The 2014 Conference on Computational Linguistics and Speech Processing
ROCLING 2014, pp. 42-57


# 利用核依賴估計來進行多軌自動混音

## Automatic Multi-track Mixing by Kernel Dependency Estimation

吳宗庭　Tsung Ting Wu
國立中央大學資訊工程研究所
Department of Computer Science and Information Engineering
National Central University
101522015@cc.ncu.edu.tw

張嘉惠　Chia-Hui Chang
國立中央大學資訊工程研究所
Department of Computer Science and Information Engineering
National Central University
chia@csie.ncu.edu.tw

### 摘要

近年來由於數位音樂的蓬勃發展，錄音器材越來越普及。使得非混音專業人士也能利用錄音界面(Audio Interface)錄製出不錯的成品；但是一旦錄製了多軌(Multi-Track Recording)就會面臨到混音(Mixing)的問題，即需要把多軌的聲音混合在同一個軌中。混音牽扯到許多音響及聲學心理學的相關技術與知識，非專業人士要混出尚可的成品有一定的難度，所以我們提出了自動多軌混音系統(Automatic Multi-track Mixing System)，希望藉由監督式學習的方式學習各軌間混音參數的調配，產生每首的基礎混音(Basic mix-down)來幫助非混音專業人士也能混出不錯的成品(Mix-down)。由於混音參數取得不易，我們會先藉由分軌及混音好的關係估計出各個混音參數，接著利用其參數進行混音模型(Model)的建立。在參數學習(Parameter Learning)方面由於每軌的混音參數是有依賴關係的(Dependency)，我們採用了核依賴估計(Kernel Dependency Estimation)[1]的參數學習(Parameter Learning)方式來預測每軌的混音參數。

### Abstract

Due to the revolution of digital music, people can create recordings in a home studio with cheaper gear. However multi-track recordings need to be mixed to combine them into one or more channels. The question is that mixing requires background knowledge in sound engineering and psychoacoustics. It is difficult to get good mixdown for non-specialist in sound engineer. In this paper, we use supervised learning method for automatically mixing multi-track recording into coherent and well-balanced piece. Due to lack of mixing parameters, first we estimate the weight of mixing parameters by using the relation between raw multi-track and mixdown. Given the mixing parameters for any music genre, we use kernel decency estimation method to create our mixing model. The experiment show KDE is

able to make a more satisfactory estimation than treating each parameter independently.
關鍵詞：核依賴估計，音樂資訊檢索，音樂製作，混音
Keywords: Kernel Dependency Estimation, Music IR, Music Production, Mixing.

# 一、 緒論

在音樂的製作(Music Production)上大致分為三個階段，創作編曲(Pre-Production)、聲音錄製(Production)、後製(Post-Production)。其中後製又分為混音(Mixing)及母帶後製(Mastering)兩部分; 混音在音樂製作上是個非常重要的過程; 其主要的工作是要把先前錄製好的多軌(Multi-Track)的聲音，如人聲、吉他、爵士鼓等聲軌混合進同一個立體聲軌(stereo channel)或單聲軌(Mono Channel)中。

近年來由於數位音樂的蓬勃發展，錄音器材越來越普及。使得非混音專業人士也能利用錄音界面(Audio Interface)錄製出不錯的成品; 但是一旦錄製了多軌(Multi-Track Recording)就會面臨到混音(Mixing)的問題，即需要把多軌的聲音混合在同一個軌中; 混音牽扯到許多音響及聲學心理學的相關技術與知識，非專業人士要混出尚可的成品有一定的難度，混出來的結果往往會照成整首歌聆聽的清晰度降低、聲音不扎實、音量落差太大、空間感不夠、聲音雜亂等問題; 而且混音的處理方式基本上會隨著樂器、音樂類型而有所不同，不同的音樂類型會有不同的混音風格(Mixing Style)，這更加增加了一般非專業人士學習混音的難度。所以要如何藉由電腦來幫助混音便是本篇論文的目標。接下來將會詳細介紹混音的相關背景知識。

## 1.1 混音(Mixing)

混音在音樂製作上是非常重要的過程，不同的混音方式在最後的成品上會有截然不同的模樣。混音的好壞會影響整首歌的表現，好的混音可以掩蓋瑕疵、放大優點，提升整體的質感。在混音的過程中，混音師(Mixing Engineer)會依照各音軌/樂器間的頻率(Frequency) 、響度(loudness) 、音色、音場定位(Panoramic Position) 、空間感等聲音元素加以調配，以讓每個音軌(track)/樂器最佳化，讓每個音軌在最後混在一起時一樣能保持清晰，保有層次，使得音樂呈現更生動、更動聽。

在開始混音前，混音師會先作混音規劃(Mixing Design)，規劃整首歌的音像(Sound Image)，決定每個音軌在整首歌裡的定位以及其重要性。如下圖 1 為一首爵士樂的混音規劃示意圖，我們可以發現在這首歌中人聲(Vocal)設計在音像的正中間，吉他(Guitar)分別落在人聲的左右; 音量方面主吉他(Lead Guitar)略大於人聲等等，混音的過程中主要就是調配這些音量(Volume)、等化(Equalization)、擺位(Pan)來達成我們的混音規劃讓整體更加和諧，讓個樂器融入其中。

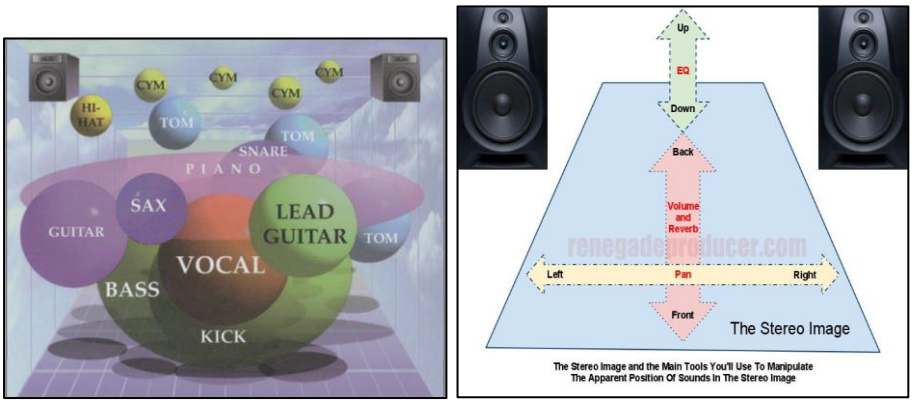


圖 1 混音設計示意圖：人聲(Vocal)居中

圖片來源：The Art of Mixing – David Gibson

## 1.2 研究動機(Motivation)

在自動混音(Automatic Mixing)的研究中，大多都是利用聲音特徵間的關係估計以及預測其混音的參數，所使用的聲音長度大多為一首歌中 30 秒的片段，最後所混音出來的成品當然也會相似於這 30 秒的片段。但是混音實際上是會因為橋段的不同而有不同的混音方式的，例如在主歌(Verse)與副歌(Chorus)的混音方式會是不一樣的，後者其橋段通常為歌曲中激昂的部份，配器使用會前者多，在音量或頻率上比例會有所不同。

