國立台北大學資訊工程學系專題報告

基於雙層複合模型與物聯網針對聽力障礙族群之室內聲音視覺化系統

Dual-Layer Composite Model and IoT-Based Indoor Sound Visualization System for the Deaf or Hard of Hearing Community

專題組員: 陳怡茜、吳育瑄、黃泓愛

專題編號: PRJ-NTPUCSIE-112-006

執行期間: 2023 年 09 月 至 2024 年 06 月

1. 摘要

對於聽障人士而言，無法辨識環境中的重要聲音事件將會導致生活不便、安全受到威脅。本次專題開發了一個可在居家環境使用的聲音可視化系統，其透過樹莓派(Raspberry Pi)嵌入式裝置上的麥克風收取音訊,利用裝置內的小型人工智慧(artificial intelligence, AI)模型，即動態 MobileNet (Dynamic MobileNet, DyMN)對音訊進行初步辨識，並使用不確定性校準方法調適辨識結果的信心值，再根據辨識結果的信心值，決定傳送至伺服器的資料型態。伺服器視資料型態得使用二層模型 YAMNet 再次分類，並對分類結果進行後處理，再傳輸最終分類結果至使用者的攜帶式裝置，例如手機及智慧手錶。在此架構下，除了能夠確保重要聲音的辨識準確度，亦能區分真實環境中多樣複雜的聲音種類。

2. 簡介

現今市面上缺少用於聽障人士的室內空間輔助系統，導致聽障人士在居家環境的方便性與安全性方面的疑慮。近年來，儘管許多智能家電系統都增加了電器的聲音通知功能，但這類系統大多著重於家電用品相關之使用，對於聽障人士的幫助有限。經過調查，聽障人士的主要需求為視覺化安全需求相關的聲音類型，例如火災警報器、警報聲等。因此，我們的系統將著重於辨識此類聲音，以排除前述智能家電系統的聲音類別侷限性。

許多研究都使用了單一的人工智慧模型進行聲音辨識，然而單一模型在聲音的可辨識類別數量及辨識的準確性之間可能無法兼顧。因此，本研究想製作一個能夠即時辨識，且可以兼顧可辨識類別數量與分類準確性的雙層人工智慧模型聲音分類系統，並透過智慧型手機，以直觀的文字將分類結果通知使用者。

倘若能利用雙層模型架構，以小型的邊緣計算裝置即時偵測周遭環境聲音，並使用其內部的第一層重要類別精確模型，對特定重要類別進行精確分類，再視分類結果不確定性高低，將不確定者重新利用伺服器端的第二層多類別模型進行分類，即可達到同時精確分類重要類別，又能分類多種聲音的目的。另外，此架構也能考量聽障人士的需求來制定重要類別，提升整體系統的實用性。

3. 專題進行方式

我們所開發的針對聽障人士使用之室內聲音辨識系統，系統架構圖如

圖 1 所示。此系統包含多個加裝麥克風之 Raspberry Pi 裝置、主機、智慧型手機、智慧型手錶。麥克風接收環境聲音並傳送至 Raspberry Pi，並使用訓練過的聲音分類模型進行辨識。

第一個聲音分類模型(以下簡稱小模型)可辨識之聲音類別較少但準確度較高，用來辨識居家常見聲音類別或是需要高準確度的重要聲音類別，如火災警報器、警報聲等，並使用不確定性校準方法在 Raspberry Pi 上進行不確定性評估。若信心值較高，便將結果直接傳送給伺服器；若信心值較低，則將所錄製的音訊檔案傳送給伺服器，並透過伺服器上的第二個聲音分類模型(以下簡稱大模型)進行分類。該模型使用大型的音訊數據集進行訓練，涵蓋了近 521 個類別的環境聲音，其能夠將小模型沒有涵蓋到的聲音類別正確分類出來。且在兩者模型的配合下，我們能夠在確保重要聲音辨識準確度的同時，兼顧真實環境中聲音的複雜性。兩者模型的分類結果皆透過伺服器傳送至使用者端，同時，智慧型手錶與手機將會顯示通知，達到通知使用者的目的。

