中華民國第 59 屆中小學科學展覽會

作品說明書

高級中等學校組　電腦與資訊學科

探究精神獎

第一名

052505

以類神經網路為輔助自動生成小提琴演奏骨架

學校名稱：臺北市立建國高級中學

作者：	指導老師：
高二 林泓毅 高二 劉峻瑋	王鼎中

關鍵詞：小提琴、演奏骨架、神經網路

得獎感言

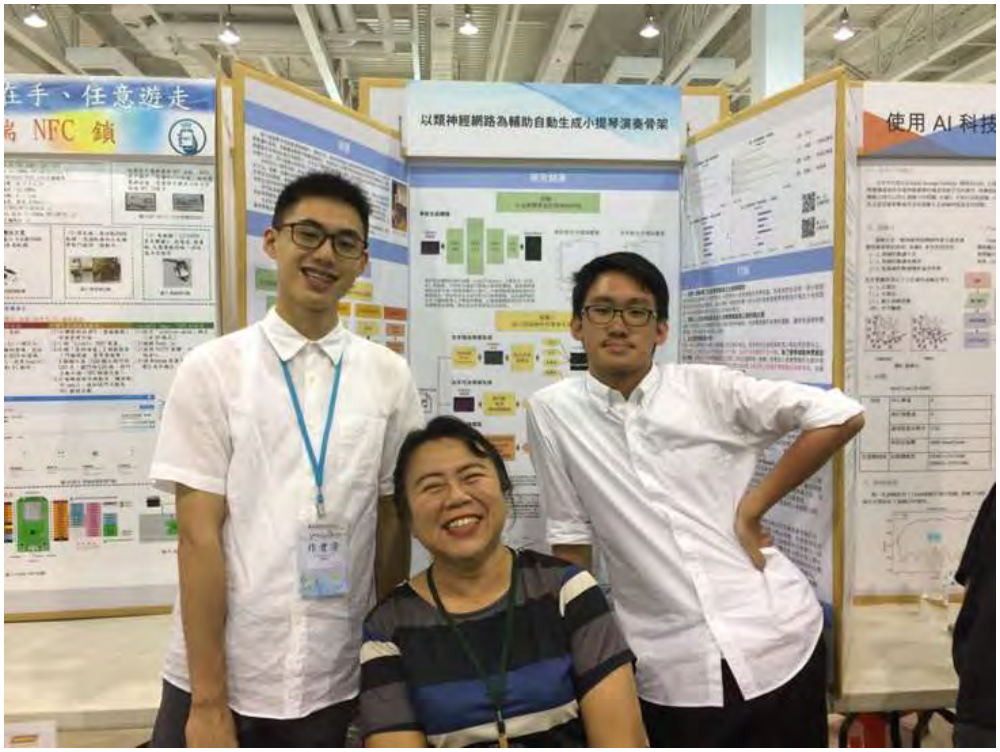
很榮幸能有機會在全國中小學科展中得到第一名的殊榮。除了得獎的喜悅之外，我們也學習到了許多在一般學校課程中所無法獲得的知識與道理。科展研究，不僅僅是發表一件自己的作品而已，在這背後，有在生活中尋找問題、實驗失敗的調整、整理研究數據、分析研究結果等等諸多無法單單從作品說明書裡面得知的酸甜苦辣。我想這些雖然並非研究中最光鮮亮麗的一面，但它卻是與傳統學校課程的最大不同，也是這趟科展之行最大的收穫。

五天四夜的科展競賽，十分緊湊。競賽過程中，要在限定的時間內對評審說明自己的研究成果並回答評審的問題。這十分考驗臨場的應變、整理資料及分析評審問題的能力。過程中不僅僅能夠自我思索作品的不足之處，還能夠與身旁同學交流，從他人的角度來看，能夠看到更多自己沒有發現的問題，並隨之改進。

在科展研究中，總會遇到許多困難，不管是在研究中遇到瓶頸，實驗花了許多時間與心力才柳暗花明，抑或是比賽前一天熬夜苦練，只為了明日場上展現完美的十二分鐘，現在看來，皆是微酸回憶、青澀過往，我們已經成長，繼續前行。

最後，感謝研究路上幫助過我們所有的人，科學從來就不是一個人完成的，願眾人同舟共濟，在前人的樹蔭，巨人的肩膀上，讓科學之路向前拓展。而我們，也在其上。

比賽前的檢錄入口與指導老師合照。



布展時與班級導師合照。

頒獎時合照。

摘要

目前在音樂動畫的領域，若動畫師要產生音樂動畫演奏影片，皆是請真人演奏，再透過感測器去取得骨架座標資料，搭配動畫製作技術進而產生演奏影片，此方法不僅耗時且耗費人力成本，若是能將此生成骨架座標的任務交給電腦自動化生成，將大幅減少時間與人力成本。此研究以小提琴為例，提出了兩種僅以音樂為基礎，透過類神經網路生成虛擬演奏者演奏骨架座標的方法，並對兩種方法生成出的演奏動作結果進行比較與討論。方法一延續先前相關論文的網路骨架，並對其做出修改；方法二為本研究自行設計的骨架生成流程。研究結果顯示方法二相較於方法一與先前相關論文能更有效地生成出合理的小提琴音樂演奏骨架。

壹、 研究動機

目前在音樂動畫的領域，皆是請真人演奏去取得骨架座標資料，搭配動畫製作技術進而產生演奏影片，此種方法不僅耗時且耗費人力成本，若是能將此生成骨架座標的任務交給電腦自動化生成，將大幅減少時間與人力成本。我們希望以類神經網路為輔助讓電腦自動生成音樂演奏骨架。

貳、 研究目的

研發出一套能僅以一段小提琴的獨奏錄音檔，生成出合理的小提琴演奏骨架的程式，以協助動畫師在音樂動畫領域能夠更有效率的製作動畫。

參、 研究設備與器材

一、軟體環境：Python3.6、Vpython、Spyder3、Tensorflow

二、訓練資料：URMP dataset(取小提琴演奏資料)、DEAM dataset

三、硬體規格：

（一）CPU：Intel(R) Core™ i7-8750H CPU @ 2.20GHz 2.21GHz

（二）記憶體：16.0GB

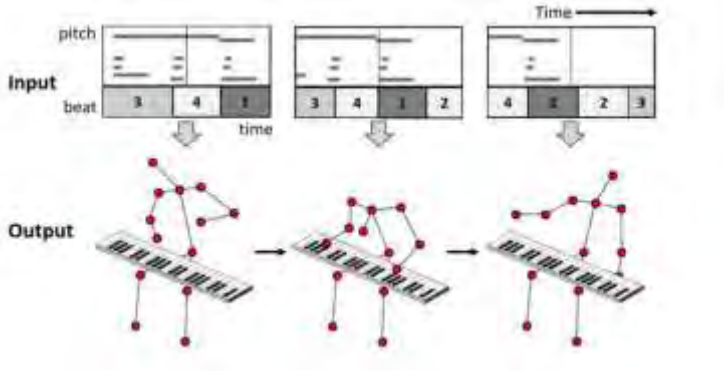
（三）GPU:：NVIDIA GeForce GTX 1070Ti

（四）作業系統：Windows 10

肆、 研究過程與方法

一、文獻探討

有研究以類神經網路生成鋼琴的動態演奏骨架 [1]，其研究是使用 LSTM 遞迴類神經網路，以音樂音高與節奏為輸入，以骨架節點座標為輸出。此研究使我們看到了遞迴類神經網路對於生成時序性的資料處理能力與由網路生成演奏骨架的可能性。



圖一、生成鋼琴演奏骨架之研究示意圖（圖擷取自論文）

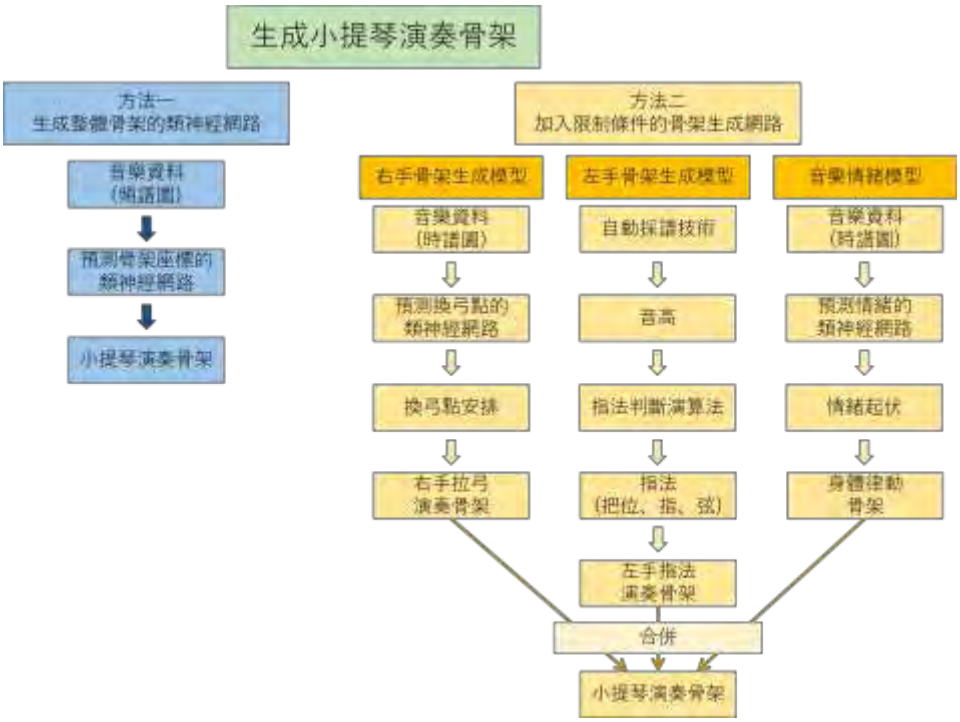
另有一研究同樣以類神經網路生成骨架 [2]，但其目標生成樂器的是小提琴與鋼琴。其研究使用 LSTM 遞迴類神經網路，以音樂為輸入，以骨架節點為輸出。此研究所展示的動態演奏骨架結果，我們關注於其生成出的小提琴演奏骨架，我們認為其生成結果並不是很好，雖然可看出其骨架隨

著音樂而有變化，但卻不斷抖動，且其弓法與指法皆不符合常理。此研究使我們看到別人的作法與仍可改進之處。

圖二、生成小提琴演奏骨架之研究示意圖（上）；

小提琴骨架生成結果展示（下）（圖擷取自論文）

二、研究方法



圖九、研究流程示意圖

以上為我們的研究流程示意圖，我們提出了兩種生成小提琴演奏骨架的方法。

第一種是以音樂資料直接生成整個小提琴演奏骨架，其方法在前面相關研究探討中，我們發現目前其他相關研究皆是以本方法去生成音樂演奏骨架，我們在第一種方法中同樣採取了與其他相關研究也使用的遞迴類神經網路，但在運算單元上我們不選擇使用 LSTM 而選擇使用 GRU，GRU 是於 2014 年被提出的遞迴類神經網路運算單元，其訓練效果與 LSTM 相似，但訓練效率更好 [3]。我們除了在網路架構上做了改變，在 Loss Function 上也做出了改變，由於我們認為在沒有受過小提琴訓練的人眼中，小提琴動態演奏骨架的合理性很大一部份在於右手拉弓的方式與時間點，因此我們對於右手三個節點乘上了較大的權重，以期望網路能對於右手的骨架預測更加重視，隨音樂有更大的擺動，進而使生成出的骨架更具合理性。

第二種方法為一套我們自行設計的骨架生成網路流程，目的在於改善我們在方法一中所看見的問題。在生成小提琴演奏骨架的問題上，其中有很大一部分的複雜度在於音樂資料與骨架間並非一對一關係，而這也是方法一所面對的困境：在現有資料量下，其網路無法學習到音樂與右手運弓骨架座標的複雜關係。因此我們將生成小提琴演奏骨架的問題拆成三個部分，分別為右手、左手與音樂情緒模型，各自以不同方法處理，並加入了一些假設（Assumption），對其設定生成規則。藉由將問題分部化，降低音樂資訊與整體骨架間的複雜度以增進類神經網路與生成骨架的合理性。

