工作日誌

目錄

決定題目的過程

◆ 找尋研究方向

因本身對音樂這領域有興趣，目前構想的研究主題想主要和音樂有關，又剛好 Machine Learning 應用到音樂領域也是非常熱門，所以想做這方面的研究。

◆ 初步研究題目

(1) 音樂風格轉換 Music Style Transfer

(2) 音樂情緒分類 Music Emotion Classification

- 目前做情緒分類大部分都是以四個象限四個情緒為主，是否可以進一步分析更多種情緒？
  →在文獻探討[6]提到的 Russell Model，可改善 4 個情緒分類過於簡易的問題。
- 音樂轉圖像？(透過情緒對應)
  →目前已有研究做應用，目前還沒想到新的點子。

(3) 音樂生成 Music Generation

- 三個方向可研究，「利用歌詞來生成音樂」、「利用音樂來生成歌詞」、「利用音樂生成音樂」
- 目前已閱讀相關文獻關於音樂生成(基於 GAN 或基於 RNN)，其中在下一頁所列的文獻探討[7]所提出的模型在音樂生成上很完整，因此目前正在思考是否還有其他和音樂生成相關的音樂元素還沒被實現的(目前已針對節奏、和旋、旋律做改善)。

◆ 論文題目：人工智慧音樂家(作詞作曲兼具)

目前朝向「音樂生成」部分，第一個部分利用歌詞來生成歌曲，中英文歌詞最大的差別是中文歌詞需要透過斷詞系統(Jieba 斷詞演算法、CKIP 中文斷詞系統)來處理這些歌詞，歌詞可以表現一首歌的情緒，基於分析後的情緒然後生成相對應情緒的歌曲。第二部份利用歌曲來生成歌詞，透過分析音樂元素，如旋律、和旋、強度等，提取音頻中重要的音樂特徵(如目錄中 5.音樂分析相關文獻 (計畫用)所示)，就可以分析出每首歌曲所代表的情緒，進而生成一首歌曲。考慮目前的文獻中，有哪些地方是需要改善的或是有哪些地方還未被實現，將可作為這篇的研究方向。

文獻探討

[1] "Style" Transfer for Musical Audio Using Multiple Time-Frequency Representations, ICLR, 2018

[2] E. Grinstein, N. Q. K. Duong, A. Ozerov, and P. Perez, “Audio style transfer,”, arXiv: 1710.11385, 2017.

[3] M. B. Mokhsin, N. B. Rosli, W. A. W. Adnan, and N. A. Manaf, “Automatic Music Emotion Classification Using Artificial Neural Network Based on Vocal and Instrumental Sound Timbres,” New Trends in Software Methodologies, Tools, and Techniques, 2014, pp. 3–14

[4] C. Lin, M. Liu, W. Hsiung and J. Jhang, "Music emotion recognition based on two-level support vector classification," 2016 International Conference on Machine Learning and Cybernetics (ICMLC), Jeju, 2016, pp. 375-389.doi: 10.1109/ICMLC.2016.7860930

[5] Chia-Hung Yeh & Wen-Yu Tseng & Chia-Yen Chen & Yu-Dun Lin & Yi-Ren Tsai & Hsuan-I Bi & Yu-Ching Lin & Ho-Yi Lin, Popular music representation: chorus detection & emotion recognition, Springer Science + Business Media, Multimedia Tools and Applications, 2014, Volume 73, Issue 3, pp. 2103–2128

[6] Hu, X., Choi, K., & Downie, J. S. (2017). A framework for evaluating multimodal music mood classification. Journal of the Association for Information Science and Technology, 68(2), 273-285.

[7] Zhu, H., Liu, Q., Yuan, N. J., Qin, C., Li, J., Zhang, K., ... & Chen, E. (2018, July). XiaoIce Band: A Melody and Arrangement Generation Framework for Pop Music. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2837-2846). ACM.

論文名稱	"Style" Transfer for Musical Audio Using Multiple Time-Frequency Representations
摘要	神經風格轉移（Gatys et al，2016）已成為使用卷積神經網絡生成不同藝術風格圖像的流行技術。最近在圖像樣式轉換方面的成功提出了一個問題，即是否可以利用類似的方法來改變音樂音頻的「風格」。在這項工作中，我們嘗試在時域中進行長時間高質量的音頻轉換和紋理合成，抓取與音樂風格相關的旋律，節奏和音色的元素，使用具有不同長度和音樂鍵作為例子。我們展示了使用隨機初始化卷積神經網絡將音樂風格的這些方面從一個片段轉移到另一個片段的能力，使用 3 種不同的音頻表示：短時傅立葉變換（STFT）的對數幅度，Mel 頻譜圖和 CQT 轉換頻譜圖，使用這些表示作為產生和修改音樂音頻內容的重要特徵的方式。我們透過仔細設計與音樂音訊的本質互補的神經網路結構，來展示每個表示法的缺點和優勢。最後，我們展示了最引人注目的「風格」轉換例子，利用這些表示的集合來說明捕捉音訊信號的不同期望特徵。
優點	1. 比較 3 種音頻的表示方式，發現使用 Mel 頻譜圖和 CQT 轉換頻譜圖可改善先前的方法，能抓取到有意義的樣式資訊。 2. 成功嘗試完全在時域中執行風格轉移。
缺點	1. 本篇在 style loss 和 content loss 沒有將計算結果列出來。
自評	1. 認為自己在專有名詞上還需多加了解，以方便了解流程圖的內容。

論文名稱	Audio Style Transfer
摘要	圖像之間的「風格轉移」最近成為一個非常活躍的研究課題，由卷積神經網絡（CNN）的力量推動，並且已成為社交媒體中非常流行的技術。本文研究了音頻領域中的類似問題：如何將參考音頻信號的風格轉換為目標音頻內容？我們提出了一個靈活的任務框架，它使用聲音紋理模型來提取表徵參考音頻風格的統計數據，然後是基於優化的音頻紋理合成來修改目標內容。與基於主流優化的視覺傳遞方法相比，所提出的過程由目標內容而不是隨機噪聲初始化，優化的損失僅僅是紋理而不是結構。這些差異被證明是我們實驗中音頻風格轉移的關鍵。為了提取感興趣的特徵，我們研究了不同的體系結構，無論是在其他任務上預先訓練，如在圖像樣式轉移中完成，還是基於人類聽覺系統設計。對不同類型的音頻信號的實驗結果證實了所提出的方法的潛力。
優點	1. 使用 4 種不同的模型(VGG-19、SoundNet、Wide-Shallow-Random network、McDermott)來比較，分別探討何者的成效較好。
缺點	1. 本文未計算出參考樣式音頻和輸出音頻之間的 Style loss，僅提供聲音檔和頻譜圖來判斷差異。 2. 實驗數據的部分較薄弱，無法明確知道該方法是否能真正能達到預期的效果。 3. 作為實驗的音頻數量不多，應該透過更多組音頻來比較測試。
自評	1. 如果未來要繼續做風格轉換這區塊，可以參考本篇的架構，針對風格特徵提取和風格轉換這兩部分。

