Evaluating Logit-Based GOP Scores for Mispronunciation Detection

評估基於 Logit 的 GOP 分數於錯誤發音偵測之效能

Aditya Kamlesh Parikh, Cristian Tejedor-Garcia, Catia Cucchiarini, Helmer Strik

Aditya Kamlesh Parikh、Cristian Tejedor-Garcia、Catia Cucchiarini、Helmer Strik

Centre for Language Studies, Radboud University, the Netherlands

荷蘭 Radboud 大學語言研究中心

aditya.parikh@ru.nl, cristian.tejedorgarcia@ru.nl, catia.cucchiarini@ru.nl, helmer.strik@ru.nl

aditya.parikh@ru.nl、cristian.tejedorgarcia@ru.nl、catia.cucchiarini@ru.nl、helmer.strik@ru.nl

Abstract 摘要

Pronunciation assessment relies on goodness of pronunciation (GOP) scores, traditionally derived from softmax-based posterior probabilities. However, posterior probabilities may suffer from overconfidence and poor phoneme separation, limiting their effectiveness. This study compares logit-based GOP scores with probability-based GOP scores for mispronunciation detection. We conducted our experiment on two L2 English speech datasets spoken by Dutch and Mandarin speakers, assessing classification performance and correlation with human ratings. Logit-based methods outperform probabilitybased GOP in classification, but their effectiveness depends on dataset characteristics. The maximum logit GOP shows the strongest alignment with human perception, while a combination of different GOP scores balances probability and logit features. The findings suggest that hybrid GOP methods incorporating uncertainty modeling and phoneme-specific weighting improve pronunciation assessment. Index Terms: GOP, logit-based GOP, mispronunciation detection, pronunciation assessment, softmax posterior probabilities

發音評估依賴於發音優良度（GOP）分數，傳統上是由基於 softmax 的後驗機率推導而來。然而，後驗機率可能會出現過度自信及音素分離不佳的問題，限制其效能。本研究比較了基於 logit 的 GOP 分數與基於機率的 GOP 分數在錯誤發音偵測上的表現。我們在兩個由荷蘭語和中文母語者所說的第二語言英語語音資料集上進行實驗，評估分類效能及與人工評分的相關性。基於 logit 的方法在分類上優於基於機率的 GOP，但其效能依賴於資料集特性。最大 logit GOP 與人類感知的對應性最強，而不同 GOP 分數的組合則在機率與 logit 特徵間取得平衡。研究結果顯示，結合不確定性建模與音素特定加權的混合 GOP 方法能提升發音評估。索引詞：GOP、基於 logit 的 GOP、錯誤發音偵測、發音評估、softmax 後驗機率

1. Introduction 1. 引言

In today’s interconnected world, globalization has led to increased movement across borders for work, education, and other opportunities. For individuals who are adapting to a new linguistic environment, learning the local language is essential for social integration, career advancement, and overall wellbeing [1]. Effective communication goes beyond knowing the vocabulary and grammar: A clear pronunciation is equally important, as it impacts intelligibility, confidence, and the ability to engage in meaningful conversation [2, 3]. Poor pronunciation can lead to misunderstandings, impede effective social interactions and even create barriers in academic and professional settings [4]. However, mastering pronunciation in a second language (L2) can be challenging. Differences between the first language (L1) and the target language often result in persistent pronunciation errors. These difficulties are further compounded by the limited tools and resources available to

language instructors for providing personalized pronunciation feedback.

在當今互聯互通的世界中，全球化促使人們為了工作、教育及其他機會跨越國界移動。對於適應新語言環境的個人來說，學習當地語言對於社會融入、職業發展及整體福祉至關重要[1]。有效的溝通不僅僅是掌握詞彙和語法：清晰的發音同樣重要，因為它影響可理解度、自信心以及進行有意義對話的能力[2, 3]。發音不佳可能導致誤解，阻礙有效的社交互動，甚至在學術和職業環境中造成障礙[4]。然而，掌握第二語言（L2）的發音具有挑戰性。母語（L1）與目標語言之間的差異常導致持續的發音錯誤。這些困難因語言教師缺乏提供個別化發音反饋的工具和資源而更加複雜。

To address these challenges, Computer-Assisted Pronunciation Training (CAPT) systems have gained popularity [5, 6]. A key component of these systems is Mispronunciation Detection and Diagnosis (MDD), which helps learners identify and correct pronunciation errors in real time [7]. Phoneme-level assessment, in particular, provides more precise feedback than broader word- or sentence-level evaluations, allowing learners to focus on specific areas for improvement [8]. One of the most widely used methods for detecting phoneme-level mispronunciations is the goodness of pronunciation (GOP) score [9].

為了解決這些挑戰，電腦輔助發音訓練（CAPT）系統逐漸受到重視[5, 6]。這些系統的一個關鍵組成部分是誤讀偵測與診斷（MDD），能夠幫助學習者即時識別並修正發音錯誤[7]。特別是音素層級的評估，比起較廣泛的詞彙或句子層級評估，能提供更精確的反饋，使學習者能專注於特定的改進區域[8]。其中，最廣泛使用的音素層級誤讀偵測方法之一是發音優劣分數（GOP）[9]。

GOP was initially introduced as a measure of pronunciation quality, estimating the probability of a phoneme and comparing it against a predefined threshold to flag mispronunciations [10]. Over time, several enhancements have improved its accuracy. Weighted-GOP [11] adjusts phoneme scores based on linguistic and acoustic factors, prioritizing phonemes prone to mispronunciation. Lattice-based GOP [12] considers multiple pronunciation possibilities using phoneme lattices, yielding more robust confidence scores. Context-aware GOP [13] incorporates phoneme transitions and durations to better capture natural pronunciation variations. More recently, multidimensional GOP features [14, 8, 15] have been introduced, leveraging richer feature representations beyond simple probability thresholds for more precise mispronunciation detection.

GOP 最初被引入作為發音品質的衡量指標，透過估計音素的機率並與預先設定的閾值比較，以標示錯誤發音[10]。隨著時間推移，數項改進提升了其準確度。加權 GOP [11] 根據語言學和聲學因素調整音素分數，優先考量易錯發的音素。基於格網的 GOP [12] 利用音素格網考慮多種發音可能性，產生更具魯棒性的信心分數。具上下文感知的 GOP [13] 融入音素轉換和持續時間，更好地捕捉自然發音變化。近期則引入多維度 GOP 特徵[14, 8, 15]，利用比單純機率閾值更豐富的特徵表示，以實現更精確的錯誤發音偵測。

Despite these advancements, GOP methods face challenges in data availability and computational efficiency [16]. Annotated pronunciation data, including prosody and fluency scores, requires expert evaluation, making large-scale collection expensive. Additionally, traditional GOP methods scale poorly with phoneme size because of an increasing number of features and increasing computational costs. This highlights the need for a fast, robust, and non-trainable GOP computation method.

儘管有這些進展，GOP 方法在資料可用性和計算效率方面仍面臨挑戰[16]。帶有韻律和流暢度分數的標註發音資料需要專家評估，使得大規模收集成本高昂。此外，傳統 GOP 方法隨著音素數量增加，特徵數量和計算成本也隨之增加，擴展性較差。這凸顯了對快速、穩健且無需訓練的 GOP 計算方法的需求。

More recently, foundation models [17, 18] trained on massive amounts of data have been used to improve MDD. These models can be fine-tuned with significantly less data, addressing some of the data availability constraints faced by traditional GOP-based methods. GOP scores can be derived from a forced alignment using Connectionist Temporal Classification (CTC)based phoneme recognition models, which rely on posterior probabilities generated by acoustic models. These probabilities are obtained through softmax normalization of model logits, which is a widely adopted method [19]. However, softmaxbased probability estimates suffer from overconfidence [20,21], inflating confidence in incorrect phoneme predictions and reducing the granularity needed to detect subtle mispronunciations. This issue is particularly problematic in phoneme recognition for children’s speech and non-native speakers, where articulatory deviations are very

common [22, 23].