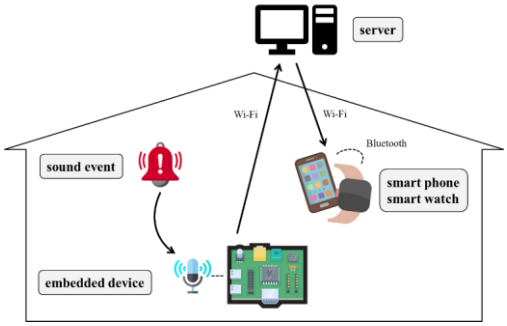


圖 1 系統架構圖

3.1 硬體架構

本系統的硬體裝置由一個樹莓派嵌入式平台及一個麥克風所組成。我們採用了 Raspberry Pi 5 作為樹莓派嵌入式平台，該型號為 Raspberry Pi 最新的版本，其 CPU 大小為 Raspberry Pi 4 Model B 的兩倍，且經過我們的實測，發現在 Raspberry Pi 5 上實現實時聲音偵測確實比在 Raspberry Pi 4 上實現速度更快且更為流暢，其 CPU 的使用率為在 Raspberry Pi 4 上實現的 0.6 左右，相當於節省了約 40%。因此我們選擇了 Raspberry Pi 5 作為系統的嵌入式平台。

在嵌入式平台上，我們放置了小模型，並搭配了 e-Kit MIC-G11 電容式麥克風，實現實時且大範圍的收音偵測。該麥克風使用了 Raspberry Pi 5 的 USB 3.0 進行傳輸，並且，在麥克風音量開到 100%時，其收音範圍長達 8 公尺。此外，該麥克風為全指向性(omnidirectional)，對於 360 度的聲響都有很高的靈敏度，能夠收錄空間中各個方向的聲音而不需特別調整角度，且錄製的聲音完整。因此我們選用它作為我們系統的麥克風。

3.2 軟體架構

3.2.1 深度學習和邊緣運算

本系統的軟體架構如圖 2 所示。軟體架構包含了訊號預處理、由 DyMN 與不確定性校準方法所結合而成的小模型、大模型 YAMNet，以及最後呈現給使用者的分類結果。

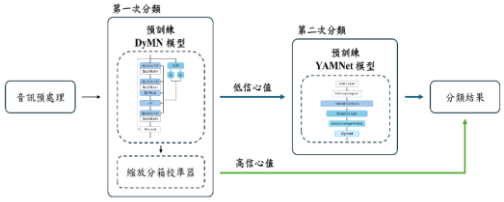


圖 2 軟體架構圖

我們在[1]所提出的 DyMN 模型中使用了與[2]中相同的 Transformer-to-

CNN 知識蒸餾方法。透過將 PaSST [3] 蒸餾至 DyMN，能夠使模型在性能優於一般 CNNs 模型的同時，也保有預測速度快、資源使用量較少，適合用於邊緣裝置的特點。其架構圖如圖 3 所示。

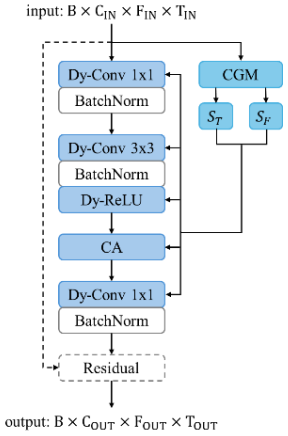


圖 3 DyMN 結構中的動態殘差反轉模塊架構圖[1]

為了提高機器學習模型的校準精度並增強其預測類別機率的效能，因此我們採用了不確定度校準器來測量 DyMN 模型(小模型)的校準誤差。我們使用[4]所提出的不確定度校準方法，該研究提出了一個用於準確估計校準誤差的縮放分箱校準器(scaling-binning calibrator)，此校準器結合了傳統的縮放方法(scaling)和具有可測量模型之校準誤差的直方圖分箱法(histogram binning)的優點，且無須採用大量的樣本即可取得低校準誤差。

