爲管弦樂設計的非即時自動翻譜系統

Offline Music-to-Score Following for Orchestral Music

呂　行

指導教授：鄭士康教授

1　緒論

1.1　研究動機

管弦樂總譜是愛樂者深入欣賞一部管弦樂作品的重要門徑，能夠一邊聆聽音樂、一邊翻閱總譜可以幫助愛樂者聽到到很多「弦外之音」。對於音樂老師或是演講者而言，能夠在播放音樂時同時顯示總譜給聽眾看，可以更容易的傳達出想要講解的內容 [7]。但是即使對於一般能夠閱讀五線譜的業餘欣賞者而言，管弦樂總譜依然很難掌握，因爲大量的樂器同時出現在同一個頁面上，音樂的主弦律又可能在不同的樂器間快速轉換。想要一邊聆聽音樂，一邊找到樂譜上正確的位置是很困難的工作。如果能由電腦來協助翻譜，可以解決譜對不上音樂的窘境。

另外在音樂剪輯軟體中，目前多用音樂的波型來顯示，只能大略看出聲音的大小，很難直接找到想要的段落。若是能利用樂譜來作爲音樂的呈現方式，則可以輕鬆的檢索、剪輯想要的段落。以上所述的應用都需要能夠找出音樂的時間點與樂譜間的對應關係。

樂譜追蹤 (Score following), 又稱爲 Music-to-Score Aligment，是一個已經研究多年的題目，但是目前現成的方法大多是針對單一樂器，對於有許多樂器同時演奏的管弦樂作品效果比較差，一般系統最多也只能準確的跟隨三個樂器的演奏。

(請見「文獻回顧」)。本研究的目的在於設計一個可以運用在管弦樂曲的樂譜追蹤系統，利用音量的特徵來有效降低運算的困難度。輸入一個現成的管弦樂錄音檔案以及對應的樂譜，可以在音樂播放的同時正確的顯示出對應的樂譜位置。

2 文獻回顧

樂譜追蹤的歷史大約從 1980 年代開始，早期的作法大多是擷取音樂的音高，將旋律視爲一連串音高的文字序列，利用 string matching 的方法來進行樂譜追蹤，例如 [8]。1997 之後逐漸開始使用動態規劃或是機率模型的方法，例如語音常用的 Dynamic Time Warping[4][3] 和 Hidden Markov Model[2]。更多的樂譜追蹤回顧可以參考 [5][1] 樂譜追蹤可以分爲即時與非即時，即時樂譜追蹤有一個即時輸入的音樂訊號，需要馬上對應到一個已知的樂譜。此類研究佔據了樂譜追蹤的大部分，MIREX 中的樂譜追蹤比賽也限定要是即時。一般的應用有音樂演出時即時幫音樂家翻譜，以及利用已知樂譜來合成電腦伴奏，讓電腦自動幫演奏家配上伴奏。研究的困難在於如何加快演算法的速度，以及當音樂家演奏錯誤或是跳過段落時如何快速檢索到樂譜上正確位置。非即時因爲可以使用到整首樂曲的資訊，因爲比較沒有速度考量，研究的重點比較放在增進解析度。目前一般樂譜追蹤研究大多會利用單一樂器，例如鋼琴，或是少數樂器的合奏來作爲結果展示。因爲樂器比較單純，實驗結果也會比較良好。本研究的主要特色在於:

1. 針對管弦樂團的多種樂器齊奏。
2. 使用音量作爲主要特徵，避開了樂器分離這類比較複雜的演算法，降低了計算的複雜度。
3. 解析度只到分段爲止，減少計算需求，在實際應用上也能提供合理的精準度。