近年來，基於大量數據訓練的基礎模型[17, 18]已被用於提升錯誤發音檢測（MDD）。這些模型可以用顯著較少的數據進行微調，解決了傳統基於 GOP 方法所面臨的一些數據可用性限制。GOP 分數可以從使用基於連結時序分類（CTC）的音素識別模型進行強制對齊中獲得，這些模型依賴於聲學模型生成的後驗概率。這些概率是通過對模型 logits 進行 softmax 正規化獲得的，這是一種廣泛採用的方法[19]。然而，基於 softmax 的概率估計存在過度自信的問題[20,21]，會誇大對錯誤音素預測的信心，降低檢測細微錯誤發音所需的細緻度。此問題在兒童語音和非母語者的音素識別中特別嚴重，因為這些語者的發音偏差非常常見[22, 23]。

To address these limitations, we propose a logit-based GOP method that directly utilizes raw logits from CTC-based models rather than softmax-normalized probabilities. Logits retain more discriminative information and avoid the gradient saturation problem inherent in softmax-based scoring [24]. We explore four logit-based metrics to enhance mispronunciation detection: Maximum Logit (GOP $\mathrm{MaxLogit}$), Mean Logit Margin [25] ($\mathrm{GOP}_{\mathrm{Margin}}$), Logit Variance ($\mathrm{GOP}_{\mathrm{LogitVariance}}$), and combined (hybrid) logit-probability $\mathrm{GOP}_{\mathrm{Combined}}$ scores. This novel approach provides a fast, robust, and non-trainable solution, crucial for real-time phoneme assessment.

為了解決這些限制，我們提出一種基於 logit 的 GOP 方法，直接利用來自 CTC 模型的原始 logits，而非經 softmax 正規化的機率。Logits 保留更多區辨資訊，並避免了 softmax 基分數中固有的梯度飽和問題[24]。我們探討了四種基於 logit 的指標以提升誤讀偵測：最大 Logit（GOP $\mathrm{MaxLogit}$）、平均 Logit 邊際[25]（$\mathrm{GOP}_{\mathrm{Margin}}$）、Logit 變異數（$\mathrm{GOP}_{\mathrm{LogitVariance}}$）以及結合（混合）logit-機率 $\mathrm{GOP}_{\mathrm{Combined}}$ 分數。這種新穎方法提供了一個快速、穩健且無需訓練的解決方案，對於即時音素評估至關重要。

Our method builds upon prior work on uncertainty quantification in pronunciation assessment, such as [26], which applies GOP-based scores to dysarthric speech. However, prior studies have not explicitly investigated the use of raw logits in GOP calculations. Our approach fills this gap by utilizing logit-based metrics to improve accuracy and reliability in pronunciation assessment. Our leading research question (RQ) is: To what extent does a logit-based GOP score enhance mispronunciation detection and improve correlation with human rater scores compared to traditional softmax-based GOP scores?

我們的方法建立在先前關於不確定性量化的研究基礎上—在發音評估中的應用，例如[26]，該研究將基於 GOP 的分數應用於構音障礙語音。然而，先前的研究尚未明確探討在 GOP 計算中使用原始 logits。我們的方法填補了這一空白，通過利用基於 logit 的指標來提升發音評估的準確性和可靠性。我們的主要研究問題（RQ）是：與傳統的基於 softmax 的 GOP 分數相比，基於 logit 的 GOP 分數在多大程度上提升了誤讀檢測的效果並改善了與人工評分者分數的相關性？

2. Methodology 2. 方法論

2.1. Definition of GOP 2.1. GOP 的定義

The GOP score, first introduced by Witt and Young [9], quantifies pronunciation quality by comparing the likelihood of a hypothesized phoneme to competing alternatives. For a phoneme p aligned to an audio segment, the original GOP formulation computes:

GOP 分數最早由 Witt 和 Young[9]提出，用於通過比較假設音素與競爭替代音素的可能性來量化發音質量。對於與音頻片段對齊的音素 p ，原始 GOP 公式計算如下：

$$\mathrm{GOP}_{\mathrm{original}}(p) = \log \frac{P(\mathbf{X} \mid p)}{\frac{1}{N}\sum_{q \in \mathcal{Q}} P(\mathbf{X} \mid q)}$$

where $P(\mathbf{X} \mid p)$ is the likelihood of the acoustic features $\mathbf{X}$ given phoneme p, and $\mathcal{Q}$ represents competing phonemes.

其中 $P(\mathbf{X} \mid p)$ 是在音素 p 下聲學特徵 $\mathbf{X}$ 的可能性， $\mathcal{Q}$ 代表競爭音素。

With deep neural networks (DNNs), GOP is derived from posterior probabilities using the negative log of the mean softmax output over aligned frames [27]:

使用深度神經網路（DNN）時，GOP 是從後驗機率導出，利用對齊幀的平均 softmax 輸出取負對數 [27]：

$$\mathrm{GOP}_{\mathrm{DNN}}(p) = -\log\left(\frac{1}{T}\sum_{t=1}^{T} P\left(p \mid \mathbf{x}_t\right)\right)$$

where $P\left(p \mid \mathbf{x}_t\right)$ is the softmax probability of phoneme p at frame t and T is the total number of frames in the phoneme segment. This equation interprets the mean softmax probability of the target phoneme as a probabilistic measure of pronunciation quality. This approach inherits softmax limitations such as overconfidence and gradient saturation.

其中 $P\left(p \mid \mathbf{x}_t\right)$ 是音素 p 在幀 t 的 softmax 機率， T 是音素區段中的總幀數。此方程將目標音素的平均 softmax 機率解釋為發音品質的機率度量。此方法繼承了 softmax 的限制，如過度自信和梯度飽和。

2.2. Logit-Based GOPS 2.2. 基於 Logit 的 GOPS

To address softmax limitations, we propose four novel metrics using raw logits, defined below.

為了解決 softmax 的限制，我們提出了四個使用原始 logit 的新指標，定義如下。

2.2.1. GOP_{MaxLogit}

This metric captures the model's peak confidence in the target phoneme p across aligned frames t_1 to t_2 :

此指標捕捉模型在對齊的幀 t_1 到 t_2 中對目標音素 p 的最高信心：

$$\mathrm{GOP}_{\mathrm{MaxLogit}}\left(p\right) = \max_{t \in [t_1, t_2]} \mathbf{l}_t^{(p)}$$

where $\mathbf{l}_t{}^{(p)}$ is the logit for phoneme p at frame t. It identifies unambiguous articulations but may emphasize transient spikes.

其中 $\mathbf{l}_t{}^{(p)}$ 是第 t 幀中音素 p 的 logit。它能識別明確的發音，但可能會強調瞬間的峰值。

2.2.2. GOP_{Margin}

This measure quantifies the average superiority of the target phoneme over its strongest competitor. For each frame, we compute the difference (or margin) between the target logit and the highest competing logit. The average of these margins over the segment indicates how well-separated the target phoneme is from other phonemes. This helps in cases where pronunciation errors cause phoneme confusion, which may not always be reflected in probability scores.

此指標量化目標音素相較於其最強競爭者的平均優勢。對每一幀，我們計算目標 logit 與最高競爭 logit 之間的差值（或邊際）。這些邊際在整個片段的平均值表示目標音素與其他音素的分離程度。這有助於處理發音錯誤導致音素混淆的情況，而這種情況不一定會反映在機率分數中。

$$\mathrm{GOP}_{\mathrm{Margin}}\left(p\right) = \frac{1}{T}\sum_{t=t_1}^{t_2}\left(\mathbf{l}_t^{(p)} - \max_{k \neq p} \mathbf{l}_t^{(k)}\right)$$

2.2.3. GOP_{VarLogit}

This metric measures the variability of the model's confidence in predicting the target phoneme across timeframes.

此指標衡量模型在不同時間框架中對目標音素預測信心的變異性。

$$\text{GOP}_{\text{VarLogit}}(p) = \frac{1}{T}\sum_{t=t_1}^{t_2}\left(\mathbf{l}_t^{(p)} - \mu_p\right)^2, \quad \mu_p = \frac{1}{T}\sum_{t=t_1}^{t_2}\mathbf{l}_t^{(p)}$$

It is computed as the variance of the raw logit values associated with the target phoneme. A low logit variance suggests that the model consistently assigns similar confidence levels to the phoneme across frames, indicating a stable and confident recognition. Conversely, a high logit variance implies fluctuating confidence, which may occur due to acoustic distortions, coarticulation effects, or phonetic ambiguity.

