行政院國家科學委員會專題研究計畫 成果報告

鳥類鳴唱聲音自動辨識與音樂曲風自動分類之研究
研究成果報告(精簡版)

計 畫 類 別 ：個別型
計 畫 編 號 ：NSC 98-2221-E-216-028-
執 行 期 間 ：98 年 08 月 01 日至 99 年 10 月 31 日
執 行 單 位 ：中華大學資訊工程學系

計 畫 主 持 人：李建興
共 同 主 持 人：連振昌、陳建宏
計畫參與人員：碩士班研究生-兼任助理人員：葉書峻
碩士班研究生-兼任助理人員：曾文宏
碩士班研究生-兼任助理人員：方仁政

報 告 附 件 ：出席國際會議研究心得報告及發表論文

處 理 方 式 ：本計畫可公開查詢

中 華 民 國 100 年 01 月 20 日

行政院國家科學委員會補助專題研究計畫▓ 成 果 報 告
□期中進度報告

鳥類鳴唱聲音自動辨識與音樂曲風自動分類之研究

計畫類別：☑ 個別型計畫　　☐ 整合型計畫
計畫編號：NSC 98-2221-E-216-028-
執行期間：2009 年 08 月 01 日 至 2010 年 10 月 31 日

計畫主持人：李建興
共同主持人：連振昌、陳建宏
計畫參與人員：葉書峻、方仁政、曾文宏

成果報告類型(依經費核定清單規定繳交)：☑精簡報告　☐完整報告

本成果報告包括以下應繳交之附件：
☐赴國外出差或研習心得報告一份
☐赴大陸地區出差或研習心得報告一份
☑出席國際學術會議心得報告及發表之論文各一份
☐國際合作研究計畫國外研究報告書一份

處理方式：除產學合作研究計畫、提升產業技術及人才培育研究計畫、列管計畫及下列情形者外，得立即公開查詢
☐涉及專利或其他智慧財產權，☐一年☐二年後可公開查詢

執行單位：中華大學資訊工程學系
中　華　民　國　100 年　01 月　16 日

摘要

本計劃提出以調變頻譜分析來自動分類音樂之曲風。首先將輸入之音樂聲音切割為一些固定長度之分析視窗，對於每一分析視窗分別對其八度音程頻譜對比值(Octave spectral contrast, OSC)、MPEG-7 之正規化聲音頻譜封包(NASE)及 MFCC 倒頻譜等之靜態(static)及動態(dynamic)特徵做調變頻譜分析，分別產生靜態 OSC 調變聲譜圖、動態 OSC 調變聲譜圖、靜態 NASE 調變聲譜圖、動態 NASE 調變聲譜圖、靜態 MFCC 倒頻譜調變聲譜圖及動態 MFCC 倒頻譜調變聲譜圖，然後對每一個調變頻譜分解成對數間距之調變頻帶，自每一調變頻帶中我們擷取調變頻譜能量值(modulation subband energy)、調變頻譜波谷值(modulation spectral valley)及調變頻譜對比值(modulation spectral contrast)，然後以主軸向量分析(PCA)來選取適當之調變頻譜特徵值並降低特徵向量維度，最後再以 LDA 來辨識決定每一分析視窗屬於每一特定類別聲音之相似度，以辨識此一輸入之音樂檔案是屬於何種類別之音樂曲風。

一. 報告內容

1. 前言

對於音樂曲風之分類，特徵值的擷取和分類器的選取將會影響分類的效果，其中特徵值的好壞，對於分類的結果有很大的影響。首先，我們可以將擷取之特徵向量分為短時距特徵(short-term features)和長時距特徵(long-term feature)兩類。

短時距特徵代表從一段較短時間(通常是一個音框)之音樂訊號中所擷取之特徵向量，一般而言是屬於較低階之音樂特徵。最常用來做為音樂曲風分類之音樂特徵可分為三類：音色(timbre)、節奏(rhythm)及音高(pitch)。音色特徵通常呈現了演奏之樂器或聲音來源之特性，譬如音樂、語音、及環境聲音等。通常較常使用之音色特徵有以下幾種：

(1) 低能量特徵 (low-energy feature, LEF)

此特徵之定義是將連續數個音框看成一個紋理視窗(texture window)，首先計算每個音框之能量值：

$$E_{RMS}(n) = \left(\frac{1}{M}\sum_{m=0}^{M-1}(x[n\times M+m])^2\right)^{1/2}$$

其中，M 表示每個音框的樣本數目，然後計算紋理視窗中所有音框能量的平均值：

$$\overline{E}_{RMS} = \frac{1}{N}\sum_{n=0}^{N-1}E_{RMS}(n)$$

其中，N 表示每個紋理視窗中的音框數目，此低能量特徵是計算在此紋理視窗中有

多少百分比之音框其能量值低於所有音框之能量平均值：

$$LEF = \frac{1}{N}\sum_{n=1}^{N} LEI(n),\ LEI(n) = \begin{cases} 1, & E_{\mathrm{RMS}}(n) \geq \overline{E}_{RMS} \\ 0, & otherwise \end{cases}$$

(2) 越零率 (zero-crossing rate)

計算經過振幅為零的水平線的次數，可以用來表示時間域上的頻率程度，其特徵擷取公式如下：

$$Z_t = \frac{1}{2}\sum_{n=1}^{N} |\, sign(x[n]) - sign(x[n-1]) \,|$$

其中 t 表示音框索引值，N 為音框大小，且 $sign(x[n]) = \begin{cases} 1, & x[n] \geq 0 \\ 0, & x[n] < 0 \end{cases}$

(3) 頻譜質心 (spectral centroid)

主要用來表示聲音訊號組成頻率的平均值，可用來表示聲音的嘹亮(brightness)程度。其計算公式如下：

$$C_n = \frac{\sum_{m=1}^{N/2} M_n[m] \times m}{\sum_{m=1}^{N/2} M_n[m]}$$

其中 C_n 為頻譜質心，N 為音框大小，n 為音框索引值(frame index)，M_n**[**m**]** 為第 m 個頻率係數之強度值。

(4) 頻譜頻寬 (spectral bandwidth)

用來描述頻譜圖的形狀，其值越小表示頻率分布越集中在頻譜質心，可以用其來描述頻率分布是集中在頻譜質心亦或是分散在各個頻率。

$$D_n = \sqrt{\frac{\sum_{m=1}^{N} (m - C_n)^2 \times M_n[m]}{\sum_{m=1}^{N} M_n[m]}}$$

(5) 頻譜滑動率 (spectral rolloff)

也是用來描述頻譜圖形狀的另一種用法。

$$\sum_{m=1}^{R_n} M_n[m] = TH \times \sum_{m=1}^{N} M_n[m]$$

其中 R_n 為頻譜滑動率，TH = 0.85 為此公式常用的數值

(6) 頻譜變遷度 (spectral flux)

用來描述前後連續兩個音框間的頻譜差異性，可用來分割聲音片段。

$$F_n = \sum_{m=1}^{N} (N_n[m] - N_{n-1}[m])^2$$

其中 N_n[m]為將 M_n[m]正規化的值，其計算公式如下：

$$N_n[m] = \frac{M_n[m]}{\sum_{m=1}^{N} M_n[m]}$$

(7) 梅爾倒頻譜係數 (Mel-frequency cepstral coefficients, MFCC)

