低資源語言的語音辨識：客語漢字與拼音模型比較
Speech Recognition for Low-resource Languages: A Comparative Study on Hakka Han Characters and Romanization

鄭宇翔
Department of Computer Science and Information Engineering, National Taitung University / Rm. SEC408, No. 369, Sec. 2, University Road, Taitung City, Taitung County 950309, Taiwan
karlboy3306@gmail.com

吳亦軒
Department of Computer Science and Information Engineering, National Taitung University / Rm. SEC408, No. 369, Sec. 2, University Road, Taitung City, Taitung County 950309, Taiwan
pu1254@gmail.com

摘要

本研究針對低資源語言的語音辨識，以客語為例進行探討。由於目前缺乏專門處理閩南語、客語及原住民族語的語音模型，本研究以 OpenAI Whisper-Medium 為基礎，並透過 LoRA（Low-Rank Adaptation）進行微調，建立兩種不同輸出形式的模型：客語漢字與客語拼音模型。實驗資料共計約 80 小時，涵蓋大埔腔與詔安腔，並分別以字元錯誤率（CER）與詞錯誤率（WER）評估模型表現。

Abstract

This study focuses on speech recognition for low-resource languages, with Hakka as the case study. Since there is currently a lack of dedicated speech models for Taiwanese Southern Min, Hakka, and indigenous languages, we adopt OpenAI Whisper-Medium as the base model and apply Low-Rank Adaptation (LoRA) for fine-tuning. Two models with different output forms were developed: a Hakka character-based model and a Hakka phonetic-based model. The experimental dataset contains approximately 80 hours of speech, covering the Dapu and Zhao'an dialects, and the models were evaluated using Character Error Rate (CER) and Word Error Rate (WER).

關鍵字：低資源語言、客語、語音辨識
Keywords: Low-resource Languages, Hakka, Speech Recognition

1 緒論

在語音辨識的研究中，高資源語言（如中文與英文）已達到相當高的準確度，而這些語言擁有大規模的語音與文字的對照語料庫，以及成熟的自然語言處理資源，相較之下，低資源語言由於缺乏大規模語料及相關工具，研究與應用的進展相對有限，導致語音辨識的效果普遍不佳。

近年來，隨著 OpenAI Whisper 等大型多語言語音模型的出現，研究者開始嘗試利用遷移學習與微調（fine-tuning）技術，將這些模型應用於低資源語言，以彌補語料不足的缺陷。Whisper 此種多語言模型使其能在缺乏資料的情況下，仍展現出一定程度的泛化能力，為低資源語言的語音辨識研究提供了新的可能性。然而，如何設計合適的標註策略與輸出格式，仍是提升辨識效能的重要議題。

對於聲調語言而言，標註方式的選擇非常關鍵，以客語為例，其並沒有統一的書寫標準：一方面可以使用漢字進行書寫，另一方面也能以羅馬拼音搭配數字標註聲調的方式呈現。兩種標註系統各具優缺點：漢字符合使用者的閱讀習慣，但存在多音字與語音與文字對應不一致的挑戰；拼音則能直接反映語音特徵，減少歧義，卻可能因使用者不熟悉而降低應用價值。

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 435–440
November 20-22, 2025

本研究以客語為例，探討在相同語音辨識架構下，分別使用客語漢字與客語拼音作為輸出標註，對模型效能所造成的差異。我們以 OpenAI Whisper-Medium 為基礎，並透過 LoRA（Low-Rank Adaptation）進行微調，建立兩種模型，並分別以字元錯誤率（Character Error Rate, CER）與詞錯誤率（Word Error Rate, WER）進行評估。透過比較兩種標註策略的實驗結果，我們期望提出一套適用於低資源語言的有效訓練流程，並提供對未來客語與其他低資源語言語音辨識研究的參考。

