2025年臺灣國際科學展覽會
優勝作品專輯

作品編號　190009

參展科別　電腦科學與資訊工程

作品名稱　語音模型逆向攻擊架構分析與防禦策略探討

得獎獎項　三等獎

印尼世界創新科學作品奧林匹亞競賽

就讀學校　國立花蓮高級中學
指導教師　趙義雄
作者姓名　林炫宇
莊家瑋

關鍵詞　逆向攻擊、語者辨識、模型安全

作者簡介

我們是來自花蓮高中的林炫宇（左）和莊家瑋（右），對電腦科學與資訊安全的研究充滿了興趣與熱忱。在一年多的專題研究過程中，我們經歷了無數次試錯與挑戰，雖然有精疲力盡的時刻，但也收穫了寶貴的經驗與樂趣。我們非常感謝在支持與指導我們的老師，讓我們能突破自我，朝著國際科展的目標努力前進！

2025 年臺灣國際科學展覽會

研究報告

區　　別： 中區

科　　別： 電腦科學與資訊工程

作品名稱： 語音模型逆向攻擊架構分析與防禦策略探討

關 鍵 詞： 逆向攻擊、語者辨識、模型安全

編　　號：

摘要

本研究中，我們對模型逆向攻擊在語音辨識系統中的影響及風險進行深入分析。隨著 Siri、Google Home 等智能助理設備在日常生活中的廣泛使用，其語者辨識系統的安全隱患引起了我們的注意。本研究目的在於深入理解模型逆向攻擊的運作機制，並探討其對語音辨識系統的攻擊效果。我們透過實施多樣化的攻擊策略，對不同的模型架構和數據處理方法進行了評估，並對人聲與非人聲的數據集進行了攻擊效果的比較。此外，我們亦實現了基於差分隱私的防禦算法，在多數模型架構下達到接近 50%的防禦效果，顯著提高攻擊代價。研究整體揭示了語音辨識系統在面對模型逆向攻擊時的脆弱性，並藉由實驗分析推論出可能的防禦策略，期待能通過策略來增強模型的安全性。

Abstract

Speaker recognition systems are susceptible to model inversion attacks that can reconstruct sensitive user data from model outputs, posing significant privacy risks. This study investigates the key factors influencing the success of such attacks and explores effective defense strategies. Our research is divided into two core components: implementing model inversion attacks and evaluating defense mechanisms.

For the attack analysis, we examine how input data representations, feature extraction techniques (e.g., Mel-frequency cepstral coefficients), dataset characteristics, and model architectural complexity impact attack performance. Our findings indicate that more complex model architectures present greater challenges for attackers.

In evaluating potential defenses, we demonstrate the efficacy of differential privacy, particularly for raw waveform and SincNet acoustic models. Furthermore, we show that adopting targeted feature extraction methods and refining model architectures can significantly enhance the resilience of speaker recognition systems against model inversion attacks.

Our study provides valuable insights into the vulnerabilities of speaker recognition systems and proposes robust defense strategies to mitigate privacy threats posed by model inversion attacks in communication technologies.

壹、 前言

一、研究動機

現在的生活中，像 HomePod 的 Siri 或 Google Home 這樣的智能助理已成為日常的一部分。這些設備能識別家庭成員的聲音並提供個性化服務，不僅增加了生活趣味，也提高了便利性。然而，這些智能助理存在安全隱患。如果有人通過模仿家庭成員的聲音，來控制他人的智能設備，可能會導致隱私泄露和安全風險。

我們好奇是否有一種方式能夠對抗這種風險，於是開始研究關於機器學習以及語者辨識的資料。此研究目前聚焦於深入理解與分析模型逆向攻擊(model inversion attack)相關的實現與改進方法。研究初步階段，我們主要探索這些攻擊是如何利用語音辨識系統的弱點進行攻擊。經由逐步深入研究，我們期望揭示這些系統中可能存在的漏洞，並為日後開發有效的防禦策略打下基礎。

二、研究目的

(一) 重現模型逆向攻擊方法，理解其對語音辨識系統的影響。

(二) 實施不同的攻擊策略，對其效果的差異進行比較。

(三) 採用不同攻擊參數，對其攻擊效果進行比較。

(四) 探索基於差分隱私的模型逆向攻擊防禦機制。

貳、 研究設備及器材

一、硬體環境

(一) 筆記型電腦 MacBook Pro 處理器 M2 Max 記憶體 32GB 一台

(二) ASUS GA502IU 處理器 AMD Ryzen 7 4800HS 記憶體 16GB 一台

(三) ASUS PRO E500 G6 處理器 intel i7-10700 記憶體 8GB 五台

(四) Z690 AORUS ELITE AX DDR4 處理器 intel i7-12700k 記憶體 128GB 一台

二、軟體環境

(一) Anaconda 環境下 Python librosa 音訊分析套組

(二) tensorflow, python = 3.8.11 訓練環境，Visual Studio Code 編譯器

(三) 調用 ART_MIface 函式庫

參、 研究過程與方法

一、研究流程

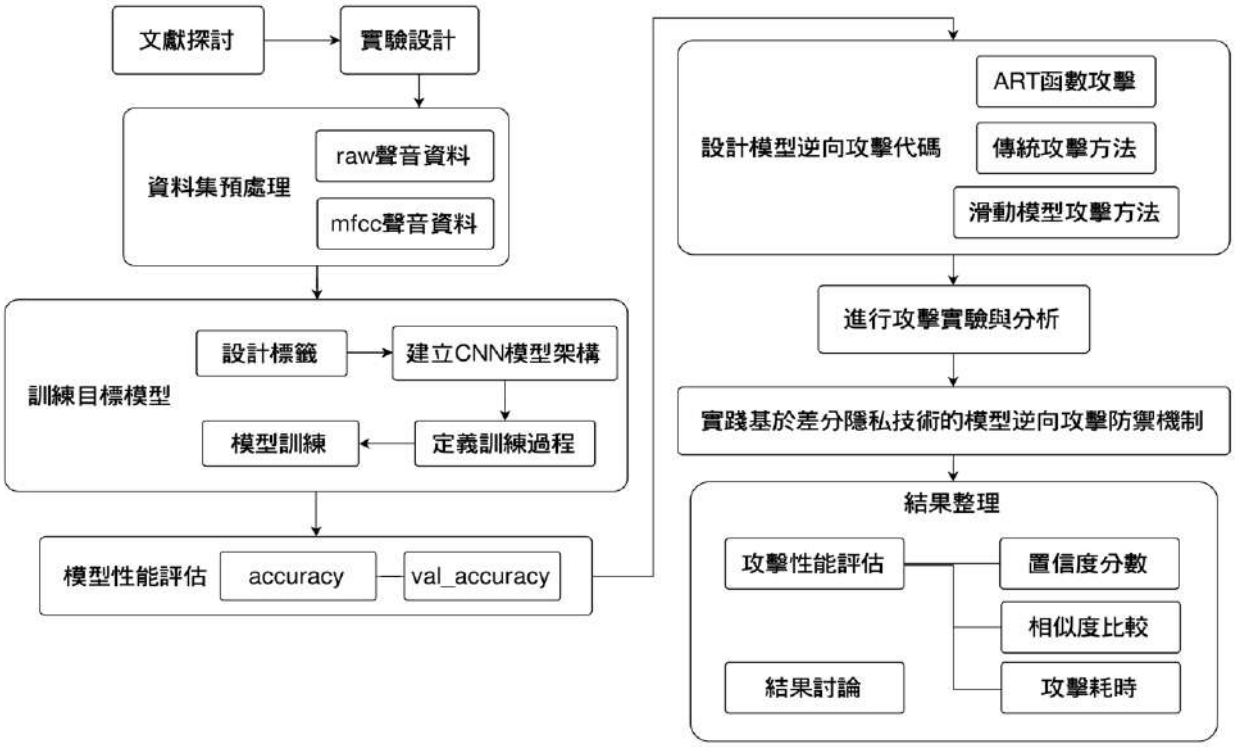

圖 3-1：研究架構圖

二、文獻探討

(一) 自動語者識別系統架構和算法

1. 主流的自動語者識別系統架構

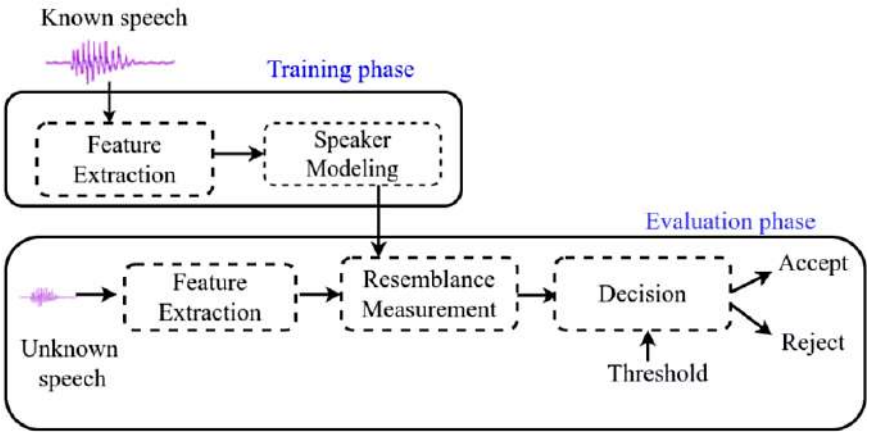

圖 3-2：語者識別系統架構圖

機器學習模型可以根據所提供的訓練數據進行自我學習，並擁有在新數據上進行預測或決策的能力。圖 3-2 為自動語者識別系統的架構圖[3]。在訓練階段，會先對已知的語音提取出特徵，然後以此建造一個語者辨識的模型。而要評估一個未知的語音時，在提取特徵後，會將其放入模型比對，然後依照語音相似程度來判斷是否為同一位語者。

2. 特徵提取技術

(1) 梅爾倒頻譜(Mel-Frequency Cepstrum, MFC)

梅爾頻率倒譜係數(Mel Frequency Cepstral Coefficients，MFCC)[4]是現今在音訊處理中廣泛使用的特徵，尤其在語音識別、語者識別和聲音檢索等領域中非常重要。此技術的想法是透過模擬人類耳朵的聲音處理機制，來提取出語音信號的特徵。MFCC 的計算過程的主要步驟為：預加重(Pre-emphasis)、分幀(Framing)、加窗(Windowing)、快速傅立葉變換(FFT)、梅爾濾波器組處理(Mel Filter Bank Processing)、對數能量(Logarithmic Energy)、離散餘弦變換(DCT)與動態特徵(Delta Features)。

(2) SincNet

SincNet 為 SPEAKER RECOGNITION FROM RAW WAVEFORM WITH SINCNET [2] 此論文中所提出用於處理語音信號的一種 CNN 架構。它的主要特點是在第一層卷積層中使用了一種參數化的 sinc 函數來卷積，實現了帶通濾波器(只允許一定範圍內的頻率信號通過)。此濾波器的低頻和高頻截止頻率是從數據中學習的，而不像傳統的 CNN 會學習每個濾波器的所有元素。因此使濾波器能夠專注於特定的頻率範圍，並加快收斂速度。

(二) 模型攻擊相關知識

1. 模型受到的攻擊類型

(1) 對抗攻擊(Adversarial Attacks)：此種攻擊涉及惡意製作的輸入，旨在誤導模型進行不正確的預測，以破壞模型的性能。

(2) 成員推斷攻擊(Membership Inference Attacks)：此種攻擊的目標是確定特定的數據點是否被用來訓練模型，以便推斷訓練數據的性質。

(3) 模型逆向攻擊(Model Inversion Attacks)：此種攻擊試圖利用模型的輸出，來不斷重建輸入的數據，直到模型輸出符合需求，以提取模型內的資訊。

2. 攻擊環境

(1) 黑盒攻擊[5]：黑盒攻擊過程中，攻擊者無法訪問模型的內部機制或參數。模型被視為一個拆不開的物件，只能從模型的輸入和輸出來調整模型的攻擊輸入。

(2) 白盒攻擊[5]：不同於黑盒攻擊，白盒攻擊的過程中，攻擊者擁有模型訪問權限，包括訪問模型的參數和各層的梯度變化資訊等。攻擊者利用這些模型內部資訊更有效地進行攻擊。

3. 模型攻擊的防禦

(1) 對抗性訓練：在訓練模型時利用不同攻擊產生更多的資料，使得模型能夠辨別受到攻擊的數據，以提高應付對抗性攻擊的抵抗力。

(2) 輸入預處理：如正規化等技術可以幫助減輕攻擊的影響。

(3) 差分隱私(Differential privacy)[6]：其為一種隱私保護技術。在此概念下，當少量的數據被添加到數據集中時，整體數據的統計結果不易明顯地發生變化。差分隱私通過添加一些隨機噪聲，使得攻擊者從輸出中難以確定數據中單個個體的資訊。然而此隱私保護會使數據因添加噪聲而不準確，所以在隱私保護與模型效能上需有所權衡。

以上為模型攻擊中常見的攻擊類型與防禦方法。其中我們的研究將以模型逆向攻擊方法為主，未來也會探討差分隱私對於模型逆向攻擊的防禦機制。

(三) 模型逆向攻擊在語者識別系統中的示例

在 Introducing Model Inversion Attacks on Automatic Speaker Recognition[1]此研究中，提出了一種新的模型逆向攻擊策略用在語者識別系統中，並將其與傳統的攻擊方法進行比較。下方將介紹此研究中對模型逆向攻擊效果的評估方式，以及所提出的攻擊策略原理。

