Whisper 微調與 Branchformer 於客語語音辨識之應用
Applying Whisper Fine-tuning and Branchformer to Hakka Speech Recognition

Yu-Sheng Huang
National Ilan University
peilun2016@gmail.com

Wei-Cheng Hong
National Ilan University
q0989323897@gmail.com

Xin-Yu Chen
National Ilan University
cxinyu153@gmail.com

Szu-Yin Lin
National Yang Ming Chiao Tung University
szuyinlin@gmail.com

摘要

本研究針對 FSR 2025 客語辨識任務，比較大型預訓練模型微調與從頭訓練兩種策略。漢字辨識部分，透過微調五種不同規模的 Whisper 模型，large-v3-turbo 在測試集達到 7.55% CER。拼音辨識部分，則比較 Branchformer 與採用 LoRA 微調的 Whisper-small，兩者在測試集的 WER 分別為 4.7%與 6.5%。在資料前處理方面主要採用速度擾動進行資料增強。

Abstract

This study addresses the FSR 2025 Hakka speech recognition task by comparing two strategies: fine-tuning large pre-trained models and training from scratch. For character (Hanzi) recognition, we fine-tuned five different scales of the Whisper model, with large-v3-turbo achieving a 7.55% CER on the test set. For Pinyin recognition, a Branchformer model was compared against a LoRA fine-tuned Whisper-small, yielding WERs of 4.7% and 6.5% on the test set, respectively. Speed perturbation was the primary method used for data augmentation in our pre-processing pipeline.

關鍵字：客語、ASR、Whisper、Branchformer
Keywords: Hakka, ASR, Whisper, Branchformer

1 Introduction

自動語音辨識(ASR)技術近年來因深度學習模型的突破而快速發展，端到端(End-to-End, E2E)模型與大型預訓練語音模型已成為主流。Formosa Speech Recognition Challenge (FSR)的主要任務是客語語音辨識，對於瀕危語言的保存具有重要意義。

回顧 2023 年的 FSR，比賽團隊針對客語 ASR 採用了多種方法。在模型架構上，E2E 模型(如 Conformer、Branchformer、Zipformer transducer 及 Hybrid CTC/Attention)被廣泛應用，以捕捉語音的時序與長距離依賴(Chang and Chen, 2023; Lu et al., 2023a; Su et al., 2023)。Whisper 與 WavLM 等大型預訓練模型也被廣泛使用(Lu et al., 2023a; Chiang et al., 2023; Huang and Tsai, 2023)，透過少量客語資料結合參數高效微調(PEFT，如 LoRA、AdaLoRA)，可達到良好的辨識效果。同時 Wav2vec2.0 與 HuBERT 等自監督學習(SSL)模型常用作前端特徵提取器，從未標記語音中學習表示(Hu and Chen, 2023; Yang et al., 2023)。

為解決資料稀缺與雜訊問題，多型態訓練(MTR)、頻譜擴增(SpecAugment)、速度擾動(Speed Perturbation)等資料擴增技術被廣泛使用，以提升模型穩健性(Chang and Chen, 2023; Yang et al., 2023; Lu et al., 2023b)。部分系統結合淺層融合(Shallow Fusion)、N-best Rescoring 等後處理方法，利用額外文本語料改善辨識結果，也有研究採用基於 BERT 的 pBERT 重新計分(Lu et al., 2023a; Yang et al., 2023)以及使用語音活性檢測(Voice Activity Detection, VAD)則用於去除靜音片段，提升辨識效率(Chen et al., 2023)。

儘管過往的比賽已經取得多方面進展，但仍存在挑戰，例如如何進一步提升模型的多樣性、縮小訓練與測試資料之間的差異，以及處理客語羅馬拼音與漢字轉換的問題。今年的比賽中，我們將延續過往的經驗與成果，探索更適合客語語音辨識的解決方案。

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 441–445
November 20-22, 2025 ©2025 Association for Computational Linguistics