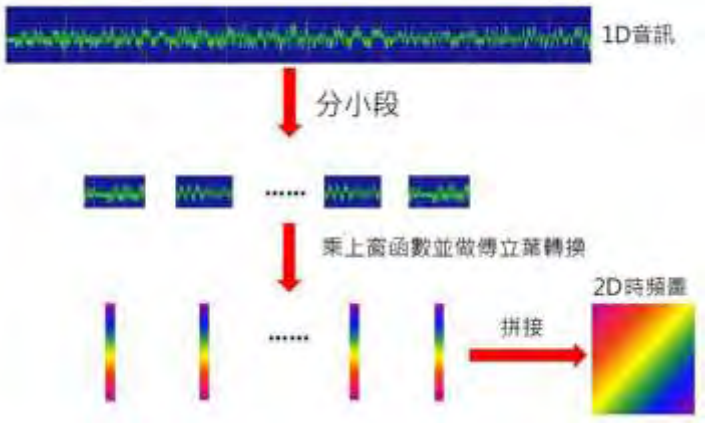
以上兩種方法中使用到的網路，皆是監督式學習，也就是需要給定輸入資料與輸出資料，以下第一部分—資料前處理會說明輸入資料與輸入的格式、種類與前處理過程。第二部分與第三部分將會介紹兩種方法的詳細流程。

（一）資料介紹與前處理

1. 音訊資料（使用 Librosa 函式庫）

Librosa 是一個用於音頻、音樂分析與處理的 python 函式庫。我們使用 Librosa 將一維的 wav 錄音檔轉換成二維以梅爾刻度表示的時譜圖。

時譜圖的轉換是藉由短時距傅立葉變換(STFT)的運算，將一段一維的音訊分許多小段，每一小段分別乘上窗函數後，利用傅立葉轉換把每一小段的音訊轉成一個一維的向量，再將每一小段轉換後的一維向量拼接後便形成了二維時譜圖。轉換方法如圖十。



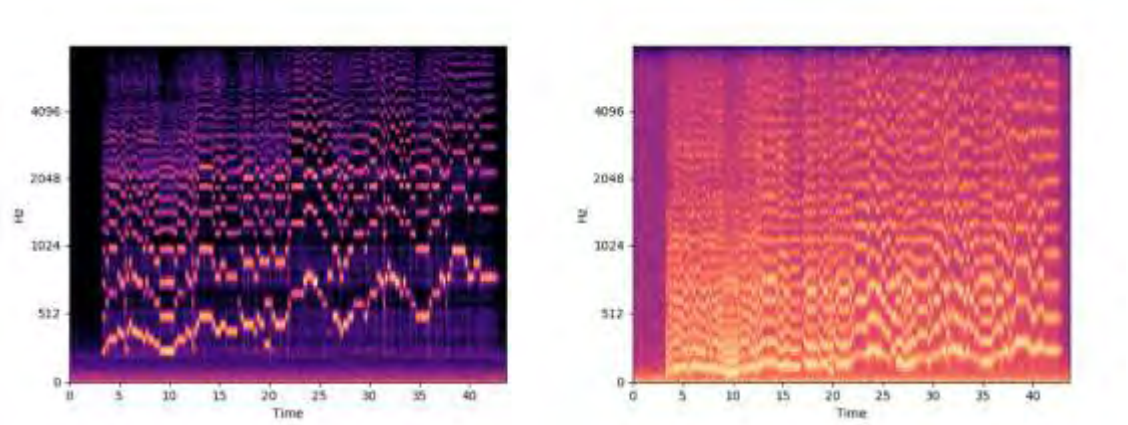
圖十、時譜圖的轉換方法與流程示意圖

梅爾刻度(Mel scale)，是一種基於人耳對等距的音高變化感官判斷而定的非線性頻率刻度。對於小提琴的演奏來說，每個半音對於骨架的動作變化是相似的，舉例來說在同一條弦上一個半音的變化對於骨架上變化也就是移動一個把位，但是一個半音的距離在音頻上卻是以等比級數成長(差一個八度為兩倍的頻率)，因此我們在這邊選擇用含有對數形式的梅爾刻度去將其頻率限縮在每個半音間的距離是相近的。梅爾刻度的轉換公式如下。

$$\mathrm{m} = 2595log_{10}\left(1+\frac{f}{700}\right) = 1127log_{e}\left(1+\frac{f}{700}\right)$$

頻譜圖的採樣頻率為 30Hz，每個採樣點為一個長度為 128 的一維向量。

$$\text{Music Spectrogram DataFormat} = (\text{AudioLength(secs)} * 30, 128)$$



圖十一、以資料集中第一首<Jupiter>的為例。經過 mel scale 處理的時譜圖（左）和未經過 mel scale 處理的時譜圖（右）。

2. 骨架座標資料（使用 Openpose 函式庫）

Openpose 是一個利用 OpenCV 和 Caffe 並以 C++寫成的函式庫，實現了即時多人骨架節點偵測，由美國卡內基美隆大學所開發 [14]。在本研究中我們使用 Openpose 抓出小提琴演奏影片中身體動態骨架的節點位置座標。

在骨架資料的前處理上，我們做了以下四個步驟。

(1)將影片裁切成相片

為了確保每一張間隔的時間都是一樣的，我們將影片分成每秒 30 張照片，對每張照片使用 Openpose 函式庫去抓身體骨架座標，一張照片共可抓到 18 個節點的 XYZ 座標。

(2)刪除 Z 座標並將座標原點標準化

我們對 34 段小提琴演奏的影片進行骨架偵測後，發現 Openpose 對於影片深度（Z 軸）的偵測不靈敏，因此將 Z 座標資料刪除。此外我們還發現每個影片的原點座標皆不相同，於是我們變換原點，改成以單個影片中所有骨架的共同質點為原點，座標轉換公式如下。其中 n 為一個影片的總幀數，k 為一幀中的骨架節點數量。

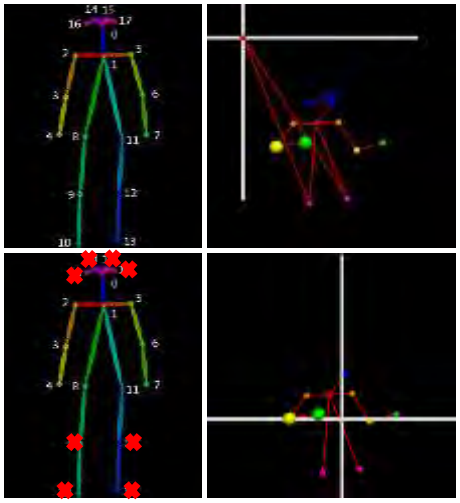
$$x_{ij}{}' = x_{ij} - \frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{k} x_{ij}\right)}{\mathrm{nk}}$$

$$y_{ij}{}' = y_{ij} - \frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{k} y_{ij}\right)}{\mathrm{nk}}$$

圖十二、Openpose 身體骨架節點標序（左）；資料庫小提琴演奏影片（中）；使用 Openpose 抓取出的原始骨架（右）。

(3)刪除節點

在檢查資料過程中，我們發現有些骨架點的資料 Openpose 並不是每張圖都有抓到，而且那些骨架點和音樂演奏並沒有太大相關，所以我們從 18 個節點資料中，只挑選了 10 個對音樂演奏較相關的骨架點當作資料。



圖十三、Openpose 身體骨架節點標序與抓取後骨架資料點（上二）；挑選後的骨架節點標序與抓取後經過步驟二與步驟三處裡的骨架資料點（下二）。

(4)骨架平滑化

因為我們在抓骨架時是將影片切成圖片，並對每張圖片使用 Openpose 去抓取圖中的骨架，因此圖片跟圖片之間的骨架沒有連貫性，有時會突然跳動、錯位。為了解決這個問題，我們對各骨架點資料使用中值濾波器（Median Filter），使骨架在連續性上是平滑的。

中值濾波器是一種非線性數字濾波器，可以去除資料中的雜訊，其原理為要處理的點設置一個固定長度的觀察窗，並取觀察窗中的中位數當該點的值。一維資料的處理運作範例如下，其中 A 為處理前的資料，A’ 為經過 Median filter 處理後的資料。

A = [2, 3, 56, 3]

A’ [1] = Median[2, 2, 3] = 2

A’ [2] = Median[2, 3, 56] = 3

A’ [3] = Median[3, 56, 3] = 3

A’ [4] = Median[56, 3, 3] = 3

A’ = [2, 3, 3, 3]

骨架資料經過以上四個步驟處理後，資料格式變為每秒三十張圖（frames），每張圖中有 10 個節點的二維 XY 軸資料，共 20 維。

$$\text{Skeleton Data Format} = (\text{VideoLength(secs)} * 30, 20)$$

3. 音高

音高取得上我們採用自動採譜技術（Automatic Music Transcription, AMT）[8]，從一段錄音檔中抓取每一時刻的頻率分佈，進而取得主旋律的音高變化。該專案計畫使用類神經網路的眾多模型中的語義分割模型（Semantic Segmentation Model），可支援多種樂器（包含小提琴）的輸入錄音檔轉換。

4. 換弓點

由於原本的 URMP 資料庫中沒有提供換弓點的資料，因此我們以骨架座標資料中的右手手腕（拉弓手）座標速度資料去換弓點時刻。我們以下列步驟生成換弓點的標籤：

（1） 骨架資料平滑化並轉為速度

由於骨架資料的處理是先將影片切成每秒 30 張的照片，再對每張照片以 Openpose 函式庫抓取其骨架座標資料，因此每張照片之間的骨架座標皆有一些誤差，若是

以未經處理的骨架座標資料直接以閾值判斷換弓點，其弓法標籤極為不準確且有許多不合理之處，不合理之處如在三十分之一秒內切換弓法。因此我們先將骨架資料做平滑化後轉為速度並以閾值判斷，詳細步驟如下：

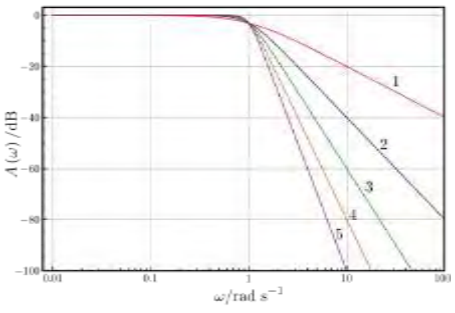
1. 以巴特沃斯濾波器（Butterworth）處理

我們先將右手手腕的 Y 座標以時間為軸轉成一維資料，並以 Butterworth 低通濾波器對其處理，使其平滑化。

Butterworth 是一種在 1930 年被提出的濾波器[11]，其特點在於在低於一定頻率範圍（Cutoff frequency）內，其曲線可達到最大限度平坦，超過其頻率範圍後振幅便會隨頻率而下降，其階數若越高則下降速度越快。

