國 立 中 央 大 學

資 訊 工 程 學 系

碩 士 論 文

小提琴演奏追蹤系統：應用音源分離結果實現即時音樂追蹤與伴奏

A Violin Performance Tracking System: Utilizing Music Source Separation Results for Real-Time Music Tracking and Accompaniment

研 究 生：林妤潔

指導教授：蘇木春 博士

中 華 民 國 一百一十三年 六 月

小提琴演奏追蹤系統：應用音源分離結果實現即時音樂追蹤與伴奏

摘要

摘要...

關鍵字：音樂資訊檢索, 音源分離, 自動伴奏, 深度學習, 線上動態時間規劃

A Violin Performance Tracking System: Utilizing Music Source Separation Results for Real-Time Music Tracking and Accompaniment

Abstract

Abstract

Keywords: Music Informaation Retrieval, Music Source Separation, Automatic Accompaniment, Deep Learning, Online Dynamic Time Warping

誌謝

誌謝...

目錄

圖目錄

頁次

表目錄

頁次

一、 緒論

1.1 研究動機

根據大學術科考試委員會 107 年至 112 年的音樂術科考試人數資料統計 [1]，樂器主修報考最多的項目分別為弦樂與鋼琴，其中弦樂主修又以小提琴佔比最高。由此可知，小提琴是很多人學習和演奏的樂器。而在眾多涉及小提琴的樂曲中，除了無伴奏小提琴曲 (例如巴赫無伴奏小提琴奏鳴曲...) 之外，幾乎都需要與其他樂器合奏，而鋼琴則是最為常見的合奏樂器，例如小提琴奏鳴曲就是由小提琴與鋼琴一同合奏的曲子，另外也有許多曲目從原始樂器編制改編為鋼琴與小提琴的合奏版本。

若是想演奏合奏曲目，就必須自行尋找或是聘請其他演奏者共同演奏，但聘請其他演奏者的價格並不便宜，通常一節伴奏 (約 50 分鐘到 1 小時) 會根據專業度的不同收費，平均收費約為台幣 600 元至 1800 元不等 [2]，因此若是找不到其他演奏者，通常只能選擇演奏自己的部分或是在網路上尋找伴奏音訊使用。網路上雖然有些演奏者會將音訊上傳到公開平台上提供大家使用，但通常這些音訊已經是混合音訊，無法按照個人習慣或練習的速度演奏，也會被音訊中演奏同一部份的聲音干擾。

因此若是可以將網路上公開的混合音訊分離成演奏音訊與伴奏音訊，並使伴奏音訊跟隨自己演奏的速度播放，這樣一來就算無法找到其他演奏者，平時也可以自主練習合奏，同時也省下了人事成本。

而近年來，音樂資訊檢索 (MIR) 這門領域隨著音樂數位化與深度學習的進步，越來越多人開始關注 MIR 領域的發展。這門領域包含音樂來

源分離 reference 、自動伴奏 reference 、音樂特徵提取 reference 、音樂信號處理 reference 、樂器辨識 reference 等子領域，而這些研究領域也被廣泛的應用在商業化的產品上，例如音樂推薦系統 reference 、音樂創作工具 reference 、音樂教育應用軟體 reference 等產品，國際音樂資訊檢索協會 (ISMIR) [3] 也從 2000 年開始舉辦一年一度的 MIR 研討會，促進相關領域的想法交流。

其中音樂來源分離與自動伴奏更是近幾年許多人關注的領域，音樂來源分離的主要目標為分離混和音訊中的各個音源 (樂器)，分離出來的音源可應用在音樂的重新混音，例如卡拉 OK、DJ 等，或作為其他問題的前處理工具 reference 。因為這項技術在音樂領域的廣泛應用，Sony 與 ISMIR 在 2021 年舉辦了音樂解混 (MDX) 競賽 [4]，大力推動了這項技術的發展，近期也延續了先前的成果舉辦了更大的聲音分離 (SDX) 競賽 [5]。而自動伴奏的主要目標為根據特定的旋律生成伴奏 reference ，或是追蹤特定的旋律生成伴奏或跟隨樂譜 reference 。這項技術已經在音樂教育 reference 、音樂創作 reference 等商業產品使用。