論文名稱	Automatic music emotion classification using artificial neural network based on vocal and instrumental sound timbres
摘要	檢測歌曲中的情感特徵仍然是各種研究領域的挑戰，尤其是在音樂情感分類（MEC）中。為了將所選擇的歌曲分類為具有特定的情緒或情緒，機器學習的算法必須足夠智能以學習數據特徵以相應地將特徵與精確情緒相匹配。到目前為止，只有少數關於 MEC 的研究利用了歌曲的聲樂部分和歌曲的樂器部分結合的音頻音色特徵。音色特徵是音樂的特性或聲音，它區分人類聲音和樂器中不同類型的聲音產生，如弦樂器，管樂器和打擊樂器。 MEC 中的大多數現有作品都是通過查看音頻，歌詞，社交標籤或兩個或更多類的組合來完成的。問題是從聲樂和樂器聲音特徵中利用音色特徵是否有助於在 MEC 中產生積極效果？因此，本研究目的在於利用人工神經網絡通過從聲樂和樂器聲音片段中提取音頻音色特徵來檢測馬來流行音樂中的情感特徵。該研究的結果將基於對聲樂和樂器音色特徵的操縱來共同改進 MEC，並且有助於音樂信息檢索，情感計算和心理學的文獻。
優點	1. 選擇 ANN 分類器，有別於以往使用 SVM 分類器。 2. 以往的有歌詞的歌曲大多是利用歌詞來判斷情緒，而本篇使用唱歌者的聲音和樂器伴奏的音色來作為分類依據。
缺點	1. 在音頻的部分只能使用 WAV 格式的音檔。 2. 分類器所選擇的分類依據只有音色。 3. 本篇目前只針對馬來西亞的音樂做為研究。
自評	1. 未來可加入更多的音樂元素作為分類的依據，畢竟影響音樂情緒的因素有很多種。 2. 本篇的準確率有機會再提升。 3. 增加訓練樣本的多樣性。

論文名稱	Music emotion recognition based on two-level support vector classification
摘要	音樂情感識別 (MER) 可以檢測音樂片段中人們固有的情感表達。 MER 有助於多媒體理解，音樂檢索和其他與音樂相關的應用。隨著近年來在線音樂內容的數量迅速擴大，最近出現了對情感檢索的需求。以計算方式確定音樂的情感內容是一項跨學科研究，不僅涉及信號處理和機器學習，還涉及對聽覺，心理學，認知科學和音樂學的理解。評估自動音樂情感檢測的一個挑戰是，目前還沒有完善的音樂情感描述情感模型。此外，由於基於聲學特徵的音樂情感識別器的透射率低，因此難以解釋由該機制產生的數據。在這項研究中，提出了一個基於領域知識預先描述的音樂流派和音樂特徵的兩級分類系統。該框架具有利用最合適的聲學信息的優點。實驗將通過衡量不同情緒表達和各種音樂線索之間的相關性來進行。為了驗證整體系統的性能，還將基於音樂特徵與地面真實情感之間的一致性來評估提議模型。
優點	1. 在特徵提取方面，使用到音樂的元素(節奏、音色、音調、動態)，透過這些元素能表達一首歌的情感。 2. 使用到特徵加權的工具(RReliefF)。 3. 採用雙層的 SVM。
缺點	1. 實驗的音頻未提供。 2. 情緒的類別在本篇只分成四種，或許可以增加情緒的多元性。
自評	1. 本文使用到特徵加權的工具(RReliefF)，畢竟音樂元素眾多，可以挑選一兩個當作主要特徵，其餘輔佐用，這樣就能更確定某一特徵的成效性或影響性。

論文名稱	Popular music representation：chorus detection & emotion recognition
摘要	本文提出了一種基於歌曲情感的流行音樂表現策略。首先，通過所提出的合唱檢測算法將一段流行音樂分解為合唱和詩歌片段。從結構化片段中提取三個描述特徵：強度，頻帶和節奏規律性，用於情緒檢測。採用分級 Adaboost 分類器來識別一首流行音樂的情感。音樂的一般情緒根據 Thayer 的模型分為四種情緒：快樂，憤怒，沮喪和放鬆。在 350 個流行音樂數據庫上進行的實驗表明，我們提出的合唱檢測的平均召回率和精確度分別約為 95%和 84%;情緒檢測的平均準確率為 92%。對具有不同歌詞和語言的封面版本的歌曲進行附加測試，結果精確率為 90%。提議方法已經由專業在線音樂公司 KKBOX Inc.測試和驗證，並且顯示出有效且有效地識別各種流行音樂的情緒的有希望的表現。
優點	1. 本篇使用自己的 database 和 MIREX 2009 的 database 來做比較，以證明自己的 database 比較好。 2. 本篇 3 個特徵提取的部份都能得到很好的結果。 3. 階層式分類器在此篇能有很精準的結果。
缺點	1. 有些地方矛盾(前面 XY 軸屬性和後面說特徵值都使用 arousal 有關) 2. Database 的內容應該要針對音樂類型有所挑選(像是舞曲部份，節奏過於相似，無法突顯特別結構) 3. 文中有些公式的參數錯誤。
自評	1. 使用音樂的元素作為分類依據，有別於以往單純只使用以文字為基礎來分類。 2. 我認為在副歌偵測和情緒偵測的準確率有機會再提升。 3. 加大資料集的規模、修改資料集內容。 4. 自己定義的資料集沒有公開內容。 5. 音頻轉成圖片的應用(音頻情緒對應到該情緒的圖片)

論文名稱	A framework for evaluating multimodal music mood classification
摘要	該研究提出了一種音樂情緒分類框架，該框架使用多個和互補的信息源，即音樂音頻，歌詞文本和與音樂片段相關聯的社交標籤。 本文介紹了每個組件的框架和全面評估。 在 18 個情緒類別的大型數據集上的實驗結果表明，結合歌詞和音頻明顯優於使用純音頻功能的系統。 自動特徵選擇技術進一步證明具有減少的特徵空間。 此外，對學習曲線的檢查表明，使用歌詞和音頻的混合系統需要較少的訓練樣本和較短的音頻剪輯，以實現與單獨使用歌詞或音頻的系統相同或更好的分類準確度。 最後但同樣重要的是，性能比較揭示了音頻和歌詞特徵在心情類別中的相對重要性。
優點	1. 使用多模式系統，就是結合歌詞和音頻的分類系統，最後的表現比單純只有音頻的分類器還要好。 2. 提出新的歌詞特徵、特徵選取的方法、fusion 的方法。 3. 3 種特徵選取的方法中，以 chi 檢驗最有效，而且平均使用 65%的訓練樣本就可以得到和其他分類器使用全部樣本一樣的準確率。 4. 歌詞搭配音頻可以利用比較少的訓練樣本和比較短的音頻長度，就可以得到和 single-source(單純使用音頻或歌詞)一樣的效果。
缺點	1. 提到利用 chi-square 的方法得到篩選過後的特徵集 BEST-chi2，但沒有說明是哪些特徵被選取到。 2. 在 2D 空間中繪製 18 種情緒類別，提到是計算情緒類別之間的相對距離，但不知道實際是如何計算的。 3. 在音頻分類上只使用 timbre 音色作為分類依據。
自評	1. 情緒模型上可以參考 Russell 情緒模型，覺得這個模型的情緒蠻多樣的，對音樂的情緒描述是足夠的，這篇是用相對距離來建立模型，我目前偏向 XY 軸使用音樂的特徵。情緒模型過多情緒可能會導致情緒重複性過高。 2. 新想法：使用歌詞特徵訓練出 model(文本分析)，可以知道這首歌是哪一種情緒，再依據情緒的結果自動編曲一首對應情緒的歌，可考慮使用 GAN。

