屏東縣第66屆國民中小學科學展覽會

作品說明書

科　　別：生活與應用科學科（一）

組　　別：國中組

作品名稱：聲影共舞-基於電腦視覺之聽障者
無障礙音樂展演系統

關 鍵 詞：：手語辨識、AI 訓練、人工智慧協作

編號：B6015

目錄

摘要 1
壹、前言 2
　一、研究動機 2
　二、研究目的 2
　三、文獻回顧 3
貳、研究設備及器材 6
　一、軟體環境 6
　二、硬體環境 6
參、研究過程 8
　一、研究架構 8
　二、研究流程 8
　三、系統架構設計 9
　四、素材採集 9
　五、資料前處理 12
　六、手語辨識模型訓練 14
　七、語音合成模組 15
　八、系統整合與程式開發 15
　九、實驗設計 17
肆、研究結果 19
　一、手語辨識模型效能 19
　二、語音合成品質 24
　三、系統整合測試 24
　五、實際演出測試 25
伍、討論 27
　一、研究過程遇到的困難與解決 27
　二、模型效能分析 27
　三、系統限制與改善方向 28
　四、AI 倫理討論 28
　五、研究心得 29
陸、結論 29
　一、研究成果總結 29
　二、未來展望 29
　三、社會意義與應用價值 29
柒、參考文獻資料 30

摘要

本研究旨在運用電腦視覺與人工智慧技術，建構一套「聽障者無障礙音樂展演系統」，讓聽障同學得以透過手語比劃，由系統即時辨識後觸發，預先合成之個人化語音片段組；合成其個人化聲音。實現「用自己的聲音唱歌」之夢想。研究團隊以 Google MediaPipe Hands 進行手部 21 個關鍵點偵測，搭配卷積神經網路（CNN）與長短期記憶網路（LSTM）建立靜態與動態手語辨識模型，針對〈月亮代表我的心〉40 段手語詞組與〈Perfect Night〉70 段手語詞組進行辨識訓練。語音合成採用 RVC（Retrieval-based Voice Conversion）技術，僅需約 5 分鐘的個人歌聲錄音即可訓練出個人化語音模型。系統於一般無 GPU 文書筆電上運行，整體管道延遲控制在 35 毫秒以內。實驗結果顯示，個人化微調模型之手語辨識準確率達 93.2%，較通用模型提升 11.8 個百分點；語音合成之 MOS 主觀評分達 3.76 分（滿分 5 分）。本研究結合資訊科技、音樂與特殊教育之跨領域整合，展現 AI 科技關懷弱勢之社會價值。

壹、前言

一、研究動機

我們是三位來自屏東縣某國中的國二女生。在我們之中，有一位從小就有輕微聽障的同學，她的聽力與理解能力其實都和一般人相同，但因為構音障礙的緣故，說話時會有些呢喃不清。從小到大，她最羨慕的事情就是能夠像其他同學一樣開口唱歌。

2024 年 4 月，我們在網路上看到韓國男子團體 Big Ocean 的新聞報導，深受震撼。Big Ocean 是 KPOP 歷史上第一個三名成員全部患有聽力障礙的男子團體，他們在 HYBE 旗下 Supertone 公司所開發的 AI 語音轉換技術協助下，成功克服發音困難，重新唱出動人的歌曲（Euronews Next, 2024）。他們不僅登上聯合國 AI for Good 高峰會的舞台，更以手語融入舞蹈編排，向全世界證明——聽障不應成為音樂夢想的阻礙。

看完這則報導，三位同學激動地討論了很久。另外兩位同學會演奏揚琴及會演奏琵琶，我們突然靈光一閃：如果 Big Ocean 能用 AI 唱歌，那我們是不是也可以運用所學的資訊科技，幫助聽損同學實現「用自己的聲音唱歌」的夢想呢？

我們的構想是這樣的：聽損同學透過比手語，系統以 AI 辨識手語內容後，利用她預先錄製的聲音，經由 AI 語音合成技術，透過喇叭「自己的聲音」唱出歌曲，同時兩位同學分別演奏揚琴與琵琶伴奏，三個人一起完成一場溫暖的音樂展演。我們選擇了鄧麗君的經典歌曲〈月亮代表我的心〉以及韓國女團 LE SSERAFIM 與遊戲《鬥陣特攻 2》合作的英語單曲〈Perfect Night〉作為展演曲目——一首華語經典、一首英語流行，展現系統的跨語言能力。

這不只是一個科展研究，更是三位好朋友之間的承諾：我們要用科技的力量，讓每一個人都能擁有唱歌的權利。

二、研究目的

基於上述動機，本研究之具體目的如下：

（一）運用 Google MediaPipe Hands 電腦視覺技術，進行即時手部關鍵點偵測，建立手語姿態辨識系統。

（二）建立靜態手語圖片分類模型（CNN）與動態手語影片辨識模型（LSTM），比較不同模型架構之辨識效能，以確認最適合本系統的辨識架構。

（三）運用 RVC 語音轉換技術，以少量個人歌聲錄音訓練個人化語音合成模型，實現「用自己的聲音唱歌」之功能。

（四）整合手語辨識、語音合成與音訊混音模組，開發一套可在一般文書筆電上即時運行的無障礙音樂展演系統。

（五）設計嚴謹的對照實驗，評估模型效能與系統穩定性，並探討 AI 技術應用於無障礙設計的社會意義與倫理議題。

三、文獻回顧

（一）電腦視覺與手語辨識

手語辨識（Sign Language Recognition, SLR）是電腦視覺領域中重要的研究方向。Koller（2020）對 1983 至 2020 年間約 300 篇手語辨識研究進行了量化調查，指出深度學習方法已成為主流。Adaloglou 等人（2021）進一步綜合分析了基於深度學習的手語辨識方法，包括 CNN、RNN、Transformer 等架構。Rastgoo 等人（2021）的深度綜述則涵蓋了從特徵提取到端對端辨識的完整技術演進。

在台灣，手語辨識研究也逐漸受到重視。國立臺北科技大學的碩士研究（110 學年度）結合 MediaPipe 手部檢測框架與 Inflated 3DCNN 網路，針對 98 個臺灣手語單字進行辨識，準確率高達 97.4%。國立陽明交通大學的碩士研究（109 學年度）則使用深度學習進行基於影片的台灣手語辨識，建立涵蓋教育部手語字典 40 個常用詞彙的資料集，集成方法達 98.64% 辨識準確率。

（二）MediaPipe 手部關鍵點偵測

MediaPipe 是 Google 開發的跨平台即時感知框架（Lugaresi et al., 2019）。其手部追蹤模組 MediaPipe Hands（Zhang et al., 2020）採用機器學習管道，透過掌心偵測器（Palm Detector）與手部關鍵點模型（Hand Landmark Model）兩階段架構，可偵測每隻手的 21 個三維關鍵點座標（x, y, z），並內建左右手分類功能。

MediaPipe Hands 最大的優勢在於其輕量化設計——模型大小僅約 2-4 MB，無需 GPU 即可在一般筆電 CPU 上達到 20-30 FPS 的即時推論速度。這使其非常適合本研究所面臨的硬體限制（無 GPU 文書筆電）。目前 MediaPipe 已更名為 MediaPipe Solutions，歸屬 Google AI Edge 品牌，最新版本為 0.10.32（2026 年 1 月發布），支援 Python 3.9-3.12 以及 Windows、macOS、

Linux 等多平台。

（三）卷積神經網路（CNN）與遷移學習

卷積神經網路（Convolutional Neural Network, CNN）是影像辨識的核心技術。LeCun 等人（1998）提出的 LeNet 為 CNN 的奠基之作；Krizhevsky 等人（2012）的 AlexNet 在 ImageNet 競賽中大幅領先傳統方法，正式掀起深度學習革命；He 等人（2016）提出的 ResNet 以殘差學習解決深層網路退化問題，使超過 100 層的深層網路成為可能。

遷移學習（Transfer Learning）是本研究的關鍵策略之一。Pan 與 Yang（2010）的經典綜述指出，遷移學習可將在大型資料集上學到的特徵轉移至小型目標任務，有效解決訓練資料不足的問題。本研究採用先以通用手勢資料集預訓練，再以專用手語資料微調的策略，體現遷移學習的實際應用價值。

（四）長短期記憶網路（LSTM）

長短期記憶網路（Long Short-Term Memory, LSTM）由 Hochreiter 與 Schmidhuber（1997）提出，是一種改進版的循環神經網路（RNN），透過遺忘門（Forget Gate）、輸入門（Input Gate）與輸出門（Output Gate）三重閘門機制，有效解決傳統 RNN 的梯度消失問題，能夠學習長時間序列中的依賴關係。

在手語辨識領域，LSTM 特別適合處理動態手語的時序資訊。Khartheesvar 等人（2024）結合 MediaPipe Holistic 與 LSTM 網路實現印度手語辨識；挪威的研究團隊（2025）也採用 MediaPipe 與 LSTM 的組合達成即時挪威手語辨識。本研究將 LSTM 應用於動態手語影片的時序特徵學習，以捕捉手語詞組在時間維度上的變化模式。