在上述的兩個領域中，音樂來源分離的研究重點主要集中在分離流行樂中的音源，自動伴奏使用的音訊資料大部分也是來自可通過 MIDI 協定傳輸的樂器。古典樂器因為音色的複雜性與資料的稀缺性，因此音源分離與自動伴奏在古典樂器的研究相對較少，目前也尚未有將音源分離的結果結合於自動伴奏的研究，因此本研究旨在開發一套專注於追蹤小提琴演奏的即時音樂追蹤系統，此系統應用音源分離技術將混合音源分離為參考音源並作為即時音樂追蹤系統的參考音訊使用。

1.2 研究目的

本研究的目的是開發一套專注於追蹤小提琴演奏的即時音樂追蹤系統，此系統分兩部分研究，第一部分為音源分離技術的研究，此研究專注於探討如何分離小提琴與鋼琴的混合音訊，作為後續音樂追蹤的參考音訊使用。研究詳細內容包含探討如何根據不同樂器的特性調整資料前處理的方式，以提升音源分離模型的效果、應用深度學習網路訓練音源分離模型並與過往的模型進行比較、利用已知資訊對分離音訊進行後處理，以獲得更乾淨的目標音源；第二部分為音樂追蹤技術的研究，此研究專注於探討如何使用不同來源的參考音訊來追蹤即時小提琴演奏的樂曲位置，並輸出對應的伴奏達到即時合奏的效果。研究詳細內容包含探討系統行程與線程的設計，平均分配系統計算資源來達到即時的效果、設計參考音訊與即時串流音訊的特徵提取方式，降低特徵失真率、設計音樂偵測模組判斷現場演奏是否開始、改良粗略估計位置模組與線上動態時間規整演算法提升追蹤位置的準確率並降低系統的計算延遲、設計決策模組決定最後輸出位置並透過數位音訊工作站輸出伴奏音訊。

1.3 論文架構

本論文分為五個章節，其架構如下：

第一章、緒論，敘述本論文之研究目的、動機以及架構。

第二章、背景知識以及文獻回顧，介紹本研究所需的背景知識，包含小提琴與鋼琴的基本演奏方式與特性分析，以及小提琴與鋼琴的音色比較，並探討目前在音源分離領域與音樂追蹤領域的研究現況。

第三章、研究方法，說明本研究細節，如整體的系統架構、音源分離模型的資料前處理與後處理、音樂追蹤系統的各個模組設計與改良細節。

第四章、實驗設計與結果，說明實驗使用的資料集、實驗設計內容以及評估方法，並對於實驗結果進行探討。

第五章、總結，對於研究結果進行總結，點出尚可改進的部分並討論研究的未來展望。

二、 背景知識以及文獻回顧

2.1 背景知識

本研究專注於開發適用於小提琴與鋼琴的系統，因此本節將介紹小提琴與鋼琴的基本知識與特性。

2.1.1 小提琴與鋼琴的演奏特性

小提琴是一種弓弦樂器，演奏方式透過左手與右手的協調配合來實現。左手透過指腹按壓琴弦來調整音高，常見的左手指法技巧有滑音、抖音等等。滑音為通過指腹沿琴弦滑動到不同位置，產生音符連續變化的音色，抖音則是通過手腕的擺動，使按壓點的觸碰面積改變產生微小的音高波動，使音符聽起來有抖動的音色；右手握弓並將弓毛貼住琴弦，透過手臂與手腕的上下來回動作使弓與琴弦摩擦並發出聲音，常見的弓法技巧有連弓、跳弓等等。連弓技巧是在一弓之內演奏多個音符，產生連貫流暢的音色。跳弓則是使弓在琴弦上跳動，產生彈性與清晰的音色。

