基於開源軟體部署的即時聲音轉台灣手語服務

指導教授：熊博安教授

專題學生：王俐晴、洪子翔

目錄：

（一）摘要

語音轉文字(Speech-to-Text)、自然語言處理(Natural Language Processing)以及計算機視覺(Computer Vision)等技術，早已是資訊領域的重要研究方向。然而，在台灣，能夠即時將語音翻譯成台灣手語(Speech-To-Sign in Taiwan)的系統尚未普及。為此，我們認為，結合這些技術打造一個促進資訊傳達的服務，將是一項極具發展潛力且深具社會價值的研究方向。

隨著社會對資訊平權的重視逐年提升，重大記者會中逐漸安排了即時手語翻譯人員，讓聽障人士即便無法聽到聲音，也能透過手語理解資訊。然而，手語翻譯的服務多集中於特定場合，面對日常生活中大量的影音內容，聽障人士獲取資訊的權利依然受到諸多限制。因此，一套專為台灣設計的即時語音翻手語服務，將有效彌補現有的資訊傳遞缺口，進一步促進資訊平權的實現。

我們計畫建立一個完整的 AI model pipeline，其主要流程包括將語音轉為文字，經由 Encoder-Decoder 模組進行語法分析與翻譯處理，生成符合台灣手語語法的動作序列，再以影像動畫的方式呈現翻譯結果。為了提升系統的靈活性與可用性，我們採用 Kserve 部署該模型至平台，利用其動態部署特性實現彈性流量處理，打造一個易於使用、資源妥善分配，且能長期維護與升級的開放式服務平台。

透過這個即時服務，能讓聽障朋友能在任何時間、任何地點使用手語翻譯，無需等待特定場合或專人協助。除此之外，我們的系統不僅能用於日常溝通場景，也適用於教育、醫療、公共服務等多種場合，進一步擴展了其應用範圍。

截至 2019 年，全台約有 12 萬聽覺機能障礙者(數據來源：衛生福利部統計處)，而視覺是他們生活的最大輔助。這項服務不僅能填補現有資訊傳遞上的缺口，更能實現真正的資訊平等，讓每一個人都能享有無障礙的溝通與理解權利。

（二）研究動機與研究問題

手語是聽障人士的主要溝通工具，而每個國家的手語都有其獨特的語法、詞彙及文化背景。例如，台灣手語（TSL）具有不同於美國手語（ASL）的語序與表達特徵。然而，現有的手語研究多集中於手語辨識（Sign Language Recognition, SLR），即從影片中辨識手語內容，並將其轉換為文字或語音(Sign-to-Speech)。雖然這些研究對於大眾理解手語有所幫助，但反向的語音轉手語（Speech-to-Sign）應用則相對匱乏，

二者對比如表一所示。

表一 雙向轉換對比 口語時態問題對中文的影響較少

情境	Speech-to-Sign	Sign-to-Speech
上下文處理	多義詞問題、補充非手勢特徵	透過語意補充口語時態、修飾語等
技術難點	手語動畫的連貫性、手語語料稀缺	除手勢動作需精確識別外 還有面部表情、視線、姿態
技術成熟度	較慢，非手勢特徵易、造成動畫不連貫、客群較小	相對成熟，電腦視覺的發展也會推進手語辨識
實時性	動畫生成延遲較大	語音合成延遲較小
輸出結果	流暢的動作序列 需額外考量速度、方向	簡單的文字語句、僅需考慮文法問題

此外，國際語言間的隔閡同樣存在於手語上。市面上的手語翻譯系統大多綁定於特定語言，且研究多著重在美國手語（ASL），缺乏對其他手語系統的支持。這種語言綁定限制導致現有技術無法滿足台灣聽障群體的需求。台灣手語擁有豐富的詞彙和獨特的語法結構(如圖一 a.b.與圖二所示)，而現有技術在語法處理、語意翻譯和手語動作生成方面缺乏針對性，翻譯結果往往不準確或不流暢，無法達到實際應用的水準。

