多模組錯誤檢測與修正的客語語音辨識系統
A Multi-Module Error Detection and Correction System for Hakka ASR

Min-Chun Hu
National Cheng Kung University Department of Computer Science and Engineering
p76131563@gs.ncku.edu.tw

Yu-Lin Xiao
National Cheng Kung University Department of Computer Science and Engineering
p76121518@gs.ncku.edu.tw

Wen-Hsiang Lu
National Cheng Kung University Department of Computer Science and Engineering
whlu@mail.ncku.edu.tw

摘要

本研究提出一個針對客語（以大埔／詔安腔為主）的自動語音辨識（ASR）後矯正系統，旨在解決低資源語言辨識錯誤率偏高的問題。客語因受限於語料規模、異體字與腔調差異，在既有的通用 ASR 模型上表現往往不佳。為此，我們首先以 Whisper Large v3 Turbo 為基底辨識模型，使用約 60 小時的大埔與詔安語料進行微調，以提升對特定腔調的適應性。在獲取 ASR N-best 候選句後，系統進一步透過多模組錯誤偵測矯正流程進行修正，包含四個主要步驟： (1) 潛在錯誤偵測，用於鎖定候選間錯誤的候選詞彙；(2) 音素混淆集偵測（Phoneme Confusion Set）: 依據音素相近關係提供可能替代詞；(3) 辭典（Lexicon）修正: 確保詞彙存在於語言使用的實際範疇中，(4) 搭配詞關聯度偵測: 利用收集之語料所建立的搭配詞關聯度來偵測錯誤詞彙。本研究所提出的矯正機制能有效補足 ASR 在低資源語言中的不足，實驗顯示經過多階段錯誤偵測矯正後，最終 CER 減少至 15.49%，減少 2.14 % ，證明該方法能有效提升語音辨識的準確率。

關鍵字：語音辨識、客語、錯誤矯正、混淆集、搭配詞

。

Abstract

This study proposes a post-correction system for Automatic Speech Recognition (ASR) targeting Hakka (with a focus on the Dapu and Zhao'an dialects), aiming to address the high error rate commonly observed in low-resource languages. Due to limitations in corpus size, the existence of variant characters, and dialectal differences, Hakka often performs poorly on general-purpose ASR models. To improve recognition, we first fine-tuned Whisper Large v3 Turbo with approximately 60 hours of Dapu and Zhao'an speech data, enhancing the model's adaptability to these specific dialects. After generating the ASR N-best candidates, the system performs a multi-module error detection and correction process consisting of four main steps: (1) potential error detection to identify suspicious words among candidates; (2) phoneme confusion set detection, which provides alternative words based on phonetic similarity; (3) lexicon-based correction to ensure that words belong to valid linguistic usage; and (4) collocation-based detection, which leverages word association scores derived from collected corpora to identify contextually inconsistent words. The proposed correction mechanism effectively compensates for the limitations of ASR in low-resource languages. Experimental results show that, after multi-stage error detection and correction, the final Character Error Rate (CER) was reduced to 15.49%, achieving a 2.14% absolute reduction, thereby demonstrating that the method can effectively enhance ASR accuracy.

Keywords: speech recognition, Hakka, error correction, confusion set, collocation

1 緒論

客語屬於低資源語言，現有公開語料相對稀缺，且漢字用法中存在異體字、詞彙多形

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 481–488
November 20-22, 2025

以及方言差異等挑戰。在自動語音辨識（ASR）的輸出結果中，常見的錯誤類型包括：（i）同音近形所造成的用字錯誤、（ii）華客語彼此搶詞引發的混淆，以及（iii）語意搭配不精確等問題。這些錯誤不僅影響辨識結果的可讀性，也限制了 ASR 系統在真實場域中的應用效益。因此，本文旨在 ASR 輸出後進行精細化的矯正，以提升整體客語辨識的準確度與實用性。

本研究所提出的 ASR 系統基於 Whisper Large v3 Turbo 的基底辨識模型，並以約 60 小時的大埔與詔安腔語料進行 fine-tuning 訓練。在獲得 N-best 候選句後，系統會依序經過多階段的錯誤偵測矯正模組：首先透過跨候選句的詞彙差異統計來標記潛在詞彙錯誤位置；其次利用混淆集提供音素近似的替換建議；再透過搭配詞分數檢查語境合理性；最後以辭典檢驗輸出，避免輸出未收錄或極罕見的詞彙。此流程兼顧可擴充性與可解釋性，適合於商業應用實務中逐步迭代改進。

