|  |  |  |  |
| --- | --- | --- | --- |
| **Instructions for ROCLING 2025 Proceedings (中文稿件需提供英文題目)** | | | |
|  | | | |
|  | | | |
|  | **First Author**  Affiliation / Address line 1  Affiliation / Address line 2  Affiliation / Address line 3  email@domain | **Second Author**  Affiliation / Address line 1  Affiliation / Address line 2  Affiliation / Address line 3  email@domain |  |
|  | | | |

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 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 () 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 ( , ) 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.

關鍵字：關鍵字1、關鍵字2

Keywords: Keyword 1, Keyword 2

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, ) 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 歸一化得到的 。然而，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 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 () ，用以捕捉模型的峰值信心；Logit 邊界 () ，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 ( ) ，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 () 與峰度 () 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 () 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 () 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 () 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。

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 指標 ()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 (​)：取音素對齊幀範圍內，目標音素 p 的 logit 序列 ​ 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。

一張含有 字型, 文字, 白色, 書法 的圖片

AI 產生的內容可能不正確。

Logit 邊界 (​)：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」。

一張含有 字型, 白色, 書法, 文字 的圖片

AI 產生的內容可能不正確。

Logit 變異數 (​)：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 (​)：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 (​)：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。

一張含有 字型, 文字, 白色, 筆跡 的圖片

AI 產生的內容可能不正確。

* 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 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 ()

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 (μk​)、變異數 (​) 和權重 (wk​)。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)

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： 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)

**3.1. 實驗設定 (Experimental Setup)**

**資料集 (Dataset):** 本研究的所有實驗皆於

**SpeechOcean762** 資料集上進行 。此資料集是一個專為發音評估研究所設計的公開語料庫，包含了由 250 位母語為中文的 L2 英語學習者所錄製的 5,000 句英語語音 。至關重要的是，該資料集提供了每個音素的標準音標 (canonical transcription) 和實際發音音標 (realized transcription) 。這讓我們可以透過程式化比對，客觀地產生「正確」或「錯誤」的標籤，作為我們分類任務的參考標準 (ground truth)，而無需依賴任何人類的主觀評分。

**評估指標 (Evaluation Metrics):** 我們將發音錯誤檢測定義為一個二元分類任務（正確 vs. 錯誤）。由於資料集中正確發音的樣本遠多於錯誤發音，存在類別不平衡問題，因此我們採用了以下指標來評估模型效能：

**準確率 (Accuracy)**、**精確率 (Precision)**、**召回率 (Recall)**、**F1-score** 以及 **馬修斯相關係數 (Matthews Correlation Coefficient, MCC)** 。其中，MCC 是一個在不平衡資料下表現特別穩健的分類效能指標，我們將其視為主要的評斷依據 。

**比較方法 (Compared Methods):** 我們將本研究提出的五類多面向統計指標，與 Parikh et al. (2025) 提出的

logit-based 基線指標進行比較，主要包括 ， ， 。

**3.2. 結果與分析 (Results and Analysis)**

**3.2.1. 主要結果**

本實驗旨在評估多種 指標在發音錯誤檢測二元分類任務上的效能。詳細的實驗數據如表1所示，所有指標均以馬修斯相關係數 () 作為主要排序依據，因為它在處理類別不平衡的數據時最為穩健。

實驗結果揭示了幾個出乎意料的發現：

1. **整體分類效能有限**：首先，從數據中可以看出，所有單獨使用的 GOP 指標，在本次實驗中的分類能力都相對有限。表現最好的指標 （峰度），其 MCC 分數也僅為 0.0819，這表明僅依靠單一維度的統計特徵，並設定一個最佳門檻值，難以完美地區分正確與錯誤的發音。
2. **高階動差表現最佳**：令人驚訝的是，在所有指標中，表現最好的是描述 logit 分佈**形狀**的高階動差特徵，特別是**峰度 (kurtosis)** 和**偏度 (skewness)**，它們在 MCC 排序中位列第一和第二。
3. **基線方法與部分新方法的權衡問題**：許多基線方法（如 ）和我們提出的部分方法（如, ）展現出極端的「高召回率、低精確率」現象。例如， 和 的召回率達到了 1.0，但其準確率卻極低（約 0.12），MCC 分數也趨近於 0。這意味著，為了找到盡可能多的錯誤發音（高召回率），這些指標的最佳門檻值將幾乎所有的音素都標記為「錯誤」，從而導致了極高的誤報率（低精確率），使其作為分類器幾乎失效。

(請見附錄一：表格1：不同 指標於發音錯誤檢測之分類效能比較)