2 Methods

2.1 ESPnet

ESPnet(End-to-EndSpeechProcessingToolkit)，是一個開源、端到端的語音處理工具包，主要基於 PyTorch 深度學習框架進行開發，並延續 Kaldi 風格的資料處理流程。而後來推出的 ESPnet2 採用 YAML 設定 + recipe 的模組化設計，易於重現實驗與切換架構。ESPnet2 的核心優勢之一是其模型庫的豐富性與靈活性，使用者可以在設定檔中輕鬆切換和配置不同的後端模型架構，例如：Transformer、Conformer、或是基於 RNN 的經典模型等。ESPnet2 針對不同的應用任務提供了不同方法，例如：自動語音辨識(ASR)、文字轉語音(TTS)、語音增強/分離(SE/SS)、語音翻譯(ST)與口語理解(SLU)等主流領域。除了內建的核心模型，ESPnet 的框架還支援最新的大型預訓練模型(如 Whisper)進行整合與支援 LoRA 微調。

2.2 Whisper

Whisper 是由 OpenAI 所開發的一套開源自動語音辨識(ASR)系統(Radford et al., 2022)，該系統經過 680,000 小時的訓練，使用多語言及多任務的監督式資料，提升系統在口音、背景噪音及技術性語言上的穩定性。且支援多國語言的語言辨識。

Whisper 模型屬於典型的 Transformer 架構(Vaswani et al., 2023)，採用 Encoder-Decoder 的 Attention 機制。模型前端包含兩層一維卷積層(濾波器大小為 3，啟用函數為 GELU)，第二層卷積的步長為 2，用於對輸入的梅爾頻譜特徵進行下採樣。輸入音訊會重採樣至 16kHz，並計算 80 維 log-magnitude Mel spectrogram，視窗大小 25 毫秒、步長 10 毫秒。處理後的數值正規化至[-1, 1]，並近似零均值。以 30 秒音訊片段為例，可得到 3000×80 的特徵矩陣，經兩層卷積後縮減為 1500×80。卷積輸出再加入位置編碼，其中編碼器使用 sinusoidal positional embedding，解碼器則使用 learned positional embedding。Transformer 區塊採用 pre-activation residual blocks(Child et al., 2019)，並於編碼器輸出端施加最終層正規化。解碼器則使用 tied input-output embeddings(Press and Wolf, 2017)。

標記器部分採用基於 GPT-2 的 Byte-Pair Encoding (BPE)，包括 tiny、base、small、medium 與 large，其中 large 分為 large-v1、large-v2 和 large-v3，且 large-v3 的整體表現最佳。針對不同語言，英文部分會直接沿用 GPT-2BPE，其餘語言的模型將重新擬合詞彙分布，但保持詞表大小不變，以避免非英文語言的過度斷詞。Whisper 的不同模型版本依層數與參數量劃分：最小的 tiny 模型包含 4 層 encoder 與 4 層 decoder，參數量為 39M；最大的 large 模型包含 32 層 encoder 與 32 層 decoder，參數量為 1550M。除了直接使用官方釋出的模型外，亦常透過微調(fine-tuning)的方式，針對特定領域或語言進行再訓練，以進一步提升模型在專業場景中的辨識效能與適應性。

3 Experiments

3.1 資料集

我們使用 Formosa Speech Recognition Challenge 2025 - Hakka ASR II 競賽官方提供的訓練資料集 FSR-2025-Hakka-train，以及熱身賽資料集 FSR-2025-Hakka-evaluation 為音檔資料集與熱身賽資料集 FSR-2025-Hakka-evaluation-key 為熱身賽資料集的標準答案。而資料集分割的部分我們將訓練資料集隨機打亂後依照 8:1:1 的比例各切成 train、dev、test 三份。訓練資料集的切分(見表 1)。

資料集切分	句數	時長(hours)
train	21879	49.64
dev	2735	6.17
test	2735	6.19

表 1. 訓練資料集切分比例

整理熱身賽資料集時，我們發現兩者以話語 ID 對齊後，FSR-2025-Hakka-evaluation 比 evaluation-key 多出 105 筆音檔而無答案。因此僅在 FSR-2025-Hakka-evaluation 與 FSR-2025-Hakka-evaluation-key 的交集 4,299 筆上計算，熱身賽資料集的配置(見表 2)。