1. 成效評估方式

在此研究的模型中，一段聲音先會經過 SincNet 提取特徵，接著經過一段深度神經網路後得出分類結果，若為正確的原始說話者，則視為成功，並以正確分類的攻擊樣本的百分比作為分類依據。

2. 攻擊策略

下方將分別介紹傳統的模型逆向攻擊策略，與此論文提出的滑動模型逆向攻擊方法。

(1) 傳統模型逆向攻擊策略

攻擊者利用機器學習模型對於預測置信度的特性，來反向推導訓練模型中的個別數據。假設有一個目標模型為 f，該模型被訓練用來將 n 維的輸入數據點 x 映射到 m 維的向量 p，p 表示每個類別的概率，即 $f: x \rightarrow p$。為了逆轉此模型，需定義一個目標函數，用以應用梯度下降法，此函數稱為成本函數 $c(x)$，基本上可代表我們與想要重建的資訊的接近程度。設定 $c(x) = 1 - p(t)$，其中 t 表示我們想要從中獲得資訊的目標。從一個隨機初始化的輸入樣本 x_0 開始，我們可計算出其成本為 $c(x_0)$。在此基礎上，便可應用梯度下降算法進行迭代，不斷對上次的輸入進行修改。目的是最小化特定類別的成本，從而使得結果的數據樣本可代表該類別。

(2) 滑動模型逆向攻擊策略

不同於傳統的模型逆向攻擊，論文的作者提出了一種新的攻擊策略：對語音數據切段，每一個區域單獨逆轉。滑動模型逆向攻擊通過迭代地逆轉重疊的數據塊來進行操作，這樣部分輸入數據已經被逆轉，因此能更接近實際的語音數據波形。

滑動模型逆向攻擊需建立一個逆轉向量並初始化，此論文研究中，使用了常態分佈隨機初始化一個長度為 ℓ 的逆轉向量 inverted [0,...,ℓ]。 接著設定一個長度固定的窗口，窗口在向量中按照固定的步長移動，針對逆轉向量上每一段窗口內的資料進行逆轉，並將逆轉的結果更新回去向量的對應段中，直到達到給定的迭代次數，或者成本函數 $c(x)$ 小於給定的閾值。最終返回去掉開頭和結尾各半窗口大小部分的逆轉向量作為輸出。

該研究的實驗結果顯示，滑動模型逆向攻擊技術能夠將反轉音頻樣本的分類準確率提高至 90%，遠超傳統模型逆向攻擊的 54%，這表明模型對攻擊的脆弱性非常高。

(四) 實驗環境與功能介紹

1. 函式庫介紹

(1) TensorFlow

TensorFlow[8]為一個由 Google 開發的開源框架，提供了一個靈活且高效的平台，能夠建立、訓練和部署各種機器學習模型。

(2) Librosa

Librosa[9]為一個用於音訊的分析處理的 Python 函式庫。常見的功能，如時頻處理、特徵提取與繪製聲音圖形等，皆包含在內。

(3) Adversarial Robustness Toolbox (ART)

ART[10]是一個針對機器學習安全性的 Python 函式庫。ART 提供的功能能夠評估、防禦、認證和驗證機器學習模型，來實現抵禦攻擊威脅的目的。ART 支援許多機器學習任務（分類、目標檢測、語音識別、生成、認證等）、所有數據類型（圖像、表格、音訊、影片等）和絕大多數的機器學習框架。

(4) Tensorflow Privacy

本實驗原先採用的防禦策略是基於 tensorflow-privacy 函式庫的差分隱私算法，這種算法主要是在訓練模型時在梯度加入噪聲機制，以此提高攻擊的難易度。然而經過我們的實驗發現，該函式庫所提供的基於梯度的差分隱私算法會大幅提升語音模型的訓練時長，且模型不易收斂。

2. 資料集介紹

(1) Timit[11]

由德州儀器、麻省理工學院和國際斯坦福研究所合作建構的連續語音資料集。採樣頻率為 16kHz，由 630 美國人說出每人給定的 10 句英語，一共 6300 個句子。在 Introducing Model Inversion Attacks on Automatic Speaker Recognition[1]論文中主要使用此資料集。

(2) Urban Sound 8k dataset[12]

由 Justin Salamon, Christopher Jacoby, Juan Pablo Bello 提供的資料集，包含 8732 個帶標籤的城市聲音，有以下 10 類：air_conditioner、car_horn、children_playing、dog_bark、drilling、enginge_idling、gun_shot、jackhammer、siren、street_music。

(3) Speaker Recognition Audio Dataset_50_speakers_audio_data [13]

由 Vibhor Jain 發布在 Kaggle 平台上。包含 50 個長度超過一小時的人聲資料集。

三、實驗設計

(一) 實驗規劃

1. **實驗一**：討論基於 MIface 函數、傳統模型逆向攻擊法、滑動模型逆向攻擊法三種攻擊手法，對於原始音訊(Raw)和 MFCC 預處理資料語音模型的攻擊效果。
2. **實驗二**：討論基於傳統模型逆向攻擊法和滑動模型逆向攻擊法，攻擊基於 Raw 架構語音模型、MFCC 架構語音模型和 SincNet 架構語音模型的攻擊效果。
3. **實驗三**：討論 TIMIT 與 50_speakers 兩種人聲資料集所訓練的目標模型，受傳統模型逆向攻擊和滑動模型逆向攻擊的攻擊效果。
4. **實驗四**：討論以類似模型架構，不同採樣長度訓練目標模型，受傳統模型逆向攻擊和滑動模型逆向攻擊的攻擊效果。
5. **實驗五**：討論 50_speakers 資料集以相同採樣長度、不同模型複雜度架構來訓練目標模型，受傳統模型逆向攻擊和滑動模型逆向攻擊的攻擊效果。
6. **實驗六**：討論 urban 與 50_speakers 資料集所訓練的目標模型，受傳統模型逆向攻擊和滑動模型逆向攻擊，分別以 urban、50_speakers 和 TIMIT 資料集內的一段音訊作為初始化樣本的攻擊效果。
7. **實驗七**：討論 50_speakers 資料集基於 Raw 架構語音模型、MFCC 架構語音模型和 SincNet 架構語音模型，加入差分隱私算法後，受傳統模型逆向攻擊和滑動模型逆向攻擊的效果，以驗證該防禦算法的可行性。
8. **實驗八**：討論 50_speakers 與 urban 資料集，以轉換成時頻譜的資料來訓練目標模型，受傳統模型逆向攻擊和 MIface 攻擊的效果。結合實驗二之結果進行分析，並討論時頻譜分類模型受模型逆向攻擊的特性。

(二) 訓練目標模型

本研究旨在評估模型逆向攻擊對語音識別模型的攻擊效果。同時討論基於白盒攻擊的 ART_MIface 函式庫和重現自 Introducing Model Inversion Attacks on Automatic Speaker Recognition[1]論文當中的傳統與滑動兩種攻擊方式，對於聲音模型的攻擊效果。

由於 Adversarial Robustness Toolbox (ART)當中 MIface 函數攻擊的目標模型在訓練時需

要使用 KerasClassifier 這個架構，因此以下實驗皆需要針對 MIface 函數訓練專屬的目標模型。

圖 3-3 為實驗一、二和三語音模型訓練架構，我們針對每個實驗訓練數個不同架構或不同資料預處理方式的模型，圖中的紫色路徑架構代表實驗一的兩種目標模型架構：以原始音訊 (Raw) 或 MFCC 預處理資料作為模型輸入；綠色路徑架構代表實驗二的三種目標模型架構：SincConv1D、MFCC_Layer 或 Raw 數據作為模型的首層。

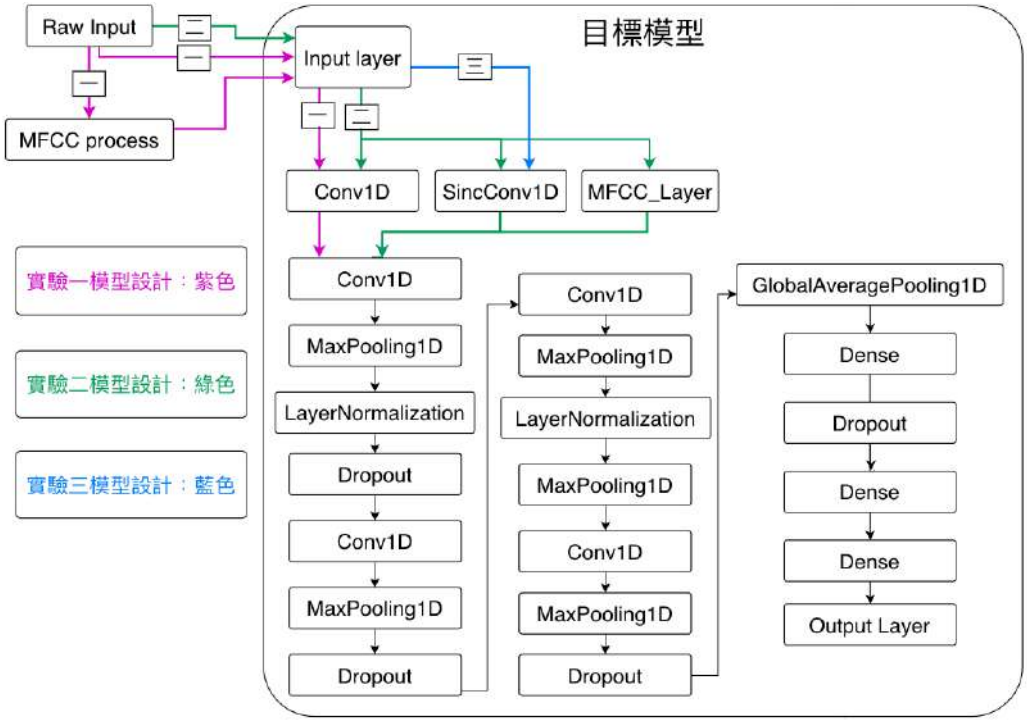

圖 3-3：實驗一、二和三語音模型訓練架構

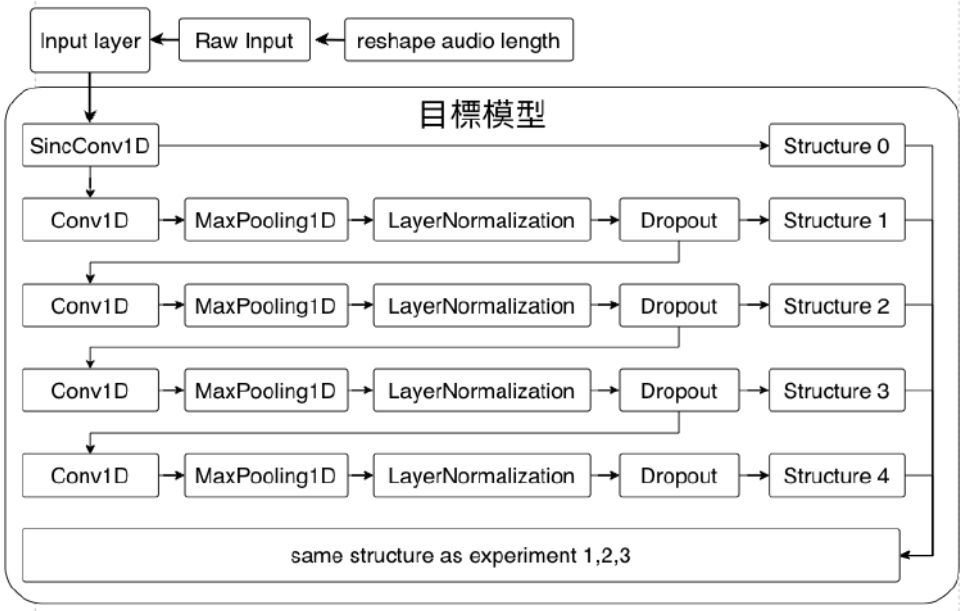

圖 3-4：實驗四、五語音模型訓練架構圖

圖 3-4 為實驗四、五語音模型訓練架構，針對實驗四，我們在 Reshape Audio Length 這一步設定不同採樣長度，作為模型的輸入；對於實驗五，我們則是選擇不同 Structure 進行訓練，討論攻擊效果。

(三) 設計實驗

1. 實驗一：討論模型以不同資料格式訓練對模型逆向攻擊的效果的影響

討論以原始音訊(Raw)和 MFCC 預處理資料兩種格式作為訓練資料集所訓練出的語音模型，使用 MIface 函數、傳統攻擊法與滑動攻擊法的攻擊效果。

本實驗採用的兩個資料集：URBANSOUND8K DATASET（簡稱為 urban）、Speaker Recognition Audio Dataset_50_speakers_audio_data（簡稱為 50_speakers）。本實驗中，我們總共訓練了 8 個模型：urban_raw、art_urban_raw、50_speakers_raw、art_50_speakers_raw、urban_mfcc、art_urban_mfcc、50_speakers _mfcc 與 art_50_speakers _mfcc

其中 Raw 訓練資料集格式為：採樣率 16000Hz，總時長 0.2s，共計 3200 個資料點；MFCC 訓練資料集格式為：將採樣率 16000Hz，總時長 0.2s，共計 3200 個資料點的資料經過 MFCC 擷取特徵後，將資料型態轉為 350 個資料點。

2. 實驗二：討論模型中首層不同特徵擷取方式對模型逆向攻擊效果的影響

討論模型首層採用 Raw 架構、MFCC 特徵擷取架構和 SincNet 特徵擷取架構所訓練出的語音模型，使用傳統攻擊法與滑動攻擊法的攻擊效果。

由於本實驗討論模型中首層不同的特徵擷取方式對模型逆向攻擊的影響，此實驗的模型輸入皆是採樣率 16000Hz，總時長 0.2s，共計 3200 個資料點的原始音訊(Raw)。我們共訓練了以下 6 個模型：urban_raw、50_speakers_raw、urban_sinc、50_speakers _sinc、urban_mfcc 與 50_speakers _mfcc

