CTC-Adapted Multi-Layer SSL Representations to Complement GOP for Taiwanese Hakka Pronunciation Assessment
CTC 適應之多層自監督語音表徵作爲 GOP 補充資訊的臺灣客語發音評估

Jian-Hua Wang
National Pingtung University
Pingtung, Taiwan
wangmahua0301@gmail.com

Yi-Chin Huang
National Pingtung University
Pingtung, Taiwan
ychuangnptu@mail.nptu.edu.tw

摘要

臺灣客語可用的專家評分語料有限，利用既有語音表徵改善發音評分是低資源語言的重要研究問題。本研究以 Sixian-Reading 語料的客語音素辨識 CTC 任務適應 XLSR-53，再凍結編碼器參數，比較自監督學習（SSL）表徵是否能爲發音良好度（GOP）特徵提供互補資訊。比較適應前後的最後層、等權平均與可學習層加權表徵。評估使用 19,668 筆認證作答、2,809 個考生識別碼，保留考生互斥的固定測試集，其餘資料進行五組訓練／驗證輪替，並使用五個訓練種子。所有條件使用相同的 Transformer 評分器架構，並分別訓練模型參數。CTC 適應在三種層聚合方式下皆提升 SSL 單獨評分的效能，表現仍低於 GOP。與 GOP 融合後，可學習層加權取得最高平均 PCC 0.809，GOP 基準爲 0.793；其平均 MSE 與 MAE 分別爲 0.424 與 0.512。

Abstract

Expert-rated speech data for Taiwanese Hakka are limited, motivating the use of existing speech representations for pronunciation assessment. We adapt XLSR-53 on the Sixian-Reading corpus using a phoneme-level CTC objective and freeze the encoder during downstream scorer training. We examine whether self-supervised learning (SSL) representations provide complementary information to goodness of pronunciation (GOP). Pre-trained and adapted representations are compared using the final layer, uniform averaging, and learnable layer weighting. Evaluation uses 19,668 certification responses associated with 2,809 candidate identifiers, a fixed candidate-disjoint test set, five train/validation rotations, and five training seeds. All conditions use the same Transformer scorer architecture with separately trained parameters. CTC adaptation improves SSL-only assessment under all three aggregation settings, although performance remains below GOP. When fused with GOP, learnable layer weighting achieves the highest mean PCC of 0.809, compared with 0.793 for GOP, with mean MSE and MAE of 0.424 and 0.512, respectively.

關鍵字：臺灣客語發音評估、發音良好度、自監督語音表徵、CTC 適應、多層表徵融合

Keywords: Taiwanese Hakka pronunciation assessment, goodness of pronunciation, self-supervised speech representations, CTC adaptation, multi-layer fusion

1 緒論

自動發音評估可爲語言學習提供可重複的評分結果 (El Kheir et al., 2023; Witt and Young, 1998)。臺灣客語可用的專家評分資料有限，從既有聲學模型與語音預訓練模型取得可轉移的發音資訊，有助於建立低資源語言的評分方法。既有客語研究以時間延遲神經網路（time-delay neural network, TDNN）產生音素分數，搭配 Transformer 進行句級發音評分 (Huang et al., 2025)。本研究沿用其發音良好度（goodness of pronunciation, GOP）特徵，評估加入自監督學習（self-supervised learning, SSL）表徵後的效能。

GOP 描述聲學模型對參考音素與候選音素的支持程度；SSL 模型則由大量未標註語音學得上下文表徵。兩者的表示方式不同，SSL 表徵可能提供 GOP 未充分描述的聲學資訊。wav2vec 2.0 與 XLSR 提供可轉移的語音表徵 (Baevski et al., 2020; Conneau et al., 2021)。

既有研究已將 SSL 表徵用於發音評分 (Kim et al., 2022)。對臺灣客語而言，仍需釐清在相

Proceedings of the 38th Conference on Computational Linguistics and Speech Processing (ROCLING 2026), pages 246–251
November 6-7, 2026 ©2026 Association for Computational Linguistics

