下一句預測：模型會根據給定的文本序列，預測下一句話。在進行預測時，輸入資料通常會是兩句帶有特殊分隔符的文本，模型會對這個文本進行預測，並判斷句子 1 是否爲句子 2 的接續句。若是，標記爲 1(IsNext)，反之，則標記爲 0(NotNext)，參見表 2。訓練數據中的正樣本爲眞正的接續句，而負樣本則是隨機或是不相關的句子。三、雙向連接模型，與傳統的由左到右或由右到左的單向語言模型，BERT 使用雙向的 Transformer 編碼器，能夠同時考慮一個詞左右兩側的單詞，模型不只能夠考慮到單詞的特性，也能考慮到前後文，捕捉詞彙之間的關係。四、單詞片段 (WordPiece Embedding) (Yonghui Wu, 2016) 分詞：BERT 使用單詞片段分詞技術對單詞進行拆分，形成更精細的單詞表示。例如：假設有三個英文單字，play, playing, player，模型會將這些詞切分成基本形式和其對應的後綴（suffix）。例如，“play” 會被拆分成基本形式 “play” 以及後綴 “ing” 或 “er”。此方法使得 BERT 能夠更好地處理未知單詞和罕見單詞，並能夠更好地理解單詞的上下文。

Sentence
It is [MASK1] to [MASK2] that
Label
$[MASK1] = important; [MASK2] = say$

表 1. 遮罩語言預測範例

Sentence
[CLS]The weather is nice today. [SEP] Let's go to the park. [SEP]
Label
IsNext
Sentence
[CLS] The weather is nice today. [SEP] I love to read books. [SEP]
Label
NotNext

表 2. 下一句模型預測範例

此外，BERT 的輸入標記由三個部分相加而形成，如圖 2 所示。一、標記嵌入 (Token Embedding)，目的是將每個字符轉換成固定的維度向量表示。二、位置嵌入 (Position Embedding)，表示每個字符的位置資訊。三、分段嵌入 (Type Embedding)，將文本信

息分割，用來表示不同的樣本。每個文字訊息中各自的標記，皆會被送入 BERT 的標記嵌入層、位置嵌入層和分段嵌入層，分別得到向量 V1, V2 和 V3，最後將這三項加起來，輸入到 BERT 模型中。

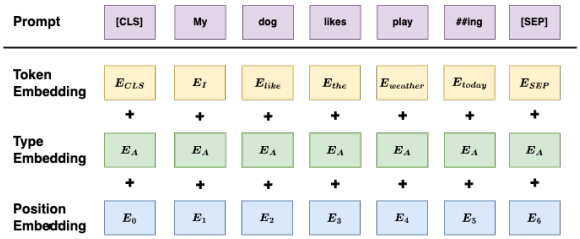

圖 2. BERT 輸入表示示意圖

3 模型架構

3.1 BERT 評分器

圖 3 爲三個評分器的模型架構圖。在進行 BERT 評分器的實驗中（圖3中的 a），我們使用 HuggingFace Transformer Library(Wolf et al., 2020)[1]所提供的預訓練模型，將一連串的標記 (Token) 進行嵌入。我們使用的公開語料集已包括人工音檔轉錄的結果，因此可將受試者回答的文字內容作爲輸入，並將其傳遞到 BERT 的編碼器層。在分類過程中，取用最後的 [CLS] 的隱藏狀態，並將其輸入到多層感知器（Multi-perceptron layer）進行分類。在訓練過程中，我們會固定 (Freeze)BERT 模型，使模型無法進行參數更新。

3.2 Wav2vec2.0 評分器

在 Wav2vec2.0 中（圖3中的 b），語音訊息透過多層卷積神經網路 (Convolution Neural Network, CNN) 進行編碼，並對生成的潛在表示法進行遮罩，輸入到 Transformer 中以建立表示法。訓練模型的過程中，使用 Gumbel Softmax 計算對比損失。本實驗，使用 HuggingFace Transformer Library(Wolf et al., 2020)[2]所提供的預訓練模型來初始化模型的配置跟音訊的前處理。受試者的回答輸入到模型後，Wav2vec2.0 會生成對應的表示法。爲了處理不同長度的音訊，我們採用平均池化方法，將原本大小爲 3 維的向量 (即批次大小、長度和隱藏層數) 轉換爲 2 維向量（批次大小、隱藏層數），最後再經過多層感知器得到該類別的等級。和 BERT 評分器相同，訓練過程中，我們也會固定住 Wav2vec 模型參數，使其在訓練過程中不被更新。

