MuFFIN：具互動階層式神經建模的多面向發音回饋模型

顏碧成，IEEE 學生會員，蔡明康，及陳柏霖，IEEE 會員

摘要

電腦輔助發音訓練 (CAPT) 透過提供即時且具指導性的回饋，協助第二語言 (L2) 學習者練習發音技巧。為了從多個面向檢視發音能力，現有的 CAPT 方法大致可分為兩類：錯誤發音偵測與診斷 (MDD) 以及自動發音評估 (APA)。前者旨在定位語音發音錯誤並提供診斷回饋，後者則著重於針對各種面向量化發音熟練度。儘管 MDD 與 APA 之間具自然互補性，研究者與實務工作者卻常將兩者視為獨立任務並採用不同的建模範式。有鑑於此，本文首先提出 MuFFIN，一種具互動階層式神經架構的多面向發音回饋模型，以共同處理 MDD 與 APA 任務。 為了在特徵空間中更好地捕捉音素之間的細微差異，本文提出了一種新穎的音素對比序位正則化機制，用以優化所提出的模型，以生成更具音素可區辨性的特徵，同時考量面向分數的序位性。此外，為了解決 MDD 中複雜的資料不平衡問題，我們設計了一個簡單但有效的訓練目標，專為以音素特定變異擾動音素分類器的輸出而量身打造，以便在考慮其誤讀特性時更好地呈現預測音素的分佈。在 Speechocean762 基準資料集上進行的一系列實驗顯示，我們的方法相較於多個最先進基線具有顯著效能，並在 APA 與 MDD 任務上展現出最先進的表現。

索引詞—Computer-assisted pronunciation training、automatic pronunciation assessment、mispronunciation detection and diagnosis、多面向與多粒度發音評估、對比式學習。

I. 引言

隨著對外語習得需求大幅增加，電腦輔助發音訓練（CAPT）的研究在全球化浪潮中引起了高度關注，並在電腦輔助語言學習（CALL）領域中佔據重要地位 [1][2]。為了彌補語言教師和學習者之間供給不足與迫切需求之間的差距，CAPT 系統已成為普遍吸引人的學習工具，將傳統的以教師為主導的教學範式轉向自我導向學習。

除了在教育與語言學習中的關鍵角色外，CAPT 系統也為高風險評量中的專業人士（例如面試官與考官）提供便利的參考，目的是減輕工作量 [3][4]、減少招募新的人力專家負擔，並達成一致且客觀的評量結果 [5][6][7]。

對於 CAPT（電腦輔助發音教學）而言，一個事實上的典型系統通常會在朗讀場景中實例化，該場景會提供給 L2 學習者一段參考文本並指示其正確發音。將學習者的語音與參考文本配對作為輸入後，CAPT 系統預期能從多面向評估學習者的口語能力，並以接近即時的反饋速度提供詳細且具診斷性的表現回饋。為此，誤讀偵測與診斷（MDD）以及自動發音評估（APA）是開發 CAPT 發音回饋模組時兩條活躍的研究脈絡。前者旨在定位語音發音錯誤並提供對應的診斷性回饋[8][9]。相較之下，後者則較著重透過多面向的發音分數來評估學習者的發音品質，反映其在特定面向或某些口語使用層次上的熟練度 [10][11]。 一種經過實證的 MDD 方法是發音優度 (GOP) 及其衍生方法 [12][13]，它們計算典型發音音素與最可能發音音素之間機率的比值。當某些音素區段的機率比低於預定閾值時，即可偵測到音素層級的錯誤發音。在另一個面向，具象徵性 APA 方法的模型通常基於表面特徵（即一組手工設計的特徵）來模擬人工評分。這些模型要麼使用分類器來預測代表學習者口語能力的整體分數 [10]，要麼使用迴歸器來估計特定發音面向的連續解析性分數，例如音素層級的準確度 [14]、詞級的詞彙重音 [15] 以及語句層級的發音品質 [16][17]。

儘管 MDD 與 APA 在性質上互補，多數現有工作仍將它們視為獨立任務，從而在 CAPT 中開發出兩個不同的回饋模組。

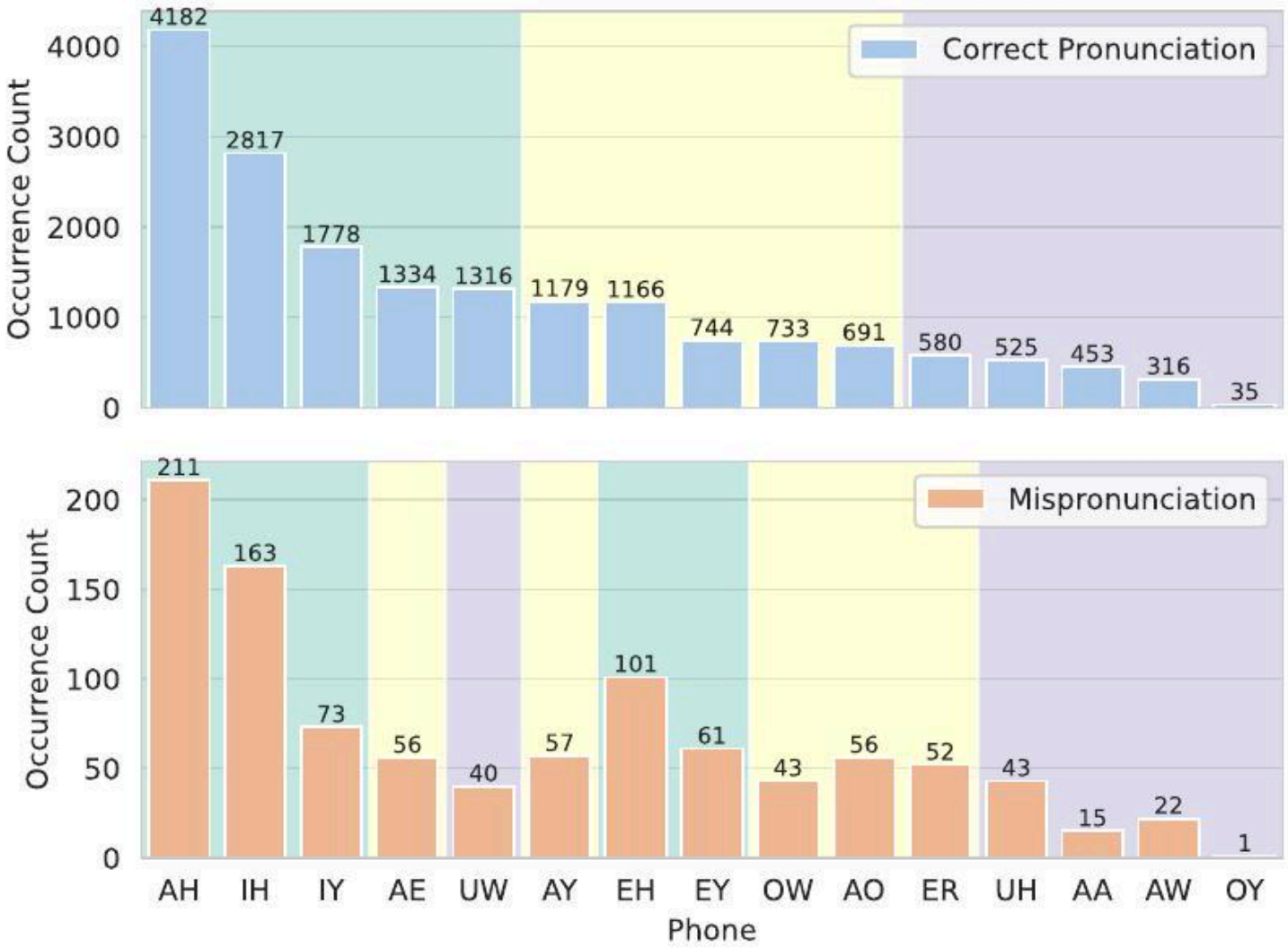


圖 1：圖 1. 在 Speechocean762 資料集中存在 MDD 的資料不平衡問題（☐ 多樣本，☐ 中等樣本，☐ 少樣本），正確發音的出現頻率明顯高於誤發音。此外，正確與錯誤發音呈現兩個明顯不同的長尾分布。

然而，先前一些研究顯示，當第二語言英語學習者的語句中頻繁出現音素層級的發音錯誤時，其可理解度與流暢度等語句層級的評估分數往往較低 [18][19][20]。基於此，本論文首先提出一種新穎的 CAPT 建模範式，稱為 MuFFIN，一個具互動分層神經結構的多面向發音反饋模型（Multi-Faceted pronunciation Feedback）。MuFFIN 透過多任務學習機制，將 MDD 與 APA 的各個獨立反饋模組統一為一個精簡且分層的神經架構。基於一個具語言層級感知的神經架構，並採用為其量身打造的 convolution-augmented Branchformer 區塊，MuFFIN 能有效捕捉不同語言粒度（即音素、詞與語句）間的互動，並在不同語言單元上保留細緻的發音線索。接著，為了在特徵空間中呈現音素間的微妙差異，我們引入了一種新穎的 phoneme-contrastive ordinal regularizer，以協助所提出的模型生成更具音素可鑑別性的特徵。 此訓練機制利用對比學習，使評分模型的音素表示能更好地與其對應之典型音素的文字嵌入對齊，同時考量回歸目標之序關係（即音素層級的準確度分數）。此外，我們探討了一個簡單但有效的訓練目標—音素特定變異（phoneme-specific variation），以緩解 MDD [21] 所帶來的資料不平衡問題。資料不平衡是 MDD 長期存在的問題，正確與錯誤發音實例之間的音素分布常常偏斜。如圖 1 所示，我們展示了 Speechocean762 資料集（見後續第 IV 節）訓練集中元音的正確與錯誤發音分布，並進一步將其分類為多次樣本（many-shot）、中等次樣本（medium-shot）與少次樣本（few-）

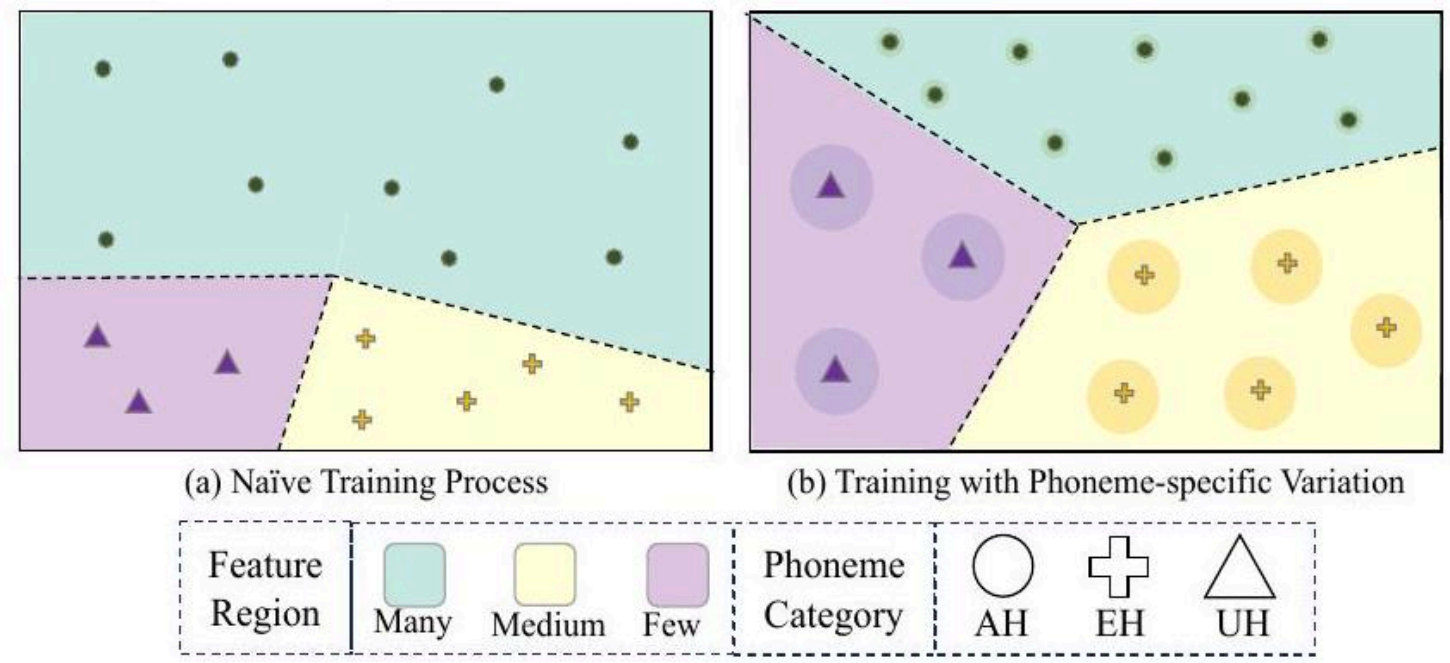


圖 2：圖 2。所提出之音素特定變異的動機。在特徵空間中，每個點代表由音素分類器所預測的一個資料實例，不同顏色表示不同類別。(a) 標準的訓練過程傾向於對多數音素類別產生偏置，導致少數音素類別被壓縮到狹窄區域。(b) 透過應用所提出的音素特定變異訓練策略，少數音素類別的特徵空間被擴展，達成更平衡的特徵分佈，同時引入發音難度來調節特徵區域。

根據其出現次數將區域劃分為不同頻次區域。很明顯，正確發音的出現次數遠高於錯誤發音。更複雜的是，正確與錯誤發音各自呈現不同的長尾趨勢。例如，在正確發音中，元音 /EH/ 與 /EY/ 被歸類為中頻區域，但在錯誤發音中則屬於高頻區域。類似地，元音 /UW/ 在正確發音中位於高頻區域，卻在錯誤發音中移至低頻區域。

通常，針對 MDD 的音素層級發音分類器所採用的單純訓練流程，容易因正確發音出現頻率較高而產生不良偏差，導致其在整個訓練過程中佔據主導地位[22]。為了補救此一情形，我們提出的音素特定變異建模策略是基於以下假設：出現次數較高的音素類別（即多數音素類別）的 logits 可能佔據較大的特徵空間，而出現次數較低的音素類別（即少數音素類別）的 logits 則被壓縮到較狹窄的區域[23]，如圖 2(a) 所示。所提出的訓練策略透過對音素預測的 logits 加入隨機抽樣的高斯噪聲來增強，其中半徑由所提出的音素特定變異決定。為了解決 MDD 複雜的資料不平衡問題，音素特定變異的建模包含兩個互補因素：數量因素與發音難度因素。前者會對多數音素類別指定較小的變異數，對少數音素類別指定較大的變異數。 相較之下，後者則根據音素的誤讀率來調節特徵區域。如此一來，如圖 2(b) 所示，兩個因素的協同作用不僅平衡了不同音素的特徵分佈，還會調整與發音困難度對應的區域。總之，本文延續了我們於 [24] 與 [25] 中所述的先前工作
並在技術內容、實驗與分析上做出顯著擴展，其主要貢獻至少有四項：

