前瞻科技與管理　14 卷 2 期，111-134 頁（2026 年 5 月）
Journal of Advanced Technology and Management Vol. 14, No. 2, pp. 111-134 (May, 2026)
DOI:10.6193/JATM.202605_14(2).0005

面向客語學習之文字轉語音與自動評分：低資源模組化整合

黃奕欽 [1,*]　洪翌翔 [2]

[1] 國立屏東大學電腦科學與人工智慧學系副教授
[2] 國立屏東大學電腦科學與人工智慧學系碩士生

摘 要

本研究針對低資源客語學習，提出以客語拼音為共同中介表示方式的模組化整合系統，涵蓋中文、客文、拼音，與示範語音等轉換，並提供學習者口說之自動發音評量，以降低跨模組錯誤傳遞並提升可診斷性。實作方法上，中文轉客文採卷積式序列到序列模型，結合強制斷詞與未知詞回填以提升客家專有詞彙的生成與翻譯輸出可用性；發音評量以發音良好度（Goodness of Pronunciation, GOP）衍生之對數音素事後機率（Log-Phone Posterior, LPP）／對數事後機率比（Log-Posterior Ratio, LPR）特徵搭配輕量 Transformer 迴歸模型，並以偏差校正處理分數分布偏態。實驗結果顯示，翻譯於內外部測試集中更為穩健，且自動評分與教師評分具良好一致性，具支援電腦輔助語言學習之應用潛力。

關鍵詞：GOP、文字轉語音、自動發音評分、低資源語音與語言處理、客語學習

[*] 通訊作者：黃奕欽
電子郵件：ychuangnptu@mail.nptu.edu.tw
（收件日期：2026 年 1 月 26 日；修正日期：2026 年 2 月 17 日；接受日期：2026 年 2 月 23 日）

Journal of Advanced Technology and Management Vol. 14, No. 2, pp. 111-134 (May, 2026)
DOI:10.6193/JATM.202605_14(2).0005

Text-to-Speech and Automatic Pronunciation Assessment for Hakka Learning: A Low-Resource Modular Integrated System

Yi-Chin Huang[1,*], Yi-Hsian Hong[2]

[1]Associate Professor, Department of Computer Science and Artificial Intelligence, National Pingtung University
[2]Master Student, Department of Computer Science and Artificial Intelligence, National Pingtung University

Abstract

This study addresses low-resource Taiwanese Hakka learning and proposes a modular integrated system that uses Hakka Pinyin as a shared intermediate representation. The system covers a pipeline from Chinese to Hakka text and then to Pinyin and demonstration speech, while providing automatic pronunciation assessment for learners' spoken utterances, thereby reducing cross-module error propagation and improving diagnosability. Specifically, Chinese-to-Hakka translation is built on a convolutional sequence-to-sequence model with forced segmentation and unknown-word backfilling to improve Hakka idiom generation and translation results. For pronunciation assessment, we adopt GOP (Goodness of Pronunciation)-derived LPP (Log-Phone Posterior)/LPR (Log-Posterior Ratio) features with a lightweight Transformer-based regression model, and apply bias calibration to handle skewed score distributions. Experimental results show that the translation component is more robust on both in-domain and out-of-domain test sets, and that the automatic scores exhibit strong agreement with teacher ratings, indicating the potential to support computer-assisted language learning.

Keywords: GOP, Text-to-Speech (TTS), Automatic Pronunciation Assessment (APA), low-resource language, Hakka language learning

* Corresponding Author: Yi-Chin Huang
E-mail: ychuangnptu@mail.nptu.edu.tw

壹、緒論

近年來語音科技在智慧助理、人機互動與教育應用上快速成熟，帶動文字轉語音（Text-to-Speech, TTS）、自動語音辨識（Automatic Speech Recognition, ASR）與電腦輔助語言學習（Computer-Assisted Language Learning, CALL）等技術在各場域落地。然而，對於客語等低資源語言（Low-Resource Language）而言，語料規模、標註成本與腔調差異等因素，使得主流以大量資料為前提的端到端模型難以直接移植並達到可用水準。更具挑戰的是，客語在實務使用中同時存在漢字書寫（客文）與羅馬拼音（如教學拼音）等不同表徵形式，導致「文字—語音」之間的對應關係更為複雜，資料呈現碎片化：一方面，現有中文文本資源豐富但缺乏穩定可用的客語對應；另一方面，客語語音資料即便存在，也常伴隨轉寫（Transcription）標準不一與人工標註不足等問題，進一步限制了整體系統的可訓練性與可評估性。

在教育情境中，客語學習者除了需要自然、可理解的示範語音以支持朗讀、跟讀與情境化練習外，更需要可量化且可解釋的自動發音評分（Automated Pronunciation Assessment, APA）機制，以支援自學與課堂口說評量。若系統僅提供語音合成輸出，學習者難以定位自身偏差；若僅提供分數而缺乏可對照的標準示範語音，回饋亦難以形成可持續的學習循環。此外，教學場域常要求示範語音能貼近特定腔調或特定說話者風格，以提升學習者接受度與課室情境一致性。基於此，本研究關注的核心問題是：在少量客語語料條件下，如何同時滿足：一、可讓使用者以中文輸入、客語輸出自然且可客製化，與二、評分可對齊人評標準且具可解釋回饋的整合式 CALL 系統。

為回應上述挑戰，本研究提出一套以「客語拼音」作為跨模組共同中介表示（Intermediate Representation, IR）的模組化回饋循環機制：系統將「中文文字到客語語音」之生成流程分解為中文轉客文（機器翻譯〔Machine Translation, MT〕）、客文轉客語拼音（字轉音〔Grapheme-to-Phoneme, G2P〕），以及客語拼音轉語音（TTS）三個模組，並進一步以同一目標拼音序列作為 APA 的共同對齊單位。此一設計與既有 CALL 系統常見的將示範合成模組與 APA 模組各自分開建置，使得對齊單位可能不一致的作法不同：本研究以拼音中介統一內容的表示方式，使生成端與評估端共享同一種發音單元，因而能降低跨模組之錯誤傳遞、提升錯誤可診斷性（能回溯到拼音／音素層的偏差位置），並提升系統的可維護性和可擴充性（可替換單一模組而不破壞整體介面）。此外，在語音生成端，本研究於少量目標語者資料下引入語者條件（如語者／風格嵌入向量），以支援示範語音之語者／風格客製化需求；在發音評估端，本研究透過語音對齊與發音良好度（Goodness of Pronunciation, GOP）家族特徵抽取，建置可對應人類評分量尺之自動化評分模型，使系統不僅能產生示範語音，也能對學習者口說提供一致且可量化之回饋，形成從示範到模仿，再到回饋三步驟之學習回饋循環。

綜合而言，本研究的貢獻可從三個面向來集中說明：

一、架構層面：提出以客語拼音為跨模組共同中介之模組化回饋循環機制，將中文輸入、示範語音生成與 APA 以相同的對齊單位整合，提升低資源條件下的可控性、可診斷性與可部署性；

二、生成層面：在低資源客語情境下建置可落地之中文到客語語音流程，並支援腔調／語者客製化；

三、評估層面：建立可與真實的專家評估一致的客語 APA 模型，並以客觀指標與主觀聽測驗證其有效性。

透過上述設計，本研究嘗試在低資源語言的技術限制與教育應用的實務需求之間取得平衡，提供一個可被重複運用與延伸的系統化框架。

本文其餘章節安排如下：第貳章回顧低資源語言之語音與語言處理相關研究；第參章詳述系統流程與各模組架構與模型訓練資料；第肆章呈現主要實驗設置與結果分析；第伍章總結研究發現並提出後續方向。

貳、文獻探討

本章回顧低資源語言之語音與語言處理相關研究，並整理文字轉寫、少量語料語音合成（TTS）與 APA 等技術脈絡，最後說明本研究之方法定位。

一、低資源語言與臺灣本土語言語料建置

低資源語言常面臨可用語料量體不足、書寫系統不一、腔調與說話人差異大等問題，使得端到端語音／語言模型在訓練穩定性、泛化與可落地性上受到限制。近年臺灣本土語言在政府推動與學界投入下，逐步建立較大規模的語音與文本資源，為客語等語言之語音與語言處理研究提供關鍵基礎。

以客語為例，Liao et al.（2023）、Liao, Kuo, Huang, Lan, Lai, and Hsu（2025）建立 Taiwanese Hakka Across Taiwan（HAT）語音語料庫，並透過 Formosa Speech Recognition Challenge（FSR）系列競賽促進客語 ASR 研究與基準化評估；相關工作包含 FSR-2023 Hakka ASR 競賽資料與說明，以及後續延伸之 FSR-2025 Hakka ASR II（新增腔調資料）等。這類公開語料與評測活動的重要性在於：一方面提供跨團隊可重現之訓練／測試切分與評估指標，另一方面也讓低資源語言可借鏡主流語音技術（如自監督預訓練）在小量標註資料條件下的效益。