鋼琴是帶有琴弦的鍵盤樂器，演奏方式透過雙手與雙腳的協調配合來實現。雙手透過指腹敲擊琴鍵，使琴槌敲擊琴弦發出聲音，其中可以透過不同的敲擊力度與速度產生不同的音量與音色。雙腳透過踩踏鋼琴下的踏板，改變聲音的音色。踏板大致可分為延音踏板與弱音踏板，延音踏板可使音符響起的時間更長，而弱音踏板則是使彈奏的音符更為柔和。

由於樂器本身的不同，使兩者樂器演奏出來的音樂效果也大不相同，小提琴透過左右手的控制，相較於鋼琴固定的琴鍵而言，更可以表現出細膩的音色變化，但對於多聲部或複雜和聲的音樂作品，鋼琴的和聲能力與音域的寬闊更能夠勝任這類的作品。

感覺這兩節可以合起來變一節就好

2.1.2 小提琴與鋼琴的音色分析

從上一節可以得知，小提琴與鋼琴的聲音是非常不同的，但我們如何分辨出哪個聲音是小提琴或是鋼琴，為什麼兩種樂器演奏同一段旋律聽起來會不一樣呢？這是因為人耳所聽到的聲音可以分為基頻與諧波，通常基頻代表的是因高，而產生出來的諧波則是表現了聲音的音色，我們拿圖來舉個例子，此圖為小提琴與鋼琴演奏同一個音高 C4 三秒鐘的特徵圖，C4 的頻率約為 262 赫茲，因此可以看到在 y=262Hz 的特徵很明顯。而諧波通常出現在整數倍的地方，所以在 262*2,262*3 也會有明顯的特徵