圖一 a.b. 「灰色的」中文手語，僅需一隻手即可表達[13]

圖二 「grey」英文手語，需雙手才能完整表達[13]

因此，本研究希望從語音到手語翻譯(Speech-to-Sign)的角度切入，開發一個即時語音轉台灣手語的系統，突破現有技術在語言綁定上的局限，並聚焦於台灣手語的

語法結構與本地化特性，填補技術空白，以推動台灣手語的普及和應用。

（三）文獻回顧與探討

我們以技術為分類，總結現有方法的特徵及優缺點如下：

1. 靜態手語生成技術

- Prillwitz 提出一套無語言依賴的符號系統 HamNoSys 來描述手語的細節特徵，可以精準的表示手型、動作、位置、方向等屬性；由於圖形具高抽象性，學習門檻較高。[9]
- Elliott 在[9]提到的 HamNoSys 的基礎上，把抽象的圖形轉換成 XML 標記語言。XML 標籤包含手型（hand shape）、手指方向（finger direction）、運動軌跡（motion trajectory）等細節，適合應用於動畫合成。[10]

2. 文字轉手語系統

將輸入的語言文字轉換為手語，再轉化成動畫來表達其結果。然而，現有大多數系統僅支持特定語言，或者尚未實現即時轉換功能。

- Fang 提供 Avatar 動畫展示，但動畫需要人工操作；支持 8 種語言但不包含中文。[1] 並提供兩種主要模式:
    1. MLSF(Multi-Language Switching Framework): 允許藉由動態添加 Encoder-Decoder 平行產生多種手語，減少語意錯亂。
    2. 與 Prompt2LangGloss，以單一 Encoder-Decoder 模組做到問題風格的提示手語，概念類似 LLM 且適合複雜自然語言輸入。

兩者都使用 Priority Learning Channel，利用強化學習損失函數和優先學習模組，量化訓練組品質與排變數組優先順序。

除此之外，與其他研究不同的是，作者不是利用現成的資料集，而是自己構建了新資料集 PROMPT2SIGN。在資料集中，特別設計以口語文本生成提示詞語(prompt words)，實現更高的效率。而在統一格式方面，他們將手語影片姿態經 OpenPose 處理後進行標準化，統一每一幀的關鍵點資訊(keypoints information)，減少冗餘資訊，並透過 seq2seq 與 text2text 模型生成結果，最終以 json 格式儲存。

資料收集方法: 2D 關鍵點 -> 3D 關鍵點 -> 壓縮姿勢資訊

- 統一格式處理辦法: 只專注在萃取姿勢和手勢資訊，減少了 80% 的資料大小（相較於原始影片）

1. 計算骨骼長度（skeleton): 以二維座標計算 2D 關鍵點之間的歐幾里得距離

$$L = \sqrt{(ax - bx)^2 + (ay - by)^2}$$

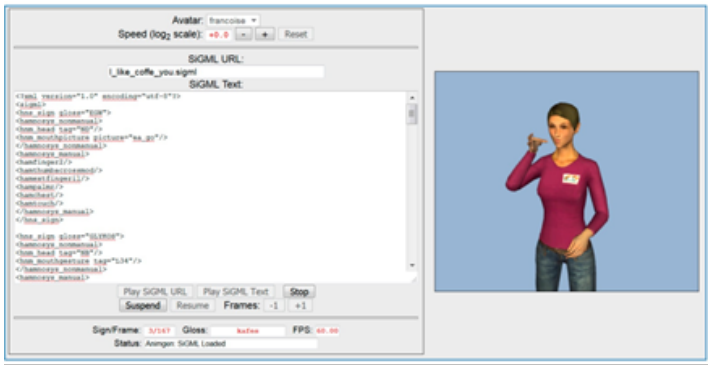
2. 計算 3D 旋轉角度: 以 2D 關鍵點計算正規劃旋轉角度

$$A_x, A_y, A_z = \frac{angle_x, angle_y, angle_z}{\sqrt{angle_x^2 + angle_y^2 + angle_z^2}}$$