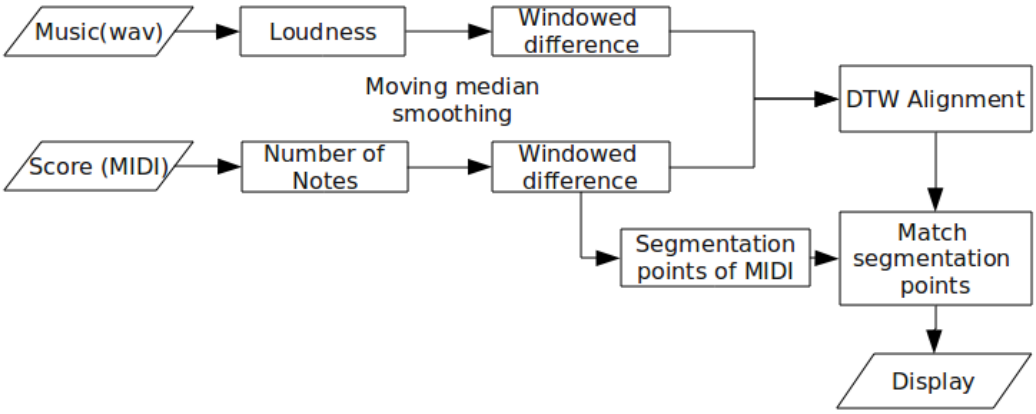


Figure 1: 流程圖

3 系統介紹

3.1 概述

參考 Figure 1。本系統分別有兩個輸入：音樂與樂譜，音樂輸入的格式是 wave 檔，樂譜則是用 MIDI[1]。從 wave 檔計算樂曲的音量，從 MIDI 計算每個節拍的音符數，這兩個訊號分別經過 moving median 平滑化去除雜訊。兩個訊號經過 windowed difference 算出音量/音符數急遽減少的點，這些點對應的就是音樂從一個段落的高潮結束進到下一個段落的分段點。將 windowed difference 過的訊號利用 Dynamic Time Warping(DTW) 來做 aligment，就可以得到兩者時間上的對應關係。但是因爲這兩個訊號間並非完美的對應，只有在分段點的對應比較有意義，因此只找出分段點。兩個分段點中間的段落假設是以穩定的速度進行，則採用線性的關係來對應，這個假設的原因會在後文說明。最後利用分段點的對應關係輸入最後的圖形顯示介面，就可以在音樂播放的同時在螢幕上顯示出相對應的樂譜。

[1]雖然 MIDI 嚴格來說不算是樂譜的格式，但是 MIDI 具有一般樂譜都有的資訊：音符出現時間、音符長度、音高等等，可以輕易代換成其他樂譜格式。雖然有許多機器可讀的樂譜格式 (如 MusicXML)，但是現階段 MIDI 算是比較成熟的格式，也有比較多現成工具可用，因此本研究還是使用 MIDI 作爲樂譜格式。

3.2 特徵擷取

傳統上使用旋律或是節拍的方法都需要能夠分離樂曲中的各個樂器，但是樂器分離需要更多的運算，而且分離後的雜訊比單一樂器的錄音大，再做旋律擷取會有很大的誤差。爲了降低這方面的運算需求，本系統採取另外的方法: 若是觀察一般管弦樂曲的樂譜，可以發現大多數樂曲會分爲幾個段落，每個段落開始時音量會逐漸上升，最後到達高潮之後再快速衰減，進入下一個段落。通常在樂曲高潮時不但每個樂器會用很強的音量演奏，而且會同時讓很多的樂器齊奏，這樣才能造成比較強烈音量效果，這種強弱對比的作曲手法基本上在大多數的管弦樂曲中都會出現。

因此本系統採取一種直覺的特徵擷取方法，從 wave 檔擷取音量，從 midi 樂譜計算每個節拍上同時演奏的音符數目，這兩者之間雖然不會完美對應，但是大致有一樣的外型。wave 檔方面直接從時域計算振幅的絕對值，作爲音量的訊號。而在 midi 方面在每個節拍上計算同時演奏的音符數目，也就是把所有樂器演奏的所有音符加起來，解析度取到八分音符 (半拍) 爲一個單位。如果是一個二分音符 (維持兩拍)，則在它維持的兩拍都要計算。

取得這兩個訊號以後，分別利用 moving median 的方法來平滑。moving median 就是從輸入訊號的開頭開始取一個 window，計算 window 內所有資料點的中位數 (把所有數字排序以後，排在中間的那個，如果是偶數個資料則取最中間兩個的平均)，之後 window 往旁邊移動一個資料點，重複以上計算，如此一來可以得到一個平滑過後的訊號。比起常用的 moving average 方法，moving median 可以有效的去除高頻的雜訊，尤其是對少數突然很高或很低的雜訊可以有效除去。(參見 Figure 2) 這個方法對於 midi 的音符數訊號特別有效，因爲樂譜上可能會出現休止符，造成訊號突然降低許多，但是從音量方面來看，聲音會持續一段時間，因此有必要把 midi 方面突然降低的雜訊消除。

3.3 分段

如同前述原因，音樂的音量漸漸推升，然後再突然衰減的點就是音樂段落的分界點。如果可以分別找出兩個訊號的分段點，再找出兩個訊號間分段點的對應關係，就可以大略的將音樂與樂譜對齊。不同於傳統方法要找出每個資料點間的對