2 相關研究

A. 低資源語言 ：低資源語言的語音辨識研究受到廣泛關注。(江宥呈,2023)提出 VoxCentum 資料集涵蓋了 137 種語言共 13,072 小時語音，指出資料集不平衡會顯著影響模型效能，而平衡語料與對比學習能有效提升泛化能力。(劉廷緯 ,2024) **提出了低資源語言的語音處理**，特別是如何在資料不足的情況下，利用 **自監督式學習 (self-supervised learning, SSL)** 來提升語音辨識 (ASR) 與語音處理效能。這些研究顯示低資源語言不僅依賴語料量，也需要設計合適的訓練與增強策略。本研究延續此方向，進一步探討 標註格式（漢字 vs. 拼音）對模型效能的影響。

B. Whisper 模型：OpenAI 所提出的 Whisper 模型，已成為多語言語音辨識的重要基礎。(呂可名,2024) 基於 Whisper 開發即時語音辨識與語者分段系統，驗證了其在多人對話與多語境的強大適應性。(Hsieh et al.,2023) 則針對台語與中文進行 Whisper 微調，利用 Common Voice 與台語戲劇資料，共約 800 小時語料，最終 CER 約 50.7%，顯示 Whisper 在低資源語言上的潛力，但仍需更多資料與後處理。另一項研究比較 Whisper 與 Wav2vec2 在台語辨識的表現，發現 Whisper 在跨語言適應上具優勢，但仍面臨書寫系統不一致的挑戰。這些研究突顯了 Whisper 在 多語言與低資源語言 環境下的強大泛化能力，也為本研究比較「漢字 vs. 拼音」提供了方法論上的基礎。

3 方法

3.1 語料

本研究使用 FSW Challenge 2025 所公開的客語語料，涵蓋大埔腔與詔安腔兩種主要方言，總長度約 80 小時。每筆語料均附有兩種標註，此設計為我們提供了直接比較不同標註策略的可能性，並可探討文字表示對語音辨識效能的影響。語料經過整理後，依照 8:1:1 的比例劃分為訓練集（80%）、驗證集（10%）、測試集（10%），並確保兩種腔調的比例在各資料集內保持平衡，以避免模型因資料分布不均而產生偏差。值得注意的是，雖然主辦方分別提供了約 40 小時的大埔腔與 40 小時的詔安腔語料，但本研究並未將兩者分開訓練，而是統一整合後進行模型訓練。此設計的原因在於：若模型僅依賴單一腔調語料可能會導致模型對特定腔調過度擬合，進而降低對其他腔調的辨識效果，透過將不同腔調混合訓練，模型能同時學習多樣化的發音特徵提升其泛化能力，使其在實際應用中面對不同腔調輸入時，仍能維持穩定的辨識表現。

3.2 音訊前處理

為確保資料一致性，所有音檔在訓練前均進行以下處理：

- 單聲道轉換：將立體聲檔案轉為單聲道，以降低計算負擔。
- 重取樣：將音訊取樣率統一至 16 kHz，與 Whisper 模型的輸入規範一致。
- 峰值正規化：對所有音檔進行正規化，以避免因音量差異過大導致訓練不穩定。

此外，本研究在訓練過程中加入 輕量級資料增強技術(data augmentation)，以模擬多樣化的語音環境，提升模型泛化能力:

A. 音高偏移(Pitch Shifting)：在不影響語義的情況下，對音訊進行小幅度隨機音高調整，使模型能夠學習到不同人說話及語境下的聲學變化，此方法能增加語音的多樣性，尤其對於有限語料的低資源語言來說，有助於提升模型的泛化能力。

B. 雜訊注入(Noise Injection)：在音訊中加入低訊噪比的高斯雜訊，以模擬真實場景中可能出現的背景噪音。由於實際應用環境（如會議、課堂、日常對話）常存在干擾聲，本研究透過此方法使模型能夠學習在雜訊下仍保持穩定辨識能力。

C. 音量縮放(Volume Scaling)：機將音訊振幅調整至原本的 0.8 至 1.25 倍，模擬不同錄音設備、錄音距離或說話音量的差異，避免模型對固定音量過度擬合，進而提升其對不同輸入條件的魯棒性。

與部分研究常見的語速變化(Speed Perturbation)不同，本研究刻意避免使用此方法。原因在於 Whisper 採用固定時間解析度的聲譜表示，若對語料進行語速改變，可能導致資料分布偏離模型的原始特徵空間，進而影響訓練穩定性，因此本研究以音高偏移作為主要的增強手段。

