The NPTU ASR System for FSR2025 Hakka Character/Pinyin Recognition: Whisper with mBART Post-Editing and RNNLM Rescoring

Yi-Chin Huang[1]*, Yu-Heng Chen[1], Jian-Hua Wang[1], Hsiu-Chi Wu[1]
[1]Department of Computer Science and Artificial Intelligence,
National Pingtung University,
Pingtung city, Taiwan
*ychuangnptu@nptu.edu.tw

Chih-Chung Kuo[2], Chao-Shih Huang[2], Yuan-Fu Liao[2†]
[2]Speech AI Research center,
National Yang Ming Chiao Tung University,
Hsinchu city, Taiwan
†yfliao@nycu.edu.tw

摘要

本文針對 FSR-2025 客語語音辨識（Hakka Automatic Speech Recognition, ASR）競賽，統整兩個子任務：（i）客語漢字（Characters），（ii）客語拼音（Pinyin）。我們提出一條統一架構：以 Whisper [1]（大型弱標註語音辨識模型）為聲學骨幹，視情況採用 LoRA（Low-Rank Adaptation [2]）進行參數高效微調；搭配 MUSAN [3]（音樂／語音／噪音資料庫）與語速（tempo/speed）擾動 [4]之資料增強；在漢字任務中加入 mBART-50 [5,6]（多語言序列到序列模型）進行文本修正；兩個任務皆以 RNNLM [7]（Recurrent Neural Network Language Model）對 N-best 候選做重評分。在漢字任務之決賽設定下，mBART 驅動之 10-best 文本修正 + RNNLM 可達成 CER（Character Error Rate）6.26%，主辦方公布最終 CER 為 22.5%；在拼音任務中，Medium 的模型相較 Large 更適合本資料規模與腔調分布，配合 10-best RNNLM 重評分，在自訂的熱身賽測試集可得 SER（Syllable Error Rate）4.65%，最終決賽公布帶腔調成績為 14.81%。此外，我們亦分析 LID（Language Identification, 腔調辨識）在不同來源（錄製／媒體）之效益。

Abstract

This paper presents our system for the FSR-2025 Hakka Automatic Speech Recognition (ASR) Challenge, which consists of two sub-tasks: (i) Hakka Characters and (ii) Hakka Pinyin. We propose a unified architecture built upon **Whisper** [1], a large weakly supervised ASR model, as the acoustic backbone, with optional **LoRA** (Low-Rank Adaptation [2]) for parameter-efficient fine-tuning. Data augmentation techniques include the **MUSAN** [3] corpus (music/speech/noise) and **tempo/speed perturbation** [4]. For the character task, **mBART-50** [5,6], a multilingual sequence-to-sequence model, is applied for text correction, while both tasks employ an **RNNLM** [7] for N-best rescoring. Under the final evaluation setting of the character task, mBART-driven 10-best text correction combined with RNNLM rescoring achieved a **CER (Character Error Rate)** of **6.26%**, whereas the official leaderboard reported **22.5%**. For the Pinyin task, the **Medium** model proved more suitable than the **Large** model given the dataset size and accent distribution. With 10-best RNNLM rescoring, it achieved a **SER (Syllable Error Rate)** of **4.65%** on our internal warm-up test set, and the official final score (with tone information) was **14.81%**. Additionally, we analyze the contribution of **LID (Language Identification)** for accent recognition across different recording and media sources.

關鍵字：客語語音辨識、資料增強、語言模型重評分、腔調辨識、文本修正

Keywords: Hakka ASR, Data Augmentation, RNNLM N-best Rescoring, Accent Identification, Text Correction

1 緒論

客語具多腔調特性，語音資源相對稀少且口語變異度高，使語音辨識等任務的建模面臨資料稀疏與跨腔穩健性的挑戰。近年大型自監督/弱標註模型（如 Whisper [1]）展現強健跨語言能力；同時，機器翻譯領域的多語序

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 518–522

列到序列模型（如 mBART-50 [5,6]）在文本層面具備校正與樣式歸一化的潛力。基於此架構，本次競賽中，我們主要設計的語音辨識系統，採取「聲學候選 → 文本修正 → 語言模型重評分」的流程，以兼顧資料規模限制與跨腔調需求。

而客語拼音 ASR 面臨腔調差異（大埔、詔安）與資料來源多樣（錄製/媒體）的雙重挑戰。大型弱標註模型 Whisper 具跨語言泛化力，但在高噪環境與腔調失配時仍需資料增強與語言模型輔助；同時，針對「不揭露腔調」的決賽情境，如何在不犧牲準確度的前提下進行腔調亦是本次競賽重點。