3. 計算 3D 座標：Q 是 3D 的目標座標，P 是以 2D 關鍵點算出的起始座標

$$Q_x = P_x + L \times A_x, Q_y = P_y + L \times A_y, Q_z = P_z + L \times A_z$$

- Efthimiou 開發了一個高彈性的希臘文 Speech-To-Sign 人機交互 GUI「Dicta-Sign」，提供使用者創建和修改手語內容的功能。以 HamNoSys 轉 SiGML 格式儲存、描述手語特徵，如圖三，用於驅動手語虛擬人偶（Signing Avatar）的動作生成。文獻中提到手語的三維特徵增加了數據存儲的複雜性，使這種標準化數據集的應用場合受限。[2]



圖三 SiGML 腳本及以此腳本動態生成的希臘手語[2]

- Sanaullah 提出 Sign4PSL 巴基斯坦手語生成框架，使用 HamNoSys 紀錄結構特徵 + SiGML 處理文本，支持即時與離線使用。實際測試時發現單詞和短語的準確率可以到 100%，但複雜句子則因文法問題準確率降至 80%。為了提高準確率需要面對如圖四中一字多義情況的挑戰。[6]

Table 10: Detected ambiguity in sentences

No.	Problematic word	Tested sentence	PSL video	PSL expert
1	Right	You are **right**	✓	✓
		It is on **right** side	X	X
2	Fly	A **fly** was sitting on window	X	X
		He **flies** away happily.	✓	✓

圖四 一字多義辨識失誤問題將導致輸出錯誤手語動畫[6]

3. 動態手語生成技術

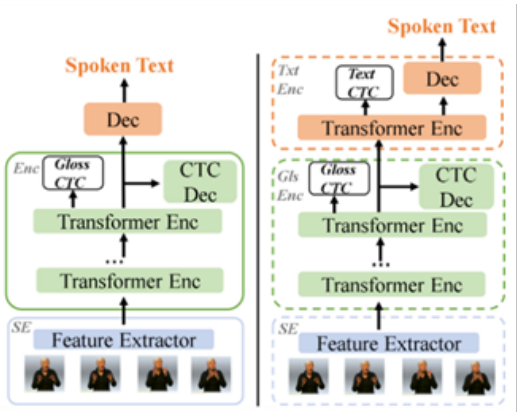
- Wang 提出 LVMCN 模型，包括 Cross-modal semantic aligner 與 Multimodal semantic comparator，計算語義一致性並進行訓練，其與其他模型之比較見圖五。PHOENIX14 資料集是一個手語翻譯(SLT) 的公開資料集，用於評估 SLT 效能；LVMCN model 處理手語影片與語言之間的複雜關係，特別是跨模態對齊和語意一致性。[3]

QUANTITATIVE RESULTS ON PHOENIX14T DATASET. '†' INDICATES THE MODEL IS TESTED BY US UNDER A FAIR SETTING.

Methods	DEV							TEST						
	B1↑	B4↑	ROUGE↑	WER↓	DTW-P↓	FID↓	MPJPE↓	B1↑	B4↑	ROUGE↑	WER↓	DTW-P↓	FID↓	MPJPE↓
Ground Truth	29.77	12.13	29.60	74.17	0.00	0.00	0.00	29.76	11.93	28.98	71.94	0.00	0.00	0.00
PT-base† [13]	9.53	0.72	8.61	98.53	29.33	2.90	41.92	9.47	0.59	8.88	98.36	28.48	3.22	51.35
PT-GN† [13]	12.51	3.88	11.87	96.85	11.75	2.98	40.63	13.35	4.31	13.17	96.50	11.54	3.33	50.80
NAT-AT [15]	–	–	–	–	–	–	–	14.26	5.53	18.72	88.15	–	–	–
NAT-EA [15]	–	–	–	–	–	–	–	15.12	6.66	19.43	82.01	–	–	–
D3DP-sign† [36]	17.20	5.01	17.94	91.51	–	2.38	39.42	16.51	5.25	17.55	91.83	–	2.63	47.65
DET [17]	17.25	5.32	17.85	–	–	–	–	17.18	5.76	17.64	–	–	–	–
G2P-DDM [37]	–	–	–	–	–	–	–	16.11	7.50	–	77.26	–	–	–
GCDM [16]	22.88	7.64	23.35	82.81	11.18	–	–	22.03	7.91	23.20	81.94	11.10	–	–
GEN-OBT [14]	24.92	8.68	25.21	82.36	10.37	2.54	41.47	23.08	8.01	23.49	81.78	**10.07**	2.97	52.90
LVMCN(Ours)	**25.79**	**9.17**	**27.29**	**76.86**	**10.25**	**1.94**	**35.11**	**24.33**	**9.36**	**26.24**	**75.43**	10.14	**2.16**	**42.54**