3. 實驗三：討論不同語音資料集對模型逆向攻擊效果的影響

討論 TIMIT 資料集與 50_speakers 資料集兩種人聲資料集受傳統模型逆向攻擊、滑動模型逆向攻擊的攻擊效果。本實驗沿用實驗二的 SincNet 特徵擷取架構，訓練語音模型，該架構與 SincNet 官網提供的 TIMIT 語音分類模型架構一致。

此實驗的模型輸入皆是採樣率 16000Hz，總時長 0.2s，共計 3200 個資料點的原始音訊(Raw)。我們共訓練了以下 2 個模型：TIMIT_sinc 和 50_speakers_sinc。

4. 實驗四：討論不同採樣長度對模型逆向攻擊效果的影響

討論 50_speakers 資料集以類似模型架構，不同採樣長度訓練目標模型，受傳統模型逆向攻擊、滑動模型逆向攻擊的攻擊效果。我們對資料集擷取不同長度的聲音資料，並將其用於訓練模型，我們總共訓練了輸入資料長度(採樣次數)為 3200、6400、12800、19200、25600、32000、38400、44800、51200 的語音分類模型，由於各資料長度差異較大，我們在盡量保持模型架構不變的前提對每個模型架構進行微幅調整，使模型達到相似性能。

此實驗中，模型輸入之音訊採樣率皆為 16000Hz，以 SincNet 架構進行訓練，攻擊時將攻擊初始化資料向量全部設為 0。

5. 實驗五：討論不同複雜度的模型架構對模型逆向攻擊效果的影響

討論 50_speakers 資料集，以相同採樣長度，不同模型架構訓練目標模型，受傳統模型逆向攻擊與滑動模型逆向攻擊的攻擊效果。我們對資料集擷取採樣率 16000Hz，長度為 64000 的音訊，以圖 3-4 中呈現的 Structure 0、1、2、3、4 五種架構進行訓練，分別代表由簡單到複雜的模型架構。

此實驗中，模型輸入之音訊採樣率皆為 16000Hz，以 SincNet 架構進行訓練，攻擊時將攻擊初始化資料向量全部設為 0。

6. 實驗六：討論不同音訊作為攻擊初始化樣本對模型逆向攻擊效果的影響

討論 urban 資料集與 50_speakers 資料集受傳統模型逆向攻擊、滑動模型逆向攻擊，分別以 urban、50_speakers、TIMIT 資料集內的一段音訊作為初始化樣本的攻擊效果。本實驗的目標模型輸入音訊皆是採樣率 16000Hz，總時長 0.2s，共計 3200 個資料點的原始音訊數據(Raw)，模型首層採用 SincNet 特徵擷取架構。我們總共訓練兩個模型：urban_sinc 和 50_speakers_sinc。

從 urban、50_speakers 和 TIMIT 三個資料集中分別挑選一筆音訊資料作為攻擊初始化樣本，並分別分析攻擊耗時、攻擊置信度分數、攻擊結果與資料集原始樣本的歐幾里得距離。

7. 實驗七：討論差分隱私防禦算法的防禦效果

討論 50_speakers 加入差分隱私算法後，受傳統模型逆向攻擊、滑動模型逆向攻擊的效果，本實驗沿用實驗二的三種架構的模型，於原模型基礎上加入差分隱私算法，以驗證該防禦算法的可行性。本實驗原先採用的防禦策略是基於 tensorflow-privacy 函式庫的差分隱私算法，這種算法主要是在訓練模型時在梯度加入噪聲機制，以此提高攻擊的難易度。然而經過我們的實驗發現，該函式庫所提供的基於梯度的差分隱私算法會大幅提升語音模型的訓練時長，且模型不易收斂。因此我們參考 One Parameter Defense - Defending against Data Inference Attacks via Differential Privacy[7]對圖像分類系統的差分隱私算法，進行簡化以降低計算時間，設計出針對語音模型的差分隱私算法，如圖 3-5 所示。該防禦算法無須重新訓練模型，僅需在原有的模型後加入一層差分隱私層便能達到防禦效果，大幅減少訓練防禦模型的訓練成本。

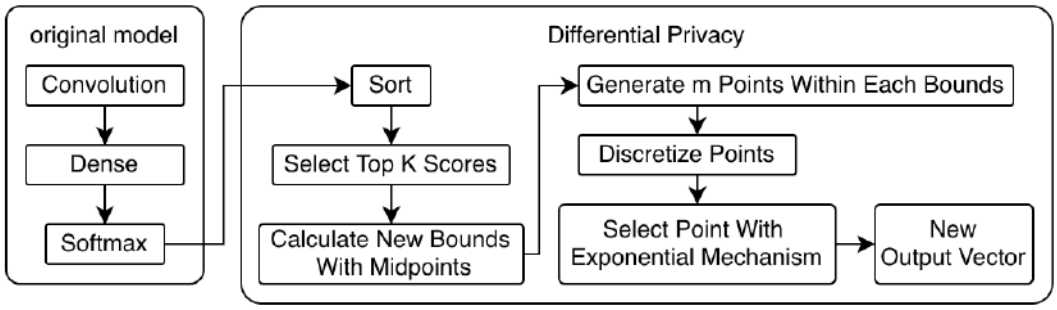

圖 3-5：差分隱私算法流程圖

8. 實驗八：討論時頻譜分類模型與單維度語音分類模型受模型逆向攻擊時的效果

討論 50_speakers 和 urban 兩個資料集，以轉換成時頻譜的資料格式訓練目標模型（模型架構如圖 3-6 所示），並分別進行 MIface 和傳統模型逆向攻擊。此實驗將攻擊所得出的時頻譜結果還原為單維度的音訊資料，並延續實驗二的評估方法，比較時頻譜分類模型攻擊結果與實驗二中的單維度語音分類模型攻擊結果，討論其差異。本實驗時頻譜輸入音訊皆是採樣率 16000Hz，總時長 2s，共計 32000 個資料點的原始音訊數據經過 stft 和 amplitude_to_db（n_fft=2048, hop_length=512）兩個步驟後轉換成(1025, 63)的時頻譜圖，其強度單位為 dB。單維度語音分類模型的各項參數則與實驗二相同。50_speakers 和 urban 資料集的音訊經過時頻譜轉換後，其最高能量約為 5dB 而最低能量約為-80dB，因此我們設計從+20 到-105dB 的初始化樣本進行攻擊實驗，以涵蓋原始資料的所有能量強度分佈。

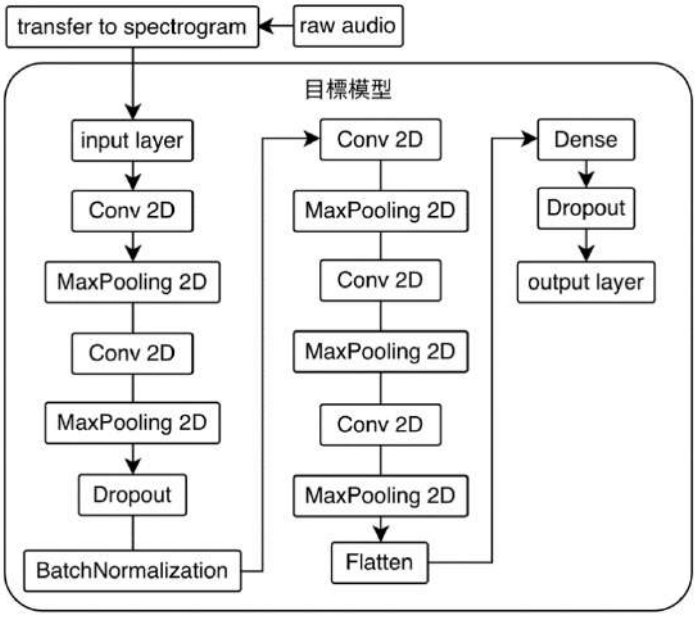

圖 3-6：時頻譜模型架構圖

(四) 模型逆向攻擊

1. 攻擊方式

我們使用的 MIface、傳統模型逆向攻擊、滑動模型逆向攻擊等三種算法皆是白盒模型攻擊。依據 Introducing Model Inversion Attacks on Automatic Speaker Recognition[1]論文當中的傳統與滑動兩種攻擊方式進行優化，並應用在我們設計的攻擊算法中。

2. 逆向攻擊演算法

如圖 3-6 所示，傳統攻擊法與滑動攻擊法的差別在於，傳統攻擊法每輪更新都會更新整段音訊樣本，而滑動攻擊則是每輪攻擊只會更新一小段音訊樣本，我們稱其為窗口，接著該窗口會向右移動一個步長的距離，如圖 3-7 所示，如此循環往復以獲得盡可能高的置信度分數，直到 loss 達到閾值即攻擊成功。

對每種攻擊方法，我們討論不同的初始化樣本，把初始化資料向量全部設為 1、0.5、0、-0.5、-1、random 與 mean 共七種參數，其中 mean 初始化樣本是指將資料集原始樣本平均的結果作為攻擊時的初始化資料，藉此討論不同初始化參數對模型逆向攻擊效果的影響。我們在經過大量測試後，發現滑動攻擊在窗口單位長度 430、步長 85 時攻擊結果置信度最高，窗口長度 410、步長 100 時攻擊耗時最短，並在每次實驗都討論此兩種滑動攻擊參數。

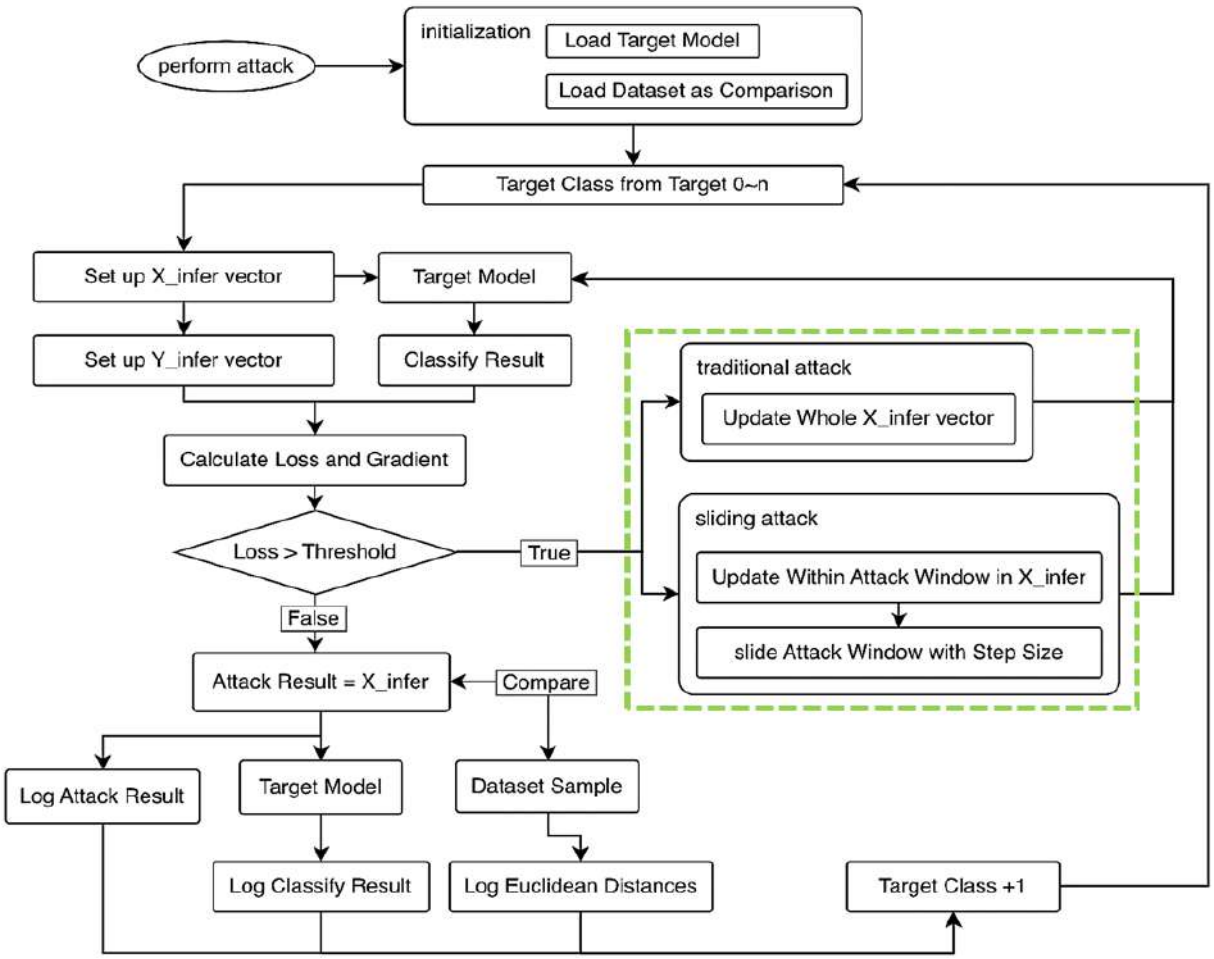

圖 3-6：逆向攻擊演算法攻擊流程圖

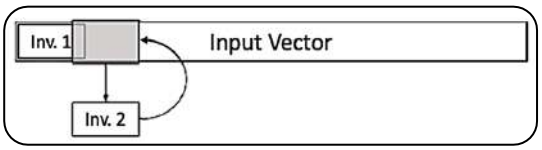

圖 3-7：滑動攻擊法攻擊方式

3. 攻擊成果評估

我們透過模型逆向攻擊算法為目標模型的每一個分類標籤(label)生成一組攻擊樣本，將此攻擊樣本作為目標模型的輸入，紀錄該分類標籤(label)的置信度分數，並紀錄攻擊該分類標籤的總耗時。同時，我們對將攻擊樣本與資料集原始樣本進行比對，計算兩者之間的歐幾里得距離，以描述其差異性。