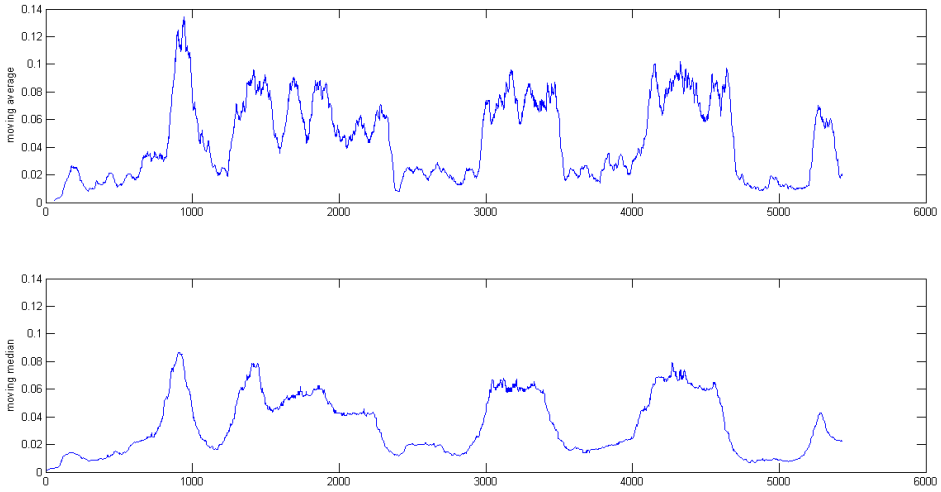


Figure 2: Moving average 與 moving median 比較圖：上圖是 moving average 的結果，下圖是 moving median 的結果

應，本系統只找分段點來做對齊有以下幾個原因：

1. 音量與音符數訊號之間並非完美的對應，硬是要找出每個資料點的對應關係容易產生出很多錯誤的對應。

2. 音樂的每個段落間速度會有變化，但是再段落裡面的速度大致上會是固定的，因此找出段落分界以後，每個段落裡面只要取平均速度，對於此翻譜系統所需的精確度就已經足夠。

3.4 Windowed Difference

爲了找音量急遽降低的時間點，必須使用差分來算出每個時間點前後音量的差。但是因爲音量衰減的長度並不固定，若是用一階差分有時候會找到很多因爲高頻雜訊產生的點。理想中段落的差異必須是音量先維持在比較高的水平，衰減下來以後再維持在較低的水平一段時間，考慮到這些比較全域的資訊，可以使用 Windowed difference（WD）的方法。此方法是再要計算的時間點的前後各取一個固定長度的 window，分別計算兩個 window 中所有資料點的 RMS 平均，再計算後面 window 減前面 window 的差值。在理想的分段點，前 window 中的值都會在比較高的水準，後 window 的資料都在比較低的水準，如此算出來的值會比段落

中間的值來的突出許多，可以明顯分辨出差異。參見 Figure 4。不同的 window 長度會影響輸出結果，本實驗中使用 2 秒鐘的 window 長度，可以明顯看出幾個分段點，不會有太多雜訊干擾。(參見**??**) 計算所有局部極小值的平均，再篩選出所有小於三倍平均的極小值，就是本系統得到的分段點。

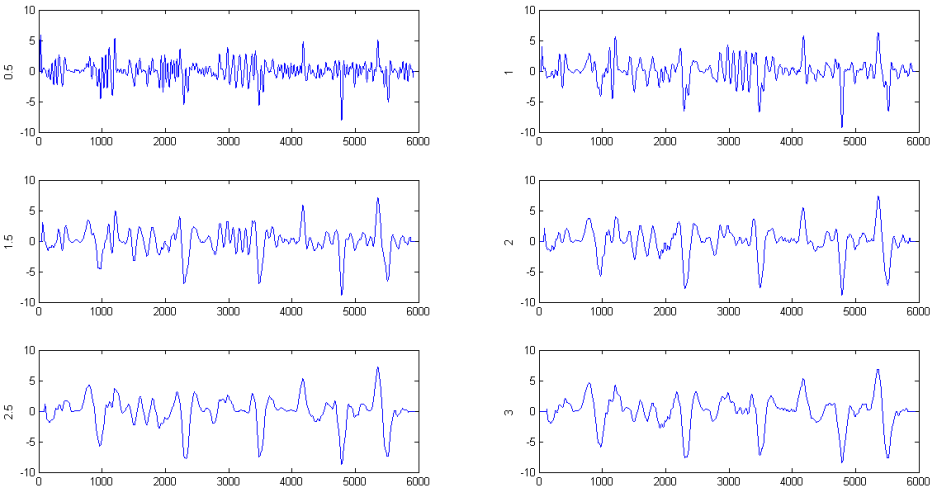


Figure 3: 不同的 window 長度對 windowed Difference 結果的影響：由上至下，由左至右分別是 0.5 秒、1 秒、1.5 秒、2 秒、2.5 秒、3 秒