它是目標音素相關原始 logit 值的變異數。低 logit 變異數表示模型在各時間框架中對該音素持續給予相似的信心水準，顯示出穩定且自信的辨識。相反地，高 logit 變異數則表示信心波動，可能因聲學失真、連音效應或語音模糊性所致。

2.2.4. GOP_{Combined}

This hybrid metric is designed to use the strengths of both logitbased and probability-based approaches to pronunciation assessment. It integrates the Mean Logit Margin, which quantifies the relative confidence of the target phoneme against competing phonemes, and the traditional GOP DNN.

此混合指標旨在結合基於 logit 與基於機率的發音評估方法的優點。它整合了平均 logit 邊際（Mean Logit Margin），用以量化目標音素相較於競爭音素的相對信心，以及傳統的 GOP DNN。

$$\text{GOP}_{\text{Combined}}(p) = \alpha \cdot \text{GOP}_{\text{Margin}}(p) - (1-\alpha) \cdot \text{GOP}_{\text{DNN}}(p)$$

where $\alpha \in [0,1]$ balances contributions. By combining these two metrics, the combined score may provide a more balanced assessment of pronunciation quality, mitigating the weaknesses of each individual measure. This hybrid approach ensures that both posterior probability ($via\text{GOP}_{\text{DNN}}$) and phoneme separability (via $\text{GOP}_{\text{margin}}$) contribute to the final score, making it more sensitive to pronunciation deviations while reducing the impact of softmax-related limitations.

其中 $\alpha \in [0,1]$ 用於平衡貢獻。透過結合這兩個指標，綜合分數能提供更均衡的發音品質評估，減輕各單一指標的弱點。此混合方法確保後驗機率（$via\text{GOP}_{\text{DNN}}$）與音素可分離性（透過 $\text{GOP}_{\text{margin}}$）皆對最終分數有所貢獻，使其對發音偏差更敏感，同時降低 softmax 相關限制的影響。

2.3. GOP Calculations 2.3. GOP 計算

In our study for forced alignment, we utilized the CTC segmentation algorithm [28], which uses an end-to-end CTC-based phoneme recognition model to determine phoneme boundaries. For the CTC-based acoustic model for phoneme recognition, we utilized an open-source fine-tuned phoneme recognition Wav2vec2.0 model [1] based on [29].

在本研究中，為了強制對齊，我們使用了 CTC 分割演算法 [28]，該演算法利用端對端的 CTC 基音素識別模型來判定音素邊界。針對基於 CTC 的音素識別聲學模型，我們採用了基於 [29] 的開源微調音素識別 Wav2vec2.0 模型 [1] 。

2.4. Datasets 2.4. 資料集

To address our RQ, we conducted experiments using two L2 English speech datasets: My Pronunciation Coach (MPC) [30] and SpeechOcean762 [31]. MPC comprises recordings of Dutch children speaking English, while the latter includes speech from Mandarin-speaking adults and children. Given its high acoustic variability-resulting from L1 transfer and inconsistent phoneme realizations [32]-non-native children's speech was a primary focus, as it presents a particularly challenging

testbed for pronunciation assessment.

為了回應我們的研究問題，我們使用兩個第二語言英語語音資料集進行實驗：My Pronunciation Coach (MPC) [30] 和 SpeechOcean762 [31]。MPC 包含荷蘭兒童說英語的錄音，而後者則包含以中文為母語的成人和兒童的語音。由於其高聲學變異性——源自第一語言轉移和不一致的音素實現 [32]——非母語兒童的語音成為主要焦點，因為它為發音評估提供了特別具挑戰性的測試環境。

MPC [30] contains speech from 124 Dutch secondary school students. Each recording includes 53 English words and 53 sentences covering a wide range of phonemes. Recordings are categorized into quality groups: Excellent, OK, Doubtful and Overloud. For this study, we used 50 OK and 21 Excellent sessions, totalling 3,130 utterances from 71 speakers (38 males, 33 females). Since MPC lacks annotated mispronunciations, we introduced simulated pronunciation errors by modify-

MPC [30] 包含 124 名荷蘭中學生的語音。每段錄音包含 53 個英語單字和 53 句涵蓋廣泛音素的句子。錄音依品質分為四組：Excellent、OK、Doubtful 和 Overloud。本研究使用了 50 個 OK 和 21 個 Excellent 的錄音，共計 71 位說話者（38 男，33 女）的 3,130 句話。由於 MPC 缺乏標註的誤讀，我們透過修改音素序列引入模擬的發音錯誤。

ing phoneme sequences. Common substitutions include replacing /ð/ $\rightarrow /\mathrm{d}/, /\theta/ \rightarrow /\mathrm{s}/, /$ æ/ $\rightarrow /\mathrm{e}/$, and diphthong simplifications such as /ei/ $\rightarrow /\mathrm{e}//$.

常見的替換包括將 /ð/ 替換為 $\rightarrow /\mathrm{d}/, /\theta/ \rightarrow /\mathrm{s}/, /$ ，/æ/ 替換為 $\rightarrow /\mathrm{e}/$ ，以及雙元音簡化，如 /ei/ 替換為 $\rightarrow /\mathrm{e}//$ 。

SpeechOcean762 [31] is an open-source corpus for pronunciation assessment, containing 5,000 English utterances from 250 native Mandarin speakers (125 adults, 125 children). Each utterance is annotated by five experts at the sentence, word, and phoneme levels, with 3,401 phonemes labelled as mispronunciations. We used all 5,000 utterances in our experiments. [2]

SpeechOcean762 [31] 是一個用於發音評估的開源語料庫，包含來自 250 位以普通話為母語的說話者（125 位成人，125 位兒童）的 5,000 句英語語音。每句語音由五位專家在句子、單字和音素層級進行標註，其中有 3,401 個音素被標記為誤讀。我們在實驗中使用了全部 5,000 句語音。 [2]

2.5. Evaluation Metrics 2.5. 評估指標

We assessed model performance using accuracy, precision, recall, F1-score, and Matthews Correlation Coefficient (MCC). Given the class imbalance in both datasets, we optimized the GOP threshold by selecting the percentile that maximized MCC. Additionally, we reported the ROC AUC score at this threshold to evaluate classification effectiveness.

我們使用準確率、精確率、召回率、F1 分數及 Matthews 相關係數（MCC）來評估模型表現。鑑於兩個資料集皆存在類別不平衡問題，我們透過選擇能最大化 MCC 的百分位數來優化 GOP 閾值。此外，我們在此閾值下報告 ROC AUC 分數，以評估分類效果。

In addition, in order to analyze GOP score distributions, we used violin plots to compare posterior probability-based, logit-based, and hybrid GOP scores across correct and mispronounced phonemes. This visualization helps determine whether a given GOP scoring method provides a clear phoneme separation, a key factor in pronunciation assessment.

另外，為了分析 GOP 分數分布，我們使用小提琴圖比較基於後驗機率、基於 logit 以及混合 GOP 分數在正確與誤讀音素間的差異。此視覺化有助於判斷特定 GOP 評分方法是否能提供清晰的音素區分，這是發音評估中的關鍵因素。

The Speechocean762 dataset includes human-annotated phoneme accuracy scores. Following prior research [31], we applied a second-order polynomial regression to model the relationship between GOP and human phoneme accuracy ratings. Performance was evaluated using Pearson Correlation Coefficient (PCC) and Mean Squared Error (MSE) to quantify prediction accuracy on the test set. Finally, phoneme-level mispronunciation error rates of the Speechocean762 dataset were also analyzed using a bar plot, comparing the GOP method with the highest PCC correlation to human-rated

phoneme accuracy.