圖五 Results on PHOENIX14 Dataset test with multi models[3]

- Tan 基於 CTC/Attention 結合 transfer learning，處理單調與非單調對齊問題，提升手語生成的準確性，特別是手語影片與口語文本間的不規則關係。其在構造中加入了詞彙(gloss)為導向的編碼器(GlsEnc)與文本為導向的編碼器(TxtEnc)，架構請見圖六。前者負責使用 CTC 作詞彙序列對齊與手語影片長度調整;後者則負責在編碼過程中重新編排表徵，以處理文本對應的非單調性。[4]



圖六 普通 shared SLT encoder(左)與 hierarchical SLT encoder(右)架構圖[4]

4. 雙向手語翻譯處理技術

- Chaudhary 所開發的 SignNetII 基於 Transformer，加上使用 Attention 機制處理輸入序列，結合手語轉文字(P2T)與文字轉手語(T2P)的網路，透過雙重學習、共同訓練，利用手語解釋的對偶性，提升兩面的翻譯效能。除此之外，在 T2P 中作者新提出了姿勢相似性嵌入式學習，用一種損失函數，確保生成之手與序列與真實手語序列盡可能相似，並與其他手語序列保持一定距離。[7]而其公式如下:

$$\underbrace{\|f(B)-f(T)\|^2}_{d(B,T)} - \underbrace{\|f(B)-f(S)\|^2}_{d(B,S)} \leq 0$$

B 為基準姿勢(Baseline);T 為目標姿勢(True);S 為錯誤姿勢(False);d 是為計算兩姿勢序列間的距離函數。如果距離條件未被滿足，損失會增加，最小化姿勢相似性的損失函數將利於 T2P 模型準確度的提升，損失函數公式如下:

$$d(B,T,S) = max((d(B,T) - d(B,S) + \alpha), 0)$$

α 是一邊界值，用於引入更強的約束條件。

5. 輔助資源

- Tseng 所籌備的台灣手語資料庫計劃書，詳細說明資料庫的設置原因與內容處理方式。此資料庫提供多達 4100 個詞彙和 560 個例句，供我們做模型訓練。［5］
- Speech-To-Text 的準確性是 Speech-To-Sign 能達成的重要基礎。[8]在 2022 年提出的 Whisper 模型能精準提取語音並輸出文本，這為後續 Transformer 的映射和手語產生提供了高品質的數據源。

6. Encoder-Decoder 內部神經網路

根據最近的研究，Transformer 和 RNN-LSTM 都是常被應用於 Speech-to-Sign 領域的深度學習模型。儘管這些技術的文獻內容尚未全面開放，但基於文獻的摘要和技術說明，我們發現這些技術在處理多模態語法結構轉換中具有卓越的潛力。這些研究為我們設計 Encoder-Decoder 的神經網路時提供了重要的參考，以找出更適合本地需求的設計框架。

- 文獻[11]中分別使用 Transformer 和 CNN 提取語音的高層特徵，並映射到手語符號，目標是優化語音數據處理和手語生成的效率，使系統能夠滿足實時應用需求。兩個做法的優缺點比較尚未公開。
- 文獻[12]結合 CNN 與 RNN(Google LeNet 語音特徵提取 ＋RNN-LSTM 動作序列建模)。其中 Google LeNet 的語音識別率為 97.82%，RNN-LSTM 模型的序列建模能力有高達 98.25%的轉換準確率，並在端對端轉換中有較低的延遲。後續我們會將此性能指標做為對比基準，測試 RNN-LSTM 是否在處理中文複雜語法結構時，仍能保持高精準度。