我們提出 MuFFIN，一個多面向的發音回饋模型，透過互動式階層神經架構共同處理 MDD 與 APA 任務。此模型代表了一種範式轉變，從分開建模 APA 與 MDD 轉向統一評估方法，為 CAPT 開啟了新途徑。
提出了一種對比式語音序位正則化器，用以將從語音導出的音素表示對齊到對應的音素層級文字嵌入，並有機地結合發音準確度分數的序位性。透過序位性與音素特性之角度，進行了一系列圖形化檢視。
據我們所知，這是首度嘗試在 MDD 中透過將音素特定的變異納入訓練過程來處理資料不平衡問題。我們的方法指出，MDD 的資料不平衡問題源自兩個交織且同等重要的因素，即訓練資料的數量與發音難度。我們的實證結果顯示，僅考慮資料數量因素來解決 MDD 的資料不平衡問題，主要會提升召回率，但會犧牲精確率。基於此分析，我們提出了一種將發音難度因素納入的訓練策略，與單獨考慮數量因素或單獨考慮發音難度因素的策略相比，能在召回率與精確率之間達到更好的平衡。
在 Speechocean762 基準資料集 [26] 上進行的大量實驗驗證了我們所提出方法的有效性，該方法在 APA 與 MDD 任務上皆提升了現有最先進方法的表現。

II. 相關工作

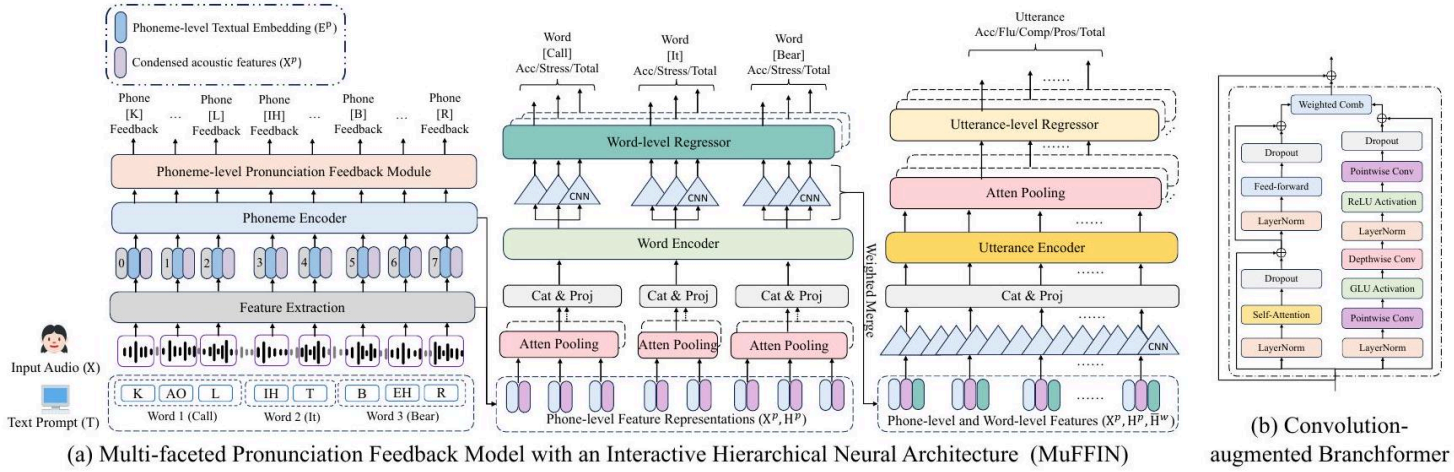
電腦輔助發音訓練（CAPT）是電腦輔助語言學習（CALL）的一個子領域，其研究與開發可追溯至 1960 年代的先驅性努力[27]，並因語音與語言技術的前所未有進展而在近期獲得重大關注[28][29][30]。根據 CAPT 的診斷性回饋，研究工作通常可分為音素層級的錯誤發音偵測與診斷（MDD）以及自動發音評量（APA），兩者多數在朗讀學習情境下發展。

A. 錯誤發音偵測與診斷

錯誤發音偵測與診斷（MDD）旨在於音素段落偵測發音錯誤，並進一步為 L2 學習者提供相應的診斷回饋[31][32]。常見的 MDD 方法可分為三類：基於發音評分、基於聽寫（dictation）、以及基於提示（prompt-based）的方法。基於發音評分的方法通常利用各種型態的信心度量來評估發音品質，透過
一個訓練良好的 ASR 系統（例如 hybrid DNN-HMM ASR 系統）。常用的度量包括但不限於音素持續時間 [33][34]、似然比 [13]、音素後驗機率 [35] 及其組合 [36]。給定一段輸入語音及其對應的典型音素序列（即音素層級的文字提示），基於發音評分的方法會先對典型音素序列中的每個音素計算發音分數。當某些音素的分數低於預先設定的閾值時，即被判定為發音錯誤的音素片段，表示其發音偏離預期。然而，基於發音評分的方法無法為被偵測出的錯誤發音片段提供診斷性回饋。為了解決此問題，基於聽寫的方法試圖將 MDD（錯誤發音偵測）定義為一個音素辨識任務，透過音素辨識器聽寫出 L2 學習者最有可能發出的音素序列。只要將聽寫結果與對應的典型音素序列比對，即可輕易識別出發音錯誤的部分。例如，Leung et al. 採用基於 CTC 的音素識別器來處理 L2 英語學習者的研究顯示，在錯誤發音檢測子任務中，其表現與基於發音評分的方法相當，而性能提升主要來自於對無聲音素區段錯誤發音的準確診斷 [37]。Yan 等人則將混合 CTCAttention ASR 模型作為聽寫模型，並透過 anti-phone 建模來捕捉帶有口音的 L2 學習者所產生的偏離（非範疇化）的音素發音 [38]。上述兩種方法皆依賴精確對齊以識別錯誤發音的片段；然而在實務應用中，將基準音素序列與 L2 學習者所發出的帶口音或不流暢語音比較時，可能會產生對齊錯誤。為此，基於 prompt 的方法利用注意力機制以端到端方式導出基準音素序列與學習者輸入語音之間的軟對齊，提供了一種有望減少對齊錯誤的可行途徑。 作為最早的嘗試之一，PeppaNet 透過 Transformer 解碼器將典型音素與學習者的語音對齊，任何差異都透過端到端神經建模在匹配度向量中捕捉到 [39]。此外，MDDGCN 為典型音素引入了基於圖的提示編碼器，旨在透過預先定義的語音圖正則化典型音素與實際發音音素之間的關係，以提高診斷準確度 [40]。

B. 自動發音評估

自動發音評估 (Automatic Pronunciation Assessment, APA) 透過提供特定發音面向的解析性分數（即連續數值）[41][42]或反映整體口語能力的整體評估（即離散類別值）[10]，來量化 L2 學習者在目標語言的發音能力。早期的 APA 研究主要集中在單一面向的評估，通常透過構建獨立的評分模組來預測能力分數，並於



(a) Multi-faceted Pronunciation Feedback Model with an Interactive Hierarchical Neural Architecture (MuFFIN)

(b) Convolution-augmented Branchformer

圖 3: Fig. 3. 所提出的多面向發音回饋模型，具有互動式分層神經架構 (MuFFIN)。 (a) 整體模型架構處理輸入語音與文本提示，分層表徵學習者的發音以生成各面向的評分。 (b) 所提出的卷積增強 Branchformer 模組作為 MuFFIN 的主幹，於不同語言層級的編碼器中運作。發音面向的準確性、流暢度、完整性與韻律分別以 Acc、Flu、Comp 與 Pros 表示。

特定語言層級，並使用各種手工設計的特徵集合。這些手工設計的特徵從學習者的輸入語音或其對應的 ASR 生成轉錄中提取，可能包括聲學特徵、被識別語言單位的置信分數、時間對齊資訊以及統計度量 [43][44]。為了全面檢視學習者的發音，近期 APA 的進展主張多面向與多粒度的發音評估，利用統一的評分模型在多個語言層級（即音素、詞與語句）針對不同面向（例如準確度、流暢度與完整性）來評估發音能力。基於此研究趨勢，Gong 等人提出了一種稱為 GOPT 的平行發音建模架構，其以 GOP 特徵為輸入並採用 Transformer encoder 作為骨幹模型，以聯合建模跨不同語言粒度的多重發音面向 [45]。遵循這一思路，3M 透過在模型的輸入 embedding 中加入韻律特徵與基於自監督學習（SSL）的特徵來擴展 GOPT，旨在實現多視角、多粒度與多面向的發音建模 [46]。儘管其表現不錯，但語句的層次結構在很大程度上被忽視。為了捕捉語句的語言層次，Do 等人提出了一個分層的 APA 模型，並探索了一種新穎的多特徵注意力層以強化評分面向之間的連結 [47]。Chao 等人引入了子音素建模，並採用深度可分離卷積層來構建分層的 APA 模型，有助於更好地在子詞層級建模局部語境線索 [48]。除了上述方法外，Gradformer 採用了粒度解耦的 Transformer 網路，先將語句的粒度分離為較低層（音素與詞層）與較高層（語句層），其中 Conformer 編碼器則聯合建模較低層的發音面向，而 Transformer 解碼器則處理一組可訓練的面向向量並與之互動

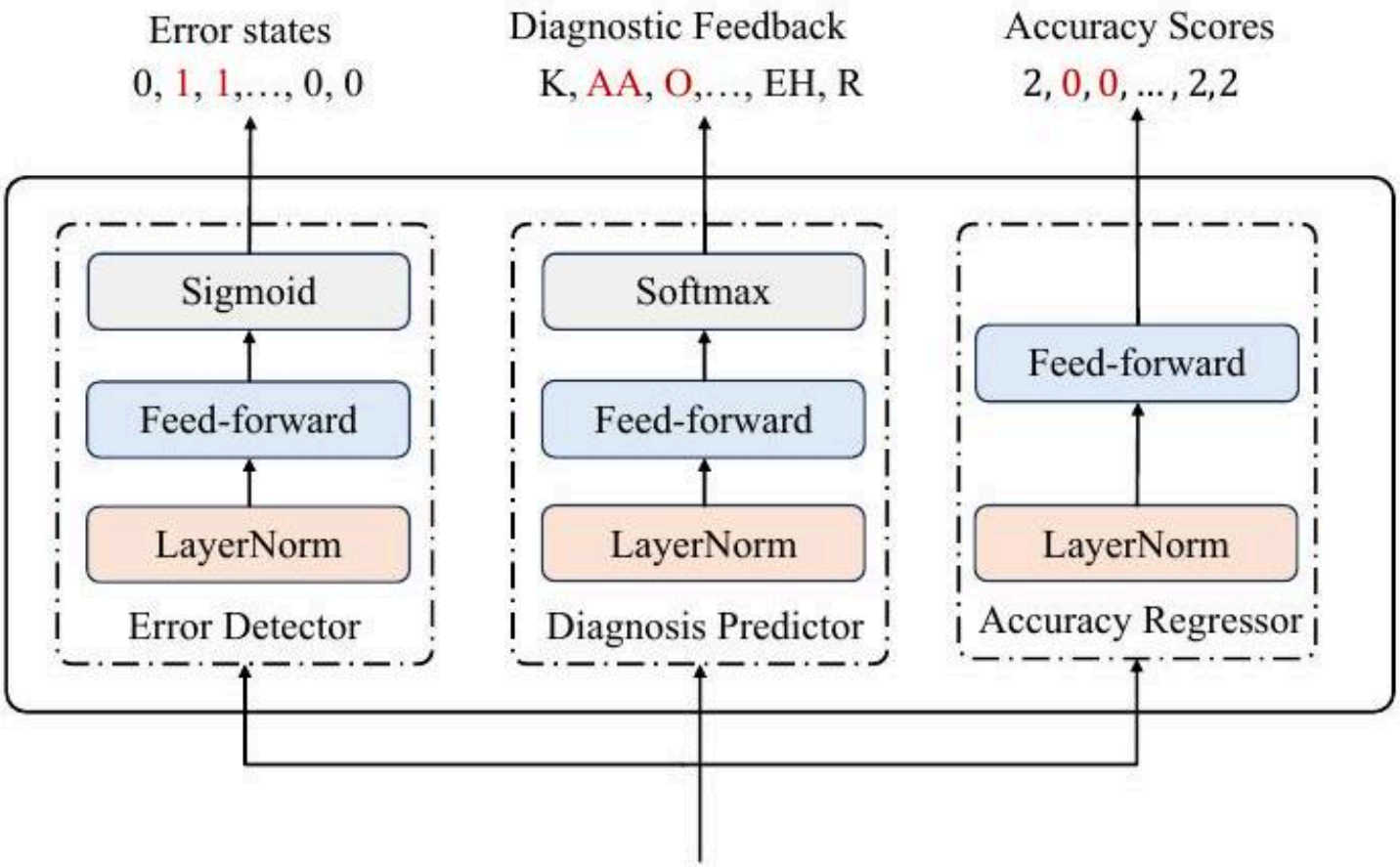


圖 4：圖 4。所提出的音素層級發音回饋模組，輸入為音素編碼器的輸出，並同時產生音素層級錯誤狀態、診斷音素及準確度分數。

與用於句子層級發音評估的編碼器輸出相結合 [42]。

III. 具交互式階層神經結構的多面向發音回饋模型

所提出的 MuFFIN 整體架構如圖 3(a) 示意，其中包含三個主要組件：音素層級建模、詞層級建模與句子層級建模。各語言層級的編碼器均採用新穎的 convolution-augmented Branchformer 區塊 [25]（如圖 3(b) 所示），該區塊由兩個分支組成，一個分支使用多頭注意力（MHA）層以捕捉超段位發音線索，另一個分支則透過一系列卷積層以擷取細粒度發音線索。此外，如圖 4 所示，提出了一個新的音素層級發音回饋模組，用以評估音素層級準確度並執行誤讀偵測與診斷。

A. 問題表述

給定一個輸入語句 U，由一個時間序列組成
給定由 L2 學習者產生的一段語音訊號 X，以及一個包含 M 個詞的參考文字提示 T，該提示基於發音詞典 [1] 被轉換為 N 個標準音素，所提出的多面向發音回饋模型旨在於不同語言粒度上估計能力分數，同時為標準音素序列定位音素層級的發音錯誤。形式上，令 $\mathrm{G} = \{p, w, u\}$ 為一組語言粒度，其中 p, w, u 分別代表音素、詞和語句層級。對於每個粒度 $g \in G$ ，我們的模型旨在預測一組面向分數序列 $\mathrm{A}^g = \left\{\mathbf{a}_1^g, \mathbf{a}_2^g, \ldots, \mathbf{a}_{N_g}^g\right\}$ ，其中 N_g 為粒度 g 上的發音面向數量。與此同時，對於標準音素序列 $\mathbf{q} = (q_1, q_2, \ldots, q_N)$ ，所提出的模型試圖檢測錯誤狀態序列 $\mathbf{e} = (e_1, e_2, \ldots, e_N)$ 並生成語音診斷序列 $\mathbf{y} = (y_1, y_2, \ldots, y_N)$ 。 e_n 與 y_n 均為針對 q_n 的音素層級發音回饋，其中 $e_n = 1$ 表示發音錯誤的音素片段， $e_n = 0$ 表示正確的音素片段，而 y_n 指示學習者實際產出的音素。

