ConPCO：利用對比序數正則化保留語音特徵的自動發音評估

顏碧成 [1] 、王義誠 [1] 、李俊廷 [1] 、林孟霞 [1] 、王馨瑋 [1] 、
趙偉成 [2] 、陳伯林 [1*]
國立臺灣師範大學 資訊工程與科學系
台灣電信公司高階技術實驗室
{bicheng, yichengwang, jtlee, berlin}@ntnu.edu.tw,
weicheng@cht.com.tw

摘要

自動發音評估（APA）旨在評估第二語言（L2）學習者在目標語言中的發音能力。現有的研究通常採用回歸模型來預測能力分數，這些模型在訓練時估計目標值，但並未明確考慮特徵空間中的音素感知。在本論文中，我們提出了一個針對基於回歸的 APA 模型的對比音素序數正則化器（ConPCO），旨在生成更具音素區別性的特徵，同時考慮回歸目標之間的序數關係。所提出的 ConPCO 首先通過對比學習，對齊 APA 模型的音素表示和音標文本嵌入。之後，通過調節特徵空間中音素類別間和類別內的距離，同時允許輸出目標之間的序數關係，來保留音素特徵。我們進一步設計和開發了一個分層的 APA 模型，以評估我們的正則化器的有效性。在 speechocean762 基準數據集上進行的一系列實驗表明，我們的方法在多個競爭基線中具有可行性和有效性。

關鍵字-電腦輔助語言學習、自動發音評估、對比學習

I. 緒論

隨著外語學習需求的激增，電腦輔助發音訓練（CAPT）系統的發展已在全球化浪潮中引起越來越多的關注。CAPT 系統旨在為二語（第二語言）學習者提供客製化且資訊豐富的回饋，讓他們能在無壓力且自主的學習環境中練習發音技能[1][2][3]。作為 CAPT 系統不可或缺的元件，自動發音評估（APA）的目標是確定二語學習者的口語能力程度，並對目標語言的特定發音面向提供詳細回饋[4][5]。

APA 系統的事實標準通常體現在朗讀學習情境中，在此情境下，二語學習者會被呈現一段文字提示，並被指示依照該提示發音[6][7]。通過對輸入語音和參考文字提示的協同處理，APA 系統預期能評估學習者的口語技能並提供即時回饋，包括整體能力（整體分數）或特定發音面向（分析性分數）。為了對學習者的發音品質提供深入回饋，近期研究致力於多面向和多粒度的發音評估，這些研究設計了統一的評分模型，以先進的平行[8][9]或階層神經架構共同評估不同語言層級（如音素、詞彙和話語）的發音能力，並考慮多元面向（如準確性、流暢度和完整性）。由於輸出目標的連續性可能是無限且無邊界的[11]，現有方法通常採用回歸損失函數，如均方誤差（MSE）作為

