國　立　清　華　大　學

碩 士 論 文

題目: 一個適用於複音音樂之音高追蹤的混成法

A Hybrid Method for Pitch Tracking of Polyphonic Audio Music

系別　資訊工程學系　　組別

學號姓名　9862576　白宗儒　(Tsung-Ju Pai)

指導教授　張智星博士　(Jyh-Shing Roger Jang)

中華民國一百年六月

摘要

一首音樂的主旋律通常很容易就能被人類辨識，尤其是歌唱類型的音樂，在這類的音樂中歌唱音高通常就是主旋律。但是要利用電腦來直接辨識音樂中的歌唱音高是一件相當困難的事情，對於電腦來說背景音樂就像是干擾人聲的雜訊。在本論文中我們使用了一個基本方法，此方法是以一連串的頻譜分析演算法所構成，大部分的演算法目的都在於提高人聲並且降低音樂，藉此使動態規劃方法的音高追蹤準確度提高。但此方法在人聲端點與頻率快速變化區域容易得到錯誤的音高，所以我們利用反轉訊號的時間軸得到相異的結果，並使用疊合找出不穩定的音高，再輔以隱藏式馬可夫模型訓練的音高抽取方法，使用投票法來對不穩定的音高進行修正。在本論文的方法裡，我們改進了基本方法在弱點區域的準確度，使得整體辨識率得到明顯的提升。

Abstract

Human can easily recognize the main melody of a piece of music, especially of a song with singing voice, because pitch of singing voice usually represents the main melody of a song. However, it is not as easy for a computer to automatically detect the singing pitch from a song, because the background music acts as an interfering signal to the singing voice. In this thesis we propose a method that is composed of a series of spectrum analysis algorithms. Most of the existing algorithms focus on enhancing singing voice while reducing the background music so that the accuracy of singing pitch extraction can be improved. But these methods tend to yield incorrect pitch values near both endpoints of singing voice and sound segments with fast-varying frequencies. We therefore reverse the time axis of the song signal and overlap it with the original signal to find the segments that have unreliable pitch values. This method is assisted with another singing pitch extraction method incorporating hidden Markov models. A voting mechanism is adopted to justify and correct the unreliable pitch values. Experimental results show that the proposed method yields a better performance than the original baseline system in terms of raw pitch accuracy.

謝誌

首先相當感謝我的父母，在求學的過程中給予經濟上與精神上的支持，有你們當我的後盾才能夠有今天的成果。再來十分感謝我的指導教授張智星教授，讓我可以選擇一直很有興趣的音樂相關題目來進行研究，興趣一直以來都是我做事情的動力，這對於研究來說有推波助瀾的作用。除此之外相當感謝許肇凌博士對我在研究上的幫助，有耐心地將完全不懂的我帶入門，這是能夠完成研究的另一個重要原因。感謝張魚哥（張雅雯）提供了論文的樣板，使我不需要花太多時間在惱人的排版工作上，間接地加快了論文的寫作速度。感謝 MIR 實驗室的每個人在碩士班兩年裡帶給我許多歡樂，碩班的部分有常常幫買點心遊戲人間的小杜（杜承恩），不太正經卻對某些東西很有一套的紅西哥（曾泓熹），幫我排憂解悶其實沒什麼嫌隙的高射砲（陳宏瑞），奮發向上的標竿羅傑洞洞（黃弈銘），懂很多卻保有赤子之心的西歪欸司（陳揚昇）。博班則有常常幫我出點子開竅的阿諾（吳明儒），實驗室守護神小龜（王崇喆），在自然語言和研究上幫助我許多的葉子（葉子雋），不辭辛勞什麼問題都可以問的 Davidson（陳亮宇），在大四音訊課時幫助我許多的 Kevin（任佳珉）。要感謝的人實在太多無法一一列上，所以最後相當感謝在求學路上與我相遇的每個人，由於和你們相遇才能夠走到今天，謝謝！

目錄

表目次

圖目次

第一章　緒論

在音樂中主旋律是最容易被人類所辨識的部分，尤其含有歌唱成分的音樂，此類音樂的主旋律多半就是歌唱的音高。歌唱音高有許多方面的應用，例如以音高為基礎的人聲抽取或是歌曲檢索，更直接一點的則有歌者音準訓練。但要在音樂裡找到旋律對於機器來說是相當困難的事情，伴奏用的背景音樂就像是雜訊一般干擾著人聲，如同語音辨識一般，我們會希望雜訊越少越好。雖然有許多研究致力於歌唱音高抽取（Singing Pitch Extraction），但是仍然無法做到完美。

自從 1999 年 Goto[1]提出第一個以統計訓練參數模型的旋律抽取方法後，越來越多的方法[2]被投注在旋律抽取的研究裡。由於 Harmonic Structure 在頻譜上很容易被辨識出來，因此有部分的研究使用 Harmonic Structure 做為歌唱旋律抽取的標示，不過在這些研究中，他們忽略了背景音樂的資訊，以致於背景音樂仍然會干擾到人聲。

而 Ryynänen[8]使用了音樂學模型來建立隱藏式馬可夫模型（Hidden Markov Model）進行歌唱旋律抽取，這個方法會判斷一首歌的調性，調性決定了關係大小調（例如 C 大調與 Am 小調），每個調性都有各自的主要和絃，有了和絃之後便可以用來來計算相鄰音符的移轉機率（Transition Probability）。而 Li 等人也一樣使用了隱藏式馬可夫模型，不同於 Ryynänen，在這個模型中的移轉機率是計算自人工所標記的正確答案。而 Hsu 等人[9]在 2009 同樣使用了隱藏式馬可夫模型，不同以往，他使用了 2-Stream 的隱藏式馬可夫模型，其中之一使用梅爾倒頻譜係數（Mel-scale Frequency Cepstral Coefficients，MFCC）來進行人聲端點偵測，另一半則使用 NSHS（Normalized Sub-harmonic Summation）來進行音高抽取。

訓練的方法好處在於訓練的資料越多，則準確率會越高，由於語音相關研究在訓練方法上有相當高的準確率，因此投注在訓練方法上的歌唱旋律抽取也相當的多。除了訓練方法之外，也有一些研究投注在頻譜分析的方法上，例如 Hsu 等

人[10]在 2010 年的 MIREX 當中使用了一連串的頻譜分析演算法，降低頻譜上不可能是歌聲的部份，並加強頻譜上歌聲的部分，最後以動態規劃的方式取得歌唱音高。這個方法有相當高的準確率，但是在頻譜上歌聲與樂器聲易混淆區域卻相對顯得較為不精準。

訓練法和頻譜分析法各有其優缺點，在本論文中將使用 Hsu 在 2010 年所提出的方法為基礎，使用倒轉音樂訊號時間軸的方法來判定不穩定區域，以方便進行修正。我們實作了一個以隱藏式馬可夫模型為概念的方法做為修正時的第三決策者，這個方法相較起 Hsu 的方法來說雖然準確度較低，但在人聲邊緣與音高快速變化區域卻有較高的準確度。我們希望透過這個方法結合訓練法與頻譜分析法，使他們能夠補足彼此的缺點，以提升整體辨識率。

本論文的第二章將介紹整個系統架構與函式功能，第三章則會展示實驗相關結果與改進成果，第四章將會對整個方法做一個結論，並且討論未來可以改進的方向。

第二章　研究方法

2.1 系統概觀

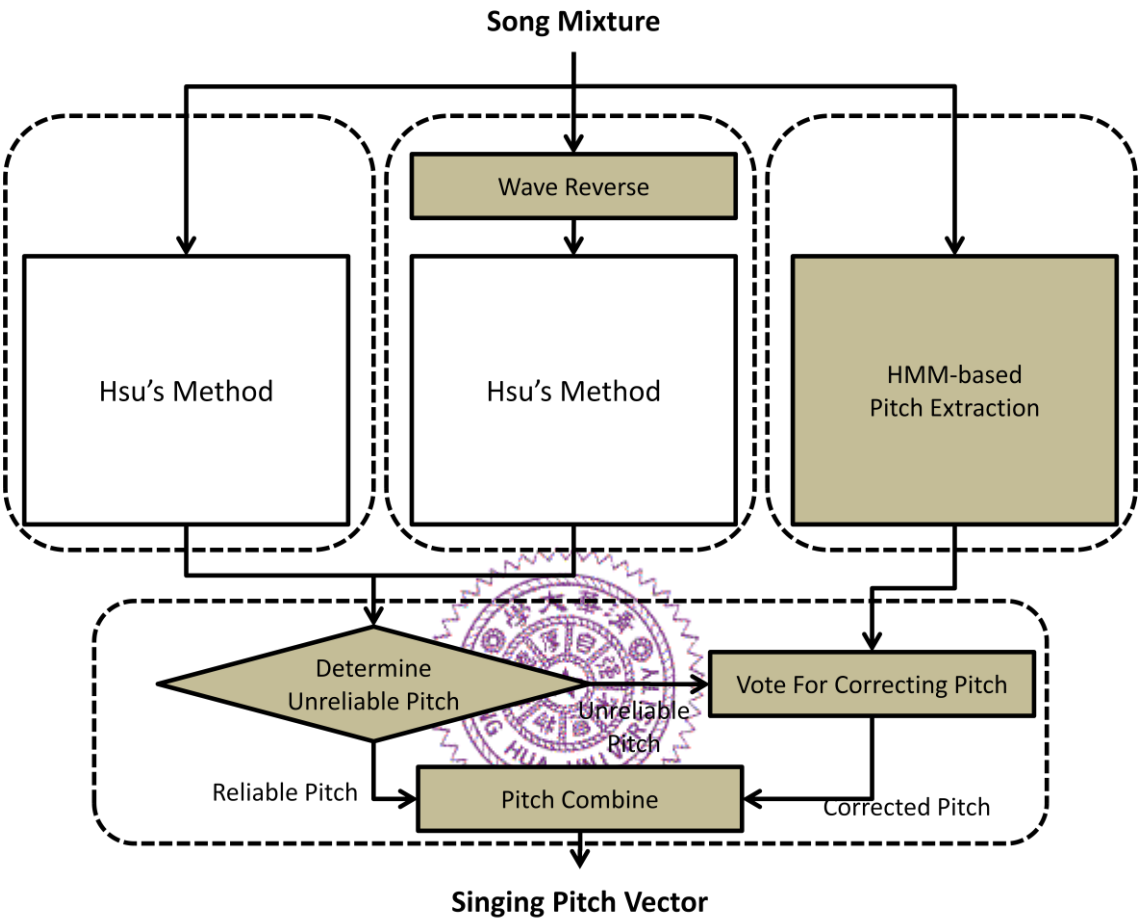


圖 2.1.1 本論文系統架構圖

圖 2.1.1 是本論文所使用的系統架構圖，大略可分成四個區塊，最左邊的區塊是 Hsu 在 2010 年 MIREX 中所使用的方法（在本論文中以 Hsu's Method 稱呼之），是一個多步驟音高抽取的系統。中間的部分增加了反轉波形（Wave Reversal），目的在於取得反向波形的結果以做為尋找判定不穩定音高時使用。下方的區塊會對正反向的 Hsu's Method 結果做疊合比對，找出可能的不穩定音高，由右方以隱藏馬可夫模型訓練為概念的方法來進行投票修正。下面的小節將會對上述所提到的區塊做詳細的介紹。