我們使用 Butterworth 的目的在於消除上面所敘述到的極快速、不合理切換弓法的情況。

參數使用：cutoff=5, frequency=30, order=4



圖十四、Butterworth 低通濾波器，其中的數字為階數。

2. 將座標轉為速度

速度比起座標位置更能判斷的出目前時刻是否

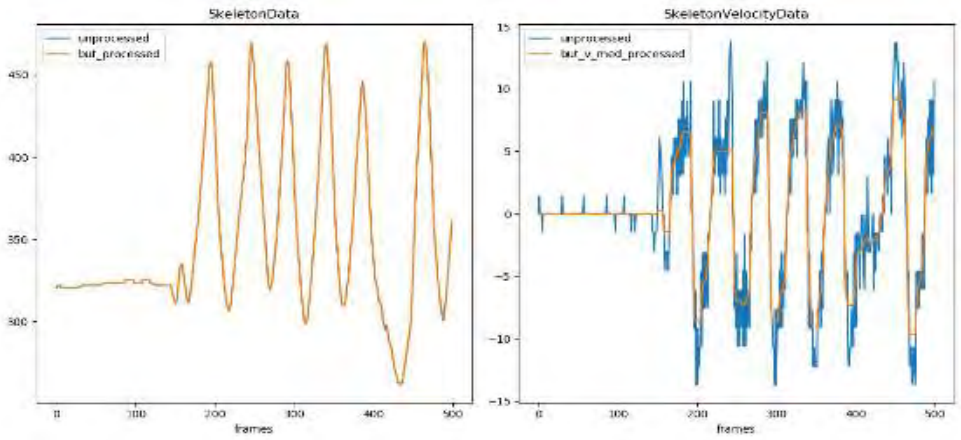
為換弓點，因此我們將骨架座標轉換成速度。轉換公式如下：

$$V_t = X_{t+1} - X_t$$

3. 以中值濾波器（Median Filter）處理

最後我們以中值濾波器對一維的速度資料進行平滑化，並選擇使用大的觀察窗，目的在於讓資料在整體性上更加平滑。

參數使用：window_size=15



圖十五、未處理與處理後的資料比較，以資料集中第一首<Jupiter>的前 500 幀為例。左圖中藍線為原骨架座標資料，橘線為經 Butteworth 低通濾波器處理過後的骨架座標資料；右圖中藍線為原骨架的骨架速度資料，橘線為經以上三步驟處理後的骨架速度資料。

（2） 以閾值判斷換弓點

我們以平滑化後的骨架速度資料，以閾值判斷現在是否為換弓點。判斷方法如下：在目前時刻，若前 10 個 frame 的速度平均與後 10 個 frame 的速度平均皆有大於閾

值且速度方向相反，則判定為換弓點。閾值設定：1

$$\text{Bowing Label Data Format} = (\text{AudioLength(secs)} * 30, 1)$$

5. 情緒標籤

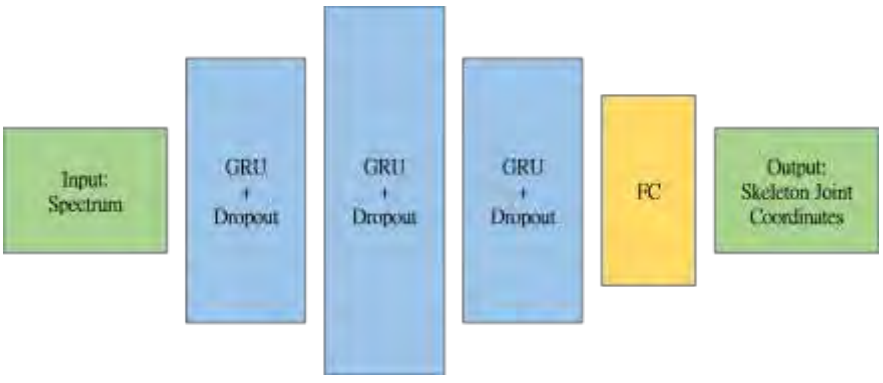
在情緒預測的模型上，我們以 DEAM dataset 的資料訓練[10]。DEAM 資料庫中含有 1802 首，橫跨多種音樂風格的曲目，每首曲目含有採樣頻率為 2Hz 的動態情緒標籤。在此資料集中的情緒標籤使用情緒分類中的 Circumplex Model [9]，含有 Arousal 跟 Valence 兩個向度，而我們使用其中 Arousal 的情緒動態標籤去訓練一個可由時頻圖為輸入預測 Arousal Value 的神經網路。Arousal Value 的值域為-1~1。

$$\text{Emotion Annotation DataFormat} = (\text{VideoLength(secs)} * 30, 1)$$

（二）方法一：生成整體骨架的類神經網路

在方法一中，我們模仿了先前對於骨架生成的論文中網路架構設計，選擇使用 GRU 而非 LSTM 的類神經網路架構，並在損失函數上做出改變，測試並討論在損失函數上對右手弓法增加權重與否（分別是下面第一項與第二項）對於骨架生成表現上的差異。

1. 輸入頻譜的 GRU 骨架預測網路



圖十六、輸入為頻譜的 GRU 網路示意圖

上圖為輸入為頻譜的 GRU 類神經網路架構設計示意圖。輸入資料點為單一時刻的頻譜，長度為 128 的一維向量，而架構上我們設計了三層 GRU，並在輸出層設了全連接層。此網路的實驗目的在於我們希望去了解單純的 GRU 網路是否能抓到音樂與骨架之間的對應關係。以下為網路之參數：

（1） 輸入層

輸入維度：(1, 128)

（2） 隱藏層一—GRU+Dropout

節點數：256、Dropout 比例：0.3

（3） 隱藏層二—GRU+Dropout

節點數：512、Dropout 比例：0.3

（4） 隱藏層三—GRU+Dropout

節點數：256、Dropout 比例：0.3

（5） 輸出層—全連接層（Fully Connected, FC）

輸出維度：(1, 20)

Loss Function 使用均方誤差（Mean-Square Error, MSE），Optimizer 使用 Adam。Loss Function 公式如下，其中 N 為骨架節點座標總數，在此網路中輸出的骨架格式是 10 個點的 XY 座標，在這邊 N 為 20：

$$\mathrm{L} = \frac{1}{N}\sum_{j=1}^{N}(y_{true}^{(j)} - y_{pred}^{(j)})^2$$

2. 輸入頻譜且注重右手弓法的 GRU 骨架預測網路

此網路與上面第一個網路架構相等，但我們在此網路的 Loss Function 上對右手（拉弓手）的肩膀、手肘和手腕這三點乘上較高的權重，因為我們認為在沒有受過小提琴訓練的人眼中，小提琴動態演奏骨架合理性的判斷依據很大一部份在於右手是否有隨音樂正確擺動，因此我們對於右手三個節點乘上了較高的權重，以期望網路能對於右手的骨架預測更加重視，使生成出的骨架更具合理性。以下為網路之參數：

（1） 輸入層

輸入維度：(1, 128)

（2） 隱藏層一—GRU+Dropout

節點數：256、Dropout 比例：0.3

（3） 隱藏層二—GRU+Dropout

節點數：512、Dropout 比例：0.3

（4） 隱藏層三—GRU+Dropout

節點數：256、Dropout 比例：0.3

（5） 輸出層—全連接層（Fully Connected, FC）

輸出維度：(1, 20)

Loss Function 使用均方誤差（Mean-Square Error, MSE），Optimizer 使用 Adam。Loss Function 公式如下，其中 N 為骨架節點座標總數，在此網路中輸出的骨架格式是 10 個點的 XY 座標，在這邊 N 為 20，其中我們在 Loss Function 中加入了我們自己設計的權重項w_j，其代表的是在第 j 個節點的權重：

$$L = \frac{1}{N}\sum_{j=1}^{N}(y_{true}^{(j)} - y_{pred}^{(j)})^2 * w_j$$

骨架節點	骨架權重w_j
0-1的XY座標	1.0
2-4的XY座標	2.5
5-8, 11的XY座標	1.0

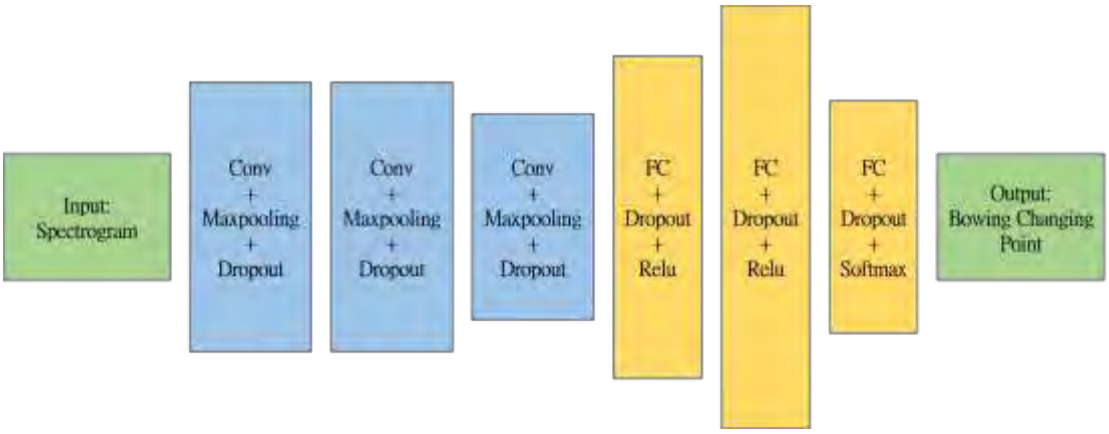
圖十七、骨架權重增加之節點示意圖（左）；骨架權重數值表（右）

（三）方法二：加入限制條件的骨架生成網路

由於由方法一所生成的演奏骨架雖然可隨著音樂起伏而有所變化，但其演奏骨架在流暢性與合理性上我們認為還有進步空間，但由於資料量的關係，這種表現我們認為已經是方法一一也就是以輸出整體骨架為目的的類神經網路一的極限，我們所取得的資料量並不足以使網路抓取到音樂與骨架之間的複雜相關性。

因此我們提出了我們自己設計的方法二。在方法二中我們將生成小提琴演奏骨架的問題拆成三個部分，分別為右手、左手的骨架生成與音樂情緒模型，各自以不同方法處理，最後再合而為一。在此方法中我們不再純粹讓類神經網路去學習音樂與小提琴演奏骨架之間的所有關聯，而是拆成兩個神經網路，分別去學習找出一段音樂中的潛在換弓點安排與找出一段音樂中的音樂起伏，在其他地方如左手指法演奏骨架、換弓點安排轉成右手運弓骨架和音樂起伏轉換為身體律動骨架，我們以我們對小提琴演奏方法的知識對其加入假設，去限制骨架的移動方法，賦予其生成規則。此假設使由方法二生成出的骨架更具備合理性。

1. 右手拉弓骨架生成



圖十八、輸入為時頻圖的換弓點預測網路

上圖為輸入為時頻圖的換弓點預測網路架構示意圖，以一段時間的時頻圖為輸入去預測成該時刻是否為換弓點。不同於方法一中的網路，我們在此網路中加入了卷積層，其目的在於希望讓 CNN 抓出一段時間的二維時譜圖中的特徵，期望藉此可以增進其正確判斷當前時刻是否為潛在換弓點的能力。其中的參數 timestep 即是指一個訓練資料所包含的時頻圖序列長度。此處的時頻圖採樣頻率為 30Hz，因此時頻圖長度的實際秒數為 timestep/30 秒。

（1） 輸入層

輸入維度：(timestep, 128, 1)、timestep：60

（2） 隱藏層一—Conv+Maxpooling+Dropout

通道數：16、卷積大小：(5,5)、池化層大小：(2,2)

激活函數：relu、Dropout 比例：0.25

（3） 隱藏層二—Conv+Maxpooling+Dropout

通道數：16、卷積大小：(5,5)、池化層大小：(2,2)

激活函數：relu、Dropout 比例：0.25

（4） 隱藏層三—Conv+Maxpooling+Dropout

通道數：8、卷積大小：(5,5)、池化層大小：(2,2)

激活函數：relu、Dropout 比例：0.25

（5） 隱藏層四—FC+Dropout

節點數：64、Dropout 比例：0.5、激活函數：relu

（6） 隱藏層五—FC+Dropout

節點數：128、Dropout 比例：0.5、激活函數：relu

（7） 輸出層—全連接層（Fully Connected, FC）

輸出維度：(2)、激活函數：softmax

Loss Function 使用分類交叉熵（Categorical Cross entropy），Optimizer 使用 Adam。Loss Function 公式如下，其中 N 為分類類別，在此網路中輸出格式一個二維向量—是否為換弓點，在這邊 N 為 2：

$$\mathrm{L} = -\sum_{j=1}^{N} y_{true}^{(j)} * \log(y_{pred}^{(j)})$$

由上述網路取得一套換弓點安排後，我們由以下流程去由換弓點轉換成右手的動態演奏骨架。

我們將拉一次滿弓（右手拉弓之最高點移動到最低點為

一次滿弓）分成 28 個影格，並依當前換弓點判斷當前骨架更新方向，詳細骨架更新方法我們在下圖以虛擬碼表示：

```
#================================================
Algorithm 1 : Right Hand Dynamic Skeleton Synthesize
#================================================

Devide a whole bowing samples into 28 frames()

Get a list of changing points data predicted from model()
BowingLengthArrangement = Compute frames of length between every two changing points()

for bowingLength in BowingLengthArrangement:

    # if bowing direction changes in 15 frames(15/30=0.5sec)
    # don't play a whole bow
    # to avoid unreasonable rapid movement

    if bowingLength < 15 frames:
        bow direction changes
        bow for bowingLength frames

    # if bowing direction changes over 15 frames
    # play a whole bow
    if bowingLength > 15 frames:
        bow direction changes
        bow a whole bow for bowingLength frames

end for
```