梅爾倒頻譜係數已經廣泛的應用在語音辨識上，事實上，梅爾倒頻譜係數在語音辨識上是非常有用的而且可以用一組頻帶來描述一段聲音訊號。人類之聽覺系統可將聲音之頻率分為一個個臨界頻帶(critical band)，位於同一臨界頻帶內之頻率聲音對人耳聽起來是相似的，梅爾倒頻譜係數之概念即是用一組梅爾濾波器來過濾每一臨界頻帶之聲音訊號，並對每一頻帶之對數能量頻譜值(logrithmic spectra)做離散餘弦轉換(discrete cosine transform, DCT)，即可求得每一音框之梅爾倒頻譜係數。

(8) 八度音程頻譜對比值 (octave-based spectral contrast, OSC)

八度音成頻譜對比值[1]是用來描述在每一個八度音程的次頻帶中，頻譜波峰和波谷的強度值之差異值，如此可以大略的反映聲音訊號之泛音(harmonic)和非泛音(non-harminic)的分佈狀況。

節奏特徵主要是描述一首音樂之節奏特性，通常是由一段音樂中的節拍統計圖(beat histogram)中擷取其節奏特徵，包括所有節拍的強度、主節拍的速度及強度、主節拍和次節拍之速度間距，以及主節拍和次節拍的相對強度值。預估主節拍速度和其對應強度的方法可參考[2, 3]。Tzanetakis 提出從一首音樂之音高統計圖(pitch histogram)中擷取音高特徵的方法[4]，其特徵包括頻率、音高強度值和音高間距。此一音高統計圖可以使用各種音高偵測演算法來統計得到[5, 6]，而旋律與泛音也廣泛地由音樂家用來研究音樂的結構，因此，Scaringella 等人提出藉由描述每一小段音樂片段之音高分佈來擷取旋律與泛音的方法[7]，此一方法類似旋律或泛音分析器，但不用事先決定較高階之音樂特性，如基頻、和弦或音樂調性。

欲描述一整首音樂之特性，通常必需將短時距的特徵向量整合在一起而構成長時距之特徵向量，整合的方式包括計算所有短時距特徵向量之平均值及標準差，或者以自我回歸模型[8]或調變頻譜[9-11]來分析。

(1) 平均值和標準差

最常被使用來整合短時距特徵向量的方法是計算所有特徵向量之平均值和標準差，假設 $\mathbf{x}_i = [x_i[0], x_i[1], \ldots, x_i[D-1]]^T$ 表示第 i 個音框之 D 維特徵向量。所有音框之特徵向量平均值與標準差之計算公式如下：

$$\mu[d] = \frac{1}{T}\sum_{i=0}^{T-1} x_i[d], \quad 0 \le d \le D-1$$

$$\sigma[d] = [\frac{1}{T}\sum_{i=0}^{T-1} (x_i[d] - \mu[d])^2]^{1/2}, \quad 0 \le d \le D-1$$

其中，T 表示所有音框的數目。然而以所有特徵向量之平均值和標準差等統計資料來描述一整首音樂並無法顯示音樂訊號隨時間變化之特性。

(2) 自我回歸模型 (autoregressive (AR) model)：

Meng 等人以 AR 模型來分析音樂訊號隨著時間變化的特性[8]，他們提出以對角自我回歸模型(diagonal autoregressive model, DAR)與多變量自我回歸模型(multivariate autoregressive model, MAR)分析來整合短時距特徵向量。在 DAR 模型裡，將每一個短時距特徵值視為一個獨立的 AR 模型，並計算所有短時距特徵值的平均值、標準差和每一個 AR 模型的回歸係數作為長時距特徵向量。在 MAR 模型中，將短時距特徵向量以一個多變量自我回歸模型來表示。MAR 模型和 AR 模型最大的不同在於 MAR 模型考慮了特徵值間之關聯性，因此，在 MAR 模型下所擷取的長時距特徵向量則包含所有短時距特徵向量的平均值、共變異數矩陣和 MAR 模型的回歸係數。

(3) 調變頻譜分析 (modulation spectrum analysis)

調變頻譜分析是要觀察沿著時間軸上頻率的變化情形，這個方法最早是由 Kingsbury 提出用來做語音的辨識[9]，其方法顯示對人類聽覺最敏感的調變頻率大約在 4 赫茲左右。Sukittanon 也使用調變頻譜分析來辨識分析音樂之內容[10]，其實驗顯示對每一次頻帶正規化後之調變頻率特徵受到旋積雜訊干擾之影響較小。Shi 同樣使用調變頻譜分析來描述音樂訊號之長時距特性[11]，用以擷取音樂之拍子速度以對不同情感之音樂類型做分類。

Tzanetakis 及 Cook [12]在其提出之音樂曲風分類系統中，使用音色、節奏及音高等音樂特徵，並且以高斯混合模型來分類。在其分類系統中，包含以下幾種音樂曲風：古

典音樂、鄉村音樂、嘻哈音樂(hip-hop)、爵士樂、搖滾樂、藍調音樂、雷鬼音樂(reggae)、流行音樂及金屬音樂(metal)等，並且將音樂曲風之分類建構成階層式架構，因此古典音樂可再細分為聖歌(choir)、管弦樂(orchestra)、鋼琴音樂及四重奏等，而爵士樂則細分為爵士樂團(bigband)、冷峻爵士樂(cool)、融合爵士樂(fusion)、鋼琴爵士樂、爵士樂四重奏及搖擺爵士樂(swing)，其實驗結果顯示以由三個高斯分佈群數組成之高斯混合模型來分類可以得到最佳之辨識率。

West 及 Cox[13]提出以音框為辨識單元之階層式音樂曲風分類系統，在其系統中以投票法則來決定一首音樂之曲風。其分類之音樂曲風包含以下幾種：搖滾樂、古典音樂、重金屬音樂(heavy metal)、鼓聲(drum)、貝斯聲(bass)、雷鬼音樂及叢林音樂等，其使用之音樂特徵包括梅爾倒頻譜係數(MFCC)及八度音程頻譜對比值(OSC)，同時比較以下幾種分類器之辨識效能：含(或不含)決策樹之高斯分類器、由三個高斯分佈群數組成之高斯混合模型及線性區別分析演算法，其實驗結果顯示以高斯混合模型且含決策樹之高斯分類器之辨識率最佳。

Xu 等人[14]以支撐向量機(support vector machine, SVM)來區分純音樂及人聲，其所使用之音樂特徵由支撐向量機之學習演算法來決定分類參數，而且其實驗結果顯示以支撐向量機來分類比起傳統之歐基里得距離公式或隱藏馬可夫模型(HMM)之分類結果還要好。

Esmaili 等人[15]使用低階之音樂特徵(如梅爾倒頻譜係數、熵值(entropy)、頻譜質心、頻寬等)以及線性區別分析演算法來辨識音樂曲風，在其系統中將音樂分類為以下五類：搖滾樂、古典音 樂、民謠(folk)、爵士樂及流行音樂等，其分類結果之正確率可以達到 93%。