2.2 Singing Pitch Extraction

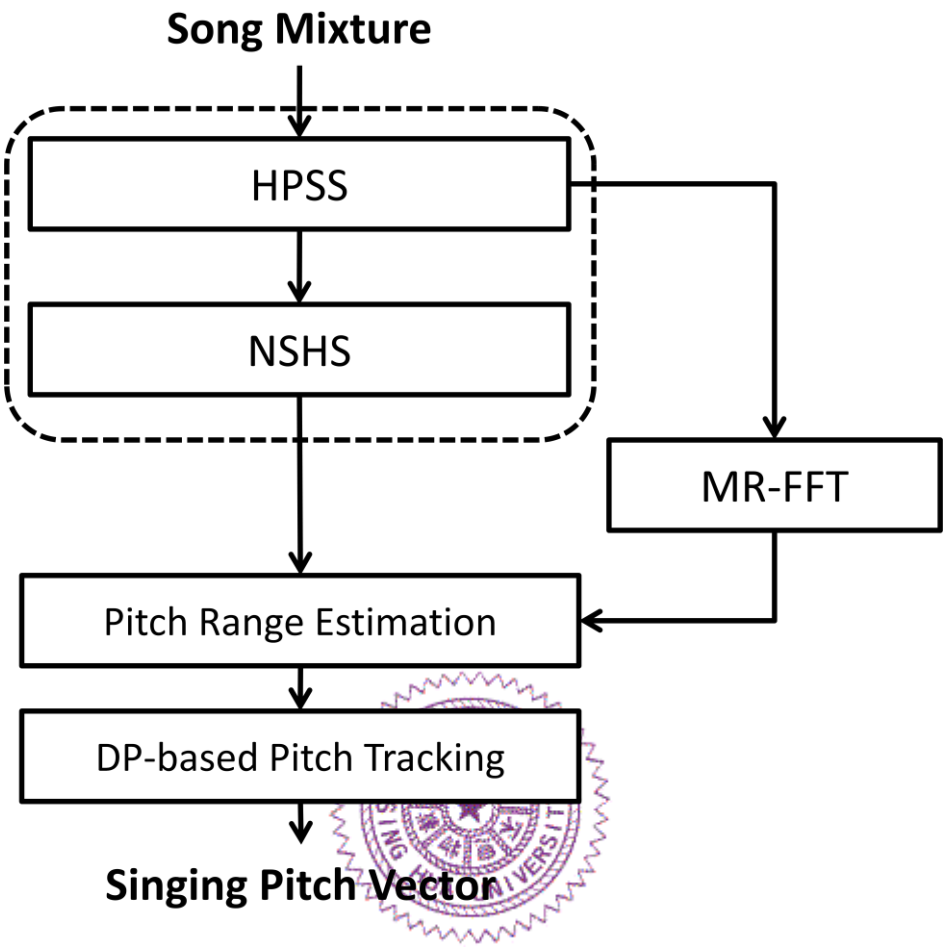


圖 2.2.1 Hsu's Method 系統架構圖

Hsu's Method 可以大略分成四個部分，消減背景音樂的 Harmonic/Percussive Sound Separation（HPSS），將頻譜上有音高與沒音高差距加大的 Normalized Sub-harmonic Summation（NSHS），推估音高趨勢路線的 Pitch Range Estimation 與最後的 DP-based Pitch Tracking。

2.2.1 Harmonic/Percussive Sound Separation

Ono 等人[13]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）。此方法假設訊號的功率譜W可以以下列式子表示：

$$H(m,f) + P(m,f) = W(m,f) \quad (1)$$

$H(m,f)$和$P(m,f)$分別為 Harmonic 和 Percussive 部分。m 和 f 則代表了音框（Frame）與頻段（Frequency Bin）的標示值，圖 2.2.2 是一個 HPSS 的簡單概念範例圖。

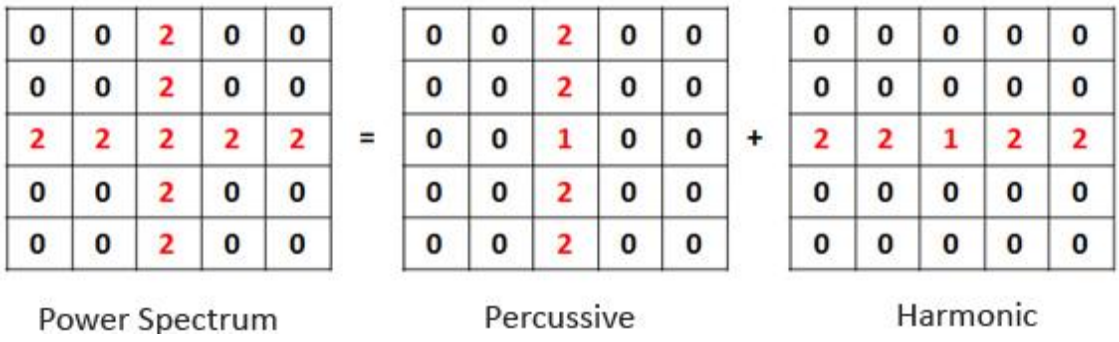


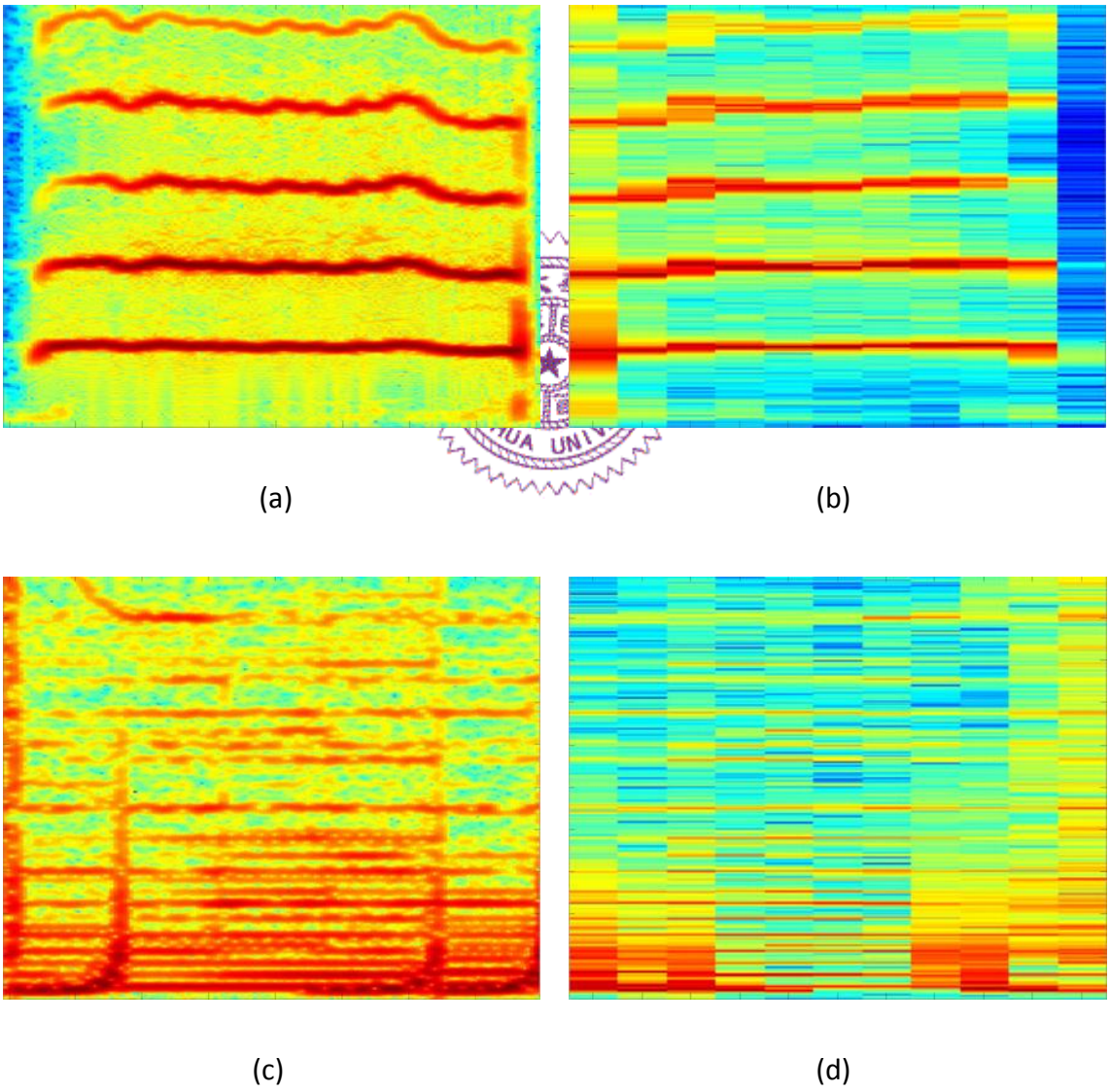
圖 2.2.2 HPSS 簡易概念

在上述的前提條件下 HPSS 可以看作是一個將下列式子化為最小值的問題：

$$\int\int \left(\frac{\partial}{\partial m}\left|H(m,f)\right|^{\gamma}\right)^2 dmdf + \int\int \left(\frac{\partial}{\partial m}\left|P(m,f)\right|^{\gamma}\right)^2 dmdf \,, \gamma \cong 0.6 \quad (2)$$

而後 Tachibana 等人[12]延伸了原始的 HPSS，將其變成一個多步驟的新版本，專門使用人聲上。最主要的原因是發現了人聲在時間方向並沒有比分類在 Harmonic 中的樂器來得平滑，另一方面又比 Percussive 中的樂器來得平滑。針對這個觀察使用了較大的 Window Size 來計算 Short Time Fourier Transform（STFT），

這個改變使得頻譜在時間方向的解析度降低，如圖 2.2.3 所示，(a)(c)的 Window Size 為 30ms，而(b)(d)為 256ms。原本難以分辨的人聲與樂器，再使用大 Window Size 後差異就會變得較為明顯。在這樣的條件下，人聲會被分到 Percussive 中，而且少了許多 Harmonic 樂器，人聲變得較為乾淨。接著只要再使用較小的 Window Size（例如 30ms），就可以將人聲與 Percussive 分開來。Hsu's Method 的目標結果是歌聲的音高，而敲擊樂器並沒有音高，因此就算留存許多的敲擊樂器也不會對結果造成影響。



(a) (b)

(c) (d)

圖 2.2.3 人聲樂器在不同傅立葉轉換框大小下的差異

2.2.2 Normalized Sub-harmonic Summation

有兩種方法在頻譜上可以使音高提取的準確率提高，一是提高可能是人聲區域的能量，二是盡量降低不可能是人聲區域的能量。Hermes[14]提出的 Sub-harmonic Summation（SHS）就是利用疊加 Harmonic 的方法來拉開有音高和沒音高的區域差距，SHS 的方程式如下：

$$H_t(f) = \sum_{n=1}^{N} h_n P_t(nf) \quad (3)$$

$H_t(f)$是在 Frame Index 為 t，Frequency 為 f 時 SHS 的值，N 是列入考慮的 Harmonic 數量，$P_t(*)$代表經過 STFT 計算後的功率譜，h_n是第 n 個 Harmonic 的權重，一般而言我們會設定$h_n = h^{n-1}, \mathrm{h} \leq 1$。

由於人聲的 Harmonic 在頻譜上損耗地相當慢，基於這個特性 Hsu's Method 將所有的 Harmonic 都列入疊加考慮，這與原始的 SHS 不同，原始的 SHS 只考慮了固定數量的 Harmonic，但如果只考慮全部的 Harmonic 而不做正規化（Normalized），則低頻疊加的結果能量會特別高，將會大大影響結果，所以在 Hsu's Method 中使用了 Normalized Sub-harmonic Summation（NSHS），其方程式如下：

$$\hat{H}_t(f) = \frac{\sum_{n=1}^{N_f} h_n P_t(nf)}{\sum_{n=1}^{N_f} h_n} \quad (4)$$

而 Harmonic 的數量N_f是由疊加時的基頻頻率而定：

$$N_f = floor\left(\frac{0.5 f_s}{f}\right) \quad (5)$$

上述式子的f_s為取樣率。 NSHS 在單一音框上的效果可以參考圖 2.2.4，明顯地看得出來原本頻譜上具有音高的部分，經過 NSHS 處理後變得較為尖細，讓有音高的部分更加準確，也拉開了有音高與無音高部分的差距。而從圖 2.2.5，圖 2.2.6

與圖 2.2.7 可以看出 SHS 與 NSHS 在頻譜上的效果差異，SHS 在低頻處明顯較 NSHS 來的雜亂，在最後的動態規劃音高追蹤上會直接地影響其準確度。