本系統之主要特色為：a) 針對客語漢字任務中，建立一套客語端對端 ASR + 文本修正 + LM 重評分之語音辨識系統；系統性比較熱身賽與決賽兩階段的資料切分/訓練差異；加入 mBART 後對於模型的影響。b)在客語拼音任務中，建立以 Whisper-Medium 為本的拼音 ASR 管線，實證其在本任務中優於 large-v3+LoRA；定量分析 MUSAN 噪音增強與 RNNLM 10-best 重評分的疊加效益；提出 avg_logprob 與 辭典查詢兩種 LID，並比較其在異質語料的效果；

2 資料與設定

2.1 原始語料

原始資料涵蓋大埔/詔安腔、男女聲，總計 123 位說話者、27,349 句、約 62.0 小時。本文後續實驗所用的統計與切分均依競賽規範與團隊內部規劃進行。決賽公開集 4,563 句 / 約 10.0 小時。

2.2 資料前處理

針對語料中的資料格式，為了單純處理可用之語料，若 CSV 欄位「備註」含「正確讀音」則刪除該筆；合音字的部分，將其移除星號（例：「來*去」→「來去」）。

熱身賽的語料處理，每個子語料挑 6 位說話者，隨機挑選 3 位語者之資料作為測試集、3 位作驗證集。整體統計：訓練集中包含 99 位/47.28 小時、驗證集中包含 12 位/7.10 小時、測試集中，包含 12 位/6.02 小時。而針對決賽的語料處理，為了最大化訓練的語料數，僅保留熱身語料的 5% 為測試，其餘與訓練集合併；另含媒體語料。總計：69.35 小時做為訓練集、0.50 小時的資料為驗證集；統計合計 69.39 小時。

資料增強的步驟中，我們採用 MUSAN 資料集[3]之語音（noise/speech/music）進行訓練模型的語料擴增，每段語音隨機混入 2–3 段樣本，SNR ∈ {5,10,15} dB；並以 0.5 倍速為間隔進行語速調整：慢速為原始語料的 0.7 至 0.95 倍，快速則為 1.05 至 1.5 倍。最終資料量為原始的 6 倍，但內容保持不變。在漢字任務中，將訓練語料在資料增強後，擴增前後的驗證集，在 large-v3+LoRA 的設定下，其 CER 由 8.31% 降至 1.98%，有顯著的下降。

而在拼音任務中，我們發現到以實際訓練語料的測試集的結果比較，顯示 Whisper-Medium 優於 Large +LoRA，因此後續一律以 Medium 為基底；並採用典型的訓練超參數：epoch = 5、batch = 4、lr = 1e-5、grad-accum = 2、warmup = 100。而加入語料擴增之後，實驗結果顯示，以原始訓練語料訓練的模型在乾淨語料上表現良好，但在含有雜音的語料上辨識效果明顯下降。相較之下，經過資料增強的模型在面對含雜音的語料時，確實能提升辨識表現(大埔 7.75% vs. 5.63%，詔安 9.28% vs 7.80%)。故後續訓練模型均以經過資料增強的語料進行。

3 模型定義與系統流程

3.1 Whisper + LoRA（聲學模型）

在客語漢字的任務中，我們透過實驗決定以 Whisper Large 為基底，其採 128 Mel bins 與擴增語言標記。訓練的步驟，採 LoRA 進行參數高效微調。主要超參數差異如下：熱身賽：epoch = 4、batch size = 4、grad_accum = 2、grad_checkpointing = False；決賽：epoch = 10、batch size = 16、grad_accum = 1、grad_checkpointing = True；共同設定則包含使用 AdamW、fp16、lr = 1e-4、warmup_steps = 1000；LoRA： r = 8、alpha=16、dropout=0.1、目標模組 {k/q/v/out_proj, fc1, fc2}。在拼音任務中，在實際訓練語料在測試集的結果比較，顯示 Whisper-Medium 在本任務中，優於 Large +LoRA（大埔 6.5% vs 6.98%，詔安 5.99% vs 10.12%）

3.2 RNNLM（語言模型）

為了提升辨識的效能，我們在客語漢字和拼音的任務中都加入了語言模型的後處理，希望能夠提升辨識的正確率。因此，我們總共收集了多個來源的文字作為語言模型的訓練，其中包含共 109,487 句，而來源包含教育部客語辭典例句、哈客網路學院[13]教材/試題與原始訓練文本。模型為 2-layer LSTM [8]（emb=512、hid=1024、dropout=0.3），訓練超參數如下：seq_len=256、batch=128、epochs 200、lr 2e-3（AdamW, β=0.9/0.98）、clip_grad 1.0。推論的部分，以 beam=5、LM 權重 0.5 作重評分。