然而，本研究同時也面臨低資源語言的典型挑戰：一方面，深度學習模型如 Whisper 在大規模語料上能展現優異的表現，但在客語這類語料有限的情境下，辨識效能往往因缺乏詞彙完整覆蓋性而受到限制；另一方面，若僅依靠人工擴充語料，則需付出龐大的人力與時間成本，而異體字及方言差異更使資料標註的一致性難以維持。因此，如何在「模型效能」強化與「語料資源不足」改善之間取得平衡，並透過多模組的後處理錯誤偵測矯正機制來彌補 ASR 的不足，即是本研究的核心挑戰與主要貢獻。

2 相關研究

2.1 語音辨識模型（Speech Recognition Model）

傳統的音素式（phoneme-based）ASR 模型（Daniel et al., 2011）在錯誤矯正上具有一項天然優勢：它們會產生音素層級的輸出，可作為辨識後矯正（post-recognition correction）的額外參照資訊。相較之下，非音素式（non-phoneme-based）End-to-End model（如 Baevski et al., 2020；Radford et al., 2023）多半是直接從語音映射到文字，缺少顯式的音素資訊。像 Kaldi 與傳統混合式 ASR 這類音素式系統，會保留音素序列與對齊資訊，因而能支援基於音素的對齊與矯正策略。然而，儘管音素式方法在錯誤矯正上具可用訊息，其整體辨識正確率通常仍不及最新的 End-to-End model。這類系統高度依賴發音詞典、音素對齊與人工詞彙資源，在跨語言或跨領域時彈性受限。在近年的 ASR model 中，Whisper（Radford et al., 2023）以其在多語言、多領域上的穩健性與廣泛實務採用特別受到關注；其在大規模有標註的平行語料上訓練，使其具備良好的泛化能力，在未特別微調的情況下亦能有不錯表現。不過，在特定領域或是特定語言(例如:客語的特殊腔調等等)，Whisper 的表現仍可能不理想。主要原因在於預訓練語料中領域相關資料稀缺，導致辨識錯誤偏多。因此，實務上常需要額外的調適或錯誤矯正機制，才能在此類高風險情境中達到可靠表現。

2.2 非音素式 ASR 的錯誤矯正（Non-phoneme-based ASR Error Correction）

在 ASR 後矯正領域，非音素式的方法亦有顯著進展。許多研究（Ma et al., 2023）不再僅依賴單一最佳假設（1-best），而是利用 N-best 假設作為輸入，以提供較豐富的候選與語境訊息，提升矯正的準確率。儘管如此，面向特定領域的 ASR 矯正仍具挑戰，尤其在專業術語與語言變體的處理上。部分研究（López-Cózar & Callejas, 2008）嘗試將語意、句法、詞彙與語境納入模型，使矯正更貼近自然語言使用情境，進而得到更精準且自然的修正。

2.3 音素式 ASR 的錯誤矯正（Phoneme-based ASR Error Correction）

近年亦有越來越多工作將音素層級訊息用於改善 ASR 後矯正。不同於傳統必須倚賴辭典與對齊的音素式 ASR，一些方法改為在辨識之後再運用音素序列來偵測與修補可能錯誤。例如，Serai et al.（2019）提出以聲學模型的後驗機率來進行抽樣，取代固定的混淆矩陣，藉此較真實地模擬 ASR 錯誤，並使錯誤行為更貼近現代 ASR 的實際狀況。另有方法（Wang et al., 2022）將音素（phonetic）與語意（semantic）資訊結合 N-best 假設共同