論文名稱	XiaoIce Band: A Melody and Arrangement Generation Framework for Pop Music
摘要	隨著音樂創作知識的發展和近期需求的增加，越來越多的公司和研究機構開始研究音樂的自動生成。然而，以前的模型在應用於歌曲生成時具有局限性，這需要旋律和排列。此外，許多與歌曲質量相關的關鍵因素，如和弦進行和節奏模式都沒有得到很好的解決。特別是，如何確保多軌音樂和諧的問題仍未得到充分發掘。為此，我們提出了一個關於流行音樂生成的重點研究，其中我們考慮了旋律生成的和弦和節奏影響以及音樂安排的和諧。我們提出了一種名為 XiaoIce Band 的端到端旋律和排列生成框架，該框架生成一個旋律音軌，其中包含幾種類型樂器演奏的幾個伴奏音軌。 具體而言，我們設計了基於 Chord 的節奏和旋律交叉生成模型 (CRMCG) 來產生具有和弦進行的旋律。我們提出了一種多樂器協同安排模型 (MICA)，它使用多任務學習來進行多軌音樂安排。最後，我們對現實世界的數據集進行了大量實驗，結果證明了小冰帶的有效性。
優點	1. 提出了一種基於音樂知識的旋律和編曲生成框架，稱為小冰樂隊 2. 可以同時生成多種樂器伴奏的旋律。 3. 旋律生成部分，提出 CRMCG 模型，利用和弦進行來引導旋律進行，以及透過節奏來學習歌曲的結構。 4. 編曲生成部分，提出 MICA 模型，在解碼器層的每一步使用其他音軌資訊來提高性能並確保多音軌的和諧性。 5. 經過大量實驗，無論是指標評估或是人工評估，本篇的系統都比其他模型有更好的性能。
缺點	1. 在 MLP cell 的部分，做法有點粗糙，應該做更深入的分析。
自評	1. 本文結合了和弦進行，提出了一個 Chord Accuracy 和弦精確度，作者沒有提到說是怎麼識別生成音樂所屬的和弦？ 2. 這篇使用的數據集沒給，導致無法客觀評估生成水準。 3. 這篇已經可以突破許多音樂生成上的限制，是否可以加入其他特徵或是輸入(目前未使用到的)，創新的點要思考一下。

論文報告 PPT

Popular music representation：chorus detection & emotion recognition
https://drive.google.com/file/d/1IGuRQNYM8EnZHIbMUs5a-Rw9KUufb3z_/view?usp=sharing

A framework for evaluating multimodal music mood classification
https://drive.google.com/file/d/1aHRgLONvdW_4WKzRmJfyvL5Pdf5jlDSI/view?usp=sharing

XiaoIce Band: A Melody and Arrangement Generation Framework for Pop Music
https://drive.google.com/file/d/1e01oJROu0jY8zmQxQcB1bUwXN9Gf5HeA/view?usp=sharing

音樂分析相關文獻 (計畫用)

音樂可以被視為一種具有獨特語法的語言，音樂作曲涉及許多音樂元素（如圖 1 所示[1]），是由各種基本元素相互結合而成的，例如音色、音高、節奏、和弦等。許多音樂元素已被應用於音樂情緒分類、音樂生成、音樂類型分類等，如表 1 所示。

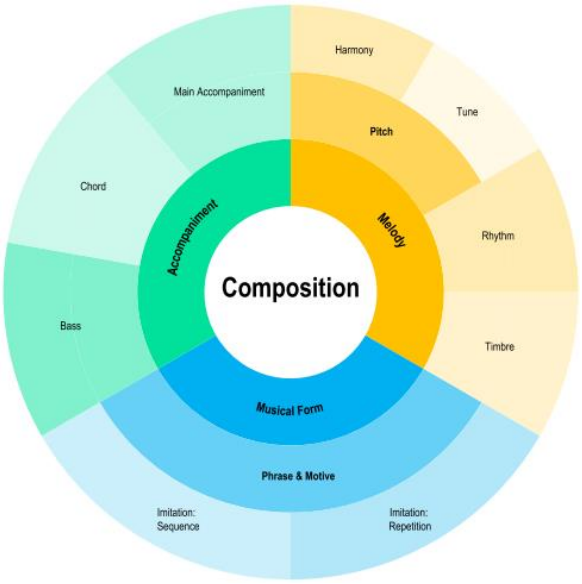

圖 1 音樂作曲的音樂元素

表 1 音樂元素相關應用

應用	文獻	使用之音樂元素
音樂情緒分類	Hu,2017[2]、Barthet,2013[3]、Yang&Chen,2012[4]	音色
	Yeh,2014[6]	強度、節奏、頻帶能量
音樂檢索系統	Kuo F,2009[7]、Mulder T,2006[8]	旋律
	Tzacheva AA,2010[9]、Fujigara H,2009[10]	音色
音樂生成系統	Zhu H,2018[11]、Monteith K,2012[12]	旋律、和弦、節奏
音樂類型分類	Johnson-Roberson C,2017[13]	音色、節奏
	Huang Y F,2014[14]	音高、音色、節奏、強度

除了綜合以上過去的文獻，本研究也加入其他元素來探討，圖 2 列出 5 個最關鍵的元素以及特定的術語和概念。音樂分析即為針對某一音頻或樂段從中分析出這些音樂元素，該如何提取音頻中重要的音樂特徵成了一大課題。

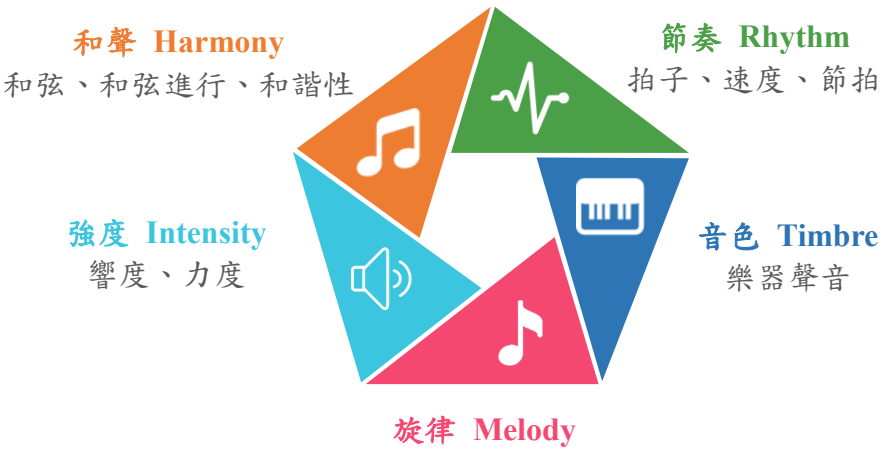

圖 2 音樂五大關鍵元素

以下將介紹關鍵音樂元素提取的方法：

3.1 旋律（Melody）

音樂的首要要素，按照一定的音高、音長的單聲部所結合的序列，它是由許多音樂基本要素結合而成，如節奏、音色、音高等。

3.2 節奏（Rhythm）

節奏隱含在旋律的表現中，可以說是音樂的骨架，音樂中的重拍和弱拍周期性地、有規律地重複進行，進而得知歌曲的結構。

節奏可細分為兩個部分：節拍（Metre）和速度（Tempo）。節拍通常用分數表示，分子表示每小節中單位拍的數目，分母表示單位拍的音符時值，例如 2/4，指的是每小節有兩拍，每拍是四分音符。速度決定了一段音樂的快慢，影響作品的情感與演奏難度（表 2），音樂速度表示法通常以每分鐘多少拍（beats per minute，BPM）作量度單位。

