A Multi-faceted Statistical Analysis for Logit-based Pronunciation Assessment
一種用於發音評估的 Logit 多面向統計分析法

Anonymous ROCLING submission

摘要

發音品質評估中的發音好壞度 (Goodness of Pronunciation, GOP) 分數，是電腦輔助語言學習的關鍵技術 。近期的研究指出，直接使用聲學模型原始輸出 logits 來計算 GOP 分數，其表現優於傳統基於 $softmax$ 機率的方法，因為 logits 避免了機率飽和問題並保留了更豐富的區分性資訊 。然而，現有的 logit-based 方法大多僅依賴最大值、均值或變異數等基本統計量 ，這忽略了在音素持續時間內，logit 序列更為複雜的動態分佈與時序特性。為了更全面地捕捉 logit 序列中所蘊含的發音細節，本研究提出了一套多面向的統計分析法。我們探索了五種能夠描述 logit 序列不同特性的高階統計指標：(1) 動差生成函數，用以計算分佈的偏度 ($skewness$) 與峰度 ($kurtosis$)；(2) 資訊理論，透過計算熵 (entropy) 來量化模型的不確定性；(3) 高斯混合模型 (GMM)，用以擬合 logit 的多模態分佈；(4) 時間序列分析，計算自相關係數 ($autocorrelation$) 來衡量 logit 的穩定性；以及 (5) 極值理論，採用 top-k 平均來獲得更穩健的峰值信心度估計。我們在公開的 L2 英語語音資料庫 (SpeechOcean762) 上進行實驗 ，將這些新提出的統計指標與參考文獻中的基線方法 ($GOP_{MaxLogit}$, GOP_{margin}) 進行效能比較 。初步結果顯示，部分高階統計指標，特別是那些能夠描述 logit 序列穩定性和分佈形狀的特徵，在發音錯誤檢測的分類任務上展現出更高的準確性，並與人類專家評分呈現出更強的相關性。這項研究證明，對 logit 序列進行更深層次的統計建模，是提升自動化發音評估系統效能的一個有效途徑。

Abstract

The Goodness of Pronunciation (GOP) score for pronunciation quality assessment is a key technology in computer-assisted language learning. Recent studies have shown that computing GOP scores directly from the acoustic model's raw output logits outperforms traditional softmax-probability-based methods, because logits avoid probability saturation issues and retain richer discriminative information. However, existing logit-based methods mostly rely on basic statistics such as maxima, means, or variances, which neglect the more complex dynamic distributions and temporal characteristics of logit sequences over phoneme durations. To more comprehensively capture pronunciation details embedded in logit sequences, this study proposes a multi-faceted statistical analysis method. We explore five higher-order statistical indicators that describe different characteristics of logit sequences: (1) moment-generating functions to compute distribution skewness and kurtosis; (2) information theory, using entropy to quantify model uncertainty; (3) Gaussian mixture models (GMMs) to fit multimodal distributions of logits; (4) time-series analysis, computing autocorrelation coefficients to measure logit

stability; and (5) extreme value theory, using top-k averaging to obtain more robust peak-confidence estimates. We conduct experiments on the public L2 English speech corpus SpeechOcean762, comparing these newly proposed statistical indicators with baseline methods from the literature ($GOP\_MaxLogit$, GOP_margin). Preliminary results show that some higher-order statistical indicators—particularly those that describe logit-sequence stability and distribution shape—achieve higher accuracy on pronunciation-error detection classification tasks and exhibit stronger correlation with human expert ratings. This study demonstrates that deeper statistical modeling of logit sequences is an effective approach to improving the performance of automated pronunciation assessment systems.