而在拼音的任務中，我們更進一步地進行分析，以 LSTM/GRU [9], 2 層，embed=512，hidden=1024 訓練字/子詞級 LM，對 ASR 輸出結果 10 個候選句進行重評分。大埔腔在 beam=10、LM weight=0.5 下，WER 由 5.65% 分別下降至 3.71%（LSTM）/ 3.37%（GRU）。因此，在拼音任務中我們採用 GRU 作為語言模型。

3.3 mBART 文本修正

在決賽中，為了進一步改善漢字任務的辨識效果，我們加入了預訓練的大型語言模型的微調，希望能夠進一步地改善辨識後的成效，在此我們採用了支援多語的 mBART[5]作為基礎模型，再透過客語的文本語料來微調後，使其有能力進行更正輸入字串的下游任務。

訓練資料來自我們所訓練出來的三種聲學模型的輸出（大埔、詔安、混合）模型推理之 5 個最佳的候選句，以採樣設定（top-k=8、top-p=0.96、temperature=0.7）產生多樣候選，構成（input, target）對。接著透過同樣客語文字資料集來對 BERT-base-Chinese 進行斷詞。斷詞器是基於 bert-base-chinese 模型，經過 10 次 epoch 訓練，其超參數 batch size = 16、lr = 2e-5。訓練的語料則為主辦方所提供之兩個腔調訓練語料。斷詞的資訊則是透過台灣客語語料庫所建置的斷詞系統，其標註的詞性標記總共 18 類[12]。

我們比較斷詞前後的語料來進行文本修正，其訓練集包含了大埔腔、詔安腔與混合腔調三種模型分別進行推理，並透過 sampling 機制（top-k=8、top-p=0.96、temperature-0.7）為每段語音生成 5 個候選句，而訓練目標則為真實的原始句子。訓練總數約有 37,407 個音檔，測試集則隨機挑選 1,000 句，結果顯示，測試集中未斷詞的更正文本後，CER 為 24.77%；加入斷詞資訊後，則可改善至 19.29%，因此我們最後採用斷詞後的語料來進行訓練/驗證。mBART 訓練採用 facebook/mbart-large-50，其超參數設定如下：beam=5、lr=5e-5、batch=16、epoch=5、weight decay=0.01。另在 CE loss 上加入 Over-Edit Penalty 抑制過度修改，透過比較不同的加權值 α 後，透過驗證集決定其權重，最終尋得最佳 0.8（驗證語料中，獲得 CER 12.91%、句錯率 SER 20.83%）。

4 系統效能分析與競賽結果討論

在本節中，我們將針對熱身賽以及決賽所使用的模型以及如何推論出最終繳交的答案進行說明。

4.1 客語漢字

針對漢字任務，在熱身賽時，我們所採取的步驟主要為資料增強 → Whisper Large-LoRA → RNNLM 重評分，這三個步驟，所獲得的結果，我們自行測試的結果為混合腔調 Large 的 9.34% CER，Medium 的 10.39%。最終，主辦方公布為 8.84%，顯示在大型的育訓練模型進行模型微調後，在較複雜的客語漢字任務中較合適。此外，再加上語言模型針對 Whisper 輸出進行重新評分後，進一步提升其效能，比 baseline 的 10.42%進一步提升。

而在決賽中，我們進一步加入 mBART 的結果進行分析，因此在推論的流程中，修改為以下的方式：資料增強 → Whisper 產生 5 個候選 → mBART 文本修正並擴增 5 個候選（共 10 個候選句）→ RNNLM 打分挑選最佳解。在此，我們自行內部用熱身賽語料的 5%進行測試時，能夠從單純僅用訓練集的 CER 為 9.20% 下降至 6.26%。最終主辦方公布的結果，則是意外的有落差(22.50%)。

4.1.1 錯誤分析與改進方向

由於在決賽公布結果後，漢字的任務效果低於預期，在詳細檢測結果後發現 mBART 模型中，有預存的 token 辭典，若是沒有看過的 token，皆使用<unk>取代，這將大幅影響我們

