The EZ-AI System for Formosa Speech Recognition Challenge 2025
針對 2025 福爾摩沙客語語音辨識競賽的 EZ-AI 辨識系統

Yu-Sheng Tsao[†], Hung-Yang Sung[‡], An-Ci Peng[‡], Jhih-Rong Guo[‡], Tien-Hong Lo[‡]
[†]EZ-AI, [‡]National Taiwan Normal University
sam@ez-ai.com.tw, redsheep@ntnu.edu.tw, anci.peng@gmail.com,
jhih.rong.guo@gmail.com, teinhonglo@ntnu.edu.tw

摘要

本研究參與 2025 客語語音辨識競賽的拼音軌及漢字軌，針對大埔腔與詔安腔兩個低資源腔調，設計並比較不同的語音辨識系統。我們的核心策略是透過跨語言遷移學習 (Transfer Learning)，有效利用相近語系的資源，並結合自監督學習 (Self-Supervised Learning, SSL) 以提升模型在拼音軌的辨識效能。在漢字軌方面，則使用 Whisper 模型並搭配 LoRA (Low-Rank Adaptation) 進行微調。爲了緩減語料不足的限制，我們採用兩種資料擴充方法：模擬對話式語音以處理多語者情境，以及利用文字轉語音 (Text-to-Speech, TTS) 生成額外的詔安腔語料。在熱身賽的結果顯示，遷移學習在拼音軌表現尤爲顯著，使系統於所有隊伍中取得平均字錯誤率 (Character Error Rate, CER) 19.57%，排名第三；在漢字軌中，Whisper 結合 LoRA 系統則達到平均 CER 6.84%，並獲得社會組第一名。本研究證明遷移學習與資料擴充能有效提升低資源語言的辨識表現，但在媒體語料的領域落差下仍存在挑戰，未來將探索語境學習 (In-Context Learning, ICL) 與熱詞建模 (Hotword Modeling) 以改善此問題。

Abstract

This study presents our system for Hakka Speech Recognition Challenge 2025. We designed and compared different systems for two low-resource dialects: Dapu and Zhaoan. On the Pinyin track, we gain boosts by leveraging cross-lingual transfer-learning from related languages and combining with self-supervised learning (SSL). For the Hanzi track, we employ pre-trained Whisper with Low-Rank Adaptation (LoRA) fine-tuning. To alleviate the low-resource issue, two data augmentation methods are experimented with: simulating conversational speech to handle multi-speaker scenarios, and generating additional corpus via text-to-speech (TTS). Results from the pilot test showed that transfer learning significantly improved performance in the Pinyin track, achieving an average character error rate (CER) of 19.57%, ranking third among all teams. While in the Hanzi track, the Whisper + LoRA system achieved an average CER of 6.84%, earning first place among all. This study demonstrates that transfer learning and data augmentation can effectively improve recognition performance for low-resource languages. However, the domain mismatch seen in the media test set remains a challenge. We plan to explore in-context learning (ICL) and hotword modeling in the future to better address this issue.

關鍵字：客家語、語音辨識、FSR-2025

Keywords: Hakka, Speech Recognition, FSR-2025

1 簡介

儘管語音辨識的研究於主流語種進展快速，欲解決的研究問題已從通用場景延伸至不同的小眾市場，但針對資源匱乏的弱勢語言，如何善加發揮仍然是一項具挑戰性的題目。

此次 2025 客語語音辨認競賽聚焦在兩個弱勢腔調：大埔腔、詔安腔。認知到這兩個腔調語料有限，欲達到較佳的辨識結果勢必會需要額外的語料參與，儘管腔調上有所不同，同屬客語語系的相似腔調語料在模式上還是會有所幫助 (Qian et al., 2024)。因此我們的做法分爲兩部分：最大化既有語料的表現，以及擴增目標語言的資料。

