中原大學

資訊工程學系

113 學年度專題實驗期末報告

環境聲音即時辨識

組員:

資訊四乙 11027222 黃彥霖

資訊四乙 11027239 邱宥文

資訊四乙 11027254 方璿瑞

指導教授:湯松年 副教授

中華民國 113 年 10 月 11 日

摘要

本專題結合 LAION-AI 的 CLAP 模型與 Raspberry Pi，設計並實現了一套即時聲音辨識系統。該系統利用預訓練模型進行特徵提取，並透過麥克風捕捉即時音訊進行音頻處理。所使用的 CLAP 模型能將音訊與文本嵌入向量進行相似度比較，以準確識別聲音的類別並評估其信心程度。為提升辨識效果，本系統採用視窗步幅（window strides）技術，逐步分析音訊段落，與傳統固定大小窗口相比，顯示出準確率提升的趨勢。該系統可應用在 Raspberry Pi 等設備上運行，展現了其在資源受限環境中的即時音頻分類能力。

關鍵字: 對比學習，機器學習，聲音辨識，視窗步幅技術，嵌入式系統，低功耗。

目次

圖目錄

表目錄

第一章 緒言

1.1 計畫之背景與重要性:

在近年來，聲音辨識的技術越來越純熟也被運用到很多的地方，像是智能家居、醫療照護、環境監控及物聯網裝置等多樣應用場景，而且需求日益增加。然而，許多現在有的聲音辨識模型雖然具有高準確性，卻常因為它高計算需求而難以在資源受限的嵌入式裝備上運行。為了應對這一挑戰，LAION-AI 所開發的 CLAP 模型使用對比式學習(Contrastive Learning)的方法能，從聲音和語音中自動學習與語意相關的特徵，因此不僅可有效支持多樣化的聲音辨識任務，還能在有限的資源下，以軟體方法和演算法提高辨識的準確率，提高辨識的精度。這種技術帶來了在低資源設備中進行高效聲音辨識的可能性。

1.2 研究目的:

本研究旨在實現並驗證一套基於 LAION-AI 的 CLAP 模型和 Raspberry Pi 的低功耗即時聲音辨識系統，目標是探索在資源有限的嵌入式設備上如何應用機器學習技術和軟體設計，進行準確的聲音分類。透過本研究，希望可以在有限的硬體設備下實行此應用並保證它一定程度的準確率。

第二章 相關文獻

1.1 LAION_CLAP 音訊辨識架構

如下圖 Figure 1:

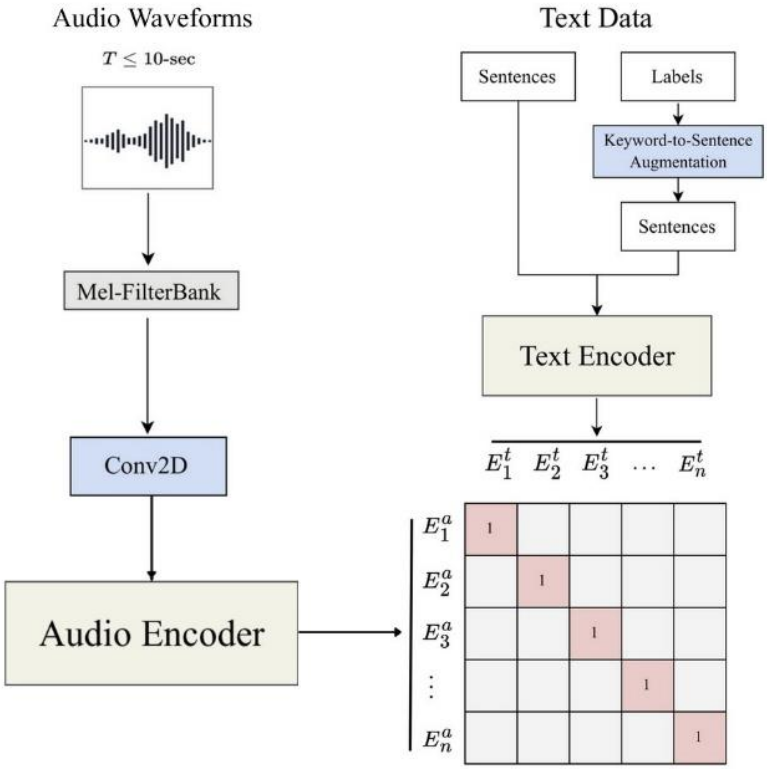

Figure 1 模型架構

1.1.1 音訊的前置處理:

為了確保所有音訊樣本具有一致的規格，避免因音訊長度不足而導致訊息遺失，先對音訊進行前置處理，採用重複循環與添加空白音訊的方式來統一音訊長度。例如，對於長度為 3 秒的音訊，將其重複三次並添加 1 秒的空白音，以填充至 10 秒的長度。過程中盡可能地重複音訊內容，以維持信號的一致性和完整性。

1.1.2 編碼器特徵提取:

Mel-Filter Bank:

使用 Mel 頻濾波器（Mel-Filter Bank）將音訊轉換為梅爾頻譜圖，以提取音訊的頻率特徵。此轉換可將音訊表示為頻率特徵圖，便於後續處理。採用卷積層（Conv2D）對梅爾頻譜圖進行處理，提取音訊的特徵，以捕捉音訊在時間和頻率上的變化。

1.1.3 音訊編碼器:

本研究採用 HTSAT 編碼器作為音訊特徵提取的核心，HTSAT 編碼器基於 Transformer 架構，能捕捉音訊在長時間範圍內的聲音關係和信息，例如背景聲音或持續變化的音訊。編碼器的輸出包括三種重要的音訊特徵：時間特徵、頻率特徵和語義特徵。時間特徵表示聲音隨著時間變化的模式；頻率特徵描述音訊中的頻率高低，並具備處理高頻和低頻噪音的能力；語義特徵則能夠捕捉音訊中的高層次信息，提取特定的事件或情境。最終編碼器將這些特徵整合，提供音訊的嵌入結果，用於後續的聲音分類或辨識。

1.1.4 文字編碼器:

文本數據來自標題(Caption)和標籤(Labels)，標籤進一步被轉化為自然語言描述，以提升音訊檢索的效果。應用關鍵字到標題增強技術（Keyword-to-Caption Augmentation），使用生成式模型對關鍵字進行語義擴展，提供更豐富的聲音特徵描述，克服標籤信息不足的問題，使每段音訊具備更詳細和自然的文本描述，增強了音訊與文本之間的對應關聯性。

使用 RoBERTa 作為文本編碼器，以有效提取文本中的語義特徵，將文本描述轉換為嵌入向量。RoBERTa 透過分詞技術將文本分解為語言文本的基本單位(Tokens)，確保模型能靈活處理新詞和罕見詞彙。每個基本單位被轉換為嵌入向量，提取詞彙的語義和語境信息。這些嵌入向量通過多層 Transformer 編碼器層進一步處理，使得每個語言文本的基本單位的嵌入能夠參考到整個文本，處理句子並學習其中的語義結構，最後提供文本嵌入結果。

1.1.5 MLP Layer:

MLP 層的設計目的是將來自不同編碼器（如文本和音訊）的嵌入特徵映射到相同維度的空間中以便進行比較。通過 MLP 層，文本和音訊的嵌入能夠在統一的嵌入空間中進行對比學習，從而使兩種特徵相互配對，減少不相關特徵而導致的干擾現象。

1.1.6 對比學習:

使用對比學習（Contrastive Learning）的方法來學習文本和音訊之間的關聯性。對比學習的是通過優化嵌入空間，使正樣本對（例如音訊與其對應的文本描述）的嵌入向量距離最小化，並使負樣本對（例如音訊與不相關的文本描述）的嵌入向量距離最大化。應用了對比損失函數（Contrastive Loss）與餘弦相似度（Cosine Similarity）兩種方法。對比損失函數提高正樣本對之間的相似度得分，並降低負樣本對的相似度得分；而餘弦相似度則用於計算音訊與文本嵌入之間的相似性，幫助模型學習嵌入空間中音訊與文本之間的關聯性。模型最終將相似度最高的對應結果作為預測輸出。