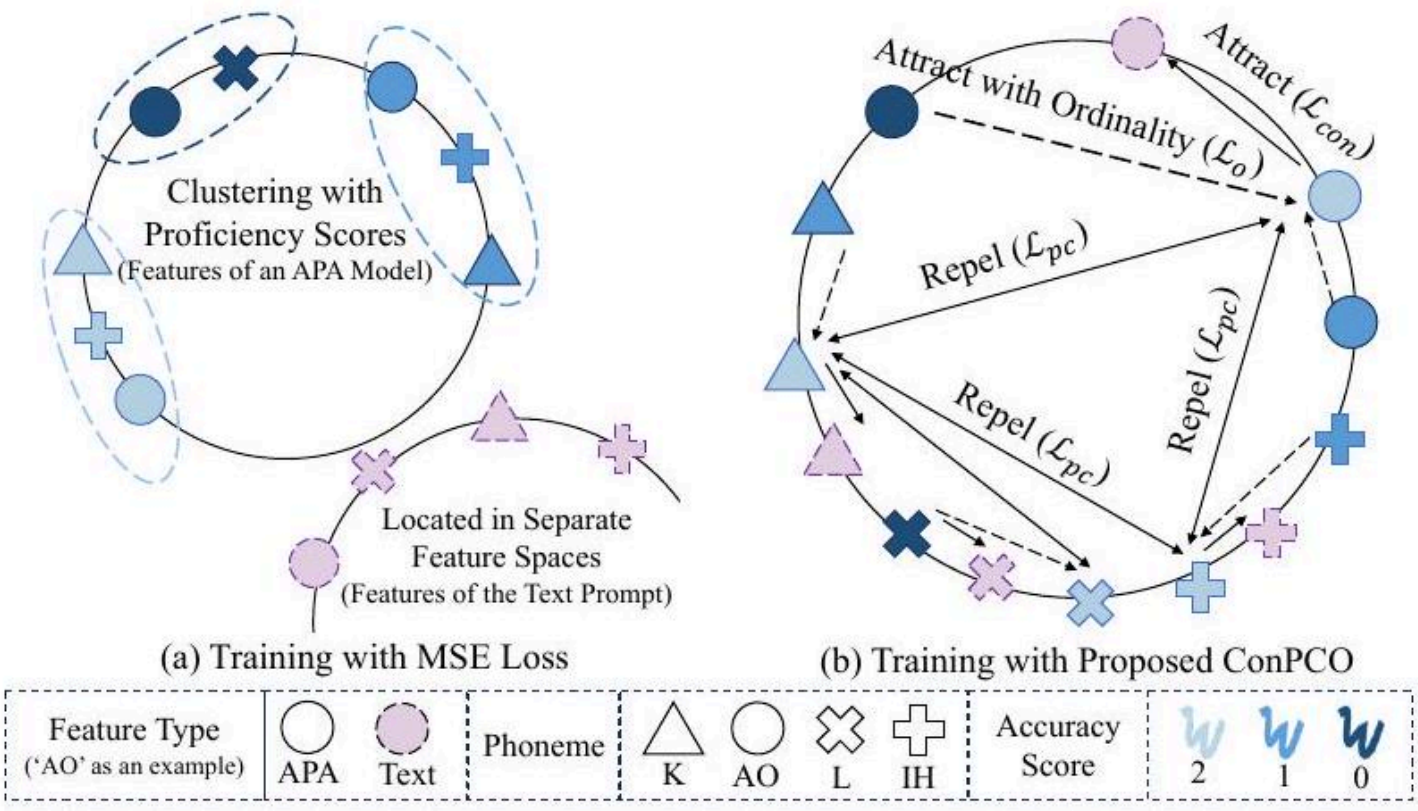
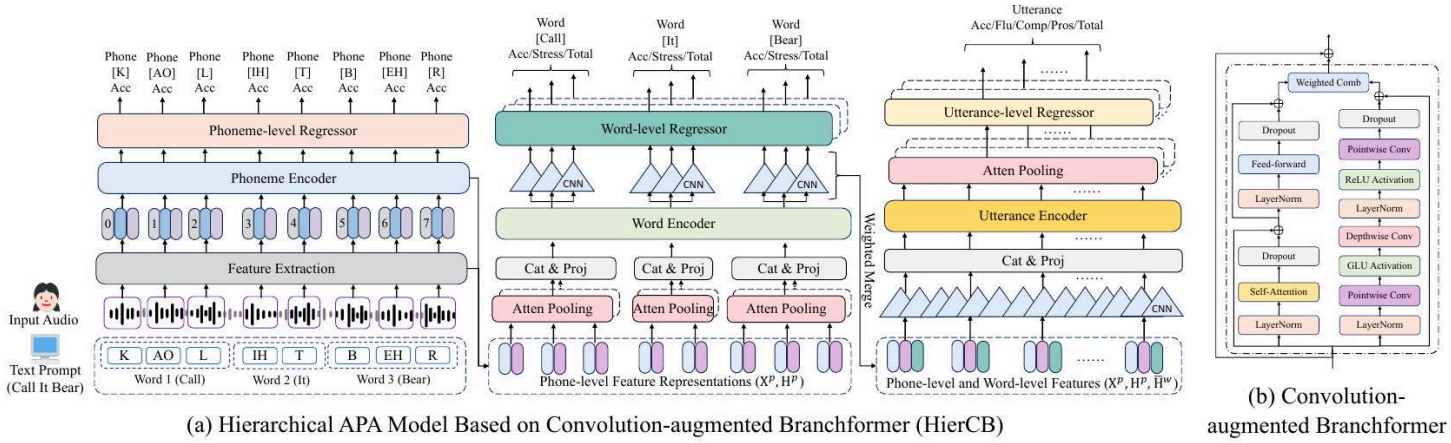


圖 1：圖 1. 我們動機的總結。如圖所示，一個使用(a) MSE 損失和(b)所提出的 ConPCO 訓練的 APA 模型。ConPCO 正則化器透過操縱特徵空間中類別間和類別內語音單位的距離，同時保留語音特徵，對齊來自異質資訊的語音特徵。

訓練目標是模仿專家的評估。儘管已取得一些有希望的成果，但語言單位的獨特特徵（例如語音資訊[12][13]和詞彙語義[14][15]）在優化過程中幾乎被忽略。

在這項研究中，我們識別出現有基於回歸的 APA 模型中的三個限制，如圖 1(a) 所示：(1) 由 APA 模型衍生的輸入語音的音素表示，以及音素級文字提示的文本嵌入，位於不同的特徵空間，這給存取音素分數同時保持音素身份意識帶來了挑戰；(2) 屬於相同熟練程度的不同音素表示被不經意地強制彼此靠近，這可能會損害與發音清晰度相關的評估任務的表現；(3) 在訓練目標的設計中幾乎忽視了回歸目標之間的序數關係，其中標籤空間中觀察到的序數行為未能在特徵空間中得到適當反映。 為了解決這些限制，我們提出了一種新穎的訓練機制，稱為對比音素序數正則化器（ConPCO），通過在特徵表示中捕捉音素特徵，同時保持回歸目標之間的序數關係，以增強基於回歸的 APA 模型。如圖 1(b) 所示，所提出的 ConPCO 通過對比損失，將 APA 模型的音素編碼器的輸出表示與音素級文字提示的嵌入對齊，這會將配對的音素表示拉近，同時將非匹配對的表示推遠。為了建模音素類別的細微差別，通過考慮其回歸目標的序數關係，將來自相同音素類別的特徵表示拉近，同時迫使不同類別的表示更加分散。此外，我們還設計了一個新穎的分層 APA 模型，命名為 HierCB，建立在新提出的卷積增強分支形成器塊之上，以證明 ConPCO 的有效性。 總結來說，本研究的主要貢獻有：(1) 就我們所知，ConPCO 是首次嘗試使用對比學習來幫助語音模型獲取語音鑑別特徵；(2) 我們進一步開發了一個簡單但有效的階層式語音模型，以驗證所提出方法的可行性