根據上一節介紹的特性 blablabla 介紹小提琴與鋼琴的音色特性帶出不同人或樂器演奏出的音色都不同、導致音源分離與音樂追蹤的難處

2.2 文獻回顧

2.2.1 音源分離相關研究

引用音源分離的相關論文並進一步討論

2.2.2 音樂追蹤相關研究

DTW、DTW 的各種變形、ODTW、RL、引用音樂追蹤的相關論文並進一步討論

三、 研究方法

3.1 系統架構

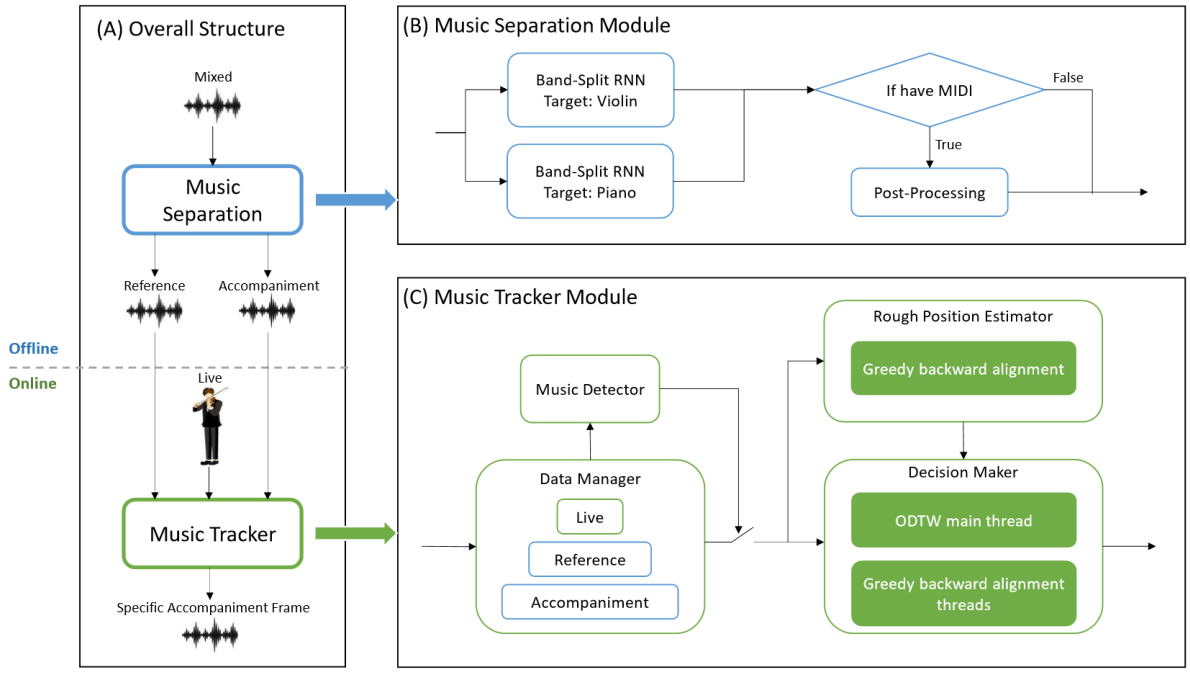

圖 3.1: 系統架構圖

圖 3.1 (A) 為本系統的整體架構，此系統可分為離線處理與線上處理兩個階段，離線處理階段會將混合音源分離成參考音訊與伴奏音訊作為線上處理階段的資料，線上處理階段接收演奏者的即時音訊，並結合參考音訊來計算目前伴奏音訊的輸出位置並播放。請注意本研究將小提琴

與鋼琴之混合音訊作為主要資料，並專注在當演奏者樂器為小提琴或鋼琴的狀況下使用本系統。

圖 3.1 (B) 為音源分離模組的細部架構，此模組包含了參考音訊 (小提琴音訊) 的分離模型與伴奏音訊 (鋼琴) 的分離模型，採用深度學習模型 Band-Split RNN reference 將混合音訊中的目標音源分離出來，最後若有適合的 MIDI 資訊，我們會對輸出做一些後處理使音訊更乾淨。關於模型的訓練與調整、後處理的作法會在 3.2 節做討論。

圖 3.1 (C) 為音樂追蹤模組的細部架構，此模組使用 Python 3.9 開發，我們使用 multiprocessing package reference 實現平行化處理，串流音訊使用 pyaudio package reference 接收處理，達到即時伴奏的效果。此模組包含四個元件，分別為 Data Manager、Music Detector、Rough Position Estimator 與 Decision Maker。Data Manager 管理所有音訊的資料與特徵、Music Detector 即時偵測音樂是否開始、Rough Position Estimator 粗略估計可能的伴奏時間位置，提供 Decision Maker 更多選擇、Decision Maker 決定最後的輸出位置，關於元件詳細的設計與改動會在 3.3 節做討論。

3.2 音源分離模組

說明介紹順序

3.2.1 Band-split RNN

介紹 Band-Split RNN 模型，內部架構等等

3.2.2 資料前處理：頻帶切割的選擇

說明小提琴與鋼琴的頻率域的不同，因此選擇切割的頻帶範圍不同

3.2.3 資料後處理：使用 MIDI 資訊濾除雜訊

3.3 音樂追蹤模組

此模組的實現是參考 reference 的 real-time music tracker，本節將分成兩個部分說明，3.3.1 節至 3.3.3 節會介紹實現音樂追蹤模組的核心方法，3.3.4 節至 3.3.7 節會介紹每個元件的設計與改良。

3.3.1 Dynamic Time Warping Algorithm

動態時間規整演算法 (DTW) [6]，是一種用來計算兩個時間序列資料之間相似度的方法。給定兩個時間序列 $X = (x_1, x_2, \dots x_N), N \in \mathbb{N}$ 和 $Y = (y_1, y_2, \dots y_M), M \in \mathbb{M}$，其中 N 和 M 為兩個時間序列的長度，$x_N$ 和 y_M 是任兩個具有時間順序且形狀相似的資料，例如語音、音樂等等，如圖 放 X 和 Y 的例子 所示。可以看到雖然在t= 圖上時間點時，兩個資料點的距離較大，但若將此點的時間放到t= 圖上時間點上，計算出來的距離就會比較小，達到更好的對齊。考慮到時間序列的長度不一定相等的情況，例如不同人演奏同一個音符可能會有不同的速度差異，因此 DTW 通常採用動態規劃的方法來計算相似度。

