Content Table

一、Abstract

通常音樂輸入訊號是 non-stationary 訊號，會跟著時間改變。也就是頻率分量通常會隨著時間改變，因此無法在單一的頻譜分析非靜態(non-stationary)的訊號，將訊號做傅立葉變換後得到的結果，並不能給予關於訊號頻率隨時間改變的任何資訊，所以利用時間-頻率分析做更有效的分析工具。
通常時頻分析有以下較為常用的方法:STFT、Gabor transform與Wigner disturbution，也是本學期上過的課程內容。
在本次研究中，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，裡面包含10個不同的樂器類別以及各30個音檔，每個音檔長度約為2~10秒不等。
透過首先對音檔預處理（Pre-Emphasis）增加信躁比，再經由STFT來進行分析，此處設置window function為20~40ms，會有每個frame的大小，每組幀在各自進行傅立葉轉換後所得的複數結果會再進行相加，可得到每個點時間與頻率變化的大小與相位。此時由於需要用到深度學習CNN，因此需使用MFCC再把音檔的特徵提取出來，放入model進行training再分類為十個不同樂器類別，最後透過把MFCC進行標準化後放進model訓練，能得到91.2%的分辨率。

二、Introduction

1. 人耳對聲音特性

在人耳的聲域範圍內，聲音聽覺心理的主觀感受主要有響度、音高、音色等特徵和掩蔽效應、高頻定位等特性。其中響度、音高、音色可以在主觀上用來描述具有振幅、頻率和相位三個物理量的任何複雜的聲音，故又稱為聲音「三要素」。

響度：

人類耳朵所能感覺到聲波的強弱稱為「響度」。聲波的響度大小，通常與聲源振動的幅度有關，振動幅度越大，響度越大。分貝（dB），則是用來表示聲波的強弱的單位。
正常人聽覺的強度範圍為0dB—140dB。超出人耳的可聽頻率範圍(即頻域)的聲音，即使響度再大，人耳也聽不出來(即響度為零)。在人耳的可聽頻域內，若聲音弱到或強到一定程度，人耳同樣是聽不到的。當聲音減弱到人耳剛剛可以聽見時，此時的聲音強度稱為「聽閾」。一般以1kHz純音為準進行測量，人耳剛能聽到的聲壓為0dB，而當聲音增強到使人耳感到疼痛時，這個閾值稱為「痛閾」。仍以1kHz純音為準來進行測量，使人耳感到疼痛時的聲壓級約達到140dB左右。

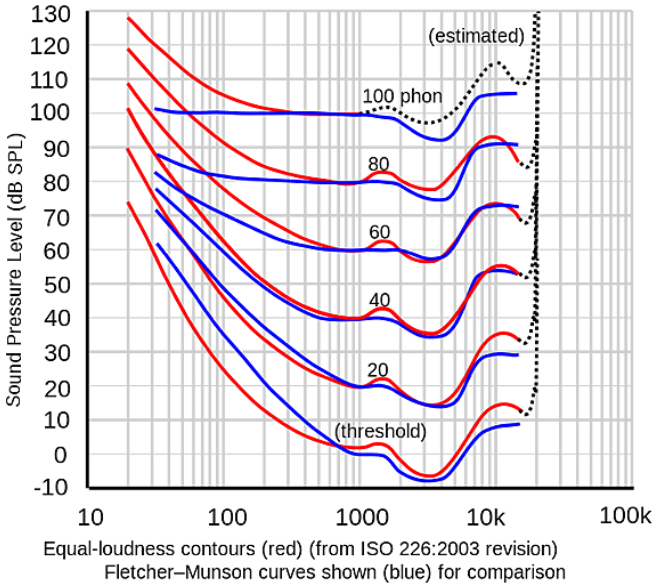
音高：

聲波的頻率高低稱為「音調」。聲波的音調高低，通常與發生體振動快慢有關，物體振動頻率越大，音調就越高。人耳對頻率的感覺同樣有一個從最低可聽頻率20Hz到最高可聽頻率別20kHz的範圍。響度的測量是以1kHz純音為基準，同樣，音高的測量是以40dB聲強的純音為基準。

音色：

「音色」又叫音品，它反映了聲波（聲音）的品質和特色。由聲音波形的諧波頻譜和包絡決定。聲音波形的基頻所產生的聽得最清楚的音稱為基音，各次諧波的微小振動所產生的聲音稱泛音。單一頻率的音稱為純音，具有諧波的音稱為複音。每個基音都有固有的頻率和不同響度的泛音，藉此可以區別其它具有相同響度和音調的聲音。不同物體發出的聲音，其音色是不同的，因此我們才能分辨不同人講話的聲音、不同樂器演奏的聲音等。

2. 等響曲線（Equal Loudness Contour）



藍色的曲線是 Fletcher與Munson當時1933年時，美國兩位在貝爾實驗室，所研究的結果，Y軸代表實際音壓強度(dB-SPL)，X軸代表不同受測頻率，圖上面的各條曲線是以每20個phon為測量單位去進行測試的結果。（phon是以 pure tone測出的聲音響度單位 ，0 phon代表人耳最小能聽到的聲音響度）。
紅色曲線是根據後來的研究人員Robinson and Dadson所測量出的曲線為基礎，再行修正，在2003年認證為ISO226標準，也就是我們現在所說的Equal Loudness Contour（等響曲線）。
以20 phon這條響度曲線來看，測試音1kHz為20dB-SPL時，60Hz的聲音要跟它聽起來一樣大聲，實際音量要達到58dB-SPL，3kHz只要15dB-SPL就一樣大聲了。