表 2 音樂速度術語表

速度術語	中文名詞	中文意義	速度範圍
Largo	廣板	寬廣的、宏偉的、莊嚴的	bpm = 40~60
Adagio	慢板	悠閒的、柔和的、緩慢的	bpm = 66~76
Andante	行板	行走的、流動的	bpm = 76~108
Moderato	中板	中庸的、不疾不徐的	bpm = 108~120
Allegro	快板	愉悅的、歡欣的	bpm = 120~168
Presto	急板	生動活潑的	bpm = 168~200
Prestissimo	最急板	快的、急的、立刻的	bpm = 200~208

音樂中會出現一些相同重複的結構，以圖 3 來說，藍色和紅色框起來的地方分別代表相同的節奏。Yeh 等人[6]提出一個判別節奏是否具有規律性，利用檢測函數計算相鄰幀的頻譜差異，再透過峰值距離的標準差來確認節奏的規律性。

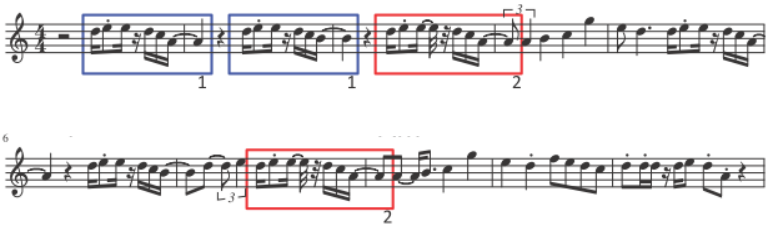

圖 3 歌曲 We Don’t Talk Anymore 樂譜之節奏

節奏常被作為一個重要的特徵依據，在音樂類型分類方面，Tzanetaki[22]提出使用節拍直方圖（Beat Histogram，BH）進行類型分類，計算時域中包絡訊號（Envelope Signal）的自相關函數，觀察自相關函數的峰值，分析音樂的潛在規律性。節拍直方圖模擬包絡訊號中表現規律性的分佈，可獲得節奏特徵；在音樂情緒分類方面，Chua, Bee Yong 等人 [23]和 Yang, Y. H 等人[24]的文獻中，可以發現歌曲情緒和節奏有高度相關，如悲傷的歌曲節奏緩慢，而憤怒的歌曲節奏快速。

3.3 強度（Intensity）

強度是指曲譜或音樂表演中音的強弱，決定於音頻的振動幅度（即振幅大小），又稱為響度、動態。Yeh 等人[6]提出強度為頻域的能量總和，每一幀的每一個傅立葉係數進行加總。Senac 等人[15]使用短時距分析（Short-term Analysis），將聲音先切成音框（Frame），每個音框長度大約 20 ms，再根據音框內的訊號來進行分析。

歌曲的強度特徵可藉由「振幅（Amplitude）」和「低能量（Low energy）」來取得。振幅是透過訊號的均方根（Root Mean Square，RMS）來量化時域中聲音波形的幅度，C. Lin 等人[16]和[25]提到可以透過計算均方根來檢測音樂片段的強度，離散時間信號的 RMS 根據音節的響度來估計能量特徵，平穩而平靜的音樂軌道的 RMS 能量低於高能量音樂。低能量表示能量低於平均值的幀所佔的百分比[22]。

3.4 音色（Timbre）

聲音的特色，又稱為音質，決定於聲音的波形，即使在同一音高和同一響度的情況下，也能讓人區分開來。音色可以分為兩種類型：樂器和聲樂。對於樂器音樂，許多樂器都有自己獨特的音調，在作曲時應予以考慮。音色的不同，代表基本週期的波形不同，若要從基本週期的波形來直接分析音色，是一件很困難的事。因此，要將每個音框進行頻譜分析（Spectral Analysis），把一個音框訊號拆解成在不同頻率的分量，然後才能進行比對或分析。在頻譜分析時，最常用的方法就是「快速傅立葉轉換」（Fast Fourier Transform，FFT），將時域（Time Domain）的訊號轉換成頻域（Frequency Domain）的訊號，進而知道每個頻率的訊號強度，如圖 4 所示。

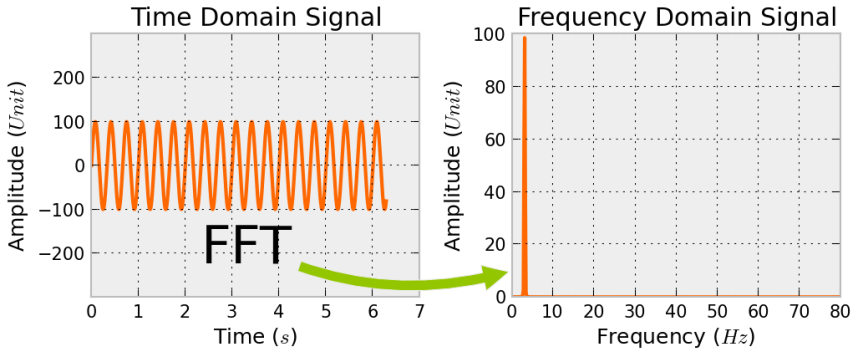

圖 4 快速傅立葉轉換

音色可以作為評估樂器分配的標準。Mokhsin 等人[17]提出音色特徵提取主要是透過三個方法：頻譜滑動率（Spectral Rolloff），測量音樂信號中的頻率偏度；頻譜質心（Spectral Centroid），用來表示聲音訊號組成頻率的平均值，也可定義為聲音的明亮度；過零率（Zero-Crossing Rate），在每個音框中，音訊通過零點的次數。Wang 等人[18]從中國江南小調提取特徵來符合函數，使用頻譜質心來作為音色的評價標準。

音色也可以作為辨別音樂和語音的標準特徵，計算短時傅立葉變換（Short-Time Fourier Transform，STFT），並針對每個短時間的聲音幀進行計算。Logan, B [19]探討了使用梅爾頻率倒譜係數(Mel-Frequency Cepstrum Coefficients，MFCC)來分離音樂和語音。表 3 列出在音色特徵提取時可使用的標準。

表 3 音色特徵的特徵型態

特徵型態	相關文獻
過零率 (Zero Crossing Rate，ZCR)	Bergstra J[25]、Morchen F[26]
頻譜質心 (Spectral Centroid，SC)	Bergstra J[25]、Morchen F[26]、Lu L[27]
頻譜滑動率 (Spectral Rolloff，SR)	Bergstra J[25]、Morchen F[26]、Lu L[27]
頻譜變遷度 (Spectral Flux，SF)	Lu L[27]
頻寬 (Spectral Bandwidth，SB)	Morchen F[26]、Lu L[27]
振幅頻譜包絡 (Amplitude Spectrum Envelop，ASE)	Kim H G[28]、Lee C H[29]
梅爾頻率倒譜係數 (Mel-Frequency Cepstrum Coefficients，MFCC)	Bergstra J[25]、Mandel M I[30]、Shen J[31]

3.5 和聲（Harmony）

和聲包括和弦（Chord）及和弦進行（Chord Progression），基本單位是和弦，前者涉及一組上下垂直同時發聲的多個音高；後者是各個和弦的先後連接，如圖 5 所示，可發現和弦進行為 F→G→Am→Em，依此順序反覆多遍。

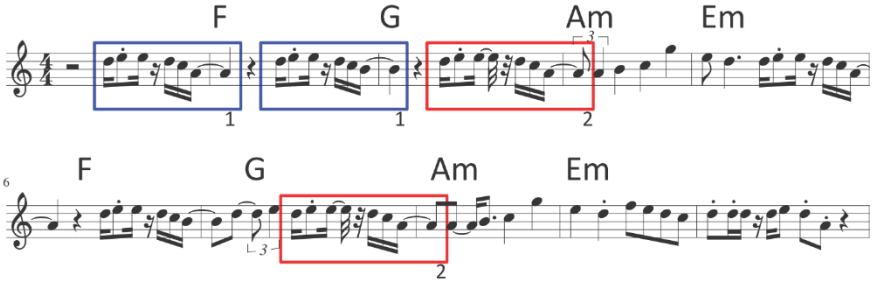

圖 5 歌曲 We Don't Talk Anymore 樂譜之和旋進行

和聲對音樂作曲至關重要，適當的音高安排可以達到和諧性，一些研究利用這概念來解決音樂作曲中的和諧問題，Marques 等人[19]提出了一種功能函數，有利於同時發生的音形成和弦，特別是主和弦和次和弦，形成和諧，Chang 和 Jiau [20] 透過旋律採用合適的和弦來解決和諧問題，但目前所提出的方法還無法達到最好的和諧性。

參考文獻

[1] Liu, C. H., & Ting, C. K. (2017). Computational Intelligence in Music Composition: A Survey. IEEE Trans. Emerging Topics in Comput. Intellig., 1(1), 2-15.