Speechocean762 資料集包含人工標註的音素準確度分數。依照先前研究 [31]，我們應用二次多項式迴歸來建模 GOP 與人工音素準確度評分之間的關係。效能評估使用皮爾森相關係數（PCC）和均方誤差（MSE）來量化測試集上的預測準確度。最後，亦透過長條圖分析 Speechocean762 資料集的音素層級誤讀率，將 GOP 方法中與人工評分音素準確度相關性最高的結果進行比較。

3. Results 3. 結果

Table 1 presents the evaluation scores for posterior probability based (first column), logit-based (second to fourth columns) and hybrid GOP scores (last column) on the MPC dataset. Of all these measures, $\mathrm{GOP}_{\mathrm{Margin}}$ achieves the highest accuracy (0.851), MCC (0.347). It also outperforms other approaches in F1-score (0.415) and precision (0.347), ensuring better mispronunciation detection while maintaining precision. However, $\mathrm{GOP}_{\mathrm{MaxLogit}}$ achieves the highest AUC at $\mathrm{MCC}_{\max}$ (0.736), making it the most effective in distinguishing correctly pronounced and mispronounced phonemes. The measure GOP $_{\mathrm{DNN}}$ shows the highest recall (0.929) but has the lowest precision (0.184), indicating it detects most mispronunciations but lacks specificity. $\mathrm{GOP}_{\mathrm{MaxLogit}}$ shows moderate performance, achieving an accuracy of 0.590 and an MCC of 0.286 . Its performance is slightly better than $\mathrm{GOP}_{\mathrm{DNN}}$ but still lower than $\mathrm{GOP}_{\mathrm{Margin}}$ in most metrics. The GOP $_{\mathrm{Combined}}$ score achieves an accuracy of 0.82 and MCC of 0.314 , showing a good balance. Its AUC at MCCmax (0.704) is competitive, suggesting it may be a promising hybrid approach.

表 1 顯示了在 MPC 資料集上基於後驗機率（第一欄）、基於 logit（第二至第四欄）及混合 GOP 分數（最後一欄）的評估分數。在所有這些指標中，$\mathrm{GOP}_{\mathrm{Margin}}$ 達到最高的準確率（0.851）和 MCC（0.347）。它在 F1 分數（0.415）和精確度（0.347）上也優於其他方法，確保在維持精確度的同時能更好地檢測誤讀。然而，$\mathrm{GOP}_{\mathrm{MaxLogit}}$ 在 AUC 上達到最高值 $\mathrm{MCC}_{\max}$ （0.736），使其在區分正確發音與誤讀音素方面最為有效。GOP $_{\mathrm{DNN}}$ 指標顯示最高的召回率（0.929），但精確度最低（0.184），表示它能檢測大多數誤讀，但缺乏特異性。$\mathrm{GOP}_{\mathrm{MaxLogit}}$ 表現中等，準確率為 0.590，MCC 為 0.286。其表現略優於 $\mathrm{GOP}_{\mathrm{DNN}}$ ，但在大多數指標上仍低於 $\mathrm{GOP}_{\mathrm{Margin}}$ 。GOP $_{\mathrm{Combined}}$ 分數達到準確率 0.82 和 MCC 0.314，展現良好的平衡。其在 MCCmax 時的 AUC（0.704）具競爭力，顯示它可能是一種有前景的混合方法。

Table 2 shows the evaluation results of different GOPbased pronunciation assessment scores of the SpeechOcean762 dataset. This table also includes PCC and MSE scores since the SpeechOcean762 dataset includes human-annotated phoneme accuracy ratings, allowing us to evaluate the correlation between GOP and expert scores. GOP $_{\mathrm{DNN}}$ achieves the highest

表 2 顯示了 SpeechOcean762 資料集中不同基於 GOP 的發音評估分數的評估結果。此表亦包含 PCC 與 MSE 分數，因為 SpeechOcean762 資料集包含人工標註的音素準確度評分，使我們能評估 GOP 與專家分數之間的相關性。GOP $_{\mathrm{DNN}}$ 達到最高

Table 1: Performance analysis on the MPC dataset

表 1：MPC 資料集上的效能分析

	$\mathrm{GOP}_{\mathrm{DNN}}$	$\mathrm{GOP}_{\mathrm{MaxLogit}}$	$\mathrm{GOP}_{\mathrm{Margin}}$	$\mathrm{GOP}_{\mathrm{VarLogit}}$	$\mathrm{GOP}_{\mathrm{Combined}}$
Accuracy 準確率	0.572	0.590	0.851	0.461	0.820
Precision 精確度	0.184	0.189	0.347	0.146	0.297
Recall 召回率	0.929	0.919	0.515	0.882	0.559
F1	0.307	0.314	0.415	0.250	0.388
MCC	0.279	0.286	0.342	0.184	0.314
AUC MCC $_{\max}$	0.730	0.736	0.702	0.648	0.704

accuracy (0.947), precision (0.333), and MCC (0.367), showing that it effectively separates correctly pronounced and mispronounced phonemes. However, its PCC scores (0.278 for low confidence and 0.295 for high confidence) are significantly lower than other logit-based approaches. This suggests that while $\mathrm{GOP}_{\mathrm{DNN}}$ can classify pronunciation errors well, it does not align well with human-annotated phoneme scores, making it less reliable for subjective pronunciation assessment. GOP $_{\mathrm{MaxLogit}}$ achieves the highest PCC scores (0.442 for low confidence and 0.456 for high confidence), outperforming all other GOP metrics in correlating with human ratings. This indicates that maximum logit values capture pronunciation quality in a way that aligns better with human perception compared to posterior probability-based GOP. It

also achieves a strong AUC at $\mathrm{MCC}_{\max}$ (0.754), reinforcing its reliability as a GOP metric. $\mathrm{GOP}_{\mathrm{Margin}}$ shows the weakest overall performance, with an MCC of only 0.174 and lower PCC scores (0.173 for low confidence and 0.191 for high confidence).

準確率（0.947）、精確率（0.333）和 MCC（0.367），顯示其能有效區分正確發音與錯誤發音的音素。然而，其 PCC 分數（低信心為 0.278，高信心為 0.295）顯著低於其他基於 logit 的方法。這表明雖然 $\mathrm{GOP}_{\mathrm{DNN}}$ 能夠良好分類發音錯誤，但與人工標註的音素分數對齊度不佳，使其在主觀發音評估中可靠性較低。GOP $\mathrm{MaxLogit}$ 達到最高的 PCC 分數（低信心為 0.442，高信心為 0.456），在與人工評分的相關性上優於所有其他 GOP 指標。這表示最大 logit 值以更符合人類感知的方式捕捉發音品質，相較於基於後驗機率的 GOP 更具優勢。它在 $\mathrm{MCC}_{\max}$ 也達到強勁的 AUC（0.754），進一步強化其作為 GOP 指標的可靠性。 $\mathrm{GOP}_{\mathrm{Margin}}$ 整體表現最弱，MCC 僅為 0.174，且 PCC 分數較低（低信心為 0.173，高信心為 0.191）。

Table 2: Performance analysis on the SpeechOcean762 dataset

表 2：SpeechOcean762 資料集上的性能分析

	$\mathrm{GOP}_{\mathrm{DNN}}$	$\mathrm{GOP}_{\mathrm{MaxLogit}}$	$\mathrm{GOP}_{\mathrm{Margin}}$	GOP $^{\mathrm{VarLogit}}$	$\mathrm{GOP}_{\mathrm{Combined}}$
Accuracy 準確率	0.947	0.925	0.741	0.894	0.843
Precision 精確度	0.333	0.257	0.089	0.195	0.139
Recall 召回率	0.466	0.571	0.672	0.621	0.642
F1	0.388	0.354	0.157	0.297	0.228
MCC	0.367	0.350	0.174	0.308	0.247
AUC MCC $_{\max}$	0.715	0.754	0.708	0.763	0.747
PCC (low conf) PCC（低信心）	0.278	0.442	0.173	0.341	0.303
PCC (high conf) PCC（高信心）	0.295	0.456	0.191	0.357	0.319
MSE	0.124	0.109	0.131	0.120	0.123

While margin-based GOP obtained most of the best metric scores in the MPC dataset, it does not generalize well to SpeechOcean762, possibly due to differences in speaker demographics and phoneme variability. In Table 2 we see that $\mathrm{GOP}_{\mathrm{VarLogit}}$ achieves high recall (0.621) but has weak MCC (0.308) and low precision (0.195), suggesting it works well for detecting mispronunciations but lacks robustness in classification. $\mathrm{GOP}_{\mathrm{Combined}}$ achieves an MCC of 0.247 and PCC scores (0.303 for low confidence and 0.319 for high confidence), indicating a balance between probability and logit-based features.