肆、 研究結果

一、攻擊結果

(一) 討論模型以不同資料格式訓練對模型逆向攻擊的效果的影響

1. 整體攻擊效果

我們發現 ART_MIface 函數的攻擊結果最差，其餘攻擊結果的平均成功率皆有達到九成以上。表 4-1 中呈現了輸入不同資料格式之目標模型受模型逆向攻擊的攻擊成功率。

表 4-1：不同資料格式之目標模型受模型逆向攻擊的攻擊成功率

50_speakers				
	MIface	traditional	sliding_410_100	sliding_430_85
Raw	0.02	0.98	0.98	0.97
MFCC	0.03	1.00	0.99	1.00
urban				
	MIface	traditional	sliding_410_100	sliding_430_85
Raw	0.13	1.00	0.99	0.99
MFCC	0.13	1.00	1.00	1.00

2. 輸入資料格式為 Raw 之模型的攻擊分析

圖 4-1 與圖 4-2 為 50_speakers 資料集和 urban 資料集以輸入資料格式為 Raw 之模型的攻擊結果與資料集原始樣本的歐幾里得距離。發現以 mean 和 random 初始化樣本的攻擊結果與資料集原始樣本的距離較小，其中又以 mean 初始化樣本的攻擊結果最接近原始樣本。

圖 4-3 為分析各初始化樣本和資料集原始樣本之間的歐幾里得距離，以 mean 為初始化樣本和資料集原始樣本最為接近；另外，觀察 random 初始化樣本與 mean 初始化樣本，這兩種初始化方式都具有波動的特性，我們認為對於語音模型的逆向攻擊，提供具波動特性的初始化參數可以使攻擊出的結果接近資料集原始樣本，同時保有較好的攻擊效率。

觀察攻擊結果與資料集原始樣本的相似度，發現 50_speakers 資料集的攻擊結果與原始樣本的相似度較 urban 低。推測是由於資料集本身的每個分類間的差異程度所致。圖 4-4 呈現了兩種資料集內各個類別之間的差異，以 urban 資料集中各類樣本之間差異程度較大，而 50_speakers 較小。我們認為，若各個類別之間差異大，攻擊過程中每一類別的特徵差異會明顯反映在梯度上，因此攻擊結果較接近資料集原始樣本。

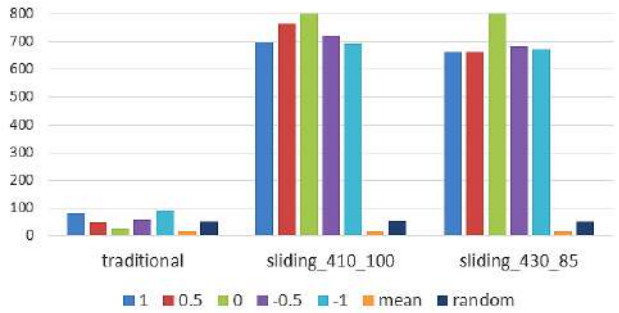

圖 4-1：50 speakers 資料集 Raw 輸入攻擊結果與原始樣本的距離

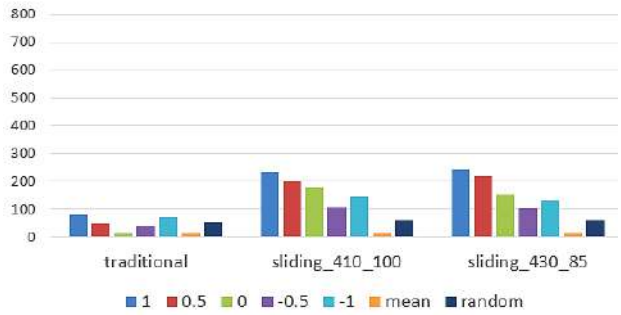

圖 4-2：urban 資料集 Raw 輸入攻擊結果與原始樣本的距離

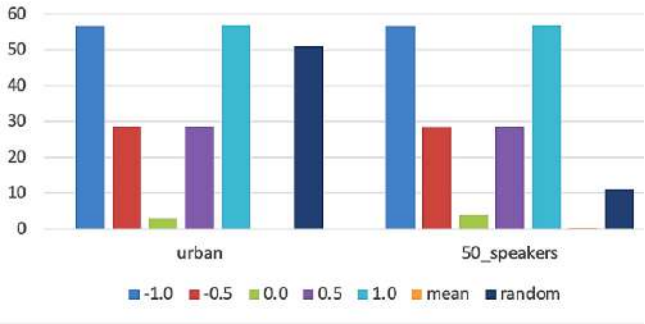

圖 4-3：資料集與初始化樣本的距離

圖 4-4：資料集中不同類別樣本的差異

3. 輸入資料格式為 MFCC 之模型的攻擊分析

當模型輸入資料為 MFCC，以不同初始化樣本進行攻擊時，我們發現除了 mean 方式之初始化樣本的攻擊結果與資料集原始樣本的距離較小，其餘初始化樣本的攻擊結果與原始樣本的距離不但偏高且數值幾乎一致，見圖 4-5 與圖 4-6。

圖 4-7 為 MFCC 處理後的資料集原始樣本依照數值所呈現的折線圖，分析此圖，我們發現存在著一種 MFCC 提取出的特徵位於最開始的極端值。圖 4-8 為以 mean 初始化樣本得出的攻擊結果樣本，相較於圖 4-9，以 0 初始化得出的攻擊結果樣本，mean 初始化樣本提供了相似的極端值，其攻擊結果也比較接近資料集原始樣本，而其他初始化方式由於不會提供此極端值特徵，因此攻擊結果都與資料集原始樣本的距離較大。

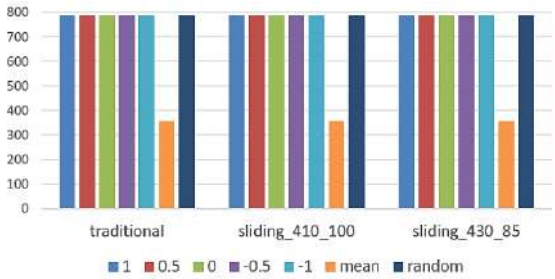

圖 4-5：50 speakers 資料集 MFCC 輸入攻擊結果與原始樣本的距離

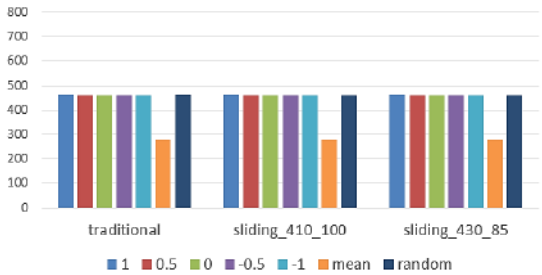

圖 4-6：urban 資料集 MFCC 輸入攻擊結果與原始樣本的距離

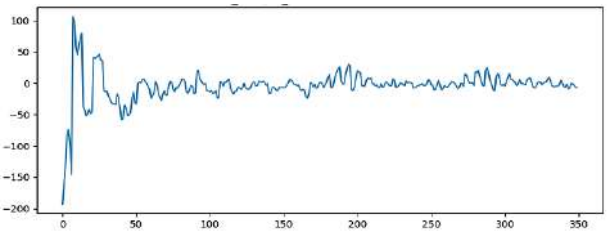

圖 4-7：MFCC 處理後的原始樣本

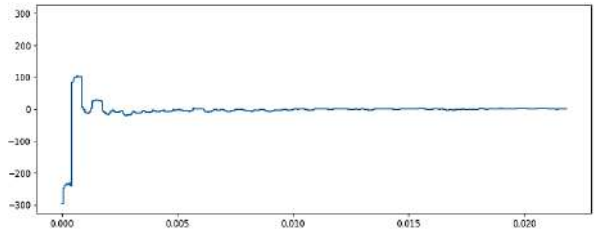

圖 4-8： mean 初始化樣本的攻擊結果

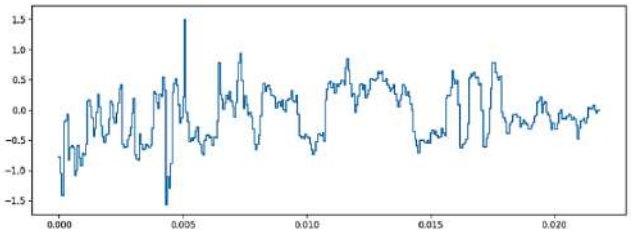

圖 4-9：0 初始化樣本的攻擊結果

(二) 討論模型中首層不同特徵擷取方式對模型逆向攻擊效果的影響

1. 模型首層採用不同架構之攻擊耗時比較

攻擊 Raw 架構、MFCC 及 SincNet 架構目標模型，如圖 4-10 所示，MFCC 架構的目標模型攻擊耗時最短，SincNet 架構次之，Raw 架構最長。觀察目標模型內各層的計算結果，Raw 架構首層擷取出長度 3072，共 80 維度的特徵資料；MFCC 架構首層擷取出長度 350 的 1 維資料，而 SincNet 架構首層擷取出的特徵，輸出形狀為長度 200 共 80 維度的資料。Raw 架構提取出的特徵量約是 SincNet 架構的 15 倍，MFCC 架構的 700 倍。三種架構的目標模型以相同初始化樣本進行攻擊，其攻擊耗時呈現 Raw 架構> SincNet 架構>MFCC 架構。模型首層擷取出的特徵越多，後續各層所獲得與輸入音訊相關的資料越多，輸入音訊的微小差異對模型分類結果的影響越明顯，攻擊該模型的難度越大，攻擊耗時越長，此現象在 50_speakers 資料集，滑動攻擊的攻擊結果尤為明顯，見圖 4-11。

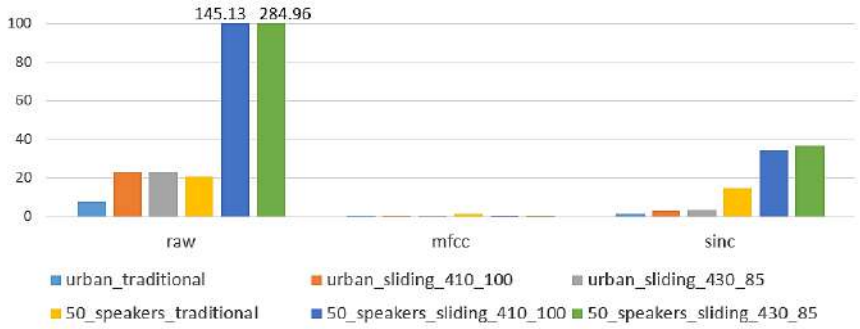

圖 4-10：不同特徵擷取架構之目標模型的攻擊耗時

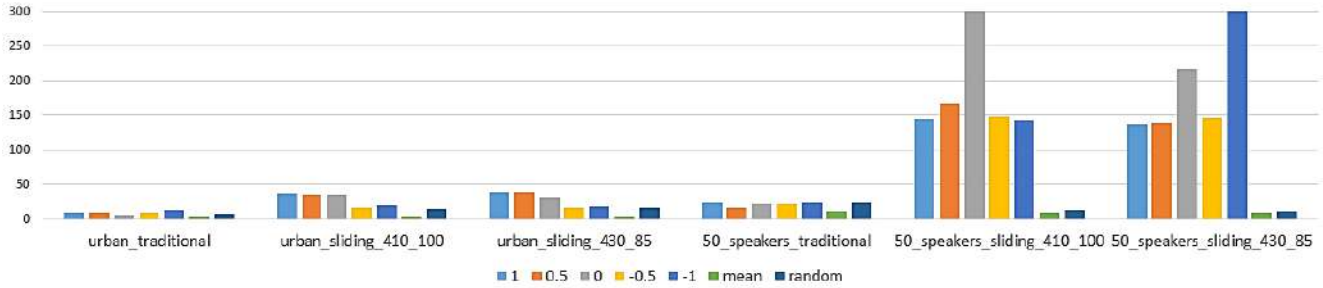

圖 4-11：Raw 特徵擷取架構之目標模型的攻擊耗時

2. 首層不同架構的目標模型，受逆向攻擊的結果與資料集原始樣本的相似度比較

我們觀察 Raw 架構、SincNet 架構和 MFCC 架構模型的攻擊結果與資料集原始樣本的歐幾里得距離，後二者攻擊結果與資料集原始樣本的距離和各初始化樣本與資料集原始樣本的距離相似，0 初始化樣本的距離最短，1 和-1 個初始化樣本的距離最大，見圖 4-12-1、4-12-2 與 4-12-3。而 Raw 架構的攻擊結果與資料集原始樣本的距離則和攻擊耗時趨勢較相關。