[2] Hu, X., Choi, K., & Downie, J. S. (2017). A framework for evaluating multimodal music mood classification. Journal of the Association for Information Science and Technology, 68(2), 273-285.

[3] Barthet, M., Fazekas, G., & Sandler, M. (2013). Music emotion recognition: From content-to context-based models. In M. Aramaki, M. Barthet, R. Kronland-Martinet, & S. Ystad (Eds.), From sounds to music and emotions (pp. 228–252). Berlin, Heidelberg: Springer.

[4] Yang, Y., & Chen, H. (2012). Machine recognition of music emotion: A review. ACM Transactions on Intelligent Systems and Technology (TIST), 3(3), Article 40, 1–30.

[5] Trohidis, K., Tsoumakas, G., Kalliris, G., & Vlahavas, I. (2008). Multi-label classification of music into emotions. In Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR) (pp. 325–330). Philadelphia: ISMIR.

[6] Yeh CH, Tseng WY, Chen CY, Lin YD, Tsai YR, Bi HI, Lin YC, Lin HY (2014) Popular music representation: chorus detection & emotion recognition. Multimedia Tools and Application 73(3):2103–2128. doi:10.1007/s11042-013-1687-2

[7] Kuo F, Shan M (2009) Music retrieval by melody style. In: Proc Int Symp on Multimed, pp 613 – 618

[8] Mulder T, Martens J, Pauws S, Vignoli F, Lesaffre M, Lenman M, Baets B, Meyer H (2006) Factors affecting music retrieval in query by melody. IEEE Trans Multimedia 8(4):728 – 739

[9] Tzacheva AA, Bell KJ (2010) Music information retrieval with temporal features and timbre. Springer Act Media Technol 6335:212 – 219

[10] Fujigara H, Goto M, Kitahara T, Okuno HG (2009) A modeling of singing voice robust to accompaniment sounds and its application to singer identification and vocal-timbre-similarity-based music information retrieval. IEEE Trans Audio Speech Lang Process 18(3):638 – 648

[11] Zhu, H., Liu, Q., Yuan, N. J., Qin, C., Li, J., Zhang, K., ... & Chen, E. (2018, July). XiaoIce Band: A Melody and Arrangement Generation Framework for Pop Music. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2837-2846). ACM.

[12] Monteith, K., Martinez, T. R., & Ventura, D. (2012, May). Automatic Generation of Melodic Accompaniments for Lyrics. In ICCC (pp. 87-94).

[13] Johnson-Roberson, C., & Sudderth, E. (2017). Content-Based Genre Classification and Sample Recognition Using Topic Models. Cited on, 100.

[14] Huang, Y. F., Lin, S. M., Wu, H. Y., & Li, Y. S. (2014). Music genre classification based on local feature selection using a self-adaptive harmony search algorithm. Data & Knowledge Engineering, 92, 60-76.

[15] Senac, C., Pellegrini, T., Mouret, F., & Pinquier, J. (2017, June). Music feature maps with convolutional neural networks for music genre classification. In Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing (p. 19). ACM.

[16] C. Lin, M. Liu, W. Hsiung and J. Jhang, "Music emotion recognition based on two-level support vector classification," 2016 International Conference on Machine Learning and Cybernetics (ICMLC), Jeju, 2016, pp. 375-389.doi: 10.1109/ICMLC.2016.7860930

[17] Mokhsin, M. B., Rosli, N. B., Wan Adnan, W. A., & Abdul Manaf, N. (2014). Automatic music emotion classification using artificial neural network based on vocal and instrumental sound timbres. In New Trends in Software Methodologies, Tools and Techniques - Proceedings of the 13th SoMeT 2014 (pp. 3-

14). (Frontiers in Artificial Intelligence and Applications; Vol. 265). IOS Press. DOI: 10.3233/978-1-61499-434-3-3

[18] Logan, B. (2000). Mel Frequency Cepstral Coefficients for Music Modeling. ISMIR.

[19] Wang, Xin & Zhan, Ying & Wang, Yuanzhong. (2015). Study on the composition rules for Chinese Jiangnan ditty. 492-497. 10.1109/ICIST.2015.7289022.

[20] M. Marques, V. Oliveira, S. Vieira, and A. C. Rosa, "Music composition using genetic evolutionary algorithms," in Proceedings of the IEEE Congress on Evolutionary Computation, 2000, pp. 714–719.

[21] C.-W. Chang and H. C. Jiau, "An improved music representation method by using harmonic-based chord decision algorithm," in Proceedings of the IEEE International Conference on Multimedia and Expo, 2004, pp. 615–618.

[22] Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. IEEE Transactions on speech and audio processing, 10(5), 293-302.

[23] Chua, Bee Yong & Monash University. Gippsland School of Information Technology (2007). Automatic extraction of perceptual features and categorization of music emotional expressions from polyphonic music audio signals.

[24] Yang, Y. H., Lin, Y. C., Su, Y. F., & Chen, H. H. (2008). A regression approach to music emotion recognition. IEEE Transactions on audio, speech, and language processing, 16(2), 448-457.

[25] Bergstra, J., Casagrande, N., Erhan, D., Eck, D., & Kégl, B. (2006). Aggregate features and a da b oost for music classification. Machine learning, 65(2-3), 473-484.

[26] Morchen, F., Ultsch, A., Thies, M., & Lohken, I. (2006). Modeling timbre distance with temporal statistics from polyphonic music. IEEE Transactions on Audio, Speech, and Language Processing, 14(1), 81-90.

[27] Lu, L., Liu, D., & Zhang, H. J. (2006). Automatic mood detection and tracking of music audio signals. IEEE Transactions on audio, speech, and language processing, 14(1), 5-18.

[28] Kim, H. G., Moreau, N., & Sikora, T. (2004). Audio classification based on MPEG-7 spectral basis representations. IEEE Transactions on Circuits and Systems for Video Technology, 14(5), 716-725.

[29] Lee, C. H., Shih, J. L., Yu, K. M., & Lin, H. S. (2009). Automatic music genre classification based on modulation spectral analysis of spectral and cepstral features. IEEE Transactions on Multimedia, 11(4), 670-682.

[30] Mandel, M. I., & Ellis, D. (2006). Song-level features and SVM for music classification. In International Symposium on Music Information Retrieval (ISMIR).

[31] Shen, J., Shepherd, J., Cui, B., & Tan, K. L. (2009). A novel framework for efficient automated singer identification in large music databases. ACM Transactions on Information Systems (TOIS), 27(3), 18.