(a) Hierarchical APA Model Based on Convolution-augmented Branchformer (HierCB)

(b) Convolution-augmented Branchformer

圖 2：所提出的階層式 APA 模型架構，建立於新穎的卷積增強型 Branchformer 編碼器區塊。依設計，(a) HierCB 階層式表示 L2 學習者的輸入語句，以及(b)所提出的卷積增強型 Branchformer 區塊。此處，準確性、流暢度、完整性和語調等發音面向分別以 Acc、Flu、Comp 和 Pros 表示。

訓練體制，透過新提出的捲積模組強化 Branchformer 模型 [17]；且(3) 在公開的 APA 資料集上進行廣泛的實驗，證實我們所提出方法的實用性，大幅提升了跨多個語言層次的多重評估效能。

II. 研究方法

A. 對比音素序數正則化器 (ConPCO)

如圖 1(b) 所示，所提出的 ConPCO 正則化器由三個數學項組成：對比項 $\mathcal{L}_{\text{con}}$ 、音素特徵項 $\mathcal{L}_{pc}$ 和序數項 $\mathcal{L}_{o}$. $\mathcal{L}_{\text{con}}$ ，旨在同時將由 APA 模型生成的音素表示和音素層級文本提示的嵌入投射到一個聯合特徵空間。 $\mathcal{L}_{pc}$ 和 $\mathcal{L}_{o}$ 共同調整特徵空間中音素類別間和類別內的距離，前者增強音素間的差異性，後者保持輸出目標的序數關係並同時保持音素內部的緊湊性。對比項。令 $\mathrm{H}^{p}=\left(\mathbf{h}_{1}^{p}, \mathbf{h}_{2}^{p}, \ldots, \mathbf{h}_{N}^{p}\right)$ 代表由 APA 模型的音素編碼器生成的話語音素表示， $\mathrm{E}^{p}=\left(\mathbf{e}_{1}^{p}, \mathbf{e}_{2}^{p}, \ldots, \mathbf{e}_{N}^{p}\right)$ 表示音素層級文本提示的文字嵌入。為獲得一組成對的音素表示 $\mathcal{M}=\left\{\left(\mathbf{z}_{i}^{p}, \mathbf{z}_{i}^{t}\right), i=1, \ldots, M\right\}$ ，我們計算 H^{p} 和 E^{p} 中每個音素類別的質心向量，隨後進行分別的線性投射。在集合 $\mathcal{M}$ 中，計算 $M \times M$ 相似性，對比項 $\mathcal{L}_{\text{con}}$ 旨在同時最大化成對音素表示的相似性並最小化不成對表示的相似性 [18][19]。對比項 $\mathcal{L}_{\text{con}}$ 包含兩個損失，具有溫度超參數 τ ，用於控制負樣本的懲罰強度：

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

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

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

其中 $\phi\left(\mathbf{z}_i^p, \mathbf{z}_j^t\right)$ 表示 ℓ_2 -歸一化向量 $\mathbf{z}_i^p$ 和 $\mathbf{z}_j^t$ 之間的點積（即餘弦相似性）。在訓練階段，集合 $\mathcal{M}$ 由每個批次構建，我們經驗性地對具有最高熟練度分數的數據實例進行採樣以計算質心向量。
音素特徵術語。音素特徵術語 $\mathcal{L}_{pc}$ 透過最小化質心向量 $\mathbf{z}_i^p$ 之間的負距離，保留音素接近性資訊：

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

其中 $\mathcal{L}_{pc}$ 相當於在優化過程中最大化音素類別之間的距離。
序數術語。為了反映迴歸目標在特徵空間中的序數關係，定義序數術語 $\mathcal{L}_o$ 以最小化特徵表示 $\mathbf{h}_i^p$ 與其對應音素質心向量 $\mathbf{z}_i^p$ 之間的距離，同時考慮熟練度分數的相對差異：

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

其中 $w_i = |C - y_i^p|$ 是每個 $\mathbf{h}_i^p$ 的緊湊性權重，反映標籤空間內的序數行為， y_i^p 表示對應的音素層級準確度分數。可調整的常數 C 設定為 3，代表最高的音素層級熟練度分數。

