國 立 交 通 大 學

電信工程研究所

碩士論文

使用 GMM 轉換之背景伴奏消除及趨勢估計之歌曲音高軌跡追蹤

Using GMM transform-based background removing and trend estimation on pitch contour tracking for singing song

研 究 生 ：林佳緯

指導教授：王逸如 博士

中華民國一百零二年 4 月

使用 GMM 轉換之背景伴奏消除及趨勢估計之歌曲音高軌跡追蹤

Using GMM transform-based background removing and trend estimation on pitch contour tracking for singing song

研究生：林佳緯　　Student: Chia-Wei Lin

指導教授：王逸如 博士　　Advisor: Dr. Yih-Ru Wang

國立交通大學

電信工程研究所

碩士論文

A Thesis

Submitted to Institute of Communication Engineering

College of Electrical and Computer Engineering

National Chiao Tung University

in Partial Fulfillment of the Requirements

for the Degree of

Master of Science

in Communication Engineering

April 2013

Hsinchu, Taiwan, Republic of China

中華民國一百零二年月

使用 GMM 轉換之背景伴奏消除及趨勢估計之歌曲音高軌跡追蹤

研究生：林佳緯　　指導教授：王逸如 博士

國立交通大學電信工程研究所碩士班

中文摘要

在音樂資料檢索中，要對音樂做任何的分類、搜尋或分析，都要先對資料庫中的音樂提取適當的描述作為比對基準，在描述一段音樂的多種基準中，音高軌跡是一種直觀有用的資訊，在有人聲歌唱的音樂中，它指的就是人類歌聲的音高變化，如何從多個音源的音樂中擷取出人類的聲音，並做自動的音高追蹤是本論文的研究重點。

本研究先使用在語音處理中常用來做語者聲音轉換的 GMM 轉換方程式，來去除音樂中的背景伴奏，將含有伴奏的特徵參數轉換成純人聲的特徵參數；接著進行人聲音高趨勢估計，目的在於預測出音高軌跡可能存在的範圍，用以縮小基頻搜尋範圍，除加快運算速度外，並可排除諧波對音高軌跡追蹤所造成的不良引響；最後以動態規劃或是直接峰值選取的方式完成音高軌跡追蹤。實驗結果顯示本研究所提出的方法和現有最好的方法成效相當。

Using GMM transform-based background removing and trend estimation on pitch contour tracking for singing song

Student: Chang-You Cai Advisor: Dr. Yih-Ru Wang

Institute of Communication Engineering
National Chiao Tung University

Abstract

In music information retrieval, a proper representation of music signals in the database is needed for music analysis, search or classification. Among all existing music representations, pitch contour is an intuitive and useful feature. In a song, pitch contour is the fundamental frequency variation of the singing voice. The main concern of this thesis is hence to automatically extract the pitch contour of singing voice from a polyphonic music.

In this study, we first employ the GMM transform function, which is popularly used in speaker voice conversion, to remove the background accompanies of polyphonic music. It transforms the characteristic features of a polyphonic music to those of vocal-only signal. Then, a trend estimation is conducted to predict the range of human's pitch contour from the vocal-only characteristic features. The trend estimate is used to restrict the searching region for speeding up the searching process as well as for eliminating the interference of harmonics on pitch tracking. Lastly, pitch contour is obtained by dynamic programming or simple peak picking. Experimental results confirmed that the performance of the proposed method was comparable to the best method existing nowadays.

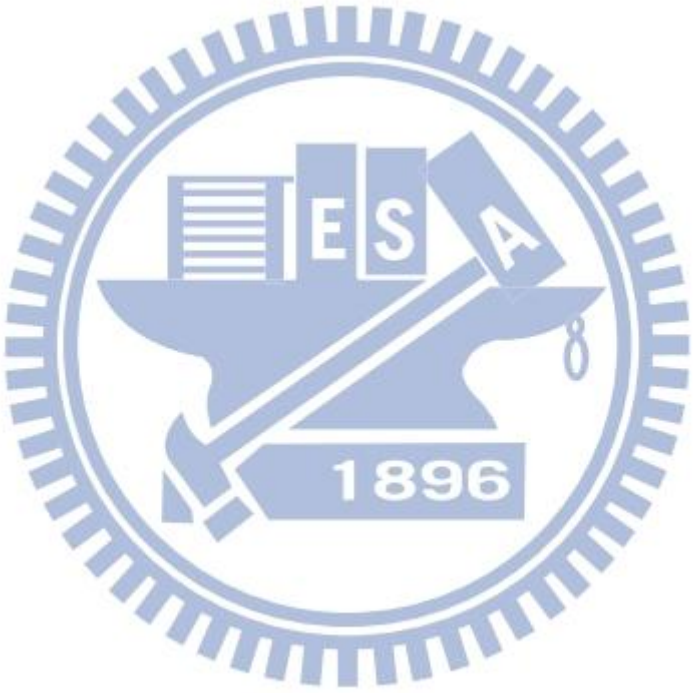
誌謝

本論文得以順利完成，有賴於諸位師長朋友的支持與鼓勵。首先感謝指導我的陳信宏教授和王逸如教授，在我漫長的研究中，持續給予忠肯的指導與建議；另外還有廖宜斌學長，這數年來亦師亦友的在研究上，從引領入門到深入研究，提供許多寶貴的建議和創意；同時也還感謝蔡昌祐學弟在同領域中一起打拼互相給予支持。

在這感謝我的家人，在這段時光給予的支持，家最終仍是每個人的避風港，讓我總能在疲累不堪時回到家裡休息，帶著滿滿的電力回到戰場。特別感謝我的女友，這段時間的扶持，因為初期研究進度的落後，內心累積了許多壓力，還有深感自己追不上大家腳步的自卑感，好幾次都在放棄或是繼續的邊緣掙扎，妳是讓我繼續前進的原因，只有妳不願意讓妳失望，為了妳我必須更努力，妳有如一盞明燈照亮我的人生路標。

本人不才，一千餘日的光陰，僅得此寥寥數筆的拙作。如有後人能從本文中得到些許啟發，即為對我最大的鼓舞。回首近四年的旅程，走得比別人久，感觸也特別多，在這段時間自己的蛻變，種種難以言喻。感謝這段時光中的貴人、朋友、家人和妳，我畢業了!

目錄



表目錄

圖目錄

第一章 緒論

1.1 研究動機

在資訊爆炸的時代中，各種音樂資訊的取得非常容易，不論是個人的收藏或是公司團體建立的大型資料庫，都比過去來的大的多，在龐大的音樂資料庫中要進行搜尋或是音樂間的比對非常不容易，理由是音樂有多種的表現方式，例如不同的曲風、歌唱者、樂器…等等，不同的表現方式使音樂的頻譜不能簡單地互相做比較。

過去音樂的資訊通常是以文字的方式來達成，例如將一個音樂的資料以作曲者、歌名、演唱者及出版商等方式儲存在資料庫中，但是這類方法仍然不夠直觀，彈性也有限。因為音樂的變化性，就算相同歌名的歌曲，可能也有多種不同曲風的版本，同一個歌手同一首歌，可能也有現場演唱和錄音室版本，很多資訊無法用單純的文字來展現。如同人類的記憶方式，最能讓人分辨一段未知的聲音到底是什麼音樂的基準，就是音樂的主旋律。

雖然音樂在頻譜上的變異度相當大，因此用來表達音樂資訊的參數應該跳脫音色伴奏的影響，對不同的歌曲做比較，最簡單的比對基準就是音高軌跡(pitch contour)，是否為同樣的歌曲，在分別擷取音高軌跡並且互相比對後，能輕易地顯現相似的程度。音高軌跡通常指稱的是音樂中主旋律的音高路徑，而主旋律在一般流行音樂中指的就是人類歌聲的部分。分析純人類歌聲的音高軌跡難度其實不高，但是如何在一個同時有各種樂器伴奏的音樂中擷取出人類歌聲的音高軌跡才是難題所在。人類的音高在音樂當中大約處在 80Hz 到 1500Hz 之間，在這個頻帶依然混雜了很多伴奏聲，如何凸顯人類的聲音特性，或是如何分離伴奏的聲音特性是本論文要探討的內容。

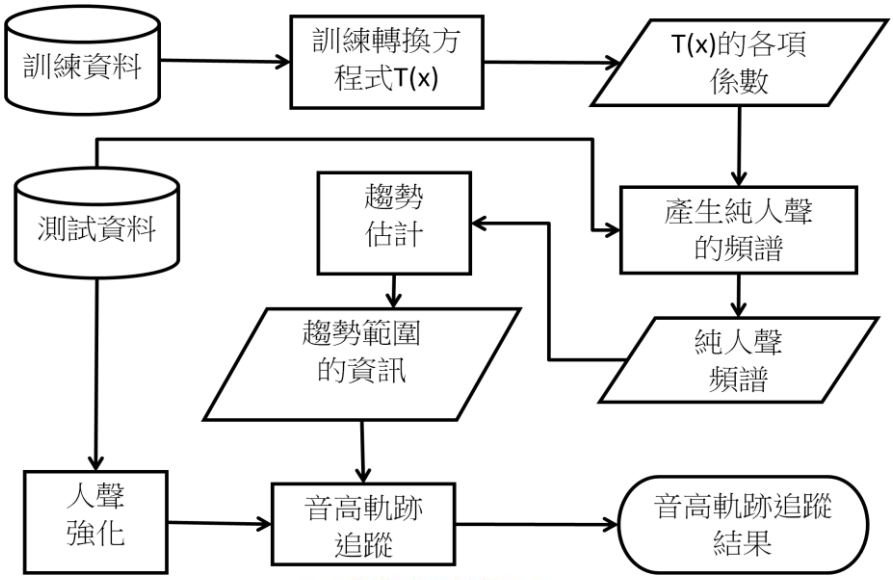
1.2 文獻回顧

在音高軌跡追蹤的研究中，[10]使用訓練參數模型的方式，這樣的研究需要大量的訓練資料，[11]則是使用了一連串的頻譜分析演算法，先使用[8]提出的Harmonic/Percussive Sound Separation (HPSS)來壓抑頻譜上伴奏的能量，再使用[9]中的Normalized Sub-harmonic Summation (NSHS)來凸顯音高的能量。另一個頻譜分析的方法是[12]，先在頻譜上建立多個片段的軌跡，並訂立一連串的規則，將屬於主旋律音高的片段軌跡連接起來。

在背景伴奏消除的研究中，[11]是利用 Harmonic/Percussive Sound Separation（HPSS）和 Normalized Sub-harmonic Summation (NSHS)的方式，壓抑頻譜上伴奏的能量，並凸顯人聲在頻譜上的能量。[13]是訓練樂器的模型，並且將主旋律從背景樂器中分離。[14]則是將整個頻譜當作訓練資料，使用 support vector machin(SVM)分類器的方式來達到分離背景伴奏的功用。[15]是將音檔提取特定參數集合，如音高，能量和諧波性質，進行一系列的頻譜群聚(spectral cluster)，對多音音樂進行音源分離，再進行主旋律的軌跡偵測。另外這領域較不同的是本論文參考了[1]聲音轉換(voice conversion)研究中常用的 GMM 轉換方式，將去除背景伴奏的步驟當作是一種聲音的轉換，以將含有伴奏的音樂特徵參數轉換成無伴奏的歌聲特徵參數。

1.3 研究方法

本論文中使用的方法可以分為三大步驟，第一是參照[1]中的 GMM 轉換法，去除音樂特徵參數中的背景伴奏資訊，並且參考[11]的系統架構，將去除伴奏的音樂特徵參數應用在第二步驟的趨勢估計，最後以趨勢估計的資訊配合原始音樂頻譜進行最後一步驟的音高軌跡預測，如圖一-1。



圖一-1：整體系統流程

其中在趨勢估計的步驟，因為諧波(harmonic)的干擾，很容易誤將高頻的能量當作是正確答案，而錯誤的趨勢估計將產生完全錯誤的音高軌跡，因此本論文第三章將著重在探討如何優化趨勢估計。在進入音高軌跡偵測前有個人聲強化步驟(Vocal enhancement)這步驟是為了讓原始頻譜上的人聲凸顯出來，在第四章以及第五章會深入討論幾個可行的方案並做比較。

