應用 Whisper 與拼音後處理的客語語音辨識
Hakka Speech Recognition with Whisper and Pinyin Post-processing for FSR-2025

Chia-Hsin Lee*, **Yung-Jun Chang***, **Jin-Yan Wu***, **and Kuan-Yu Chen**
Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology
{d11415004, m11315045, m11315085, kychen}@mail.ntust.edu.tw

Abstract

本研究爲參加 FSR-2025 客語語音辨識挑戰賽（Hakka ASR II）的技術報告，旨在推進客語自動語音辨識技術的發展。由於客語屬於低資源語言，且存在多種腔調，語音辨識面臨高度挑戰。我們以 Whisper-large-v2 爲骨幹模型，設計兩階段訓練流程：首先利用「Hakka Across Taiwan (HAT)」語料庫進行模型調適，以捕捉客語的一般聲學特徵；其次在賽事方提供的 60 小時腔調語料上進行微調，以增強對目標資料的適應性。實驗發現，直接輸出客語漢字可達到良好的字錯率（CER），但由於腔調差異與拼音規則變化多，拼音任務表現顯著下降。爲解決此問題，我們以漢字模型的編碼器初始化拼音模型，並提出結合 RoBERTa 漢字轉拼音、腔調判斷與字典修正的後處理模組，期望可以在比賽中提升辨識的成效。

1 Introduction

FSR-2025 客語語音辨識挑戰賽（Hakka ASR II）旨在推動台灣客語自動語音辨識技術，資料同時涵蓋朗讀與自然口語場景，並分成漢字與拼音兩種辨識任務，評估指標分別採用字元錯誤率（charactor error rate, CER）與音節錯誤率（syllable error rate, SER）。開發高品質客語語音辨識系統（automatic speech recognition, ASR）具備三方面意義：**(1)** 語言保存：完整蒐集與標註各腔調語料，有助文化傳承與方言研究；**(2)** 數位包容：提供客語族群友善的語音互動介面；**(3)** 研究價值：客語的多調系統與低資源特性，是檢驗新型 ASR 方法的重要試金石。

本研究以 Whisper-large-v2 爲骨幹，我們首先比較使用不同的訓練資料量對於後續實驗成效的影響。在這部分的研究中，我們發現在客語漢字以及客語拼音任務中，單純提高訓練資料的規模會有不同結果。所以我們後續採用漢字模型的 Encoder 作爲模型的初始權重，發現可以更穩定的提升客語拼音的成效。此外，雖然 Whisper 在進行客語漢字的輸出時能獲得不錯的表現，但由於客語屬於低資源語言，且存在多種腔調，這造成了 Whisper 在輸出客語拼音時準確率明顯下降。爲了解決此一問題，我們提出一套後處理方法，期望可以在比賽中提升辨識的成效。

2 Whisper

Whisper (Radford et al., 2023) 爲 OpenAI 所提出，採用 Transformer Encoder-Decoder 架構的語音辨識模型。Whisper 基於 weakly supervised 的策略，以多達 680K 小時的語音資料進行訓練，這使得 Whisper 在多語言、多任務乃至於噪聲環境下的各種任務均具有強大的魯棒性（robustness）。得益於上述諸多的優點，許多開發低資源語音辨識模型的研究者也傾向使用 Whisper 作爲預訓練模型。

利用 Whisper 進行單一語言的訓練最常見的策略是全參數微調，不過對於資源缺乏的語言，容易產生過擬合的問題。有鑑於此，近期對於資源缺乏的語言多採用 LoRA、Prompt tuning 的策略 (Qian et al., 2024)，僅訓練少量的參數，讓 whisper 可以學習新語言的特徵。除此之外，透過數據增強來增加訓練資料也是常見的方法，由於文字語料的取得難易度遠低於語音語料，以 TTS 等語音合成策略增加可用的訓練資源也行之有年 (Gokay and Yalcin, 2019)。我們的實驗將以 Whisper 預訓練模型作爲骨幹架構，此外爲了追求最佳辨識率，我們採用全參數微調的策略，以盡可能利用全部的參數。

3 方法

爲了讓 Whisper 學習客語語音辨識任務，我們採用兩階段的訓練流程。在第一階段中，我們利用臺灣客語語音資料庫（HAT，Hakka Across Taiwan）(Liao et al., 2023) 對 Whisper

