二、研究計畫內容：

(一) 摘要

本計畫旨在探討如何運用深度學習技術處理語音側錄與 Deepfake 偽造音辨識，特別著重處理具 VoIP(Voice over Internet Protocol, VoIP)及 PSTN(Public Switched Telephone Network, PSTN)或其他網路傳輸的通道效應下語音欺騙（spoofing）所引發的資安問題。為了解決此問題，我們以 ASVspoof(Automatic Speaker Verification Spoofing And Countermeasures Challenge, ASVspoof)系列競賽最新的 Spoofing and Anti-Spoofing 語料庫為基礎，其中最近幾年特別注重模擬通道效應的語料收集。本計畫透過端點偵測確定音訊的開始和結束，我們選擇採用總體經驗模態分解法 (Ensemble Empirical Mode Decomposition, EEMD)，將語音資料分解成多個本質模態函式（IMF）。接著，利用基因演算法 (Genetic Algorithm, GA) 計算各 IMF 的權重，以此去除偽造音，其中包含主要通道效應成分，有助於提升具通道效應之真假音的辨識效能。

在特徵值計算方法中，我們將以兩種不同的方法進行訓練並研究分析，分別採用梅爾倒頻譜參數（Mel-Scale Frequency Cepstral Coefficients, MFCC）和 CQT(Constant-Q Transform, CQT)方法，此兩種特徵是在 ASVspoof 競賽中效果較好的特徵。另外在訓練過程中，我們會依照過去的研究針對不同語音的情況設計不同的語音辨識模型。首先，針對側錄音我們將使用 ResMax 模型及 transfer learning 的策略，先在大規模語音數據上預訓練 ResMax 模型，並且為了減少模型的複雜性，同時保持性能，以便可實現邊緣計算我們引入深度可分離卷積，再將其權重遷移到目標模型中，經過預訓練好的 ResMax 模型將嵌入到我們的模型結構中，再透過長短時記憶網路 (Long Short-Term Memory, LSTM) 進行模型參數最佳化。另一方面，針對 Deepfake 偽造音，我們將在辨識模型中使用對 Deepfake 偽造音效果較好的 ResNet18 模型，同樣的為了減少模型的複雜性，同時保持性能，我們引入深度可分離卷積，以達到最佳的輕量化模型。

最後為了更深入了解模型的預測基準，我們引入 Grad-CAM(Gradient-weighted Class Activation Mapping, Grad-CAM) 分析方法。所設計的系統的辨識效果將以 ASVspoof2021 資料庫來進行實驗，並將應用到現實層面中，結合生成式 AI 製造的 Deepfake 偽造音，進行欺騙音辨識，以驗證我們所提出的方法在解決語音欺騙問題上的有效性。

(二) 研究動機與研究問題

在科技的飛速進步下，語音辨識技術已廣泛應用於生活中的各個領域，包括語音導航、室內裝置控制、語音解鎖、語音助理等。然而，這種便利也伴隨著相應的資安問題，特別是在語音側錄與 Deepfake 偽造音的情境下。語音辨識技術面臨許多資安挑戰，首先於 Deepfake 偽造音中，近年來，隨著 AI 模仿音技術的進步，帶給人更多的便利外，卻也導致更多的資安風險，AI 模仿音之技術被不肖人士拿來賺取不法之財，其中最影響一般民眾的為在手機通話中可能聽到極為相似自己家人或朋友聲音的情況。這種現象的背後可能隱藏著嚴重的資安風險，因為有心人士可以利用這樣的技術進行語音詐騙。以此為例，許多人可能因認為通話中的聲音來自親友而受到欺騙，可能導致財務損失，甚至危及生命安全。

另一方面在語音側錄攻擊可能帶來的身份竊取和私人資料泄露問題。語音辨識相較於其他生物特徵容易被竊取，這使得語音辨識的安全性一直受到威脅。目前，即使語音辨識技術已廣泛應用，但由於安全考量，其在公共場所、身分認證或具通道效應的裝置上的使用仍然受到限制。