1.4 章節概要

本論文一共分為六章，各章節內容分配如下：

第一章 緒論：介紹本論文之研究動機與方向。

第二章 使用轉換方程式去除伴奏特徵:說明如何使用 GMM 轉換法去除背景伴奏特徵。

第三章 使用預測的特徵參數於趨勢估計：使用第二章得到的參數進行趨勢估計。

第四章 人聲強化項目整理：介紹並討論幾種人聲強化的方法。

第五章 修改訓練參數以及二階段趨勢估計：再次優化本論文使用的方法與流程。

第六章 結論與未來展望。

第二章 使用轉換方程式去除伴奏特徵

本章討論以聯合高斯機率密度(Gaussian mixture model, GMM)的轉換方程式，應用於去除多音音樂(polyphonic music)特徵參數中的伴奏成分，僅保留人聲歌唱的特徵資訊，這些資訊於後續章節中，將應用於音高軌跡的趨勢估計和音高軌跡的追蹤(tracking)步驟。

2.1 高斯轉換方程式訓練

聯合高斯機率密度的轉換方程法廣泛的應用於聲音轉換(voice conversion)研究中，本論文參考[1]中提出的方法，使用多音的特徵參數訓練 GMM 轉換方程式，訓練後的方程式輸入混音的特徵參數可得到去除伴奏的特徵參數，利用這些特徵參數分別對混音和純人聲的特徵參數做歐幾里得距離的計算，以驗證去除伴奏的成效。

2.1.1 高斯機率密度轉換方程式

定義 $\boldsymbol{s}=\{s_1, s_2, \dots s_L\}$和$\boldsymbol{v}=\{v_1, v_2, \dots v_L\}$是獨唱(solo)和有伴奏(accompanied)的音樂特徵向量，L 是音檔的總音框(frame)數，要定義一個轉換方程式 $T(\cdot)$，使$\sum_{l=1}^{L}[T(v_l)-s_l]^2$有最小值。根據 MMSE 法則當$T(\boldsymbol{v})=E\{\boldsymbol{s}|\boldsymbol{v}\}$時為最佳解，$E\{\cdot\}$表示期望值，推導過程參考[2]。因此假設$\boldsymbol{s}$和$\boldsymbol{v}$是高斯隨機變數，則方程式可以表示成：

$$T(\boldsymbol{v})=E\{\boldsymbol{s}|\boldsymbol{v}\}=\boldsymbol{\mu}_s+\boldsymbol{\Sigma}_{sv}\boldsymbol{\Sigma}_v^{-1}(\boldsymbol{v}-\boldsymbol{\mu}_v) \tag{2-1}$$

其中$\boldsymbol{\mu}_s$表示$\boldsymbol{s}$的平均值，$\boldsymbol{\mu}_v$表示$\boldsymbol{v}$的平均值，$\boldsymbol{\Sigma}_{sv}$表示$\boldsymbol{s}$和$\boldsymbol{v}$的交互共變異矩陣，$\boldsymbol{\Sigma}_v$表示$\boldsymbol{v}$的

共變異矩陣。

接著定義$z_l = \left[s_l^{\mathrm{T}} v_l^{\mathrm{T}} \right]^{\mathrm{T}}$，for $1 \le l \le L$，其中上標 T 表示矩陣轉置，因此z_l的平均值和變異數可以表示成：

$$\boldsymbol{\mu}_z = \begin{bmatrix} \boldsymbol{\mu}_s \\ \boldsymbol{\mu}_v \end{bmatrix} \text{ 和 } \boldsymbol{\Sigma}_z = \begin{bmatrix} \boldsymbol{\Sigma}_z & \boldsymbol{\Sigma}_z \\ \boldsymbol{\Sigma}_{sv}^{\mathrm{T}} & \boldsymbol{\Sigma}_z \end{bmatrix} \tag{2-2}$$

方程式定義完成後，輸入多音的特徵參數序列$\boldsymbol{x} = \{x_1, x_2, \ldots x_l, \ldots\}$，可以利用(2-1)式將它轉換為去除伴奏的後的特徵參數序列。

2.1.2 聯合高斯機率密度轉換方程式

承上一小節，$\boldsymbol{s}$和$\boldsymbol{v}$之間的關係事實上並不是簡單線性關係，z_l應該使用聯合高斯密度(GMM)的方式來表達而非單一高斯密度。參數的表達應加入每個維度的權重$w^{(m)}$，上標 m 是各 mixture 的標號，經由 EM (Expectation-Maximization)演算法可以得到各 mixture 的權重，(2-2) 式的改變如下：

$$\boldsymbol{\mu}_z^{(m)} = \begin{bmatrix} \boldsymbol{\mu}_s^{(m)} \\ \boldsymbol{\mu}_v^{(m)} \end{bmatrix} \text{和} \boldsymbol{\Sigma}_z^{(m)} = \begin{bmatrix} \boldsymbol{\Sigma}_s^{(m)} & \boldsymbol{\Sigma}_{sv}^{(m)} \\ \boldsymbol{\Sigma}_{sv}^{(m)^T} & \boldsymbol{\Sigma}_v^{(m)} \end{bmatrix} \tag{2-3}$$

承上，在輸入多音的特徵參數序列$\boldsymbol{x} = \{x_1, x_2, \ldots x_l, \ldots\}$，(2-1)式可以改寫為：

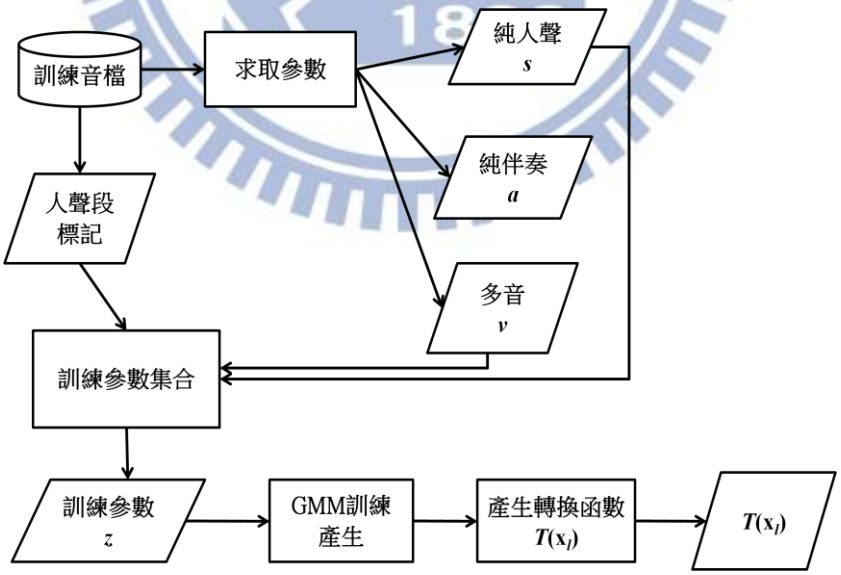
$$T(x_l) = \sum_{m=1}^{M} P(m|x_l) \left[\boldsymbol{\mu}_s^{(m)} + \boldsymbol{\Sigma}_{sv}^{(m)} \boldsymbol{\Sigma}_v^{(m)^{-1}} \left(x_l - \boldsymbol{\mu}_v^{(m)} \right) \right] \tag{2-4}$$

其中 M 是聯合高斯使用的 mixture 數，

$$P(m|x_l) = \frac{w^{(m)} N(x_l; \boldsymbol{\mu}_v^{(m)}, \boldsymbol{\Sigma}_v^{(m)})}{\sum_{m'=1}^{M} w^{(m')} N(x_l; \boldsymbol{\mu}_v^{(m)'}, \boldsymbol{\Sigma}_v^{(m')})} \tag{2-5}$$

$$N\left(x_l; \boldsymbol{\mu}_v^{(m)}, \boldsymbol{\Sigma}_v^{(m)}\right) = \frac{1}{(2\pi)^{\frac{D}{2}} \left|\boldsymbol{\Sigma}_v^{(m)}\right|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\left(x_l - \boldsymbol{\mu}_v^{(m)}\right)^{\mathrm{T}} \boldsymbol{\Sigma}_v^{(m)^{-1}} \left(x_l - \boldsymbol{\mu}_v^{(m)}\right)\right\} \tag{2-6}$$

基於上面之敘述就完成了轉換方程式的定義，方程式的作用是可以輸入多音的特徵參數，經過方程式轉換成純人聲(背景伴奏去除)的特徵參數，整體訓練流程如圖二-1，首先對訓練音檔求取需要的特徵參數，這裡訊號分成三種：$\boldsymbol{s}$代表純人聲、$\boldsymbol{a}$代表純伴奏、$\boldsymbol{v}$代表混和的訊號。在形成訓練參數$\boldsymbol{z}$的過程，為了避免沒有人聲的時間區段造成訓練過程的誤解，所以僅保留有人聲的時間區段來做 EM 訓練，最後得到的各項係數建構成轉換方程式$T(\boldsymbol{x})$。



圖二-1：訓練轉換方程式流程

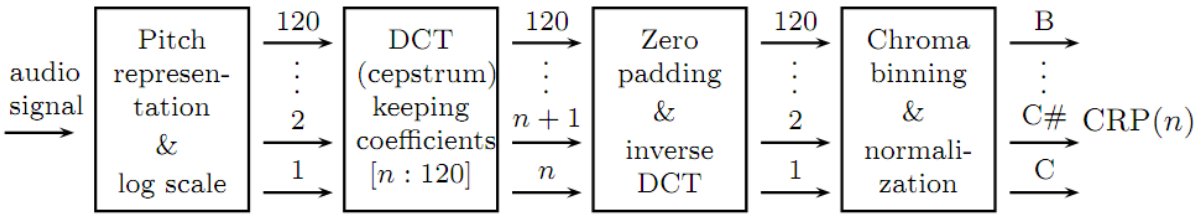
2.2 特徵參數設定

在[1]中使用的特徵參數是 MFCC 參數，MFCC 廣泛的應用在語音處理的研究中，對於音色的特徵有較佳的詮釋。本論文中將使用一種變化型的參數 PFCC (pitch-frequency cepstral coefficients)，是將 MFCC 程序中根據梅爾頻率(Mel-frequency)設計的 filter banks 替換成根據 MIDI 頻率設計的 filter banks，增加頻率上的解析度。MIDI 是音樂上的一種頻率計算單位，因應人耳對頻率的敏感程度，MIDI 每一度間是對數關係，通常以音樂中的中央 La (A4)作為基準，而中央 La 的頻率是 440Hz，在 MIDI 上的標號是 69，因此頻率和 MIDI 有這樣的轉換式：

$$m = 69 + 12 \times \log_2\left(\frac{f}{440}\right) \tag{2-7}$$

其中 m 表示 MIDI，f 表示頻率。

在[3]中發表了一種求取色度參數(chroma)的 toolbox，它應用 PFCC 於頻譜分析，用來產生一種高強健(robust)的色度參數叫做 CRP (chroma DCT-reduced log pitch)。圖二-2 是 CRP 的產生流程，先將聲音信號經過 120 個以 MIDI 頻率做設計的 filter banks，每個代表一度 MIDI 的能量，再以 DCT 轉換成倒頻譜的形式稱為 PFCC；PFCC 的低維度代表著音色相關的特性，因此只取 n 維度以上的 PFCC，並將 n 維度以下補零，以 IDCT 將倒頻譜轉回頻譜，並將 120 維度中分別屬於同一色度的各自疊加，產生 12 維度的 CRP 參數，本研究中只取用 PFCC 作為替代 MFCC 參數的用途。



圖二-2：求取 CRP 以及 PFCC 的步驟

圖片來源：[4]的 fig.1

2.3 驗證

2.3.1 環境設定

MFCC 使用 20 個 Mel-scale filter banks， 獲得 20 維參數，PFCC 如 2.2 節所述，使用 120 個 MIDI-scale filter banks 獲得 120 維參數，兩種參數取樣窗口皆是 32 ms，並且每次取樣位移 10ms。GMM 使用 32 個 mixture，以 k-means 當作訓練演算法，使用源自於 voicebox [6]的 code 來做運算。