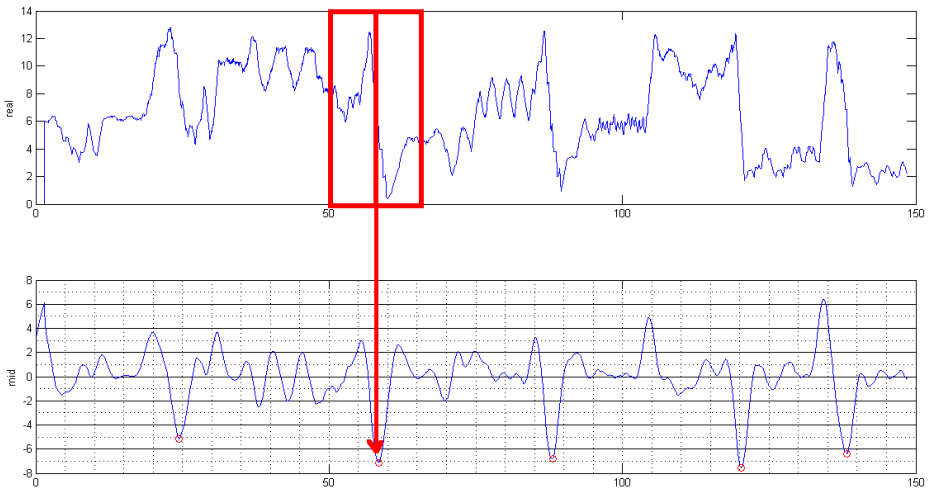


Figure 4: Windowed Difference 示意圖：上圖是原始輸入訊號 (音符數)，下圖是經過 WD 處理後的數據；紅框代表 window(未按比例)

3.5 Dynamic Time Warping

Dynamic Time Warping(DTW) 是一種常用在語音的演算法，目的是要找出兩個資料序列的最佳對應關係。如果將其中一個序列做部份的伸縮 (warping)，想辦法對應到另外一個序列，就可以找到兩個序列的時間對應關係，這也是 DTW 這個名稱的來源。如果計算兩個序列中對應點之間的距離 (Euclidean distance)，距離越短代表兩個序列越相似，距離越長代表越不相似。因此問題就變成要如何調整兩個序列間的對應關係，讓兩個序列間的距離函數 (或常被稱爲「成本函數 (cost function)」) 可以最小。DTW 一開始要算出兩個序列的累積成本矩陣 (accumulated cost matrix)。如果我們有兩個時間序列 $p(t)$ 與 $q(t)$， p 的第 i 個值與 q 的第 j 個值之間的距離就可以定義爲

$$|p(i) - q(j)|$$

要找出從起始點 $(0,0)$ 開始一直到 (i,j) 的累積成本，利用下列的遞迴式：

$$D(i,j) = |p(i) - q(j)| + min \begin{pmatrix} D(i-1,j), \\ D(i-1,j-1), \\ D(i,j-1) \end{pmatrix}.$$

也就是選擇 $(i-1,j),(i-1,j-1),(i,j-1)$ 三點中累積成本最低的點。算出所有 i,j 組合的成本，就形成一個累積成本的矩陣。假設兩個序列在 $(0,0)$ 與 p,q 的最後一點都對應，就可以在矩陣中找出一條連接此兩點成本最低的路徑，此路徑就是兩個序列的最佳對應路徑。(參見 Figure 5)

觀察前述兩個經過 windowed difference 的訊號，雖然並非很完美的對應，但是高點與低點大致上是相合的。因此利用 DTW 來找出這兩個訊號的對應關係。輸入的兩個訊號經過正規化 (normalized)，也就是 (DW 後的訊號)/(訊號絕對值的最大值)。因爲除了分段點以外，兩個訊號並非完美的對應，所以分段點外的對應關係並沒有很大的意義。因此只從 MIDI 找出 MIDI 的分段點，利用 DTW 找出的對應關係，去找到對應的 wav 分段點。如此一來就可以找出兩個訊號分段點得時間對應關係，就可以輸入下一個階段的顯示介面。