再者歌曲音樂類型也會影響混音的方式，同樣是爵士鼓在流行歌與爵士樂中音色與其占有的比例也會有所不同，同樣的樂器在不同的音樂類型中會有不同的音色以及角色，例如，在爵士樂中鼓手也可以是樂曲中的主角，在流行歌中主角往往是人聲或是電吉他。所以若利用 30 秒片段建立的模型來套用在整首歌曲的混音將會造成整首歌較平淡無味; 在 Jeffrey Scott et al.[2]的訪問中也有提到，混音其實是個別的(Case By Case)，混音師基本上都會依照各音軌的聲響、音樂類型不同而有不同的處理方式，再者音樂感受是非常主觀的，不同的人會有不同的偏好，所以比較難去訂出一個通則(General Rule)來進行混音。

一個混音模型包括音量、頻率(Equalization)、樂器擺位(Panning)等多個參數，大多數的研究都是獨立的去預測各軌的混音參數，如多線性迴歸，利用各軌特徵間的關係去建立各軌的迴歸模型。但實際上每軌間的混音參數是有依賴關係的(Dependency)，舉例來說有一軌的音量上升勢必就會有一軌的音量下降，這樣整首歌的音量才不會忽大忽小，若是對各軌獨立去建立其模型的話，最後的成品將會喪失其依賴關係。所以在參數預測的方法選用上我們認為需要考慮到依賴性的問題。

由以上三點，本篇論文採用對不同的音樂類型不同的橋段，利用和依賴估計(Kernel Dependency Estimation)[1]的方法來建立最後的模型。首先我們會先需要使用者先提供一些該歌曲的訊息，例如音樂形態、橋段、分軌的樂器標籤等等，接著在對個別的橋段套用該音樂類型的核依賴模型預測出混音參數後即完成混音。(圖 2)

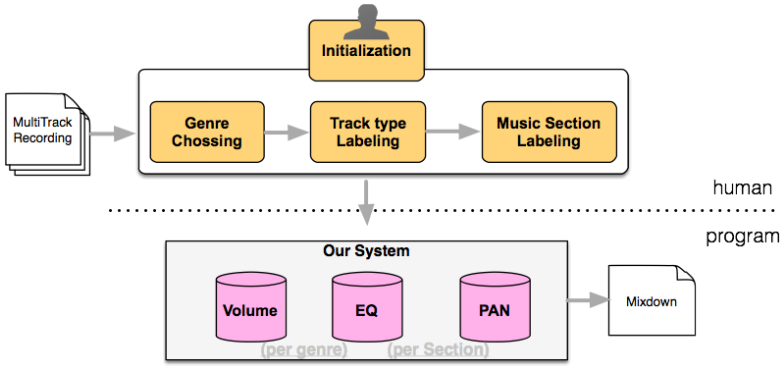


圖 2 系統使用概念圖

本篇論文的架構如下，第二章將會討論自動混音的相關研究，第三章則會介紹本篇論文的研究方法及所使用的資料集，第四章為實驗，我們會對我們的混音模型去做交叉驗(Cross validation) 來評估模型的正確性以及依賴性的功用，第五章為結論以及未來工作。

# 二、 相關研究

在介紹混音參數預測模型之前，我們必須設法取得一些歌曲的混音參數以作為訓練資料，不過混音參數在實際上是難以取得的，由於混音師使用的軟體不同，不同的器材/軟體會有不同的設定、基準及刻度，這讓得到混音參數這一類的資訊變得非常困難，且混音師在混音時也鮮少會把參數記錄下來。所以在多軌混音(Multitrack Mixing)的相關研究中大部份的研究著重於如何估計出混音參數，包括：音量(Volume)、頻率(Frequency)、動態(Dynamic)等等。另一部份的研究著重在如何建立混音參數模型。

## 2.1. 混音基本元素(Basic Factor of Mixing)

在混音的過程中音量平衡是件很重要的事，混音師(Mixing Engineer)會調整每個音軌間彼此之間的音量(圖 1)，決定各軌在這首歌中的音量比例，即是決定各軌在音像(Image)的前後順序; 若其中一音軌的音量比其他音軌還要大很多的話，整首歌將會聽起來頭重腳輕。

在音色修正的過程中，混音師(Mixing Engineer)會用到等化器(Equalizer, EQ)的工具來去對每個音軌的頻率做修正調整。例如我們發現吉他的音色跟其他的樂器相比太亮太尖銳，以至於無法融入這首歌中，我們就可以用等化器去對吉他的高頻部分做衰減(cut)。頻率間的平衡在混音過程中也是重要的過程之一。

樂器擺位是將錄製好的聲音訊號放置於新的雙聲道或多聲道的聲場(Sound Field)。由於我們一般音響設備的環境基本上是以雙聲道為主(stereo)，雙聲道的混音可以在聆聽上增加平面的聽感，而不是只有一點;所以我們在混音的時候會決定各樂器的擺位，看是要擺在中間還是擺在靠近左聲道/左喇叭還是右聲道/右喇叭等等來增加整體的空間感，而非全部的樂器都擠在一起。在實作上即是分別調整左聲道與右聲道的音量，如下(1)式

$$Left\_output = cos(p)*input$$
$$Right\_output = sin(p)*input \quad (1)$$

其中 *p* 為偏離中央點的角度，*Left_output, Right_output* 分別為左聲道右聲道的輸出。

## 2.2. 混音參數估計(Mixing Parameter Estimation)

如前一節研究動機所提到，由於混音參數難以取得，我們需要利用原始分軌和混音成品(Mixdown)間的關係來估計出每首歌的混音參數。在 Jeffrey Scott et al. [2]的研究中，他們採用了訪問(interview)的方式，訪問了線上的混音師(Mixing Engineer)了解他們如何混音、如何處理聲音，利用訪問後得到的一些通則來當作他們最後混音模型建立的依據。在大部分自動混音(Automatic Mixing)的研究中[3, 4]，在混音參數估計上大多是假設其分軌和混音成品是聲音特徵(Sound Feature)的線性組合關係，如圖 3 所示 $X_i$ 為第 i 聲軌的特徵向量，$\beta_i$ 為第 i 軌之混音參數權重，y 為混音成品的特徵向量。利用最小平方法的方式(Least Square Method)最小化 y 至 col(X)平面的距離來做各音軌混音參數權重估計。