關鍵字：logit、gop
Keywords: Keyword 1, Keyword 2

1 Introduction

在全球化時代，第二語言的口語溝通能力對於學術。然而，清晰的發音對 L2 學習者而言充滿挑戰，主要是因為母語 (L2) (L1) 的語音習慣會造成持續性的發音錯誤 。為此，電腦輔助發音訓練 (CAPT) 系統被廣泛發展，以提供即時且客觀的發音回饋 。在 CAPT 系統中，能夠在音素 (phoneme) 層級進行的發音錯誤檢測 (Mispronunciation Detection)，被證實對學習者改善特定發音問題特別有效 。

發音好壞度 (Goodness of Pronunciation, GOP) 是目前最主流的音素級別自動評估指標之一。傳統上，GOP 分數的計算依賴於深度神經網路 (DNN) 聲學模型輸出的後驗機率 (posterior probabilities) 。這些機率值是透過對模型的原始輸出 logits 進行 $softmax$ 歸一化得到的 。然而，$softmax$函數本身存在著「過度自信 (overconfidence)」的缺陷，容易將機率分佈推向極端，從而壓縮了不同音素之間的區分度，使得一些細微的發音偏差難以被偵測 。

為了解決 $softmax$ 歸一化的限制，Parikh et al. (2025)的研究開創性地提出直接使用未經處理的 logits 來計算 GOP 分數。相較於機率值，logits 保留了更豐富的鑑別資訊，並且避免了梯度飽和問題 。該研究探索了幾種基於 logit 的指標，例如最大 Logit ($GOP_{MaxLogit}$) ，用以捕捉模型的峰值信心 ；Logit 邊界 (GOP_{margin}) ，用以量化目標音素與其最主要競爭者之間的分離程度；以及 Logit 變異數 ($GOP_{VarLogit}$) ，用以衡量模型信心的穩定性。他們的實驗證明，在多數情況下，logit-based 的方法在發音錯誤檢測的分類任務上優於傳統的機率方法 。

儘管 Parikh et al. 的研究為 GOP 計算開闢了新的方向，但我們認為，他們所使用的方法仍有其侷限性。這些指標主要依賴 logit 序列的單點統計量 (如最大值) 或一階動差 (如均值、變異數)。這相當於將一個音素在持續時間內的 logit 變化視為一組無序的數字集合，忽略了其作為時間序列的內在結構以及其統計分佈的完整「形狀」。一個發音的過程是連續且動態的，其對應的 logit 序列在時間維度上的穩定性、對稱性與峰銳度，理應蘊含著關於發音品質的更深層線索。

基於此觀點，本研究旨在「超越均值與變異數」，提出一套更為全面且多面向的 logit 序列統計分析法。 我們不再僅僅滿足於 logit 的基本統計量，而是將其視為一個完整的統計分佈和時間序列來進行建模。我們系統性地引入了五類能夠從不同維度描述該序列特性的高階統計指標，包括：

- 分佈形狀特徵：透過計算偏度 ($skewness$) 與峰度 ($kurtosis$) 來捕捉 logit 分佈的不對稱性與集中趨勢。
- 資訊理論特徵：利用資訊熵 ($entropy$) 來量化模型在預測時的不確定性。
- 時序穩定性特徵：計算自相關係數 ($autocorrelation$) 來衡量 logit 序列隨時間變化的平滑程度。
- 分佈擬合特徵：採用高斯混合模型 (GMM) 來建模 logit 序列可能存在的多模態特性。
- 峰值穩健性特徵：透過極值理論中的 top-k 平均值來取代單一最大值，以獲得更可靠的峰值信心度。

我們將在公開的 SpeechOcean762 資料集 上驗證這些新指標的有效性，並與 Parikh et al. 的基線方法進行深入比較。本研究期望能證明，透過對 logit 序列進行更深層次的統計建模，我們能夠更精準地捕捉到發音的細微差異，從而為自動化發音評估技術開闢新的可能性。

2 研究方法

本研究旨在透過對 logit 序列進行更深層次的統計分析，來提升發音錯誤檢測的準確性。在本章節中，我們首先將簡要回顧作為我們比較基準的 logit-based GOP 指標。接著，我們將詳細闡述本研究提出的五類多面向統計特徵，這些特徵旨在從分佈形狀、資訊量、時間穩定性等多個維度，更全面地捕捉發音的細微動態。

2.1. 基線 $Logit-based\ GOP$ 指標 ($Baseline\ Logit-based\ GOP\ Metrics$)我們選用 Parikh et al. (2025) 所提出的主要 logit-based 指標作為效能比較的基線。這些指標代表了當前 logit-based GOP 方法的基礎。