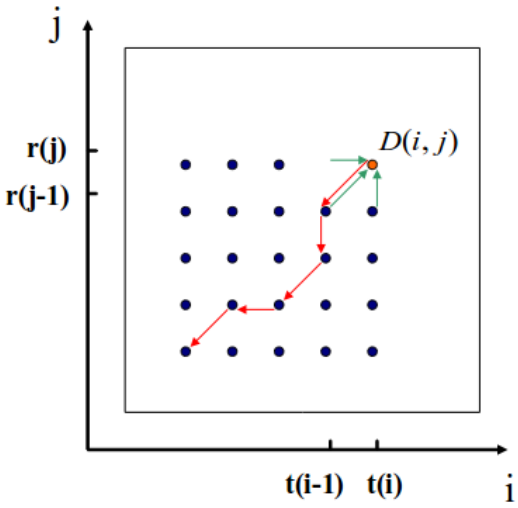


Figure 5: DTW 示意圖, 來源: 張智星，" 音訊處理與辨識"，網路線上課程，可由作者之網頁 http://www.cs.nthu.edu.tw/ jang 連結到此線上教材。

3.6 顯示介面

通過前述步驟取得時間的對應關係以後，必須要在正確的時間點顯示出正確的樂譜。由與之前只算出幾個關鍵時間點的對應，在這些時間點中間的區域則是採取線性的關係來計算，也就是在兩個時間點之間算出平均速度，利用此平均速度來顯示樂譜。有別於傳統的書籍式一頁一頁的樂譜，本系統把樂譜編排成類似卷軸式的一個長條形，由左至右延伸，這樣的好處是樂譜的由左端算起的長度就直接是樂譜的時間。螢幕就是一個視窗，在樂譜上滑動，而滑動速度隨之前算出的速度來控制。使用這種方式的話，如果另外一個好處就是顯示的樂譜圖檔可以很輕易的抽換。由於缺乏一個現成的數位樂譜顯示方法，本實驗使用的是 MIDI 常用的 Pianoroll 顯示方法，每個音符條的 y 軸高度表示音高，長度代表音符延續的時間。只要利用一些商業軟體將 MIDI 轉換成五線譜，或是將印刷樂掃描、剪貼成長條形，就可以簡單的抽換掉 pianoroll 的圖檔。

4 實驗結果與分析

實驗的測試檔案利用莫札特的 A 大調豎笛協奏曲 (Mozart: Clarinet Concerto in A Major, K.622) 的第一樂章的開頭兩分鐘 (管弦樂團齊奏)，midi 檔利用網路上找

到的此曲的 midi 檔，利用 midi 轉換成樂譜的軟體檢查過，與正確的樂譜沒有太大的差異。Fiugre 6顯示 wave 的音量與 midi 的音符數，可以看出兩者間的波型相當相似，尤其是幾個音量突然降低的區域特別明顯。 但是這兩個訊號還是有太多的