我們首先將自定義資料集中的音訊輸入至訓練好的小模型，並擬和一選定函數，表示未校準機率和真實標籤之間的關係。再根據函數值將輸入空間進行分箱，然後對於每個箱，計算該箱內函數值的平均值，這些平均值將作為相應輸入範圍的重新校準機率。最後透過產生的重新校準機率設置縮放分箱校準器，讓輸入給模型的音檔在輸出時能根據校準器的校準精度改變輸出的不確定性評估。

在辨識聲音時，如果小模型對於所分類出的類別信心值經過校準後仍然過低，就會將所錄製的音訊透過網路傳遞給在伺服器上的第二個模型 YAMNet [5]，又稱為大模型，如圖 4 所示。該模型使用大規模的音訊資料集 Audioset 進行訓練，其可分類的聲音種類高達 521 種，涵蓋了幾乎日常生活中可能發生的聲音類別，讓有些重要但可能不常出現的聲音也能夠受到辨識。最後的分類結果將再次透過伺服器傳送到用戶端，達到通知目的。在大模型與小模型的配合下，我們能夠確保重要聲音辨識的準確度及辨識速度，同時兼顧真實環境中聲音的複雜性。

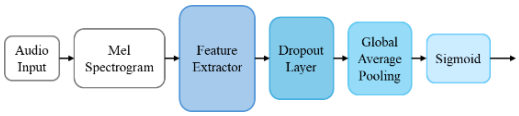


圖 4 YAMNet 架構圖[5]

3.2.2 聲音種類評估

在小模型部分，我們參考了研究[6]先前對聽障人士的聲音需求所做的調查[7]，他們統計了 201 名失聰或聽力障礙 (deaf or hard-of-hearing, DHH)參與者對行動和穿戴式裝置感興趣的聲音，發現緊急警報（例如警報、安全關鍵聲音）和針對聽障人士自己的聲音(例如有人喊叫到他們的名字時，系統可以偵測到並透過手錶告知)獲得了最高優先級。而 Mielke 等人[8]訪問了六位 DHH 人員，確定危險警報、電話鈴聲和警笛是工作場所環境中首選的聲音。與研究[9]的作法類似，我們採用了兩個音訊資料集，分別為 ESC50[10]、

ReaLISED [11]，並從中選用 13 種家庭環境中常見的聲音，包含了狗叫聲、敲門、說話、常見警報聲響等等，如表 1 所示。

表 1 資料集選用類別

ESC50			ReaLISED
dog	crying_baby	vaccum_cleaner	speech
cat	laughing	clock_alarm	walking
water_drops	door_wood_knock	glass_breaking	
pouring_water	washing_machine		

為使資料集中的音訊能夠具備相同的時長，且讓 13 種音訊類別在資料集中能夠平均分布，分別對從 ESC50、ReaLISED 資料集提取出的特定類別音訊進行處理：

(1) ESC50：自 ESC50 提取 11 種類別，每種類別各有 50 筆音訊，每筆音訊時長 5 秒。從原先的 50 筆音訊中，進行變調(pitch shift)來產生額外的 50 筆音訊並。

(2) ReaLISED：自 ReaLISED 提取 2 種類別，每種類別各有數量不等的音訊，將每種類別隨機抽取 100 筆音訊。另外，對於時長不足 5 秒的音訊，重複音訊使其時長達到 5 秒。

經過處理後的混合資料集，每個類別各有 100 筆音訊資料，每個音訊資料時長為 5 秒，且為單一格式(16kHz、32 位元、單聲道)。對於處理後的資料集音訊，參考小模型[1]的方法計算輸入特徵，並重新訓練模型以產生家庭環境當中常見聲音的分類器。