從這張等響曲線我們可以解讀出幾件事：

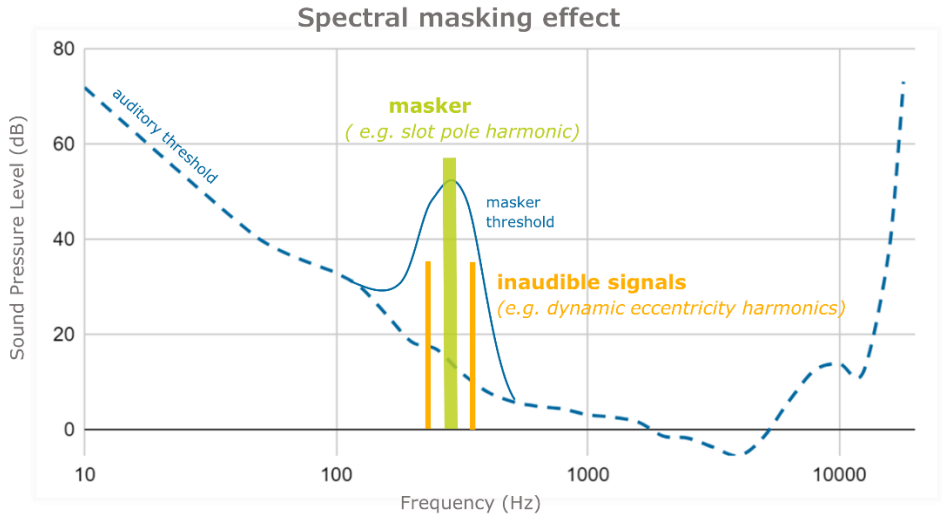
- 1kHz以下，越低的頻率要越大聲才能聽起來有同等響度。
- 2kHz～5kHz之間為人耳最敏感的區域，而且人耳在低音量時比起高音量時對此區域敏感。
- 1kHz~2 KHz間人耳對音量的敏感度會稍差些。
- 6kHz以上，人耳的敏感度會逐步下降，但比起低頻率來說，音量大小對人耳低頻的敏感度影響高於5kHz以上頻率。

整體來說，**人耳對於中頻的敏感度較佳**，高低頻在音量小時，人耳較不靈敏，但隨著音量變大，人耳對各頻率的反應差異性就慢慢變小了。

3. 人耳遮蔽效應

可分為頻率遮蔽與時域遮蔽兩大類:

(1)頻率遮蔽

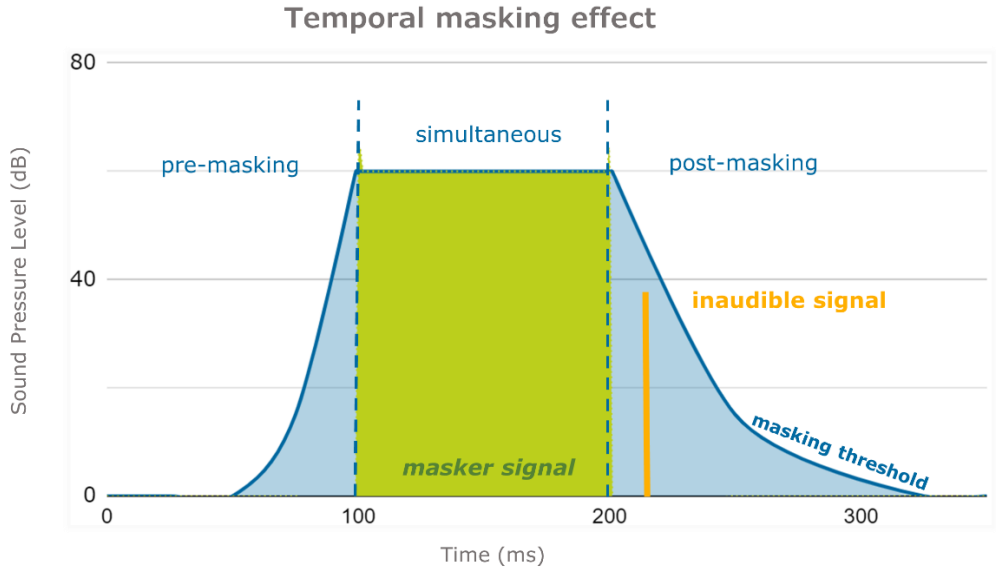


可分為兩種情況，一種是若能量過大則會遮蔽另外一個能量較小的聲音，假設遮蔽音是單一頻率的純音，它的遮蔽效果會隨著音量變大，遮蔽的頻率範圍也會變大，在頻域中 400Hz 能量強度約 58dB 的聲音訊號會對鄰近的二組聲音訊號產生遮蔽效應，因此對於聲音訊號而言，能量的強度將會影響所能遮蔽的範圍，當能量愈強時，所能遮蔽的範圍也會相對越大。

另一種情況是如果聲音強度相等，則頻率低的聲音會遮蔽到頻率高的聲音。

一般來說，弱純音離強純音越近就越容易被掩蔽；低頻純音可以有效地掩蔽高頻純音，但高頻純音對低頻純音的掩蔽作用則不明顯。

(2)時域遮蔽



如果兩個聲音在時間上特別接近，人類在分辨它們的時候也會有困難。例如如果一個很強的聲音後面緊跟著一個很弱的聲音，後一個聲音就很難聽到。但是如果在第一個聲音停止後過一段時間再播放第二個聲音，純音若間隔5毫秒以上，後一個聲音就可以聽到。

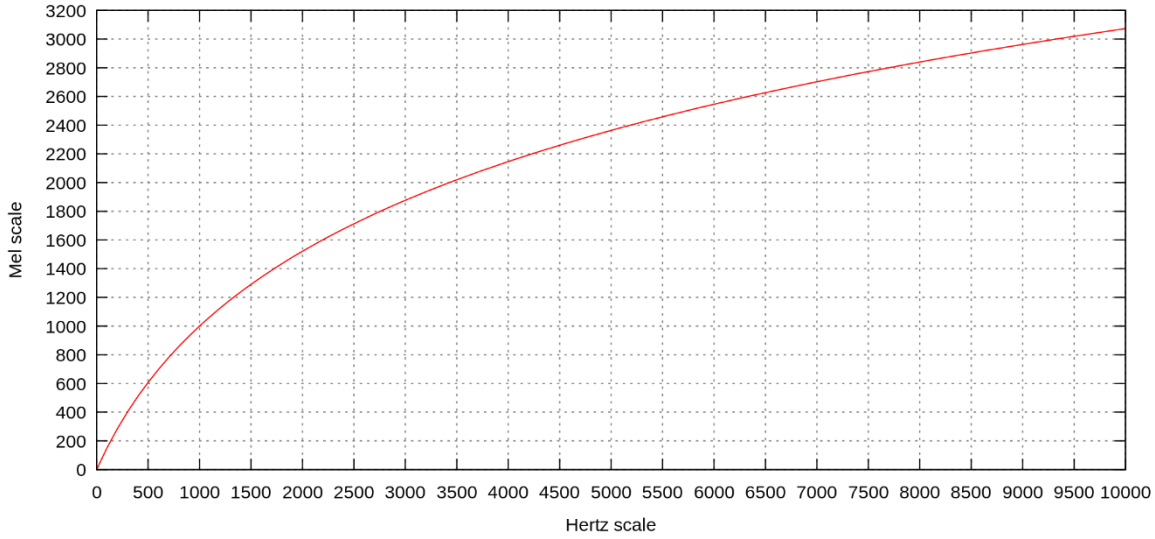
當然如果在時序上反過來效果是一樣的，如果一個較弱的聲音出現在一個較強的聲音之前而且間隔很短，那個較弱的聲音也聽不到。