[1]huggingface.co/bert-base-uncased
[2]huggingface.co/facebook/Wav2vec2-base

3.3 融合 BERT 與 Wav2vec2.0 評分器

本實驗除了分別使用專門針對文字和語音的模型進行評估之外，更進一步探討整合兩種模型的方法，以評估同時考慮文字和語音資訊對英語口語評測的有效性，參見圖3中的 c。爲此，我們先將 BERT 評分器和 Wav2vec2.0 評分器各自的輸出結果透過線性組合的方式整合，再輸入到多層感知器進行分類。透過上述的整合，過程不僅能平衡文字和語音資訊的相對重要性，更可以彌補單一模型的限制，從而得到精確的綜合評分結果。

4 實驗設定

4.1 資料集

本次實驗所使用的資料集，爲國際亞洲英語受試者語料庫 (ICNALE)(shikawa, 2023) 的公開語料集，語料集使用的評分框架基於歐洲語言參考框架（CEFR），包含了從 A2 到 B2 的受試者，與部分母語人士。受試者國籍涵蓋中國、香港、印度尼西亞、日本、南韓、巴基斯坦、菲律賓、新加坡、泰國和台灣。受試者 CEFR 等級評估方法，主要基於受試者最初參與的詞彙量測驗表現。同時，也會收集他們在托福、多益、雅思等國際認可的英語能力測試中的成績。綜合評估方法不僅強調多元的評估策略，也能反映了受試者在不同英語能力方面的表現，以利評估受試者的 CEFR 等級。

語料庫的口說分爲兩個部分，獨白和對話，在本次的實驗中，只使用到獨白的部分，語料總共有 4332 個回答，每個回答的答題時間介於 36 秒到 69 秒之間。題目內容，請受試者闡述他們認爲打工的重要性以及對於在餐廳裡吸菸有什麼看法。爲了與基線方法比較時，能保持一致性，在數據集的切分上參考了基線所使用的資料切割方式 (Stefano Bannò, 2022)。數據集中，訓練集共包含 3898 個回答，而開發集和測試集則各有 217 個回答。資料被劃分爲五個類別，分別爲 A2、B1-1、B1-2、B2 以及母語者等級 (NS)。詳細的標籤分布參見表 3。

	Train	Dev	Test	Total
A2	299	16	17	332
B1_1	792	44	44	880
B1_2	1681	94	93	1868
B2	586	33	33	652
Native(NS)	540	30	30	600
Total	3898	217	217	4332