Bagci 及 Erzin [16]建構一個以音框為辨識單元之音樂曲風分類系統，在其系統中，以 Inter-Genre Similarity (IGS)模型來偵測一些無效之音框並且將其排除以提高辨識的正確率。欲判定一個音框是否無效，其系統先對每一種音樂曲風建構其 GMM 模型，然後以此 GMM 模型進行辨識，所有辨識錯誤之音框將被視為無效之音框，而辨識正確之音框則用以更新 GMM 模型之參數，此外對於那些辨識錯誤之之無效音框也以一個 GMM 模型來表示其分佈狀況。其所使用之音樂特徵包括 13 個梅爾倒頻譜係數、4 個頻譜形狀特徵(頻譜質心、頻譜滑動率、頻譜變遷度及越零率)，此外還包括以上特徵之一階及二階導函數係數，而其音樂資料庫中總共有 10 種音樂曲風：藍調音樂、古典音樂、鄉村音樂、迪斯可舞曲、嘻哈音樂、爵士樂、金屬音樂、流行音樂、雷鬼音樂及搖滾樂等，

當每一個音框長度為 30 秒且每一高斯混合模型中有 48 個高斯分佈群數時，其分類結果之正確率可以達到 88.6%。

Umapathy 等人[17]採用局部區別基底(local discirminant bases, LDB)來分析兩種不同音樂類型之差異性，並且將具有最大差異性之 LDB 節點視為特徵值。首先，對一音樂訊號以小波轉換建構其五層架構之小波轉換封包樹(wavelet packet tree)，然後自小波轉換封包樹中擷取二種新的特徵值(頻帶之能量分佈及不穩定指標)以評評估兩種不同音樂類型之 LDB 節點之差異性，其所採用之特徵值有 30 個，為自前 15 個有最大差異性之 LDB 節點之基底向量係數之能量值及變異數。其實驗結果顯示其所提出之 LDB 特徵向量結合梅爾倒頻譜係數，再加上線性區別分析演算法，在第一層區分人為或自然聲音(artificial and natural sound)時之正確率為 91%，在第二層區分樂器聲或汽車聲(instrumental and automobile sound)以及人聲或非人聲(human and nonhuman)時之正確率為 99%，在第三層區分以下幾組聲音之正確率為 95%：鼓聲、笛聲或鋼琴聲(drum, flute, and piano)，飛機聲音或直昇機聲音(aircraft and helicopter)，男性語音或女性語音(male and female speech)，動物聲音、鳥類鳴聲或昆蟲叫聲(animals, birds, and insects)。

Holzapfel和Stylianou[18]利用非負矩陣分解演算法(Nonnegative Matrix Factorization, NMF)來描述音樂之音色並做為音樂風格分類之特徵。首先，對音樂訊號劃分為一個個固定大小之視窗，再利用傅立葉轉換計算其頻譜資訊，最後使用NMF演算法將每一視窗之頻譜資訊分解成為一組頻譜基底向量，將此組頻譜基底向量視為此一視窗之特徵向量。因此，假設每一種音樂風格之所有訓練視窗之頻譜，透過NMF演算法分解可求出一序列之頻譜基底向量，然後以所有頻譜基底向量建構一GMM模型來進行分類。在其實驗時使用了兩種常用的資料庫，一種是由Tzanetakis等人所建構的資料庫，其中共分為10個類別，每個類別包含了100首長度為30秒的音樂片斷；另一種則是ISMIR2004的資料庫，其資料庫分為6個類別。實驗結果顯示當GMM模型之高斯分佈之數目設為10時可得到最好的辨識率，分別達到74.0%及83.5%。

Song 及 Zhang[19] 應用半監督式的資訊融合系統於音樂風格之分類。由於監督式的方法具有較高的辨識效果，但是需要花費大量的時間在手動標記的工作上；而非監督式的方法則省去了手動標記的動作，但是其辨識效果較差；因此，作者採用了半監督式的方法，主要之目的是為了減少手動標記的工作且提高辨識之正確率。在其系統中，使用了資訊融合的架構以整合每一個單一特徵以提高辨識率。其中，要如何決定融合的權重，以EM演算法來學習調整資訊融合之參數。在辨識上則以兩音樂片段之相似程度做

為分類之依據,作者使用正規化最小平方法(Regularized Least Square, RLS)針對資訊融合之架構來計算相似度,藉以避免過度強調單一特徵之相似度計算且能擷取各個音樂風格之特徵。其分類之音樂資料庫是使用ISMIR2004之資料庫,此系統融合了梅爾倒頻譜係數(MFCC)、波動圖樣(Fluctuation Pattern)及頻譜統計圖(Spectrum Histogram)等特徵之相似度的資訊,其辨識率可達到84.77%。

Silla等人[20]應用機器學習的方法於音樂風格之分類。其系統主要是利用集成(Ensemble)的概念來整合對於時間分解及空間分解之分類辨識結果。在時間分解的架構上,是將音樂訊號分為開始、中間及結尾之長達30秒鐘的片段,針對各個片段來擷取其音樂特徵(包括節拍、音色及音高等),其目的是希望對於訊號隨著時間的變化有更好的描述,最後的結果則是由三個音樂片段之結果採投票的方式決定。在空間分解的架構上,則將一序列的二元分類器之辨識結果加以組合,用來解決多元類別之音樂風格分類的問題。其中,採用了一對多(One Against All, OAA)及循環式(Round Robin, RR)之兩種演算法來加以達成,在最後的決定階段,OOA演算法是使用事後機率(posteriori probability)的方式決定,而RR演算法則採用投票的方式;其所使用的分類器包含:古典決策樹(classic decision tree)、*K*鄰近分類器(instance-based *K*-NN classifier)、貝氏分類器(Naïve-Bayes classifier)、多層感知類神經網路及支撐向量機(SVM)。此系統所使用的資料庫為拉丁音樂資料庫,其資料庫分為10類,共有3160首歌。當RR演算法應用於SVM分類器時,擁有最佳之辨識率為65.06%。

Cataltepe等人[21]使用MIDI檔案及音訊特徵應用於MIDI檔案之音樂風格分類。在利用MIDI檔案之辨識上,則將擷取不同長短之MIDI檔案片段轉換成字串,再使用正規化壓縮距離(Normalized Compression Distance, NCD)計算兩MIDI字串之距離做辨識。而利用音訊特徵之辨識時,則是將MIDI檔案先轉換為音訊檔案,再針對音訊檔案擷取其音色、節奏及音高等特徵,利用LDA分類器及*K*-NN分類器做辨識。此系統使用McKay及Fujinaga所提出之資料庫,而結合了MIDI檔案及音訊特徵其正確率可達到93%。

2. 研究目的與研究方法

本計畫應用調變聲譜圖於音樂曲風之自動辨識分類,調變聲譜圖主要是對八度音程頻譜對比值(Octave spectral contrast, OSC)、MPEG-7 之正規化聲音頻譜封包(NASE)及MFCC 倒頻譜等之靜態(static)及動態(dynamic)特徵做調變頻譜分析,分別產生靜態 OSC 調變聲譜圖、動態 OSC 調變聲譜圖、靜態 NASE 調變聲譜圖、動態 NASE 調變聲譜圖、

