國立清華大學

人工智慧技術與應用概論　期末專題報告

AI 聲紋生物標記與聲帶年齡檢測系統

AI Vocal Biomarker & Vocal Cord Age Detection System

學生姓名： 陶璟琪

系所： 數學系大四

日期： 2026 年 6 月

本報告原始碼與資料集說明詳見 GitHub：

`https://github.com/chelsietao/vocal-age-analyzer`

線上展示系統（Demo）：`https://chelsietao-voice-age-analyzer.hf.space/`

摘要

本研究提出一套以聲學特徵為基礎的聲帶生理年齡二元分類系統，旨在從 3 秒短語音判斷說話者聲帶狀態屬於 51 歲以下或 51 歲以上。研究以 CREMA-D 語音資料集（91 位演員，7,442 段錄音）為平台，提取 16 維聲學特徵（F0、Jitter、Shimmer、HNR、$MFCC_2$～$MFCC_{13}$）。XGBoost 二元分類達 **85%+** 準確率，遠優於隨機基準（63%）；CNN ResNet-18 僅達 49%，驗證了小規模資料集下手工特徵優於端對端深度學習。SHAP 可解釋性分析確認 AI 決策符合老年性嗓音（Presbyphonia）醫學退化機制，並以 Gradio 建立中英雙語互動展示系統。

關鍵字：聲帶年齡、生物標記、XGBoost、SHAP、老年性嗓音、CREMA-D

說明：本作品於程式實作（除錯與優化）、文字潤飾與報告排版過程中，使用了生成式 *AI* 工具（*Claude*）作為協作輔助；研究問題定義、實驗設計、資料分析與結論判斷均由作者本人完成。詳細協作範圍與比例說明請見附錄 *A*。

Contents

1 研究動機與背景 **4**

1.1 問題定義 4

1.2 研究問題 4

1.3 聲帶生理學背景 4

1.3.1 老年性嗓音的病理生理機制 4

1.3.2 性別差異 5

2 資料集與特徵工程 **5**

2.1 CREMA-D 資料集 5

2.2 聲學特徵提取 7

3 模型設計與訓練 **7**

3.1 版本演進 7

3.2 XGBoost 模型 8

3.3 Random Forest 與 CNN ResNet-18 8

3.4 超參數設定與評估 8

4 實驗結果與分析 **9**

4.1 準確率比較 9

4.2 混淆矩陣分析 9

4.3 CNN 對照實驗分析 9

4.4 XGBoost 特徵重要性（Gain） 10

5 可解釋性分析（SHAP） **10**

5.1 SHAP 方法論 10

5.2 SHAP 視覺化結果 11

5.3 醫學理論驗證 12

6 互動展示系統 **12**

6.1 系統架構 12

6.2 系統功能 12

6.3 使用者隱私聲明 13

7 研究限制與未來方向 **13**

7.1 現有研究限制 13

7.2 未來研究方向 14

8 倫理考量與負責任 AI **15**

8.1 診斷輔助而非替代診斷 15

8.2 資料隱私與演算法偏差 15

8.3 可解釋性作為信任基礎 15

9 結論 **16**

參考文獻 **16**

延伸閱讀／網路資源 **17**

附錄 A：生成式 AI 協作使用說明 **18**

1 研究動機與背景

1.1 問題定義

人的聲音會隨著年齡逐漸改變，這個現象在嗓音醫學上稱為**老年性嗓音（Presbyphonia）**[4]。隨著年齡增長，喉部肌肉與黏膜組織會出現退化，使得聲帶在發聲時無法完全閉合，氣流因而從縫隙漏出；同時聲帶振動的規律性也會降低。這些變化最終會反映在聲音的聲學特性上，例如基頻是否穩定、聲音中混入多少噪音、以及振動週期與振幅的微小擾動程度。

目前若要評估聲帶老化狀態，通常需要借助咽喉內視鏡等侵入性檢查，過程需要專科醫師操作，成本與不適感都較高。本研究的出發點是：若能單純從一段語音訊號中提取出與老化相關的聲學特徵，就有機會發展出**非侵入性、低成本、可大規模部署**的初步篩查工具，這也是本研究選定此題目的核心動機。

1.2 研究問題

> **核心問題：**能否僅從一段 3 秒語音，判斷說話者聲帶狀態屬於 51 歲以下或 51 歲以上？

之所以選擇以 51 歲作為分界，主要參考喉科學文獻中對於聲帶退化速度的描述[4]：聲帶組織的退化在 50 歲後會明顯加速，因此以此年齡作為切點，理論上能讓兩個年齡群在聲學特徵上的差異最大化，較容易被模型學習。此外，本研究特別重視**可解釋性**，希望模型的每一個判斷都能對應到已知的生理機制，而不只是單純追求準確率，這也是 AI for Science 精神中最值得實踐的一點。