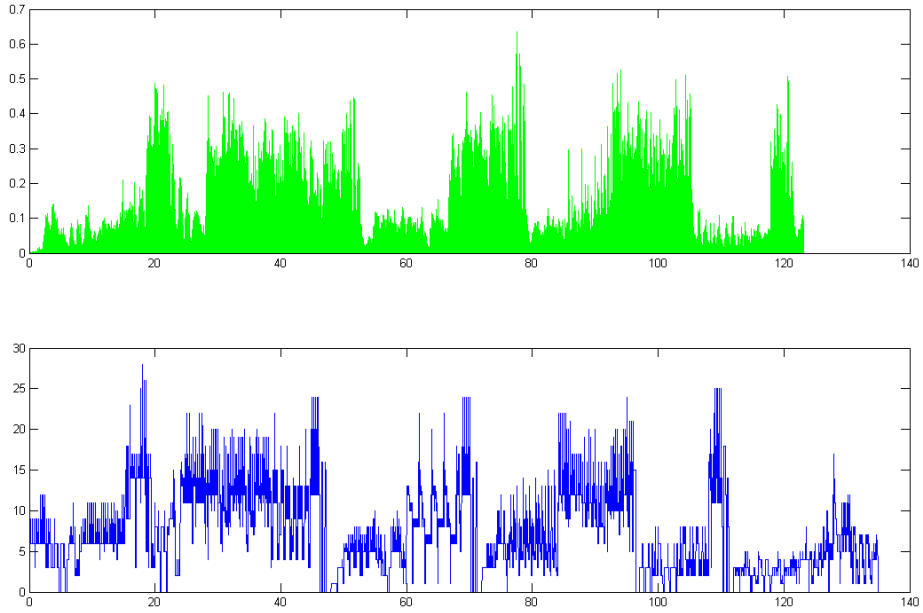
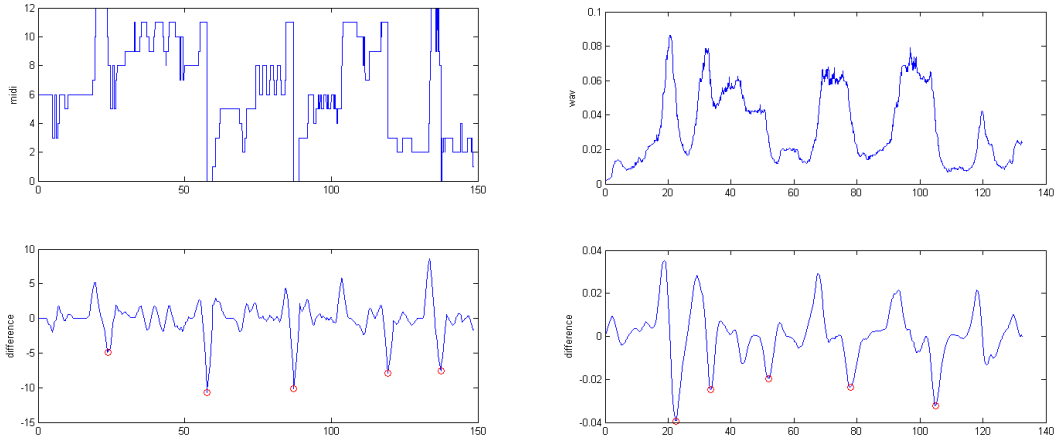


Figure 6: 輸入信號：上圖是 wav 音量，下圖是 MIDI 音符數。橫軸爲時間 (秒)

雜訊，利用 moving median 平滑過後更可以看出兩者的相似性。再經過 windowed difference，可以發現幾個分段點的值都特別負，紅圈標起來的是系統自動找出的分段點。請參考 Figure 7 接著用 DTW 來找出兩者的對應關係，如 Figure 8所示。

從 midi 中找出分段點以後，利用 Figure 8的對應關係，就可以對應出 wav 中對應的時間點。結果顯示在 Figure 9。圖中的紅線連結代表對應的關係，但是並沒有很精確的對到最低點位置，這是因爲圖形在繪製的時候爲了減少電腦負擔，訊號都有降低採樣率，導致些許的誤差。但是可以看到分段點都有正確的對應。爲了評估結果的準確度，必須要用手動標出所有時間點的對應關係。但是因爲用人工來聽解析度有限，很難與實驗數據比較，因此配合本系統的顯顯示介面改用另外一種評估方式。假設音樂與樂譜是完全對應的話，當樂譜捲軸在捲動時，在畫面上畫一條游標線，正確的樂譜就會在剛好的時間點通過游標線。但是因爲存在有誤差，所以樂譜可能會超前或是落後理想的游標線，若是畫出這個誤差的範圍，則會得到一個從理想游標線往左右張開的長方形區域，也就是誤差的範圍。實驗