同層聚合條件下，音素辨識適應對評分效能的影響，以及融合 GOP 後的改善是否在不同訓練種子下維持一致。因此，本研究以客語音素辨識的連接時序分類（Connectionist Temporal Classification, CTC）適應 XLSR-53，使用相同的下游評分器架構，比較適應前後的最後層、24 層等權平均及可學習層加權表徵，並評估與 GOP 融合後的評分效能。評估採考生互斥的固定測試集、五組訓練／驗證輪替與五個訓練種子，並以考生層級誤差檢查模型差異。

2 相關研究

GOP 類方法利用聲學模型衡量音素發音品質 (Witt and Young, 1998)。後續研究將深層聲學特徵用於發音評分 (Lin and Wang, 2021)，GOPT 則使用 Transformer 建模 GOP 序列，進行多面向與多粒度評估 (Gong et al., 2022)。在臺灣客語上，Huang et al. (2025) 使用 TDNN 產生的 68 維 GOP 特徵搭配 Transformer 進行句級評分。本研究沿用相同的 GOP 特徵定義，作爲 SSL 表徵與特徵融合的比較基準。

SSL 已應用於發音錯誤偵測與評估 (Xu et al., 2021; Kim et al., 2022)。Kim et al. (2022) 先以 CTC 適應 SSL 模型，再利用不同 Transformer 層的表徵，結合文本資訊與雙向長短期記憶網路進行發音評分。語音辨識適應與層選擇因此成爲 SSL 發音評分的重要比較條件。Chiu et al. (2024) 於 SUPERB 語音辨識任務比較可學習層選擇與多個 SSL 模型的表徵融合。

近期研究亦整合 GOP 與 SSL 資訊。JCAPT 依強制對齊的音素邊界聚合 SSL 表徵，再與 GOP 結合 (Yang et al., 2025)；HiPPO 使用以 CTC 爲基礎的 GOP 與 SSL 特徵進行階層式發音評估 (Yan et al., 2025)。這些研究提供特徵整合與層聚合的設計依據；本研究在臺灣客語上，固定下游評分器架構與層聚合條件，直接比較 CTC 適應前後的表徵及其融合效能。

3 研究方法

整體流程如圖 1 所示。語音經 TDNN 與 XLSR-53 取得 GOP 及 SSL 表徵。評分器依實驗條件使用單一或融合表徵，結合參考音素與位置資訊預測句級分數。

3.1 GOP 與 SSL 特徵

令參考音素序列爲 $P = (p_1, \ldots, p_n)$。參考音素序列沿用既有 GOP 資料，評分時需提供語音與對應的參考序列。固定的 TDNN 產生音素層級 GOP 特徵 $G_t \in \mathbb{R}^{68}$，由 34 維音素後驗機率對數值（log-phone posterior, LPP）與 34 維對數後驗機率比值（log-posterior ratio, LPR）串接，依參考音素位置排列 (Huang et al., 2025)。

SSL 表徵由 XLSR-53 取得；該模型有 24 層 Transformer，每層輸出 1,024 維。原始預訓練編碼器記爲 PT（pretrained）；經客語音素辨識的 CTC 適應 (Graves et al., 2006) 後記爲 FT（fine-tuned）。下游評分使用 Transformer 層的輸出，編碼器參數在評分訓練時保持固定。

第 l 層音框序列經自適應平均池化（adaptive average pooling），調整爲參考音素數量 n，得到 $Z^{(l)} = (z_1^{(l)}, \ldots, z_n^{(l)})$。此處按序列索引與 GOP 配對，未使用音素起訖時間，故池化位置可能與實際音素邊界有偏差。最後層條件使用 $Z^{(24)}$；多層條件使用 $Z^{(1)}, \ldots, Z^{(24)}$。

