促進大學社區化之智慧互動導覽機台
結合即時情緒辨識、台語語音辨識與生成
A Smart Interactive Guidance Kiosk for Promoting University–Community Integration Integrating Real-Time Emotion Recognition, Taiwanese Speech Recognition, and Speech Generation

林奕安 陳偉峻 林威逸 羅岳均
國立聯合大學 資訊工程學系
苗栗市南勢里聯大2號
{U1124001, U1124024, U1024044 } @nuu.edu.tw

摘要

本專題開發一套「長者友善參訪導覽 App」，以降低老年族群在校園參訪中的使用門檻。系統整合台語語音辨識(Whisper 微調)、中文與台語語音生成(Bert-VITS2 與 mms-tts-nan)、情緒辨識(ResNet 量化部署於邊緣裝置)，使長者能透過語音與系統互動，並依其情緒狀態調整回應方式。後端以 FastAPI 建置，提供統一 API、設備配對與資料管理；前端 App 則具備導覽、景點提示與語音助理功能。整體系統成功提升長者於校園參訪時的可及性與友善度。

關鍵字：台語語音辨識、語音合成、KV260、長者友善導覽、智慧校園導覽 App

Abstract

This project develops an “Elder-Friendly Campus Tour App” designed to lower the barrier of use for older adults during campus visits. The system integrates Taiwanese Hokkien speech recognition (fine-tuned Whisper), Mandarin and Taiwanese Hokkien speech synthesis (Bert-VITS2 and mms-tts-nan), and emotion recognition (ResNet quantized and deployed on an edge device), enabling seniors to interact with the system through voice while receiving responses tailored to their emotional state.

The backend is built with FastAPI, providing unified APIs, device pairing, and data management; the frontend app offers tour guidance, attraction prompts, and voice assistant functions.

Overall, the system enhances accessibility and user-friendliness for seniors during campus tours.

Keywords: Taiwanese Hokkien speech recognition, speech synthesis, KV260, elder-friendly tour, smart campus tour app

一、 簡介

1.1 研究背景

智慧校園導覽仍停留在單向資訊提供，缺乏對表情、情緒等非語言訊息的感知，互動冷漠且不便於視障或不熟悉打字者，並預設使用者精通華語與數位操作，排斥部分族群，造成資訊落差與文化隔閡。本研究主張新一代智慧導覽應理解本土語言、感知使用者狀態，並結合社區大學化精神，拓展至在地社群，促進多元交流與共學，打造包容共享的智慧學習環境。

1.2 研究目的

承接上述研究背景，本研究旨在設計並實作一套具備「本土語言」與「情緒感知能力」的智慧導覽機台，將傳統的資訊查詢轉化為更具備人情味的互動夥伴。

1.3 相關研究

1.3.1 Whisper 微調

如文獻 [1]所指出，Whisper 在低資源語言(每種語言約10小時訓練資料)的情況下 Whisper 如何使用「參數高效的微調方法(PEFT)」大幅降低 WER，並比較性能與成本。該研究微調7種少數語言，並使用 Large-v2模型進行微調實驗，微調後錯誤率(WER)顯著下降。

1.3.2 ResNet 架構

如文獻 [2]所指出，ResNet 在圖像分類具有極高的準確性與穩定性，如表 所示，使用相同深度 ResNet 的分類誤差是較小的。此外，作者亦證實 ResNet 在 CIFAR-10等中小型圖像資料集上同樣表現優異，如圖 一所示。因此，本專題使用 ResNet 作為情緒辨識模型的主要架構。

表 一：Error rates on ImageNet validation.

