**中 國 文 化 大 學**

**資 訊 工 程 學 系**

**資 訊 系 統 專 題**

**情緒聲音辨識處理系統**

**學 生：方建智**

**康宇翔**

**曾敬峰**

**指導教授：劉仲鑫**

**中 華 民 國 114年5月**

**情緒聲音辨識處理系統**

**專題學生：方建智、康宇翔、曾敬峰**

**指導教授：劉仲鑫 博士**

**中國文化大學 資訊工程學系**

# 摘要

本研究旨在探討人的情緒聲音辨識處理系統，設計並實現了一個基於自然語言處理和機器學習的情緒聲音辨識系統，該系統能夠高效地將人員的語音輸入後經過模型辨識情緒類別，我們的系統集成了先進的語音辨識模型，並使用了多種微調技術，如LoRA和Prefixing，以優化其性能。系統功能包括以MFCC萃取聲音之特徵再進行模型訓練，最終可以達到99%以上之準確率辨識人的情緒類別。

**關鍵詞：**聲音辨識、語音轉錄、自然語言處理

指導教授\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_(簽名)

**Emotion voice recognition processing system**

**Student: Yu-Xaing Ｋang, Jian-Zhi Fang, Jing-Fong Zeng**

**Advisor: prof.Chung-Hsin Liu**

**Department of Computer Science and Information Engineering**

**Chinese Culture University**

# Abstract

This study aims to explore the emotional voice recognition processing system. A system based on natural language processing and machine learning for emotional voice recognition is designed and implemented. This system efficiently recognizes emotional categories from voice input using a model. Our system integrates advanced speech recognition models and employs various fine-tuning techniques, such as LoRA and Prefixing, to optimize its performance. The system's functionality includes extracting sound features using MFCC and then training the model, ultimately achieving over 99% accuracy in recognition of emotional.

**Keywords:** Voice Recognition, Speech Transcription, Natural Language Processing

目 錄

[摘要 2](#_Toc168064953)

[Abstract 3](#_Toc168064954)

[表 目 錄 5](#_Toc168064955)

[第 1 章 研究動機與目的 7](#_Toc168064956)

[1.1 研究動機 7](#_Toc168064957)

[1.2 研究目的 7](#_Toc168064958)

[第 2 章 文獻探討 9](#_Toc168064959)

[2.1 人工智慧在聲音辨識之應用 9](#_Toc168064960)

[2.2 聲音特徵提取 12](#_Toc168064961)

[2.3 深度學習(Deep Learning) 17](#_Toc168064962)

[第 3 章 研究內容、方法與工作項目 24](#_Toc168064963)

[3.1 研究方法設計 24](#_Toc168064964)

[3.2 使用工具 25](#_Toc168064965)

[第 4 章 預期成果 26](#_Toc168064966)

[第5章 人力配置 31](#_Toc168064967)

[第6章 未來展望：醫療聲音辨識處理系統 32](#_Toc168064968)

[參考文獻 35](#_Toc168064969)

# 表 目 錄

# 圖 目 錄

# 研究動機與目的

## 研究動機

情緒聲音辨識技術在客服領域的研究動機和目的主要包括以下幾方面：首先，通過快速分析客戶的語音，提高服務效率和準確性，特別是在面對大量使用時。其次，可以減少人力資源負擔。再次，提升使用者體驗，通過便捷的語音操作減少等待時間和不便。情緒聲音辨識技術還能增強情緒狀況監測，對於需要情緒支持的個體提供持續的情緒變化追蹤。特別是在特殊狀況時，無接觸的情緒篩查有助於減少直接接觸的風險。研究目的包括開發高效的情緒識別工具、提升遠程服務技術、改善使用者與服務提供者之間的互動、推動數據分析與管理、促進個性化服務以及支持心理健康和社會研究。這些目標旨在利用先進技術提升各領域服務的效率、質量和覆蓋範圍，最終改善人們的福祉與生活品質。

## 研究目的

情緒聲音辨識技術在客服或精神診斷領域的研究目的包括以下幾個方面：首先開發高效的診斷工具，通過快速分析和識別情緒的特徵聲音，提供輔助判斷，協助客服人員或是診斷精神疾病時，預先判定對方情緒時，在與對方溝通時可以不受遠程地點限制，或是減少面對面接觸，能夠輔助人員即時做出對應措施。此外，改善人員溝通時的互動，創建智能化的語音助手，理解和回應對方需求，提供更佳自然的互動體驗。推動情緒數據分析和管理，利用語音辨識技術自動轉錄和分析，支持大數據分析客戶即時情緒，促進個性化服務，通過語音數據的個性化分析，提供針對每個客戶的客制化服務方案。

# 文獻探討

## 人工智慧在聲音辨識之應用

聲音辨識在現代社會應用廣泛，例如常見對著智慧音箱講「Hi Siri」即可以語音互動，或是像是智能客服等，在我們生活上該技術滲透率很高，過去較為常見且成熟的方法為隱含式馬可夫模型或是高斯混合模型，當今由於機器學習或深度學習很廣泛，並且因為開源，取得容易，適合各種規模的機構運用。以下簡介國內人工智慧在聲音辨識之應用研究：

1. 運用類神經網路做聲音之辨識

將時域之聲音檔案，這些檔案主要是生活中各種聲音，透過特徵萃取，取得其中特徵後並轉化成向量，再將這些參數運用類神經網路做訓練，計算屬於不同類別之機率。與傳統模型GMM相比，使用神經網路模型錯誤率較低，與傳統方式比較其錯誤率少0.05。

一張含有 文字, 螢幕擷取畫面, 字型, 圖表 的圖片

自動產生的描述

**圖 1研究流程示意圖**

1. 運用道路事故的聲音資料

將其中碰撞聲以及剎車聲取出，每個音檔以0.5秒做分割，以短時距傅立葉轉換產出頻譜圖如圖2. 2，在套用YOLOv7的模型，進行類別之訓練，進一步實現透過聲音訊號來辨別道路事件的模型。