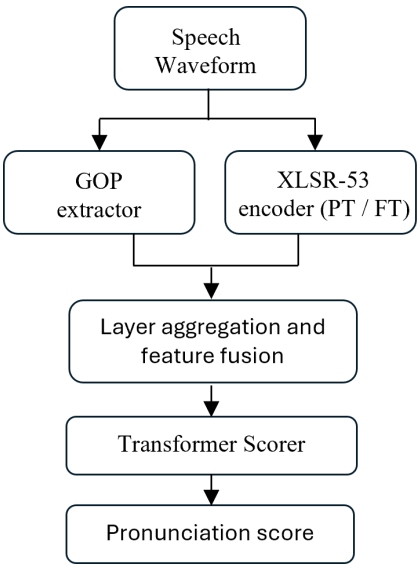

圖 1. TDNN 產生 GOP 特徵，XLSR-53 提供 SSL 表徵；依實驗條件使用單一表徵或融合，由 Transformer 評分器輸出句級發音分數。各條件使用相同架構，參數分別訓練。評分器另使用參考音素及位置嵌入；特徵標準化與多層聚合見第 3.2 節。

3.2 特徵標準化與多層聚合

每組訓練／驗證輪替對 GOP 及各種 SSL 表徵分別以訓練資料估計一組純量平均值 μ_b 與標準差 σ_b，其中 b 表示該次使用的輸入表徵。統計涵蓋訓練語句的有效音素位置與全部特徵維度，補零位置不參與估計。特徵元素 x 的標準化爲

$$\widetilde{x} = \frac{x - \mu_b}{\sigma_b}. \tag{1}$$

GOP 的統計量涵蓋有效位置的全部 68 維特徵；驗證與測試資料沿用該次輪替的訓練統計，五個種子共用同一組統計。

等權平均（uniform averaging, AVG）先對原始表徵取平均：

$$z_t^{\mathrm{AVG}} = \frac{1}{24}\sum_{l=1}^{24} z_t^{(l)}. \tag{2}$$

再以平均後的 1,024 維 SSL 特徵估計平均值與標準差。可學習層加權（learnable layer-weighted sum, LWS）則先以完整 24 層堆疊估計一組 SSL 統計，標準化各層後加權：

$$\begin{aligned} w_l &= \frac{\exp(\alpha_l)}{\sum_{k=1}^{24}\exp(\alpha_k)}, \\ s_t^{\mathrm{LWS}} &= \sum_{l=1}^{24} w_l \widetilde{z}_t^{(l)}. \end{aligned} \tag{3}$$

24 個可訓練係數 α_l 初始化爲零，權重由下游評分損失學習，對所有語句與位置共用。

3.3 融合與 Transformer 評分器

單一表徵模型使用標準化後的 GOP 或 SSL。最後層與 AVG 融合模型將 GOP 與 SSL 向量串接：

$$F_t = [\widetilde{G}_t; s_t] \in \mathbb{R}^{1092}, \tag{4}$$

其中 s_t 爲標準化的最後層或 AVG。LWS 融合模型則使用 $[s_t^{\mathrm{LWS}}; \widetilde{G}_t]$。僅使用 SSL 的模型仍保留參考音素資訊。

輸入先以線性層投影至 24 維，再加入 34 維參考音素獨熱編碼（one-hot）向量的線性投影與可學習位置嵌入。參考序列的有效音素 ID 爲 2–33，ID 1 用於尾端補零。評分器採 Transformer 架構 (Vaswani et al., 2017)，三個區塊均採前置層正規化（pre-LayerNorm），各含兩個注意力頭及 96 維前饋網路。最後經層正規化（LayerNorm），對有效位置取平均，再由線性分數頭輸出句級分數。輸入序列上限爲 38，本次評分使用的輸入序列均未超過此限制。注意力遮蔽補零位置，輸出未使用 sigmoid 或分數截斷。

輸入維度爲 D 時，輸入投影含 $24D+24$ 個參數，其餘評分器模組合計 23,497 個，總計 $24D + 23,521$。LWS 另加 24 個可訓練係數；凍結的編碼器與標準化統計不計入可訓練參數。