然而，語料建置仍需處理語者、錄音條件與腔調差異帶來的資料偏移（Domain Shift），並兼顧教學情境所需的「可對應文本—拼音—語音」鏈結與標記一致性。此類一致性要求也直接影響下游的文字轉寫、語音合成與 APA。

二、文字轉寫與中文轉換至資源稀缺之語言

文字層轉換在本研究系統中扮演「把中文語意映射到客語書寫形式」的關鍵入口。若採端到端中文 TTS 直接生成客語語音，錯誤難以診斷且不易控制；因此不少低資源情境會採模組化策略，先完成文字層翻譯／轉換，再以規範化的 IR（例如拼音或音素）串接語音端模型。

在神經機器翻譯（Neural Machine Translation, NMT）領域，Transformer 架構（Vaswani et al., 2017）因其自注意力機制而成為主流；在低資源條件下，子詞切分（如位元組對編碼

〔Byte Pair Encoding, BPE〕）（Sennrich, Haddow, and Birch, 2016）可有效降低辭典未包含詞彙（Out-of-Vocabulary, OOV）問題並提升詞彙覆蓋。對於資料量較小但句長不長的任務，全卷積式序列到序列（Fully Convolutional Sequence-to-Sequence, ConvS2S）（Gehring, Auli, Grangier, Yarats, and Dauphin, 2017）亦提供可並行化訓練與較穩定的優化特性，並常透過 Fairseq（Ott et al., 2019）等工具實作與重現。

而在本研究範疇中，從低資源中文轉換少數資源稀缺語言的文字翻譯，其挑戰常集中在：（一）專有名詞與固定搭配難以對齊；（二）斷詞邊界不穩定導致關鍵詞拆分；（三）平行語料不均衡造成模型偏向高頻樣本。本研究所提出之對策包含：一致化斷詞與子詞切分、以詞表／術語表約束解碼、資料擴增（回譯、同義改寫）與後處理回填等，以提升關鍵詞的可生成性與系統可用性。

三、客文書寫系統到客語拼音（G2P）

G2P 是連結文字處理與語音處理的重要橋樑。相較於英語等拼寫較規範的語言，客語存在腔調差異、外來詞與多種書寫習慣（漢字借寫、方言字、異體字）等現象，使得「同字異音」與「同音異字」更為常見，增加 G2P 的不確定性。

低資源 G2P 常見兩類策略：其一為字典／規則導向方法（Bisani and Ney, 2008），優點是可解釋、可控且在字典覆蓋率高的情況下表現穩定；其二為資料所訓練的序列模型（遞迴神經網路〔Recurrent Neural Network, RNN〕、Transformer、卷積等）（Yao and Zweig, 2015），可利用上下文做語境判別並補足字典未涵蓋的詞彙。實務上常採混合式設計：先以神經模型預測拼音，再利用拼音合法集合檢核、置信度門檻與腔調別字典回退（Fallback）進行後處理，以降低不可用輸出並提升系統穩定性（洪翌翔，2023）。

此外，G2P 的輸出若同時作為 TTS 與 APA 的共同介面，則需要特別重視符號集合一致性（例如教育部〔2012〕客語拼音方案）、連讀／變調標記策略，以及與強制對齊（Forced Alignment）所用音素集合的對應關係；否則容易造成下游訓練與推論不一致。

四、少量語料語音合成（低資源 TTS）

神經式 TTS 近年以聲學模型（輸入為文字／音素，而輸出聲音的梅爾頻譜）整合聲碼器（從頻譜轉換為聲音的波形）的兩階段架構為主流。早期代表如 Tacotron 2（Shen et al., 2018）以自迴歸解碼生成梅爾頻譜，品質佳但推論速度慢且可能出現漏字／重複；後續非自迴歸架構（FastSpeech/FastSpeech 2）（Ren et al., 2019, 2020）透過長度調節（Length Regulator）與顯式預測時長，顯著提升推論速度並改善穩定性。

在聲碼器方面，生成對抗網路（Generative Adversarial Network, GAN）類方法（如 HiFi-GAN〔Kong, Kim, and Bae, 2020〕）能以較低計算成本產生高擬真語音，常被用於低資源情境中快速建立可用的語音輸出。若要支援多語者或客製化音色，常引入語者嵌入向量（Speaker Embedding）（Jia et al., 2018）或參考編碼器（Reference Encoder）（Wang et al., 2018）以解耦說話的內容與語者特徵。

低資源 TTS 的核心困難在於標註／乾淨錄音不足與說話人／腔調變異。常見作法包含：遷移學習（以大型語言或相近語言預訓練再微調〔Chen, Tu, Yeh, and Lee, 2019〕）、多語者／多語言聯合訓練、資料增強（如速度／音高擾動〔Ko, Peddinti, Povey, and Khudanpur, 2015〕）與嚴謹前處理。當需要將拼音與語音做時間對齊以供 APA 或分析使用時，亦可借助 Kaldi（Povey et al., 2011）或 Montreal Forced Aligner（MFA）（McAuliffe, Socolof, Mihuc, Wagner, and Sonderegger, 2017）等工具產生可訓練的強制對齊，降低人工作業成本。

五、APA 與專家評分之一致性

CALL 中的 APA 目標是：在不依賴人工評審的條件下，對學習者的口說品質給出可量化且可解釋的回饋。傳統方法多基於 ASR 模型計算 GOP，其核心概念是比較「目標音素」相對於其他競爭音素的支持度，並以對數似然比或後驗機率比形成片段級分數；Witt and Young（2000）的工作奠定了 Phone-Level GOP 用於互動式語言學習的基礎，而後續研究亦針對門檻設定、錯誤類型與穩健性進行更完整的分析（Kanters, Cucchiarini, and Strik, 2009）。

近年深度學習與自監督語音表徵（如 Wav2vec 2.0〔Baevski, Zhou, Mohamed, and Auli, 2020〕、HuBERT〔Hsu, Bolte, Tsai, Lakhotia, Salakhutdinov, and Mohamed, 2021〕）大幅提升低資源 ASR 的可用性，亦促使 APA 從「單一 GOP 分數」走向「多特徵、序列建模與多層級評分」。例如，將片段級 GOP、對數音素事後機率（Log-Phone Posterior, LPP）、對數事後機率比（Log-Posterior Ratio, LPR）等特徵輸入 Transformer 進行序列整合，可同時考慮跨音段的依賴關係並輸出句子層級連續分數；Gong, Chen, Chu, Chang, and Glass（2022）提出基於 GOP 的 Transformer（Goodness of Pronunciation Transformer, GOPT）亦展示 GOP 特徵搭配 Transformer 的效益，並嘗試多面向（Accuracy、Fluency 等）與多粒度（Phone/Word/Utterance）共同建模。此外，國立臺灣師範大學團隊亦提出結合多粒度特徵與神經序列模型之自動口說評分方法，並探討資料不平衡與多模態特徵整合對評分效能之影響（林孟欣、王馨偉、羅天宏、陳柏琳、趙偉成，2023；Peng, Wang, Chen, and Chen, 2023）。

在專家評分之一致性的研究方向中，語音評分常受到評審嚴格度、量尺離散化、題目難度與錄音品質影響，導致標籤具有噪聲與偏移。文獻中常見作法包括：對評分進行校正（例如以開發集估計全域或分段偏移）、採用排序／相關係數作為評估指標，以及在模型訓練時使用適合的損失函數或不確定性建模，以提升自動分數與人評的可對齊性。

綜上所述，低資源語言的可落地系統往往需要在「資料可得性」與「系統可診斷性」間取得平衡。本研究採模組化流程，先以中文 → 客文文字轉換確保語意層對應，再以客語拼音作為跨模組共同介面，串接語音合成與 APA，兼顧可控性與可擴充性。另一方面，APA 則以可解釋的 GOP 家族特徵為基礎，結合序列建模迴歸句子層分數，以對齊教學場域的人評量尺並提供可量化回饋。

參、系統架構與方法

一、系統概述

本研究提出一套低資源客語的模組化閉環系統，將中文輸入依序轉換為客文、客語拼音，並生成示範語音；同時以相同目標拼音作為評量依據，對學習者口說提供可量化回饋。此設計的關鍵在於以「客語拼音」作為跨模組的共同 IR，使生成端與評估端共享同一對齊單位，降低錯誤傳遞並提升可診斷性（流程如圖 1 所示）。

由於篇幅限制，且客文轉拼音（G2P）與語音合成（TTS）之核心方法主要延續既有研究與既有工具鏈（以少量修改以符合本研究之拼音介面與腔調需求），本文僅保留必要之介面定義、關鍵設計決策與最終效能指標；完整模型超參數、訓練細節與前處理流程則移至補充資料。相對地，本文著墨於中文轉客文在低資源下的可控改善，以及以拼音為中介的 APA 設計與校正策略，作為本研究之主要貢獻。