*These authors contributed equally to this work

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 454–458

#	Models	Dataset		CER (%)	
		pretraining	finetuning	dev	test
1	large-v2		賽事訓練語料	1.5	42.8
2	large-v2		HAT	1.3	35.1
3	large-v2	HAT	賽事訓練語料	1.0	19.8

Table 1: 客語漢字辨識結果（dev：錄製語料；test：媒體語料）

#	Models	Dataset		WER (%)	
		pretraining	finetuning	dev	test
1	large-v2		賽事訓練語料	6.9	53.0
2	large-v2		HAT	7.9	90.1
3	large-v2	HAT	賽事訓練語料	5.7	26.6

Table 2: 客語拼音辨識結果（dev：錄製語料；test：媒體語料）

模型進行初步訓練，以便模型能夠學習客語語音的基本特徵。接著，我們利用賽事方所提供約 60 小時的訓練語料進行微調，以提升模型對目標資料的適應能力。最後，我們對模型輸出的辨識結果進行後處理，並將其作爲最終的系統輸出提交。

3.1 語音辨識模型訓練

我們使用 Whisper large-v2 進行訓練，該模型爲各 32 層的 Transformer Encoder-Decoder 架構，約 1.5B 的參數量。前置訓練中我們以臺灣客語語音資料庫 (Hakka Across Taiwan，HAT) 語料庫，及賽事方提供共 60 小時之大埔與詔安腔作爲訓練資料。HAT 資料集是中華計算語言學學會推出之語料庫，其中包含豐富的語音標註資料，包含錄製語音、媒體語音以及語音合成的語音，腔調主要以海陸以及四縣腔爲主，此外我們使用 SpecAugmentation (Park et al., 2019) 進行資料增強。我們訓練兩個 Whisper 模型，分別輸出客語漢字與拼音兩種結果，以對應比賽的兩種任務。驗證集（development set）和測試集（test set）分別是熱身賽釋出的錄製語料及媒體語料。作爲額外的實驗，我們另行訓練了僅以大埔與詔安腔作爲訓練資料的模型進行比較。

我們的模型是使用 ESPNet toolkit (Watanabe et al., 2018) 進行實現，詳細的實驗設置列舉如下：

- Model: Whisper large-v2
- Learning rate: 5×10^{-5}
- Optimizer: AdamW
- Epoch: 50
- Save strategy: Top3

每個模型皆共訓練 50 epoch，保留成績最佳的三個 checkpoints 作爲最終模型。

3.2 結果比較

Tables 1和 2呈現我們的實驗結果。我們首先發現，在客語漢字以及拼音的場合中，單純提高訓練資料的規模會造成不同結果。客語漢字的 CER 隨著語料的增加有顯著的改善 (見 Table 1，編號 1、2)；然而對於客語拼音而言，資料的增加反而造成 WER 的提升 (見 Table 2，編號 1、2)。我們推測此現象來自於客語拼音的規則複雜。除了對 Whisper 的 Decoder 而言，從頭學習客語拼音的難度高於客語漢字外，根據客家委員會推出之《客家語拼音方案》所描述，異於客語漢字，不同客語拼音對同一語句的拼音規則也因腔調、聲調而異，導致模型學習的困難。

綜合上面論述，爲了利於模型學習客語拼音，第二階段的模型訓練我們以漢字模型的 Encoder 作爲模型的初始權重，幫助 Decoder 能夠更穩定的學習客語拼音的規則。也就是說，Table 1中的編號 3 是由編號 2 模型繼續進行調適而成。Table 2中編號 3 的系統，是使用 Table1中編號 2 的 Encoder 與 Table 2中編號 2 系統的 Decoder 串接後再進行調適而得。

模型的第二階段，我們透過先前已學習客語語音基本特徵的 Whisper large-v2 模型進行再訓練，可見儘管初步訓練中使用的語料多爲海陸及四縣腔，仍有助於模型更好的掌握語音資訊，在以大埔與詔安腔爲主的驗證集與測試集中漢字與拼音均取得最優的結果 (見 Table 1、2，編號 3)。此外，客語拼音方面，成績的進步也證實了若 Encoder 能夠持續提供較優的語音特徵，將有助於 Decoder 學習更爲複雜的拼音規則。