文本修正的效能，我們統計了熱身賽與決賽中有多少 token 會受此影響，發現分別為 3,163 與 4,134 個 token 都被強制更換成<unk>。而文本修正的模型會直接忽略這些<unk>而導致最終影響辨識客語漢字的結果。若這些 token 都能被正確辨識的狀況下，我們賽後測試 Byte-fallback 後，會提升約 5%的正確率(從自行測試的 22.97%降到 17.54%)，但其效果跟主辦方公布的 baseline (17.13%)還是十分相近，可見還是有很大的進步空間。

錯誤主要的面向，一方面可能是斷詞的效果在辨識錯誤的前提下，無法將其修正，第二就是訓練文本修正的模型還不夠完善，可採用 無需固定詞彙表的序列生成模型，例如 ByT5 [10]或 CANINE [11]，其以 byte-level 或 character-level 方式建模，能夠處理 <unk> token 之問題。另外，也有研究是採取 專用的文本修正與重評分策略。以 ByT5 或 T5 為基礎的序列到序列後編輯模型，可針對 ASR 輸出中出現 <unk> 的位置進行上下文預測修補；同時，於重評分階段導入 Transformer-LM [13] 以取代傳統 RNNLM，並在評分函式中加入 <unk> 懲罰項，可進一步減少未知詞候選被選中的機率。

4.2 客語拼音

在客語拼音的任務中，在熱身賽的階段，我們分別對不同的腔調進行各自模型與混和模型的訓練與測試。結果顯示，大埔腔與詔安腔，在原始的語料中，經過資料增強加入噪音後，的確有增加其強健性，大埔腔對於熱身賽的測試從 7.75% 降至 5.63%；詔安腔亦從 9.28% 降至 7.80%（在 Whisper Medium 的設定下）。進一步加入拼音的 RNNLM 熱身賽的大埔腔測試語料從 5.65%降至 3.37%。因此，在熱身賽的階段，我們最終繳交的版本便是採用各自腔調的聲學模型，再加上語言模型的後處理的結果。最終，包含兩個腔調語料的辨識結果，主辦方公布為 13.44%。

4.2.1 個別腔調進行判斷

由於本次競賽中具有兩種腔調，分別為大埔腔跟詔安腔，其拼音的音節組成與音調的調號有些許差異，因此可能導致辨識效果不穩定，再加上決賽的語料並不會提供腔調的標籤，因此，在此我們測試了混和訓練模型跟個別訓練模型，觀察各自的現象。其中，混合腔調的模型基本上就是採兩個腔調的語料子集合進行合併後訓練，並同時比較 Whisper Large 跟 Medium 的差異。結果顯示混合模型在 large 的狀況下，其拼音的 WER 分別為 13.06% (大埔)以及 19.76% (詔安)；而 medium 的狀況下則是 5.64% (大埔) 以及 8.80% (詔安)。雖然合併語料訓練 medium 較 Large 可獲得較佳的辨識率，但比腔調單獨訓練的模型來說差異並不大，因此最終仍以腔調分開之 medium 模型作為最終模型。

為了能達到最佳的個別腔調的判斷，以符合決賽的需求，因此我們提出了兩組分辨腔調的方法。第一個方法相對單純，採用 Whisper 模型在進行辨識任務時，所輸出拼音序列的平均機率 (avg_logprob) 的數值作為模型選定。並比較兩個腔調的數值大小來做為選擇腔調的依據，若兩者數據相同時，根據前述的實驗結果看起來，混合模型有一定的準確度，因此便採用混合模型來處理此狀況。我們使用熱身賽的所有語料進行腔調的辨識，結果得出，詔安腔的正確率為 93.95%、大埔腔的正確率為 97.89%，且透過腔調辨識後，使用其辨識模型所辨識出來的拼音音節錯誤率分別為 11.67%(大埔)和 19.53%。其中雖然大埔腔的效果相較直接使用混合模型的表現好(15.89%)，但詔安腔的混合模型反而較進行腔調辨識後的結果來得較好(17.27%)，因此，此方式無法保證在未知腔調的狀態下，挑選各自腔調或是混合模型來得好。

為了得到更穩定的腔調辨識結果，以幫助挑選到合適的模型，我們希望透過兩個腔調的漢字常用辭典的方式來幫助挑選。因此，我們將在漢字任務中辨識出來的結果，透過一個預先訓練好的客文字斷詞器將其斷詞，接著透過查詢教育部台灣客語辭典[12]中，其提供的大埔腔以及詔安腔的辭典，分析斷詞後的字詞對應到兩個腔調的辭典中，何者的

測試語料	Large Hybrid	Medium hybrid	Medium Accent	Best
大埔腔 (5 hr)	28.25%	15.89%	**13.50%**	**10.70%**
詔安腔 (5 hr)	31.68%	17.27%	**16.48%**	**16.39%**

表格 1: 熱身賽拼音任務的個別腔調與混合模型結果分析.

