應用於電腦輔助發音訓練的多模態語音特徵融合架構

運用大語言模型與注意力機制，提升發音評估的準確性與診斷能力

研究背景與挑戰

傳統CAPT系統限制

- 評估結果缺乏診斷性
- 僅能判斷對錯，無法提供細節分析
- 特徵融合方法過於簡化
- 無法捕捉發音品質的細微差異

學習者需求

- 精準的發音錯誤定位
- 詳細的改進建議
- 個人化學習回饋
- 即時性能評估

整體研究目標

發音錯誤檢測

Mispronunciation Detection (MDD)

精確識別學習者發音中的具體錯誤位置與類型，提供針對性的修正建議

自動發音評估

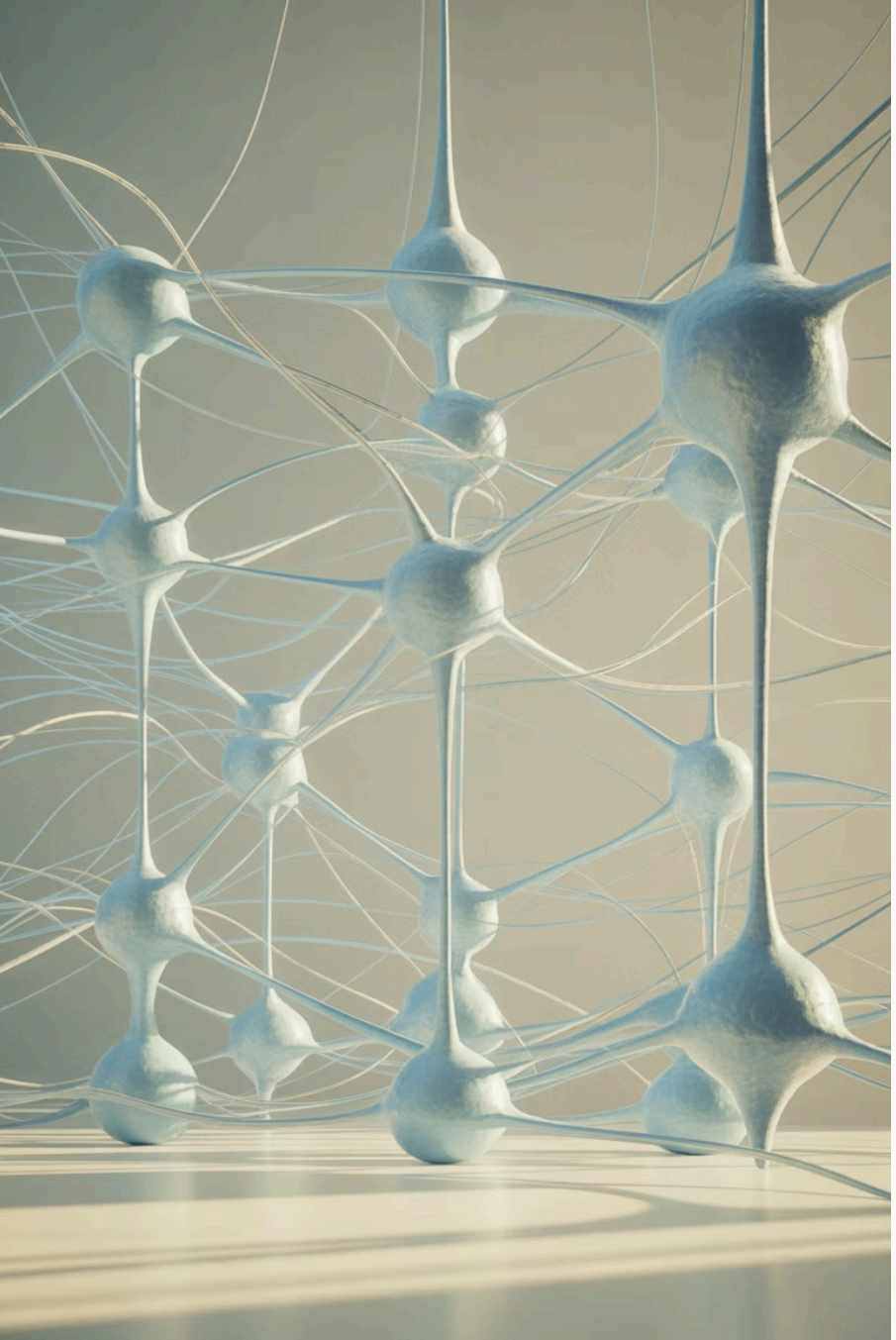
Automatic Pronunciation Assessment (APA)

