Optimizing Whisper Parameters and Training Data Processing for Formosa Speech Recognition Challenge 2025 - Hakka ASR II

李振暐　國立陽明交通大學資訊工程學系　howardlee.cs13@nycu.edu.tw
郭勝威　國立陽明交通大學資訊工程學系　guosw.cs13@nycu.edu.tw
鄭安喆　國立陽明交通大學資訊工程學系　andrewzheng.cs13@nycu.edu.tw
陳秉華　國立陽明交通大學資訊工程學系　binghua.cs13@nycu.edu.tw
劉逸安　國立陽明交通大學資訊工程學系　whitecat.cs13@nycu.edu.tw

Abstract

This paper presents the development and experimental process of our system for the Formosa Speech Recognition Challenge 2025 (Hakka ASR). The proposed system is built upon the OpenAI Whisper model. We achieved significant performance improvements for the Sixian dialect of Hakka through dataset preprocessing and model fine-tuning. In the warm-up evaluation, our system achieved a Character Error Rate (CER) of 10.51% on the character recognition track and a Syllable Error Rate (SER) of 14.72% on the pinyin recognition track. In the final evaluation, our system achieved a Character Error Rate (CER) of 11.21% on the character recognition track and a Syllable Error Rate (SER) of 15.08% on the pinyin recognition track.

Abstract

本文旨在記錄參與 2025 年福爾摩沙語音辨識挑戰賽（Hakka ASR）的實驗過程。我們所提出的系統以 OpenAI 的 Whisper 模型爲基礎，透過對官方提供的客語四縣腔資料集進行預處理與模型微調，顯著提升了辨識效能。在熱身賽的評估中，本系統於漢字辨識任務達到 10.51% 的字元錯誤率（CER），並在拼音辨識任務中達到 14.72% 的音節錯誤率（SER）；而在決賽評估中，本系統於漢字辨識任務達到 11.21% 的字元錯誤率（CER），並在拼音辨識任務中達到 15.08% 的音節錯誤率（SER）。

Keywords: Whisper, Hakka, Denoise

1 Introduction

1.1 深度學習與語音技術的發展

隨著深度學習 (deep learning) 的快速發展，語音相關技術在近幾年有許多突破，特別是在語音辨識 (Automatic Speech Recognition, ASR)，深度神經網路的應用大幅提升模型的準確性，現在的 end-to-end Transformer 模型逐漸取代傳統 ASR 系統。傳統模型通常拆成多個部分分開訓練在進行整合；相較之下 end-to-end 模型直接能從輸入的聲音訊號學習對應的文字輸出，簡化了訓練流程，降低了系統設計的複雜性，同時也增加了辨識準確度。

在許多 end-to-end 模型當中，近年來最具代表性的就是 OpenAI 所發表的 Whisper-ASR (1) 其系統從網路上蒐集眾多資料訓練而成。本篇研究選擇以 Whisper-ASR 作爲基礎模型，並在此基礎上進行針對客語語料的微調以及評估。同時，透過一系列實驗，如在資料加上 SpecAugment(2) 資料增強技術，以及嘗試應用 LoRA(3) 進行參數調整，以期待進一步提升模型的辨識準確度。

1.2 客語在台灣的發展

客語在台灣隨著漢語的普及，逐漸趨向弱勢，在語言傳承方面，現除了國民教育體系中的本土語言課程以外，透過逐漸發展的深度學習語音辨識來建立資料庫，甚至進一步結合相關技術提供學習資源，可做爲客語傳承的重要方法。

1.3 Formosa Speech Recognition Challenge 2025

此競賽聚焦於台灣客語語音辨識的相關研究，藉由提供客語語音轉拼音文字資料集，提供語音辨識模型開發的相關資源。比賽分爲兩部分，Track 1 以字元錯誤率 (CER) 來評估漢字輸出；Track 2 以音節錯誤率 (SER) 評估拼音輸出。透過這兩種評估指標，有效比較不同模型在漢字與拼音輸出上的辨識效果。

2 Methodology

2.1 Dataset

本研究所使用的資料主要來自兩部分。第一部分是競賽主辦方提供的官方訓練集，第二部分則是熱身賽的資料。我們將熱身賽資料全部納

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 471–475

入最終訓練，以最大化資料利用率。詳細的資料統計如表 1 所示。總計約 72 小時的語料，涵蓋了多樣的說話者與內容，爲模型訓練提供了穩固的基礎。

Table 1: 訓練資料集統計 (Statistics of the Training Dataset)

Data Source	Duration (hrs)	Utterances
Train Set	62.02	27,349
Warm-up (Speech)	8.01	3,458
Warm-up (Media)	2.22	946
Total	**72.25**	**31,753**

