Improving Low-Resource Speech Recognition with Whisper-MoE and Synthetic Data Augmentation: A Case Study on Hakka
基於 Whisper-MoE 與合成資料增強的低資源語音辨識改進研究：以客家話為例

Yuan-Chi Hsu
National Kaohsiung University of Science and Technology
Department of Electrical Engineering
F113154165@nkust.edu.tw

Liang-Chun Fang
National Kaohsiung University of Science and Technology
Department of Electrical Engineering
F113154169@nkust.edu.tw

Hong-Jie Dai
National Kaohsiung University of Science and Technology
Department of Electrical Engineering
hjdai@nkust.edu.tw

摘要

本研究其目的於如何提高在低資源特定種族的客家語音辨識能力，本團隊透過微調不同基底的 Whisper（如原生與針對中文微調的 Belle 模型）的方式進行實驗，我們發現在客家文字和拼音的任務中微調不同基底的模型會有各自較好且不同的實驗結果。本團隊為了更進一步提升模型的準確率，實驗了在 Whisper Encoder 的 attention 層中的 q、k、v 線性層替換成混和專家模型結合 RoLA，及合成的語音有來自不同風格的聲音以及不同的講話速度，結果顯示任務一中字元錯誤率下降了 0.73%，任務二中字詞錯誤率下降了 0.02 %。可以從結果確認說調整模型架構與在少量的方言語料中有策略地使用少量合成語音是可以提升模型的辨識能力。

Abstract

The objective of this study is to improve speech recognition performance for low-resource Hakka, a language spoken by a specific ethnic group. Our team conducted experiments by fine-tuning different base versions of Whisper (e.g., the original model and the Mandarin-focused Belle model). We found that fine-tuning on different bases yielded distinct advantages and varying results in Hakka character and phonetic recognition tasks. To further enhance model accuracy, we experimented with replacing the q, k, and v linear layers in the attention blocks of the Whisper encoder with a mixture-of-experts model combined with RoLA. In addition, we augmented the training data with synthesized speech generated with diverse voice styles and varying speaking rates. The results showed a 0.73% reduction in character error rate for Task 1 and a 0.2% reduction in word error rate for Task 2. These findings confirm that both architectural adjustments to the model and the strategic use of limited synthetic speech data in low-resource dialect corpora can effectively improve recognition performance.

關鍵字：客家語音辨識、Whisper、Data Augmentation
Keywords: Hakka ASR, Whisper, Data Augmentation

1 Introduction

近幾年隨著算力資源的進步，語音辨識模型一直不斷地創新及突破，在主要的語言如英文、中文、法文等都有著非常好的表現。但台灣屬於一個多元文化的國家，光是方言就有國語、台語、客語還有各種不同的原住民語。但隨著時代演變，在生活中基本上越來越少出現這些方言，因此有相關研究希望能藉由收集語音資料並開發相關語言模型讓這些文化得以延續。隨著 seq2seq 架構的推出，與後來的 Attention (Vaswani, Shazeer et al. 2017) 機制讓模型可以去考慮更多上下文，在這樣的條件下產生的 Transformer 架構，廣泛應用在機器翻譯、語音辨識、圖像識別。本研究是使用當前在語音辨識（Automatic Speech Recognition，簡稱 ASR）主流的模型 Whisper (Radford, Kim et al. 2023)，透過大量資料、不

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 446–449

同語音任務像 ASR、Speech Translation 及單一網路架構，使得 Whisper 可以執行多種語言的任務如語音辨識、語音翻譯。在有著基礎的預訓練模型的基礎之上，透過 Formosa Speech Recognition Challenge 2025 的資料集微調成可以辨識大埔腔、詔安腔的客家語音辨識模型。

本研究還運用了多種策略去提升模型的性能，在任務上採用修改原始 whisper 架構新增 MoE-RoLA 讓模型在混雜的情況下，可以更正確的判斷語音的正確腔調，並使用對應的專家模型進行語音辨識。另外為了可以更好的提升模型的性能，我們使用了 VoxHakka(Chen, Lee et al. 2024) TTS 系統進行語音資料的擴增，合成的文本內容來自教育部臺灣客語辭典(教育部) 之詞彙，語音的風格多變，避免模型發生過擬和。

2 Approach

2.1 Whisper

ASR 領域中 Whisper 的表現亮眼，其重要的原因是他經過 68 萬小時的標註資料，進行了監督式學習，在英文的表現能力上達到跟人類一樣的能力。此舉證明了透過學習大量且多元的資料，可以提升模型在對於口音、噪音的適應性。另外 Whisper 有一個特點就是他擁有強大的適應力，因為本身有著強健的英語能力，透過遷移式的方式使其對其他語言也能夠熟練，甚至是方言(Chen, Huang et al. 2023)。