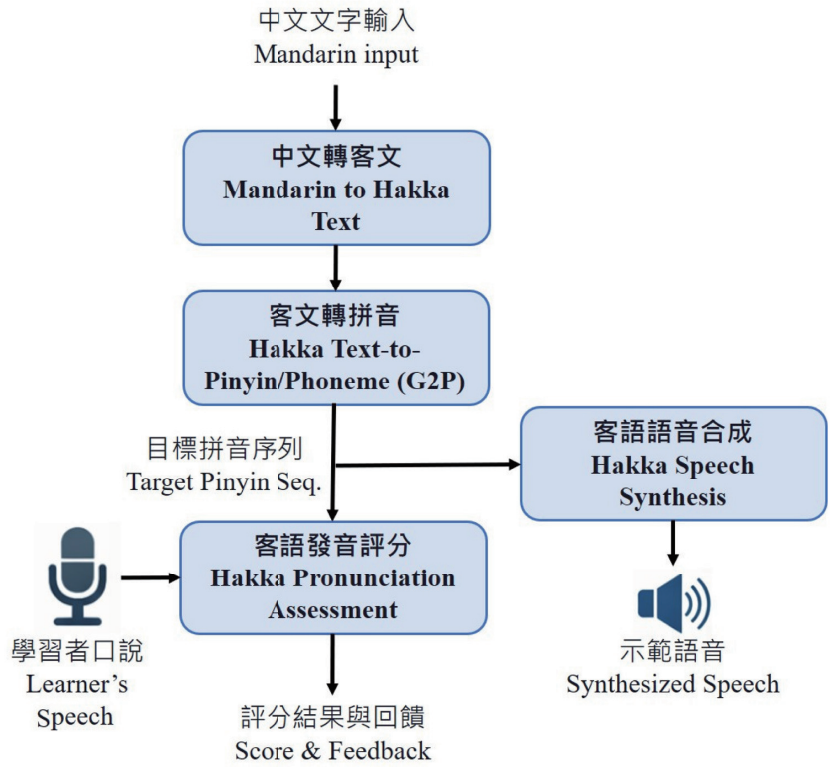

圖 1　本研究系統流程與主要技術模組（拼音作為跨模組共同介面）

資料來源：本研究自行產製。

二、MT 模組：序列轉序列之機器翻譯架構

MT 採用之 ConvS2S 作為基礎架構。模型以卷積編碼器（Encoder）擷取來源序列局部上下文表徵，並由具注意力機制的卷積解碼器自迴歸生成目標客文序列。相較於循環式模型，卷積架構具較佳的平行化效率；本文將其作為固定骨幹，後續實驗聚焦於斷詞一致性與未知詞處理等可控因子對翻譯品質與系統可用性的影響。

三、MT 模組：強制斷詞（Forced Segmentation）與翻譯改善機制

低資源翻譯中，詞邊界不穩定與專用詞對齊不足常導致關鍵詞彙無法生成，並進一步放大到下游轉寫與語音模組。本研究提出「強制斷詞」機制：針對可直接對應客語詞單位之中文片段，於斷詞階段固定其邊界，並在客文端同步施作對應的固定切分，以維持平行語料在詞元（Token）粒度上的對齊一致性。值得注意的是，本研究並非以查字典式之人工規則翻譯取代神經翻譯，而是將強制斷詞嵌入至神經模型的訓練資料生成流程中，使模型仍以平行語料學習來源句與目標句的條件分布。換言之，強制斷詞改變的是模型所觀測到的詞元序列表示，而非改變 Seq2seq（Sequence to Sequence）的學習目標本身。實作上，若以 $\tilde{x}^{zh}$ 表示原始中文句、x^{zh} 表示斷詞後送入模型之序列，則原先的一般斷詞可寫為 $x^{zh} = g_{seg}(\tilde{x}^{zh})$；加入強制斷詞後，斷詞函數可視為受詞彙約束的映射：

$$x^{zh} = g^{seg}_{force}(\tilde{x}^{zh}; V_{force})$$

其中，V_{force} 為欲保留為固定詞邊界的片段集合。本研究實作上，進一步考量神經翻譯的學習，來自來源與目標之間的序列對齊，而非單向詞典查詢。因此，除了在中文端加入強制斷詞外，也同步對客文端斷詞系統加入相應的強制斷詞處理，以維持平行語料在詞元粒度上的對齊一致性，期望同時達成「整體翻譯品質」與「專用詞彙可產生性」的兼顧。

此外，為避免未知詞（<Unk>）進入下游造成不可轉寫，本研究於解碼端採兩層處理：先在集束搜尋（Beam Search）中對未知詞施加懲罰以降低產生機率；若仍出現未知詞，則利用注意力對齊訊號回溯最可能對應的來源詞元進行回填，以將不可用符號轉為系統可接受的可讀輸出，降低跨模組錯誤擴散。

四、客文轉拼音模組

客文轉拼音模組在本系統中作為跨子系統的共同介面：上游 MT 輸出的客文句子，需先轉為客語拼音序列，才能一致地串接語音合成與 APA。本研究採監督式模型學習客文字／詞到拼音的映射，並以教育部（2012）客語拼音方案定義之合法符號集合進行輸出約束；其餘模型結構與超參數僅作必要交代，重點結果統一於實驗章節中呈現。

實作上，客文轉拼音的模組，由於任務較為單純，故採用雙向長短期記憶網路（Bi-

directional Long Short-Term Memory, Bi-LSTM）建模，並以拼音字典輔助修正不可靠的拼音輸出。在訓練目標上，本研究將拼音轉寫視為監督式序列標註／預測問題，最小化詞元層級交叉熵以學得模型參數 θ。

五、G2P 拼音模組：拼音錯誤偵測與字典輔助修正

由於低資源資料下模型易在未見詞與專有名詞上產生不穩定輸出，本研究加入字典輔助修正作為後處理：當模型輸出不符合拼音合法性或落在低可信區間時，回退至腔調別拼音字典之查詢結果，以降低不可用輸出並提升整體穩定性。

六、客語語音合成模組：客製化語者之客語語音

本研究以非自迴歸式 TTS 架構產生示範語音，並以拼音作為內容條件以降低書寫差異造成的不可控性。為支援不同示範音色，本模組引入語者嵌入向量以達成多語者合成；聲學模型採 FastSpeech2 類框架，聲碼器採 HiFi-GAN 還原波形。其餘訓練細節與超參數配置統一於第肆節實驗設置中交代。

七、客語 APA 模組：自動化口說評量流程

在完成客語語音合成模組後，系統已具備由目標拼音序列產生對應客語語音之能力，可支援教材朗讀、示範音檔生成與多語者語音輸出等應用。然而，在客語教學與學習情境中，僅提供示範語音尚不足以支撐有效的發音學習歷程；學習者更需要結合量化評分與錯誤指引的即時回饋機制，以形成完整的學習閉環。因此，本研究於語音合成模組之後，進一步設計「客語 APA 模組」，以相同的目標拼音作為評量依據，對學習者口說語音進行分析，並輸出可對應教學尺度的總分（以及必要時之分項回饋）。透過整合語音示範與 APA 模組，系統得以實現一套涵蓋輸入、示範、模仿與回饋的閉環式學習機制，從而同時滿足語言教學中對示範、練習與評量的需求。

在此模組中，輸入為學習者語音與對應之目標音素／拼音序列，輸出為句子層級之連續分數。此評分模組採用 GOP 家族方法作為可解釋的聲學證據來源，並以輕量 Transformer 迴歸器整合時序特徵，兼顧低資源語言的可訓練性與推論效率。

GOP 的關鍵前提是：必須先知道「目標音素」在學習者語音中的時間區段，才能在該區段內計算目標音素與競爭音素的相對支持度。為此，本研究先以客語語音訓練一個 ASR 聲學模型，並僅用於產生強制對齊與音框層級（Frame-Level）的音素之後驗機率；該聲學模型在推論時不參與分數預測本體，因此額外開銷主要集中在「特徵萃取」階段，而評分器本身仍維持單次前向傳遞即可完成評分。在實作上，我們採用時間延遲神經網路（Time Delay Neural Network, TDNN）聲學模型來產生對齊與後驗分布，作為供 GOP 特徵計算的前處理。

在 GOP 架構下，對於被對齊到目標音素 p 的觀測片段 O_p，傳統 GOP 可視為與目標音素相對於最具競爭力音素的長度正規化對數似然比；而在以深度神經網路（Deep Neural Network,

DNN）後驗為主的設定中，則可等價地以目標音素的 LPP 與競爭音素的 LPP 最大值之差來表達，如式（1）所示：

$$GOP(p \mid o) = LPP(p \mid o) - \max_{q \in Q} LPP(q \mid o) \quad (1)$$

其中，LPP 定義為在對齊區段內所有音框的平均對數後驗；若以 o_t 表示第 t 個聲學音框，則 LPP 可近似為式（2）所示：

$$LPP(p \mid o) \approx \frac{1}{T} \sum_{t} \log P(p \mid o_t),\ P(p \mid o_t) = \sum_{s \in p} P(s \mid o_t)) \quad (2)$$