一張含有 螢幕擷取畫面, Rectangle 的圖片

自動產生的描述

**圖 2研究流程示意圖**

1. 利用收集的聲音資料

藉由訊號前處理轉換後找到相對應的特徵並加以分析，傳統之聲音訊號處理方式會用一些特徵如頻率、振幅來進行分類，但若是特徵無法找到有效的差異，使得分類準確率無法有效提升。因此該研究使用人工智慧中的卷積神經網路模型來建立聲音的分類模型。先運用梅爾倒頻譜係數原理將聲音訊號轉換成頻譜圖再匯入訓練模型進行訓練。其中卷積神經網路克服了傳統聲音識別只藉由單一特徵進行分類導致不準確問題。梅爾倒頻譜係數將數位的聲音訊號經由頻率之間的差異進行特徵提取。由於資料有些不平衡，因此藉由資料擴增的方式來改善分類模型之準確率，由實驗結果顯示能夠提升分類的準確性30%。

一張含有 文字, 螢幕擷取畫面, 字型, 行 的圖片

自動產生的描述

**圖 3研究流程示意圖**

## 聲音特徵提取

聲音可視為是一個連續的函數，若要進行訊號分析可以用以下方式進行特徵萃取：

1. 傅立葉轉換：

將聲音訊號以頻率域的方式呈現。轉換後聲音的本質特性還在，會將聲音訊號以「頻率」的方式來表達。傅立葉轉換本身為一種數學上的線性積分變換方式，藉由傅立葉轉換將週期函數轉換為另一種函數。傅立葉轉換根據輸入的訊號類型可以分成如表2.1：

|  |  |
| --- | --- |
| 方法轉換方法 | 圖形 |
| 非週期性連續訊號傅立葉變換 | 一張含有 設計 的圖片  自動產生的描述 |
| 週期性連續訊號傅立葉級數 | 一張含有 廚房用具, 餅乾成型切割刀 的圖片  自動產生的描述 |
| 非週期性離散訊號離散時域傅立葉變換 | 一張含有 扳手, 工具 的圖片  自動產生的描述 |
| 週 期 性 離散訊號離散傅立葉變換 |  |

**表2. 1 傅立葉形式的圖**

在分析聲音時，會「短時距分析」（Short-term Analysis）

為主，在短時間內是相對穩定的。

詳細說明如下：

1. 音框

將聲音訊號切成數個音框，其中音框長度約為20 ms ~ 30ms之間，若音框太大，比較無法找到其中聲音特性隨時間的變化。反之，若太小會無法找出聲音本身的特徵，因此音框的用途主要是能夠切割成數段有週期性的聲音，各音框相鄰之間，可以保有重疊，藉此保有資料的連續性，一般來說重疊的部分約為音框長度的 1/2 到 2/3 不等，實務上亦可依照需求來做調整。

1. 特徵提取

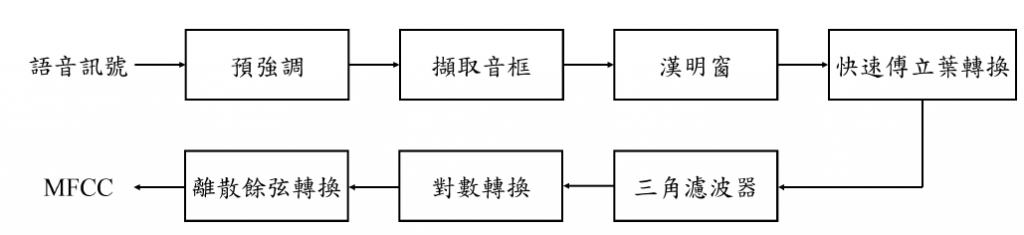
將訊號音框化後，針對各個音框進行特徵提取，除了常見藉由聲音基本資訊，例如音量、音高或音質的方法，也有如過零率(zero-crossing rate, ZCR)、MFCCs 進行特徵提取參數…等方式做為特徵提取的方式。

1. 分析

根據音框的特徵提取後，針對各音框的特徵資訊，進行分析與辨識。

1. 梅爾倒頻譜係數：

MFCCs 在1980 年由 Davis 和 Mermelstein 共同所提出的，從那之後 MFCCs在語音識別領域被廣泛的利用，包括了語音識別，聲音識別，環境聲音識別等[4]，其流程參照下圖：



**圖 4 MFCC流程示意圖**

1. 預強調(Pre-Emphasis)

聲當音收集完成之後，訊號處理時頻率較高的部分會產生能量衰減的情形，因此在進行辨識或比對前，會將聲音訊號先經過一個高通濾波器，目的是為了提高高頻的訊號，避免後續進行轉換時發生問題，如公式(2)：

𝑦(𝑡) = 𝑥(𝑡) − 𝛼𝑥(𝑡 − 1) , 0.9 ≤ 𝛼 ≤ 1.0 (2)

其中𝑥(𝑡)為原先的波型訊號α𝑥(𝑡−1)為經過高通濾波器訊號，通常α介於0.9到1之間，此方步驟主要是將聲音訊號的高頻進行放大，放大後的聲音訊號變得較尖銳清脆，但音量會降低。

1. 音框化(Frame Blocking)

連續的完整聲音訊號資料因為時間過長導致快速變化，因此在分析訊號時，通常會將聲音訊號以 N 個取樣點為切割數量，將連續訊號切割分塊，我們將這些區塊稱為音框(Frame)，標準音框大小為 25ms，舉例說明，取樣頻率 44.1kHz 的聲音資料，因此取樣點為 1103 個，而在各音框間有一段重疊區域，目的在於增加各音框間的連續性，為 1/2 到 2/3 不等。

1. 漢明窗(Hamming Window)：

將分割的音框乘上一個漢明窗，目的在於增加各音框之間的連續性，避免後續傅立葉轉換時產生訊號不連續而產生雜訊，

一張含有 行, 繪圖, 圖表, 斜率、斜坡 的圖片

自動產生的描述

**圖2.5 漢明窗**

1. 特徵提取參數(Feature Extraction)