第三章 系統設計

1.1 硬體環境

硬體:Raspberry Pi 4 Computer Model B

作業系統: Raspberry Pi OS (64-bit)，

硬體設備:外接麥克風

1.2 軟體設計

1.2.1 開發環境:

python 3.10

PyTorch 1.11.0

1.2.2 軟體架構:

音頻輸入與前處理階段，首先收集環境聲音並將其從 float32 轉換為模型可接受的 int16 格式（.wav）。減少嵌入式系統運作時的成本，轉換後的音訊數據會存入佇列中等待辨識。基於即時聲音辨識的需求進行開發，目的在提供高效的音訊處理和分析。

1.2.3 視窗步幅技術

採用窗口和步幅(如下圖 Figure 2)的方式對環境音訊進行分割，窗口大小預設為 4 秒，步幅大小為 3 秒。窗口（window）用於將音訊分割成相同大小的區塊，每個區塊可以獨立提取特徵；步幅（stride）則是相鄰窗口之間的滑動距離，決定了窗口的重疊程度。假設有三種音訊分別為 A、B 和 C，空白為無聲音，每一方格為一秒，如下圖，傳統方法會因為固定視窗而忽略 B 的聲音，但透過視窗步幅的設

計，可以捕捉到邊緣的音訊，因此 B 特徵能夠被偵測。

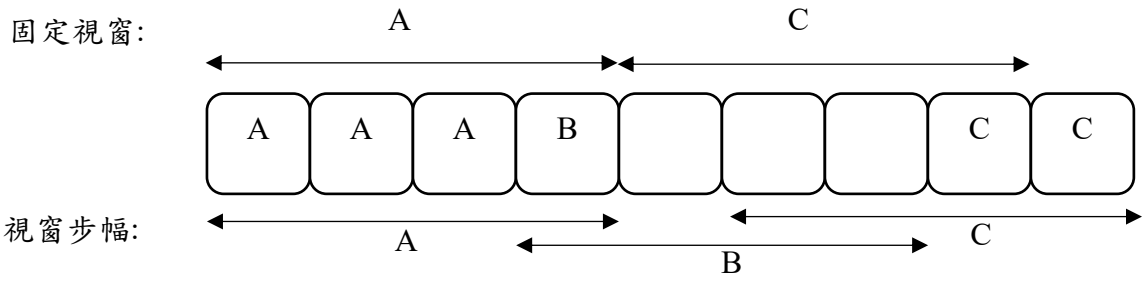

Figure 2 視窗步幅示意圖使

1.3 使用者介面

如下圖 Figure 3，本系統能根據使用者的操作時間，對環境聲音進行即時標註與分類，提供聲音辨識資訊。使用者能夠掌握各類聲音出現的時間區段，並可辨識每種聲音的開始與結束時刻。若使用者希望了解在特定期間內各聲音類別的出現的頻率，可透過選擇 Generate Results 按鈕生成報告。系統將自動產生完整的分析結果，方便使用者進行回顧和確認。