4. 聽覺臨界頻帶(Critical Band)

當我們改變窄頻帶聲音刺激(narrowband sound stimulus)時，其聲音成分若跨越某一頻率，則聽覺上會感到有差異，而在一頻率範圍內，則感覺不到差異，這個頻率範圍稱臨界頻帶。

5. 梅爾刻度(Mel Scale)

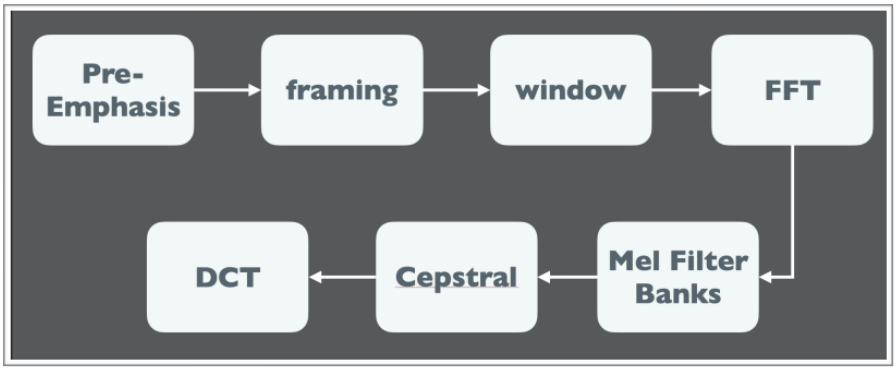
是一種非線性刻度單位，表示人耳對等距音高(pitch)的感官變化，參考點定義是將1000Hz，且高於人耳聽閾值40分貝以上的聲音信號，定為1000mel。在頻率500Hz以上時，人耳每感覺到等量的音高變化，所需要的頻率變化隨頻率增加而愈來愈大。



人類對不同頻率語音有不同的感知能力：對1kHz以下，與頻率成線性關係，對1kHz以上，與頻率成對數關係。頻率越高，感知能力就越差。

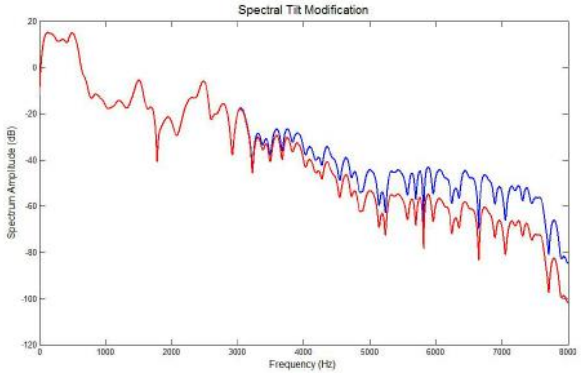
6. MFCC

(0) Flow Chart



(1) 預加重(Pre-Emphasis)

語音信號往往會有頻譜傾斜（Spectral Tilt）現象:



由上圖能看到高頻部分的幅度會比低頻部分的小，預加重在這裡就是起到一個平衡頻譜的作用，增大高頻部分的幅度。以現實層面來說，是為了消除發聲過程中聲帶和嘴唇的效應，來補償語音信號受到發音系統所壓抑的高頻部分。而整體而言就是為了補償高頻的振幅。

Pre-Emphasis能使用一階濾波器來實現:

$$y(t)=x(t)-\alpha x(t-1),\quad 0.95<\alpha<0.99$$

對信號做一階差分時，高頻部分（變化快的地方）差分值大，低頻部分（變化慢的地方）差分值小，就能達到平衡頻譜振幅的作用。

(2)分幀(Framing)

由於下一步我們要使用FFT，但通常音樂信號中的頻率會隨時間變化，所以我們希望透過framing讓一段短時間的音檔不太會隨著時間改變，才能讓我們實作FFT。一般設置幀長取20ms~40ms，通常取樣點 N 的值是 256 或 512 ，而且根據人耳所能聽到的極限與奈奎斯定理，因此，音檔的取樣速率通常會是8kHz或是16kHz，以 8 KHz 來說，若frame長度為 256 個取樣點，則對應的時間長度是$\frac{256}{8000}\cdot 1000=32$ ms 即通常framing會落在20~40ms。

(3)加上窗函數(windowing)

在做STFT時會有很多種不同的window function，此處能應用的window function一樣有很多不同的function可以供使用者做選擇，例如rectangular window function ,hanning window ,hamming window等。而在此處由於前面已經有經過framing，並且我們在做fft時都會假設這個frame內的signal會是週期性訊號，但有可能把所有frame加在一起的時候會導致在frame左右邊的不連續，因此就需要透過window來加強frame左端和右端的連續性，通常會選擇採用 Hamming window:

$$w(n)=0.54-0.46\,cos(\frac{2\pi n}{N-1})$$

(4) 快速傅立葉轉換(FFT)

對framing並windowing後的各幀信號再進行N點傅里葉變換得到各幀的頻譜。N為每幀的採樣點，通常情況下N的值為256或512。這也叫STFT（Short-Time Fourier-Transform）。

一般DFT的時間複雜度為 O(N^2)，在取樣率大的時候會造成計算上的負擔，因此可以透過快速傅立葉轉換(FFT)演算法，將分解計算量在合併，以將時間複雜度降到O(NlogN)，大幅減少計算量。

以下為FFT原理:
先說明DFT的形式:

$$X[k]=\sum_{n=0}^{N-1}x[n]W_N^{nk},0\le k\le N-1 \qquad W_N^{nk}=e^{-j\frac{2\pi k}{N}n}$$