	句數
FSR-2025-Hakka-evaluation-key	4299
FSR-2025-Hakka-evaluation	4404

表 2. 熱身賽資料集的配置

3.2 評估方式

模型評估採用字元錯誤率(Character Error Rate, CER)、詞錯誤率(Word Error Rate, WER)與句子錯誤率(Sentence Error Rate, SER)。CER 表示語音轉寫內容在字元層級的準確性，能夠反映模型對於每個字的辨識效果，用於衡量模型在實際應用上的轉寫精確度。WER 則為詞錯誤率(Word Error Rate, WER)，以詞為單位，計算預測結果相較於正確答案所發生的替代、刪除與插入錯誤總數來衡量轉寫的準確性。而 SER 是計算模型整句預測是否與正解完全一致，若句子中有任一字元錯誤則判為錯誤，為更嚴格的指標，評估在完整語句層級的表現。

3.3 客語漢字

本節比較 Whisper 系列多個規模的模型在客語語音辨識任務中的微調表現，所使用之模型包含 tiny、base、small、medium 與 large-v3-turbo。

模型訓練使用 HuggingFace Transformers 與 PyTorch 架構實作，並整合 Whisper 提供之 WhisperProcessor，進行特徵擷取、分詞與標註等前處理工作。主要訓練參數設定如下：音訊採樣率為 16kHz，最大音訊長度限制為 30 秒，文字最大生成長度為 128。訓練採用 mini-batch 大小 16，並進行 4 次梯度累積以模擬較大批次；學習率設定為 1e-5，總訓練週期為 10epochs。在模型選擇上，以驗證集 CER 分數最低之 checkpoint 作為最佳模型並保存。

3.4 客語漢字實驗結果

使用 Whisper 不同規模的模型進行客語漢字辨識實驗，以下整理各模型在測試集上的漢字辨識結果(見表 3)

Dataset	Model	CER	SER
測試集	tiny	28.37	82.78
	base	15.28	67.71
	small	11.89	57.26
	medium	39.41	56.86
	large-v3-turbo	7.55	33.81

表 3. 各模型在測試集中的漢語辨識結果

從測試集的結果可觀察到，隨著模型規模增大，CER 整體呈現下降趨勢，其中 large-v3-turbo 在測試集上取得最佳表現，CER 為 7.55%，SER 也降至 33.81%，顯示其對客語漢字辨識具有較佳能力。

在比賽提供的熱身賽資料集上的漢字辨識表現結果(見表 4)。

Dataset	Model	CER	SER
熱身賽資料集	tiny	33.18	84.28
	base	25.76	73.71
	small	22.27	65.22
	medium	13.61	49.85
	large-v3-turbo	22.78	66.18

表 4. 各模型在熱身賽資料集中的漢語辨識結果

在比賽提供的熱身賽資料集中，表現最佳的反而是 medium 模型，其 CER 為 13.61%，SER 為 49.85%；而 large-v3-turbo 在該資料集的 CER 為 22.78%，表現不如預期。

根據比賽官方網站熱身賽的說明，baseline 採用的模型為 large-v3-turbo。儘管本研究針對該模型進行微調，但在熱身賽資料集上的 CER 表現仍未能超越 baseline 的 10.42%，顯示模型在特定資料上辨識能力仍有提升的空間。

造成 CER 上升的可能因素包括語者差異、背景噪音干擾、語速變化及資料分布不均等問題。

3.5 客語拼音

本研究在客語拼音部分比較兩種基於 ESPnet 所實現模型，分別為從頭訓練的 Branchformer(CTC/Attention)混合訓練，以及對大型預訓練模型 Whisper 進行參數高效微調。兩種方法皆採用了速度擾動 (Speed Perturbation)作為共通的資料增強手段，將訓練語音以 0.9、1.0 及 1.1 三種不同語速進行資料增強。