圖十九、以換弓點標籤生成右手拉弓動態骨架程式虛擬碼

2. 左手指法骨架生成

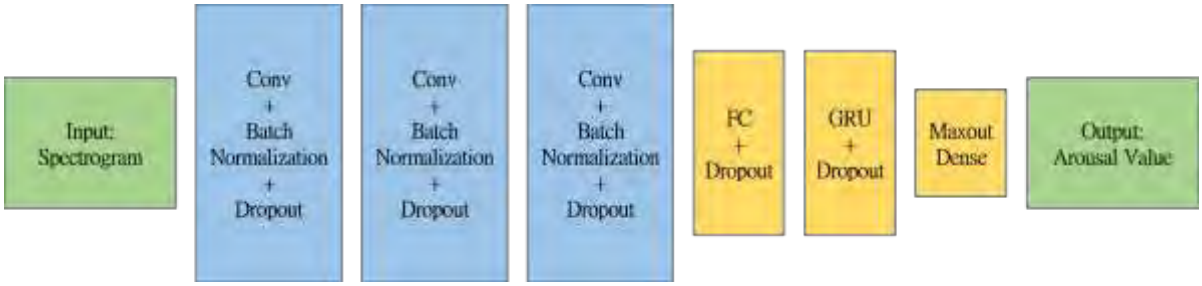
我們利用自動採譜技術（Automatic Music Transcription, AMT）[8]取得音高，並且由音高去判斷左手的位置，由於我們在抓取骨架座標時精細度僅抓到手腕而沒有到手指，因此整個骨架生成的判斷在於由音高判定左手手腕位置（把位）與運弓角度（第幾弦），然而小提琴的同一個音常常會有超過一種按壓的方法，因此在音高生成左手骨架時我們使用貪婪演算法，以低把位為優先考量，也就是在音與音變換時能不換把位就不換，若是必須換則選擇最近的把位按法。音高與其演奏按壓方法對應詳見附錄一。詳細骨架更新方法我們在下圖以虛擬碼表示：

```
#================================================
Algorithm 2 : Left Hand Dynamic Skeleton Synthesize
#================================================

Notes = Get a list of note from AMT model()

for note in Notes:

    if note does not change:
        do not update skeleton

    if note changes:

        # position check
        if I can play it without changing position:
            do not update skeleton

        if I must changing position to play the note
            play the note on the nearest position

        # string check
        play the string according to the position
```

圖二十、以音高更新左手指法動態骨架程式虛擬碼

3. 音樂情緒模型

我們希望讓骨架除了左右手的骨架生成外，其身體能如真人演奏一樣隨音樂節拍與起伏前後搖擺。我們認為身體的擺動應該與 Arousal Value 有關，並在一篇論文中得到驗證[13]，其顯示 Arousal 值與頭擺動加速度的相關係數為 0.75，與身體傾斜的相關係數為-0.70。



圖二十一、輸入為時頻圖的情緒辨識網路架構設計示意圖

上圖為輸入為時頻圖的情緒辨識網路架構設計示意圖，此網路設計參考相關論文 [12]。在此網路中我們設計以輸入一段時間的時頻圖去預測該時段每一個時刻的 Arousal Value。其中的參數 **timestep** 為一段時間的時頻圖長度。這邊的時頻圖取樣頻率為 2Hz，因此所取的時頻圖長度實際秒數即為 **timestep/2** 秒。

（1） 輸入層

輸入維度：(timestep, 64, 1)

（2） 隱藏層一—Conv+BatchNormalization+Dropout

通道數：8、卷積大小：(3,3)

激活函數：relu、Dropout 比例：0.75

（3） 隱藏層二—Conv+BatchNormalization+Dropout

通道數：8、卷積大小：(3,3)

激活函數：relu、Dropout 比例：0.75

（4） 隱藏層三—Conv+BatchNormalization+Dropout

通道數：8、卷積大小：(3,3)

激活函數：relu、Dropout 比例：0.75

（5） 隱藏層四—Reshape

Reshape 大小：(timestep, 64*8)

（6） 隱藏層五—FC+Dropout

節點數：8、Dropout 比例：0.5

（7） 隱藏層六—GRU+Dropout

節點數：16、Dropout 比例：0.5

（8） 輸出層—MaxoutDense

輸出維度：(timestep, 1)、nb_feature=8

Loss Function 使用均方誤差（Mean-Square Error, MSE），Optimizer 使用 Adam。Loss Function 公式如下，其中 N 為由時頻圖序列長度，由此比較每一瞬間預測的 Arousal 值與原本資料中的 Arousal 值之差距，在這邊 N 為 timestep：

$$\mathrm{L} = \frac{1}{N}\sum_{j}^{N}(y_{true}^{(j)} - y_{pred}^{(j)})^2$$

取得情緒標籤後，我們以下列演算法將 Arousal 值轉換為

身體傾斜與頭的擺動。在頭的擺動部分，我們將其設定為以一個小節為單位的前後擺動，第一下強拍（Downbeat）為向前擺動，其後弱拍向後擺動。我們以 madmom 函式庫抓取強拍時間點。詳細音樂情緒的身體骨架的更新方法如下。

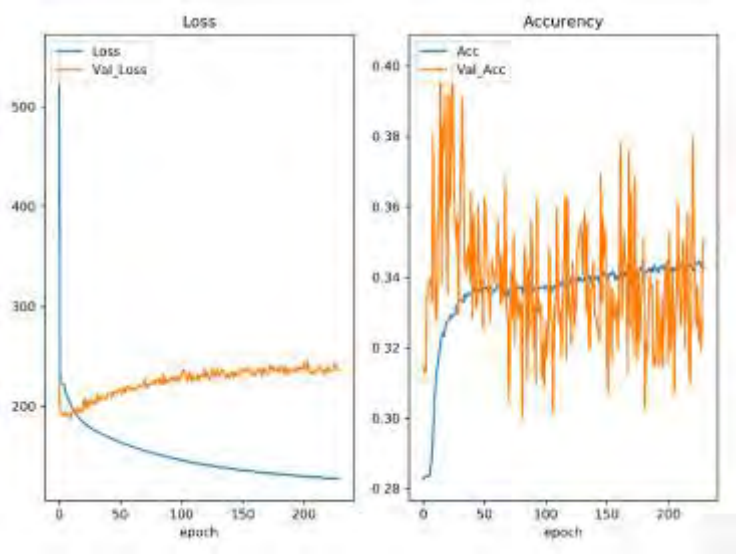
```
#=============================================================
Algorithm 3 : Body Emotion Expression Dynamic Skeleton Synthesize
#=============================================================

Arousal = Get a list of Arousal Value predicted from model
Downbeat_pos = Get downbeat position via madmom

# body tilt
for arousalValue in Arousal:
    update body tilt angle according to arousalValue


# head acceleration
for downbeat in Downbeat_pos:
    compute mean of arousal value between downbeat positions
    set the extent and velocity of the head according to arousalValue
    update head skeleton
```

圖二十二、以 Arousal 更新身體情緒動態骨架程式虛擬碼

伍、 研究結果

一、方法一：生成整體骨架的類神經網路

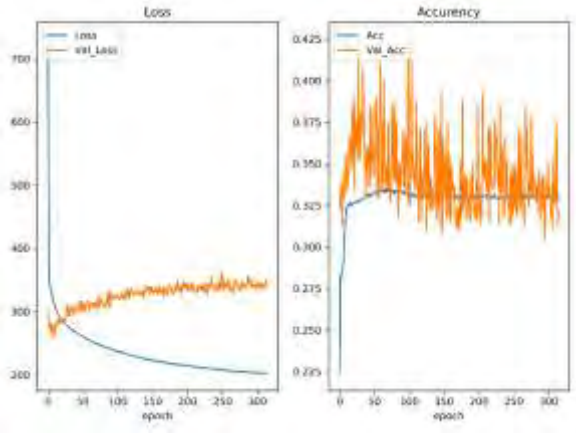
（一）輸入頻譜的 GRU 骨架預測網路



圖二十三、輸入為頻譜的 GRU 網路訓練結果

訓練過後神經網路由圖二十三中可看出，其預測能力在訓練資料集（藍線）隨著訓練時間增加逐漸變好，而在測試資料集（橘線）中則極不穩定。實際將預測出的骨架座標視覺化後便會發現，骨架在動作上能隨著時間與音樂起伏而有所改變，但會有不合常理快速抖動的現象。

（二）輸入頻譜且注重右手弓法的 GRU 骨架預測網路

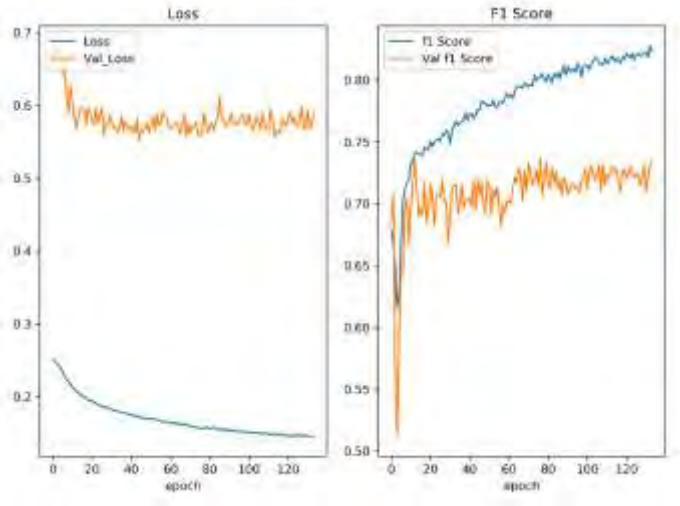


圖二十四、輸入頻譜且注重右手弓法的 GRU 骨架預測網路訓練結果

網路訓練情況從圖二十四中可看出，其訓練過程與第一個網路相似，但由於有對右手骨架加權的關係，實際將預測後的骨架視覺化便可以發現，相較於沒有增加權重的網路，其生成出的右手骨架對於音樂的起伏更為敏感，右手骨架移動幅度更大，但仍會有不合理快速抖動的現象。