B. 互動式層級神經建模

音素層級建模。對於一個輸入語句，我們首先擷取各種發音特徵，以在音素層級描述 L2 學習者的發音品質，然後將這些特徵串接並進行投影以獲得一序列的濃縮聲學特徵 X^p 。特徵擷取過程表述如下：

$$\mathrm{X}^p = \mathrm{Linear}_p\left(\left[\mathrm{E}^{\mathrm{GOP}}; \mathrm{E}^{\mathrm{Dur}}; \mathrm{E}^{\mathrm{Eng}}; \mathrm{E}^{\mathrm{SSL}}\right]\right),$$

其中 $\mathrm{Linear}_p(\cdot)$ 為單層前饋層， $\mathrm{E}^{\mathrm{GOP}}$ 為基於發音良好度（GOP）的特徵，包括對數音素後驗（LPP）與對數後驗比（LPR）[12][14]。 $\mathrm{E}^{\mathrm{Dur}}$ 與 $\mathrm{E}^{\mathrm{Eng}}$ 為與時長與能量統計相關的韻律特徵[49][50]，而 $\mathrm{E}^{\mathrm{SSL}}$ 為基於自監督學習（SSL）的特

徵[46]。接著我們將音素層級的文字嵌入 E^p 加到 X^p ，並經過音素編碼器以取得面向表示 $\mathrm{H}^p = \left(\mathbf{h}_1^p, \mathbf{h}_2^p, \ldots, \mathbf{h}_N^p\right)$ ：

$$\begin{aligned} \mathrm{H}_0^p &= \mathrm{X}^p + \mathrm{E}^p, \\ \mathrm{H}^p &= \mathrm{PhnEnc}\left(\mathrm{H}_0^p\right). \end{aligned}$$

此處， E^p 是將 $\mathbf{q}$ 輸入至一個以音素為層級的提示編碼器後生成的，該編碼器包含音素與位置嵌入層。PhnEnc(.) 由堆疊 3 個以卷積增強的 Branchformer 區塊組成。

接著，發音反饋模組在 H^p 的基礎上構建，用以估計多面向的發音反饋，包含三個組成部分：錯誤偵測器、診斷預測器，以及準確度分數回歸器。錯誤偵測器是一個二元標註模型，預測錯誤狀態 $\hat{e}_n$ ，用以表示是否將 $\mathbf{q}$ 的第 n 個音素判定為誤讀：

$$\mathrm{P}_{\mathrm{det}}\left(\hat{e}_n \mid \mathbf{q}, \mathrm{X}\right) = \mathrm{Sigmoid}\left(\mathrm{Linear}_{\mathrm{det}}\left(\mathbf{h}_n^p\right)\right),$$

其中 $\mathrm{Linear}_{\mathrm{det}}(\cdot)$ 是接著層正規化（layer normalization）的一個線性層。診斷預測器執行序列式的多類別標註流程，以推導第 n 個標準音素的診斷反饋機率分布為：

$$\mathrm{P}_{\mathrm{diag}}\left(\hat{y}_n \mid \mathbf{q}, X\right) = \mathrm{Softmax}\left(\mathrm{Linear}_{\mathrm{diag}}\left(\mathbf{h}_N^p\right)\right),$$

其中 $\mathrm{Linear}_{\mathrm{diag}}(\cdot)$ 用於將隱藏維度轉換為發音字典的大小。最後，音素層級的準確度分數由準確度分數回歸器估計。
字詞層級建模。針對字詞層級的評估，引入了一個字詞層級的注意力池化，用以從其構成的音素產生字詞表示向量，此池化以一個 1-D depth-wise convolution 層為實例，接著是一個 MHA 層與一個平均運算。字詞層級的輸入表示 X^w 是透過分別將 X^p 與 H^p 輸入字詞層級注意力池化，然後以線性投影打包在一起所計算得出：

$$\begin{aligned} \widehat{\mathrm{X}}^w, \widehat{\mathrm{H}}^w &= \mathrm{AttPool}_{w_1}\left(\mathrm{X}^p\right), \mathrm{AttPool}_{w_2}\left(\mathrm{H}^p\right), \\ \mathrm{X}^w &= \mathrm{Linear}_w\left(\left[\widehat{\mathrm{X}}^w; \widehat{\mathrm{H}}^w\right]\right). \end{aligned}$$

接著，字詞層級的文字嵌入 E^w 被加到 X^w ，並使用一個字詞編碼器來產生字詞層級的情境化表示 H^w ：

$$\begin{aligned} \mathrm{H}_0^w &= \mathrm{X}^w + \mathrm{E}^w, \\ \mathrm{H}^w &= \mathrm{WordEnc}\left(\mathrm{H}_0^w\right), \end{aligned}$$

其中 E^w 是透過字詞與位置嵌入層將文字提示 T 映射為對應的嵌入序列所得到，而 WordEnc(.) 由 2 個結合捲積的 Branchformer 區塊組成。最後，對 H^w 分別施行三個不同的 1-D depth-wise convolution 層以產生字詞層級的面向表示（即 $\mathrm{H}^{w_1}, \mathrm{H}^{w_2}$ 與 H^{w_3} ），然後再由對應的字詞層級回歸器將其轉換為發音分數序列。
話語層級建模。對於話語層級的評估，我們首先透過對逐幀 SSL 基礎特徵在時間維度上進行平均池化，提取出話語層級的 SSL 基礎特徵 $\overline{\mathrm{E}}^{SSL}$ 。接著，我們以加權組合合併 $\mathrm{H}^{w_1}, \mathrm{H}^{w_2}$ 與 H^{w_3} ，以獲得詞級表示 $\overline{\mathrm{H}}^w$ 。一系列話語層級輸入表示 H_0^u 是先對 $\mathrm{X}^p, \mathrm{H}^p$ 與 H^w 應用一維深度可分離卷積層，然後進行串接與線性投影所得到。因此，利用一個話語編碼器來產生具上下文資訊的表示 H^u ：

$$\begin{aligned} \overline{\mathrm{H}}^w &= \mathrm{Merge}\left(\mathrm{H}^{w_1}, \mathrm{H}^{w_2}, \mathrm{H}^{w_3}\right), \\ \mathrm{H}_0^u &= \mathrm{Linear}_u\left(\left[\mathrm{DC}_1\left(\mathrm{X}^p\right); \mathrm{DC}_2\left(\mathrm{H}^p\right); \mathrm{DC}_3\left(\overline{\mathrm{H}}^w\right)\right]\right), \\ \mathrm{H}^u &= \mathrm{UttEnc}\left(\mathrm{H}_0^u\right), \end{aligned}$$

其中 $\mathrm{Merge}(\cdot)$ 為加權平均操作 [51]， $\mathrm{UttEnc}(\cdot)$ 為單一經卷積增強的 Branchformer 區塊， $\mathrm{DC}_1(\cdot), \mathrm{DC}_2(\cdot), \mathrm{DC}_3(\cdot)$ 為各自不同的一維深度可分離

卷積層，且每層的核大小為 3。之後，在 H^u 之上構建五個獨立的注意力池化模組以產生話語層級的面向表示向量。這些特徵接著透過殘差連接與 $\overline{\mathrm{E}}^{\mathrm{SSL}}$ 結合，並透過各自的回歸器轉換為話語層級的面向分數。
訓練目標。所提出模型的訓練目標由 APA 與 MDD 的損失計算得出：
$\mathcal{L}_{MuFFIN} = \mathcal{L}_{APA} + \mathcal{L}_{MDD}$

APA 損失是從不同粒度層級收集而來的均方誤差（MSE）損失的加權總和：

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

其中 $\mathcal{L}_{p^{jp}}, \mathcal{L}_{w^{jw}}$ 、 $\mathcal{L}_{u^{ju}}$ 分別為音素層級、詞彙層級和語句層級針對不同面向的損失，而 N_p, N_w 、 N_u 表示每個粒度層級上面向的數量。另一方面，MDD 的訓練目標包含誤讀偵測 $\mathcal{L}_{\text{det}}$ 與診斷 $\mathcal{L}_{\text{diag}}$ 任務：

$$\mathcal{L}_{MDD} = \mathcal{L}_{det} + \mathcal{L}_{diag},$$

$$\mathcal{L}_{det} = -\sum_{n=1}^{N} \log \mathrm{P}_{det}\left(\hat{e}_n = e_n \mid \mathbf{q}, \mathrm{X}\right),$$

$$\mathcal{L}_{diag} = -\sum_{n=1}^{N} \log \mathrm{P}_{diag}\left(\hat{y}_n = y_n \mid \mathbf{q}, \mathrm{X}\right),$$

其中 $\mathcal{L}_{\text{det}}$ 與 $\mathcal{L}_{\text{diag}}$ 分別表示用於訓練偵測器與預測器的負對數概似（negative log-likelihood）。

IV. 對比音素序列正則化器

為了為多面向的發音評估模型生成更能區分音素的特徵，我們提出了對比音素序關正則器（ConPCO），其由三個數學項組成：對比項 $\mathcal{L}_{\text{con}}$ 、音素特徵項 $\mathcal{L}_{pc}$ 與序關項 $\mathcal{L}_{o}$ 。 $\mathcal{L}_{\text{con}}$ 旨在同時將由發音評估模型產生的音素表示與音素層級文字提示的嵌入投影到共同的特徵空間。 $\mathcal{L}_{pc}$ 與 $\mathcal{L}_{o}$ 調整音素間及音素內類別之間的距離，其中前者增強音素間差異，後者透過序關係提升音素內的緊密性。所提出的 ConPCO 正則器表述如下：

$$\mathcal{L}_{ConPCO} = \mathcal{L}_{con} + \mathcal{L}_{pc} + \mathcal{L}_{o}$$

對比項。令 $\mathrm{H}^p = \left(\mathbf{h}_1^p, \mathbf{h}_2^p, \ldots, \mathbf{h}_N^p\right)$ 表示由發音評分模型中的音素編碼器對一段語音產生的音素表示序列， $\mathrm{E}^p = \left(\mathbf{e}_1^p, \mathbf{e}_2^p, \ldots, \mathbf{e}_N^p\right)$ 表示由音素層級提示編碼器生成的典型音素的文字嵌入。接著，先對 H^p 與 E^p 分別應用線性投影，然後為每個音素類別計算質心向量，由此得到一組配對的音素表示 $\mathcal{M} = \left\{\left(\mathbf{z}_i^p, \mathbf{z}_i^t\right), i = 1, \ldots, M\right\}$ 。如圖 5 所示，接著從 $\mathcal{M}$ 推導出 $M \times M$ 相似度，並且...

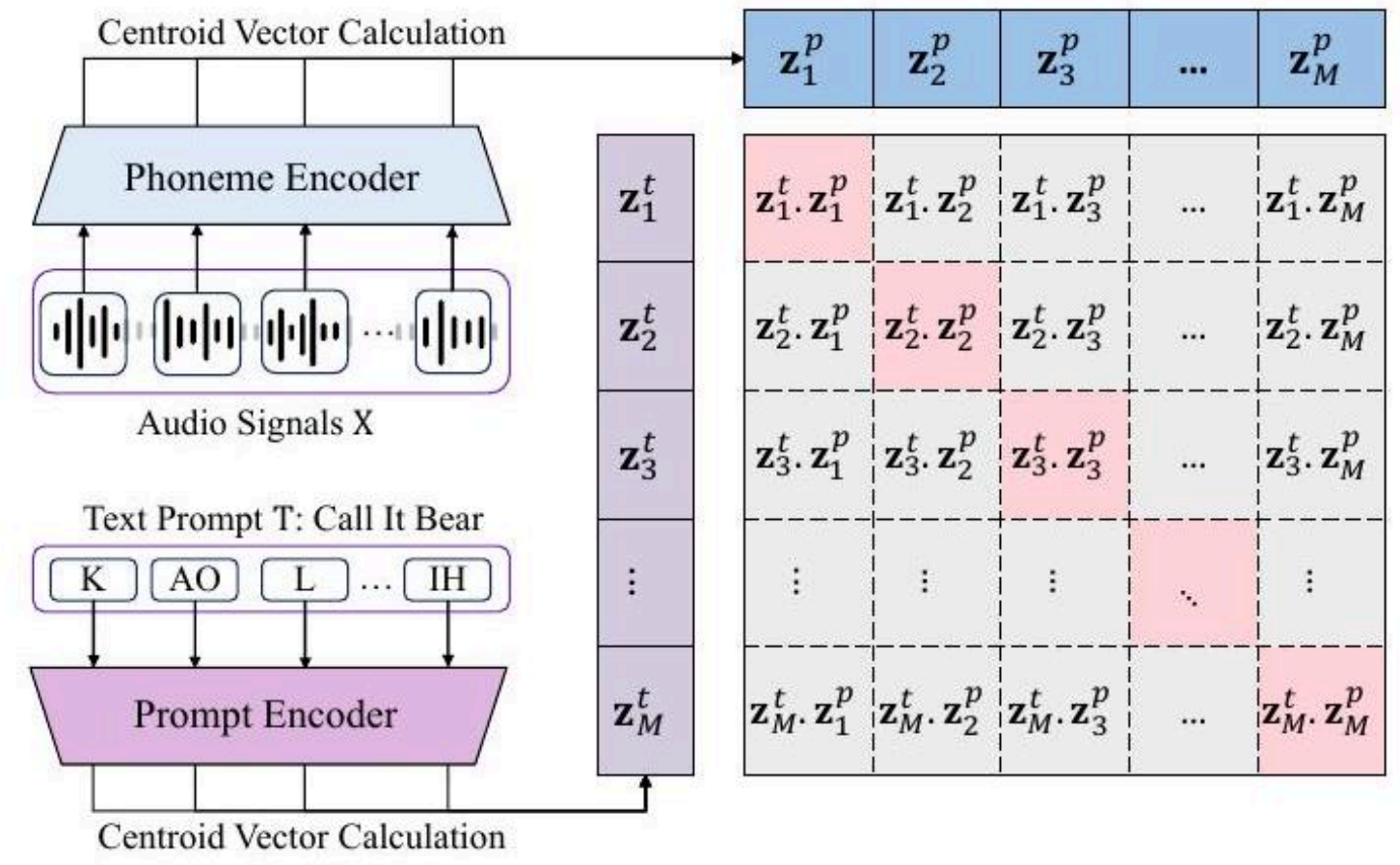


圖 5：圖 5。對比項 $\mathcal{L}_{\text{con}}$ 計算過程的視覺化。

對比項 $\mathcal{L}_{\text{con}}$ 旨在最大化配對音素表示之間的相似度，同時最小化非配對表示之間的相似度 [52][53]。對比項 $\mathcal{L}_{\text{con}}$ 包含兩個損失，並有一個溫度超參數 τ 用以控制對負樣本處罰的強度：

$$\mathcal{L}_{\text{con}} = \mathcal{L}_{p2t} + \mathcal{L}_{t2p}$$

$$\mathcal{L}_{p2t} = -\frac{1}{M}\sum_{i=1}^{M}\log\frac{\exp\left(\phi\left(\mathbf{z}_i^p, \mathbf{z}_i^t\right)/\tau\right)}{\sum_{j=1}^{M}\exp\left(\phi\left(\mathbf{z}_i^p, \mathbf{z}_j^t\right)/\tau\right)}$$

$$\mathcal{L}_{t2p} = -\frac{1}{M}\sum_{i=1}^{M}\log\frac{\exp\left(\phi\left(\mathbf{z}_i^t, \mathbf{z}_i^p\right)/\tau\right)}{\sum_{j=1}^{M}\exp\left(\phi\left(\mathbf{z}_i^t, \mathbf{z}_j^p\right)/\tau\right)}$$