我們認為 SincNet 和 MFCC 架構都經過特殊算法擷取輸入音訊的特徵，目標模型學習到各種聲音的特徵，攻擊結果與攻擊初始化樣本的差異不大，攻擊過程只需要擬合出各個分類的某些明顯特徵便可攻擊成功，因此攻擊結果與原始資料的差異較大，且受各初始化樣本與資料集原始樣本的距離影響。但是 Raw 架構的目標模型僅採用卷積提取特徵，對於輸入資料的細節判定也更嚴格，需要擬合所有細部特徵，因此，初始化樣本的選擇對攻擊效果的影響更明顯。我們認為，不同初始化樣本的攻擊效果與資料集特性密切相關。以 50_speakers 資料集為例，0 初始化樣本所需攻擊耗時較長，而 urban 資料集則是以 1 初始化樣本的攻擊耗時最長。對 Raw 架構的攻擊中，攻擊結果與資料集原始樣本的距離和圖 4-11 的攻擊耗時具正相關，初始化樣本對攻擊效果的影響在 Raw 架構下尤為明顯。

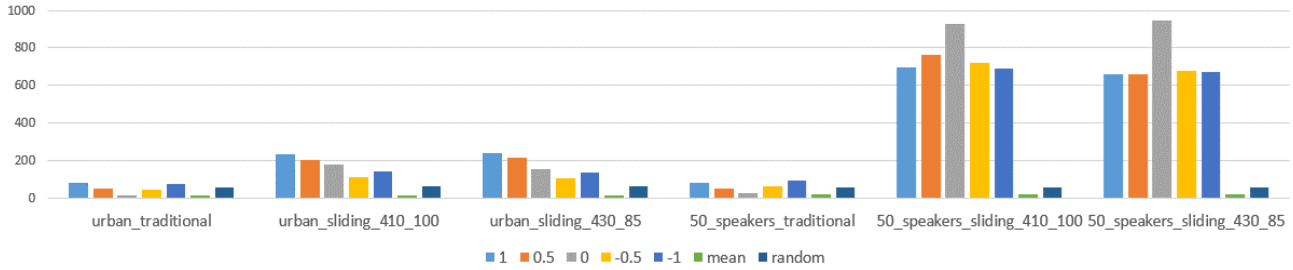

圖 4-12-1：Raw 特徵擷取架構之目標模型攻擊結果與原始樣本的距離

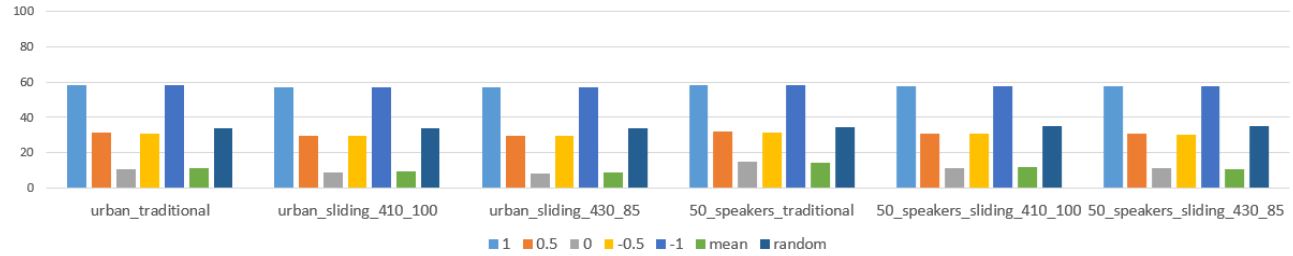

圖 4-12-2：MFCC 特徵擷取架構之目標模型攻擊結果與原始樣本的距離

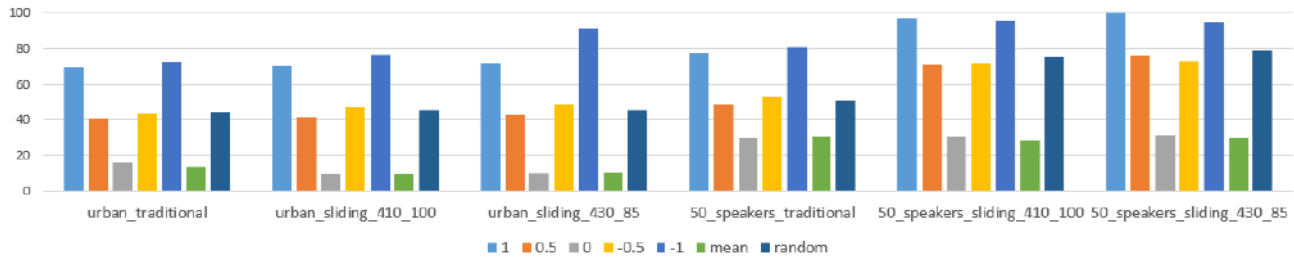

圖 4-12-3：SincNet 特徵擷取架構之目標模型攻擊結果與原始樣本的距離

(三) 討論不同語音資料集對模型逆向攻擊效果的影響

TIMIT 和 50_speakers 同樣為人聲資料集，分析兩種資料集目標模型的攻擊結果，發現 TIMIT 的攻擊結果相較於 50_speakers，其攻擊置信度稍高，與資料集原始樣本的歐幾里得距離稍短，如圖 4-13-1 和圖 4-13-2 所示。

分析 TIMIT 與 50_speakers 兩資料集的樣本差異，由圖 4-14 所示，50_speakers 資料集中各個類別之間相較於 TIMIT 差異程度較小。此現象與實驗一的結果相似，由於各個類別之間差異大，攻擊過程中每個類別的特徵差異明顯，反映在梯度上，導致 TIMIT 攻擊結果的置信度較高，且攻擊結果與資料集原始樣本的距離也較接近。

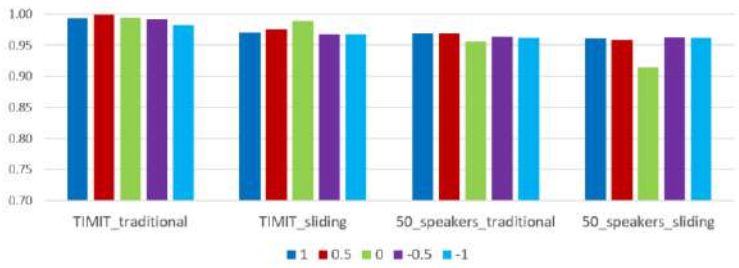

圖 4-13-1：不同人聲資料集攻擊置信度分數

圖 4-14：資料集中不同類別樣本的差異

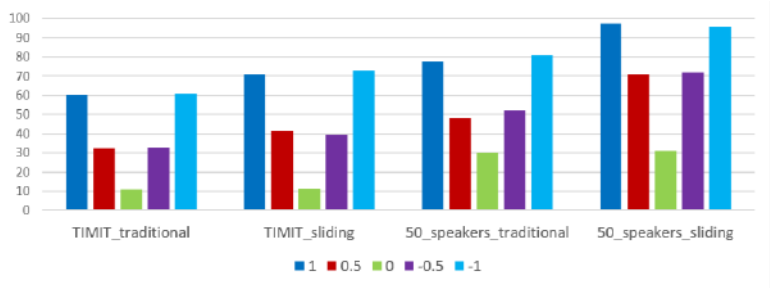

圖 4-13-2：不同人聲資料集攻擊結果與原始樣本的距離

(四) 討論不同採樣長度對模型逆向攻擊效果的影響

在圖 4-15-1 與圖 4-15-2 中，觀察對不同音訊長度模型的攻擊效果，我們發現攻擊耗時和攻擊結果與資料集原始樣本的歐幾里德距離具正相關。該現象於 urban 資料集尤為明顯，其攻擊耗時在音訊採樣長度為 25600 到 32000 間出現峰值。

這一發現與預期中「音訊長度越長，攻擊耗時越長」的假設不同，同時，我們也觀察到，攻擊耗時的變化量與歐幾里得距離的變化量不具有正比關係，而是趨勢相似。攻擊結果與資料集原始樣本間距離的峰值主要受資料集特性影響，如 50_speakers 資料集的距離和耗時峰值出現在採樣長度 12800 到 25600 之間。兩個資料集的攻擊耗時於 25600 到 32000 這兩個採樣長度上較高，這可能是語音資料的通性，在採樣 2.6 到 3.0 秒時，攻擊成本較高。

另外，我們也發現滑動攻擊法的攻擊耗時隨音訊長度增加變化不大。由於滑動攻擊在攻擊過程中每次更新的範圍由窗口長度決定，因此攻擊過程的計算量不隨音訊採樣長度變化。同時在圖 4-15-3，我們觀察到傳統攻擊在音訊長度為 3200 到 12800 攻擊耗時較長，攻擊置信度分數較低，該攻擊方法在還原短音訊時較滑動攻擊效果差。我們認為，還原短音訊時，輸入的微小變化對模型輸出結果影響較大，滑動攻擊針對小範圍的攻擊手法具優勢。但當攻擊目標長度增加時，每段資料的小幅差異對攻擊結果的影響就不明顯，傳統攻擊這種單位時間

多次更新的策略反而更具優勢。

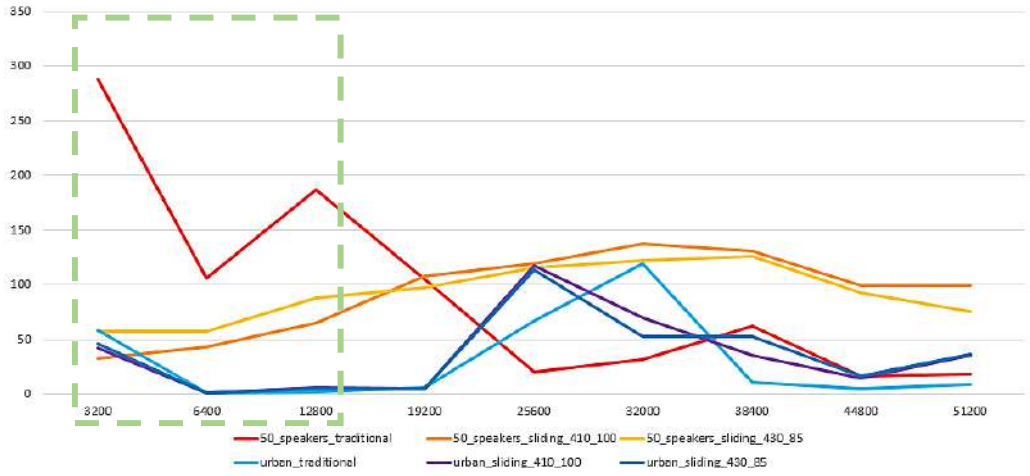

圖 4-15-1：不同採樣長度的模型攻擊耗時

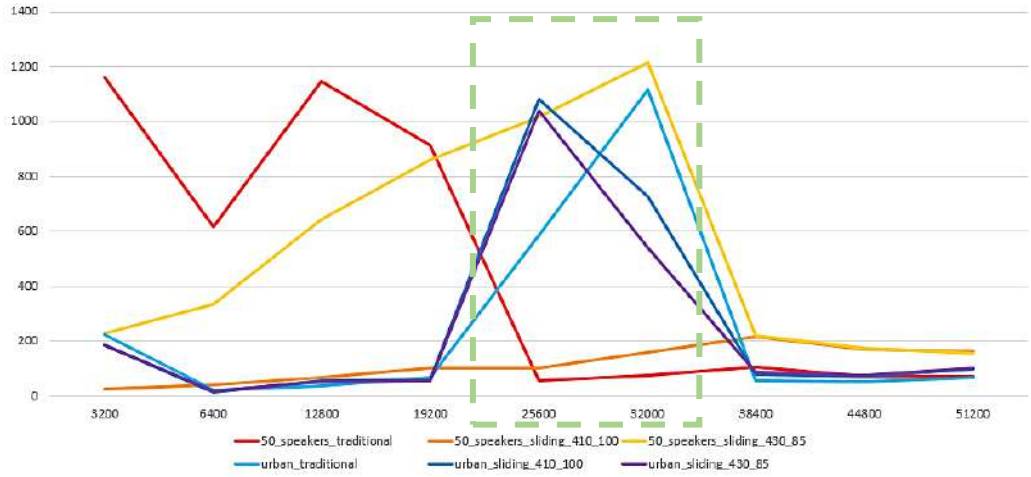

圖 4-15-2：不同採樣長度的模型攻擊結果與原始樣本的距離

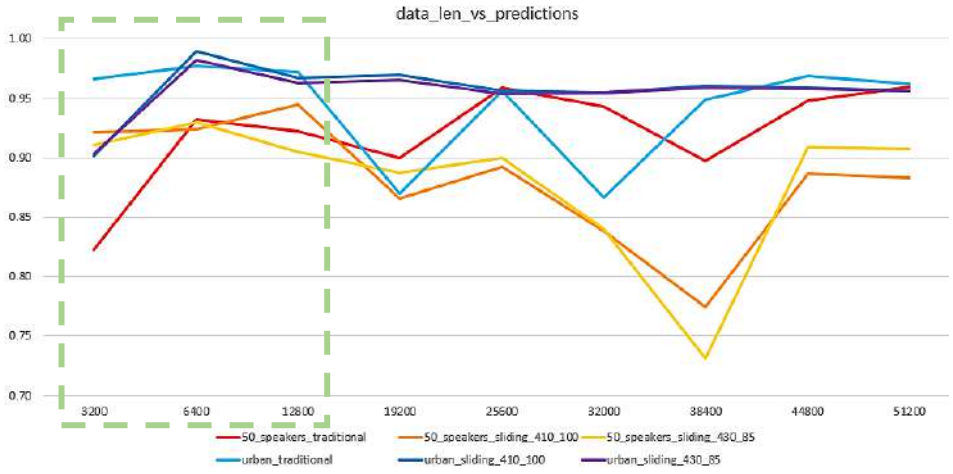

圖 4-15-3：不同採樣長度的模型攻擊置信度分數

不同採樣長度的目標模型，攻擊結果與資料集原始樣本的歐幾里德距離相似時，傳統攻擊下音訊採樣長度越長，攻擊耗時越短，而滑動攻擊則沒有這樣的趨勢。距離計算時，若兩

段音訊單位長度的距離一致，音訊長度越長，整段音訊的距離應呈現線性成長。同理，當音訊的距離相同時，音訊長度越長，單位音訊之間的距離就會變短。由於單位音訊之間的距離縮短，傳統攻擊便可以在更少的回合數完成擬合，但是滑動攻擊下，攻擊目標長度增加時，單位音訊每次更新的間隔也會增加，攻擊耗時不會隨攻擊目標採樣長度增加有太大的區別。