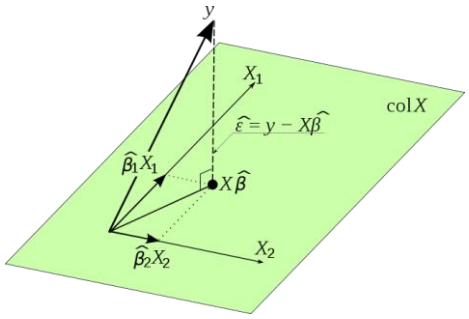


圖 3 最小平方法

由於採用線性組合的假設，每軌間彼此要是線性獨立(Linearly independent)的，這會影響最後聲音特徵的選用以及音軌的選擇，選擇含有較多串音的音軌(例如錄製 overhead 時會連大鼓小鼓等其他鼓組的聲音一併錄進)，在實務上因為包含多個樂器的聲音，使得音軌間彼此並非獨立會造成其估計結果會有誤差。

## 2.3. 混音參數預測

參數預測大多是採用機器學習(Machine Learning)上參數預測(Parameter Prediction)的方法，如[3] [5]採用了多線性迴歸(Multiple Linear Regression)方法，利用大量的 sample 建立最後的回歸模型，利用最佳化的方法去最小化分軌及成品間的尤拉距離(Euclidean Distance); 在[3]中同時利用了 Linear Dynamic System 的方法，將混音參數視為一個潛藏的狀態(Latent State)，利用聲音特徵當作動態系統的輸出。另外在[6]採用了不一樣的方式，利用"例子混音"(Mixing by Example)。其概念類似理髮師的概念，混音前提供例子供使用者選擇，最後藉由複製該例子的混音設計(Copy Mixing Design)的方式來達成混音。

## 三、 研究方法

本系統主要流程如下圖 4 所示，我們會先把原始的分軌錄音檔依照使用者提供該音樂的資訊做前處理，接下來做聲音特徵的擷取(Feature Extraction)以便之後模型(Model)的訓練及測試。由於不同的類型、不同橋段的音樂會有不同的混音方式，在模型建立時我們會特別依照不同的音樂類型建立個別的混音參數模型，在依不同的橋段去建立模型，如圖 5 我們會對 ROCK 的的音樂類型建立 Intro、Verse、Chorus，POP 音樂類型也建立其三個橋段的模型，以此類推。訓練(Training)的部分有兩大步驟; 第一步驟是混音參數估計(Parameter Estimation)，由於混音參數難以取得，原始訓練資料中也無此資訊，我們會先利用原始分軌錄音和混音成品(Mixdown)做最小平方法估計(Least Square Estimation)，藉此來估計出訓練資料中每首歌的混音參數的權重(Weight of Mixing Parameter) 。

第二步驟是核依賴估計(Kernel Dependency Estimation)的模型建立，我們會利用每首歌的混音參數權重及特徵向量來當作訓練核依賴估計(Kernel Dependency Estimation)的依據，訓練好的模型將會用來預測各個混音參數的權重。最後依照模型預測出的權重進行混音。

在接下來的章節我們會詳細介紹各步驟的做法，在章節 3.1 會先對我們所使用的資料集(Data Set)做介紹以及前處理的部分。章節 3.2 會介紹我們如何利用最小平方法來估計各個混音參數及為何需要做混音參數估計。章節 3.3 我們會介紹核依賴估計(下面簡稱 KDE)的核心概念以及在本篇論文的問題中該如何設計。

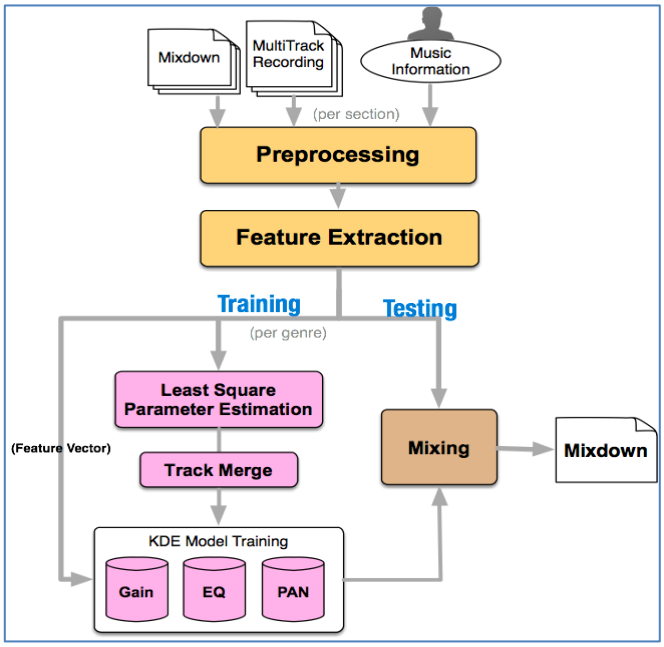


圖 4 系統架構圖

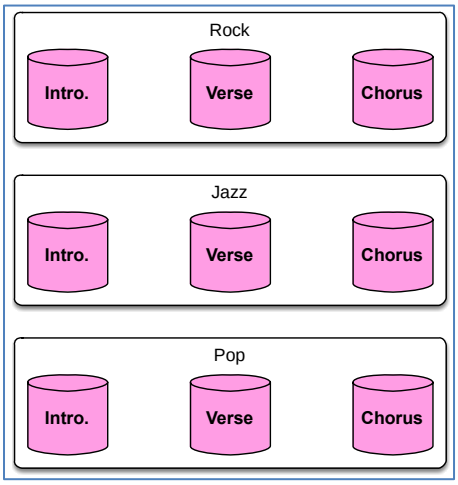


圖 5 混音參數模型

## 3.1 資料集 (DATASET)

我們使用的資料集(Data Set)是來自國外一本關於混音的專門書籍”Mixing Secrets for Small Studio”[7]，此書有提供多首原始分軌檔案給讀者用於混音練習用，其含括的音樂類型搖滾、爵士、鄉村等多種音樂類型，如下表 1 所示，此書將相似的音樂類型分成四大類。此資料集較特別的點在於提供的聲音檔案長度是整首歌(Full Multitrack)，一般以往音樂資訊探勘(Music Information Retrieval)研究所使用的資料集大多是 20~30 sec 的長度，鮮少有提供整首歌的資料集。此特點有利於幫助我們對於不同的音樂橋段(Music Section)去建立不同的模型，來讓最後的模型能更適用於實際的情況，此資料集所提供的混音成品(Mixdown)一樣也是整首歌的長度。檔案格式為無損 WAV 檔(uncompressed WAV files, 24bit and 44.1 kHz sample rate)。