模型	層數	Top-1 Error(%)	Top-5 Error(%)
Plain	34	28.54	10.02
ResNet	34	25.03	7.76

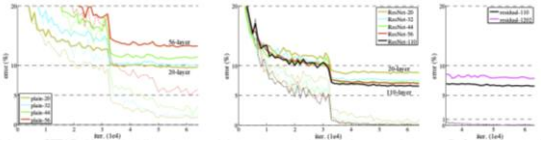

圖 一：CIFAR-10在 Plain 與 ResNet 模型的誤差比較

1.3.3 ResNet 模型

如文獻 [3]所指出，作者等人提出 ResNet18在資源受限的邊緣裝置上，經過量化後，可以達到約1.9倍的推論加速，如表 所示。此外，在文獻的量化實驗顯示在達到1.9倍推論加速的同時，其精度下降不到1%，證明 ResNet18在邊緣環境具備良好的即時性與效能表現。

表 二：ResNet 在邊緣設備上的推論時間比較

優化方式	推論時間(ms)
Baseline	1120
核心最佳化	588
圖最佳化	509
最佳化+INT 8量化	369

二、 系統內容

2.1 系統架構

本專題旨在服務老人而展開，因應老年人不擅長使用手機因此使用 Whisper 進行語音辨識，後傳給語音模型生成導覽服務，與此同時根據顧客表情判斷使用者心情，修改導覽服務內容，最後依使用者需要選擇使用中文或台語語音生成回應用戶，解決老人不識字的問題。

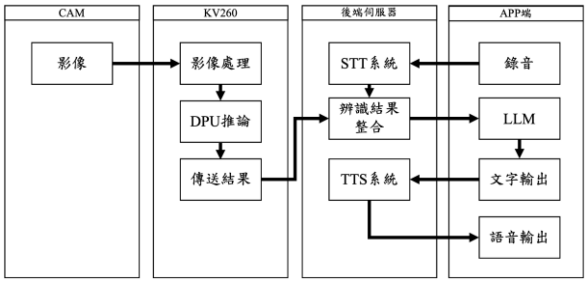

圖 二：系統架構圖

三、 系統功能

3.1 情緒辨識系統

圖 三：情緒辨識系統流程圖

3.1.1 情緒辨識模型

本系統之情緒辨識模型採用 TensorFlow 的 ResNet 架構，主要原因為 Vitis-AI 對 PyTorch 訓練模型支援度較低，因此選用 TensorFlow 作為開發框架。模型輸入為48x48像素8-bit 彩色影像，輸出為三類情緒的機率向量。

資料集透過水平翻轉、旋轉、對比度調整等方式進行資料增強，以提升對光線與姿勢變化的辨識能力。模型前段包含卷積、批次正規化與 ReLU，後段加入殘差模塊以改善梯度消失並提升準確率。特徵萃取後轉換為一維向量，再接上全連接層與 Softmax 完成分類。

表 三：情緒辨識模型

指標	數值
Accuracy	0.7664
平均 F1分數	0.7624

Weighted F1 (以樣本數作為權重)	0.7680

訓練採用 Adam 最佳化器與交叉熵損失函數，並加入 Early Stopping 避免過擬合。完成訓練後，使用 Vitis-AI 進行量化與編譯，生成可在 KV260 DPU 上執行的.xmodel 檔。

表 四：各類別測試表

類別	Precision	Recall	F1分數	測試集樣本數
生氣	0.6756	0.7328	0.7031	958
開心	0.8640	0.7984	0.8300	1281
平靜	0.7494	0.7591	0.7542	1233

整體模型表現如表 三，模型在另外分出的測試集上之整體準確率為76.64%。值得注意的是，平均 F1分數與加權 F1分數相當接近，表示模型並沒有出現明顯的誤差。

各類別表現如訓練採用 Adam 最佳化器與交叉熵損失函數，並加入 Early Stopping 避免過擬合。完成訓練後，使用 Vitis-AI 進行量化與編譯，生成可在 KV260 DPU 上執行的.xmodel 檔。

表 四，模型對開心的辨識能力最強，F1分數達到0.8300，代表系統能知道使用者是滿意得到的答案的。平靜的表現則中規中矩，比較差的是生氣的情緒，可能是生氣的臉部特徵不夠明顯導致。

3.1.2 多數決機制

本系統將固定時間內所取得的辨識結果次數進行統計，優先選擇辨識到最多次的情緒作為最終結果。若有多個情緒次數相同，則進一步比較他們的平均預測機率，選擇較高者作為最終結果。當累積的結果數量超過設定值(8幀)、與上次傳送之情緒不同、超過信心度、冷卻時間結束時，系統會將該結果與 APP 端 ID 封裝成 JSON 訊息，透過 HTTP POST 請求傳送至後端伺服器。