因此，本研究計畫的動機在於探討語音側錄及 Deepfake 偽造音的資訊安全問題，並尋找解決方案以降低語音辨識技術的資安疑慮。我們計劃結合語音信號的處理和深度學

習技術，特別針對偽裝音(spoofing)進行辨識，以確保語音辨識在使用時能夠準確判斷語音是否為側錄或 Deepfake 偽造音，同時提高其在公共場所、身分認證或具通道效應的裝置上的可應用性與判斷的準確性。這將有助於解決資訊安全問題，提升語音辨識技術在日常生活中的實用性。

(三) 文獻回顧與探討

對於欺騙音(spoofing)的研究是一個相對較新的研究主題，它是一種利用 AI 技術，模擬特定對象的聲音，製造高仿真的與音訊號，用於欺騙自動化語者驗證(ASV)系統或人類聽者的一種行為，例如社會事件中「台灣刑事局破獲首起 AI 語音詐欺案-警方盤點至少有 70 人受騙上當[17]」與關於語音側錄與 Deepfake 偽造音相關論文[1-9]。而語音辨識研究的步驟大致分為準備語音資料庫，進行語音訊號分析，語音前置處理，訓練語音模型及進行辨識等。

近年來，自動語者身分驗證 (Automatic Speaker Verification, ASV) 系統的研究不斷進行。從 2015 年開始，引入了一個名為 ASVspoof 系列的競賽[5-9]，該競賽是首個專注於自動語者身分驗證欺騙和對策的挑戰賽。該競賽提供了真實錄製和欺騙合成的語料庫，被廣泛用於 ASV 相關研究，其中最新的語料庫引入了通道效應下的真假音，而本計畫即以此語料庫作為實驗的主要資料來源。

擁有語料庫後，我們使用經驗模態分解法（Ensemble Empirical Mode Decomposition, EEMD）以提取人聲，同時消除偽造音傳輸通道效應。EEMD 是由黃鍔博士 (Dr. Norden Huang) 於 1998 年提出的語音訊號分析方法[10]，其將語音訊號分為多個本質模態函式(IMF)，透過排列組合這些 IMF，以取得更優越的語音比對效果。然而在執行 EMD 的分解過程中耗時冗長。因此，當分解出的語音訊號的標準差（SD）值介於 0.2 到 0.3 之間時，我們將此訊號視為 IMF。若 SD 值不在 0.2 到 0.3 之間，則將語音訊號以前一次求得 IMF 所得訊號扣除平均線，反覆執行此步驟直到獲得 IMF。接著，本計畫運用基因演算法 (Genetic Algorithm, GA) 來計算每個 IMF 的權重值。將 EEMD 與 GA 結合，重新組合從語音資料庫中分解出的 IMF 並取得最佳權重值以消除通道效應，進一步提升語音資料庫的辨識效能。

在特徵計算上，本計畫預定採用的語音特徵值擷取方法有兩種，將透過實驗比較出較好的特徵擷取方法，第一種方法為目前語音辨識常見的梅爾倒頻譜參數 (Mel-Scale Frequency Cepstral Coefficients, MFCC)，此法考慮人耳對於不同頻率的感受度差異，使參數能夠相較其他語音特徵值擷取法更接近人耳的聽覺[11]。第二種方法為 CQT(Constant-Q Transform, CQT)，它是一種能夠更有效捕捉語音信號豐富頻率特徵的方法。CQT 採用變寬的頻率區段，使其更符合人耳對聲音高低的感知。相較於 MFCC，CQT 能夠更細緻地捕捉語音信號的細節，尤其在高頻段的表現相對較優[9]。

在本研究之語音辨識模型訓練中，我們參考過往研究中，較好的方法，並結合過去實驗中常用的方法，來提升辨識結果的準確度，並針對側錄音與 Deepfake 偽造音提供兩種不同的方法進行訓練。(如圖 1 所示)