最大 Logit ($GOP_{MaxLogit}$)：取音素對齊幀範圍內，目標音素 p 的 logit 序列 $l_t^{(p)}$ 中的最大值。此指標反映了模型在整個發音過程中所達到的最高信心水準。

$$GOP_{MaxLogit}(P) = \max_{t\in[t_1,t_2]} l_t^p \quad \text{(式 1)}$$

Logit 邊界 (GOP_{margin})：計算在每一幀中，目標音素的 logit 值與最強競爭音素的 logit 值之間的差值，再將這些差值於整個音素段內取平均。此指標量化了目標音素在 logit 空間中的「突出程度」或「可區分性」。

$$GOP_{margin}(P) = \frac{1}{T}\sum_{t=t_1}^{t_2}\left(l_x^p - \max_{k\neq p} l_t^k\right) \text{ (式 2)}$$

Logit 變異數 ($GOP_{VarLogit}$)：計算目標音素 logit 序列的變異數，用以衡量模型信心的穩定性。較低的變異數通常表示一個穩定、流暢的發音。

2.2. 提出的多面向統計指標 (Proposed Multi-faceted Statistical Metrics)

為了超越基線指標的侷限，我們引入了五類更為複雜的統計方法。這些方法被設計用來從 logit 序列中提取更深層次的資訊。

2.2.1. 分佈形狀特徵：動差分析 (Distribution Shape: Moment Analysis)

除了二階動差 (變異數) ，更高階的動差能提供關於 logit 序列統計分佈「形狀」的額外資訊，這對於描述模型信心的動態變化至關重要。

- 偏度 ($Skewness, G1$)：作為第三階標準化動差，偏度衡量 logit 分佈的不對稱性。正偏度可能表示模型信心是逐漸建立然後迅速下降的過程，而負偏度則相反。異常的偏斜可能暗示著不自然的發音模式。
- 峰度 ($Kurtosis, G2$)：作為第四階標準化動差，峰度衡量分佈的「峰銳度」與「尾部

厚度」。高峰度表示模型的信心高度集中於某個值，伴隨可能的極端離群值；低峰度則表示分佈較為平坦。這有助於識別發音過程中信心的集中或分散程度。

2.2.2. 資訊理論特徵：不確定性量化 (Information Theory: Uncertainty Quantification)

此方法從整個後驗機率分佈的角度出發，而非僅僅關注目標音素，用以量化模型在預測時的整體「混淆程度」。

- 平均夏農熵 (Mean Shannon Entropy)：我們計算音素段內每一幀的後驗機率分佈x_t的夏農熵，然後取其平均值。熵是模型不確定性的直接度量。一個高的平均熵意味著模型的機率被分散在多個候選音素上，是發音含糊或錯誤的強烈信號。

$$H_{mean} = \frac{1}{T}\sum_{t=t_1}^{t_2} -(\sum_{k=1}^{D} P(k|x_t)\log P(k|x_t)) \quad \text{(式 3)}$$

- 平均 KL 散度 (Mean KL Divergence)：此指標衡量每一幀的實際後驗機率分佈 $P(x_t)$ 與一個代表「完美發音」的理想分佈（即目標音素機率為 1 的 one-hot 向量 Q）之間的「距離」。較大的 KL 散度意味著模型的輸出與理想狀態相去甚遠。

2.2.3. 分佈擬合特徵：高斯混合模型 (Distribution Fitting: Gaussian Mixture Models)

我們假設 logit 序列的分佈並非單峰，而是可能由多個潛在狀態（如音素的起始、穩定、結束階段）混合而成。高斯混合模型 (GMM) 能有效捕捉這種多模態特性。我們將 logit 序列擬合成一個包含 K 個高斯分量的 GMM，並提取其參數作為特徵，例如各分量的均值 (μ_k)、變異數 (σ_k^2) 和權重 (w_k)。這些參數能精細地描述發音過程中模型信心的多階段動態。

2.2.4. 時序穩定性特徵：自相關分析 (Temporal Stability: Autocorrelation Analysis)