3.1.3 KV260上之即時推論性能

量化後的情緒辨識模型在 KV260的 DPU 上進行純推論可達318.7FPS，顯示 DPU 對卷積運算具有極高的加速效果。然而完整系統仍需進行 USB 攝影機擷取、人臉偵測(Haar Cascade)、裁切與縮放等前處理，其中 Haar Cascade 於 CPU 執行，形成主要瓶頸，使整體效能降至5.02FPS。

儘管加入人臉偵測後 FPS 明顯下降，但由於人類表情變化速度不高，5FPS 仍足以滿足即時辨識需求。此外，系統搭配多數決與冷卻時間機制，可提升輸出穩定性並降低伺服器負載。

表 五：情緒辨識系統平均處理時間

項目	方法	平均處理時間(毫秒)
擷取畫面	CPU	2.56
臉部偵測(未量化)	CPU	184.41
圖片預處理	CPU	0.93
DPU 推論	DPU	3.40
傳送結果	CPU	0.26

3.2 台語 STT

本系統整合了微調(fine-tuning)後的 Whisper 台語語音辨識模型，能有效處理使用者輸入的台語語音並自動轉換成對應的文字內容。整體流程如下圖 四。

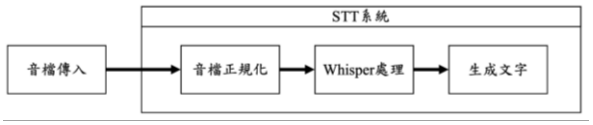

圖 四：STT 架構圖

在本專題中，我們採用 Transformers 框架對語音識別模型 Whisper 進行微調，並選用 Common Voice 23.0–台灣話(閩南語) [4]作為主要訓練資料，一共有23小時了訓練資料。

由於 Whisper 提供多種模型以供微調，因此嘗試多種模型大小 small、medium 用同一份資料集結果如圖 五，微調選擇效果最好的 medium。

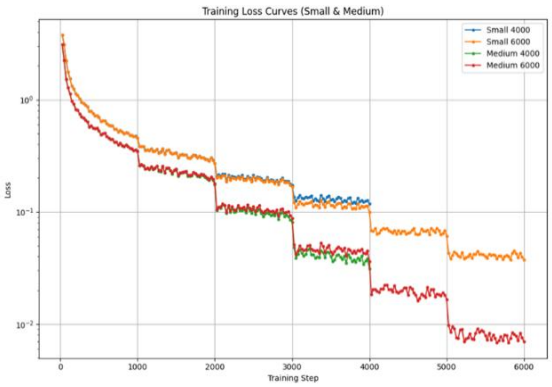

圖 五：Train loss 折線圖

3.3 TTS

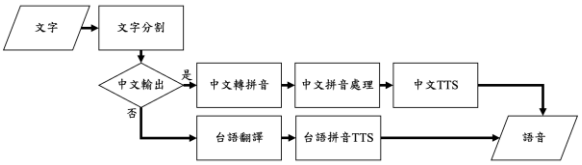

圖 六：TTS 流程圖

3.3.1 中文 TTS

本研究以 Bert-VITS2 [5]為中文語音合成框架，能生成高品質且貼近真實語音的輸出。

在 TTS 執行時，系統將文章分段，先用 BERT 編碼器提取語義以改善韻律與情感，再交由 VITS 生成語音。分段合成並加入停頓，使朗讀更自然、穩定。

3.3.2 台語 TTS

採用 facebook/mms-tts-nan [6]作為台語(閩南語)語音合成模型。此模型以羅馬拼音(Pe̍h-ōe-jī)作為輸入，最後將取得的台語拼音輸入語音模型後輸出台語語音。

3.4 Server

Server 使用 FastAPI 建立後端 API，作為前端與各種模型之間的統一介面如圖 十一。Server 提供標準化的 HTTP REST API，負責接收請求、呼叫對應模型做推論(或轉換)、並回傳統一格式的結果如圖 七。

STT http://localhost:8001/transcribe	TTS http://localhost:8001/text
取得情緒 http://localhost:8001/emotion-from-text	APP配對 http://locallgst:8001/pair
情緒分析結果 http://localhost:8001/emotion	設備配對 http://localhost:8001/pair_device