2.3.2 音樂資料庫

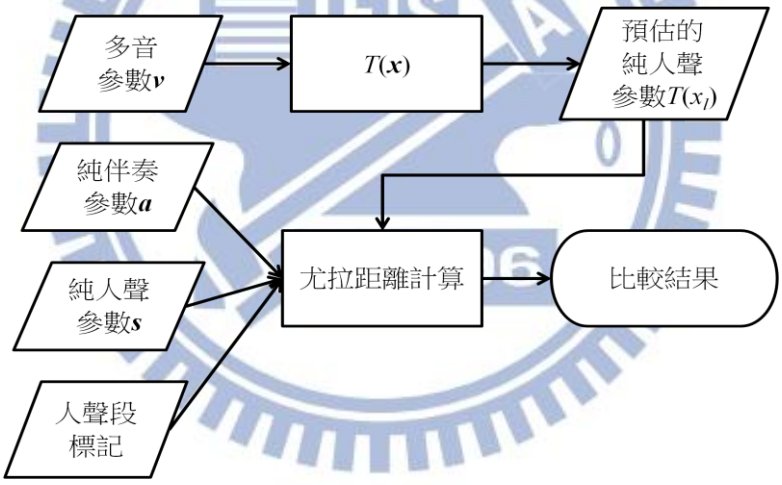
在音樂資料庫方面使用 MIR-1k，由清大 MIR 實驗室提供[5]，這是一個公開的音樂資料庫，內有 1000 段取樣頻率 16 kHz，16 bits 的無壓縮音檔，並以左右聲道分別伴奏和人聲。資料庫中男女音檔總長度比例約為 6:4，每段音檔平均 8 秒鐘，總長度 2.22 小時，並且有人工標記之人聲音高正確答案。本論文的實驗中取男女各 21 分鐘作為訓練資料，共約 253,455 個參數向量，而剩餘的音檔中隨機挑出 300 段音檔做為測試資料。

2.3.3 驗證結果

驗證步驟以及驗證結果如圖二-3 和表二-1 所示，

表二-2 驗證方法是計算預測出來特徵參數$T(x_l)$與混音的參數$\boldsymbol{v}$和人聲的參數$\boldsymbol{s}$之間的歐幾里得距離，合理狀況是$T(x_l)$與$\boldsymbol{s}$距離較近，與$\boldsymbol{v}$較遠才符合本論文想要去除伴奏的目的。如表二-1 和

表二-2 所示，不論是在縱觀、內部測試或是外部測試，不論使用 MFCC 還是 PFCC，歐幾里得距離的大至小排序都是“$\boldsymbol{v}$ vs. $\boldsymbol{s}$”、“$\boldsymbol{v}$ vs. $T(x_l)$”、“$\boldsymbol{s}$ vs. $T(x_l)$”，“$\boldsymbol{v}$ vs. $\boldsymbol{s}$”最大是因為多音參數比純人聲多了許多伴奏的特徵，“$\boldsymbol{v}$ vs. $T(x_l)$”第二大是因為預測的參數是由多音參數經過轉換方程式而來，其中仍然會殘留部分沒有被完全消去的伴奏特徵，“$\boldsymbol{s}$ vs. $T(x_l)$”最小則符合期望預測參數和純人聲最接近。



圖二-3：驗證流程

表二-1：使用 MFCC 做參數的各項歐幾里得距離

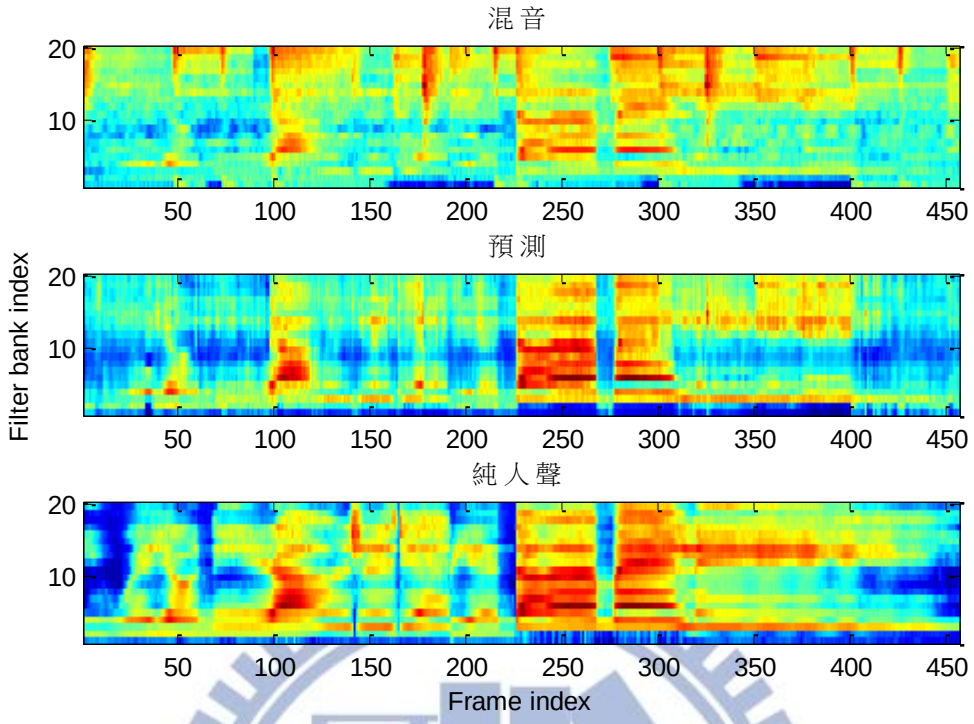
	Total	Inside	Outside
$\boldsymbol{v}$ vs. $T(x_l)$	61.2139	58.007	64.4768
$\boldsymbol{s}$ vs. $T(x_l)$	29.7823	26.0417	33.7753
$\boldsymbol{v}$ vs. $\boldsymbol{s}$	62.7763	61.2276	66.091

表二-2：使用 PFCC 參數的各項歐幾里得距離

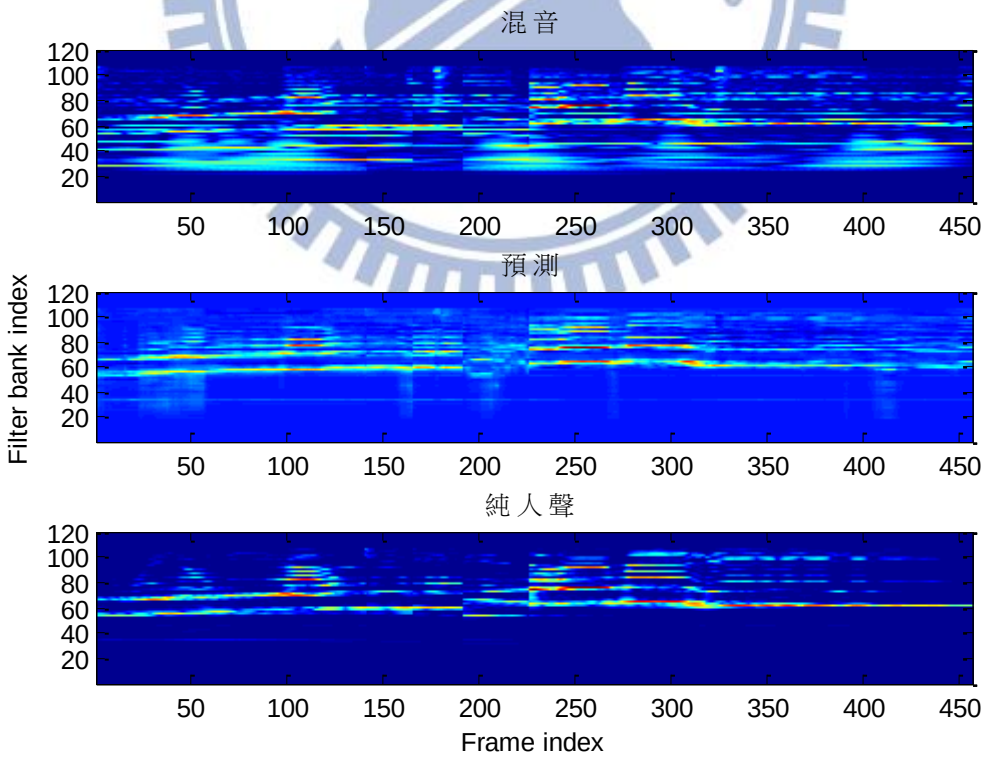
	Total	Inside	Outside
$\boldsymbol{v}$ vs. $T(x_l)$	5.7884	2.7819	3.0065
$\boldsymbol{s}$ vs. $T(x_l)$	4.1725	1.6594	2.513
$\boldsymbol{v}$ vs. $\boldsymbol{s}$	6.8169	3.2488	3.5681

2.3.4 討論

承 2.4.2 小節，既然使用兩種參數效果都如同預測，那使用 PFCC 的優勢在哪? 請見圖二-4 和圖二-5，圖中縱軸的單位是 filter bank 的編號，橫軸是 frame 的編號。不管 MFCC 還是 PFCC，在產生步驟中都有一個步驟是把對數能量頻譜做 DCT 轉換，而在產生預測參數後，為了頻譜分析的用途，必須將參數做 IDCT 轉換回對數能量頻譜，如圖二-4 這是使用 MFCC 的範例，因為 MFCC 的 Mel-scale filter banks 只有 20 個，因此在轉回頻譜表示法後頻域的解析度嚴重不足，而圖二-5 是使用 PFCC 的，因為使用了 120 個 filter banks，並且都是剛好針對 MIDI 規格做設計，所以對數能量頻譜的頻域解析度較佳，在“預測”的欄位可以看到跟“純人聲”欄位一樣有清楚的音高軌跡，與“混音”相比“預測”則是消去了很多伴奏造成的干擾，這也就是本論文參數選用 PFCC 而不是 MFCC 的重要原因。



圖二-4：使用 MFCC 做參數的對數能量頻譜



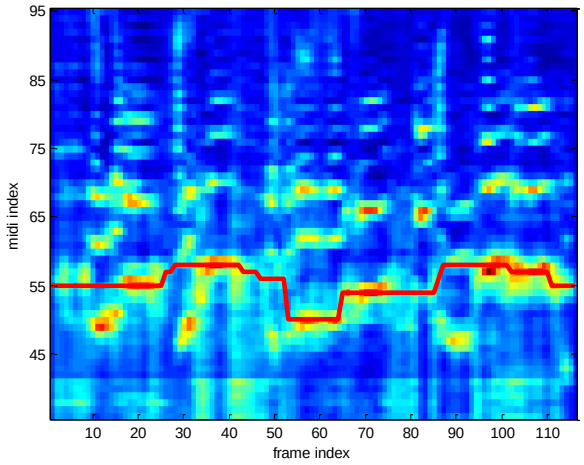
圖二-5：使用 PFCC 做參數的對數能量頻譜

第三章 使用趨勢估計之音高軌跡追蹤

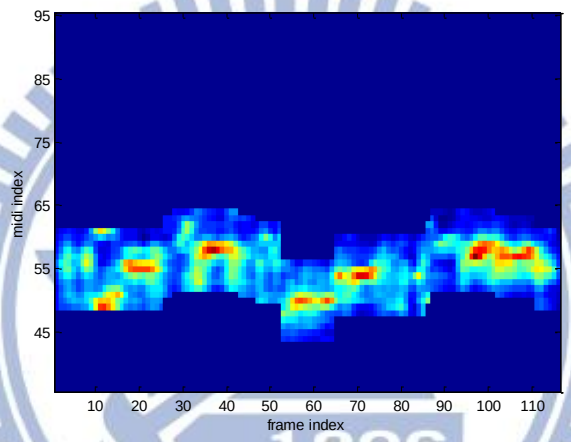
本章承接第二章，經由轉換方程式所產生的特徵參數$T(x_l)$ (往後簡稱為Tx)，利用 IDCT 轉換回頻譜型式，這個頻譜可以視為頻率解析度較低的純人聲的頻譜(往後簡稱為S_{Tx})。參考[7]中的 Hsu's method，在音高軌跡追蹤之前，有一個步驟名為趨勢估計，用意在於大略的估計音高可能的範圍，一來加速最後一步驟音高追蹤的速度，二來可以避免使用 Dynamic Programming (DP)的方法時，因為高頻諧波的干擾而"拉走"正確的音高軌跡。S_{Tx}正好可以應用在這個步驟所需的歌聲頻譜。