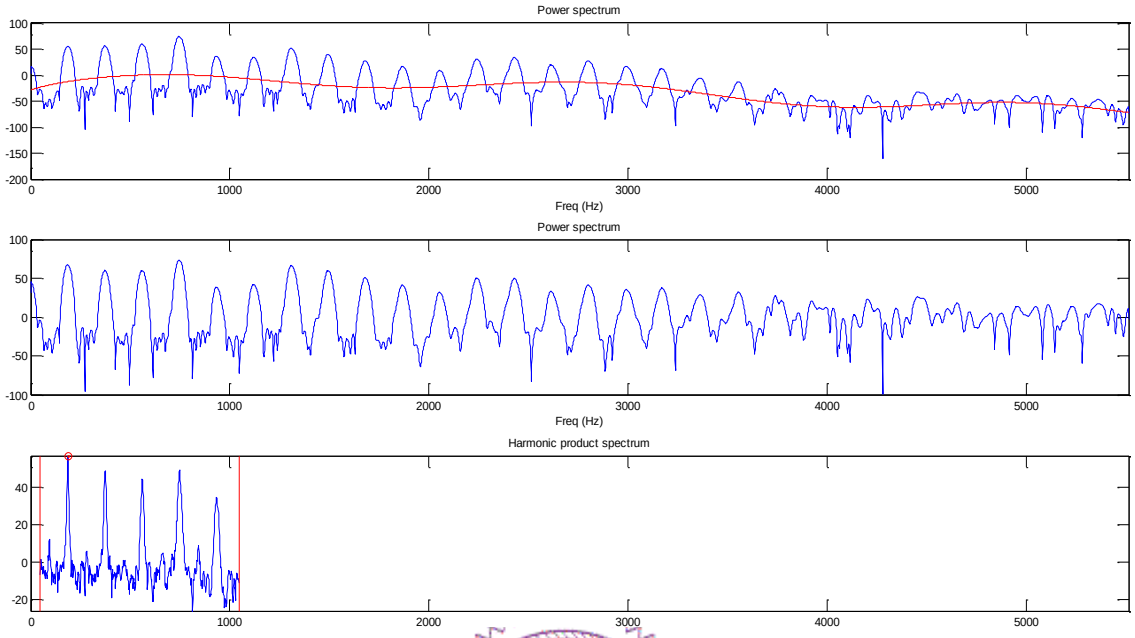


圖 2.2.4 單一音框 NSHS 效果，最上方的圖代表原始的 Power Spectrum，中間為 Power Spectrum 扣除 10 次多項式逼近後的結果，最下面則為最後的 NSHS 結果

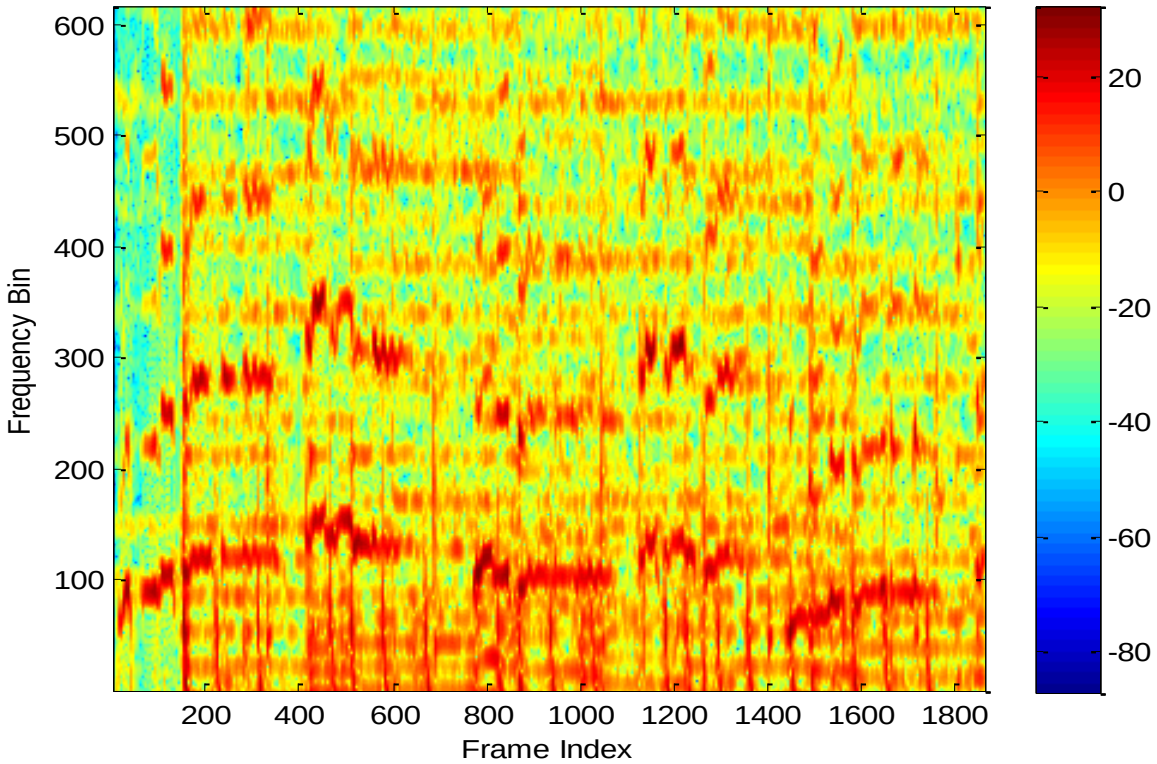


圖 2.2.5 原始頻譜

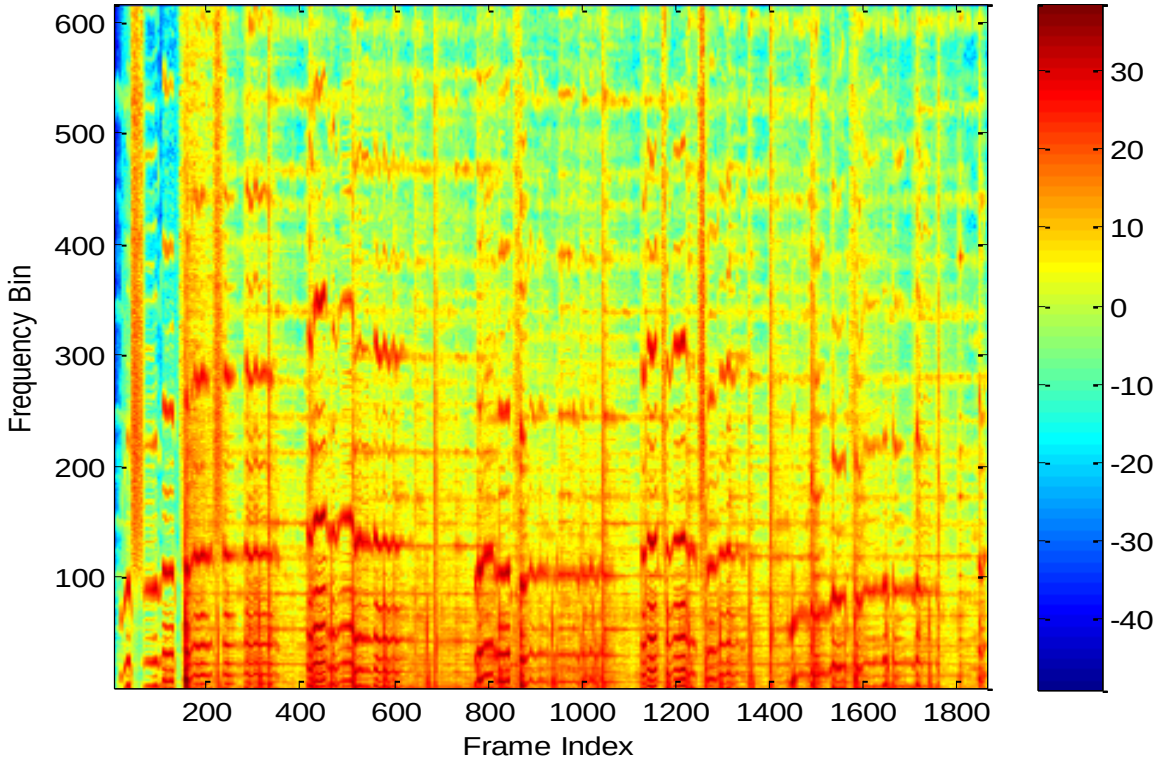


圖 2.2.6 SHS 效果

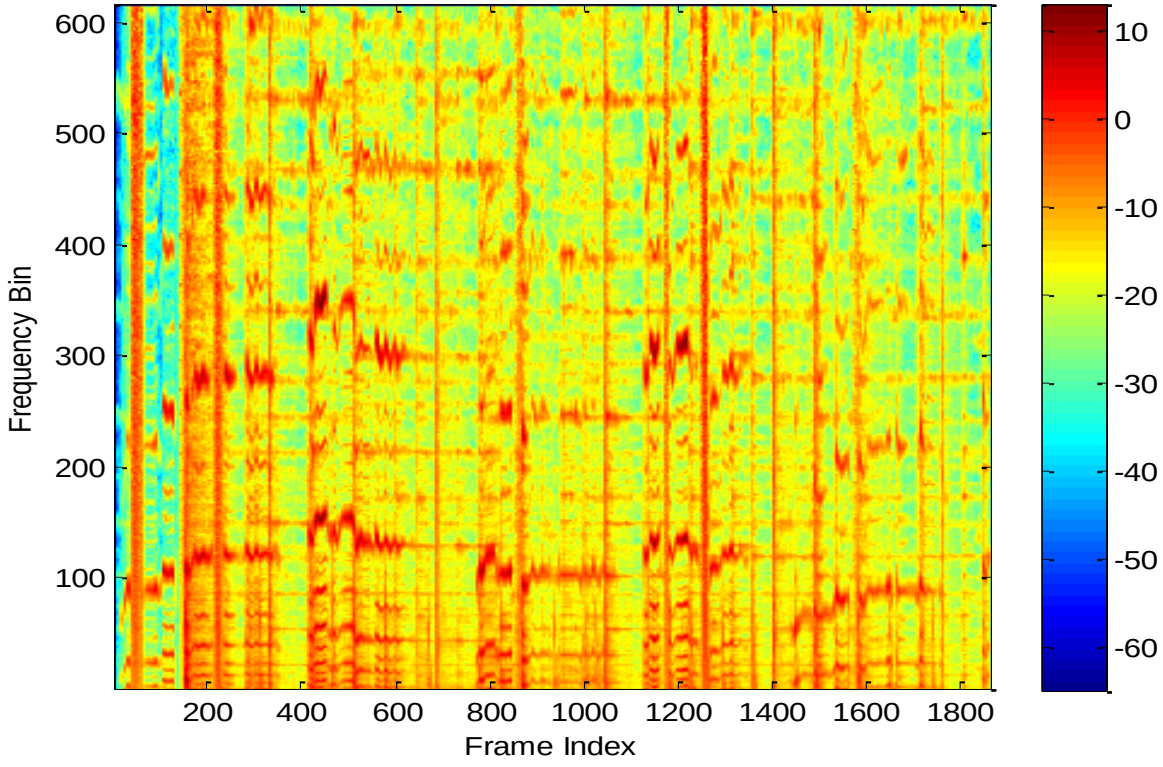


圖 2.2.7 NSHS 效果

2.2.3 Trend Estimation and Pitch Extraction

Trend Estimation 的目的在於建立一個遮罩，使頻譜留下歌聲音高行進的可能範圍，這對於最後進行動態規劃取得歌聲音高有莫大的幫助，因為在沒有指定範圍的情況下，電腦無法直接分辨出基頻與倍頻的差異，會使動態規劃音高追蹤時抓取到倍頻的音高，這不是我們所樂見的。Hsu's Method 借助 Dressler 所提出的 Multi-resolution Fast Fourier Transform（MR-FFT）來刪除頻譜上不可靠的峰值，這些峰值多半不是來自於主要的聲音訊號（例如：雜訊），其結果如圖 2.2.8，頻譜看起來會較為「乾淨」。