4 實驗設定

4.1 語料與資料切分

發音評分使用客家委員會提供的去識別化客語認證作答語料，共 19,668 筆語音與 2,809 個考生識別碼。每筆由兩位評分者給予 0–4 分，平均後形成間隔 0.5 的九個分數等級。整句人工分數爲預測目標，資料切分與統計分析以考生識別碼分組。

輪替	訓練 筆數／考生	驗證 筆數／考生
1	12,582 ／ 1,797	3,149 ／ 450
2	12,588 ／ 1,799	3,143 ／ 448
3	12,583 ／ 1,797	3,148 ／ 450
4	12,588 ／ 1,798	3,143 ／ 449
5	12,583 ／ 1,797	3,148 ／ 450

表 1. 固定測試集下的五組訓練／驗證輪替；測試集固定爲 3,937 筆作答、562 位考生。

集合	筆數	小時	說話者
訓練	82,652	243.78	124
驗證	9,721	29.47	15
測試	11,465	33.19	17

表 2. Sixian-Reading 的 CTC 適應切分。

以考生爲單位保留固定測試集，共 3,937 筆作答、562 位考生。其餘 15,731 筆作答分成五個互斥子集，每次使用四組訓練、一組驗證，形成五組訓練／驗證輪替。固定測試集在所有實驗中保持不變；各次訓練、驗證與測試的考生互不重疊，比例約爲 64:16:20。

CTC 適應使用 Hakka Across Taiwan（HAT）語料中的四縣腔朗讀資料 Sixian-Reading (Liao et al., 2023)，處理後共 103,838 筆、306.44 小時、156 位說話者。訓練、驗證與測試分別含 124、15、17 位說話者，識別碼互不重疊。

CTC 目標由參考讀音依音節至音素對照表產生，採不含聲調的音素序列。輸出共 34 類：ID 0 爲空白符號（blank）、ID 1 爲靜音符號（SIL）、ID 2–33 爲實際目標音素。起始檢查點爲 `facebook/wav2vec2-large-xlsr-53`，建立 34 類 CTC 輸出頭，連同前端特徵編碼器更新全部參數。音訊以 16 kHz 單聲道輸入。適應訓練共 10 輪，批次大小爲 2，學習率爲 10^{-5}、權重衰減爲 10^{-4}，使用混合精度與種子 42；最佳化器爲 AdamW，前 10% 步數暖身，後續線性降低學習率。第 10 輪的驗證／測試音素錯誤率（phoneme error rate, PER）爲 0.86% ／ 0.85%。計算 PER 時，先以貪婪解碼（greedy decoding）產生音素序列，再將音素替換、刪除與插入的總數除以參考音素總數。此 PER 僅反映 Sixian-Reading 朗讀語料內、以無聲調參考音素爲目標的辨識結果，不代表跨語料或一般客語音素辨識的效能。

模型	可訓練參數	MSE ↓	PCC ↑	MAE ↓
GOP	25,153	0.455 ± 0.0014	0.793 ± 0.0003	0.523 ± 0.0019
PT-SSL	48,097	0.986 ± 0.0107	0.439 ± 0.0099	0.816 ± 0.0051
FT-SSL	48,097	0.559 ± 0.0058	0.737 ± 0.0032	0.589 ± 0.0032
PT-SSL-AVG	48,097	0.861 ± 0.0450	0.541 ± 0.0351	0.753 ± 0.0213
FT-SSL-AVG	48,097	0.514 ± 0.0028	0.763 ± 0.0013	0.565 ± 0.0037
PT-SSL-LWS	48,121	0.852 ± 0.0648	0.546 ± 0.0480	0.744 ± 0.0312
FT-SSL-LWS	48,121	0.515 ± 0.0049	0.763 ± 0.0021	0.563 ± 0.0012
GOP+PT-SSL	49,729	0.480 ± 0.0059	0.779 ± 0.0032	0.542 ± 0.0042
GOP+FT-SSL	49,729	0.442 ± 0.0071	0.799 ± 0.0033	0.520 ± 0.0068
GOP+PT-SSL-AVG	49,729	0.470 ± 0.0020	0.784 ± 0.0010	0.534 ± 0.0012
GOP+FT-SSL-AVG	49,729	0.433 ± 0.0027	0.804 ± 0.0011	0.516 ± 0.0024
GOP+PT-SSL-LWS	49,753	0.457 ± 0.0013	0.792 ± 0.0005	0.524 ± 0.0010
GOP+FT-SSL-LWS	49,753	**0.424±0.0029**	**0.809±0.0016**	**0.512±0.0019**