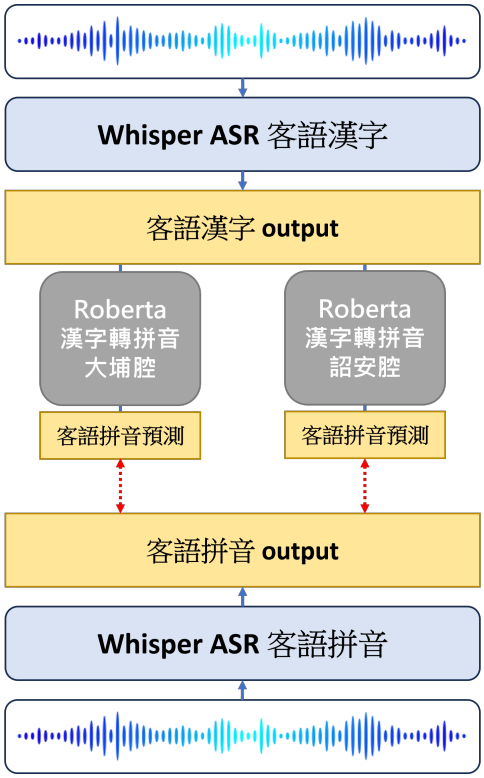

Figure 1: 整體流程：Whisper ASR 先輸出漢字，再透過 RoBERTa 漢字轉拼音模型輸出不同腔調的拼音，最後根據與 ASR 拼音的比較結果判斷腔調。

3.3 爲何需要後處理

客語屬於低資源語言，且存在多種腔調。在我們的實驗中發現，Whisper ASR 在輸出客語漢字時能獲得不錯的表現，但在直接輸出客語拼音時準確率明顯下降。原因在於同一個漢字，可能因腔調不同而對應到不同的拼音。因此，我們提出一套基於漢字的腔調判斷與拼音修正流程，以提升最終拼音輸出的準確度。系統流程如 Figure 1所示。

3.4 方法流程

3.4.1 漢字轉拼音模型訓練

我們首先依照腔調（如大埔腔、詔安腔）將資料集分開，並整理爲「漢字—拼音」配對資料。透過自訂的資料處理程式，將每個字對應到相應的拼音標籤，再使用 RoBERTa 預訓練模型 (Liu et al., 2019)，進行逐字的標記分類訓練，讓模型學會將漢字正確轉換爲拼音（含聲調）。每種腔調獨立訓練一個模型。

我們使用的中文預訓練 RoBERTa* 模型，是已經針對中文語料進行擴展式訓練 (Cui et al., 2021)。內容包含：

- **Masked Language Modeling**：隨機遮罩輸入中的部分字詞，並讓模型學習預測被遮蔽的字。
- **Whole Word Masking, (WWM)**：對於多字組成的詞彙（例如「臺灣科技大學」），遮罩時會同時遮罩整個詞，而非僅遮罩單一字，提升模型學習詞級語義的能力。

因此，在中文任務上（如詞性標註、序列標記、問答任務等）可以展現出比傳統 BERT 更優異的表現。

基於這個中文預訓練 RoBERTa 模型，我們採用 HuggingFace Transformers 框架進行訓練，以成爲漢字轉拼音模型。具體設置如下：

- Model：`hfl/chinese-roberta-wwm-ext`
- Learning rate：5×10^{-5}
- Optimizer：AdamW（使用 HuggingFace 預設）