在這裡我們觀察能發現每項需要N次複數乘法和N-1次複數加法，若有N項，則會需要N^2次複數乘法以及N(N-1)次複數加法，其複雜度為$O(N^2)$。

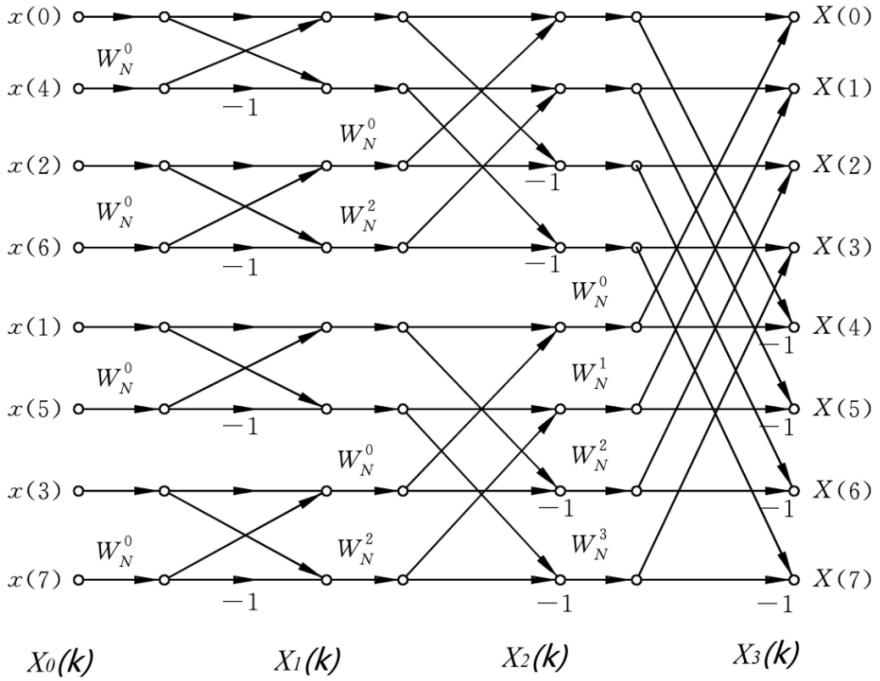
將離散傅立葉變換公式拆分成奇偶項，則前 N/2 個點可以表示為：

$$\begin{aligned}X[k]&=\sum_{r=0}^{\frac{N}{2}-1}X[2r]W_N^{2rk}+\sum_{r=0}^{\frac{N}{2}-1}X[2r+1]W_N^{(2r+1)k}\\&=\sum_{r=0}^{\frac{N}{2}-1}X[2r]W_{\frac{N}{2}}^{rk}+W_N^k\sum_{r=0}^{\frac{N}{2}-1}X[2r+1]W_{\frac{N}{2}}^{rk}\\&=A[k]+W_N^kB[k]\ ,k=0,1,\dots,\frac{N}{2}-1\end{aligned}$$

同理，後 N/2 個點可以表示為：

$$X[k+N/2]=A[k]-W_N^kB[k]\ ,k=0,1,\dots,\frac{N}{2}-1$$

因此，後 N/2 個點的值可以通過計算前 N/2 個點時的中間過程值確定。
對 A[k] 與 B[k] 繼續進行奇偶分解，直至變成 2 點的 DFT，此過程可以避免重複計算，也就能實現快速離散傅立葉變換（FFT）。
由這個拆分方法能畫成以下的蝴蝶圖
此為八個點FFT的蝴蝶圖:



能發現經過計算幾次乘法和加法，便可完成離散傅立葉變換過程，而不用對每個資料逐一計算。
每一個蝶形單元運算時，都會進行一次乘法和兩次加法。
而我們定義每分割一次，稱為一級運算，並且每一級中，均有 N/2 個蝶形單元。

故完成一次 FFT 所需要的乘法次數和加法次數分別為：

$$N/2 \cdot \log_2 N,\ \ N \cdot \log_2 N$$

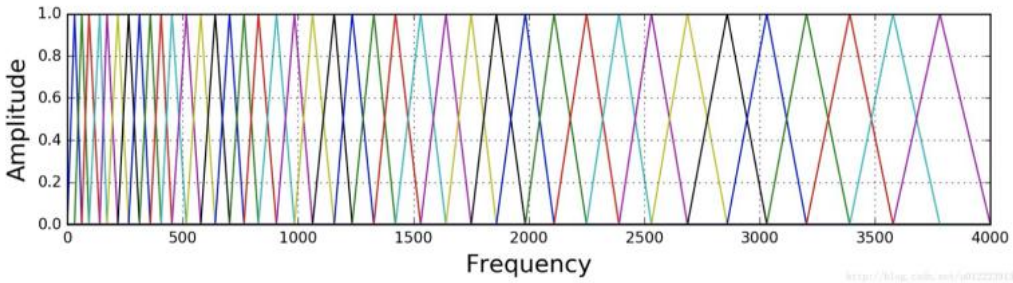
複雜度就從原本DFT的$O(N^2)$降為變成O(NlogN)

(5) 梅爾濾波器（Mel Filter Banks）

在做完短時傅立葉之後，會先通過了一個特殊的梅爾濾波器。這個濾波器的樣子見下圖。他由很多的三角型濾波器組成，每個濾波器在中心頻率響應都是 1 ，一般會設定40個這樣的三角形濾波器組成梅爾濾波器，並且利用通過梅爾濾波器的頻譜功率映射到梅爾刻度上。可以注意到梅爾刻度的濾波器組就是為了模擬人耳的靈敏程度，在低頻區域有很多的濾波器，他們分佈比較密集，但在高頻區域，濾波器的數目就變得比較少，分佈很稀疏。而經過實驗後發現人耳對於頻率 f 的感受是呈對數變化的，大致能表示為下列式子:

$$Mel(f) = 2595 * log(1 + \frac{f}{700}) = 1125 * ln(1 + \frac{f}{700})$$

此濾波組如下圖所示:



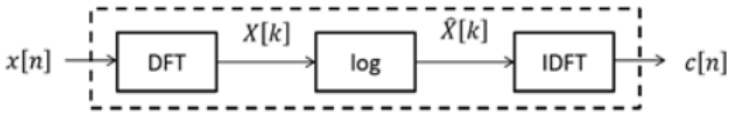
在低頻部分，人耳感受是比較敏銳

在高頻部分，人耳的感受就會越來越不靈敏

三角形filter:對頻譜進行平滑化，並消除諧波的作用，突顯原先語音的共振峰。

(6)倒譜分析（Cepstrum Analysis）

・Cepstrum Analysis Flow



$$c[n] = \mathcal{F}^{-1}\{\log|\mathcal{F}\{x[n]\}|\}$$

$$c[n] = \sum_{n=0}^{N-1} \log\left(\left|\sum_{n=0}^{N-1} x[n]e^{-j\frac{2\pi}{N}kn}\right|\right) e^{j\frac{2\pi}{N}kn}$$