表 3. 固定測試集上的評估結果。每個訓練種子先平均五組訓練／驗證輪替的測試指標，再報告五個種子的平均值 ± 樣本標準差；平均值保留三位小數，標準差保留四位。所有融合條件的 GOP 與 SSL 特徵分別標準化，粗體爲全表最佳平均值。

模型比較	考生 ΔMAE	r_{rb}	Holm p
GOP+FT-SSL − GOP+PT-SSL	−0.0220	−0.2202	$\mathbf{2.45 \times 10^{-5}}$
GOP+FT-SSL-LWS − GOP	−0.0114	−0.0596	0.4414
GOP+FT-SSL-LWS − GOP+FT-SSL	−0.0075	−0.0512	0.4414
GOP+FT-SSL-LWS − GOP+FT-SSL-AVG	−0.0039	−0.1276	**0.0263**

表 4. 562 位測試考生的雙尾 Wilcoxon 檢定，四組比較一起做 Holm 校正。ΔMAE 爲左側模型減右側模型，負值表示左側模型的 MAE 較低；r_{rb} 負值表示左側模型誤差傾向較低。考生等權差值與語句等權差值可能略有不同。

4.2 下游訓練與效能評估

各模型在五組訓練／驗證輪替中使用訓練種子 42–46，共訓練 25 次；種子控制參數初始化、資料載入與訓練擾動。下游以 Adam 最小化均方誤差（mean squared error, MSE），訓練 50 輪，批次大小爲 32；初始學習率爲 0.001，$(\beta_1, \beta_2) = (0.95, 0.999)$，權重衰減爲 5×10^{-7}，dropout 爲 0。前 100 步爲暖身階段，自第 20 輪起每五輪將學習率減半。訓練時於有效特徵位置加入 $[-0.1, 0.1]$ 的均勻雜訊，依最低驗證 MSE 選擇模型。

效能指標爲 MSE、平均絕對誤差（mean absolute error, MAE）與皮爾森相關係數（Pearson correlation coefficient, PCC）。每個種子先平均五組輪替的測試指標，再對五個種子平均值計算平均與樣本標準差。所有模型使用相同切分與種子，配對比較對應相同輪替與訓練種子。差值與相對變化率均由未四捨五入的指標計算。 AVG 的訓練雜訊加入層平均後的 SSL 特徵；LWS 則在各層 SSL 特徵加入雜訊後加權。兩種設定的 GOP 特徵均只加入一次雜訊。

4.3 考生層級統計比較

本節的 PT、FT、FT-AVG 與 FT-LWS 均指與 GOP 融合的模型。四組事後比較爲 FT 最後層對 PT 最後層、FT-LWS 對 GOP、FT-LWS 對 FT 最後層，以及 FT-LWS 對 FT-AVG。考生層級檢定使用每位考生在 25 次評估中的平均 MAE，共 562 組配對值。使用雙尾 Wilcoxon 符號等級檢定，四組 p 值以 Holm 方法校正 (Holm, 1979)，門檻爲 0.05，並報告等級雙列相關（rank-biserial correlation）效果量 r_{rb}。比較依測試結果選定，作探索性解讀；檢定以已訓練模型爲條件，考生爲配對單位。