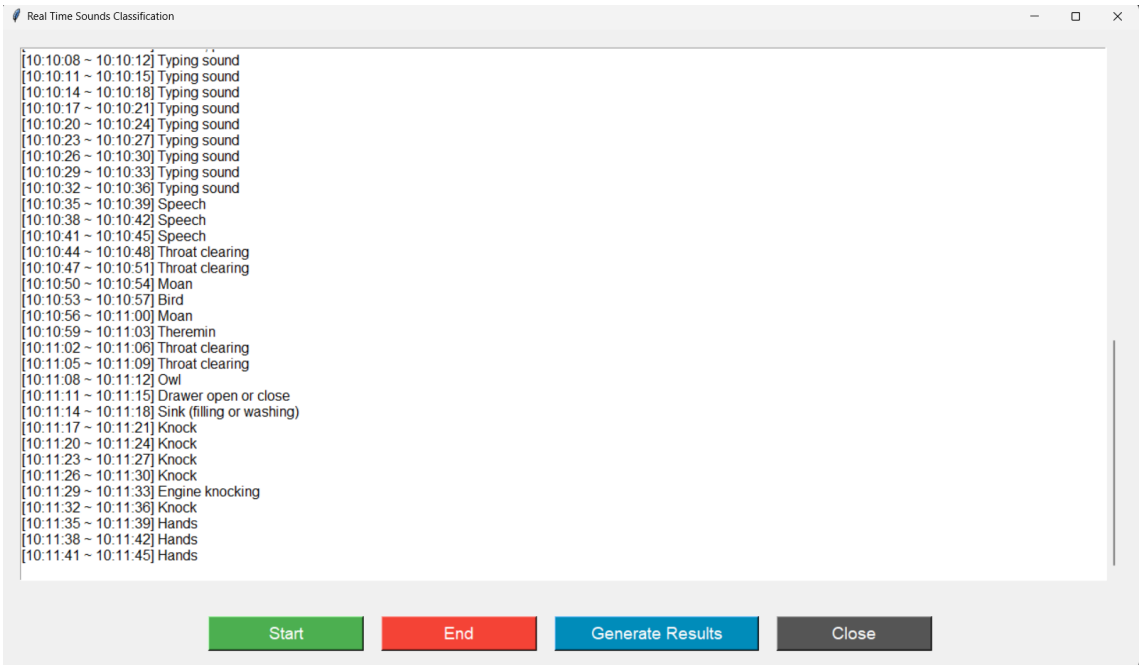

Figure 3 使用者介面

1.4 分析使用者環境聲音生成圖表

為了提供使用者能夠更好地理解環境中的聲音頻率，透過圖形和文字的方式將分析結果呈現給使用者，使其能查看特定時間段內各種聲音的出現次數及連續出現時間。幫助使用者清楚辨識主要環境中的聲音，了解哪些聲音最常出現或持續最久，進一步識別噪音來源，下面將介紹網頁包含的分析項目。

1.4.1 功能介紹:

考量到聲音的出現具有不可預測性，有些聲音僅出現於極短時間，而另一些則持續時間較長，本研究基於出現頻率及持續出現時間的雙重分析方法，協助使用者更全面地了解並反映周圍環境中的聲音特徵。

1.4.2 聲音出現頻率分析:

基於先前的音訊辨識結果，統計在特定時間段內出現的不同種類聲音，幫助使用者輕鬆理解最頻繁出現的聲音類型。如下圖 Figure 4:

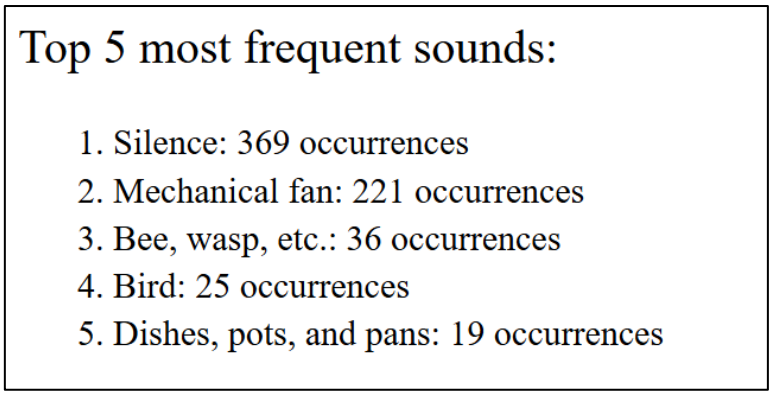

Figure 4 最高頻率出現聲音

1.4.3 連續出現次數分析:

分析每種聲音的連續出現時間，找出那些持續時間最長的聲音類型，以協助使用者了解環境中可能存在的持續噪音源。如下圖 Figure 5:

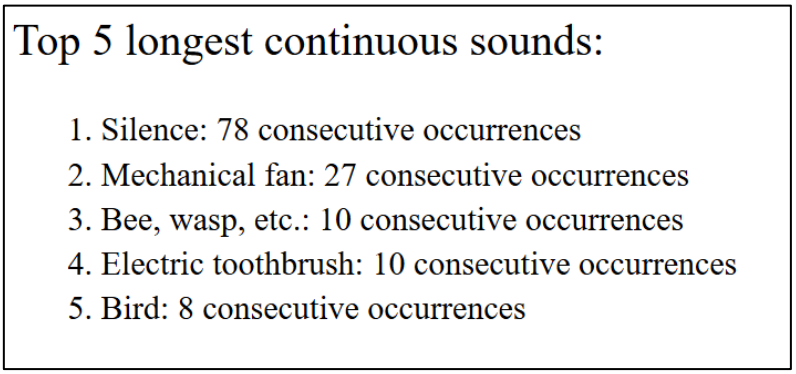

Figure 5 連續出現次數最高的聲音

1.4.4 圖形化展示:

透過圖形化展示，將聲音數據轉化為直觀的模式，以便使用者能快速掌握關鍵資訊並提升數據的可讀性。長條圖(如下圖 Figure 6)提供了展示不同類別之間差異的方式，顯示聲音特徵的分佈；圓餅圖(如下圖 Figure 7)則清晰呈現各類別所占的比例，理解各種聲音在整體中的出現比例及其差異。

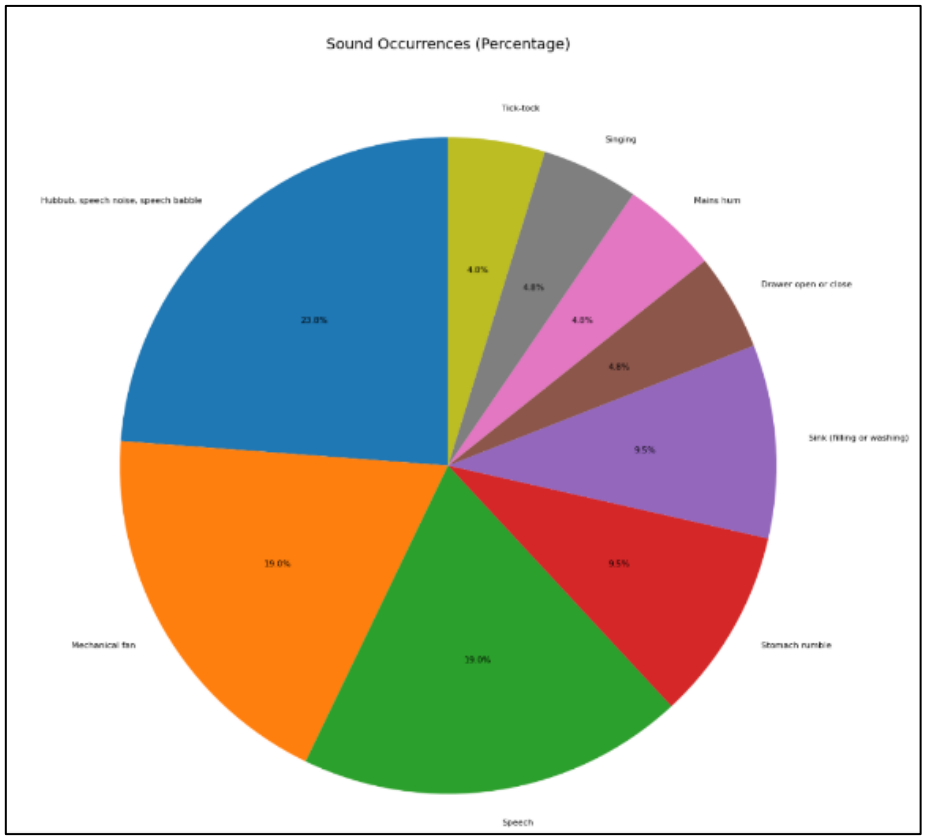

Figure 6 出現次數統計(圓餅圖)