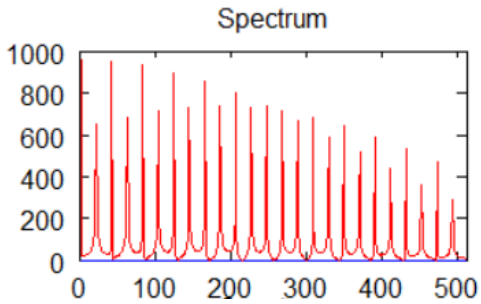
- **DFT:**前面在FFT有介紹過，能參考(4)，而在此MFCC flow中即是 FFT+Mel filter 當作input

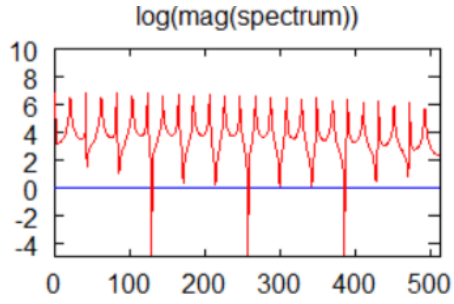
- **對數振幅譜:**

對於頻率而言分為很多種不同的頻譜分析方式，其中一種即是對數振幅譜。

對數振幅譜中各譜線的振幅都作了對數計算，所以其縱坐標的單位是dB（分貝）。這個變換的目的是使那些振幅較低的成分相對高振幅成分得以拉高，以便觀察掩蓋在低幅噪聲中的周期信號。

例如下圖中從一般頻譜信號，對magnitude取log能得到右圖





這樣就能進行觀察低幅噪聲中的周期信號。會有以下好處:

(i)取log讓 spectrum 看來更有週期性，便於 IFFT 找出 fundmental pitch

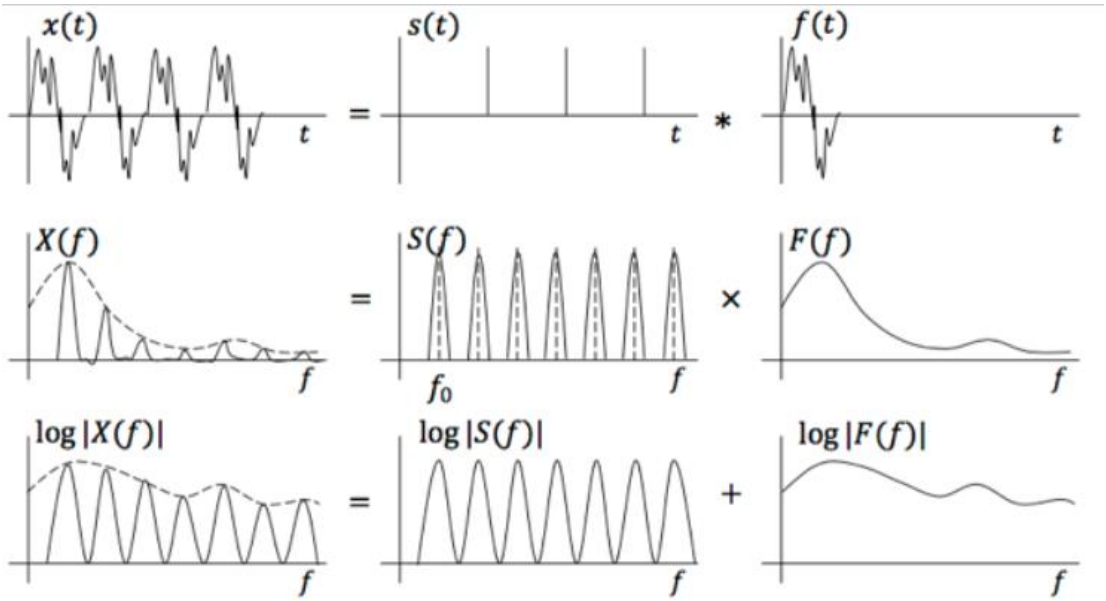
(ii)讓 time domain convolution or frequency domain filtering 轉換 cepstrum domain 的 addition。

-spectrum特性:

下面是一個語音的頻譜圖。峰值就表示語音的主要頻率成分，我們把這些峰值稱為共振峰（formants），而共振峰就是攜帶了聲音的辨識屬性，用它就可以識別不同的聲音。

以下圖為例:會是由x(t)=s(t)+f(t)，所組成的，X(f)代表的是x(t)在頻譜上的樣子;F(f)是Spectral Envelope，S(f)則是由Spectral details。

若我們要進行深度學習，我們要提取的不僅僅是共振峰的位置，還得提取它們轉變的過程。所以我們提取的是頻譜的包絡（Spectral Envelope）。這包絡就是一條連接這些峰點的平滑曲線，也就是F(f)。



再來就需要討論，如何將Envelope與Spectrum detail分離出來，在這透過IDFT來實現，若是在MFCC中就是使用DCT來還原。

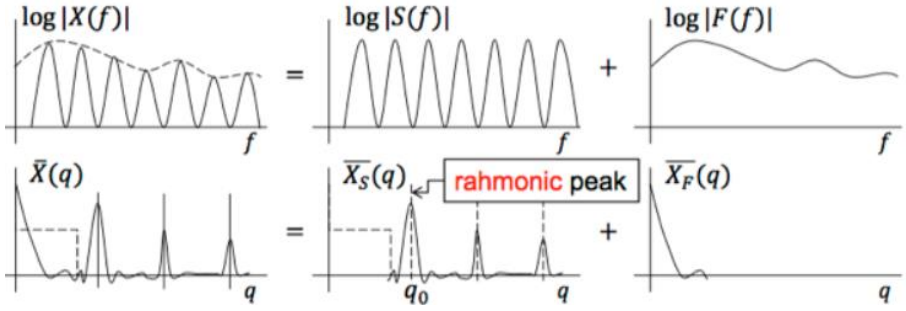
- IDFT

首先，音樂信號不是單頻，其具有豐富的泛音，也就是說可以想像成 frequency domain 是有週期性，例如: 1st fundamental, 2nd harmonic, 3rd harmonic, etc。透過取log後，我們能獲得更加具有週期性的頻域。