靜態 MFCC 倒頻譜調變聲譜圖及動態 MFCC 倒頻譜調變聲譜圖，然後我們將每一個調變頻譜分解成對數間距之調變頻帶，接著自每一調變頻帶中擷取三種調變特徵值：調變頻譜能量值、調變頻譜波谷值、及調變頻譜對比值。然後以主軸向量分析演算法(PCA)來選取適當之調變頻譜特徵值並降低特徵向量維度，最後再以線性區別分析演算法(LDA)來辨識決定每一分析視窗屬於每一特定類別聲音之相似度，以辨識此一輸入之音樂檔案是屬於何種類別之音樂曲風。

本計劃之音樂曲風自動分類辨識系統包含訓練階段和辨識階段兩部分，訓練階段是由三個主要模組所組成: 調變頻譜特徵擷取、主軸向量分析演算法、及線性區別分析演算法。辨識階段是由四個主要模組所組成: 調變頻譜特徵擷取、主軸向量分析轉換、線性區別分析轉換、和分類。

2.1 調變頻譜特徵擷取

首先，我們先建構聲音訊號之聲譜圖，包括 OSC 聲譜圖、NASE 聲譜圖及 MFCC 倒頻譜聲譜圖，然後對各式各樣聲譜圖以調變頻譜分析來描述其隨時間之變化趨勢以擷取辨識特徵。

2.1.1 靜態 OSC 聲譜圖

OSC 是用來描述一音樂訊號之頻譜特性[1]，首先將音樂訊號依據八度音程之觀念將其分解為 B 個(本計劃中 $B = 9$)次頻帶，每一次頻帶之頻率範圍請參考表一，然後分別計算每一次頻帶之頻譜波峰和波谷的強度值，一般而言，頻譜波峰主要反映聲音訊號之泛音(harmonic)成份，而波谷相當於非泛音(non-harminic)或雜訊成份，因此頻譜波峰值和波谷值的之差異值可以大略的反映聲音頻譜的對比分佈狀況。

對於一聲音訊號，我們先將其切割成一個個音框，然後以傅立葉轉換得到每一音框之聲音頻譜，接下來以八度音程之帶通濾波器(octave scale band-pass filter) 將一音框之聲音頻譜分解為 B 個次頻帶，然後再對每一次頻帶計算其頻譜對比特徵。假設($x_{b,1}, x_{b,2}, \ldots, x_{b,N_b}$)代表第 b 個次頻帶之強度頻譜，N_b 代表所有位於第 b 個次頻帶中之傅立葉轉換係數之數目，假設此一次頻帶之強度頻譜已經依據其強度值由大至小排序過，也就是說 $x_{b,1} \geq x_{b,2} \geq \ldots \geq x_{b,N_b}$，第 b 個次頻帶之頻譜波峰及波谷的強度值就可以下列公式來預估：

$$Peak_b = \log(\frac{1}{\alpha N_b}\sum_{i=1}^{\alpha N_b} x_{b,i}),$$

$$Valley_b = \log(\frac{1}{\alpha N_b}\sum_{i=1}^{\alpha N_b} x_{b,N_b-i+1}),$$

其中α為鄰近區之參考因子(本計劃中設$\alpha = 0.2$)，第 b 個次頻帶之頻譜對比值可定義為

$$SC_b = Peak_b - Valley_b.$$

對每一音框，我們取所有次頻帶之頻譜波谷值($Valley_b$, $1 \le b \le B_o$)及頻譜對比值(SC_b, $1 \le b \le B_o$)為此一音框之 OSC 特徵向量，然後我們將所有音框之 OSC 係數(包含所有次頻帶之頻譜波谷值及頻譜對比值)沿著時間軸串接起來構成二維之影像圖，稱為 OSC 聲譜圖。

表一. 八度音程之每一次頻帶之頻率範圍 (Sampling rate = 44.1 kHz)

Subband	Low Frequency (Hz)	High Frequency (Hz)
1	0	100
2	100	200
3	200	400
4	400	800
5	800	1600
6	1600	3200
7	3200	6400
8	6400	12800
9	12800	22050

2.1.2 靜態 NASE 聲譜圖

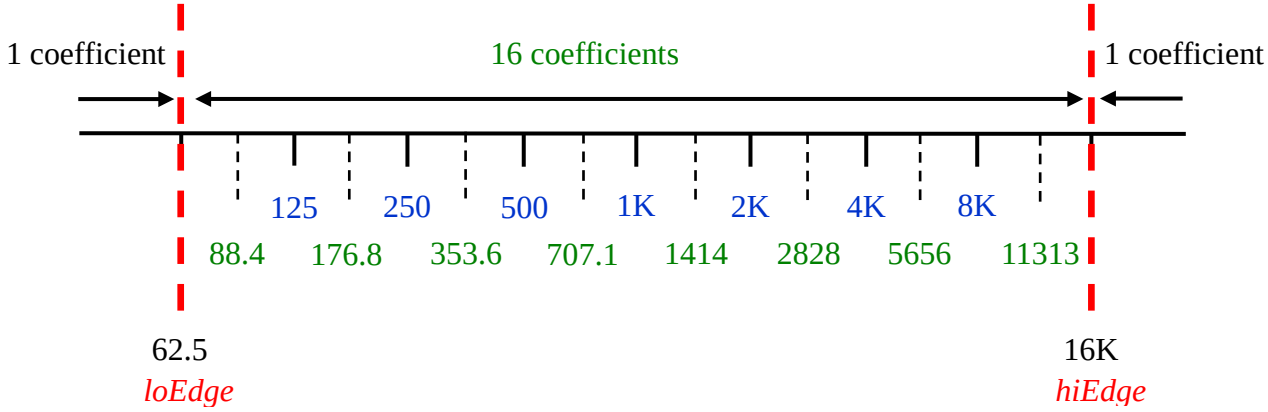
在MPEG-7標準中，是以對數之頻率間格來描述音訊訊號的頻譜圖。由於人類對於頻率的敏感度是呈現對數之對應關係，所以我們使用對數頻率來取頻率間距，如此可以兼顧描述性與簡潔性。聲音頻譜封包(audio spectrum envelope, ASE)在MPEG-7標準裡普遍用於表示原始聲音訊號中每一頻帶之功率頻譜[22, 23]，主要是描述介於*loEdge* (預設62.5Hz)與*hiEdge* (預設為16000Hz)間的頻譜資訊，將介於*loEdge*與*hiEdge*間的頻率再分解為*B*個頻帶，而每一頻帶的頻寬解析度是以八度音(octave)解析度為基準，以1000Hz為中心上下區分，總計介於[*loEdge, hiEdge*]間之頻帶數目為 $B_N = 8/r$，其中r是八度音的解析度，其範圍是介於1/16倍八度音至8倍八度音之間：

$$r = 2^j \text{ octaves}, \ -4 \le j \le 3$$

在本計劃中，我們擬採用之r值為1/2，因此$B_N = 16$，而每一頻帶之邊界頻率(f_{edge})的公式如下：

$$f_{edge} = 2^{rm} \times 1000$$

其中m是整數。此外又加上兩個額外的頻帶，一個為0Hz到*loEdge*的頻帶能量總合，一個為*hiEdge*到取樣頻率一半的頻帶能量總合，因此整個頻譜範圍可分解為 (B_N+2) 個頻帶，圖一為一個八度音解析度之邊界頻率f_{edge}分隔圖。