表二 議題方法優缺點比較

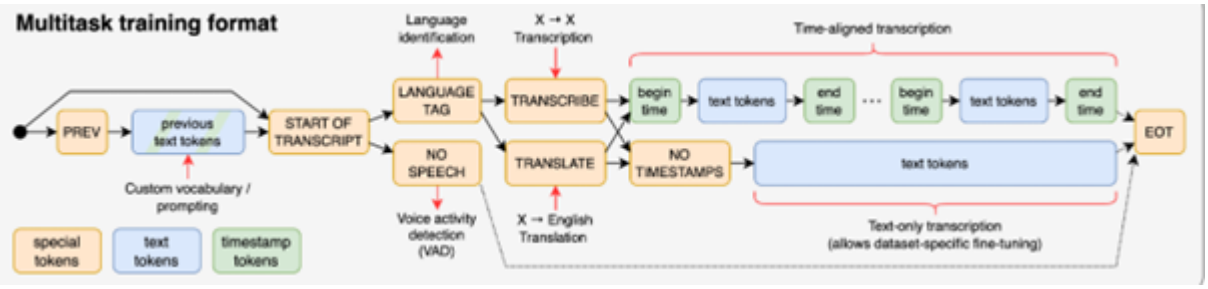
研究議題	方法類型	代表文獻	優點	缺點
手語資料集產生方法	Encoder-Decoder	[1]	靈活性、上下文感知能力強	資源要求大 黑盒子
SLT 語意對應一致性	LVMCN	[3]	實現細粒度(grained）的跨模態對齊、解決詞語動作不一致、生成動作較自然	未探討其缺點
提升 SLT 生成的準確性	Joint CTC/Attention	[4]	處理非單調性對齊	具對特定資料的依賴性，限制模型的擴展與提升

(四) 研究方法及步驟

1. 語音轉文字：

- 工具：Whisper API（Python SDK）是 OpenAI 提供的即時語音轉文字工具，內部已使用 Transformer，可以得到文字序列。

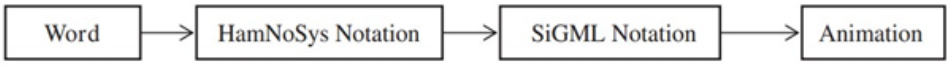
- 訓練/調試：測試不同語音樣本以確保模型的文字生成準確性。
- 後處理：使用 NLP 模型(BERT)作語法調整。



圖七 whisper 模型修改流程圖[14]

2. 文字序列轉換成手語序列：

綜合[2]、[6]可以歸納出，傳統轉換方式(如圖八所示)不外乎先將文字換成 HamNoSys 圖形表示法後，再轉成計算機語言 SiGML，如此一來便能生成手語動畫。



圖八 Sign generation workflow 流程圖[6]

實時性和自動化是本研究的首要目標。HamNoSys + SiGML 方法作為一種符號驅動的手語生成框架，在靜態手語表示方面具有較高的準確性和標準化優勢，但其符號化過程需要專業手語學者的人工設計，對語言間的擴展性有限。而 Encoder-Decoder 方法不僅能直接從語音或文本中學習語法和語義特徵，還能利用大規模數據自動優化翻譯過程，尤其適合處理多義詞等高語義需求的應用場景。

基於以上緣由，本研究將採用 Encoder-Decoder 方法動態產生手語資料集。

表三 手語資料集產生策略比較

手語資料集產生方法	Encoder-Decoder	HamNoSys + SiGML
代表文獻	[1]、[7]	[2]、[6]
架構	深度學習(RNN/Transformer)	基於符號表示和規則的結構化模型
處理能力	依訓練模型動態學習語義和語法規則，支持多義詞和上下文分析	使用標準化符號，適合靜態手語生成和對精確度有嚴格要求的情境
實時性	訓練後可高效生成手語，但推理流程速度浮動大	SiGML 可以用來低延遲生成手語動畫
擴展性	支持多語言模型訓練	需為不同語言擴展資料集，適應性較低