3.2.3 Raspberry Pi 與使用者之間的資料傳送

我們以 Flask 為框架實現 Raspberry Pi 與伺服器之間及伺服器與使用者之間的資料傳送。Raspberry Pi 會將實時收到的資料以.json 格式上傳至區域網路，而伺服器端則會將資料存入本機的資料庫內。若資料為分類結果則直接存於伺服器上；若為音檔則交由大模型進行分類，其結果也將存於伺服器上。

Raspberry Pi 並不會記錄或傳送事件結束的時間，因此我們默認事件持續直到出現新的事件或是靜音，伺服器所儲存的聲音結果皆為事件的起始，如圖 5 所示。

圖 5 Raspberry Pi 傳送資料至伺服器示意圖

最終將聲音事件的辨識結果、時間、地點及準確度透過網路傳送至使用者，其資料傳輸格式如表 2~4 所示。使用者端則會隨時向伺服器端檢查是否有新的聲音事件，若有則發通知於智慧型手錶及手機。

表 2 Raspberry Pi 傳送事件結果至伺服器

Event Name	Start Time	Accuracy	Volume	Raspberry Pi ID
Alarm	2014/04/01 17:20	0.95	40 dB	0001

表 3 Raspberry Pi 傳送音檔至伺服器

Audio	Start Time	Volume	Raspberry Pi ID
20240401172030.wav	2014/04/01 17:20	40 dB	0001

表 4 伺服器傳送至用戶端

Event Name	Start Time	Accuracy	Volume	Location
Alarm	2014/04/01 17:20	0.95	40 dB	living room

3.2.4 應用程式架構

我們在智慧手機及智慧手錶上呈現聲音事件辨識結果，在兩種攜帶式裝置上撰寫 APP，讓使用者可以查詢

當下的聲音辨識情形及歷史紀錄，智慧型手機 APP 的使用者操作如下，APP 操作介面如圖 6 所示：

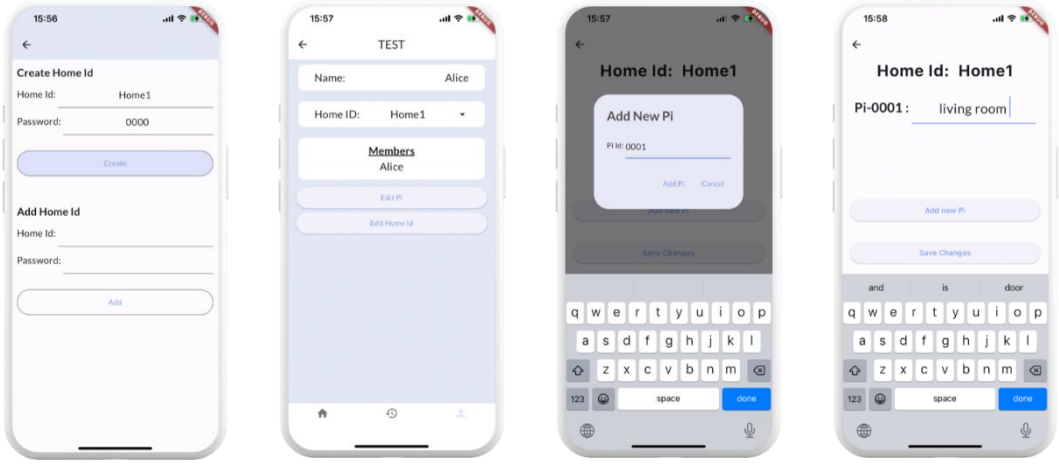
(1) 新增家庭：使用者可新增或加入多個家庭，並於使用者介面選取目前所在的家庭。

(2) 新增 Raspberry Pi 裝置：使用者可透過輸入 Raspberry Pi 的 ID 將 Raspberry Pi 裝置加入家庭。並自定義 Raspberry Pi 的名稱，如果未定義則默認為 Pi ID。

