多語預訓練模型 Whisper 在低資源語言中的挑戰與限制探討：以客語語音辨識為例

Challenges and Limitations of the Multilingual Pre-trained Model Whisper on Low-Resource Languages: A Case Study of Hakka Speech Recognition

Pei-Chi, Lan
Department of Japanese Language and Culture
Soochow University
Taiwan
lizza63159@gmail.com

Hsin-Tien, Chiang
Department of Data Science
Soochow University
Taiwan
nataliechiang92@gmail.com

Ting-Chun, Lin
Department of Data Science
Soochow University
Taiwan
Janelin454@gmail.com

Ming-Hsiang, Su
Department of Data Science
Soochow University
Taiwan
huntfox.su@gmail.com

摘要

本研究以客語語音辨識競賽為案例，探討多語預訓練模型 Whisper 在低資源語言環境下的實務表現與限制。於熱身賽階段，研究團隊（G 組）的官方成績為漢字字元錯誤率（Character Error Rate, CER）75.58%，拼音錯誤率（Syllable Error Rate, SER）100.97%；而在決賽階段，CER 與拼音錯誤率（Word Error Rate, WER）皆達 100%。透過對系統設計與執行流程的回顧分析，我們歸納出三項主要問題來源：(1) 長語音處理策略失當，僅保留首段進行解碼導致內容截斷；(2) 解碼階段的語言提示固定為「中文」，與客語辨識目標不符；(3) 資料對齊與提交檔生成流程缺乏系統化檢核，且評估設定不當。根據這些觀察，我們提出可重複應用的實務準則，涵蓋長語音處理、語言設定一致性確認與資料提交流程檢查等面向。研究結果顯示，在低資源語言的語音辨識任務中，若資料品質與流程設計未妥善控管，即使使用先進的多語預訓練模型，其效能亦可能嚴重退化。本研究強調資料與流程管理在系統開發中的關鍵角色，並為後續改進與結果重現提供具體參考。

Abstract

This study investigates the practical performance and limitations of the multilingual pre-trained model Whisper in low-resource language settings, using a Hakka speech recognition challenge as a case study. In the preliminary phase, our team (Group G) achieved official scores of 75.58% in Character Error Rate (CER) and 100.97% in Syllable Error Rate (SER). However, in the final phase, both CER and Word Error Rate (WER) reached 100%.

Through a retrospective analysis of system design and implementation, we identified three major sources of failure: (1) improper handling of long utterances, where only the first segment was decoded, causing content truncation; (2) inconsistent language prompting, fixed to “Chinese” instead of the Hakka target; and (3) lack of systematic verification in data alignment and submission generation, combined with inadequate evaluation setup.Based on these findings, we propose a set of practical guidelines covering long-utterance processing, language consistency checking, and data submission validation. The results highlight that in low-resource speech recognition tasks, poor data quality or flawed workflow design can cause severe degradation of model performance. This study underscores the importance of robust data and process management in ASR system development and provides concrete insights for future improvements and reproducibility.

關鍵字：客語、語音辨識、低資源語言、長語音處理、語言提示、資料對齊、評估指標

Keywords: Hakka, speech recognition, low-resource language, long-audio processing, language prompting, data alignment, evaluation metrics

1 研究背景

近年來，隨著多語自監督學習模型（如 Whisper、XLS-R、wav2vec 2.0）陸續問世，語音辨識技術已逐漸從高資源語言（例如英語、中文普通話）擴展至低資源語言與方言。然而，這類模型的效能仍高度依賴語料的品質與規模，以及

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 512–517
November 20-22, 2025

標註方式與模型設定的一致性。當資料稀疏、腔調多樣或語音特徵差異顯著時，即使採用強大的預訓練模型，也可能因流程細節錯誤而導致辨識失準。

臺灣客語屬於典型的低資源語言，其內部分化為多個腔調（如四縣、海陸、大埔、詔安等），在聲學與詞彙層面皆具有顯著差異。雖然政府與學術界已推動語料蒐集與文字化工程，但可直接用於自動語音辨識（ASR）訓練的開放語料仍相對有限。因此，客語 ASR 的開發不僅需面對低資源問題，亦須同時處理多腔調資料整合、語音長度差異與標註一致性等挑戰。