圖 七：HTTP 呼叫網址

3.4.1 設備配對

運作流程如圖 八，當 APP 端發起重新配對請求時，Server 會接收該請求，並在30秒內將已提交配對申請的設備進行分配，將可用設備指派給對應的 APP。

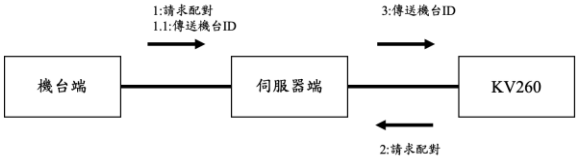

圖 八：配對流程圖

3.4.2 日誌管理

此功能整合各個服務的輸出訊息，將原本分散、難以追蹤的紀錄集中處理，不僅提升除錯效率，也讓整體系統的運作狀態更加清晰易懂，結果如圖 九。

圖 九：TTS 日誌

3.4.3 Ngrok 外網連線

本專題透過 ngrok 將本機伺服器映射至可公開存取的 HTTP 網址，使前端在無固定 IP 或雲端主機的情況下，仍能與後端服務順利溝通。為檢驗其運行效果，我們設計約20－30字的測試文稿，共20句，並記錄每次執行的處理時間，再取其平均值，以作為系統效能的客觀評估指標。最終統計結果如表 六。

表 六：平均處理時間

API 項目	平均處理時間
STT	6.898 (s)
TTS 中文	2.403 (s)
TTS 台語	3.095 (s)

3.5 APP 問卷地圖功能

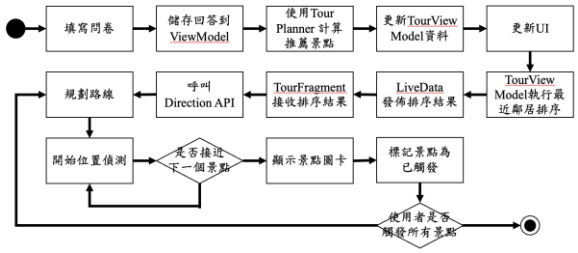

圖 十：規劃路線活動圖

3.5.1 首頁與智慧導覽介紹

首先**錯誤! 找不到參照來源。**展示了的主畫面。整體設計以使用者友善為核心，其中「開始導覽」可引導使用者進入問卷流程以規劃個人化行程；**錯誤! 找不到參照來源。**「生成智慧導覽介紹」則會依據已生成的路線產生語音導覽內容。

圖 十一：首頁畫面

圖 十二：導覽介紹

3.5.2 問卷頁與行程頁

當使用者點擊「開始規劃」後，便會進入問卷流程。展示的個人化問卷介面採用一次一題的全螢幕設計。若對系統推薦的行程不滿意，也可在行程頁中直接手動調整景點。

圖 十三：問卷

圖 十四：行程頁

3.5.3 導覽頁面與景點圖卡觸發

選定景點後,APP 會根據使用者當前位置,自動計算並排序這些景點的訪問順序。排序完成後開始地圖繪製,並行呼叫 Directions API 取得各段實際步行路徑。取得路徑後解碼並依照訪問狀態以不同顏色的 Polyline 標示於地圖上:灰色表示已完成路段,藍色為當前導航路段,黑色為未來待訪問路段,形成清晰的導航路線。系統會持續追蹤使用者位置,當靠近景點小於 50 公尺時即判定為抵達,自動標記該景點為已訪問並更新路線顏色,同時彈出對應的景點資訊卡片。

圖 十五：導覽頁面

圖 十六：景點卡觸發

3.6 助理聊天室

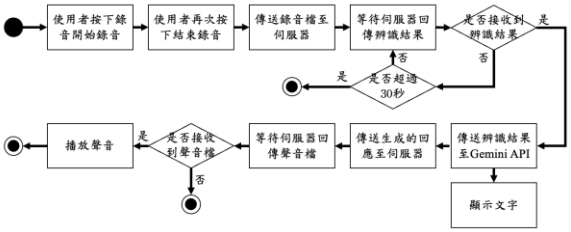

圖 十七：語音助理活動圖

3.6.1 Prompt 機制與固定問答集

我們設計了特定 Prompt 如圖 十八，讓語音助理只提供校園資訊，避免不當回應。遇到超出能力的問題，助理會誠實說明角色限制。此外，系統會先比對使用者問題與固定問答集，即使語音辨識有誤，也能捕捉核心意圖並匹配正確問答，確保常見問題快速準確回應。