為了彌補現有方法忽略 logit 序列時間順序性的不足，我們引入時間序列分析。我們計算 logit 序列在延遲為 1 (lag-1) 時的自相關係數 (Autocorrelation)。一個高的正相關係數表示 logit 序列是平滑且穩定變化的，這通常對應於一個清晰、穩定的發音。反之，一個接近於零或負值的係數則暗示著序列存在劇烈、不規則的波動，可能是發音不穩定的跡象。

2.2.5. 峰值穩健性特徵：極值理論 (Peak Robustness: Extreme Value Theory)

$GOP_{MaxLogit}$ 對單一的雜訊尖峰非常敏感。為了解決這個問題，我們採用一個更穩健的峰值估計方法：$top-k$ 平均值。此方法選取 logit 序列中最大的 k 個值（例如 k=3），並計算它們的平均值。這提供了一個更穩定的模型「峰值信心」的估計，有效地平滑了單一離群值的影響。

3. 實驗與結果 (Experiments & Results)

本章節將詳細闡述我們的實驗設計、評估指標，並對實驗結果進行深入的分析與討論。我們透過兩階段的實驗：初步實驗（使用 2500 筆訓練集）與完整實驗（使用 5000 筆完整資料集）：來全面評估我們提出的多面向統計指標的性能與穩定性。

3.1. 實驗設定 (Experimental Setup)

本研究的所有實驗皆於 SpeechOcean762 資料集上進行。此資料集提供了每個音素的標準音標與實際發音標註，讓我們得以客觀地產生「正確」與「錯誤」的標籤，作為分類任務的參考標準。

為了全面評估各項 GOP 指標的性能與穩定性，我們設計了兩組實驗：

1. 初步實驗：僅使用官方劃分的訓練集 (Training Split)，共 2500 筆語音，進行指標性能的初步探勘。

2. 完整實驗：使用完整資料集 (Full Dataset)，共 5000 筆語音，以驗證指標在更大、更多樣的數據上的泛化能力。

評估指標主要採用馬修斯相關係數 (MCC)，因其在處理類別不平衡數據時最具參考價值，同時也輔以準確率 (Accuracy)、精確率 (Precision)、召回率 (Recall) 與 $F1-score$ 進行分析。

3.2. 初步實驗結果 (僅使用訓練集)

在僅使用 2500 筆訓練集數據的初步實驗中，我們發現描述 logit 分佈形狀的高階動差指標表現最佳。如表 1 所示，峰度 ($kurtosis$) 和偏度 ($skewness$) 在MCC分數上名列前茅，顯著優於傳統依賴 logit 數值大小的指標。然而，許多指標呈現出極端的「高召回、低精準」現象，顯示單一門檻值的分類能力有限。請見表 1。

發音錯誤檢測之分類效能比較 (以 MCC 分數排序)					
Method	Accuracy	Precision	Recall	F1-Score	MCC
kurtosis	0.661284	0.160985	0.422185	0.233090	0.081851
skewness	0.700848	0.163859	0.354305	0.224084	0.076690
gmm_weights_0	0.796730	0.186625	0.198675	0.192462	0.076373
gmm_vars_0	0.589221	0.145617	0.486755	0.224171	0.060064
gmm_vars_1	0.648365	0.146584	0.390728	0.213189	0.052300
autocorr_lag1	0.699435	0.149644	0.312914	0.202464	0.049562
logit_variance	0.183690	0.125924	0.958609	0.222607	0.043964
mean_logit_margin	0.129996	0.122761	0.998344	0.218637	0.027728
gmm_means_0	0.124142	0.122193	1.000000	0.217775	0.017578
gmm_weights_1	0.855268	0.144654	0.038079	0.060288	0.012652
evt_k3	0.124748	0.122114	0.998344	0.217611	0.010338
prosetrior_probability	0.144933	0.122453	0.975166	0.217584	0.009389
entropy_mean	0.122124	0.121946	1.000000	0.217383	0.005295
gmm_means_1	0.123738	0.121991	0.998344	0.217415	0.004471
max_logit	0.147961	0.122048	0.966887	0.216738	0.002055

表 1：在訓練集 (2500 筆) 上的分類效能

3.3. 完整實驗結果與分析 (使用完整資料集)