此外，為了讓模型不只知道目標音素有多好，也能掌握目標音素相對其他音素的競爭關係，本研究同時採用 LPR 作為特徵。LPR 定義為任一競爭音素 p_j 與目標音素 p_i 的 LPP 差，如式（3）所示：

$$LPR(p_j \mid p_i, o) = LPP(p_j \mid o) - LPP(p_i \mid o) \quad (3)$$

直觀上，若目標音素發音正確，則 $LPP(p_i \mid o)$ 會顯著高於競爭音素，對應的 LPR 多為負值；若發音有誤時，LPR 可能趨近 0 甚至為正。基於此，本研究以對齊結果為索引，對每個音素片段萃取 LPP 與 LPR，並沿通道（Channel）維度串接成 $T \times C$ 的時序表示，再以線性投影壓縮到較小維度（投影維度設定 $d = 24$），以降低後續序列模型的參數需求並提升訓練穩定性。

在得到融合的 LPP/LPR 時序嵌入向量後，本研究以 Transformer 編碼器作為句子層級評分器，將「逐音素／逐片段的發音證據」整合為一個連續分數。相較於以池化為主的淺層迴歸（如線性或多層感知機〔Multi-Layer Perceptron, MLP〕），Transformer 的自注意力能在不顯著增加推論成本的情況下建模跨片段的依賴關係，使模型同時看見局部錯誤與整體一致性，較符合教師在句子層級評分時會參照「多個音素共同決定整體可懂度」的評分行為。

在架構細節上，本研究所設計的評分器採用純編碼器（Encoder-Only）Transformer 架構，首先將投影後的序列嵌入送入多層自注意力與前饋網路堆疊，以整合跨音段之辯識訊息；其後透過序列聚合策略，如採平均池化以取得句子層級表示，並接續一線性層輸出連續評分值。為在參數量受限的前提下維持對關鍵聲學片段的建模能力，模型引入可學習的位置嵌入與前置正規化（Pre-Normalization）（於殘差連接前施行層正規化〔Layer Normalization〕）等設定，以提升注意力整合之穩定性與有效性。整體設計以「小參數量、但能有效利用自注意力進行全域整合」為核心原則，使評分模組在作為整體系統之一環時，仍能兼顧訓練與推論的計算可負擔性。

由於教學場域常使用固定量尺（例如 0 ~ 4 分且以 0.5 為間距）呈現評分結果，本研究

以連續迴歸輸出為主，再將其四捨五入映射到離散集合。此外，為減少標計分布不均造成的系統性偏差，我們亦採用以開發集估計的偏差校正（Bias Calibration），包含全域偏移或依預測落點選取的分段偏移，以在不增加模型推論成本下改善分數一致性，其效益與比較將統一保留於第伍節結果分析中呈現。

八、語音資料集分析

本研究採模組化分段式架構，各子系統雖使用不同資料來源，但需共享一致的 IR（客語拼音）以維持可追溯的對應鏈。因此資料建置重點不僅在於規模，更在於：文本層的斷詞一致性、拼音層的合法集合約束，以及語音層的可對齊標記品質。各模組資料的總句數／總時數／語者數彙整如表 1 所示。

表 1　各模組使用之資料集規模統計

模組名稱	資料類別	語者／評審人數	總句數	總時長（小時）
中文轉客文（MT）	平行文本	—	北四縣 30,790； 南四縣（擴增後 32,727）	—
客文轉拼音（G2P）	文字—拼音對	—	北四縣 9,207； 南四縣 22,408	—
語音合成（TTS）	語音—拼音對	客語 3 人 （中文 4 人）	客語 17,703 （中文 14,534）	客語 24.43 （中文 19.71）
發音評分（APA）	語音—人評分數	2	19,668	—

註：MT：機器翻譯（Machine Translation）；G2P：字轉音（Grapheme-to-Phoneme）；TTS：文字轉語音（Text-to-Speech）；APA：自動發音評分（Automated Pronunciation Assessment）。

資料來源：本研究自行產製。

本研究之資料來源以「可落地」為原則：中文—客文平行文本主要來自哈客網、教育部本土語言資源與客語語料庫等公開來源，並依腔調分流建置北四縣與南四縣語料；客文—拼音配對以具拼音標註之教材與競賽朗讀文本為主，並以拼音合法集合進行清理；語音合成與 APA 資料皆以拼音作為統一標記，以確保後續對齊、合成與評分之可追溯性。各模組完整來源、清理規則與前處理流程因篇幅限制省略，必要時可於補充資料提供。此外，本章後續實驗之資料規模均以表 1 為準；為避免重複，第肆節不再逐段重述資料統計。

肆、實驗設置與結果分析

本章依系統模組化流程，分別評估四項子任務：中文至客文翻譯、客文至拼音轉寫、拼音至語音合成，以及客語 APA。為確保各模組最佳設定能以實驗結果客觀決定，本研究針對每一模組設計多組候選配置（包括斷詞懲罰值、語料擴增策略、特徵組合與後處理方法），並採用通用且可重現的評估指標進行量化比較，最後依測試集表現選定最終配置。

一、MT 模組之評估

（一）實驗設置

本研究之 MT 翻譯模組以 Fairseq 之全卷積式 Seq2seq（ConvS2S）作為固定骨幹，並在模型容量與訓練策略一致下，聚焦比較低資源客語情境的可控因子：腔調分流建模、目標端斷詞策略、解碼端未知詞懲罰，以及南四縣語料擴增對跨域泛化之影響。

模型設定方面，嵌入向量（Embedding）與位置嵌入向量（Position Embedding）維度皆為 256；編碼器採 4 層卷積區塊（卷積核尺寸〔Kernel Size〕= 3），解碼器（Decoder）採 3 層卷積區塊（卷積核尺寸 = 3）；訓練採標籤平滑交叉熵（Label-Smoothed Cross Entropy）（$\varepsilon = 0.1$）與 Adam 最佳化演算法，並以反平方根學習率排程（Inverse Square Root Learning Rate Schedule）進行最佳化。

資料規模彙整如表 1。另因南四縣原始平行語料較小，本文建置擴增版本（擴增後 32,727 句）以量化其對外部測試域之泛化增益。

（二）測試語料

評估資料分為兩域：內部測試集（與訓練資料同來源、未納入訓練）與外部童話故事測試集（不同翻譯者／語域）。外部測試集取自客家委員會童話故事集（客家委員會，n.d.-a），北四縣／南四縣皆以相同中文輸入、不同腔調客文參考譯文進行評估；兩域之句數與字數統計沿用原始資料分割設定。

（三）評估指標與解碼參數搜尋

翻譯品質以候選譯文與參考譯文之重疊與編輯距離衡量：回報萊文斯坦編輯距離（Levenshtein Distance, LD）、雙語評估替補指標（Bilingual Evaluation Understudy-4, BLEU-4）與以召回率為導向的摘要評估指標 （Recall-Oriented Understudy for Gisting Evaluation, ROUGE）-1/2/L。考量外部測試僅有單一參考譯文且翻譯風格差異較大，本文同時回報內部與外部兩域結果，以避免單一語域偏誤。

在解碼端，本研究針對各模型進行未知詞懲罰（Unknown Word Penalty）之網格搜尋。具體而言，對每一個訓練完成之模型，於集束搜尋解碼時掃描一組懲罰值（本文表格以 “Unk Penalty” 呈現，候選範圍為 -10～-5），並以 BLEU/ROUGE 的最佳點決定該模型之預設解碼設定。此設計可避免不同斷詞策略所造成的 OOV 分布差異，使得固定懲罰值產生較差的翻譯結果。實作上，在集束搜尋的步驟中，先得到下一步各詞彙的對數機率（Log-Probabilities）向量計算公式為 $\ell_t(v) = \log P(v \mid y_{<t}, x)$，在每一時間步 t 對未知詞的對數機率額外扣除一個常數懲罰 λ_{unk}（對應 Unk Penalty），得到調整後的打分。其中，x 為給定的來源序列，而 $y_{<t}$ 為目前已生成的目標前綴，v 為下一個要生成的詞元（$v \in V$，V 為目標詞彙表）。

翻譯任務中，斷詞（Tokenization/Segmentation）被視為低資源條件下影響模型訓練與對齊穩定性的關鍵前處理，因此實驗採取「中文端固定斷詞器、客文端更換斷詞策略」的設計，以避免同時變動來源端與目標端造成不可歸因的差異。實作上，每一組平行語料在進入模型訓練之前，均先對中文句執行固定的中文斷詞；而客文句則依實驗條件分別以三種不同方法