其中 $\phi\left(\mathbf{z}_i^p, \mathbf{z}_j^t\right)$ 是 ℓ_2 經標準化的向量 $\mathbf{z}_i^p$ 與 $\mathbf{z}_j^t$ 之間的點積（餘弦相似度）。在訓練期間， $\mathcal{M}$ 由每個 batch 建構，其中我們經驗性地從資料中抽取具有最高熟練度分數的實例來計算質心向量。

語音音素特徵項。語音音素特徵項 $\mathcal{L}_{pc}$ 透過最小化質心向量 $\mathbf{z}_i^p$ 之間的負距離來保留音素相似性資訊：

$$\mathcal{L}_{pc} = -\frac{1}{M(M-1)}\sum_{i=1}^{M}\sum_{i\neq j}\left\|\mathbf{z}_i^p - \mathbf{z}_j^p\right\|_2$$

其中 $\mathcal{L}_{pc}$ 等同於在最佳化過程中最大化音素類別之間的距離。

序數項。為在特徵空間反映回歸目標的序數關係，定義序數項 $\mathcal{L}_o$ 以最小化特徵表示 $\mathbf{h}_i^p$ 與其相對應、帶有能力分數相對差異的音素質心向量 $\mathbf{z}_i^p$ 之間的距離：

$$\mathcal{L}_o = \frac{1}{N}\sum_{i=1}^{N} w_i \|\mathbf{h}_i^p - \mathbf{z}_i^p\|_2,$$

其中 $w_i = |C - y_i^p|$ 為每個 $\mathbf{h}_i^p$ 的緊湊性權重，用以反映標籤空間中的序數行為，且 y_i^p 表示對應的音素層級正確率分數。可調常數 C 設為 3，代表最高正確率分數加上一小幅餘量。

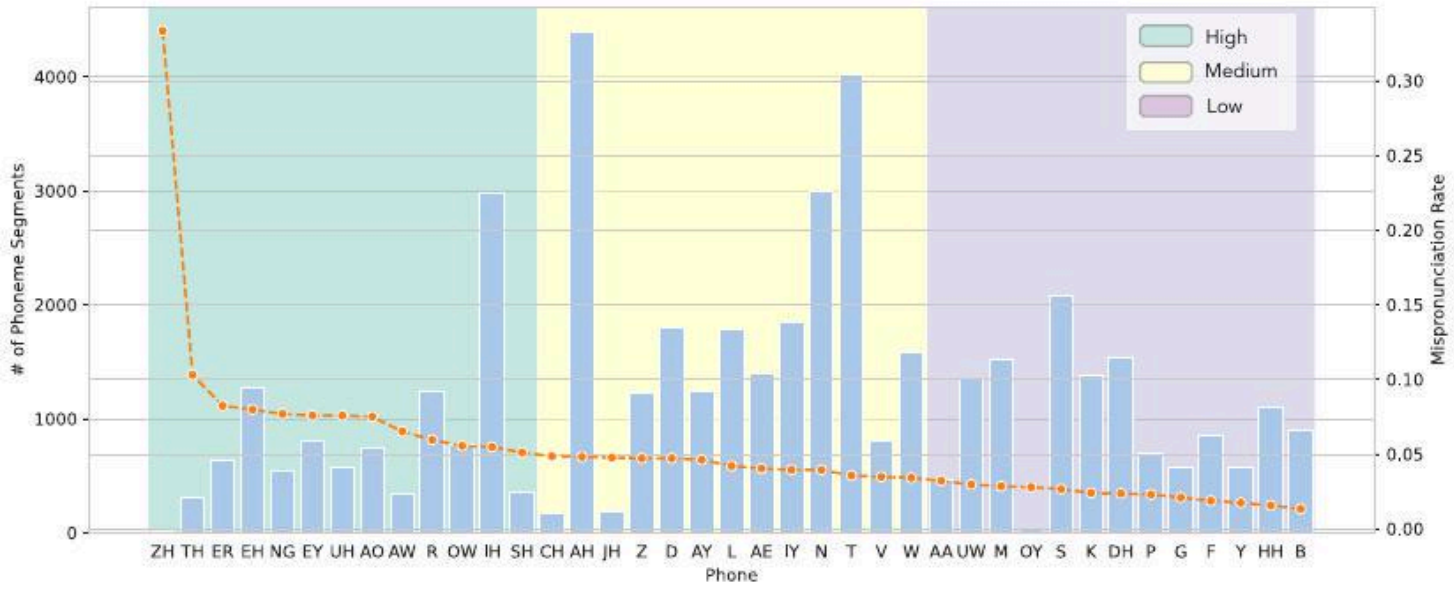


圖 6：Fig. 6. Speechocean762 的音素統計資訊，包括出現次數與對應的誤讀率。

IV. 音素特定變異

為了在兼顧發音困難度的同時平衡預測音素的分佈，由音素預測器產生的音素預測 logits 會加入隨機採樣的高斯噪音，噪音半徑由音素相關的變異量決定。為此，所提出的訓練方案「音素特定變異（PhnVar）」由兩個因子組成：資料量因子與發音困難度因子。資料量因子會對多數音素類別指派較小的變異，對少數類別則指派較大的變異；而發音困難度因子則根據誤讀率調節特徵區域。形式上，我們重訪式 (5)，並將第 n 個典型音素被預測為診斷音素 k 的機率表達出來，該機率源自 softmax 函數：

$$p_k^n = \frac{\exp\left(g_k^n\right)}{\sum_{i=1}^{M}\exp\left(g_i^n\right)}.$$

這裡， g_k^n 是由 Linear $^{\text{diag}}\left(\mathbf{h}_N^p\right)$ 所產生的 logit 向量 $\mathbf{g}^n = \left(g_1^n, g_2^n, \ldots, g_M^n\right)$ 中第 k 個音素的 logit，其中 M 是音素類別的數量。我們接著以音素特定的變異性來擴增 logits，該變異性定義為資料量因子 QF_k 與音素發音難度因子 DF_k 的加權 p ，對於音素 k 的權重係數分別為 α 與 β ：

$$\hat{g}_k^n = g_k^n + \delta(\sigma) \times \exp\left(\frac{\alpha \times \log\left(\mathrm{QF}_k\right) + \beta \times \log\left(\mathrm{DF}_k\right)}{\alpha + \beta}\right),$$

其中 $\delta(\sigma)$ 代表均值為零、標準差為 σ 的高斯分布。於我們的實驗中， α 與 β 均設為 1。資料量因子定義為在對數尺度上操作後正規化的反向音素頻率：

$$\mathrm{QF}_k = \frac{c_k}{\max_i c_i}; \quad c_k = \log \frac{\sum_{i=1}^{M} q_i}{q_k},$$

其中 q_k 為音素類別 k 中的實例數。發音難度因子則表示為正規化的誤讀率：

$$\mathrm{DF}_k = \frac{d_k}{\max_i d_i}; \quad d_k = \frac{mp_k}{mp_k + cp_k},$$

其中 mp_k 與 cp_k 分別為音素類別 k 的誤讀與正確發音實例數。

表 1：表 I
Speechocean762 中 APA 任務的統計資料

自動發音評估				
粒度	面向	分數區間	計數	
			訓練	測試
音素	準確率	[0, 2]	47,076	47,369
	準確率	[0, 10]	15,849	15,967
字詞	重音			
語句	準確度 完整性	[0,10]	2,500	2,500
	流暢度			
	韻律			
	總分			

表 2：表 II
Speechocean762 中 MDD 任務的統計資料

發音錯誤偵測與診斷			
類型	描述	計數	
		訓練	測試
正確性	所發出的音素與典型音素對齊	45,088	45,959
刪除	省略了一個典型音素	450	396
替換	一個典型音素被誤發為其他音素	914	593
非類別性錯誤	所發音素不存在於 CMU 發音字典中	488	332
口音錯誤	一個標準音位發音正確，但帶有明顯口音	136	89

IV. 實驗設置

A. 實驗資料與評估指標

資料集。實驗系列在 Speechocean762 資料集上進行，該資料集為公開可得，專為電腦輔助語言學習 (CALL) 研究設計 [26]。此資料集包含 250 名以中文為母語的 L2 學習者所發的 5,000 段英語錄音。訓練集與測試集大小相等，各自包含 2,500 個語句。對於 APA 任務，發音能力分數在多種語言粒度上針對不同的發音面向進行評估，APA 任務的統計資料彙總如表 I。對於 MDD 任務，音位標籤遵循 CMU 發音詞典的定義，該詞典包含一組 39 個標準音位。在 Speechocean762 中，對於準確度分數低於 0.5 的音位區段，人工標註為誤讀標籤，並被歸類為四種類型：刪去 (deletion)、替換 (substitution)、非類別錯誤 (non-categorical error) 與口音錯誤 (accented error)。表 II 彙總了 MDD 任務的音位區段統計資料。
評估指標。APA 的主要評估指標採用 Pearson 相關係數 (PCC)，用以衡量預測分數與實際標準答案之間的線性相關性

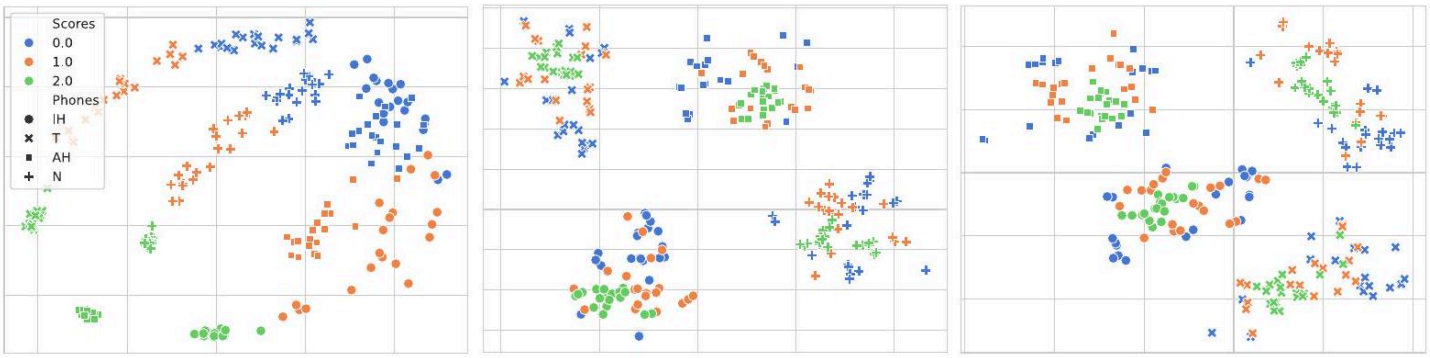


圖 7：圖 7. 從 MuFFIN（以音素特徵與序數項，即 $\mathcal{L}_{pc}$ 和 $\mathcal{L}_{o}$ 訓練）擷取的音素表示視覺化。每個資料點的顏色表示音素層級的正確率分數，形狀則表示對應的音素類別。我們展示了 (a) 原始 MuFFIN 模型的音素表示 H^{p} 、(b) $\mathrm{MuFFIN} + \mathcal{L}_{pc}$ ，以及 (c) $\mathrm{MuFFIN} + \mathcal{L}_{pc} + \mathcal{L}_{o}$ 。

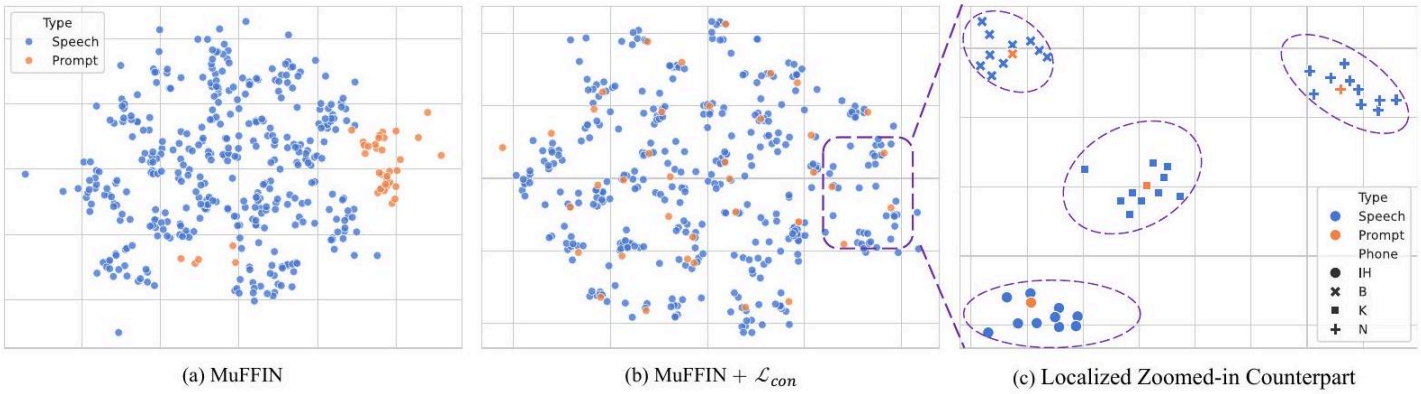


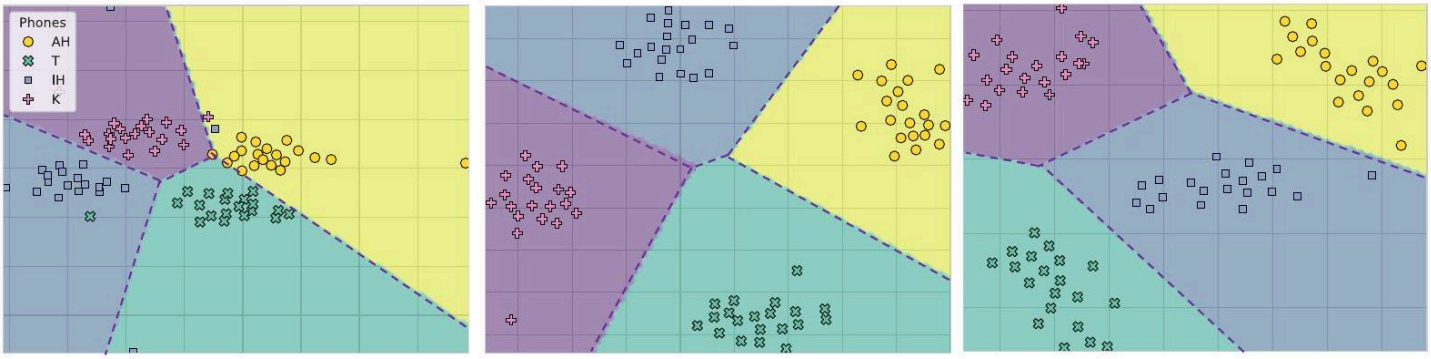
圖 8：圖 8. 音素表示視覺化，藍色與橘色點分別代表由 MuFFIN 的音素編碼器與音素層級提示編碼器所產生的特徵表示 H^{p} 與 E^{p} 。特徵表示分別顯示為 (a) 原始 MuFFIN 模型、(b) $\mathrm{MuFFIN} + \mathcal{L}_{\mathrm{con}}$ ，以及 (c) 一個局部放大的對應圖。