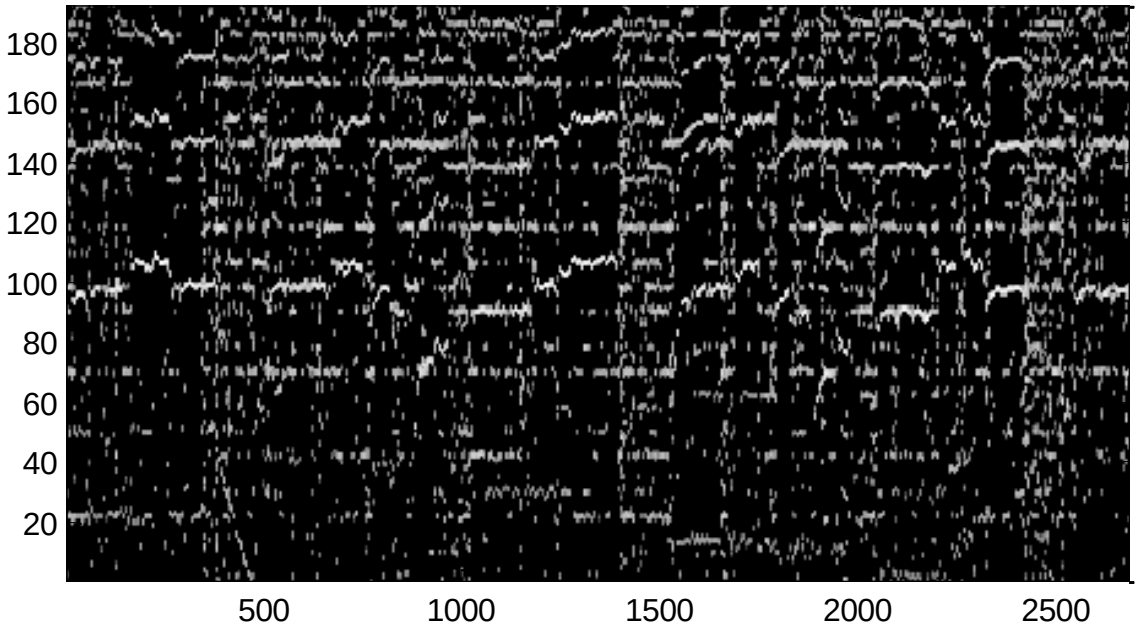


圖 2.2.8 MR-FFT

由於頻段有其極限，頻率在頻譜上的解析度永遠不夠多，所以會出現洩溢的狀況，在頻譜上看起來訊號的能量會像是水墨畫渲染開來，而不是集中在某一個頻段上。透過觀察每個峰值的音框瞬間頻率（Instantaneous Frequency），找出瞬間頻率相近（小於 0.2 個半音）峰值，選擇能量（Magnitude）最強的一個留下，將其餘的刪去。但頻譜上仍然殘留有許多不需要的峰值，有些是未被刪去的不可靠峰值，更多的部分則是 Harmonic 的峰值。而我們知道基頻（*F0*）只會是最低的頻

率，根據這個定則刪除基頻的 Harmonic（Overtone Deletion），結果如圖 2.2.9 所示，這時候已經可以觀察出歌聲音高大略的路徑了。

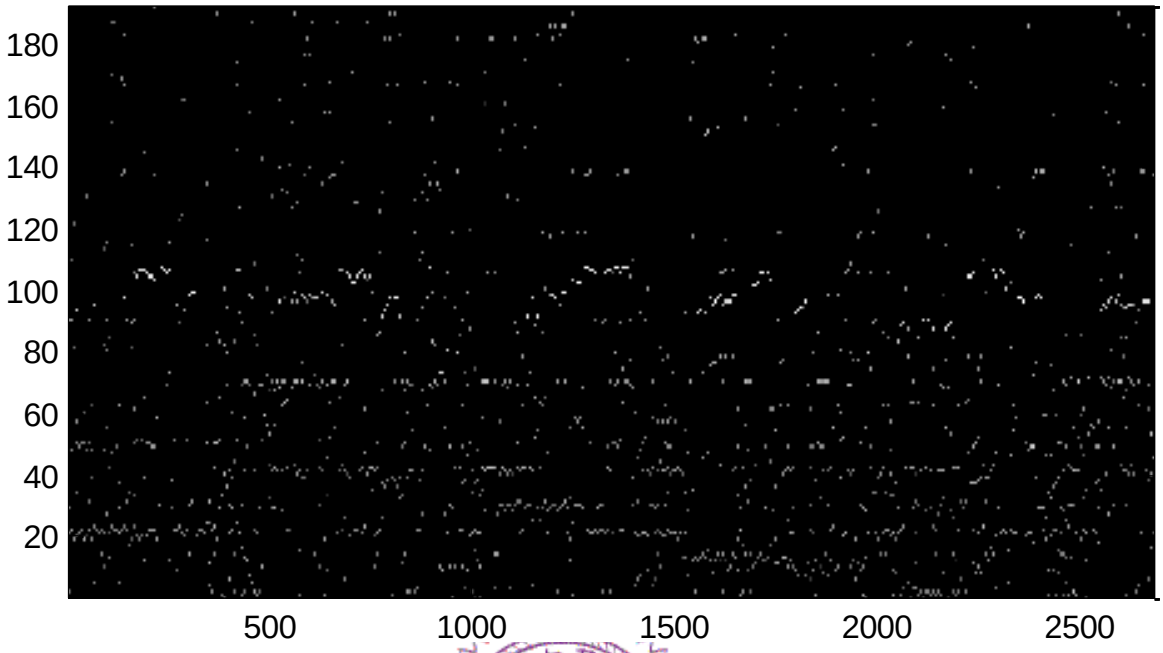


圖 2.2.9 Overtone Deletion

但為了更加精準地找出音高的趨勢，Hsu's Method 使用了 T-F Block 降低頻譜解析度，T-F Block 是含有一段時間與一段頻率範圍的矩形區塊，且每個區塊在時間與頻率方向的交疊（overlap）都是 1/2。若以 MR-FFT 產出的頻譜$x[m,f]$為基礎，T-F Blocks 的結構可以被定義為：

$$b(T,F)=\sum_{m=0}^{M_T-1}\max_{f\in[0,M_F-1]}x[m+TL_T,f+FL_F] \quad (6)$$

上述式子中，T 與 F 代表了 T-F Block 在時間軸與頻率軸上的位置，M_F與M_T為一個 T-F Block 所擁有的 Frame Index 數和頻段數，L_T和L_F代表在時間與頻率上平移的大小。在經過 T-F Blocks 來降低解析度後在頻譜上使用 DP，便可以得到含有音高的能量在頻譜上大略的趨勢，如圖 2.2.10 中的黑色粗線，為此 DP 所計算出來的音高能量趨勢。

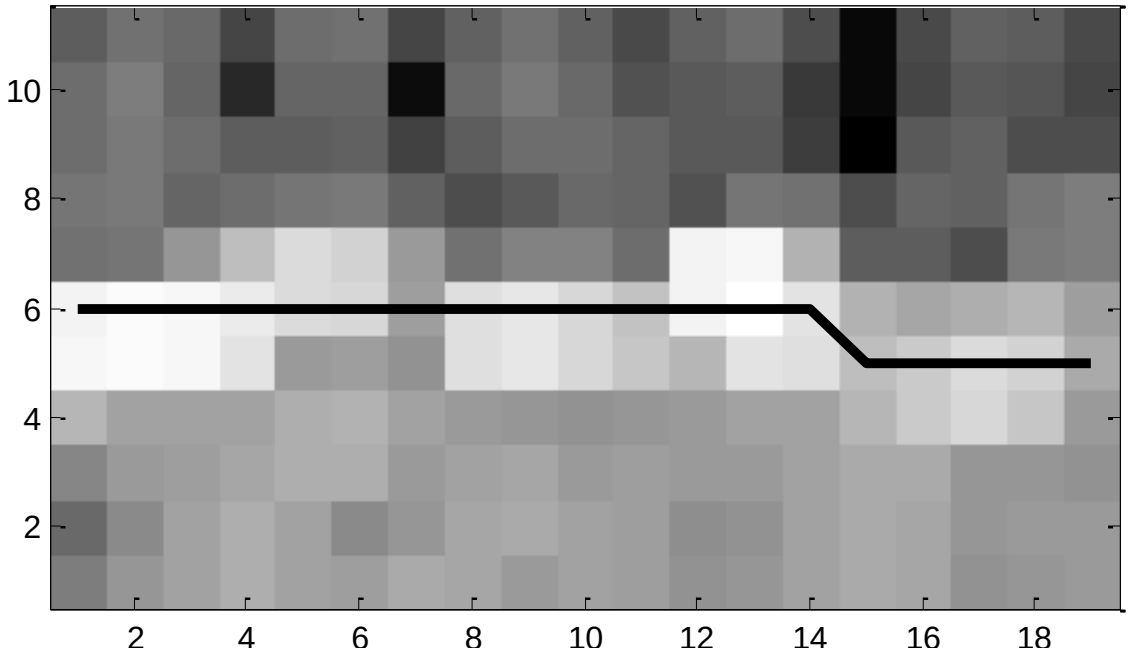


圖 2.2.10 Trend Estimation

由於 T-F Block 包含的音框與頻段相當多，一個在路徑上的 T-F Block 中的能量可能大部分是來自於與他交疊的其他 T-F Block（因為時間與頻率方向的交疊都是 1/2），因此我們希望將與路徑在頻率方向相鄰的 T-F Block 也列入考慮：

$$\begin{cases} \left[f_{T,F}^{lower}, f_{T,F}^{upper} + w\right], if\ b(T,F-1)/b(T,F+1) < 0.8 \\ \left[f_{T,F}^{lower} - w, f_{T,F}^{upper}\right], if\ b(T,F+1)/b(T,F-1) < 0.8 \\ \left[f_{T,F}^{lower} - \frac{w}{2}, f_{T,F}^{upper} + \frac{w}{2}\right], \qquad else \end{cases} \tag{7}$$

上式中 w 在 Hsu's Method 的原始設定為 4 個半音。若上方的 T-F Block（頻率較高）能量較下方的 T-F Block，則路徑大小向上擴張 w，反之則向下擴張 w。若上下的能量差不多，則分別擴張 $\frac{w}{2}$，得到如圖 2.2.11 上的粗線範圍，即為 Hsu's Method 所預估人聲的能量行進範圍。

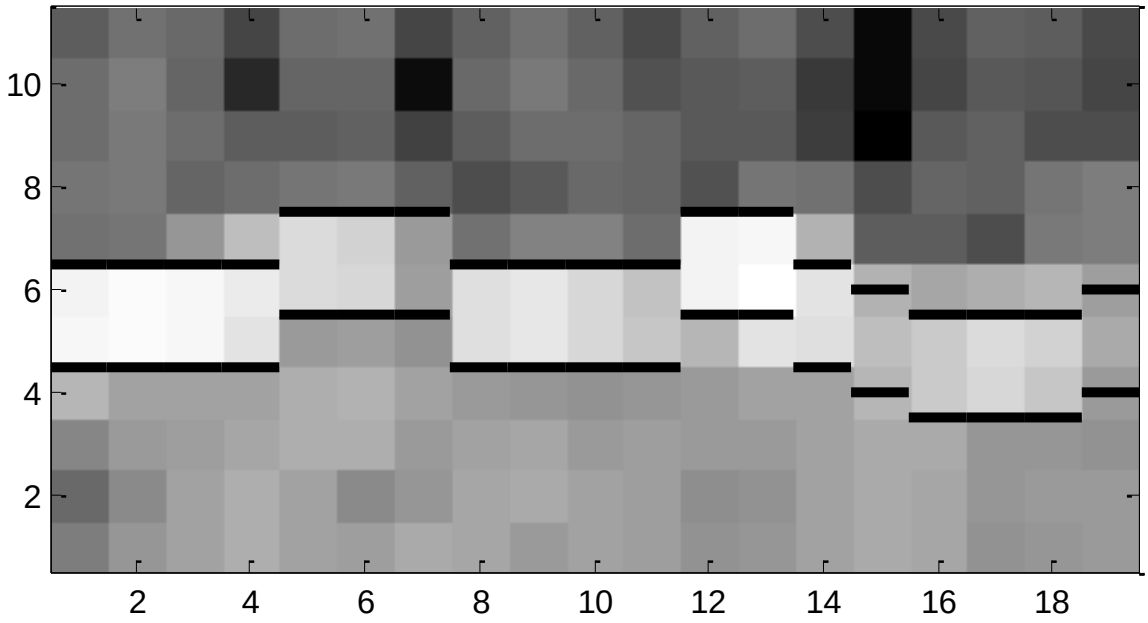


圖 2.2.11 Pitch Range Estimation

確定歌聲音高可能存在的區域後，使用 DP 的方法在加載了趨勢路徑的頻譜上進行音高追蹤，便得到最後的音高結果（如圖 2.2.12 所示，黑色線段即為 DP 方法所判斷的歌聲音高）。