3.4.2 腔調判斷

至此，我們已具備三種模型：

1. Whisper ASR 直接輸出拼音。
2. Whisper ASR 輸出漢字。

*https://huggingface.co/hfl/chinese-roberta-wwm-ext

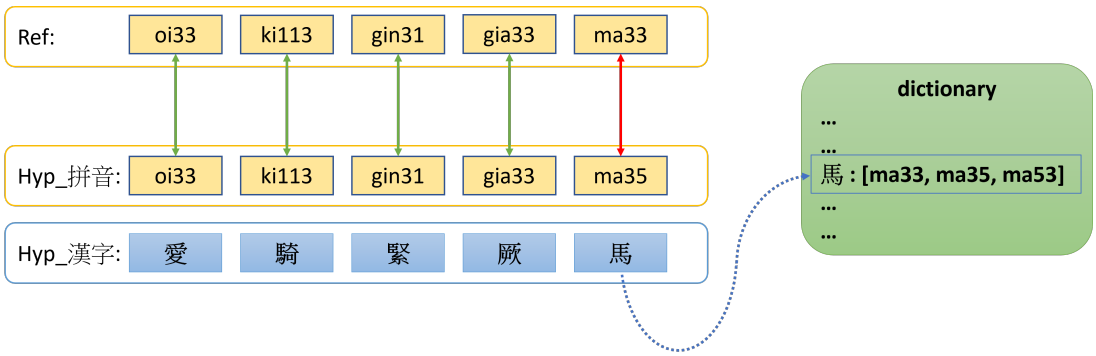

Figure 2: 字典修正流程：alignment 偵測到替換錯誤（例：「ma35」對「ma33」），再由字典提供候選拼音，選擇與參考輸出最相近者作修正。

Table 3: 熱身賽資料集媒體語料的 WER 結果比較。

系統	WER
Whisper ASR（直接拼音）	0.265
Whisper ASR（漢字 → 拼音）	0.266
＋ 字典修正（最終）	**0.264**

3. 各腔調的漢字轉拼音模型。

我們首先採用 Whisper ASR 辨識輸出漢字，再分別利用大埔腔與詔安腔的漢字轉拼音模型轉換成拼音序列，並與 Whisper ASR 辨識輸出的拼音進行 WER 計算。我們選擇 WER 較低的拼音，做爲這個語句的腔調。

3.4.3 字典比對修正

即便在同一腔調中，單一漢字仍可能對應到多種拼音，導致模型選擇錯誤。因此，我們整理出一份字典，將每個漢字對應到其在資料中出現過的所有拼音變體。在後處理過程中，若對齊 (alignment) 結果中出現替換錯誤 (Substitution Error)，我們會從字典中取出候選拼音，逐一比較 CER（字元錯誤率），並選擇最接近參考輸出的拼音進行修正。Figure 2展示了字典修正流程。

3.5 實驗結果

3.5.1 資料集

我們使用熱身賽資料中錯誤率較高的媒體語料進行實驗，並以 WER（詞錯誤率）作爲評估指標。

3.5.2 結果比較

實驗結果如 Table 3 所示。Whisper 直接輸出拼音的 WER 爲 0.265。透過漢字轉拼音並進行腔調判斷後，WER 爲 0.266。最後加入字典修正機制後，WER 降至 0.264。

4 Acknowledgment

This work was supported by the National Science and Technology Council (NSTC) of Taiwan under Grants NSTC 112-2636-E-011-002, NSTC 112-2628-E-011-008-MY3, and NSTC 113-2640-B-002-005. Additional support was provided by the "Empower Vocational Education Research Center" at the National Taiwan University of Science and Technology (NTUST) through the Featured Areas Research Center Program, as part of the Higher Education Sprout Project funded by the Ministry of Education (MOE), Taiwan. The authors also thank the National Center for High-Performance Computing, National Applied Research Laboratories (NARLabs), Taiwan, for providing essential computational and storage resources.

References

Yiming Cui, Wanxiang Che, and Ting Liu et al. 2021. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:2103.00492*.

Ramazan Gokay and Hulya Yalcin. 2019. Improving low resource turkish speech recognition with data augmentation and tts. In *2019 16th International Multi-Conference on Systems, Signals Devices (SSD)*, pages 357–360.

Yuan-Fu Liao, Shaw-Hwa Hwang, You-Shuo Chen, Han-Chun Lai, Yao-Hsing Chung, Li-Te Shen,

Yen-Chun Huang, Chi-Jung Huang, Hsu Wen Han, Li-Wei Chen, Pei-Chung Su, and Chao-Shih Huang. 2023. Taiwanese hakka across taiwan corpus and formosa speech recognition challenge 2023 - hakka asr. In *2023 26th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6.

Yinhan Liu, Myle Ott, Naman Goyal, and Jingfei Du et al. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. In *Interspeech 2019*, pages 2613–2617.

Mengjie Qian, Siyuan Tang, Rao Ma, Kate M. Knill, and Mark J.F. Gales. 2024. Learn and Don't Forget: Adding a New Language to ASR Foundation Models. In *Interspeech 2024*, pages 2544–2548.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. 2018. Espnet: End-to-end speech processing toolkit. In *Interspeech 2018*, pages 2207–2211.