(五) 討論不同複雜度的模型架構對模型逆向攻擊效果的影響

分析 50_speakers 資料集以不同複雜度的架構訓練目標模型，受模型逆向攻擊的結果，如圖 4-16-1、圖 4-16-2 與圖 4-16-3，攻擊耗時以及攻擊結果與資料集原始樣本的距離隨著模型複雜度增加而上升，而置信度則隨之下降。攻擊成功率也隨模型複雜度上升而下降。

面對越複雜的模型，攻擊過程中的梯度變化較不明顯，可能出現梯度屏蔽的問題。由於攻擊過程仰賴梯度信息更新攻擊樣本，當梯度變化不明顯時，攻擊就需要更長的時間達到擬合。與此同時，模型中的參數變多也會導致每輪攻擊的計算量增加，這也是攻擊耗時增加的原因。比對傳統和滑動兩種攻擊方式，由於傳統攻擊的特性，每輪攻擊都會更新攻擊樣本的所有採樣點，單位採樣點更新頻率遠高於滑動攻擊，在梯度變化不明顯的情況下，傳統攻擊能對單位樣本點更新更多次，攻擊效果較佳。

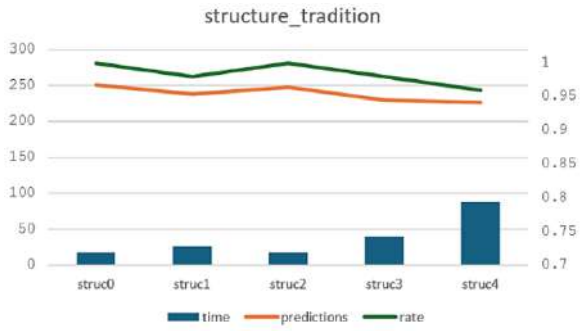

圖 4-16-1：不同模型複雜度的傳統攻擊耗時

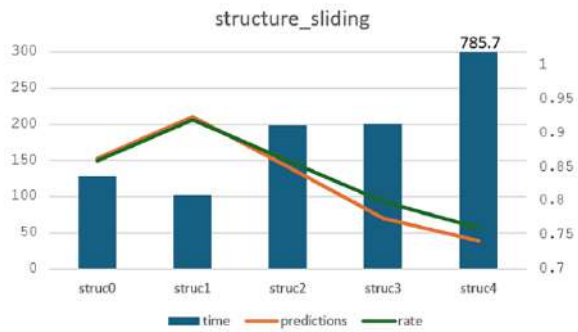

圖 4-16-2：不同模型複雜度的滑動攻擊耗時

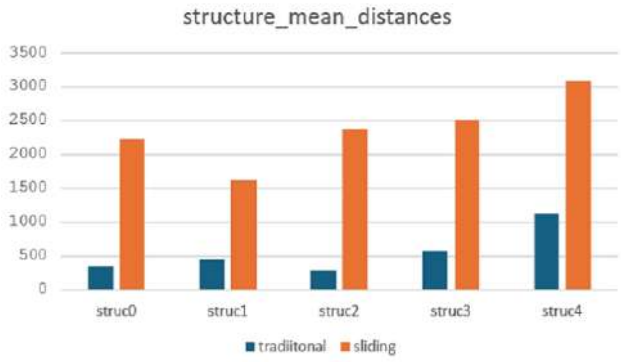

圖 4-16-3：不同模型複雜度的攻擊結果與資料集原始樣本的距離

滑動攻擊在有限回合數的攻擊中，單位採樣點的更新次數較少，因此攻擊結果與資料集原始樣本的距離大於傳統攻擊。另外，我們也發現模型複雜度、攻擊耗時與置信度並非完全正相關，如滑動攻擊的架構 2 和 3，雖然置信度與成功率下降，耗時卻變化不大。滑動攻擊下的架構 0 相較架構 1 表現差，不符合架構 1~4 的趨勢。這與訓練目標模型時模型收斂的性能有關，其整體趨勢仍遵守模型攻擊耗時與置信度隨模型複雜度增加的趨勢。

(六) 討論不同音訊作為攻擊初始化樣本對模型逆向攻擊效果的影響

圖 4-17-1、圖 4-17-2、圖 4-17-3 和圖 4-17-4 為針對 50_speakers 與 urban 兩資料集，以不同音訊資料作為初始化樣本和 mean、random 初始化樣本，共五種初始化方式的攻擊結果比較。50_speakers 目標模型的攻擊結果中， urban 音訊初始化樣本的耗時最長、置信度最低，推測是由於 50_speakers 和 TIMIT 同為人聲資料集，資料集原始樣本與 urban 資料集差異較大；同理，由於 50_speakers 資料集屬於人聲資料，故以 TIMIT 音訊初始化樣本在 50_speakers 目標模型當中效果較佳。同時，在 urban 目標模型的攻擊結果中則是以 TIMIT 音訊初始化樣本耗時最長、50_speakers 音訊初始化樣本置信度最低，此亦符合以上歸納。然而，對比 mean 和 random 兩種初始化樣本的攻擊結果，多數情況下，採用一段聲音作為初始化樣本的攻擊耗時較採用 mean 和 random 初始化樣本的攻擊耗時長，攻擊置信度分數也較差。

觀察圖 4-17-5 中，攻擊結果與資料集原始樣本的距離可發現，50_speakers 目標模型中，以音訊作為攻擊初始化樣本的攻擊結果與資料集原始樣本的距離高於 mean 和 random 兩種初始化方式，尤其是 urban 音訊作為初始化樣本時，距離最大。這顯示音訊資料較不適合用於攻擊初始化，由於音訊資料樣本自身便具有個別特徵，其在該音訊特徵上在進行增減以擬合出目標模型的各個類別，相互影響下，使攻擊結果與資料集原始資料距離越來越遠。

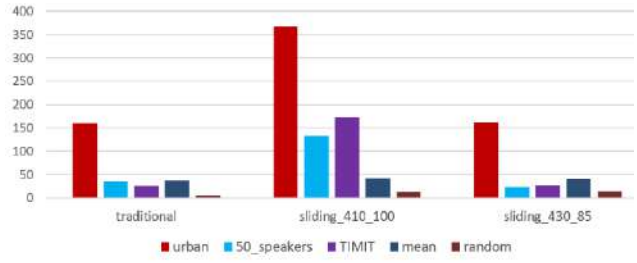

圖 4-17-1

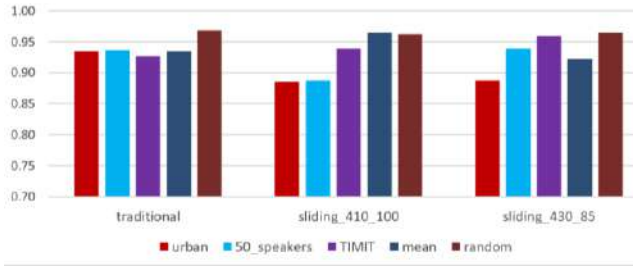

圖 4-17-2

50 speakers 目標模型以不同音訊資料作為初始化樣本的攻擊耗時(左)與置信度(右)

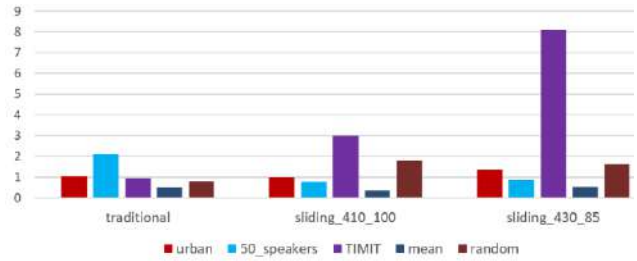

圖 4-17-3

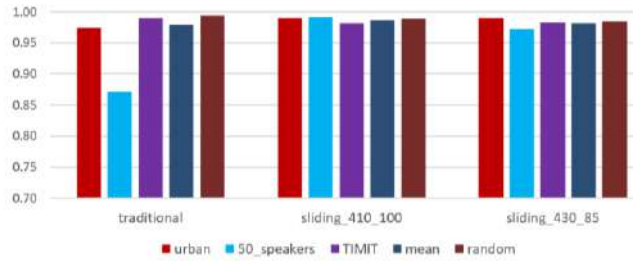

圖 4-17-4

urban 目標模型以不同音訊資料作為初始化樣本的攻擊耗時(左)與置信度(右)

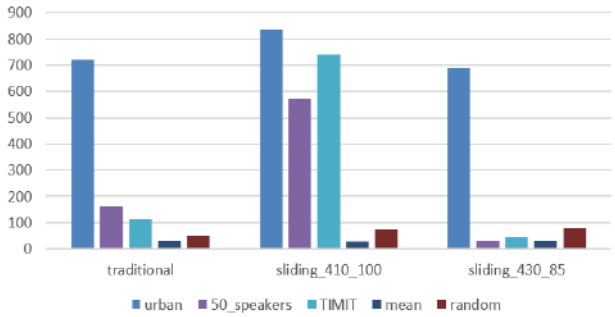

圖 4-17-5：50 speakers 目標模型以不同音訊作為初始化樣本的攻擊結果與原始樣本的距離

二、語音模型模型設計與防禦策略

綜合實驗一到六的各項發現，在設計模型時，可以考慮以下幾點，提升語音模型面對模型逆向攻擊時的防禦效果：

(一) 特徵提取架構：模型採用 Raw 架構的特徵擷取方式可以最大程度增加攻擊成本（攻擊耗時），且攻擊結果與資料集原始樣本差距較大，然而該架構的模型訓練耗時較長，若考慮模型訓練耗時與防禦性能，可以選擇 SincNet 架構，該架構訓練耗時明顯較 Raw 架構低，攻擊結果與原始樣本的差距僅次於 Raw 架構。

(二) 資料性質：採用類別相似度較高的資料集，可以增加攻擊結果與原始樣本的差異性，同時增加攻擊耗時，降低置信度分數。

(三) 音訊採樣長度：採用採樣長度 25600 或 32000 的音訊作為模型輸入可以使攻擊結果與原始樣本的差距較大，攻擊成本增加（攻擊耗時）。

(四) 模型複雜度：在不過擬合的情況下，可以多新增幾層卷積層，有效利用複雜模型對模型逆向攻擊本身具備的防禦效果，或者採 Tanh 或 Sigmoid 作為激勵函數，以造成梯度屏蔽或梯度消失的現象，提升模型防禦效果。

由於第二點提到的資料集性質防禦方法在應用層面較難實踐，可能仰賴其他資料預處理方式，但這些操作容易造成模型不收斂並降低模型性能，因此我們稍微簡化實踐方向，改為降低模型中各類別對輸出向量的影響程度。我們基於單一參數差分隱私算法，經過調適，簡化了 delta 參數和效用函數的計算方法，在保證語音模型分類性能的前提下，顯著提高了該模型的防禦性能，並對該算法的防禦性能進行了驗證。

三、差分隱私算法驗證

如圖 4-18-1 所示，加入差分隱私算法的目標模型，其 accuracy 相比原模型有些微下降，然其下降幅度接不超過原模型的 5%，顯示我們採用的差分隱私算法並不會大幅降低原本模型的分類性能。另外，如圖 4-18-2，我們也發現，加入差分隱私算法會使模型的計算耗時增加，增加比率隨模型架構有所不同。

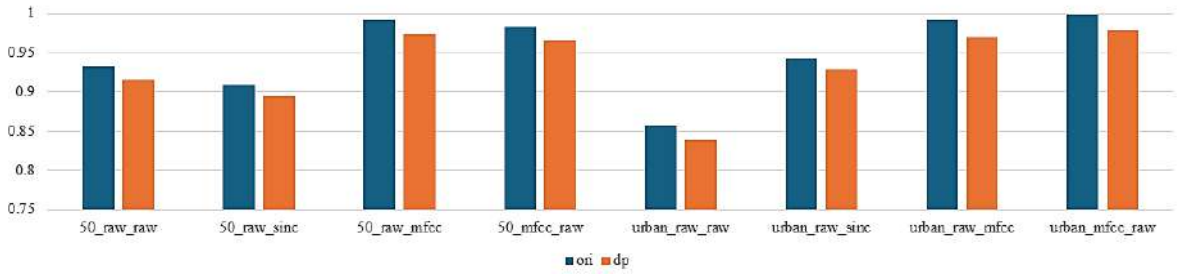

圖 4-18-1：差分隱私模型與原模型的分類性能(accuracy)比較

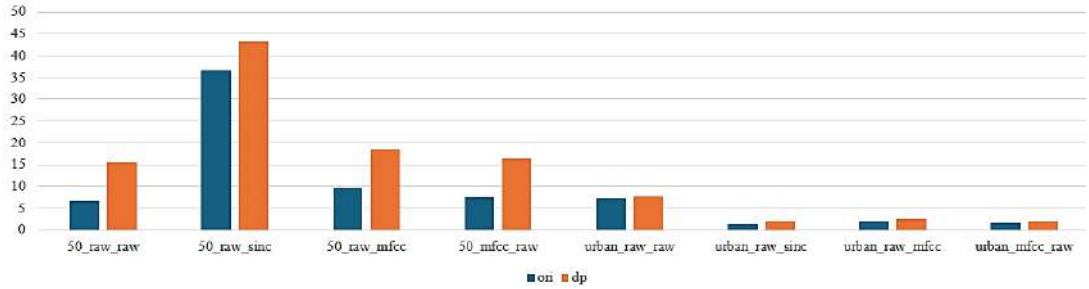

圖 4-18-2：差分隱私模型與原模型預測 12500 筆樣本的計算耗時

如圖 4-18-3 與 4-18-4，無論採用哪種攻擊方法，對模型的攻擊的耗時皆明顯上升，置信度下降。圖 4-18-5 中 Raw 和 SincNet 架構的差分隱私模型攻擊成功率下降至 50%左右，顯示該算法可以成功抵禦近半數的類別遭到攻擊，而 MFCC 架構的差分隱私模型防禦效果較差。