由於在傳統的傅立葉轉換時，我們必須將連續的聲音資料進行積分，因此處理上非常繁瑣，而本文使用的是快速傅立葉轉換(Fast Fourier Transform, FFT)，頻域相較於時域能夠依據能量的分佈來辨識不同聲音的特性，由於各個音框為短時間訊號，因此處理上較快速且簡單，提高轉換的效率，接著將轉換完的頻譜能量乘上一組三角濾波器，並求得每一個濾波器所輸出的對數能量(Log Energy)，其中這三角帶通濾波器，必須依據「梅爾頻率」（Mel Frequency），就有如人的聽力系統，模擬人耳在對於不同的頻率資料時有著不同的敏感度，將各頻率間的資料進行過濾，只提取所需的特徵，各組濾波器並沒有平均分布的在符合梅爾頻率的三角濾波器的數量 M 通常設定為 20~40，而濾波器之間的間隔也隨著頻率的提高而減少變成稀疏，如圖2.6 所示。

一張含有 文字, 螢幕擷取畫面, 行, Rectangle 的圖片

自動產生的描述

**圖2.6 三角濾波器**

其中橫軸為頻率，縱軸為頻率響應，在各個濾波器中，頻率響應都為 1，梅爾頻率是依據人耳對於頻率的感受程度而計算的。依據人耳的特性，當聲音頻率越低時，三角濾波器越密集，則能保留人耳可分辨的聲音資料；反之當聲音頻率越高時，三角濾波器越稀疏。在語音識別系統常使用 MFCCs 做為特徵提取的方法，會去考量到音調的變化，而在本研究並不會受到輸入語音的音調不同而有所影響，其最主要目的，是根據不同的三角濾波器的個數所產生的頻譜圖之間的差異，做為聲音分類的特徵，並探討不同三角濾波器數量之間的差異。

1. 離散餘弦轉換(Discrete cosine transform, DCT)

經常用於訊號處理和圖像處理，對訊號和圖像進行壓縮，實際就是對每幀數據在進行一次降維，由三角濾波器的梅爾頻譜加權濾波器相乘而獲得，而m是第m個三角形濾波器係數，最後得到我們整個流程所要提取的梅爾倒頻譜係數。公式如下:

C[n] 是第n個MFCC係數

S'[m] 是對數梅爾頻譜

M 是濾波器組的數量

m代表梅爾濾波器的索引

n 是MFCC係數的索引

α(n) 是歸一化因子

綜合以上簡言之，梅爾頻率倒譜係數是一種模擬人耳對不同頻率敏感度的頻率尺度。在這種尺度下，低頻部分的變化較為敏感，而高頻部分的變化較為不敏感。因而適合用於語音辨識。語音訊號經過預強調 (Pre-emphasis)，目的是加強語音訊號中的高頻成分，因為語音信號的高頻部分通常較弱。這樣可以平衡語音中不同頻率範圍的能量，使後續的處理過程能夠更好地捕捉到語音的特徵，以此來突顯高頻的部份，再將訊號轉變為音框(Frame)，並對每一個音框乘上一漢明窗(Hamming Window) 來增加音框的連續性，接下來經由快速傅立葉轉換(Fast Fourier Transform,FFT) 將訊號從時域轉換到頻域上，再將得到的能量頻譜乘上M個三角帶通濾波器(Triangular Bandpass Filters)，獲得每一個濾波器輸出的對數能量(Log energy)，將上述的M個對數能量透過離散餘弦轉換(Discrete Cosine Transform, DCT) 最終算得梅爾頻率倒譜係數。

綜合以上簡言之，梅爾頻率倒譜係數考慮人耳對於不同頻率會有不同敏感度的特性，因而適合用於語音辨識。語音訊號經過預強調 (Pre-emphasis) 來突顯高頻的部份，再將訊號轉變為音框(Frame)，並對每一個音框乘上一漢明窗(Hamming Window) 來增加音框的連續性，接下來經由快速傅立葉轉換(Fast Fourier Transform,FFT) 將訊號從時域轉換到頻域上，再將得到的能量頻譜乘上M個三角帶通濾波器(Triangular Bandpass Filters)，獲得每一個濾波器輸出的對數能量(Log energy)，將上述的M個對數能量透過離散餘弦轉換(Discrete Cosine Transform, DCT) 最終算得梅爾頻率倒譜係數。

另外一種常見的聲音特徵萃取手法為非負矩陣分解(Non-negative Matrix Factorization)(NMF)，該方法可以達到數據降維和特徵學習之效果，將一個非負矩陣分解為兩個非負矩陣的乘積。這兩個矩陣的乘積能夠近似原始矩陣，並且這些分解後的矩陣通常有助於發現數據中的潛在結構和模式。其數學如下：

其中W是一個 m×r 的矩陣，代表「基底」或特徵矩陣，H 是一個 r×n 的矩陣，代表在基底上加權的係數。其主要特色如下：

1. 非負約束：

NMF要求所有矩陣元素都是非負的，即大於或等於零。這使得NMF能夠揭示數據中的積極關係或結構，並且在某些應用中（如圖像處理、文本分析等）這一點尤其有用，因為數據本身通常是非負的。

1. 解釋性強：

由於NMF將原始數據矩陣分解成兩個低維矩陣，這兩個矩陣的元素通常可以解釋為某些隱含的因素或特徵。例如在文本分析中，這些特徵會對應於潛在的主題。

1. 降維：

通過選擇較小的維度，NMF實現了數據的降維。這有助於去除噪音並突出數據中的核心結構。

NMF之應用如下：

1. 文本挖掘與主題建模：

在文本分析中，NMF可以用來從文檔-詞語矩陣中提取主題。每個主題可以由一些詞語的加權組成，這些詞語通常反映了該主題的關鍵內容。

1. 聲音事件檢測：

NMF還可以應用於聲音事件的檢測和分類。例如，對於環境噪聲監測系統或智能監控系統，NMF可以用來從音頻中提取出不同的聲音事件，如門鈴聲、車輛聲或人聲，並將它們與背景噪音分離開來，從而更準確地檢測和識別特定的聲音事件。

1. 音頻特徵萃取：

NMF也用於音頻分類任務中，例如，對音樂片段、語音錄音或環境聲音進行分類。通過將音頻信號分解為基本成分，NMF可以提取有用的特徵，這些特徵可以用來進行後續的分類或聚類。例如基於音頻特徵的情感識別、語者識別等。