(a) MIDI (b) wav

Figure 7: 上排是經過 moving median 平滑過的訊號，下排是再經過 windowed differnce 過的信號

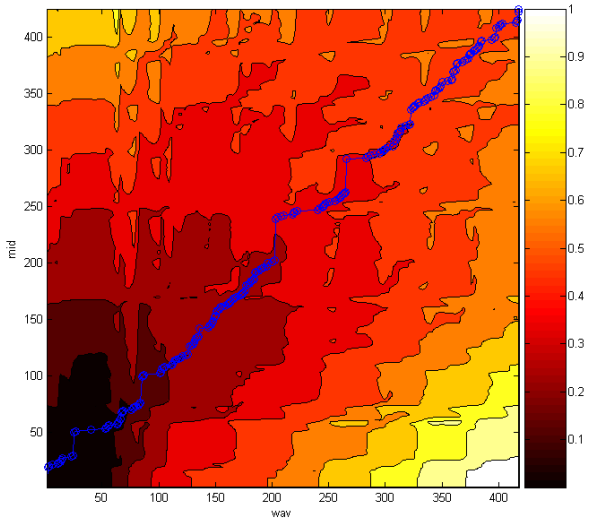
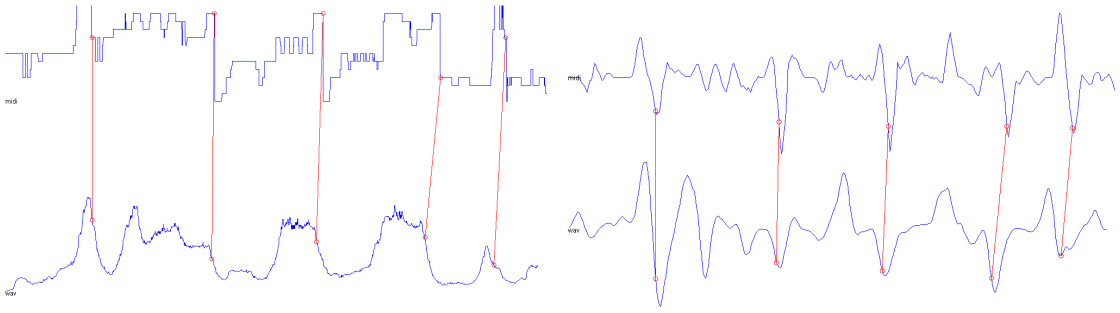


Figure 8: 成本矩陣：橫軸是 wav 訊號，縱軸是 MIDI 訊號；顏色越深表示成本越低，藍線表示最佳路徑



(a) moving median 後的訊號　　(b) 再經 WD 後的訊號

Figure 9: DTW 找出的分段點對應關係

者可以先假設出一個誤差範圍，然後觀察正確的樂譜是否都出現在這個誤差範圍內，如果不是可以再放大或縮小直到決定出正確的誤差範圍。Figure 10畫出的是從理想游標線 $\pm 2sec$ 的誤差範圍，但實際上觀察正確的樂譜會出現在 0 秒 (理想線) 至 -2 秒 (落後兩秒) 間的範圍，所以誤差大約是 2 秒。相較之下，如果用固定速率跑完全曲 (完全不做 aligment)，在放到這段樣本的最後的時候大約會累積到 -4 秒的誤差。

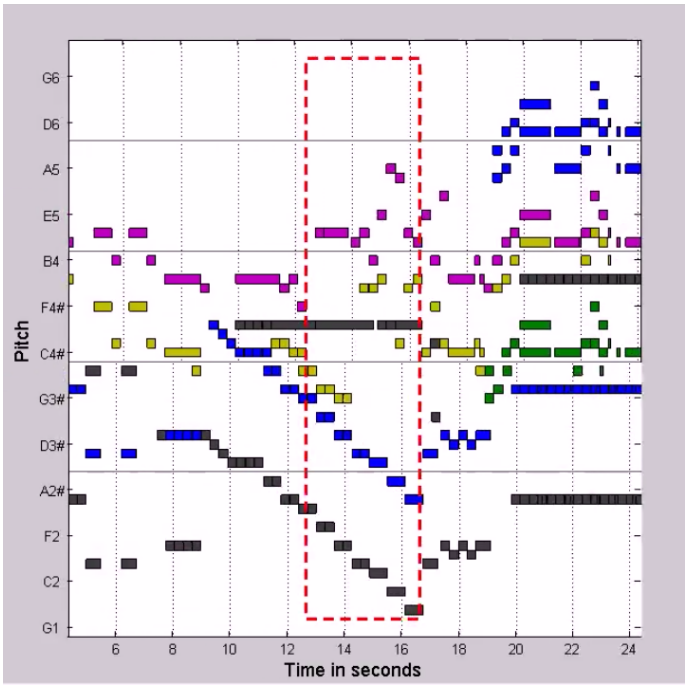


Figure 10: 顯示介面，紅色虛線框代表 $\pm 2sec$ 的誤差範圍，若是樂譜與音樂的誤差在兩秒以內，正確的樂譜理應出現在此範圍內

若是觀察 wav 的音量波型，會發現其實音量衰減的區段並不是像 MIDI 一樣斜

率很大，而是延續一小段時間，利用 windowed difference 找出來的大約是在這段時間的中間，但是這樣會讓 MIDI 有點落後，因爲 wav 已經衰減到一半，如果能稍微做一點偏移可以更精準的對齊。

5 未來工作

本實驗並沒有處理到獨奏樂器的問題，在協奏曲之類有出現獨奏的樂曲中，當只有獨奏樂器（小提琴、木管等等）在演奏時，音符數會維持在 1 2 之間，會有一段時間無法找到任何分段點。但是目前單樂器的演算法已經相當成熟，若是發現有一段時間音符數都很少，可以讓系統切換到使用傳統的單樂器演算法。