比例較高，進而採用該腔調辨識模型來進行辨識。表格 1 則為我們透過熱身賽的語料進行測試的結果，總共比較三種模型，分別是 Whisper Large 的混合腔調模型、Whisper Medium 混合腔調模型、以及透過辭典輔助的腔調辨識後，在以該腔調辨識 Whisper Medium 的結果；此外，我們亦將使用腔調標記所辨識出來的結果(以 Best 標註)放入表中以做為參考。從結果可得知，透過斷詞再加辭典的方式，可以在詔安腔獲得近似最佳的結果，而大埔腔雖然沒有像詔安腔如此明顯，但也是目前最佳的效果，因此我們便採用此方案來辨識決賽的客語拼音任務。為了進一步增加訓練語料以改善辨識效率，在拼音的決賽模型中，我們保留少量的大埔與詔安腔的語料(各 400 筆錄製語料和 100 筆媒體語料)做為測試集，其餘皆放入訓練之中。最終拿此測試集去測試我們的拼音模型任務的結果為 5.87% (medium)，以及 5.43% (medium +腔調辨識)，並作為最終繳交的模型，主辦方公布決賽的 WER 為 14.81%。

5 結論

本文針對 FSR-2025 客語 ASR 兩子任務提出一條統一且可擴充的辨識流程：以 Whisper 作為聲學骨幹，配合 MUSAN 與語速擾動提升雜訊與腔調下的強健性；於漢字任務中加入 mBART-50 文本後編輯；兩任務皆以 RNNLM 進行 10-best 重評分。實驗顯示：在拼音任務上，Whisper Medium 較 Large-v3+LoRA 更適合本競賽資料規模與腔調分布，透過 GRU-LM 重評分可顯著降低 WER；在漢字任務上，mBART-50 可進一步降低內部評測的 CER，但最終決賽表現受限於字彙覆蓋與 <unk> 之影響。錯誤分析指出，未知詞的處理為主要瓶頸。我們的改進方向包括：採用 byte/character-level 後編輯模型（如 ByT5、CANINE）以消弭 <unk>，以 Transformer-LM 取代傳統 RNNLM 進行重評分並加入 <unk> 懲罰。整體而言，本系統所提出的模組化流程在低資源、多腔調場景具備良好可遷移性，亦為未來客語 ASR 的系統化改良提供可重現的基準。

致謝

本研究承蒙國立陽明交通大學人工智慧語音研發中心與產學創新研究學院之支持與協助，在此謹致謝忱。

參考文獻

[1] Alec Radford, Jong Wook Kim, Tao Xu, et al. 2023. Robust Speech Recognition via Large-Scale Weak Supervision (Whisper). *arXiv:2212.04356*.

[2] Edward J. Hu, Yelong Shen, Phillip Wallis, et al. 2022. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv:2106.09685*.

[3] Daniel Snyder, Guoguo Chen, and Daniel Povey. 2015. MUSAN: A Music, Speech, and Noise Corpus. *Proc. Interspeech*.

[4] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. Audio Augmentation for Speech Recognition via Speed Perturbation. *Proc. Interspeech*.

[5] Yinhan Liu, Jiatao Gu, Naman Goyal, et al. 2020. Multilingual Denoising Pre-training for Sequence-to-Sequence. *ACL* (mBART).

[6] Yuqing Tang, Chau Tran, Xian Li, et al. 2020. Multilingual Translation with Extensible Multilingual Pre-training and Finetuning. *arXiv:2008.00401* (mBART-50).

[7] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernocký, and Sanjeev Khudanpur. 2010. Recurrent Neural Network Based Language Model. *Interspeech*.

[8] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*.

[9] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, et al. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation (GRU). *EMNLP*.

[10] Linting Xue, Noah Constant, Adam Roberts, et al. 2022. ByT5: Towards a Token-Free Future with Byte-Level Models. *TACL*.

[11] Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. CANINE: Pre-training an Efficient Tokenization-Free Encoder for Language Representation. *TACL*.

[12] 教育部臺灣客家語常用詞辭典（線上資源）。網站：Ministry of Education, Taiwan Hakka Dictionary (accessed 2025).

[13] 哈客網路學院（線上教材/試題資源，作為文本來源與詞表參考）。Website: Hakka Online Academy (accessed 2025).