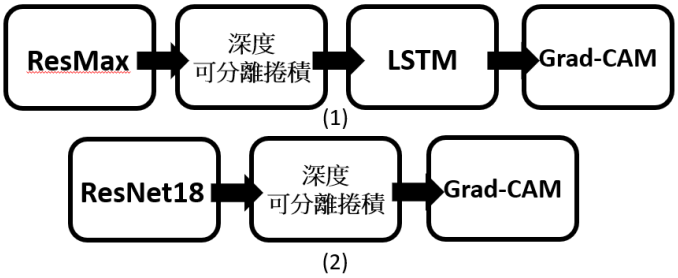

圖 1　上圖(1)為側錄音辨識模型設計，下圖(2)為 Deepfake 偽造音辨識模型設計

首先，針對側錄音的辨識模型中，我們引入了 ResMax 模型，此為一種 ResNet 模型，該架構結合了 ResNet 的跳躍連接概念和 LCNN 的最大特徵圖概念。這種結合的方法在過去 Kwak et al.的研究中證實，針對側錄音它可以減少模型的複雜性，同時保持高準確性[9]。並且為了維持模型的性能並減少複雜性，我們引入了深度可分離卷積。這種卷積操作以更高效的方式進行特徵提取，減少模型的參數數量，同時提高運算效率。深度可分離卷積的使用有助於在保持模型性能的同時減輕計算成本[9]。接下來，我們運用長短時記憶網路（LSTM）來處理語音特徵。LSTM 被引入以處理語音信號的時間序列特性，改進模型對上下文的理解，解決迴圈神經網路(Recurrent Neural Network, RNN)無法長期依賴的問題（分析過去出現的字詞以預測遺漏的字句）[12]，進而增強模型對語音特徵的捕捉和預測能力。

其次針對 Deepfake 偽造音，我們參考 Pan et al.的研究發現，當我們運用了 Resnet18 模型，以透過跳躍連接（skip connections）解決梯度消失的問題[19]，可提升 Deepfake 偽造音的辨識效能並且為了維持模型的性能並減少複雜性，我們引入了深度可分離卷積[18-19]，相對於側錄音辨識模型的設計，Deepfake 偽造音辨識模型更加重視輕量化模型訓練。

這些技術的整合使得我們的模型能夠更好地應對偽造音辨識中的複雜情境，提高辨識效果、輕量化模型和泛化性能。

(四) 研究方法及步驟

本計畫的研究方法及步驟如圖 2 所示，首先要建立語音側錄及 Deepfake 偽造音資料庫，再以總體經驗模態分解法(Ensemble Empirical Mode Decomposition, EEMD)將語音資料分解成多個本質模態函式（IMF），並結合遺傳演算法(Genetic Algorithm, GA)來提出具有側錄音及 Deepfake 偽造音特質的語音以消除通道效應。接下來，為了找出較佳的特徵計算的方法，本計畫分別對語音訊號進行語音前置處理後使用 MFCC 與 CQT 方法做特徵值運算並比較辨識效果。

而在辨識模型中，如同前一節所描述為了可以針對不同情況達到更好的辨識效果與輕量化模型，因此將針對不同欺騙音使用不同方法，針對側錄音將額外引入 transfer learning 的策略來訓練辨識模型，首先在大規模語音數據上預訓練 ResMax 模型，並將其原有的卷積層替換為深度可分離卷積，再將其權重遷移到目標模型中，並透過 LSTM 可以將 ResMax 的參數最佳化。而針對 Deepfake 偽造音辨識模型，將引入 Resnet18 模型結合深度可分離卷積，可提升辨識率。

最終便能夠透過 Grad-CAM 分析方法判斷模型辨識效果，最終可運用此模型判斷語音測資是否為側錄的語音。

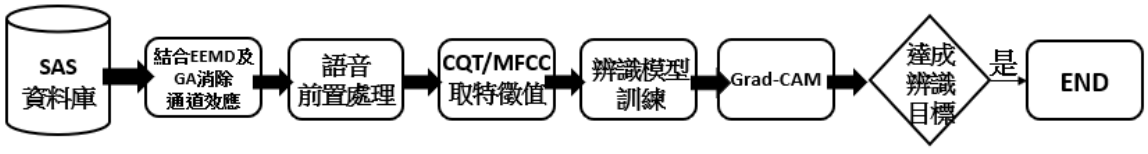

圖 2　本計畫研究方法及步驟示意圖

以下說明各個步驟的細節:

1. 實驗資料庫

本研究的資料將採 ASVspoof 系列競賽中所提供的 Spoofing and Anti-Spoofing(SAS)語料庫。由於最新的比賽—ASVspoof 2023 年的資料庫尚未發布，因此將採用 ASVspoof 2019 與 ASVspoof 2021 的語料庫，另外相對於 2019 的資料庫 2021 為多個貢獻者提供的，