首先我們嘗試的方法為 Branchformer (CTC/Attention)混合訓練，模型前端將原始音訊轉換為 FBANK 聲學特徵，並在頻譜圖上應用 SpecAugment 進行進一步的資料增強。模型的核心架構由一個包含 12 個區塊的 Branchformer 編碼器與一個包含 6 個區塊的 Transformer 解碼器所組成。訓練過程中，採用混合式 CTC/Attention 的訓練方法，將 CTC 損失與注意力導向的交叉熵損失進行加權，並使用標籤平滑作為正規化手段。優化器選用 Adam，搭配 WarmupLR 學習率。在解碼階

段，系統使用寬度為 20 的波束搜尋，並將 CTC 分數與解碼器分數進行聯合解碼。為了進一步提升辨識準確率，額外訓練並使用一個基於 4 層 Transformer 的 BPE-300 子詞級語言模型。該語言模型在解碼階段透過二次評分的方式被整合進系統，其權重被設定為 1.0，以優化最終輸出的語法與流暢度。

第二種方法採用參數高效微調(PEFT)策略。與方法一不同，在前端直接以原始音訊波形作為輸入。模型的核心是使用 OpenAI 的預訓練模型 Whisper small 版本作為基礎編碼器與解碼器。為了更好的輸出效果，我們採用了 LoRA(Low-Rank Adaptation)技術進行微調。凍結 Whisper 模型的原始權重，僅在 Transformer 注意力機制的 query, key, value 層中注入可訓練的低秩矩陣。模型的訓練目標僅為注意力導向交叉熵損失。在解碼階段，系統採用寬度為 10 的波束搜尋。

3.6 客語拼音實驗結果

漢字拼音部分，我們嘗試三種模型組合作為比較，以下整理各方法在測試集上的漢字拼音辨識結果(見表 5)

Dataset	Model	WER	SER
測試集	BRF	4.9	38.5
	BRF+ LM	4.7	38.7
	WSP_SM + LoRA	6.5	37.4

表 5. 各方法在測試集中的拼音辨識結果

其中 Branchformer(BRF, CTC/Attention)方法的表現為 WER 4.9%、SER 38.5%；在 BRF 上加入 Transformer 的 BPE-300 子詞級語言模型進行二次重評分(BRF+LM)後，WER 下降至 4.7%，但 SER 略升至 38.7%，顯示 LM 有助於 WER 修正，對整句完全正確的比例未必同步提升。Whisper small 採用 LoRA 微調(WSP_SM+LoRA)其表現 WER 為 6.5%，高於 BRF 系列，但 SER 為 37.4%，為三者最佳，反映其較強的語言模型能提高句子完整度。

接著為各方法在比賽提供的熱身賽資料集上的辨識拼音辨識表現結果(見表 6)。

Dataset	Model	WER	SER
熱身賽資料集	baseline		23.4
	BRF	30.3	71.7
	BRF+ LM	54.4	99.0
	WSP_SM + LoRA	35.0	58.6

表 6. 各方法在熱身賽資料集中的拼音辨識結果

在熱身賽資料集中，官方 baseline 表現最佳，SER 為 23.4。而 BRF 模型訓練後，WER 為 30.3、SER 71.7，整體落後 baseline。進一步加入外部語言模型(BRF + LM)後，WER 升至 54.4、SER 幾近 99.0，我們認為可能是 LM 權重設定過高所致，導致解碼分數被 LM 主導而產生錯誤預測。相較之下，採用 Whisper small + LoRA 的參數高效微調，WER 35.0、SER 58.6，雖未優於 baseline，但 SER 顯著低於 BRF 的 71.7，顯示大型預訓練模型的穩定度較佳。