產生詞元序列。後續模型架構與訓練流程保持一致，並在解碼端再搭配前述的未知詞懲罰值掃描，觀察不同斷詞策略下的最佳設定與其翻譯品質表現。

中文端斷詞統一採用中央研究院 CkipTagger（Li and Ma, 2019），以固定來源端詞元化（Tokenization）。目標端斷詞之基準系統採國立政治大學團隊於 GitHub 公開之基於 Jieba 客文斷詞器（ldkrsi, 2018），其在 Jieba 框架下以客語資源替換詞庫與統計模型，作為比較其他斷詞策略之參照組。

主要比較的方法為本研究所提出的強制斷詞方法，首先於內部測試集（與訓練資料同翻譯者風格之測試域）比較兩種斷詞策略：以基於 Jieba 客文斷詞作為基準條件，以及在相同斷詞框架上加入強制斷詞之條件。兩者皆採相同模型骨幹與訓練設定，並於解碼端掃描未知詞懲罰值，以各指標表現決定該設定下之最佳懲罰值。

分析結果如表 2 所示，由表中可觀察到，在內部測試集中，Jieba 斷詞的最佳懲罰值大約落在 -7 附近，而強制斷詞的最佳懲罰值則集中於 -8。此外，強制斷詞相較 Jieba 斷詞在 BLEU 與 ROUGE 指標上均呈現明顯提升。此結果顯示，在低資源訓練資料下，強制斷詞透過固定特定片語的詞邊界，能使來源端（中文）與目標端（客文）的詞元粒度更一致，從而提供模型更集中、可重現的對齊訊號，進而影響在生成關鍵片語時較不易因詞界歧義而產生重複或錯誤斷詞。另一方面，Jieba 斷詞在不同指標下的最佳懲罰值分布較不整齊，反映其斷詞結果可能更依賴個別句子的切分型態，導致模型所學得的對齊關係較分散，進而使解碼端調整未知詞懲罰值時，品質增益不具一致性；此現象可視為目標端詞元化不穩定時常見的現象。

表 2　北四縣未知詞懲罰比較表：內部測試集（Jieba 斷詞 vs. 強制斷詞）

Unk Penalty	LD ↓	BLEU-4 ↑	ROUGE-1 ↑	ROUGE-2 ↑	ROUGE-L ↑
-10	4,149	0.4393	0.7794	0.6164	0.7731
-9	3,632	0.4725	0.7944	0.6373	0.7882
-8	3,378	0.4821	0.7996	0.6416	0.7929
-7	3,255	0.4836	0.7926	0.6351	0.7857
-6	3,214	0.4846	0.7883	0.6309	0.7813
-5	3,288	0.4788	0.7827	0.6244	0.7758
Unk Penalty	LD ↓	BLEU-4 ↑	ROUGE-1 ↑	ROUGE-2 ↑	ROUGE-L ↑
-10	3,325	0.4902	0.7886	0.6315	0.7825
-9	3,109	0.5108	0.7995	0.6478	0.7931
-8	3,101	0.5118	0.8018	0.6521	0.7951
-7	3,129	0.5074	0.7979	0.6477	0.7912
-6	3,225	0.5005	0.7909	0.6408	0.7838
-5	3,296	0.4928	0.7850	0.6341	0.7784

註：LD：萊文斯坦編輯距離（Levenshtein Distance）；BLEU：雙語評估替補指標（Bilingual Evaluation Understudy）；ROUGE：以召回率為導向的摘要評估指標（Recall-Oriented Understudy for Gisting Evaluation）。

資料來源：本研究自行產製。

在完成內部測試域的比較後，本研究進一步於外部童話故事測試集（不同翻譯者、不同語域風格之測試語料（客家委員會，n.d.-a）比較同樣兩種斷詞條件，以檢驗其跨域泛化能力。結果如表 3 所示，無論採用 Jieba 或強制斷詞，外部測試集的分數皆明顯低於內部測試集；此差距並不表示模型在外部語料上無法運作，而主要顯示出機器翻譯的客觀評估在僅有單一參考譯文（Single Reference）下的限制：正常情況下，翻譯評估較理想的作法應提供多組可接受參考答案，以涵蓋同義改寫與不同翻譯風格；然而本研究外部童話故事僅提供單一參考譯文，使得即使在候選譯文語意合理，N 元語法（N-Gram）重疊仍可能偏低，因此對外測試分數相對下降屬合理現象。值得注意的是，即使在更嚴格的外部測試域下，強制斷詞相較 Jieba 仍維持較佳的指標表現，且其最佳懲罰值穩定落在 -8；此結果支持強制斷詞所帶來的詞邊界一致化，不僅能改善同風格測試域的表現，也能在跨翻譯風格情境下提供一定程度的泛化增益。

表 3　北四縣未知詞懲罰比較表：外部童話故事測試集（Jieba 斷詞 vs. 強制斷詞）

Unk Penalty	LD ↓	BLEU-4 ↑	ROUGE-1 ↑	ROUGE-2 ↑	ROUGE-L ↑
-10	4,917	0.2946	0.6835	0.4794	0.6752
-9	4,307	0.3420	0.7099	0.5121	0.7026
-8	4,092	0.3555	0.7105	0.5171	0.7027
-7	3,933	0.3661	0.7087	0.5199	0.7017
-6	3,899	0.3652	0.7005	0.5123	0.6932
-5	3,963	0.3558	0.6957	0.5051	0.6882
Unk Penalty	LD ↓	BLEU-4 ↑	ROUGE-1 ↑	ROUGE-2 ↑	ROUGE-L ↑
-10	3,910	0.3675	0.6978	0.5096	0.6917
-9	3,716	0.3847	0.7117	0.5260	0.7046
-8	3,665	0.3906	0.7146	0.5300	0.7077
-7	3,735	0.3818	0.7086	0.5243	0.7009
-6	3,787	0.3725	0.7033	0.5171	0.6953
-5	3,980	0.3549	0.6905	0.5027	0.6818

註：LD：萊文斯坦編輯距離（Levenshtein Distance）；BLEU：雙語評估替補指標（Bilingual Evaluation Understudy）；ROUGE：以召回率為導向的摘要評估指標（Recall-Oriented Understudy for Gisting Evaluation）。

資料來源：本研究自行產製。

為了比較若能夠有較為完善的斷詞正確性的狀況下，我們嘗試採用了使用大量人工所建置的臺灣客語語料庫系統（客家委員會，n.d.-b）所提供的斷詞（並包含詞性標注），該語料庫由客家委員會委託國立政治大學規劃建置（葉秋杏、賴惠玲，2023），迄今已系統化收錄涵蓋書面與口語之客語語料，累積規模達逾 600 萬字，並由專家進行文字校訂；其中口語語料另經人工聽打與時間碼標記，使文字內容可與音訊片段對應。理論上可以作為一個具代表性的「高一致性斷詞基準」，用以對照通用中文斷詞與低資源情境下的強制斷詞策略。

將語料庫斷詞納入同樣的未知詞懲罰值掃描框架後（如表 4），可觀察到其整體表現優

於 Jieba，並多數情況略優於僅加入強制斷詞之設定。此結果顯示：當斷詞器具較佳詞邊界品質與詞彙覆蓋時，能提供更穩定的詞元對齊訊號以改善翻譯品質。

此外，本研究亦測試在語料庫斷詞基礎上再加入強制斷詞，結果顯示其表現與「Jieba + 強制斷詞」相近，但不優於單純語料庫斷詞；個別例句亦可觀察到，過度施加詞邊界約束可能使模型偏向逐字生成（如生成「食晚飯」而非「食夜」），反而削弱語料庫斷詞原先透過大量語料所獲得的固定搭配泛化能力。此結果指出：強制斷詞在低資源、詞界不穩定時能提供有效補救，但當斷詞器本身已具良好詞邊界品質與詞彙覆蓋時，額外的硬性約束未必帶來增益，甚至可能對泛化造成干擾，因此在不同資料規模與斷詞品質情境下應採取差異化策略。

針對南四縣腔的分析，由於原始平行語料規模偏小，除了針對解碼端的未知詞懲罰值進行最佳值的掃描外，另外設計「未擴增 vs. 擴增」兩種訓練語料配置，以檢驗擴增對外部測試域的泛化效益。在控制模型骨幹、訓練策略與斷詞框架一致，在此僅實驗最佳的斷詞設定，也就是語料庫斷詞的結果。

表 4　北四縣（語料庫斷詞）未知詞懲罰比較表：內部測試集和外部測試集

Unk Penalty	LD ↓	BLEU-4 ↑	ROUGE-1 ↑	ROUGE-2 ↑	ROUGE-L ↑
-10	3,309	0.5148	0.8068	0.6569	0.8008
-9	2,936	0.5286	0.8128	0.6678	0.8068
-8	2,912	0.5292	0.8103	0.6659	0.8045
-7	2,956	0.5225	0.8067	0.6613	0.8005
-6	3,023	0.5136	0.7993	0.6524	0.7929
-5	3,076	0.5094	0.7940	0.6469	0.7875
Unk Penalty	LD ↓	BLEU-4 ↑	ROUGE-1 ↑	ROUGE-2 ↑	ROUGE-L ↑
-10	3,634	0.3917	0.7167	0.5334	0.7118
-9	3,457	0.4071	0.7281	0.5475	0.7229
-8	3,444	0.4084	0.7287	0.5487	0.7235
-7	3,501	0.4032	0.7219	0.5429	0.7169
-6	3,632	0.3907	0.7109	0.5320	0.7053
-5	3,778	0.3749	0.6983	0.5184	0.6923