B. 基於卷積增強分支型網絡的階層式 APA 模型（HierCB）

我們提出的 APA 模型的整體架構如圖 2(a) 所示，由三個主要元件組成：音素級建模、詞彙級建模和語言級建模。在不同建模階段的每個編碼器採用新穎的卷積增強分支型網絡區塊，如圖 2(b) 所示。
卷積增強分支型網絡。歸功於可變範圍的上下文建模能力，分支型網絡 [17] 比其他先進的神經網絡模型 [20][21] 更適合構建階層式 APA 模型。分支型網絡區塊由兩個平行分支組成，其中一個分支通過多頭自注意力（MHA）模組捕捉全域上下文，而另一個分支則通過帶有卷積門控機制（cgMLP）[22] 的多層感知器模組學習局部上下文。為了有效表示更低粒度（即音素和詞彙單元）的發音特徵，所提出的卷積增強分支型網絡區塊將原始架構中的 cgMLP 層替換為卷積模組。如圖 2(b) 所示，我們的卷積模組以包括點態卷積和門控線性單元函數（GLU）[23] 的門控機制開始。在上述模組之上堆疊了一個核心大小為 3 的一維深度卷積層，用於捕捉局部資訊，然後通過圖層標準化並由修正線性單元（ReLU）函數激活。之後應用另一個點態卷積，並加入丟棄層以防止過擬合。另一個分支保留 MHA 模組。整個區塊被構建為殘差網絡，這兩個分支通過加權平均操作進一步融合 [17]。
階層式 APA 建模。對於二語學習者的輸入話語，我們首先提取各種發音特徵，以描繪其在音素層級的發音品質，然後將這些特徵串接並投影，以獲得一系列濃縮的聲學特徵 X^p 。特徵擷取過程可以表示為：

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

其中 $\mathrm{Linear}_p(\cdot)$ 是一個線性層， E^{GOP} 是基於發音優良度（GOP）的特徵 [24]， $\mathrm{E}^{\mathrm{Dur}}$ 和 $\mathrm{E}^{\mathrm{Eng}}$ 是持續時間和能量統計的韻律特徵 [25][26]，而 E^{SSL} 是基於自監督學習（SSL）的特徵 [9]。接著，我們將音素層級的文字嵌入 E^p 加入 X^p ，並採用音素編碼器來獲得方面表示 H^p ：

$$\mathrm{H}_0^p = \mathrm{X}^p + \mathrm{E}^P$$
$$\mathrm{H}^p = \mathrm{PhnEnc}\left(\mathrm{H}_0^p\right)$$

在此， E^P 是通過將音素層級的文字提示傳遞到音素和位置嵌入層而生成的，而音素編碼器 PhnEnc(.) 是由 3 個增強卷積的 Branchformer 區塊堆疊而成。接下來，在 H^p 之上建立回歸頭，以存取音素準確度分數。

對於詞彙層級的評估，我們首先使用詞彙層級的注意力池化，從其組成音素中衍生出詞彙表示向量，這是通過一維深度卷積層實現，並接著使用具有平均運算的多頭注意力（MHA）層。詞彙層級的輸入表示 X^w 是通過分別將 X^p 和 H^p 傳遞到詞彙層級注意力池化而計算的。然後，結果表示透過線性投影打包在一起：

$$\widehat{\mathrm{X}}^{w} = \mathrm{AttPool}_{\mathrm{w}_1}\left(\mathrm{X}^{p}\right),$$
$$\widehat{\mathrm{H}}^{w} = \mathrm{AttPool}_{\mathrm{w}_2}\left(\mathrm{H}^{p}\right),$$
$$\mathrm{X}^{w} = \mathrm{Linear}_{w}\left(\left[\widehat{\mathrm{X}}^{w}; \widehat{\mathrm{H}}^{w}\right]\right).$$

字詞層級的文字嵌入 E^{w} 被添加到 X^{w} ，並採用詞彙編碼器生成字詞層級的上下文表示 H^{w} ：

$$\mathrm{H}_{0}^{w} = \mathrm{X}^{w} + \mathrm{E}^{w}$$
$$\mathrm{H}^{w} = \mathrm{WordEnc}\left(\mathrm{H}_{0}^{w}\right)$$

其中 E^{w} 通過詞彙和位置嵌入層將文字提示映射到對應的嵌入序列，且 WordEnc(•) 由 2 個增強卷積的 Branchformer 塊組成。最後，在 H^{w} 上執行三個不同的一維深度卷積層，以生成字詞層級的方面表示（即 $\mathrm{H}^{w_1}, \mathrm{H}^{w_2}$ 和 H^{w_3} ），然後透過相應的字詞層級回歸頭將其轉換為發音分數序列。

對於話語層級的評估，我們首先使用加權平均合併 $\mathrm{H}^{w_1}, \mathrm{H}^{w_2}$ 和 H^{w_3} 以獲得字詞層級的輸出表示 $\overline{\mathrm{H}}^{w}$ 。然後，一維深度卷積層分別疊加在 $\mathrm{X}^{p}, \mathrm{H}^{p}$ 和 $\overline{\mathrm{H}}^{w}$ 上。結果輸出經由線性投影進一步組合，形成話語層級輸入表示的序列 H_{0}^{u} 。接著，利用話語編碼器生成話語層級的上下文表示 H^{u} ：

$$\overline{\mathrm{H}}^{w} = \mathrm{Merge}\left(\mathrm{H}^{w_1}, \mathrm{H}^{w_2}, \mathrm{H}^{w_3}\right),$$
$$\mathrm{H}_{0}^{u} = \mathrm{Linear}_{u}\left(\left[\mathrm{DC}_1\left(\mathrm{X}^{p}\right); \mathrm{DC}_2\left(\mathrm{H}^{p}\right); \mathrm{DC}_3\left(\overline{\mathrm{H}}^{w}\right)\right]\right),$$
$$\mathrm{H}^{u} = \mathrm{UttEnc}\left(\mathrm{H}_{0}^{u}\right),$$