在低資源語言的研究領域中，使用預訓練模型是熱門且相對簡單的手段，如何能夠在有限資源中有效的學習也是此領域的一大重點 (Piñeiro-Martín et al., 2024)。方法上可以是：

- 尋找類似語系，經大規模語料訓練過的先進模型，對其解碼器進行遷移訓練

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 476–480

- 使用自監督模型如 wav2vec, WavLM 或 HuBERT 並訓練其進行下游的語音辨識任務 (Zhao and Zhang, 2022)

- 基於語音辨識基礎模型如 Whisper (Radford et al., 2023)，因其訓練所用的語料以及任務設定，使得 Whisper 能夠快速地適應不同的語料標記，並達到足夠強健的辨識結果這些方法讓稀少資源的語料也能善用既有的優異語音特徵，從有限的標記中達到較理想的辨識效果。

另外我們觀察到在測試語料中的媒體語料子集具有比較複雜的語音環境，如對話、噪音、遠場等性質，以及可能測試語料與訓練語料的領域差異導致表現不佳，我們參考過往研究與比賽經驗，合成相似性質的語料進行訓練，以改善辨識結果。

2 策略與方法

2.1 K2 與 SSL

K2 爲 Kaldi (Povey et al., 2011) 作者所建立團隊進行開發的語音處理框架，具辨識效果良好、易操作、運算快速、節省資源等優勢，並且在中文語系的常見語料都有預訓練模型可供快速實驗；另外在自監督模型上，也有對應的研究 (Yang et al., 2024) 能夠套用如 wav2vec、HuBERT 等模型，進行下游任務的訓練，故在漢字賽軌，我們會先已 HuBERT + RNN-T 的方式訓練，將資料集擴增的策略在此模型上做初步的嘗試。

在拼音賽軌，由於 Whisper 最相近於客語的語系爲中文，但解碼器在該語系已經被訓練至對漢字比較拿手，微調其輸出拼音，又或是重置解碼器都是相對次優的做法，所以拼音我們會使用 K2 zipformer 於 WenetSpeech 預訓練的模型，重置其解碼器使其輸出拼音。

2.2 Whisper

Whipser 爲 OpenAI 所發表的語音辨識基礎模型，使用常見的 Transformer 架構，訓練在 68 萬小時自網路蒐集、多數來自 Youtube 影片的多語言語料，訓練任務爲轉錄與轉譯（至英文）。由於語料的多樣性，Whisper 對於常見的環境變異都有良好的強健性，欲微調相似語系時也只需較少的語料就能有所改善，至今仍是熱門的語音辨識模型。

然而此模型若想要在中文上有可靠的辨識能力（準確度高於八成），至少得選用參數量 small 以上的模型，而訓練更大的模型卻伴隨著更長的訓練時間，不利於比賽的實驗迭代，故我們會先在 K2 探索適合的語料設定，再套用至強健性較佳的 Whisper，並使用 AdaLoRA (Zhang et al., 2023) 技術降低訓練的運算成本。

Corpus	Spks.	Sents.	hrs.
Train			
Dapu	64	12197	31.43
Zhaoan	59	15152	30.59
Eval (Pilot test)			
Studio - Dapu	10	1304	4.01
Studio - Zhaoan	11	2154	4.00
Media - Dapu	-	445	1.08
Media - Zhaoan	-	501	1.13
HakkaCouncil			
Reading - Sixian	208	-	396
Reading - Hailu	151	-	300
TTS - Zhaoan			
OOV	9	682	9.65
E-Learning	9	124588	136

Table 1: 比賽的訓練與測試語料，與 TTS 語料的統計資訊

Unique Words	Train	Eval (OOV)	
		Studio	Media
dapu	5771	2323 (230)	1552 (483)
zhaoan	4911	2221 (57)	1317 (275)

