改善多細粒度的發音評測上資料不平衡的問題
Addressing the issue of Data Imbalance in Multi-granularity Pronunciation Assessment

Meng-Shin Lin 林孟欣, Hsin-Wei Wang 王馨偉, Tien-Hong Lo 羅天宏, Berlin Chen 陳柏琳
國立臺灣師範大學資訊工程學系
Department of Computer Science and Engineering,
National Taiwan Normal University
{61147077s, hsinweiwang, teinhonglo, berlin} @ntnu.edu.tw

Wei-Cheng Chao 趙偉成
中華電信研究院前瞻科技研究所
Advanced Technology Laboratory, Telecommunication Laboratories,
Chunghwa Telecom Co., Ltd., Taiwan
weicheng@cht.com.tw

摘要

自動發音評測 (Automatic Pronunciation Assessment, APA)是在量化非母語(L2)學習者在某種語言中發音的熟練程度。然而隨著技術的發展 APA 已經可以評測多個發音細粒度如音素層級、單字層級和語句層級及發音準確度、流利度、重音等多個面向。然而目前的 APA 方法使用均方誤差 (Mean Squard Error, MSE)損失函數，但在每個細粒度的標籤都存在資料高度不平衡的問題，這會影響模型的泛化能力和公平性，MSE 會低估稀有的標籤，但現有的研究卻很少涉及數據不平衡的問題。因此在本研究中，我們參考了在視覺分類建模中使用的類平衡損失函數 ，使用重新採樣的方式及加入一個可訓練的變數，縮小了在不平衡的回歸任務中，訓練集和測試集不匹配的程度。而我們在 speechocean762 資料集上評估我們的方法，這個資料集上字詞層級顯示出明顯不平衡的標籤，而我們的實驗結果顯示，在這個不平衡的資料集上，我們實驗的結果明顯獲得改善。

Abstract

Automatic Pronunciation Assessment (APA) aims to quantify non-native (L2) learners' pronunciation proficiency in a specific language. With technological advancements, APA now evaluates various aspects of pronunciation, from phoneme level to sentence level, including accuracy, fluency, stress, and more. However, current APA methods rely on the Mean Squared Error (MSE) loss function, which struggles with imbalanced labels across different levels of granularity. This imbalance affects model generalizability and fairness, as MSE tends to underestimate rare labels. Despite these issues, existing research has not adequately addressed data imbalance. To address this gap, we draw inspiration from class-balanced loss functions in visual classification. Our approach involves resampling and introducing a trainable variable to narrow the gap between training and testing sets in imbalanced regression tasks, aiming to alleviate label imbalance effects in APA. Evaluating our method on the Speechocean762 dataset, known for significant word-level label imbalance, we observe remarkable enhancements in performance. Our proposed approach shows promise in tackling challenges stemming from imbalanced data in automatic pronunciation assessment.

關鍵字：自動發音評測、資料不平衡、回歸損失函數

Keywords: Automatic Pronunciation Assessment, data imbalanced, regression loss function

1 介紹

電腦輔助發音訓練 (computer-assisted pronunciation training , CAPT)系統越來越受歡迎，並被用於各種用例，例如減輕教師的工作量 (Bannò et al, 2022)，發音評測線上課程 (Mehri, 2021)，學習者能夠練習他們的語言技能，以及其他 (Ai, 2015)。電腦輔助發音訓練 (Computer-assisted pronunciation training, CAPT）近年來吸引了人們大量的關注，透過利用許多機器學習的技術展示了令人印象深刻的成果 (Shi, 2020; Li, 2017; Korzekwa, 2022)。

自動發音評測(APA)是一種常見的方法在 CAPT 系統中。自動發音評測很常用於非母語(L2)學習者學習陌生的語言。通常非母語學習者（L2）會朗讀接收到的文本提示，而自動發音評測會根據文本提示和接收到的 L2 學習者的語音資料進一步的去評測學習者的口說能力，並即時的對學習者給出指導性的回饋。

隨著詳細回饋的需求增加，近期的研究根據不同細粒度在多個面向(如:重音、流利度、韻律和準確性等)進行評估發音(Sancinetti, 2022; Tepperman, 2005)。有人嘗試使用單一模型 (Arias, 2010; Gong, 2022) 並行預測在各細粒度級別上不同面向的發音程度評估，以取代分別採用多個模型進行評估的做法。其中基於 Transformer 的發音評估模型 GOPT（Goodness of pronunciation feature-based transformer）（Gong, 2022），有效地運用分段級別（segmental-level）特徵，也就是發音優良度 Goodness of pronunciation (GOP)，在發音評估任務上有重大的貢獻。