本研究以 **ROCLING 2025 客語語音辨識競賽**為案例，嘗試在有限的開源語料與模型條件下建構可運作的客語 ASR 系統。研究團隊於熱身賽中使用 Whisper 模型，雖能產生一定可辨識的輸出，但在決賽階段，系統表現卻不盡理想，官方評分之 **字元錯誤率（CER）與拼音錯誤率（WER）皆達 100%**。此結果提供了一個重要的反思契機——模型性能並非僅受限於資料量或模型規模，而更容易受到流程一致性、語料對齊、語言設定與評估機制等實務因素的深刻影響。

本文旨在以此失敗案例為出發點，檢視低資源語言 ASR 系統在實作層面的潛在風險，並透過事後分析歸納出可重複應用的檢核準則，期能為後續客語及其他低資源語言的研究提供經驗基礎與改進參考。

2 語料與資料集

本研究所使用之語料與任務皆源自 ROCLING 2025 客語語音辨識競賽。競賽主辦單位提供經整理的官方客語語音資料集，內容涵蓋多位說話者、不同腔調及多樣錄音環境。所有語料均由客家委員會「臺灣客語語音資料庫」授權使用，取樣率為 16 kHz、單聲道、16-bit PCM 編碼，並附有對應之轉寫文字（客語漢字及拼音）。研究團隊僅使用主辦方提供的資料，未額外引入外部語料或語言模型，以確保與官方評測條件一致。

2.1 熱身賽資料集

熱身階段所使用的語料為 FSR-2025-Hakka-evaluation，內容包含錄製語料與媒體語料兩部分，總時長約 10 小時。語料來源為客家委員會之「臺灣客語語音資料庫」，涵蓋大埔腔與詔安腔兩種腔別，並分為男、女聲語者共 21 人。

錄製語料共計 3,458 句、約 44,744 字元、8.0 小時；媒體語料則包含大埔腔與詔安腔各約 1 小時，共 946 句、約 26,883 字元。錄音環境與設備多樣，音檔中保留部分雜訊與腔調差異，以模擬真實語音使用情境並提升模型的泛化能力。

2.2 決賽資料集

決賽階段所使用的語料為 FSR-2025-Hakka-final-release，同樣出自客家委員會「臺灣客語語音資料庫」。該語料總時長約 10 小時，包含 4,563 句語音、約 91,642 字元。音檔以亂數命名之 WAV 檔形式提供，格式為單聲道、16-bit PCM、16 kHz 取樣率。

語料涵蓋多位說話者與兩種腔調（大埔腔、詔安腔），錄音使用多種麥克風與環境條件，部分音檔長度達 20 秒。由客語教師監聽審核，僅於讀錯字時修正，保留自然發音與環境音以維持真實語音特徵。

2.3 任務定義與限制條件

本競賽的核心任務為：給定一段客語語音，系統需自動產生對應之「客語漢字」與「客語拼音」轉寫結果，並盡可能降低錯誤率。

主辦單位未提供額外語言模型或外部語料，系統須完全依賴官方訓練資料中的語音與文字對應關係進行學習。此設定可用於評估多語預訓練模型（如 Whisper）在低資源語言環境下的實際辨識能力與資料依賴程度。

3 方法

本研究以開源多語自監督模型 Whisper 為基礎，探討其在客語語音辨識任務中的實作流程與性能表現。整體方法包含五個部分：(1) 模型架構、(2) 資料前處理、(3) 訓練設定、(4) 語言提示與長語音處理策略、(5) 評估指標與分析方法。以下將逐一介紹。

3.1 模型架構

Whisper 模型為由 OpenAI 發表之 Encoder–Decoder 結構，預先以多語音資料（超過 680,000 小時）進行訓練，能同時執行語音辨識與翻譯任務。本研究採用其開源權重進行微調。

熱身階段使用 whisper-tiny 模型，以驗證可行性；決賽階段則改用 whisper-medium，期望獲得更佳的聲學表徵能力。

3.2 資料前處理

語音樣本經 WhisperFeatureExtractor 處理後，轉換為 16 kHz 的對數梅爾頻譜（log-Mel spectrogram）。標註部分取自主辦單位提供之「客語漢字」欄位，並以 WhisperTokenizer 進行編碼。為維持多腔調資料的一致性，未額外進行拼音正規化處理。訓練資料經由 Dataset.map() 生成語音與標註特徵對應，並於訓練前隨機混合不同語者，以降低說話者差異造成的偏差。