他們使用了不同的語音合成和語音轉換算法來產生欺騙的語音。真實的語音是由 200 位人類實際錄製（包含 100 位男性，100 位女性），並且涵蓋了 10 種語言。欺騙的語音是由真實的語音修改而來，或者是由文本合成而成。每個任務的欺騙語音都有不同的攻擊方法和難度[13]。

2. 提取偽造音，消除通道效應

針對不同的任務(側錄音與 Deepfake 偽造音)及不同通道(VoIP 及 PSTN)傳輸，本計畫結合 EEMD 與 GA，將通道效應消除，以提取偽造音，而後計算不同特徵值，並以前述不同模型進行辨識，詳細流程如圖 3 所述。

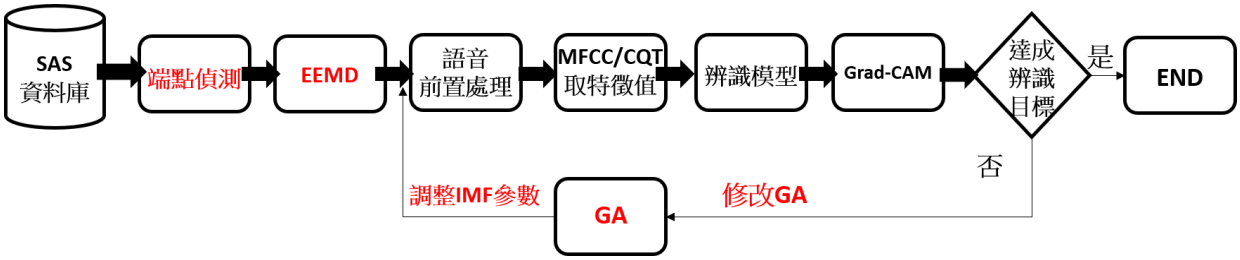

圖 3　消除通道效應提取偽造音步驟

(1) 端點偵測

端點偵測識別語音訊號的開始和結束點，以進一步處理和分析。時域法簡單且廣泛，但抗雜訊較差。頻域和混合法複雜但準確。本研究選擇時域法以降低計算負擔，儘管抗雜訊較弱，但在研究情境下合理。

(2) 總體經驗模態分解法(Ensemble EMD)訊號分解：

語音訊號具有間歇性，進行 EMD 分解時常遭遇模態混雜問題，即同一 IMF 中可能包含不同尺度訊號，或相同尺度訊號分布於不同 IMF 中。Wu and Huang [14]改進了 EMD，提出 Ensemble EMD 方法，以白噪音消除模態混雜。在 Ensemble EMD 中，須事先選定白噪音強度和加總平均次數。將 m 筆獨立但強度相同的白噪音與 n 筆原始訊號相加，經 m 次 EMD 分解後取平均，有效降低模態混雜，避免不同時間尺度訊號出現在同一 IMF。本計畫選用 Ensemble EMD 取得 IMF，並進行深度學習模型的訓練與側錄語音辨識。

(3) 基因演算法 GA

本計畫以改良式基因演算法進行 IMFi 權值 wi 的最佳化。染色體編碼使用實數表示每個 wi，wi 為相對本質模態函式 IMFi 的權重，n 為模態函式個數。改良式基因演算法可使染色體交配分散在整個搜尋空間上，避免集中於特定區域以得到更佳的染色體。 本計畫的適應值計算使用語音欺騙辨識率，辨識率越高的染色體存活率越高。

3. 語音前置處理

在語音前置處理中，本計畫所使用的前置處理步驟如圖 4 所示。

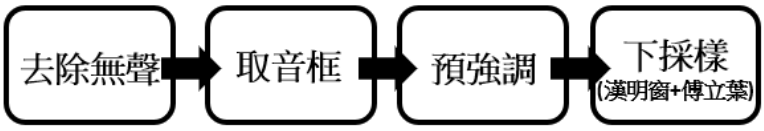

圖 4　語音前置處理步驟

4. 梅爾倒頻譜參數(MFCC)抽取特徵值:

由於本計畫想比較兩種計算特徵值的方法，以獲得更好的效果，因此會分成 MFCC 與 CQT 兩種特徵值做同步研究，以下將先解釋 MFCC 的方法。

首先梅爾頻率與一般頻率的轉換公式如下(式(4.1)及(4.2))：

$$\text{mel(f)} = 2595 * log_{10}\left(1 + \frac{\text{f}}{700}\right) \tag{4.1}$$

$$\text{f} = 700 * \left(10^{\frac{mel}{2595}} - 1\right) \tag{4.2}$$

其中 mel 為梅爾頻率， f 為一般頻率。藉由式(4.1)及(4.2)由梅爾頻率推導出梅爾三角帶通濾波器(4.3)：

$$f_m = \frac{N}{f_s} * f\left(mel(f_l) + m * \frac{\text{mel(f}_\text{h}) - mel(f_l)}{M+1}\right) \tag{4.3}$$

其中 f_s 為語音訊號的取樣頻率，f_l 為濾波器中最低的頻率，f_h 為濾波器中最高的頻率，f_m 為梅爾三角帶通濾波器的中心頻率。訊號經過帶通濾波器後的數值對應如下(4.4)：

$$A_m(k) = \begin{cases} 0, & k < f_{m-1} \\ \frac{k - f_{m-1}}{f_m - f_{m-1}}, & f_{m-1} \le k \le f_m \\ \frac{f_{m+1} - k}{f_{m+1} - f_{m-1}}, & f_m \le k \le f_{m+1} \\ 0, & f_{m+1} < k \end{cases} \tag{4.4}$$

訊號經由梅爾三角帶通濾波器後取得人耳聽覺比較敏感的頻率，且將取得 $A_m(k)$的值與每個頻率的能量加總相乘後再取對數，如下列式子(4.5)：

$$\text{Y(m)} = \log\left\{\sum_{k=f_{m-1}}^{f_{m+1}} |X(k)|^2 \times A_m(k)\right\} \tag{4.5}$$

其中 X(k) 為經由快速傅立葉轉換(Fast Fourier Transform, FFT)運算後所獲得的頻域訊號，$A_m(k)$為經由公式(4.4)帶通濾波器後所對應的數值。

最後，進行離散餘弦轉換 (Discrete Cosine Transform, DCT)，如公式(4.6)：

$$\text{D(n)} = \frac{1}{M}\sum_{m=1}^{M} Y(m) \cos\frac{\pi n\left(m - \frac{1}{2}\right)}{M} \tag{4.6}$$

MFCC 特徵值通過以上一連串的取樣轉換後即可獲得。

5. 常量 Q 變換(Constant Q Transform,CQT)抽取特徵值

針對所需處理的語音，已完成語音前置處理、取音框、預強調、漢明窗處理以及傅立葉轉換後，而接著為 CQT 實作步驟:

A. 常數 Q 變換：將傅立葉轉換後的頻域訊號進行 CQT。CQT 的特點是其頻率解析度與頻率成比例，這使得在低頻區域有更高的頻率解析度，而在高頻區域有更高的時間解析度。

B. 特徵抽取：從 CQT 的結果中抽取特徵。

C. 後處理 ： 對抽取的特徵進行可能需要的後處理。

CQT 的基本形式可以表示為(5.1):

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-\frac{j2\pi kn}{N}} \tag{5.1}$$

其中，x(n) 是輸入信號，N 是信號的長度，k 是頻率索引，X(k) 是 CQT 的結果。

這裡的頻率索引 k 並不是線性的，而是按照一個常數 Q 因子進行尺度變換的。具體來說，對於每個 k，其對應的頻率 fk 可以表示為(5.2)：

$$f_k = f_0 \times (2^{\frac{1}{Q}})^k \tag{5.2}$$

其中，f0 是參考頻率（通常選擇為最低頻率），Q 是常數 Q 因子，決定了頻率解析度的變化速度。

6. 辨識模型

根據過去的研究中，在 Kwak e tal.的研究中證實，針對側錄音，使用 ResMax 可以減少模型的複雜性，同時保持高準確性[9]，而針對 Deepfake 偽造音，Pan et al.的研究中發現使用 ResNet18 將有更佳的效果，因此我們將依據不同情況這記不同的語音辨識模型。

a. 語音側錄辨識模型

側錄音是指未經授權或未經許可而在數位通信系統中截取、監聽或記錄音訊的行為。這種行為可能涉及在電話、VoIP 通話、網路音訊傳輸或其他數位通信渠道中，以某種方式監聽或錄製對話、音訊或其他數位資訊。為此，我們將針對以上可能的側錄音數據建構如圖 5 之模型。