表 1 音樂類型統計表

| Genre | # of Song |
|---|---|
| **Alt Rock / Blues / Country Rock / Indie / Funk / Reggae** | 7 |
| **Rock / Punk / Metal** | 17 |
| **Pop / Singer-Songwriter** | 10 |
| **Acoustic / Jazz / Country / Orchestral** | 8 |
| **Total** | 42 |

資料集前處理的部分，由於此資料集涵蓋的音樂類型包括 Alt Rock/Blues、Rock/Punk、Pop/Singer-Songwriter、Acoustic/Jazz/Country 等，每個音樂類型所使用的配器會略有不同，例如在爵士樂中管樂器的使用比例會比搖滾樂來的高。這樣會造成之後無法將此資料集套用至我們的模型中。所以為了解決此問題我們統計了各音樂類型所使用的配器，決定出每個類型基本音軌(Basic Track)如表 2，我們定義了 12 種音軌的形態。前處理的部分會先對分軌做音樂格式上的轉換，待每首歌的混音參數權重估計出來後會做基本軌的合併(Track Merge)，將同一類型的音軌依照其權重先合併成該類型基本音軌，將同一類型的檔案先合併成一個;例如在某一首歌中吉他錄了兩把，我們會先將這兩把的錄音合併成一個，以方便之後訓練及測試該類型的音樂。

表 2 基本軌表

| 12 Basic Track Type | | | | | |
|---|---|---|---|---|---|
| (1)Kick | (2)Snare | (3)Hihat | (4)TOM | (5)DrumRoom | (6)OVERHEAD |
| (7)PERCUSSION | (8)BASS | (9)Electric Guitar | (10)Acoustic Guitar | (11)LeadVox. | (12)BackVox |

## 3.2 混音參數估計 (PARAMETER ESTIMATION)

我們的系統最終的目的是要希望藉由訓練資料(Training Data)來對每一種混音參數建立一個模型，藉由每一軌的特徵來預測其參數的值。所以訓練的過程中勢必需要原始

的分軌檔案及最後各混音的參數來進行監督式學習(Supervised Learning)。但是實際上混音參數的資訊是非常難取得的。由於混音師使用的軟體不同，不同的器材/軟體會有不同的設定、基準及刻度，而且混音師在混音時也鮮少會把參數記錄下來，這讓得到混音參數這一類的資訊的取得變得非常困難。為了之後的監督式學習，我們必須先估計出資料集中每首歌的混音參數，利用原始分軌檔案(Raw Multi-track)及最後混音成品(Mixdown)來估計出每首歌的混音參數，針對每一首歌求得其混音參數當作之後監督式學習(Supervised Learning)的依據。

為了從原始分軌檔案估計出其混音參數，我們假設原始分軌與最後混音的成品(Final Mix)的關係是一個線性的組合(Linear Combination)，如下(2)式為一首歌的線性組合關係。

$$\alpha_1 U_1+\alpha_2 U_2 +.....+\alpha_k U_k=V \qquad (2)$$

$\alpha_i$ 為第 $i$ 軌的混音參數權重，$U_i = [u_{1i}, u_{2i}, \dots, u_{Ni}]^T$為第 $i$ 軌特徵向量，$V$為最後混音結果的特徵向量(Feature Vector)，每一軌抽取 $N$ 個 frames 做為其代表。其中在不同的混音參數會用不同的聲音特徵，例如在音量參數方面會採用聲音的方均根來當作衡量的依據，頻率參數方面會採用聲音的頻譜(Spectrum)等等。利用此線性組合的關係，我們可以利用最小平方法(Least Square Method)來估計出混音參數 α 的數值，最小平方法(Least Square Method)是以觀測值 U 與預測值 $\hat{U}$ 之差的平方和作為最佳化的目標函數(Objective Function)。以音量參數為例，令 $u_{Nk}$ 為第 $k$ 軌第 $n$ 個音框(Frame)的方均根值(RMS)，每個音框長度約為 20 毫秒，則(1)即可表示為

$$\begin{bmatrix} u_{11} & u_{12} & u_{13} & \cdots & u_{1k} \\ u_{21} & u_{22} & \cdots & \cdots & u_{2k} \\ \vdots & \vdots & & & \vdots \\ \vdots & \vdots & & & \vdots \\ u_{N1} & \cdots & \cdots & \cdots & u_{Nk} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \vdots \\ \alpha_k \end{bmatrix} \approx \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ \vdots \\ v_N \end{bmatrix} \qquad (3)$$

在混音時由於混音的參數眾多，本篇論文只討論了其中三個最重要參數：Volume、Frequency、Pan，在估計不同的參數時我們會使用不同的聲音特徵。接下來會對各參數所用的特徵來做介紹

## 音量 (VOLUME)

音量參數也被稱作增益(Gain)，主要在控制每個音軌間的音量使得整體音量達成平衡。本篇論文使用了方均根(Root Mean Square)的方式去測量同一個音框(Frame)中各音軌的聲壓(Sound Pressure Level)，以此作為音量參數的特徵向量，寫成矩陣形式如(4)式。$G_k$ 為第 k 軌的權重，N 為音框的長度。

$$\begin{bmatrix} RMS_{11} & RMS_{12} & RMS_{13} & \cdots & RMS_{1k} \\ RMS_{21} & RMS_{22} & \cdots & \cdots & RMS_{2k} \\ \vdots & \vdots & & & \vdots \\ \vdots & \vdots & & & \vdots \\ RMS_{N1} & \cdots & \cdots & \cdots & RMS_{Nk} \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ \vdots \\ g_k \end{bmatrix} \approx \begin{bmatrix} V_{RMS_1} \\ V_{RMS_2} \\ \vdots \\ \vdots \\ V_{RMS_N} \end{bmatrix} \qquad (4)$$

### 頻率(Frequency)

頻率參數即是控制整首歌中各音軌在頻率上的平衡。在混音過程中為了修正音軌的音色或頻率時會用到等化器來幫助我們對聲音的頻率作調整，也就是說頻率參數即是等化器(Equalizer)參數。在設計等化器時會先將整個頻譜切成多段(Multi-band)，如切成三塊的話即是高頻、中頻、低頻。接著選出各頻段的中心頻率(Center Frequency)及頻寬(bandwidth)後即完成設計。本篇論文在頻率參數方面一樣採用多頻段等化方式(Multi-band Equalization)來模擬實際等化器的操作，即是把頻率參數估計的問題切成了多個聲音參數的子問題。如下圖 6 所示。