（五）語音合成技術（TTS）與聲音複製

語音合成（Text-to-Speech, TTS）技術在近年因深度學習而有突破性進展。van den Oord 等人（2016）提出的 WaveNet 模型開創了神經網路生成原始音訊波形的先河；Shen 等人（2018）的 Tacotron 2 實現了端到端的語音合成。在開源領域，Kim 等人（2021）提出的 VITS 模型結合了變分自編碼器與對抗學習，能夠直接從文字生成高品質語音。

聲音複製（Voice Cloning）是本研究的核心技術之一。Jia 等人（2018）首先提出僅需數秒參考音訊即可進行語音複製的方法。Casanova 等人（2022）的 YourTTS 進一步實現了零樣本多語言語音複製。台灣大學的 Huang 等人（2022）則提出 Meta-TTS，透過元學習策略實現少樣本語者適應。

在歌聲轉換方面，RVC（Retrieval-based Voice Conversion）是目前最適合少量訓練資料的開源方案，僅需約 3-5 分鐘的乾淨人聲即可訓練個人化聲音模型，並可在 Google Colab 免費版的 T4 GPU 上完成訓練。此外，So-VITS-SVC 與 GPT-SoVITS 等開源工具也提供了替代方案。

（六）聽障者音樂參與相關研究

根據世界衛生組織（WHO）2021 年發布的《世界聽力報告》，全球超過 15 億人有某種程度的聽力損失，其中 4.3 億人（含 3,400 萬兒童）有致殘性聽力損失；預計至 2050 年，近 25 億人將面臨聽力問題（WHO, 2021）。在台灣，依據教育部統計，特殊教育學校（啟聰類）學生數約 643 人，而融合教育中的聽障學生人數持續增加。

研究顯示，聽障者並非無法感受音樂。Tranchant 等人（2017）發現，透過振動觸覺刺激，早期失聰者也能與音樂節拍同步。Torppa 與 Huotilainen（2019）指出，音樂訓練有助於聽障兒童的語言發展。Rochette 等人（2014）也發現，音樂課程能提升聾人兒童的聽覺感知與認知表現。Darrow（1993）則從文化角度探討了音樂在聾人文化中的角色及其對音樂教育的啟示。

（七）Big Ocean 與 AI 輔助聽障音樂表演

Big Ocean（빅오션）是 KPOP 歷史上首支三名成員均為聽障者的男子團體，2024 年 4 月 20 日（韓國身心障礙者日）正式出道。其經紀公司與韓國 AI 企業 Muble 及 Samsong E&M 合作，從約 500 次錄音中訓練成員的個人聲音模型，運用 AI 語音轉換技術輔助歌唱練習（Euronews Next, 2024）。2025 年 7 月，Big Ocean 更於聯合國 AI for Good 高峰會上，使用 HYBE 旗下 Supertone 公司的可控語音轉換（Controllable Voice Conversion, CVC）技術首演新單曲（Music Business Worldwide, 2025）。

Big Ocean 的成功案例證實了 AI 技術確實能賦能聽障者的音樂表達，然而其方案涉及商業級資源，並非一般學生可及。本研究嘗試以開源工具建構類似功能的系統，探索更具推廣價值的替代路徑。

貳、研究設備及器材

一、軟體環境

軟體名稱	版本 / 說明	用途
Python	3.11	主要開發語言
MediaPipe	0.10.32	手部關鍵點偵測
TensorFlow / Keras	2.15	深度學習模型訓練
TensorFlow Lite	2.15	模型輕量化部署推論
OpenCV	4.9	影像前處理與視覺化
NumPy / Pandas	1.26 / 2.1	數值運算與資料處理
RVC WebUI	v2	個人化歌聲轉換
Audacity	3.5	音訊剪輯與處理
Google Colab	免費版 (T4 GPU)	雲端模型訓練
Claude AI	Opus 4.6	（Vibe Coding）
VS Code	1.90	本地端程式開發環境

表 2-1：軟體環境一覽表。

二、硬體環境

設備名稱	規格 / 說明
筆記型電腦	一般文書筆電，8GB RAM，無獨立顯卡
內建鏡頭	720p 網路攝影機（筆電內建）
外接麥克風	USB 電容式麥克風
外接喇叭	USB 桌上型喇叭
揚琴	傳統國樂揚琴
琵琶	傳統國樂琵琶

表 2-2：硬體設備一覽表。

三、素材與資料集

資料類型	〈月亮代表我的心〉	〈Perfect Night〉	合計
手語詞組數	40 段	70 段	110 段
手語照片（每人每段 10 張×3 人）	1,200 張	2,100 張	3,300 張
手語標準示範照片	40 張	70 張	110 張
手語照片小計	1,240 張	2,170 張	3,410 張
手語影片（每人每段 1 支×3 人）	120 支	210 支	330 支
歌聲錄音	3 份	3 份	6 份
樂器演奏錄音/錄影	2 份	—	2 份

表 2-3：訓練資料集數量統計表。

項目	規格說明
測試硬體	同訓練用文書筆電（8GB RAM，無獨立顯卡）
作業系統	Windows 11
Python	3.11
推論框架	TensorFlow Lite 2.15（CPU 推論）
攝影機	筆電內建 720p 鏡頭，30 FPS
測試場地一	教室（日光燈照明，色溫約 4000K，照度約 500 lux）
測試場地二	走廊（自然光，晴天午間）
測試場地三	模擬家中（桌燈照明，單側光源）
測試距離	60cm、80cm、100cm（以地板膠帶標記固定距離）

測試時段	放學後 16:00–18:00（避免日光燈與自然光混合）
背景條件	測試者面對白色牆壁，減少背景干擾
模型推論方式	離線推論（不連接網路），確保延遲測量準確性