3.1 實驗流程

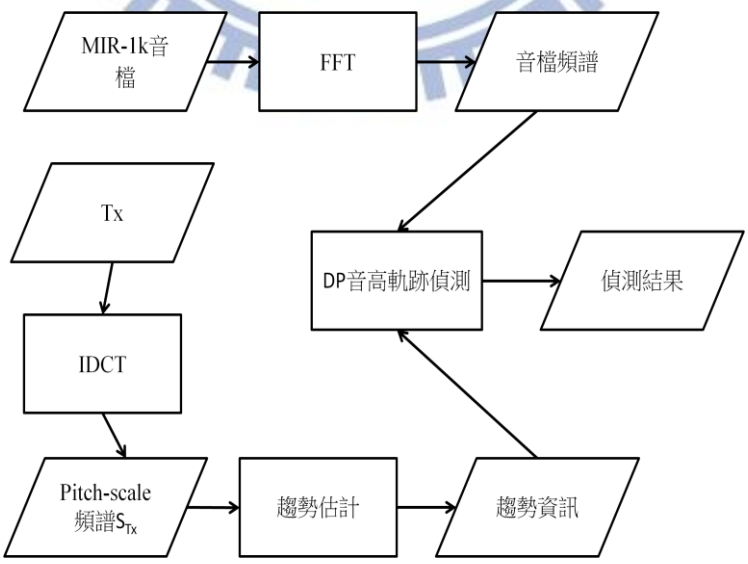
"趨勢估計"的目的就是估計出人類歌聲音高軌跡所在的頻帶，第一步要先有去除伴奏的純歌聲頻譜，第二步將純歌聲的頻譜先降低時間上的解析度(為了得到比較平滑的音高性質)，第三步以 DP 的方式在歌聲的頻譜上尋找一條"分數"最高的路徑，並且"擴展"這條路徑，形成一個帶狀的範圍，這個範圍就認為是音高軌跡可能出現的頻帶，在這部分趨勢的寬度是 12 個半元音，也就是一個"八度"的範圍。在第二章中利用 GMM 轉換方程法得到了去除背景伴奏的純歌聲特徵參數Tx，這是一種倒頻譜(cepstrum)的特徵參數，而在經由 IDCT 轉換後得到的頻譜稱為S_{Tx}，方可應用在趨勢估計中。圖三-1 是利用 DP 在S_{Tx}上尋找路徑的範例，圖三-2 是將這條路徑擴展成帶，並遮蔽其它頻帶完成趨勢估計的步驟的範例。進行趨勢估計後得到趨勢資訊，再將原始頻譜在趨勢估計的範圍內以 DP 進行音高軌跡預測，以得到結果。音高軌跡的"正確率"計算方式是把預測的音高軌跡對比上正確答案，若差異在於 t 個半元音(semitone)之間則算正確，在此忽略人聲偵測的步驟所以只計算有人聲段的音高正確率(raw pitch accuracy)，最後統計整個測試資料的結果就是音高軌跡的正確率，t 的大小關係系統的容忍度，通常取 0.5 當作一個客觀衡量標準。正確率運算流程如圖三-3。



圖三-1：在 S_{Tx} 上找出一條路徑



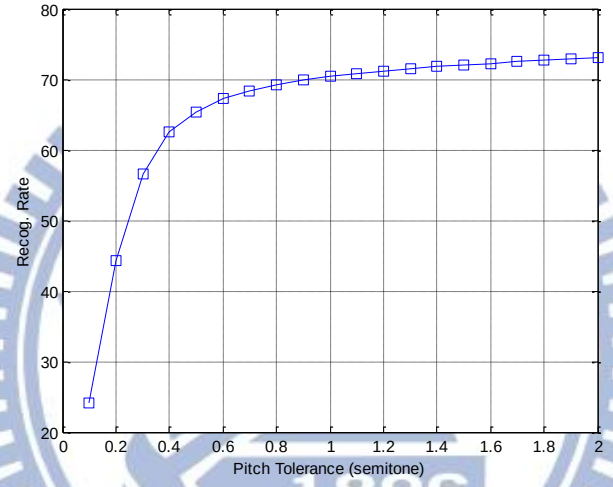
圖三-2：以 S_{Tx} 上的路徑為準擴展出帶狀範圍



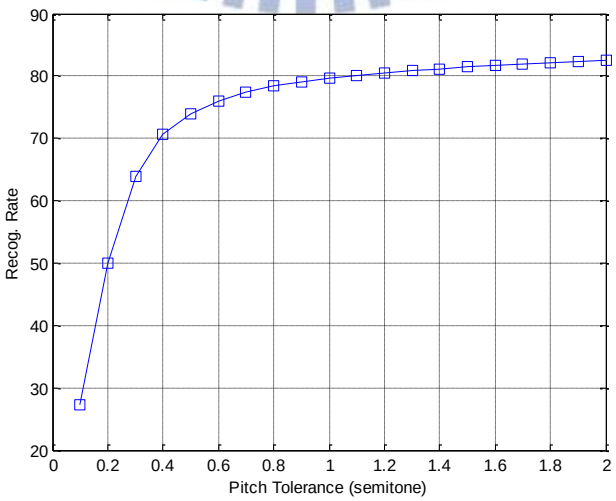
圖三-3：應用 *Tx* 於趨勢估計中並計算音高軌跡正確率

3.2 DP 優化

首先看到兩張圖三-4 和圖三-5，圖三-4 是應用 *Tx* 於趨勢估計後的音高軌跡正確率縱軸是正確率，橫軸是容忍度也就是 *t*，關注於容忍度 0.5 處，正確率約為 67%，對於整體系統來說相當差，但是看到圖三-5，去除部分正確率過低的音檔，容忍度 0.5 處的正確率來到 74%，可見若是校正這些正確率過低的音檔可以對整體成效有顯著的提升。



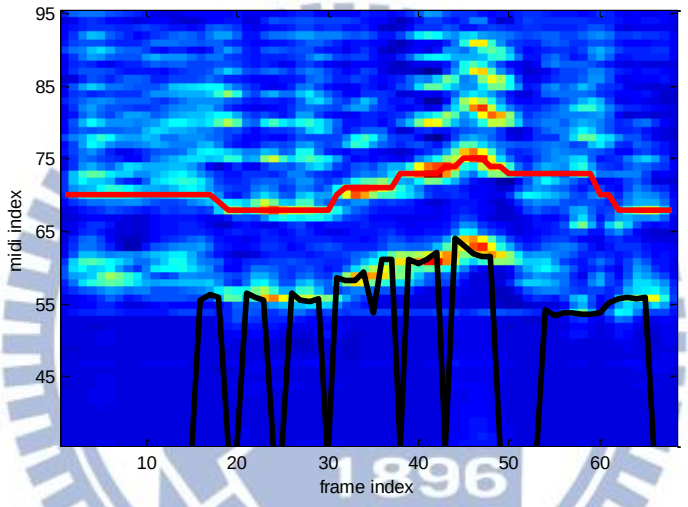
圖三-4：應用 *Tx* 於趨勢估計後，音高軌跡的正確率



圖三-5：應用 *Tx* 於趨勢估計後，音高軌跡的正確率(去除正確率過低的音檔)

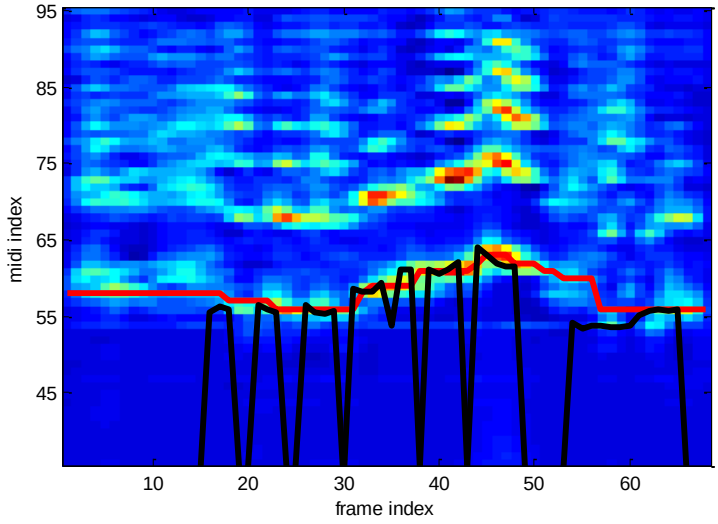
3.2.1 選取適當的 backtracking 點

觀察正確率低下的音檔後，發現大部分的錯誤是源自於 DP 趨勢軌跡選取錯誤，如圖三-6，這個範例中紅色的線是 DP 選取的高分路徑，而黑線是正確答案(黑線歸零的地方表示沒有人聲)，由圖可知因為諧波的干擾，導致 DP 選錯了範圍，選到了較高頻帶的部分。

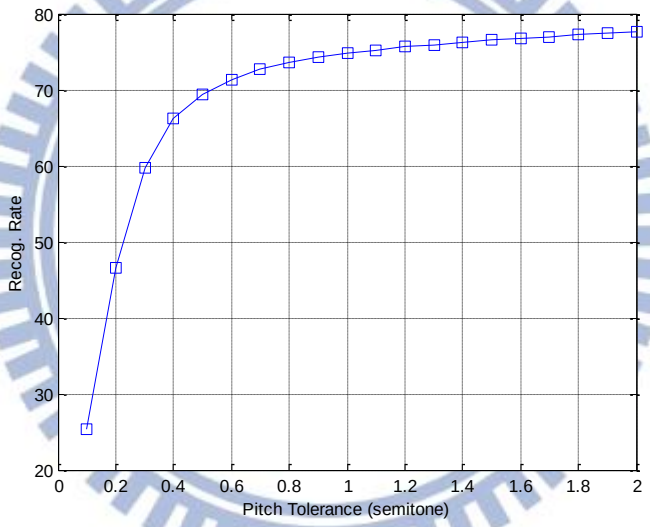


圖三-6：錯誤的 DP 路徑範例

要解決這個問題可以先調整，DP 的 backtracking 點。常態下 backtracking 都是從分數累加最高的點開始，但是為了避免如圖三-6 的錯誤，本論文改為選擇由低頻往高頻遇到的第一個最大值作為 backtracking 點，這也因為S_{Tx}在低頻區域幾乎沒有任何能量，所以遇到第一個能量集中的區域就會是人聲的頻帶範圍。如圖三-7，這是一個修正後的範例，而總體正確率提升至 70%，如圖三-8。



圖三-7：修正 Backtracking 點後的範例

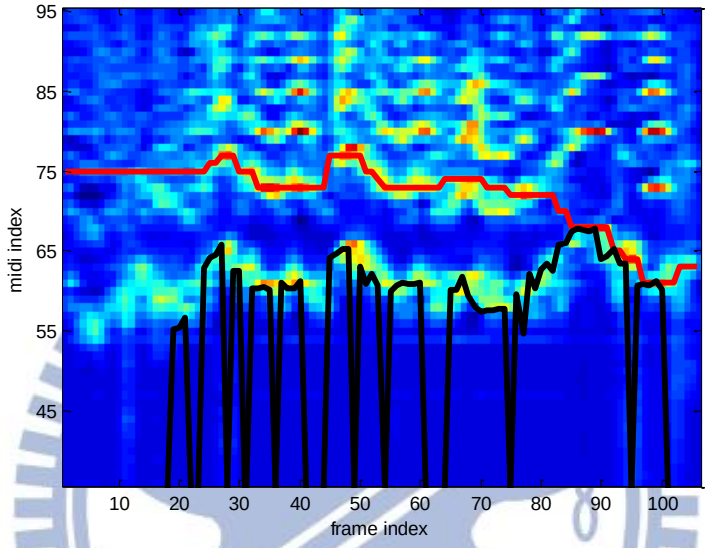


圖三-8：修正 Backtracking 點後的正確率

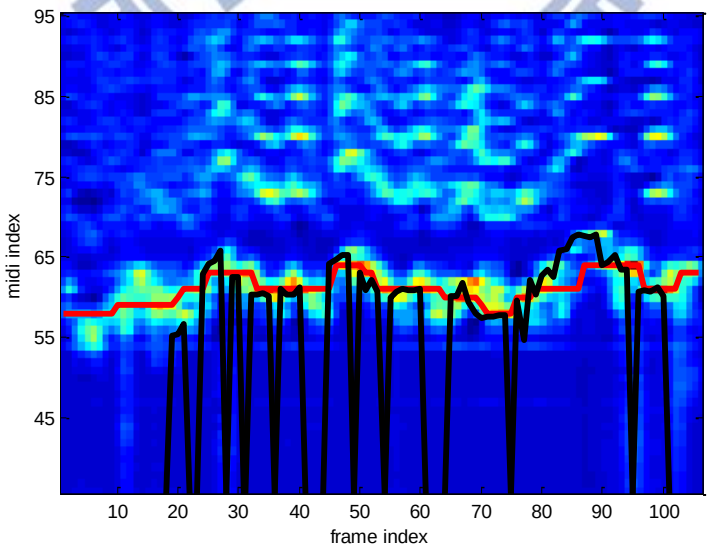
3.2.2 諧波壓抑

承上一小節，糾正 backtracking 點後仍然有些音檔正確率低下，原因是諧波(Harmonic)干擾過於強烈，就算改變了 backtracking 的起始點，仍然會被“拉走”，如圖三-9的例子。解決辦法是壓抑諧波部分的能量，但是要優先決定哪個頻帶開始才該被壓抑，因此首先必須定義出主調(main pitch)，計算方式是將S_{Tx}在時間方向做累加，然後選擇