表 3. ICNALE 語料集分佈

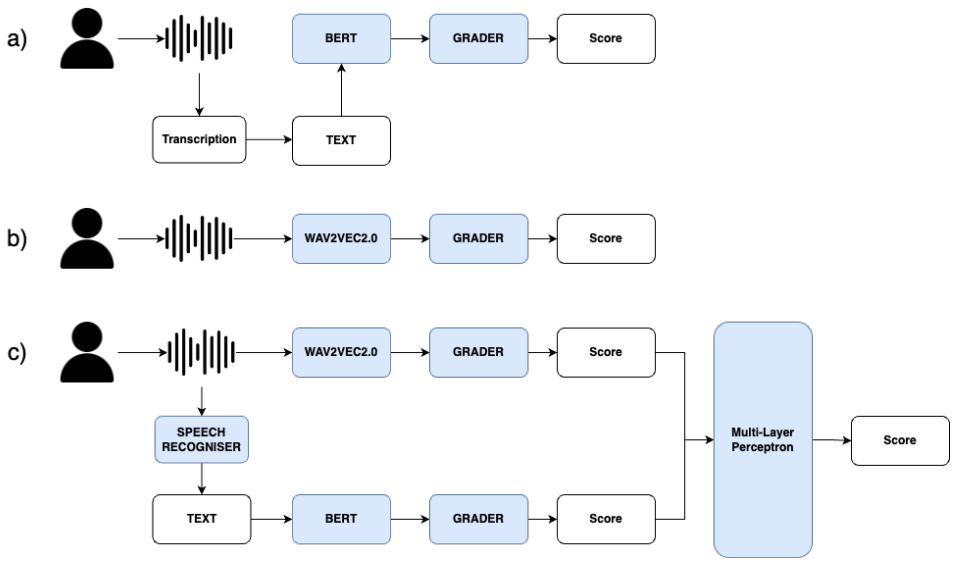

圖 3. 本研究使用的三種模型：a) BERT-based 評分器, b) Wav2vec2-based 評分器, c) 融合 BERT 和 Wav2vec2 評分器

	Epochs	Learning Rate	Dropout
BERT	600	5e-5	-
Wav2vec2	8	1e-5	0.2
BERT+ Wav2vec2	8	1e-5	0.1

表 4. 三個評分器的超參數配置

4.2 任務評估指標

在口說評測的分類任務中，我們使用準確率與 Weight F1 作爲評估指標。準確率 (Accuracy) 可以了解模型在整個資料集上正確的分類表現。Weighted F1 能更全面地評估模型性能，即使資料呈現不平衡的狀態，也能夠兼顧模型在每個類別的表現。計算方式如下：

$$Accuracy = \frac{TP + TN}{Total\ Number\ of\ Samples} \quad (3)$$

$$precision = \frac{TP}{TP + FP} \quad (4)$$

$$recall = \frac{TP}{TP + FN} \quad (5)$$

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} \quad (6)$$

$$Weighted\ F1 = \frac{\sum_{i=1}^{N} w_i \cdot F1_i}{\sum_{i=1}^{N} w_i} \quad (7)$$

其中 TP、TN、FP、FN 分別代表四種可能的預測情況：True Positive(TP) 將正確預測爲正確；True Negative(TN) 將錯誤預測爲錯誤；False Positive(FP) 將錯誤預測爲正確；False Negative(FN) 將正確預測爲錯誤。

4.3 實驗設定

本研究爲多類別分類任務，考量到模型間的輸入資料類型不同，因此我們對不同評分器使用了不同的實驗設定，請參見表 4。在 BERT 評分器的架構中，首先得到 BERT 的表示法，並將其輸入到由三個具有 768 個神經元和三個具有 128 個神經元的全連接層，再輸入到輸出層。其中，輸出層包含 5 個神經元並使用 softmax 作爲激活函數。訓練過程中，損失函數使用交叉熵 (Cross-Entropy) 訓練最小誤差，並使用 AdamW 作爲優化器。此外，批次大小設爲 256，學習率設定爲 5e-5，最大序列長度限制爲 256，整個訓練過程共進行 600 次迭代。

Wav2vec2.0 評分器在獲得 Wav2vec2 的表示法後，將其輸入到由 768 個神經元組成的全連接層，隨後連接到 5 個神經元的輸出層，並使用 softmax 作爲激活函數。訓練過程中，使用交叉熵作爲損失函數，並使用 AdmaW 作爲模型的優化器，其他訓練參數包括，批次大小設爲 4，梯度累積步數爲 2，丟失率（dropout）設爲 0.2，學習率爲 1e-5，並進行總共 8 次的迭代。

在融合 BERT 與 Wav2vec2.0 評分器中，首先獲得經由 BERT 評分器以及 Wav2vec2 評分器輸出的結果，並經過簡單的線性組合後，

得到一個綜合的等級資訊，而這項結果將輸入到由三個 768 神經元的全連接層和 5 個神經元的輸出層。訓練階段同樣使用交叉熵作爲損失函數，以 AdamW 爲優化器，並設定批次大小 4，梯度累積步數 4，丟失率 0.1，學習率 1e-5，共進行 8 次迭代。