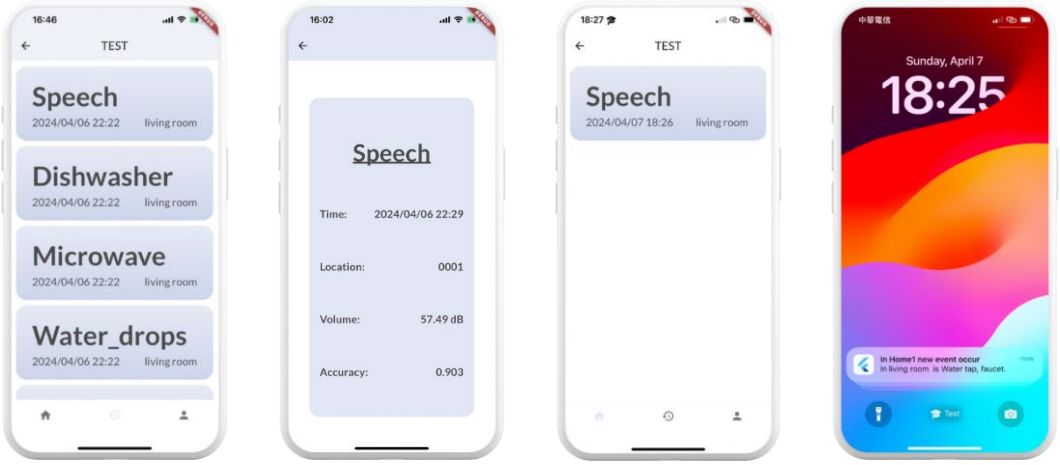
(3) 查看歷史事件：用戶可於歷史頁面查看大致歷史事件，上方為事件名稱，下方則為時間與自定義的 Raspberry Pi 名稱。

(4) 查看事件細節：用戶可透過點選事件查看事件細節，包含事件名稱、時間、地點（自定義 Raspberry Pi 名稱）、音量、準確度。

(5) 查看即時事件：用戶可於主頁面查看即時事件，家庭內有幾個 Raspberry Pi 則會有幾個事件按鈕位於主頁面，以顯示每個 Raspberry



(a) 新增/加入家庭頁面　(b) 選取家庭頁面　(c) 新增裝置頁面　(d) 裝置名稱設定頁面



(e) 歷史頁面　(f) 事件細節頁面　(g) 主頁面(即時事件)　(h) 新事件通知

圖 6 智慧手機應用程式的使用者操作介面

Pi 裝置目前的事件。在主頁面一樣可以透過點選事件查看事件細節。

(6) 通知：每當發生新的事件（不包括靜音）便會發送通知告知使用者，現在位於哪個位置有什麼事件發生。

此研究的預期結果為建構完善的聲音可視化系統，即時接收聲音並在攜帶式裝置上顯示聲音經由雙層分類模型分類後的結果，如圖 6(e)-(h)所示。模型架構可為 13 個重要聲音類別並預計提供 90%以上的準確度，且能有效辨識 521 種聲音類別。

4. 主要成果與評估

對於我們所提出的系統，我們做了一些實驗來量測軟體架構中的小模型的性能。在測試集上，我們的小模型取得了 87.9%的準確度。其中在 13 種類別及靜音類別的準確度如圖 7 所示。其中多數類別皆達到 80%以上的準確度。我們推測走路(walking)及水滴聲(water_drops)是因音訊資料之音量相對小，易受噪音影響，而致準確度較低。

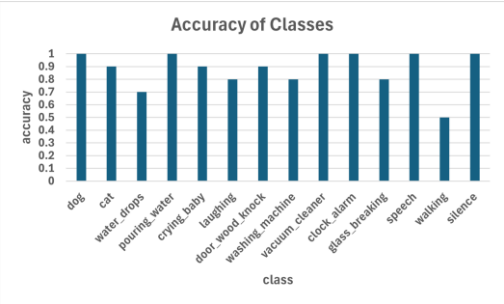


圖 7 小模型於測試集之各類別準確度

5. 結語與展望

對於本作品之總結說明我們所提出的聲音視覺化系統，能夠有效增加聽障人士在居家環境的方便性與安全性。此系統使用了方便設置的嵌入式裝置作為硬體裝置，除了成本低廉之外也擁有環境的可適應性，只要部署裝置於插頭附近即可使用，易於設置於各類居家環境。而體積較小的硬體裝置也能增加隱蔽性，進而減少使用者的心理負擔。