1. 語音分離和增強：

在語音處理中，NMF被用來處理具有背景噪聲的語音信號。比如，當語音信號與背景噪音（如街道噪聲或其他對話）混合時，NMF可以用來提取和增強語音部分，並消除或減少噪音。

## 機器學習(Machine Learning)

機器學習範疇很多主要分監督式學習與無監督式學習，監督式學習就包含線性回歸、邏輯回歸、支持向量機(SVM)、決策樹、隨機森林，主要特點是需要有標註來進行訓練，無監督式學習有K-均值聚類、層次聚類、主成分分析，不需要標註即可以訓練模型，本研究參考李氏[10]使用SVM進行聲音的分類。支持向量機於 1992 年由 Boser、Guyon及Vapnik等人所提出，主要的概念就是希望能在一群由 2 種不同類別混合而成的資料群集中，找到一個最佳的超平面(hyperplane)將這些資料區分開來，以 2 維資料為例，有兩群不同的資料(o 與×)分佈，可用1條直線將其分成2類如圖2.12，若在3維甚至多維空間中這條直線則可想像成平面如圖2.13，故稱作超平面。此外，希望這條區分的超平面與兩群資料的距離(又稱邊距，margin)越遠越好，如此在進行分類時也就 越 容 易 判 別 ， 這 樣 的 超 平 面 則 稱 作 最 佳 超 平 面 (optimal hyperplane)，而與最佳超平面平行並且最靠近兩邊資料之平面稱為支持平面(support hyperplane)，在支持平面上的資料即稱為支持向量(support vector)。

一張含有 鮮豔, 圓形, 水, 設計 的圖片

自動產生的描述

**圖2.12二維平面最佳超平面示意圖[4]**

一張含有 兒童藝術, 圖表, 螢幕擷取畫面, 設計 的圖片

自動產生的描述

**圖2.13三維空間最佳超平面示意圖[4]**

在解題時，可以假設

X1=[2,3] y1=1

X2=[1,1] y2=1

X3=[-1,-2] y3=-1

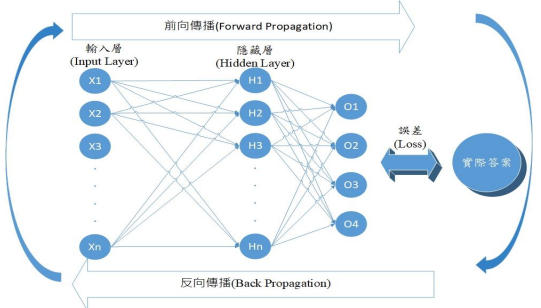
假設平面為w1X+w2Y+b=0，可以得出下列算式

SVM的常見方法是通過對偶問題來求解。對偶問題可以利用拉格朗日乘子法來解，將原始的最小化問題轉換為對偶的最大化問題。具體過程包括求出Lagrange乘子，並進行優化，這是通過求解二次規劃來完成的。來求最佳化問題，最終解決下列算式即可以得到w1及w2 。

## 深度學習(Deep Learning)

深度學習為本研究會使用得核心技術，這項技術在2006年，因為深度學習之父 Geoffrey Hinton 提出深度信念網路(Deep Belief Networks, DBN)，使用非監督的提前訓練方法來優化網路權重的初始值，接著進行權值的些微調整(Fine-Tune)後讓多層神經網路能夠真正的被實現，而當時的深度神經網路則被稱之為深度學習。在 2012 年Hinton 學生使用圖形處理器(Graphics Processing Unit，GPU)運算深度學習模型，使其運算效率高於原本中央處理器(Central Processing Unit，CPU)運算的 70 倍以上。此後深度學習技術應用於各大領域，包含金融股市預測、醫療保健用品、氣象預測、客戶分析、醫療影像辨識和智慧交通等產生十分深遠且巨大的影響。其最具代表性的深度學習為AlphaGo，AlphaGo 在圍棋領域擊敗各路的高手。現在深度學習架構應用的領域十分廣泛普及，例如音訊識別、語音識別、影像識別、自然語音處理、生物資訊學與生活眾多相關領域。

深度學習是模擬人類神經網路的運作方式。由於人腦結構非常複雜，預估具有 860 億個神經元及超過 100 兆條的神經相連，形成的網路比最先進的超級電腦還要強大。但為方便以電腦模擬，將神經元分為多層次，來模擬神經網路模型。在人工神經網路(Artificial Neural Network, ANN)實現中，每層裡頭各個神經元都有所謂的節點，而節點都相互的連結，神經元和節點通常具有隨著學習進行而調整的權重，權重增加或減小做為連接處訊號的強度。神經元可以具有閾值，使得僅當總訊號超過該閾值時才發送訊號。通常，神經元聚集成各層，不同的層可以對它們的輸入執行不同的變換。神經網路模型通常會有 1 個輸入層、1 個輸出層、隱藏層(可多層)，以下我們將逐一介紹各層，如圖 2.7所示。



**圖2.7 神經網路收斂流程**

1. 輸入層(Input Layer)

由多個神經元所組成，主要功能是接收輸入資訊，圖2.7中以 X1、X2…Xn為輸入層各輸入節點，一般為數值型的資料，但依據不同的神經網路模型，所輸入的型態也就不一樣，如數值型的輸入資料、以圖片做為輸入資料。

1. 隱藏層(Hidden Layer)

介於輸入層和輸出層之間的神經網路層，主要功能是透過非線性轉換函數運算，由神經元(Neuron)組成，透過前向傳播(Forward propagation) 計算出我們的輸出值，最終將值傳至輸出層，如圖 5 中以 H1、H2…Hn表示為隱藏層中各隱藏節點。通常隱藏層的層數不容易決定，須由多次嘗試運算找到最佳解。使用過多的層數會導致模型收斂過慢或無法收斂；而較少的層數則無法應付較困難的問題。最更新我們隱藏層的參數，經過運算所得到的預測值，我們藉由縮小預測值與實際標記值(Label)，更新我們隱藏層間的參數，最終可訓練出一組權重(weights)值，提供給輸出層進行評估。

1. 輸出層(Output Layer)