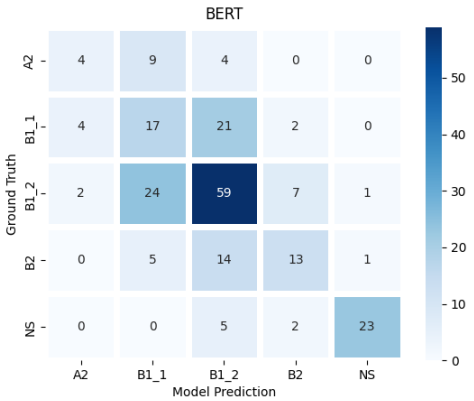

圖 4. BERT 評分器

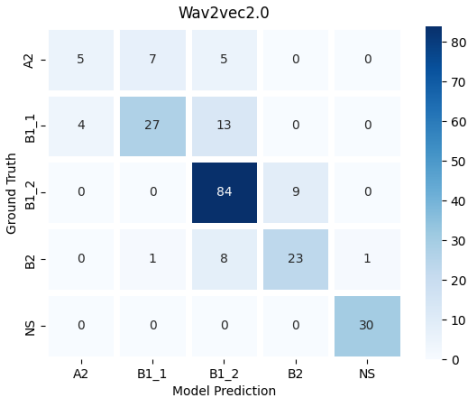

圖 5. Wav2vec2.0 評分器

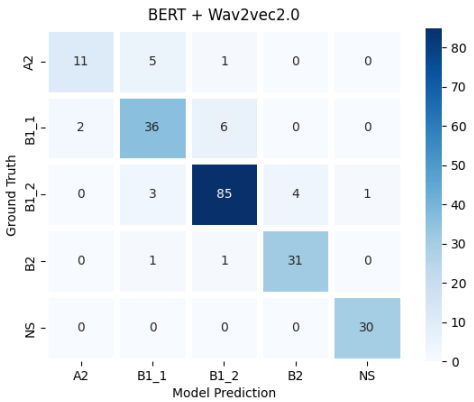

圖 6. 融合 BERT 和 Wav2vec2 評分器

4.4 實驗結果與討論

本研究與基線模型進行比較 (Stefano Bannò, 2022)。在實驗設定上，除了部分超參數和 Wav2vec2.0 的預訓練模型之外，微調方法基本相同。根據表5的結果，我們提出的整合文字與語音資訊的方法達到了 88.9% 的準確率，明顯高於原本的方法。

我們在 ICNALE 的資料集上共嘗試三次實驗，建立三個評分器，分別是 BERT 評分器，Wav2vec2 評分器以及融合 BERT 和 Wav2vec2 評分器。表5顯示，單獨的 Wav2vec2 評分器在準確率、Weight F1、Micro F1 和 Macro F1 各方面均優於單獨的 BERT 模型。然而，融合 BERT 和 Wav2vec2 的評分器在所有評估指標中，皆明顯高於其他兩者，準確率達到 88.94%，表示模型的整合方法能更全面地捕捉文字和語音信息，從而獲得更準確且有效的評測結果。此外，透過 Macro F1 和 Micro F1 的比較，還可以看出融合 BERT 和 Wav2vec2 的評分器不僅能考慮到整體性，在不同的類別之間也達到了良好的平衡。

圖 3、圖 4 和圖 5 呈現了三個評分器在每個 CEFR 等級下的混淆矩陣。根據表6所呈現的結果，不論在 BERT 評分器還是 Wav2vec2 評分器中，相較於 A2、B1-1、B1-2 和 B2，模型在母語者 (NS) 的分類表現最佳。我們推測是因爲母語人士與中低程度的非母語英語受試者之間存在明顯的英語水平差異。因此，模型能夠捕捉到這種差異，並有效地對其進行分類。另外，BERT 和 Wav2vec 模型中，分類效果不佳可能與訓練資料中不同類別的數量差異有關，資料不平衡會影響模型的區分能力。例如，A2 與 B1-2 之間的訓練資料相差超過 1000 筆（參見表 3），導致 A1 類別在這兩種模型中都難以被準確區分。

本研究提出融合文字和語音的模型，能有效的處理上述的問題，在圖 6 的混淆矩陣中，能夠看到明確的對角線，分隔出五個類別。儘管在 A2、B1-1、B1-2 之間仍存在一些模糊的區分，但是針對 A1 級別的受試者，本模型的分類效果也明顯優於其他兩個模型，請參見表 6。展現出融合文字與語音對於單一資訊的模型有一定的提升效果。