雖然基於邊際的 GOP 在 MPC 資料集中獲得了大多數最佳指標分數，但它在 SpeechOcean762 上的泛化能力不佳，可能是由於說話者人口統計和音素變異性的差異。在表 2 中，我們看到 $\mathrm{GOP}_{\mathrm{VarLogit}}$ 達到高召回率（0.621），但 MCC 較弱（0.308）且精確度低（0.195），這表明它在檢測誤讀方面表現良好，但分類的穩健性不足。 $\mathrm{GOP}_{\mathrm{Combined}}$ 達到 MCC 為 0.247，PCC 分數（低信心為 0.303，高信心為 0.319），顯示在機率和 logit 基礎特徵之間取得平衡。

To better interpret the performance results, Figure 2 visualizes the distribution of correctly pronounced and mispronounced phonemes across both datasets. $\mathrm{GOP}_{\mathrm{DNN}}$ shows a wide distribution overlap between correct and mispronounced phonemes in both cases (both graphs of Figure 2), with close means and high variance, making it less effective for error distinction. In contrast, $\mathrm{GOP}_{\mathrm{MaxLogit}}$ achieves better separation, especially in MPC (first graph of Figure 2), though overlap increases in SpeechOcean762 (second graph of Figure 2), leading to higher variability. The best-performing score in MPC, GOP $^{\mathrm{Margin}}$, effectively classifies children's speech (first graph of Figure 2), but struggles with greater distribution overlap in SpeechOcean762 (second graph of Figure 2). $\mathrm{GOP}_{\mathrm{VarLogit}}$ shows high variability and significant overlap in both cases, making it unreliable. $\mathrm{GOP}_{\mathrm{Combined}}$ balances probability- and

為了更好地解釋性能結果，圖 2 視覺化了兩個資料集中正確發音和誤發音音素的分佈。 $\mathrm{GOP}_{\mathrm{DNN}}$ 在兩種情況下（圖 2 的兩個圖表）均顯示正確與誤發音音素分佈重疊廣泛，平均值接近且變異度高，使其在錯誤區分上效果較差。相較之下， $\mathrm{GOP}_{\mathrm{MaxLogit}}$ 達成了較佳的分離效果，尤其是在 MPC（圖 2 的第一個圖表）中，儘管在 SpeechOcean762（圖 2 的第二個圖表）中重疊增加，導致變異度較高。MPC 中表現最佳的分數 GOP $^{\mathrm{Margin}}$ 有效地分類了兒童語音（圖 2 的第一個圖表），但在 SpeechOcean762 中因分佈重疊較大而表現不佳（圖 2 的第二個圖表）。 $\mathrm{GOP}_{\mathrm{VarLogit}}$ 在兩種情況下均顯示高變異度和顯著重疊，因而不可靠。 $\mathrm{GOP}_{\mathrm{Combined}}$ 在機率與

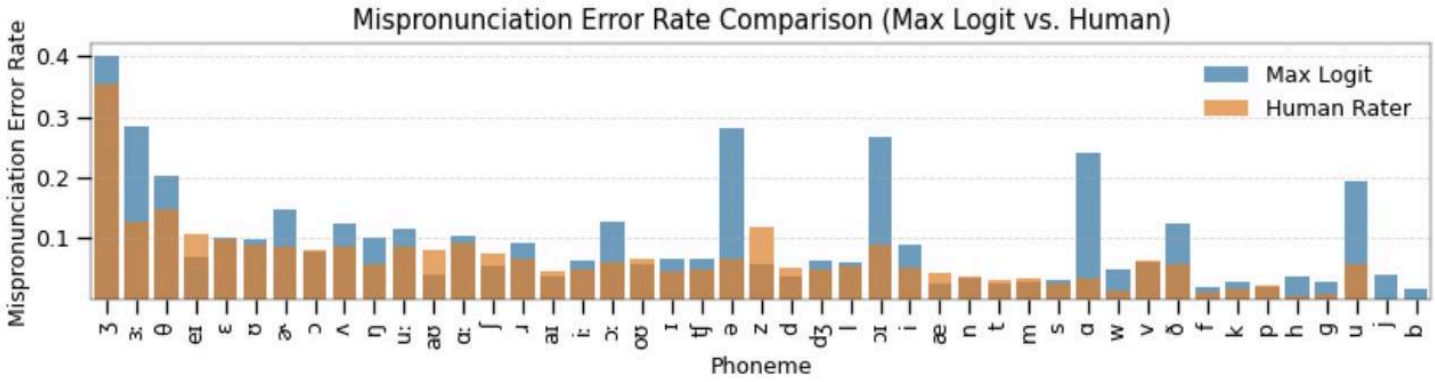

Figure 1: Comparison of mispronunciation error rates by phoneme (Max Logit vs. human rater) in the SpeechOcean762 dataset. Blue bars show GOP $^{\text{MaxLogit}}$ -predicted error rates, while red bars indicate human-rated phoneme accuracy.

圖 1：SpeechOcean762 資料集中按音素比較誤發音錯誤率（最大 Logit 與人工評分）。藍色條顯示 GOP $^{\text{MaxLogit}}$ 預測的錯誤率，紅色條則表示人工評分的音素準確率。

logit-based approaches across both datasets, reducing overlap compared to GOP_{DNN}. It presents a trade-off between classification accuracy from posterior probabilities and human correlation from logit-based scores. While we incorporated GOP $^{\text{Margin}}$ in $\text{GOP}_{\text{Combined}}$, $\text{GOP}_{\text{MaxLogit}}$ could also be considered, leaving the optimal choice undecided.

基於 logit 的方法在兩個資料集上均降低了與 GOP_{DNN} 的重疊。這在後驗機率的分類準確度與基於 logit 分數的人類相關性之間呈現出一種權衡。雖然我們在 $\text{GOP}_{\text{Combined}}$, $\text{GOP}_{\text{MaxLogit}}$ 中納入了 GOP $^{\text{Margin}}$ ，但也可考慮其他方法，最佳選擇尚未決定。

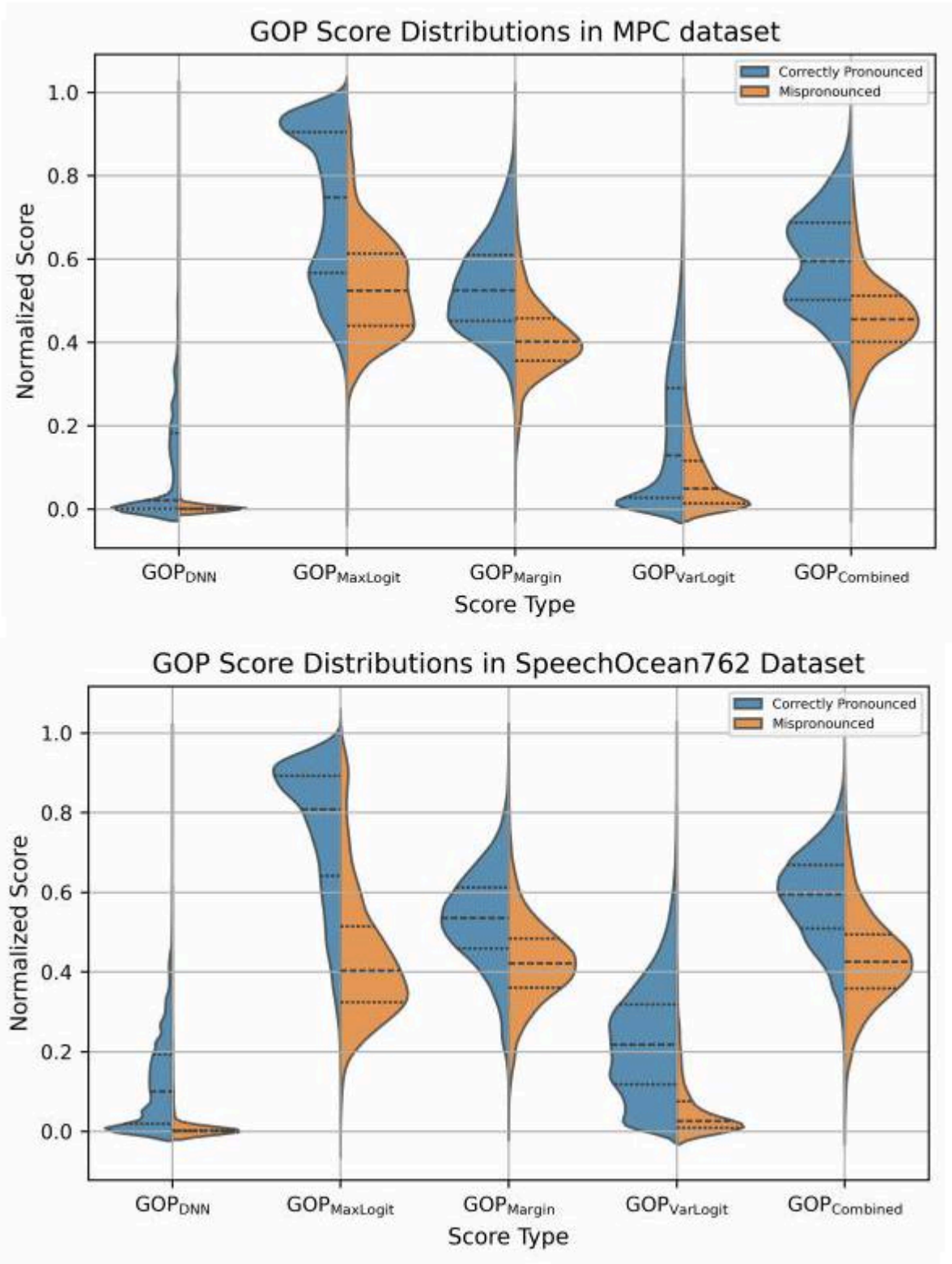

Figure 2: Comparison of GOP score distributions across MPC and Speechocean762 datasets.

