摘要

Abstract

在資訊通訊科技發達的21世紀裡，音樂是反映人類現實生活情感的一種藝術，藉由音樂我們能振奮人心並擊退抑鬱，由此可見音樂對於人類的日常是息息相關的，而音樂除了單純的聆聽欣賞外，不外乎也很多人是從事相關產業，在近幾年科技的發展已經生活化地運用在各個領域，當然音樂這方面也是不例外，所以我們想從一首歌曲中辨識出其樂器的聲音和音色，從音樂片段中找出與資料庫相符的音色特徵乃是一項及其具有挑戰性並具有延伸性的問題。因此本計畫的目的，就是開發一個可以輔助音樂創作者的音樂音色解析系統。

其功能包刮，

1. 音訊檔案解析成各別音色音訊。
2. 支援轉成MIDI檔案方便使用者導入iVST作運用。
3. 以深度學習技術實作層次分析來強化音色分離的精準度，蹦同時依分類的深度給予音色資料庫能解析更多不同樂器。

本計畫書利用統計模型來分析比對音框的異同，並使用一種特徵係數，『MFCC』以提升大數據的強健性，來結合CNN深度學習技術，經由不同「失真」程度的樣本進行交互比對，挑戰高容錯率的系統。

研究動機與問題

Motivation

由於大部分想要踏入音樂創作或是表演家的人並不見得都有絕對音感，對於剛接觸或是經驗不足的人，就會需要花很多錢及精力尋找鋼琴譜或是聘請一些比較有經驗的人幫忙寫譜，這種情況大多都成了一種無形的門檻，所以我們想藉由AI與訓練這種龐大的音色模型來有效分析個人化數據並讓輸出的數據可以讓我們做有效的利用，不僅可以減少時間上的成本，也可以使音樂產業帶來更多元的風格，而用雲端服務或是手機端服務可以讓使用者能隨時隨地使用音色辨識服務。

現今科技日新月異使得近來AI機器學習技術日漸普及與普遍衍生運用到各項領域中，音樂領域中也漸漸的可以藉由AI技術輔助使音樂創作者在領域中增加產值，從創作、辨識…等都有AI技術的蹤影，使得音樂各項方面有效的被運用並加以效益最大化，以至於讓音樂創作領域可以投入更多風格上的突破，也由於音樂領域曲風的多元性，讓音樂每個風格使用的樂器及音色越來越難辨識，所以我們將要嘗試把固定時間軸內的track做音色解析辨識(unmixing)。

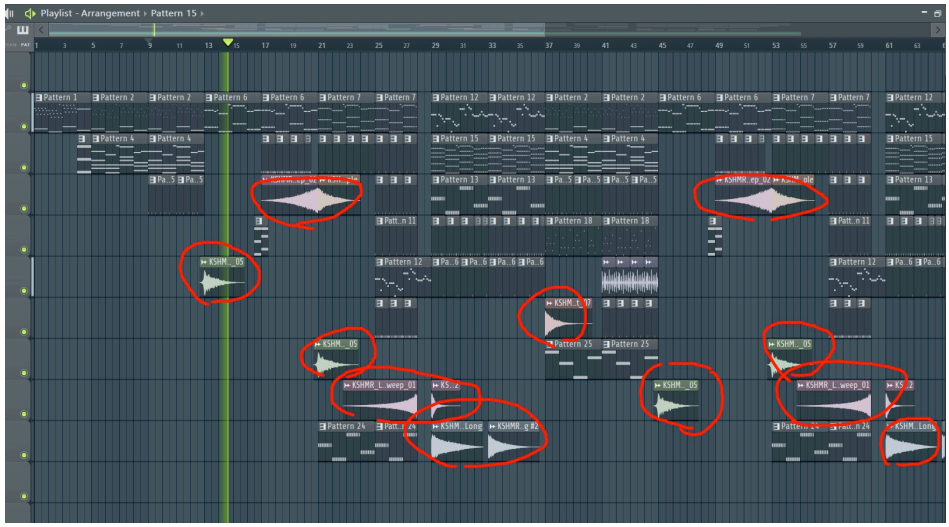
且由於目前現今開源的工具大多數都是使用MIDI檔來做音源輸入，然而依賴MIDI檔案作為訓練的主要資料會有幾點缺點及問題，在MIDI檔案的取得並不好蒐集，絕大部分都是沒有公開的，且多數公開的檔案有些風格有大量的即興，也有可能會遇到沒有標示音符強度或拍點的情況，為了讓這問題能更深入的探討，我們將擴大可支援的音樂檔案格式，此時就會有音訊品質的問題，想要保有乾淨、清脆的聲音品質是相當重要的，但是隨著音樂的格式眾多，音樂的品質也隨之不同，例如無損的音頻格式（CD，WAV，FLAC）、有損音頻格