音樂生成相關文獻 (計畫用)

近幾年來，隨著音樂創作知識的發展和近期需求的增加，越來越多的公司和研究機構開始研究音樂的自動生成，音樂生成是一項具有挑戰性的任務，目前文獻中已經提出了各種方法，不管是使用深度學習或是普遍的機器學習方法都有將音樂生成的核心方法展現出來，但是，目前的方法在音樂生成時仍具有侷限性，因此，本研究之音樂生成文獻探討將分為兩個部份：(1)深度學習自動音樂生成之方法回顧、(2)音樂生成之侷限性。

- **方法回顧**

1. 馬可夫模型(Markov models)

Papadopoulos & Wiggins[1]所發表的論文將過去自動化產生音樂之方式進行分類，將電腦自動作曲方式分成六大類，而馬可夫模型被分為「數學模型(Mathematical models)」，是一種典型的數據驅動統計方法。

Chordia 等人[2]、Pachet 等人[3]、Fernández 等人[4]使用馬可夫模型來生成音樂；Whorley 等人[5]應用基於多視點方法的馬爾可夫模型來生成具有四部協調的音樂。Kaliakatsos-Papakostas 等人[6]使用隱馬爾可夫模型(HMM)結合中間固定和弦來約束，該組合稱為約束隱馬爾可夫模型(CHMM)。約束是由一個演算法過程指定的，該過程的音樂結構有較高的表現水準，或者可以手動引入。Eigenfeldt 等人[7]提出了一種用於音樂作曲家識別的加權馬爾可夫鏈模型，透過對連續的音符對進行建模，可以抓取單聲道音樂作品的結構片斷資訊，並考慮音高和音程(即兩個音之間的距離)，發現加權馬爾可夫鏈方法比簡單馬爾可夫鏈模型具有更好的處理效果，這個過程確保了作曲家的特徵識別，因此，加權馬爾可夫鏈模型可以用作生成模型。

綜合上述的應用，使用馬可夫模型可成功地生成音樂，時間複雜度低，可達到即時運算效能，但缺點是需要手動提取特徵，這種較傳統的方法需要大量的人力和領域的相關知識。

2. 長短期記憶模型(Long Short-Term Memory，LSTM)

Eck 等人[8]探討了過去大多數遞歸神經網絡在音樂生成中遇到的問題，並探討 LSTM 是否可以學習和弦結構和旋律結構，然後在生成歌曲時使用該學習到的結構，透過兩個實驗，結果證明，與 Mozer[9]相比，基於 LSTM 的音樂生成模型成功地學習了藍調音樂形式的整體結構，並且能夠以這種風格創作出新穎的旋律。

Chu 等人[10]提出一個深度學習的 LSTM 模型(hierarchical recurrent network)，利用 100 個小時的流行歌曲的 midi_dataset 進行訓練(Midi collection https://goo.gl/4moEZ3)。在這模型裡整合了一些音樂相關的 general knowledge，在 RNN 模型中分成 4 個結構，模型設計和合成的過程中也考慮了 scale 和 chord 等音樂因素，模型結構如下圖 1 所示。

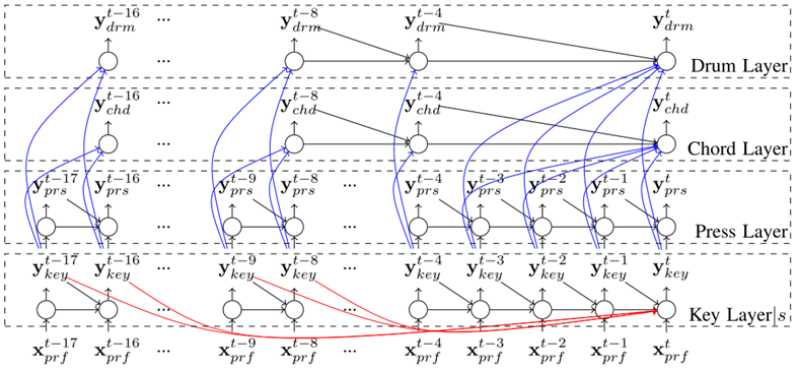

圖 1　Chu 等人之模型框架

本篇還展示了一些應用，比如說生成音樂的過程中同時生成跳舞的小人（如下圖 2），以及嘗試用 neural image captioning 的辦法生成歌詞或者朗誦。

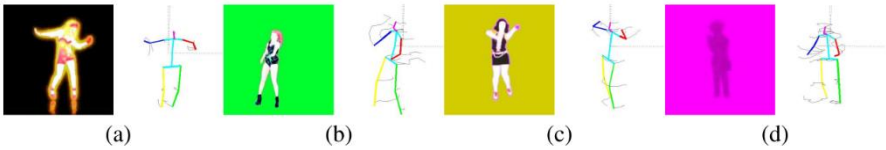

圖 2　生成音樂之跳舞小人

Johnston[11]使用 LSTM 循環網絡並透過訓練音樂集合來生成音樂，音樂集合中都以文本形式 ABC 表示，將音樂文件中所有先前的字符作為輸入，以此為基礎預測下一個字符。透過不斷地執行此操作，生成了新歌曲，並測試不同類型的架構，也可以生成音樂。這篇所提出的方法只適用於簡單的歌曲，因為更複雜的複音歌曲不能以正確的格式標記。

使用 LSTM 生成音樂可以改良使用 RNN 生成音樂時缺乏整體結構的缺點，雖然 RNN 可以學習逐個音符生成，甚至可以重現一段樂句，但是生成的結果缺少整體性，主要原因是 RNN 無法追蹤音樂結構上的時間事件。

3. 遞歸神經網絡(Recurrent Neural Networks，RNN)

Hadjeres 等人[12] 提出一種基於 RNN 的「深度巴哈(DeepBach)」的生成模型，該模型能夠透過使用類似吉布斯採樣(Gibbs sampling) 程序來產生四部分的合唱類似於巴哈的合唱音樂作品，號稱能模仿巴哈的四部聖詠的風格來創作新曲。使用 352 首由巴哈創作的四聲部聖詠，對類神經網路模型進行訓練，接著再讓模型模仿巴哈，重新編寫和聲旋律，下圖 3 為使用 DeepBach 神經網絡來預測女高音的架構圖。

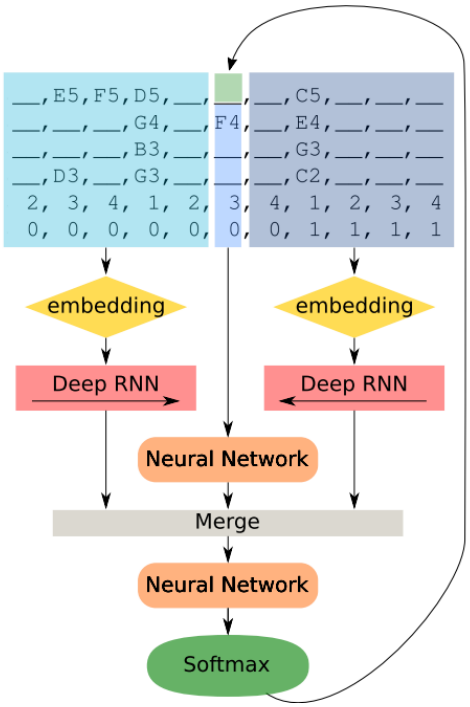

圖 3　DeepBach 預測女高音的架構圖

DeepBach 運用深度學習技術來學習音樂當中的規則，然而，一般的深度類神經網路只考慮當下的輸入資料，而在音樂的旋律是有連貫性的，因此 DeepBach 加入 LSTM 的技術，LSTM 也是由類神經網路構成，但加上了控制閘，因此具有記憶體(Mermory)的特性，可以保留先前的資訊，並控制輸入和輸出，如此便解決了連貫性的問題。