二、方法二：加入限制條件的骨架生成網路

（一）換弓點預測網路

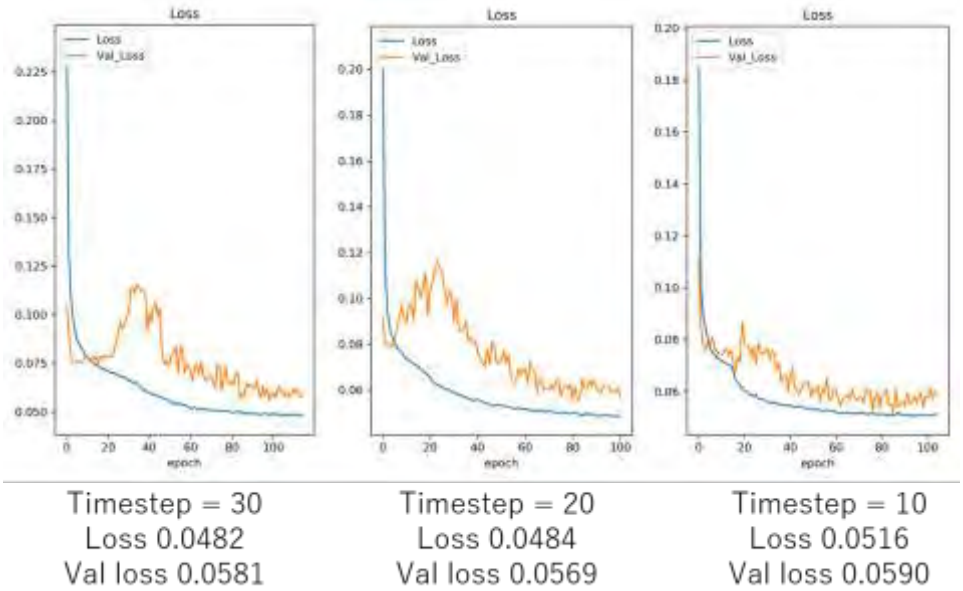


圖二十五、輸入為時頻譜的換弓點預測網路訓練結果

從上圖中的訓練過程與準確率顯示出我們所設計與訓練的換弓點預測網路是有能力對於一段時間的時頻圖音樂資訊去正確判斷是否為換弓點的。

實際對網路的預測能力進行測試，在輸入資料上我們測試了兩種類型的資料，分別是測試資料集內的音訊資料與請真人演奏並實際錄音的音訊資料。將上述兩種音訊資料輸入進換弓點預測網路進行預測，再經過於上文中所提到的演奏骨架流程使網路預測結果視覺化後，我們發現網路在兩種測試資料上皆是有能力找出一套前潛在且合理的換弓點安排的，其弓法切換時間點幾乎都是在音與音切換的時間點上。我們認為小提琴相較於其他樂器本身就是演奏多樣性大的樂器，網路的目的並非在於去學習與原資料集中演奏者使用一模一樣的演奏方法，而是在於學習找出一套合理的換弓點安排，因此雖然從圖二十五看來，網路在測試資料集上的 F1 score 只有達到約七成，但就最後的生成結果看來我們認為是已經足夠好且可被接受的。

（二）音樂情緒模型



圖二十六、輸入為時頻圖的情緒辨識網路訓練結果

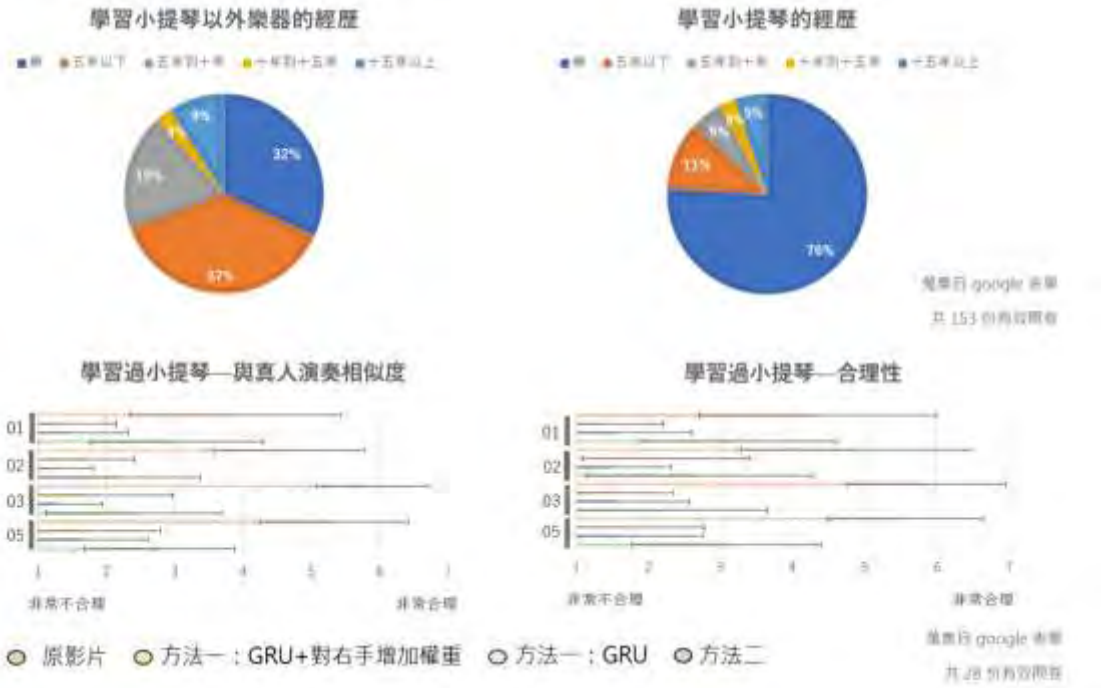
我們測試以不同時頻圖長度作為輸入，去比較網路的表現，最後選用是 timestep 為20的模型。

（三）合併後的演奏骨架

方法二中經三部分分別生成並合併所成最終生成的整體骨架，由於我們以我們對小提琴演奏的背景知識，對其加入了許多假設，對其骨架更新方法加上了許多規則與限制，其整體動態骨架相較於由方法一所生成的骨架合理性與真人相似度上皆大幅提高。

三、主觀問卷分析

對於我們所生成的小提琴動態演奏骨架是否合理，目前並沒有客觀標準可以對我們的結果進行評斷，我們以 google 表單製作了一份主觀問卷，邀請受測者協助我們評斷合理性與真人相似度，以七點量表進行評分，一分為非常不合理（非常不相似），七分為非常合理（非常相似）。



圖二十七、受測者學習樂器經歷（上）；主觀問卷評分結果之平均與標準差（下）

我們總共收集到153份有效問卷，其中大約有四分之一是沒有學習過小提琴的，而有三成是完全沒有學習過小提琴以外樂器的。針對有學習過小提琴的受測者所給出的分數進行分析，我們發現在四首曲子裡，方法二在合理性和與真人相似度的分數上雖然皆比不上原影片，但皆明顯優於方法一。問卷網址附於附錄二。

陸、 討論

一、方法一中 GRU 骨架預測網路有無對注重右手弓法之比較與探討

經由主觀問卷與視覺化後，我們發現由方法一所生成的骨架，無論是否有針對右手增加權重皆能夠隨著音樂的起伏而有所變化，但兩者皆仍有許多不合理的快速抖動，且在合理性和與真人演奏相似度上，分數明顯低於原影片與方法二。

對此我們提出了幾種可能的原因：

（一）輸入資料與輸出資料間相關性過於複雜

無論是有無對右手增加權重，從由頻譜去生成骨架座標資料的結果看來，雖然骨架在動作上能隨著時間而有所改變，但骨架動作會有快速抖動或是骨架比例不符合人體等不合理的情形。經由此實驗，發現我們所設計的 GRU 網路並無法抓取到音樂資料與骨架座標之間的關係，因此我們認為輸入資料與輸出資料間相關性過於複雜，導致在現有資料量下此網路無法抓取到兩者關聯。

（二）問題複雜度太高

在小提琴的演奏骨架當中，其所包含的資訊面向可大致分為左手的指法（把位）、右手的弓法（上下弓）與音樂情緒（身體的

擺動)。在方法一與其他先前研究中皆是透過類神經網路，由音樂資訊去直接生成同時含有不同面向資訊的骨架，然而神經網路無法一次顧及每一面向，導致生成出的骨架任一方面皆沒做好、皆不合理。對於問題複雜性太高的問題，我們才提出了方法二去對骨架運作方法加上了規則，並以上述三面向分部處理，對抓取音樂資料與小提琴骨架關係的問題做了簡化。

二、方法一與方法二生成骨架結果之比較與探討

方法一與方法二的比較中，也就是一次生成整個演奏骨架與分部分生成演奏骨架的比較，我們發現由於在方法二中我們有加入我們對於小提琴演奏的背景知識，其骨架的生成有一部分被我們所給予的假設所限制，因此方法二相較於方法一，從視覺化與主觀問卷評分來看，其生成的音樂演奏骨架在合理性與真人演奏相似度上皆更為合理，且較少有如方法一所生成的演奏骨架所呈現不合理抖動的情況。

三、未來展望

在未來的研究中，我們有以下幾點預計要嘗試與加入的方向：

（一） 增進左手指法的安排多樣性

在左手生成指法的部分，目前我們是貪婪演算法去計算此刻音高的按法，策略是能不換把位就不換把位，對於小提琴的初學者來說一開始在練習時的確是以此方法演奏。但隨著演奏技巧的上升，在真實的小提琴演奏中，往往會有為了音色的需求而選擇更換不同的把位的情況，使演奏在音色豐富度與技巧難度方面上升。我們預計在未來研究中使用隱藏式馬可夫模型（Hidden Markov Model），以把位切換與否作為隱藏狀態，以音樂資訊做為觀察狀態，處理在不同情況下把位更

換的問題。其構想啟發於吉他指法安排的相關研究論文 [7]。

（二） 增進動畫生成之精細度

目前在將骨架資料視覺化與 Demo 上，我們是以 Vpython 函式庫處理，其在畫面處理與生成上僅有一些基本的幾何圖案可以使用。我們預計在未來學習3D 建模，增進 Demo 與生成動畫的精細度，同時也是為了增加此技術在未來的應用性。

柒、 結論

在本研究中我們提出了兩種僅以音樂資料為基礎，透過類神經網路自動生成音樂演奏骨架的方法。方法一沿用先前其他研究論文的網路架構，雖然在損失函數上有對右手增加權重，但其骨架生成結果仍然與先前研究結果情形相似，骨架不斷抖動且不具合理性，與真人演奏骨架仍有一段差距。為了增加合理性，我們提出了自行設計的方法二，透過將問題限縮到預測骨架並對生成骨架過程以背景知識加入骨架變化規則，其成果合理性相較於方法一與先前論文結果大幅增加。

回應到我們最開始的研究目的，目前我們做出了一套流程，能僅以小提琴獨奏錄音檔為輸入，生成合理的小提琴演奏骨架。

捌、 參考資料與其他

[1] Bochen Li, Akira Maezawa, Zhiyao Duan (2018). Skeleton Plays Piano: Online Generation of Pianist Body Movements from MIDI Performance.

[2] Eli Shlizerman, Lucio Dery, Hayden Schoen, Ira Kemelmacher-Shlizerman (2017). Audio to Body Dynamics.

[3] Cho, Kyunghyun; van Merrienboer, Bart; Gulcehre, Caglar; Bahdanau, Dzmitry;

Bougares, Fethi; Schwenk, Holger; Bengio, Yoshua (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation.

[4] Wentao Zhu, Cuiling Lan, Junliang Xing, Wenjun Zeng, Yanghao Li, Li Shen, Xiaohui Xie (2016). Co-Occurrence Feature Learning for Skeleton Based Action Recognition Using Regularized Deep LSTM Networks.

[5] Yong Du, Wei Wang, Liang Wang (2015). Hierarchical Recurrent Neural Network for Skeleton Based Action Recognition.

[6] Ricardo M. Araujo, Gustavo Graña, Virginia Andersson (2013). Towards Skeleton Biometric Identification Using the Microsoft Kinect Sensor.

[7] Matt McVicar, Satoru Fukayama, Masataka Goto (2014). Auto Lead Guitar: Automatic Generation of Guitar Solo Phrases in the Tablature Space.

[8] https://github.com/BreezeWhite/Music-Transcription-with-Semantic-Segmentation

[9] Russell James (1980). A circumplex model of affect.

[10] Mohammad Soleymani, Anna Aljanaki, and Yi-Hsuan Yang (2018). DEAM: MediaEval Database for Emotional Analysis in Music.

[11] S. Butterworth (1930). On the Theory of Filter Amplifiers

[12] Miroslav Malik, Sharath Adavanne, Konstantinos Drossos, Tuomas Virtanen, Dasa Ticha, Roman Jarina (2017). Stacked Convolutional and Recurrent Neural Networks For Music Emotion Recognition.

[13] Birgitta Burger, Suvi Saarikallio, Geoff Luck, Marc R. Thompson, Petri Toiviainen (2013). Relationship Between Perceived Emotions in Music and Music-Induced Movement.