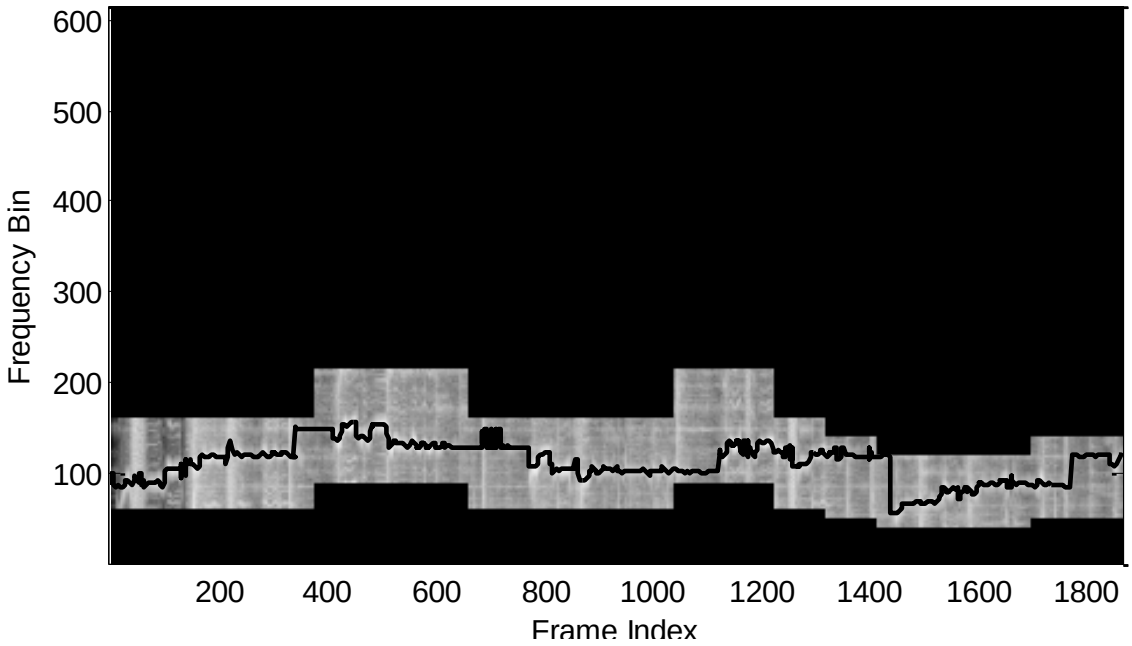


圖 2.2.12 DP-based Pitch Extraction

2.3 以 HMM 為基礎的旋律抽取法

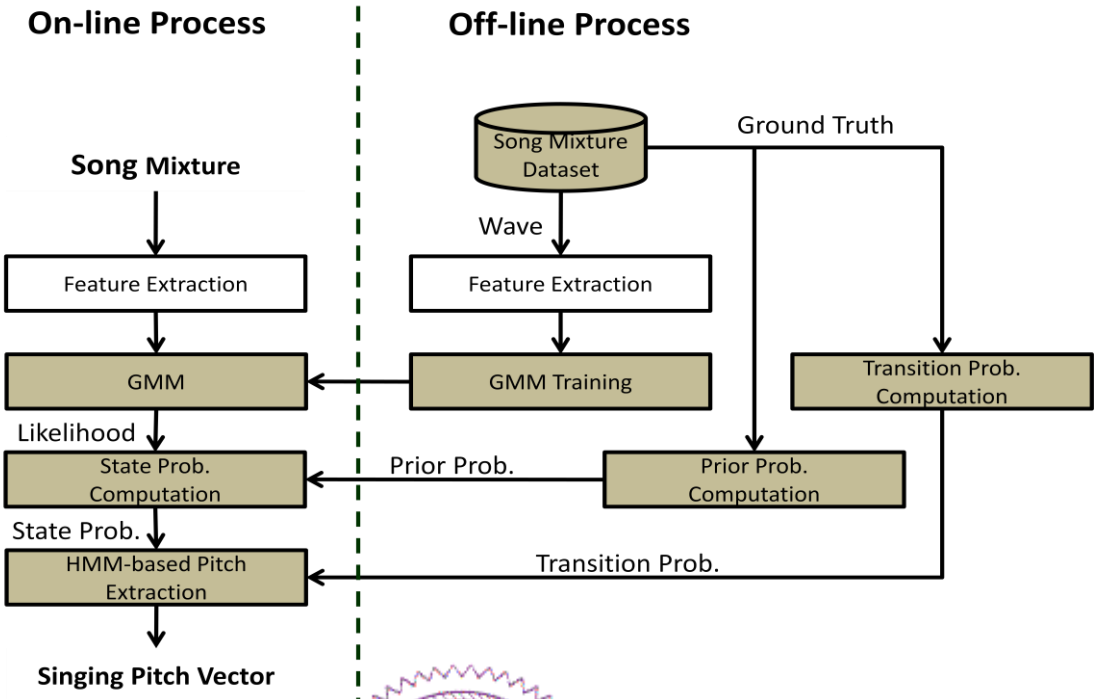


圖 2.3.1 HMM-based Pitch Extraction System Overview

除了原始的 Hsu's Method 外，在本論文的系統裡還實作了另一種音高抽取法，其架構圖如圖 2.3.1 HMM-based Pitch Extraction System Overview，這個方法是透過訓練的方式來取得歌聲的音高。我們在這個方法中使用了隱藏式馬可夫模型（Hidden Markov Model, HMM）的概念，可以將音高的行進視作在不同的狀態間移轉。但是若以音高做為 HMM 的狀態，狀態將有會無限多個，所以我們使用頻段作為 HMM 的狀態，而一個頻段的大小可以由下列式子推算出：

$$L_{bin} = \frac{fs}{L_{fft}} \quad (8)$$

L_{bin}為一個頻段所包含的頻率帶大小，fs為訊號波的取樣率，L_{fft}為進行傅立葉轉換時的框大小。以本論文的參數設定為例（傅立葉轉換框大小為 4096），L_{bin}大約為 3.9063，以人聲頻率範圍除以L_{bin}，便可以知道我們的 HMM 所含有的狀態數量，統計所有歌曲每個頻段前後的分布狀況後，便可以計算出狀態間的移轉機率

為：

$$p(S_r|S_{r-1}) = \frac{N_{r-1,r}}{N_{r-1}} \quad (9)$$

$p(S_r|S_{r-1})$為狀態 *r-1* 移轉至狀態 *r* 的機率，$N_{r-1,r}$為狀態 *r-1* 移轉至狀態 *r* 的總數，N_{r-1}則代表從狀態 *r-1* 移轉出的狀態總數。整個模型到這裡只完成了一半，我們可以將每一個狀態都視作一個高斯混合模型，每個音框都可以利用模型得到似合率來建立狀態的內容。

在 2.2.2 裡 Hsu's Method 使用 NSHS 拉開有音高與無音高部分的差距，此時人聲在頻譜上的能量通常都會很大，這有助於提高最後進行 DP 時的準確度。利用這個特性，我們在訓練法中使用了每個音框經過 NSHS 處理後的前兩高的峰值與峰值的頻段當作特徵，每一個音檔都會得到一個特徵矩陣$S = \{x_1, \dots, x_k, \dots\}$，其中$x_k = \{p_{k,1}, i_{k,1}, p_{k,2}, i_{k,2}\}$，$p_{m,n}$代表第$m$個音框經過 NSHS 計算後的第$n$高峰值，$i_{m,n}$為其所在頻段。我們便可以藉此對每個頻段訓練一個高斯混合模型，每一個高斯混合模型都使用對角協方差矩陣（Diagonal Covariance Matrices），模型參數初始來自於 k-means 演算法，爾後透過 Expectation Maximization（EM）取得，本論文 EM 所使用的疊代（Iteration）為 20。我們最後的目標是透過 HMM 得到音高向量$R = \{r_0, \dots, r_t, \dots\}$：

$$R = arg\max_R \left\{ p(r_0)p(x_0|s_0) \prod_t \{p(s_t)p(x_t|s_t)p(s_t|s_{t-1})\} \right\} \quad (10)$$

$p(x_t|s_t)$是從 GMM 模型所求得狀態s_t的似合率，$p(s_t|s_{t-1})$是從狀態s_{t-1}移轉到s_t的機率。$p(s_t)$為狀態s_t的事前機率（Prior Probability），來自於 Dataset 中頻段的分布統計。整個問題我們可以視作是在由事前機率與似合率所構成的表上來找出最佳路徑，而移轉機率則用來當作 DP 時的 Penalty。最後便可以得到訓練方法所抽取的音高向量。圖 2.3.2 為訓練方法所建立出來的動態規劃地圖，黑色的線段為音

高追蹤的結果。

從圖 2.3.2 以狀態機率建立的結合移轉機率後的音高追蹤結果可以發現除了我們所設定的人聲分布範圍外（40 個半音到 74 個半音）其他頻段都是不可能移轉到的區域，但第一個頻段是例外，由於我們使用人工標記的標準答案來做訓練，這些人工標記的區域含有大量的 0，也就是無人聲的部分。在本論文中並沒有排除這些為 0 的部分，所以使得音高有可能移轉至 0。換言之，這個訓練的方式包含了人聲端點偵測（Vocal Detection）。因此在隱藏式馬可夫模型訓練法的結果中會出現許多的「破洞」，為了填補這些破洞，本論文使用了內插的方式來進行填補，以直線連結破洞兩端的音高點，但這是一個較為不可靠的方法，第四章將會描述一個較佳的解決方式。