儘管取得重大進展，但相關研究卻很少針對資料極其不平衡的面向進行設計，導致在發音評估任務上效果未臻完美。然而不平衡的數據集可能會導致模型在訓練過程中過度擬合到多數的類別。因此解決不平衡的問題對於量化評估非常的重要。然而單一模型針對多面向及多細粒度並行評分情況下，各個面向之間存在極大的差距,，進而阻礙了在真實教育情境中應用相關的評分模型(Basuki, 2018)。因此我們需要取相應的方法來處理這些數據不平衡的問題。以確保評估結果的質量和準確度。

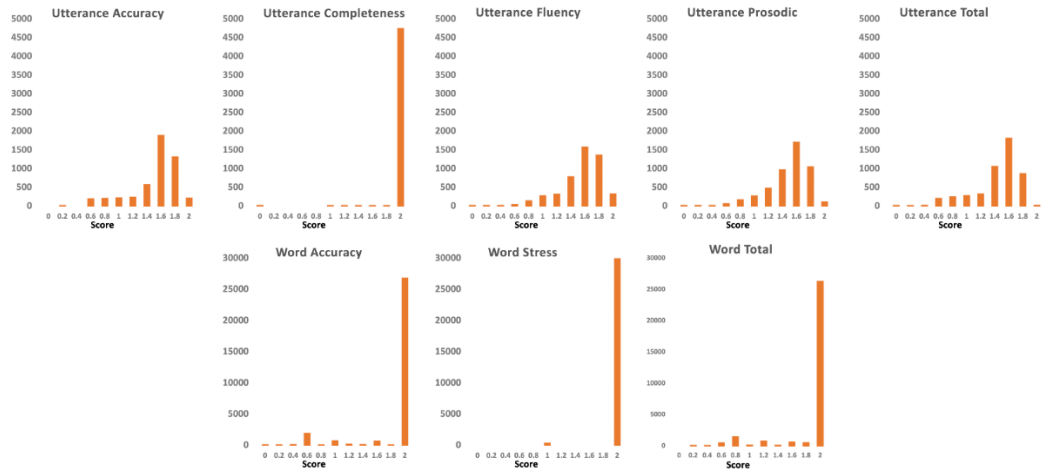
原先 GOPT 的模型架構使用 Mean Square Error (MSE)作為損失函數，但在回歸訓練中 MSE 會低估稀有的標籤。在此研究中，我們參考了在視覺分類建模 (Ren, 2022)中使用的類別平衡損失函數 Batch-based Monte-Carlo (BMC)作為我們的損失函數，BMC 是基於批次的蒙地卡羅方法。在近期深度學習任務中，訓練時的標籤可能具有非常高維度且具有複雜的基礎分佈。由於對分佈建模的約束，對訓練時的標籤進行解析表達可能具有挑戰性。因此(Ren, 2022)使用 Monte Carlo Method (MCM)的方法來近似訓練時的標籤，而 BMC 不需要對訓練標籤做額外的前處理就可以克服在多面向及多細粒度發音評估中因為數據的不平衡而引起評測效能下降的問題。而在此視覺分類的研究中 (Ren, 2022)，模型只基於一種細粒度計算損失函數，而在我們認為應該為不同細粒度計算不同的損失函數，因此在我們的研究中我們針對三種細粒度做不同的損失函數進而去預測不同細粒度及不同面向的標籤。

我們使用廣泛用於發音評測中的公開資料集 Speechocean762 (Zhang, 2021)，作為我們的測試語料。在此資料集上我們使用了我們上述所參考到的平衡損失函數。根據我們的觀察發現某些面向 (如完整性和重音) 和某些細粒度 (如單字層級) 的數據分佈具有高度不

