

國立台北大學資訊工程學系專題報告

人聲的分離與樂器聲的重建

專題組員：古明翰、曾品翔

專題編號：PRJ-NTPUCSIE-106-002

執行期間：106 年 09 月至 107 年 05 月

1. 摘要

當你很想要唱歌的時候，但並不希望有歌手的聲音，卻發現網路上並沒有 KTV 版的音樂，這時你會怎麼做？你可能會直接跟著歌手一起唱了。於是我們產生出把人聲與樂器聲分離的想法，如此一來我們就有專屬於自己的「KTV」了。而分離出來的人聲是不是還可以換成不同的樂器，使其具有別種風格。

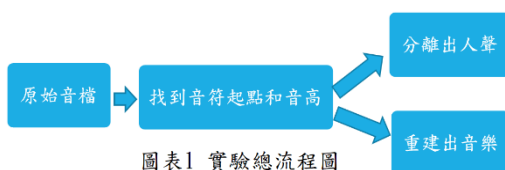
2. 簡介

每個人與每一種樂器都有其獨特的聲音，在本次實驗當中，我們選擇使用鋼琴作為我們研究的主要樂器，為常見且主要的樂器之一，具有頻率穩定且音色具高辨識度等優點。藉由分析鋼琴與人聲頻率的不同，找出屬於鋼琴的音高，以此重建出純鋼琴的樂曲，並將鋼琴聲從與人聲混合的歌曲中去除，而獲得人聲。