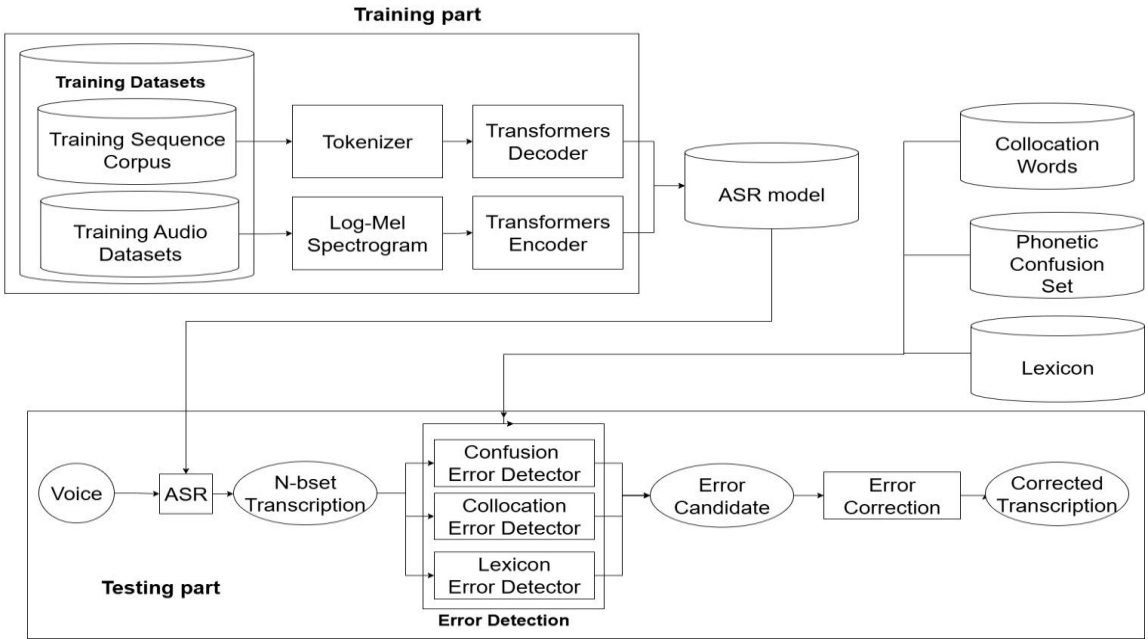

圖 1:系統架構

使用，用以偵測「聽起來合理但語意不通」的錯誤，進一步提升矯正準確率。

3 研究方法

3.1 系統架構

- Training part

在訓練階段，本研究以客語（大埔／詔安）語料為基礎，建構資源較稀少的語音辨識模型。首先語音資料被轉換為 Log-Mel Spectrogram，文字資料則經過 Tokenizer 轉為符號序列，兩者分別進入 Transformer Encoder 與 Transformer Decoder 進行訓練(參考圖 1)。透過編碼器與解碼器的互動，完成語音與文字的對應學習，最終得到基於 Whisper Large v3-turbo 微調之 ASR model。此模型能夠捕捉客語聲學與語言特徵，為後續測試與錯誤檢測模組奠定基礎。

- Testing part

本研究設計之客語（大埔／詔安）語音辨識矯正系統包含三大特徵(參考圖 1)，分別為自動語音辨識（ASR）、潛在錯誤偵測、混淆集合（Confusion Set）、語意搭配檢測（Collocation）、辭典比對（Lexicon）以及最終輸出。整體流程如圖 1 所示。

在 ASR 階段，系統以 Whisper large v3-turbo 為核心，並利用 60 小時的大埔與詔安語料進行微調。於解碼過程中，模型會產生多個候選句輸出（N-best），提供後續模組進一步比對與修正。

接下來的錯誤偵測模組透過候選之間的對齊與統計分析，辨識在相同位置上出現多種不同字詞的情況，並將其標記為潛在錯誤點。其內部包含三個子模組：(1) 音素混淆偵測模組(2) 辭典比對模組(3) 語意搭配檢測模組。

最後，錯誤候選會進入 Error Correction 模組。該模組依據錯誤檢測所提供的候選字詞，結合辭典與語境約束，挑選出最合適的修正結果，最終生成 Corrected Transcription，提供較為穩健且符合語意的客語逐字稿輸出。

3.2 語音前後處理

在進行語音辨識前，系統會先進行前處理。首先，利用 Silero-VAD 偵測並擷取有效的語音區段，去除靜音與雜訊部分，確保輸入訊號的品質。接著，所有音訊皆統一重取樣為 16 kHz 並轉換為單聲道，以符合模型的輸入需求。此外，輸出候選詞之文字亦需正規化處理，包含將簡體字轉換為繁體字，以及針對異體字進行客語漢字替換，以避免因文字形式不一致而導致後續比對失敗。