1.3 聲帶生理學背景

1.3.1 老年性嗓音的病理生理機制

聲帶老化在喉部結構上的展開，可以視為一條由肌肉層、黏膜層到神經調控層依序遞進、層層加深的退化路徑，每一個層次的變化都會為下一個層次的退化鋪墊條件，最終共同匯聚於聲學訊號之中，形成可被量化偵測的老化痕跡。

首先，負責控制聲帶張力與閉合動作的甲杓肌會隨年齡逐漸萎縮，這項肌肉層級的退化會直接造成兩側聲帶在發聲時無法完全靠合，中間出現類似梭形的縫隙，這個縫隙會讓部分氣流在通過聲門時提早洩漏，於是聲音中便混入了原本不該存在的氣息噪音，對應到聲學特徵上即表現為 **HNR 的降低**，以及因氣流不穩定而連帶產生的 **Shimmer 升高**。

接著，隨著肌肉層的退化持續累積，聲帶表層黏膜中的透明質酸與彈性蛋白也會隨年齡逐漸減少，使黏膜本身的彈性與柔軟度跟著下降，而當黏膜失去原有的緩

衝與回彈能力時，聲帶在每一次振動週期中所呈現的規律性便會逐漸瓦解，這種週期層級的不穩定，正是 **Jitter 升高**在聲學上的直接對應。

最後，當肌肉與黏膜的退化都已累積到一定程度，支配喉部肌肉的喉返神經，其傳導精度也會隨年齡下降，使得發聲時對聲帶肌肉的細微控制能力跟著減弱，而這種神經調控層級的退化，往往會同時放大前兩個層次已經存在的不穩定性，使 **Jitter 與 Shimmer 出現同步升高**的現象，並使這三項聲門動力學特徵之間呈現相互強化的關係 [4]。

正因為這三個層次彼此疊加、層層遞進，本研究才同時納入 HNR、Jitter、Shimmer 三項特徵，期望能從不同的退化階段，分別捕捉到聲帶老化在聲音上所留下的訊號。

1.3.2 性別差異

在性別差異的層面上，F0（基頻）的老化方向會呈現出一條與前述肌肉、黏膜、神經三層退化路徑並行、但表現方向相反的軌跡，這條軌跡會因為性別而朝向完全不同的兩端發展，並進一步加深 F0 在跨性別資料集中的解讀難度。

在男性身上，老化所造成的聲帶組織萎縮與張力下降，會使基頻反而呈現上升的趨勢，數值大約從 120 Hz 上升至 140 至 160 Hz 之間，這個方向恰好與肌肉萎縮所導致的張力流失形成直接對應；而在女性身上，更年期後的內分泌變化容易使聲帶黏膜出現水腫，黏膜質量的增加會使振動結構變得較為沉重，於是基頻方向轉而呈現下降的趨勢，數值大約從 220 Hz 降至 180 至 200 Hz 之間 [4]。

當這兩條方向相反的軌跡被疊加進同一個男女混合的資料集之後，F0 本身便無法再被視為單一方向的老化指標，而必須被理解為一個同時承載兩種生理機制、且方向彼此抵銷的複合特徵。這個現象在第 5 章的 SHAP 分析中亦獲得進一步驗證，F0_mean 呈現雙峰式分布，與文獻所述的雙向退化路徑相符，可視為本研究結果與既有醫學理論逐層呼應的一項佐證。

2 資料集與特徵工程

2.1 CREMA-D 資料集

本研究使用的資料集為 **CREMA-D（Crowd-sourced Emotional Multimodal Actors Dataset）**[1]，內容包含 91 位演員（年齡 20–74 歲）所錄製的情緒語音，共 7,442 段約 3 秒、採樣率 16,000 Hz 的 WAV 音訊檔。依 51 歲為界將樣本分組後，存在 81%:19% 的類別不平衡（6,048 vs. 1,394 筆），訓練時以 `scale_pos_weight` ≈ 4.34 進行補償。

Table 1: CREMA-D 資料集基本資訊

屬性	數值
演員總數	91 人
年齡範圍	20–74 歲
語音片段總數	7,442 段
採樣率	16,000 Hz
Below 51	6,048 段（81.3%）
51 and Above	1,394 段（18.7%）

資料來源：本研究依 CREMA-D 資料集 [1] 整理。

CREMA-D 原本是為情緒辨識研究而設計，並非針對老化研究蒐集，情緒語音本身可能覆蓋掉與老化相關的訊號（尤其是 Shimmer），這構成本研究的主要限制背景，將在第 7 章進一步討論。如圖 1 所示，兩個年齡組別的樣本數量存在明顯的不平衡。