5 實驗結果與討論

5.1 CTC 適應與 SSL 單獨評分

表 3 顯示，CTC 適應在三種 SSL 表徵設定下皆提升發音評分效能。最後層表徵的 PCC 由 0.439 提升至 0.737，FT 在各種層聚合條件下的 MSE 與 MAE 也均低於 PT。適應後表徵與 GOP 的效能差距因而縮小。

模型比較	ΔMAE ± SD	改善種子
FT − PT	-0.0218 ± 0.0096	5/5
FT-LWS − GOP	-0.0114 ± 0.0014	5/5
FT-LWS − FT	-0.0078 ± 0.0057	5/5
FT-LWS − FT-AVG	-0.0039 ± 0.0030	4/5

表 5. 融合模型的配對 MAE 差。FT ／ PT 均指與 GOP 融合的版本；每個種子先平均五組輪替，SD 爲五個種子配對差的樣本標準差。ΔMAE 爲左側模型減右側模型，負值表示左側模型的 MAE 較低。差值以未四捨五入的指標計算。

多層聚合進一步改善適應後表徵的評分效能，FT-AVG 與 FT-LWS 的平均 PCC 均爲 0.763。FT-AVG 的三項指標在 25 組配對中均優於 FT 最後層；FT-LWS 與 FT-AVG 的 MSE 與 PCC 差距皆小於 0.001。所有僅使用 SSL 的模型仍低於 GOP。

5.2 融合效能與跨種子變異

三種 PT 融合的平均 MAE 均高於 GOP；對應的 FT 融合則均低於 GOP，且三項平均指標均優於 PT。各組 PT ／ FT 對照具有相同的下游可訓練參數量。

FT-LWS 融合在本次比較中取得最低平均 MSE 與 MAE，以及最高平均 PCC。相對 GOP，MSE 降低 6.76%、PCC 增加 0.0162、MAE 降低 2.17%。在 25 組相同輪替與訓練種子的配對結果中，MSE 與 PCC 均優於 GOP，MAE 有 24 組較低。這些結果描述固定測試集下的訓練穩定性；考生層級 MAE 差異經 Holm 校正後未達顯著（表 4）。FT 最後層融合的平均 MAE 較 GOP 降低 0.0036，五個種子中有兩個的輪替平均 MAE 低於 GOP。

表 5 的配對差顯示，FT-LWS 相對 GOP 的 MAE 在五個訓練種子下均降低；考生間的誤差差異另以表 4 檢驗。

5.3 考生誤差與多層權重

考生層級比較見表 4。FT 最後層融合相對 PT 最後層融合，以及 FT-LWS 相對 FT-AVG 的 MAE 差異，Holm 校正後 p 值分別爲 2.45×10^{-5} 與 0.0263。FT-LWS 相對 GOP 及 FT 最後層融合的差異均未達 0.05 門檻。

各條件 25 個最佳模型的平均層權重中，FT-SSL-LWS 以第 16 層最高（7.545%），融合 GOP 後則以第 7 層最高（6.059%）。兩種輸入設定學得不同的層權重分布。各層共用同一組 SSL 標準化統計，這些權重仍可能受到原始層間尺度差異影響。

5.4 限制

SSL 按參考音素數量池化，時間位置與 GOP 的音素邊界可能不同。AVG 與 LWS 同時改變層權重、標準化階段及聚合後的雜訊強度，所得差異反映整體配置的效果。融合模型的可訓練參數量高於僅使用 GOP 的模型，容量增加也可能影響效能；相同維度的隨機特徵或參數量匹配的 GOP 模型可作爲後續對照。

CTC 的 PER 僅涵蓋同一朗讀語料內的無聲調音素辨識。跨語料辨識效能、CTC 與評分語料的說話者重疊情形，以及兩位評分者的一致性，均需額外資料評估。

6 結論

客語音素辨識的 CTC 適應改善 XLSR-53 表徵的發音評分效能。最後層、等權平均與可學習層加權三種條件下，適應後表徵皆優於原始預訓練表徵，單獨使用仍未超越 GOP。與 GOP 融合後，FT-LWS 取得最高平均 PCC 0.809，並降低 MSE 與 MAE；其 MAE 在五個訓練種子下均改善，考生層級 MAE 差異經 Holm 校正後未達顯著。