2.2 Evaluation Metric

我們遵循比賽規定，採用字元錯誤率 (Character Error Rate) 與音節錯誤率 (Syllable Error Rate)，兩者公式如下：

$$ErrorRate = \frac{S + D + I}{N}$$

S 代表被替換的字元 (音節)，D 代表被刪除的字元 (音節)，I 代表被插入的字元 (音節)，而 N 代表總字元 (音節) 數。錯誤率越低，表示模型效能越佳。

2.3 Fine-tuning Whisper

儘管 Whisper 模型在多種語言上表現優異，但對於客語等特定低資源語言或領域，其詞彙覆蓋與聲學特徵適應性仍有提升空間。因此，我們採用微調（Fine-tuning）策略，在我們的客語資料集上進一步訓練預訓練好的 Whisper 模型。此舉能使模型學習客語獨特的發音、詞彙與語法結構，從而有效降低辨識錯誤率。

2.4 Data Augmentation

爲了解決資料量有限可能導致的過擬合問題，並提升模型的泛化能力，我們探索了多種資料增強（Data Augmentation）技術，模擬眞實世界的語音變化。

2.4.1 SpecAugment

SpecAugment (2) 是一種在時頻譜（Spectrogram）上進行遮蔽的有效增強技術。我們主要採用其中兩種策略：

- **Time Masking**: 在時間軸上隨機遮蔽一小段連續的訊框，模擬語音中短暫的停頓或遮蔽。
- **Frequency Masking**: 在頻率軸上隨機遮蔽一段連續的頻帶，增強模型對部分頻率資訊損失的魯棒性 (Robustness)。

2.4.2 Audio Concatenation

觀察到訓練集中單個語音檔案的平均長度較短，我們設計了語音拼接策略（Audio Concatenation），以期待讓模型更好地適應長語音輸入。具體作法爲：隨機選取數個（本實驗設爲 3 個）短音檔，將其拼接成一個新的、更長的音檔，其對應的標註也相應拼接。

2.4.3 Speed Perturbation

爲了模擬不同說話者的語速差異，我們對音檔進行速度微擾（Speed Perturbation）。透過改變音訊的採樣率來實現加速（如 1.1 倍）或減速，但不改變其音高。

2.4.4 Noise Injection

原始訓練集的錄音環境相對純淨。爲了提升模型在眞實噪音環境下的表現，我們在部分音檔中加入了高斯白噪音（Gaussian Noise），噪音的強度根據原始訊號的振幅進行設定。

2.5 LoRA

隨著模型規模的增大，完整的微調（Full fine-tuning）對計算資源的需求也急劇增加。爲此，我們嘗試了低秩適應（Low-Rank Adaptation, LoRA）(4) 技術。其核心思想是凍結預訓練模型原有的權重 W_0，並在模型特定層（如 Transformer 的 attention 層）旁注入一個可訓練的低秩矩陣 $\Delta W = AB$。原始的前向傳播 $h = W_0 x$ 被修改爲：

$$h = W_0 x + \Delta W x = W_0 x + ABx$$

其中，$W_0 \in R^{n \times m}$，$A \in R^{n \times r}$ 和 $B \in R^{r \times m}$ 是可訓練的低秩矩陣，且秩 $r \ll \min(n, m)$。如此，需要更新的參數數量從 $n \times m$ 大幅減少至 $(n + m) \times r$，從而顯著降低了訓練成本。

2.6 Denoising as Preprocessing

我們在決賽資料的語音部分含有噪音，而當我們用降噪工具進行處理後，發現即使原始的語音與降噪後的聽起來沒有太大差別，輸出也會不一樣。我們推論是因爲兩個音檔在轉成梅爾頻譜後會有明顯的差異。因此，我們決定將訓練集的資料也進行降噪以確保一致性。

3 Experiments

3.1 Experimental Setup

在本研究的所有實驗中，我們皆以 `whisper-medium` 作爲基礎模型（base model），並在 NVIDIA V100 GPU 上進行訓練。爲確保比較的公平性，所有模型的訓練週期（epoch）均設定爲 5 次，批次大小（batch

size）爲 4，學習率（learning rate）則固定爲 5×10^{-5}。

我們使用的訓練資料集結合了主辦方提供的官方資料與熱身賽資料。爲了評估模型效能，我們將熱身賽資料集以 60% 與 40% 的比例進行切分，其中 40% 的部分作爲我們的測試集。所有實驗結果均在此測試集上進行評估，並以字元錯誤率 (Character Error Rate, CER) 與音節錯誤率（Syllable Error Rate, SER）作爲主要指標。