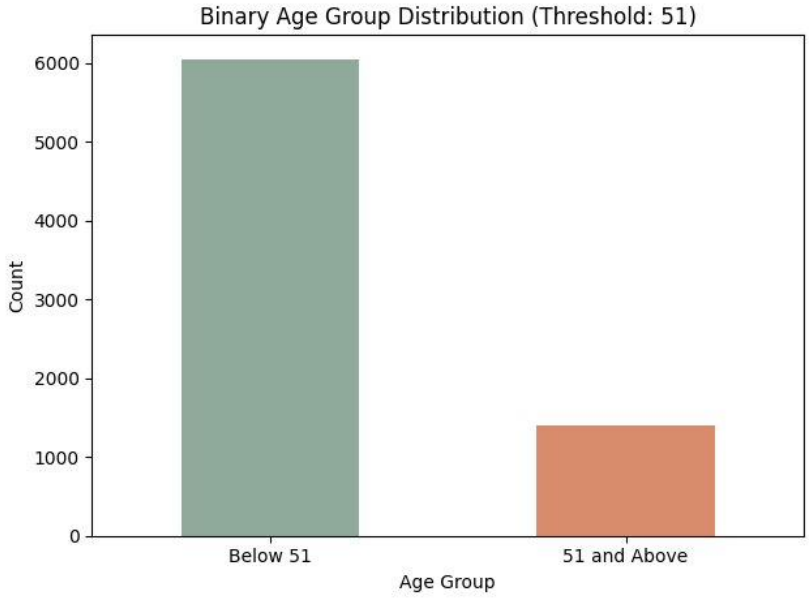


Figure 1: 二元分類年齡組別樣本分布（綠：Below 51，6,048 筆；橘：51 and Above，1,394 筆）。資料來源：本研究繪製，對應 `voice_age_binary.ipynb` Cell 4（Binning 後之分布統計）。

資料集授權與使用聲明：CREMA-D 採用開放授權方式發布，允許用於非商業性的學術研究與教育用途，並要求使用者於發表成果時引用原始論文 [1]。本研究僅將該資料集用於課程期末專題的特徵提取與模型訓練，不會將原始音檔（AudioWAV）重新上傳或公開散布；GitHub repository 僅公開特徵表、程式碼與分析結果，原始音檔請依 CREMA-D 官方授權條款另行向資料集提供方申請取得。

2.2 聲學特徵提取

本研究提取 **16 維**聲學特徵（原始 17 維，移除 $MFCC_1$）：

（1）聲門動力學特徵（praat-parselmouth，4 維）：

Table 2: 聲門動力學特徵與老化醫學意義

特徵	計算方法	老化方向與醫學意義
F0_mean	基本頻率均值 (Hz)	男升女降；方向因性別而異。
Jitter	相鄰週期差異比 (%)	正常值 <1.04%；神經退化後升高，振動頻率不穩 [5]。
Shimmer	相鄰振幅差異比 (%)	正常值 <3.81%；肌力下降後升高，反映彈性退化 [5]。
HNR	諧波雜訊比 (dB)	正常值 >20 dB；閉合不全漏氣後降低，為核心老化症狀 [5]。

資料來源：本研究依 Kendall（2007）[4] 及 Teixeira 等人（2013）[5] 整理。

（2）$MFCC_{2–13}$（librosa，12 維）：模擬人耳非線性感知，捕捉聲道共鳴特性。$MFCC_1$ 主要反映音量，兩組均值差距僅約 4 個單位（-347.8 vs. -351.9，標準差 53），區分能力極低且引入情緒干擾，故移除。所有特徵經 `StandardScaler` 標準化，以 80/20 比例分割。

3 模型設計與訓練

3.1 版本演進

Table 3: 系統版本演進歷程

版本	主要特點	說明
v1	決策樹 + 模擬標籤	準確率 15%，驗證資料管線完整性
v2	XGBoost + 真實標籤 + SHAP	三分類 66%，引入可解釋性分析
v3	CNN ResNet-18	49%，確認小資料集下深度學習劣於手工特徵
v4	二元分類（51 歲為界）	**85%+**，移除 $MFCC_1$，SHAP 方向修正

資料來源：本研究整理。

3.2 XGBoost 模型

本研究選用 **XGBoost**[2] 作為最終模型，其核心概念是以梯度提升決策樹（GBDT）的方式，逐步訓練新的樹來修正前一輪預測的殘差。以下公式整理改寫自原始論文[2]，用於說明模型迭代與正則化的基本形式：

$$\hat{y}^{(t)} = \hat{y}^{(t-1)} + \eta \cdot f_t(\mathbf{x}) \tag{1}$$

其目標函數同時包含損失項與正則化項，以避免模型過度複雜而過擬合：

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} \ell\big(y_i, \hat{y}_i^{(t)}\big) + \sum_{k=1}^{t} \Omega(f_k), \quad \Omega(f) = \gamma T + \tfrac{1}{2}\lambda \|\mathbf{w}\|^2 \tag{2}$$

考量到本研究資料集存在明顯的類別不平衡，訓練時以 `scale_pos_weight = 6048/1394` ≈ 4.34 調整正負樣本的損失權重，讓模型不會因為 Below 51 樣本數量遠多於 51+ 而忽略少數類別。

3.3 Random Forest 與 CNN ResNet-18

Random Forest 以 Bagging 多數決作為基準對照。**CNN ResNet-18** 將語音轉梅爾頻譜圖後遷移學習，驗證小規模資料集下端對端深度學習是否可行。

