# AI生成語音辨識及應用

B103040044 林廷宇
B103040045 楊貽婷
B103040047 周安

# 01
# 前言


兴AI语音诈骗
受害者听不出是机器人！


AI語音機器人
您好 我們是理財調查小組
透過大家一起討論交流
台北市 詐騙集團AI語音機器人
激似真人！詐團省成本 AI機器人騙投資
TVBS NEWS
掌握新聞脈動 ▶ 訂閱TVBS NEWS頻道


好友微信借錢應急
原來是
AI換臉擬聲假扮
科技商被騙走430萬元
過數後致電朋友即知「中伏」
急截回336萬元


王仁甫粉絲被騙300萬籲政府出面！
曝新手法「用女兒聲音AI行騙」

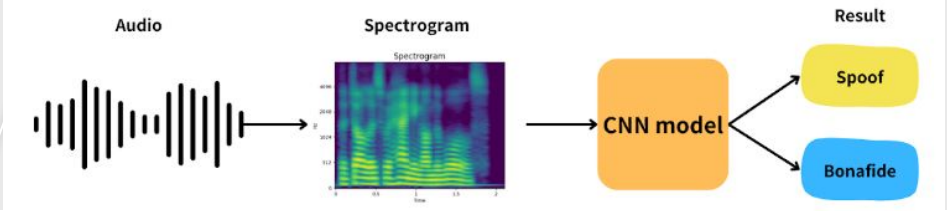
# 產品願景


情境一: **防止生成語音電話詐騙**
情境二: **防止生成語音破解聲紋門鎖**

# 02

# 作品介紹

# 實作流程圖




蒐集語音資料集

訓練CNN模型

利用Vitis-AI中的AI quantizer將模型參數量化成8位元的定點數

利用AI compiler將AI模型映射至硬體指令集，得到xmodel

設計C語言程式碼，內容有使用Vitis-AI Library中的API_2，呼叫DPU執行xmodel，做AI推論，並且自行設計前處理跟後處理方式，設計要將推論後結果的如何應用

編譯此C語言程式碼，得到執行檔

將xmodel與執行檔放上KV260板子

麥克風收音，經由前處理得到頻譜圖，開始運行

# 中英文語音資料集

英文資料集

- ASVspoof 2019

中文資料集

- CFAD
- AISHELL-3
- LJSpeech
- Text-To-Speech自行設計文本生成語音
- Speech-To-Speech:拿真實語音生成AI合成語音
- 利用open source code來使用他的模型生成語音