[14] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, Yaser Sheikh (2018). OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields

附錄一、小提琴音高與其對應按法（弦—把位—指）

G3	1-1-0					C5	2-3-4	2-4-3	3-1-2	3-2-1	
G3#	1-1-1					C5#	2-3-4	2-4-3	3-1-2	3-2-1	
A3	1-1-1					D5	2-4-4	3-1-3	3-2-2	3-3-1	
A3#	1-1-1					D5#	2-4-4	3-1-3	3-2-2	3-3-1	
B3	1-1-2	1-2-1				E5	3-1-4	3-2-3	3-3-2	3-4-1	4-1-0
C4	1-1-3	1-2-2	1-3-1			F5	3-2-4	3-3-3	3-4-2	4-1-1	
C4#	1-1-3	1-2-2	1-3-1			F5#	3-2-4	3-3-3	3-4-2	4-1-1	
D4	1-1-4	1-2-3	1-3-2	1-4-1	2-1-0	G5	3-3-4	3-4-3	4-1-2	4-2-1	
D4#	1-1-4	1-2-3	1-3-2	1-4-1	2-1-1	G5#	3-3-4	3-4-3	4-1-2	4-2-1	
E4	1-2-4	1-3-3	1-4-2	2-1-1		A5	3-4-4	4-1-3	4-2-2	4-3-1	
F4	1-3-4	1-4-3	2-1-2	2-2-1		A5#	3-4-4	4-1-3	4-2-2	4-3-1	
F4#	1-3-4	1-4-3	2-1-2	2-2-1		B5	4-1-4	4-2-3	4-3-2	4-4-1	
G4	1-4-4	2-1-3	2-2-2	2-3-1		C6	4-2-4	4-3-3	4-4-2	4-5-1	
G4#	1-4-4	2-1-3	2-2-2	2-3-1		C6#	4-2-4	4-3-3	4-4-2	4-5-1	
A4	2-1-4	2-2-3	2-3-2	2-4-1	3-1-0	D6	4-3-4	4-4-3	4-5-2	4-6-1	
A4#	2-1-4	2-2-3	2-3-2	2-4-1	3-1-0	D6#	4-3-4	4-4-3	4-5-2	4-6-1	
B4	2-2-4	2-3-3	2-4-2	3-1-1		E6	4-4-4	4-5-3	4-6-2	4-7-1	

附錄二、主觀問卷網址

此主觀問卷分為四個版本，此網址會隨機導向任一問卷。

此問卷中的方法二未加入音樂情緒模型，僅有左右手骨架生成模型。

https://skeletonsynthesizing.droppages.com

【評語】052505

此作品研發出一套能僅以一段小提琴的獨奏錄音檔來生成出合理的小提琴演奏骨架的程式，藉以協助動畫師在音樂動畫領域能夠更有效率的製作動畫。此團隊所提出的方法，與現存的最新技術，有顯著的差異，將骨架生成分左右手，並加入情緒模型。此作品技術有深度，且具創新性。

這是非常有趣且有用的應用，希望可以繼續發展，更進一步將設定的把位及拉弓限制拿掉，但還能維持骨架移動的穩定度。

作品海報

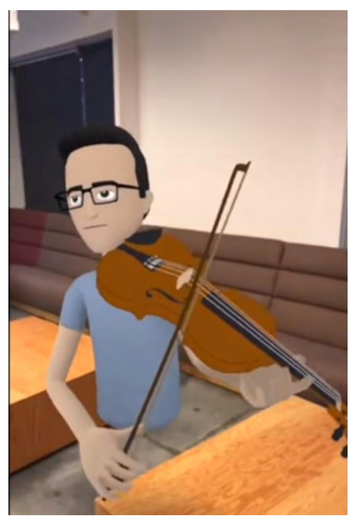
摘要

現今虛擬歌手與虛擬演唱會盛行，啟發於此，我們希望能做出虛擬音樂家與虛擬演奏會。而目前在音樂動畫的領域，若動畫師要製作動畫演奏影片，皆是請真人演奏再透過感測器去取得骨架座標資料，搭配動畫製作技術進而產生演奏影片，此方法不僅耗時且耗費人力成本。若是能將此生成骨架的任務交給電腦自動化生成，將大幅減少時間與人力成本。

此研究以小提琴為例，提出了兩種僅以音樂為基礎，藉由類神經網路技術的輔助，自動生成演奏骨架的方法。架構一延續先前相關論文的網路架構，但結果不如預期。因此本研究提出架構二—加入限制條件的骨架生成網路，運用類神經網路找出合理的演奏弓法，並以我們賦予的生成規則生成演奏骨架。此外，為使生成的演奏骨架更接近真人演奏，本研究在架構二所生成的骨架基礎上再加入了音樂情緒的表達。

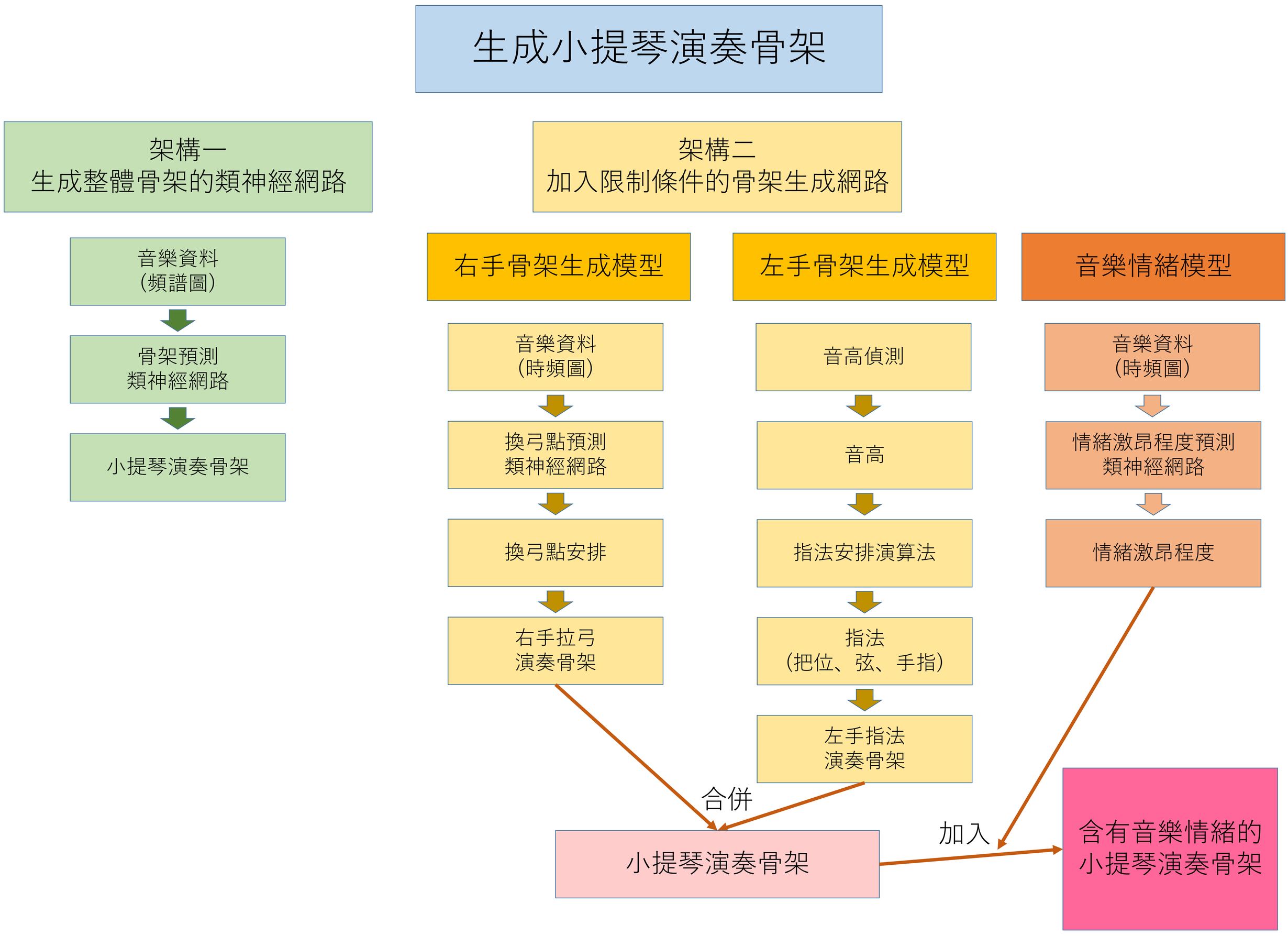
研究結果顯示架構二相較於架構一與先前相關論文能更有效地生成出合理的小提琴音樂演奏骨架。且有音樂情緒表達者比無音樂情緒者又更接近真人演奏。

生成的演奏動作骨架在未來應用上用途甚廣，透過虛擬演奏者，不但能輔助小提琴初學者學習，更能讓音樂藉由虛擬演奏者的視覺化，增進人類的音樂享受體驗，最終達到虛擬演奏會的目標。