其中 $\mathrm{Merge}(\cdot)$ 是一個加權平均操作， $\mathrm{UttEnc}(\cdot)$ 是一個單一的卷積增強分支前饋塊，而 $\mathrm{DC}_1(\cdot)$ 、 $\mathrm{DC}_2(\cdot)$ 和 $\mathrm{DC}_3(\cdot)$ 是不同的一維深度卷積層，每個層的卷積核大小為 3。最後，接著五個獨立的注意力池化模組，每個模組都有不同的回歸頭，用於推導相應的話語級發音分數。
訓練目標。多面向與多粒度的發音評估訓練目標 $\mathcal{L}_{\mathrm{APA}}$ ，是透過從不同粒度層級蒐集的均方誤差（MSE）損失所計算的加權總和。此外，ConPCO 正則化器被納入最佳化過程中：

$$\mathcal{L} = \mathcal{L}_{APA} + \mathcal{L}_{ConPCO}$$

表一：speechocean762 語料庫統計資訊

粒度	方面	分數區間	計數次數	
			訓練	測試
語音音素	準確度	[0, 2]	47,076	47,369
單字	準確性 壓力 總計	在零到十之間	15,849	15,967
語音	準確度 完整性 流暢度 語調 總分	在零到十之間	2,500	2,500

$$\mathcal{L}_{APA} = \frac{1}{N_p}\sum_{j_p}\mathcal{L}_{p^j p} + \frac{1}{N_w}\sum_{j_w}\mathcal{L}_{w^j w} + \frac{1}{N_u}\sum_{j_u}\mathcal{L}_{u^j u}$$
$$\mathcal{L}_{\mathrm{ConPCO}} = \lambda_{\mathrm{con}}\mathcal{L}_{\mathrm{con}} + \lambda_{pc}\mathcal{L}_{pc} + \lambda_{o}\mathcal{L}_{o}$$

其中 $\mathcal{L}_{p^j p}, \mathcal{L}_{w^j w}$ 和 $\mathcal{L}_{u^j u}$ 分別是在不同層面上的音素級、詞彙級和語句級損失； N_p, N_w 和 N_u 分別標記音素、詞彙和語句層面的特徵數量；而 $\lambda_{con}, \lambda_{pc}$ 和 λ_o 是可調節的參數，用於控制不同數學項的影響力。

三、實驗

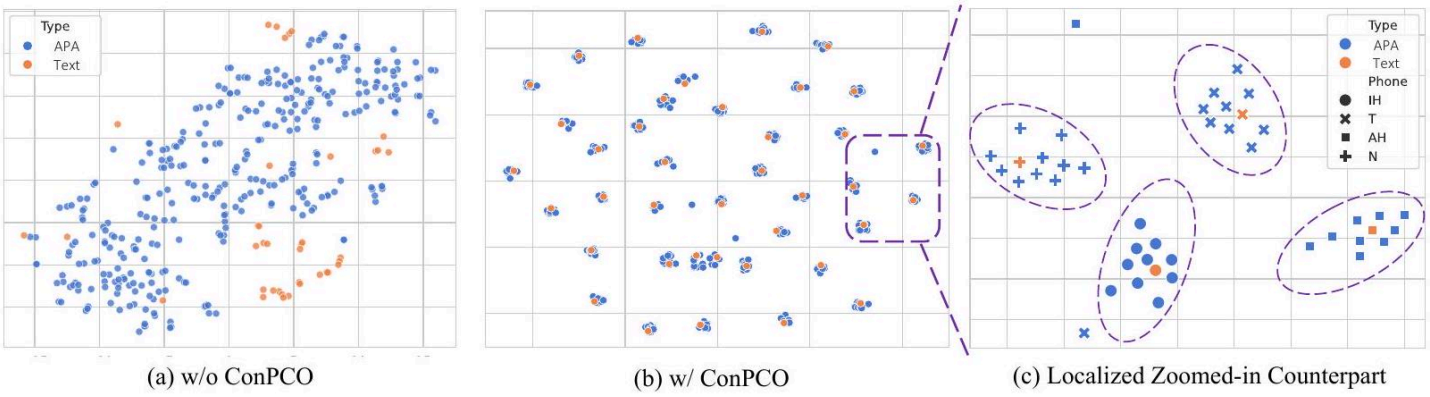
A. 實驗設置

資料集。我們在 speechocean762 資料集上進行了自動發音評估（APA）實驗，這是一個專門為 APA 研究設計的公開可用資料集[27]。該資料集包含 250 位以普通話為第二語言的學習者所錄製的 5,000 個英語語音。訓練集和測試集大小相等，各有 2,500 個語音樣本，並在多個語言粒度和不同層面上評估發音熟練度。表 I 總結了實驗資料集的統計資訊。
實施細節。對於輸入特徵擷取，能量和持續時間統計資訊遵循[25][26]中建議的處理流程。基於自監督學習（SSL）的特徵是從三個預訓練的聲學模型中擷取，包括 Wav2vec2.0 [28]、WavLM [29]和 HuBERT [30]，特徵來自最後一層的輸出[9]。這些擷取的幀級特徵隨後根據音素時間戳彙總為音素級特徵。編碼器中使用的多頭注意力層和注意力池化機制採用 1 個注意力頭和 24 個隱藏單元。關於訓練配置，我們遵循[16]中報告的設定，進行 5 次獨立試驗，每次試驗包含 100 個世代，並使用不