3.3 模型架構與 LoRA 微調

本研究以 OpenAI Whisper-Medium 模型作為基底。Whisper 是一種基於 Transformer 編碼－解碼器架構的多語言語音辨識模型，具備跨語言的強大泛化能力，特別適合用於低資源語言的研究。然而直接微調完整模型需要大量運算資源，因此本研究採用 Low-Rank Adaptation (LoRA) 技術進行參數高效化的調整。LoRA 的優點在於僅需訓練少量附加參數，顯著降低訓練成本，同時保留模型對其他語言的泛化能力。

在此基礎上，本研究設計了兩組實驗模型：

1. 漢字模型：輸出客語漢字，學習率設定為 5e-5，訓練 10 個 epoch。選擇較低學習率與較長訓練週期，目的是讓模型在有限語料下能更穩定地擬合字元級輸出。
2. 拼音模型：輸出帶數字聲調的拼音，學習率設定為 1e-3，訓練 5 個 epoch。由於拼音單位較漢字單純，模型較容易收斂，因此選擇較高的學習率與較短的訓練週期，以加速收斂並避免過擬合。

3.4 評估指標

為比較不同標註策略，本研究設計兩種實驗：Track 1（漢字）：以字元錯誤率 (CER) 作為評估指標，並計算模型輸出與標註的差異。Track 2（拼音）：以音節錯誤率 (SER) 作為評估指標，計算模型輸出與標註在音節的差異。計算公式如下：

$$錯誤率 = \frac{S + D + I}{N}$$

其中：

S：替換數（模型輸出錯誤的單位數）

D：刪除數（模型輸出缺少的單位數）

I：插入數（模型輸出多餘的單位數）

N：參考標註的總單位數

在 Track 1（CER）中，單位為 漢字字元；在 Track 2（SER）中，單位為 帶有數字調的拼音音節。

範例（CER）

參考標註（漢字）:「客語」(共 2 個字)

系統輸出： 「語」(1 個字)

$$S = 0, D = 1, I = 0, N = 2$$

$$CER = (0 + 1 + 0)/2 = 0.5（50\%）$$

範例（SER）

參考標註（拼音）：ng31 ngied54（2 個音節）

系統輸出： ng31 ngid54（2 個音節，第二音節調號錯誤）

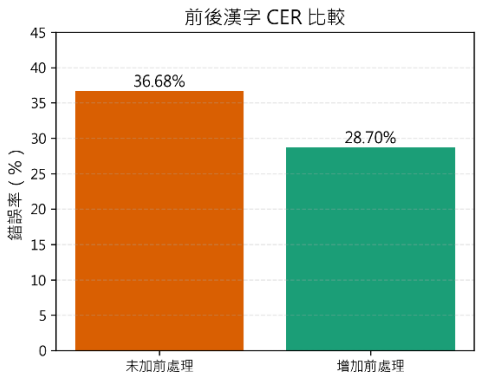
$$S = 1, D = 0, I = 0, N = 2$$

$$SER = (1 + 0 + 0)/2 = 0.5（50\%）$$

4 實驗結果

4.1 客語漢字模型結果

本研究首先針對客語漢字模型在未進行前處理時進行測試，計算整體字元錯誤率(Character Error Rate, CER)。結果顯示，模型的 CER 為 36.68%。在經過前處理後 CER 下降到了 28.70% 。(圖一)