3.3 訓練設定

模型以交叉熵損失函數（cross-entropy loss）進行序列到序列學習。熱身階段的訓練設定為：max_steps = 500、learning_rate = 1e-5、batch_size = 4，未啟用驗證集（evaluation_strategy = "no"）。決賽階段因硬體資源受限且資料量增加，設定調整為 max_steps = 300、learning_rate = 5e-6、batch_size = 1，並啟用 eval_steps = 200 以保存最佳權重。

3.4 語言提示與長語音處理

在微調與推論過程中，**WhisperProcessor** 皆設定 language = "chinese"、task = "transcribe"。此策略在多語模型中常見，但於本任務中造成語言提示與實際語音不符，使客語語音被模型誤判為中文。此外，為避免記憶體溢出，決賽實作僅保留音檔的首段（約 30 秒以內）進行推論，未採用重疊解碼或片段合併。此設計雖可降低運算負擔，卻導致長語音內容被截斷，成為模型性能嚴重退化的主要原因之一。

3.5 評估指標

官方評分以漢字字元錯誤率（Character Error Rate, CER）與拼音錯誤率（Syllable Error Rate, SER）為主要指標；內部分析則另採用字元層級 CER 與字詞層級 WER 進行比較。由於客語標註未進行分詞，WER 在此情境下容易高估實際錯誤率，因此本研究以 CER 作為主要評估指標，以反映系統在不同階段的整體辨識穩定性。

4 結果與分析

本節呈現熱身賽與決賽兩階段之實驗結果，並探討模型退化的原因。

4.1 熱身賽結果

在多腔調（大埔腔與詔安腔）資料上，模型能基本完成語音轉寫。官方評分顯示，G 組在學生組中取得漢字 CER = 75.58%、拼音 SER = 100.97%（請見圖一，G 組部分）。雖然音節層級錯誤率偏高，但部分輸出仍能保留與原句相近的字形或聲韻組合，顯示模型已學得一定程度的聲學對應關係。

主要誤差來源為腔調混用與標註不一致。若以字元層級進行分析，錯誤多集中於同音或近音字的替換，顯示模型在聲學辨識上具有限度的區辨能力，但在語言層級仍受資料品質影響。

4.2 決賽結果

進入決賽後，系統性能明顯退化。官方面板顯示，G 組之 CER 與 WER 均為 100%（請見圖二，G 組部分）。分析顯示，此極端結果並非模型「學壞」，而是由流程錯誤所導致的系統性失效。

主要問題可歸納為四點：

(1) 長語音截斷問題：程式僅保留音檔首段進行推論，導致句尾與後段內容完全遺失。

(2) 語言提示不一致：解碼端固定設定 language = "chinese"，使模型傾向輸出中文或雜訊字元。

(3) 資料對齊與提交流程缺乏驗證：決賽資料以資料夾掃描方式生成 CSV，可能造成音檔與標註對應錯位。

(4) 評估指標選擇不當：未分詞的 WER 在中文及客語語境下容易誇大插入與刪除錯誤，造成評估偏差。

4.3 討論

綜合以上觀察，熱身與決賽之間的落差顯示，模型效能對資料流程極為敏感。當語言提示與標註不一致、長語音處理不當或資料對齊錯位時，系統可能完全失效。

此結果說明，低資源語言 ASR 的瓶頸不僅在於資料稀缺，更在於流程設計的正確性與一致性。若能在實務層面建立自動化檢核機制，例如對齊驗證、語言一致性檢查與重疊解碼策略，應能顯著提升低資源語言系統的穩定性與可重現性。