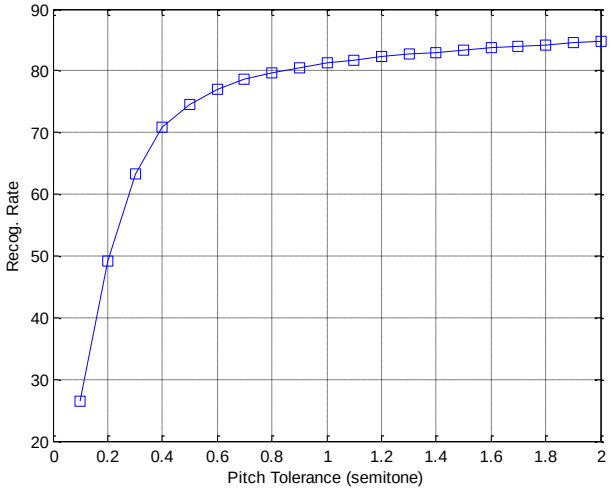
累加結果中，有最大累加能量的作為主調，但是同樣因為受到諧波的干擾，最大值可能發生在主調倍頻的頻帶，因此應該選擇由低頻往高頻第一個遇到的峰值做為主調。選擇主調後測試以高於主調 *M* 個 semitone 的地方都乘上 *N* 倍，在這裡經由 try-&-error 得到 *M* 等於 4，*N* 等於 0.5，而圖 3-9 的例子經由壓抑諧波後成為圖三-10。圖三-11 則是經過 3.2.1 和 3.2.2 提到的改進後正確率的趨勢圖，容忍度 0.5 處正確率約為 74%。



圖三-9：諧波干擾過於強烈的範例



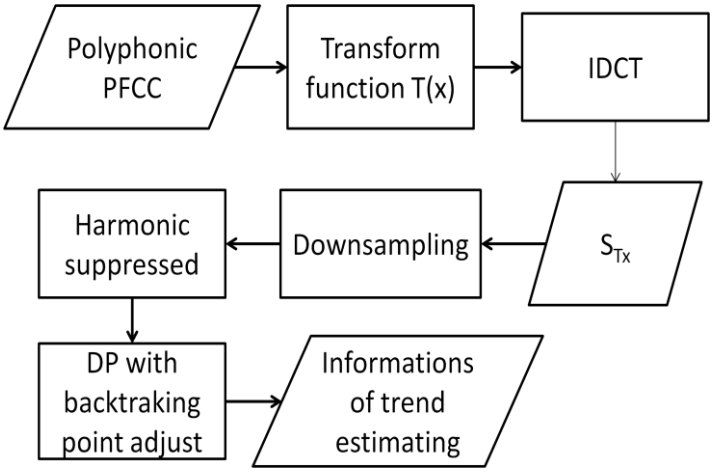
圖三-10：經過諧波壓抑後的範例



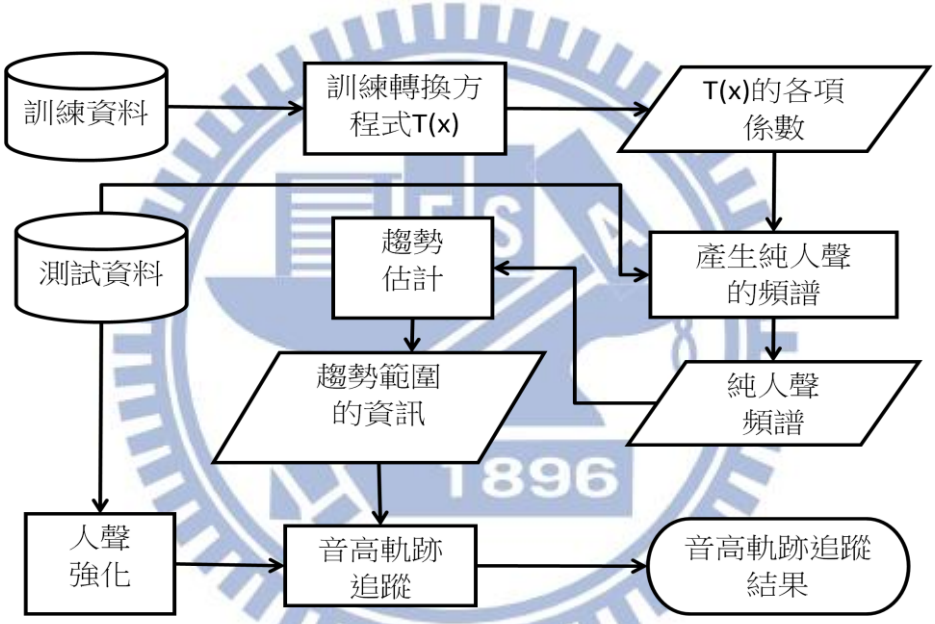
圖三-11：經過諧波壓抑後的正確率

3.3 討論

經由前述的兩個步驟，選擇特定的 backtracking 起始點和壓抑諧波干擾，如圖三-12 的趨勢估計流程圖，能把正確率從 67%提升到 75%，而[7]中的 Hsu's method 正確率為 80%，這其中的差距來自於 Hsu's method 還有一個步驟"人聲強化"，在擁有趨勢估計的資訊後，對音檔頻譜做 DP 時，原始的頻譜仍然存在很多伴奏的干擾，Hsu's method 使用 HPSS 和 NSHS 來強化原始頻譜中的人聲部份。HPSS 是一種聲音分離的技術，使人聲能夠與伴奏大致分開，NSHS 則是利用人聲在頻譜上損耗較慢的特性，累加每個頻率自己的數個倍頻，藉以凸顯諧波結構強烈的人聲。因此若要再進一步提升正確率，必定要有"人聲強化"這個步驟，如圖三-13，除了趨勢估計，還有加入"人聲強化"，這才會是完整的系統。



圖三-12：趨勢估計完整流程



圖三-13：完整系統流程

第四章 人聲強化項目整理

承接第三章的完整系統如圖三-13，在擁有趨勢估計的資訊後，尚需要人聲強化這個單元，而在[7]中是用 HPSS 和 NSHS 來達到目的。以下會討論幾種目前嘗試中效果較佳的方法，並且可以於表四-1 中看到它們與[7]中 Hsu's method 的正確率比較(正確率的計算方式同 3.1 小節，並且只關注 t 等於 0.5 的狀況)。Hsu's method 即是參考資料[7]中的方法，開頭標記Tx表示使用本論文第三章的趨勢估計方法，originSpec 表示使用原始頻譜，沒有做其他人聲強化的處理，top25bandinS_{Tx}表示從頻譜中只取出部分頻帶，這些頻帶是依據S_{Tx}(由Tx做轉換產生的人聲頻譜)每個 frame 中前 25 個能量最大的頻帶來做選擇。

表四-1：各種人聲強化之正確率比較

Title	Raw pitch accuracy rate
Hsu's method	80.18%
Tx_HPSS_NSHS	78.09%
Tx_originSpec	74.59%
Tx_originSpec_localmaxOnly	74.81%
Tx_originSpec_localmax&others× 0.5	75.06%
Tx_top25bandinS_{Tx}	74.84%

4.1 使用 HPSS 和 NSHS 的人聲強化法

Ono 等人[8]利用訊號在水平和垂直方向平滑度資訊，提出了將音樂訊號分成在時間方向平滑的 Harmonic sound 與在頻率方向平滑的 Percussive sound 兩個部分的方法，

是為 Harmonic/Percussive Sound Separation（HPSS）。HPSS 可以經由不同的分析窗口，得到不同成效的 Harmonic 和 Percussive sound，以較大的分析窗口可以得到較清晰且較純淨的 Harmonic Sound ，反之較小的分析窗口可以得到較純粹的 Percussive sound。[8] 實驗發現人聲在時間方向並沒有比被分類在 Harmonic sound 中的樂器來得平滑，但又比被分到 Percussive sound 中的樂器來的平滑，因此先使用較大的分析窗口(200ms)做 STFT 增加時間解析度，在這樣的條件下，人聲會被分到 Percussive sound 中，再使用較小的分析窗口(30ms)就可以將人聲分到 Harmonic sound 中與純 Percussive sound 分離開來。

Hermes [9]提出的 Sub-harmonic Summation（SHS）則是利用疊加諧波的方法來拉開有音高和沒音高的區域差距，如式(4-1)，$H_t(f)$為疊加固定諧波個數的結果，t 為 frame index，f 為 frequency bin，$P_t(\cdot)$代表經過 STFT 後的頻譜圖，$\boldsymbol{h}_n$ 是第 n 個諧波權重，通常設定 $h_n = h^{n-1}$, for $h \le 1$，這裡設定 h=0.98，N 為考慮的諧波總個數。NSHS 則是因應頻譜上低頻會疊加過多諧波的能量，因此考慮疊加諧波的總數，並且正規化，所以不會有低頻獨大的狀況發生，如式(4-2)。

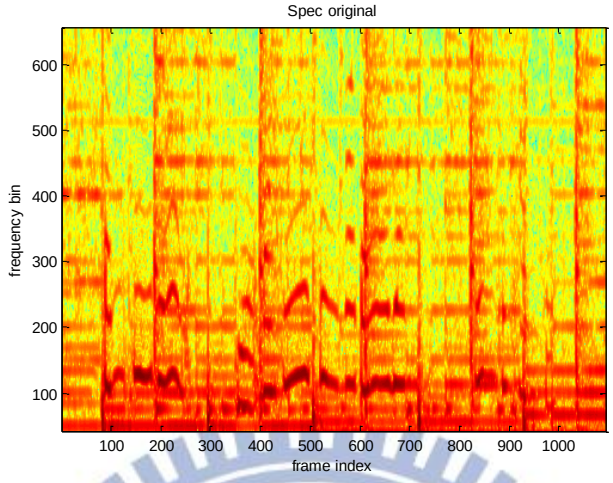
$$H_t(f) = \sum_{n=1}^{N} h_n P_t(nf) \tag{4-1}$$

$$\hat{H}_t(f) = \frac{\sum_{n=1}^{N_f} h_n P(nf)}{\sum_{n=1}^{N_f} h_n} \tag{4-2}$$

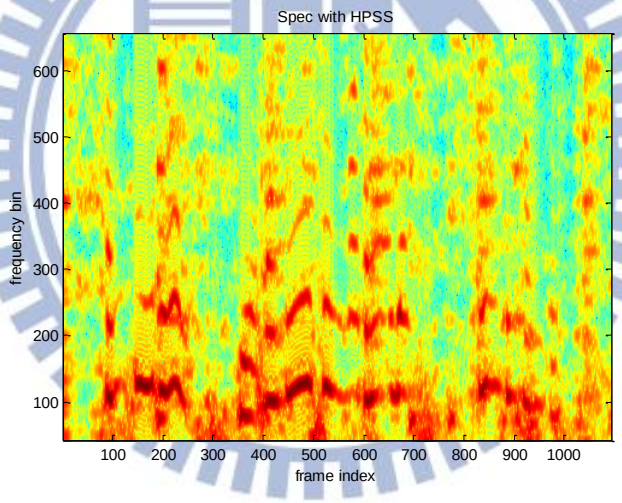
其中 $N_f = floor(\frac{0.5\,fs}{f})$ (4-3)