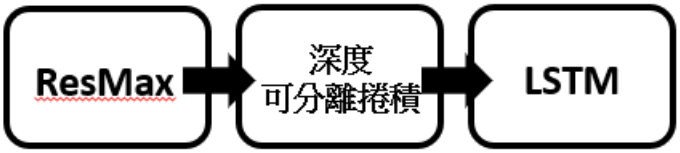

圖 5　語音側錄辨識模型步驟

(1) ResMax

在本研究中，對於語音辨識模型的構建，首先引入了 ResMax 模型，採用 ResNet 的跳躍連接概念和 LCNN 的最大特徵圖概念。這種融合的結構被證實能夠減少模型的複雜性，同時維持高準確性。ResMax 模型的預訓練優勢使其在大規模語音數據上學習通用特徵，提供了更強的基礎。而以下將是我們參考論文[9]所設計的 ResMax 的方法進行設計，如圖 6。

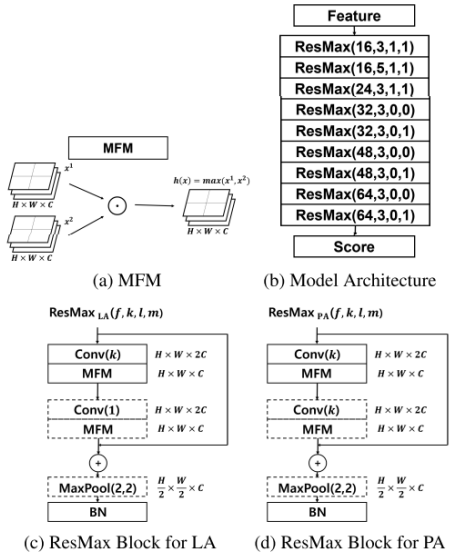

圖 6　ResMax 模型設計[9]

圖 6 中 (a) 代表一個 MFM 層；(b) 描述了整個模型架構；(c) 和 (d) 分別代表 ResMaxLA 和 ResMaxPA，在 PA 和 LA 數據上構建整個模型的基本模塊。這些模塊有四個參數：f 是 ResMax 濾波器的數量，k 是卷積層中的核大小 (k, k)。如果 ResMax 模塊有一個額外的卷積層後面跟著一個 MFM 層（在 (c) 和 (d) 中虛線 Conv 和 MFM 模塊），則 l 為 1，否則為 0。如果 ResMax 模塊有一個額外的最大池化層（在 (c) 和 (d) 中虛線 MaxPool 模塊），則 m 為 1，否則為 0。

(2) 深度可分離卷積

由於當我們使用更深的 ResMax 模型，就會使模型的容量愈大，為了達到輕量化模型的目的，我們使用深度可分離卷積層(Depthwise separable convolution)，它是由深度卷積(Depthwise convolution)以及逐點卷積(Pointwise convolution)所組成(如圖 7)。

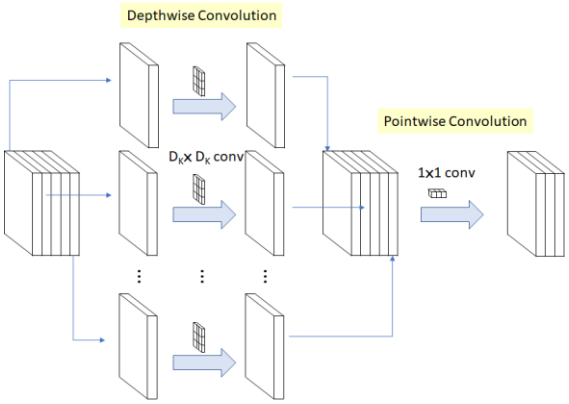

圖 7　深度可分離卷積層示意圖[18]

i. 深度卷積(Depthwise convolution):是 channel-wise 具有 DKxDK 維度的空間卷積。 透過將 channel 分組，例如將 3x3 的 kernal map 將每個 channel 的空間特徵收集起來，接著用 1x1 的 conv.從深度方向將這些特徵(Depthwise feature)生成進一步特徵。這樣的計算量會比原來的卷積少一半以上。

ii. 逐點卷積(Pointwise convolution) :實際上是透過 1x1 卷積來改變維度。

iii. 運算成本:

透過深度可分離卷積的運算成本如式子(6.1)。

$$D_K \times D_K \times M \times D_F + M \times N \times D_F \times D_F \tag{6.1}$$

透過標準卷積的運算成本如式子(6.2)。

$$D_K \times D_K \times M \times N \times D_F \times D_F \tag{6.2}$$

因此以此方式減少的計算量為如式子(6.3):

$$\frac{D_K \times D_K \times M \times D_F + M \times N \times D_F \times D_F}{D_K \times D_K \times M \times N \times D_F \times D_F} = \frac{1}{N} + \frac{1}{D_K \times D_K} \tag{6.3}$$

因此可以得出，透過深度可分離捲積可以達到輕量化模型的目的。

(3) 長短時記憶網路 LSTM

LSTM 專注於解釋按照時間序的特徵，處理語音信號的時序性資料。這樣的結合使得模型更全面地捕捉靜態和動態特徵，提高了整體效果。

b. Deepfake 偽造音辨識模型

以下為 Deepfake 偽造音辨識模型設計步驟示意圖(圖 8)。

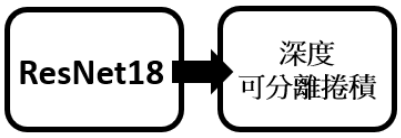

圖 8　Deepfake 偽造音辨識模型設計步驟

(1) ResNet18

ResNet（殘差網路）是一種深度神經網路架構，其關鍵在於殘差塊的設計。每個殘差塊由兩個 3x3 的卷積層組成，卷積後接批量規範化層和 ReLU 激活函數，然後將輸入直接與卷積後的輸出相加。這樣的設計使得殘差塊更容易學習殘差映射，即輸出與理想映射的差異，並且加速了訓練過程。

而 ResNet18 包含四個模塊，每個模塊包含若干個殘差塊。模塊的通道數逐漸增加，而高和寬逐漸減半。最後一個模塊後接全局平均池化層和全連接層，用於分類任務。ResNet18 具有 18 個主要層，包括卷積和全連接層，但不包括池化和批量規範化層。該網絡模型在圖像識別任務中廣泛應用，具有良好的性能和可解釋性。(如圖 9)

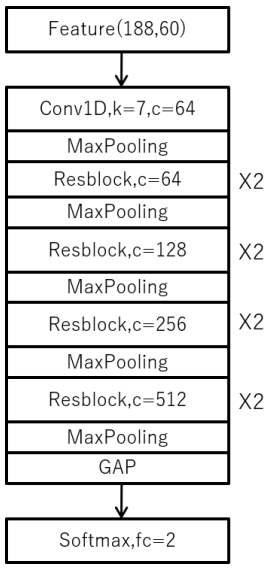

圖 9　ResNet 模型設計示意圖

(2) 深度可分離卷積

此方法同語音側錄辨識模型中深度可分離卷積之說明，在此的目的為，因使用深層的 ResNet18 模型，就會使模型的容量愈大，為了達到輕量化模型的目的，因此使用深度可分離卷積優化。

(五) 預期結果

主要目標:

(1)完成一個辨識率 90%以上的語音側錄音及 Deepfake 偽造音之辨識系統。(高辨識率)

(2)達成輕量化模型

(3)快執行速度

(4)可去除 VoIP 及 PTRN 傳輸效應，提高遠端偽造音辨識效能。

目標細節:

(1) 完成訊號預處理。

(2) 完成兩種特徵值計算並比較出較佳的方法，分別是 MFCC 與 CQT。

(3) 完成 ResMax 與深度可分離卷積運算與 LSTM 訓練。

(4) 完成 ResNet18 與深度可分離卷積運算訓練。

(5)完成 EEMD 結合 GA 進行傳輸通道效應去除工作

(六)需要指導教授指導內容

(1) 需教授指導經驗模態分解法基本操作。

(2) 需教授指導卷積類神經網路與深度可分離卷積運用於語音之基本操作。

(3) 需教授指導梅爾倒頻譜參數與 CQT 抽取特徵值之基本操作。

(4) 需教授指導長短期記憶模型之基本操作。

(5) 需教授指導 ResMax 與 ResNet18 方法。

(6) 需教授指導實驗的進行方式及內容。

(七) 參考文獻

[1] H. SHIM, J. Jung, H. Heo, S. Yoon, H. Yu, “Replay spoofing detection system for automatic speaker verification using multi-task learning of noise classes,” arXiv preprint arXiv:1808.09638, 2018.