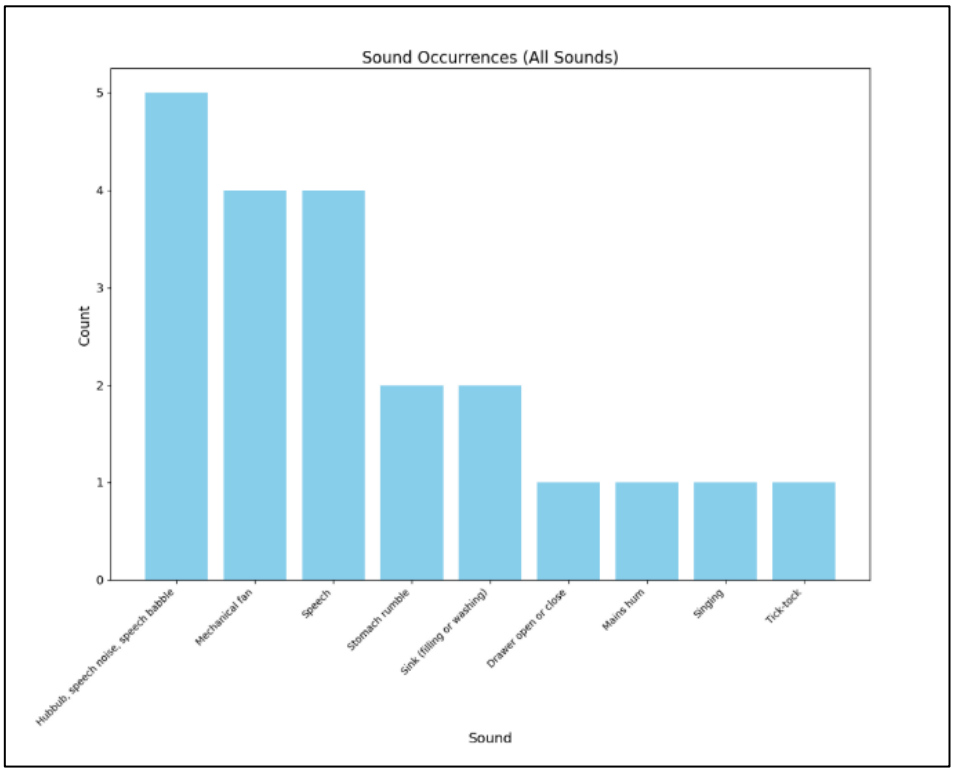

Figure 7 出現次數統計(長條圖)

第四章 實驗與討論

1.1 實驗與結果分析

此研究進行了兩種不同的實驗來評估系統的性能。

1.1.1 電腦與樹梅派準確率之差異：

本實驗從多個開源音訊資料集中隨機選取音訊檔案，透過外部裝置進行播放，再分別使用電腦與 Raspberry Pi 的麥克風進行錄音與辨識。

Table 1 電腦與樹梅派準確率之差異

環境	電腦	Raspberry Pi 4
準確度	77.81%	60.26%

1.1.2 傳統固定式窗口與視窗-步幅比較：

本次測試完全排除了即時收音的干擾，比較傳統固定式窗口與視窗-步幅模式的效果，去判斷是否準確率有增加。

Table 2 固定視窗和視窗-步幅模式比較

環境	固定視窗	視窗-步幅模式
準確度	77.81%	83.44%

1.2 數據分析

準確率差異實驗結果(Table 1)顯示，電腦在辨識準確度方面均優於 Raspberry Pi。差異主要為 Raspberry Pi 的外接麥克風在收音過程中容易受到環境雜訊的干擾，導致辨識準確度下降，電腦內建的聲音處理系統表現出色，能在雜訊較少的情況下更精確地進行音訊辨識。

傳統固定式窗口與視窗-步幅比較的實驗結果(Table 2)，傳統固定式窗口模式將音訊分割為相同大小的區塊進行分析，但由於缺乏靈活性，容易遺漏重要的音訊特徵，尤其是在邊緣的聲音。相較之下，視窗步幅模式透過在相鄰窗口之間設置重疊區域，使得模型能夠更精細地捕捉音訊中的細微變化，特別是在多樣化且隨機性強的環境聲音中，辨識準確度顯著提高。視窗步幅模式還能有效解決同一時間段內多種類型音訊重疊的問題，為音訊辨識提供了更豐富的特徵訊息同時有效減少特徵遺漏的現象發生。