同的隨機種子。在本研究中，主要評估指標是皮爾森相關係數（PCC），用於衡量預測分數和真實分數之間的線性相關性。我們報告了音素級準確性的均方誤差（MSE）值。實驗的詳細資訊可在 https://github.com/bicheng1225/ConPCO 上查看。

B. 實驗結果

異質編碼器生成的語音音素表示的定性視覺化。為了定性地檢驗 ConPCO 是否將從 APA 模型獲得的語音音素表示與音素級文本提示的文字嵌入對齊，我們從所提出的 HierCB 中獲取學習到的表示 H^p 和 E^p ，然後使用 t-SNE 演算法進行投影和視覺化，如圖 3 所示。比較圖 3(a)和圖 3(b)，我們可以觀察到所提出的正則化器有效地將這兩種類型的語音音素表示投影到共享的特徵空間，呈現出更加一致的分佈。更進一步，圖 3(c)中給出了放大的視圖，說明異質特徵表示通過根據其音素類別聚集特徵，有效地保留了音素特徵。
在較細微粒度上的效能評估。表二呈現了在音素和字詞層級粒度的評估結果，分為兩個部分：第一部分報告了



(a) w/o ConPCO (b) w/ ConPCO (c) Localized Zoomed-in Counterpart

圖 3：使用所提出的階層式 APA 模型生成的語音單位表徵的視覺化，其中藍色和橙色點分別是從語音單位編碼器 H^p 的輸出和語音單位層級文字嵌入 E^p 所得到的投影特徵表徵。

本文展示了僅依賴 GOP 特徵的 APA 模型結果，第二部分呈現了整合 SSL 基礎特徵作為輸入的模型結果。為了進行公平比較，我們使用方程式(6)所示的特徵擷取流程重現 GOPT 模型，並採用輸入特徵 X^p 作為模型輸入（GOPT-SSL）。我們還報告了 HierCB 的一個變體，其中編碼器層採用 Branchformer 區塊（HierBFR），並提出了一個音素級正則化器 PCO，這是 ConPCO 的特殊情況，其中方程式(19)中的參數設置為（$\lambda_{\mathrm{con}}, \lambda_{pc}, \lambda_o$）變為（$0, 1, 1$）。

首先，一個普遍觀察是，所提出的 HierCB 模型在音素與詞彙層級的發音評估各方面，表現均優於先前提出的 APA 模型。具體而言，相較於 Gradformer (GFR)、3M 及 GOPT-SSL 模型，HierCB 在 PCC 分數上分別實現了最高達 $2.95\%, 3.75\%$ 與 3.28% 的平均提升幅度。關於音素層級的正則化器 PCO 與 ConPCO，我們觀察到兩者皆能提升 HierCB 的音素準確度，並為詞彙層級評估帶來額外效益。值得注意的是，我們的 ConPCO 方法能顯著提升 HierCB 的表現，達成最低音素層級均方誤差，並在各項評估任務中展現最佳效能。其次，得益於大規模預訓練聲學模型，GOPT-SSL 在各類發音評估任務中持續超越當前主流方法（包括 GOPT 與 LSTM）。此外，強基線方法 3M 在多數評估任務中表現優於 GOPT-SSL，唯獨在詞級重音評估中例外。可能的解釋在於，3 M 在神經建模中整合了特定語音線索（如元音與輔音嵌入），這些線索在音素層級建模中提供細緻資訊，並透過平行 APA 模型架構進一步提升詞彙層級評估的效能。最後，相較於分層式 APA 模型（HierCB、HierBFR 與 HiPAMA），HiPAMA 表現遜於後兩者，而 HierCB 則展現優於 HierBFR 的性能。相較於 HiPAMA，HierCB 與 HierBFR 皆受益於基於 SSL 的特徵，透過 Branchformer 的雙分支架構，利用超音段發音線索並從不同語句粒度中提取獨特特徵。此外，在所有評估任務中，所提出的 HierCB 表現始終優於 HierBFR。 此優勢可歸功於我們策略性設計的卷積增強式分支形成器模組，該模組透過一系列卷積層有效建模細微發音特徵。
話語層級效能評估。表三報告了話語層級發音評估的實驗結果。所提出的 HierCB 在大多數評估中都優於或與具有平行架構的 APA 模型相當，特別是在完整性評估中分別為 3 M 和 GOPT-SSL 提升了 0.352 和 0.387。我們將話語完整性評估的性能提升歸因於所提出的捲積增強 Branchformer 區塊，它有效地捕捉了局部和全域發音線索，同時動態調節不同粒度的組合權重。透過音素層級正則化器 ConPCO，HierCB 相較於基礎模型持續提升效能，尤其是在話語層級