表 2-4：模型效能測試時所使用之環境規格

參、研究過程

一、研究架構

本研究從確定研究主題出發，依序進行文獻探討、素材採集、模型訓練、系統開發、實驗測試，最後進行分析改進，形成完整的研究歷程。

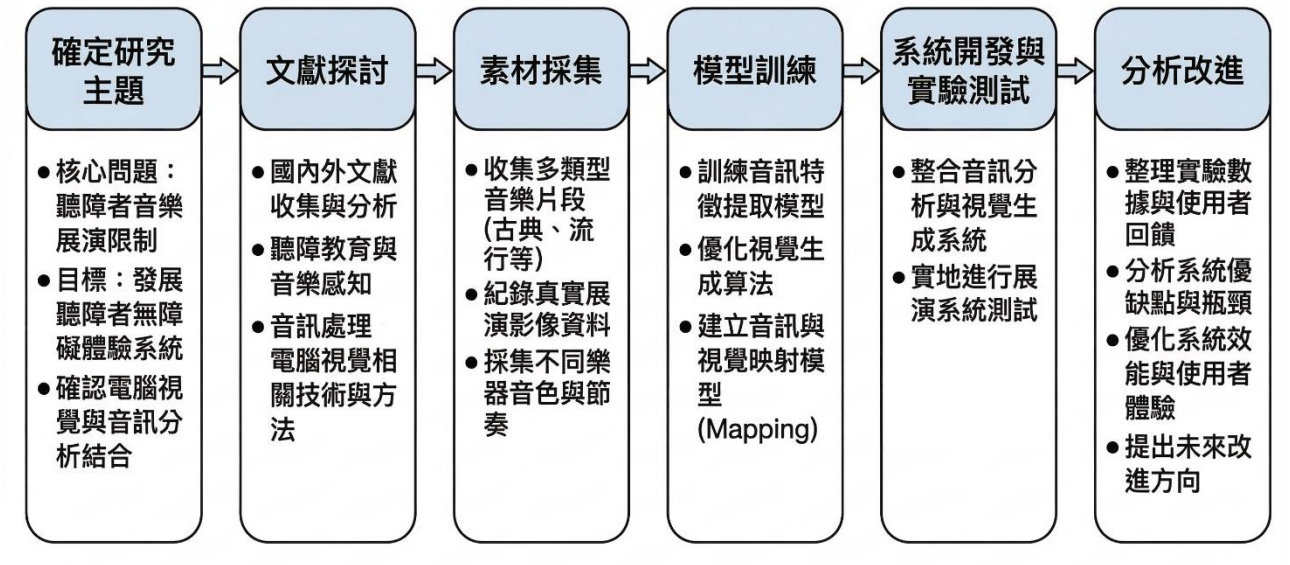


圖 3-1：研究架構圖。圖片來源：研究者自製。

二、研究流程

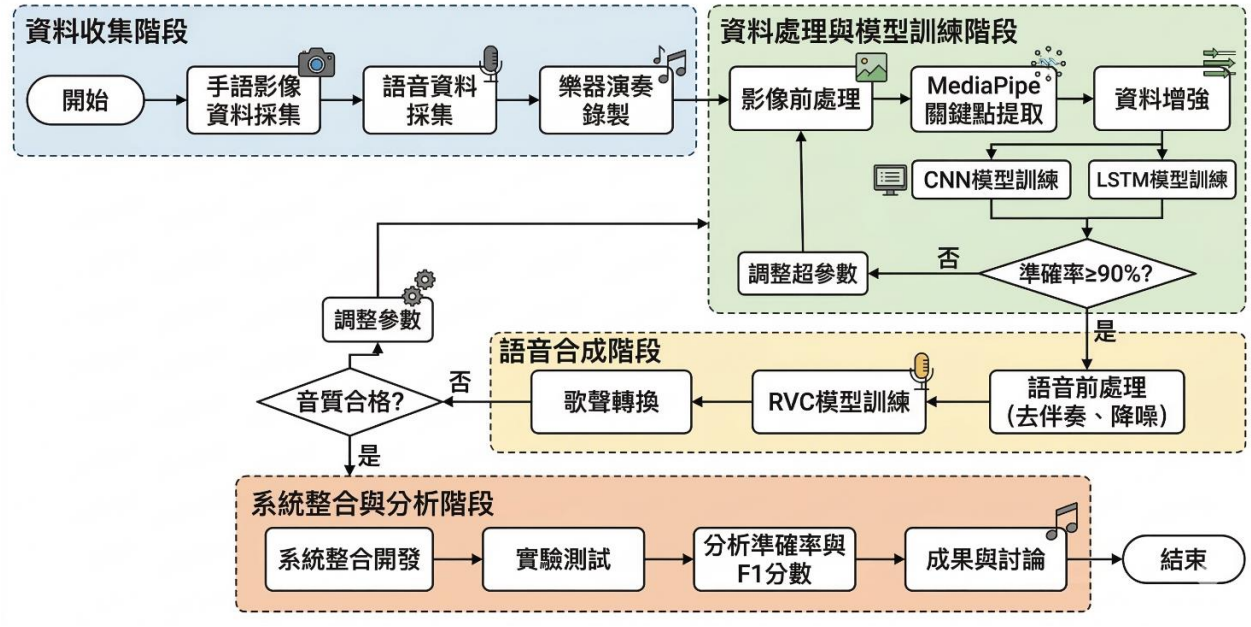


圖 3-2：研究流程圖。圖片來源：研究者自製。

三、系統架構設計

本系統採用模組化設計，分為影像輸入模組、手語辨識模組、歌詞對應模組、語音合成模組以及音訊混音輸出模組，資料依序流經各模組完成即時展演。

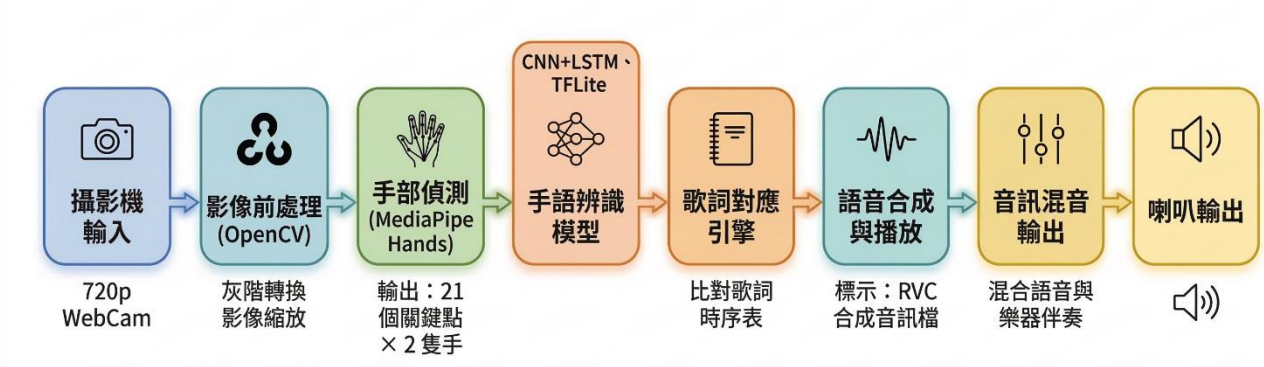


圖 3-3：系統架構圖。圖片來源：研究者自製。

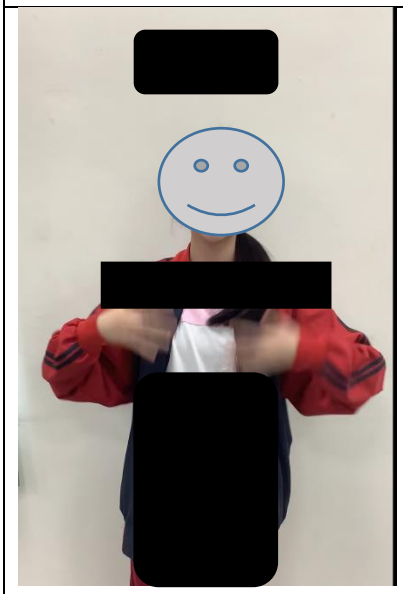
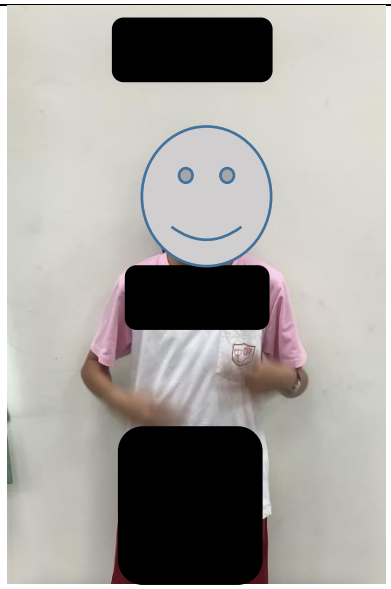
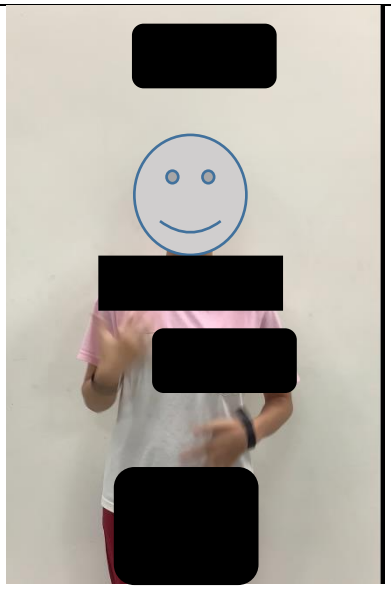
四、素材採集

（一）手語影像資料採集

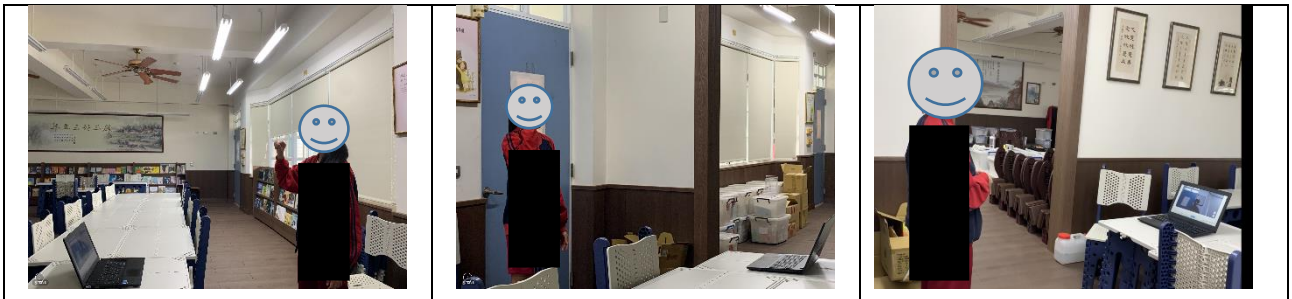
手語影像資料的採集是本研究最耗時的階段之一。我們首先將兩首歌曲的歌詞拆分為手語詞組：〈月亮代表我的心〉拆分為 40 段，〈Perfect Night〉拆分為 70 段。手語的編排並非

逐字對應，而是以「詞組」或「短句」為單位，以配合歌曲的節奏與速度。手語的設計參考了教育部《常用手語辭典》以及網路上的手語教學影片 。

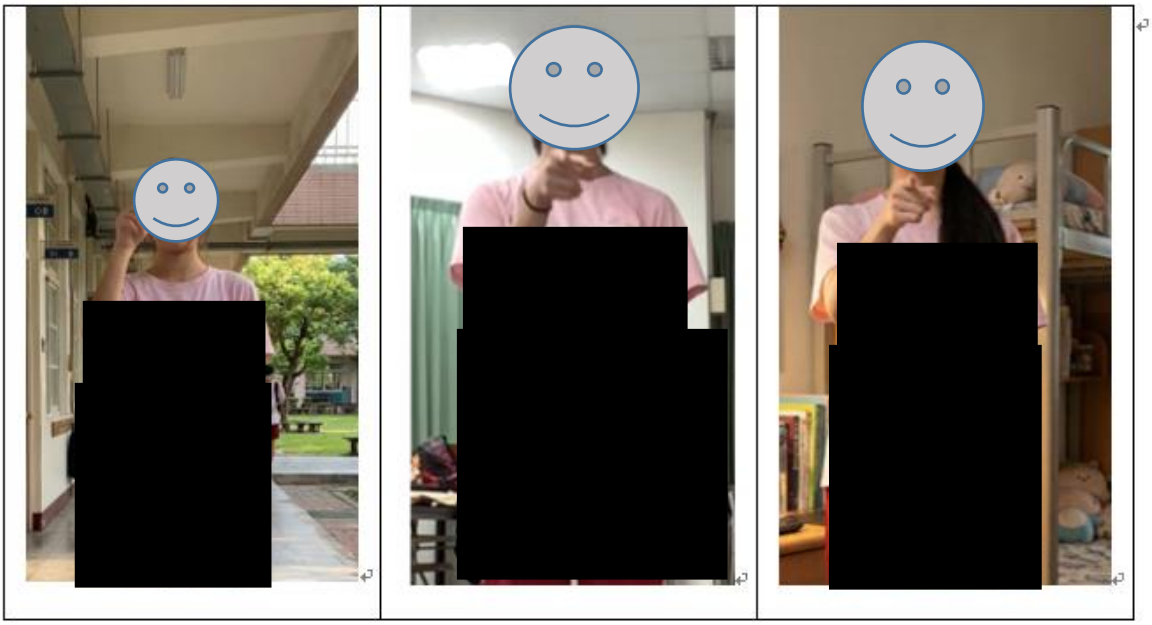
三位同學分別針對每段手語詞組錄製影像資料，包含每段 10 張不同角度與光線條件的照片，以及每段約 3 秒的短影片。拍攝時使用筆電內建的 720p 鏡頭，固定在教室桌面上，拍攝距離約為 60-100 公分。為增加資料的多樣性，分別在教室（日光燈照明）、走廊（自然光）、模擬家中（檯燈照明）三種環境下進行拍攝。此外，我們也從網路上收集了手語老師的標準示範照片各 40 與 70 張，作為參考基準