3.3 潛在錯誤區域偵測

潛在錯誤偵測模組以漢字為最小對齊單位，針對多個候選輸出進行逐字對齊，並統計每個位置的輸出詞彙多樣性。若在相同位置上出現兩種以上的不同結果，即判定為潛在錯誤區域。此設計利用了 N-best 解碼輸出在錯誤區域往往呈現高分歧度的特性，藉由辨識候選間的差異，能夠有效捕捉可能的錯誤字詞，並作為後續修正模組的輸入依據。接著系統將透過三個錯誤偵測模組來進行處理: (1)音素混淆偵測模組:根據音近或形近關係，建立潛在替代詞的候選集，(2) 辭典比對模組: 整合客語辭典作為修正依據(3) 語意搭配檢測: 透過大規模文本計算詞與詞之間的相關性，評估候選詞或片語與句內其他詞語之間的搭配合理性。

3.4 錯誤檢測

錯誤檢測模組的核心在於辨識 ASR 輸出中潛在的異常字詞，作為後續矯正的基礎。系統透過多候選序列的比對與對齊，觀察在同一位置出現高分歧度的情況，並將其標記為可疑區段。此方法能有效捕捉音素混淆、語境不符或辭典外詞彙等錯誤來源，縮小矯正範圍並提升整體修正的精準性。

3.4.1 音素混淆偵測模組

本模組中，系統同時考慮字級與短語級的混淆情境，以建立潛在的替代候選。於字級層面，系統根據客語的聲母、韻母、鼻化、送氣與腔口等音素差異，為每一潛在錯誤字生成音近的候選字詞。例如，聲母之間的變異如 s/ts/tsh、p/ph/f 以及 n/l 等，皆是常見的聲母混淆來源。於短語層面，則考慮整體語境下的音近短語，當某些詞組在語音或語意上高度相似時，即納入候選。例如，「貧」與「鼻」在拼音上為 pin113 與 pi53，在實際辨識過程中容易因發音相近而產生混淆。此時，音素混淆集合（Phoneme Confusion Set）(參考表 1)便能提供可能的替代字詞，協助系統在後續的偵測與修正階段進行判斷。

漢字及音素	說明
貧↔鼻 pin113↔pi53	拼音相近，常在口語中混淆
呼↔福 fu113↔fug21	拼音相近，常在口語中混淆
過↔擱 go53↔gog21	拼音相近，常在口語中混淆

表 1 : Phoneme confusion error 範例

3.4.2 錯誤詞彙偵測模組

在錯誤詞彙偵測階段，系統藉由客語辭典進一步過濾與修正候選詞。若原始辨識候選詞不在辭典中，但其混淆候選詞存在於辭典內，則系統傾向將輸出修正為候選詞。舉例來說(參考表 2)，當系統將「頒獎」誤辨識為「班獎」時，由於「班獎」並非辭典中的詞彙，而「頒獎」則是存在於辭典中，系統便會修正為「頒獎」，當作矯正候選。此外，為確保修正結果的合理性，系統遵循「從左至右、最長片語優先」的原則，並避免修正片段之間的互相重疊。此外，系統也能新增詞彙，例如，當輸入為「新聞高」時，辭典中並無此詞，但其音近的「新聞稿」是合理詞彙。此時將「新聞稿」新增至辭典中，並將辨識結果修正為「新聞稿」，以便後續辨識與修正能更準確地處理此類詞彙。

原詞 -> 修正	理由
新聞高 → 新聞稿	辭典無"新聞高"，但近似音"新聞稿"存在辭典中
讀立 → 獨立	辭典無"讀立"，但近似音"獨立"存在辭典中
班獎 → 頒獎	辭典無"班獎"，但近似音"頒獎"存在辭典中

表 2 : 辭典修正

3.4.3 語意搭配詞檢測

在大埔與詔安腔的口語逐字稿中，部分辨識錯誤並非單純由於語音近似或句法不合所導致，而是源自於詞語搭配上的語境不合理。例如，若辨識結果為「蠶窟的事實」，則顯得語意不通，因為「蠶窟」與「事實」並非合理組合；相對地，「殘酷」與「事實」則經常並置使用，語境上更為自然。同樣地，若辨識結果為「學曉天光日當晝會辦一場考

試」也會顯得語意不通順，因為「學曉」與「考試」的無搭配詞關聯度，但「學校」往往與「考試」共同出現，形成高頻搭配，進而將「學曉」替換成「學校」。為了系統化偵測此類語境層面的不一致，本研究在偵測流程中引入搭配詞錯誤檢測模組。