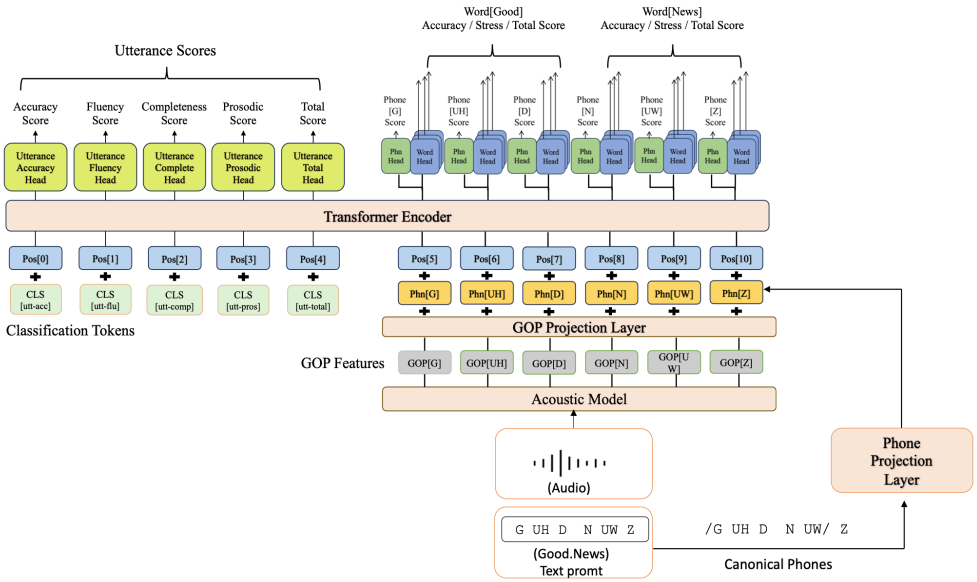
平衡的問題 (如圖一)，由於他們的標籤分數密集分佈在高分的區域，所以會使得低分的標籤容易被預測為高分的標籤。這些高度不平衡的面相相較於其他面向會得到較劣質的表現。所以我們基於 GOPT 的架構改進訓練的損失函數。我們針對三種不同細粒度分別應用可訓練參數的 BMC 損失函數。研究結果顯示，在明顯不平衡的面向及細粒度上獲得顯著的改善，從而減少了在不同面向及細粒度評估效果的差距，而值得注意的是，我們增強的效果是在不進行任何資料增強或架構建模的的情況下實現的。

2 相關研究

回歸任務的研究，早期的研究((Chawla, 2002 ; Cui, 2019 ; Wang ,2017)側重於為稀有標籤重新



圖一：資料分佈的長條圖;第一列為語句層的資料分佈，第二列為單字層級的資料分佈。



圖二：發音評估模型(GOPT)架構圖。

資料集是影響監督分類和回歸很重要的因素，所以資料集不平衡的問題是一直以來受到積極討論的問題，尤其是在視覺和文本分類的任務中 (Padurariu, 2019; He, 2009)。近期針對資料不平衡的分類任務的研究可以分為是重新取樣(Chawla, 2002)和重新加權(Cui, 2019 ; Wang, 2017)的作法，重新取樣通過重複或刪除現有的資料來對資料進行過多的樣本或過低的樣本進行處理。重新加權則是將罕見的樣本分配更大的損失權重，反之將頻繁出現的樣本分配給較小損失權重，以達到平衡資料的效果。最近缺乏針對資料不平衡的採樣和合成新樣本。近期在影像處理的任務(Ren , 2022)上，有人針對不平衡回歸提出平衡策略，他們總共提出三種損失函數的方法來解決不平衡資料的方法，分別是 (1) GAI (GMM-based Analytical Integration) 是一種基於高斯混合模型（GMM）的分析積分方法。為了使積分計算變得可行，他們將訓練的標籤表示為一個高斯混合模型。使用 GMM 的主要優點是兩個高斯分布的乘積仍然是一個未經歸一化的高斯分布。(2) BNI (Bin-based Numerical Integration) 是一種基於區間劃分的數值積分方法，主要應用在單一維度的標籤

空間。它利用核密度估計（KDE）來估算不同區間中標籤的密度，進而進行數值積分。這種方法將標籤空間分成均勻的區間，然後使用 KDE 在每個區間的中心點估算標籤的概率密度函數，從而計算積分。這種方法可以幫助處理數值積分的問題，特別是在標籤空間不均衡的情況下，提供了一種有效的解決方案。(3) BMC (Batch-based Monte-Carlo) 他是基於批量的蒙地卡羅的損失函數，該方法不需要訓練標籤分佈的先驗知識，因此可以快速地應用在實際應用中。而在此研究中，我們使用 BMC 來解決我們在發音評估中資料不平衡的問題。發音評估的任務也同樣面臨資料標記不平衡的問題，因此我們試圖應用了 BMC 改善當前具有代表性的發音評估模型。

