數位音樂擷取系統之發展

林育弘
助理教授兼科主任

主要聯絡方式

康寧醫護暨管理專科學校 視訊傳播科

臺北市 114 內湖區康寧路三段 75 巷 137 號

電話: (02)2632-1181 分機 240

email: yhlin@webmail.knjc.edu.tw

摘要

Abstract

數位化音樂在現代人的娛樂生活中已經是不可或缺的媒介，但要在龐大的音樂/音訊資料庫中搜尋特定的資料是一項艱鉅的工作。過去十年來，在語音辨識領域中常用的隱式馬可夫模型及 MFCC 音訊特徵分析方法也被運用在音樂/音訊資料庫的分析，然而辨識的準確率始終無法提昇。本文將探討和回顧音樂/音訊特徵值分類的決策，以及近來備受矚目的學習理論和支撐向量機理論，探討未來可能的發展。

Digital music is now a must media for the daily entertainment. However, it is tedious to find the exact music/audio files in a huge multimedia database. In the past ten years, Hidden Markov Model and Mel-Frequency Cesptral Coefficients have been used to analyze music and audio other than just for speech with limited success. In this paper, new approaches such as Learning Theory and Support Vector Machines for the decision scheme regarding the retrieval of music/audio will be reviewed.

關鍵字：音樂/音訊特徵值、擷取、支撐向量機

Keywords：music/audio feature, retrieve, support vector machines

前言

由於數位化音樂技術的演進和硬體設計的創新，加上寬頻網路的推波助瀾之下，mp3 音樂在近年來大行其道，根據統計，幾乎 60-80%的 ISP 頻寬是被用來下載 mp3 音樂。在 2000 年時，提供下載服務的 Napster 曾擁有一百五十七萬使用者，到了 2001 年，被下載過的音樂更高達六千萬首，預估到 2008 年全球線上音樂下載的市值將達 20 億美金，許多廠商都在覬覦這塊大餅。蘋果電腦也看好這塊市場，轉向經營線上音樂，推出 mp3 撥放機 iPod 和 iTune，廣受消費大眾的歡迎，目前已成為該公司的主力產品之一，在 2004 年四月的統計，iPod 佔了數位音樂播放機市場 49%的銷售率。同時，全世界每月所發行的音樂 CD 亦達 4000 張，目前累計的音樂 CD 已達四百萬張，去年(2003 年)共賣出了七億五千萬張 CD，要在如此龐大的音樂資料庫中搜尋音樂資料是一項艱鉅的工作。

同時，數位化的潮流也衝擊到電視和電影的製作流程，以往在製作和後製(post production)階段配音時所需的音效(sound effect)和音樂，也改以 CD 資料庫或線上下載的方式提供影視製作公司購買，以便讓音效工程師做進一步的編輯和剪輯。位於美國好萊塢的 Sounddogs.com 是第一家在線上販售音效檔案的公司，自 1997 年五月開張以來，它的資料庫已經蒐集了二十萬個音訊檔案，客戶只要先註冊為會員，便可以隨時搜尋並購買所需的音效和配樂，而不用再自行錄製，可以節省相當的成本，使用該公司產品而完成的影片包括「鐵達尼號」、「征服情海」、「綠色奇蹟」等，不過目前搜尋的方式仍以文字和該公司的分類為主。

另一股視訊數位化的趨勢是在家庭中發生的，即所謂的數位娛樂，美國消費市場中的數位錄放影機(digital video recorder, DVR)是近幾年來的新產品，以佔有率最高的 TiVo 為例，它已不再使用磁帶作為儲存的媒介，而是採用特殊規格的硬碟，具有網路通訊的功能，只要接上電話線，每天便會定時下載最新的電視節表，消費者可以搜尋並選擇喜歡的節目作預約錄影。但經過一段時日，當累積了一定數量的節目內容之後，要找到某個精采片段，猶如海底撈針，因此利用音訊資料作為分析和註解視訊資料也是熱門的研究題材，由於視訊資料較複雜，目前以運動和新聞節目的分析較為成功。

音訊特徵的分析和分類(classification)在上述的系統中佔了重要的角色，另外，系統的決策架構和演算法也決定了擷取時的準確度。

技術發展背景

無論是線上音樂或音效資料庫，現行所採用的擷取(retrieve)技術都是以音樂/音訊檔案的本質(如取樣頻率和量化位元數)或文字描述(如作曲家、歌手、發行日期、錄音日期和出版商等)作為搜尋比對的依據。更上一層，能自動萃取出音訊本身特徵值的內涵擷取(content-based retrieval)技術尚在萌芽階段，最先提出完整架構並實作的是美國 Muscle Fish 公司(Wold et al, 1996)，他們認為人們辨識不同聲音的方法有下列四種:

- 比喻法—說明某一種聲音和已知的一種或一組聲音是同一類的
- 聲響/知覺特徵—用一般易懂的物理特性來分類，如音色(timbre)、亮度(brightness)、響度(loudness)和音調(pitch)
- 主觀特色—使用個人的語言來描述聲音，這需要一些專業的訓練
- 擬聲法(onomatopoeia)—產生類似音質的聲音來描述所要的聲音，例如「轟轟」來描述飛機群的聲音

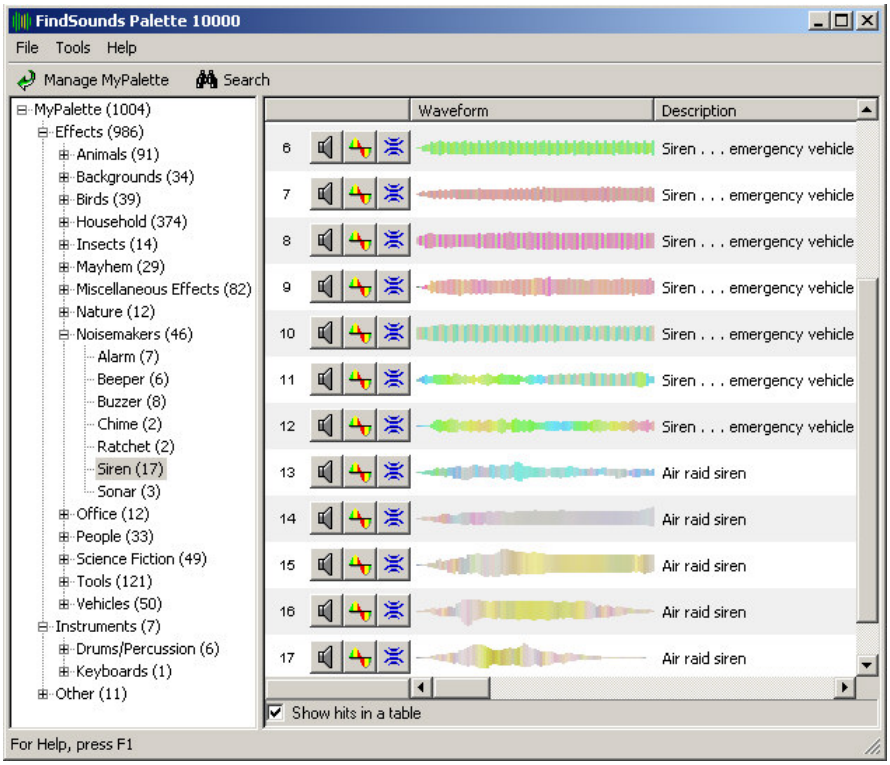

圖 1 FindSounds 的介面，利用顏色代表音訊資料中的變化

除了 Muscle Fish 之外，另一家公司 FindSounds.com，則主攻音訊資料的分割(segmentation)和網路上音訊檔案的搜尋，他們設計了一套調色盤(palette)來描述不同的聲音，如圖 1。

一般而言，設計音樂/音訊資料庫系統時，首先將音訊資料中的每一個框(a frame)的聲音作特徵分析，產生一組參數(包括聲調、長度、響度和音色)，然後在所謂的特徵空間(feature space)中使用統計的方法，將每個音訊檔案作分類。特徵的選用攸關整個音樂/音訊擷取系統的性能，除了 Muscle Fish 團隊所提出的特徵群組之外，後續許多研究(Foote, 1997;Foote, 1999;Li et al, 2001)提出其他不同的特徵群組，可以概分為知覺特徵和 MFCC(Mel-Frequency Cesptral Coefficients)。理論上，越多特徵群組會提升分類和擷取的準確度，但牽涉到系統效率表現的問題，不可能囊括所有的參數，一般常用的特徵度量如表 1 所列。

知覺特徵群組	MFCC 群組
● 音量，或稱響度和短時能量 (volume/loudness/short time energy) ● 全頻譜功率(total spectrum powers) ● 頻帶能量(band energy)或次頻帶功率 (sub band power) ● 亮度(brightness)，或稱 centroid frequency/medium frequency ● 頻帶(bandwidth) ● 調音(pitch)，或基頻(fundamental frequency) ● 零穿越率(zero-crossing rate)	● 由傅立葉轉換功率係數計算而來，然後透過含 19 個三角帶通濾波器排 (triangular bandpass filter bank)，其頻率間隔為固定值 ● Delta MFCC ● Autocorrelation MFCC ● LPC cepstral coefficient ● Delta LPC