Suno Bark [Source code]. https://github.com/suno-ai/bark

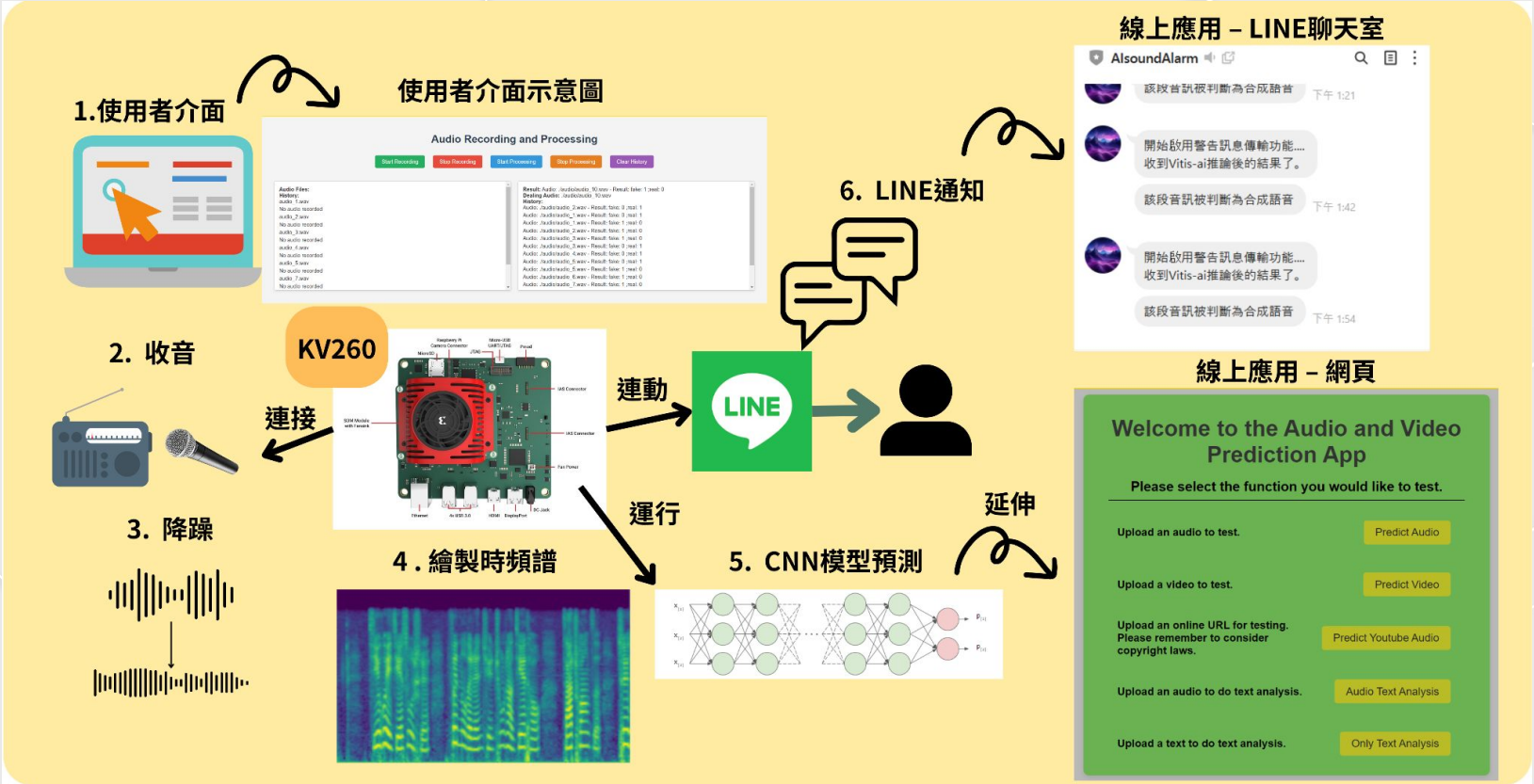
# 作品規格

**KV260特點:**

**1. DPU(Deep learning process unit): Xilinx開發的硬體加速器**

**2. 可以結合多媒體 : 麥克風、螢幕、錄影鏡頭**

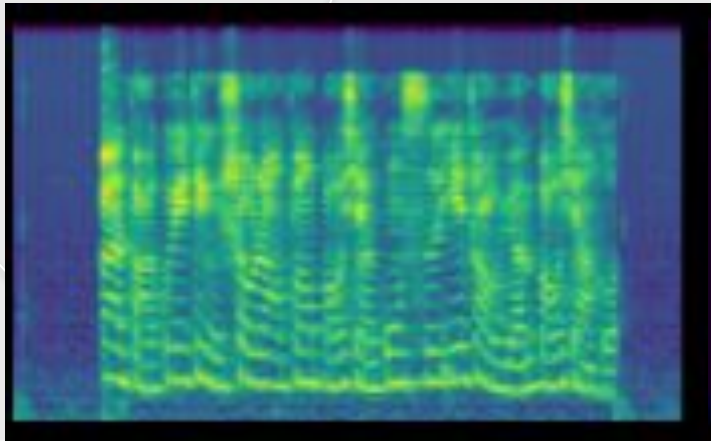
**3. 可以自行設計前處理、後處理方式**



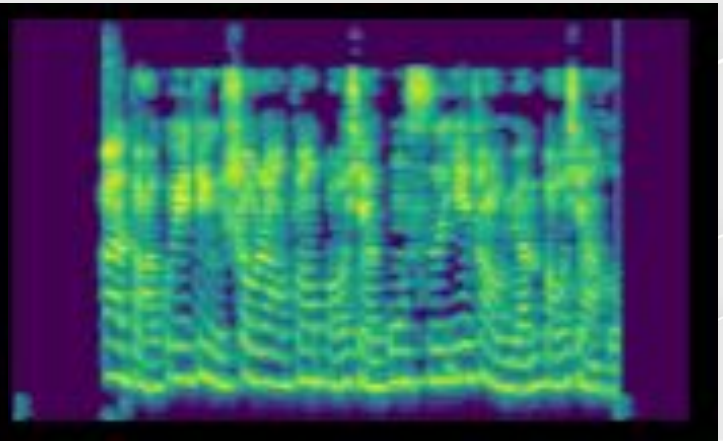
# 前處理-對音檔的處理

1. SoX 降躁語音
因為經過錄製，音檔跟原始音檔有所變化，包含音量變小、噪音變多。

**原始音檔：**



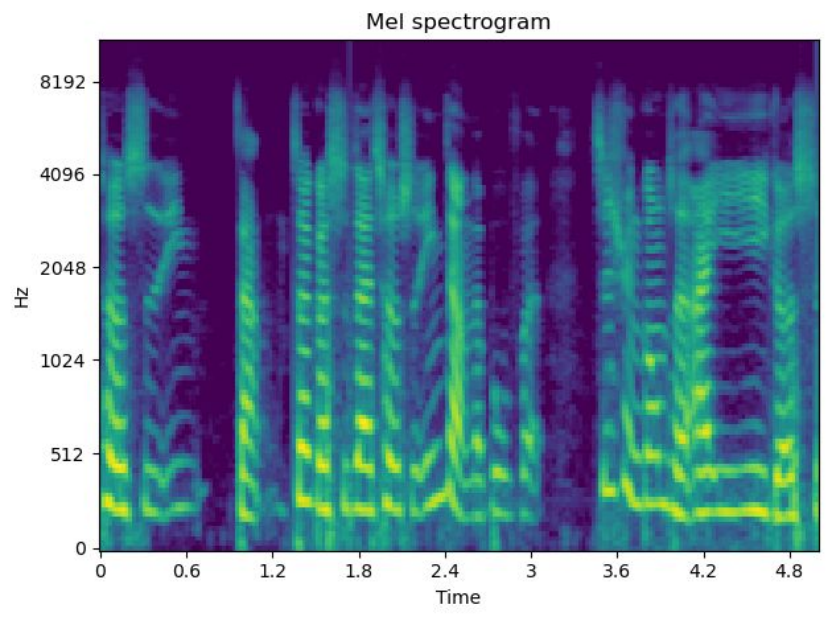