圖 2：MPC 與 Speechocean762 資料集中 GOP 分數分佈的比較。

Finally, to analyze the alignment between GOP-based mispronunciation detection and human-rated phoneme accuracy, we investigated whether the GOP scoring method with the highest correlation to human ratings, $\mathrm{GOP}_{\mathrm{MaxLogit}}$, effectively identifies mispronounced phonemes. Figure 1 compares phonemelevel mispronunciation error rates predicted by $\mathrm{GOP}_{\mathrm{MaxLogit}}$ with human annotator ratings from the SpeechOcean762 dataset. Discrepancies were further examined by computing the difference in mispronunciation rates between GOP $^{\mathrm{MaxLogit}}$ predictions and human ratings, highlighting

phonemes where the model was overconfident, underconfident, or well-aligned (Figure 1).

最後，為了分析基於 GOP 的誤讀偵測與人類評分音素準確度之間的一致性，我們調查了與人類評分相關性最高的 GOP 計分方法 $\mathrm{GOP}_{\mathrm{MaxLogit}}$ 是否能有效識別誤讀音素。圖 1 比較了由 $\mathrm{GOP}_{\mathrm{MaxLogit}}$ 預測的音素層級誤讀錯誤率與 SpeechOcean762 資料集的人類標註評分。透過計算 GOP $\mathrm{MaxLogit}$ 預測與人類評分之間誤讀率的差異，進一步檢視了模型過度自信、信心不足或與人類評分良好對齊的音素（圖 1）。

Phonemes where the GOP $\mathrm{MaxLogit}$ overestimates mispronunciations (meaning the model assigns significantly higher mispronunciation rates than human raters) include (top 5): / / / / / a/, $/\mathrm{oI}/$, $/\mathrm{s:}/$, and $/\mathrm{u}/$. These phonemes are frequently flagged as mispronounced by the model, despite human raters considering them correctly pronounced in most instances. This suggests

GOP $\mathrm{MaxLogit}$ 高估誤讀的音素（意指模型分配的誤讀率顯著高於人工評審）包括（前五名）：/ / / / / a/， $/\mathrm{oI}/$, $/\mathrm{s:}/$ ，以及 $/\mathrm{u}/$ 。這些音素經常被模型標記為誤讀，儘管人工評審在大多數情況下認為它們發音正確。這顯示

that $\mathrm{GOP}_{\mathrm{MaxLogit}}$ is overly confident for these phonemes, possibly misinterpreting slight articulatory variations as errors. Conversely, phonemes where $\mathrm{GOP}_{\mathrm{MaxLogit}}$ underestimates mispronunciations (bottom 5), meaning human raters perceive more errors than the model detects, include /æ/, /ʃ/, /eɪ/, /av/, and /z/. Some phonemes exhibit strong agreement between $\mathrm{GOP}_{\mathrm{MaxLogit}}$ predictions and human-rated error rates, indicating well-aligned phoneme classification. These phonemes are /b/, /t//, /iː/, /dʒ/, and /aː/.

$\mathrm{GOP}_{\mathrm{MaxLogit}}$ 對這些音素過於自信，可能將輕微的發音變異誤判為錯誤。相反地， $\mathrm{GOP}_{\mathrm{MaxLogit}}$ 低估誤讀的音素（後五名），意指人工評審察覺的錯誤多於模型偵測到的錯誤，包含 /æ/、/ʃ/、/eɪ/、/av/ 和 /z/。部分音素在 $\mathrm{GOP}_{\mathrm{MaxLogit}}$ 預測與人工評審錯誤率間展現高度一致，顯示音素分類相當準確。這些音素為 /b/、/t//、/iː/、/dʒ/ 和 /aː/。

4. Discussion and Conclusion

4. 討論與結論

In this work, we have analyzed differences in probability-based and logit-based GOP for pronunciation assessment across two datasets, MPC and SpeechOcean762. To answer our RQ, our findings indicate that logit-based methods achieve a better classification performance than probability-based GOP; however, their effectiveness depends on the characteristics of the dataset. GOP DNN consistently obtains high recall, but low precision, indicating its tendency to over-detect mispronunciations, as seen in its high overlap between correctly pronounced and mispronounced phonemes. In contrast, logit-based methods ($\mathrm{GOP}_{\mathrm{MaxLogit}}$) demonstrate better phoneme separation (Fig. 2).

在本研究中，我們分析了基於機率與基於對數幾率（logit）的 GOP 在兩個資料集 MPC 和 SpeechOcean762 中用於發音評估的差異。針對我們的研究問題（RQ），結果顯示基於 logit 的方法在分類表現上優於基於機率的 GOP；然而，其效能取決於資料集的特性。GOP DNN 持續獲得高召回率，但精確度較低，顯示其傾向於過度偵測誤讀，這可從正確發音與誤讀音素間的高度重疊看出。相較之下，基於 logit 的方法（ $\mathrm{GOP}_{\mathrm{MaxLogit}}$ ）展現出更佳的音素分離效果（圖 2）。

A key insight is that $\mathrm{GOP}_{\mathrm{MaxLogit}}$ aligns best with human ratings, achieving the highest PCC scores (Table 1), while $\mathrm{GOP}_{\mathrm{DNN}}$, despite strong classification performance, does not correlate well with expert judgments (Table 2). This suggests that maximum logit values better capture pronunciation quality from a perceptual standpoint. GOP $\mathrm{varLogit}$ shows high variability (Table 1 and 2), making it less reliable, while $\mathrm{GOP}_{\mathrm{Combined}}$ balances probability and logit-based information but still shows some overlap.

一個重要的見解是， $\mathrm{GOP}_{\mathrm{MaxLogit}}$ 與人工評分的對應度最高，達到最高的 PCC 分數（表 1），而 $\mathrm{GOP}_{\mathrm{DNN}}$ 雖然分類表現強勁，卻與專家判斷的相關性不佳（表 2）。這表明最大 logit 值從感知角度更能捕捉發音品質。GOP $\mathrm{varLogit}$ 顯示出高度變異性（表 1 與表 2），使其可靠性較低，而 $\mathrm{GOP}_{\mathrm{Combined}}$ 則在機率與 logit 資訊間取得平衡，但仍存在部分重疊。

To the best of our knowledge, the highest PCC score reported on the SpeechOcean762 dataset is 0.69 [33]. However, this was achieved using a multidimensional MDD model that calculates GOP scores while incorporating additional aspects of speech in the SpeechOcean762 dataset. In contrast, our approach focuses solely on logit-based GOP scoring, making direct comparisons with these other methodologies challenging. Future research should focus on reducing reliance on forced alignment, which can introduce errors due to acoustic variability in child and non-native speech, contributing to high recall but low precision.

據我們所知，SpeechOcean762 資料集上報告的最高 PCC 分數為 0.69 [33]。然而，該成績是使用一個多維度的 MDD 模型達成，該模型在計算 GOP 分數時結合了 SpeechOcean762 資料集中語音的其他方面。相較之下，我們的方法僅專注於基於 logit 的 GOP 評分，因此與這些其他方法直接比較存在困難。未來的研究應著重於減少對強制對齊的依賴，因為強制對齊可能因兒童及非母語語音的聲學變異性而引入錯誤，導致高召回率但精確度低。

In conclusion, our logit-based methodology offers a modelagnostic framework for any CTC-based acoustic model using a threshold-based approach. However, results vary across datasets, with $\text{GOP}_{\text{Margin}}$ performing best on MPC and GOP $_{\text{MaxLogit}}$ on SpeechOcean762, underscoring the role of acoustic variability.