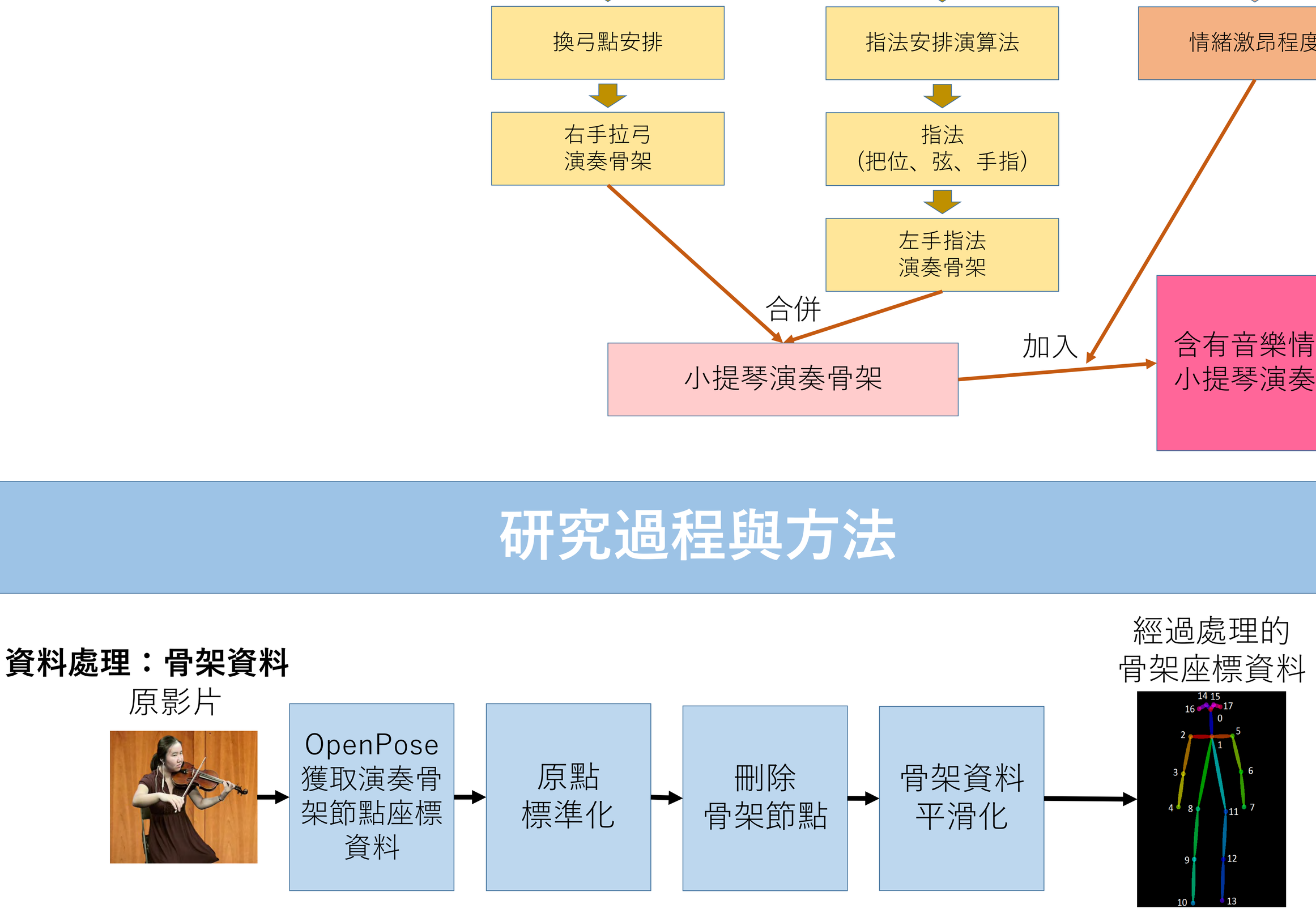


虛擬演奏示意圖

研究流程



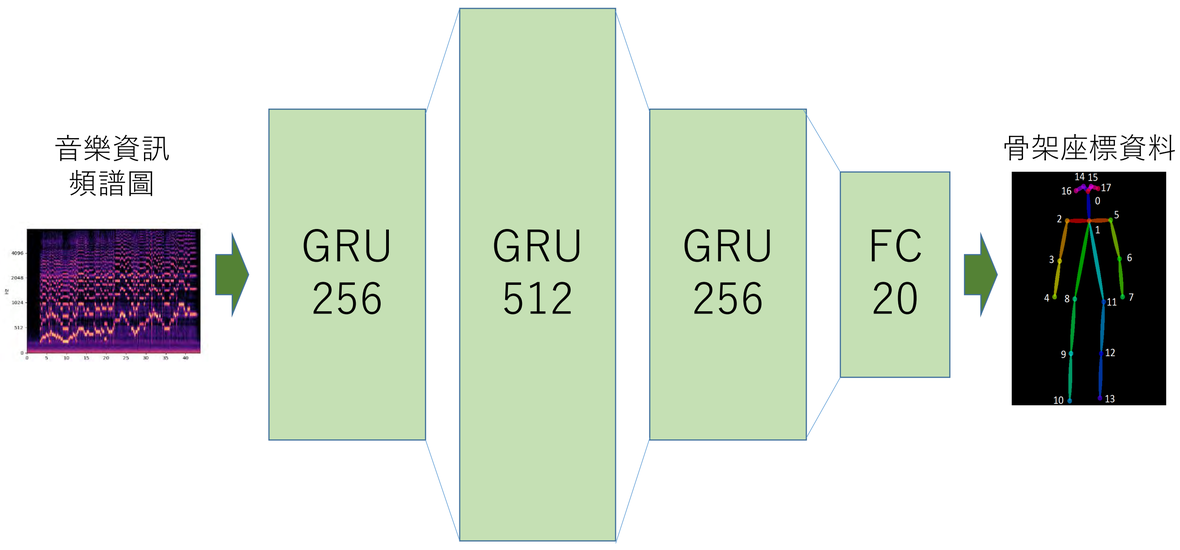
研究過程與方法



研究結果

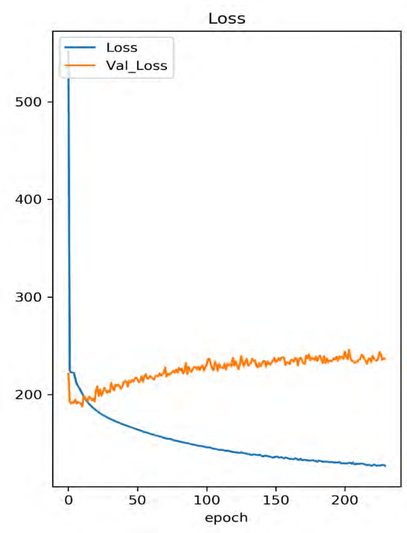
架構一
生成整體骨架的類神經網路

骨架生成網路

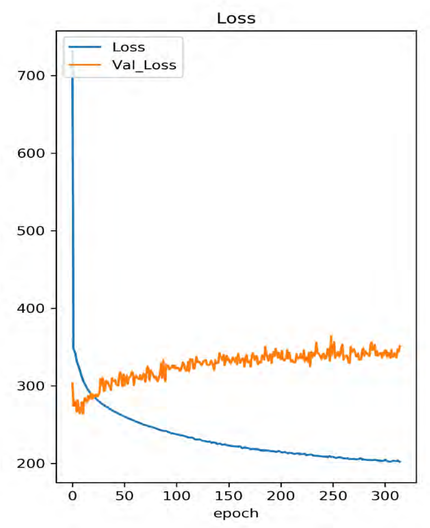


我們在此網路的損失函數（Loss Function）上對右手（拉弓手）的肩膀、手肘和手腕這三點乘上較高的權重，因為我們發現在沒有受過小提琴訓練的人眼中，小提琴動態演奏骨架的合理性很大一部份在於右手拉弓的方式與時間點，因此我們對於右手三個節點乘上了較高的權重，以期望網路能對於右手的骨架預測更加重視，使生成出的骨架更具合理性。

無針對右手增加權重



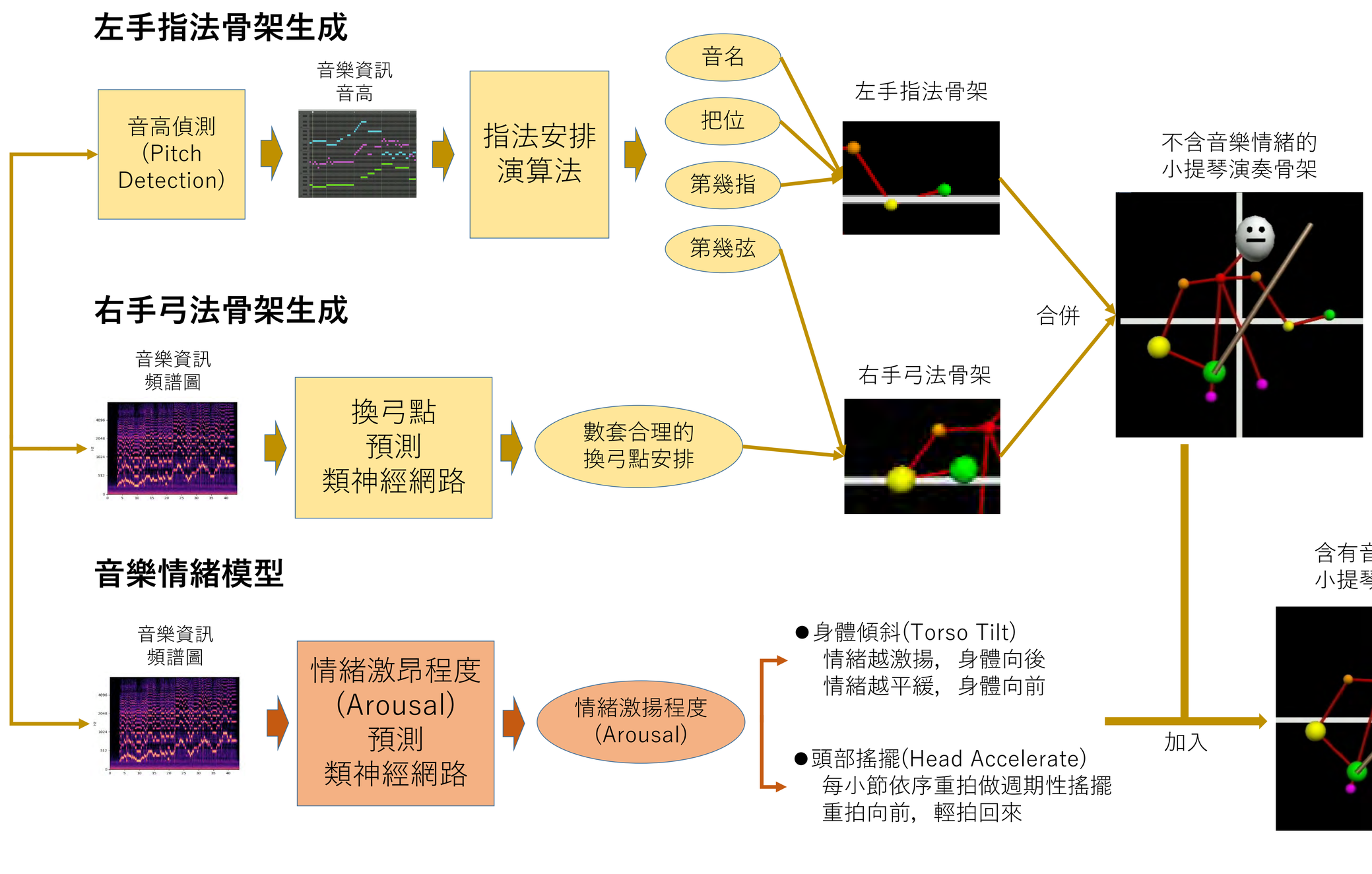
有針對右手增加權重



架構一網路訓練損失函數(Loss Function)圖。
藍線為訓練資料集而橘線為測試資料集 。

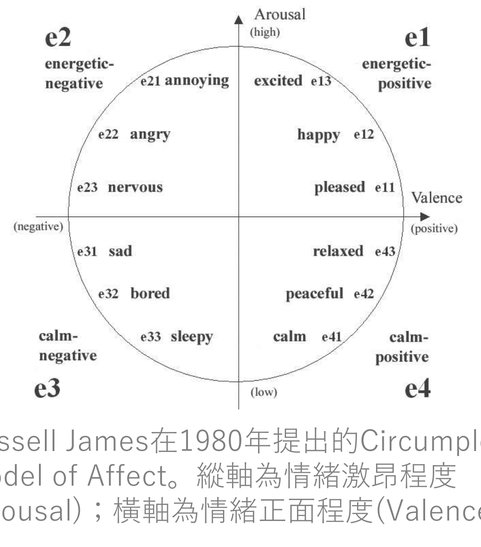
架構二
加入限制條件的骨架生成網路

左手指法骨架生成



由於由**架構一**所生成的演奏骨架雖然可隨著音樂起伏而有所變化，但其演奏骨架在流暢性與合理性上我們認為還有進步空間，且由於**資料量**的關係，這種表現我們認為已經是架構一—也就是以輸出整體骨架為目的的類神經網路—的極限，我們目前所取得的資料量**並不足以使網路抓取到音樂與骨架之間的複雜相關性**。

因此我們提出了我們自己設計的**架構二**。在方法二中我們將生成小提琴演奏骨架的問題**拆成三個部分以簡化問題**，分別為右手、左手的骨架生成與音樂情緒模型，各自以不同方法處理，最後再合而為一。在此方法中我們不再純粹讓類神經網路去學習音樂與小提琴演奏骨架之間的所有關聯，而是拆成兩個神經網路，分別去**學習找出一段音樂中的潛在換弓點安排**與**學習找出一段音樂中情緒激昂程度的變化**。在其他地方如左手指法演奏骨架、換弓點安排轉成右手運弓骨架與音樂情緒的表達方式，我們則以我們對小提琴演奏方法的知識對其加入假設，去限制骨架的移動方法，**賦予其生成規則**。此假設使由架構二生成出的骨架更具備合理性。