註：LD：萊文斯坦編輯距離（Levenshtein Distance）；BLEU：雙語評估替補指標（Bilingual Evaluation Understudy）；ROUGE：以召回率為導向的摘要評估指標（Recall-Oriented Understudy for Gisting Evaluation）。

資料來源：本研究自行產製。

由表 5 可見，在未擴增訓練語料的情況下，南四縣模型於外部童話測試域的較佳表現集中於未知詞懲罰值 -6 ~ -7 附近。其中 BLEU-4 於 -6 取得最高值 0.2991，ROUGE 指標則在 -7 時達到相對高點。此結果顯示，在語料規模偏小且詞彙覆蓋受限的條件下，解碼端對未知詞的抑制強度若過高（例如 Unk Penalty = -10 更極端）容易導致模型以不精確的已知詞替代原本應對應的低頻詞彙，使外部測試域的 N 元語法一致性下降；相對地，Unk Penalty = -7 在未知詞數量控制與可接受譯文形成之間取得較佳平衡。

在加入語料擴增後（表 5 下半部分為擴增語料版本結果），外部童話測試域的整體指標呈現一致提升，且最佳點同樣穩定落於 Unk Penalty = -7。相較未擴增版本，BLEU-4 由 0.2991 提升至 0.3105，ROUGE-1/2/L 亦同步小幅上升，並伴隨 LD 亦呈現改善現象，顯示擴增主要貢獻在於提升低頻詞彙與固定搭配的可觀測性，使模型在外部風格差異下仍能生成更接近參考譯文的詞彙選擇與片語結構。值得注意的是，Unk Penalty 的最佳值在擴增前後皆保持一致（-7），代表擴增改善了模型本身的可生成能力，而非僅是改變未知詞與已知詞的競爭關係。

二、G2P 拼音模組之評估

本研究於北四縣與南四縣分別訓練客文轉拼音模型，採用上下文式序列模型搭配字典輔助修正。其中，神經模型用以處理語境依賴之讀音判別（如常見破音字），字典則作為未見字與確定讀音之回退保底，以降低錯誤擴散至下游語音模組。

結果顯示，採用同一超參數設定（Embedding = 64, BiLSTM = 64, Layer = 1）並加入字典修正後，南四縣整體準確率達 0.9659、北四縣達 0.9712。整體而言，此混合式策略能在低資源情境下兼顧可重現性與準確率。

表 5　南四縣未知詞懲罰比較表：外部童話故事測試集（未擴增訓練語料 vs. 擴增訓練語料）

Unk Penalty	LD ↓	BLEU-4 ↑	ROUGE-1 ↑	ROUGE-2 ↑	ROUGE-L ↑
-10	5,599	0.2209	0.6099	0.3885	0.6004
-9	5,055	0.2667	0.6429	0.4327	0.6328
-8	4,726	0.2929	0.6637	0.4582	0.6537
-7	4,630	0.2988	0.6670	0.4620	0.6563
-6	4,619	0.2991	0.6623	0.4600	0.6514
-5	4,676	0.2900	0.6522	0.4512	0.6411
Unk Penalty	LD ↓	BLEU-4 ↑	ROUGE-1 ↑	ROUGE-2 ↑	ROUGE-L ↑
-10	5,051	0.2509	0.6201	0.4060	0.6105
-9	4,581	0.2921	0.6544	0.4494	0.6451
-8	4,400	0.3074	0.6680	0.4647	0.6583
-7	4,337	0.3105	0.6715	0.4673	0.6614
-6	4,436	0.3039	0.6656	0.4607	0.6547
-5	4,549	0.2931	0.6577	0.4521	0.6470

註：LD：萊文斯坦編輯距離（Levenshtein Distance）；BLEU：雙語評估替補指標（Bilingual Evaluation Understudy）；ROUGE：以召回率為導向的摘要評估指標（Recall-Oriented Understudy for Gisting Evaluation）。

資料來源：本研究自行產製。

三、客語語音合成模組之評估

語音合成採非自迴歸式架構（FastSpeech2 類聲學模型 + HiFi-GAN 聲碼器），並以語者

嵌入向量支援多語者併訓。主觀聽測以平均意見分數（Mean Opinion Score, MOS）衡量自然度與語者相似度；在 20 段客語合成音檔、19 位客語母語者評分下，客語自然度 MOS 為 4.32/5，顯示已足以支援教材朗讀與示範音檔生成。

補充分析顯示：跨語言合成時，客語語者以中文內容合成之自然度／相似度（3.03/2.70）低於中文語者（3.93/4.12），仍有改善空間；四縣腔連讀變調測試中，51 處變調位置之「無問題」判定占比逾九成；合成效率方面，端到端可達約 57.16 倍即時，符合互動式教學應用需求。

四、APA 模組之評估

本研究之 APA 以「句子層級連續分數」為主要輸出。輸入為學習者語音與其對應之目標音素／拼音序列，輸出為連續分數 $\hat{y} \in R$，以貼近教學場域之整體口說評量需求。為支援等第回饋，本文亦將連續分數四捨五入至 0.5 間距集合 G = {0,0.5, ..., 4.0} 成離散分數（Rounded Score），並同時回報連續與離散兩類指標，以兼顧「可回饋性」與「可比較性」。

（一）實驗設置

為計算 GOP 家族特徵所需的強制對齊，本研究先訓練一套僅用於對齊與產生音框層級 Phoneme Posterior 的聲學模型（不直接做評分預測）。對齊語料共 686 位語者（男 256 人、女 430 人，年齡 17 ~ 75 歲），採語者獨立的方式來切分為訓練集與評估集，比例約 10:1。訓練與評估規模如表 6 所示；聲學模型在評估集之音素錯誤率（Phoneme Error Rate, PER）約 5%，可提供穩定對齊品質以支援後續特徵萃取。

表 6　對齊用語料規模（Alignment Corpus）

切分	語者數	總時長（小時）	句數	音節數（Syllables）
訓練集	617	279.59	94,696	1,668,876
評估集	69	27.10	9,225	168,976
合計	686	306.69	103,921	1,837,852

資料來源：本研究自行產製。

如前所述，APA 資料約 20,000 筆詞句（Utterances）。每筆由兩位具資格教師依 0 ~ 4 分量尺（0.5 間距）評分，標籤取兩位教師分數平均形成連續標記。資料切分採非特定語者（Speaker-Independent）的 8:1:1（訓練集〔Train〕／驗證集〔Valid〕／測試集〔Test〕），並使各切分的分數分布相匹配。由於測試集分布偏向高分區間（2.0 ~ 4.0，占 77.4%），本文另設計偏差校正以處理分布偏態造成的系統性偏誤。

在相同輸入特徵（LPP、LPR 及其融合）與相同訓練目標（均方誤差〔Mean Squared Error, MSE〕損失）設定下，本文聚焦比較基於 Transformer 評分器之設計差異，包含：1. Vanilla Transformer（Van）：精簡版編碼器作為基準；2. Proposed Transformer（Pro）：加入可學習位置嵌入（Positional Embedding）、前置正規化殘差設計（層正規化）以及注意力

權重調整（Attention-Map Reweighting），以強化序列聚合能力並提升句級分數與人評之一致性。此外，為支援教學情境的等第回饋，本文亦比較兩種基於 Transformer 離散化策略：1. 四捨五入迴歸（Rounded Regression, ProR）：將迴歸輸出取整至 0.5 間距集合；2. 九類 Transformer 分類器：直接進行九分類（0, 0.5, ..., 4.0）。

（二）實驗結果與討論

如表 7 所示，基於 Transformer 模型於測試集之迴歸結果顯示：以 Vanilla Transformer 而言，融合 LPP + LPR 相較單獨使用 LPP 或 LPR 皆能獲得較佳表現（MSE 由 0.5105/0.4896 降至 0.4699，皮爾森相關係數〔Pearson Correlation Coefficient, PCC〕提升至 0.7876），顯示「目標音素支持度」（LPP）與「目標—競爭差距」（LPR）具有互補性。進一步採用 Proposed Transformer 後，融合特徵同樣達到最佳（MSE = 0.4356, 平均絕對誤差〔Mean Absolute Error, MAE〕= 0.5094, PCC = 0.8031），整體優於 Vanilla 對應設定（MSE = 0.4699, PCC = 0.7876）。