總結來說，我們的基於 logit 的方法為任何基於 CTC 的聲學模型提供了一個模型無關的框架，採用閾值式方法。然而，結果因資料集而異，其中 $\text{GOP}_{\text{Margin}}$ 在 MPC 上表現最佳，而 GOP $_{\text{MaxLogit}}$ 在 SpeechOcean762 上表現最佳，凸顯了聲學變異性的影響。

5. Acknowledgements 5. 致謝

This publication is part of the project Responsible AI for Voice Diagnostics (RAIVD) with file number NGF.1607.22.013 of the research programme NGF AiNed Fellowship Grants which is financed by the Dutch Research Council (NWO).

本出版物為負責任語音診斷 AI（Responsible AI for Voice Diagnostics, RAIVD）計畫的一部分，計畫編號為 NGF.1607.22.013，隸屬於由荷蘭研究委員會（NWO）資助的 NGF AiNed Fellowship Grants 研究計畫。

6. References 6. 參考文獻

[1] A. Kuschel, N. Hansen, L. Heyse, and R. P. Wittek, “Combining language training and work experience for refugees with lowliteracy levels: a mixed-methods case study,” Journal of International Migration and Integration, vol. 24, no. 4, pp. 1635-1661, 2023.

[1] A. Kuschel, N. Hansen, L. Heyse, 和 R. P. Wittek，「結合語言訓練與工作經驗於低識字率難民：一項混合方法個案研究」，《國際移民與融合期刊》，第 24 卷，第 4 期，頁 1635-1661，2023 年。

[2] R. Walker, E.-L. Low, and J. Setter, “English pronunciation for a global world,” Oxford, October 2021, Last visited: 2025-02-10. [Online]. Available: https://centaur.reading.ac.uk/101017/

[2] R. Walker, E.-L. Low, 和 J. Setter，「全球化世界的英語發音」，牛津，2021 年 10 月，最後訪問日期：2025-02-10。[線上]。可取得於：https://centaur.reading.ac.uk/101017/

[3] J. Jenkins, The phonology of English as an international language. Oxford University Press, 2000.

[3] J. Jenkins，《作為國際語言的英語音韻學》。牛津大學出版社，2000 年。

[4] J. Zoss, “What do adult english learners say about their pronunciation and linguistic self-confidence?” MinneTESOL Journal, 2016.

[4] J. Zoss，「成人英語學習者如何看待他們的發音與語言自信？」MinneTESOL Journal，2016。

[5] H.-W. Hsu, “An examination of automatic speech recognition (asr)-based computer-assisted pronunciation training (capt) for less-proficient efl students using the technology acceptance model,” International Journal of Technology in Education, vol. 7, no. 3, pp. 456-473, 2024.

[5] H.-W. Hsu，「基於自動語音識別（ASR）的電腦輔助發音訓練（CAPT）對較不熟練英語為外語學生之技術接受模型檢驗」，International Journal of Technology in Education，第 7 卷，第 3 期，頁 456-473，2024。

[6] M. Amrate and P. hua Tsai, “Computer-assisted pronunciation training: A systematic review,” ReCALL, no. 1, pp. 22-42, 2024.

[6] M. Amrate 與 P. hua Tsai，「電腦輔助發音訓練：系統性回顧」，ReCALL，第 1 期，頁 22-42，2024。

[7] N. Alrashoudi, H. Al-Khalifa, and Y. Alotaibi, “Improving mispronunciation detection and diagnosis for non-native learners of the arabic language,” Discover Computing, vol. 28, no. 1, p. 1, 2025.

[7] N. Alrashoudi、H. Al-Khalifa 與 Y. Alotaibi，「提升阿拉伯語非母語學習者的誤發音偵測與診斷」，Discover Computing，第 28 卷，第 1 期，頁 1，2025。

[8] X. Cao, Z. Fan, T. Svendsen, and G. Salvi, “A Framework for Phoneme-Level Pronunciation Assessment Using CTC,” in Interspeech 2024, 2024, pp. 302-306.

[8] X. Cao, Z. Fan, T. Svendsen, 和 G. Salvi，「使用 CTC 的音素層級發音評估框架」，發表於 Interspeech 2024，2024，頁 302-306。

[9] S. M. Witt, “Use of speech recognition in computer-assisted language learning.” Ph.D. dissertation, University of Cambridge, 2000.

[9] S. M. Witt，「語音識別在電腦輔助語言學習中的應用」，劍橋大學博士論文，2000。

[10] S. Kanters, C. Cucchiarini, and H. Strik, “The goodness of pronunciation algorithm: a detailed performance study,” in Speech and Language Technology in Education (SLaTE 2009), 2009, pp. 49-52.

[10] S. Kanters, C. Cucchiarini, 和 H. Strik，「發音優劣算法：詳細的性能研究」，發表於 Speech and Language Technology in Education (SLaTE 2009)，2009，頁 49-52。

[11] J. van Doremalen, C. Cucchiarini, and H. Strik, “Using non-native error patterns to improve pronunciation verification,” in Interspeech 2010, 2010, pp. 590-593.

[11] J. van Doremalen, C. Cucchiarini, 和 H. Strik，「利用非母語錯誤模式改進發音驗證」，發表於 Interspeech 2010，2010，頁 590-593。

[12] Y. Song, W. Liang, and R. Liu, “Lattice-based gop in automatic pronunciation evaluation,” in 2010 The 2nd International Conference on Computer and Automation Engineering (ICCAE), vol. 3, 2010, pp. 598-602.

[12] Y. Song, W. Liang, 和 R. Liu，「基於格點的自動發音評估中的 GOP」，收錄於 2010 年第二屆國際計算機與自動化工程會議（ICCAE），第 3 卷，2010 年，第 598-602 頁。

[13] J. Shi, N. Huo, and Q. Jin, “Context-aware goodness of pronunciation for computer-assisted pronunciation training,” in Interspeech 2020, 2020, pp. 3057-3061.

[13] J. Shi, N. Huo, 和 Q. Jin，「具上下文感知的發音優劣評估用於電腦輔助發音訓練」，收錄於 Interspeech 2020，2020 年，第 3057-3061 頁。

[14] H. Do, W. Lee, and G. G. Lee, “Acoustic feature mixup for balanced multi-aspect pronunciation assessment,” CoRR, vol. abs/2406.15723, 2024. [Online]. Available: https://doi.org/10. 48550/arXiv.2406.15723

[14] H. Do, W. Lee, 和 G. G. Lee，「用於平衡多面向發音評估的聲學特徵混合」，CoRR，卷 abs/2406.15723，2024 年。[線上]。可取得：https://doi.org/10.48550/arXiv.2406.15723

[15] Y. Gong, Z. Chen, I.-H. Chu, P. Chang, and J. Glass, “Transformer-based multi-aspect multi-granularity non-native english speaker pronunciation assessment,” in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 7262-7266.

[15] Y. Gong, Z. Chen, I.-H. Chu, P. Chang, 和 J. Glass，「基於 Transformer 的多面向多粒度非母語英語發音評估」，收錄於 ICASSP 2022-2022 IEEE 國際聲學、語音與訊號處理會議（ICASSP），2022 年，第 7262-7266 頁。

[16] Y. El Kheir, “Mispronunciation detection with speechblender data augmentation pipeline,” PhD Thesis Report, KTH Royal Institute of Technology, Stockholm, Sweden, 2023, Last visited: 2025-0210. [Online]. Available: https://urn.kb.se/resolve?urn=urn:nbn:se: kth:diva-339940

[16] Y. El Kheir，「使用 speechblender 資料增強流程的誤讀偵測」，博士論文報告，瑞典斯德哥爾摩 KTH 皇家理工學院，2023，最後訪問時間：2025-02-10。[線上]。可取得：https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-339940

[17] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, “Xls-r: Self-supervised cross-lingual speech representation learning at scale,” in Interspeech 2022, 2022, pp. 2278-2282.