在此處我們把頻域上的波形再次進行分析:

此時做 IDFT 可以得到 fundamental frequency (pitch) 且不受 missing fundamental 影響。也就是 cepstrum 不是只看單一頻率 (fundamental), 而是看 fundamental 和 harmonics 的頻差決定 pitch (qo)

如下圖所示:



我們能發現經過IFFT後，Envelope的位置就會位於低頻處，而其他harmonics的成分會為在高頻，我們也就能分離出這兩種成分了。

- DCT(Discrete cosine transform)

在MFCC中，透過DCT來實現逆轉換倒time domain上的過程，40個經過梅爾濾波器對數能量Ek帶入DCT，求出L階的Mel-scale Cepstrum 參數，這裡 L 通常取 12。離散餘弦轉換公式如下：

$$C_m = \sum_{k=1}^{N} cos[m * (k - 0.5) * p/N] * E_k, m = 1, 2, \ldots, L$$

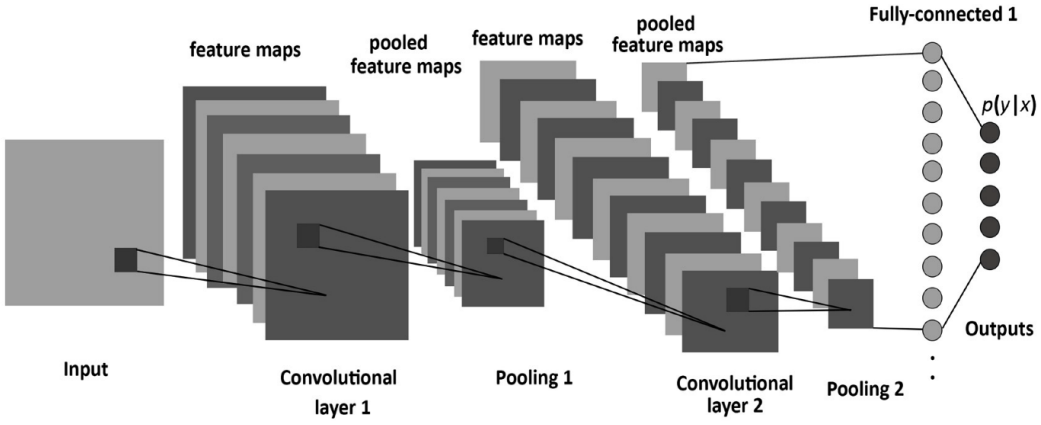
其中 Ek 就是代表上述的frequency domain中所算出來的三角濾波器和頻譜能量的內積值，N 是三角濾波器的個數。

從整個倒譜分析後，我們就能以原本有的對數能量Ek+12個由DCT所算出來的MFCC參數，做為表示MFCC特徵的13維matrix。

7. CNN

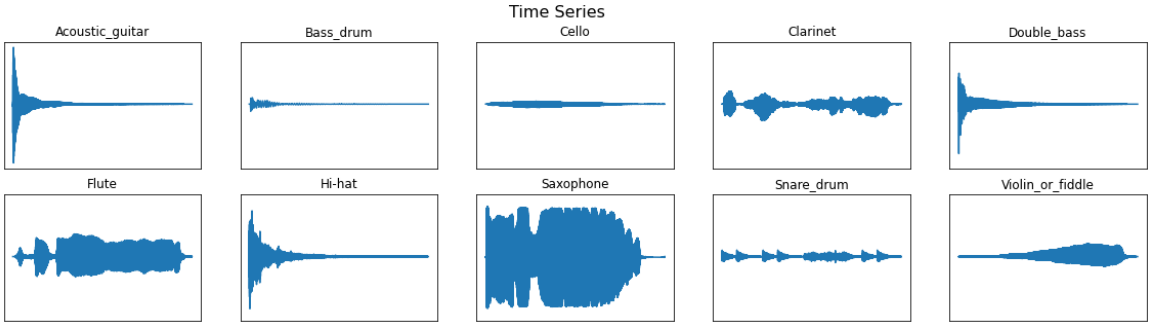
卷積神經網路（Convolutional Neural Network, CNN）是一種前饋神經網路，由一個或多個卷積層和頂端的全連通層組成，同時也包括關聯權重和池化層。這結構使得卷積神經網路能夠利用輸入資料的二維結構。與其他深度學習結構相比，卷積神經網路在圖像和語音辨識方面能夠給出更好的結果。這一模型也可以使用反向傳播演算法進行訓練。相比較其他深度、前饋神經網路，卷積神經網路需要考量的參數更少，因此成為深度學習模型中，大眾常用的方法之一。

一般的model就會如下圖所示:



三、實驗

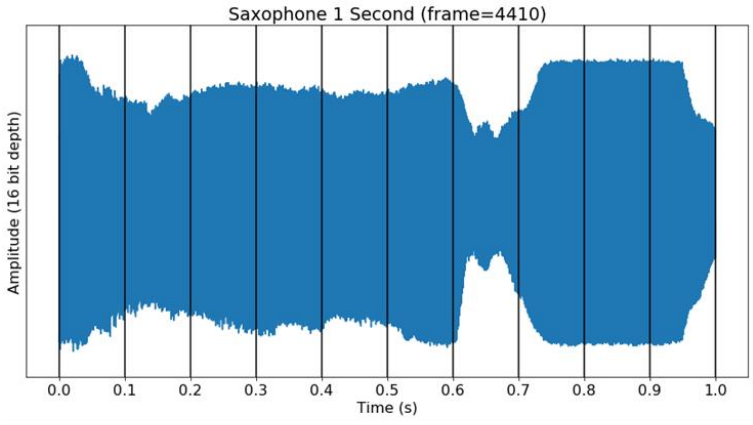
1.首先，Magenta這個資料庫的音檔的十個樂器種類分別是Acoustic quitar,Bass drum,Cello,Clarinet,Double bass,Flute,Hi-hat,Saxophone,Snare drum,Violin/fiddle，先將十種樂器load，先做pre-emphasis後畫出time Domain和frequency domain。