**經過SoX降躁處理過後：**



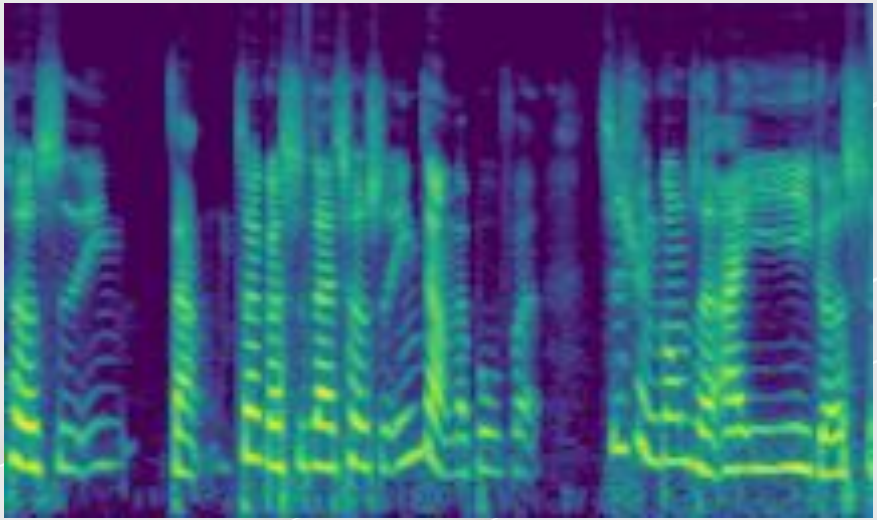
# 前處理-對音檔的處理

2. C++ 模擬Python librosa繪製Mel_spectrogram，使在板上繪製圖片只需不到一秒(原先需要大於20秒)，成功達到**即時性**。

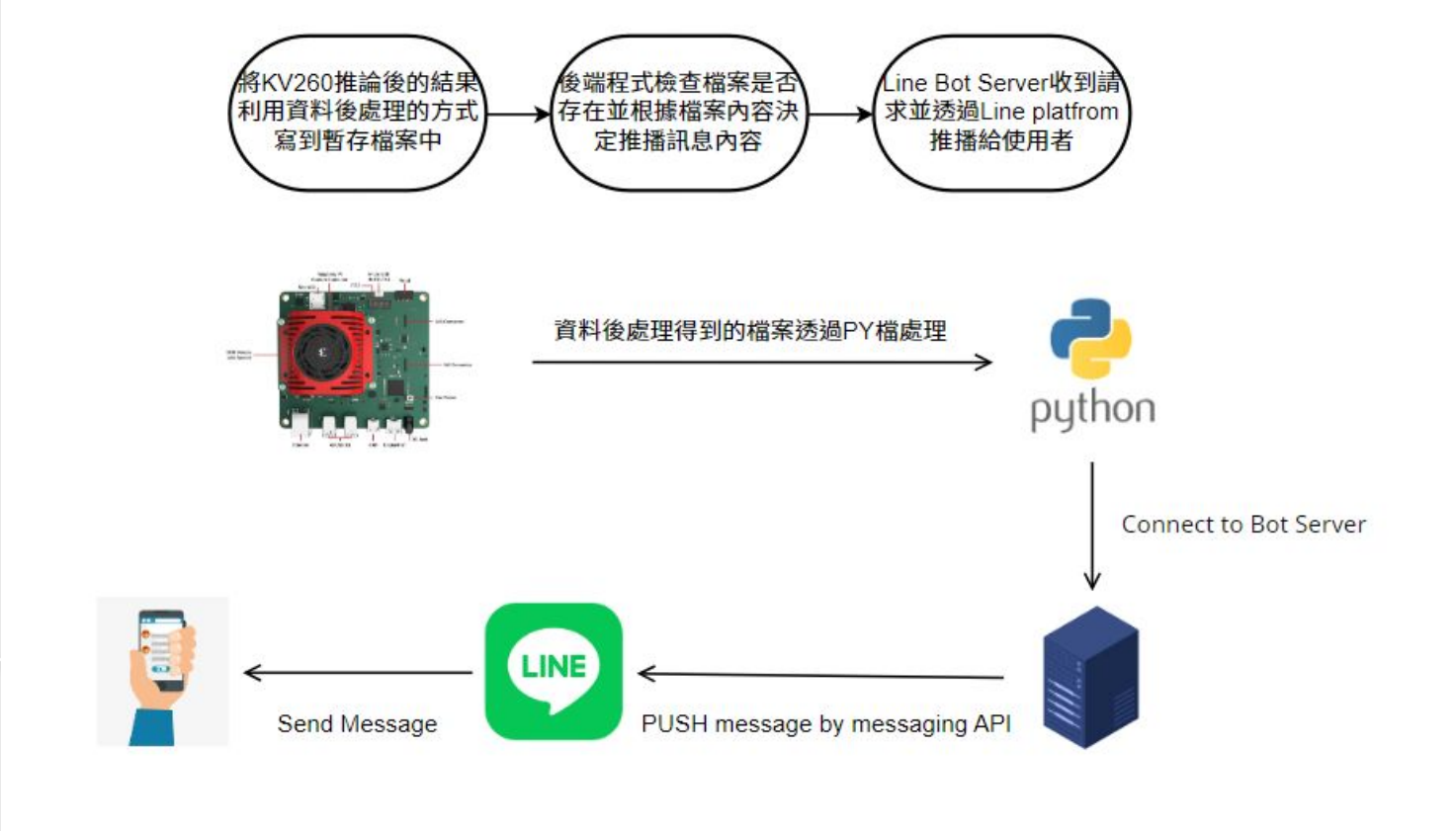
**Python librosa 產生的時頻譜圖**




**利用C++模擬，產生的時頻譜圖**



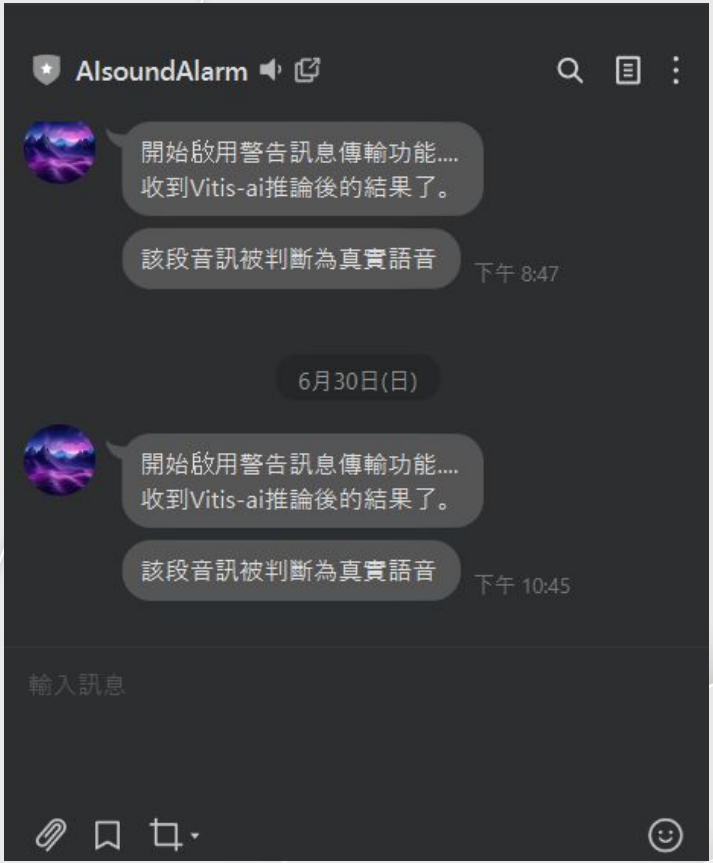
# 後處理-LineBot 介紹



# 03

# 作品成果

# LineBot 結果


AIsoundAlarm
開始啟用警告訊息傳輸功能....
收到Vitis-ai推論後的結果了。
該段音訊被判斷為真實語音
下午 8:47
6月30日(日)
開始啟用警告訊息傳輸功能....
收到Vitis-ai推論後的結果了。
該段音訊被判斷為真實語音
下午 10:45
輸入訊息

# 模型在軟體上的準確率 -英文部分 (資料集來源：ASVspoof2019)

- 3M_ENG相較於原先參考的模型，增加了中間卷積層的 Channel數

| | 訓練集 | 驗證集 | 測試集 |
|---|---|---|---|
| 真實語音數量 | 2500 | 1000 | 1000 |
| 生成語音數量 | 2500 | 1000 | 1000 |
| 總和 | 5000 | 2000 | 2000 |

| | 驗證集<br>準確率 | 測試集 (Eval)<br>準確率 | 測試集 (Eval)<br>Recall | 測試集 (Eval)<br>Precision | 測試集<br>(Eval)F1-score |
|---|---|---|---|---|---|
| EfficientCNN[1] | - | - | - | - | 94.14 |
| CNN_ENG | 96.65% | 90.20% | 86.10% | 93.79% | 89.78 |