圖一. 八度音頻帶濾波器(頻譜解析度 r = 1/2)

NASE 在 MPEG-7 標準中是針對每一個音框之 ASE 係數轉換至分貝之刻度後做正規化之動作，然而對一段聲音訊號而言，可能包含了許多音框，所以我們將所有音框之 NASE 係數沿著時間軸串接起來構成二維之影像圖，稱為 NASE 聲譜圖。我們取所有音框之 NASE 係數值($NASE(b)$, $0 \le b \le B_N + 1$)及 RMS 值($R(b)$)沿著時間軸串接起來構成二維之影像圖，稱為 NASE 聲譜圖。

2.1.3 靜態 MFCC 倒頻譜聲譜圖

梅爾倒頻譜係數已經廣泛應用於語音辨識上[24-26]，事實上，梅爾 (mel)是用以表示人類聽覺系統對一個音調(tone)感覺上的音高或頻率的計算單位，在人類聽覺系統中，對於一個音調的實際頻率(physical frequency)之反應並不是完全呈線性變化，而實際頻率和梅爾頻率之間的對應關係在頻率低於 1 KHz 時是呈線性變化，但在高頻的部份則是呈現對數變化，而實際頻率和梅爾頻率之間的對應關係數學式如下：

$$mel = 2595 \log_{10}(1 + \frac{f}{700}),$$

$$f = 700(10^{\frac{mel}{2595}} - 1),$$

其中 f 代表實際的頻率值。人類之聽覺系統可將聲音之頻率分為一個個臨界頻帶(critical band)，位於同一臨界頻帶內之頻率聲音對人耳聽起來是相似的，因此我們可以用一組濾波器來過濾每一臨界頻帶之訊號，另外每一個臨界頻帶的頻寬會隨著頻率值而改變。

對每一分析視窗我們計算梅爾倒頻譜係數為每一音框之特徵向量，然後將所有音框之梅爾倒頻譜係數沿著時間軸串接起來構成二維之影像圖，稱為 MFCC 倒頻譜聲譜圖。

2.1.4 動態聲譜圖(dynamic spectrogram)

假設第 t 個音框之 MFCC、OSC 及 NASE 之特徵向量(稱為靜態特徵向量)分別為：

$$\mathbf{x}_t^{MFCC} = [MFCC_t(0), MFCC_t(1), \cdots, MFCC_t(L\text{-}1)]^{\mathrm{T}},$$

$$\mathbf{x}_t^{OSC} = [OSC_t(0), OSC_t(1), \cdots, OSC_t(2B_O\text{-}1)]^{\mathrm{T}},$$

$$\mathbf{x}_t^{NASE} = [R_t, NASE_t(0), NASE_t(1), \cdots, NASE_t(B_N + 1)]^{\mathrm{T}},$$

其中 L 為 MFCC 特徵向量之長度，B_O 為 OSC 之濾波器數目，B_N 為 NASE 之濾波器數目。而 MFCC、OSC 及 NASE 之動態特徵值 ΔMFCC、ΔOSC 及 ΔNASE 之定義如下：

$$\Delta MFCC_t(k) = MFCC_{t+1}(k) - MFCC_{t-1}(k),\ 0 \le k < L,$$

$$\Delta OSC_t(k) = OSC_{t+1}(k) - OSC_{t-1}(k),\ 0 \le k < 2B_O,$$

$$\Delta NSAE_t(k) = NASE_{t+1}(k) - NASE_{t-1}(k),\ 0 \le k < 2B_N,$$

我們再將動態特徵值串接起來分別構成 MFCC、OSC 及 NASE 之動態特徵向量。

2.1.5 調變頻譜分析(modulation spectral analysis)

調變頻譜特徵主要用來描述一聲音訊號中某一特定頻率(或頻帶)隨時間變化之過程。首先，對時間域之聲音訊號做傅立葉轉換後可以得到在頻率域中之短時間聲音的頻譜係數。若對一特定頻率追蹤其隨時間變化之軌跡，稱之為流動頻譜(running spectrum)，因此所有的頻率之流動頻譜則構成二維之資料，而對流動頻譜進行頻率分析之結果就稱為調變頻譜。

在本計劃中，我們將此一調變頻譜之觀念應用於不同之特徵向量，包括 OSC 係數、NASE 係數及梅爾倒頻譜係數等，計算各種靜態及動態特徵向量之調變頻譜，然後將調變頻譜切割成數個調變頻帶(如表二)，再從每一個調變頻帶內擷取特徵值，然後將所有

係數之調變頻帶特徵值沿著時間軸串接起來構成二維之影像圖，稱為調變聲譜圖，包含靜態 OSC 調變聲譜圖(modulated static OSC spectrogram, MSOSC)、動態 OSC 調變聲譜圖(modulated dynamic OSC spectrogram, MDOSC)、靜態 NASE 調變聲譜圖(modulated static NASE spectrogram, MSNASE)、動態 NASE 調變聲譜圖(modulated dynamic NASE spectrogram, MDNASE)、靜態 MFCC 倒頻譜調變聲譜圖(modulated static MFCC cepstrogram, MSMFCC)及動態 MFCC 倒頻譜調變聲譜圖(modulated dynamic MFCC cepstrogram, MDMFCC)。

表二. 調變頻帶之頻率範圍

Filter number	Modulation frequency index range	Modulation frequency interval (Hz)
0	[0, 2)	[0, 0.33)
1	[2, 4)	[0.33, 0.66)
2	[4, 8)	[0.66, 1.32)
3	[8, 16)	[1.32, 2.64)
4	[16, 32)	[2.64, 5.28)
5	[32, 64)	[5.28, 10.56)
6	[64, 128)	[10.56, 21.12)
7	[128, 256)	[21.12, 42.24]

首先，假設 $\mathbf{x}_n = [x_n(1), x_n(2), \ldots, x_n(D)]^{\mathrm{T}}$ 表示擷取自第 n 個音框之特徵向量，此特徵向量可以是第 n 個音框之 FFT 頻譜強度值、OSC 特徵向量、NASE 特徵向量或 MFCC 特徵向量，沿著時間軸對相同特徵值之連續 W 個音框做 FFT 轉換，即可得到其調變頻譜係數：

$$M_t(m,d) = \sum_{n=0}^{W-1} x_{(t\times W/2)+n}(d) e^{-j2\pi\frac{m}{W}n}, \quad 0 \le m < W, \quad 0 \le d < D,$$