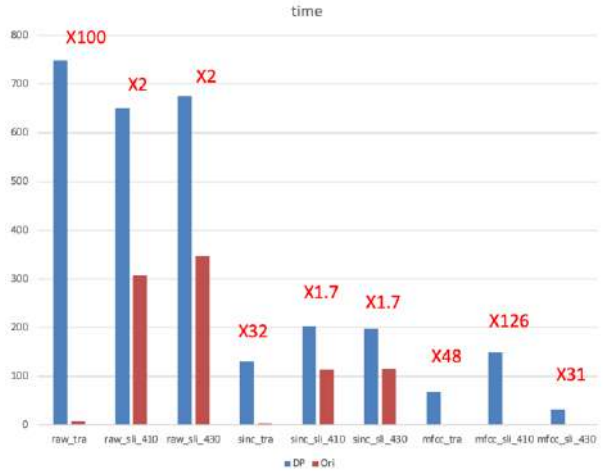

圖 4-18-3

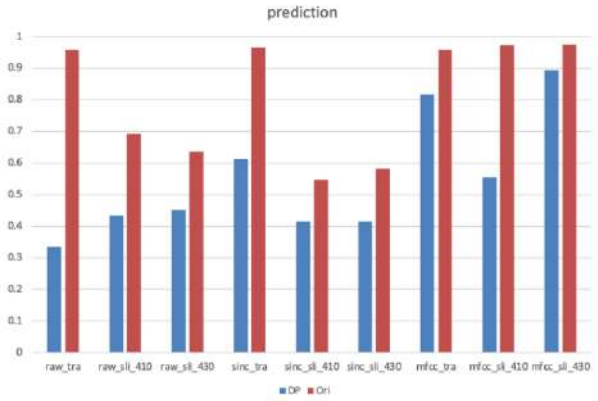

圖 4-18-4

差分隱私模型與原模型的攻擊耗時（左）攻擊置信度分數（右）比較

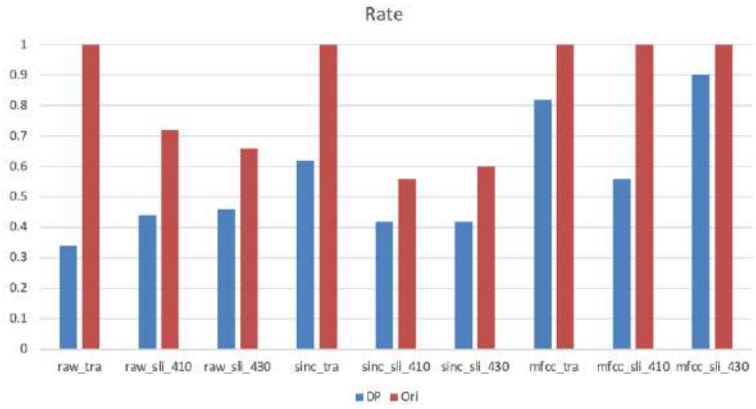

圖 4-18-5：差分隱私模型與原模型的攻擊成功率

由於傳統攻擊法攻擊原模型的攻擊耗時較低，加入差分隱私算法後，其攻擊耗時增加比例最大，顯著增加攻擊成本，同時，置信度與成功率下降也最多。而滑動攻擊法攻擊耗時也有明顯增加，該算法對於兩種攻擊手段都有不錯的防禦性能。另外，不同的特徵擷取方式間，則以 Raw 架構的耗時增加及置信度下降最為明顯。對比實驗二的發現，模型首層提取的特徵量確實對模型的攻擊難易度有影響，加入差分隱私算法後，同樣出現 Raw 架構攻擊置信度最低，MFCC 架構攻擊置信度最高的情況。

四、時頻譜分類模型

從表 4-2 可以發現，即便時頻譜分類模型屬於一種二維影像分類模型，Miface 函數的攻擊耗時仍比傳統攻擊更高，而攻擊成功率最高僅 6%，這與實驗一中的假設相背，顯示 Miface 函數的攻擊效果與模型輸入資料是否為圖像格式並無關聯。

表 4-2 時頻譜分類模型的攻擊效果

50_speakers				
	time	prediction	rate	d-vector_distance
traditional	13.48	0.96	1.00	0.97
MIface	375.57	0.03	0.03	51.25
urban				
	time	prediction	rate	d-vector_distance
traditional	35.15	0.95	1.00	1.00
MIface	362.19	0.07	0.06	48.52

我們檢視攻擊過程中的各項指標變化，發現攻擊過程中，傳統攻擊算法與 MIface 算法的梯度變化明顯不同，如圖 4-19 所示，無論任何攻擊初始化方式，傳統攻擊法的在第一輪攻擊時，計算出的梯度明顯高於 MIface 函數計算出的梯度。這也導致 MIface 函數攻擊過程中對攻擊樣本的更新幅度極小，難以攻擊出模型中各個分類的時頻譜資料特性，導致其攻擊耗時長而攻擊成功率低。

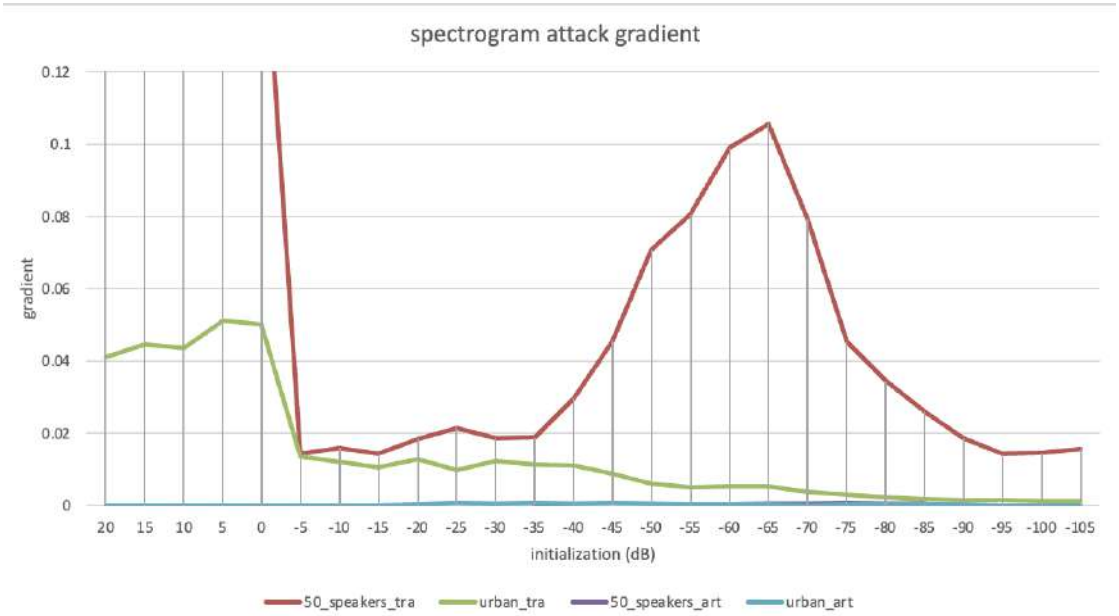

圖 4-19：不同初始化樣本第一輪攻擊的梯度資訊

進一步比較 MIface 攻擊和傳統攻擊的梯度計算函數，MIface 梯度計算前對輸入的攻擊樣本先進行了一次自定義預處理步驟，才進行反向傳播計算梯度，最後進行梯度剪裁；而傳統攻擊在梯度計算時直接進行反向傳播，我們認為 MIface 函數的自定義預處理和梯度剪裁是導致攻擊過程中梯度偏低的重要關鍵。

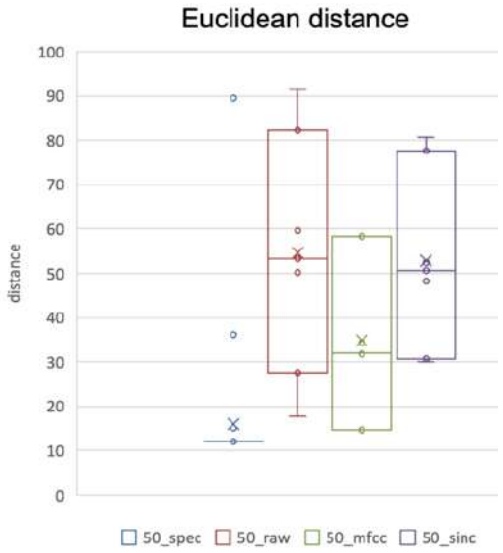

圖 4-20-1：傳統攻擊下時頻譜模型與單維語音模型攻擊結果與原始樣本的差異性

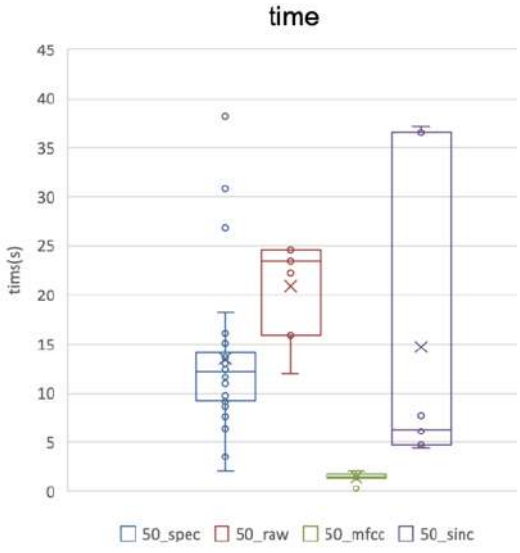

圖 4-20-2：傳統攻擊下時頻譜模型與單維語音模型攻擊耗時

同時，我們也加入實驗二的模型攻擊結果進行對比，如圖 4-20-1 和 4-20-2 所示，時頻譜模型的攻擊結果，明顯較實驗二中三種單維度語音分類模型的攻擊結果，更接近資料集原始樣本，我們認為是因為時頻譜相較於單維度語音多出了能量資訊和頻率資訊的維度。單維度

語音模型攻擊過程中提供的特徵僅從語音資料的波形與時間關係衍生而來，而時頻譜模型則進一步提供了語音的能量、時間與頻率資訊的衍生特徵，透過這些更詳細的特徵，攻擊演算法便可以還原出更接近資料集原始樣本的攻擊結果。攻擊耗時方面，時頻譜的攻擊耗時僅高於 MFCC 架構的模型。

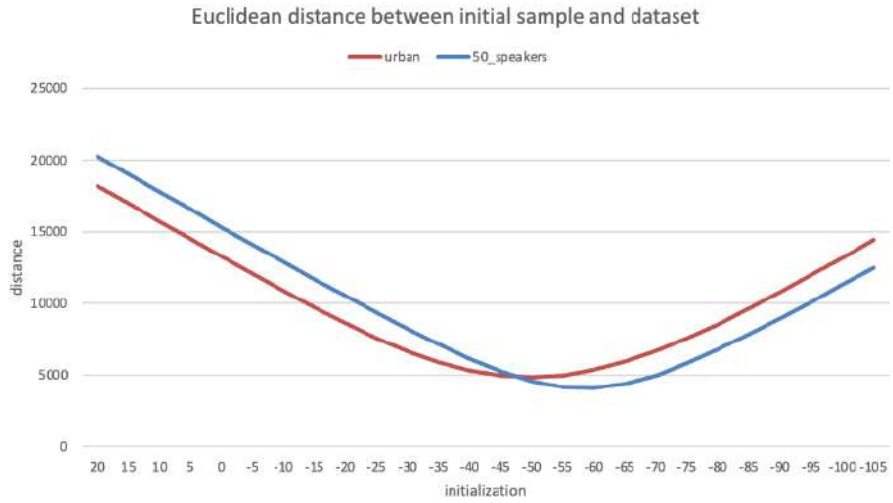

圖 4-21-1：時頻譜攻擊初始化樣本與資料集的歐幾里得距離

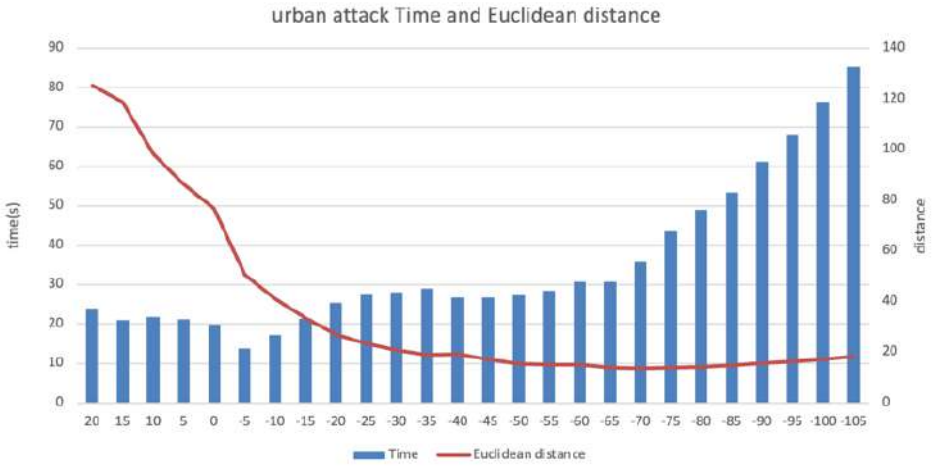

圖 4-21-2：urban 資料集時頻譜攻擊結果與原始樣本的
歐幾里得距離和攻擊耗時於不同初始化樣本的變化

觀察圖 4-21-1 和圖 4-21-2 則可以發現，攻擊初始化樣本與資料集的歐幾里得距離愈小時，攻擊效果愈佳。在 -40dB 到 -65dB 的初始化樣本之間出現初始化樣本與資料集樣本距離的最小值，反應在攻擊結果上，該區間的攻擊耗時最短，攻擊結果與原始樣本的歐幾里得距離最小。該現象亦出現在實驗二中 SincNet 架構與 MFCC 架構以 mean 和 random 初始化的攻擊結果，顯示無論是單維度語音分類模型或時頻譜分類模型皆具有該特性。