建立多維度評估機制，從流暢度、準確性、韻律等角度全面分析發音品質

語音評分

Speech Scoring

提供標準化分數系統，支援學習進度追蹤與能力水準認證



研究策略：深化特徵工程

計畫初期將專注於**深化特徵工程**（Deepening Feature Engineering），這是提升CAPT系統性能的關鍵基礎。傳統方法往往依賴簡單的特徵組合，無法有效捕捉語音信號中的複雜模式。

我們的創新方法將開發專為CAPT應用設計的多模態特徵融合架構，能夠識別並量化發音品質的細微變化，為後續的錯誤檢測與評估提供更豐富的資訊基礎。

特徵工程的深化將直接影響系統診斷能力的提升，這是實現精準發音訓練的核心技術突破點。

核心技術架構

01

大語言模型基底

採用Titans架構作為特徵融合核心，提供強大的語義理解能力

02

多源特徵輸入

整合logit-based特徵與韻律特徵，捕捉發音的多維度資訊

03

語義層級融合

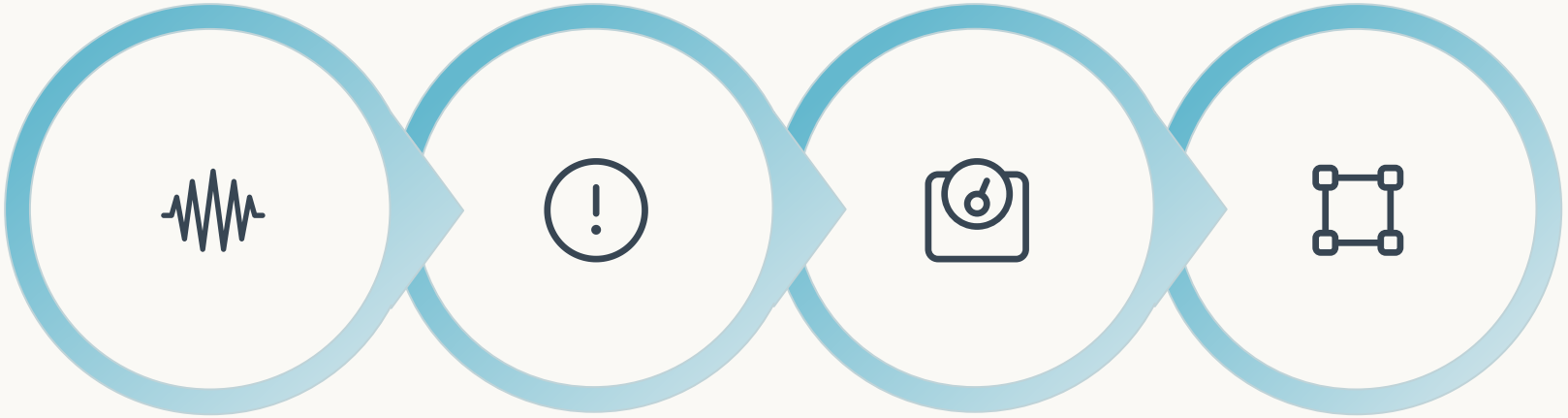
生成高維度融合特徵向量，保留發音品質的細節資訊

04

診斷性輸出

提供具體的錯誤分析與改進建議，而非僅有對錯判斷

注意力機制：智慧型特徵聚合器



傳統的特徵聚合方法（如簡單平均）會遺失重要的時序資訊。我們的注意力機制能夠：

- 自動識別與發音品質最相關的音訊片段
- 為不同時間點的特徵分配適當的重要性權重