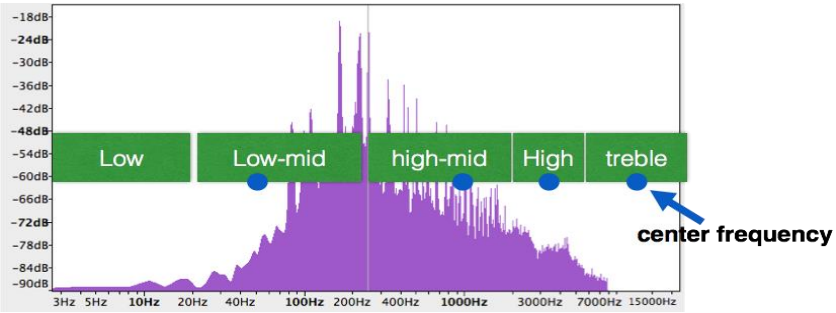


圖 6 多頻段頻譜

在進行頻率參數估計時，我們會先用快速傅立葉轉換(Fast Fourier Transform)先得到各音軌的頻譜(Spectrum)，接著把各音軌的頻譜依照我們預先分段的頻率分別去解迴歸問題(Regression Problem)，以估計出各頻段在各軌間的平衡參數。如下(5)式

$$\begin{aligned}
&\alpha_{Treble_1}U_{Treble_1} + \alpha_{Treble_2}U_{Treble_2} + \cdots + \alpha_{Treble_k}U_{Treble_k} = V_{Treble} \\
&\alpha_{High_1}U_{High_1} + \alpha_{High_2}U_{High_2} + \cdots + \alpha_{High_k}U_{High_k} = V_{High} \\
&\alpha_{H\text{-}mid_1}U_{H\text{-}mid_1} + \alpha_{H\text{-}mid_2}U_{H\text{-}mid_2} + \cdots + \alpha_{H\text{-}mid_k}U_{H\text{-}mid_k} = V_{H\text{-}mid} \\
&\alpha_{L\text{-}mid_1}U_{L\text{-}mid_1} + \alpha_{L\text{-}mid_2}U_{L\text{-}mid_2} + \cdots + \alpha_{L\text{-}mid_k}U_{L\text{-}mid_k} = V_{L\text{-}mid} \\
&\alpha_{Low_1}U_{Low_1} + \alpha_{Low_2}U_{Low_2} + \cdots + \alpha_{Low_k}U_{Low_k} = V_{Low}
\end{aligned} \tag{5}$$

### 樂器擺位(PANNING)

樂器擺位即是決定該音軌在左聲道及右聲道之音量比例，如式(1)。在本篇論文我們將擺位參數的問題轉化成前一章節音量參數估計的問題，即左聲道做一次音量參數估計，右聲道做一次音量參數估計，這樣即可決定該音軌在左右聲道之比例，如下式(6)、(7)所示。

$$\alpha_1 u_{L1} + \alpha_2 u_{L2} + \ldots\ldots + \alpha_k u_{Lk} = V \tag{6}$$

$$\alpha_1 u_{R1} + \alpha_2 u_{R2} + \ldots\ldots + \alpha_k u_{RK} = V \tag{7}$$

## 3.3 核依賴估計模型建立(Kernel Dependency Estimation Model)

估計完資料集中每首歌的三種混音參數(Volume、EQ、Panning)後，我們有了每首歌的特徵向量(X)，及每首歌的混音參數(Y)，即可藉由學習出 X, Y 間的關係建立我們最後的混音模型。在模型建立的部分我們會依照不同的音樂類型，不同的音樂橋段建立一組模型，其中一組模型包括了音量模型兩個(左聲道、右聲道)，頻率模型五個(5

sub-band)。利用此模型建立的方法讓最後的成品更能應用在實際的歌曲中，訓練模型的樣本數我們自各訓練歌曲中隨機抽取 3000 個樣本來當作我們的訓練資料。如圖 7，$u_{nk}$ 為該模型的第 n 個樣本的第 k 軌的聲音特徵值，$\alpha_{nk}$ 為第 n 個樣本之第K軌的混音參數權重值;K 為分軌的個數，在本篇論文為 12。

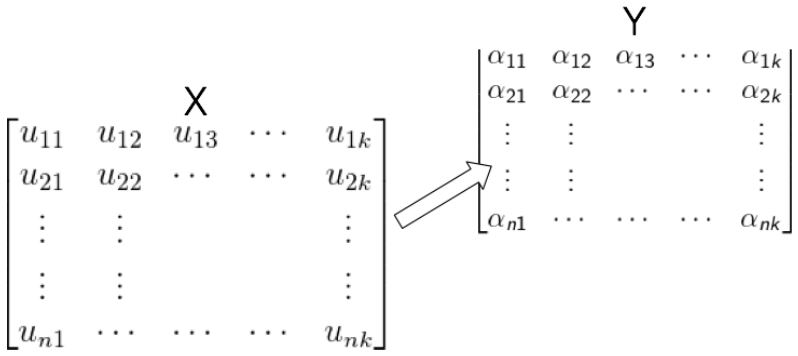


圖 7 混音參數預測示意圖

這個問題基本上是可以利用常見的參數預測(Parameter Prediction)的方法分別去求解，例如多線性迴歸(Multiple Linear Regression)、動態線性系統(Linear Dynamic System); 但是由於在混音時每一軌間的參數間是有互相影響的，其中一軌的音量變大後勢必其他軌會變小聲些，彼此是依賴關係的(dependent)。若用各別訓練模型來預測我們的混音參數的話，我們將會得到 k 個獨立的迴歸模型，這對於混音結果可能會產生較不和諧的影響。所以本篇論文採用了核依賴估計的方式來解決上述的問題，用核依賴估計訓練出我們最後的模型。接下來將會先介紹核依賴估計的基本精神。

### 3.3.1 核依賴估計(Kernel Dependency Estimation)

核依賴估計(以下簡稱 KDE)是一種用於尋找輸入 X 與輸出 Y 間依賴關係的學習架構，KDE 的流程如下圖 8 所示，步驟主要分為三個步驟：

1. 投影(Projection)：對輸出 Y 做主成份分析(Principal Component Analysis)，將輸出 Y 也就是參數向量 $y_i \in R^k$ 投影至 m 個主成份(Principal Component)所形成的空間上，即對輸出 Y 做降維的動作，投影至較低的維度下形成 Y’。

2. 學習(Learning The Map)：對每個基礎原件(Principal Component)j，$1 \le j \le m$ ，我們會學習一個對應函數(Mapping Function)$r_j(X)$，對應函數 r 會將會把 X 對應至 Y’第 j 個基礎原件，即是在較低的空間上去解 m 個迴歸問題。