其中 $M_t(m, d)$表示第 t 個分析視窗之調變頻譜，m 代表調變頻率索引值。接著我們將調變頻譜分為 J 個對數間距之調變頻帶(modulation subband)，每個調變頻帶之調變頻率分佈範圍可參考表二($J = 8$)，每個調變頻帶我們擷取三種調變特徵值：調變頻譜能量值(modulation subband energy, MSE)、調變頻譜波谷值(modulation spectral valley, MSV) 及調變頻譜對比值(modulation spectral contrast, MSC)。調變頻譜能量值之定義如下：

$$MSE(j,d) = \sum_{m=\Phi_{j,l}}^{\Phi_{j,h}} |M_t(m,d)|$$

其中 $\Phi_{j,l}$ 及 $\Phi_{j,h}$ 分別表示第 j 個調變頻帶之下界頻率索引值和上界頻率索引值。同時定義調變頻譜波谷值及調變頻譜波峰值(modulation spectral peak, MSP)如下：

$$MSP(j,d) = \max_{\Phi_{j,l} \le m < \Phi_{j,h}} |M(m,d)|$$

$$MSV(j,d) = \min_{\Phi_{j,l} \le m < \Phi_{j,h}} |M(m,d)|$$

調變頻譜波峰值與調變頻譜波谷值之差異值即是調變頻譜對比值：

$$MSC(j,d) = MSP(j,d) - MSV(j,d)$$

因此所有特徵值之所有調變頻譜對比值、調變頻譜波谷值及調變頻譜能量值可構成三個 $D \times J$ 之矩陣，每一矩陣可視為二維之影像圖，統稱為調變聲譜圖。為了降低特徵向量維度，我們對於每一調變頻譜矩陣之每一列及每一行計算平均值及標準差為特徵向量：

$$\mathbf{f}^{row} = [u_{MSC}^{row}(0), \sigma_{MSC}^{row}(0), u_{MSV}^{row}(0), \sigma_{MSV}^{row}(0), \cdots, u_{MSC}^{row}(D-1), \sigma_{MSC}^{row}(D-1), u_{MSV}^{row}(D-1), \sigma_{MSV}^{row}(D-1)]^{\mathrm{T}}$$

$$\mathbf{f}^{col} = [u_{MSC}^{col}(0), \sigma_{MSC}^{col}(0), u_{MSV}^{col}(0), \sigma_{MSV}^{col}(0), \cdots, u_{MSC}^{col}(J-1), \sigma_{MSC}^{col}(J-1), u_{MSV}^{col}(J-1), \sigma_{MSV}^{col}(J-1)]^{\mathrm{T}}.$$

然後再將每一列及每一行之特徵向量串接起來成為每一調變頻譜矩陣之特徵向量：

$$\mathbf{f} = [(\mathbf{f}^{\mathbf{row}})^{\mathrm{T}},\ (\mathbf{f}^{\mathbf{col}})^{\mathrm{T}}]^{\mathrm{T}}.$$

最後我們將每一個特徵值正規化來使得其特徵值範圍是介於 0 與 1 之間：

$$\hat{f}(n) = \frac{f(n) - f_{\min}(n)}{f_{\max}(n) - f_{\min}(n)},$$

其中，$f(n)$為第 n 個特徵值量，$\hat{f}(n)$ 為正規化後之徵值量，$f_{max}(n)$和 $f_{min}(n)$為第 n 個特徵值之最大值和最小值。

2.2 主軸向量分析演算法(principal component analysis, LDA)

PCA 是先計算所有訓練資料之特徵向量的平均變異數矩陣 $E[\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}]$之 eigenvalue 及 eigenvector [27]，並以 eigenvector 當作基底來做線性轉換，而 eigenvalue 的大小可以決定其對應之 eigenvector 轉換後之特徵所保留之資訊量大小，eigenvalue 越大表示資料作線性轉換後，特徵的變異數值會越大，而變異數的大小又表示了分佈的寬廣，資料分佈越廣表示所保留之資訊量越大，也就是說，以 eigenvalue 值較大之 eigenvector 做為線性轉換之基底，轉換後的特徵分佈範圍會比以 eigenvector 較小的 eigenvector 轉換後的分佈範圍來得大。PCA 之進行步驟如下：

步驟 1：計算平均向量

$$\mathbf{m} = E[\boldsymbol{X}]$$

其中 $\boldsymbol{X}$是所有訓練資料之集合，$X = \{\mathbf{x}_i \mid i = 0 \ldots N\}$，$\mathbf{m}$ 是所有訓練資料的平均向量，N是訓練資料的數量。

步驟 2：令平均向量為 **0**

$$\mathbf{x}_i^{'} = \mathbf{x}_i - \mathbf{m}$$

步驟 3：求取平均變異數矩陣，**C**

$$\mathbf{C} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i^{'}(\mathbf{x}_i^{'})^{\mathrm{T}}$$

步驟 4：求取變異數矩陣 **C** 的 eigenvalue 及 eigenvector 並將其依 eigenvalue 值由大至小重新排序

步驟 5：設定臨界值α_{PCA}(表示所要保留的資訊量程度)，以計算轉換後維度 *d*

$$\sum_{i=1}^{d}\lambda_i \geq \alpha_{PCA} \times \sum_{i=1}^{D}\lambda_i$$

其中λ_i表示第 *i* 大之 eigenvalue，*D* 為轉換前之維度

步驟 6：以所保留之 *d* 個 eigenvector 對所有資料作線性轉換

$$\mathbf{x}_{\mathrm{PCA}} = \mathbf{A}_{\mathrm{PCA}}^{T}\mathbf{x}_i^{'}$$

其中 $\mathbf{A}_{\mathrm{PCA}}$ 為此 *d* 個較大 eigenvector 構成之 PCA 轉換矩陣。

2.3 線性區別分析演算法(linear discriminant analysis, LDA)

LDA 演算法之目的是將一個高維度的特徵向量轉換成一個低維度的向量，並且增加辨識的準確率[27]，LDA 演算法主要處理不同類別間的區別程度而不是用於不同類別之表示方式。LDA 演算法的主要精神是要把同類之間的距離最小化，並且把不同類別之間的距離給最大化，所以，必需決定一個轉換矩陣來將高維度的特徵向量轉換成低維度向量，透過這樣的轉換我們能夠增強不同類別之間的差異性。最常使用的轉換矩陣主要依據 Fisher criterion J_F 來求得：

$$J_F(A) = tr((A^{\mathrm{T}}S_W A)^{-1}(A^{\mathrm{T}}S_B A))$$

其中，S_W 和 S_B 分別代表的是同類別之散佈矩陣(within-class scatter matrix)和不同類別之散佈矩陣(between-class scatter matrix)，而同類別之散佈矩陣的公式如下：

$$S_W = \sum_{j=1}^{C}\sum_{i=1}^{N_i}(\mathbf{x}_i^j - \boldsymbol{\mu}_j)(\mathbf{x}_i^j - \boldsymbol{\mu}_j)^{\mathrm{T}}$$

而 $\mathbf{x}_i^j$ 代表在類別 *j* 中的第 *i* 個特徵向量，$\boldsymbol{\mu_j}$ 為第 *j* 類的平均向量(mean vector)，*C* 為類別的數目，N_j 為類別 *j* 裡的特徵向量個數。而不同類別之散佈矩陣公式如下：

$$S_B = \sum_{j=1}^{C} p_j(\boldsymbol{\mu}_j - \boldsymbol{\mu})(\boldsymbol{\mu}_j - \boldsymbol{\mu})^{\mathrm{T}},$$

$\boldsymbol{\mu}$ 為所有類別的平均向量，p_j 為第 *j* 類的事先機率。LDA 演算法的目的是要去求出能夠使不同類別之散佈矩陣和同類別之散佈矩陣的比值為最大值轉換矩陣 $\mathbf{A}_{\mathrm{LDA}}$，而其維度大小為 $n \times d$：

$$\mathbf{A}_{\mathrm{LDA}} = \arg\max_{A} \frac{tr(A^{\mathrm{T}} S_B A)}{tr(A^{\mathrm{T}} S_W A)}.$$