[1] Subramani, N., & Rao, D. (2020). Learning Efficient Representations for Fake Speech Detection. Proceedings of the AAAI Conference on Artificial Intelligence, 34(04), 5859-5866. https://doi.org/10.1609/aaai.v34i04.6044

# 模型在軟體上的準確率 -中文部分 (資料集來源：各方蒐集)

- CNN_big相較於原先參考的模型，增加了中間卷積層的 Channel數

資料集包含：

1. 自行生成、蒐集的中文語音當作合成語音　2. CFAD、AISHELL-3中的真實中文語音

| | 訓練集 | 驗證集 | 測試集 |
|---|---|---|---|
| 真實語音數量 | 2200 | 500 | 300 |
| 生成語音數量 | 1900 | 473 | 274 |
| 總和 | 4100 | 973 | 574 |

| | 驗證集<br>準確率 | 測試集<br>準確率 | 測試集<br>Recall | 測試集<br>Precision | 測試集<br>F1-score |
|---|---|---|---|---|---|
| CNN_CH | 99.49% | 99.13% | 98.54% | 99.63% | 99.08 |

AISHELL-3: https://www.aishelltech.com/aishell_3

# 在KV260上直接輸入時頻譜圖片 準確率給DPU推論-英文部分

**隨機在測試集中抓取 100張圖片，抓取多組**

| | test_ENG_100_1 | test_ENG_100_2 | test_ENG_100_3 | 平均 |
|---|---|---|---|---|
| **準確率** | **85%** | **88%** | **90%** | **87.67%** |

# 在KV260上直接輸入時頻譜圖片 準確率給DPU推論-中文部分

**隨機在測試集中抓取 100張圖片，抓取多組**

| | test_CH_100_1 | test_CH_100_2 | test_CH_100_3 | 平均 |
|---|---|---|---|---|
| **準確率** | 99% | 98% | 99% | 98.67% |

# 透過麥克風收音做判斷的準確率 -英文部分

**隨機在測試集中抓取 100段語音**

| | Predicted | | |
|---|---|---|---|
| Actual | | Bonafide | Spoof |
| | Bonafide | 48 | 2 |
| | Spoof | 11 | 39 |

| | Accuracy | Precision | Recall | F1-score | Number |
|---|---|---|---|---|---|
| Bonafide | 87% | 81.36% | 96% | 88.08 | 50 |
| Spoof | | 95.12% | 78% | 85.71 | 50 |

# 透過麥克風收音做判斷的準確率 -中文部分

**在測試集中測試共 574段語音**

| | Predicted | | |
|---|---|---|---|
| Actual | | Bonafide | Spoof |
| | Bonafide | 270 | 30 |
| | Spoof | 52 | 222 |

**小結:**

- **透過麥克風收音，增加了一些錄音上的雜音與其他因素，使錄製到的音檔與原始音檔有所偏差。才導致準確率略為下降。**

| | Accuracy | Precision | Recall | F1-score | Number |
|---|---|---|---|---|---|
| Bonafide | 85.71% | 83.86% | 90% | 86.82 | 300 |
| Spoof | | 88.1% | 81.02% | 84.42 | 274 |