伍、 討論

一、具有波動特性的初始化樣本可使攻擊結果更接近資料集原始樣本。

對於輸入資料格式為 Raw 的目標模型而言，採用 mean 和 random 兩種初始化樣本可以使攻擊結果更接近資料集原始樣本，這是因為其波動的特性。

二、輸入資料格式為 MFCC 之模型，採用 mean 初始化樣本的攻擊結果與資料集原始樣本距離最近。

mean 初始化樣本為攻擊提供了 MFCC 的極端值特徵，所以其攻擊結果比較接近資料集原始樣本，然而其他的初始化不會提供此極端值特徵，因此攻擊結果與資料集原始樣本的距離較大。

三、首層特徵提取架構會影響模型對輸入音訊的敏感程度，進而影響攻擊的難易程度和攻擊樣本與資料集原始樣本之間的差異。

Raw 架構保留了最多特徵資訊，模型對輸入變化較為敏感，因此攻擊難度較高且攻擊耗時也最長。MFCC 和 SincNet 架構通過特徵提取算法，模型只需學習顯著特徵，攻擊結果只需擬合部分特徵就能攻擊成功，故攻擊結果與原始數據的差異較大，但攻擊耗時較短。

四、資料集中各類別樣本之間的差異性會影響模型逆向攻擊的效果。

類別差異較大的資料集(如 TIMIT 和 urban)會使攻擊生成的樣本置信度更高、與資料集原始樣本的歐幾里得距離更短。是因為當類別差異較大時，特徵差異在會更明顯的反映在梯度上，有利於生成更逼真的攻擊樣本。

五、不同的音訊採樣長度對模型逆向攻擊的耗時和攻擊結果與資料集原始樣本的歐幾里得距離正相關。

攻擊耗時的變化量與歐幾里得距離的變化量不具有正比關係，而是趨勢相似，攻擊結果與資料集原始樣本距離的峰值主要受資料集特性影響，50_speakers 資料集的距離峰值出現在接近採樣長度 12800 到 25600 間，而 urban 資料集則在採樣長度 25600 到 32000 間出現峰值。兩種資料集在採樣 2.6 到 3.0 秒的時候，攻擊耗時較長，攻擊成本較高，我們認為這可能是語

音資料的通性。

六、不同採樣長度的目標模型，攻擊結果與資料集原始樣本的歐幾里德距離相似時，音訊採樣長度增加，傳統攻擊耗時降低，滑動攻擊耗時無明顯變化。

隨著音訊長度增加，單位音訊之間的距離會縮小，需要較少的迭代次數即可攻擊成功，因此傳統攻擊耗時會降低。對於滑動攻擊而言，每次更新的範圍由固定的窗口大小決定，當攻擊目標音訊長度增加時，單位音訊的更新間隔也會增加，因此攻擊耗時不會隨音訊長度的增加而顯著變化。

七、滑動攻擊適合攻擊短音訊，傳統攻擊則適合攻擊長音訊。

當音訊長度較短時，微小的變化對模型輸出的結果影響較大，滑動攻擊針對小範圍的攻擊手法具優勢。但當攻擊目標長度增加時，輸入資料的小幅變化對攻擊結果的影響就不具明顯影響，傳統攻擊單位時間多次更新的策略反而更具優勢。

八、隨著模型架構的複雜度增加，模型逆向攻擊效果變差。

更複雜的模型導致梯度變化不明顯或梯度遮蔽問題，同時模型參數增多也提升每輪攻擊的計算量，攻擊難度上升，導致模型逆向攻擊耗時和攻擊結果與資料集原始樣本的歐幾里得距離增加，而置信度和攻擊成功率則下降。

九、對於複雜的模型，傳統攻擊的結果較滑動攻擊佳。

傳統攻擊每輪攻擊更新所有數據點，在模型複雜時更有優勢，而滑動攻擊在有限迭代次數下，每個樣本點更新較少，攻擊效果相對較差。

十、以音訊作為攻擊初始化，音訊的性質差異會影響攻擊結果：

使用與目標模型資料集相似的音訊資料作為攻擊初始化（如：50_speakers 目標模型搭配 TIMIT 音訊作為初始化）的攻擊耗時較短，攻擊置信度分數較高，然而，其攻擊效果仍比以 mean 和 random 兩種初始化樣本進行攻擊的效果差。這是由於音訊資料本身便具有個別特徵，在該音訊特徵上再進行增減以擬合出目標模型各個類的特徵，兩者相互影響下，會使攻擊結果與資料集原始資料距離越來越遠，因此音訊資料不適合用於攻擊初始化。

十一、 以差分隱私算法作為語音模型的防禦機制效果佳：

加入差分隱私算法後，目標模型的分類性能僅有微幅下降，不超過 5%。同時，該算法顯著增加了攻擊耗時，降低了攻擊置信度和成功率。特別是對於 Raw 和 SincNet 架構，其攻擊成功率下降至約 50%，展現良好的防禦效果。相比之下，MFCC 架構的防禦效果較差。總體而言，我們設計的差分隱私算法有效提升了模型的防禦性能。

十二、 Miface 攻擊算法的梯度算法使其不適合做為語音與時頻譜分類模型的攻擊算法

MIface 是為人臉分類模型的模型逆向攻擊所設計。然即便時頻譜模型是一種二維影像分類模型，使用 MIface 函數進行攻擊時，其攻擊耗時明顯比傳統攻擊更長，攻擊成功率僅 6%。它在時頻譜資料上的攻擊效果並不比在單維語音資料上更佳，這與 MIface 函數中梯度計算的自定義預處理步驟和梯度剪裁有關，使得攻擊過程中梯度偏低，更新幅度有限，難以攻擊成功。

十三、 時頻譜模型的攻擊結果較單維語音分類模型更接近資料集原始樣本

時頻譜模型提供了語音的能量、時間與頻率等衍生特徵，而單維度語音模型僅能利用波形與時間關係的特徵，使得攻擊算法在時頻譜模型能重建出更接近原始樣本的資料。

陸、 結論與未來展望

一、結論

(一) 具有波動特性的初始化參數可使攻擊生成的樣本更接近資料集原始樣本。

(二) 輸入資料格式為 MFCC 之模型，採用 mean 初始化樣本的攻擊結果與資料集原始樣本距離最近。

(三) 首層特徵提取架構會影響模型對輸入變化的敏感程度，進而影響攻擊的難易程度和攻擊樣本與資料集原始樣本之間的差異。

(四) 資料集中各類別樣本之間的差異性會影響模型逆向攻擊的效果，類別差異較大的資料集會使攻擊生成的樣本置信度更高，與原始樣本的歐幾里得距離更短。

(五) 不同的音訊採樣長度對模型逆向攻擊的耗時和攻擊樣本與資料集原始樣本的歐幾里得距離正相關，在採樣 2.6 到 3.0 秒時，攻擊成本較高。

(六) 目標模型採樣長度不同且攻擊結果與資料集原始樣本的歐幾里德距離相似時，音訊採樣長度增加，傳統攻擊耗時降低，滑動攻擊耗時無明顯變化。

(七) 滑動攻擊適合攻擊短音訊，傳統攻擊則適合攻擊長音訊。

(八) 隨著模型架構的複雜度增加，模型逆向攻擊效果變差，傳統攻擊法攻擊複雜模型具優勢。

(九) 使用與目標模型資料集相似的音訊資料作為攻擊初始化，可獲得較佳的攻擊效果。但相比 mean 和 random 初始化樣本的攻擊結果，使用音訊資料作為初始化樣本的攻擊效果較差，攻擊結果與資料集原始樣本的距離較大。

(十) 採用我們設計的差分隱私算法對 Raw 和 SincNet 架構的語音模型有顯著防禦效果。

(十一) Miface 攻擊算法的梯度算法使其不適合做為語音與時頻譜分類模型的攻擊算法。

(十二) 時頻譜模型的攻擊結果較單維語音分類模型更接近資料集原始樣本。

二、未來展望

(一) 針對不同的採樣長度和模型複雜度，本研究揭示了不同攻擊方法的適用場景以及語音模型的共通特性，望能結合本研究的發現，開發綜合性能較佳的攻擊算法。

(二) 針對不同性質的聲音資料集進行攻擊實驗，深入探討資料集特性對模型逆向攻擊的影響，並試圖提出系統性解決方法，提高攻擊結果與資料集原始樣本的相似度。

(三) 未來加入一種資料集的預處理方式，降低資料集各個類之間的差異性，並探討這種防禦手法的可行性。

(四) 未來嘗試透過 GAN 技術，將攻擊結果還原為人耳可識別的聲音訊號，並探討該操作對攻擊結果與原始樣本的差異性影響。

(五) 未來嘗試將預訓練 GAN 音訊生成模型微調，實現模型逆向攻擊，並與傳統梯度攻擊方法進行比較。

柒、 參考資料及其他

[1] Karla Pizzi, Franziska Boenisch, Ugur Sahin, Konstantin Bottinger (2023, Jan 9)。Introducing Model Inversion Attacks on Automatic Speaker Recognition。Fraunhofer AISEC, Germany; Technical University Munich, Germany。https://arxiv.org/pdf/2301.03206.pdf

[2] Mirco Ravanelli, Yoshua Bengio∗ (2019, Aug 9)。SPEAKER RECOGNITION FROM RAW WAVEFORM WITH SINCNET。Mila, Université de Montréal, ∗CIFAR Fellow。https://arxiv.org/pdf/1808.00158.pdf

[3] Tom Bäckström, Okko Räsänen, Abraham Zewoudie, Pablo Pérez Zarazaga, Liisa Koivusalo, Sneha Das, Esteban Gómez Mellado, Mariem Bouafif Mansali, Daniel Ramos (2022)。Introduction to Speech Processing, 2nd Edition。https://speechprocessingbook.aalto.fi/Recognition/Speaker_Recognition_and_Verification.html

[4] 謝采峰 (2021, Jan 3)。梅爾倒頻譜係數。聲音專有名詞筆記。https://hackmd.io/@-WoU5yJxR-2n0RdDCnLwDg/ryflpKVhP#梅爾倒頻譜係數 MFCC，Mel-Frequency-Cepstral-Coefficients

[5] Wai-Kit Tang (2020, Mar 3)。測試員所需知識:黑盒測試、白盒測試、灰盒測試之間的區別。Alvin's Blog 部落格。https://blog.vvtitan.com/2020/03/測試員所需知識黑盒測試、白盒測試、灰盒測試之/

[6] Elise Devaux (2022, Dec 21)。What is Differential Privacy: definition, mechanisms, and examples。Statice。https://www.statice.ai/post/what-is-differential-privacy-definition-mechanisms-examples

[7] Dayong Ye, Sheng Shen, Tianqing Zhu*, Bo Liu and Wanlei Zhou One Parameter Defense - Defending against Data Inference Attacks via Differential Privacy https://arxiv.org/abs/2203.06580

[8] Aladdin Persson (2020, Aug 8)。TensorFlow 2.0 Beginner Tutorials。YouTube。https://youtube.com/playlist?list=PLhhyoLH6IjfxVOdVC1P1L5z5azs0XjMsb&feature=shared

[9] Python 使用 librosa 分析聲音訊號、音樂檔案教學與範例 (n.d)。Office 指南。
https://officeguide.cc/python-librosa-package-analysis-music-and-audio-tutorial-examples/
[10] beat-buesser (2018)。Trusted-AI / adversarial-robustness-toolbox。GitHub。
https://github.com/Trusted-AI/adversarial-robustness-toolbox
[11] Michael Fekadu (2019)。DARPA TIMIT Acoustic-Phonetic Continuous Speech。kaggle。
https://www.kaggle.com/datasets/mfekadu/darpa-timit-acousticphonetic-continuous-speech
[12] Justin Salamon, Christopher Jacoby, Juan Pablo Bello (2014, Nov)。URBANSOUND8K DATASET。https://urbansounddataset.weebly.com/urbansound8k.html
[13] Vibhor Jain (2019)。Speaker Recognition Audio Dataset。kaggle。
https://www.kaggle.com/datasets/vjcalling/speaker-recognition-audio-dataset

【評語】190009

對語音攻擊防禦模型有深入探討, 提出多樣模型架構與實驗, 敘述與實驗尚屬完整。

本作品在 AI Cybersecurity 領域，具有實用性發展潛力，建議可繼續研究。