主要分析分類結果以及評估權重的資訊，如圖 5 中以 O1、O2、O3表示為輸出層中各輸出節點，節點數量也代表著分類的數量。

綜合上述各層介紹，神經網路透過前向傳播(Forward Propagation)透過神經元與權重的交叉運算，預測出解答與實際值比對，當預測與實際值相異時稱為相異值為損失(Loss)，依據 Loss 值進行反向傳播求導數，並進行更新神經網路的權重，反覆進行直到模型收斂。

在聲音分類中有許多可行的方法，如：基於機器學習中支持向量機(Support Vector Machine, SVM)進行呼吸聲音的分類，以決策樹進行聲音分類，皆為機器學習的方法。但是在對於環境聲音中如果類別單純時能夠有著準確的分類結果，但當分類種類呈倍數增加時，將大大影響了支持向量機與決策樹的分類效果，因此我們將考慮使用深度學習的方法，利用模型進行學習訓練，藉由學習特徵分類出生因資料。

常見的深度學習架構包含：多層感知器(Multilayer Perceptron, MLP)、人工神經網路(Artificial Neural Network, ANN)、深度神經網路(Deep Neural Network, DNN)、卷積神經網路 CNN(Convolutional Neural Network)、遞迴神經網路 RNN(Recurrent Neural Network)…等。

在深度學習裡，如基於傳統 ANN 的聲音分類方法[24]，基於 RNN的長短期記憶網路(Long short-term memory, LSTM)的城市聲音分類，對於聲音分類，使用各種分類模型進行訓練的技術已相當的純熟以集多元。而本論文選擇 CNN 做為分類模型的原因是由於影像識別模型的特性，已經證實了有很好的分類效果。我們能夠將轉換後的聲音頻譜圖以完整時間區間進行保留，並匯入到 CNN 模型裡進行模型訓練，在 CNN 模型中輸入為一個二維矩陣，圖型資料就是一個典型二維矩陣。因此能夠符合其模型的輸入，並將輸出的結果進行準確性的驗證，達到分類預測的能力。比起傳統利用聲音特性所建立的環境聲音分類預測模型，此方法大大提升了準確率。卷積神經網路:

在各種深度學習都有屬於自己的特性，CNN 為 DNN 之衍生，比起 DNN，CNN 多出了卷積層(Convolutional Layer)與池化層(Pooling Layer)兩層，這兩層與一個全連結層組合，此全連結層即為 DNN 的輸出層，最後並使用 Softmax 函數來輸出分類結果，如圖2.8 所示。

一張含有 圖表, 文字, 行, 方案 的圖片

自動產生的描述

**圖2.8 CNN示意圖**

由於卷積神經網路的特性擅長於圖片的識別，透過上述的卷積層 與池化層可以將高維度的輸入劃分為多個低維度的特徵。根據此方式 學習特徵後，可做為辨識用的分類器。而若要處裡非圖像辨識有關的 資料時，則需將輸入資料轉換成一個二維矩陣，透過此方式讓卷積神 經網路可以廣泛的應用在許多領域，以藉此來分析環境聲音的分類。 CNN 各層運作的說明將於下列各小節中進行說明。

卷積層：在卷積運算(Convolution)就是將原始圖片的與特定的卷積核做卷積運算(以符號⊗表示)，如圖 7 的數值為例，其中中間的卷積核(3\*3)會隨機產生，並根據輸入圖像的大小確定評估值，如同掃描的方式，如圖2.9所示。將卷積核從圖片的最左上方的像素逐一掃描，我們可以發現使用此方式掃描矩陣，而在輸入圖片中卷積核無法完整的將聲音資料進行掃描，導致掃描出來的圖片找不到特徵，因此將每個像素皆為卷積核的正中心，當卷積核掃描到圖片的邊緣像素時，我們會以補零的方式進行，目的在於讓圖片保持原本輸入的大小，防止卷積後圖片越卷越小的情形

一張含有 圖表, 數字 的圖片

自動產生的描述

**圖2.9**

池化層：在完成卷積層的特徵提取之後，將進入到 Pooling(池化層)，常見池化層為平均池化層與最大池化層，其中又以最大池化層最常見。其原理是透過特徵圖提取其最大值，一般取決於池化的大小，並設定它的步長，最大池化層主要用於提取重要特徵、加速收斂並縮小特徵圖尺寸。

全連接層：提取完特徵後，將特徵帶入到全連接層。基本上全連接層的部分就是將之前的結果平坦化之後接到最基本的分類，如圖2.10所示。

一張含有 文字, 行, 圖表, 螢幕擷取畫面 的圖片

自動產生的描述

**圖2.10**

# 研究內容、方法與工作項目

## 研究方法設計

本研究參考文獻中的做法，採納MFCC做數據的預處理，將聲音訊號換算成向量後，再分割成訓練及測試集資料，再帶入不同模型，以準確率來評估模型優劣。使用之數據來源為https://github.com/marcogdepinto/emotion-classification-from-audio-files/tree/master/TESS\_Toronto\_emotional\_speech\_set\_data，該作者提供很多公開且珍貴之數據，並且其數據有一定之根據性及可靠度。

preprocessb.com/marcogdepinto/emotion-classification-from-audio-files/tree/master/TESS\_Toronto\_emotional\_speech\_set\_data

MFCC

語音輸入

分為train及splitom/marcogdepinto/emotion-classification-from-audio-files/tree/master/TESS\_Toronto\_emotional\_speech\_set\_data

CNN模型建立m/marcogdepinto/emotion-classification-from-audio-files/tree/master/TESS\_Toronto\_emotional\_speech\_set\_data

SVM模型建立m/marcogdepinto/emotion-classification-from-audio-files/tree/master/TESS\_Toronto\_emotional\_speech\_set\_data

辨識結果討論m/marcogdepinto/emotion-classification-from-audio-files/tree/master/TESS\_Toronto\_emotional\_speech\_set\_data

**圖 5研究流程示意圖**

在訓練過程中會使用測試資料測試模型是否過擬合狀況發生，若有則會用dropout方式關閉部份節點，因為數據有咳嗽聲以及呼吸聲，預計會以其中一個數據集建立模型後，再以這份模型去套用在另一份數據集，試著不同參數例如學習率測試是否可以直接套用或是可以用較少的步驟訓練，達到遷移學習的效果。