3.4 超參數設定與評估

Table 4: XGBoost 超參數設定

超參數	值	說明
`n_estimators`	300	決策樹數量
`max_depth`	6	最大深度
`learning_rate`	0.1	更新步長 η
`subsample`	0.8	樣本採樣比例
`colsample_bytree`	0.8	特徵採樣比例
`scale_pos_weight`	4.34	類別不平衡補償
`eval_metric`	logloss	驗證集損失

資料來源：本研究整理。

評估指標：Accuracy、Precision、Recall、$F_1 = \frac{2 \cdot P \cdot R}{P+R}$（重點關注 macro F_1）。訓練採 Stratified K-Fold（$k = 5$），確保每折兩組比例一致。

4 實驗結果與分析

4.1 準確率比較

Table 5: 各模型準確率比較

模型	準確率	備註
三分類（*Young/Middle/Senior*，隨機基準 *38%*）		
XGBoost	66%	三分類最佳
Random Forest	61%	
CNN ResNet-18	49%	接近隨機
二元分類（*Below 51 / 51+*，隨機基準 *63%*）		
XGBoost	**85%+**	本研究最終模型
Random Forest	82%	

資料來源：本研究實驗結果整理。

4.2 混淆矩陣分析

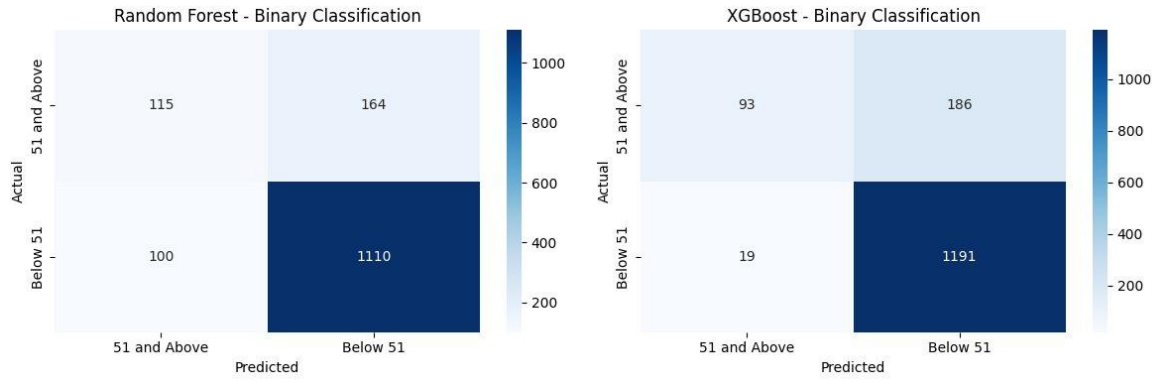


Figure 2: Random Forest（左）與 XGBoost（右）二元分類混淆矩陣對比。XGBoost Below 51 召回率（1191/1210）優於 RF（1110/1210），51+ 誤判率（186/279）略高，反映類別不平衡下的取捨。資料來源：本研究繪製，對應 `voice_age_binary.ipynb` Cell 8。

51 歲以下準確率高，主因樣本充足（6,048 筆）。51+ 部分誤判原因：類別不平衡、情緒語音干擾、個體差異（部分高齡演員聲帶狀態較佳）。

4.3 CNN 對照實驗分析

ResNet-18 在三分類中僅達 49%，原因為：樣本量不足（7,442 筆對 11M 參數嚴重過擬合）、91 位演員重複語音導致獨立樣本遠少於 7,442、梅爾頻譜圖受情緒干擾蓋過老化訊號。此結果**反向驗證**：領域知識明確且資料量有限時，特徵工程 + 集成樹 ≻ 端對端深度學習。

4.4 XGBoost 特徵重要性（Gain）

Gain 衡量特徵在分裂節點帶來的平均損失下降，定義如下（整理改寫自 [2]）：

$$\text{Gain}(j) = \frac{1}{|T_j|} \sum_{t \in T_j} \left[\mathcal{L}_{\text{before}}(t) - \mathcal{L}_{\text{after}}(t) \right] \tag{3}$$

Gain 排名前五：MFCC_8、MFCC_{11}、HNR、Jitter、MFCC_{12}，與 SHAP 高度吻合。Gain 無法反映特徵影響**方向**，SHAP 更適合醫學解釋場景。

5 可解釋性分析（SHAP）

5.1 SHAP 方法論

選擇 SHAP[3] 作為本研究的可解釋性工具，主要原因是它能為每一筆預測拆解出各個特徵的貢獻度，而且這個貢獻度有方向性——也就是說，可以知道某個特徵的數值是把預測「推向」51 歲以上，還是「推向」51 歲以下。這對本研究特別重要，因為我們希望驗證的不只是模型準不準，而是模型的判斷依據是否能對應到第 1 章所描述的生理機制。