式（MP3, WHA, AAC）等等, 再處理這些雜訊的方法中，以目前所參考得文獻當中，我們想試著使用梅爾頻譜倒頻譜係數，其最大的優點在於，於梅爾倒頻譜上的頻帶是均勻分布於梅爾刻度上的，也就是說，這樣的頻帶會較一般我們所看到、線性的倒頻譜表示方法，和非線性的音色系統更為接近。例如：我們在音訊壓縮的技術中，便常常使用梅爾倒頻譜來處理。但以實際效果為準，會試著將失真率降到最低。

音樂解析需要著重在音源分離，如下圖所示，在一段時域中，將其分離為多個頻域時，可能會出現一些音效部分（下圖紅圈表示），是否干擾分離的準確性，有待進行探討研究。

其次是特徵參數擷取部分，對樂器音色特徵的認識是相當低重要。音樂是特殊的，例如鋼琴，因為他們的聲音不是一致的。例如在鋼琴上每七个白键是一个循環，分別叫做CDEFGAB，也就是我們詳知的”Do Re Mi Fa So La Si”，過一個循環數標+1，又叫升八度，即C1-> C2-> C3-> C4。每升高八度，音頻高一倍，依此類推。要強調的是，不管其頻率如何，樂器音色是不變的。

而根據此特性，盡可能研究並塞選出合適的特徵參數，最大化其深度學習的效果，每當一段音源經過處理後，根據其特徵參數跟資



料庫進行相似度的分類，並完成其目的。

研究問題

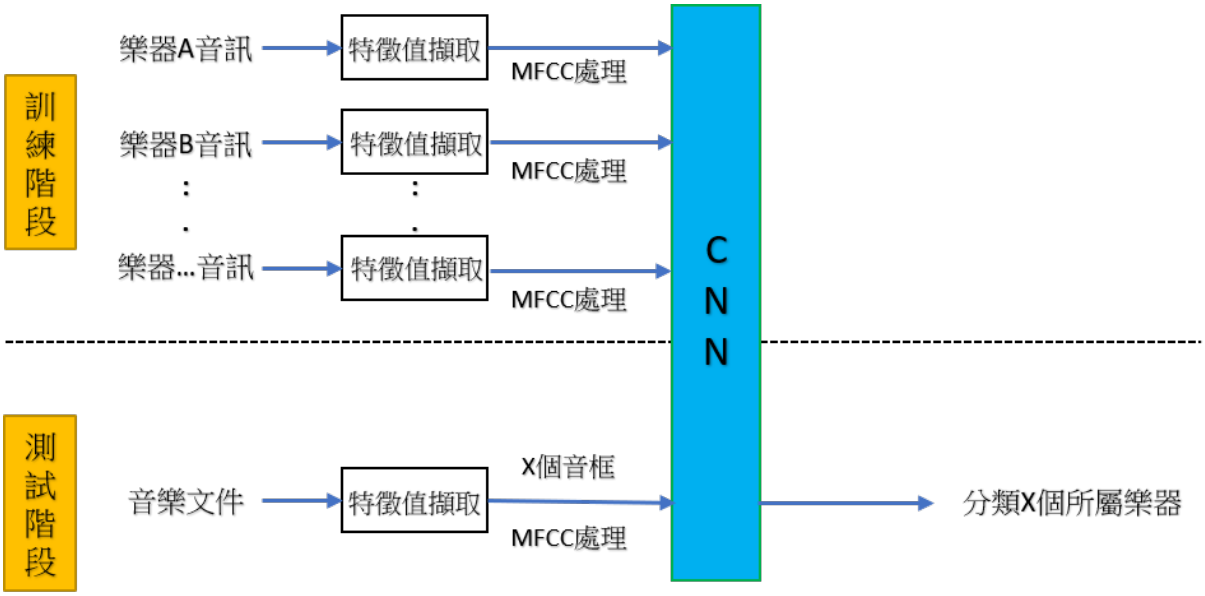
Problems

在音樂會中，我們總是能在眾多的聲音中，能夠辨識出我們熟知、想聽的聲音。但在過去中，機械想要做到這一點是相當困難的。幸運得是，近年來語音技術日益顯著，尤其是2019年，Google更是突破了機器無法擷取單獨音訊的門檻，創發了新技術名為—「Looking to Listen」人聲辨識[1]。此項突破，助長了我們心中名為希望的嫩芽。