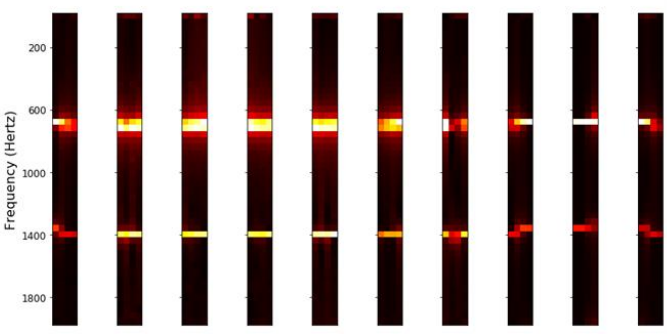


(time series)

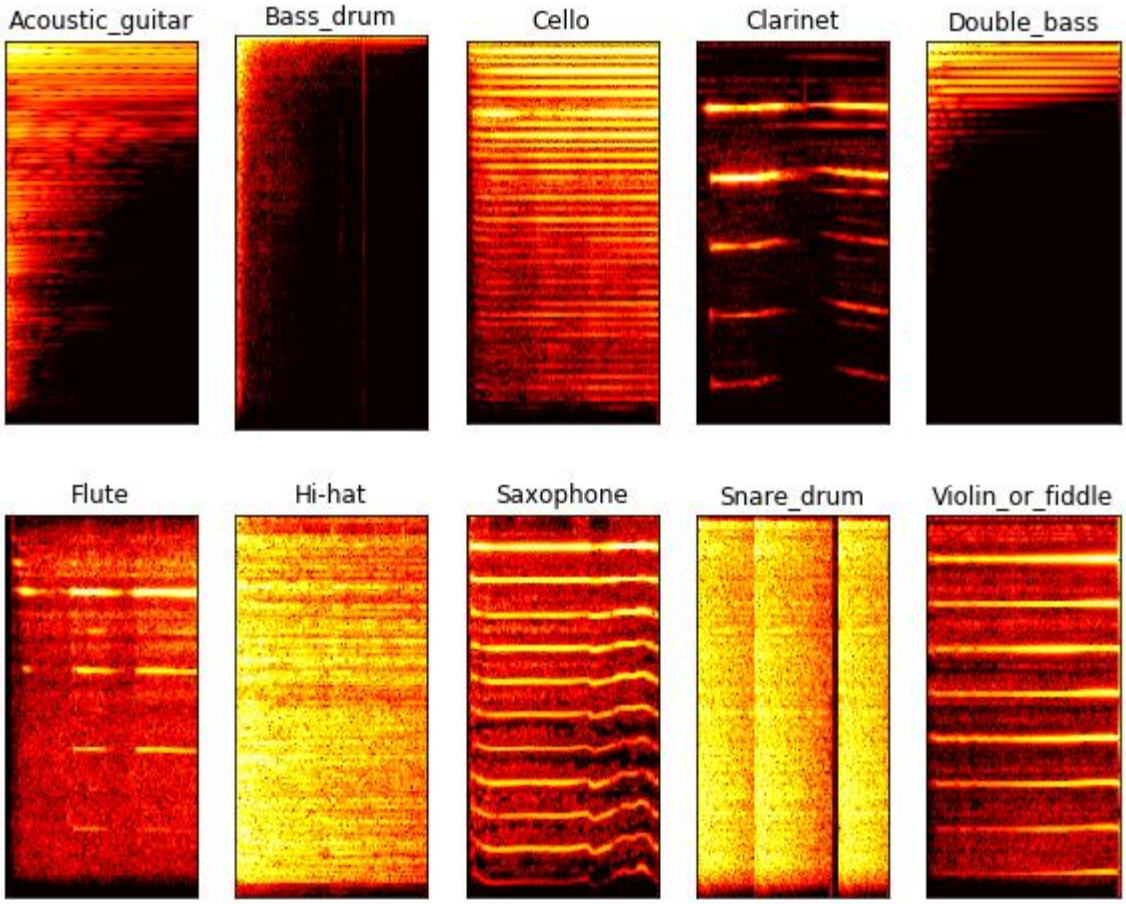
(frequency domain)

2.再來，按照framing，先定義出window大小設為0.1s，並且對framing完的time domain做FFT

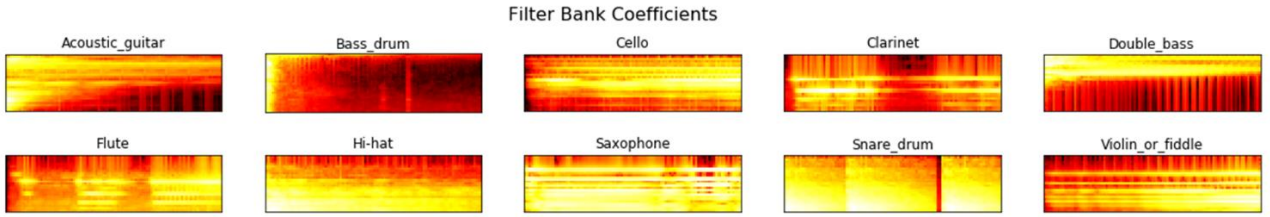




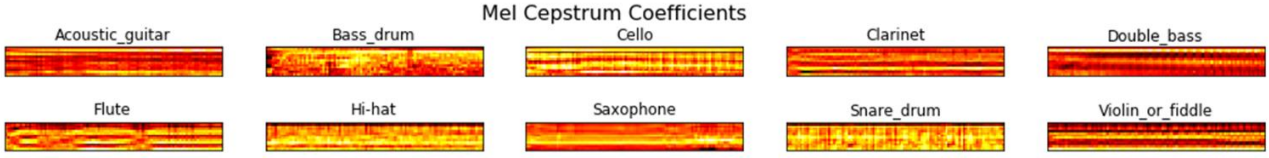
3. 以下是十種樂器經過FFT的結果:



4.再來把經過FFT的頻譜放入Mel-filter中，並取log



5.最後是經過DCT所得到的MFCC倒譜分析



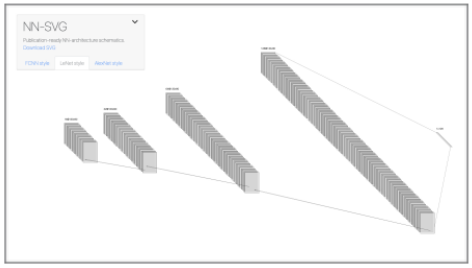
6.CNN辨識

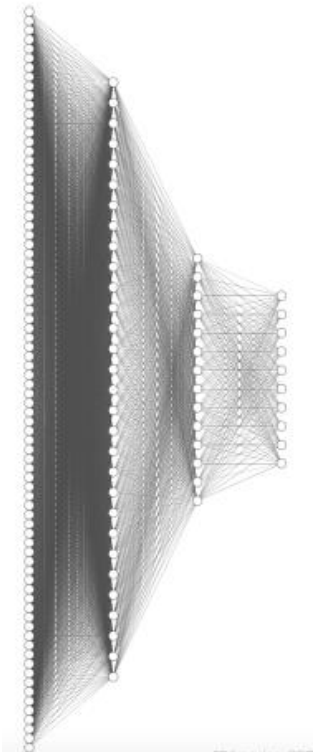
在資料庫中，有一個 .json 檔，其中標注了每一個 .wav 檔的聲音屬性，其中這些樂器：bass、brass、flute、guitar、keyboard、mallet、organ、reed、string、synth_lead

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

使用的 CNN 模型構造為先做四個卷積層，一個池化層，然後做一個 Dropout 來避免 over-fitting 的發生。再來再接著做三層全連接層，最後輸出層為 10 個樂器種類。

第一張圖為卷積層示意圖，而第二張圖為全連階層示意圖:

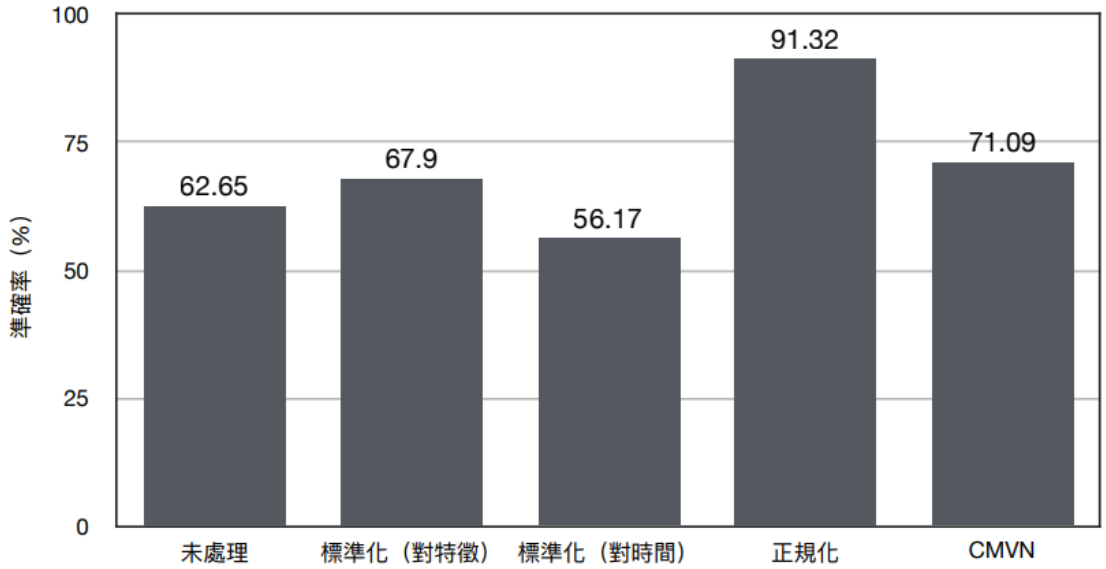




四、結果與分析

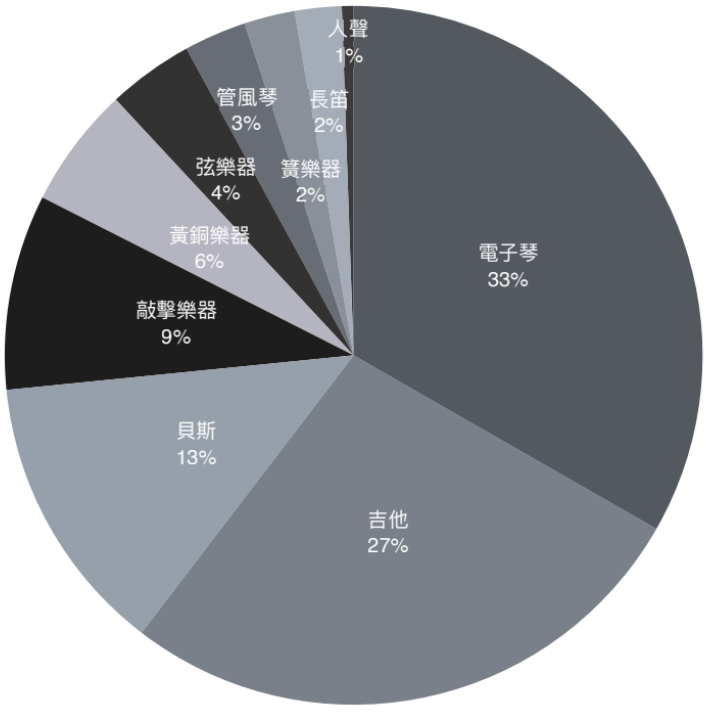
由於訓練集裡的音檔在做錄製時並不會將整個檔案結構填滿聲音訊號，會造成在做MFCC特徵提取時提取到沒有聲音訊號的部分，就會導致計算結果出現誤差。因此，必須先將 .wav 檔做去頭去尾的動作，也就是將沒有聲音的部分去除掉。

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（standardization, 一維）、倒譜均值再方差歸一化（CMVN, 二維）、正規化（normalization, 二維）以下是各個方法的準確率:



由此可知，正規化之後所表現的準確率特別突出。因此我們將該訓練模型保存下來同時，我們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）。

並且將錯誤的部份分離出來進行分析。就可以發現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。可能是三者在一些音準上的音色會有較高的相似度，也就導致了特徵相似，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。

五、參考文獻

[1] Speech Technology: A Practical Introduction Topic: Spectrogram, Cepstrum and Mel-Frequency Analysis

[2] Taylor, 2009, ch. 12; Rabiner and Schafer, 2007, ch. 5; Rabiner and Schafer, 1978, ch. 7

[3] Mirlab: audio signal processing: speech Feature Mfcc

[4] Beth Logan, Mel Frequency Cepstral Coeffiecents for Music Modeling

[5] Róisín Loughran, Jacqueline Walker , Michael O'Neill , Marion O'Farrell: The Use of Mel-frequency Cepstral Coefficients in Musical Instrument Identification