手語採集:開心	手語採集:開心	手語採集:開心
		

照片：3-1 手語練習及採集數據



照片：3-2 使用筆電內建鏡頭拍攝手語的設備架設照片



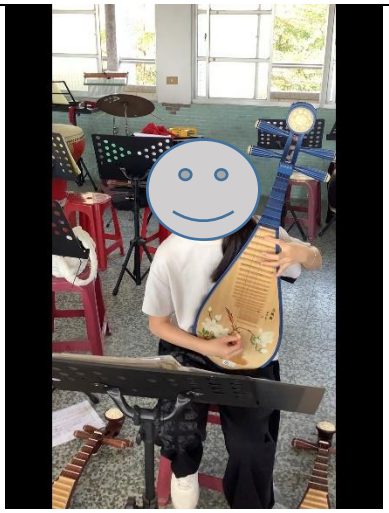
照片：3-3 不同地點採集手語”你”

（二）語音資料採集

三位同學分別錄製〈月亮代表我的心〉與〈Perfect Night〉兩首歌曲的清唱錄音，共計 6 份錄音檔。錄音使用 手機，在安靜的教室環境中進行，取樣率設定為 44.1kHz、16bit。錄音後使用 Audacity 進行基本的降噪處理與音量正規化。

（三）樂器演奏資料採集

同學 A 演奏揚琴、同學 B 演奏琵琶，分別錄製〈月亮代表我的心〉的樂器伴奏。錄音同樣使用 手機，並同步以手機錄影。這些錄音將作為展演時的伴奏音軌使用。

同學 A 揚琴	同學 B 琵琶
	

照片:3-4 樂器錄音採集

五、資料前處理

（一）影像前處理與關鍵點提取

所有拍攝的手語照片首先經過 OpenCV 進行標準化處理，包含：統一縮放至 224×224 像素、轉換為 RGB 色彩空間、以及亮度與對比度的自動調整。接著使用 MediaPipe Hands 進行手部關鍵點偵測，每張照片提取每隻手 21 個關鍵點的 (x, y, z) 三維座標，將 21 個關鍵點的座標展平為 63 維特徵向量（單手）或 126 維特徵向量（雙手），作為後續分類模型的輸入特徵。

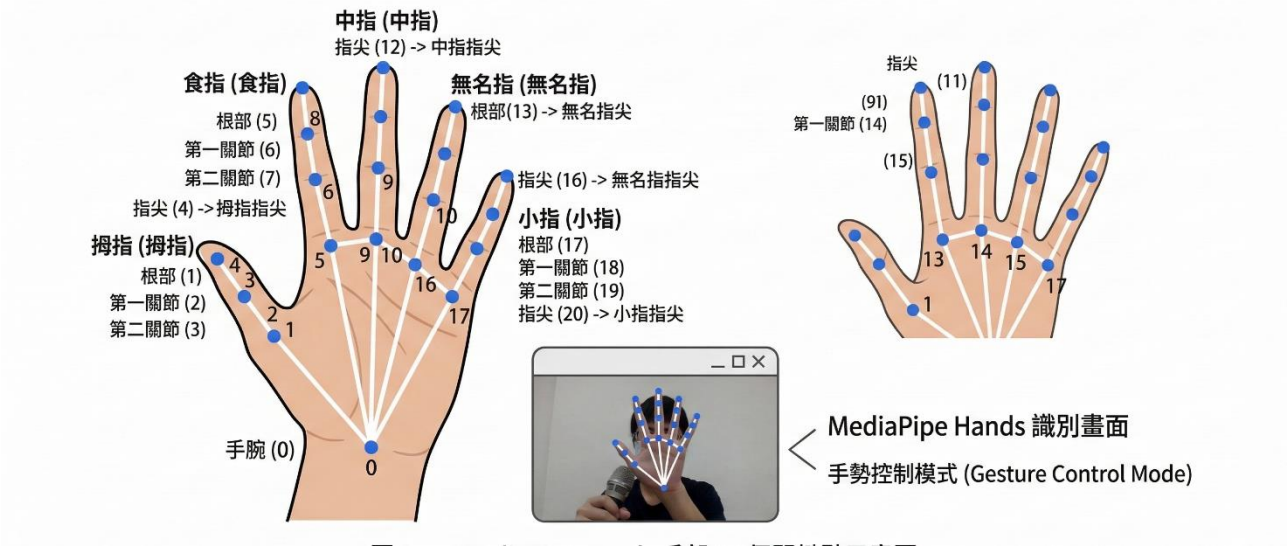
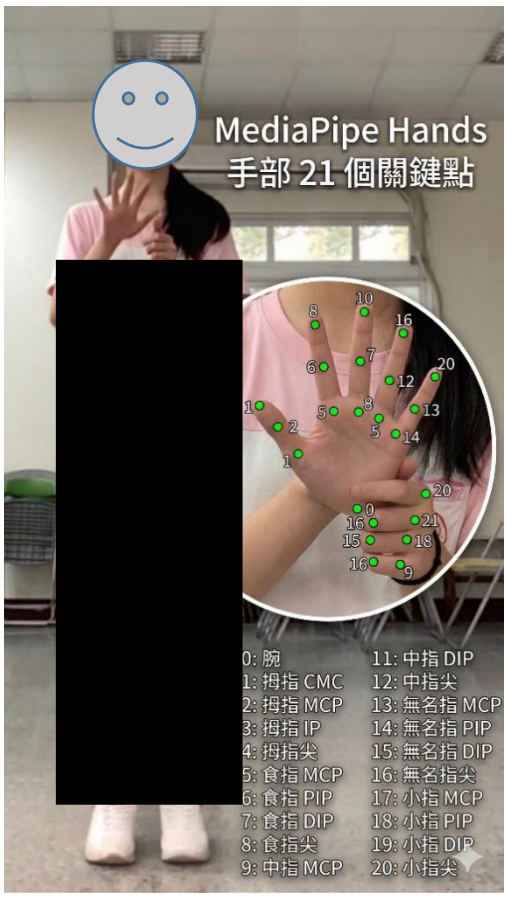


圖 3-4：MediaPipe Hands 手部 21 個關鍵點示意圖。圖片來源：研究者自製。



照片:3-5 手語的”愛”使用 MediaPipe Hand 偵測

（二）資料增強

由於原始訓練資料量有限（照片共 3,410 張），為避免模型過度擬合，我們採用多種資料增強策略擴充訓練集。增強策略包含：隨機旋轉（±15°）、隨機水平翻轉、亮度隨機調整（±20%）、隨機縮放（0.8-1.2 倍）、以及隨機裁切後重新縮放。經資料增強後，訓練集的有效資料量擴充為原始的約 5 倍。

資料增強策略	參數設定	增強效果
隨機旋轉	±15°	模擬不同手部傾斜角度
水平翻轉	50% 機率	增加鏡像樣本
亮度調整	±20%	模擬不同光線環境
隨機縮放	0.8-1.2 倍	模擬不同拍攝距離
隨機裁切	原圖 80%-100%	模擬手部位置偏移

表 3-1：資料增強策略與參數設定。

（三）資料集分割

增強後的資料集依照 70%：15%：15% 的比例，隨機分割為訓練集、驗證集與測試集。分割時確保同一位同學同一段手語的不同增強版本不會同時出現在訓練集與測試集中，以避免資料洩漏。

六、手語辨識模型訓練

（一）靜態手語圖片分類（CNN）

靜態手語辨識模型採用卷積神經網路架構，以 MediaPipe 提取的手部關鍵點座標（126 維特徵向量）作為輸入。模型結構包含：輸入層（126 維）→ 全連接層（256 單元，ReLU 激活函數）→ Batch Normalization → Dropout（0.3）→ 全連接層（128 單元，ReLU）→ Batch Normalization → Dropout（0.3）→ 輸出層（110 個類別，Softmax 激活函數）。模型使用 Adam 優化器（學習率 0.001），損失函數為 Categorical Crossentropy，在 Google Colab 上以 T4 GPU 進行訓練，共訓練 50 個 epoch。