- 動態調整聚合策略，適應不同的發音模式
- 生成更具鑑別力的特徵向量

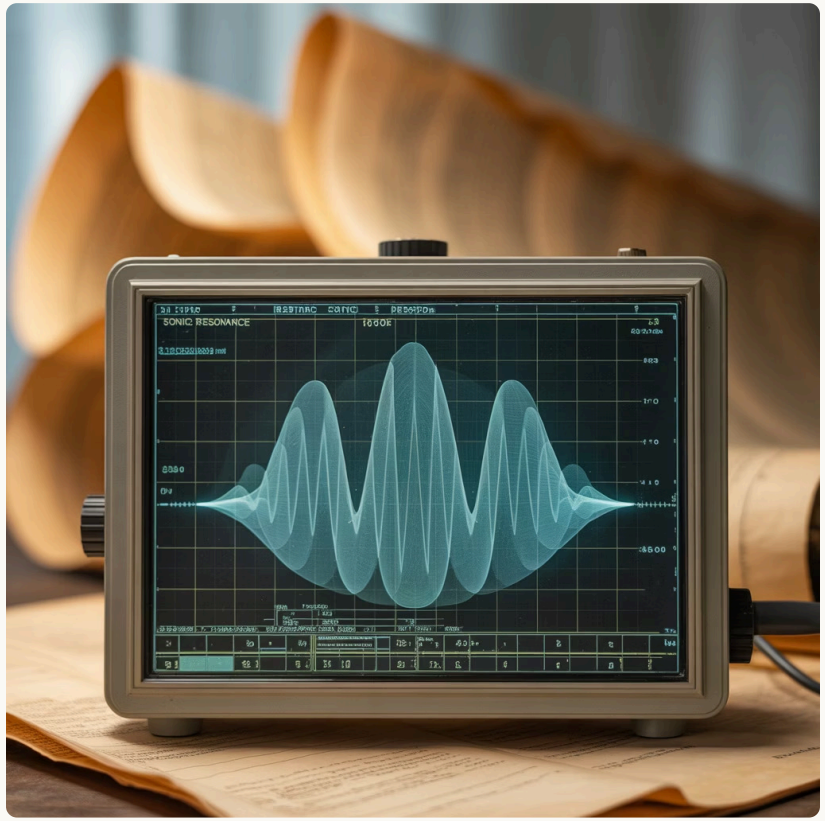
多模態特徵整合

Logit-based特徵

- 音素辨識信心度
- 發音清晰度指標
- 語音品質評估

韻律特徵

- 音調變化模式
- 節奏與停頓
- 語調起伏
- 重音分佈



融合架構將這些異質特徵轉換為統一的語義表示，讓系統能夠理解發音的完整脈絡，而非僅依賴單一維度的資訊進行判斷。

實驗設計與驗證方法

1

對照組設計

比較三種特徵聚合方法的效能差異

2

統一評估框架

使用相同的下游分類器確保公平比較

3

量化分析

採用MCC分數作為主要評估指標

平均聚合

傳統方法，對所有時間框架等權重處理

最大聚合

選取最顯著的特徵值，可能遺失重要資訊

注意力聚合

我們提出的方法，動態權重分配

研究意義與未來展望

學術貢獻

- 首創CAPT領域的LLM特徵融合方法
- 建立注意力機制在語音評估的新典範
- 提供可重現的實驗設計框架

產業應用價值

- 提升語言學習App的教學效果
- 降低人工評估成本
- 支援大規模個人化教學

> 這項研究將為電腦輔助語言學習開啟新的技術可能性，讓每位學習者都能獲得如同專業教師般精準的發音指導。

未來研究將擴展至多語言支援、情境化評估，以及與虛擬實境技術的整合，打造更沉浸式的語言學習體驗。