表 7　基於 Transformer 模型之迴歸與離散比較（於測試集中）

模型	特徵	MSE ↓	MAE ↓	PCC ↑	ACC ↑（離散）
Vanilla Transformer	LPR	0.4896	—	0.7766	—
	LPP	0.5105	—	0.7686	—
	LPP + LPR	0.4699	—	0.7876	—
Proposed Transformer	LPR	0.4518	0.5219	0.7956	—
	LPP	0.4715	0.5385	0.7847	—
	LPP + LPR	0.4356	0.5094	0.8031	—
Proposed Rounded Regression（ProR）	LPP + LPR	0.4629	0.4982	0.7898	32.03%
九類 Transformer 分類器	LPP + LPR	0.5740	0.5740	0.7713	39.37%

註：MSE：均方誤差（Mean Squared Error）；MAE：平均絕對誤差（Mean Absolute Error）；PCC：皮爾森相關係數（Pearson Correlation Coefficient）；ACC：準確率（Accuracy）；LPP：對數音素事後機率（Log-Phone Posterior）；LPR：對數事後機率比（Log-Posterior Ratio）。

資料來源：本研究自行產製。

若將本任務視為分類任務的狀況下，九類 Transformer 分類器可達 39.37%，高於取整迴歸的 32.03%；然而分類器在連續一致性指標上明顯較弱（MSE/MAE 較高、PCC 較低），且其預測分布更傾向集中於 2 ~ 3 分區間、極端分數較稀少，反映訓練標籤偏態對分類器決策邊界的影響。綜合而言，若應用目標為 CALL 教學場景中的連續分數為主要輸出，必要時可離散化為等第的方式，本研究建議以 Transformer 迴歸作為主要輸出機制，並以離散化作為呈現層策略，以維持較佳的序位一致性與可校準性。

另外，我們發現由於測試集分數分布偏向高分區間，高品質詞句在訓練中相對稀缺，未校正之 Transformer 迴歸器對高分樣本容易呈現系統性低估。全局偏差（Global Bias）（+ 0.1）僅帶來有限改善；相較之下，類別特定偏差校正（Class-Specific Bias Calibration）可顯著提升連續與離散表現。以最佳迴歸模型（Pro/ProR, LPP + LPR）為例，校正後 MAE 由 0.5094 降

至 0.4123，PCC 由 0.8031 升至 0.8948（如表 8），離散 ACC 亦提升至 34.41%（如表 9），並優於九類 Transformer 分類器（ACC = 39.37%）。統計檢定方面，+0.1 的 MAE 改善不顯著（Wilcoxon $p = 0.409$），而類別特定偏差校正之 MAE 降幅高度顯著（$p = 8.02 \times 10 \text{ ^ } -58$），顯示在偏態標籤下，針對預測落點進行的輕量校正層能更有效修正分數相依偏差（Score-Dependent Bias），使模型輸出更貼近教師量尺。本文亦嘗試以重採樣方式增加少數分數區間樣本比例，但整體表現反而退化（例如 MSE = 0.667、MAE = 0.566、ACC = 35.30%）。推測原因為重複樣本可能放大標註雜訊或造成過擬合，未能提供極端分數的多樣化證據。因此，本研究更傾向採用 Transformer 迴歸模型搭配依分數區間校正的方式，來作為資料分布偏移的修正策略。

表 8　類別特定偏差校正對 Transformer 迴歸一致性之影響（測試集）

模型	MAE ↓	PCC ↑
ProR（LPP + LPR）	0.5094	0.8031
ProR + b_bin（LPP + LPR）	0.4123	0.8948

註：MAE：平均絕對誤差（Mean Absolute Error）；PCC：皮爾森相關係數（Pearson Correlation Coefficient）；ProR：四捨五入迴歸（Proposed Rounded Regression）；LPP：對數音素事後機率（Log-Phone Posterior）；LPR：對數事後機率比（Log-Posterior Ratio）。

資料來源：本研究自行產製。

表 9　基於 Transformer 離散指標比較（測試集，Macro 指標）

設定	ACC ↑	MP ↑	MR ↑	MF1 ↑
ProR（LPP + LPR）	32.03	0.386	0.290	0.296
ProR + 0.1（LPP + LPR）	34.41	0.374	0.285	0.294
ProR + b_bin（LPP + LPR）	45.34	0.407	0.444	0.417
九類 Transformer 分類器	39.37	0.318	0.335	0.303

註：ACC：準確率（Accuracy）；MP：巨觀精確率（Macro-Averaged Precision）；MR：巨觀召回率（Macro-Averaged Recall）；MF1：巨觀 F1 分數（Macro F1-Score）；ProR：四捨五入迴歸（Proposed Rounded Regression）；LPP：對數音素事後機率（Log-Phone Posterior）；LPR：對數事後機率比（Log-Posterior Ratio）。

資料來源：本研究自行產製。

（三）案例分析

在此我們以實際範例「吃晚飯」對應之客語詞彙「食夜」作為案例，來呈現低資源情境下斷詞效果對翻譯品質之影響（如表 10 所示）。在基礎斷詞設定下，模型生成的結果，將同時生成「食」與「食夜」，導致重複翻譯，顯示詞邊界不穩定時，模型難以在「食 + 名詞」與特定詞彙（食夜）之間做出一致選擇。相較之下，導入強制斷詞後，系統更傾向輸出「食夜」作為整體對應詞，降低關鍵詞失真。另一方面，以大量資料訓練之語料庫斷詞在此類固定搭配詞上具有更佳泛化性，能在保留關鍵詞的同時維持語意一致。因此，當上游斷詞能穩定維持詞彙邊界時，可有效抑制關鍵詞翻譯錯誤，並避免後續 G2P/TTS 受到錯誤輸入影響而擴散。

表 10 不同斷詞方式對於翻譯結果之影響

斷詞設定	原始中文	中文斷詞	客文翻譯輸出	錯誤類型／觀察
Hakka Jieba	找一間飯館吃晚飯	找 一 間 飯館 吃 晚飯	尋 一 家 飯館 食 食夜	關鍵詞重複（食 + 食夜同時生成）
強制斷詞	找一間飯館吃晚飯	找 一 間 飯館 吃晚飯	尋 一 間 食堂 食夜	關鍵詞保留；局部詞彙替換（飯館 → 食堂）
語料庫斷詞	找一間飯館吃晚飯	找 一 間 飯館 吃 晚飯	尋 一 間 飯館 食夜	關鍵詞與語意較穩定

資料來源：本研究自行產製。

而低資源翻譯常見問題之一為未知詞或不可靠詞彙進入輸出，進而使下游 G2P/TTS 無法穩定處理。為減少此狀況對下游模組之錯誤擴散，本研究在解碼端採用注意力對齊回填策略：當輸出序列出現未知詞時，利用 Seq2Seq 注意力矩陣回溯該輸出位置在來源端最可能對應的詞元，並以該來源詞元進行回填，使輸出維持可讀且可被後續 G2P/TTS 處理。表 11 中以「ㄖ渴去喝茶」為輸入範例，顯示輸出詞「ㄖ渴」與「去」呈現明確的一對一對齊，而輸出短語「食茶」同時對齊至「喝」與「茶」，反映模型將「喝茶」視為整體語意單位進行轉換。此對齊訊號使系統在遇到未見詞或語料不足導致的低信心度輸出時，仍能以來源詞回填產生可用結果，降低不可用符號傳遞至下游造成串接失敗的風險。整體而言，本策略提供一種輕量且可解釋的機制，使模組化流程在低資源情境下具較佳穩定性。

表 11 未知詞處理之步驟與對應結果

步驟	輸入／輸出	說明
中文輸入	ㄖ渴 去 喝 茶	—
初始客文輸出	ㄖ渴 去 <Unk>	低資源下可能出現未知詞
未知詞懲罰觸發	避免 <Unk> 作為候選	降低 <Unk> 被選中的機率
注意力回填	以來源注意力對齊回填「茶」→「食茶」	用來源詞回填成可讀輸出
最終客文輸出	ㄖ渴 去 食茶	避免 <Unk> 擴散到下游

註：<Unk> 表為未知詞。
資料來源：本研究自行產製。

伍、結論與建議

本研究以低資源客語為對象，提出一套以拼音作為 IR 的模組化閉環系統，將中文輸入依序轉換為客文、拼音並生成示範語音，同時提供可解釋的 APA 回饋，以支援教學與自學情境之需求。在中文轉客文方面，本文以強制斷詞與未知詞處理策略改善低資源下的詞彙對齊與可用輸出比例，提升外部測試域的穩健性；在 APA 方面，本文以 GOP 家族特徵結合迴歸模型輸出連續分數，並以人評一致性作為主要檢核指標，使回饋更具可解釋性與可追溯性。

在實務的教學與管理上，本系統可作為口說作業之自動化前置評量工具，先提供一致化的句級分數與可診斷線索，協助教師在大量作業下進行快速分流（例如優先針對低分或高不確定樣本進行人工複審），降低重複性評分負擔並提升評量一致性。對學習者而言，除總分外，系統可回傳音素／音節層級的錯誤指標（例如 LPR 較高分所對應之發音音素），使回饋更具指向性。於部署層面，拼音中介使各模組可替換並維持介面一致，利於在不同腔調、不同教材或不同校務平臺中擴充與維護使用。