[2] N. Evans, J. Yamagishi, T. Kinnunen, Z. Wu, F. Alegre and P.D. Leon, “Speaker Recognition Anti-spoofing,” In S. Marcel, S.Z. Li and M.S. Nixon (Eds.), Handbook of Biometric Anti-Spoofing (pp.125-146). Germany, Springer, 2014.

[3] S. H. Mankad, S. Garg, “On the performance of empirical mode decomposition-based replay spoofing detection in speaker verification systems,” Prog Artif Intell 9, 325–339 (2020).

[4] X. Xiao, X. Tian, S. Du, H. Xu, E. S. Chng, H. Li, “Spoofing speech detection using high dimensional magnitude and phase features: The NTU approach for ASVspoof 2015 challenge,” in Sixteenth Annual Conference of the International Speech Communication Association, 2015.3

[5] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, “Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge,” in 16th Annual Conference of the International Speech Communication Association, 2015.

[6] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, “The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection.”, 2017.

[7] H. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, K. A. Lee, and J. Yamagishi, “Asvspoof 2017 version 2.0: meta-data analysis and baseline enhancements,” in Odyssey 2018-The Speaker and Language Recognition Workshop, 2018.

[8] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, “Asvspoof 2019: Future horizons in spoofed and fake audio detection,” arXiv preprint arXiv:1904.05441, 2019.

[9] I.-Y. Kwak, S. Kwag, J. Lee, Y. Jeon, J. Hwang, H.-J. Choi, J.-H. Yang, S.-Y. Han, J. H. Huh, C.-H. Lee, and J. W. Yoon, “Voice spoofing detection through residual network, max

feature map, and depthwise separable convolution," IEEE Access, vol. 11, pp. 49140–49152, 2023

[10] N.E. Huang. "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," in Proc. R.Soc.London, 1998, pp. 903-995.

[11] 張智星,"12-2 MFCC" Internet: http://mirlab.org/jang/books/audiosignalprocessing/speechFeatureMfcc_chinese.asp?title=12-2%20MFCC, Feb. 2, 2020.

[12] ITREAD, "RNN 與 LSTM 之間的介紹和公式梳理" Internet: https://www.itread01.com/content/1549269013.html, Feb 4, 2019.

[13] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, H. Delgado, "ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection," in Proc. ISCA, 2021, pp. 1-5.

[14] Z. Wu and N.E. Huang, "Ensemble Empirical Mode Decomposition：A Noise Assisted Data Analysis Method,"Adaptive Data Analysis, vol. 1, no. 1 ,pp 1–41,2009.

[15] 洪川程與潘欣泰(民 107)。經驗模態分解法應用在類神經網路之語音情緒辨識(碩士論文)。取自台灣博碩士論文加值系統。

[16] 張智星(民 85). "Audio Signal Processing and Recognition (音訊處理與辨識)" Internet: http://mirlab.org/jang/books/audioSignalProcessing/, Feb. 5, 2005.

[17] 黃靖文(2023). "她被 AI 語音詐騙，甜美女聲應答超流利，假投顧拐走 2000 萬！ 7 成難辨真偽…如何能擊破「完美騙局」" Internet: https://ynews.page.link/ZeCbV

[18] Moris(2021). "MobileNetV1 — Depthwise Separable Convolution (Light Weight Model)" Internet: https://medium.com/image-processing-and-ml-note/mobilenetv1-depthwise-separable-convolution-light-weight-model-6124286ff291

[19] J. Pan, S. Nie, H. Zhang, S. He, K. Zhang, S. Liang, X. Zhang, and J. Tao, "Speaker recognition-assisted robust audio deepfake detection," Proc. Interspeech 2022, vol. 2022, pp. 4202-4206, 2022