下列依序是圖四-1：原始的頻譜，圖四-2：做過 HPSS 後的頻譜，和圖四-3：做過 HPSS&NSHS 後的頻譜，可以看到在經過兩次處理後，比原始頻譜變得清晰且人聲在頻

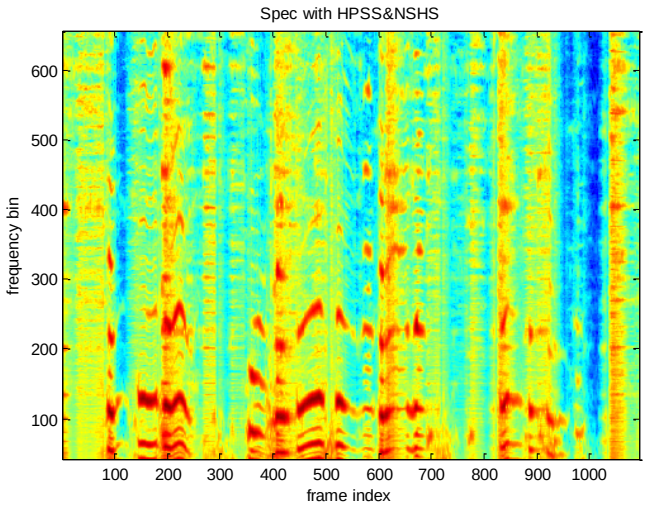
譜中的能量被凸顯，圖四-4 是一個Tx_HPSS_NSHS 的辨識範例，圖中藍線是預測的音高軌跡。



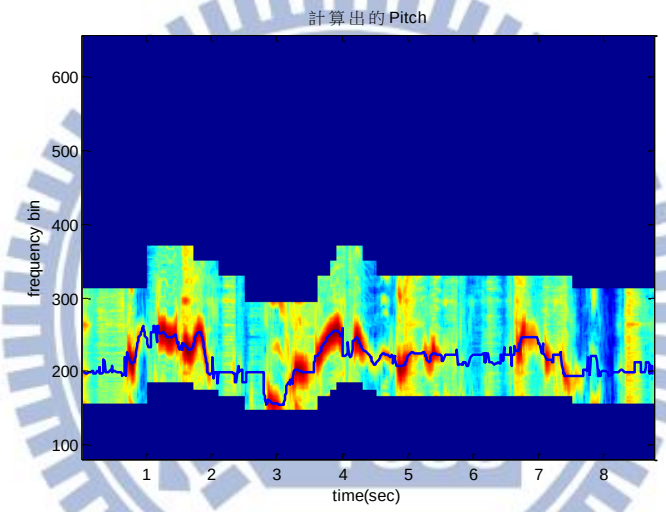
圖四-1：原始頻譜



圖四-2：做 HPSS 後的頻譜



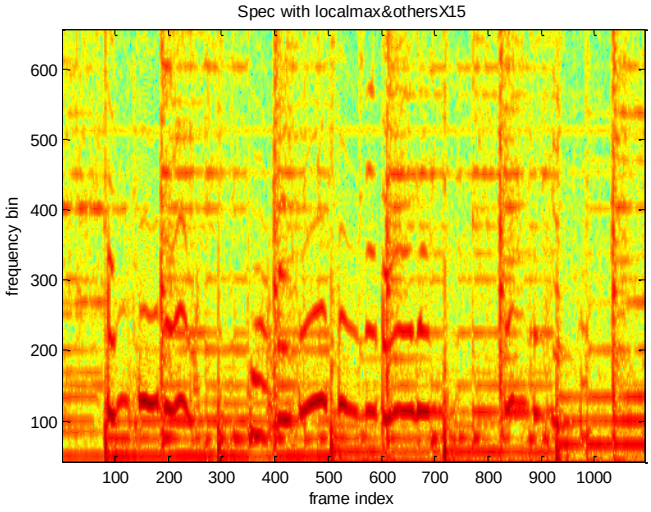
圖四-3：做 HPSS&NSHS 後的頻譜



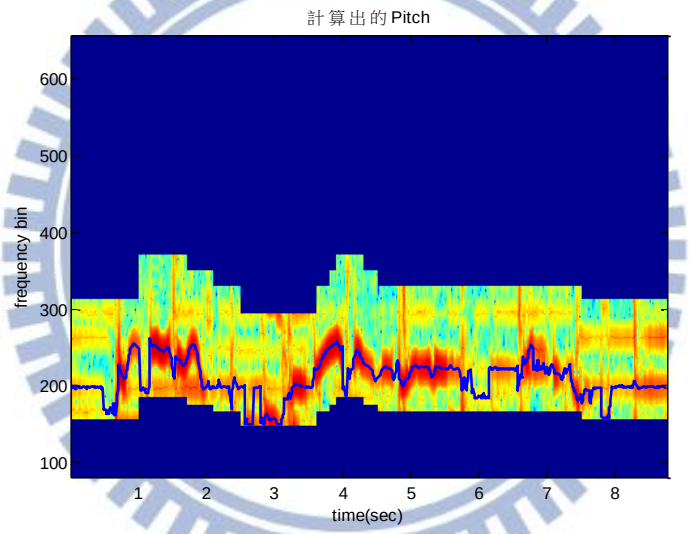
圖四-4：Tx_HPSS_NSHS 的範例

4.2 使用原訊號頻譜的 localmax

考慮到有人聲歌唱的部分會有較強的能量，因此取原始頻譜每個 frame 的 localmax，並且將其他部分壓抑乘上 0.5，這樣雖然仍有許多大於人聲能量的諧波干擾，但是趨勢估計可以濾除大部分的干擾，如圖四-5 和圖四-6，是處理後的頻譜和 Tx_originSpec_localmax&others×0.5 的一個範例。相對於圖四-4，圖四-6 因為有較多的干擾，所以頻率軌跡的波動也較大，正確率因此較低。



圖四-5：localmax&others× 0.5處理後的頻譜

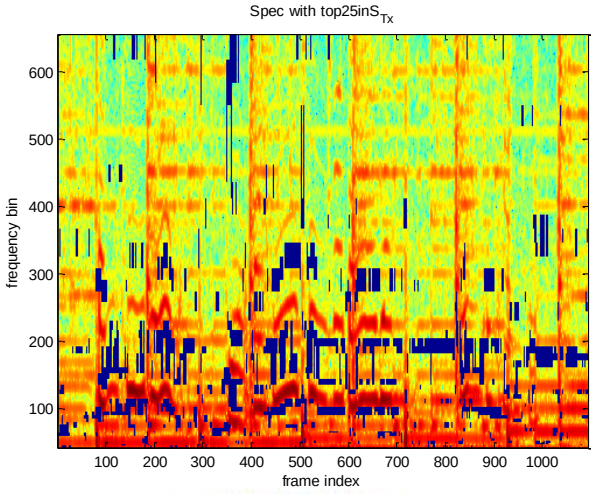


圖四-6：Tx_originSpec_localmax&others× 0.5的範例

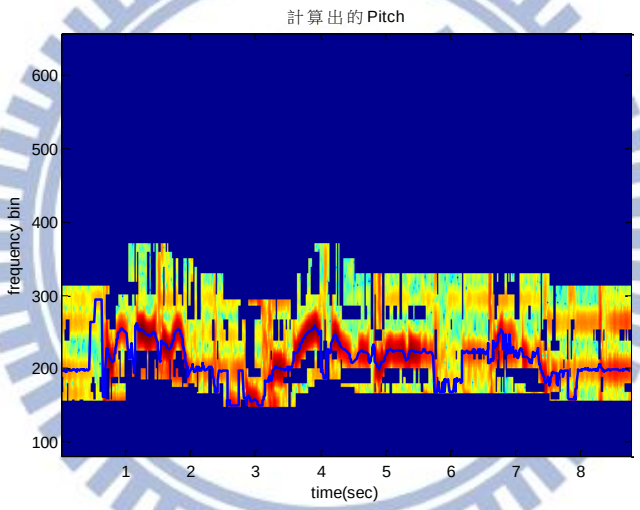
4.3 選取 S_{Tx} 中較高能量的頻帶為基準

由於本論文是使用Tx產生的資訊來進行趨勢估計，因此希望能更有效率的應用Tx衍生的其他資訊來幫助整體系統運行。S_{Tx}是由Tx做 IDCT 轉換得到的頻譜，特性就是有近似純人聲的頻譜能量分布，因此從S_{Tx}的每個 frame 中選出能量最大的前二十五個頻帶，並且對原始頻譜上保留相同的頻帶，理由是S_{Tx}上都是人聲的資訊，在原始頻譜尚保留相同的頻帶希望能夠過濾出人聲的資訊。圖四-7 是保留特定頻帶後的頻譜，可以看

到很多部位被挖空，那是被過濾掉的部分。圖四-8 是一個音高軌跡辨識結果的範例。



圖四-7：選擇 top25bandinS_{Tx} 後的頻譜

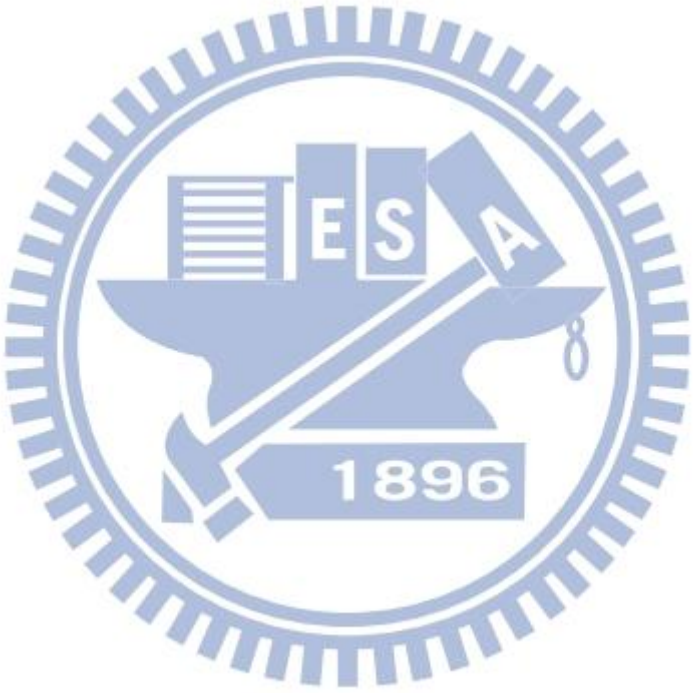


圖四-8：Tx_top25bandinS_{Tx} 的範例

4.4 討論

回到表四-1，可以看到使用 HPSS 和 NSHS 效果顯著優於其他項目，畢竟是有較完整的理論基礎而得到的系統性作法。使用 localmax 能夠高於只使用原始頻譜，可以證明去除頻譜中不必要的能量分布，有助於音高軌跡預測，因為音高軌跡近乎只發生在 localmax 上。Localmax 的方法中，壓抑不是 localmax 的能量和完全只取 localmax，差別在只取 localmax 的方法頻譜中會多了許多“空洞”，這些空洞不利於 DP 的路徑搜尋，有

時會因為空洞導致 DP 的軌跡繞了遠路才能回到正確的軌跡上。使用 top25bandinS_{Tx}則是跟純用 localmax 不相上下，因為實質上也是一種取 localmax 的思維，但是相較於直接在原始頻譜上取 localmax，top25bandinS_{Tx}保留的能量點更多，空洞少，也是正確率微高於純用 localmax 的原因。