表二：各方法的實驗結果
音素和詞彙層級的發音評估

輸入特徵	模型	音素分數		字詞分數（PCC）		
		MSE	PCC	準確率	重音	總計
GOP	長短期記憶網絡 [8]	0.089	0.591	0.514	0.294	0.531
	GOPT [8]	0.085	0.612	0.533	0.291	0.549
	GFR [6]	0.079	0.646	0.598	0.334	0.614
	HiPAMA [10]	0.084	0.616	0.575	0.320	0.591
SSL	GOPT-SSL	0.081	0.640	0.584	0.352	0.603
	3M [9]	0.078	0.656	0.598	0.289	0.617
	HierBFR	0.082	0.639	0.591	0.300	0.609
	HierCB	0.076	0.680	0.630	0.355	0.645
	+PCO [16]	0.078	0.688	0.648	0.347	0.622
	+ConPCO	0.071	0.701	0.669	0.437	0.682

表3：各種方法在語句級發音評估的實驗結果

模型	語句分數（皮爾森相關係數）				
	準確率	詞彙	流暢度	語調	總分
3M [9]	0.760	0.325	0.828	**0.827**	0.796
GOPT-SSL	0.748	0.290	0.817	0.807	0.778
階層式類別平衡模型	0.772	0.677	0.827	0.822	0.796
+ 對比序數正則化模型	**0.780**	**0.749**	**0.830**	0.823	**0.803**

注意：準確率和完整性分別簡寫為 Acc. 和 Comp.
這個結果突顯了在自動發音評估（APA）模型中保留語音特徵的重要性，這對於與發音清晰度相關的發音評估很有幫助。

四、結論

在本論文中，我們提出了一種新穎的訓練機制 ConPCO，旨在學習語音感知的表徵，同時在學習的特徵空間中保留回歸目標之間的序數關係。此外，我們還開發了一個分層的自動發音評估模型，以驗證所提出的正則化器的有效性。我們的方法的實用性已通過 speechocen762 基準資料集上的廣泛實驗得到驗證。
局限性與未來工作。所提出的方法僅限於「朗讀」學習場景，在某種程度上缺乏對所提供評估結果的可解釋性。在未來的工作中，我們計劃在開放性回應場景中檢驗所提出的方法，在這些場景中，學習者可以自由地說話或回應給定的任務或問題 [31][32]。此外，可解釋的發音回饋問題也留作未來的擴展。

五、 致謝

本研究部分由玉山銀行透過編號為 202408-NTU-02 的補助支持。論文中的任何發現與意涵不一定代表贊助單位的觀點。

參考文獻

[1] P. M Rogerson-Revell，「電腦輔助發音訓練（CAPT）：當前議題與未來方向」，RELC 期刊，第 52 卷，第 189-205 頁，2021 年。
[2] A. V. Moere 和 R. Downey，「語言評量的科技與人工智慧」，第二語言評量手冊，第 341-358 頁，2016 年。
[3] M. Eskenazi，「教育中口語語言科技的概述」，語音通訊，第 51 卷，第 832-844 頁，2009 年。
[4] S. Bannò、B. Balusu、M. Gales、K. Knill 和 K. Kyriakopoulos，「第二語言英語口語的特定視角評估」，收錄於語音處理會議論文集（INTERSPEECH），第 4471-4475 頁，2022 年。
[5] N. F. Chen 和 H. Li，「電腦輔助發音訓練：從發音評分到口語學習」，收錄於亞太信號與資訊處理協會年會暨研討會（APSIPA ASC）論文集，第 17 頁，2016 年。
[6] 裴鴻超、方海、羅曉及徐小松，「Gradformer：一個多面向、多粒度發音評估架構」，收錄於《IEEE/ACM 音訊、語音及語言處理期刊》，第 32 卷，第 554-563 頁，2024 年。
[7] 顏柏全、李俊廷、王育成、王弘偉、羅天豪、許育祺、趙偉誠及陳柏，「運用階層式 Transformer 與預訓練策略的有效發音評估方法」，收錄於《計算語言學協會會議論文集》（ACL），第 1737-1747 頁，2024 年。
[8] 龔毅、陳澤、朱以弘、張鵬及葛拉斯，「基於 Transformer 的多面向、多粒度非母語英語發音評估」，收錄於《IEEE 國際聲學、語音及信號處理會議》（ICASSP），第 7262-7266 頁，2022 年。
[9] 趙方安、羅天豪、吳彤一、宋雨庭及陳柏，「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」，收錄於《亞太訊號與資訊處理協會年會暨研討會》（APSIPA ASC），第 575-582 頁，2022 年。
[10] H. Do、Y. Kim 和 G. G. Lee，「具多面向注意力的階層式發音評估」，收錄於 IEEE 國際聲學、語音與信號處理會議（ICASSP）論文集，第 1-5 頁，2023 年。
[11] Y. Yang、K. Zha、Y. Chen、H. Wang 和 D. Katabi，「深入研究不平衡迴歸」，收錄於機器學習國際會議（PMLR）論文集，第 11842-11851 頁。
[12] P. C. English、J. Kelleher 和 J. Carson-Berndsen，「針對語音特徵對 wav2vec 2.0 嵌入進行領域知情探測」，收錄於語