5 結論與未來展望

本研究提出了融合 BERT 和 Wav2vec2 評分器模型，透過整合文字和語音特徵，以及簡單的線性組合，彌補各自模型缺乏的部分特徵。從 ICANLE 資料集的實驗結果證實，我們發現融合 BERT 和 Wav2vec 的模型不僅結合了

	Accuracy(%)	Weighted F1	Micro F1	Macro F1
BERT	53.45	0.53	0.51	0.54
Wav2vec2	77.88	0.77	0.72	0.78
BERT + Wav2vec2	**88.94**	**0.88**	**0.87**	**0.89**

表 5. 三個評分器的分類表現

Precision	A2	B1-1	B1-2	B2	NS
BERT	0.40	0.31	0.57	0.54	0.92
Wav2vec2	0.56	0.77	0.76	0.72	**0.97**
BERT + Wav2vec2	**0.85**	**0.80**	**0.91**	**0.87**	**0.97**

表 6. 三個評分器的在各個標籤之精確率比較

兩項重要特徵，在效能方面也展現出超越了單一模型的表現。本研究僅是對於英語口說評測的初步實驗，未來將繼續發展更加嚴謹的英語口說評測模型，讓模型更趨完善。

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. arXiv:2006.11477.

Lei Chen, Jidong Tao, Shabnam Ghaffarzadegan, and Yao Qian. 2018. End-to-end neural network based automated speech scoring. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6234–6238.

Sitong Cheng, Zhixin Liu, Lantian Li, Zhiyuan Tang, Dong Wang, and Thomas Fang Zheng. 2020. ASR-Free Pronunciation Assessment. In *Proc. Interspeech 2020*, pages 3047–3051.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. 2019. Universal transformers. In *International Conference on Learning Representations*.

Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. 2020. Linguistic features for readability assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–17, Seattle, WA, USA → Online. Association for Computational Linguistics.

Zhiyun Fan, Meng Li, Shiyu Zhou, and Bo Xu. 2021. Exploring wav2vec 2.0 on Speaker Verification and Language Identification. In *Proc. Interspeech 2021*, pages 1509–1513.

Eesung Kim, Jae-Jin Jeon, Hyeji Seo, and Hoon Kim. 2022. Automatic Pronunciation Assessment using Self-Supervised Speech Representation Learning. In *Proc. Interspeech 2022*, pages 1411–1415.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179.

Jakob Uszkoreit Łukasz Kaiser Noam Shazeer Alexander Ku Dustin Tran Niki Parmar, Ashish Vaswani. 2018. Image transformer. arXiv:1802.05751.

Linkai Peng, Kaiqi Fu, Binghuai Lin, Dengfeng Ke, and Jinsong Zhan. 2021. A Study on Fine-Tuning wav2vec2.0 Model for the Task of Mispronunciation Detection and Diagnosis. In *Proc. Interspeech 2021*, pages 4448–4452.

Leonardo Pepino, Pablo Riera, and Luciana Ferrer. 2021. Emotion Recognition from Speech Using wav2vec 2.0 Embeddings. In *Proc. Interspeech 2021*, pages 3400–3404.

S. shikawa. 2023. *The ICNALE Guide: An Introduction to a Learner Corpus Study on Asian Learners' L2 English*.

Marco Matassoni Stefano Bannò. 2022. Proficiency assessment of l2 spoken english using wav2vec 2.0. arXiv:2210.13168.

Hsiang-Sheng Tsai, Heng-Jui Chang, Wen-Chin Huang, Zili Huang, Kushal Lakhotia, Shu-wen Yang, Shuyan Dong, Andy Liu, Cheng-I Lai, Jiatong Shi, Xuankai Chang, Phil Hall, Hsuan-Jui Chen, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. 2022. SUPERB-SG: Enhanced speech processing universal PERformance benchmark for semantic and generative capabilities. pages 8479–8492, Dublin, Ireland. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

Zhifeng Chen Quoc V Le Mohammad Norouzi Wolfgang Macherey Maxim Krikun Yuan Cao Qin Gao Klaus Macherey et al. Yonghui Wu, Mike Schuster. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv:1609.08144.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 19–27.