1.3 裝置效能分析

在處理速度方面，電腦對音訊檔案的辨識時間約為 0.5 秒，而 Raspberry Pi 需要 1.8 秒來處理相同的音訊檔案，顯示出兩者在硬體性能上的顯著差異。尤其是在即時收音的情境中，Raspberry Pi 的延遲更為明顯，影響其在實時應用中的效能。考量到效能瓶頸的問題，因此把窗口設為 4，步幅設為 3 秒，讓樹梅派在辨識當中不會因為過量的數據堆積在佇列中而導致延遲產生。

1.4 雜訊與環境影響分析：

實驗結果顯示，環境噪音對辨識結果有影響，尤其是在 Raspberry Pi 中，收音品質受到背景噪聲（如風扇聲、背景談話聲等）的影響而大幅降低。為了改善這一問題，未來的系統開發可以考慮加入自適應濾波器或應用深度學習的降噪技術，或是更換陣列式的麥克風，提高收音效果，以減少背景噪音對辨識的干擾。

1.5 模型表現分析：

在錯誤分析時發現，某些聲音類別的辨識表現不佳，主要是由於訓練資料集中這些類別的樣本數量不足。這樣的情況無論在即時收音還是預錄音檔的實驗中均有發生。針對這一問題，可以通過增加資料集的多樣性和樣本量，或者利用數據增強技術，特別是針對那些少見或聲音複雜的樣本，以提升模型的辨識能力。

第五章 結論與未來方向

根據兩種實驗的結果，此音訊辨識系統在預錄音檔的辨識準確度上表現出色，不論是使用 Windows 還是 Raspberry Pi，都能達到較高的辨識水準。但是當系統需要即時收音進行辨識時，硬體設備的差異對結果有明顯影響，尤其是 Raspberry Pi 的外接麥克風在噪音較多的環境下會導致辨識準確度下降。若訓練資料集中某些聲音的樣本過少甚至缺失，無論是即時辨識還是預錄音辨識，都可能導致誤判情形發生。透過增加訓練數據集的多樣性和樣本量，以提高模型在各種聲音類別上的辨識準確度，尤其是對於少見或容易混淆的聲音環境，調整模型的參數設置或引入更多先進的演算法，也有助於增強系統的辨識精度。考慮在系統中加入自適應濾波器或深度學習的降噪技術，以減少環境噪音對辨識結果的干擾，特別是針對 Raspberry Pi 等嵌入式設備的應用，這樣的改善將有助於提高在嘈雜環境中的辨識效果與穩定性。

參考文獻

[1] J. Salamon, C. Jacoby and J. P. Bello, "UrbanSound8k Dataset", Urban Sound Datasets.
Available: https://urbansounddataset.weebly.com/urbansound8k.html

[2] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.
Available: https://arxiv.org/abs/2211.06687

[3] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "HTS-AT: A Hierarchical Token-Semantic Audio Transformer for Sound Classification and Detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.
Available: https://arxiv.org/abs/2202.00874

[4] K. J. Piczak. "ESC: Dataset for Environmental Sound Classification." Proceedings of the 23rd Annual ACM Conference on Multimedia, Brisbane, Australia, 2015.
Available: http://dx.doi.org/10.1145/2733373.2806390

[5] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: an open dataset of human-labeled sound events" *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829-852, 2022.
Available: https://arxiv.org/abs/2010.00475

[6] B. Elizalde, S. Deshmukh, and H. Wang, "Natural Language Supervision for General-Purpose Audio Representations," arXiv, 2023. [Online]. Available: https://arxiv.org/abs/2309.05767

[7] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "Clap: Learning audio concepts from natural language supervision," in *ICASSP 2023 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1-5
Available: https://arxiv.org/abs/2309.05767