分數。依照先前研究，音素層級的準確度以平均平方誤差（MSE）報告。另一方面，對於 MDD 任務，評估指標遵循文獻[9]中的評分規範。具體來說，錯誤發音偵測子任務使用召回率（RE）、精確率（PR）與 F1 分數（F1）進行評估，而錯誤發音診斷子任務則以診斷錯誤率（DER）、誤拒率（FRR）、誤接納率（FAR）和音素錯誤率（PER）來評估。

B. 實作細節

特徵擷取。對於發音特徵擷取，採用與我們先前研究相同的 GOP 特徵、能量與持續時間統計量 [24][25]。基於 SSL 的特徵擷取遵循文獻 [46] 建議的處理流程，特徵從預訓練語音模型的輸出中擷取，包括 Wav2vec2.0 [54]、WavLM [55] 與 HuBERT [56]。基於 SSL 的特徵與能量特徵均在幀等級擷取，然後根據將學習者語音與參考文本進行強制對齊所得到的音素區段時間戳，彙整為音素等級表示。擷取出的音素等級發音能力特徵共計 3,164 維，包含 84 維的 GOP 特徵 $\mathrm{E}^{\mathrm{GOP}}$ 、7 維的能量統計 $\mathrm{E}^{\mathrm{Eng}}$ 、1 維的持續時間數值 $\mathrm{E}^{\mathrm{Dur}}$ ，以及 3,072 維的基於 SSL 的特徵 $\mathrm{E}^{\mathrm{SSL}}$ 。

訓練設定。關於訓練設定，我們遵循文獻 [24][25] 中報導的設定，每個實驗由 5 次獨立試驗組成，每次試驗以不同的隨機種子執行 100 個 epoch。在每次試驗中，模型使用初始學習率為 $1\mathrm{e}-3$ 的 Adam 優化器訓練，批次大小為 25。使用學習率調度器，在整體損失連續 10 個 epoch 未下降時將學習率衰減 0.1 倍。此外，我們的模型依照文獻 [41] 所述的預訓練策略以預訓練模型進行初始化。報告的實驗結果為 5 次試驗的平均值，評估基準為最小音素層級 MSE。
模型設定。音素層級、詞層級與語句層級的編碼器（即 PhnEnc(•)、WordEnc(•)、 $\mathrm{UttEnc}(\cdot)$ ）分別由 3、2 與 1 個經卷積增強的 Branchformer 區塊組成 [25]。在每個編碼器區塊內，自注意力分支以單頭注意力層實作，之後接兩層前饋層。自注意力與前饋層的隱藏維度均為 24。同時，卷積分支由一個核為 1×3 的深度可分離卷積層與一個核為 1×1 的逐點卷積層組成，兩者均為 24 通道。為了聚合詞層級與語句層級的特徵，注意力池化模組由一個深度可分離卷積層與一個單一-



(a) MuFFIN (b) MuFFIN + PhnVar w/o DF (c) MuFFIN + PhnVar

圖 9：圖 9. MuFFIN 模型在提出的音素特異性變異（PhnVar）方案下，診斷預測器之音素層級 logits 與決策邊界視覺化。我們展示診斷預測器的音素 logits，分別為 (a) 原始 MuFFIN 模型、(b) MuFFIN + PhnVar w/o DF（即未考量發音困難度因子的 PhnVar 變體）、以及 (c) MuFFIN + PhnVar。

head 自注意力層，其中捲積層具有 24 個通道，核大小為 1×3 ，注意力層的隱藏維度為 24。此外，我們將投影層（即 $\mathrm{Linear}_p(\cdot), \mathrm{Linear}_w(\cdot)$ 與 $\mathrm{Linear}_u(\cdot)$ ）的隱藏維度設為 24。在訓練階段，式 (14) 中可調參數 $\mathcal{L}_p, \mathcal{L}_w$ 與 $\mathcal{L}_u$ 設為 3、1 與 1（分別），而式 (20) 和 (21) 中的溫度因子 τ 設為 1。

透過 MuFFIN 的誤讀偵測與診斷。為了偵測誤讀片段，MuFFIN 採用基於發音打分的範式，將音素層級錯誤偵測器的輸出作為誤讀片段的指標。當對應的指標超過預先定義的閾值時，該音素片段即被識別為誤讀。隨後，偵測到的誤讀片段會被送入音素層級預測器以產生診斷結果。為確保偵測器與預測器之間的一致性，我們在預測器的 softmax 計算過程中遮蔽正規音素（即文字提示的音標轉寫）。

C. 比較方法

我們將所提出的模型（MuFFIN）與三類發音評估模型進行比較。1) 單面向發音評估：Lin2021 是一個分層的 APA 模型，採用音素層級的表面特徵作為輸入，並在語句層級評估準確度分數 [57]。Kim2022 則倚賴從預訓練聲學模型提取的分層語境表示，以在語句層級衡量流暢度或韻律等口說能力 [30]。2) 多面向與多粒度發音評估：對於具有平行神經結構的評估模型，GOPT 和 LSTM 是顯著的模型，兩者都消耗一序列的 GOP 特徵，並同時在音素、單字與語句層級產生一組能力分數 [45]。3M 在 GOPT 的輸入特徵中增強了基於 SSL 的特徵以捕捉超節段的發音線索，同時整合語音學特徵以強化音素層級的文字資訊 [46]。至於分層模型，HierGAT 設計了一個具語言層級感知的模型，採用一系列圖注意力神經網路，進一步強化各面向之間的關聯性。
與面向注意力機制 [58]。3MH 作為先前 APA 的最先進方法，採用 3M 作為骨幹模型，並引入超音素（sup-phoneme）建模，以在音素與詞彙層級之間的語言階層中捕捉更細緻的發音特徵 [48]。Gradformer (GFR) 將語言階層解耦為兩個子層級，分別為較低層（音素與詞）與較高層（話語）。Conformer 編碼器對較低語言層級的面向進行建模，而 Transformer 解碼器處理一序列可學習的面向向量，並與編碼器輸出互動以評估話語層級的面向 [42]。3) 多面向發音評估：Ryu2023 提出一個統一的模型架構，透過在預訓練聲學模型之上獨立堆疊基於 CTC 的音素辨識器與一組回歸器，將音素識別與發音評估共同優化 [59]。JAM 在 3M 基礎上進一步整合了音素分類器，根據輸入的典型音素預測診斷音素，並透過利用電磁發音描記（EMA）特徵以捕捉第二語言學習者的發音動作，進一步提升 MDD 的效能 [60]。

V. 實驗結果

A. 質性分析

在實驗開始時，我們首先分析 Speechocean762 的音素統計，揭示每個音素的資料量與發音難度之間的複雜關係。接著，我們透過一系列質性視覺化檢視所提出之音素層級正則化項的有效性。

Speechocean762 的音素統計。在圖 6 中，報告了 speechocean762 資料集中音素片段的出現次數（藍色長條）及其各自的誤讀率（橘色點），其中音素依誤讀率排序，並根據誤讀率分為三個互不重疊的子集：高（誤讀率高於 5.1%）、中（誤讀率介於 5.1% 與 3.4% 之間）以及低（誤讀率低於 3.4% ）區域。

在圖 6 中，可以明顯看出音素的出現次數與其相應的發音錯誤率呈現不同的分佈模式。例如，高出現次數的

表 3：表 III
MuFFIN 與比較方法在 SpeechOcean762 的 APA 任務上的效能評估

Model	音素層級 準確率		詞彙層面			語句層面				
	MSE↓	PCC ↑	正確率 ↑	重音 ↑	總計 ↑	正確率 ↑	組成 ↑	流暢度 ↑	韻律 ↑	總分 ↑
Lin2021 [57]	-	-	-	-	-	-	-	-	-	0.720
Kim2022 [30]	-	-	-	-	-	-	-	0.780	0.770	-
LSTM [45]	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 [45]	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)
3M [45]	0.078	0.656	0.598	0.289	0.617	0.760	0.325	0.828	0.827	0.796
	(0.001)	(0.005)	(0.005)	(0.033)	(0.005)	(0.004)	(0.141)	(0.006)	(0.008)	(0.004)
GFR [42]	0.079	0.646	0.598	0.334	0.614	0.732	0.318	0.769	0.767	0.756
	(0.001)	(0.004)	(0.006)	(0.013)	(0.006)	(0.005)	(0.139)	(0.006)	(0.004)	(0.003)
HierGAT [58]	0.073	0.683	0.648	0.327	0.663	0.798	0.531	0.840	0.833	0.821
	(0.001)	(0.004)	(0.003)	(0.011)	(0.002)	(0.002)	(0.047)	(0.002)	(0.002)	(0.002)
3MH [48]	0.071	0.693	0.682	0.361	0.694	0.782	0.374	0.843	0.836	0.811
	(0.001)	(0.004)	(0.005)	(0.098)	(0.007)	(0.003)	(0.115)	(0.003)	(0.004)	(0.004)
Ryu2023 [59]	-	-	-	-	-	0.719	-	0.775	0.773	0.743
JAM [60]	0.076	0.664	0.622	0.241	0.638	0.773	0.205	0.831	0.829	0.805
	(0.002)	(0.001)	(0.012)	(0.034)	(0.005)	(0.007)	(0.080)	(0.004)	(0.004)	(0.004)
MuFFIN	0.063	0.742	0.705	0.315	0.714	0.807	0.768	0.841	0.832	0.830
	(0.002)	(0.006)	(0.004)	(0.033)	(0.004)	(0.003)	(0.049)	(0.004)	(0.004)	(0.002)

報告的結果包含在 5 次獨立實驗試驗上計算的平均 PCC 分數及標準差。Acc. 與 Comp. 分別指發音面向中的準確性與完整性。所提出的 MuFFIN 在除話語流暢度之外的所有指標上，相較於 3 MH，均達到較高的 PCC 分數，並進行近似隨機化檢定（$p < 0.001$）。

音素 (例如 $/\mathrm{AH}/, /\mathrm{T}/$ 和 $/\mathrm{N}/$) 被發現在中等錯誤發音率區域。相反地，一些低出現率的音素（例如 /ZH/、/TH/ 和 /NG/）常常與高錯誤發音率相關。基於此，為了減緩 MDD 任務所面臨的資料不平衡問題，所提出的音素特定變異包含兩個新穎的調節項：數量因子與發音難度因子。前者平衡音素的特徵分佈，後者則依據錯誤發音率調整特徵的分散程度。

音素表示的質性視覺化—音素特性與序列項。在第二組實驗中，我們基於所提出的 APA 模型圖形化檢視音素特性項與序列項（即 $\mathcal{L}_{pc}$ 與 $\mathcal{L}_{o}$ ）的影響。如圖 7 所示，我們從測試集擷取音素表示 H^{p} ，並將每個資料點依其音素類別（以形狀表示）及相對的發音準確度分數（以顏色表示）進行視覺化。

從圖 7(a)可觀察到，儘管 MuFFIN 共同優化了音素辨識與評量任務，但產生的音素表示仍不可避免地依據音素層級的正確率分組，無法在特徵空間中明確捕捉音素間的細微差異。當以 $\mathcal{L}_{pc}$ 訓練 MuFFIN，如圖 7(b) 所示，則可獲得具音素判別力的特徵，表示會依各自的音素類別分散。然而，單純分離特徵表示會遺漏序關係，可能阻礙發音

表 4：Table IV

MuFFIN 在音素特定變異訓練方案下，加入音素層級正則化器的效能評估

MuFFIN			音素分數	詞級分數		
PhnVar	$\mathcal{L}_{\mathrm{con}}$	$\mathcal{L}_{\mathrm{pco}}$	正確率	準確率	重音	總計
-	-	-	0.742	0.705	0.315	0.714
V	-	-	0.746	0.704	0.310	0.714
V	V	-	**0.749**	0.707	0.314	**0.718**
V	-	V	0.745	0.703	0.296	0.713
V	V	V	0.747	**0.708**	**0.341**	**0.718**

Acc. 指發音面的準確性。

評量任務。為回應此一情形，$\mathcal{L}_{pc}$ 與 $\mathcal{L}_{o}$ 的協同作用提供了補救，使音素表示能同時反映源自其正確率的類別差異與序關係，如圖 7© 所示。具體而言，整合 $\mathcal{L}_{o}$ 會在每個音素類別內導致成對距離與音素層級正確率之間出現更強的相關性，當正確率降低時在特徵空間中產生向外擴散的現象。根據這些觀察，在 MuFFIN 的訓練過程中納入 $\mathcal{L}_{pc}$ 與 $\mathcal{L}_{o}$ 能夠大幅提升音素表示的可區辨性，並同時在特徵空間中反映出預測正確率的序關係。

針對所提出對比項的音素表示之定性視覺化。接著，為了以定性方式評估對比項 $\mathcal{L}_{\mathrm{con}}$ 是否將由語音導出的表示（以藍色標示）與其

對應於音素區段的文字向量嵌入（以橘色標示），我們在圖 8 中於測試集上視覺化了 MuFFIN 的表示 H^{p} 與 E^{p} 。比較圖 8(a) 與圖 8(b) 後，我們觀察到所提出的 $\mathcal{L}_{\mathrm{con}}$ 能有效地將這兩種音素表示投射到共享的特徵空間，從而呈現出更一致的分佈。再進一步，圖 8© 提供了放大檢視，突顯出對比項不僅將異質的音素表示與相應的文字嵌入對齊，還保留了各音素類別的音素特有特徵。

語音單位(logits)的定性視覺化以說明所提出之音素特定變異(PhnVar)。最後，為了定性評估所提出的 PhnVar 訓練方案之有效性，我們視覺化了音素 logits 及診斷預測器的決策邊界。如圖 9 所示，我們比較了以 PhnVar 訓練之 MuFFIN 模型與其變體（即未考慮發音難度因子項的 PhnVar）。此外，視覺化的音素（$/\mathrm{AH}/\mathrm{h}/\mathrm{T}/, /\mathrm{IH}/$ 、$/\mathrm{K}/$ ）皆從測試集以均勻抽樣取得，其出現次數分別為 $4.4\ \mathrm{K}, 4\ \mathrm{K}, 3\ \mathrm{K}$ 和 1.3 K，錯誤發音率則分別為 $4.80\%, 3.55\%, 5.46\%$ 和 2.39% 。