## 使用工具

本研究使用google提供的colab之python編譯器，做深度學習演練。

系統主要分三個架構：

1. 數據前處理

將聲音訊號變成數字資訊。

1. 模型建置模組

利用不同深度學習架構

1. 結果輸出模組

將程式寫成一份exe檔案，可以批次預測並輸出成一份html檔案。

**表 3.1方法技術工具表**

|  |  |
| --- | --- |
| 套件 | 用途 |
| Keras | 實現深度學習 |
| Librosa | 處理聲音訊號 |
| sqlite | 紀錄結果 |

# 預期成果

本研究規劃案，預期可以得到下列之成果：

1. 與傳統之聲音辨識比較

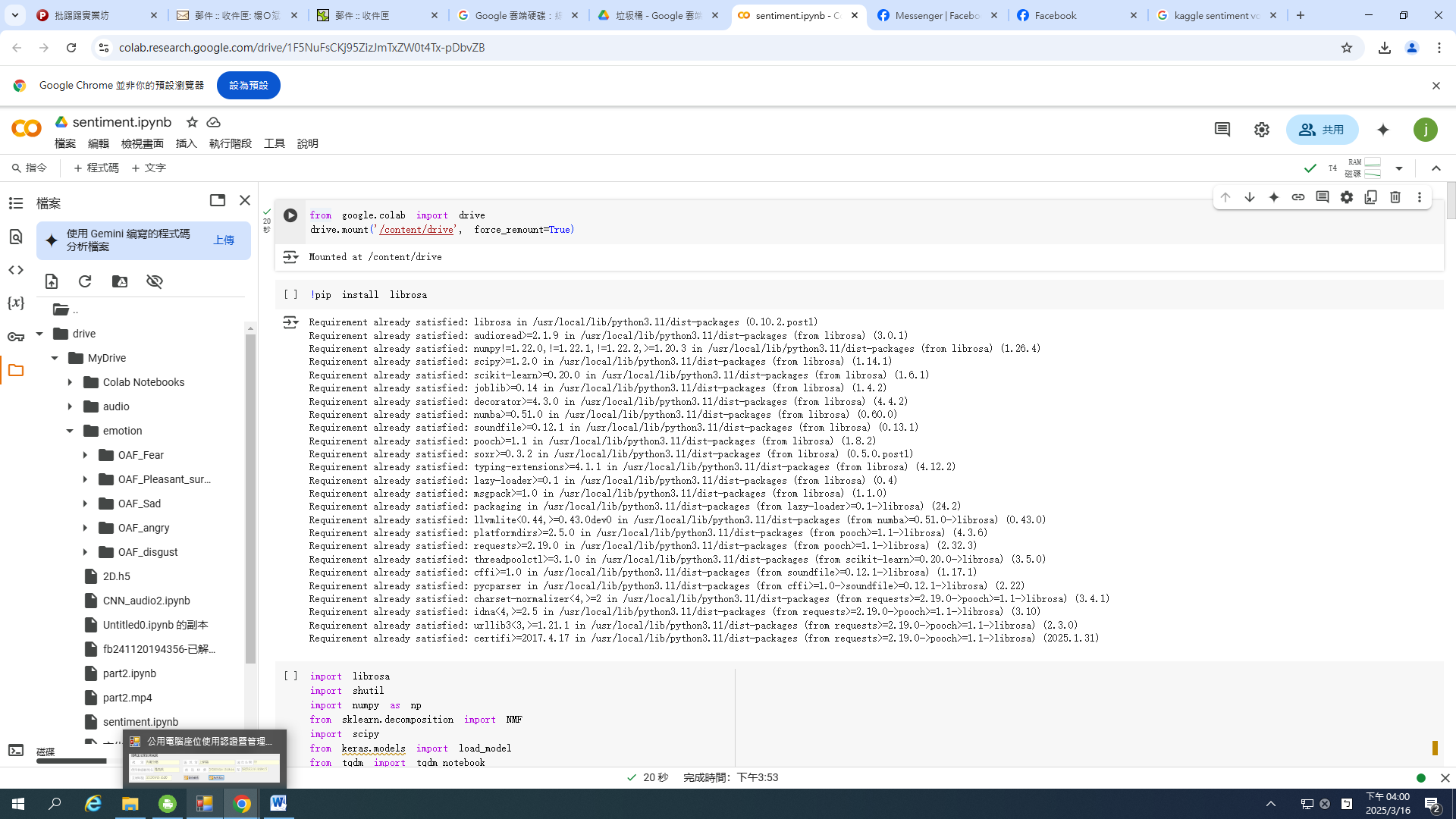
本研究採用MFCC萃取出該聲音之特徵，並且將這些特徵使用機器學習或是深度學習來訓練，期望準確率能達到80%以上或更高，希望未來該程式可以真正應用在客服或領域。

1. 探究是否可應用遷移式學習

本數據有五種情緒聲音目錄如下圖，探究是否可能訓練完這五種聲音後，將此模型權重應用其他情緒聲音，再與單獨訓練這些情緒聲音之模型做比較，若效果相似，表示可以用更短的時間進行其他情緒聲的模型建立。

1. 分析各種情緒聲音是否有特徵

原始檔之聲音類別如下圖，若用人去聽來分類，非常曠日廢時，並且需要很專注反覆地聽取才能辨析其中差異，本研究希望將聲音訊號數位化，嘗試尋找是否在某些頻率上或繪製成頻譜圖後可以找到一些特徵，讓我們了解是什麼樣之差異來辨識出情緒聲音之類別。