3.2 Effectiveness of Data Augmentation

爲了驗證不同資料增強技術對模型的影響，我們設計了一系列的對比實驗。

3.2.1 SpecAugment

首先，我們評估了 SpecAugment 的效果。實驗中，我們對每一筆輸入資料以 50% 的機率應用 SpecAugment，其中時間遮罩（Time Masking）的參數設爲 30，頻率遮罩（Frequency Masking）的參數設爲 15。如表 2及表 3 所示，僅加入 SpecAugment 就讓模型的 SER 從 9.51% 降至 9.04% 及讓 CER 從 3.58% 降至 3.45% ，顯示此技術能有效提升模型的泛化能力。

Table 2: SpecAugment 實驗結果比較 (拼音)

Configuration	SER (%)
Baseline (Original Data)	9.51
+ SpecAugment	**9.04**

Table 3: SpecAugment 實驗結果比較 (漢字)

Configuration	CER (%)
Baseline (Original Data)	3.58
+ SpecAugment	**3.45**

3.2.2 Other Augmentation Techniques

接下來，我們在 SpecAugment 的基礎上，進一步疊加其他三種增強方法：語音拼接（Audio Concatenation）、語速改變（Speed Perturbation）與噪音注入（Noise Injection）。

- 語音拼接: 隨機挑選 3 個音檔拼接成 5000 筆新資料。
- 語速改變: 將語速調整爲 1.1 倍。
- 噪音注入: 加入標準差爲 0.005 的高斯白噪音。

實驗結果如表 4 及表 5 所示。我們發現，語音拼接策略帶來了最佳效果，將 SER 與 CER 分別進一步降低至 8.96% 及 2.93%。然而，在拼音的部分，語速改變反而使模型表現略微下降，噪音注入的影響則相對中性。而在漢字部分，語速改變與噪音注入的表現皆些許上升。

Table 4: 多種資料增強技術實驗結果 (拼音)

Configuration	SER (%)
+ SpecAugment (Baseline)	9.04
+ SpecAugment + Audio Concatenation	**8.96**
+ SpecAugment + Speed Perturbation (1.1x)	9.61
+ SpecAugment + Noise Injection	9.06

Table 5: 多種資料增強技術實驗結果 (漢字)

Configuration	CER (%)
+ SpecAugment (Baseline)	3.45
+ SpecAugment + Audio Concatenation	**2.93**
+ SpecAugment + Speed Perturbation (1.1x)	3.18
+ SpecAugment + Noise Injection	3.12

3.3 Comparison of Base Models and LoRA

我們進一步比較了不同尺寸的 Whisper 模型，以及在大型模型上應用 LoRA 技術對效能的影響。本階段實驗包含以下設定：(i) 使用 `whisper-large-v2` 與 `whisper-large-v3-turbo` 進行微調；(ii) 在 `whisper-medium` 與 `whisper-large-v3` 上採用 LoRA 進行參數高效微調。爲加速實驗迭代，所有模型均僅訓練 3 個 epoch。各模型的 LoRA 參數（alpha, rank）如表 6 所示。

實驗結果顯示，`large-v2` 在客語資料集上的表現最佳（SER = **8.78%**），相較於 `large-v3-turbo`（SER = 9.95%）略有優勢。然而，僅透過 LoRA 微調的模型（`medium` 與 `large-v3`）效能仍顯著落後完整微調，顯示雖然 LoRA 能有效降低訓練參數量，但在本資料集上仍難以達到相同的準確度。

Table 6: 不同基礎模型與 LoRA 實驗結果 (拼音)

Model / Method	Alpha	Rank (r)	SER (%)
whisper-large-v3-turbo	–	–	9.95
whisper-large-v2	–	–	**8.78**
whisper-medium (LoRA)	128	256	22.24
whisper-large-v3 (LoRA)	128	256	21.80

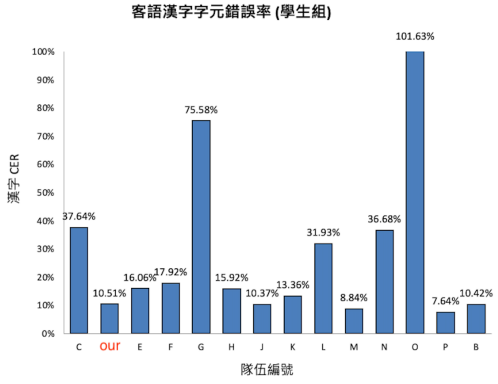
綜合以上所有實驗與時間成本的考量，我們最終決定採用以 `whisper-medium` 爲基礎，並僅使用 SpecAugment 作爲資料增強策略的模型，同時加入降噪處理以提升模型在含噪環境中的適應力。