Table 2: 訓練與測試語料的詞目統計，括號中爲遺失字數量

2.3 資料擴增

在 K2 的初步實驗中，我們發現儘管對於錄音室測試語料的漢字辨識能力已能達到九成以上的正確率，在媒體測試語料上卻不到三成；同樣的情況也發生在 Whisper 的結果上，尤其是紹安腔的部分，與大埔腔的字錯誤率差了大約 2.5 倍。我們進一步分析發現，媒體語料相較於錄音室語料會有較多的遺失字 (Out-of-Vocabulary, OOV)，在紹安腔所以我們使用近期常見於稀少資源語料的擴增做法：透過 TTS 合成額外語料 (Chen et al., 2023)，來試圖提高在媒體語料上的表現。在這裡我們使用 FormoSpeech 團隊的 TTS 模型 yourtts-htia-240704[1] 進行語料的生成。

另外，由於媒體語料的組成大多爲對話式的語音，若只用朗讀型的語料訓練，模型在遇到語者的語音重疊或是被打斷時，辨識結果會產生明顯衰退，所以我們在 K2 的訓練額外合成了對話式的語料，以提高媒體語料的辨識率。

[1]https://huggingface.co/formospeech/yourtts-htia-240704

Exps. (CER%)	Studio			Media			Total Avg.
	Dapu	Zhaoan	Avg.	Dapu	Zhaoan	Avg.	
Train	6.78	6.46	6.62	73.98	80.01	77.00	41.81
+ft speed&reverb	6.15	5.64	5.90	68.48	77.77	73.13	39.51
Train&Conversation	6.12	5.25	5.69	63.71	74.92	69.32	37.50
Train&Conversation (TTS)	3.80	3.52	3.66	62.34	57.37	59.86	31.76

Table 3: 漢字軌於 k2 框架進行的實驗結果

Exps. (CER%)	Studio			Media			Total Avg.
	Dapu	Zhaoan	Avg.	Dapu	Zhaoan	Avg.	
FormoSpeech/hakka	9.47	29.95	19.71	14.56	41.19	27.88	23.79
+ft Train	1.11	2.22	1.67	8.71	21.13	14.92	8.29(6.84)
+ft Train & OOV (TTS)	1.13	3.26	2.20	7.85	21.11	14.48	8.34

Table 4: 漢字軌於 Whisper 的實驗結果，總平均欄位的括號爲官方所回報之結果

3 實驗設定

3.1 資料集

除了決賽的結果會加入熱身賽的測試語料進行訓練外，其他實驗的訓練語料皆不包含測試語料。這些語料的統計資料如表 1。

TTS 語料進一步分成兩種，我們先使用客語能力認證的文字語料進行一般性用詞的語音生成，但發現只用這個領域的文字語料並不足以改善媒體測試語料的辨識率，我們便仔細檢視媒體語料的標記，並與客家詔安腔字典進行比較，如表 2，鎖定分詞結果不在訓練語料的句子進行合成。雖然就統計上來看大埔腔在 OOV 的字詞比例較多，但由於詔安腔的表現較差，故我們優先以詔安腔進行語料的合成。

3.2 硬體與參數

在本次的大部分實驗中，我們採用兩款不同型號的 GPU 進行運算，分別爲 NVIDIA GeForce RTX 3090 與 4090。不論是在 k2 框架下或是使用 Whisper，所需時間均大約爲 18 至 24 小時。在 k2SSL 的實驗上我們參考原始論文的訓練參數，訓練最多 200 週期 (epoch)，再挑選收斂的區間進行權重平均後，使用貪婪搜尋法 (Greedy Search) 進行解碼。

對於 Whisper 的類型挑選，我們站在巨人的肩膀上，使用 FormoSpeech 團隊所公開的 whisper-large-v3-taiwanese-hakka[2] 模型作爲基底，此模型使用台灣最常見的六種客語腔調進行微調訓練，直接對這次比賽的測試語料辨識就已經具有不錯的表現，我們即固定使用這個模型做爲基準，進一步使用這次比賽語料進行微調，另外也參考了先前比賽的報告，加入資料擾動如：速度、音高變動與空氣吸收 (AirAbsorption) 以適應測試語料媒體子集的聲學環境，最多訓練 10 個週期。推論則挑選驗證集損失最低的單一檢查點，大部分收斂落在 3~5 週期左右。