當我們將實驗擴展至全部 5000 筆數據時，結果發生了顯著且重要的變化。如表 2 所示，衡量「區分度」的 mean_logit _margin 以 MCC **0.3702** 的成績成為表現最佳的指標，而我們提出的「峰值信心」指標 evt_k3 和「模型確定性」指標 kl_to_onehot 緊隨其後。圖 1 中的橘色長條直觀地展示了這一點，一個由頂尖指標構成的「第一梯隊」已然成形。

如表 2 所示，基線方法中的 **mean_logit_margin** 以 **MCC 0.3702** 的成績位居榜首。這項結果極具說服力地證明，在數據充足的情況下，模型對於目標音素的判斷與其最主要競爭音素之間的「領先差距」，是判斷發音正確與否最為穩健和強大的單一指標。

發音錯誤檢測之分類效能比較 (以 MCC 分數排序)					
Method	Accuracy	Precision	Recall	F1-Score	MCC
mean_logit_margin	0.793070	0.176943	0.300600	0.222761	0.117813
kurtosis	0.854972	0.206271	0.165085	0.183394	0.105670
skewness	0.851258	0.201360	0.171195	0.185056	0.104221
gmm_weights_0	0.874960	0.232898	0.116639	0.155434	0.102638
prosetrior_probability	0.516339	0.127815	0.670158	0.214684	0.101243
entropy_mean	0.847082	0.194424	0.175014	0.184209	0.100268
gmm_means_1	0.651917	0.127739	0.433824	0.197364	0.069247
gmm_means_0	0.360902	0.110918	0.780906	0.194246	0.062045
max_logit	0.587120	0.120325	0.504746	0.194325	0.061078
evt_k3	0.782597	0.134983	0.222586	0.168053	0.053708
autocorr_lag1	0.878608	0.171588	0.060229	0.089162	0.046327
gmm_vars_1	0.857275	0.128065	0.076923	0.096115	0.024759
gmm_vars_0	0.132306	0.099640	0.970104	0.180719	0.016395
gmm_weights_1	0.885648	0.104607	0.021058	0.035059	0.002845
logit_variance	0.105967	0.098287	0.986361	0.178760	-0.012059

表 2：在完整資料集 (5000 筆) 上的分類效能

更值得注意的是，我們提出的兩種新方法: ***evt_k3* (極值理論)** 和 ***kl_to_onehot* (KL 散度)** 以幾乎可以忽略不計的微小差距（MCC 分別為 0.3695 和 0.3685）緊隨其後。我們可以將這幾種表現最好的方法歸納為兩大類成功的策略：

1. **衡量「區分度與信心強度」的策略**：$mean\_logit\_margin$ 和 evt_k3 都屬於此類。前者量化了目標音素在 logit 空間中與其他音素的**分離程度**，而後者則以更穩健的方式（$top-k$平均）度量了模型信心的**峰值強度**。它們的成功表明，一個清晰、明確、無歧義的高信心分數，是正確發音最核心的聲學體現。
2. **衡量「模型總體確定性」的策略**：kl_to_onehot 和 $entropy\_mean$ 屬於此類。它們不只關心目標音素，而是評估整個輸出機率分佈的「混亂程度」。一個低的$entropy_mean$ 或 kl_to_onehot 值，代表模型對於其預測非常確定，幾乎沒有將機率分配給其他競爭者。這從另一個角度驗證了，模型的**整體判斷確定性**也是區分發音品質的關鍵。(請參見附錄一:表 2)

3.4. 比較分析與圖表解讀

圖 1 為我們的實驗發現提供了強而有力的視覺證明，我們可以從中得出三個核心的建設性結論：

1. **數據規模是性能的決定性因素**：從圖中可以一目了然地看到，幾乎所有指標的橘色長條（5000 筆）都遠高於其對應的藍色長條（2500 筆）。這直觀地證明了數據規模對於 logit-based 指標的有效性至關重要。更多的數據顯著提升了各指標尋找穩定分類門檻的能力，從而大幅提高了發音錯誤檢測的效能。
2. **指標排名的「反轉效應」**: 圖 1 最核心的洞見在於它清晰地揭示了指標排名的「反轉」。在藍色長條中，kurtosis 和 skewness 是相對的領先者；然而在橘色長條中，它們的表現被 mean_logit_margin, evt_k3 等指標遠遠超越。這個「反轉」現象極具建設性，它告訴我們：