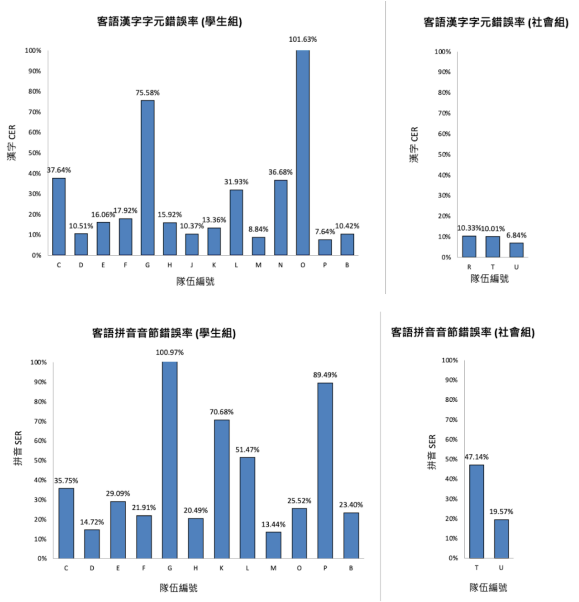

圖 1　熱身賽結果

資料來源：ROCLING 2025 客語語音辨識競賽官方網站 <https://sites.google.com/nycu.edu.tw/fsw/home/challenge-2025?authuser=0>

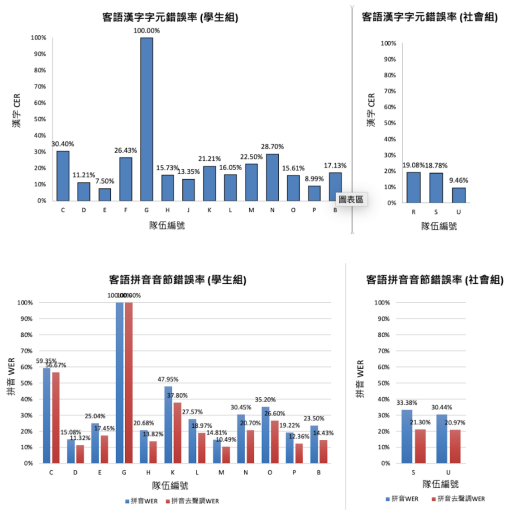

圖 2　決賽結果

資料來源：ROCLING 2025 客語語音辨識競賽官方網站 <https://sites.google.com/nycu.edu.tw/fsw/home/challenge-2025?authuser=0>

5　錯誤分析與討論

5.1　系統性錯誤來源

經追蹤程式版本、訓練記錄與提交流程後，本研究歸納出三項主要失效原因。

(1) 長語音截斷與內容遺失

決賽版本的推論程式為了節省 GPU 記憶體，僅保留每段音檔的首段進行推論。由於決賽語料中平均句長遠高於熱身賽（部分超過 15 秒），此做法直接導致語音後半段未被解碼。系統因此在句尾完全無輸出，導致 CER 與 SER 均接近 100%，屬於結構性錯誤（systemic failure）。

(2) 語料誤用與語言提示不一致

決賽階段的訓練與推論誤將熱身賽語料（大埔、詔安混合）用於模型更新，而正式決賽資料集為獨立語料，兩者語句與錄音來源完全不重疊。此一語料錯置導致模型無法對應測試集語音內容，即使運行正常也不可能產生正確轉寫。此外，Whisper 模型解碼端語言提示（language token）固定為 chinese，與客語實際語音不符，造成生成結果出現大量非預期語言。在熱身階段的輸出中，即可觀察到模型同時產生英語、日語甚至泰文拼音符號，顯示其在語言識別與字集選擇上受到提示錯置的嚴重干擾。此現象突顯多語預訓練模型在低資源語言中的脆弱性：當語言提示與聲學特徵不符時，模型傾向「回退」至高資源語言的字表與音系。

(3) 資料對齊與提交流程的不可驗證性

與熱身賽使用固定 CSV 對應不同，決賽程式改以資料夾掃描方式自動生成提交清單。若檔名排序、語者 ID 或清單順序不一致，音檔與文字對應即可能錯位。由於未設置提交前的比對驗證（例如雜湊或樣本抽查），部分輸出結果可能與官方標準答案對不上，進一步放大系統性錯誤。