在聲音視覺化的部分，我們使用了智慧型手機作為系統與使用者之間的媒介，考量到智慧型手機的高普及性，無須另外購買其他穿戴裝置也可使用。此外，智慧型手機的便利性也讓使用者可以更加容易使用此系統，不限於居家中的特定位置。

對於智慧型手機上的應用程式，我們撰寫了能夠方便操作的 App，透過操作介面即可新增居家環境、新增 Raspberry Pi 裝置、查看歷史事件、查看事件細節、查看即時事件，為了讓使用者能夠及時了解聲音事件的發生，App 也附有通知功能，讓整體系統更加完善。

6. 銘謝

非常感謝指導教授提供大量資源與豐富的研究經驗分享，以及實驗室學長姐的協助，讓我們在進行專題的路上更加順利。最後也十分感謝組員們的合作，才能使我們的專題能如期完成。

7. 參考文獻

[1] F. Schmid, K. Koutini, and G. Widmer, “Dynamic Convolutional Neural Networks as Efficient Pre-trained Audio Models,” 2023, *arXiv:2310.15648*.

[2] F. Schmid, K. Koutini, and G. Widmer, “Efficient Large-Scale Audio Tagging Via Transformer-to-CNN Knowledge Distillation,” in *Proc*. *ICASSP 2023*, Rhodes Island, Greece, Jun.4-10, 2023, pp. 1–5.

[3] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, “Efficient Training of Audio Transformers with Patchout,” 2022, *arXiv:2110.05069*.

[4] A. Kumar, P. Liang, and T. Ma, “Verified Uncertainty Calibration,” 2019, *arXiv:1909.10155.*

[5] Google. “yamnet” kaggle.com. https://tfhub.dev/google/yamnet/1 (accessed Jan.30, 2024).

[6] D. Jain, K. Mack, A. Amrous, M. Wright, S. Goodman, L. Findlater, and J. E. Froehlich, “HomeSound: An Iterative Field Deployment of an In-Home Sound Awareness System for Deaf or Hard of Hearing Users,” in *Proc*. *CHI’20*, Honolulu, HI, USA, Apr.25-30, 2020, pp. 1 – 12.

[7] L. Findlater, B. Chinh, D. Jain, J. Froehlich, R. Kushalnagar, and A. C. Lin, “Deaf and Hard-of-hearing Individuals’ Preferences for Wearable and Mobile Sound Awareness Technologies,” in *Proc*. *CHI’19*, Glasgow, Scotland UK, May.4-9, 2019, pp. 1–13.

[8] M. Mielke and R. Brück, “A Pilot Study about the Smartwatch as Assistive Device for Deaf People,” in *Proc*. *ASSETS'15*, Lisbon, Portugal, Oct.26-28, 2015, pp.301–302.

[9] F. Rong, “Audio Classification Method Based on Machine Learning,” in *Proc*. *ICITBS*, Changsha, China, 2016, pp. 81–84.

[10] K. J. Piczak, “ESC: Dataset for Environmental Sound Classification,” in *Proc*. *MM’15*, Brisbane, Australia, Oct.26-30, 2015, pp. 1015–1018.

[11] I. Mohino-Herranz, J. García-Gómez, M. Aguilar-Ortega, M. Utrilla-Manso, R. Gil-Pita, and M. Rosa-Zurera. “Real-Life Indoor Sound Event Dataset (ReaLISED) for Sound Event Classification (SEC)” Zenodo. https://zenodo.org/records/6488321 (accessed Jan.30, 2024).