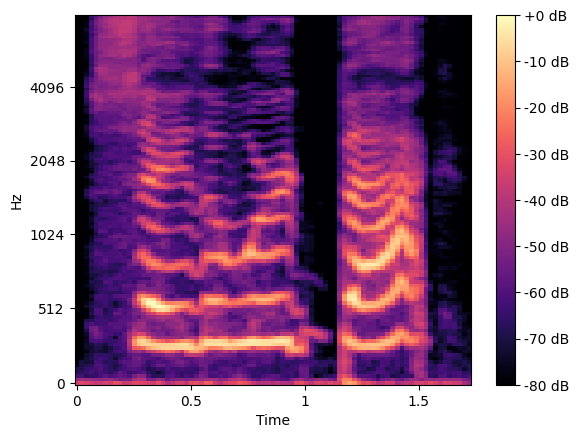


**圖4.1 圖片之目錄示意圖**

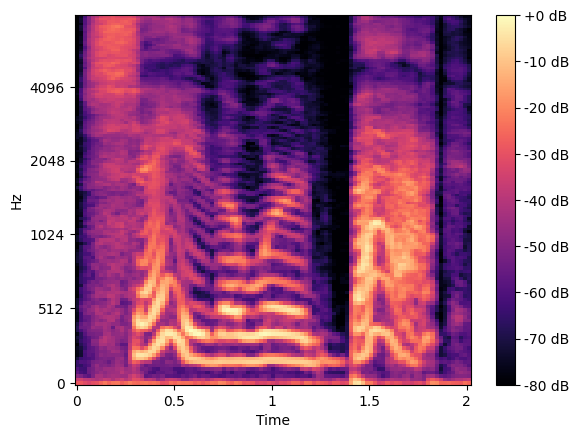
1. 程式進行封裝

實際應用時，會希望一批次可以處理大量檔案，因此會嘗試使用pyinstaller將程式碼包成一個exe檔案，將來在應用時可以一鍵執行產出結果，預計以HTML檔案形式呈現報告。

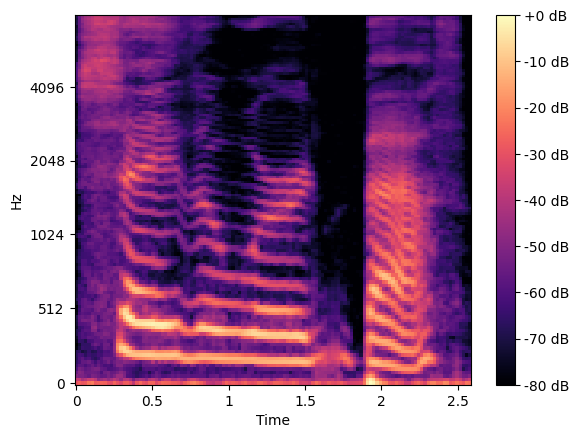
初步產出的結果有將聲音訊號轉成圖譜如圖4.2~4.6，會將該圖以類神經網路形式進行訓練，類神經網路構造如圖4.7，以此進行訓練之後，再計算測試集的準確率，在未優化超參數的情況下，初步算得之混淆矩陣如圖4.8，



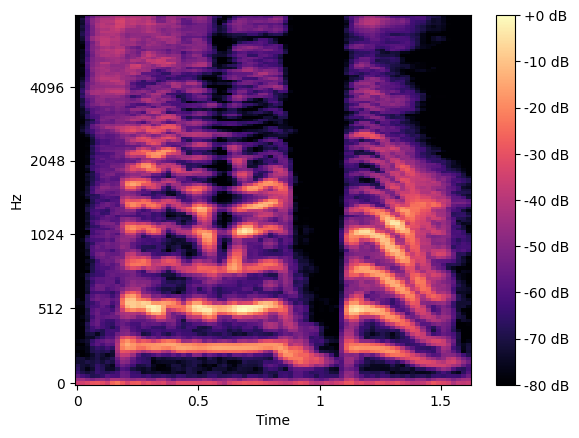
**圖4.2 MFCC產出之fear圖**



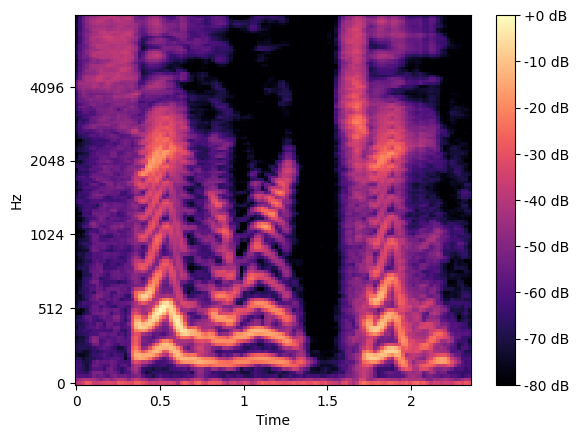
**圖4.3 MFCC產出之Pleasant\_surprise圖**



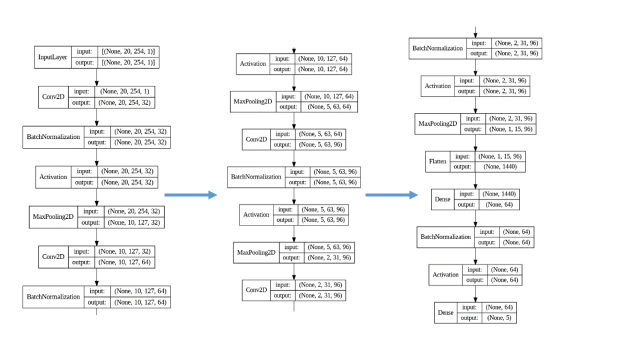
**圖4.4 MFCC產出之sad圖**



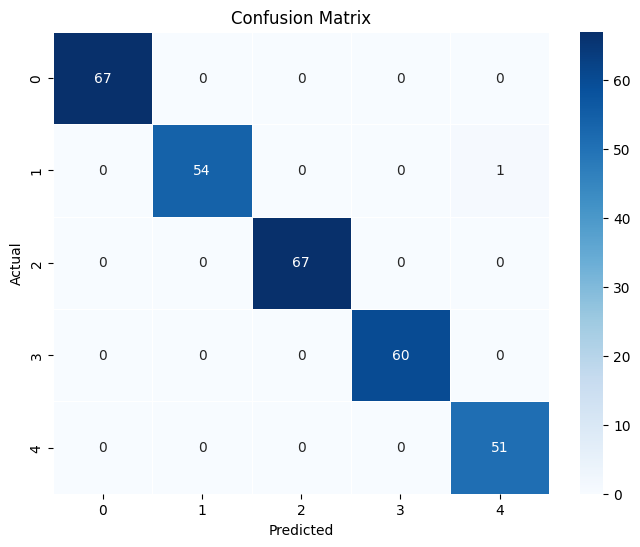
**圖4.5 MFCC產出之angry圖**



**圖4.6 MFCC產出之** **disgust圖**



**圖4.7 模型架構**



**圖4.8 CNN預測混淆矩陣**

# 第5章 人力配置

**表 5.1工作人員工作分配圖**

|  |  |
| --- | --- |
| 姓名 | 工作內容 |
| 方建智 | SVM模型建立及優化 評估  CNN模型建立及優化 評估  遷移學習 |
| 康宇翔 | 模型建立vgg及優化評估  輸出靜態網頁報表 |
| 曾敬峰 | 模型建立mobilenet及優化評估  類神經網路特徵分析可視化 |
| 全體成員 | 蒐集與分析資料、製作報告、建立資料庫、設計系統、功能實現、系統測試 |