[2]https://huggingface.co/formospeech/whisper-large-v3-taiwanese-hakka

4 實驗結果

4.1 漢字軌

4.1.1 K2SSL 實驗

初步實驗我們採用 K2SSL 研究中的 zipformer-based HuBERT 模型 (由 HuBERT-base-ls960 衍生) 作爲編碼器訓練 RNN-T 系統進行辨識，如表 3，僅使用比賽訓練資料的話，雖然能在錄音室語料上達到接近九成五的辨識率，但在會議語料上卻僅有兩成左右，即使進一步增加資料的擾動，改善的程度也有限。

觀察媒體語料的組成後，我們將訓練語料加上擾動，產生模擬對話情境的語料再次訓練。模擬對話情境相較僅使用朗讀風格的訓練語料有更爲明顯的改善，但在媒體語料的部分，詔安腔的表現則明顯弱於大埔腔。故我們蒐集詔安腔能力測驗的文字語料，使用 TTS 產生詔安腔的合成語料後，再次模擬對話情境進行訓練，在詔安腔媒體語料降低了 17% 的字錯誤率，並也一併改善了錄音室語料的辨識率。

然而媒體語料的整體辨識率仍不及五成，我們推測由於 HuBERT-base 因僅訓練在 Librispeech 的朗讀語料，仍不具有足夠的強健性處理複雜的聲學情境，因此接下來我們會使用 Whisper 進行。

4.1.2 Whisper

實驗結果如表4，我們將 FormoSpeech 團隊所微調的模型作爲基準值，使用其直接對測試

Exps. (CER%)	Studio			Media			Total Avg.
	Dapu	Zhaoan	Avg.	Dapu	Zhaoan	Avg.	
Zipformer-HuBERT	4.55	19.15	11.85	35.50	59.34	47.42	29.64
Wenet-Zipformer +ft '23, '25 train	5.77	10.78	8.69	20.20	40.29	31.37	17.17
Wenet-Zipformer +ft 客委會->'23, '25 train	5.46	10.36	8.31	21.07	38.75	30.90	16.76
Whisper +ft Train	8.32	17.24	12.78	26.56	31.52	29.04	20.91(19.60)

Table 5: 拼音軌的實驗結果，總平均欄 (Total Avg.) 的括號內數字爲官方所回報之結果

語料進行辨識，在大埔腔的錄音室與媒體語料均有接近九成的辨識率，得益於 Whisper 對複雜聲學環境的強健性，詔安腔則相對較爲弱勢，所以在 Whisper 的實驗上我們仍然是聚焦在改善詔安腔的辨識結果。

使用比賽的訓練語料進行微調後，在錄音室語料上就有大幅度的改善，兩個腔調的平均字錯誤率從 19.71% 下降至 1.67%，推測是基底模型在訓練時詔安腔語料不足的關係；媒體語料也從平均 27.88% 下降至 14.92%，儘管如此，媒體語料的詔安腔錯誤率仍居高不下，即使我們進一步針對分詞後的 OOV 去產生合成語料，也僅僅是讓大埔腔的辨識結果稍微改善，詔安腔的改善仍然有限。

對此，針對媒體語料進一步分析錯誤結果，應是媒體語料含有比例不少的專有名詞，導致即使模型已經在不同聲學環境、額外的 OOV 合成語料上訓練了，面對專有名詞依然是無法妥善的辨識。

4.2 拼音軌

我們使用在漢字軌上較爲有效的策略訓練拼音軌的模型：在 k2 框架上採用自監督模型或是預訓練模型，並適時增加語料，考慮到儘管腔調不同，拼音書寫均爲一致。在 Whisper 則是直接使用比賽語料進行訓練。