3. 專題進行方式

我們的專題分為兩個部分：

1. 將具有人聲與鋼琴聲的歌曲中分離出人聲
2. 重建出純鋼琴的歌曲



圖表1 實驗總流程圖

要達成分離與重建，需要先找出音符的起點及音高。

(一) 找出音符的起點

A. 錄製音檔與琴鍵

以取樣率 44,100 Hz

樣本位元數 16 bits

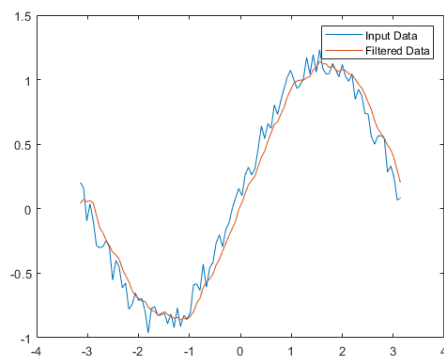
單聲道作為取樣的標準，屬於高品質音樂，包含完整的音檔資訊。

錄製每個琴鍵的聲音。

B. 去除雜音

錄製音檔時會有雜音，而人聲中氣音也會使影響波形。

以 $[1, 6, 15, 20, 15, 6, 1]/64$ 的平均濾波器做低通濾波，可以讓圖形的平滑以便判斷。



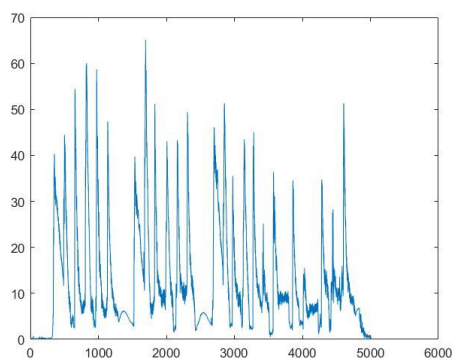
圖表 2 去除雜音前後的波形

C. 計算音量

聲波的振幅高低代表音量的大小，將音檔切成多個短時距音框做分析，每

個音框為 20ms，包含足夠的資訊，且強度變化不會過大。

由聲音訊號的平方和計算出能量，可以有效的將大訊號增強。



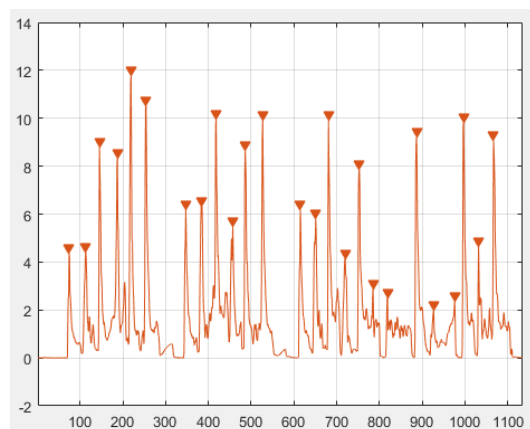
圖表 3 訊號能量圖

D. 尋找音起點

音的起點可以由聲音的能量得知，當敲擊琴鍵時能量會迅速上升，便可以藉由找出能量的高峰位置得出音符的起點位置，如圖表 4。

找高峰位置我們利用以下兩項條件，第一，每個高峰必須大於 1.3 倍的平均音量。

第二，藉由樂譜節拍數計算出間距，每個高峰的距離必須大於間距。



圖表 4 起點位置

(二)計算音高

A. 計算自相關係數

自相關 (Autocorrelation)，也叫序

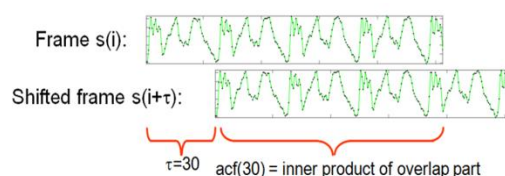
列相關，是一個訊號於其自身在不同時間點的互相關。非正式地來說，它就是兩次觀察之間的相似度對它們之間的時間差的函數。

$$acf(\tau) = \sum_{i=0}^{n-1-\tau} s(i)s(i+\tau)$$

圖表 5 自相關函數

當自相關係數(ACF 值)愈高時，表示其相關性也愈高，並且如果訊號有週期性，其 ACF 值也具週期性。人聲加鋼琴聲為準週期性訊號，因此我們可以利用 ACF 值來找出音訊的基本週期，經過音訊之取樣率轉換後並可以得到聲音的音高(基本頻率)。

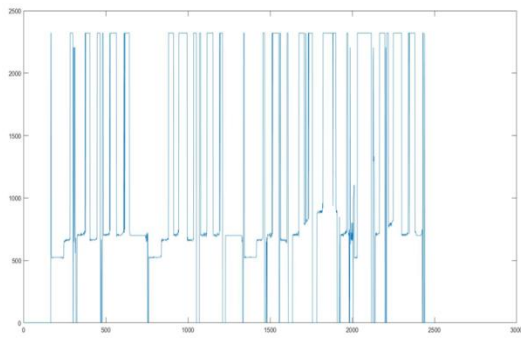
我們以 20 毫秒作為音框大小，取樣率為 44,100 赫茲，因此一個音框包含 882 個樣本，每次移動 1 個樣本數，與原音框的重疊處相乘，計算該音框之 ACF 值，如圖表 6。



圖表 6

B. 找出鋼琴的頻率

一般成年男性的人聲基本頻率為 85 至 180Hz，成年女性則為 165 至 255Hz。唱歌的頻率只比說話頻率高一點。而鋼琴的基準音中央 La 就有 440 赫茲，由此可知人聲和鋼琴的頻率相比有一段差距。因此我們將可以限定 ACF 取值的範圍將人與鋼琴的頻率分離出。圖表 7 中極端值為人聲的週期，其餘為鋼琴的週期。



圖表 7 週期圖

分離出人聲



圖表 8 分離出人聲流程圖

(一)壓低樂器聲

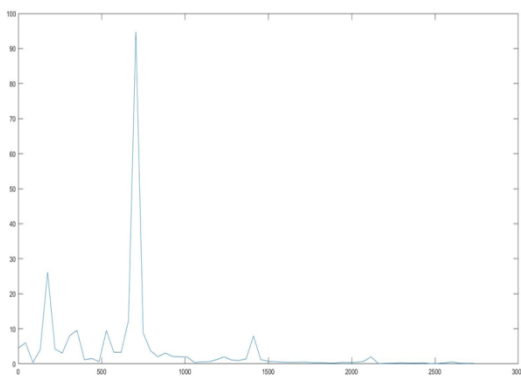
A. 傅立葉轉換

兩個音的起點之間，為一個音符。將每個音切成小音框，做傅立葉轉換，從時域轉換為頻域。

$$F_u = \frac{1}{N} \sum_{x=0}^{N-1} \exp \left[-2\pi i \frac{xu}{N} \right] f_x$$