**3.2.2. 討論與分析**

這次的實驗結果雖然絕對數值不高，但其**相對排行**為我們提供了極具價值的洞見：

* **為何傳統強指標失效？** 像 和 這樣的指標，在我們的實驗中表現不佳，呈現出極端的召回率-精確率權衡。這強烈暗示，對於我們的數據來說，正確發音和錯誤發音的 logit 分佈可能存在著巨大的重疊。沒有一個簡單的門檻值，可以僅僅依據 logit 的「高度」（最大值）或「領先差距」（邊界）就有效地將兩者分開。這與參考論文的結果有所不同，可能反映了數據集、聲學模型或特徵提取流程中的細微差異。
* **為何「形狀」比「高度」更重要？** 本次實驗最有趣的發現是， (峰度) 和 (偏度) 成為了相對最好的指標。這說明了什麼？這可能意味著，一個錯誤的發音，其 logit 序列的特徵，**不一定是整體數值更低，而是其分佈的「形狀」更為異常**。
  + 例如，一個不確定的發音可能導致模型信心在過程中劇烈波動後又回歸，形成一個「尖銳」的峰（高 ）。
  + 或者，一個猶豫的發音可能讓模型信心緩慢上升但在結尾時非常確定，形成一個不對稱的分佈（高 ）。
  + 這個發現支持了我們最初的假設：**超越均值和變異數，去分析 logit 序列的完整分佈特性，是具有潛力的方向。**
* **對未來研究的啟示**： 這次的結果表明，單一特徵的分類能力是有限的。然而， 和 的相對成功，為我們指明了方向。未來的研究不應再尋找單一的「最佳」GOP 指標，而應考慮**將這些不同的特徵（特別是描述形狀、穩定性和區分度的特徵）結合起來**，輸入一個更強的機器學習分類器（如 SVM、梯度提升樹等）。例如，一個同時考慮 和 的模型，可能會比單獨使用其中任何一個都更強大。

4結論 (Conclusion)

本研究旨在探索超越傳統 指標的可能性，透過引入一套更多面向的統計特徵，以期能更精準地進行發音錯誤檢測。我們基於現有的 logit 分數（如最大值、邊界值）僅依賴基本統計量的觀察，提出並實作了五類能夠描述 logit 序列完整分佈與動態特性的高階指標，涵蓋了分佈形狀（動差）、模型不確定性（資訊理論）、時間穩定性（自相關分析）等多個維度。

我們在 SpeechOcean762 資料集上進行了客觀的二元分類實驗。實驗結果揭示了一個深刻且出乎意料的發現：在我們的實驗設定下，所有單一特徵分類器的絕對性能都相對有限，這表明僅依靠任何一種 logit 統計特徵都難以完美地解決發音錯誤檢測的挑戰。

然而，從各指標的**相對表現**來看，本研究最重要的貢獻在於發現：**描述 logit 分佈「形狀」的高階動差指標，特別是峰度 (kurtosis) 與偏度 (skewness)，展現出比傳統依賴「高度」或「區分度」的指標（如 max\_logit 和 mean\_logit\_margin）更優越的分類潛力。** 這個結果強烈暗示，一個錯誤的發音所引起的模型信心變化，其特徵可能更多地體現在 logit 序列分佈的異常形狀上（例如更尖銳、更不對稱），而非單純的數值降低。

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

本次研究的主要限制在於我們僅評估了每種統計指標**獨立**作戰時的性能。實驗結果表明這條路徑的潛力有限。基於本次研究的發現，我們規畫了幾個未來可能的研究方向：

1. **特徵融合建模**：最直接的下一步是將本次實驗中表現突出的多個特徵（如 kurtosis, skewness）以及其他有潛力的指標**結合起來**，共同作為一個更強大的機器學習分類器（例如邏輯迴歸、支持向量機或梯度提升樹）的輸入，我們預期組合特徵後的模型性能將遠超任何單一特徵。
2. **模型依賴性分析**：本研究的發現是基於單一的 Wav2Vec2 聲學模型。未來的研究可以探討我們提出的這些統計指標，在不同的聲學模型（如 Whisper 或其他自監督模型）上是否依然有效。
3. **錯誤類型深入分析**：針對 kurtosis 和 skewness 的相對成功，未來可以進行更深入的錯誤分析，探討它們是否對特定類型的發音錯誤（例如，替代、刪除或插入）特別敏感。

總而言之，本研究雖然未能找到單一的「完美」GOP 指標，但卻成功地為 logit-based 發音評估方法開闢了一個新的視角：**從分析 logit 的「數值」，轉向分析 logit 的「形狀」**。我們相信，這將是未來提升自動化發音評估系統性能的一個關鍵方向。

References

Alfred. V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling, volume 1*. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual.* American Psychological Association, Washington, DC.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

附錄一

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 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 |

表格1：不同 GOP 指標於發音錯誤檢測之分類效能比較