圖 9 的觀察重點如下。首先，出現頻率較高的音素的 logits 傾向佔據較大的特徵空間，而出現頻率較低的音素則被壓縮到較窄的區域。這可從音素 $/\mathrm{K}//\mathrm{IH}//\mathrm{T}/$ 到 $/\mathrm{AH}/$ 的特徵區域逐漸增大的現象看出，與它們各自的出現頻率一致。接著，在圖 9(b)中可見，使用 PhnVar 變體（即 PhnVar w/o DF）訓練 MuFFIN 會導致特徵區域分佈較為均勻，與音素出現頻率無關。然而，僅考慮資料量因子來調整特徵空間仍無法捕捉誤讀分佈。因此，我們的 PhnVar 另外納入了發音困難度因子。圖 9(c)視覺化了以 PhnVar 訓練的 MuFFIN 之音素 logits，特徵區域依照音素誤讀率劃分，區域大小依序由大到小為 /IH/、/AH/、/T/ 與 /K/。

B. 自動發音評估的效能

在本小節中，我們轉而評估 MuFFIN 在發音評量上的表現，並將其與多個最先進模型進行比較以驗證其有效性。相應結果列於表 III。我們先討論音素與詞彙層級的實驗結果，然後再進入語句層級的評量。

音素與詞彙層級的評量表現。我們首先在表 III 中評估音素與詞彙層級的評量表現，從中可得出以下觀察。首先，所提出的 MuFFIN 在大多數發音評量任務上，以顯著差距優於其他 APA 模型，唯獨詞彙層級的重音例外。具體而言，MuFFIN 在音素層級的準確度上表現突出，PCC 分數分別較 4.9% 及 5.9% 提升。

相較於先前的模型 3 MH 與 HierGAT，我們將這些效能提升歸因於所提出的多面向音素層級發音回饋模組，該模組共同優化 APA 與 MDD 任務，從而鼓勵音素編碼器在評估發音分數時學習區別性的音素身分。就單字層級評估而言，MuFFIN 在大多數發音面向上整體表現良好。然而，在單字層級重音方面，我們的模型與 GFR 及 HierGAT 表現相當，且落後於 3 MH 。造成較差表現的可能原因是 3 MH 採用子音素建模，在音素與單字層級之間建立一個偽（增強）語言學階層，從而有助於更好地呈現超音段資訊以利單字層級的評估。

其次，我們轉而評估具備平行神經架構的強基線模型表現（表 III 中的第二組）。與 LSTM 和 GOPT 相比，3M 脫穎而出，顯示出有前景的方法，其優勢來自於有效利用基於 SSL 的特徵來減緩 L2 學習者語音資料的稀缺問題，同時封裝長距離發音特徵。透過將電磁發音描記（electromagnetic articulography）特徵加入 3M 的輸入，JAM 在大多數發音評估任務中略微提升了表現。最後，在具有進階神經架構的 APA 模型（表 III 中的第三組）中，3MH 獲得最佳表現，受益於階層式建模方法與深度導向卷積層的協同效應。然而，與 MuFFIN 相比，3MH 在功能上有所限制，因為它僅以多面向分數評量發音能力，缺乏子音節（phoneme）層級的診斷回饋。

句子層級的評估表現。針對表 III 中的句子層級發音評估表現，MuFFIN 在大多數面向上獲得最高的表現。與 3 MH 相比，MuFFIN 在句子層級準確度的 PCC 分數提升了 2.5% 、在句子層級總分提升了 2.9% ，並在句子層級流暢度與韻律上達到相當的表現。MuFFIN 在句子層級完整度評估（該指標反映一個句子中正確發音單字的比例）上也取得顯著改善。此一提升歸因於在 APA 模型內對 MDD 任務的聯合訓練，因而使 MuFFIN 能夠定位發音錯誤的片段並識別學習者語音中的對應音素。透過利用具音素區辨性的表示，MuFFIN 能夠經由客製化且具階層感知的神經架構，將從音素到句子層級評估的細緻資訊有效傳遞。

接著，與其他強力基線模型相比，Lin2021 在表現上落後於數個針對多重發音面向訓練的 APA 模型（表 III 的第二組），顯示單一面向評估模型未能透過多任務學習來利用面向間的相依關係，因而造成較差的效能。在後續工作中，Kim2022 嘗試以 SSL 為基礎的特徵取代傳統的 ASR 驅動特徵，帶來顯著改善並達到與 3 M 相近的表現。與多面向發音評估模型（即 JAM 與 Ryu2023）比較時，JAM 在大多數語句層級評估面向上表現較佳。

表 5：表 V
MuFFIN 與比較方法在 MDD 任務上的效能評估

Model	誤讀偵測			發音錯誤診斷			PER (%) ↓
	RE (%) ↑	PR (%) ↑	F1 (%) ↑	FAR (%) ↓	FRR (%) ↓	DER (%) ↓	
Ryu2023 [59]	91.60	26.90	41.50	-	-	-	9.93
JAM [60]	34.76	61.10	45.01	64.32	0.58	45.23	2.81
MuFFIN	64.33	66.89	65.99	35.67	0.97	60.97	2.36
w/ PhnVar	68.37	67.60	67.98	31.63	1.01	58.82	2.33

在誤讀檢測子任務中，我們表現最佳的模型（MuFFIN+PhnVar）以顯著更好的表現超越了基礎模型（MuFFIN）（$p < 0.001$）。

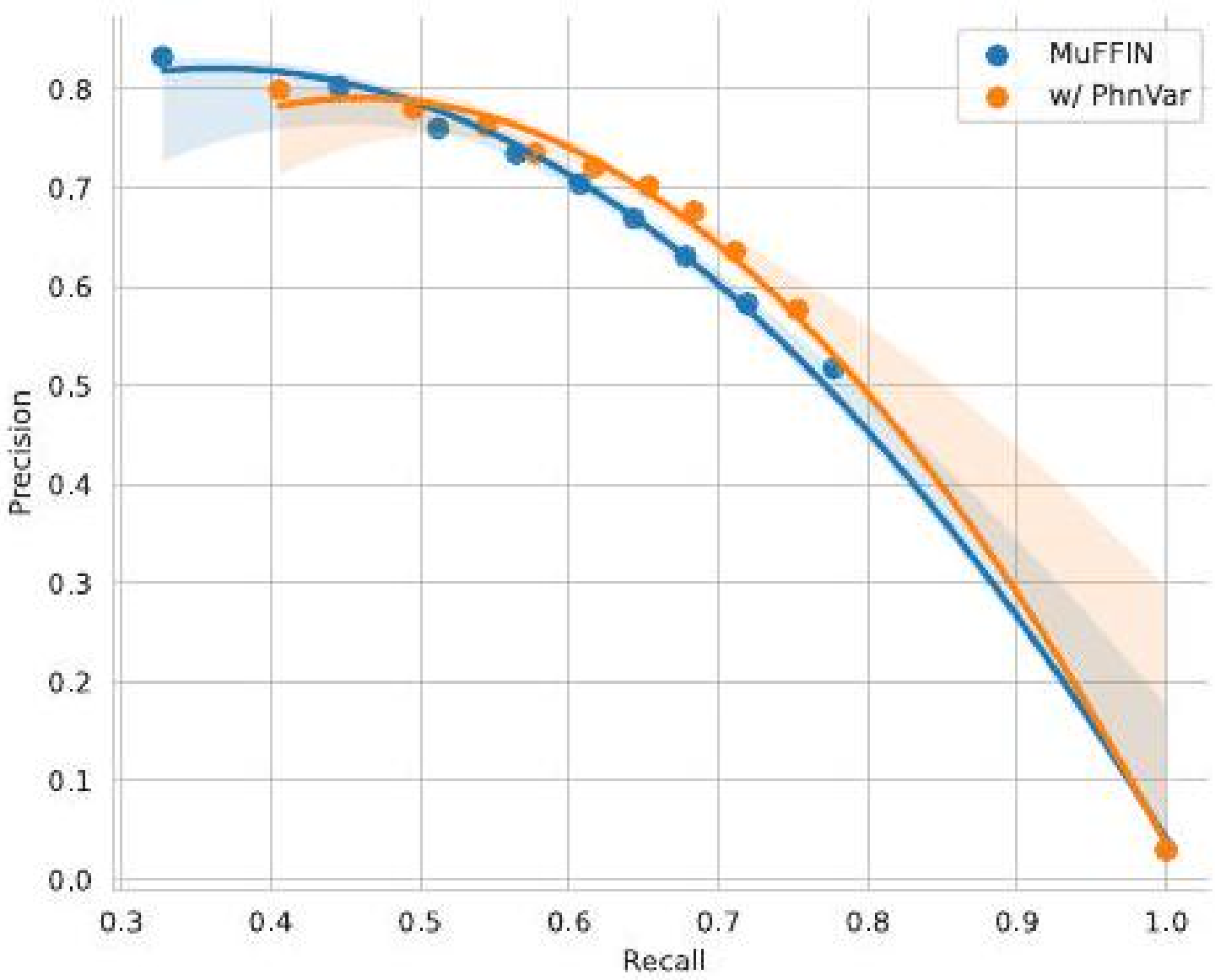


圖 10：圖 10. 使用 PhnVar 訓練的 MuFFIN 的精確度-召回率曲線。

此改進可能源於新穎地使用了細緻的音素層級特徵，包括 GOP 特徵、韻律統計量與 EMA 特徵。
音素特定變異與 ConPCO 的有效性。最後，我們檢驗了所提出的訓練方案——音素特定變異（PhnVar）以及對比音位序數正則化器（ConPCO）——在發音評估上的有效性。在表 IV 中，我們專注於音素與詞層級的評估表現，因為我們經驗上發現所提出的正則化器並不會對語句層級的發音評估造成不利影響，表現要麼略有提升，要麼至少與原始 MuFFIN 模型 [24][25] 相當。此外，ConPCO 被分解為對比項（ $\mathcal{L}_{\text{con}}$ ）與音位序數正則化器（ $\mathcal{L}_{pco} = \mathcal{L}_{pc} + \mathcal{L}_{o}$ ），兩者皆與 PhnVar 結合用於訓練 MuFFIN。

從表 IV 可以觀察到，我們提出的訓練方案 PhnVar 在音素層級準確度上比基礎模型提升了 0.7% 。隨後，在 PhnVar 訓練策略下加入音素層級正則化項（即 $\mathcal{L}_{\text{con}}$ 與 $\mathcal{L}_{\text{pco}}$ ）有利於發音評量，這從音素與字詞層級評量任務上持續或提升的結果可見一斑。此外，對比損失項主要提升了音素層級準確度與字詞層級總分的表現。相較之下，音素序數正則化器則傾向於略微提升或維持 vanilla MuFFIN 模型的表現。另外，以 ConPCO 訓練的 MuFFIN 在字詞層級評量任務上達到最佳表現（如表 IV 最後一列所示）。

C. 錯誤發音偵測與診斷的表現

在本小節中，我們評估 MDD 在多面向發音評量模型中的效能。由於 APA 與 MDD 的聯合優化在電腦輔助語言教學（CAPT）領域仍較少被探討，因此在下列實驗中可供比較的相關研究僅有少數。據我們所知，Ryu2023 是首個嘗試開發多面向發音回饋模型的研究，而 JAM 則代表近期的後續工作。MDD 任務的主要結果彙整於表 V。接著，表 VI 更深入探討了資料不平衡所固有的問題。

MDD，展示了所提出的音素變異訓練（PhnVar）方案的有效性。

MDD 的效能評估。為了使用 MuFFIN 偵測誤讀，我們利用音素層級錯誤檢測器的輸出來識別誤讀片段。具體來說，我們首先收集訓練集中所有音素片段的檢測器輸出。接著透過格點搜尋選擇一個全域閾值，步進為 0.1，範圍為 $[0.0, 1.0]$ 。在

MDD 實驗開始時，我們在圖 10 中呈現了 MuFFIN 與以 PhnVar 訓練的 MuFFIN 的精確度-召回率曲線。對於我們的模型（即 MuFFIN 與 $\mathrm{MuFFIN} + \mathrm{PhnVar}$）此全域閾值設為 0.4。與先前基於聽寫的方法（例如 Ryu2023 與 JAM）透過比對辨識到的音素與典型音素之間的不一致來偵測誤讀不同，MuFFIN 採用基於分數的方式，透過閾值調整來偵測誤讀，從而更易於適應各種不同的第二語言學習者。

如表 V 所示，MuFFIN 在誤讀偵測子任務上勝過其他方法，於 F1 分數與精確度上表現優異。此外，以音素特定變異（PhnVar）訓練 MuFFIN，相較於基礎模型在所有評估指標上皆有顯著提升。此增益在圖 10 中進一步說明：橘色曲線（MuFFIN+PhnVar）於精確度-召回率曲線下面積上超越藍色曲線（MuFFIN）。接著，與其他基線方法比較時，Ryu2023 在以 CTC 為基礎的音素辨識器之上取得最高的召回率，但在誤讀偵測任務上精確度偏低。此限制與以聽寫為基礎之 MDD 模型所報導的缺點一致[37][38]，因為模型效能本質上受限於音素辨識率。JAM 則非透過直接的自由音素辨識流程，而是建立在 3M 之上，透過將音素分類器附加至音素層編碼器來偵測學習者語音中的誤讀。 相應結果在精確度指標上表現可觀，但在召回率上仍有困難。與 JAM 相比，我們的 MuFFIN 在誤讀檢測子任務的所有指標上皆達到更優異的表現。這些發現共同突顯了所提出基於評分的方法在多面向發音回饋上的有效性。

在誤讀診斷子任務中，我們的方法在 FAR 與 PER 指標上達到可觀的表現。

表 6：表 VI
以 MuFFIN 探討 MDD 中的不平衡問題

群組	指標	平均 PER (%) ↓			平均召回率 (%) ↑			平均精確度（%） ↑			平均 F1 分數（%） ↑		
出現次數	類型	許多	中等。	少數	許多	中等	少量	大量	中等	少量	大量	中等	少量
	MuFFIN	1.19 (0.31)	2.27 (0.14)	10.93 (25.79)	70.19 (7.76)	75.53 (15.61)	68.56 (22.75)	50.84 (12.31)	53.42 (19.05)	69.50 (24.79)	57.53 (6.95)	60.14 (14.59)	67.77 (21.43)
	PhnVar	1.41 (0.34)	2.41 (1.51)	9.38 (20.19)	61.90 (6.22)	64.04 (17.41)	68.97 (24.66)	66.82 (9.81)	66.27 (16.03)	72.70 (24.14)	63.86 (5.80)	62.71 (10.64)	69.05 (9.38)
	不含 DF	1.45 (0.34)	2.42 (1.63)	11.22 (25.83)	64.71 (8.43)	68.62 (23.62)	67.99 (24.77)	61.27 (12.99)	58.87 (14.96)	61.83 (22.68)	63.02 (7.94)	59.23 (14.34)	62.38 (19.55)
	不含 QF	1.28 (0.39)	2.09 (1.36)	9.89 (21.73)	57.55 (6.39)	62.72 (22.32)	60.02 (23.27)	69.22 (9.90)	64.04 (19.84)	76.22 (24.85)	62.58 (6.80)	61.23 (18.60)	65.22 (20.65)
誤讀率	類型	高	中位數	低	高	中位數	低	高	醫學。	低	高	醫學。	低
	MuFFIN	10.62 (25.82)	1.69 (1.01)	2.08 (2.41)	77.43 (12.19)	69.12 (8.91)	67.73 (23.77)	61.70 (16.77)	55.83 (20.91)	56.23 (24.37)	66.95 (10.94)	59.81 (14.18)	58.68 (20.39)
	PhnVar	9.30 (20.13)	1.82 (0.91)	2.15 (2.36)	70.51 (14.30)	62.19 (10.57)	62.22 (24.78)	70.88 (12.35)	67.97 (16.37)	66.95 (23.08)	69.10 (6.93)	64.22 (10.94)	62.30 (20.47)
	不含 DF	10.86 (25.75)	1.80 (0.92)	2.42 (3.68)	71.33 (12.52)	73.52 (16.62)	59.16 (26.29)	66.10 (14.05)	54.20 (14.12)	61.67 (21.04)	66.77 (8.10)	60.79 (10.93)	57.08 (20.64)
	不含 QF	9.27 (21.65)	1.70 (1.06)	2.29 (3.72)	66.23 (14.28)	55.39 (16.96)	58.67 (23.25)	71.94 (12.39)	70.43 (21.93)	67.11 (23.18)	67.29 (8.42)	61.04 (17.27)	60.69 (20.74)