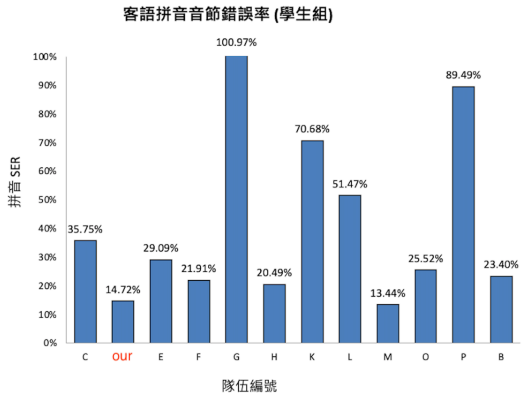
4 Conclusion

此篇研究以 Whisper-ASR 爲基礎，針對台灣客語的語音資料進行優化，透過資料前處理，SpecAugment 資料增強、語音拼接、語速調整、加入噪音以及 LoRA 技術等方法，嘗試提升模型的效能，由於目前我們所訓練的資料大多都是乾淨的聲音，所以這些資料增強無法明顯看出模型適應不同環境，但我們也自己設想，在什麼樣的環境下，哪種資料增強技術對最後的辨識結果會有較大的提升。熱身賽的成績如表 7所示。

Table 7: 熱身賽成績

Type	Character	Pinyin
CER(%)	10.51	X
SER(%)	X	14.72

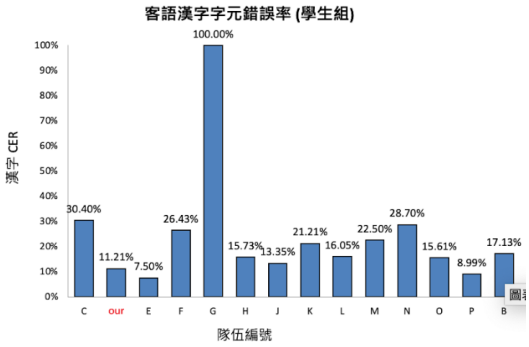


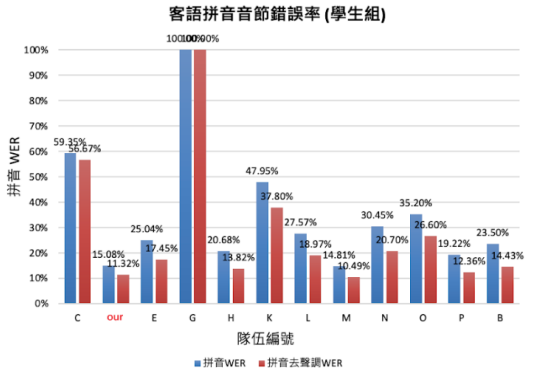


在決賽的測試資料中，我們發現語音檔案中含有大量雜音。爲了應對這個情況，我們採取的策略是：先將訓練資料進行降噪處理，再訓練一個新模型，接著也對主辦方提供的決賽資料進行同樣的降噪，最後再用訓練好的模型進行預測。決賽的成果如表 8 所示。

Table 8: 決賽成績

Type	Character	Pinyin
CER(%)	11.21	X
SER(%)	X	15.08
SER(去聲調)(%)	X	11.32





本研究初步展示了 Whisper 對於不同語言的可調整性，透過此開發以及實驗，提供了可行的方向，協助保存與推廣逐漸式微的台灣客語，未來可以朝向蒐集各種情境下的客語語料，來讓模型能夠適應各種不同的環境，以期在語言科技與文化傳承上做出更多貢獻。

Acknowledgment

感謝顏安孜教授在本研究過程中提供 GPU 資源，並感謝國家高速網路與計算中心提供環境讓我們使用，感謝 Formosa Speech Recognition Challenge 2025 主辦方提供訓練資料，讓本研究得以圓滿完成。

References

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

[2] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le, "SpecAugment: A

Simple Data Augmentation Method for Automatic Speech Recognition," in *Proceedings of Interspeech 2019*, pp. 2613-2617, 2019. DOI: 10.21437/Interspeech.2019-2680.

[3] Zheshu Song, Jianheng Zhuo, Yifan Yang, Ziyang Ma, Shixiong Zhang, and Xie Chen. 2024. *LoRA-Whisper: Parameter-Efficient and Extensible Multilingual ASR.* arXiv preprint arXiv:2406.06619.

[4] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. *LoRA: Low-Rank Adaptation of Large Language Models.* In ICLR.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. *Attention is All you Need.* In NeurIPS.