首先使用歐基里德距離方法計算 X 與 Y 所有時間點的距離，建立距離矩陣 $C \in \mathbb{R}^{N \times M}$，解釋 DTW 的三個限制: 邊界限制、連續性、單調性

放圖：兩個特徵序列例子、cost matrix + warping path 介紹 DTW 演算法、說明使用 librosa package 的 DTW

3.3.2 Online Dynamic Time Warping Algorithm

介紹 ODTW 演算法、最早提出的作者、解釋不同的更動、

3.3.3 Greedy Backward Alignment Method

解釋為何需要、從候選位置中挑選最好的 path、說明參數的設定

3.3.4 Data Manager 音訊的特徵擷取

說明非串流音訊與串流音訊的特徵擷取方法、放圖

3.3.5 Music Detector Block

說明計算平均振幅判斷靜音片段、使用 DTW 計算對齊成本、兩個 threshold 的設置、放圖

3.3.6 Rough Estimator Block

使用低解析度特徵、如何計算 cost matrix、Greedy 的設置與做法、決定輸出的機制、放圖

3.3.7 Decision Maker Block

使用高解析度特徵、cost matrix 的計算清楚解釋 ODTW main thread、清楚解釋 Greedy threads、清楚解釋輸出的轉換放圖

四、 實驗設計與結果

4.1 音源分離評估

本節將比較兩個不同的音源分離模型，Aug4mss [7] 與本研究使用的 Band-Split RNN [8] 在小提琴與鋼琴混合音訊資料上的表現。

4.1.1 音源分離資料集

由於版權問題，多軌錄音檔通常不會提供每個錄音音軌的資料，因此本研究蒐集並整合了多個公開資料集作為訓練資料，公開資料集包含由 John Thickstun 等人提供的 MusicNet [9], [10]、Muneratti Ortega 等人提供的 Expressive Solo Violin [11] 與 Dong, Hao-Wen 等人提供的 Bach Violin Dataset [12]。

MusicNet 資料集為收集了 330 個古典音樂錄音的大資料集，錄音包含鋼琴獨奏、小提琴獨奏、大提琴獨奏、長笛獨奏、鋼琴與樂器合奏、管樂合奏與弦樂合奏等組合，我們從資料集中挑選出鋼琴獨奏與小提琴獨奏的音訊資料作為訓練資料集的一部分。Expressive Solo Violin 資料集是由專業的小提琴家在同一天錄製九個不同曲子片段的音檔，每一個片段會以不同的音樂表達方式演奏三次，並使用多個電容式麥克風同時錄音。我們採用了資料集中所有的音訊資料 (81 個片段) 作為訓練資料集的一部分。Bach Violin Dataset 為整合了高品質公開錄音的巴赫小提琴獨奏奏鳴曲 (BWV 1001-1006) 資料集，其中包含 17 位小提琴家在不同的演奏

場所下錄製的資料，我們採用有提供音訊檔案的資料作為訓練資料集的一部分。

整合完畢的訓練資料集如表 4.1所示，小提琴獨奏音檔的總時長為 5 小時 24 分 47 秒，鋼琴獨奏音檔的總時長為 15 小時 06 分 37 秒。因此我們使用隨機混合的方式來平均資料，隨機選取兩種音檔 10 秒鐘的片段混合作為一筆訓練資料，共取 2000 筆，驗證資料則是隨機取 100 筆。

表 4.1: 整合的訓練資料集

	小提琴音源	鋼琴音源
總時長	5hr 24min 47sec	15hr 6min 37sec
Channel	Mono	Mono
音訊格式	WAV	WAV

為了公平比較評估結果，評估資料集我們採用 [7] 所提供的公開整合資料集進行評估，此資料集是 Chiu Ching-Yu 等人從公開資料集 MedleyDB [13] 挑選小提琴與鋼琴的錄音所製作的評估資料集，一共有 16 首音檔。