在比較 PhnVar 變體（不含 DF 與不含 QF）時的效能提升以粗體顯示，而與未經 PhnVar 訓練的原始 MuFFIN 相比的改進則以底線標示。「Mispron. Rate」表示誤發音率。

然而，召回率與 FRR 及 DER 指標之間似乎存在權衡。具體而言，與 JAM 相比，MuFFIN 在召回率上更高且 PER 較低，但在 FRR 與 DER 兩項上表現較差。這結果意味著我們的模型能偵測出更多誤發音段落，但以診斷準確性為代價。我們將此問題留作未來研究方向。
系統性檢視 MDD 中的資料不平衡問題。在下列章節中，我們探討 MDD 中源自音素片段之兩個交織因素的資料不平衡問題，即資料數量與發音困難度。為了將這兩個因素解開，我們將音素片段分為兩組，並針對每個子集報告相應的音素錯誤率（PER）、召回率、精確度與 F1 分數，並給出其平均值與標準差。如表 VI 所示，在第一組中，音素片段依其出現次數被分為多樣（出現次數超過 1.3 K）、中等（出現次數介於 1.3 K 與 0.6 K 之間）與少量（出現次數低於 0.6 K）三個區域。相反地，在第二組中，音素片段依其誤讀率被分類為高（誤讀率高於 5.1%）、中等（誤讀率介於 5.1% 與 3.4%）與低（誤讀率低於 3.4%）三個發音錯誤區域（參見圖 6）。 此外，本組實驗旨在評估 MuFFIN 的上限表現，分析 MDD 的不平衡問題並檢驗所提出的音素特定變異量（PhnVar）的效用。為此，從訓練資料中預留出 500 個句子作為驗證集，剩餘 2,000 個句子用於模型訓練。此驗證集被設計以涵蓋每個音素的正確與錯誤發音，並用以透過最大化精確率-召回率曲線下面積來決定音素特定的閾值。

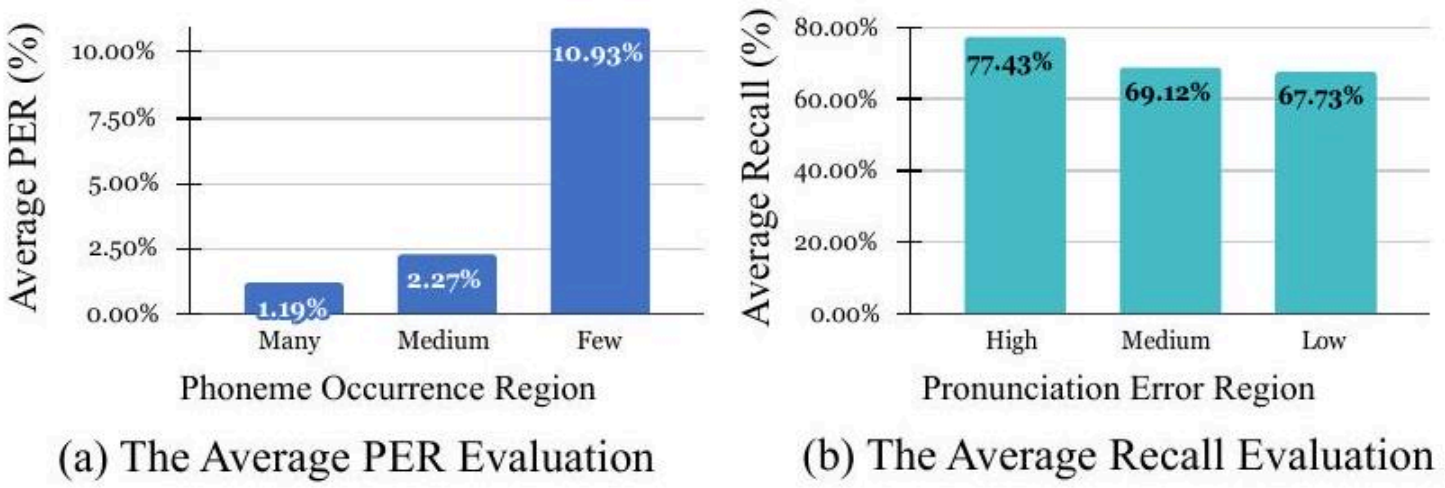


圖 11：圖 11 的效能評估突顯了 MDD 的資料不平衡問題。我們的長條圖顯示 (a) 資料量因素主要影響依出現次數分組之音素子集的平均 PER，及 (b) 發音難易度因素顯著影響依誤讀率分組之音素子集的平均召回率。

根據這些音素特定閾值，我們接著在 Speechocean762 測試集上評估 MDD 的表現。

為了突顯資料不平衡問題，圖 11 先基於 MuFFIN 模型呈現兩組長條圖，其中圖 11(a) 顯示依出現次數分組的音素子集之平均 PER，而圖 11(b) 則顯示依誤讀率分組的平均召回率。在圖 11(a) 中，我們可以觀察到 MuFFIN 的平均 PER 從多樣本區間顯著上升至少樣本區間。此觀察與先前關於資料不平衡學習的研究結果一致[22][23]，顯示出現頻率在音素辨識準確率上扮演重要角色。基於經驗風險最小化的音素分類器之天真的訓練過程，難免會使模型偏向多數音素類別，因而在少數類別上表現不佳。除了資料量問題外，圖 11(b) 顯示發音難度因素會導致平均召回率呈穩定下降趨勢，因為
音素子集的錯誤發音率從高轉為低。這表明那些較少被錯誤發音的音素片段對發音錯誤偵測構成更大的挑戰。基於這些觀察，所提出的 PhnVar 訓練方案整合了兩個數學項，以同時考量資料量與發音困難度。

接著我們探討 PhnVar 中各個組件在處理 MDD 不平衡問題上的效能。表 VI 的消融研究是透過從所提出的 PhnVar 中排除資料量因子（w/o QF）或發音困難度因子（w/o DF）來進行。如表中所示，我們在這兩組中觀察到一致的效能趨勢：當考慮資料量因子（w/o DF）時，MuFFIN 模型在 recall 上有提升；而當納入發音困難度因子（w/o QF）時，precision 則獲得改善。進一步同時考慮兩個因子時，PhnVar 在 F1-score 上取得顯著提升。這些結果指出，單純平衡音素預測的 logits 會促使 MDD 模型偵測到更多的發音錯誤；然而這會提高 recall 率，同時導致 precision 評估下降。一個合理的解釋是：誤讀率在所有音素間並非均勻分布。 作為補救方法，所提出的 PhnVar 同時考量資料量與發音難度兩大因素，在召回率與精確率之間取得平衡，以達到最佳的 F1 分數。此外，與原始 MuFFIN 模型相比，使用 PhnVar 進行訓練能顯著提升精確率與 F1 分數的表現，增益在出現頻率為中等與少量的音素子集，以及誤讀率為中等與低等的子集中尤為明顯。

D. 對 APA 與 MDD 目標之消融研究

在本小節中，進行一系列消融研究，以分析不同訓練目標對 MuFFIN 在發音準確度與誤讀偵測表現上的有效性。
MuFFIN 中多粒度發音評估的有效性。在這組實驗中，我們先訓練不含 MDD 任務的 MuFFIN，然後逐步加入不同語言層級的發音評估任務。表 VII 報告了針對發音準確性評估的 PCC 分數，涵蓋了在音素、單字、語句層級的評估任務，以及跨粒度的組合。從表 VII 我們觀察到，以多粒度方式訓練的 MuFFIN 相較於本文比較的任何單一粒度評估模型皆能達到更優的結果，這表示在語句的語言層級中，評估任務之間存在強烈的關聯。例如，以多粒度目標訓練的 MuFFIN（即 Phone+Word 及 Phone+Word+Utt.）分別優於其單一粒度的對應模型（即 Word Only 及 Utt. Only），分別提升了 14% 與 13% 。此外，參數量的比較顯示語句層級的評估模型（即 Utt. Only 與 Phone+Word+Utt. 的參數量顯著大於其他不同粒度的評估模型

表 7：表七
MuFFIN 中 APA 訓練目標不同粒度的消融研究

訓練目標	參數	發音面向 (PCC 分數)		
		音素準確度	詞彙準確率	語句準確率
僅語句	541 K	-	-	0.782
僅字詞	248 K	-	0.674	-
僅語音	126 K	0.715	-	-
+ 單字	249 K	0.724	0.688	-
+ 發音	608 K	**0.726**	**0.687**	**0.807**

在此表中，MuFFIN 指不包含 MDD 任務的訓練，Utt. 表示該語句。

表 8：表八
MuFFIN（含 APA 與 MDD 目標）之訓練粒度消融研究

訓練目標	APA 任務			MDD 任務		
	Phone（音素）Acc.（準確度）	單字準確率	句子準確率	F1 (%)	RE (%)	PR (%)
僅 MDD	-	-	-	62.71	65.67	60.33
+發話 (Utt.)	-	-	0.787	63.34	63.49	63.45
+詞 (Word)	-	0.681	-	64.46	66.27	62.86
+音素 (Phone)	0.717	-	-	**66.26**	**69.06**	63.77
+ 詞 (Word)	0.741	0.696	-	66.04	67.08	65.36
+ Utt.	**0.742**	**0.705**	**0.807**	65.99	64.33	**66.89**

Utt. 表示語句 (utterance)，Acc. 表示發音準確度。

組合（例如 Phone Only、Word Only 與 Phone+Word）。我們將此歸因於平均池化特徵 $\overline{\mathbf{E}}^{\mathrm{SSL}}$ 與語句層迴歸器之間的殘差連接。

聯合訓練 MDD 與 APA 對 MuFFIN 的效用。我們接著分析在多面向發音評估模型中，MDD 與 APA 各自訓練目標的貢獻。表 VIII 詳列 MuFFIN 在同時處理 MDD 與 APA 任務時的表現，其中針對 MDD 的評估指標與不同粒度下發音準確度的 PCC 分數皆有報告。細看表 VIII，我們有以下觀察：1) 將 APA 任務整合進來的多面向發音模型（例如 MDD+Utt., MDD+Word, 與 MDD+Phone）在所有 MDD 評估指標上均持續優於僅以 MDD 訓練的模型（即 MDD Only），顯示出同時建模 MDD 與 APA 的協同效果。2) 在這些多面向發音模型中，以音素層級評估與 MDD 任務共同訓練的模型（MDD+Phone）在 recall 指標上表現最佳。3) 關於發音評估的表現，從表 VII 與表 VIII 的觀察顯示，加入 MDD 任務可維持或略微提升發音準確度。 最後，發音評估效能的主要提升來自於在多個語言層次整合多樣化評估任務。

VI. 結論

在本文中，我們提出一個新穎的多面向發音回饋模型，命名為 MuFFIN，旨在從多個角度評估學習者的發音品質，包括跨越各種語音面向的發音評估。

語言學層次，以及音素層級的誤讀偵測與診斷。提出了一種新穎的對比性音素序數正則化器（contrastive phonemic ordinal regularizer），使 MuFFIN 能在考量音素層級準確度分數序數性質的同時，產生更具音素辨識性的特徵。此外，為了解決 MDD 的複雜資料不平衡問題，我們提出一個簡單但有效的訓練方案，透過對音素分類器的輸出施加音素特定的擾動來實現。此方法在納入發音難度考量的同時，有效平衡預測音素的分布。我們在 speechocen762 基準資料集上進行大量實驗以驗證該方法的實用性。提出的對比性音素序數正則化器亦已透過一系列圖形化視覺化結果進行詳細檢驗。此外，本研究為首次嘗試從資料量與發音難度兩個面向著手解決 MDD 的資料不平衡問題。實證結果顯示，我們的模型在 APA 與 MDD 任務上均優於部分最先進的方法。

限制與未來工作。本方法的侷限在於其依賴於「朗讀」學習情境，且在某種程度上對所提供評估結果缺乏可解釋性。此類腳本化語音的評估無法反映學習者在真實溝通中的口語能力。此外，實驗資料集僅包含中文普通話學習者，可能妨礙對具有其他口音學習者的泛化能力與適用性。未來工作中，我們計畫在口語評量上檢驗所提方法，使學習者能自由發言或對給定任務或問題作答 [61]。此外，可解釋發音反饋的議題亦將作為未來延伸研究。

References

[1] L. Davis, and J. M. Norris, “Assessing second language speaking at ETS: Introduction,” Routledge, 2025. 1-18.
[2] A. Van Moere and R. Downey, “Technology and artificial intelligence in language assessment,” in Handbook of Second Language Assessment, D. Tsagari and Banerjee J., Eds., 2017, pp. 341-357.
[3] P. M R.-Revell, “Computer-assisted pronunciation training (CAPT): Current issues and future directions,” RELC Journal, vol. 52, pp. 189205, 2021.
[4] K. Evanini, and X. Wang，「Automated speech scoring for non-native middle school students with multiple task types」，收錄於 Proc. Interspeech, 2013，頁 2435-2439。
[5] Y. K. Singla, A. Gupta, S. Bagga, C. Chen, B. Krishnamurthy, and R. R. Shah，「Speaker-conditioned hierarchical modelling for automated speech scoring」，收錄於 Proc. Int. Conf. Inf. Knowl. Manag., 2021，頁 1681-1691。
[6] K. Evanini and X. Wang，「Automated speech scoring for Nonnative middle school students with multiple task types」，收錄於 Proc. Interspeech, 2013，頁 2435-2439。
[7] K. Evanini, M. C. Hauck, and K. Hakuta，「Approaches to automated scoring of speaking for K-12 English language proficiency assessments」，收錄於 ETS Research Report Series，頁 1-11，2017。
[8] D. Korzekwa, J. Lorenzo-Trueba, T. Drugman, and B. Kostek，「Computer assisted pronunciation training—Speech synthesis is almost all you need」，Speech Commun., vol. 142, pp. 22-33, 2022.
[9] K. Li, X. Qian, and H. Meng，「Mispronunciation detection and diagnosis in L2 English speech using multi-distribution deep neural networks」，IEEE/ACM Trans. Audio Speech Lang. Process., vol. 25, no. 1, pp. 193-207, 2017.