3 方法

我們採用的發音評估模型是 GOPT (如圖二)，GOPT 是基於 Transformer 架構並基於 GOP (Goodness of Pronunciation)的特徵，並行預測多面向和多細粒度的分數。我們使用了公開可用的資料集 Speechocean762，這個資料集包含一種音素層集、三種字詞層級和五個語句層級的標籤，包含正確性、流利度、完整度、韻律等多面向的標籤。GOPT 的目標是通過分析音頻輸入及其對應的規範轉錄來進行發音評估。該過程涉及使用聲學模塊獲取幀級音素後驗概率，然後在音素級進行強制對齊，並將其轉換為 84 維的發音優良度（GOP）特徵。這些特徵通過稠密層投影到 24 維。同時，使用一位熱編碼生成規範音素嵌入，同樣投影到 24 維，與 GOP 特徵相同。這些投影特徵以及 24 維的位置嵌入一起輸入到 Transformer 編碼器中。為了捕獲句子級表示，模型在音素級輸入序列中添加了可訓練的[cls]標記，類似於 BERT。這些[cls]標記的 Transformer 編碼器輸出用作對應的句子級表示。訓練過程涉及多任務學習，使用分別針對每個音素、單詞和句子標籤的回歸頭。這些回歸頭添加在與其對應級別的 Transformer 輸出之上。該段解釋了對每個評估任務使用均方誤差（MSE）損失，並將分數標準化為共同尺度。最終的損失是每個粒度（句子、單詞和音素）的損失之和。

Speechocean762提供了豐富的標籤資訊，主要用於多面向的評估任務。對於每個非母語學習者的語音資料，此資料集包含語句級別、字詞級別和音素級別各種面向的分數標籤。而在音素級別的得分在 0-2 之間，而字詞和語句級別的得分在0-10之間。而在GOPT的模型中我們重新調整字詞層級和語句層級將他們標籤分數的範圍重新調整為 0-2 {0,0.2,...,2.0}。儘管這個資料集促進了多面向及多細粒度的發音評估研究，但所提供的得分標籤是不平衡的，特別的是在字詞階級和語句階級裡的完整度都出現偏向高得分的分佈(如圖一)，在圖一我們可以看出在語句層級中完整度的資料和字詞層級有嚴重的資料不平衡的問題。

而在此實驗中我們分別使用 Mean Square Error (MSE)和 Batch-based Monte-Carlo (BMC)的方法作為損失函數：

$$MSE(p_{tar}, p_{pred}) = \| p_{tar} - p_{pred} \|_2^2 \quad (1)$$

p_{tar}指的是目標的標籤，p_{pred}指的是預測標籤，$\|\cdot\|$指的是 L2 norm。

BMC 是基於批次的 Monte Carlo Method (MCM) ，透過在訓練時隨機取樣來近似訓練時的標籤。

$$BMC = -\log N(y_t; y_p, \sigma_{noise}^2 \mathrm{I}) + \log \sum_{i=1}^{N} N(y_{(i)}; y_p, \sigma_{noise}^2 \mathrm{I})) \quad (2)$$

BMC 可以重新被表示為 Softmax 的數學式：

$$BMC_x = -\log \frac{e^{(-\|y_p - y_t\|_2^2/\alpha)}}{\sum_{y' \in B_y} e^{(-\|y_p - y'\|_2^2/\alpha)}} \quad (3)$$

其中x 指的是不同層級的細粒度。B_y是指在訓練時的批次$B_y = \{y_{(1)}, y_{(2)} \ldots y_{(N)}\}$而 N 是指批次的大小。

$$\alpha = 2\sigma_{noise}^2 \quad (4)$$

σ_{noise}是我們設定為一個低敏的參數，並且在模型訓練期間和訓練的標籤一起優化。x為我

們分別使用三種細粒度計算 BMC 損失函數，分別是音素級別、字詞級別和語句級別。

4.2 實作細節

在 GOPT 模型架構中，我們使用 DNN-HMM 聲學模型來提取 84 維的 GOP 特徵。這個聲學模型基於 Factorized time-delay neural network (TDNN-F)，並使用 Librispeech 960 小時的數據在 Kaldi 進行訓練。為了評估我們應用的損失函數的有效性，我們將 GOPT 的所有訓練超參數與中的設定保持一致。並且確保實驗結果的可靠性，我們使用不同的 random seed 重複了五次獨立的實驗，每個實驗包含 100 個 epochs。學習率是 1e-3。根據訓練集上的 Person Correlation Coefficient (PCC)性能，實驗結果都是基於第五次獨立實驗的最後一個 epoch 所得的結果。