(圖一)

CER 仍偏高的原因我們發現為以下兩點:

A. 多音字現象

客語中存在大量的多音字現象，即同一漢字對應多個不同的發音與語義。例如：

- 「著」可讀作 *tok5*(表示「穿著」)或 *zok8*(表示「正在」);
- 「會」可讀作 *voi5*(能夠)或 *hoi5*(開會/聚會)。在語音辨識中，模型需從語音特徵正確選擇對應的漢字，但由於上下文有限以及語料不足，模型常會出現替換錯誤。例如，輸出「正在」時，可能誤判成「穿著」，造成 CER 提升。相較之下，拼音標註方式能更精準地對應聲學特徵，避免了多音字歧義。

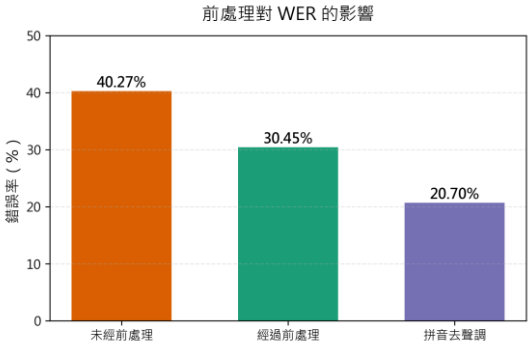
B. 語料規模有限

本研究所使用的語料訓練總量約為 60 小時，對於現代深度學習語音模型而言仍屬於小規模資料。雖然語料已涵蓋大埔腔與詔安腔，但在詞彙分布上仍顯不足：

- 常見詞：如「食」、「飲」、「行」等詞彙模型能學習得較好；
- 稀有詞彙：例如專有名詞、方言特殊用語，出現頻率極低，導致模型在遇到測試集中的新字時錯誤率升高。這種情況尤其會影響漢字模型，因為字表龐大(漢字數量遠多於拼音單位數量)，有限的資料無法充分涵蓋所有字形，導致模型對少見字的預測準確率明顯下降。

4.2 客語拼音模型結果

本研究接著針對**客語拼音模型**進行測試，計算整體**詞錯誤率(Word Error Rate, WER)**。拼音模型未經前處理的 **WER 為 40.27%**，而經過前處理後的 WER 為 30.45，拼音去聲調的 WER 為 20.70。(圖二)



(圖二)

以下整理出我們觀察到的錢處理過後顯著的影響：

A. 資料前處理的重要性

前處理包含語音正規化、去除異常符號及資料一致化，能夠有效減輕模型受到資料雜訊的影響，讓訓練更集中於語音與拼音對應關係。拼音去聲調後 WER 顯著下降，顯示聲調雖具語義區分作用，但若模型尚未能正確捕捉其音韻特徵，反而會降低辨識準確度。這意味後續研究可針對聲調特徵進行專門建模，如增加 tone embedding 或 tone-aware acoustic feature。

B. 字詞規模較小

漢字的字表可能高達數千甚至上萬個字，對低資源語料而言，許多字在訓練集中出現頻率極低，模型難以學習。相對而言，拼音的音節組合有限。例如，客語的聲母、韻母與聲調的組合數量遠少於漢字總量，詞彙表規模縮小至數百個單位即可涵蓋主要發音。這樣的差異讓拼音模型在訓練時更容易收斂，並在測試階段遇到陌生語音時仍能正確對應到既有音節單位，降低了替換錯誤與刪除錯誤的機率。

4.3 漢字與拼音模型比較

綜合兩種標註策略的結果我們可得知以下觀察：

1. **效能比較**：拼音模型（WER 20.70%）明顯優於漢字模型（CER 28.70%）。這顯示拼音作為中介表示更貼近聲學特徵，有助於模型學習與收斂。
2. **實用性比較**：雖然拼音模型效能更佳，但輸出結果對一般使用者不直觀，閱讀成本高；相反地，漢字模型雖然錯誤率較高，但輸出內容更符合使用習慣，應用潛力較大。