4.1.2 音源分離結果比較

列出與 Open-UMX 的模型訓練出來的結果比較使用 cSDR 標準 表 4.2 表 4.3

表 4.2: N=250 模型 SDR 結果比較

分離目標樂器	模型	前處理方法	SDR
Violin	Aug4mss(paper)	Random-mixing	1.08
		Wet	0.73
		Chroma	1.54
		Correlation	1.56
		NonSilence	1.27
	Aug4mss(retrain)	Random-mixing	2.05
	Band-Split RNN	**Random-mixing**	**11.458**
Piano	Aug4mss(paper)	Random-mixing	7.43
		Wet	8.48
		Chroma	7.47
		Correlation	9.66
		NonSilence	8.76
	Aug4mss(retrain)	Random-mixing	10.95
	Band-Split RNN	**Random-mixing**	**15.619**

表 4.3: N=2000 模型 SDR 結果比較

分離目標樂器	模型	前處理方法	SDR
Violin	Aug4mss(paper)	Random-mixing	3.84
		Wet	4.48
		Chroma	3.82
		Correlation	4.19
		NonSilence	3.03
	Aug4mss(retrain)	Random-mixing	5.811
	Band-Split RNN	**Random-mixing**	**12.136**
Piano	Aug4mss(paper)	Random-mixing	13.46
		Wet	11.76
		Chroma	12.52
		Correlation	11.37
		NonSilence	12.82
	Aug4mss(retrain)	Random-mixing	13.514
	Band-Split RNN	**Random-mixing**	**16.358**

4.1.3 頻帶切割對於分離結果的影響

4.2 音樂追蹤評估

4.2.1 不同速度下的追蹤結果

4.2.2 使用音源分離之音訊做為參考的追蹤結果

4.2.3 不同系統參數設定下的追蹤結果

五、 總結

5.1 未來展望

參考文獻

[1] C.A.P.E. “Statistics on the number of applicants for music subjects over the years.” (2024), [Online]. Available: https://www.cape.edu.tw/statistics/ (visited on 05/07/2024).

[2] 我是江老師.“鋼琴伴奏月薪？一小時賺多少？到底都在做什麼？.” (2020), [Online]. Available: https://youtu.be/8MBaTBXLzEw?t=100 (visited on 05/07/2024).

[3] ISMIR. “International society for music information retrieval.” (2024), [Online]. Available: https://ismir.net/ (visited on 05/06/2024).

[4] Y. Mitsufuji, G. Fabbro, S. Uhlich, and F.-R. Stöter, *Music Demixing Challenge 2021*, 2021.

[5] G. Fabbro, S. Uhlich, C.-H. Lai, *et al.*, “The Sound Demixing Challenge 2023 Music Demixing Track,” *arXiv e-prints*, arXiv:2308.06979, arXiv:2308.06979, Aug. 2023.

[6] P. Senin, “Dynamic time warping algorithm review,” *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA*, vol. 855, no. 1-23, p. 40, 2008.

[7] C.-Y. Chiu, W.-Y. Hsiao, Y.-C. Yeh, Y.-H. Yang, and A. Wen-Yu Su, “Mixing-Specific Data Augmentation Techniques for Improved Blind Violin/Piano Source Separation,” *arXiv e-prints*, arXiv:2008.02480, arXiv:2008.02480, Aug. 2020.

[8] Y. Luo and J. Yu, “Music Source Separation with Band-split RNN,” *arXiv e-prints*, arXiv:2209.15174, arXiv:2209.15174, Sep. 2022.

[9] J. Thickstun, Z. Harchaoui, and S. M. Kakade, “Learning features of music from scratch,” in *International Conference on Learning Representations (ICLR)*, 2017.

[10] J. Thickstun, Z. Harchaoui, D. P. Foster, and S. M. Kakade, “Invariances and data augmentation for supervised music transcription,” in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018.

[11] F. J. Muneratti Ortega, *Expressive solo violin*, 2021.

[12] H.-W. Dong, C. Zhou, T. Berg-Kirkpatrick, and J. McAuley, *Bach violin dataset*, 2021.

[13] R. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. Bello, “Medleydb: A multitrack dataset for annotation-intensive mir research,” Oct. 2014.