表一：在不同損失函數設定下各個細粒度及面相的實驗結果。分別呈現音素層級的損失表現及和三個層級（音素、單詞和語句層級）的 PCC 分數。

Loss Function Setting	Phoneme Score		Word Score (PCC)			Utterance Score (PCC)				
	Loss	PCC	Accuracy	Stress	Total	Accuracy	Completeness	Fluency	Prosodic	Total
[1] $MSE_{phn/word/utt}$ (Baseline)	**0.09**	**0.61**	0.53	0.29	0.55	0.71	0.16	**0.75**	**0.76**	0.74
[2] $BMC_{phn/word/utt}$	0.12	0.52	0.49	0.25	0.49	0.71	0.32	0.75	0.75	0.74
[3] BMC_{phn}+BMC_{word}+BMC_{utt}	0.09	0.60	0.53	0.30	0.55	0.72	**0.40**	0.75	0.75	0.74
[4] MSE_{phn}+BMC_{word}+BMC_{utt}	**0.09**	**0.61**	**0.54**	**0.31**	**0.56**	**0.72**	0.37	**0.75**	**0.76**	**0.75**

4 實驗

4.1 資料集

我們使用 Speechocean762 資料集，Speechocean762 是一個設計給發音評測的免費公開資料集，此資料集總共包含 5000 句英語語句，由 250 位母語非英語且帶有中國口音的學習者所朗讀而成，然而 Speechocean762 提供了非常豐富的標籤資訊。每個語句提供五種語句級別面向的分數，包含正確性、流利度、完整度、韻律、和四個面向的總分而分數的範圍在 0-10 分，在此資料集的語句中正確性的評分標準為句子整體的發音準確程度，完整度的評分標準為在句子中單字是否發音良好，流利度的評分標準為有無明顯地停頓或結巴，韻律的評分標準為是否有穩定的說話速度正確的腔調及節奏說話。而每個單字提供三種單字層級不同面向的分數，分別是正確性、重音和兩個面向的總分而分數的範圍也是 0-10 分，然而 Speechocean762 也提供了音素層級的分數，分數範圍是 0-2，然而在模型中，我們重新規範了語句層級和單字層級的分數讓他們的範圍變成 0-2，讓他們跟音素層級的分數在同一個規範裡。而訓練集包含 2,500 句語句、15,849 個單字和 47,076 個音素。然而測試集也包含 2,500 句語句、15,967 個單字和 47,369 個音素。Speechocean762 包含多種面向及多細粒力度的標籤分數，並將此資料集來評估 BMC 對資料不平衡的影響。

5 實驗結果

我們研究的結果在表一，MSE(表一的損失函數設定[1])是我們的基線方法。表一的損失函數設定[2]是我們在所有細粒度層級都使用同一個損失函數及可訓練的參數σ_{noise}。得到的結果顯示，雖然只有資料極度不均的「完整度」受到改善，但已經可以發現 BMC 對於處理資料不平衡已有改善。由於不同細力度的資料分布表現不一致，我們認為如果三種細粒度層級都使用同一個損失函數及訓練參數，會使得整體的效能下降。因此我們進一步根據三種不同細粒度的層級分別去計算不同的 BMC 損失函數及調整可訓練參數(設定如表一的設定[3])，實驗結果可以發現因為音素層級沒有資料不平衡的問題，所以使用 BMC 去計算損失函數反而會過度重疊，而在 GOPT 模型裡因為是使用音素層級進而去對模型訓

練單字層級和語句層級的標籤，因此音素層級的表現會影響單詞及句子層級的訓練成效，導致其他兩個層級表現不如預期。因此我們調整在音素層級的損失函數，維持使用 MSE 去計算損失函數(設定如表一的設定[4])，其他發生資料不平衡層級則是進一步使用 BMC 計算損失函數，根據在表一設定[4]的結果顯示在音素層級使用 MSE，而單字層級及語句層級使用 BMC 可以發現，資料不平均的層級及面向都獲得了改善。

6 結論

在此研究中，我們參考在視覺分類建模中使用的類平衡損失函數，並將此損失函數用來改善多細粒度發音評測模型中資料不平衡的問題。將此損失函數用在同一個模型中平行預測三種細粒度不同面向的分數，我們分別對三種細粒度做 BMC 損失函數，實驗結果表明，在多細粒的模型下類平衡的損失函數可以獲得改善並且使用 BMC 損失函數也沒有使原本平衡資料的效果變差。

7 參考文獻

S. Bannò et al., “L2 proficiency assessment using self-supervised speech representations,” arXiv preprint arXiv:2211.08849, 2022.