此模組採用多個通用主題語料庫，包括教育、醫療、公共／新聞以及日常生活等領域，並透過目前蒐集到的客語文本（如新聞稿、廣播逐字稿、教材與口語語料）計算雙詞之間的搭配詞關聯度。其核心概念是比較兩個詞在語料中同時出現的頻率，與它們各自單獨出現的頻率進行對比，用以衡量兩詞的相關性。分數越高，代表兩個詞在語境中越常被搭配使用，也越符合語境搭配。

例如，「殘酷」與「事實」通常具有較高的搭配詞關聯度，因此為自然的詞語組合；相反地，「蠶窟」與「事實」的關聯度明顯偏低，語意上難以成立，系統會傾向將「蠶窟」修正為「殘酷」。同樣地，「學校」與「考試」的搭配頻繁度很高，因此在候選包含「學曉／學校」的情況下，系統會選擇語境上更合理的「學校」。

3.5 錯誤矯正

完成錯誤偵測後，系統進入錯誤矯正階段。此階段的重點是整合各模組所提供的候選詞，依據語音特徵、詞彙合法性與語境搭配等面向進行比對與篩選，輸出最符合語言使用情境的結果。矯正並非單一規則，而是多種訊息的綜合判斷。例如：

- 音素混淆集合：若將「貧民窟」誤辨為「鼻民窟」，可依音近關係修正，恢復合理詞彙。
- 辭典比對：當輸出為「該係吾夢相」時，透過辭典判定後修正為「該係吾夢想」，以回復正確語義。
- 在搭配詞的情境下，若辨識結果出現「學曉考試」，因「學曉」與「考試」缺乏語境關聯，而「學校」與「考試」則是高頻搭配，系統便能將「學曉」修正為「學校」。

透過這些不同來源的訊息整合，錯誤矯正模組能顯著提升逐字稿的自然度與準確性，為最終輸出提供更穩健的保證。

4 實驗

本研究的 ASR 模型基於 OpenAI Whisper large-v3-turbo，並在客語大埔與詔安語料上進行微調，所得到的模型命名為 Hakka_dapu_za 在微調過程中，輸入採用 16 kHz 的 mel-spectrogram，解碼則使用 top-k 與 top-p 採樣策略以生成 N-best 候選序列，同時透過 Silero VAD 去除靜音與雜訊，確保輸入訊號的品質。為了驗證本研究方法的有效性，我們同時設置了 baseline 進行比較：baseline 模型為 Whisper large-v3-turbo，未經針對客語語料進行額外調整。藉由與 baseline 的對照，我們可以清楚評估微調與錯誤矯正模組對於辨識準確率的實際貢獻。

4.1 資料集

在語種與腔調方面，本研究聚焦於客語大埔腔與詔安腔。訓練語料主要採用客語競賽所提供的約 60 小時大埔與詔安語音資料（涵蓋日常對話、新聞播讀、教學講述等），並以此對 OpenAI Whisper large-v3-turbo 進行微調。語料同時涵蓋兩種腔調，並盡量在性別、年齡及錄音條件（錄音室／半自然環境）上保持平衡。所有音檔在前處理階段均被轉換為單聲道 16 kHz，以符合模型輸入需求，此外，本研究在錯誤偵測與修正模組中，分別建立三種輔助資源，拼音字典、辭典比對、以及搭配詞關聯度：

(1)拼音字典:根據語料所附的拼音與漢字，建立客語漢字對羅馬拼音的映射字典。此資源用於將客語漢字逐字轉換為拼音，並支援大埔與詔安常見用字。若遇到未收錄的字則標記為 NULL。

(2) 客語辭典:整合教育部客語辭典、大埔腔辭典與詔安腔辭典，建立詞條對應拼音的資料庫，避免輸出極少見的詞彙。

(3) 搭配詞關聯度計算:利用蒐集之文本計算詞與詞之間的關聯度，衡量語境中的搭配合理性。資料來源包含教育、醫療、公共／新聞及日常生活等領域的文本（涵蓋客語辭典、哈客平台文章、客語朗讀材料等，共計 16,324 篇以客語漢字撰寫的文章）。矯正時不限於相鄰詞，系統會在全句上下文中進行雙向比對（A→B 與 B→A）。