後續工作可先統一 AVG 與 LWS 的標準化及訓練雜訊設定，並依音素邊界聚合 SSL 表徵，以檢驗層聚合方式的效益。加入參數量匹配的 GOP 對照，可進一步分析 SSL 資訊與模型容量對效能的影響。語音編碼器亦可延伸至 HuBERT (Hsu et al., 2021) 及加入說話者解耦機制的 ContentVec (Qian et al., 2022)，比較不同表徵對客語發音評分的影響。

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460.

Sheng-Chieh Chiu, Chia-Hua Wu, Jih-Kang Hsieh, Yu Tsao, and Hsin-Min Wang. 2024. Learnable layer selection and model fusion for speech self-supervised learning models. In *Interspeech 2024*, pages 3914–3918.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. In *Interspeech 2021*, pages 2426–2430.

Yassine El Kheir, Ahmed Ali, and Shammur Absar Chowdhury. 2023. Automatic pronunciation assessment—a review. In *Findings of the Association for Computational Linguistics: EMNLP*

2023, pages 8304–8324, Singapore. Association for Computational Linguistics.

Yuan Gong, Ziyi Chen, Iek-Heng Chu, Peng Chang, and James Glass. 2022. Transformer-based multi-aspect multi-granularity non-native English speaker pronunciation assessment. In *ICASSP 2022 – IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7262–7266.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 369–376.

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Yi-Chin Huang, Yu-Heng Chen, Chih-Chung Kuo, Chao-Shih Huang, and Yuan-Fu Liao. 2025. A GOP-based automatic pronunciation scoring system for Taiwanese Hakka using Transformer regression models. In *2025 28th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6. IEEE.

Eesung Kim, Jae-Jin Jeon, Hyeji Seo, and Hoon Kim. 2022. Automatic pronunciation assessment using self-supervised speech representation learning. In *Interspeech 2022*, pages 1411–1415.

Yuan-Fu Liao, Shaw-Hwa Hwang, You-Shuo Chen, Han-Chun Lai, Yao-Hsing Chung, Li-Te Shen, Yen-Chun Huang, Chi-Jung Huang, Wen-Han Hsu, Li-Wei Chen, Pei-Chung Su, and Chao-Shih Huang. 2023. Taiwanese Hakka Across Taiwan corpus and Formosa speech recognition challenge 2023 – Hakka ASR. In *2023 26th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6. IEEE.

Binghuai Lin and Liyuan Wang. 2021. Deep feature transfer learning for automatic pronunciation assessment. In *Interspeech 2021*, pages 4438–4442.

Kaizhi Qian, Yang Zhang, Heting Gao, Junrui Ni, Cheng-I Lai, David Cox, Mark Hasegawa-Johnson, and Shiyu Chang. 2022. ContentVec: An improved self-supervised speech representation by disentangling speakers. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 18003–18017. PMLR.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008.

Silke M. Witt and Steve J. Young. 1998. Performance measures for phone-level pronunciation teaching in CALL. In *ETRW on Speech Technology in Language Learning (STiLL)*, pages 99–102.

Xiaoshuo Xu, Yueteng Kang, Songjun Cao, Binghuai Lin, and Long Ma. 2021. Explore wav2vec 2.0 for mispronunciation detection. In *Interspeech 2021*, pages 4428–4432.

Bi-Cheng Yan, Hsin-Wei Wang, Fu-An Chao, Tien-Hong Lo, Yung-Chang Hsu, and Berlin Chen. 2025. HiPPO: Exploring a novel hierarchical pronunciation assessment approach for spoken languages. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 810–823. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Tzu-Hsuan Yang, Yue-Yang He, and Berlin Chen. 2025. JCAPT: A joint modeling approach for CAPT. In *Proceedings of the 10th Workshop on Speech and Language Technology in Education (SLaTE)*, pages 96–100.