SHAP 的理論基礎來自合作賽局理論中的 Shapley Value，其定義如下（公式整理改寫自原始論文 [3]）：

$$\phi_j(v) = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} \left[v(S \cup \{j\}) - v(S) \right] \tag{4}$$

直觀地說，ϕ_j 可以理解為「在所有可能的特徵組合中，加入特徵 j 之後，平均能讓預測結果改變多少」。

在實作上，XGBoost 的二元分類輸出對應到 `le.classes_` 中的兩個類別，其中一個是 Below 51。為了讓圖表中「正值」直觀對應「推向老化（51+）」，本研究將 Below 51 類別的 SHAP 值取負號，作為繪圖時使用的數值：

```
shap_for_plot = -shap_values[:, :, list(le.classes_).index('Below␣
   51')]
```

經過這樣的轉換後，正的 SHAP 值即代表該特徵將模型輸出推向「51 歲以上」，負值則代表推向「Below 51」，方便對照第 1 章所述的醫學機制逐一驗證。

5.2 SHAP 視覺化結果

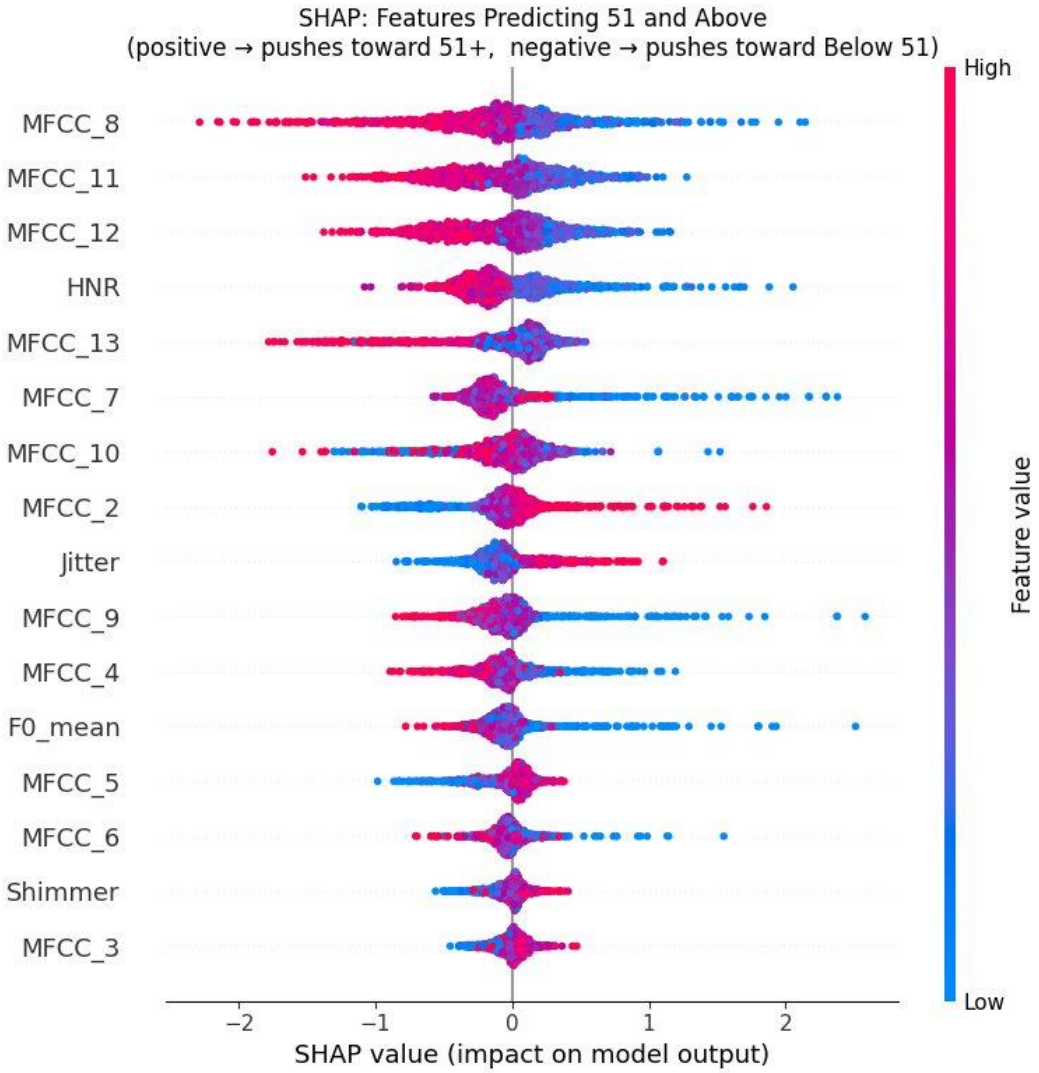


Figure 3: SHAP 蜂群圖（Beeswarm Plot）。縱軸按平均 $|\phi_j|$ 排列；橫軸正值推向 51+，負值推向 Below 51；紅/藍代表特徵值高/低。資料來源：本研究繪製，對應 `voice_age_binary.ipynb` Cell 9 （SHAP 分析，方向已修正為 $-$shap_values$[:, :, \text{below_idx}]$）。