在評估部份，有 1272 名受試者，半數為音樂愛好者、四分之一為受過專業音樂或作曲訓練的音樂家。播放 400 個音樂片段(100 個取自巴哈原曲、300 個來自「深度巴哈」及另外兩種模型生成)，必須判斷聽到的片段是由巴哈或是由機器所作的。結果顯示，由「深度巴哈」生成的曲子，約有一半受試者認為是巴哈的作品。

「深度巴哈」的技術看來也能用於學習不同作曲家與風格的音樂，巴哈的風格容易掌握，因為四聲部聖詠雖然曲目眾多，但結構嚴謹、規則明確、聲部單純。如果要訓練機器進一步

模仿更複雜多元、更無固定規則的音樂風格，以目前的方法來說，尚未有研究做到。

Google Magenta [13]透過學習 MIDI 數據，使用深度遞歸神經網絡(Deep Recurrent Neural Network，DRNN)創建了鋼琴音樂，但是這種方法只能處理單軌音樂。為了解決單音軌的問題，Chu 等人[10] 提出了一種生成流行音樂的框架，使用一個分層遞歸神經網絡，層次結構和層次結構編碼我們關於流行音樂如何組成的先前知識，低層產生旋律，而較高層產生鼓和弦。Zhu 等人[14]提出一個端到端旋律和編曲生成的框架，該框架生成一個旋律音軌，其中包含多樂器演奏的伴奏音軌。設計了兩個框架，一個是基於和旋的節奏和旋律交叉生成模型(Chord based Rhythm and Melody Cross-Generation Model，CRMCG)，利用和旋進行來引導旋律；另一個是多音軌音樂編曲的多樂器聯合編曲模型(Multi-Instrument Co-Arrangement Model，MICA)，使用多任務學習來進行多音軌編曲。最後在數據集進行了大量實驗，結果證明了小冰帶有很好的效能。

Jaques 等人[15]提出 RL-Tuner 架構，目的是控制使用者約束的旋律生成，該體系結構如下圖 4 所示，包含兩個深度 Q 網絡體系結構和兩個循環網絡(RNN)體系結構。圖中 a (Action)代表下一個音符、s (State)代表已經作出的曲子、r (Reward)代表 Music Theory 及音樂上下文。Note RNN 的目標是預測和生成下一個音符在旋律數據集上進行訓練，其方式與 Eck 等人[8]中使用 RNN 生成旋律的實驗類似。使用 Reward RNN 製作固定副本並用於強化學習(基於 MIDI 數據集的給定旋律的音符對數機率來分配獎勵)。Q Network 是從當前生成的(部分)旋律中選擇下一個音符(a)。Q Network 與另一個 Q Network 並行訓練，稱為 Target Q Network，該網絡估計增益的值，並從 Note RNN 學到的內容進行初始化。

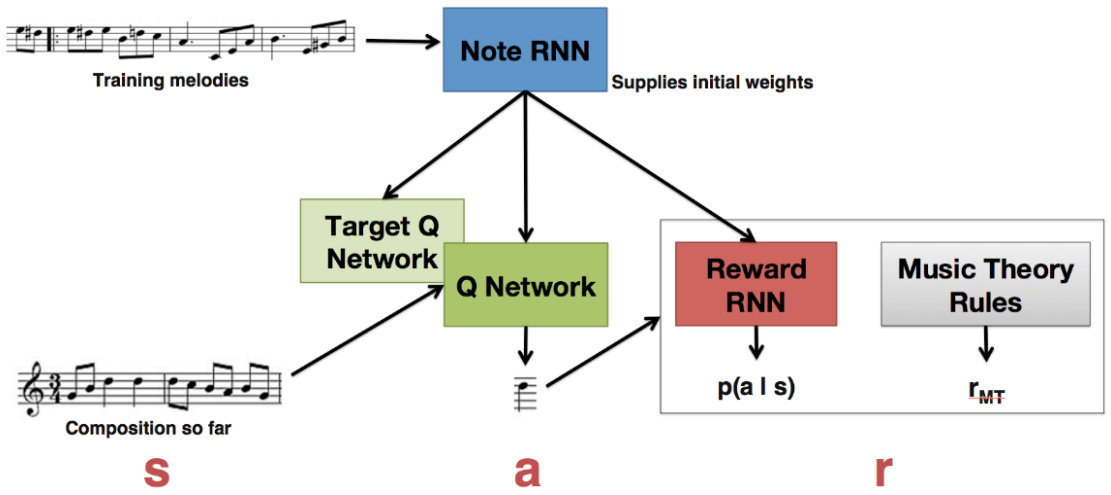

圖 4　RL-Tuner 架構

4. 生成對抗網絡(Generative Adversarial Network，GAN)

Mogren[16] 針對以往音樂計算研究中使用符號特徵(Symbolic Representation)的不足，反而是計算機更容易理解數字表達，以及 GAN 的優勢，提出了一種 LSTM/RNN 的 GAN 網絡。訓練預測數據使用作者下載的古典 MIDI 音樂，以 Tone length, Frequency, Intensity 和 Timing 作為特徵。生成網絡結構為 2 層單向 LSTM，對抗網絡為 2 層雙向 LSTM，每次生成指定長度x88 音階數據。生成音樂的評估使用韻律學的方式，根據 Polyphony(兩個音同時彈奏的頻率)、Scale consistency(標準音程的比例)、Repetitions(音符組合重複的頻率)、Tone span(整段音樂的最低最高音階差)四個方面計算。圖 5 為 C-RNN-GAN 框架，生成器(G)產生連續數據事件序列，訓練鑑別器(D)用來區分真實音樂數據和生成的數據。

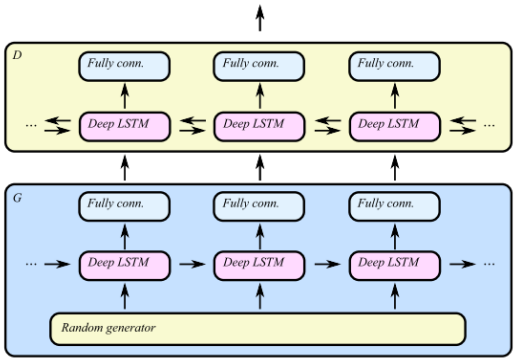

圖 5　C-RNN-GAN 框架

但這個模型存在一些缺點，單音軌的音樂生成效果比較穩定，多音軌的結果聽起來很奇怪，此外，只能生成鋼琴曲，還未支援其他樂器的音樂生成。

Doog[17]提出一個在 GAN 框架下生成多軌序列的新型生成模型(圖 6 所示)，MuseGAN 的目標是生成具有和聲和節奏結構，可以生成多軌道(樂器)的複調音樂。包含了音樂的全自動生成模塊和伴奏生成模塊，多個生成器解決互相依賴的多音軌旋律生成，使用深度 CNN 來生成多軌鋼琴音軌(piano-rolls)。