3. 預測(Prediction)：對於一個新的輸入 x’，我們可以利用先前學習好的對應函數求出 y’’ = [ $r_1(x')\ r_2(x')\ldots.r_m(x')$]，$y'' \in R^m$。最後再將 Y’’投影回原本的空間上，即求出 X’所對應的Y值。

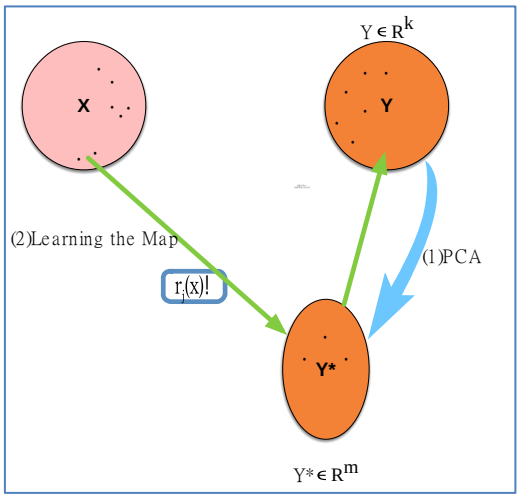


圖 8 KDE 示意圖

用於本篇論文的主題來說，我們的X就是我們每首分軌的聲音特徵向量，Y 就是我們原先估計出來的混音參數 $Y \in R^k$，在 KDE 的過程中我們會先將 Y 降維至 m 維的 PCA 空間上，接著在 m 維的空間上我們去學習 m 個基礎原件的對應函數。在對應函數方面我們使用了 ridge regression 的方法來解決 m 個迴歸問題。在做預測時，新的一首歌分軌特徵向量 x'，會先利用先前學習的 m 個對應函數 $r_j$ 求出 y''， $y'' \in R^m$，接著再將 y'' 投影回原本的 k 維空間上即求出各軌混音參數權重。

## 四、 實驗

實驗分為三個部分作討論，第一部分的實驗是評估由核依賴估計所建立的模型的正確性; 第二部分是評估 KDE 依賴性(Dependency)的效果，比較不同的 m 值對模型正確性的影響程度; 第三部分是評估不同的音樂類型混音方式的差異性。實驗評估的方式是利用一次挑一個交叉驗證對同一音樂類型的歌做均方誤差(Mean Square Error)評估。均方誤差的計算方式如下：

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{\alpha}_i - \alpha_i)^2 \qquad (8)$$

$\hat{\alpha}$代表由KDE模型所預測的混音參數權重，$\alpha$是由參數估計得出實際的混音參數權重值。

### 4.1 一次一個交叉測試(Leave-one-out Cross Validation)

表 3 為每個音樂類型副歌音量參數的一次一個交叉驗證(Leave-one-out Cross Validation)的結果，其值為 KDE 模型所預測的值與實際混音參數權重的均方誤差值(Mean Square Error)，測試時會先對測試歌曲做隨機抽樣 3000 個樣本來做測試，m 值皆為 8。由表可知由 KDE 所訓練出的模型均方誤差平均大約落在 0.129 左右的數值，預測出的權重有著還不錯的準確度，但由於是採用隨機抽樣的方式，若抽到的聲音樣本為較

安靜的樣本(即是說在該樣本的當時有較多軌音量是趨近於 0)時會導致均方誤差值飆高，形成 Outlier，如搖滾類中的第 17 首的結果。

表 3 副歌音量模型交叉驗證結果

| | Rock/Metal | POP | Jazz/Country | Alt Rock/Funk |
|---|---|---|---|---|
| song1 | 0.137 | 0.043 | 0.145 | 0.027 |
| song2 | 0.148 | 0.224 | 0.005 | 0.087 |
| song3 | 0.192 | 0.057 | 0.018 | 0.040 |
| song4 | 0.172 | 0.030 | 0.018 | 0.053 |
| song5 | 0.205 | 0.056 | 0.138 | 0.038 |
| song6 | 0.035 | 0.070 | 0.037 | 0.426 |
| song7 | 0.166 | 0.067 | 0.017 | 0.018 |
| song8 | 0.135 | 0.055 | | |
| song9 | 0.052 | | | |
| song10 | 0.120 | | | |
| song11 | 0.199 | | | |
| song12 | 0.040 | | | |
| song13 | 0.092 | | | |
| song14 | 0.048 | | | |
| song15 | 0.269 | | | |
| song16 | 0.748 | | | |
| song17 | **2.175** | | | |
| **Mean** | 0.290 | 0.075 | 0.054 | 0.099 |

圖 9 是其中一種音樂類型中副歌(Chorus)交叉驗證的詳細圖表，X 軸代表其隨機抽樣測試的聲音樣本(Sound Sample)， Y 軸為其所對應均方誤差，圖表中同樣顏色的線條代表同一首歌的聲音樣本。可以發現 KDE 所訓練出的模型在對同一首歌的聲音樣本時會有一致效果。

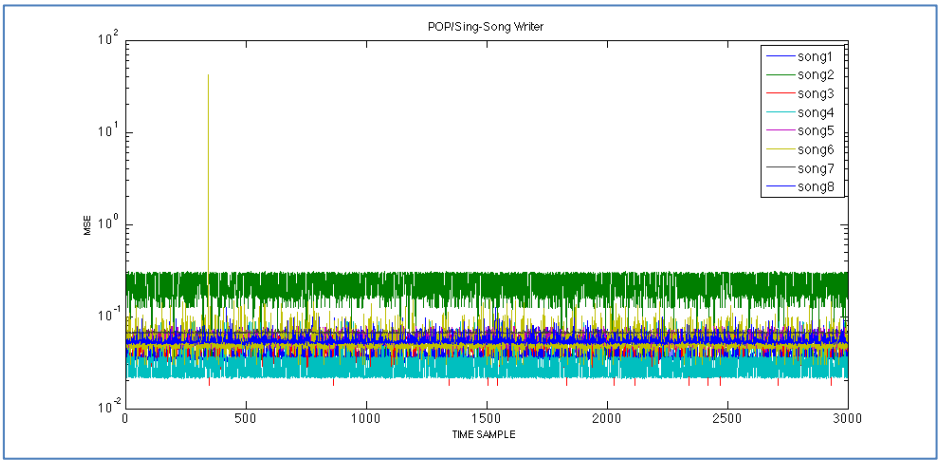


圖 9 POP/Sing-Song writer

在另一方面由音軌的觀點來看，如圖 10，我們可以發現在 Kick、DrumRoom、Overhead 等音軌上各類型會有較大的誤差出現，其原因要歸咎於在實際上多軌同步錄音(Multi-track Recording)時，單一軌也會參雜著其他軌的聲音(串音)，如 Overhead、DrumRoom 等軌會包含 kick、snare 等其他鼓組的樂器。此原因違反了當初研究方法一開始的假設: 我們假設分軌即混音成品間是個線性組合的關係，分軌間必須要是線性獨立; 但由於串音的因素會導致估計及預測有效果不佳的情形。