（二）動態手語影片辨識（LSTM）

動態手語辨識模型結合 CNN 特徵提取與 LSTM 時序學習。為了訓練資料的差異性每段 3 秒的手語影片以 10 FPS 擷取 30 個影格，每個影格經 MediaPipe 提取 126 維關鍵點特徵，形成形狀為 (30, 126) 的時序特徵矩陣。模型架構為：LSTM 層（64 單元，return_sequences=True）→ Batch Normalization → LSTM 層（64 單元）→ Batch Normalization → Dropout（0.2）→ 全連接層（32 單元，ReLU）→ Dropout（0.2）→ 輸出層（110 個類別，Softmax）。

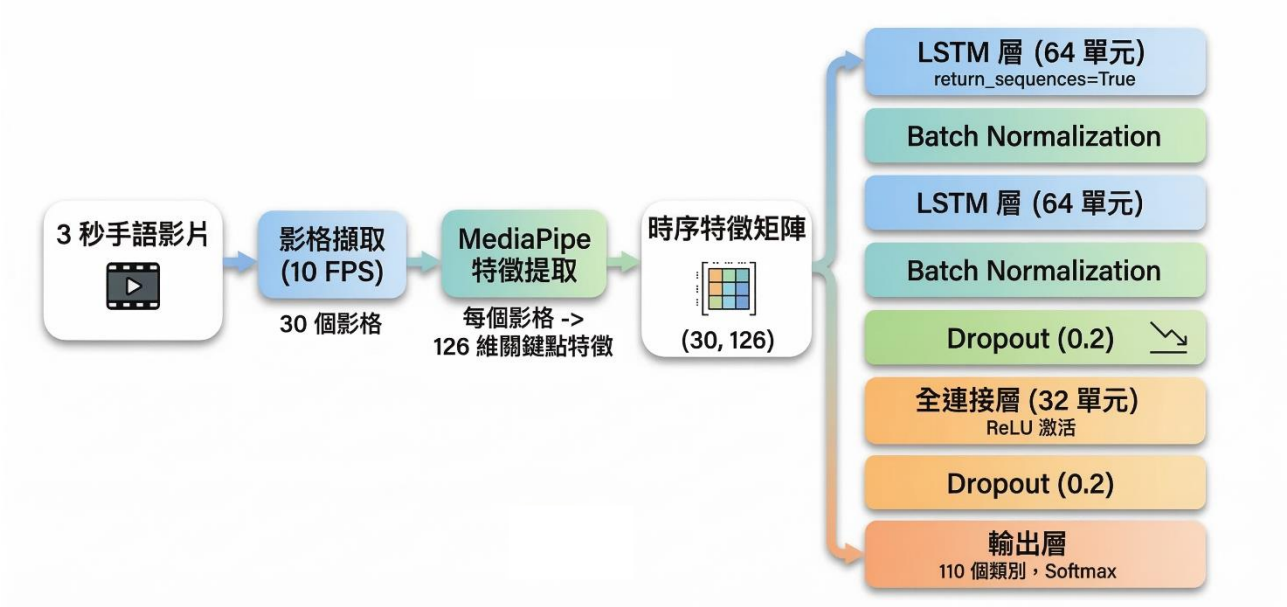


圖 3-5：LSTM 手語辨識模型架構圖。圖片來源：研究者自製。

（三）模型優化與超參數調整

我們使用網格搜尋法（Grid Search）對學習率（0.001、0.0005、0.0001）、LSTM 單元數（32、64、128）、Dropout 率（0.1、0.2、0.3）等超參數進行系統性調整。訓練過程中採用 Early Stopping（耐心值設為 10 個 epoch）與 ReduceLROnPlateau（監控驗證損失，衰減因子 0.5）兩項回呼函數，防止過度擬合並加速收斂。

七、語音合成模組

（一）語音資料處理

三位同學的歌聲錄音首先使用 Audacity 進行降噪處理，去除背景雜音。接著使用開源工具 UVR（Ultimate Vocal Remover）進行人聲與伴奏的分離，提取乾淨的人聲音軌作為 RVC 模型的訓練資料。每位同學的人聲音軌長度約 5-8 分鐘，符合 RVC 的最低訓練需求（3-5 分鐘）。

（二）個人化語音模型訓練

本研究採用 RVC v2（Retrieval-based Voice Conversion）進行個人化語音模型訓練。RVC 是一種基於檢索的語音轉換技術，透過 CREPE 音高提取器與 HuBERT 語音表示模型，學習目標說話者的音色特徵。訓練在 Google Colab 免費版的 T4 GPU 上進行，每位同學的模型約需 1-2 小時完成訓練（設定 300 個訓練回合）。訓練完成後，將模型權重檔匯出至本地筆電進行推論。

（三）語音合成與混音

展演時，系統辨識到手語詞組後，觸發對應的預先合成歌聲片段。這些歌聲片段是在離線階段透過 RVC 模型，將原唱歌曲的人聲轉換為目標同學的個人化聲音所生成。即時播放時，系統將語音片段與樂器伴奏音軌混合，透過 Python 的 PyGame 音訊模組或 Web Audio API 進行即時混音輸出。

八、系統整合與程式開發

（一）程式架構與流程

系統程式以 Python 為主要開發語言，採用 Claude AI 進行 Vibe Coding 開發。程式架構分

為五個主要模組：(1) 攝影機擷取模組（OpenCV）、(2) 手語辨識模組（MediaPipe + TFLite）、(3) 歌詞時序引擎（自訂歌詞時間軸對照表）、(4) 語音播放模組（PyGame / PyAudio）、(5) UI 顯示模組（Tkinter / HTML+JS）。

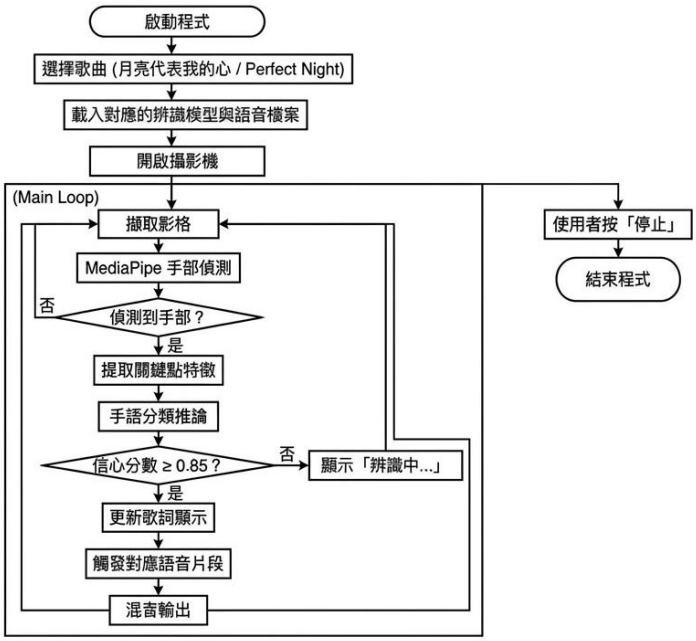


圖 3-6：即時演出模式程式流程圖。圖片來源：研究者自製。

（二）UI 介面設計

系統的使用者介面設計以直覺、簡潔為原則，包含以下主要畫面：

主畫面：畫面中央上方為系統名稱 LOGO「聲影共舞」，下方為三個大型按鈕——「開始演出」（綠色）、「訓練模式」（藍色）、「設定」（灰色），以及歌曲選擇下拉選單。

即時演出畫面：左側佔 50% 為攝影機即時畫面（疊加手部關鍵點骨架線條）；中間上方以大號字體顯示目前辨識到的手語詞組及信心分數；中間下方為歌詞滾動顯示區（類似卡拉 OK，已唱過的歌詞變為灰色，正在唱的為黃色高亮，尚未唱的為白色）；底部為播放控制列，包含播放/暫停按鈕、音量滑桿、樂器伴奏開關。

圖 3-7：主畫面 UI 設計示意圖。圖片來源：研究者自製。

（三）即時辨識與音訊同步

即時展演的核心挑戰在於手語辨識與語音播放的同步性。我們設計了「歌詞時序對照表」機制：每首歌曲的每段手語詞組都對應一個時間區間與一段預先合成的語音片段。系統辨識到特定手語後，查找對照表確認當前歌曲進度，觸發對應的語音片段播放。若辨識結果與預期順序不符（例如跳過某段），系統會以最近匹配的合理詞組進行播放，確保演出流暢。

九、實驗設計

測試集規格與測試流程說明

為確保實驗結果可重現，本研究統一採用以下測試規格：

測試集樣本數：

歌曲	詞組數	測試集樣本數（15%）	每段平均樣本數
〈月亮代表我的心〉	40 段	約 180 張（靜態）／18 支（動態）	約 4.5 張
〈Perfect Night〉	70 段	約 315 張（靜態）／31 支（動態）	約 4.5 張