4.2 初始模型比較與選擇

在本研究的實驗設計中，第一步需要確定最適合作為後續研究基礎的初始模型。我們針對此部分進行了兩種設定的比較。在實際測試時，我們將初始的 60 小時語料進行分割，其中 50 小時作為訓練集，10 小時作為測試集。第一種設定是以 Whisper Large v3 Turbo 模型為基底，並先在約 800 小時的「海陸四縣」語料上進行微調。這樣的模型理論上在跨腔調任務上可能具備一定的泛化能力。然而實驗結果顯示，在此設定下模型的 CER（字錯誤率）為 16.91%，表現並不理想。第二種設定則是直接使用 Whisper Large v3 Turbo 的預訓練版本，在相同測試條件下其 CER 為 15.46%，較優於前者。

表格中 M1 為 Whisper Large v3 Turbo 先以海陸／四縣語料做基底，再以本研究 50 小時 大埔＋詔安語料微調。M2 為 Whisper Large v3 Turbo 預訓練版直接以本研究 50 小時 大埔＋詔安語料微調（不混入其他腔調）。

	M1	M2
CER	16.91%	15.46%

表 3: 模型準確率比較

根據表 3 可見，當在初始模型中混入不同腔調的語料（如海陸四縣）時，會造成負遷移效應，進而影響辨識精度，反而對目標腔調（大埔與詔安）的辨識任務產生負面影響。因此，本研究最終選擇不納入其他腔調的語料，而是直接以 Whisper Large v3 Turbo 的預訓練版本作為後續微調的基底模型，並專注於大埔與詔安的語料，期望藉由更聚焦的語言特性來獲得更高的辨識準確率。

在確定使用 M2 模型當作後續語音辨識模型。在此設定下，模型於客語競賽提供的熱身賽語料上表現的 CER 為 17.63% 。與未經微調的模型相比已提升 2.3%，但辨識結果中仍然存在部分錯誤，特別是來自音素相近的混淆、語境搭配不當以及異體字詞的使用等。這些問題若僅依賴單純的 ASR 模型仍不易解決，因此我們將本研究提出的後校正流程套用於模型的輸出，以進一步提升準確度與可讀性。本小節的實驗結果說明了初始模型選擇的重要性。避免跨腔調語料的干擾是提升精度的關鍵，而專注於大埔與詔安語料的設定，則為後續矯正模組的發揮提供了最佳的基礎。

4.3 矯正效果與分析

在確定基底模型後，本研究將完整的三階段矯正流程應用於 ASR 輸出，並選用 2025 年客語語音競賽所釋出的熱身賽語料進行測試，該資料集共計約 10 小時，涵蓋大埔與詔安腔的多樣語音內容，包括日常對話、新聞播報等等，能有效模擬真實使用情境。經過矯正處理後，系統的最終字 CER 從 17.63%降低至 15.49%，整體提升幅度為 2.14%。這樣的結果清楚顯示，本研究設計的多模組矯正流程，能有效修正音素混淆、語境搭配不當，以及辭典外詞彙的錯誤，並在低資源語言環境中展現顯著的實用價值。

表格中的 M2 代表未經矯正模組處理後之模型，M3 代表經矯正模組處理之模型。

	M2	M3
CER	17.63%	15.49%

表 4: 矯正模組前後之模型準確率比較

從具體例子(如表 5)來看，矯正模組能成功修正如「學曉 → 學校」這類音近字詞混淆，並利用考試相關語境進行合理替換；又如「讀立 → 獨立」，雖然音同，但「讀立」並不存在於辭典中，因此系統最終將其修正為合法的「獨立」，提升了輸出結果的可讀性與正確性。

原始辨識結果(M2)	修正後結果(M3)	原因
學曉天光日當晝會辦一場考試	學校天光日當晝會辦一場考試	“學 曉”和”學校”音近可用搭配詞根據考試進行矯正
厥等在世界个盡頭過讀立个生活	厥等在世界个盡頭過獨立个生活	獨立 音同讀立，但”讀立”不在辭典內，將它修正成”獨立”

表 5: 矯正範例說明

4.4 討論

總體而言，實驗結果呈現兩個重點：第一，在初始模型的選擇上，避免將其他腔調語料納入訓練是必要的，因為這樣能降低跨腔調干擾帶來的負遷移；第二，在模型的基礎上，再結合我們提出的矯正流程，可以顯著改善模型的準確率，使其在客語（大埔、詔安）的辨識任務中展現更穩健的性能。