圖表 9 傅立葉方程式

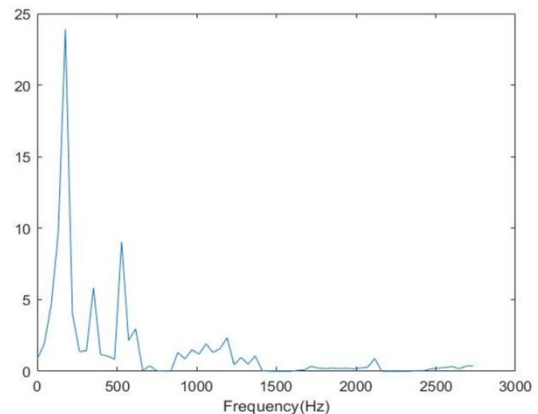
因為鋼琴的頻率和人聲頻率有明顯區隔，在傅立葉轉換的頻譜中就可以分分出二者，頻率高者為鋼琴，低者為人聲，如圖表 10 所示。



圖表 10 傅立葉頻譜

將鋼琴的能量壓低，以圖表 10 為例，鋼琴的頻率大約在 700Hz 所以會將

700, 1400, 2100……，等基頻、倍頻的能量壓低。



圖表 11 壓低後的頻譜

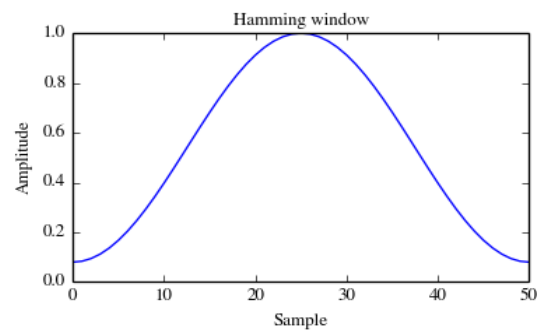
最後利用傅立葉反轉換，將頻域轉換回時域，鋼琴音會有明顯的減少，但是人的聲音也會有所受損。

B. 優化人聲

每段小音框做完傅立葉反轉換後連接起來音訊會有”啵啵啵”的雜音，是因為在頻域做過修改後轉回時域，音框的連接處並不連續。

因此將音框做 50 %的 Overlap，並以“Hamming Window”連結，使其聽起來更平滑。

漢明窗(Hamming Window)擁有將主瓣(main lobe)加寬的效果，並且將音框的 2 端壓低的效果，藉由套用 Hamming Window 可以有效削弱音框連結處的雜訊。