如表三所示，使用 Encoder-Decoder 製作手語數據集更符合需求

表四 Encoding 模式比較

方式	MLSF（多語言切換框架）	Prompt2LangGloss（提示到語言詞彙）	SignNet T2P(文字轉手語)
功能	直接翻譯文字成手語姿勢	理解複雜輸入，生成手語姿勢	轉換文字序列成對應手語姿態序列
特性	像字典/抽屜	類大型語言模型(LLM)	指標嵌入學習
輸入方式	純文字	複雜語言提示（e.g.問問題形式）	文字嵌入(文字序列換嵌入向量)、位置編碼
優點	高效、穩定、擴展性高	理解複雜輸入	快速、高效、可擴展性
缺點	無法直接處理複雜提示	效率可能低	計算資源需求高、對輸入資源品質敏感
著重點	翻譯效率、穩定性	複雜輸入的理解能力與發展潛力	手語姿態精確生成

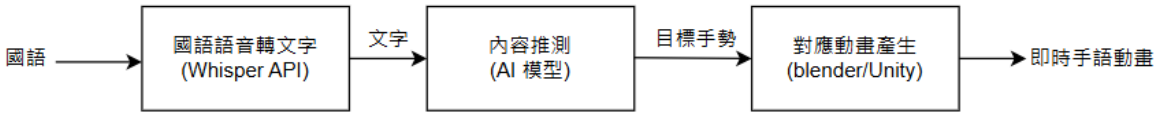
基於表四之比較，由於即時服務需要穩定的輸出、效率與易擴展性。雖然手語姿態的準確輸出也很重要，但我們希望模型對輸入也是有容忍度的，在權衡準確度與輸出速度下，我們選擇較為折衷的 MLSF 作為使用 Encoder-Decoder 的模式。

- 工具：TensorFlow/Keras 或 PyTorch，用於構建 Encoder-Decoder 模型。Encoder 負責提取文字特徵，Decoder 生成對應的手語動作關鍵點序列(手型、位置、方向)。
- 數據處理：利用本校資源(台灣手語線上辭典)結合 OpenPose 提取的關鍵點進行模型訓練。