以下針對三類錯誤分析說明:

(1) 字詞過短或缺乏上下文的詞彙：(參考表 6)如「教師」「教育」等單詞或短語，測試音檔過短等的問題。這類詞通常需要更長的句子上下文（如搭配「課綱／考試／授課／學校」等）才有足夠訊號做決策。

(2)字詞消失:(參考表 6)模型直接遺漏了某些應有的詞彙，使輸出序列缺少關鍵的語義成分。與一般的替換錯誤或音近詞混淆不同，字詞消失並非來自候選詞之間的錯配，而是序列生成本身的缺陷。這種情況在基於 Transformer 的架構中特別常見，可能導致部分詞彙在輸出時被忽略。由於後端的矯正模組主要依賴「候選對照」與「上下文推斷」，一旦關鍵詞未被輸出，就無法建立映射關係，自然也無法補回遺漏詞語。因此，字詞消失問題通常需要透過前端 ASR 模型的改進來解決，例如增強聲學建模、調整解碼策略或引入更強的語言模型，而非單純依靠後端矯正流程。

(3) 專有名詞辨識錯誤：(參考表 6)屬於難以矯正的情境。人名，地名，組織名等等屬於專有詞彙，通常不在語言模型或辭典的高頻詞範疇內，加上聲韻結構多樣（如壽氏麗），容易被誤辨為發音近似或隨機組合的詞（如受勢力）。由於候選詞缺乏正確對應，加上人名在語境中往往缺少語意輔助，因此後端矯正難以將錯誤修復為正確的人名。

辨識結果	標準答案	原因
教師	教授	語境過短單詞修正較困難
厥等夢想使得食晝	行到半爛燦厥等夢想使得食晝	字詞消失無法補回遺漏詞
十四歲个沙百裡	十四歲个沙伯利	特殊人名錯誤缺乏候選詞對應

表 6 : 無法矯正範例

5 結論

總結來說，本研究針對客語（大埔與詔安腔為主）語音辨識的準確率提升，提出了一套結合 N-best 潛在錯誤偵測、Confusion Set、Collocation 與 Lexicon 的多階段錯誤矯正方法，實驗顯示經過多階段錯誤矯正後最終 CER 減少至 15.49%，減少 2.14 % ，證明該方法能顯著提升語音辨識的準確率與可用性，並為低資源語言的 ASR 矯正研究提供了一個具體且可行的解決方案。

未來研究方向主要著重於語言資源的擴充與優化。目前系統所依賴的辭典與搭配詞庫雖已涵蓋一般日常語境，但在專業領域（如醫療、教育、公共服務等）仍存在不足，導致在處理專業詞彙或專門語境時，系統的穩定性與準確率可能受到限制。若能進一步蒐集並整合專業語料，持續擴充辭典與搭配詞庫，將能有效提升矯正模組對於專業詞彙的辨識與修正能力，進而提升模型可靠性與泛化能力。

References

Povey, D., Ghoshal, A., Boulianne, G., et al. (2011). The Kaldi speech recognition toolkit. In Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2011).

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In Advances in Neural Information Processing Systems (NeurIPS 2020).

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In Proceedings of the 40th International Conference on Machine Learning (ICML 2023).

Serai, P., Wang, P., & Fosler-Lussier, E. (2019). Improving speech recognition error prediction for

modern and off-the-shelf speech recognizers. In Proceedings of IEEE ICASSP 2019.

Guo, J., Wang, M., Qiao, X., Wei, D., Shang, H., Li, Z., Yu, Z., Li, Y., Su, C., Zhang, M., Tao, S., & Yang, H. (2023). UCorrect: An unsupervised framework for automatic speech recognition error correction. In Proceedings of IEEE ICASSP 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (NeurIPS 2017).

Yeh, C.-F., & Lee, L.-S. (2015). An improved framework for recognizing highly imbalanced bilingual code-switched lectures with cross-language acoustic modeling and frame-level language identification. IEEE/ACM Transactions on Audio, Speech, and Language Processing.

Wei, V. J., Wang, W., Jiang, D., Song, Y., & Wang, L. (2024). ASR-EC benchmark: Evaluating large language models on Chinese ASR error correction. arXiv preprint arXiv:2412.03075.