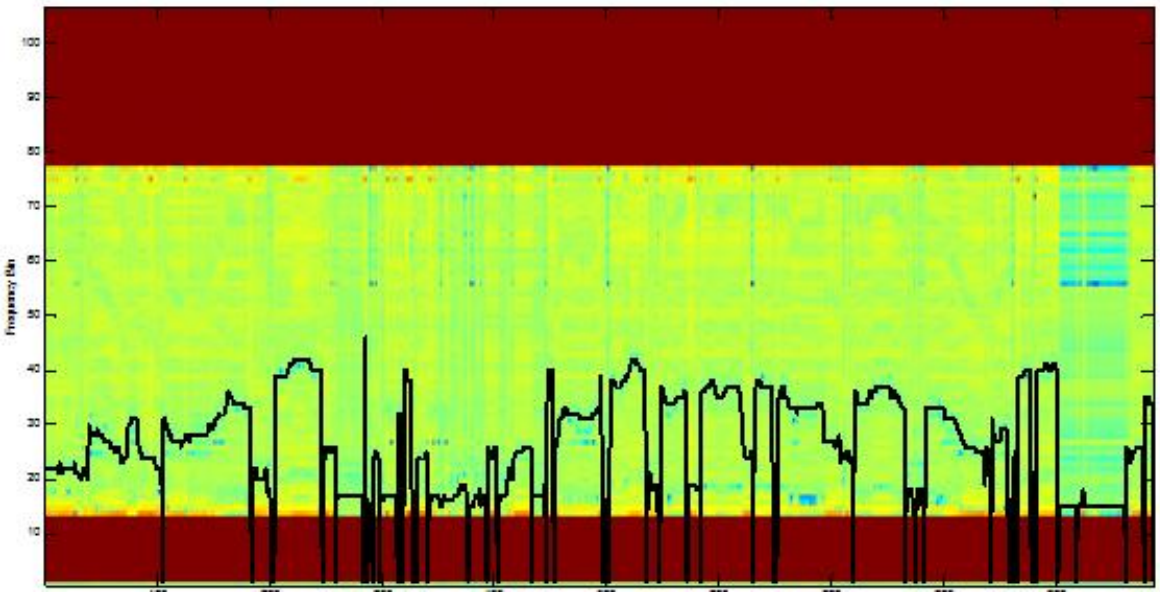


圖 2.3.2 以狀態機率建立的結合移轉機率後的音高追蹤結果

2.4 不穩定音高區域判定

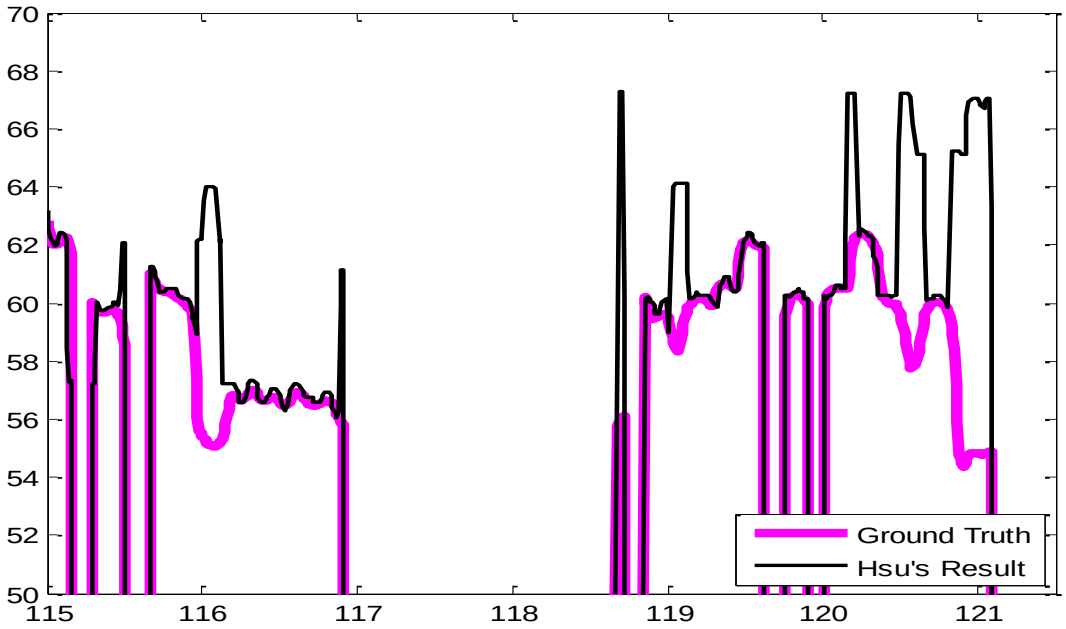


圖 2.4.1 Hsu's Method 結果與人工標記標準答案交疊結果

Hsu's Method 雖然已經是準確度相當高的方法（MIREX 2010 Melody Extraction 項目裡辨識率平均最高），但是仍然存在一些缺點。以圖 2.4.1 來看（本圖使用人工標記的人工端點標準答案修改 Hsu's Method 的結果，目的是使錯誤的區域更加明顯），較粗的線為人工所標記的標準答案，較細的線則為 Hsu's Method 判定的歌聲音高，當時間軸上的某點有兩條線時，便是發生錯誤的地方。我們可以輕易地觀察到 Hsu's Method 在歌聲邊緣與歌聲頻率差距大時會發生錯誤，這表示 Hsu's Method 在面對歌聲強度與頻率快速變化的區域時得到的結果變得不可靠（Unreliable）。

在本論文裡，我們在 Hsu's Method 前反轉了訊號的時間軸來產生不同的結果。但在反轉前必須要去除掉訊號中多餘的部分，如圖…所示。在訊號的末端通常會殘餘一些無法成為音框的部分，這些殘餘的訊號在反轉時間軸後會使得音框的內容產生偏移，在最後進行評估時將因為時間偏移會造成誤差。再去除掉殘餘的部分後便可以將訊號的時間軸反轉。反轉時間軸將對各步驟造成不同程度的影響，最

後反向的結果就會與正向的結果出現差異。有了正反波形的音高後，本論文使用疊合的方式來做不穩定音高判定，找出兩個結果差距過大的區塊（在本論文中設定為 0.5 個半音），即為本論文所需要的不穩定音高區域。我們將在第三章討論不穩定音高區域判定的準確度。圖 2.4.2 為圖 2.4.1 範例做疊合尋找不穩定音高區域的範例，每兩條直線所包的區域為判定之不穩定音高區域。

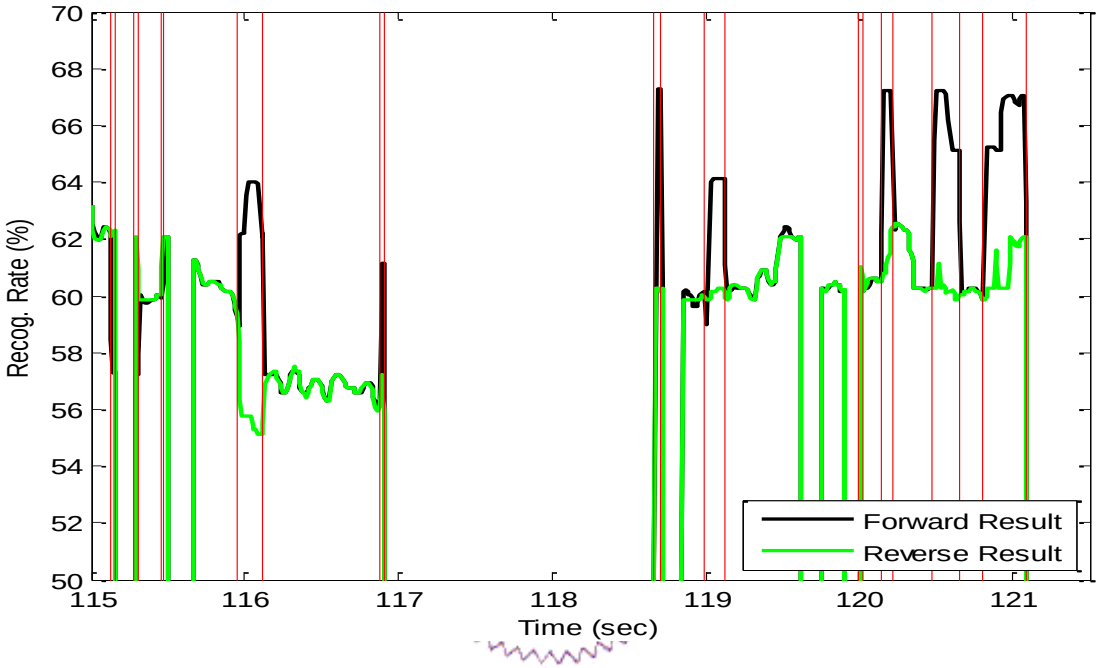


圖 2.4.2 疊合尋找不穩定音高區域範例

透過這個方式我們找出了需要「重新判決」的部分，這時我們讓上一個小節所提到的訓練方法產生結果來做為修正的第三決策者，便是投票修正法。這個方法是利用投票來修正，若兩個分歧的結果其中之一與第三決策者交疊，那麼可能是正確結果的可能性就大大增加，我們就採用票數較多的音高當作修正後的結果，若三者音高皆不相同，則使用訓練法的結果。簡化之後就是直接覆蓋第三決策者的音高，當作修正後的結果。

不過取用 NSHS 作為特徵仍然有部分缺失（在本論文第三章會詳細說明），由於使用 NSHS 做為訓練時的特徵，所以有時會抽取到頻譜上其他能量較高的鋒值（可能是獨奏樂器，或是背景音樂的人聲合聲）。為了盡量降低這個問題所造成的

影響，我們為投票修正法訂立了一個規則，若訓練法的結果與 Hsu's Method 兩個結果其中之一相差過大（本論文設定為超過 13 個半音），則選擇中間的結果。最後修正完的音高向量會與不需要修正的音高向量結合產生出整段音樂的音高結果。

修正的結果可以參考圖 2.4.3，可以發現原本在圖 2.4.1 上錯誤的部份已經被修正了不少，雖然仍有部分錯誤，但是相較起原本的 Hsu's Method，在原本容易出錯的人聲端點與人聲頻率快速變化區域，修正後的結果明顯正確許多。

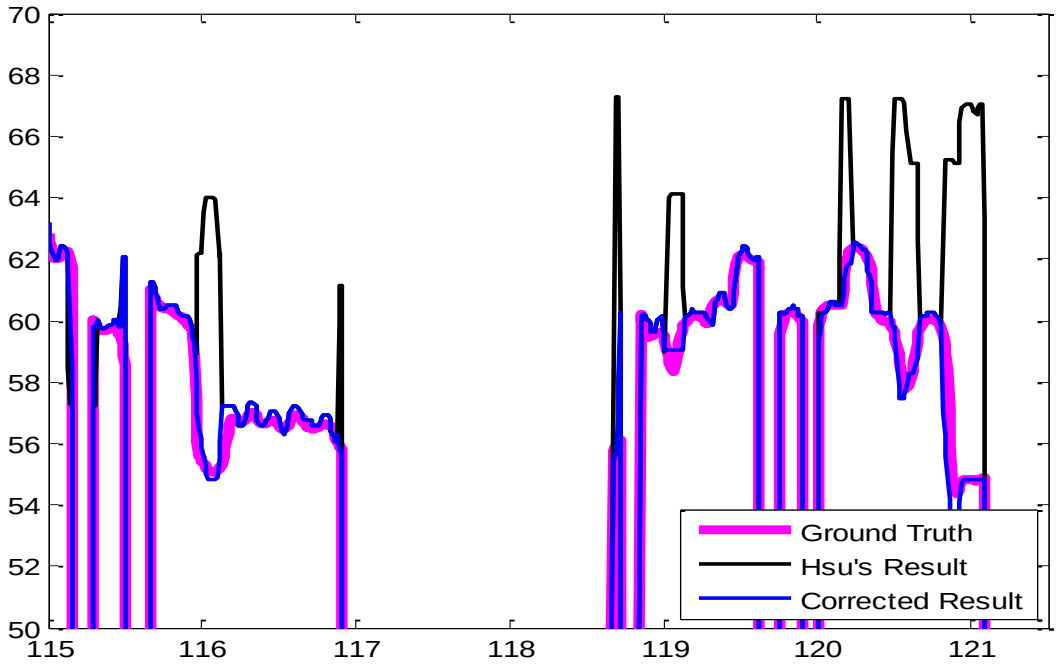


圖 2.4.3 修正後的音高與標準答案交疊結果

第三章　研究結果與分析

	MIR-1k hidden	MV Mixed
Hsu's Method	81.07%	84.48%
Hsu's Method (Reverse)	79.32%	83.96%
HMM-based Pitch Extraction	75.43%	82.21%
Corrected Result (ERR)	83.50% (12.83%)	87.21% (17.59%)

表 3.1 音高容忍度 0.5 下各方法的辨識率與提升率

我們使用了三個不同的 Dataset 來做分析，第一個是 MIR-1k，這是一個公開，專門設計給歌聲分離的 Dataset。總共含有 1000 個取樣率 16kHz，解析度 16-bit 的無壓縮音檔。MIR-1k 中的音檔皆為雙聲道，分別為純音樂與純人聲，本論文中使用此 Dataset 部分資料（這會在 3.2 小節詳述）來當作 HMM-based Pitch Extraction 的訓練資料。第二個為 MIR-1k Hidden，這是一個非公開的 Dataset，做為測試資料使用，共含有 374 個規格與 MIR-1k 相同的音檔。上述所提到的 Dataset 在本論文裡皆以信噪比（Signal to Noise Ratio）為 0dB 的方式混合，成為單軌混有人聲與音樂的音檔。

最後一個為尚在建置中未命名的 Dataset，暫以 MV Mixed 稱之。MV Mixed 共含有 18 首由卡拉 ok 帶歌曲所分割而成的 53 個音檔，取樣率為 44.1kHz，解析度 16-bit。音檔皆為雙聲道，與市面上卡拉 ok 帶一樣，為純音樂軌和人聲混音樂軌所組成。在作為研究結果分析之前，我們有針對此 Dataset 做前處理，首先是降低取樣率，我們利用刪除取樣點的方法，將原始的 44.1kHz 取樣率降低為 16kHz，接著使用頻譜相減法（Spectrogram Subtraction）消去部分背景音樂（以人聲混音樂軌減去純音樂軌），並成為單軌音檔。

在這個章節裡所提到的辨識率計算方式皆為 Raw Pitch Accuracy，也就是不考慮人聲端點偵測的結果，標準答案為 0 的區域全部忽略。而 Hsu 的 Hsu's Method 方法原始辨識率可以參考表 3.1，除此之外圖 3.1 為兩個測試用的 Dataset 辨識率隨著音高容忍度變化的曲線。