3. 手語動作呈現：

- 工具：Blender（3D skeleton model 設計），Unity（動畫生成）。
- 測試：模擬多種手語動作，驗證動作的流暢性和準確性。

通過上述技術，可大致將系統的聲音轉換為手語流程劃分為圖九之步驟。

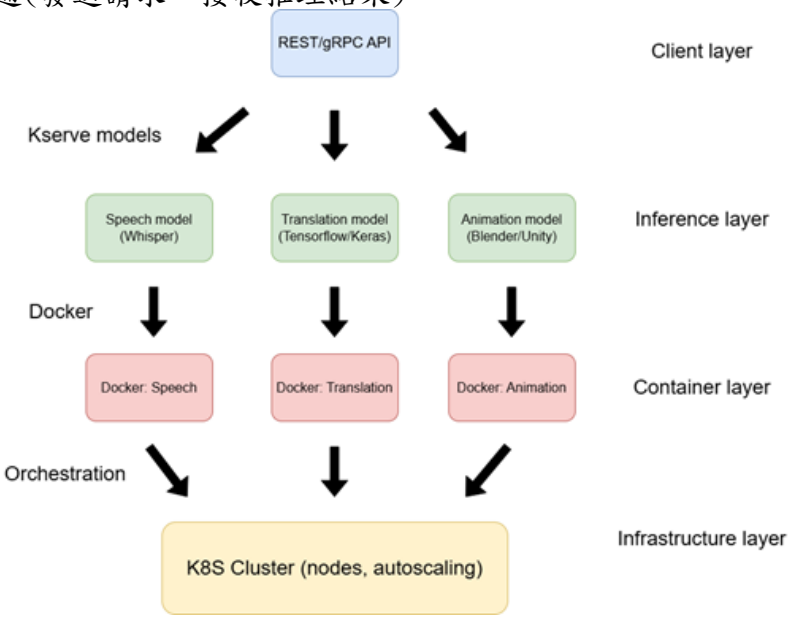


圖九 AI model pipeline 示意圖

4. Kserve 部屬：

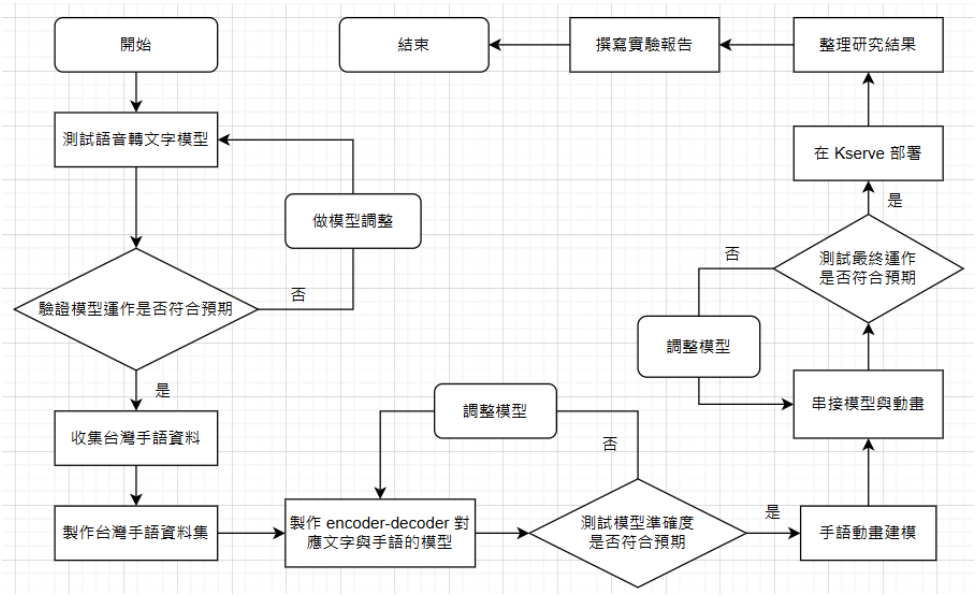
- 工具：Kserve（多模態推理服務部署），Docker（容器化每個模型），Kubernetes（實現多節點集群管理）。Kserve 支持多模態模型的推理，我們計畫以此設計一個 multi-model pipeline 將功能模組串聯起來。即部署 3 個 Kserve 模型分別處理語音、翻譯和動畫，架構如圖九。
- 測試：模擬高迸發場景，確保 Autoscaling 功能正常運作，並通過負載測試評估系統性能。

- 貢獻：將 WebSocket 整合進 Kserve 的服務中。修改 Kserve 的配置，使其支持 WebSocket 協議，讓 Kserve 作為伺服器，Blender 作為客戶端，做到即時性的溝通(發送請求、接收推理結果)。



圖十 Kserve 部署架構圖

依圖十一的流程所示，得到文字序列後，OpenPose 負責從現有的台灣手語資料中提取標準化關鍵點數據，並標註資料以供 Decoder 學習。利用 Encoder 提取文字特徵和語法轉換，Decoder 在生成對應的手語動作關鍵點座標序列；接著將關鍵點序列輸入至後續的動作生成模組中，搭配 Unity 產生手語動畫。



圖十一 研究流程表

(五) 預期結果

實現即時將語音轉換為台灣手語，並用動畫呈現的服務。首先，我們會在本地端對本服務的手語辨識準確率進行測試，比照文獻[6]靜態手語生成的準確率，我們期

望在複雜語句(非 S+V+O 結構的句子)下能達到至少 80%的準確率。

準確率達標後我們會將服務藉由 Kserve 部署至雲端，針對運行效能進行測試、整合 WebSocket 進 Kserve，使 Kserve 可以在與其他軟體協作時做到更即時性的操作，並在社群中提供操作範例，將成果回饋給開源社群。