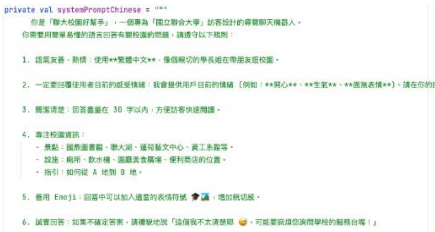

圖 十八：中文角色設定提示詞

3.6.2 AI 助理頁面

AI 助理頁面提供語音開關與語言切換設定，中央為對話紀錄區，底部有文字與語音輸入功能。使用者可打字或錄音提問，系統透過 Whisper API 辨識語音，再由 Gemini API 生成回覆，最後透過後端伺服器朗讀指定語言，提供流暢的互動體驗。

圖 二十五：AI 助理頁面

四、 結論

目前，聯合大學所開發的 長者友善參訪導覽 App 已具備基本的服務功能台語文字辨識與語音生成：讓使用者能透過台語文字或語音互動，降低操作難度，提升使用便利性;地圖導覽功能：提供校園地圖與路線引導，幫助參訪者順利瀏覽校園各處，並能即時獲得位置資訊，能夠為行動不便或不熟悉智慧型手機操作的老年使用者提供便利。透過以上功能，App 不僅解決了老年使用者在數位操作上的困難，也落實了友善校園的設計理念，實際滿足了提升校園參訪體驗與資訊無障礙的需求。

4.1 未來目標與改進

4.1.1 完善台語 TTS 功能

- 將現有台語文字轉語音流程改進為 完全離線運作，減少對網路連線的依賴。
- 將台語整合進 Bert-VITS2 模型，以生成更自然、流暢且情感豐富的台語語音。

4.1.2 擴展多語言支持

- 除台語外，未來計畫加入客語等其他地方語言，以照顧更多不同語言背景的長者使用者，提升系統的普適性與友善度。

4.1.3 完善資料庫管理

- 優化語音與文字資料庫，避免重複生成相同語句。降低系統資源消耗，提高運算效率，確保長時間、多使用者使用下的穩定性與效能。
- 透過上述改進，本系統將能提供更 完整、多元、穩定且貼心的長者友善參訪導覽服務，並進一步落實智慧校園與無障礙設計理念。

4.1.4 情緒感測系統

- 加入更多情緒，使系統更加了解使用者當前的情緒。
- 自訓練臉部偵測模型，透過量化實作在 KV260上，使臉部偵測的處理時間降低。

五、 專題作品影音平台網址

https://youtu.be/Chma9zJ6weg

參考文獻

[1] Y. Liu and D. Qu,, “Parameter-efficient fine-tuning of Whisper for low-resource speech recognition,” 於 *2024 5th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT)*, Nanjing, China, 2024.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, “Deep Residual Learning for Image Recognition,” 2015.

[3] Ziheng Jiang, Tianqi Chen, Mu Li, “Efficient Deep Learning Inference on Edge Devices,” Stanford, CA USA, 2018.

[4] Mozilla, “Common Voice Scripted Speech 23.0 - Taiwanese (Minnan),” Mozilla, 15 9 2025. [線上]. Available: https://datacollective.mozillafoundation.org/datasets/cmflnuzw6c692o9c40agnyeyf. [存取日期: 16 11 2025].

[5] fishaudio, “Bert-VITS2,” fishaudio, 2025. [線上]. Available: https://github.com/fishaudio/Bert-VITS2?tab=readme-ov-file. [存取日期: 16 11 2025].

[6] facebook, “Massively Multilingual Speech (MMS): Chinese, Min Nan Text-to-Speech,” facebook, [線上]. Available: https://huggingface.co//mms-tts-nan. [存取日期: 16 11 2025].