5 結論與未來展望

5.1 結論

本研究以客語為例，探討低資源語言語音辨識中不同標註策略對模型效能的影響。我們基於 OpenAI Whisper-Medium 模型，透過 LoRA 微調建立兩種模型：輸出客語漢字與客語拼音的模型。實驗結果顯示：

1. **漢字模型** 的字元錯誤率 (CER) 為 **28.70%**，顯示在文字輸出上仍受限於書寫不一致、多音字現象以及語料不足。
2. **拼音模型** 的詞錯誤率 (WER) 僅為 **20.70%**，效能顯著優於漢字模型，因為拼音標註與聲學特徵的直接對應、詞彙表規模小等原因。
3. 雖然拼音模型在效能上優勢明顯，但漢字模型在應用層面更具可讀性與實用性，因此這兩種策略各有優缺，未來也應考慮整合多模型以同時兼顧準確率與使用者需求。

5.2 未來展望

基於上述研究成果，我們提出以下未來方向：

1. **拼音轉漢字模組**：結合拼音模型與漢字轉換系統，透過語言模型或字典資源進行後處理，提升輸出的可讀性。
2. **跨腔調擴展**：納入更多客語方言（如四縣腔、海陸腔），驗證模型在多樣化腔調下的泛化能力。
3. **語料擴增**：蒐集更大規模的客語語音與標註，並透過自動增強技術 (如:非監督學習) 補足現有不足。
4. **模型比較與優化**：嘗試更小或更大的 Whisper 模型版本，以及其他低資源語言專用架構，進一步驗證效能差異。
5. **應用場景實驗**：將模型部署於真實應用，如客語教學平台、語音輸入法或語言保存工具，檢驗其實際效益與使用者接受度。

6 參考文獻

A. Liu, W. (2025). *Enhancing Efficiency and Reliability in Automatic Speech Recognition Systems* (Doctoral dissertation, The Chinese University of Hong Kong (Hong Kong)).

B. 陳昇德. (2025). 基於大型語音模型的模型壓縮技術應用於邊緣運算裝置. *淡江大學機械與機電工程學系碩士班學位論文*, 1-48.

C. 龙禹辰, 勾智楠, 陈宇欣, & 秦乐. (2025). 基于大语言模型的多任务生成式重构对话情绪识别. *Application Research of Computers/Jisuanji Yingyong Yanjiu*, *42*(7).

D. 劉廷緯. (2024). 更高效的語音處理: 低資源情境下的自監督式學習. *臺灣大學電信工程學研究所學位論文*, 1-179.

E. 陳元瑞. (2020). 藉助跨語言聲音單位對映之遷移學習達成使用低資源之端到端語音合成及辨識. *國立臺灣大學資訊工程學系學位論文*, 1-77

F. Hsieh, Y. C., Lyu, K. M., & Lyu, R. Y. (2023). 運用基於生成預訓練轉換器架構的 OpenAI Whisper 多語言語音辨識引擎之台語及華語語音辨識之實作. In *35th Conference on Computational Linguistics and Speech Processing, ROCLING 2023* (pp. 210-214). The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

G. Wang, S., Yang, C. H., Wu, J., & Zhang, C. (2024, April). Can whisper perform speech-based in-context learning?. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 13421-13425). IEEE.

H. 江宥呈. (2023). 中低資源語言之語音語料蒐集及語言辨識之分析研究. *國立臺灣大學電機工程學系學位論文*, 1-64.

I. Haxhibeqiri, J., De Poorter, E., Moerman, I., & Hoebeke, J. (2018). A survey of LoRaWAN for IoT: From technology to application. Sensors, 18(11), 3995.

J. Vangelista, L. (2017). Frequency shift chirp modulation: The LoRa modulation. *IEEE signal processing letters*, *24*(12), 1818-1821.