各樂器之間的音量大小並不相同，例如一支木管樂器的音量大概等同於一整個小提琴聲部，因此針對譜上的不同樂器的音符數做不同比例的加權，可能可以提昇音符數與音量之間的相似性，更容易比較。

本系統目前是使用非即時的處理，如果可以改成即時的話，就可以有更多的應用，例如現場演出即時翻譜。還有目前使用的是 piano roll 作爲樂譜顯示，對於一般使用者而言並非慣用的樂譜格式，但是因爲缺乏能夠將 midi 轉成五線譜圖檔的工具程式（必須能與目前所寫程式整合），必須利用一些商用軟體來轉換，還需要一些手動調整。但是這個問題只要有一個通用、開放的數位樂譜格式出現，應該就可以很輕易的解決。關於樂譜的討論可以參考 [6]

6 結論

大量的樂器同時演奏對於樂譜追蹤而言是很麻煩的問題，目前的方法大多無法有效的追蹤交響樂，或者是需要複雜的運算。利用音樂中音量與樂譜上音符數的對應關係，可以大略的找出音樂與樂譜的時間對應關係。利用 Dynamic Time Warping 的方法，可以在很快的時間、很少的運算，達到可接受的樂譜追蹤結果。雖然本系統目前的解析度不夠高，但是對於一般音樂欣賞的用途而言已經足夠。這個系統提供初步的追蹤，如果在結合一些現有的樂譜追蹤演算法，可以再針對段落中的細節進行更細部的對齊，可能可以得到更好的效果。

7 感謝

感謝鄭士康教授願意接納我做這個專題，在擬定題目、研究方法、治學心態等等各方面都給予我很多指導。也感謝 JCMG 的所有學長姐：志鴻、鴻欣、韋安、御仁、馨文、傳祐、鼎棋、如江針對我的報告給予許多建議，還有從他們的報告中我也學到很多。也感謝所有曾經參加過 JCMG 的各路朋友，他們的批評指教都讓我受益良多。

附錄：實用資源

1. Midi Toolbox
   https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/miditoolbox/
   Matlab 下處理 MIDI 的一些實用函數，除了可以把 MIDI 解譯成易於理解的格式外，還提供許多剪輯、抽取音軌、調整速度、分析旋律、分析節拍、分析音高分佈等等功能。

2. Keeping Score
   http://www.keepingscore.org/
   舊金山愛樂交響樂團出品的音樂欣賞軟體，在播放音樂的同時會顯示出樂譜，樂譜上還會標示主題旋律、調性、樂曲結構、作曲特色、演出錄影，點擊樂譜上的連結還可以看到講解樂曲特色的影片。是相當完整整合各種電腦音樂功能的系統，值得參考。

3. 張智星教授教學網站 http://neural.cs.nthu.edu.tw/jang/
   清大張智星教授的教學網站，裡面有不少好用的 Matlab Toolbox 以及 Corpora

4. Audio Signal Processing and Recognition by Roger Jang (張智星)
   http://neural.cs.nthu.edu.tw/jang/books/audioSignalProcessing/
   電子書，介紹一些常用的基礎知識與技巧。

References

[1] Andres Arzt. *Socre Following with Dynamic Time Warping An Automatic Page-Turner*. PhD thesis, Vienna University of Technology.

[2] Arshia Cont, Place Igor Stravinsky, and Place Igor Stravinsky. Score Following at Ircam.

[3] Simon Dixon and Gerhard Widmer. MATCH : A MUSIC ALIGNMENT TOOL CHEST. In *ISMIR*, number Ismir, pages 1--6, 2005.

[4] Robert Macrae and Simon Dixon. POLYPHONIC SCORE FOLLOWING USING ON-LINE TIME WARPING, 2008.

[5] Nicola Orio, Serge Lemouton, and Diemo Schwarz. Score Following : State of the Art and New Developments. pages 36--41, 2003.

[6] Diemo Schwarz. Requirements for Music Notation regarding Music-to-Score Alignment and Score Following. pages 1--16, 2003.

[7] 馬定一, 林浩棟, and 陳恆佑. 電腦多媒體在樂曲分析教學上的應用. In *2006 International Workshop on Computer Music and Audio Technology*, pages 152--156, 2006.

[8] 黃國庭. 自動翻譜系統 *A Note Follower for MIDI-Keyboard*. Master thesis, National Taiwan University, 2007.