# 第6章 未來展望：情緒聲音辨識處理系統

隨著技術的進步和醫療需求的增長，醫療聲音辨識處理系統在未來有著廣闊的發展前景。以下是對其未來展望的詳細分析：

(1) 提高系統精度

隨著更多情緒相關聲音數據的收集與標註，情緒聲音辨識系統的機器學習模型將不斷優化，進而顯著提升系統的精度和穩定性。這些數據可以來自各種環境，如電話客服、情緒監測平台等，涵蓋不同的情緒表達和個體特徵。特別是利用深度學習中的先進算法，如卷積神經網絡（CNN）和長短期記憶網絡（LSTM），能有效從複雜的聲音信號中提取情緒特徵，提升情緒識別的準確性。結合多模態數據融合技術例如搭配視訊時的影像，將進一步增強系統的綜合判斷能力，實現更精準的情緒分析。

(2) 強化實時處理能力

為了應對即時反饋和快速響應的需求，未來將引入高效的硬件加速技術，如使用GPU（圖形處理單元）或專用的AI加速芯片，顯著提升系統的實時處理能力。這樣，情緒聲音辨識系統可以在各類場景中快速響應，提供即時的情緒識別和反饋，尤其是在客服中心或情緒干預的環境中，能快速識別客戶情緒，提供精準的回應建議。邊緣計算技術的應用將使得系統能在設備端進行部分處理，減少延遲，提升用戶體驗。

(3) 提升用戶體驗

為了讓使用者更方便地操作，未來將開發更加直觀且友好的用戶界面，並整合語音交互、情緒報告和數據可視化等功能，幫助客服人員或心理輔導員更好地理解和運用系統提供的信息。結合各種數字平台，診斷結果可以實現自動記錄與共享，提升服務效率和協作性。隨著語音助手技術的發展，使用者能夠通過自然語言與系統互動，獲得個性化的情緒管理建議和健康服務。

(4) 法規和倫理考量

隨著技術的發展，情緒聲音辨識系統的應用需要特別注意隱私保護和數據安全問題。未來將結合相關法律法規，確保系統的運行符合道德與法律要求，保護使用者的隱私權和數據安全。例如，應用加密技術、數據匿名化處理等，防止數據洩露和未經授權的訪問。同時，建立透明的數據使用政策和監管機制，增強公眾對技術的信任，確保系統的公平性與可追溯性。

(5) 多模態學習和適應能力

情緒聲音辨識系統除了使用聲音資料，也可以融入視訊時產生的圖片資料進行更精準的預測，以應對不斷變化的需求和環境。這意味著系統將定期更新模型和數據庫，學習新的情緒模式和應對方法，保持識別的先進性和準確性。通過自適應算法，系統可以根據不同用戶的個體特徵和歷史數據，提供量身定制的情緒應對建議。這一持續學習的能力將幫助系統快速適應新興的情緒模式和特殊情境。

(6) 推動精神疾病相關研究

情緒聲音辨識系統在推動情緒研究和公共應用方面也具有重要潛力。通過收集和分析大量的情緒聲音數據，系統可以幫助研究人員識別情緒的早期特徵和趨勢，並開發新的情緒管理方法。這些數據還可以用於公共服務監測和心理健康研究，幫助預測和識別情緒波動的社會趨勢，提升社會安全與心理健康。例如，在需要遠程情緒輔導的情境中，情緒聲音辨識系統可以幫助減少接觸，提高服務的可及性與安全性。

總結來說，情緒聲音辨識處理系統在未來將具有巨大的發展潛力。隨著技術創新和跨領域合作，這一技術將在提升服務質量、優化資源分配、增強用戶體驗和改善心理健康管理等方面發揮關鍵作用，最終實現全面提升人們的情緒健康和生活品質。

總之，情緒聲音辨識處理系統在未來具有廣闊的發展前景。通過技術創新和多方合作，這一技術將在提升醫療質量、優化資源分配和提高患者滿意度等方面發揮重要作用，最終實現改善患者健康

# 參考文獻

[1] 劉佳鑫。「基於深層類神經網路之 聲音事件偵測系統」。碩士論文，國立臺北科技大學電子工程系研究所。

[2] 蔡佩樺。「機器學習聲音辨識應用於車禍偵測之研究」。碩士論文，逢甲大學資訊電機工程碩士在職學位學程

[3] 張永霖。「基於卷積神經網路的聲音分類機制」。碩士論文，朝陽科技大學資訊與通訊系，2021。

[4] M. A. Hossan, S. Memon and M. A. Gregory, "A novel approach for FCC feature extraction," The 4th International Conference on Signal Processing and Communication Systems, pp. 1-5, Gold Coast, QLD, Australia, 13-15 Dec. 2010.

[5] 林蓬榮, “規劃網站品牌的競爭策略,” 管理雜誌, vol. 326, pp. 127-129, 民90.

[6] M. Neely, “Optimal pricing in a free market wireless network,” Wireless Networks (10220038), vol. 15, no. 7, pp. 901-915, 2009.

[7] 吳克振, “非優質品牌應該如何應變,” 設計, vol. 99, pp. 110, 民90，6月.

[8] T. Wasserman, “Zsystems' Linux System Cracks $199 Price Tag,” Brandweek, vol. 41, pp. 20, 2000.

[9] T. Wasserman, “Zsystems' Linux System Cracks $199 Price Tag,” Brandweek, vol. 41, no. 18, pp. 20, 2000.