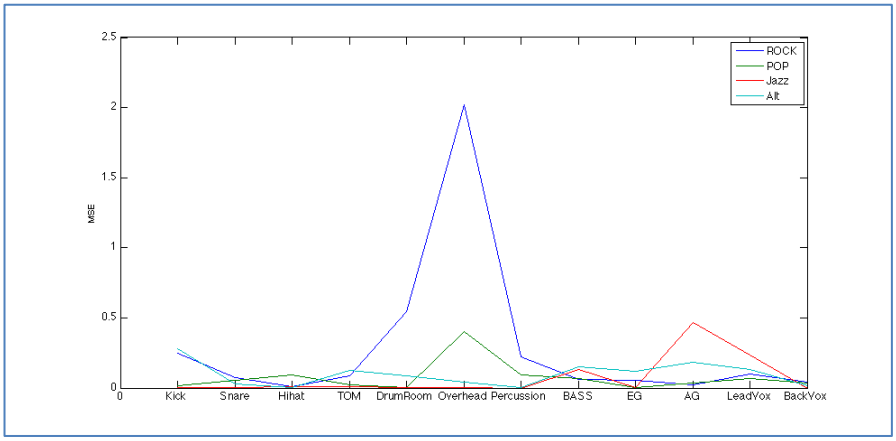


圖 10 音軌比較圖

在頻率(Equalization)模型方面我們同樣也作了交叉驗證，由表中可知大致上各分軌的準確率約為 0.015，顯示在頻率參數預測上有著較好的表現。

表 4 頻率模型交叉驗證

| | Rock/Metal | POP | Jazz/Country | Alt Rock/Funk |
|---|---|---|---|---|
| Kick | 0.011 | 0.019 | 0.003 | 0.008 |
| Snare | 0.043 | 0.015 | 0.011 | 0.018 |
| Hihat | 0.003 | 0.001 | 0.020 | 0.021 |
| TOM | 0.008 | 0.017 | 0.013 | 0.007 |
| DRUMROOM | 0.013 | 0.004 | 0.036 | 0.001 |
| OVERHEAD | 0.017 | 0.028 | 0.003 | 0.046 |
| PERCUSSION | 0.008 | 0.006 | 0.025 | 0.005 |
| BASS | 0.011 | 0.007 | 0.004 | 0.017 |
| EG | 0.006 | 0.019 | 0.024 | 0.005 |
| AG | 0.000 | 0.004 | 0.015 | 0.006 |
| LEADVOX | 0.021 | 0.013 | 0.005 | 0.025 |
| BACKVOX | 0.002 | 0.003 | 0.032 | 0.017 |
| **Mean** | **0.012** | **0.011** | **0.016** | **0.015** |

## 4.2 KDE 方法的效果(Effect of KDE Method)

第二部分的實驗是要來評估 KDE 即其依賴性的成效，首先我們討論了不同的 m 值所帶來的影響，結果如下圖 11，為 ROCK 在副歌時不同 m 值的均方差，X 軸為不同的 m 值，Y 為所對應其混音。我們可以發現當 m 值越低 KDE 模型會有較好的結果，較低的 m 值也可加速 KDE 的計算(eg.PCA space 從 $R^{12}$ 降低為 $R^{8}$)。這結果也顯示了依賴性對於混音參數預測的幫助。

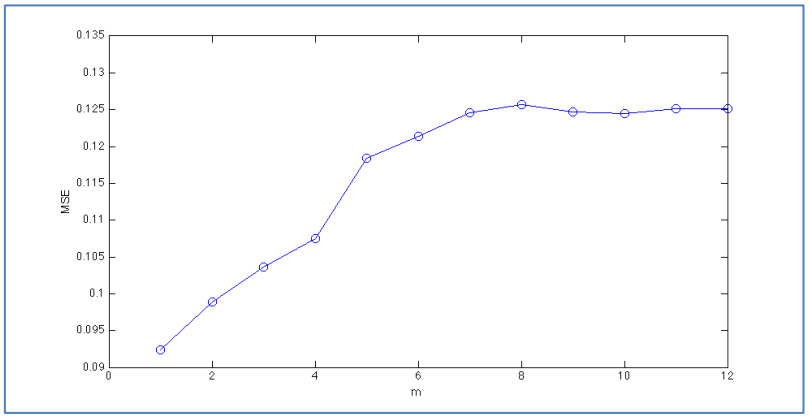


圖 11 不同 m 值比較

接著我們實作了相關研究中[3]所使用的多線性迴歸(無考慮依賴性)與本篇論文的 KDE(考慮依賴性)的方法做比較，結果如下圖 12，可發現在大多數的軌上 KDE 比多線性迴歸有較好的表現。顯示其依賴性估計方式較能考量各軌之間的平衡。

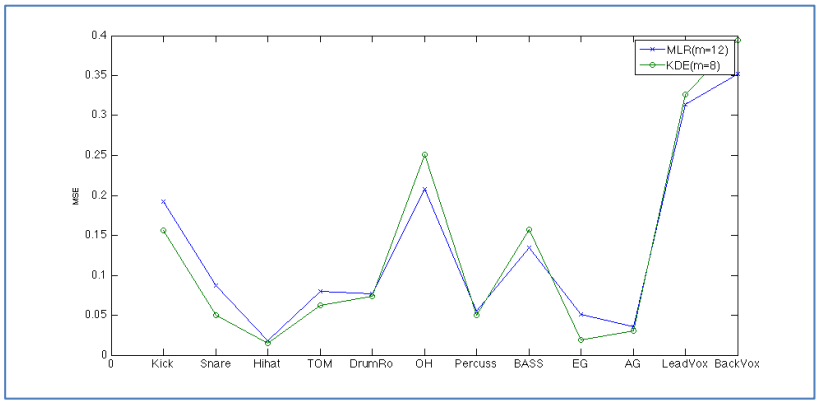


圖 12 KDE 與 MLR 比較圖

## 4.3 跨類型測試(Cross Genre Testing)

在第一部分的實驗模型的建立以及測試都侷限在同一個音樂類型中，在第三部分的實驗，我們想要評估不同的音樂風格的音樂是否有其混音特色，我們將會套用由不同類型所訓練出的模型來看看其效果是否有差別。結果如下圖 13，X 軸分別是先前定義的基本軌，Y 軸是其對應的均方誤差值，當中測試資料為 ROCK 這一類的歌，一共有 17 首，圖中的線條為分別用 POP、jazz、Alt Rock 所訓練出的模型套用至 ROCK 類別的測試結果。我們可以發現將別的類別的模型套用在不同類型的歌時會導致模型的正確度下降、誤差增大，如圖中的 POP 與 JAZZ 的結果，兩類別的模型套用在 ROCK 音樂上其結果顯示不太適合。我們也發現音樂類型相似的歌其混音方式會較相近如圖中 ROCK 與 Alt ROCK 的均方誤差值較為接近。經由此實驗我們可以得知不同的音樂類型有其不同的混音方式，所訓練出來的模型有其獨特性。