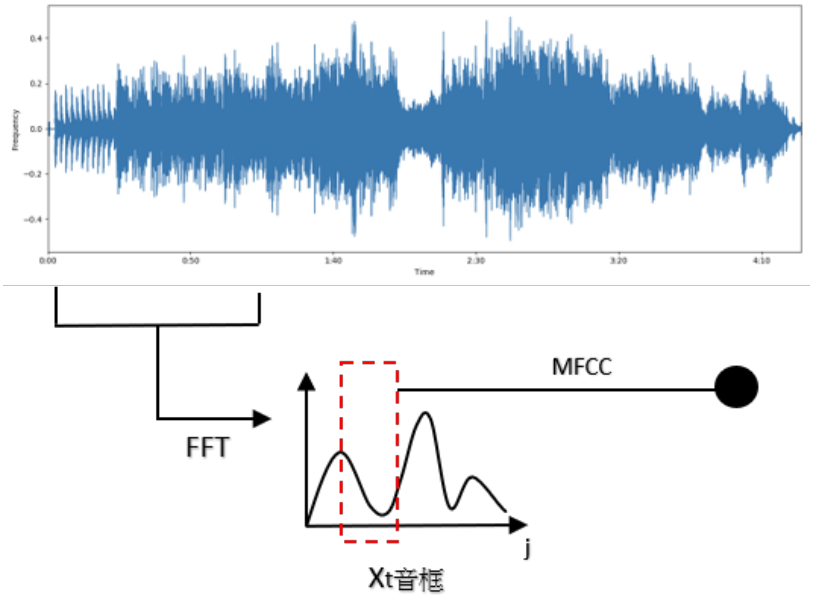
近年來說，我們所提到的音訊辨識，大都指的是人聲和背景音樂的分離[2]，使用一般的開源軟體如spleeter[3]和goldwave[4]，但是這樣並不能滿足所有的音樂創作者，所以如果將背景音樂進行更細部的區分，展示其樂器種類以及其樂理，將有更多人受惠。

就一般的音訊事件偵測系統，大多數是使用（高絲混合模型）GMM + MFCCs 架構來訓練，但是我們將使用CNN架構，主要是因為CNN具有以下特性(1)對音訊事件再輸入參數序列中的位置，具有時間與頻帶上的平移不變性，可容忍音訊事件在時間與頻譜上的變異 (2)自我訓練如何擷取最佳化的音訊事件特徵參數，因此可以避免需要專業知識，才能設計出合適的音訊參數工程問題。讓CNN能透過直接輸入頻譜參數，讓CNN自動去探索，除了MCFFs外，還有哪些特徵參數對音訊事件效能最好，有待進行探討。

如圖（一）所示為本音樂解析系統架構，藉由樂器辨識(instrument recognition)及音源分離(source separation)來做為主要步驟，其中音源分離的主要的功能在於分離audio音訊中的各種樂器，再取各段音訊的特徵值來做分類(classification)，其流程在以下小節做詳細說明; 而圖（二）所說明的是，一段音訊經過處理時，所呈現出來的概略圖。



圖（一）架構圖



圖（二）訊號流程圖

研究方法與步驟

Method and Step

以下是本計畫書步驟流程（參照圖（一））：

1. 音軌特徵擷取：

在音樂訊號經過短時間快速傅立葉轉換（Fast Fourier Transform, FFT），將會多個音框序列（第一軸式頻率，第二軸是時間），將會把各段分離後的譜圖做重點特徵擷取。並令Xt為第t個音框訊號,而在其音框中分j段絕對震幅頻譜序。

2. 梅爾倒頻譜係數（Mel-Frequency Cepstral Coefficients, MFCC）:

將$X_{t,j}$頻譜映射（mapping）至梅爾刻度上，利用三角窗函數（triangular overlapping window），並能求得每個梅爾刻度輸出得到的對數能量$S_{log[j]}$, 在對數能量壓縮後經由離散餘弦轉換（discrete cosine transform , DCT）取得目的是希望將訊號轉換為倒頻譜係數。其主要用意在於減少維度，有助於在儲存共變異矩陣時資料的縮減，增加辨識率。

$$c_i = \sum_{m=1}^{M} s_{\log}[m] \cos\left[\frac{\pi i\,(m-0.5)}{M}\right]$$

j

j

其中C_i為MFCC特徵向量，M為濾波器數量。

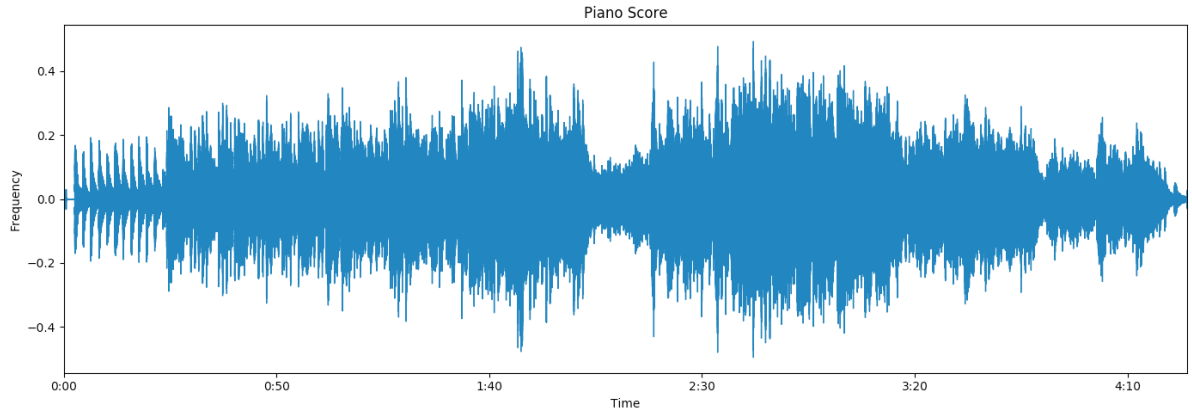
3. 卷積神經網路學習（CNN）：

CNN學習是一種對於資料進行表徵學習的一種演算法，我們希望藉由學習使系統可以分辨各種樂器頻率的不同，我們會先設定多種樂器作為標準，取出音樂中的特徵後，再以學習的方式使系統找出最正確的答案。

本階段含以下兩個部分：訓練(Train)與測試(Test)，由於樣本數重大，預估分別為80%及20%

訓練階段中：任務在於建立每一首音源的音域，再將其拆分為多個頻域，此時會有兩階段要執行，一是特徵參數擷取，將聲紋轉換為有利於分析的特徵向量，二是統計模型的建立，利於區分類別。

測試階段中：任務在於分辨輸入的一段音源與資料庫中的樂器頻域進行分類，此階段中，也包含特徵擷取，將其轉換特徵向量，計算其與其他範本的相似率，貼上其類別標籤。



製作出一套可在雲端與手機上使用的音樂解析系統，其可以精準的辨識各種音色、音階等相關數據，讓音樂人能做有效的利用，減少不必要的時間，為音樂產業帶來更多元的發展，也讓沒有音樂天分的人能享受到玩音樂的樂趣。

系統透過梅爾倒頻譜係數將音樂分類與調整，再由深度學習對資料做演算，比對各種樂器的頻率，將音樂的音色、音階與所使用何種樂器清楚表達出來。

製作音樂解析系統時，除了學習程式撰寫，完成一套可以分析出各項音樂數據的軟體，還包括了與組員間的互相合作、研究切磋，以及學習獨立思考、解決問題的能力，希望能從中學到更多相關知識與技能。

預期結果

Final

[1]https://buzzorange.com/techorange/2018/04/17/google-ai-looking-to-listen/Google AI濾人環境音
[2]https://arxiv.org/abs/1804.03619
Looking to Listen技術
[3]http://www.goldwave.com/release.php
Goldwave是一款音樂編輯軟體 用於改變音高 音軌 錄製
[4]https://github.com/deezer/spleeter
Spleeter軟體
[5] http://www.speech.cs.cmu.edu/15-492/slides/03_mfcc.pdf

參考文獻

Final

computing Mel Frequency Cepstral Coefficients (MFCC)