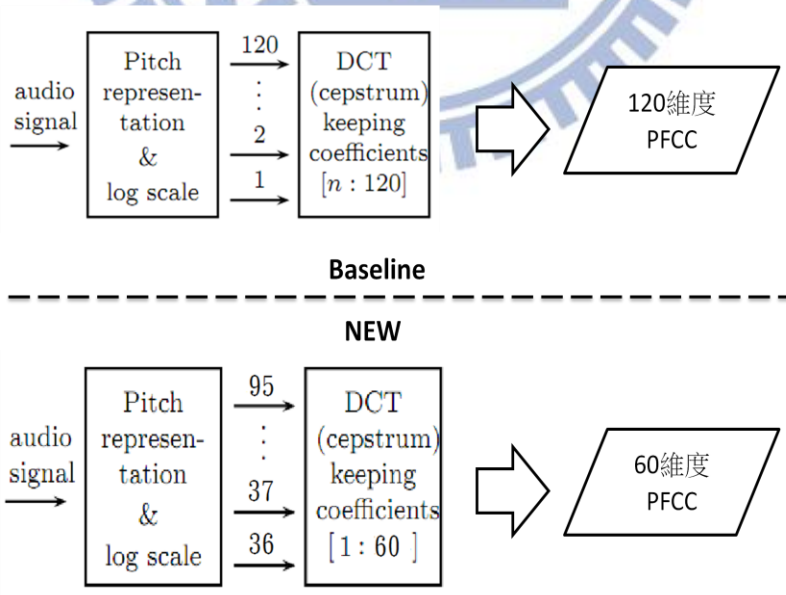


第五章 修改訓練參數以及二階段趨勢估計

基於前述章節的實驗方式，發現S_{Tx}的使用上只用到人類歌聲會出現的範圍(一般歌聲不包括極端的男女高低音)，也就是 MIDI 36 到 95，大約是 65Hz 到 1900Hz，那麼在訓練轉換方程式的時候只要能夠產生 MIDI 36 到 95 這段範圍的S_{Tx}就好，因此在對於訓練參數精簡化，去除不必要的干擾，僅對人類歌聲會出現的範圍做訓練就好。

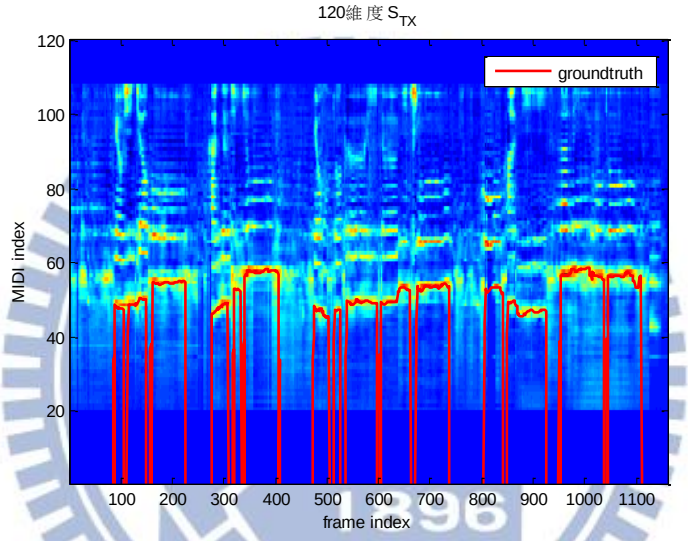
5.1 訓練參數修改

訓練過程如圖五-1，全部的步驟與過去相同，只在求參數的部分做小幅度更動。原本使用的訓練是將 120 維度的 MIDI 頻譜以 DCT 轉換得到 PFCC 參數，現在只取 36 到 95 維度，並同樣做 DCT 轉換得到 PFCC 參數，利用這樣的步驟達到去除非人聲頻段對訓練過程的干擾和錯誤訓練。

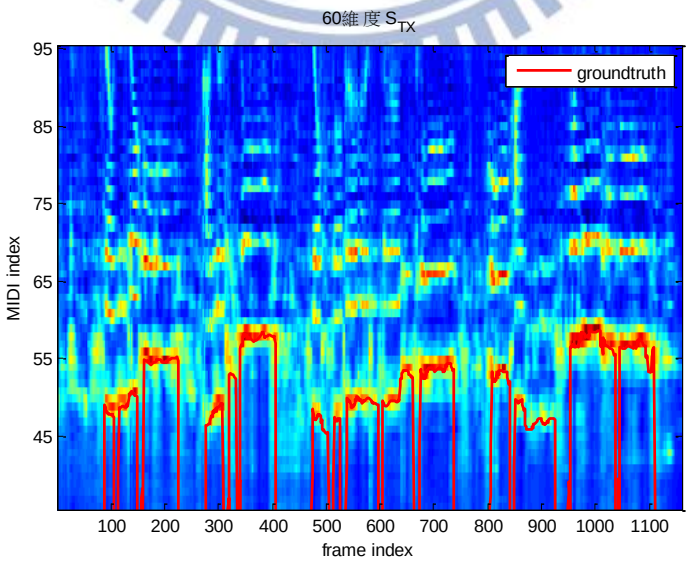


圖五-1：訓練 Tx 過程

圖五-2 和圖五-3 是原本 120 維度的S_{Tx}和修改後只剩下 60 維度的S_{Tx}，由 5-2 圖中可看到在原本 120 維度下，不在人類歌聲會出現的頻段中(36 到 95 之外)也有少量的能量響應，那都是樂器造成的，但這不是我們所關心的，我們關注的應該是頻段中同時可以有人類歌聲和樂器聲音的部分，因此對使用的參數純化，在 GMM 轉換的過程也較不容易使演算過程有不必要的錯誤訓練。在 3.2.2 小節所提到的參數 *M* 和 *N*，此時經由 try-&-error 發現分別在 *M*=9 和 *N*=0.5 時會有最佳結果。



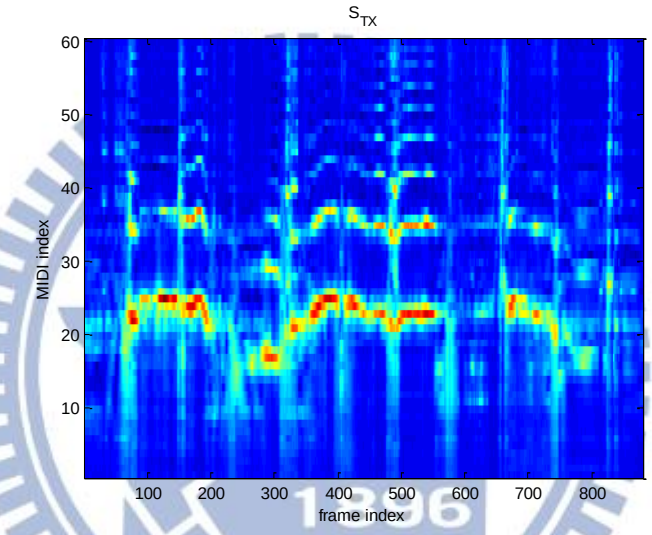
圖五-2：120 維度的 S_{Tx}



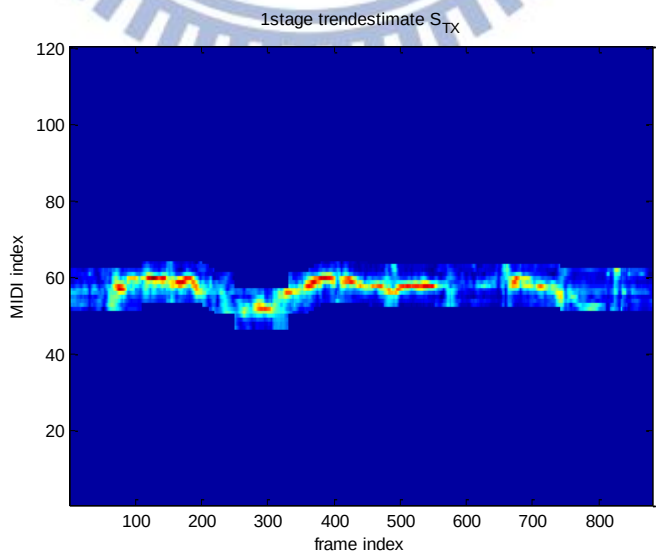
圖五-3：60 維度的 S_{Tx}

5.2 二階段趨勢估計方法

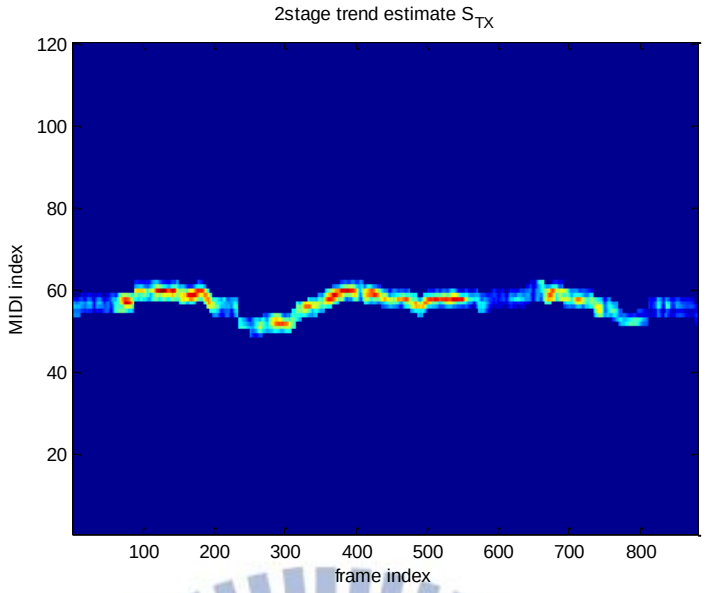
承接過去研究，希望能夠更深入的運用訓練後的產物S_{Tx}，因此設計新的趨勢估計方法。二階段趨勢估計就是在原始的趨勢估計步驟後，以相同的方法，先以 DP 尋找一條高分路徑，並擴展路徑成為範圍，但這次的寬度縮小為 5 個半元音，以得到更精準的趨勢估計結果。如圖五-4 是原始的S_{Tx}，圖五-5 是第一階段趨勢估計後的S_{Tx}，圖五-6 是第二階段趨勢估計後的S_{Tx}。



圖五-4：原始的 S_{Tx}

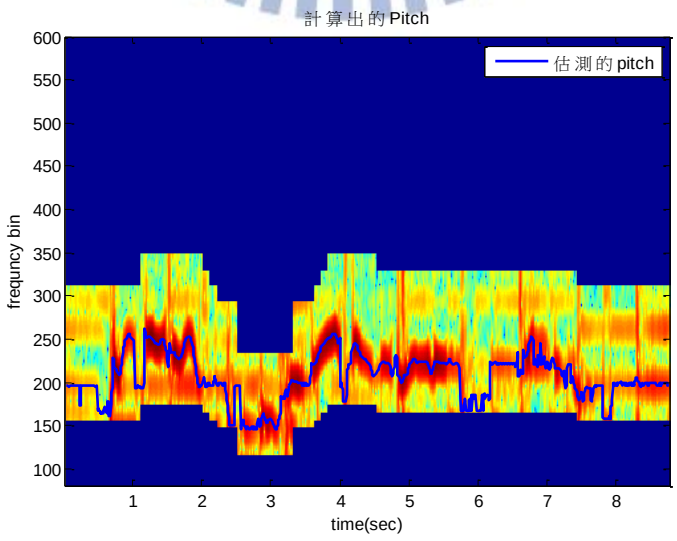


圖五-5：第一階段趨勢估計後的 S_{Tx}

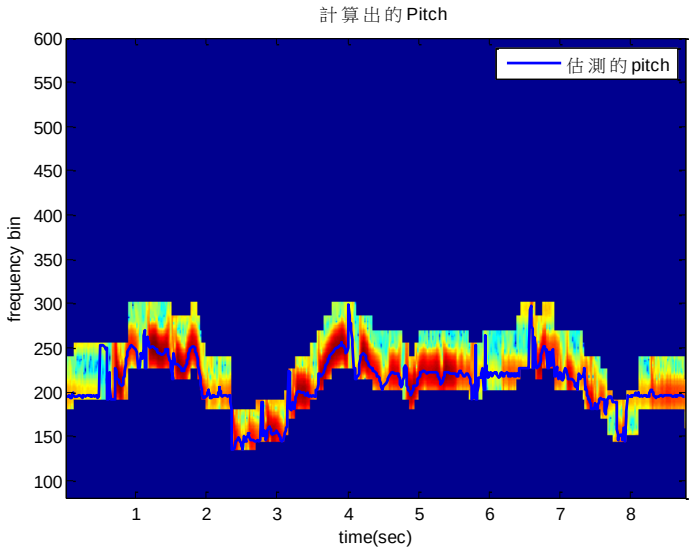


圖五-6：第二階段趨勢估計後的 S_{Tx}

在前面章節研究中發現，最後步驟的 pitch tracking 使用 DP 時，因為 DP 的限制，使 pitch 軌跡過度的平滑，導致無歌聲段進入有歌聲段時會有一小段估計在“爬”到正確 pitch 位置的現象，這樣會導致整體的 raw pitch accuracy 下降。因應這樣的現象，二階段的趨勢估計後，相信趨勢範圍應具有良好的準確性，因此可以直接對於在趨勢範圍內的頻譜每個 frame 取最大值當作 pitch 值，而不用進行 DP。圖五-7 和圖五-8 分別是“一階段趨勢估計配上 DP” 和“二階段趨勢估計配上取最大值” 的 pitch 偵測範例。