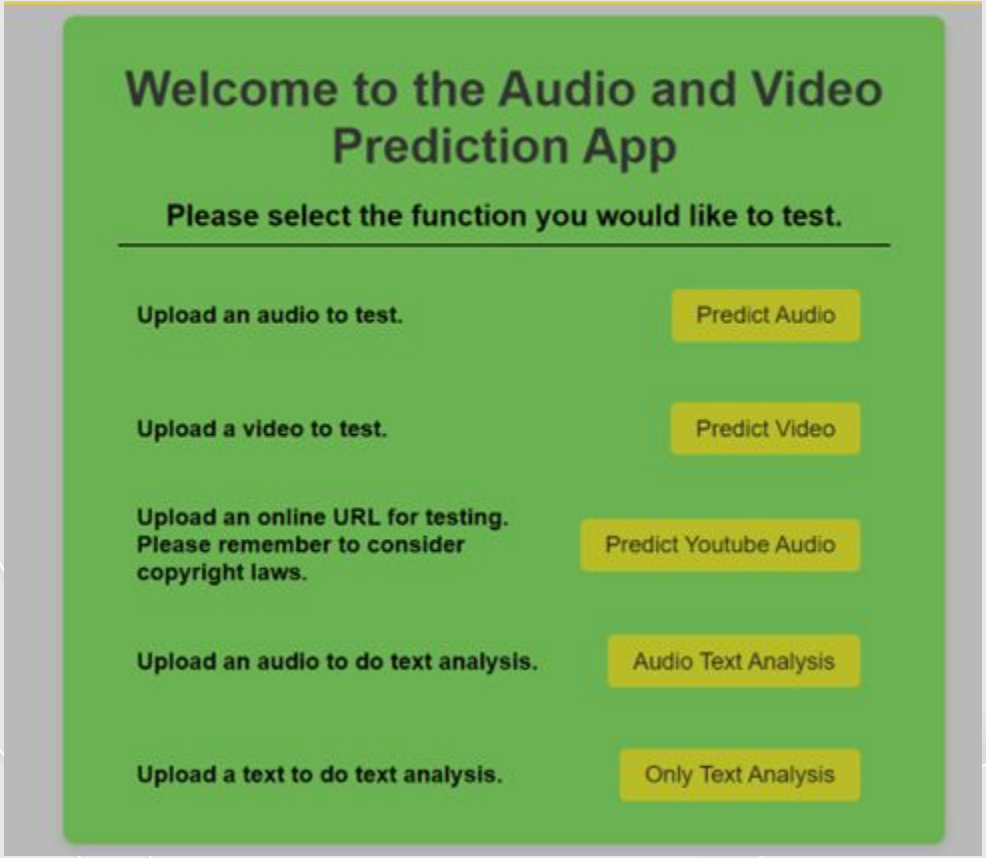
# 04

# 線上應用

# 簡介

除了嵌入式系統以外，我們也將**同樣功能的模型套用到其他應用上**，如LINE Bot和網頁，提供給使用者上傳欲辨識的內容，然後我們的後端會將辨識結果傳回去給使用者，與嵌入式系統的不同之處為模型運作的運算能力取決於**後端電腦上的運行能力**，使用者只須要依照指示進行操作即可，當之後要開發出更廣的受眾時，未來就必須嘗試架設可以持續運行的伺服器使用者隨時使用。

# 線上應用


Welcome to the Audio and Video Prediction App
Please select the function you would like to test.
Upload an audio to test.
Predict Audio
Upload a video to test.
Predict Video
Upload an online URL for testing. Please remember to consider copyright laws.
Predict Youtube Audio
Upload an audio to do text analysis.
Audio Text Analysis
Upload a text to do text analysis.
Only Text Analysis


AIsoundAlarm
00:06
選擇語言
請選擇你要辨識的語言
注意!! 判斷合成語音的過程需要一點時間且你選擇的語言會影響結果
中文
英文
其他
是否執行文本分析?
請選擇是否執行對語音內容進行潛在詐騙內容分析
是 (執行)
否 (不執行)

# 線上應用新增功能
# LLM - Gemini API
# 詐騙風險評估

## 詐騙分析報告

**一、簡要概述：**

輸入文字是關於草的描述，描述了草的柔軟、彎曲等特性，並帶有擬聲詞。

**二、詐騙可能性：**

極低。

**三、詐騙類型：**

無。

**四、分析原因：**

1. 輸入文字描述的是草的特性，並無任何與詐騙相關的訊息。
2. 文本中沒有任何誘導性或欺騙性字眼。
3. 文本整體缺乏詐騙常用的手法，例如提供虛假訊息、引誘受害者提供個人資料等。