此一轉換矩陣，可經由求出 $S_W^{-1} S_B$ 的 eigenvectors 來得到，而 $\mathbf{A}_{\mathrm{LDA}}$ 之(C-1)個行向量為前(C-1)個最大 eigenvalue 值所對應之 eigenvector。

在我們決定出最佳的轉換矩陣 $\mathbf{A}_{\mathrm{LDA}}$ 後，我們以 $\mathbf{A}_{\mathrm{LDA}}$ 將每一 n 維的特徵向量轉換為(C-1)維之向量。令 $\mathbf{x}_{\mathrm{PCA}}$ 為 PCA 轉換後之特徵向量，則 LDA 轉換後之特徵向量如下:

$$\mathbf{x}_{\mathrm{LDA}} = \mathbf{A}_{\mathrm{LDA}}^{\mathrm{T}} \mathbf{x}_{\mathrm{PCA}}.$$

2.4 分類辨識

在辨識的部份中，我們先將輸入之音樂訊號擷取其調變特徵向量，以辨識此一輸入之音樂訊號是屬於何種音樂曲風。假設輸入之音樂訊號之特徵向量為 $\mathbf{x}$，計算此一向量和每一辨識種類之聲音代表特徵向量($\mathbf{x}_k$, $1 \le k \le N$, 其中 N 為資料庫中聲音之種類數目)之間的距離，在這裡的距離公式是歐基里德距離(Euclidean distance)，最終辨認之音樂種類代表編碼 s 可由下列公式來決定：

$$s = \arg\min_{1 \le k \le N} d(\mathbf{x}, \mathbf{x}_k)$$

2.5 多個分類器整合(multiple classifiers fusion)

事實上，沒有一個特徵向量對所有的輸入之音樂類別都能得到最佳的辨識結果，也就是說，某些特徵向量對於特定的音樂類別有著較好的辨識率，但是對另一種類之音樂可能以另一種特徵向量可以得到更好的辨識率。因此，我們以資訊融合方法來整合各種特徵向量之辨識結果以提高辨識率。

假設 $\mathbf{x}_{\mathrm{MSMFCC}}$, $\mathbf{x}_{\mathrm{MSOSC}}$, $\mathbf{x}_{\mathrm{MSNASE}}$ 代表靜態之調變特徵向量，$\mathbf{x}_{\mathrm{MDMFCC}}$, $\mathbf{x}_{\mathrm{MDOSC}}$, $\mathbf{x}_{\mathrm{MDNASE}}$ 則代表動態之調變特徵向量，首先我們可以將靜態及動態之調變特徵向量分別串接起來構成靜態及動態之組合調變特徵向量 $\mathbf{x}_{\mathrm{MSCOMB}}$ 及 $\mathbf{x}_{\mathrm{MDCOMB}}$：

$$\mathbf{x}_{\mathrm{MSCOMB}} = [\mathbf{x}_{\mathrm{MSMFCC}}^{\mathrm{T}}, \mathbf{x}_{\mathrm{MSOSC}}^{\mathrm{T}}, \mathbf{x}_{\mathrm{MSNASE}}^{\mathrm{T}}]^{\mathrm{T}}.$$

$$\mathbf{x}_{\mathrm{MDCOMB}} = [\mathbf{x}_{\mathrm{MDMFCC}}^{\mathrm{T}}, \mathbf{x}_{\mathrm{MDOSC}}^{\mathrm{T}}, \mathbf{x}_{\mathrm{MDNASE}}^{\mathrm{T}}]^{\mathrm{T}}.$$

接著我們分別以 MSMFCC、MDMFCC、MSOSC、MDOSC、MSNASE、MDNASE、MSCOMB 及 MDCOMB 之特徵向量 $\mathbf{x}_{\mathrm{MSMFCC}}$、$\mathbf{x}_{\mathrm{MDMFCC}}$、$\mathbf{x}_{\mathrm{MSOSC}}$、$\mathbf{x}_{\mathrm{MDOSC}}$、$\mathbf{x}_{\mathrm{MSNASE}}$、$\mathbf{x}_{\mathrm{MDNASE}}$、$\mathbf{x}_{\mathrm{MSCOMB}}$、及 $\mathbf{x}_{\mathrm{MDCOMB}}$ 計算兩音樂檔案之特徵距離 d_{MSMFCC}、d_{MDMFCC}、d_{MSOSC}、

d_{MDOSC}、d_{MSNASE}、d_{MDNASE}、d_{MSCOMB}、及 d_{MDCOMB}，兩音樂檔案之最終特徵距離為所有個別特徵向量距離之總合：

$$d_{TOTAL} = d_{MSMFCC} + d_{MDMFCC} + d_{MSOSC} + d_{MDOSC} + d_{MSNASE} + d_{MDNASE} + d_{MSCOMB} + d_{MDCOMB}$$

3. 實驗結果與討論

在實驗中所使用之音樂資料庫為 2004 年音樂曲風分類競賽(*ISMIR2004 Music Genre Classification Contest*)所使用之音樂資料庫[28]，此資料庫中有 1458 首音樂檔案，其中有一半 729 首音樂檔案用於訓練，另外一半 729 首音樂檔案用於辨識，這些音樂檔案之取樣頻率為 44100 Hz，壓縮之位元率為 128 kbps，音訊範圍大小為 16 bits 且為立體聲之 MP3 檔案，在本實驗中，我們先將每一壓縮檔案轉換為 44100 Hz、16 bits 之單聲道音樂檔案。這些音樂檔案總共分為六種類別：古典音樂(*Classical*)、電子音樂(*Electronic*) 、爵士/藍調音樂(*Jazz/Blue*)、重金屬/龐克音樂(*Metal/Punk*)、搖滾/流行音樂(*Rock/Pop*)、及世界音樂(*World*)，總計用於訓練及辨識之古典音樂檔案分別有 320/320 首，電子音樂檔案分別有 115/114 首，爵士/藍調音樂檔案分別有 26/26 首，重金屬 /龐克音樂檔案分別有 45/45 首，搖滾/流行音樂檔案分別有 101/102 首，世界音樂檔案分別有 122/122 首。

為了與 2004 年音樂曲風分類競賽之參賽者之實驗結果比較，我們實驗中也是採用相同 50:50 之訓練檔案及辨識檔案比例，但是因為每一音樂類別之檔案數目不盡相同，因此其整體之辨識率定義如下：

$$CA = \sum_{1 \le c \le C} P_c \times CA_c,$$

其中 P_c 為第 c 種音樂類別之出現機率，CA_c 為第 c 種音樂類別之辨識率。

表格三比較各種調變特徵向量之辨識率，由此表格可以看出將所有特徵向量整合來計算距離時可以得到最佳之辨識率(87.79%)。

表三.各種調變特徵向量之辨識率

Feature Set	*CA* (%)
MSMFCC+MSOSC+MSNASE	83.81
MDMFCC+MDOSC+MDOSC	82.17
MSMFCC+MSOSC+MSNASE+MDMFCC+MDOSC+MDOSC	86.15
MSMFCC+MSOSC+MSNASE+MSCOMB	86.83
MDMFCC+MDOSC+MDOSC+MDCOMB	82.03
MSMFCC+MSOSC+MSNASE+MSCOMB+ MDMFCC+MDOSC+MDOSC+MDCOMB	87.79