5.3 醫學理論驗證

Table 6: SHAP 結果與醫學理論對照驗證

特徵	SHAP 觀察	醫學預期	驗證
HNR	低 HNR（藍）推向 51+	老化後閉合不全，HNR 降低	✓ 一致
Jitter	高 Jitter（紅）推向 51+	神經退化後升高	✓ 一致
F0_mean	SHAP 方向混合	男升女降，相反	✓ 符合
Shimmer	SHAP 值接近零	職業訓練掩蓋退化	△ 訊號弱
$MFCC_{8–12}$	影響力最高特徵群	聲道共鳴老化反映中階 MFCC	✓ 合理

資料來源：本研究依 SHAP 分析結果及相關文獻 [4, 5] 整理。

HNR 方向完全符合「老化 → 閉合不全 → 漏氣 →HNR 降低」因果鏈 [4]。Shimmer 訊號弱係 CREMA-D 演員職業訓練掩蓋老化退化，屬資料集選擇偏差。

6 互動展示系統

6.1 系統架構

本研究以 **Gradio** 建立端到端互動展示系統 `https://github.com/chelsietao/vocal-age-analyzer`，資料流如下：

麥克風錄音 / 上傳 WAV ⟶ 特徵提取（16 維）⟶ XGBoost 推論 ⟶ SHAP 計算 ⟶ 圓餅機率圖 + SHAP 條形圖 + 文字解讀

6.2 系統功能

- **輸入**：麥克風即時錄音或上傳 WAV（建議 3 秒）
- **輸出 1**：預測類別文字及各類別機率
- **輸出 2**：圓餅機率圖（顯示預測信心度）
- **輸出 3**：SHAP 條形圖（橘色推向 51+，藍灰色推向 Below 51）
- **輸出 4**：F0、Jitter、Shimmer、HNR 臨床意義說明
- **底部科普框**：Presbyphonia、MFCC、SHAP 原理說明
- **中英雙語**：支援中文／ English 切換

6.3 使用者隱私聲明

展示系統會請使用者即時錄音或上傳一段約 3 秒的語音檔，這段語音屬於使用者本人新提供的生物特徵資料，與訓練資料（CREMA-D）的隱私性質不同，因此額外說明如下：

- 使用者上傳或錄製的語音僅用於即時特徵提取與模型推論，計算完成後即從記憶體中釋放，**不會被儲存、紀錄或用於再訓練**。
- 系統介面會以中英雙語明確告知上述資料處理方式，使用者可自行決定是否提供語音樣本。
- 系統輸出僅為統計性的風險指標，不構成醫療診斷，相關說明亦會顯示於介面中（詳見第 8.1 節）。

7 研究限制與未來方向

7.1 現有研究限制

將整個研究流程從資料來源、樣本結構到模型表現逐層回顧後，可以發現本系統的限制其實彼此牽連，前一項限制往往是後一項限制成立的條件，呈現出一條層層推進的因果鏈。

首先，CREMA-D 本身是為情緒辨識研究所蒐集的語料庫，其錄音設計的出發點，從一開始就不是為了凸顯老化相關的聲學變化，而是為了讓演員表現出各種情緒，這個源頭上的設計差異，使得情緒語音中混入的訊號，會與老化訊號疊加在同一段聲音之中，即使本研究選用了理論上對老化較為敏感的特徵，這層干擾仍會持續存在，並進一步影響到後續的樣本分布。

接著，當樣本依 51 歲為界進行分組之後，51+ 樣本僅佔全體的 18.7%，這個源自資料蒐集設計的不平衡，會直接限制模型對少數類別的學習機會，即使訓練時引入 `scale_pos_weight` 進行權重補償，也只能在既有樣本分布之下調整損失函數的計算方式，無法從根本上補足少數類別在特徵空間中所缺乏的多樣性，而這又會與接下來的演員樣本特性產生疊加效應。

再者，CREMA-D 中的演員皆為職業演員，長期的發聲訓練使其聲帶健康狀況普遍優於同齡的一般民眾，於是模型在 51+ 這個樣本量本就稀少的類別中，所學習到的老化樣貌，便進一步被侵蝕，向著較為健康、退化程度較輕的一端偏移，當模型未來應用於一般人群時，這種雙重偏移會使其準確率面臨下降的風險。

與此同時，由於 F0 的老化方向會因性別而呈現相反的趨勢，將兩性樣本混合訓練，等同在同一個特徵維度中疊加了兩種方向相反、彼此抵銷的訊號，這層性別效

應未被分離的限制，會與前述的類別不平衡及樣本偏差相互交織，共同壓縮模型可以從資料中提取的有效資訊量。

最後，回到最基礎的樣本規模上來看，91 位個體所構成的資料集，本身在統計意義上仍屬有限，這個源頭上的規模限制，正是 CNN 等需要大量資料的方法在本任務中表現不佳的根本原因，也是前述四項限制之所以層層放大、難以單獨解決的共同基礎。