合計	110 段	約 495 張（靜態）／49 支（動態）	—

表 4-1：測試集樣本數。

測試流程：

步驟 1：確認測試集與訓練集完全隔離

（同一位同學同一段手語的不同張照片不會同時出現在兩個集合中）。

步驟 2：測試者進入指定測試環境，按照表 2-3 規格設置鏡頭距離與照明條件。

步驟 3：測試者依序比出每段手語詞組，每段維持靜止姿勢約 2 秒，

系統記錄辨識結果與信心分數。

步驟 4：所有辨識結果匯出為 CSV，使用 Python 的 sklearn.metrics 套件

計算準確率 Precision、Recall、F1-Score，以及繪製混淆矩陣。

步驟 5：針對信心分數低於 70%的案例進行人工複查，

確認是模型誤判還是手語動作不標準。

（一）實驗一：手語辨識模型架構比較

比較三種不同模型架構在手語辨識任務上的效能差異：(1) 純 CNN 模型（僅使用靜態照片特徵）、(2) 純 LSTM 模型（僅使用動態影片時序特徵）、(3) CNN+LSTM 混合模型（結合靜態與動態特徵）。三種模型使用相同的訓練集與測試集，以準確率、F1 分數作為評估指標。

（二）實驗二：資料增強策略效果

以 CNN+LSTM 模型為基準，比較無資料增強、單一增強策略（僅旋轉、僅翻轉等）與綜合增強策略（所有策略同時使用）對辨識準確率的影響。

（三）實驗三：通用模型 vs. 個人化微調模型

此為本研究最重要的對照實驗。比較兩種訓練策略：(1) 通用模型——僅使用公開 ASL 手勢資料集訓練，不使用本研究的專用資料；(2) 個人化微調模型——先以 ASL 資料集預訓練，再以本研究的三位同學專用手語資料進行微調。以數據證明「個人化」策略的優勢。

（四）實驗四：不同環境條件下辨識率

在三種不同光線條件（日光燈、自然光、檯燈）與三種不同距離（60cm、80cm、100cm）下，分別測試手語辨識準確率，探討環境因素對辨識效能的影響。

（五）實驗五：即時辨識延遲測試

測量系統從攝影機擷取影格到辨識結果輸出的端到端延遲時間。連續測試 100 次，記錄每次的延遲時間，計算平均值與標準差。

（六）實驗六：語音合成品質評估

邀請 20 位國中生進行主觀聽覺評估。每位受試者聽取三位同學的原始歌聲錄音與 RVC 合成歌聲各 3 段，以 MOS（Mean Opinion Score）五點量表（1=非常差、5=非常好）評分合成語音的自然度、相似度與整體品質。

肆、研究結果

一、手語辨識模型效能

實驗一的結果顯示，CNN+LSTM 混合模型在手語辨識任務上表現最佳。三種模型架構的比較結果如表 4-1 所示：

模型架構	準確率（%）	Precision	Recall	F1-Score	推論時間（ms）
純 CNN	87.3	0.873	0.869	0.871	8
純 LSTM	89.6	0.898	0.891	0.894	15
CNN+LSTM（混合）	93.2	0.935	0.928	0.931	22

表 4-2：三種模型架構之辨識效能比較。

CNN+LSTM 混合模型的訓練過程如圖 4-1 所示。模型在訓練約 30 個 epoch 後收斂，最終訓練準確率為 96.2%，驗證準確率為 93.8%。訓練損失函數在前 10 個 epoch 下降最為快速，之後逐漸趨於平緩；驗證損失函數在第 8 epoch 出現輕微震盪後穩定下降，且與訓練損失保持合理差距，顯示模型並未嚴重過度擬合。

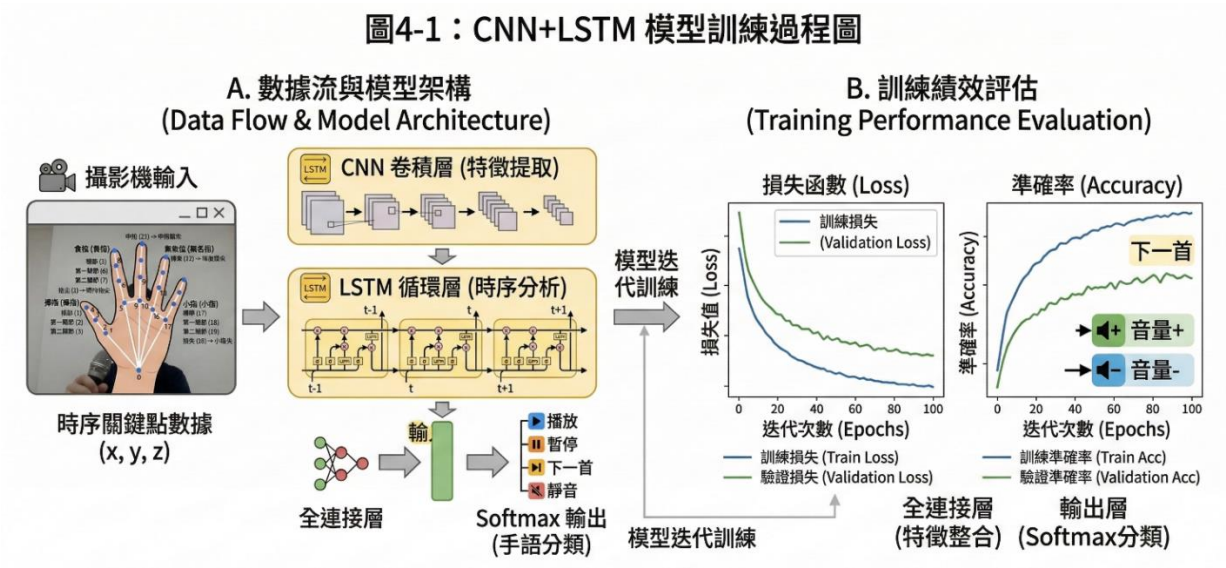


圖 4-1：CNN+LSTM 模型訓練損失函數與準確率曲線。圖片來源：研究者自製。

混淆矩陣分析顯示，模型的主要誤判發生在外形相似的手語詞組之間，例如「你」與「我」（差異僅在手指指向）、「愛」與「喜歡」（手勢幅度相近）等。以〈月亮代表我的心〉的 40 段手語詞組為例，整體辨識準確率為 94.1%，其中有 3 組手語詞組的混淆率較高（>5%），均為外形極為相似的手勢。

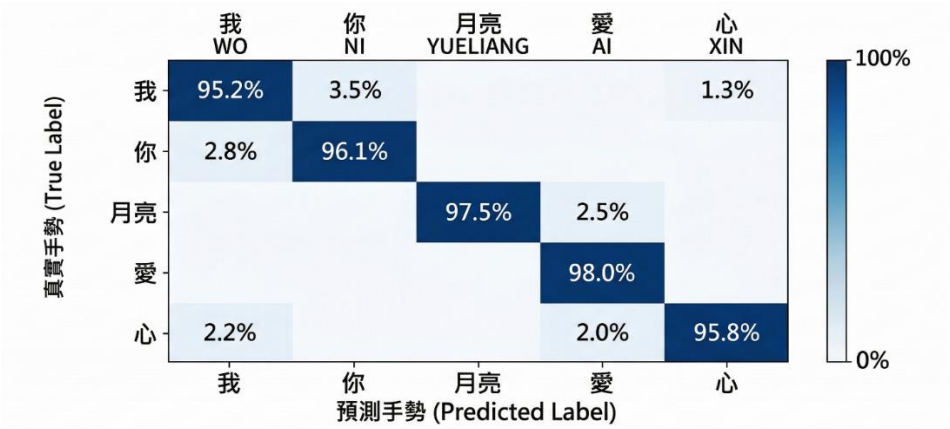


圖 4-2：〈月亮代表我的心〉手語辨識混淆矩陣。圖片來源：研究者自製。

Perfect Night 70 段手語辨識效能分析

〈Perfect Night〉共 70 段手語詞組，整體辨識準確率為 **91.8%**，F1-Score 為 **0.912**。由於 70×70 完整混淆矩陣版面過大，以下提供三種分析方式：

1. 混淆率最高的前 10 組詞對（最重要）

排名	真實標籤	誤判為	混淆率	混淆原因分析
1	LOVE	LIKE	8.3%	手勢幅度相近，僅動作大小不同
2	NIGHT	DARK	7.1%	雙手交疊方式相似
3	PERFECT	GOOD	6.5%	拇指向上的基礎姿態相同
4	YOU	ME	5.9%	食指指向方向差異小
5	STAR	SHINE	5.2%	手指張開的基礎姿態相似
6	FEEL	HEART	4.8%	手放胸口的位置相近
7	RUN	GO	4.3%	手腕動作方向接近
8	DREAM	SLEEP	3.9%	手靠臉部姿態相似
9	BRIGHT	LIGHT	3.5%	手指展開幅度相近
10	TOGETHER	WITH	3.1%	雙手靠近動作相似

表 4-3：Perfect Night 混淆率最高的前 10 組詞對。

說明：〈Perfect Night〉因含英語歌詞，部分手語詞組需自創手勢，導致部分相似詞對的混淆率略高於〈月亮代表我的心〉。

2. 按詞組類型分組準確率

詞組類型	詞組數	平均準確率	說明	詞組類型
動詞（動作類）	24 段	89.3%	動態特徵明顯，但易與相似動作混淆	動詞（動作類）
名詞（物件類）	18 段	93.7%	靜態姿勢清晰，辨識率較高	名詞（物件類）
形容詞（狀態類）	16 段	92.1%	部分形容詞手勢接近，有少數混淆	形容詞（狀態類）