4 Conclusion

本研究在漢字辨識部分，我們微調不同規模 Whisper 模型。在自行切分的測試集中，large-v3-turbo 模型憑著模型參數規模的優勢，取得 7.55%的最佳字元錯誤率(CER)。在拼音辨識部分，我們比較 Branchformer 模型與基於 Whisper small 進行 LoRA 微調的模型。實驗結果顯示 Branchformer(CTC/Attention)在測試集上的詞錯誤率(WER)為 4.9%，加入外部語言模型後進一步下降至 4.7%；相較之下，Whisper small + LoRA 的 WER 為 6.5%，雖然略高於 Branchformer，但在句子錯誤率(SER)上則達到 37.4%，比 Branchformer 的 38.5%還低，其在句子完整度上更具優勢。

在未來研究我們認為可朝以下方向進行改善：加入背景噪音處理機制，提升模型在實際環境音下的辨識穩定性；可引入資料增強技術，如聲音混合、頻譜遮蔽(SpecAugment)等，提升模型對聲音變異的適應能力。也可針對腔調進行分流訓練或採用語者標註進行語者自適應，提升在不同腔調與跨語者場景下的表現一致性。透過上述策略，有望提升模型泛化能力，提高模型在真實應用場景中的實用性與準確率。

References

Hsiu-Jui Chang and Wei-Yuan Chen. 2023. The DMS-ASR System for the Formosa Speech Recognition Challenge 2023. In Jheng-Long Wu and Ming-Hsiang Su, editors, *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 377–

379, Taipei City, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Po-Kai Chen, Bing-Jhih Huang, Chi-Tao Chen, Hsin-Min Wang, and Jia-Ching Wang. 2023. Enhancing Automatic Speech Recognition Performance Through Multi-Speaker Text-to-Speech. In Jheng-Long Wu and Ming-Hsiang Su, editors, *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 371–376, Taipei City, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Ming-Hsiu Chiang, Chien-Hung Lai, and Hsuan-Sheng Chiu. 2023. WhisperHakka: A Hybrid Architecture Speech Recognition System for Low-Resource Taiwanese Hakka. In Jheng-Long Wu and Ming-Hsiang Su, editors, *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 390–396, Taipei City, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating Long Sequences with Sparse Transformers. arXiv:1904.10509 [cs].

Hong-Jie Hu and Chia-Ping Chen. 2023. NSYSU-MITLab Speech Recognition System for Formosa Speech Recognition Challenge 2023. In Jheng-Long Wu and Ming-Hsiang Su, editors, *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 380–385, Taipei City, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Yi-Chin Huang and Ji-Qian Tsai. 2023. Whisper Model Adaptation for FSR-2023 Hakka Speech Recognition Challenge. In Jheng-Long Wu and Ming-Hsiang Su, editors, *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 423–427, Taipei City, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Hao-Chien Lu, Chung-Chun Wang, Jhen-Ke Lin, and Tien-Hong Lo. 2023a. The NTNU ASR System for Formosa Speech Recognition Challenge 2023. In Jheng-Long Wu and Ming-Hsiang Su, editors, *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 397–402, Taipei City, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Yuan-Hsiang Lu, Chung-Yi Li, and Zih-Wei Lin. 2023b. The Taiwan AI Labs Hakka ASR System for Formosa Speech Recognition Challenge 2023. In Jheng-Long Wu and Ming-Hsiang Su, editors, *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 403–408, Taipei City, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Ofir Press and Lior Wolf. 2017. Using the Output Embedding to Improve Language Models. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 157–163, Valencia, Spain. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. arXiv:2212.04356 [eess].

Jia-Jyu Su, Dong-Min Li, and Chen-Yu Chiang. 2023. A preliminary study on Hakka speech recognition by using the Branchformer. In Jheng-Long Wu and Ming-Hsiang Su, editors, *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 409–413, Taipei City, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention Is All You Need. arXiv:1706.03762 [cs].

Tzu-Ting Yang, Hsin-Wei Wang, Meng-Ting Tsai, and Berlin Chen. 2023. The NTNU Super Monster Team (SPMT) system for the Formosa Speech Recognition Challenge 2023 - Hakka ASR. In Jheng-Long Wu and Ming-Hsiang Su, editors, *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 414–422, Taipei City, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).