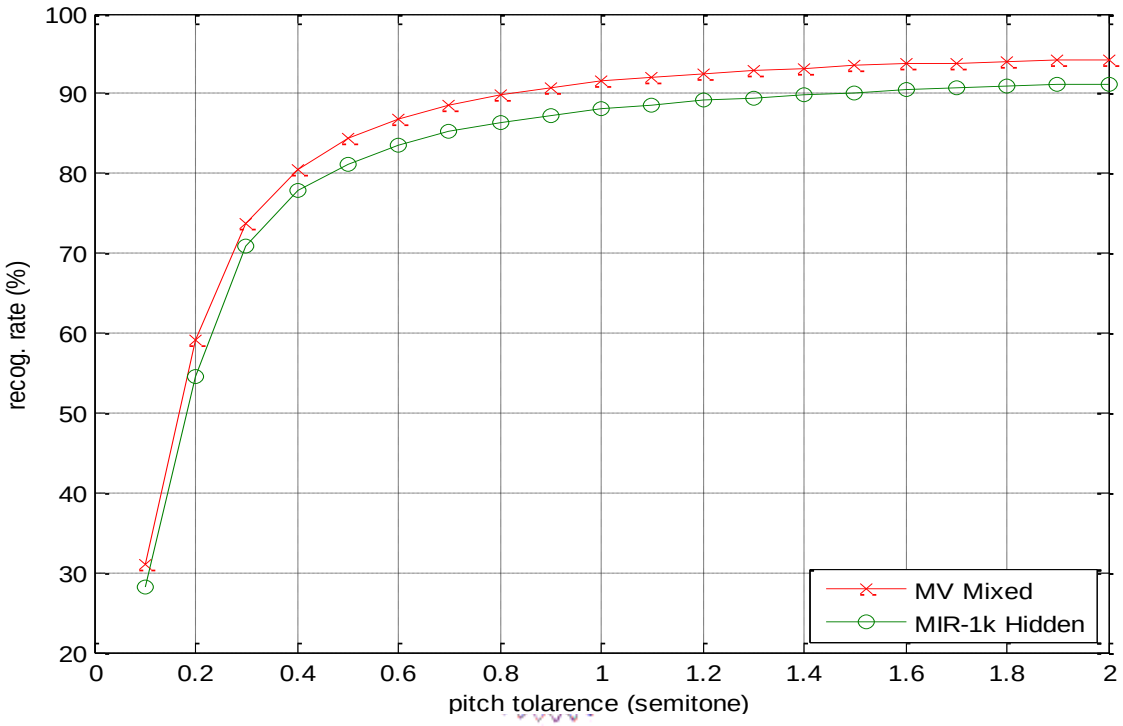


圖 3.1 Hsu's Method 使用於 MIR1k-hidden 與 MV Mixed 之辨識率曲線

3.1 隱藏式馬可夫模型訓練方法分析

我們使用 MIR-1k 來當作兩個測試 Dataset 的訓練資料。所有的參與訓練的音檔都會經過 HPSS 與 NSHS 的動作，以人工標記的標準答案建立各個音高點之間的頻段移轉機率（Transition Probability）表。並且從每個音框取出前兩個最高的峰值做為訓練特徵，為每個隱藏式馬可夫模型的狀態建立 64 個高斯分布的高斯混合模型。

對於訓練的方法來說，訓練資料好不好會直接影響結果，而 MIR-1k 最初建立時並不是以成為訓練資料為目標，在 Dataset 中勢必存在不少低品質的音檔或標準答案。所以我們首先對 MIR-1k 做內測試（Inside Test），使用該 Dataset 中的 1000 個音檔訓練後再使用同樣的音檔做辨識率的測試。結果內測試的辨識率並不是相當的高，於是我們刪去在內測試中辨識率低於一定程度的音檔，最後留存 659 個音檔做為訓練資料。

訓練方法在兩個 Dataset 的辨識率可以參考表 3.1 HMM-based Pitch Extraction 的部分，可以發現辨識率相較起 Hsu's Method 並不算高。其中一個原因在 2.3 已有提過，雖然我們使用了內插的方式填補破洞，卻沒有根本的解決問題，因為破洞兩端的音高準確度也被影響了，所以無法改善太多。除此之外，因為使用了 NSHS 的結果做為高斯混合模型的訓練特徵，所以在遇到歌曲背景有合聲或是樂器比例較重的訊號時，訓練方法也容易出現錯誤。

圖 3.1.1 顯示訓練法在兩個測試 Dataset 中隨著音高容忍度變化的辨識率曲線。

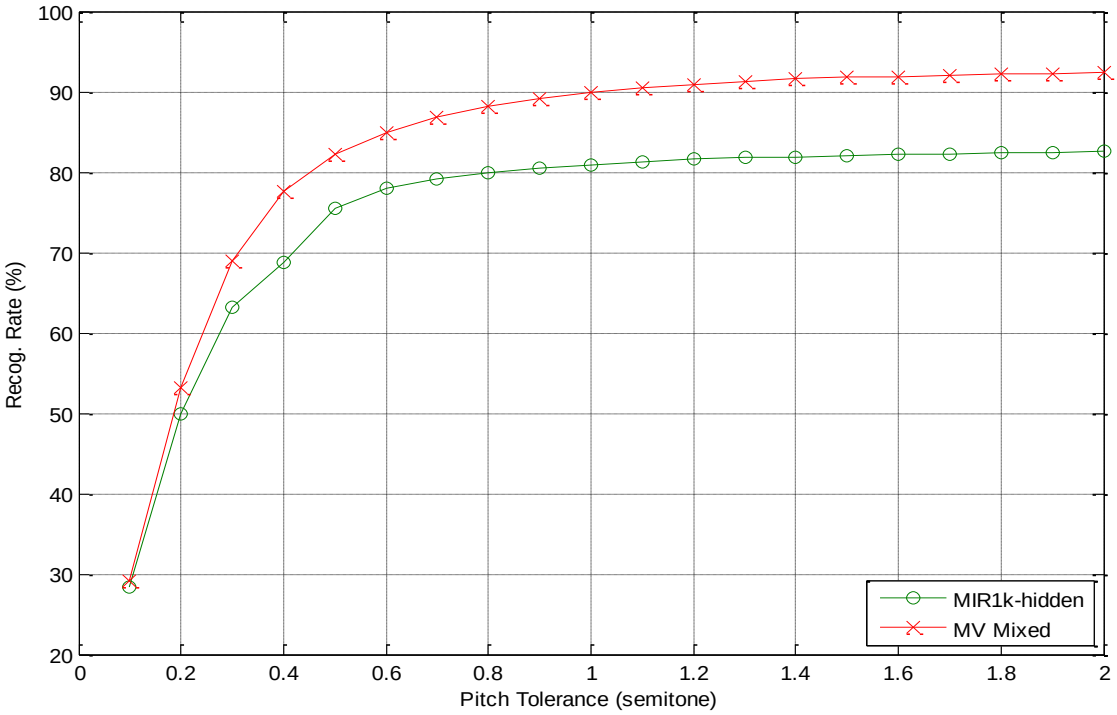


圖 3.1.1 HMM 訓練法使用於 MIR-1k Hidden 與 MV Mixed 之辨識率曲線

3.2 不穩定音高判定分析

在不穩定音高判定中，我們利用反轉訊號的時間軸來使 Hsu's Method 獲得不一樣的答案，並使用疊合的方式來找出正反向 Hsu's Method 答案不同的地方。疊合判定的結果與標準答案比對將會分成兩個部份，一是需要修正區域的部份，另一部份則是不需要修正的區域，兩個部份所佔比例直接地影響了修正後的辨識提升率。需要修正的區域越大，辨識提升率就會越高。反之若不需要修正的區域越大，修正後反而產生錯誤的機率就會提高，將使得辨識率下降。為了得到最高的辨識率，我們面臨了取捨（Trade-off）的問題，需要進行修正的音高可以分成下面三個部份：

$$N = N_c + N_n + N_{ae} \quad (11)$$

N 代表被判定需要修正的總數，N_c代表在進行修正後被改正的數量（不包含原本正確，修正後亦為正確的部份），N_n代表修正之後與原始 Hsu's Method 結果無異的數量，而N_{ae}為進行修正後被改錯的部份。辨識率會隨著N_c上升而提高，隨著著N_{ae}上升而降低，此時門檻值的設定就變得重要。門檻值若設得較低取得的N_c就會比較多，但同時N_{ae}的數量也會增加，這將使得辨識率下降。若反過來將門檻值得較高，則這個方法將會越來越不喜歡進行修正，辨識率自然也會降低。

因此我們使用 MIR-1k 進行內測試來決定門檻值。結果如圖 3.2.1 門檻值作用於 MIR-1k 內測試上之辨識率曲線可以看得出來辨識率峰值的修正門檻值為 0.5。因此我們就採用 0.5 做為修正門檻值來對另外兩個 Dataset 做測試。

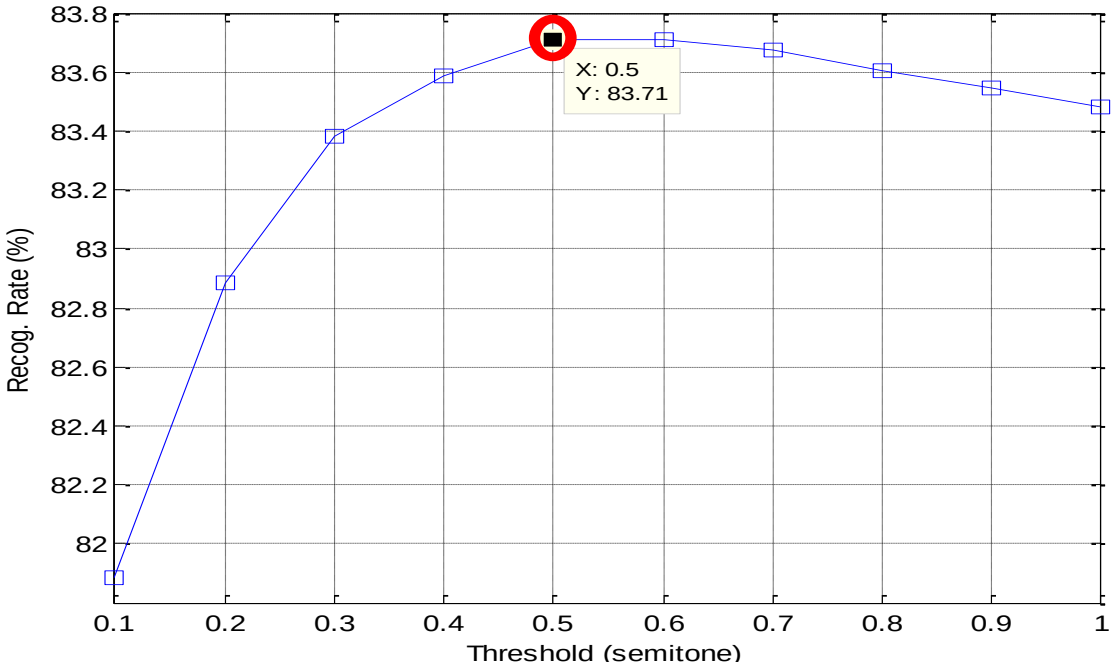


圖 3.2.1 門檻值作用於 MIR-1k 內測試上之辨識率曲線

3.3 整體結果

整個過程最後的辨識率可以參考表 3.1 Corrected Result 的部份，可以發現對比起 Hsu's Method 的結果，兩個 Dataset 辨識率都有明顯的提升，錯誤降低率（Error Reduction Rate）分別為 12.83%與 17.59%。其中很明顯地可以發現，不管是在哪個方法下，使用在 MV Mixed 的效果都比 MIR-1k Hidden 來的好。這是由於 MV Mixed 中大部分的音檔都是錄音室等級，而 MIR-1k Hidden 與 MIR-1k 都是使用一般指向性麥克風，所以錄音品質較差。除此之外 MIR-1k Hidden 中有不少音檔含有一個以上的人聲，對於非常倚靠 NSHS 的方法來說，多人聲使他們的準確度大為下降。最後的原因是兩個 Dataset 格式的差異，MIR-1k 與 MIR-1k Hidden 都是一軌純人聲加上一軌純音樂，進行測試或訓練時將兩軌混音，MV Mixed 含有一軌音樂混人聲與一軌純音樂，在進行測試前可以先通過頻譜相減的步驟，雖然兩軌的音樂並沒有完全相同，但是經過這個步驟後，頻譜上許多干擾人聲的部分都會被降低，尤其是在有和聲的音檔上，相減可以將和聲對主要歌聲的影響降低。