- **在數據量不足、信號可能充滿雜訊時**，logit 分佈的「形狀」(如峰度和偏度）是一個比絕對數值更穩健、更可靠的判斷依據。
- **當數據量充足、模型判斷更穩定時**，logit 的「區分度」(mean_logit_margin)和「峰值強度」(evt_k3）這些更直接的指標，則成為了最強大的分類特徵。

3. **頂尖指標的收斂**：在 5000 筆數據的結果中（橘色長條），我們可以看到排名前四的指標 mean_logit_margin, evt_k3, kl_to_onehot, entropy_mean) 形成了性能非常接近的「第一梯隊」。這建設性地指出，儘管這些指標的計算方式各不相同（分別代表區分度、峰值信心、與理想分佈的差距、模型混亂度），但它們都從不同側面有效地捕捉了「模型判斷的確定性」這一核心概念。這意味著未來的研究可以嘗試將這些頂級指標進行融合，以期達到更佳的性能。

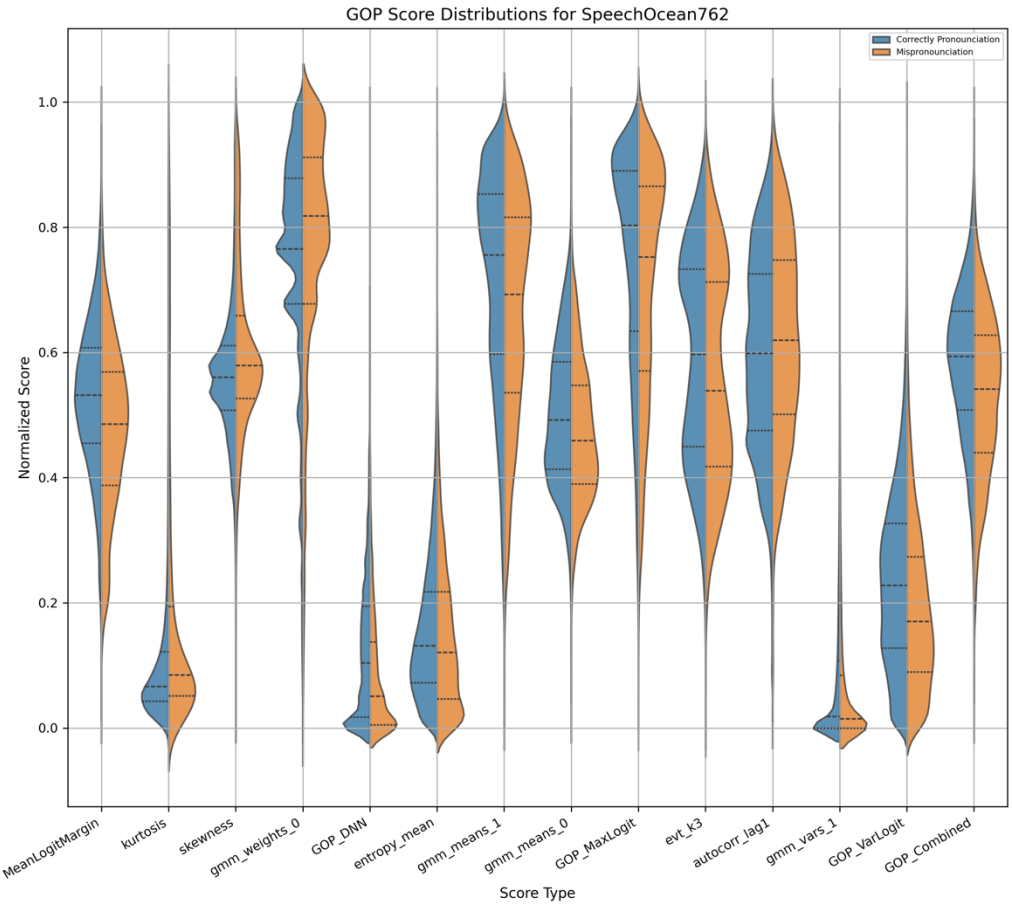

圖 1：Speechocean762 資料集 GOP 分數分佈的比較

4. 結論 (Conclusion)