由於 Whisper 解碼器的設計，將其重置再訓練將會喪失訓練過大量語料的優勢，故我們沿用原本的設定，微調中文語言讓他能夠輸出拼音。而 k2 模型因沒有這類限制，所以我們能夠直接訓練其解碼器輸出拼音。

實驗結果如表5，在錄音室語料上，兩種 k2 模型的拼音辨識結果都比 Whisper 更加準確；在媒體語料方面，即使 Whisper 因預訓練語料，比起 Zipformer-HuBERT 表現更穩定，但其優勢並不如漢字軌一般明顯，一旦換上 WenetSpeech 預訓練過的 Zipformer (下稱 Wenet-Zipformer)，只需針對拼音解碼的模型在整體的辨識效果上仍比較理想。如果我們兩階段的先將 Wenet-Zipformer 用客委會[3]的資料微調，再微調至 2023 & 2025 年的比賽資料，能進一步改善模型的辨識結果，在熱身賽的測試資料上達到平均字錯誤率 16.76%。

4.3 熱身賽結果

因爲熱身賽的時程關係，繳交的時候我們在漢字與拼音軌均使用 Whisper 的結果進行投稿，漢字軌錯誤率 6.84% 取得了社會組及所有隊伍的第一名，而拼音軌則取得了錯誤率 19.57%，位居所有隊伍的第三名。

4.4 決賽結果

考慮到決賽的語音可能也會與媒體測試語料相似，我們將 75% 的媒體語料加入漢字軌 Whisper 的訓練，訓練過程的評估指標則使用錄音語料與剩下的媒體語料計算，加入部分媒體語料後的測試集可以觀察到明顯的改善，若使用決賽語料去評估加入媒體語料前後的辨識結果之字元差異，也能得到 10% 左右的差異結果，故我們使用這顆 75% 媒體語料的模型進行漢字軌辨識的結果提交，得到 CER 9.46% 的成績，位居所有隊伍的第三名。拼音軌我們使用 Wenet-Zipformer 進行提交，儘管拼音軌應能直接的辨識出不同腔調的拼音序列，但訓練時並無納入媒體測試語料以及客委會媒體語料，或許導致模型在複雜聲學環境仍不夠強健，最終拿到了拼音 WER 30.44% 的成績，位於所有隊伍的第八名。

5 結論與展望

此次比賽我們參考過去的實驗結果與經驗，透過分析標記並增加合成語料，試圖改善詔安腔媒體語料存在過多遺失字與專有名詞，導致漢字軌辨識率居高不下的情況，不過因爲增加的絕大多數都屬於領域外資料，改善有限。未來我們會試著探討語境學習 (Incontext Learning) 或是熱詞等方式進行擴增或調校，改善領域外資料的辨識效果。

[3]https://www.aclclp.org.tw/doc/hat_brief_c.pdf

References

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 *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).

Andrés Piñeiro-Martín, Carmen García-Mateo, Laura Docio-Fernandez, María del Carmen López-Pérez, and Georg Rehm. 2024. Weighted cross-entropy for low-resource languages in multilingual speech recognition. In *Interspeech 2024*, page 1235—1239. ISCA.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, et al. 2011. The kaldi speech recognition toolkit.

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 *International conference on machine learning*, pages 28492–28518. PMLR.

Yifan Yang, Jianheng Zhuo, Zengrui Jin, Ziyang Ma, Xiaoyu Yang, Zengwei Yao, Liyong Guo, Wei Kang, Fangjun Kuang, Long Lin, et al. 2024. k2ssl: A faster and better framework for self-supervised speech representation learning. *arXiv preprint arXiv:2411.17100*.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*.

Jing Zhao and Wei-Qiang Zhang. 2022. Improving automatic speech recognition performance for low-resource languages with self-supervised models. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1227–1241.