[10] S. Bannò, B. Balusu, M. Gales, K. Knill, and K. Kyriakopoulos，「View-specific assessment of L2 spoken English」，in Proc. Interspeech, 2022, pp. 4471-4475.
[11] N. F. Chen, and H. Li，「Computer-assisted pronunciation training: From pronunciation scoring towards spoken language learning」，in Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf., 2016, pp. 1-7.
[12] S. M. Witt 和 S. J. Young，「Phone-level pronunciation scoring and assessment for interactive language learning」，Speech Commun., 第 30 卷，頁 95-108，2000。
[13] S. Sudhakara、M. K. Ramanathi、C. Yarra 和 P. K. Ghosh，「An improved goodness of pronunciation (GOP) measure for pronunciation evaluation with DNN-HMM system considering hmm transition probabilities」，收錄於 Proc. Interspeech，2019，頁 954-958。
[14] W. Hu、Y. Qian、F. K. Soong 和 Y. Wang，「Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers」，Speech Commun., 第 67 卷，頁 154-166，2015。
[15] L. Ferrer、H. Bratt、C. Richey、H. Franco、V. Abrash 和 K. Precoda，「Classification of lexical stress using spectral and prosodic features for computer-assisted language learning systems」，Speech Commun., 第 69 卷，頁 31-45，2015。
[16] E. Coutinho 等人，「評估非母語英語使用者的韻律：量測與特徵集合」，收錄於 Proc. Lang. Resour. Eval. Conf., 2016, 頁 1328-1332。
[17] C. Cucchiarini 等人，「利用自動語音辨識技術對第二語言學習者流暢度之量化評估」，J. Acoust. Soc. of Am., 2000。
[18] C. Zhu, T. Kunihara, D. Saito, N. Minematsu, 及 N. Nakanishi，「自動預測日本英語學習者口語產出之詞彙與音素可懂度」，收錄於 Proc. IEEE Spoken Lang. Technol. Workshop, 2022, 頁 1029-1036。
[19] N. F. Chen, D. Wee, R. Tong, B. Ma, 及 H. Li，「大規模描述歐洲裔說話者所說之非母語國語：基於 icall 之分析」，Speech Commun., 卷 84, 頁 46-56, 2016。
[20] W. Li, N. F. Chen, S. M. Siniscalchi, and C.-H. Lee，「使用軟目標音調標籤與基於 BLSTM 的深度音調模型改善非母語學習者之中文聲調誤讀偵測」，IEEE/ACM Trans. Audio Speech Lang. Process., vol. 27, pp. 2012-2024, 2019.
[21] K. Fu, J. Lin, D. Ke, Y. Xie, J. Zhang, B. Lin，「使用簡易資料增強技術之全文字依賴端到端誤讀偵測與診斷」，2021, arXiv preprint arXiv:2104.08428.
[22] A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, and S. Kumar，「透過 logit 調整進行長尾學習」，收錄於 Proc. Int. Conf. on Learn. Representations, 2021.
[23] Y. Wang, J. Fei, H. Wang, W. Li, T. Bao, L. Wu, R. Zhao, Y. Shen，「為長尾語義分割平衡 logit 變異」，收錄於 Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2023, pp. 19561-19573.
[24] B.-C. Yan、H.-W. Wang、Y.-C. Wang、J.-T. Li、C.-H. Lin 與 B. Chen，「為序數回歸保留音素區別：自動發音評估的新型損失函數」，收錄於 Proc. IEEE Autom. Speech Recognit. Underst. Workshop，2023，頁 1-7。
[25] B.-C. Yan、H.-W. Wang、Y.-C. Wang、J.-T. Li、W.-C. Chao、B. Chen，「ConPCO：利用對比序數正則化保留音素特徵以進行自動發音評估」，收錄於 Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.，2025，頁 1-5。
[26] J. Zhang、Z. Zhang、Y. Wang、Z. Yan、Q. Song、Y. Huang、K. Li、D. Povey 與 Y. Wang，「Speechocean762：一個用於發音評估的開源非母語英語語料庫」，收錄於 Proc. Interspeech，頁 3710-3714，2021。
[27] E. B. Page，「電腦評分論文中的統計與語言策略」，收錄於 Proc. Conf. Comput. Linguistics，1967，頁 1-13。
[28] M. Wu、K. Li、W.-K. Leung 及 H. Meng。「基於 Transformer 的端到端誤讀偵測與診斷」，收錄於 Proc. Interspeech, 2021, 第 3954–3958 頁。
[29] S. Bannò 及 M. Matassoni。「使用 wav2vec 2.0 進行 L2 口語英語能力評估」，收錄於 Proc. IEEE Spoken Lang. Technol. Workshop, 2022, 第 1088–1095 頁。
[30] E. Kim、J.-J. Jeon、H. Seo 及 H. Kim。「使用自監督語音表示學習的自動發音評估」，收錄於 Proc. Interspeech, 2022, 第 1411–1415 頁。
[31] B.-C. Yan、H.-W. Wang、Y.-C. Wang 及 B. Chen。「用於誤讀偵測與診斷的發音特徵之有效圖形化建模」，收錄於 Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2023, 第 1–5 頁。
[32] J. Shi、N. Huo 與 Q. Jin，「用於電腦輔助發音教學的情境感知發音良好度」，收錄於 Proc. Interspeech, 2020, 頁 3057-3061。
[33] Q.-T. Truong、T. Kato 與 S. Yamamoto，「使用 F0 與強度輪廓加權距離之 L2 英語詞語韻律自動評估」，收錄於 Proc. Interspeech, 2018, 頁 2186-2190。
[34] C. Graham 與 F. Nolan，「將發音速率作為口語語言評估指標」，收錄於 Proc. Interspeech, 2019, 頁 3564-3568。
[35] S. Mao、Z. Wu、R. Li、X. Li、H. Meng 與 L. Cai，「將多任務學習應用於聲學-音素模型以進行 L2 英語語音之誤讀偵測與診斷」，收錄於 Proc. IEEE Int. Conf. on Acoust., Speech, Signal Process., 2018, 頁 6254-6258。
[36] B. Lin, L. Wang, X. Feng, and J. Zhang，「L2 發音之多粒度自動評分」，見 Proc. Interspeech. Assoc., 2020, 頁 3022-3026。
[37] W.-K. Leung, X. Liu and H. Meng，「基於 CNN-RNN-CTC 之端到端誤發音檢測與診斷」，見 Proc. IEEE Int. Conf. on Acoust., Speech, Signal Process., 2018, 頁 8132-8136。
[38] B.-C. Yan, M.-C. Wu, H.-T. Hung, and B. Chen，「一種運用新穎反音素建模之 L2 英語語音端到端誤發音檢測系統」，見

Proc. Interspeech, 2020, 頁 3032-3036。
[39] B.-C. Yan, H.-W. Wang, and B. Chen，「Peppanet：結合語音、語音學與聲學提示之有效誤發音檢測與診斷」，見 Proc. IEEE Spoken Lang. Technol. Workshop, 2023, 頁 1045-1051。
[40] B.-C. Yan, H.-W. Wang, Y.-C. Wang, 和 B. Chen，「基於圖形的發音特徵有效建模以進行誤讀檢測與診斷」，收錄於 Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2023, 頁 1-5。
[41] B.-C. Yan, J.-T. Li, Y.-C. Wang, H.-W. Wang, T.-H. Lo, Y.-C. Hsu, W.C. Chao, 和 B. Chen，「一種結合分層 transformers 與預訓練策略的有效發音評估方法」，收錄於 Proc. Annu. Meet. Assoc. Comput. Linguist., 2024, 頁 1737-1747。
[42] H.-C. Pei, H. Fang, X. Luo 和 X.-S. Xu，「Gradformer：一個用於多面向多粒度發音評估的框架」，IEEE Trans. Audio Speech Lang. Process., 卷 32, 頁 554-563, 2024。
[43] P. Muller, F. De Wet, C. Van Der Walt, 和 T. Niesler，「自動評估精通 L2 語者的口語能力」，收錄於 Proc. Workshop Speech Lang. Technol. Educ., 2009, 頁 29-32。
[44] H. Franco、H. Bratt、R. Rossier、V. Rao Gadde、E. Shriberg、V. Abrash 與 K. Precoda，「Eduspeak：用於電腦輔助語言學習應用之語音識別與發音評分工具包」，Lang. Test., 卷 27，期 3，頁 401–418，2010。
[45] Y. Gong、Z. Chen、I.-H. Chu、P. Chang 與 J. Glass，「基於 Transformer 的多面向多粒度非母語英語發音評估」，於 Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2022，頁 7262–7266。
[46] F.-A. Chao、T. H. Lo、T. I. Wu、Y. T. Sung 與 Berlin Chen，「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」，於 Proc. Asia-Pac. Signal Inf. Process. Assoc. Annu. Summit Conf., 2022，頁 575–582。
[47] H. Do、Y. Kim 與 G. G. Lee，「具有多面向注意力的分層發音評估」，於 Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2023，頁 1–5。
[48] F.-A. Chao, T.-H. Lo, T.-I. Wu, Y.-T. Sung, B. Chen，「一種用於多面向與多粒度發音評估的分層情境感知建模方法」，發表於 Proc. Interspeech, 2023, 頁 974-978。
[49] C. Zhu, T. Kunihara, D. Saito, N. Minematsu, N. Nakanishi，「日本英語學習者口語中詞與音素可懂度的自動預測」，發表於 Proc. IEEE Spoken Lang. Technol. Workshop, 2022, 頁 1029-1036。
[50] Y. Shen, A. Yasukagawa, D. Saito, N. Minematsu, and K. Saito，「基於可解釋網路並利用發聲量與發音品質優化之 L2 英語流暢度預測」，發表於 Proc. IEEE Spoken Lang. Technol. Workshop, 2021, 頁 698-704。
[51] Y. Peng, S. Dalmia, I. Lane, and S. Watanabe，「Branchformer：用於擷取語音辨識與理解之局部與全域上下文的平行 MLP-Attention 架構」，發表於 Proc. Int. Conf. Mach. Learn., 2022, 頁 17627-17643。
[52] A. Radford 等人，「從自然語言監督學習可轉移的視覺模型」，見 Proc. Int. Conf. Mach. Learn., 2021, 頁 8748–8763。
[53] B. Elizalde、S. Deshmukh、M. A. Ismail 與 H. Wang，「CLAP：從自然語言監督學習音訊概念」，見 Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2023, 頁 1–5。
[54] A. Baevski、H. Zhou、A. Mohamed 與 M. Auli，「Wav2vec 2.0：一個用於語音表徵自監督學習的框架」，見 Proc. Adv. Neural Inf. Process. Syst., 2020, 頁 12449–12460。
[55] S. Chen 等人，「WavLM：用於全棧語音處理的大規模自監督預訓練」，IEEE J. Sel. Topics Signal Process., 第 16 卷, 頁 1505–1518, 2022。
[56] W.-N. Hsu 等人, “HuBERT: 透過對隱單位的遮蔽預測進行自我監督語音表示學習,” IEEE/ACM Trans. Audio, Speech, Lang. Process., 頁 3451-3460, 2021.
[57] B. Lin, 和 L. Wang, “用於自動發音評估的深度特徵轉移學習,” 收錄於 Proc. Interspeech, 2021. 頁 4438-4442.
[58] B.-C. Yan 和 B. Chen, “一種用於發音評估的有效分層圖注意力網路建模方法,” IEEE/ACM Trans. Audio Speech Lang. Process., 卷 32, 頁 3974-3985, 2024.
[59] H. Ryu, S. Kim, 和 M. Chung, “一個結合發音評估與誤讀偵測與診斷之多任務學習的聯合模型,” 收錄於 Proc. Interspeech, 2023, 頁 959-963.
[60] Y.-Y. He, B.-C. Yan, T.-H. Lo, M.-S. Lin, Y.-C. Hsu, 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」，收錄於 Proc. Asia-Pac. Signal Inf. Process. Assoc. Annu. Summit Conf., 2024,
[61] J. Park and S. Choi，「Addressing cold start problem for end-to-end automatic speech scoring」，收錄於 Proc. Interspeech, 2023, 頁 994-998。

顏必承（Bi-Cheng Yan）分別於 2025 年與 2017 年取得國立臺灣師範大學資訊工程學系的博士及碩士學位。2017 至 2020 年間任職於台北北投的華碩電腦股份有限公司（ASUSTeK Computer Inc.）。目前研究興趣包括電腦輔助語言學習（CALL）、語音辨識與語音增強。作者為超過 20 篇學術論文的作者或共同作者。

蔡明康於 1997 年自國立交通大學應用化學系取得理學士學位，並於 2005 年在美國匹茲堡大學取得計算化學博士學位。2005 至 2007 年間，他在美國太平洋西北國家實驗室（Pacific Northwest National Laboratory, Richland）進行博士後研究，隨後於 2007 至 2010 年間在美國布魯克海文國家實驗室（Brookhaven National Laboratory, Upton）從事博士後研究。2010 年起加入國立臺灣師範大學化學系，現任教授。迄今已發表及共同發表超過 50 篇學術論文。其研究興趣包括複雜化學系統之多尺度模擬、分子與材料之虛擬設計，以及針對化學資訊學應用之語言型人工智慧模型開發。

Berlin Chen Berlin Chen (M’04) 於 1994 年和 1996 年分別在國立交通大學電腦科學與資訊工程學系取得學士與碩士學位，並於 2001 年在國立臺灣大學電腦科學與資訊工程學系取得博士學位。他於 1996 年至 2001 年間任職於中央研究院資訊科學研究所（臺北），之後於 2001 年至 2002 年間任職於國立臺灣大學通訊工程研究所。2002 年起加入國立臺灣師範大學電腦科學與資訊工程研究所（臺北）。現任該校電腦科學與資訊工程學系教授。Chen 教授的研究興趣主要包括語音辨識與自然語言處理、多媒體資訊檢索、電腦輔助語言學習，以及一般的機器學習領域。

收稿：2025 年 5 月 18 日；修訂：2025 年 7 月 8 日及 2025 年 9 月 12 日；接受：2025 年 9 月 21 日。協助審稿並核准發表本篇文章之副主編為 Dr. Gutierrez-Osuna Ricardo。（通訊作者：Berlin Chen。）Bi-Cheng Yan 與 Berlin Chen 為國立臺灣師範大學資訊工程學系成員，
國立臺灣師範大學，臺北市 11677（電子郵件：80847001s@ntnu.edu.tw；berlin@ntnu.edu.tw）。Ming-Kang Tsai 為國立臺灣師範大學化學系成員，臺北市 11677（電子郵件：mktsai@ntnu.edu.tw）。

[1] CMU 字典：http://www.speech.cs.cmu.edu/cgi-bin/cmudict