iSurePlug Pro聲音辨識 AI 模型建立專案規劃

本專案預計執行時間：

環境聲音辨識：2023年12月

人聲語音辨識：2024年 1月

整體實機測試：2024年2月

1. 聲音處理
   1. 聲音種類

我們區分聲音為環境聲音和人聲語音兩種。由於兩個使用的模型不同，需要各自獨立判別。

在實際執行時，我們是以一秒鐘為一個單位，目前先假設

* 處理環境聲音小於0.3 秒，
* 處理人聲語音小於 0.4秒。
  1. 聲音格式

我們目前系統的要求，

* 聲音的取樣頻率為16KHZ,
* 每個取樣為16-BIT，
* 雙聲道或是單聲道輸入，但是資料處理為單聲道
* 每段聲音片段為 1 Second.
  1. 需要判別的環境聲音

語音的部分，因為是用另外一種MODEL，在這裡先不規劃，但也會用同樣的概念來實現。

|  |  |  |
| --- | --- | --- |
|  |  | 重要等級, 5: 最重要，1不重要 |
| 廚房餐廳 | 廚房火警警報器 | 5 |
|  | 煙霧偵測器/瓦斯偵測器 | 5 |
|  | 燒水聲音（笛聲） | 5 |
|  | 倒水聲？ |  |
|  | 熱水壺注水聲 | 2 |
|  | 熱水瓶出水聲 | 4 |
| 浴室 | 馬桶沖水聲 | 4 |
|  | 水龍頭洗臉台水聲 | 3 |
|  | 水龍頭儲水水聲 | 3 |
|  | 蓮蓬頭水聲 | 2 |
|  | 流水聲 | 3 |
| 人聲（非語音） | 咳嗽聲 | 4 |
|  | 打噴嚏聲 |  |
|  | 呻吟聲 | 不好做 |
|  | 喘息聲 | 不好做 |
| 鈴聲 | LINE 鈴聲 | 3 |
|  | 微信鈴聲 | 3 |
|  |  |  |
| 其他 | 腳步聲 | 2 |
|  | 關門聲 | 2 |
|  | 重摔聲 | ？ |
|  |  |  |
|  |  |  |
|  | 以下應該使用語音模型 |  |
| 求救聲 | 救命（國，台，客）Help，日文，粵語 | 5 |
|  | 是，對（國，台，客），はい， Yes, | 4 （做確認用） |
|  | 不是， |  |
|  |  |  |

目前先要求能夠辨識上面用黃色標誌的5種聲音：

* 火警警報器，
* 煙霧偵測器/瓦斯偵測器，
* 燒水笛聲，
* 電熱水瓶出水聲，
* 馬桶沖水聲。
* 咳嗽聲
  1. 聲音資料Golden Samples

我們可以從網路下載各種語音的檔案，作為我們測試的Golden Samples.

Golden Samples 可能也需要調整音量，模擬距離影響聲音的情形。

之後如果有判斷不對的聲音，也必須放入Golden Samples 在重新測試時使用。

* 1. 混音資料

我們需要把不同聲音混合起來，看看是否能夠正確偵測。

我們需要把環境聲音和人聲語音混合起來，看看是否能夠正確偵測。

舉例來說，我們需要測試水聲混上人求救的聲音，模擬人在浴室洗澡滑倒求救的情境。

這個可能在比較後面再測試。

1. 環境聲音識別實現過程

最終實際模型（不是在PC上執行），希望能夠

* SRAM的使用，不超過100K，
* 固定的資料（ROM/FLASH）不超過500K,
* 執行時間不超過 0.3 second
  1. 第一階段：建立操作模型

本階段作為模型的訓練學習階段，利用資料檔案的各個音檔來做為訓練的數據，產生可以使用的模型。

學習開始時，需要可以指定期待的模型大小。首次學習可以假設有比較大的記憶體，之後的幾次學習，會逐漸縮小SRAM 和 Flash 的允許大小。Layer 的層數也需要可以指定。

|  |  |  |
| --- | --- | --- |
| 執行平台 | PC |  |
| RAM 需求 | 無限制 |  |
| ROM/FLASH | 無限制 |  |
| 輸入聲音來源 | 檔案 | （Golden Samples） |
| 執行時間 | 不限 |  |

學習過程必須能夠自動化，利用script 或是程式，自動將每個Golden Samples音檔載入然後進行學習，最後產生模型的資料，並且產生報表（可利用Excel），顯示目前Golden Samples 的測試結果。使用Golden Samples 需要準確率達到 99%。

* 1. 第二階段：測試原始模型

本階段作為模型的測試階段，將產生出來的模型，來作為AI 判斷的依據。

每個模型，需要能夠知道需要的RAM 大小（包括static 和 dynamic）, 以及常數所需要的大小。

|  |  |  |
| --- | --- | --- |
| 執行平台 | PC |  |
| RAM 需求 | 無限制 |  |
| ROM/FLASH | 無限制 |  |
| 輸入聲音來源 | 檔案 | （Real Samples） |
| 執行時間 | 不限 |  |

測試的過程必須能夠自動化，利用script 或是程式，自動將每個Real Samples音檔載入然後進行測試， 並且產生測試結果的報表（可利用Excel）。使用Real Samples 需要準確率達到 95%。如果無法達到，需要將失敗的音檔，放入第一階段的 Golden Samples，重新跑訓練過程來得到新的模型。

* 1. 第三階段：優化模型

本階段作為模型的優化階段。每次優化，希望能夠把需要的 RAM/ROM 大小減少 1/3 到 1/2. 或是其他方式（減少Layer）。

每個模型，需要能夠知道需要的RAM 大小（包括static 和 dynamic）, 以及常數所需要的大小。

|  |  |  |
| --- | --- | --- |
| 執行平台 | PC |  |
| RAM 需求 | 無限制 |  |
| ROM/FLASH | 無限制 |  |
| 輸入聲音來源 | 檔案 | （Real Samples） |
| 執行時間 | 不限 |  |

測試的過程必須能夠自動化，利用script 或是程式，自動將每個Real Samples音檔載入然後進行測試， 並且產生測試結果的報表（可利用Excel）。使用Real Samples 需要準確率達到90%。如果無法達到，需要將失敗的音檔，放入第一階段的 Golden Samples，重新跑訓練過程來得到新的模型。

* 1. 第四階段：實際驗證模型

本階段作為實際系統實現的測試，需要能夠將整個MODEL 放入，透過I2S 收音，然後進行判斷，結果可以靠亮燈，或是透過 UART 傳送結果。

|  |  |  |
| --- | --- | --- |
| 執行平台 | ARM M4 | Artery AT32F435/288MHz |
| RAM 需求 | 100K |  |
| ROM/FLASH | 500K |  |
| 輸入聲音來源 | I2S 收音 | 現場測試 |
| 執行時間 | < 0.3 sec |  |

如果測試不盡理想，則需要將聲音錄檔，留作新的Golden Samples，再重複第一階段開始重新產生Model的資料。

1. 人聲語音識別實現過程

實現過程類似環境聲音識別，也是先用比較大的模型測試，再逐步減小模型尺寸和計算所需時間。

人聲需要分男聲，女聲，小孩聲（如果可以），老人聲（如果可以）。

語系需要有：

* 國語，閩南語，客家話 （可以一起打包訓練）
* 英語：包括美式英語，英式英語，菲律賓英語
* 日文
* 粵語：（需要George 提供）

詞彙要有

* 救命
* 是/對

模型是整體打包好，還是各語系分開，可能要實際測試才知道。

1. 其他