5.2　評估指標與低資源語言特性

本競賽的評分方式為：熱身賽採用 CER 與 SER，決賽則以 CER 與 WER 作為主要指標。以下結果與討論均以主辦單位公布之成績為準。由於語料未經分詞處理，WER 與 SER 在計算插入與刪除時容易高估實際錯誤，因此本研究於比較趨勢時亦同時參考 CER，以獲得較穩定之評估結果。

值得注意的是，決賽階段的 CER 與 WER 雖皆達 100%，其原因並非模型退化或學習失效，而是源於語料誤用與流程錯置所造成之對齊錯誤。此結果顯示，在低資源語言任務中，資料流程設計與評估一致性對系統效能具高度敏感性與決定性影響。

5.3　低資源語言 ASR 的實務啟示

本研究的失敗案例揭示了低資源語言 ASR 在實務層面的三項主要風險。

首先是**流程治理**。資料輸入、對齊、訓練與評估各階段皆需設立驗證機制，例如版本控制、檔案雜湊、樣本抽查與提交前比對，以避免隱性錯配或資料汙染。

其次為**語言提示一致性**。多語模型在低資源語言上的正確性高度依賴提示設定；若提示語言與輸入語音不符，模型傾向回退至其訓練頻率較高的語言（如中文或英語），導致不可預期的混語輸出。

最後是**語料範圍與重疊檢核**。在資料拆分與競賽實驗中，須確認訓練集與測試集之間無重疊或錯配；於低資源語言環境中，即便少量語料誤用，也可能造成整體模型失效。

綜上所述，這些檢核與控制策略遠比單純增加資料量更能有效提升系統的穩定性與實驗可重現性。

5.4　反思

本研究顯示，「模型能力不足」並非低資源語言語音辨識系統失效的唯一原因。在此類任務中，資料準備與流程一致性往往對最終結果具有更關鍵的影響力。

本次競賽中，模型性能的崩潰並非源於技術退步，而是由實驗設計與語料治理的疏漏所導致。此案例突顯，低資源語言研究除了演算法創新之外，更需重視資料品質與流程管理。

未來的客語 ASR 研究應將語料完整性驗證與流程可追溯性納入標準研究流程，並建立跨團隊共用的檢核框架，使低資源語言研究能在可重現、可對比的條件下持續進步。

6 實務檢核清單

本研究的經驗顯示，在低資源語言 ASR 的開發過程中，系統性能的崩潰往往源自流程設計與資料管理中的細節錯誤，而非模型能力的不足。

為降低此類失效發生的機率，我們整理出四個層面的實務檢核準則，期能作為後續研究與競賽實作的參考依據。

6.1 語料與對齊層面

(1) 資料一致性檢查

在導入任何語料前，應檢查其版本、來源與編碼格式是否與訓練目標一致。特別是在競賽或多階段任務中，必須確認訓練集與測試集互不重疊，以避免誤用前一階段的資料，造成結果偏差。

(2) 檔名與索引驗證

建議於預處理階段建立雜湊（hash）或索引比對機制，以確保音檔與文字標註之間的對應正確。若提交清單由自動化程式產生，應於輸出前進行隨機抽樣檢查，以防止對齊錯誤。

(3) 多語與腔調標註

在處理多腔調資料時，應明確標示腔別資訊（如大埔、詔安），並於訓練過程中進行條件化控制，以降低模型混淆不同聲學分布的風險。

6.2 模型設定與推論層面

(1) 語言提示一致性

多語模型在推論階段應確認 language token 與實際語音語言相符。若使用 Whisper 或類似架構，建議關閉固定語言參數（forced_decoder_ids = None），或透過自動語言識別（Language ID）機制動態調整，以避免模型受到錯誤語言提示的影響。

(2) 長語音處理策略

推論階段不應僅取音檔首段進行辨識。建議採用滑動視窗或重疊解碼策略，並於後處理階段進行片段合併與時間序校正，以確保長語音內容的完整性與準確性。

(3) 中途監控與早期警示

在訓練與推論過程中應設置開發集（dev set）或樣本監控機制，以即時偵測模型異常輸出。若模型於早期階段出現錯誤語言（如英語、日語或亂碼符號），應立即中止訓練並檢查語言提示與標註格式是否一致。