在 Whisper 中也有嘗試其他微調後的模型進行二次微調其中包含 LianjiaTech 開發的中文語音模型 Belle-whisper-larger-v3-zh(以下簡稱 Belle)，該模型是基於 Whisper-large-v3 進行微調，透過在語音頻譜上的時間軸上分別隨機做 mask 及透過模擬的方式增加噪音使辨識及泛化性能力能夠進一步的提升。

2.2 Mixture of Expert

Mixture-of-Experts (MoE) (Shazeer, Mirhoseini et al. 2017)是一種提升模型任務多元性的架構，透過引入多個專家並由 gating network 動態選擇部分專家參與計算。這種設計能在推理成本近似固定的情況下，大幅增加模型的參數規模與表達能力。在語音處理中，MoE 特別適合多語言與多方言場景，因為不同專家可以捕捉不同語言或腔調的特徵，而 gating network 則能根據輸入自適應分配最合適的專家，從而改善低資源語言的辨識效果。

2.3 MoE-RoLA

MoE-RoLA 是將 MoE 與 Rank-One Low-rank Adaptation (以下簡稱 RoLA) (Hu, Shen et al. 2022) 結合的一種參數高效化方法，架構如圖 1。其核心概念是在預訓練模型的特定層數（如自注意力投影層）中，保留原始權重不變，並在其上引入多個低秩增量專家 (RoLA experts)。每個專家僅包含極少量可訓練參數，而 gating network 根據輸入特徵動態選擇或加權專家輸出。這種設計同時具備 MoE 的專家分工能力 與 RoLA 的參數高效特性：MoE 機制允許不同專家專注於不同語言、方言或任務條件，而 RoLA 保證每個專家增量極小，大幅降低微調成本。透過 MoE-RoLA，模型能在保持參數高效的前提下，動態適應輸入特性，特別適合低資源或多樣化的語音場景。

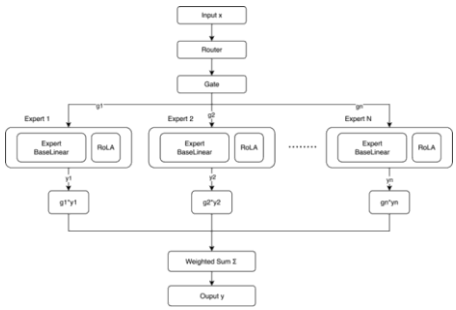

圖 1、MoERoLA 架構

2.4 VoxHakka 文字轉語音系統

VoxHakka 是一個專為臺灣客家語音文字轉語音（Text-To-Speech，簡稱 TTS）的系統，可進行 6 種腔調的合成: 四縣、海陸、大埔、饒平、詔安與南四縣。此系統基礎是 YourTTS(Casanova, Weber et al. 2022)框架，此框架的特點是可針對多語言、多說話者進行 TTS 的開發，VITS(Kim, Kong et al. 2021)的架構使我們可以將文字轉換為高品質、自然的語音。由於 VoxHakka 缺乏開源的資料，對於資料的蒐集提出一種有效率的策略，利用網路爬蟲結合 ASR 資料清理技術，確保建立的資料集品質。由於現階段的臺灣客家語音合成，沒有四縣腔外的公開 TTS 系統可進行比較，因此測試上採用比較平均意見分數（CMOS）評估聽眾在自然度、發音準確、聲調正確性。VoxHakka 系統在自然度方面表現優於其他模型，聲調正確性上接近人類。

3 Experiments

本研究所使用的硬體規格如 Table 1 所示。

CPU	i7-13500
GPU	RTX4090 24G
RAM	128G

Table 1、電腦規格

3.1 Data and Model

3.1.1 Dataset

在本研究中我們的資料集來自以下:

HAT-Vol2: 此資料由比賽主辦方所提供，包含了 100 個語者總時長 80 小時的資料，包含了 27,349 個樣本的訓練集與 3,458 個樣本驗證集。並將資料集分割為 8:2 的訓練資料與驗證資料。

Generated: 透過使用 Huggingface 提供的 TTS API 生成資料，文本內容為教育部臺灣客語辭典提供的大埔腔、詔安腔詞目，將文字內容轉換成語音，從而生成了約 5 小時的語音資料。

3.1.2 Model Configuration

模型配置上，由於兩者任務目標的語系不同，故在模型選擇上有差異。對於客家文字任務，我們使用 Huggingface 平台上的 BELLE-2/Belle-whisper-large-v3-zh，而對於客家拼音任務我們使用 whisper/medium，以上模型皆進行全參數微調，以及MoERoLA 的模型調整。拼音任務有使用資料增強的方式進行微調。Table 2 為 MoERoLA 的主要配置，EXPERTS 為 MOE 所生成的專家數量，該專家數量不一定要等同於語系數量，多出的專家可以學習更細緻的分工（例如不同說話人、口音、噪音條件、語速），RANK 決定增量權重的複雜度（RoLA 固定為 1，僅捕捉一個方向），而 ALPHA 則控制這個增量在最終模型中的影響力大小。