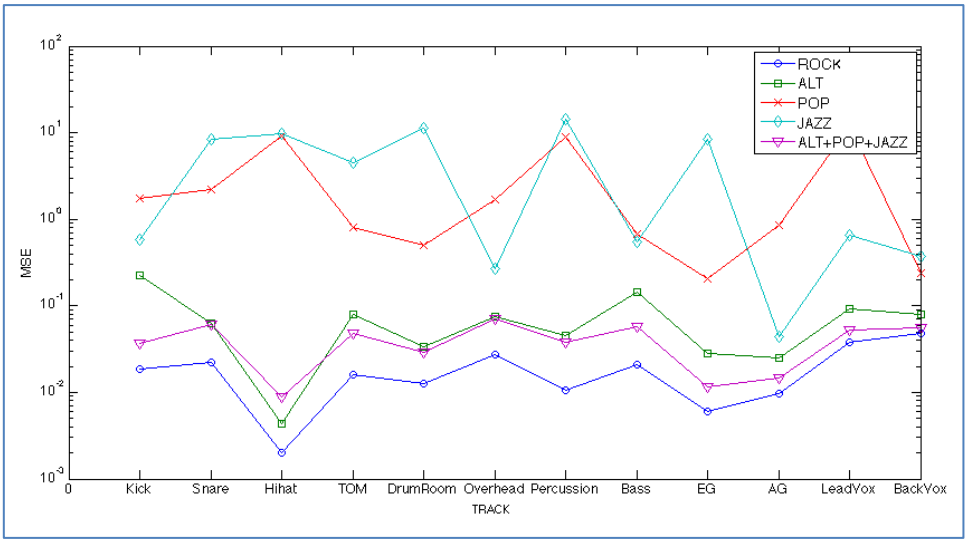


圖 13 類別交叉驗證

## 五、 結論與未來工作

在音樂製作上混音是非常重要的一個過程，音樂成品品質好壞，取決於混音是否混得好而且混音牽涉到許多音響及聲學心理學的相關技術與知識，非專業人士要混出尚可的成品有一定的難度，所以本篇論文提出一個利用監督式學習來進行自動多軌混音的系統。與其他篇相關研究不同的是其主要的核心方法是核依賴估計(Kernel Dependency Estimation)，利用混音參數間的依賴性(dependency)，來做混音參數的預測; 另一個與其他相關論文不同的是訓練的單位不同，由於混音其實是非常個別的(Case By Case)，混音師基本上都會依照各音軌的的聲響、音樂類型不同而有不同的處理方式; 所以在本篇論文的模型建立的過程我們會依照不同的音樂類型以及橋段建立不同的模型，以方便最後實際上的利用。由實驗結果得可知，不同的音樂類型其混音方式是有其獨特性的。

未來工作方面，由於本篇論文有分成四大音樂類型去做模型建立，導致各類別訓練資料有偏少的傾向，未來希望合併其他資料集以補足其數量;再者由於混音參數是非常難取得的，本篇論文是採用估計的方式估計其資料集中的混音參數權重，未來可以改用相關研究的估計方式或是實際收集混音參數以讓最後的混音模型能更貼近實際上的情況，例如建構一個線上混音系統的方式讓使用者實際混音記錄其混音參數。另外在 KDE 方法的部分由於本篇未套用其核函數(Kernel function)的部分，未來可以在做投影前先套用核函數來提升混音模型的效果。實驗部分評估的對象也可以新增跟資料集的混音成品做比較試著看看目前的混音模型與實際上的差距如何。

## 參考文獻

[1] J. Weston, O. Chapelle, A. Elisseeff, B. Scholkopf, and V. Vapnik, "Kernel Dependency Estimation," *Neural Information Processing Systems,* 2002.

[2] J. Scott and Y. E. Kim, "Instrument Identification Informed Multi-track Mixing,"

*International Society for Music Information Retrieval,* 2013.

[3] J. Scott and Y. E. Kim, "Analysis of Acoustic Features of Automated Multi-Track Mixing," *International Society for Music Information Retrieval,* 2011.

[4] D. BARCHIESI and J. REISS, "Reverse Engineering of a Mix," *Audio Engineering Society,* 2009.

[5] D. Barchiesi and J. Reiss, "Automatic Target Mixing Using Least-squares Optimization of Gains And Equalization Settings," *Digital Audio Effects(DAFx-09),* 2009.

[6] H. Katayose, A. Yatsui, and M. Goto, "A Mix-Down Assistant Interface with Reuse of Examples," *Automated Production of Cross Media Content for Multi-Channel Distribution,* 2005.

[7] Mike.Senior, "Mixing Secrets for the Small Studio," 2011.

[8] D. Ward, J. D. Reiss, and C. Athwal, "Multi-track mixing using a model of loudness and partial loudness," 2012.

[9] Kolasinski and Bennett, "A Framework for Automatic Mixing Using Timbral Similarity Measures and Genetic Optimizatio," *Audio Engineering Society,* 2008.

[10] S. H. Nielsen and E. Skovenborg, "Evaluation of Different Loudness Models with Music and Speech Material," *Audio Engineering Society,* 2004.

[11] B. C. J. Moore, B. R. Glasberg, and M. A. Stone, "Why Are Commercials so Loud? ' Perception and Modeling of the Loudness of Amplitude-Compressed Speech," *J. Audio Eng. Soc,* 2003.

[12] Balster and Alex, "Audio Control Facilities in Modern Recording Studios," *Audio Engineering Society,* 1972.

[13] N. Montecchio and A. Cont, "Accelerating The Mixing Phase In Studio Recording productions By Automatic Audio Alignment," *International Society for Music Information Retrieval,* 2011.

[14] E. R. R. 128, "Algorithms To Measure Audio Programme Loudness And True-peak Audio Level," *EBU,* 2010.