音學、語音學和形態學計算研究研討會（SIGMORPHON）論文集，第 83-91 頁，2022 年。
[13] V. Zouhar、K. Chang、C. Cui、N. B. Carlson、N. R. Robinson、M. Sachan 和 D. R. Mortensen，「PWESuite：語音詞嵌入及其促進的任務」，收錄於計算語言學、語言資源與評估聯合國際會議（LREC-COLING）論文集，第 13344-13355 頁，2024 年。
[14] 傅志，周衛，徐佳，周浩，以及李蕾。「超越遮罩語言模型的語境表徵學習」，收錄於計算語言學協會年會論文集（ACL），第 2701-2714 頁，2022 年。
[15] A. 博拉，M. P. 巴曼，以及 A. 阿威卡爾，「詞嵌入方法是否穩定？我們是否應該在意？」，收錄於超文本與社交媒體 ACM 會議論文集，第 45-55 頁，2021 年。
[16] 顏伯蒼、王鴻文、王奕誠、李建廷、林志宏，以及陳柏，「保留語音特徵以進行序數迴歸：自動發音評估的新型損失函數」，收錄於 IEEE 自動語音辨識與理解研討會（ASRU）論文集，第 1-7 頁，2023 年。
[17] 彭毅，S. 達爾米亞，I. 蓮，以及 S. 渡邊。「分支前饋網絡：捕捉語音辨識與理解中的局部與全域脈絡的平行 MLP-注意力架構」，收錄於機器學習國際會議論文集（PMLR），第 162 卷，第 17627-17643 頁，2022 年。
[18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, 和 Ilya Sutskever，「從自然語言監督學習可遷移的視覺模型」，收錄於機器學習國際會議論文集（PMLR），第 139 卷，第 8748-8763 頁，2021 年。
[19] B. Elizalde, S. Deshmukh, M. A. Ismail, 和 H. Wang，「CLAP：從自然語言監督學習音訊概念」，收錄於 IEEE 國際聲學、語音和信號處理會議（ICASSP）論文集，第 1-5 頁，2023 年。
[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N Gomez, Ł. Kaiser, 和 I. Polosukhin，「注意力就是一切」，收錄於神經信息處理系統會議（NeurIPS）論文集，第 5998-6008 頁，2017 年。
[21] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, 和 R. Pang，「Conformer：用於語音識別的卷積增強轉換器」，收錄於語音技術會議（INTERSPEECH）論文集，第 5036-5040 頁，2020 年。
[22] J. Sakuma、T. Komatsu 和 R. Scheibler，「用於自動語音辨識的可變長度輸入多層感知器架構」，arXiv 預印本 arXiv:2202.08456，2022 年。
[23] Y. N. Dauphin、A. Fan、M. Auli、D. Grangier，「使用門控卷積網路進行語言建模」，收錄於機器學習國際會議論文集（PMLR），第 70 卷，第 933-941 頁，2017 年。
[24] S. M. Witt 和 S. J. Young，「互動式語言學習中的音素層級發音評分與評估」，《語音通訊》，第 30 卷，第 95-108 頁，2000 年。
[25] C. Zhu、T. Kunihara、D. Saito、N. Minematsu、N. Nakanishi，「日本英語學習者口語發音的可理解性自動預測」，收錄於 IEEE 口語語言技術研討會（SLT），第 1029-1036 頁，2022 年。
[26] 楊嵩、阿亞諾、塩崎大輔、村上信明和齋藤和夫，「基於可解釋網絡的 L2 英語流暢度預測最佳化：結合發音量和發音質量」，發表於 IEEE 口語技術研討會（SLT），第 698-704 頁，2021 年。
[27] 張俊、張智、王毅、燕子、宋強、黃毅、李凱、大衛·波維和王豔，「SpeechOcean762：用於發音評估的開源非母語英語語音語料庫」，發表於語音交互會議（INTERSPEECH），第 3710-3714 頁，2021 年。
[28] 巴耶夫斯基、周浩、穆罕默德和奧利，「Wav2vec 2.0：語音表徵自監督學習框架」，發表於神經信息處理系統國際會議（NIPS），第 12449-12460 頁，2020 年。
[29] 陳思等，「WavLM：面向全棧語音處理的大規模自監督預訓練」，IEEE 信號處理精選專題期刊，第 1505-1518 頁，2022 年。
[30] 許偉年等人，「HuBERT：通過隱藏單元遮蔽預測進行自監督語音表示學習」，IEEE/ACM 語音、音訊與語言處理交易期刊，第 3445-3460 頁，2021 年。
[31] 王毅等人，「邁向自發性英語口語的自動評估」，《語音通訊》，第104卷，第47-56頁，2018年。
[32] 朴智賢和崔星浩，「解決端到端自動語音評分的冷啟動問題」，收錄於語音交互會議（INTERSPEECH）論文集，第 994-998 頁，2023 年。

通訊作者。