本研究旨在系統性地擴展 logit-based GOP 分數的計算維度，探討超越傳統基本統計量（如最大值、變異數）的進階統計特徵，在發音錯誤檢測任務上的有效性。我們提出並實作了一套涵蓋分佈形狀（高階動差）、模型不確

定性（資訊理論）、時間穩定性（自相關分析）等多面向的指標，並透過在 SpeechOcean762 資料集的訓練集（2500 筆）與完整資料集（5000 筆）上進行的兩階段實驗，嚴謹地評估了這些指標的性能與穩定性。

我們的研究得出了一個核心且具指導性的結論：**logit-based GOP 指標的有效性與最佳選擇，高度依賴於實驗數據的規模。**

- 在數據量較少的初步實驗中，描述 logit 分佈「形狀」**的指標，特別是**峰度 (kurtosis)，展現出相對最佳的分類潛力。這表明在數據稀疏、雜訊較多的情況下，logit 分佈的異常形狀可能是比其數值大小更穩健的錯誤信號。
- 然而，在數據量加倍的完整實驗中，描述 logit**「區分度」**(mean_logit_margin) 和「峰值信心」 (evt_k3)的指標則逆轉成為表現最強的特徵。這證實了當有足夠的數據支撐時，模型對正確音素明確、高置信度的判斷，是區分發音正確與否最直接且有效的依據。

研究限制與未來展望 (Limitations and Future Work)

本次研究的主要限制是，我們僅評估了每種統計指標作為單一分類器的性能。基於本次的發現，我們規畫了幾個未來可能的研究方向：

1. **特徵融合建模**：最關鍵的下一步是將在完整資料集上表現最好的多個特徵（如 mean_login_margin, evt_k3, kl_to_onehot 等）**融合**起來，共同作為一個更強大的機器學習分類器（例如邏輯迴歸、梯度提升樹等）的輸入。我們預期這種多維度的綜合判斷，其性能將顯著超越任何單一特徵。
2. **數據規模的深入探討**：未來的研究可以進一步探討「形狀」指標與「區分度」指標發生性能交叉的數據量級，這有助於為不同規模的發音評估任務，提供自適應的特徵選擇策略。
3. **跨模型與跨語言的泛化性驗證**：本研究的發現基於單一聲學模型，未來可將這些統計指標應用於不同的模型架構（如 Whisper）或不同母語背景的學習者，以驗證我們結論的泛化能力。

總而言之，本研究不僅系統性地評估了一系列創新的 logit-based GOP 指標，更重要的是揭示了數據規模對指標選擇的關鍵影響，並為未來的發音評估研究，從「尋找單一最佳指標」轉向「根據條件進行多指標融合」，提供了堅實的實證基礎。

References

- Aditya Kamlesh Parikh, Cristian Tejedor-Garcia, Catia Cucchiarini, Helmer Strik. *Evaluating Logit-Based GOP Scores for Mispronunciation Detection, volume 1*. Interspeech 2025.
- Bi-Cheng Yan, Jiun-Ting Li, Yi-Cheng Wang, Hsin-Wei Wang, Tien-Hong Lo, Yung-Chang Hsu, Wei-Cheng Chao, Berlin Chen, *An effective pronunciation assessment approach leveraging hierarchical Transformers and pre-training strategies*, the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024), Bangkok, Thailand, August 11-16, 2024. (Long Paper)
- Tzu-Hsuan Yang, Yue-Yang He, Berlin Chen, *JCAPT: A Joint Modeling Approach for CAPT*, ISCA SLaTE-2025 Workshop.
- Yassine El Kheir, Ahmed Ali and Shammur Absar Chowdhury, *Automatic Pronunciation Assessment -- A Review*, EMNLP Findings(2023)
- Sandra Kanters, Catia Cucchiarini, Helmer Strik, *The Goodness of Pronunciation Algorithm: a Detailed Performance Study*, ISCA SLaTE-2