**五、防範建議：**

對於任何可疑的訊息，請保持謹慎，不要輕易相信陌生人或提供個人資訊。

**六、結論：**

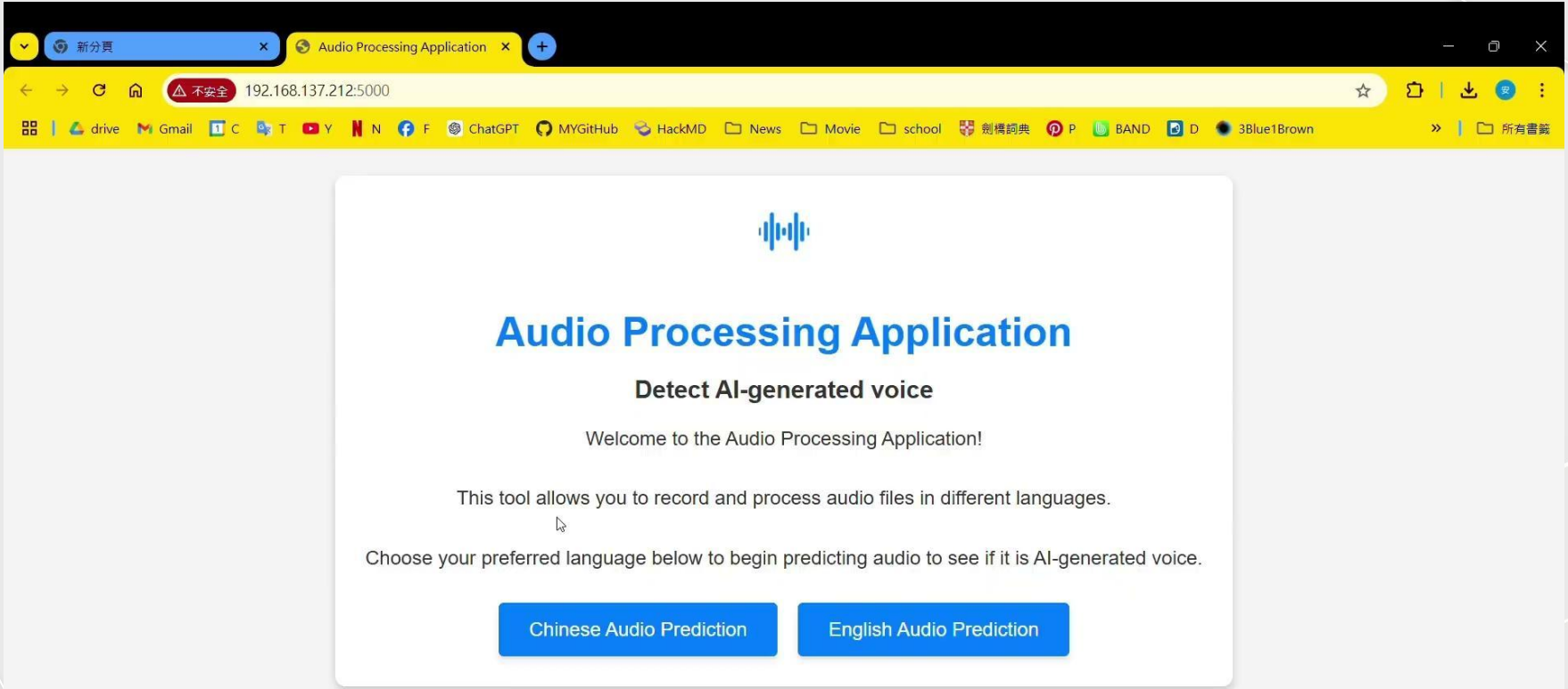
輸入文字並無任何詐騙跡象，屬於正常的文字描述。

下午 10:13

# 05

# DEMO

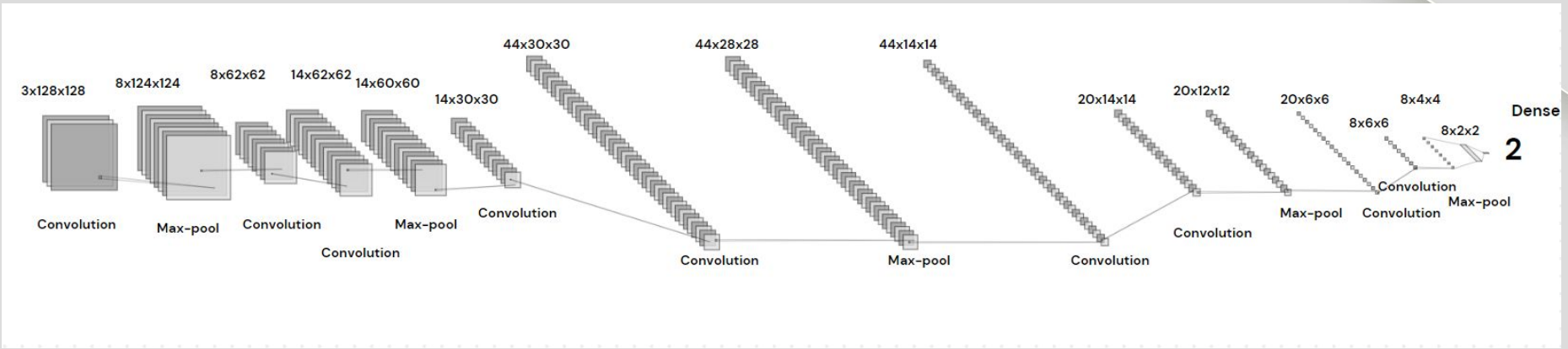
# DEMO VIDEO https://youtu.be/tRtrg5NVRNE