[17] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, 及 M. Auli，「Xls-r：大規模自監督跨語言語音表示學習」，發表於 Interspeech 2022，2022，頁 2278-2282。

[18] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” IEEE/ACM Trans. Audio, Speech and Lang. Proc., vol. 29, p. 3451-3460, Oct. 2021. [Online]. Available: https://doi.org/10.1109/TASLP.2021.3122291

[18] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, 及 A. Mohamed，「Hubert：透過隱藏單元遮蔽預測的自監督語音表示學習」，IEEE/ACM 音訊、語音與語言處理匯刊，卷 29，頁 3451-3460，2021 年 10 月。[線上]。可取得：https://doi.org/10.1109/TASLP.2021.3122291

[19] S. Sudhakara, M. K. Ramanathi, C. Yarra, and P. K. Ghosh, “An improved goodness of pronunciation (gop) measure for pronunciation evaluation with dnn-hmm system considering hmm transition probabilities.” in INTERSPEECH, vol. 2, 2019, pp. 954-958.

[19] S. Sudhakara, M. K. Ramanathi, C. Yarra, 及 P. K. Ghosh，「考慮 HMM 過渡機率的 DNN-HMM 系統中改良的發音優劣度（GOP）評估指標」，發表於 INTERSPEECH，卷 2，2019，頁 954-958。

[20] G. Pereyra, G. Tucker, J. Chorowski, L. Kaiser, and G. Hinton, “Regularizing neural networks by penalizing confident output distributions,” 2017. [Online]. Available: https://openreview.net/ forum?id=HkCjNI5ex

[20] G. Pereyra, G. Tucker, J. Chorowski, L. Kaiser, 和 G. Hinton，「透過懲罰自信輸出分布來正則化神經網路」，2017 年。[線上]。可取得於：https://openreview.net/forum?id=HkCjNI5ex

[21] H. Wei, R. Xie, H. Cheng, L. Feng, B. An, and Y. Li, “Mitigating neural network overconfidence with logit normalization,” in International conference on machine learning. PMLR, 2022, pp. 23631-23644.

[21] H. Wei, R. Xie, H. Cheng, L. Feng, B. An, 和 Y. Li，「利用 logit 正規化減輕神經網路過度自信」，收錄於國際機器學習會議論文集。PMLR，2022 年，第 23631-23644 頁。

[22] X. Xie and T. F. Jaeger, “Comparing non-native and native speech: Are 12 productions more variable?” The Journal of the Acoustical Society of America, vol. 147, no. 5, pp. 3322-3347, 2020.

[22] X. Xie 和 T. F. Jaeger，「比較非母語與母語語音：12 個發音是否更具變異性？」《美國聲學學會期刊》，第 147 卷，第 5 期，3322-3347 頁，2020 年。

[23] J. L. Preston, J. R. Irwin, and J. Turcios, “Perception of speech sounds in school-aged children with speech sound disorders,” in Seminars in speech and language, vol. 36, no. 04. Thieme Medical Publishers, 2015, pp. 224-233.

[23] J. L. Preston, J. R. Irwin, 和 J. Turcios，「學齡兒童語音障礙者對語音聲音的感知」，收錄於《語言與語音研討會》，第 36 卷，第 04 期。Thieme Medical Publishers，2015 年，224-233 頁。

[24] X. Li, X. Li, D. Pan, and D. Zhu, “On the learning property of logistic and softmax losses for deep neural networks,” in

Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 04, 2020, pp. 4739-4746.

[24] X. Li, X. Li, D. Pan, 和 D. Zhu，「關於深度神經網路中邏輯斯蒂和 softmax 損失的學習特性」，發表於 AAAI 人工智慧會議論文集，第 34 卷，第 04 期，2020 年，頁 4739-4746。

[25] J. Weng, Z. Luo, S. Li, N. Sebe, and Z. Zhong, “Logit margin matters: Improving transferable targeted adversarial attack by logit calibration,” IEEE Transactions on Information Forensics and Security, vol. 18, pp. 3561-3574, 2023.

[25] J. Weng, Z. Luo, S. Li, N. Sebe, 和 Z. Zhong，「Logit 邊際的重要性：透過 logit 校準提升可轉移目標對抗攻擊」，IEEE 資訊鑑識與安全期刊，第 18 卷，頁 3561-3574，2023 年。

[26] E. J. Yeo, K. Choi, S. Kim, and M. Chung, “Speech intelligibility assessment of dysarthric speech by using goodness of pronunciation with uncertainty quantification,” in Interspeech 2023, 2023, pp. 166-170.

[26] E. J. Yeo, K. Choi, S. Kim, 和 M. Chung，「利用發音優良度及不確定性量化評估構音障礙語音的可懂度」，發表於 Interspeech 2023，2023 年，頁 166-170。

[27] W. Hu, Y. Qian, and F. K. Soong, “A new DNN-based high quality pronunciation evaluation for computer-aided language learning (CALL),” in Interspeech, 2013, pp. 1886-1890.

[27] W. Hu, Y. Qian, 和 F. K. Soong，「一種基於 DNN 的高品質發音評估，用於電腦輔助語言學習（CALL）」，發表於 Interspeech，2013 年，頁 1886-1890。

[28] L. Kürzinger, D. Winkelbauer, L. Li, T. Watzel, and G. Rigoll, “Ctc-segmentation of large corpora for german end-to-end speech recognition,” in International Conference on Speech and Computer. Springer, 2020, pp. 267-278.

[28] L. Kürzinger, D. Winkelbauer, L. Li, T. Watzel, 和 G. Rigoll，「德語端對端語音識別的大型語料庫 CTC 分段」，收錄於國際語音與計算機會議，Springer，2020，頁 267-278。

[29] Q. Xu, A. Baevski, and M. Auli, “Simple and effective zero-shot cross-lingual phoneme recognition,” in Interspeech 2022, 2022, pp. 2113-2117.

[29] Q. Xu, A. Baevski, 和 M. Auli，「簡單且有效的零樣本跨語言音素識別」，收錄於 Interspeech 2022，2022，頁 2113-2117。

[30] C. Cucchiarini, W. Nejjari, and H. Strik, “My pronunciation coach: Improving english pronunciation with an automatic coach that listens,” Language Learning in Higher Education, vol. 1, no. 2, pp. 365-376, 2012.

[30] C. Cucchiarini, W. Nejjari, 和 H. Strik，「我的發音教練：利用自動教練聆聽來改善英語發音」，《高等教育語言學習》，第 1 卷，第 2 期，頁 365-376，2012。

[31] J. Zhang, Z. Zhang, Y. Wang, Z. Yan, Q. Song, Y. Huang, K. Li, D. Povey, and Y. Wang, “speechocean762: An open-source nonnative english speech corpus for pronunciation assessment,” in Interspeech 2021, 2021, pp. 3710-3714.

[31] J. Zhang, Z. Zhang, Y. Wang, Z. Yan, Q. Song, Y. Huang, K. Li, D. Povey, 和 Y. Wang，「speechocean762：一個用於發音評估的開源非母語英語語音語料庫」，收錄於 Interspeech 2021，2021，頁 3710-3714。

[32] R. Gretter, M. Matassoni, D. Falavigna, A. Misra, C. Leong, K. Knill, and L. Wang, “Etlt 2021: Shared task on automatic speech recognition for non-native children’s speech,” in Interspeech 2021, 2021, pp. 3845-3849.

[32] R. Gretter, M. Matassoni, D. Falavigna, A. Misra, C. Leong, K. Knill, 和 L. Wang，「ETLT 2021：非母語兒童語音自動識別共享任務」，發表於 Interspeech 2021，2021，頁 3845-3849。

[33] F.-A. Chao, T.-H. Lo, T.-I. Wu, Y.-T. Sung, and B. Chen, “A hierarchical context-aware modeling approach for multi-

aspect and multi-granular pronunciation assessment," in Interspeech 2023, 2023, pp. 974-978.

[33] F.-A. Chao, T.-H. Lo, T.-I. Wu, Y.-T. Sung, 和 B. Chen，「一種分層上下文感知模型方法，用於多面向及多粒度的發音評估」，發表於 Interspeech 2023，2023，頁 974-978。

1 https://huggingface.co/facebook/wav2vec2-xlsr-53-espeak-cv-ft

2 https://github.com/Aditya3107/GOP_logit.git