7.2 未來研究方向

承接上述限制之間環環相扣的關係，後續研究若要逐步突破現有瓶頸，較為合理的路徑是先從資料來源的根本問題著手，再逐層處理樣本結構與模型設計層面的限制。

首先，最根本的突破口在於建立或取得真正以老化為設計目標的語音語料庫，例如採用 TORGO、SVD 等資料集，或進一步蒐集臨床端的真實語料進行遷移學習，唯有當訓練資料的源頭不再受情緒語音干擾，後續特徵所承載的訊號才有機會更貼近真實的老化現象。

在資料來源問題逐步改善之後，可以進一步透過多任務學習的方式，讓模型在同一個訓練過程中，同時學習預測年齡組別與情緒類別，藉由這種同步學習的設計，模型有機會自行學會將兩種訊號分離，從而降低情緒語音對老化判斷的干擾，而這層分離能力，又會為後續的個人化分析奠定基礎。

當模型具備分離不同訊號的能力之後，若能取得同一個體在不同年齡段的縱向錄音，便可進一步建立個人化的老化軌跡，這種縱向追蹤的設計，能夠將原本橫斷面的群體比較，轉化為針對個體生理變化的長期觀察，並為臨床應用提供更貼近個案的參考依據。

在語音資料的限制逐步緩解之後，可以再嘗試將語音與臉部影像等其他生理訊號進行多模態融合，透過不同模態之間的互補資訊，為老化判斷提供更全面的依據，而這種多模態架構，最終仍須回到臨床場域中接受檢驗。

因此，整條研究路徑的最後一步，是與耳鼻喉科合作，以內視鏡檢查結果作為黃金標準，對模型在各個階段所累積的改進進行整體驗證，確認資料來源、訊號分離、個人化追蹤與多模態融合等層層推進的設計，最終是否真正提升了模型對聲帶老化狀態的判斷效度。

8 倫理考量與負責任 AI

8.1 診斷輔助而非替代診斷

本系統在定位上屬於**篩查輔助工具（screening aid）**，其輸出結果應被理解為一種統計層面的風險指標，這個指標的存在，目的是在大規模、低成本的情境下，先行標示出聲帶狀態可能值得進一步關注的對象，再由耳鼻喉科醫師透過內視鏡等專業檢查，對這些對象進行確認與診斷，形成一條由初步篩查到專業確診、逐層收斂的流程。系統介面亦會明確告知使用者，訓練資料來自情緒語音資料集，本身存在已知偏差，輸出結果僅供參考，不應被視為醫療結論。

8.2 資料隱私與演算法偏差

本研究在資料隱私的處理上，依照資料來源的不同，可以拆分為訓練資料與展示系統使用者資料兩個層次，這兩個層次彼此獨立，卻又共同構成本系統在隱私面向上需要說明的整體圖像。

在訓練資料的層次上，CREMA-D 是公開的學術資料集，其授權範圍允許用於非商業學術研究，本研究僅使用其中經由特徵提取所得的聲學特徵，不涉及對演員個人的額外身分識別，亦不會重新分發原始音檔，這層處理確保了訓練資料的使用，始終停留在授權範圍之內。

在展示系統使用者的層次上，現場所收集的語音屬於生物特徵資料，可能受到 GDPR 與台灣個人資料保護法等規範保護，為了回應這層風險，展示系統採本地即時運算，使用者提供的語音於推論完成後立即釋放、不予儲存，這項處理方式已於第 6.3 節說明，並會呈現於介面中，使使用者在錄音之前，便能清楚了解資料將如何被處理。

在這兩個層次之上，訓練資料本身仍存在若干潛在偏差，這些偏差會由資料來源逐層向下影響模型的泛化能力，包括樣本以北美英語母語者為主所形成的種族偏差，職業演員聲帶健康狀況普遍優於一般人群所形成的職業偏差，以及樣本中男女比例與各年齡層分布未必均衡所形成的性別偏差，這些偏差與第 7.1 節所述的研究限制相互對應，共同構成模型在實際應用時必須被持續監控的風險來源。

8.3 可解釋性作為信任基礎

SHAP 在本研究中所扮演的角色，可以視為一條串連模型輸出與醫學理論的橋樑，這條橋樑使得每一筆預測，都能被拆解為各個特徵的貢獻度，並逐一對應至第 1 章所描述、醫學上可驗證的生理機制，而不僅僅停留在一個輸出數值上。當 SHAP 顯示低 HNR 將預測推向 51 歲以上時，臨床醫師便可以依據其對 Presbyphonia 的專業知識，獨立檢視這個對應關係是否合理，並在必要時提出質疑或反駁，這種可被專

