THE VOICE

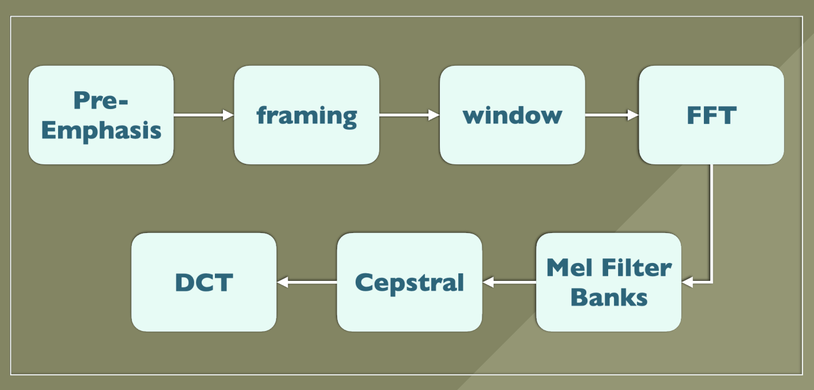
學生：黃新評、李建陽、陳睦炘、蔡嘉倫

♪. 摘要

我們的專題是開發音頻分析功能以識別樂器的特徵。另外，我們使用了「Magenta 計畫」中的數據庫，這是由 Google 提供的有關聲音識別的一項計畫。

♪.研究步驟

一. MFCC 特徵提取



我們使用倒頻譜的分析方式來完成對於聲音特色之特徵提取，詳細步驟見上圖。其關鍵在於它可以利用分幀與加窗格的方式進行多次的短時距傅立葉轉換將音色隨著時間的變化記錄下來，再利用梅爾倒頻譜的濾波器模擬人耳聽覺濾出我們最希望提取出的特徵。

二. 訓練模型

利用 tensorlfow 作為深入學習的基礎，以及 CNN 模型來完成訓練。並且使用 MFCC 處理完的資料作為 CNN 的訓練參數。另外，值得一提的是，關於給我們 AI 做訓練用的數據庫，是來自Google 與 Deepmind 合作開發的一項名為「Magenta」的計畫，那其中就有免費的開源資料庫供我們使用。總共大約有30萬組的資料可以使用，因此對我們來說是非常夠用的。其中左圖即為我們神經元輸出端的十筆結果（即代表會將音色辨識為這十種結果之一）。

Index	ID
0	bass
1	brass
2	flute
3	guitar
4	keyboard
5	mallet
6	organ
7	reed
8	string
9	synth_lead
10	vocal

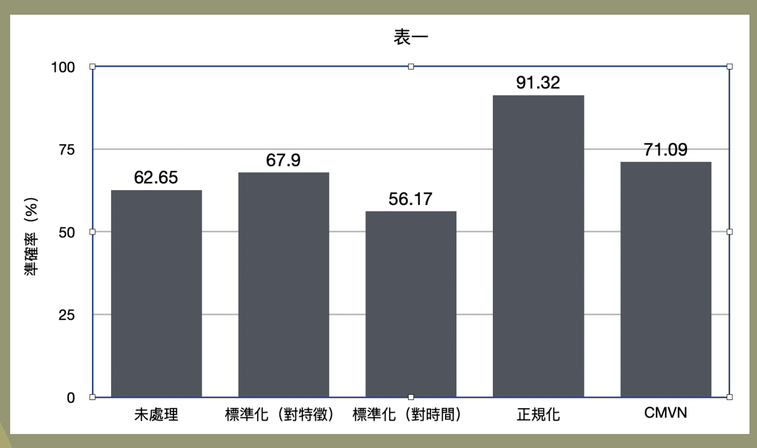
三. API 介面

我們建立了一個網站，讓一般使用者可以自由上傳音樂檔案。並且透過 Flask 伺服器上傳檔案到本地端進行分析，之後將結果實時回傳到網頁顯示出來。

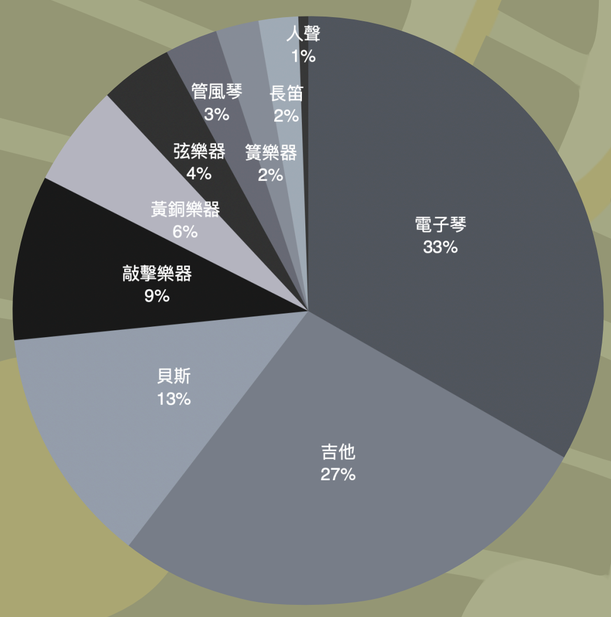
♪.結果分析

在做過多次嘗試之後，可以發現的是：CNN 模型本身的架構相比 MFCC 特徵提取之後的處理，在準確率的影響上相對甚小。因此我們的分析著重於在不同的 MFCC 特徵的處理方式如何影響準確率。MFCC 的處理方式有：標準化（standardization, 一維）、倒譜均值再方差歸一化（CMVN, 二維）、正規化（normalization, 二維）。

在選擇嘗試的取樣長度為 0.1 秒後。經過實驗得到效果最好是正規化的方式（下圖）。



之後將測試集丟入模型裡做預測，最終表現出的準確度為 86%（3523份 / 4076份）。同時將錯誤的部份分離出來進行分析。進而其中哪些樂器的錯誤率是最高的。下圖即為各項樂器之錯誤量（總錯誤量為 553）。



從圖中很明顯地看出，電子琴、吉他、貝斯的錯誤量相對高非常多。我們認為是三者在一些音準上的音色會有較高的相似度，容易造成誤判。例如吉他彈很低音與貝斯彈很高音的聲音會非常像，就算是人耳也幾乎很難分辨出來。