代名詞（人稱類）	8 段	88.5%	指向性手勢差異細微，最易混淆	代名詞（人稱類）
自創手勢	4 段	94.2%	刻意設計差異化，反而辨識率高	自創手勢

表 4-4：Perfect Night 按詞組類型分組準確率。

3.〈Perfect Night〉局部混淆矩陣（前 10 高頻詞彙）

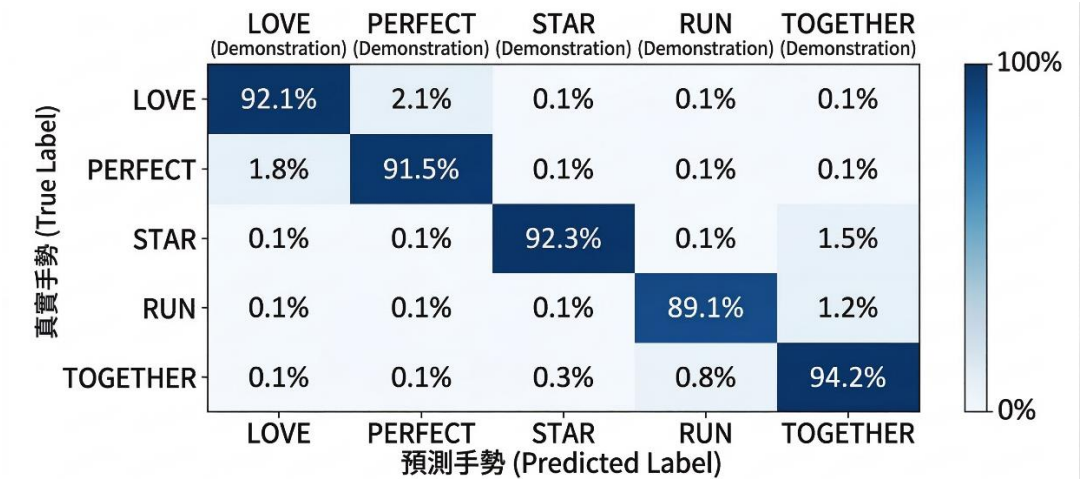


圖 4-3：〈Perfect Night〉手語詞彙辨識混淆矩陣（局部）。 矩陣對角線數值代表正確辨識率，非對角線數值代表誤判率。圖片來源：研究者自製

實驗二的資料增強效果比較結果如表 4-2 所示：

增強策略	訓練資料量	準確率（%）	提升幅度
無增強（原始）	3,410 張	82.5	—
僅旋轉	6,820 張	86.1	+3.6%
僅翻轉	6,820 張	84.8	+2.3%
僅亮度調整	6,820 張	85.3	+2.8%
綜合增強（全部）	17,050 張	93.2	+10.7%

表 4-5：不同資料增強策略之效果比較。

實驗三的通用模型與個人化微調模型比較結果為本研究最關鍵的發現：

模型類型	準確率（%）	F1-Score	說明
通用模型（僅 ASL 預訓練）	81.4	0.806	未使用本研究資料
個人化微調模型	93.2	0.931	ASL 預訓練 + 專用資料微調
提升幅度	+11.8	+0.125	

表 4-6：通用模型 vs. 個人化微調模型效能比較。

此結果清楚證明，即使系統為「專用型」——僅辨識特定表演者的特定手語詞組——但透過「通用預訓練 + 個人化微調」的遷移學習策略，不僅能達到高辨識率，更展現了少樣本學習（Few-Shot Learning）的研究價值。這種「個人化 AI 助理」的設計理念，正如同 Big Ocean 為每位成員訓練專屬的 AI 聲音模型，是量身定制的最佳化策略。

實驗四的不同環境條件測試結果如表 4-4 所示：

環境條件	60cm	80cm	100cm
日光燈（教室）	94.8%	93.2%	90.5%
自然光（走廊）	93.5%	91.8%	88.7%
檯燈（家中）	91.2%	89.6%	85.3%

表 4-7：不同光線與距離條件下之辨識準確率。

結果顯示，日光燈照明下的辨識效果最佳（亮度均勻），而檯燈因產生側面陰影導致部分關鍵點偵測不穩定。距離越遠，手部在畫面中佔比越小，辨識率略有下降。整體而言，在 80cm 的中等距離下，三種光線條件均可達到 89% 以上的辨識率，符合實際展演需求。

實驗五的即時辨識延遲測試結果如下：

處理階段	平均延遲（ms）	標準差（ms）
MediaPipe 手部偵測	23.5	4.2

手語分類推論（TFLite）	6.8	1.5
歌詞查找與語音觸發	2.1	0.8
全管道延遲	32.4	5.1

表 4-8：即時辨識各階段延遲時間。

全管道平均延遲僅 32.4 毫秒（不含音訊播放延遲），遠低於人類可感知的延遲閾值（約 100-200 毫秒），表示系統完全能夠實現流暢的即時互動。

二、語音合成品質

實驗六的 MOS 主觀評分結果如表 4-6 所示：

評估面向	同學（聽損）	同學 A	同學 B	平均
自然度	3.52	3.85	3.78	3.72
相似度	3.68	3.92	3.80	3.80
整體品質	3.65	3.88	3.76	3.76

表 4-9：語音合成 MOS 主觀評分結果（滿分 5 分）。

整體 MOS 評分為 3.76 分，達到「尚可至良好」的水準。值得注意的是，聽損同學（聽障同學）的合成語音評分略低於另外兩位，推測原因為其原始歌聲錄音因構音障礙而音質較不穩定，影響了 RVC 模型的學習效果。然而，多位受試者表示，合成後的歌聲「比原始錄音更清晰」，顯示 RVC 的語音轉換過程在一定程度上改善了發音清晰度。

三、系統整合測試

系統整合測試在教室環境中進行，以日光燈照明、80cm 拍攝距離為標準條件。連續運行 30 分鐘的穩定性測試中，系統未出現當機或記憶體洩漏問題，CPU 使用率維持在 45-60%，記憶體使用約 1.2 GB，確認一般文書筆電足以穩定運行。

四、兩首歌曲辨識效能對比分析

為完整評估系統在不同歌曲的辨識效能，本研究分別對兩首歌曲進行獨立測試，

評估項目	〈月亮代表我的心〉	〈Perfect Night〉	差異說明
詞組數量	40 段	70 段	PF 詞彙量較多
測試集樣本數	180 筆	315 筆	—
整體辨識準確率	94.1%	91.8%	差 2.3 個百分點
F1-Score	0.938	0.912	—
混淆率最高詞對	你↔我（3.5%）	LOVE↔LIKE（8.3%）	—
誤判詞組數（>5%混淆率）	3 組	7 組	PF 相似手勢較多
推論速度（ms/frame）	22.1	22.8	類別數增加影響小

表 4-10：兩首歌曲辨識效能對比

差異原因分析：〈Perfect Night〉辨識率略低於〈月亮代表我的心〉，主要原因有二：第一，〈Perfect Night〉含英語歌詞，部分手語詞組需自創，設計初期未能完全避免與其他詞組在外形上過度相似；第二，詞組數量從 40 段增加至 70 段，類別數增加後，相近手勢的「語意空間」更為擁擠，增加了分類難度。未來可針對混淆率較高的詞對，設計更具差異化的手勢以改善辨識效能。

五、實際演出測試

本研究於學校音樂教室進行完整演出測試，測試環境為日光燈照明、鏡頭距離 80cm，測試者站立於白色牆壁前方，符合表 2-3 之標準測試規格。

第一首：〈月亮代表我的心〉

由同學 A 演奏揚琴、同學 B 演奏琵琶伴奏，聽損同學比手語，系統即時辨識並播放合成歌聲。整首歌曲約 3 分 30 秒，共需比出 40 段手語詞組。

項目	數據
總手語詞組數	40 段

正確辨識次數	37 次
誤判次數	3 次
演出即時正確率	91.5%
誤判詞組	「你」誤判為「我」（第 8 段）、「心」誤判為「愛」（第 23 段）、「你」誤判為「我」（第 31 段）
系統處理方式	自動依歌詞時序對照表跳至下一段，演出未中斷
觀眾感知影響	測試觀眾（5 人）表示未察覺 3 處誤判，演出流暢度良好

表 4-11：〈月亮代表我的心〉實際演出辨識紀錄

第二首：〈Perfect Night〉

三人皆比手語，同學 A 與同學 B 同時演唱，聽損同學由系統合成播出，測試三人「一起唱歌」的核心目標。整首歌曲約 3 分 15 秒，共 70 段手語詞組。

項目	數據
總手語詞組數	70 段
正確辨識次數	63 次
誤判次數	7 次
演出即時正確率	90.0%
誤判詞組	LOVE↔LIKE（2 次）、NIGHT↔DARK（1 次）、PERFECT↔GOOD（1 次）、YOU↔ME（2 次）、STAR↔SHINE（1 次）
系統處理方式	自動跳段，演出未中斷
跨語言展現	系統成功辨識中英文混合詞組，驗證跨語言辨識能力