超參數	文字 \ 拼音任務
EXPERTS	6 \ 2
RANK	16 \ 32
ALPHA	32 \ 64
DROPOUT	0.05

Table 2、MoERoLA 參數設定

3.1.3 Fine-tuning detail

本研究在字元任務及拼音任務方面為獨立實驗，但使用方法及依賴類似的，皆以 Dataset 作為資料集型態，訓練模型以 Hugging Face Transformers 套件為基礎進行調整。Table 3 顯示字元任務中訓練使用的超參數設定，包含了 Batch、Early Stop、混合精度、學習率、餘弦退火及優化器等設置，Table 4 則顯示了在拼音任務中訓練使用的超參數。

超參數	Value
Batch	4
學習率	1e-5
優化器	adamw_bnb_8bit
EarlyStop	5
混合精度	FP16
餘弦退火	500

Table 3、字元任務訓練超參數設定

超參數	Value
Batch	8
grad Accum Steps	2
學習率	2.5e-4
優化器	adamw_bnb_8bit
Early Stop	5
混合精度	FP16
餘弦退火	500

Table 4、拼音任務訓練超參數設定

3.2 Evaluation of Character Track

在字元實驗中研究了在純微調模型及嵌入MoERoLA 調整各項參數後分數的變化如 Table 5，可發現 Expert 的參數調整對與模型的分數變化並沒有太大的影響，但其中發現在驗證集方面原本在內部測試成績 CER 為 2.76 %，但在官方的熱身賽測試集 CER 結果卻為 13.36 %，表示該模型對於不同類型的資料泛化性不足，希望藉由調整模型架構可以加強整體模型的效能。

Model	CER
Belle	2.70 %
Belle+ 2 Expert	**1.97 %**
Belle+ 6 Expert	2.27 %

Table 5、字元任務驗證集測試分數

3.3 Evaluation of Pinyin Track

在拼音任務中將僅透過 HAT-Vol2 微調的 whisper/medium 作為我們的 baseline，Table 6 展示在拼音任務 whisper 微調與兩個應用策略後的分數變化可以觀查到 WER 達到 7.22%。緊接著為了讓模型可以學到更多的客家語特徵，我們透過合併了 HAT-Vol2 與 TTS 合成的語音資料進行微調，但效能反而下降非常嚴重，推測原因是合成語料與原先資料差異太大，導致模型無法正確學習。最終嵌入MoERoLA 的 whisper/medium 效果會比 baseline 好，WER 下降了 0.2%。

Model	WER
Whisper/medium	7.22 %
Whisper/medium + Data Augmentation	36.13%
Whisper/medium + 2 Expert	**7.20 %**

Table 7、拼音任務驗證集測試分數

4 Conclusion

本次研究透過 whisper ASR 架構去進行客家語音辨識系統的開發，實驗結果表明此架構僅透過微調即可在新語言資料集上達到非常好的效果。由於熱身賽與決賽錄音中可以明顯感覺有較多環境音與不同風格的內容，因此若無法開發具有泛化能力強的模型即無法再比賽拿到高分，因此未來我們希望可以再去鑽研更多不同的資料資強技巧，尤其是在增加噪音特徵到原始資料的情況，去增強模型的性能，使模型在實際場域中可以達到實驗時一樣的效能。總而言之，我們採取的混和專家策略確實提升了模型的效能。此外透過外部擴增的資料集無法有效提升性能，推測原因是內容與目標資料差異太大。在驗證集上客家文字 CER 降低了 0.73%，客家拼音 WER 降低了 0.02%。結果表明透過專家系統可以有效提升混和不同腔調的客家語音辨識模型性能。

References

Casanova, E., et al. (2022). Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. International conference on machine learning, PMLR.

Chen, L.-W., et al. (2024). VoxHakka: A Dialectally Diverse Multi-Speaker Text-to-Speech System for Taiwanese Hakka. 2024 27th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), IEEE.

Chen, P.-K., et al. (2023). Enhancing Automatic Speech Recognition Performance Through Multi-Speaker Text-to-Speech. Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023).

Hu, E. J., et al. (2022). "Lora: Low-rank adaptation of large language models." ICLR **1**(2): 3.

Kim, J., et al. (2021). Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. International Conference on Machine Learning, PMLR.

Radford, A., et al. (2023). Robust speech recognition via large-scale weak supervision. International conference on machine learning, PMLR.

Shazeer, N., et al. (2017). "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer." arXiv preprint arXiv:1701.06538.

Vaswani, A., et al. (2017). "Attention is all you need." Advances in neural information processing systems **30**.

教育部．"台灣客語辭典." from https://hakkadict.moe.edu.tw/.