未來的研究方向可從三方面延伸：其一，擴充跨腔調與跨領域平行語料並導入更細緻的專名／多義詞處理；其二，在 TTS 端加入更完整的客觀與主觀評測（如 MOS、梅爾倒頻譜距離〔Mel Cepstral Distance, MCD〕等）並探索自監督表徵輔助之低資源合成；其三，於評分端加入錯誤類型診斷（如音節／聲調層級）與個人化校正，以提升教學回饋的可操作性。

參考文獻

Peng W.-H.、Wang H.-W.、Chen S.、Chen B.，2023，〈結合 BERT 與 Wav2vec 2.0 提升第二外語受試者之自動英語口說評測〉，收錄於《ROCLING 2023：第 35 屆自然語言與語音處理研討會》，臺北：中華民國計算語言學學會，頁 98-105。

林孟欣、王馨偉、羅天宏、陳柏琳、趙偉成，2023，〈改善多細粒度的發音評測上資料不平衡的問題〉，收錄於《ROCLING 2023：第 35 屆自然語言與語音處理研討會》，臺北：中華民國計算語言學學會，頁 134-140。

客家委員會，n.d.-a，〈客語口說故事〉，《哈客網路學院》，https://elearning.hakka.gov.tw/voice_story/index.html（瀏覽日期：2026 年 1 月 24 日）。

客家委員會，n.d.-b，《臺灣客語語料庫》，https://corpus.hakka.gov.tw/（瀏覽日期：2026 年 1 月 24 日）。

洪翌翔，2023，《運用少量客家語料建置中文文字轉客語語音系統之研究》，國立屏東大學電腦科學與人工智慧學系碩士論文。

教育部，2012，《客家語拼音方案使用手冊》，臺北：教育部。

葉秋杏、賴惠玲，2023，〈從語料庫建構探討臺灣客語難字、缺字與異體字議題〉，《臺灣語文研究》，18（1），頁 135-183。doi:10.6710/JTLL.202304_18(1).0003

Baevski A., Zhou H., Mohamed A., and Auli M., 2020, “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations,” in *NIPS’20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY: Curran Associates, 12449-12460.

Bisani M. and Ney H., 2008, “Joint-Sequence Models for Grapheme-to-Phoneme Conversion,” *Speech Communication*, 50(5), 434-451. doi:10.1016/j.specom.2008.01.002

Chen Y.-J., Tu T., Yeh C.-C., and Lee H.-Y., 2019, “End-to-End Text-to-Speech for Low-Resource Languages by Cross-Lingual Transfer Learning,” in *20th Annual Conference of the Interna-*

tional Speech Communication Association (INTERSPEECH 2019), Red Hook, NY: Curran Associates, 2075-2079. doi:10.21437/Interspeech.2019-2730

Gehring J., Auli M., Grangier D., Yarats D., and Dauphin Y. N., 2017, "Convolutional Sequence to Sequence Learning," *Journal of Machine Learning Research*, 70, 1243-1252.

Gong Y., Chen Z., Chu I.-H., Chang P., and Glass J, 2022, "Transformer-Based Multi-Aspect Multi-Granularity Non-Native English Speaker Pronunciation Assessment," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Los Alamitos, CA: IEEE Computer Society Press, 7262-7266. doi:10.1109/ICASSP43922.2022.9746743

Hsu W.-N., Bolte B., Tsai Y.-H. H., Lakhotia K., Salakhutdinov R., and Mohamed A., 2021, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 29, 3451-3460. doi:10.1109/TASLP.2021.3122291

Jia Y., Zhang Y., Weiss R. J., Wang Q., Shen J., Ren F., Chen Z., Nguyen P., Pang R., Moreno I. L., and Wu Y., 2018, "Transfer Learning from Speaker Verification to Multispeaker Text-to-Speech Synthesis," in *NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems*, Red Hook, NY: Curran Associates, 4485-4495.

Kanters S., Cucchiarini C., and Strik H., 2009, "The Goodness of Pronunciation Algorithm: A Detailed Performance Study," paper presented at the ISCA International Workshop on Speech and Language Technology in Education, SLaTE 2009, Warwickshire, UK.

Ko T., Peddinti V., Povey D., and Khudanpur S., 2015, "Audio Augmentation for Speech Recognition," in *6th Annual Conference of the International Speech Communication Association (INTERSPEECH 2015)*, Red Hook, NY: Curran Associates, 3586-3589. doi:10.21437/Interspeech.2015-711

Kong J., Kim J., and Bae J., 2020, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," in *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY: Curran Associates, 17022-17033.

ldkrsi, 2018/6/10, "jieba-Hakka," *GitHub*, https://github.com/ldkrsi/jieba-Hakka (accessed January 26, 2026).

Li P.-H. and Ma W.-Y., 2019, "CKIPTagger," *GitHub*, https://github.com/ckiplab/ckiptagger (accessed January 26, 2026).

Liao Y.-F., Hwang S.-H., Chen Y.-S., Lai H.-C., Chung Y.-H., Shen L.-T., Huang Y.-C., Huang C.-J., Han H. W., Chen L.-W., Su P.-C., and Huang C.-S., 2023, "Taiwanese Hakka across Taiwan Corpus and Formosa Speech Recognition Challenge 2023—Hakka ASR," in *2023 26th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, Los Alamitos, CA:

IEEE Computer Society Press, 190-195. doi:10.1109/O-COCOSDA60357.2023.10482979

Liao Y.-F., Kuo C.-C., Huang C.-S., Lan Y.-S., Lai H.-C., and Hsu W.-H., 2025, "Taiwanese Hakka across Taiwan Corpus and Formosa Speech Recognition Challenge 2025—Dapu & Zhao'an Accents," in *37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025)*, Kerrville, TX: Association for Computational Linguistics, 427-434.

McAuliffe M., Socolof M., Mihuc S., Wagner M., and Sonderegger M., 2017, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *18th Annual Conference of the International Speech Communication Association (INTERSPEECH 2017)*, Red Hook, NY: Curran Associates, 498-502. doi:10.21437/Interspeech.2017-1386

Ott M., Edunov S., Baevski A., Fan A., Gross S., Ng N., Grangier D., and Auli M., 2019, "fairseq: A fast, Extensible Toolkit for Sequence Modeling," in *2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, Stroudsburg, PA: Association for Computational Linguistics, 48-53. doi:10.18653/v1/N19-4009

Povey D., Ghoshal A., Boulianne G., Burget L., Glembek O., Goel N., Hannemann M., Motlíček P., Qian Y., Schwarz P., Silovský J., Stemmer G., and Veselý K., 2011, "The Kaldi Speech Recognition Toolkit," paper presented at 2011 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU 2011), Waikoloa, HI.

Ren Y., Hu C., Tan X., Qin T., Zhao S., Zhao Z., and Liu T.-Y., 2020/6/8, "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech," *arXiv*, https://doi.org/10.48550/arXiv.2006.04558 (accessed January 26, 2026).

Ren Y., Ruan Y., Tan X., Qin T., Zhao S., Zhao Z., and Liu T.-Y., 2019, "FastSpeech: Fast, Robust and Controllable Text to Speech," in *33rd International Conference on Neural Information Processing Systems (NeurIPS 2019)*, Red Hook, NY: Curran Associates, 3171-3180.

Sennrich R., Haddow B., and Birch A., 2016, "Neural Machine Translation of Rare Words with Subword Units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA: Association for Computational Linguistics, Vol. 1, 1715-1725. doi:10.18653/v1/P16-1162

Shen J., Pang R., Weiss R. J., Schuster M., Jaitly N., Yang Z., Chen Z., Zhang Y., Wang Y., Skerrv-Ryan R., Saurous R. A., Agiomvrgiannakis Y., and Wu Y., 2018, "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Los Alamitos, CA: IEEE Computer Society Press, 4779-4783. doi:10.1109/ICASSP.2018.8461368

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser Ł., and Polosukhin I., 2017, "Attention Is All You Need," in *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, Red Hook, NY: Curran Associates, 6000-6010.

Wang Y., Stanton D., Zhang Y., Skerry-Ryan R. J., Battenberg E., Shor J., Xiao Y., Jia Y., Ren F.,

and Saurous R. A., 2018, “Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis,” *Proceedings of Machine Learning Research*, 80, 5180-5189.

Witt S. M. and Young S. J., 2000, “Phone-Level Pronunciation Scoring and Assessment for Interactive Language Learning,” *Speech Communication*, 30(2-3), 95-108. doi:10.1016/S0167-6393(99)00044-8

Yao K. and Zweig G., 2015, “Sequence-to-Sequence Neural Net Models for Grapheme-to-Phoneme Conversion,” in *6th Annual Conference of the International Speech Communication Association (INTERSPEECH 2015)*, Red Hook, NY: Curran Associates, 3330-3334. doi:10.21437/Interspeech.2015-134