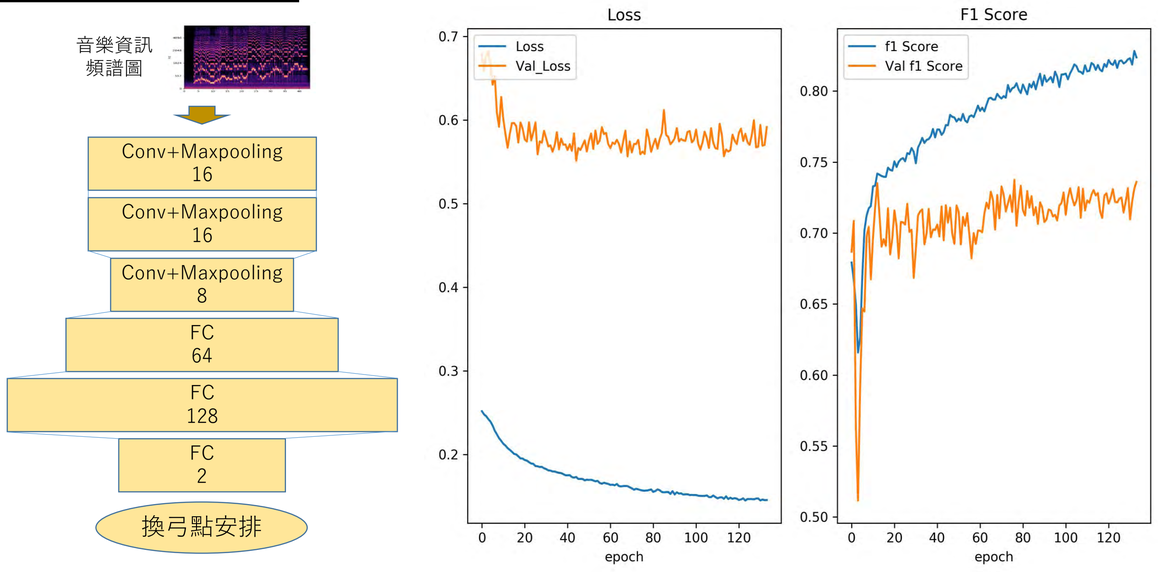
Russell James在1980年提出的Circumplex Model of Affect。縱軸為情緒激昂程度(Arousal)；橫軸為情緒正面程度(Valence)。

	Arousal	Valence
TorsoTilt	−.70***	.29
HandDist	.07	.22
FootDist	−.31	−.38
HeadAcc	.75***	−.36
HandAcc	.76***	−.18
FootAcc	.68***	.02
MoveArea	.30	.31
Fluidity	−.84***	.45*
RotRange	.14	.55**
MoveComp	.08	.42

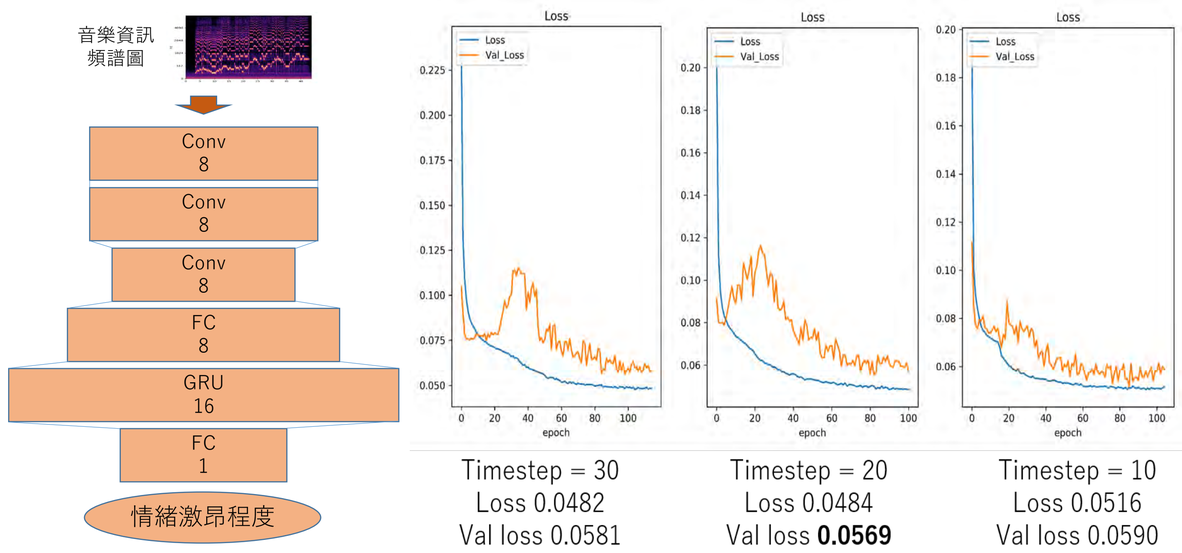
* $p < .05$,** $p < .01$, *** $p < .001$

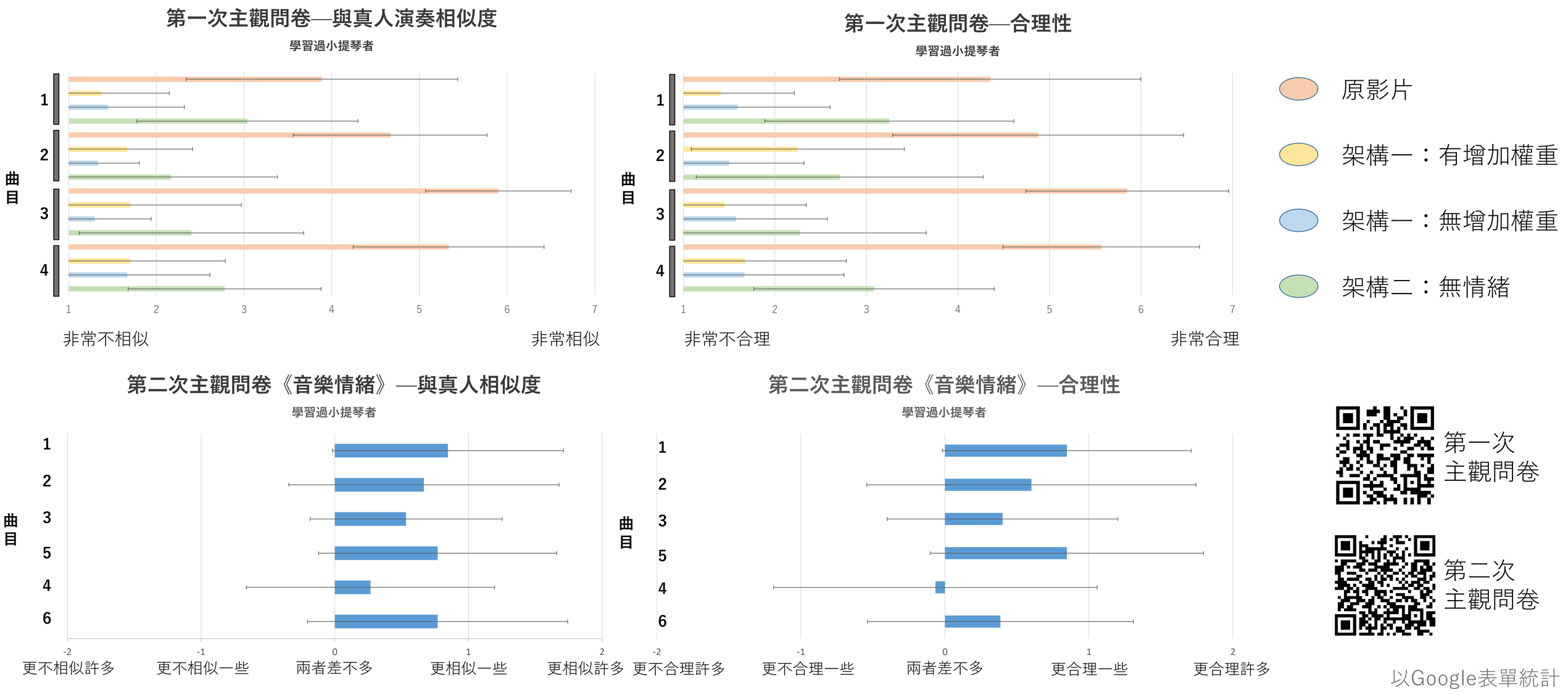
身體動作與情緒之關係。研究指出身體傾斜(TorsoTilt)與情緒激昂程度(Arousal)成顯著負相關；頭部搖擺(HeadAcc)與情緒激昂程度(Arousal)成顯著正相關。圖擷取自論文[3]。

換弓點預測網路



情緒激昂程度預測網路





討論

一、架構一與架構二生成骨架結果之比較與探討

我們發現由於在架構二中我們有加入我們對於小提琴演奏的背景知識，其骨架的生成有一部分被我們所給予的背景知識所限制，因此架構二相較於架構一其生成的音樂演奏骨架更為合理且少有如架構一所生成的演奏骨架所呈現不正常抖動的情況。

二、架構二之生成骨架是否加入音樂情緒結果之探討與比較

架構二中我們加入以情緒激揚程度出發的音樂情緒後，其身體與頭的音樂性擺動，讓骨架演奏時整體視覺上不再死板而更接近真人演奏。

三、從主觀問卷結果分析

在第一次的主觀問卷—架構一與架構二的比較中指出，受測者在合理性與和真人相似度的面向上，給予架構二的分數大多遠高於給架構一的分數，但皆不及真實影片的分數。**為了使骨架能夠更接近真人演奏**，我們加入了音樂情緒模型。在第二次的主觀問卷—架構二有無加入音樂情緒的比較中指出，受測者大多認為有情緒的演奏骨架在合理性與與真人相似度上皆優於無情緒的演奏骨架。此結果使我們生成的小提琴演奏骨架**再次更接近真人演奏**。

結論

未來展望

一、增進左手骨架的安排多樣性

在左手生成指法的部分，目前我們是指定某個音要用某條弦某個把位去演奏，也就是將音高與左手骨架變成一對一的關係，目的是使此左手骨架的生成簡單化，但在真實的小提琴演奏中，不同的音樂家對於同一段音樂往往會有不同的詮釋，而這也是小提琴骨架生成的挑戰性與困難所在。因此我們預期在未來的研究中加入左手指法的安排問題（Arrangement），使架構二在左手骨架生成上能有更有多樣性，使其整體演奏骨架能更接近真人的演奏。

二、將生成模型從規則導向(Rule-Based)轉為資料導向(Data-Based)

目前我們在架構二中皆是以我們所設計的骨架變化規則去生成演奏骨架，但此舉違反目前人工智慧的走向，也就是加入太多人為控制，結果導致生成的演奏骨架在我們賦予的規則之下顯得相較死板。我們之所以會使用規則導向(Rule-Based)，主要原因是目前網路上的開源資料庫不足，在資料量與可信度皆不足的情況下，我們選擇以人為給予的骨架變化規則去生成演奏骨架。
在未來研究中，我們希望能夠藉由取得更多可信度高的開源資料庫，藉此漸漸將人為控制去除，使整體生成骨架能更具有發展性與變化性。

在本研究中我們提出了兩種僅以音樂資料為基礎，透過類神經網路自動生成音樂演奏骨架的方法。架構一沿用先前相關研究論文的生成方法，雖然我們在損失函數上做出改動，但其骨架生成結果仍然與先前研究結果情形相似，骨架不斷抖動且不具合理性，與真人演奏骨架仍有很大一段差距。為了使骨架更接近真人演奏，我們提出了自行設計的架構二，將整體生成流程分成三個小模型以簡化問題，透過類神經網路為輔助與我們所賦予的骨架變化規則，其成果從兩階段的主觀問卷中皆被證實在合理性和與真人相似度的面向上，相較於架構一與先前論文結果大幅增加。回應到我們最開始的研究目的，我們設計出了一套流程與技術，能僅以小提琴獨奏錄音檔為輸入，生成合理與較先前相關研究更接近真人演奏的的小提琴演奏骨架。

參考資料

[1] Bochen Li, Akira Maezawa, Zhiyao Duan (2018). Skeleton Plays Piano: Online Generation of Pianist Body Movements from MIDI Performance.
[2] Eli Shlizerman, Lucio Dery, Hayden Schoen, Ira Kemelmacher-Shlizerman (2017). Audio to Body Dynamics.
[3] Birgitta Burger, Suvi Saarikallio, Geoff Luck, Marc R. Thompson, Petri Toiviainen (2013). Relationship Between Perceived Emotions in Music and Music-Induced Movement.
[4] Miroslav Malik, Sharath Adavanne, Konstantinos Drossos, Tuomas Virtanen, Dasa Ticha, Roman Jarina (2017). Stacked Convolutional and Recurrent Neural Networks For Music Emotion Recognition.