（六）需要指導教授指導內容

教授長年積極推動開源專案的理念，近期教授更受邀參與「教育部開源人才培育計畫」的國家型計畫，致力於培養具備實戰能力與國際競爭力的開源技術人才。在校內，教授不僅擔任專題老師帶領學生直接參與開源貢獻，更親自設計專業課程介紹開源工具，提升全體學生的技術實力、開源素養與協作能力。

此外，「即時語音轉台灣手語」的實現涉及語音辨識、自然語言處理、手語生成等多個技術領域，每一環節都需要結合扎實的 Domain Knowledge。教授對深度學習與影像處理的專業背景，以及對開源框架（如 KServe）的熟悉，使我們在專題執行過程中能獲得寶貴的建議和技術支持。

在教授的指導下，我們可以學到與開源社群正確互動的技巧，還能深入體會嚴謹的治學、研究態度。教授在 MLops 領域擁有豐富的實戰經驗，希望教授能在我們研究遇到困難時能以其豐富的經驗提出建議，幫助我們在解決問題的同時，從中精進實務技能，為未來的研究奠定穩固的基礎。

（七）參考文獻

[1] FANG, Sen, et al. SignLLM: Sign Languages Production Large Language Models. arXiv preprint arXiv:2405.10718, 2024.

[2] Efthimiou, Eleni, et al. "Sign Language Technologies and the Critical Role of SL Resources in View of Future Internet Accessibility Services." Technologies, vol. 7, no. 1, 2019, p. 18. https://doi.org/10.3390/technologies7010018.

[3] WANG, Xu, et al. Linguistics-Vision Monotonic Consistent Network for Sign Language Production. arXiv preprint arXiv:2412.16944, 2024.

[4] TAN, Sihan, et al. Improvement in Sign Language Translation Using Text CTC Alignment. arXiv preprint arXiv:2412.09014, 2024.

[5] Tsay, Jane. 2019. Taiwan Sign Language Online Dictionary: Construction and Expansion. In the Proceedings of the 2nd ILAS Annual Linguistics Forum – National Language Corpora: Design and Construction, 85-110. Taipei: Institute of Linguistics, Academia Sinica.

[6] SANAULLAH, Muhammad, et al. A real-time automatic translation of text to sign language. Computers, Materials & Continua, 2022, 70.2.

[7] Chaudhary, Lipisha et al. “SignNet II: A Transformer-Based Two-Way Sign Language Translation Model.” IEEE transactions on pattern analysis and machine intelligence vol. 45,11 (2023): 12896-12907. doi:10.1109/TPAMI.2022.3232389

[8] Radford, Alec, et al. "Robust Speech Recognition via Large-Scale Weak Supervision." arXiv:2212.04356 (2022).

[9] S. Prillwitz, R. Leven, H. Zienert, T. Hanke and J. Henning, “Hamburg notation system for sign languages: An introductory guide, HamNoSys Version 2.0, Signum, Seedorf, Germany, 1989.

[10] R. Elliott, J. Glauert, V. Jennings and J. Kennaway, “An overview of the SiGML notation and SiGML signing software system,” in Proc. 4th Int. Conf. on Language Resources and

Evaluation, Lisbon, Portugal, pp. 98–104, 2004.
［11］D. Rai, N. Rana, N. Kotak and M. Sharma, "Real-Time Speech to Sign Language Translation Using Machine and Deep Learning," 2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2024, pp. 1-5, doi: 10.1109/ICRITO61523.2024.10522437.
［12］Jing Zhang. 2024. Research on Speech Information to Sign Language Translation Based on 1D-GoogLeNet and LSTM. In Proceedings of the 2024 9th International Conference on Cyber Security and Information Engineering (ICCSIE '24). Association for Computing Machinery, New York, NY, USA, 707–712. https://doi.org/10.1145/3689236.3691489
［13］SpreadTheSign 多語言手語辭典
［14］whisper github page https://github.com/openai/whisper