圖五-7：一階段趨勢估計配上 DP



圖五-8：二階段趨勢估計直接取最大值

5.3 實驗結果

新舊訓練方法的幾個數據做比較如表五-1，Baseline 表示之前的方法使用 120 維度訓練的轉換方程式，New 表示使用 60 維度訓練的轉換方程式。正確率的計算方式同 3.1 小節，計算 raw pitch accuracy，將預測的答案對比上正確答案，差距在 0.5 半元音的預測點算為正確，並且忽略無人聲的段落。音樂資料庫的使用同 2.3.2，訓練音檔共 480 段，測試音檔共 300 段，每段音檔約 8 秒。可以看到精簡後的訓練參數提升了趨勢估計的效果，使正確率皆較表四-1 中列出的方法有所提升。

表五-1：新舊訓練方法的比較

	Baseline	New_1stage_DP
Hsu's method	80.18%	
Tx_HPSS_NSHS	78.09%	78.79%
Tx_originSpec	74.59%	75.11%
Tx_originSpec_localmax&others× 0.5	75.06%	75.46%
Tx_top25bandinS_{Tx}	74.84%	75.48%

使用二階段趨勢估計的效果呈現在表五-2，可以看到使用二階段的總體辨識率又比原始的一段式趨勢估計來的好，而二階段趨勢估計下不使用 DP 優於使用 DP 的狀況。另外也呈現在一階段趨勢估計就不用 DP 的結果，說明在趨勢估計夠精準時，才能以峰值選取的方法來決定音高軌跡。

表五-2：新舊訓練方法與二階段估計的比較

	Baseline	New_1stage_DP	New_1stage_noDP	New_2stage_DP	New_2stage_noDP
Tx_HPSS_NSHS	78.09%	78.79%	78.06%	77.65%	79.6%
Tx_originSpec	74.59%	75.11%	74.15%	77.11%	78.12%
Tx_originSpec_localmax &others× 0.5	75.06%	75.46%	74.78%	77.21%	78.01%

接著如同表五-3，針對新舊訓練方法在趨勢估計步驟的影響對兩個部分作了統計，在正確率方面，可看到正確率低於 50%以下的音檔隨著方法精進有所減少，因此能提升整體正確率，在覆蓋率的部分有相同的結果，趨勢估計的範圍中涵蓋的正確答案少於 50%的音檔隨著方法有所減少。但是也發現在低於 25%和 10%的部分不管是正確率還是正確答案覆蓋率幾乎沒有變動，表示這幾個音檔是現在的步驟還無法處理的特例，因為在內部的參數調校時是以整體的最大成效為目標來調整，因此對於特例的狀況可能要另訂標準或是暫時予以擱置。

表五-3：新舊訓練方法的在趨勢估計上的影響

	Baseline	New_1stage_DP	New_2stage_noDP	New_2stage_noDP_H&N
Raw pitch accuracy	74.59%	75.11%	77.11%	79.6%
Number of raw pitch accuracy < 50%	89	78	66	37
Number of raw pitch accuracy < 25%	12	13	13	10
Number of raw pitch accuracy < 10%	6	6	5	5
正確答案覆蓋率	91.73%	92.59%	90.75%	90.75%
覆蓋率低於 50%的音檔數	28	22	19	19
覆蓋率低於 25%的音檔數	10	10	10	10
覆蓋率低於 10%的音檔數	4	4	5	5

*Baseline：基本方法 PFCC 還是使用 120 維度 ，無人聲強化

New_1stage_DP：新的 PFCC 使用 60 維度，做一階段的趨勢估計使用 DP，無人聲強化

New_2stage_noDP：同上但是做二階段的趨勢估計，並且不使用 DP，無人聲強化

New_2stage_noDP_H&N：同上但是頻譜部分有做 HPSS 和 NSHS 處理

第六章 結果分析與未來展望

6.1 趨勢估計結果

本論文中提出以 GMM 轉換方程式的方法來去除背景伴奏，並且使用 PFCC 來當作特徵參數，最後再提出使用二階段趨勢估計的方法。在表五-3 中可以看到每個項目的正確答案涵蓋率都在 90%以上，這個數值代表著最後計算正確率時的極限，因為只有在趨勢估計中被選到的範圍會進入下一步驟的音高追蹤。在這個部分 90%算是一個令人滿意的結果，表示大部分的正確答案都被包含在其中。要隱憂的則是在表五-3 中可以看到正確答案覆蓋率較低的音檔數量幾乎沒有減少太多，可見這些特例是目前方法不能處理的，而在趨勢估計優化的步驟中(如第三章中所提到的 *M* 和 *N*)，各項系統參數是選擇能夠使整體最佳效果的解，因此難以對特例做出微調，然而流行音樂百百種，特例必然存在，這是往後該面對的問題。

6.2 正確率結果

本論文中最好的效能是表五-3 最後一欄“New_2stage_noDP_H&N”，是使用二階段趨勢估計，並且將原始頻譜做 HPSS 和 NSHS 處理，最後不使用 DP，直接在趨勢範圍中找每個 frame 的最大值當作預測答案，正確率是 79.6%，但仍然略輸 Hsu’s method 的 80.12%，這兩種方法不同的是本論文提出的方法是以參數訓練為基礎進行的，而 Hsu’s method 完全是以訊號處理的角度來建構系統。如表六-1 顯示訓練資料量對系統正向效益的影響，而本論文中使用的訓練資料量總長度只有 40 分鐘左右(表中 C 欄位)，以現實流行音樂來說大約只有 10 首歌的長度，實則非常的少，在我們增加訓練資料至 60 分鐘後，正確率提升到 80.68% (表中 D 欄位)，由此可預期在未來擁有更多的音樂資料時，本論文的方法能夠有更好的效能。

表六-1：訓練參數量與正確率和涵蓋率的影響

	A	B	**C**	D
Raw pitch accuracy	66.97%	73.59%	**79.6%**	80.68%
Number of raw pitch accuracy < 50%	199	124	**37**	33
Number of raw pitch accuracy < 25%	97	46	**10**	4
Number of raw pitch accuracy < 10%	43	13	**5**	1
正確答案覆蓋率	79.63%	84.57%	**90.75%**	91.16%
覆蓋率低於 50%的音檔數	146	79	**19**	19
覆蓋率低於 25%的音檔數	64	29	**10**	4
覆蓋率低於 10%的音檔數	30	11	**5**	1

*A 共使用 64727 特徵向量，總長約 10 分鐘
B 共使用 126430 特徵向量， 總長約 20 分鐘
C 共使用 253455 特徵向量， 總長約 40 分鐘
(本論文最終使用的資料量)
D 共使用 365780 特徵向量，總長約 60 分鐘

參考資料

[1] Wei-Ho Tsai and Hao-Ping Lin “Background Music Removal Based on Cepstrum Transformation for Popular Singer Identification”, IEEE Trans. on Audio, Speech, and Language Processing, vol. 19, no. 5, July 2011.

[2] 宋柏毅 “以韻律模型為基礎之中文韻律轉換研究”, 交通大學碩士論文, 2009.

[3] Sebastian Ewert and Meinard Muller “Chroma Toolbox: Matlab Implementations for Extracting Variants of Chroma-Based Audio Features”, ISMIR 2011.

[4] Sebastian Ewert, Meinard Muller and Michael Clausen “Towards Timbre-Invariant Audio Features for Harmony-Based Music”, IEEE Trans. on Audio, Speech and Language Process., Vol. 18, No. 3, pp.649 -662, 2010.

[5] Chao-Ling Hsu and Jyh-Shing Roger Jang, “https://sites.google.com/site/unvoicedsoundseparation/mir-1k” , (n.d.).

[6] Mike Brookes, “http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html”,(n.d.).

[7] 白宗儒 “一個是用於複音音樂之音高追蹤的混乘法”, 清華大學碩士論文, 2011.

[8] N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka, and S. Sagayama, “Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram”, Proc. of EUSIPCO, 2008.

[9] Hermes, D. J. (1988). "Measurement of pitch by subharmonic summation," J. Acoust. Soc. Am. 83, 257-264.

[10] M. Goto, “A Real-Time Music Scene Description System: Predominant-F0 Estimation Detecting Melody and Bass Lines in Real-World Audio Signals”, Speech Communication, vol. 43, no. 4, pp.311–329, 2004.

[11] Chao-Ling Hsu and Roger Jang, “Singing Pitch Extraction at Mirex 2010”, The Music Information Retrieval Evaluation Exchange (MIREX), 2010.

[12] Justin salamon and Emilia Gomez, “Melody Extraction from Polyphonic Music Signals using Pitch Contour Charateristics”, IEEE Trans. on Audio, Speech, and Language Processing, 2012

[13] A. Klapuri, “Multiple fundamental frequency estimation by summing harmonic amplitudes,” in Proc. 7th Int. Conf. on Music Inform. Retrieval,Victoria, Canada, Oct. 2006

[14] G. Poliner and D. Ellis, “A classification approach to melody transcrip-tion,” in Proc. 6th Int. Conf. on Music Inform. Retrieval, London, Sep.2005

[15] M. Lagrange, L. G. Martins, J. Murdoch, and G. Tzanetakis, “Normalized cuts for predominant melodic source separation processing,” IEEE Trans. on Audio, Speech and Language Process., vol. 16, no. 2, pp.278–290, Feb. 2008

[16] Yannis Panagankis ,Constantine Kotropoulos and Gonzalo R. Arce, “l_1-Graph Based Music Structure Analysis”, 12th International Society for Music Information Retrieval

Conference (ISMIR), 2011.

[17] Joan Serr`a, Emilia G´omez, Perfecto Herrera and Xavier Serra, “Chroma Binary Similarity and Local Alignment Applied to Cover Song Identification”, IEEE Trans. on Audio, Speech, and Language Processing, 2008.

[18] Cynthia C.S. Liem and Alan Hanjalic, “Cover Song Retrieval: a Comparative Study of System Component Choices”, 10th International Society for Music Information Retrieval Conference(ISMIR), 2009.

[19] D. Ellis and G. Poliner “Identifying Cover Songs with Chroma Features and Dynamic Programming Beat Tracking”, IEEE Trans. Conf. on Acoustics, Speech and Signal Processing(ICASSP), 2007.

[20] Müller and Meinard, “Information Retrieval for Music and Motion (Chapter 3)”, 2007.

[21] Frank Kurth and Meinard Müller, “Efficient Index-Based Audio Matching” IEEE Trans. on Audio, Speech, and Language Processing, Vol. 16, No. 2, 2008.

[22] Riccardo Miotto and Nicola Orio, “A Music Identification System Based on Chroma Indexing and Statistical Modeling”, ISMIR Content-Based Retrieval, Categorization and Similarity, 2008.

[23] D. Ellis, “Classifying Music Audio with Timbral and Chroma Features,” International Society for Music Information Retrieval Conference(ISMIR), 2007.

[24] George Tzanetakis and Perry Cook, “Musical Genre Classification of Audio Signals,” IEEE Trans. on Audio, Speech, and Language Processing, Vol. 10, No. 5, 2002.

[25] Yuxiang Liu, Qiaoliang Xiang, Ye Wang and Lianhong Cai, “Cultural Style Based Music Classification of Audio Signals” IEEE Trans. Conf. on Acoustics, Speech and Signal Processing(ICASSP), 2009.

[26] D. Ellis, “Beat Tracking by Dynamic Programming”, J. New Music Research, Special Issue on Beat and Tempo Extraction, Vol. 36, No. 1, pp. 51-60, 2007.