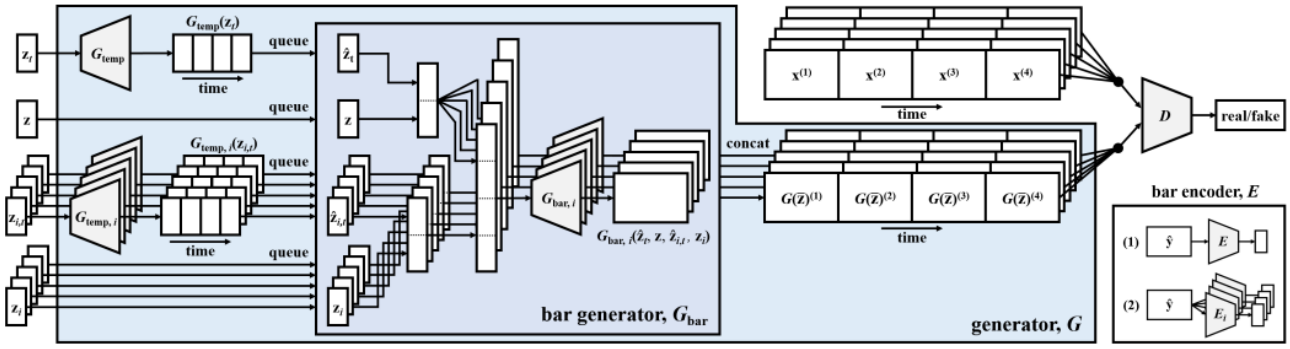

圖 6　用於多軌順序數據生成的 MuseGAN 模型的系統圖

把 MuseGAN 音樂自動生成的模型簡化，如圖 7 所示。兩組生成器，一組於協調音軌間的和弦關係，一組用於組織音軌內部的旋律生成。

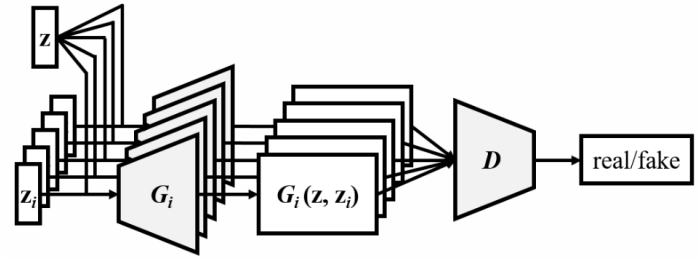

圖 7　簡化 MuseGAN 模型

此模型使用 Lakh Pianoroll Dataset(LPD)進行訓練，這是一種新的多軌鋼琴數據集，並採用無監督式方法。所提出的模型能夠從頭開始或通過伴隨用戶給出的軌道來生成音樂，可生成由低音、鼓、吉他、鋼琴和弦樂曲目組成的流行歌曲樂段。

- **侷限性**

目前已提出與音樂生成相關的不同模型，表 1 比較與音樂生成相關的模型，G 為生成、Mt 為多音軌、M 為旋律、Cp 為和旋進行、Ar 為編曲、Sa 為可歌唱性。

表 1　比較音樂生成模型

模型	*G*	*Mt*	*M*	*Cp*	*Ar*	*Sa*	*?(H)*	*?( )*
Markov music [18]	✓		✓					
Music unit selection [19]			✓					
Magenta [13]	✓		✓					
Song from PI [10]	✓	✓	✓			✓		
DeepBach [12]	✓		✓	✓				
MuseGAN [17]	✓		✓		✓			
GANMidi [20]	✓		✓					
XiaoIce Band [14]	✓	✓	✓	✓	✓	✓		
C-RNN-GAN [16]	✓	✓						

眾多的音樂生成模型均可有不錯的性能，但在音樂生成上仍有一些限制。本研究在侷限性的文獻探討部份考慮以下幾點：

1. 控制(Control)：音調一致性、重複音符的最大數量、節奏…等
2. 結構(Structure)：音樂結構呈現
3. 創造力(Creativity)：部份模型所生成的音樂多為模仿原音樂
4. 多樣性(Multi)：多音軌音樂
5. 和諧性(Harmony)：整體性的音樂

參考文獻

[1] Papadopoulos, G., & Wiggins, G. (1999, April). AI methods for algorithmic composition: A survey, a critical view and future prospects. In AISB Symposium on Musical Creativity (Vol. 124, pp. 110-117). Edinburgh, UK.

[2] Chordia, P., Sastry, A., & Şentürk, S. (2011). Predictive tabla modelling using variable-length markov and hidden markov models. Journal of New Music Research, 40(2), 105-118.

[3] Pachet, F., & Roy, P. (2011). Markov constraints: steerable generation of Markov sequences. Constraints, 16(2), 148-172.

[4] Fernández, J. D., & Vico, F. (2013). AI methods in algorithmic composition: A comprehensive survey. Journal of Artificial Intelligence Research, 48, 513-582.

[5] Whorley, R., & Conklin, D. (2015, June). Improved iterative random walk for four-part harmonization. In Mathematics and computation in music (pp. 64-70). Springer, Cham.

[6] Kaliakatsos-Papakostas, M. A., Epitropakis, M. G., & Vrahatis, M. N. (2011, April). Weighted Markov Chain model for musical composer identification. In European Conference on the Applications of Evolutionary Computation (pp. 334-343). Springer, Berlin, Heidelberg.

[7] Eigenfeldt, A., & Pasquier, P. (2012). Creative agents, curatorial agents, and human-agent interaction in coming together. Proceedings of Sound and Music Computing, Copenhagen, 181-186.

[8] Eck, D., & Schmidhuber, J. (2002). A first look at music composition using lstm recurrent neural networks. Istituto Dalle Molle Di Studi Sull Intelligenza Artificiale, 103.

[9] Mozer, M. C. (1994). Neural network music composition by prediction: Exploring the benefits of psychoacoustic constraints and multi-scale processing. Connection Science, 6(2-3), 247-280.

[10] Chu, H., Urtasun, R., & Fidler, S. (2016). Song from PI: A musically plausible network for pop music generation. arXiv preprint arXiv:1611.03477.

[11] Johnston, L. (2016). Using LSTM Recurrent Neural Networks for Music Generation.

[12] Hadjeres, G., Pachet, F., & Nielsen, F. (2016). Deepbach: a steerable model for bach chorales generation. arXiv preprint arXiv:1612.01010.

[13] Casella, P., & Paiva, A. (2001, September). Magenta: An architecture for real time automatic composition of background music. In International Workshop on Intelligent Virtual Agents (pp. 224-232). Springer, Berlin, Heidelberg.

[14] Zhu, H., Liu, Q., Yuan, N. J., Qin, C., Li, J., Zhang, K., ... & Chen, E. (2018, July). XiaoIce Band: A Melody and Arrangement Generation Framework for Pop Music. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2837-2846). ACM.

[15] Jaques, N., Gu, S., Turner, R. E., & Eck, D. (2017). Tuning recurrent neural networks with reinforcement learning.

[16] Mogren, O. (2016). C-RNN-GAN: Continuous recurrent neural networks with adversarial training. arXiv preprint arXiv:1611.09904.

[17] Dong, H. W., Hsiao, W. Y., Yang, L. C., & Yang, Y. H. (2018). MuseGAN: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In Proc. AAAI Conf. Artificial Intelligence.

[18] Van Der Merwe, A., & Schulze, W. (2011). Music generation with Markov models. IEEE MultiMedia, 18(3), 78-85.

[19] Bretan, M., Weinberg, G., & Heck, L. (2016). A Unit Selection Methodology for Music Generation Using Deep Neural Networks. arXiv preprint arXiv:1612.03789.

[20] Yang, L. C., Chou, S. Y., & Yang, Y. H. (2017). MidiNet: A convolutional generative adversarial network for symbolic-domain music generation. arXiv preprint arXiv:1703.10847.