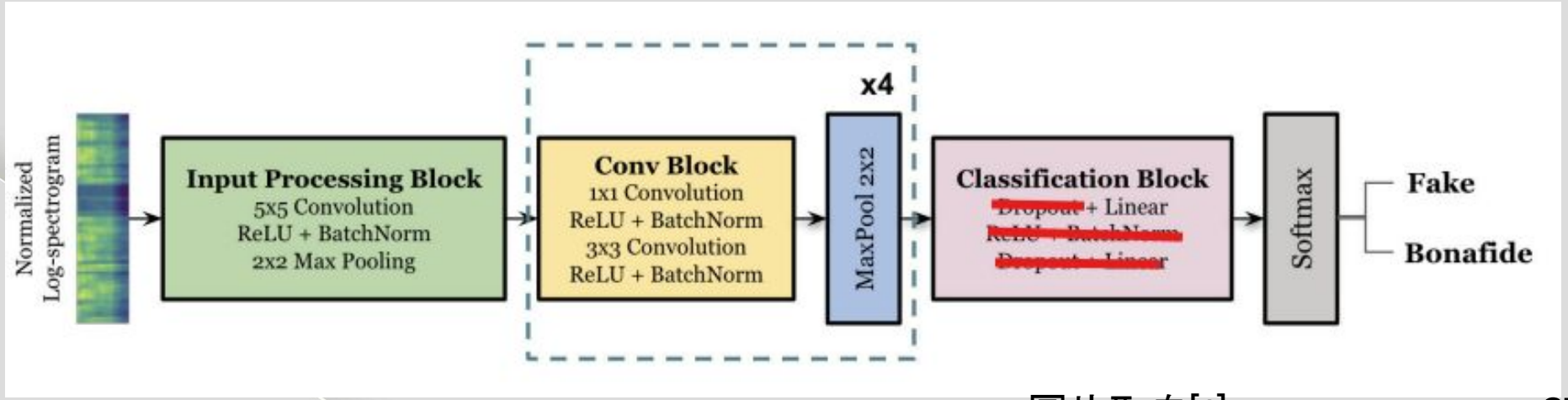



進入使用者頁面-首頁　　選擇語言

# Thank you for your listening

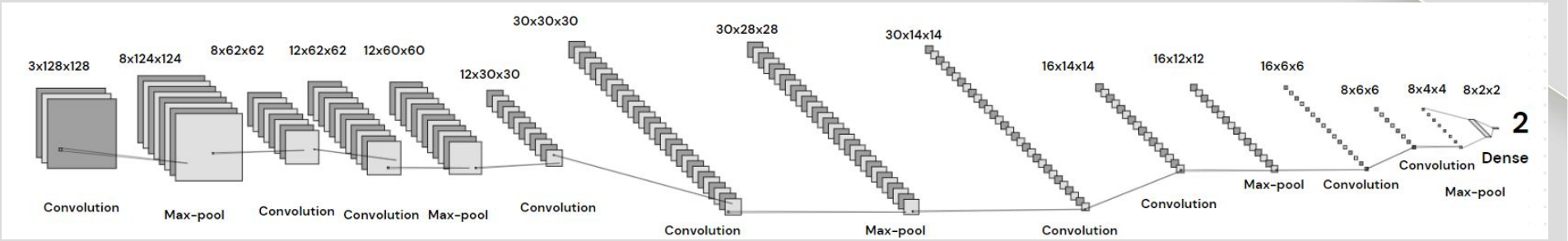
# 英文模型架構

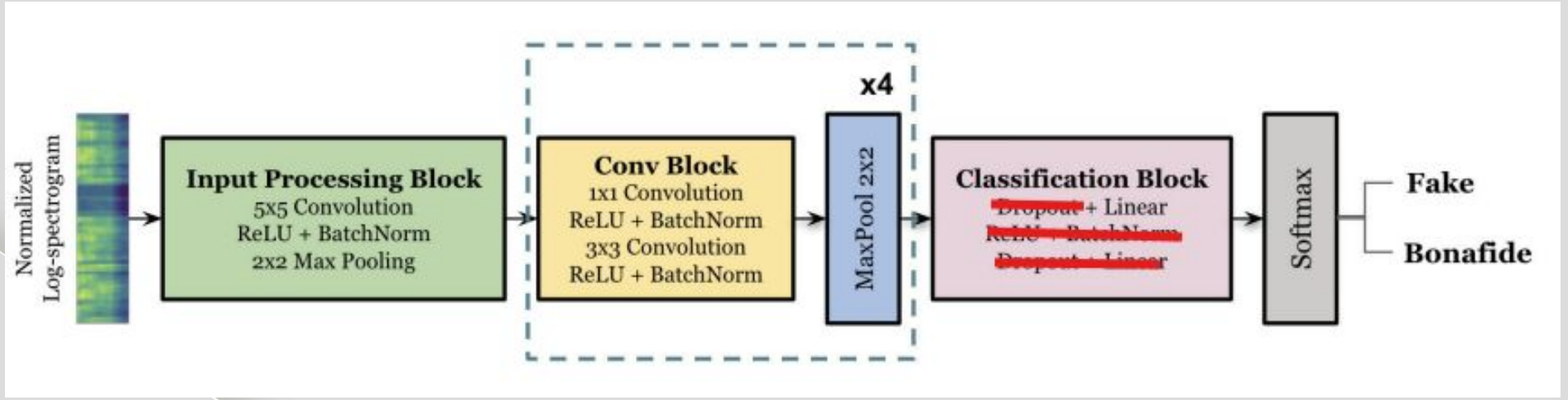



圖片取自[1]

# 中文模型架構




圖片取自[1]

# 運作流程

1. 按下執行收音程式，開始收音
2. 輸入終止符號，代表停止收音，錄音結束
3. 音源每五秒切段，轉換成一張一張的Spectrogram
4. 對個別圖片進行運算、預測
5. 採用多數決，決定這是否為合成語音

# 作品運作流程圖


slidesgo