圖表 12 漢明窗示意圖

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

圖表 13 漢明窗方程式

重建純鋼琴



圖表 14 重建純鋼琴流程圖

(一)找出代表音

兩個音的起點間為一個音，每一個音皆是由無數個小音框所組成，每個小音框都有自己的頻率，而頻率不盡相同，所以採用投票的方式決定這一段音樂的頻率，以此重建音檔。

將每個小音框利用鋼琴的頻率公式，如圖表 15，將每個音框轉換到對應的鋼琴琴鍵上，鋼琴有 88 鍵，所以就有 88 個“投票選項”最後用票數最多的音作為代表。

$$semistone = 49 + 12 \times \log_2\left(\frac{Hz}{440}\right)$$

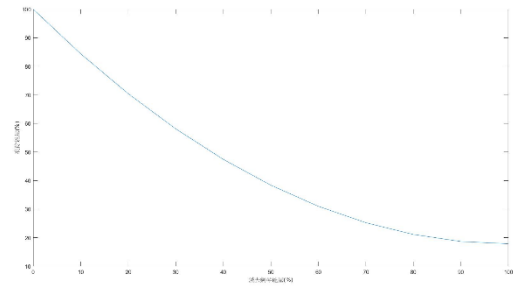
圖表 15 鋼琴的琴鍵頻率

此時我們有了音的起點以及其音高，再利用錄製的鋼琴音，藉由這些資訊重建出純鋼琴的音檔。

4. 主要成果與評估

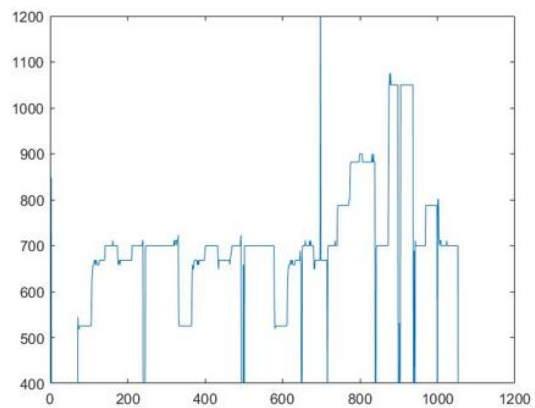
(一)分離出人聲

透過原音檔和壓低琴聲後的能量大小計算出去除鋼琴的比例，若將琴聲壓低過多會使人聲受損明顯，導致失真的情形嚴重；而壓低太少則鋼琴音明顯，分離出的效果差。經實驗測試，當去除比例達到 82% 時，效果最好，人聲的受損並不嚴重，而屬於鋼琴的頻率幾乎消失，只剩下敲擊琴鍵時的撞擊聲。

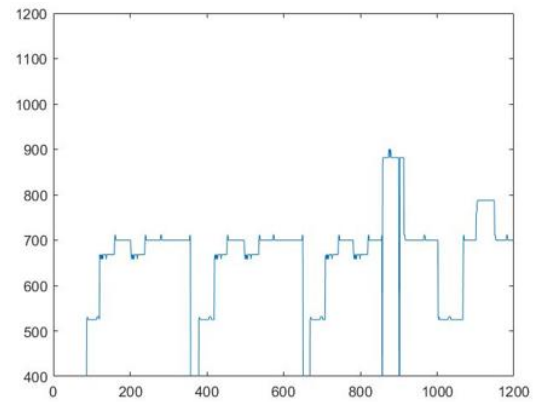


圖表 16 壓低琴聲的比例

(二)重建出鋼琴聲



圖表 17 正確的鋼琴頻率圖



圖表 18 計算出的鋼琴頻率圖

尋找音符起點的正確率 84%

錯誤的討論：

人的音量十分平穩並沒有明顯的衰減，有時甚至會隨著情緒起伏而逐漸增強，而鋼琴的音量則在彈下的瞬間即迅速衰減。二者相互影響下，有可能出現額外的高峰導致判斷錯誤。

計算音高的正確率 62%

錯誤的討論：

(1)人聲和鋼琴聲的交互影響下，會出現許多極端值，能夠用來”投票”的音所剩不多。

(2)人聲與樂器聲為 2 個頻率不同的 sin 波，兩個 sin 波相加，所計算出來的頻率會有所偏移，獲得並非鋼琴或人聲的頻率，而是介於兩者之間的數值。

(3)鋼琴本身的高音強度不夠。

(4)ACF 方法對於多種聲音混合的判斷較差，可能用傅立葉轉換去獲得頻率資訊較佳。

結論：

只要人聲和樂器聲的頻率有所分別，就可以過濾出人聲。

純鋼琴重建的成果好壞，和聲音起點、人聲息息相關。

5. 結語與展望

語音是人與人溝通最主要也最方便的方式，而語音的應用範圍也十分廣泛，雖然現在音訊處理並不是那麼熱門，但卻是非常重要的一環，各個項目都有機會使用到。我們希望透過這一次的專題，藉由實作的方式，將課程中所提到的理論以及方法，更加深入的了解體會，並且找出自己的盲點，驗證知識。

這次專題在重建純鋼琴的部分效果不如預期，未來希望能以機器學習的方式，精確的找出音符的起始點與音高，使鋼琴波形的判斷更準確，並且加入多種樂器的混合，使其更豐富。

6. 銘謝

我們要感謝我們的專題指導教授，在這一年的時間對我們耐心的指導與幫

助，從我們訂下了這個題目以後，給了我們開始的方向，並且告訴我們哪裡可以找到相關的資料，當我們遇到難題，百思不得其解時，也會為我們指點迷津。另外，在這一年中，教授除了對我們的專題進行指導外，也會關心我們的課業以及健康狀況，讓我們感到十分溫暖，非常榮幸專題能與這樣的老師一同研究。

7. 參考文獻

[1] Lawrence R. Rabiner and Ronald W. Schafer, 2011, "Theory and Applications of Digital Speech Processing"

[2] 台大張智星教授之課程網頁：
<http://mirlab.org/jang/books/audioSignalProcessing>

[3] Wolfram MathWorld - Fourier Transform
<http://mathworld.wolfram.com/FourierTransform.html>