圖 3.3.1 與圖 3.3.2 為兩個 Dataset 整體結果在不同音高容忍度下的辨識率曲線，可以明顯地看出與原本辨識率曲線的差異。且在不同的音高容忍度下，此方法還是穩定地提升了辨識率。

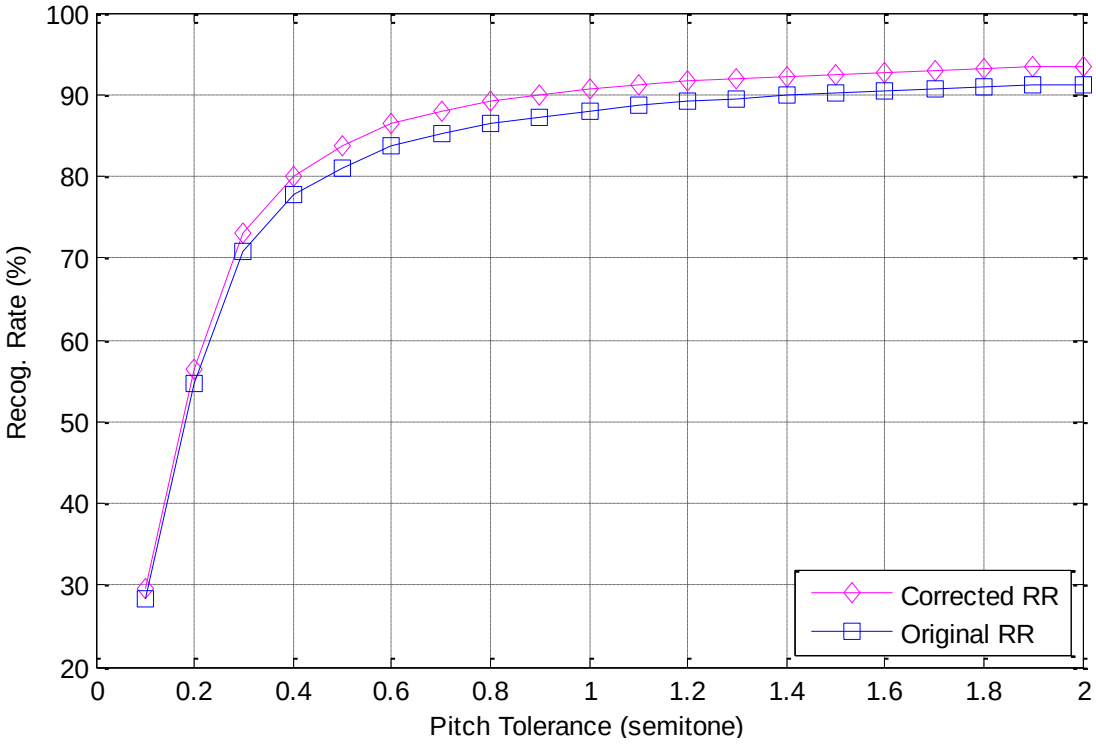


圖 3.3.1 進行修正後 MIR1k-hidden 之辨識率曲線

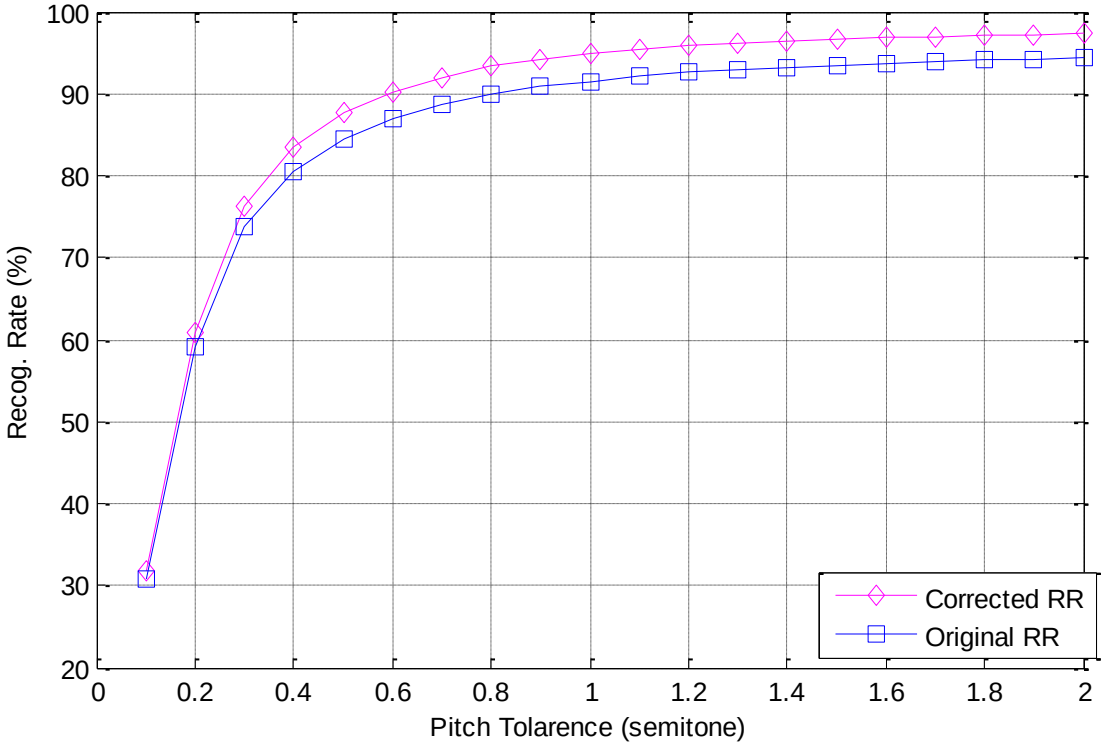


圖 3.3.2 進行修正後 MV Mixed 之辨識率曲線

第四章　結論與未來研究方向

在本論文裡，我們以準確率高的 Hsu's Method 為基礎，利用倒轉輸入波形的時間軸使 Hsu's Method 結果產生差異，疊合後的不同處即為可能需要修正的區域。本論文另外實作了一個以隱藏式馬可夫模型為概念的訓練方法，使用 NSHS 結果的前兩高峰值做為高斯混合模型的訓練特徵，並且以 Dynamic Programming 的方式來求取事前機率，移轉機率與似合率的連乘最大值，這個訓練方法被用來做為投票修正時的第三決策者。這個尋找並修正的方法使得 Hsu's Method 的準確率更加精進，在兩個測試用的 Dataset 裡提升了相當的辨識率。

透過不穩定音高判定，我們找出了可能需要修正的區域，利用反轉訊號時間軸來製造差異，在 Hsu's Method 中原始設定的音框並不大，且取樣率為 16kHz，這使得一個音框所包含的資訊並不多，也不容易造成極大的差異，所以在一些相對來說沒那麼不穩定的區域上無法正確地判定。而不穩定音高判定同時也帶來了一些額外問題，例如不需要修正的區域也被包含在其中，帶來了錯誤增加的風險。在這個方法之下，修正用的第三決策者就變得相當重要，本論文所實作的訓練法以隱藏式馬可夫模型為概念，用頻帶當作模型的狀態，不如 Hsu's Method 有 MR-FFT 輔助獲得原始的頻率，在頻率解析度永遠不足的狀況下，只能夠取得移轉的頻帶向量並轉換成粗糙大略的音高向量。再來便是訓練資料的不足，在 MIR-1k 中雖然有 1000 個音樂片段，但是我們刪去了部分有問題的片段後僅存 659 個訓練資料，訓練資料的不足使得訓練模型變得不穩定，在進行音高追蹤時容易有大起大落的狀況。在碰觸到訓練資料中不包含的樂風時（MIR-1k 全為流行音樂），也容易出現錯誤。

有鑒於這些缺點，在這裡提出未來可能的精進方向。首先是不穩定音高判定 ，未來可以採用較大的音框，較大的音框能夠容納的資訊較多，比起較小的音框更能夠凸顯正反兩個結果的差異。

在訓練方法的部分，我們採用 NSHS 結果的前兩高峰值當作特徵，效果並不是相當得好，常常使得結果差異 12 個半音，因此可以考慮採用其他的特徵，例如 MR-FFT 的結果，Dressler 所提出的 MR-FFT 可以刪去頻譜上不穩定的峰值，使得頻譜變得較為乾淨，人聲更為明顯。相較 NSHS 的結果，MR-FFT 後的結果不容易抓取到低頻區樂器當作特徵，或許可以解決差異 12 個半音的問題。除此之外，我們也可以增加訓練資料的數量，讓訓練的模型更加穩定。

若要訓練法可靠，訓練資料也必須要有一定的數量，建立一個龐大數量且可靠的 Dataset 是一件困難的事情，所以我們也可以選用其他較為可靠的方法來當作第三決策者，如此一來便可以壓低增加錯誤率，使辨識率更加提高。

最後我們可以從訓練法中觀察發現，人聲音高在變化時，相鄰的音高差異不會太大，這對未來的研究方向有極大的幫助。在第二章分析 Hsu's Method 時，錯誤發生時人聲音高的平滑度會出現極大的變化，透過這項觀察我們或許可以去限制 Hsu's Method 在進行動態規劃音高追蹤時的跳動幅度。但由於 Hsu's Method 只有考慮 Raw Pitch，加上純音樂的平滑度是我們所無法掌控的，與人聲銜接時可能會有一段音高上的落差，若單純限制平滑度可能會使人聲端點更加容易出現錯誤。所以或許可以加上隱藏式馬可夫模型訓練方法來進行人聲端點偵測，忽略無人聲部分的平滑度，減低人聲端點可能產生的錯誤。

第五章　　參考文獻

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

[2] Y. Li and D. L. Wang, “Detecting Pitch of Singing Voice in Polyphonic Audio,” *IEEE ICASSP*, pp.17–20, 2005.

[3] G. E. Poliner and D. P. W. Ellis, “A Classification Approach to Melody Transcription,” *6th ISMIR*, pp.161-166, 2005.

[4] K. Dressler, “An Auditory Streaming Approach on Melody Extraction,” *Extended abstract for 7th ISMIR*, 2006.

[5] C. Cao, M. Li, J. Liu and Y. Yan, “Singing Melody Extraction in Polyphonic Music by Harmonic Tracking,” *8th ISMIR*, 2007.

[6] V. Rao and P. Rao, “Melody Extraction Using Harmonic Matching,” *Extended abstract for 9th ISMIR,* 2008.

[7] J.-L. Durrieu, G. Richard and B. David, “An Iterative Approach to Monaural Musical Mixture De-soloing,” *IEEE ICASSP*, pp. 105-108, 2009.

[8] M. Ryynänen and A. Klapuri, "Transcription of the Singing Melody in Polyphonic Music," *7th ISMIR*, pp. 222-227, 2006.

[9] Chao-Ling Hsu, Liang-Yu Chen, Jyh-Shing Roger Jang, and Hsing-Ji Li, “Singing Pitch Extraction From Monaural Polyphonic Songs By Contextual Audio Modeling and Singing Harmonic Enhancement”, *International Society for Music Information Retrieval, Kobe, Japan, Oct. 2009.*

[10] Chao-Ling Hsu and Roger Jang, “SINGING PITCH EXTRACTION AT MIREX 2010,” *The Music Information Retrieval Evaluation Exchange,* 2010.

[11] K. Dressler, “Sinusoidal extraction using an efficient implementation of a multi-resolution FFT,” *DAFx*, pp. 247–252, 2006.

[12] H. Tachibana, T. Ono, N. Ono, and S. Sagayama, “Melody line estimation in homophonic music audio signals based on temporal-variability of melody source”, IEEE ICASSP, pp. 425-428, 2010

[13] 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,” *Proceedings of EUSIPCO*, 2008.

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