業者逐層檢驗的特性，正是建立**有依據的信任（evidence-based trust）**的核心所在，也是 AI 系統未來在醫療場域中，邁向可信賴部署的重要基礎。

9 結論

本研究成功建立聲帶生理年齡二元分類系統，主要貢獻如下：

1. **高準確率**：XGBoost 達 85%+，較隨機基準提升 22 個百分點。
2. **可解釋性驗證**：SHAP 確認 AI 決策與 Presbyphonia 醫學文獻一致，特別驗證 HNR 降低與 Jitter 升高的核心地位。
3. **方法論比較**：三模型對照驗證小規模資料下手工特徵 + 集成樹 $\succ$ 深度學習。
4. **應用展示**：Gradio 雙語系統為非侵入性聲帶篩查提供概念驗證。

本研究體現 **AI for Science** 核心精神：不僅追求準確率，更強調可解釋性與醫學理論忠實度，為醫療 AI 的可信賴部署奠定基礎。

參考文獻

References

[1] Cao, H., Cooper, D.G., et al. (2014). *CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset*. IEEE Trans. Affective Computing, 5(4), 377–390.

[2] Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD 2016), 785–794.

[3] Lundberg, S.M., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. Advances in Neural Information Processing Systems 30 (NeurIPS 2017), 4765–4774.

[4] Kendall, K. (2007). *Presbyphonia: A Review*. Current Opinion in Otolaryngology & Head and Neck Surgery, 15(3), 137–140.

[5] Teixeira, J.P., Oliveira, C., & Lopes, C. (2013). *Vocal Acoustic Analysis—Jitter, Shimmer and HNR Parameters*. Procedia Technology, 9, 1112–1122.

延伸閱讀／網路資源

以下連結為撰寫過程中參考的科普與技術說明資源，性質與上述學術文獻 [1, 2, 3, 4, 5] 不同，故另列於此，不計入正式引用編號。

機器學習演算法：

- IBM：什麼是 XGBoost？https://www.ibm.com/topics/xgboost
- IBM：什麼是 Random Forest？https://www.ibm.com/topics/random-forest

可解釋性 AI：

- IBM：什麼是可解釋性 AI？https://www.ibm.com/topics/explainable-ai
- SHAP GitHub 專案頁面（含說明與範例）：https://github.com/shap/shap

聲學特徵與語音處理：

- Wikipedia：Mel-frequency cepstrum（MFCC 說明）：https://en.wikipedia.org/wiki/Mel-frequency_cepstrum
- Wikipedia：Jitter（聲音／訊號處理意義）：https://en.wikipedia.org/wiki/Jitter

嗓音醫學背景知識：

- Mayo Clinic：嗓音疾病總覽（Voice disorders）：https://www.mayoclinic.org/diseases-conditions/voice-disorders/symptoms-causes/syc-20353022
- Merck Manual：喉部疾病（Laryngeal Disorders）：https://www.merckmanuals.com/professional/ear,-nose,-and-throat-disorders/laryngeal-disorders

附錄 A：生成式 AI 協作使用說明

本作品於製作過程中使用生成式 AI 工具（Claude，Anthropic）協作，茲說明使用比例與範圍如下：

- **由作者自行完成（核心研究內容）**：研究主題與問題定義（以 51 歲為分界）、特徵工程方向（包含移除 $MFCC_1$ 的判斷依據）、實驗設計與版本演進規劃、模型訓練結果之分析與解讀、SHAP 結果與醫學理論的對照驗證、研究限制與未來方向之規劃，以及全文之最終定稿與內容把關。
- **AI 協作部分（依產出類型估算比例）**：
  - **程式碼**：全專案（含三分類訓練、二元分類、CNN 對照實驗、Gradio 展示系統共四份 Notebook）約 600–800 行程式碼中，AI 協助進行除錯、語法修正與效能優化的部分約佔 20–25%；主要集中於 SHAP 方向公式的修正、XGBoost 參數調整建議，以及 Gradio 介面的排版與雙語切換邏輯。核心的特徵工程邏輯、模型選擇與評估流程設計由作者自行撰寫。
  - **書面報告文字**：全文約 16 頁、共九章。AI 協作主要集中於文字潤飾、段落結構調整與 LaTeX 排版（約佔全文篇幅 10–15%）；各章節之研究內容、數據分析、SHAP 與醫學理論的對照論述，以及第 1、7、8 章之論述脈絡，均由作者撰寫初稿後再請 AI 協助校對與排版調整。
  - **附屬文件**：README（中英文）、簡報大綱等文件之初稿與雙語翻譯，AI 協作比例較高（約 50% 以上），主要因為這類文件性質偏向資訊整理與格式轉換，內容均取自報告本文已完成之研究成果。
- **聲明**：上述比例為作者依製作過程回顧後之估算，並非精確統計數值。所有 AI 協作內容均經作者檢視、理解並修改後採用，研究的核心判斷、資料分析結果與結論皆為作者本人之學術成果，作者對全文內容負完整學術與著作責任。