Mehri Kamrood, A., Davoudi, M., Ghaniabadi, S., & Amirian, S. M. R. (2021). Diagnosing L2 learners' development through online computerized dynamic assessment. Computer Assisted Language Learning, 34(7), 868-897.

Ai, R. (2015). Automatic pronunciation error detection and feedback generation for call applications. In Learning and Collaboration Technologies: Second International Conference, LCT 2015, Held as Part of HCI International 2015, Los Angeles, CA, USA, August 2–7, 2015, Proceedings 1 (pp. 175-186). Springer International Publishing.

Li, W., Siniscalchi, S. M., Chen, N. F., & Lee, C. H. (2016, March). Improving non-native mispronunciation detection and enriching diagnostic feedback with DNN-based speech attribute modeling. In 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 6135-6139). IEEE.

Tepperman, J., & Narayanan, S. (2005, March). Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners. In Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. (Vol. 1, pp. I-937). IEEE.

Sancinetti, M., Vidal, J., Bonomi, C., & Ferrer, L. (2022, May). A transfer learning approach for pronunciation scoring. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6812-6816). IEEE.

Arias, J. P., Yoma, N. B., & Vivanco, H. (2010). Automatic intonation assessment for computer aided language learning. Speech communication, 52(3), 254-267.

Gong, Y., Chen, Z., Chu, I. H., Chang, P., & Glass, J. (2022, May). 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) (pp. 7262-7266). IEEE.

Chao, F. A., Lo, T. H., Wu, T. I., Sung, Y. T., & Chen, B. (2022, November). 3m: An effective multi-view, multi-granularity, and multi-aspect modeling approach to english pronunciation assessment. In 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) (pp. 575-582). IEEE.

Ren, J., Zhang, M., Yu, C., & Liu, Z. (2022). Balanced mse for imbalanced visual regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7926-7935).

Eskenazi, M. (2009). An overview of spoken language technology for education. Speech communication, 51(10), 832-844.

Cucchiarini, C., Strik, H., & Boves, L. (2000). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. The Journal of the Acoustical Society of America, 107(2), 989-999.

Li, K., Wu, X., & Meng, H. (2017). Intonation classification for L2 English speech using multi-distribution deep neural networks. Computer Speech & Language, 43, 18-33.

Zhang, J., Zhang, Z., Wang, Y., Yan, Z., Song, Q., Huang, Y., ... & Wang, Y. (2021). speechocean762: An open-source non-native english speech corpus for pronunciation assessment. arXiv preprint arXiv:2104.01378.

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. Intelligent data analysis, 6(5), 429-449.

Padurariu, C., & Breaban, M. E. (2019). Dealing with data imbalance in text classification. Procedia Computer Science, 159, 736-745.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on knowledge and data engineering, 21(9), 1263-1284.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

Estabrooks, A., Jo, T., & Japkowicz, N. (2004). A multiple resampling method for learning from imbalanced data sets. Computational intelligence, 20(1), 18-36.

Do, H., Kim, Y., & Lee, G. G. (2023). Score-balanced Loss for Multi-aspect Pronunciation Assessment. arXiv preprint arXiv:2305.16664.

Shi, J., Huo, N., & Jin, Q. (2020). Context-aware goodness of pronunciation for computer-assisted pronunciation training. arXiv preprint arXiv:2008.08647.

Korzekwa, D., Lorenzo-Trueba, J., Drugman, T., & Kostek, B. (2022). Computer-assisted pronunciation training—Speech synthesis is almost all you need. Speech Communication, 142, 22-33.

Basuki, Y. (2018). The use of drilling method in teaching phonetic transcription and word stress of pronunciation class. Karya Ilmiah Dosen, 1(1).

Cui, Y., Jia, M., Lin, T. Y., Song, Y., & Belongie, S. (2019). Class-balanced loss based on effective number of samples. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 9268-9277).

Wang, Y. X., Ramanan, D., & Hebert, M. (2017). Learning to model the tail. Advances in neural information processing systems, 30.