6.3 評估與報告層面

首先，應以 CER 作為主要評估指標，並輔以 SER 或 WER。對於非分詞語言（如中文、客語），建議採用字元層級的 CER 作為主要報告基準；若同時呈現 SER 或 WER，則須明確說明所使用的分詞策略與音節對應準則，以確保結果具可比性。

其次，建議提升錯誤分析的透明度。除了總體指標外，應附上具代表性的錯誤樣本，以呈現模型在替換、刪除與插入三類錯誤中的具體表現。此舉有助於揭示模型偏誤來源，並為後續改進提供依據。

最後，應強化評估指標的可重現性。研究者需在報告中標明所使用的評估腳本、版本與計算公式，使不同團隊得以重現並對照結果，避免因工具差異而造成評估偏差。

6.4 流程治理與版本控管

(1) 版本追蹤與紀錄保存：

所有訓練、推論與評估流程皆應透過 Git 或同等工具進行版本控管，並完整記錄關鍵參數，如 random seed、套件版本及執行命令列設定，以確保實驗可重現性。

(2) 自動化日誌與錯誤追蹤：

應建立自動化日誌系統以追蹤模型訓練過程與評估結果，並在出現異常時能快速回溯與定位，降低流程錯誤造成的資訊遺失風險。

(3) 流程文件化與可移植性：

建議將整體實驗流程以 Notebook 或 Shell 腳本形式保存，確保研究結果可由他人重現、驗證或延伸，促進低資源語言社群的資料共享與協作。

本研究總結的檢核原則顯示，低資源語言 ASR 的成功關鍵不僅在於模型選擇，更取決於語料治理與實驗流程的可驗證性。

7 結論

本研究以參與 ROCLING 2025 客語語音辨識競賽為案例，探討多語預訓練模型（Whisper）於低資源語言情境下的實務挑戰與失效機制。雖然最終結果顯示系統在決賽中完全失效（CER/SER 皆達 100%），但此極端結果反而揭示了影響低資源

語音辨識的關鍵因素：資料一致性、語言提示設定、長語音處理與流程治理。

研究過程中，我們發現語料誤用、語言提示錯置及缺乏對齊驗證等問題，足以使模型在語音仍可辨識的情況下輸出全錯結果。這說明在低資源語言任務中，流程錯誤的影響程度可遠超過模型本身的效能差距。因此，建立嚴謹的資料與流程檢核機制，比單純調整超參數或擴充模型規模更能有效提升系統穩定性與可靠性。

根據錯誤分析與實務反思，我們提出四項改進方向：
(1) 制定語料與訓練資料版本管理流程；
(2) 建立語言提示與標註一致性檢核；
(3) 引入長語音重疊解碼與分段策略；
(4) 強化評估可重現性與錯誤追蹤機制。

這些原則可作為後續客語 ASR 系統與其他低資源語言研究的基礎。

未來工作將聚焦於三個面向。首先，採用自動語言偵測結合語音特徵對齊的方法，以動態調整語言提示。其次，擴充跨腔調語料，建立能覆蓋大埔與詔安兩腔的平衡訓練集。最後，設計可視化分析工具，用以即時追蹤訓練過程中的語言漂移與對齊錯誤。

總結而言，本研究雖以「失敗案例」為出發點，但其貢獻在於提供一份可驗證、可重現、可借鑑的經驗報告，說明在低資源語音辨識領域中，「失敗」本身亦是推動技術成熟的重要養分。

References

許勝銘. (2007). 大詞彙客語語音辨識系統之初步研究 國立臺灣科技大學]. 臺灣博碩士論文知識加值系統. 台北市. https://hdl.handle.net/11296/r2d95q

羅丞邑. (2011). 以資料探勘之技術解決線上客語語音合成系統中多音字發音歧義之研究 國立中興大學]. 臺灣博碩士論文知識加值系統. 台中市. https://hdl.handle.net/11296/v5422z

吳治翰. (2012). 國語、客語及瑞典語三語言語音辨識系統之設計研究 國立中山大學]. 臺灣博碩士論文知識加值系統. 高雄市. https://hdl.handle.net/11296/aa5v2k

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In Proceedings of the 40th International Conference on Machine Learning (Vol. 202, pp. 28492–28518). Proceedings of Machine Learning Research. Available at https://proceedings.mlr.press/v202/radford23a.html