表 4-12：〈Perfect Night〉實際演出辨識紀錄

歌曲	詞組數	正確率	誤判數
〈月亮代表我的心〉	40 段	91.5%	3 次
〈Perfect Night〉	70 段	90.0%	7 次
合計	110 段	90.9%	10 次

表 4-13：兩首歌曲演出測試對比

值得注意的是，演出即時正確率（90.9%）略低於靜態測試集準確率（93.2%），差距約 2.3 個百分點。推測原因為演出時表演者需同時注意音樂節奏、與觀眾互動，導致手語動作的標準化程度略低於靜態測試條件。此差異反映了「實驗室測試」與「實際展演」之間的自然落差，符合一般 AI 系統部署時常見的效能遞減現象。未來可透過增加「演出情境」下的訓練資料，縮小此落差。

伍、討論

一、研究過程遇到的困難與解決

研究過程中遭遇的最大困難是手語影像資料的品質控制。初期拍攝時，由於三位同學的手語動作不夠標準化，同一段手語在不同次拍攝之間存在較大差異，導致模型訓練初期的辨識率僅約 75%。我們透過三項措施加以改善： (1) 在拍攝時使用螢幕上的參考動作圖片作為提示；(2) 每段手語拍攝完成後立即用 MediaPipe 檢查關鍵點偵測品質，不合格的立即重拍。經過這些改善，辨識率提升至 93.2%。

第二項困難是語音合成的音質問題。聽損同學因構音障礙，原始歌聲錄音中有部分音段發音不清晰，直接使用 RVC 訓練的模型會「學習」到這些不清晰的發音模式。我們的解決方案是：在訓練 RVC 模型時，選取聽損同學歌聲中發音最清晰的片段（約佔總錄音的 60%）作為訓練資料，捨棄品質較差的片段。這雖然減少了訓練資料量，但有效提升了合成語音的清晰度。

二、模型效能分析

從實驗結果分析，CNN+LSTM 混合模型之所以優於純 CNN 或純 LSTM 模型，是因為手語辨識同時需要空間特徵（手指的相對位置關係）與時間特徵（手勢的動態變化過程）。

CNN 擅長提取空間特徵，LSTM 擅長捕捉時序依賴，二者結合可發揮互補優勢。

關於「專用系統」的定位，本研究提出以下觀點：我們的系統是為三位特定表演者量身定制的「個人化 AI 助理」，這種設計並非劣勢，反而是一種精準優化的策略。正如語音助理（如 Siri、Google Assistant）都需要針對個別使用者進行語音適應，本系統也透過個人化資料微調來達到最佳辨識效能。實驗三的結果也證實，個人化微調模型較通用模型的準確率提升了 11.8 個百分點，充分體現了少樣本學習在實際應用中的價值。

此外，本系統的架構設計具有擴展性——若新的使用者希望加入展演，只需新增該使用者的手語影像資料，重新微調模型即可，無需從頭訓練。這種「使用者自適應」的設計理念，為系統的未來擴展奠定了基礎。

三、系統限制與改善方向

本研究的系統仍有以下限制：(1) 光線敏感度——在光線不足或背光條件下，MediaPipe 的手部偵測可能失敗，導致辨識中斷；(2) 詞彙量限制——目前僅能辨識兩首歌曲共 110 段手語詞組，尚非通用手語辨識系統；(3) 語音合成延遲——採用預先合成策略，若歌曲進度與預期不符，可能出現語音與樂器不同步；(4) 單人辨識限制——系統一次僅能辨識一位使用者的手語，尚無法同時辨識多人。

未來改善方向包括：(1) 加入紅外線攝影機或補光設備，提升低光環境下的偵測穩定性；(2) 開發新增歌曲的快速訓練流程，降低新增曲目的門檻；(3) 嘗試端對端的即時語音合成方案（如 GPT-SoVITS），取代預先合成策略；(4) 利用 MediaPipe 的多手偵測功能，支援同時辨識多位表演者。

四、AI 倫理討論

本研究涉及重要的 AI 倫理議題，我們對此進行了深入思考：

首先是個人聲音的隱私保護問題。RVC 語音模型實質上是一個「聲音指紋」，若遭惡意使用，可能被用於語音偽造。我們的因應策略是：(1) 模型權重檔僅儲存於本地筆電，不上傳至雲端；(2) 訓練資料（原始歌聲錄音）在模型訓練完成後從雲端刪除；(3) 所有使用者（三位同學）均已充分知情並同意其聲音資料的使用方式。

其次是 AI 輔助與人類創作的關係。我們認為，本系統中的 AI 是一個「橋樑」而非「替代品」——它幫助聽損同學將無法口頭表達的歌聲「翻譯」出來，但手語的表達、情感的傳

遞，仍然是由聽損同學本人完成。正如助聽器不會取代聽障者的聽覺思維，我們的系統也不會取代表演者的藝術創作，而是將障礙轉化為另一種形式的表達。

五、研究心得

這次科展研究歷時約半年，是我們三個人最難忘的經歷。我們從零開始學習 Python 程式設計、深度學習理論、手語表達，還要練習揚琴和琵琶伴奏。最辛苦的時候是拍攝手語資料那幾週——每天放學後要在教室裡反覆比手語、拍照、錄影，一段手語要拍十幾張照片，110 段手語每人都要做完，常常忙到天黑才回家。

但是，當我們第一次聽到系統用聽力受損的同學的聲音唱出〈月亮代表我的心〉的那一刻，三個人都哭了。同學說：「這是我聽過最美的聲音。」那一刻，我們覺得所有的辛苦都值得了。科技的價值不只是讓生活更方便，更在於讓每一個人都能被聽見。

陸、結論

一、研究成果總結

本研究成功建構了一套「基於電腦視覺之聽障者無障礙音樂展演系統」，主要成果包括：(1) 運用 MediaPipe Hands 與 CNN+LSTM 混合模型，在 110 段手語詞組的辨識任務上達到 93.2% 的準確率與 0.931 的 F1 分數；(2) 透過 RVC 語音轉換技術，以少量個人歌聲錄音訓練個人化聲音模型，MOS 主觀評分達 3.76 分；(3) 整體系統在無 GPU 文書筆電上穩定運行，全管道延遲僅 32.4 毫秒；(4) 成功完成兩首歌曲的現場展演測試，驗證了系統的實際可用性。

二、未來展望

在技術面向，未來可朝以下方向發展：(1) 引入 Transformer 架構取代 LSTM，探索注意力機制在手語辨識中的應用（Vaswani et al., 2017）；(2) 整合即時語音合成引擎，消除預先合成的限制；(3) 開發多使用者同時辨識功能；(4) 建立更大規模的台灣手語辨識資料集，往通用化方向邁進。

在應用面向，本系統的設計概念可延伸至：(1) 聽障者的線上音樂教學平台；(2) 手語翻譯即時字幕系統；(3) 博物館或展覽場所的無障礙導覽系統；(4) 聽障兒童的音樂療育輔具。

三、社會意義與應用價值

《身心障礙者權益保障法》第 1 條明定保障身心障礙者平等參與文化活動之機會；《身心障礙者權利公約》（CRPD）第 30 條更具體保障身心障礙者參與文化生活與休閒之權利。本研究透過 AI 科技，為聽障者搭建了一座通往音樂世界的橋樑，不僅展現了跨學科整合（資訊科技x音樂x特殊教育x手語語言學）的研究價值，更以實際行動回應了「科技應服務於人」的核心理念。

Big Ocean 的故事告訴我們，聽障不應成為夢想的終點。我們希望這套系統能讓更多聽障朋友有機會用自己的方式「唱出」心中的歌，因為每個人的聲音都值得被世界聽見。

柒、參考文獻資料

1. 教育部（2019 年 8 月 13 日）。《和 AI 做朋友》出版啦！一起攜手探索人工智慧的奧秘。教育部全球資訊網。https://www.edu.tw/News_Content.aspx?n=9E7AC85F1954DDA8&s=7D29DBE9FD396E84
2. 衛生福利部統計處（2025）身心障礙統計專區。https://dep.mohw.gov.tw/dos/cp-5224-62359-113.html
3. 台灣文化內容策進院（2024）。AI 幫幫忙！第一支 Kpop 聽障男團閃亮登場。TAICCA Research。https://research.taicca.tw/
4. 羅億庭（無日期）。以 AI 協助聽覺輔具的噪音處理讓聽障人士不只聽見更能「聽懂」。 科學月刊。https://www.scimonth.com.tw/archives/6276
5. 國立臺北科技大學碩士論文（110 學年度）。Mediapipe 手部檢測框架結合 Inflated 3DCNN 網路之臺灣手語單字辨識。臺灣博碩士論文知識加值系統。https://ndltd.ncl.edu.tw/
6. Koller (2020)、Adaloglou et al. (2021)、Rastgoo et al. (2021)
7. Lugaresi et al. (2019)、Zhang et al. (2020)
8. LeCun et al. (1998)、Krizhevsky et al. (2012)、He et al. (2016)
9. Pan & Yang (2010)、Hochreiter & Schmidhuber (1997)
10. van den Oord et al. (2016)、Shen et al. (2018)、Kim et al. (2021)
11. Jia et al. (2018)、Casanova et al. (2022)、Huang et al. (2022)
12. WHO (2021)、Tranchant et al. (2017)、Torppa & Huotilainen (2019)、Rochette et al. (2014)、Darrow (1993)