表 1 音訊特徵值群組

當完成了特徵的分析，建構了特徵空間和向量，接下來要進行的步驟是分類，系統發展者可以事先將音訊資料分類好，或者由系統自動產生分類，然後產生出各類的特徵向量，好的分類演算法會產生可鑑別度高的群聚(cluster)邊界，未來在搜尋比對時才不會有模稜兩可的困境產生。分類的演算法可以概分為兩種方式:參數化(parametric)和非參數化。前者需要知道訊號的統計模型和模型參數的機率分佈，例如貝式分類法

(Bayesian classification)；由於後者缺乏訊號的統計模型，通常必須要有訓練組資料，作為找尋最佳決策規則的種子，有人使用類神經網路(neural networks)的方法，但成效並不佳。近年來被廣泛討論的學習理論(learning theory)和支撐向量機(support vector machine, SVM)方法，相較之下比較能找到最佳的相似度。

支撐向量機分類法

支撐向量機的基本理念是將在音訊特徵空間中非常接近的兩個群聚作明確的分離，也就是說特徵向量空間Ψ內:

$$\Psi=\{(x^m, d^m) / m=1,....M\} \quad 其中 \ x^m \in \Re^n 和 \ d^m \in \{+1,-1\}$$

x^m 為特徵向量，$(x^m, +1)$被視為"正"事件，$(x^m, -1)$則被視為"負"事件，要能分離這兩類，我們必須找到一個超平面H，由向量w和偏差b來表達並符合下列條件:

$$H(w, b)= \{ x \in \Re^n / w^T x + b = 0\}$$

為了達到可鑑別度高的分離，超平面乃依據 Lagrange 最佳化原則被放在距離兩組向量最遠的地方，而支撐超平面的向量 w 便被稱為支撐向量。當所要處理的超平面是非線性的，就必須使用核心映照函數(kernel mapping function)，把音訊特徵向量轉換到 SVM 的"特徵空間"，如此可以簡化分離時所產生的可鑑別度低的問題。一般所使用的核心映照函數有多項式(polynomial)、高斯徑向基函式(Gaussian radial base function)或指數徑向基函式(exponential radial base function)，要採用哪一種函式則視應用而定。在完成訓練組資料之後，SVM 處理新的辨識的要求時，會找出送交樣本和群聚間的最小距離。

目前在音樂資料庫中，利用多層次分類，而且在不同層次採用特定的特徵向量，SVM 在處理語音和清唱資料分離的準確度可達 99%以上(Schuller, et al, 2004)。在一般的音訊(含音樂)資料庫，研究結果顯示 SVM 和不同特徵向量的組合，比起傳統的 nearest neighbor 和 nearest center 方法，在辨識準確度方面有相當程度的提升(Guo and Li, 2003)。

結論及未來發展方向

從音訊資料庫中尋找音訊資料是非常複雜的過程，SVM 雖然指出了一個可行的方向，它的運算比較快，並且能解決一些模稜兩可的分類困境，但是仍有許多待努力的地方，例如在演算過程中的一些分類標準仍須依賴嘗試錯誤或經驗值，而且效率的好壞仍

是非常應用導向的，也就是說整個流程的細節都必須經過精密的設計，才能達到實用的水平，發展新的特徵向量是另一個可能(Davy and Godsill, 2002)，然而要把人類分辨音訊的能力轉化成機器的能力還有很長的路，畢竟電腦只懂得 0 和 1 的符號。

參考資料

(1) Wold, E., Blum, T., Keislar, D., and Wheaton, J., Content-Based Classification, Search, and Retrieval of Audio, IEEE Multimedia Mag., Vol. 3, pp.27-36(1996).
(2) Foote, J., Content-Based Retrieval of Music and Audio, SPIE(1997).
(3) Foote, J., An Overview of Audio Information Retrieval, Multimedia System, Vol. 7, pp2-11(1999).
(4) Guo, G. and Li, S. Z., Content-Based Audio Classification and Retrieval by Support Vector Machines, IEEE Transactions on Neural Networks, Vol. 14, pp.209-215(2003).
(5) Schuller, B., Rigoll, G., and Lang, M., Discrimination of Speech and Monophonic Singing in Continuous Audio Streams Applying Multi-Layer Support Vector Machines, Taipei, ICME(2004)
(6) Davy, M. and Godsill, S. J., Audio Information Retrieval: A Bibliographical Study, CUED/F-INFENG/TR.429, Cambridge University(2002)