表格四比較我們所提出之方法及 2004 年音樂曲風分類競賽之前五名參賽者，還有其他具備相同實驗設定之論文，由此表格中我們可以發現我們所提出之方法得到最佳之辨識率(87.79%)，比 2004 年音樂曲風分類競賽之優勝者(84.07%)還高 3.72%。

表四、 對於 2004 年音樂曲風分類競賽之音樂資料庫之辨識率比較

References	*CA*
Our proposed approach	87.79%
Our previous approach [29]	86.83%
Y. Song *et al.* [15]	84.77%
T. Lidy & A. Rauber [12]	79.70%
E. Pampalk (winner)	84.07%
K. West (2nd rank)	78.33%
G. Tzanetakis (3rd rank)	71.33%
T. Lidy & A. Rauber (4th rank)	70.37%
D. Ellis & B. Whitman (5th rank)	64.00%

二. 參考文獻

[1] D. N. Jiang, L. Lu, H. J. Zhang, J. H. Tao, and L. H. Cai, “Music type classification by spectral contrast feature”, in *Proc. of IEEE Int. Conf. on Multimedia & Expo*, Vol. 1, pp. 113-116, 2002.

[2] M. E. P. Davies and M. D. Plumbley, “Beat tracking with a two state model”, in *Proc. Int. Conf. on Acoustic, Speech, and Signal Processing (ICASSP)*, Vol. 3, pp. 241-244, 2005.

[3] W. A. Sethares, R. D. Robin, and J. C. Sethares, “Beat tracking of musical performance using low-level audio feature”, *IEEE Trans. on Speech and Audio Processing*, Vol. 13, No. 12, Mar. 2005, pp. 275-285.

[4] G. Tzanetakis, A. Ermolinskyi, and P. Cook, “Pitch Histogram in Audio and Symbolic Music Information Retrieval”, in *Proc. IRCAM*, 2002.

[5] T. Tolonen and M. Karjalainen, “A computationally efficient multipitch analysis model”, *IEEE Trans. on Speech and Audio Processing*, Vol. 8, No. 6, pp. 708-716, Nov. 2000.

[6] R. Meddis and L. O’Mard, “A unitary model of pitch perception”, *Journal of the Acoustical Society of America*, Vol. 102, No. 3, pp. 1811-1820, Sep. 1997.

[7] N. Scaringella, G. Zoia and D. Mlynek, "Automatic genre classification of music content: a survey", *IEEE Signal Processing Magazine*, Vol. 23, Issue 2, pp.133 - 141, Mar 2006.

[8] A. Meng, P. Ahrendt, J. Larsen, and L. K. Hansen, ”Temporal feature integration for music genre classification”, *IEEE Trans. on Audio, Speech and Language Processing*, Vol. 15, No. 5, pp.1654-1664, July 2007.

[9] B. Kingsbury, N. Morgan, and S. Greenberg, “Robust speech recognition using the modulation spectrogram“, *Speech Communication*, Vol. 25, No. 1, pp.117-132, 1998.

[10] S. Sukittanon, L. E. Atlas, and J. W. Pitton, “Modulation-scale analysis for content

identification", *IEEE Trans. on signal processing*, Vol. 52, No. 10, pp. 3023-3035, Oct. 2004.

[11] Y. Y. Shi, X. Zhu, H. G. Kim and K. W. Eom, "A tempo feature via modulation spectrum analysis and its application to music emotion classification", in *Proc. of 2006 IEEE International Conference on Multimedia and Expo (ICME)*, pp.1085-1088, July 2006.

[12] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals", *IEEE Trans. on Speech and Audio Processing*, Vol. 10, No. 3, pp. 293-302, July 2002.

[13] K. West and S. Cox, "Features and classifiers for the automatic classification of musical audio signals", in *Proc. of International Symposium on Music Information Retrieval* (*ISMIR*), 2004.

[14] C. Xu, N. C. Maddage, and X. Shao, "Automatic music classification and summarization", *IEEE Trans. on Speech and Audio Processing*, Vol. 13, No. 3, pp. 441-450, May 2005.

[15] S. Esmaili, S. Krishnan, and K. Raahemifar, "Content based audio classification and retrieval using joint time-frequency analysis", in *Proc. of 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 5, pp.V - 665-8, May 2004.

[16] U. Bagci, and E. Erzin, "Automatic classification of musical genres using inter-genre similarity", *IEEE Signal Processing Letters*, Vol. 14, No. 8, pp. 521-524, Aug. 2007.

[17] K. Umapathy, S. Krishnan, and R. K. Rao, "Audio signal feature extraction and classification using local discriminant bases", *IEEE Trans. on Audio, Speech and Language Processing*, Vol. 15, No. 4, pp.1236-1246, May 2007.

[18] A. Holzapfel, and Y. Stylianou, "Musical genre classification using nonnegative matrix factorization-based features", *IEEE Trans. on Audio, Speech, and Language Processing*, Vol. 16, No. 2, pp. 424-434, Feb. 2008.

[19] Y. Song, and C. Zhang, "Content-based information fusion for semi-supervised music genre classification", *IEEE Trans. on Multimedia*, Vol. 10, No. 1, pp. 145-152, Jan. 2008.

[20] C. N. Silla Jr., A. L. Koerich, and C. A. A. Kaestner, "A machine learning approach to automatic music genre classification", *J. Braz. Comp. Soc.*, Vol. 14, No. 3, pp. 7-18, Sep. 2008.

[21] Z. Cataltepe, Y. Yaslan, and A. Sonmez, "Music genre classification using MIDI and audio features", *EURASIP Journal on Applied Signal Processing* , Vol. 2007, No. 1, pp. 150-157, Jan. 2007.

[22] H. G. Kim, N. Moreau, and T. Sikora, "Audio classification based on MPEG-7 spectral basis representation", *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 14, No. 5, pp. 716-725, May 2004.

[23] H. G. Kim, N. Moreau, and T. Sikora, *MPEG-7 Audio and Beyond: audio content indexing and retrieval*, Wiley, 2005.

[24] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[25] R. Vergin, D. O'Shaughnessy, and A. Farhat, "Generalized Mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition", *IEEE Trans. on*

Speech and Audio Processing, Vol. 7, No. 5, pp. 525-532, Sep. 1999.

[26] J. W. Picone, "Signal modeling techniques in speech recognition", *Proceedings of the IEEE,* Vol. 81, pp. 1215–1247, 1993.

[27] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. New York:Wiley, 2000.

[28] http://ismir2004.ismir.net/ISMIR_Contest.html.

[29] C. H. Lee, J. L. Shih, K. M. Yu, and H. S. Lin, "Automatic Music Genre Classification Based on Modulation Spectral Analysis of Spectral and Cepstral Features", *IEEE Trans. on Multimedia*, Vol. 11, No. 4, June 2009, pp. 670-682.

三. 計畫成果自評

本計劃完成音樂曲風之自動分類系統，能夠根據音樂的性質事先將音樂曲目分類為不同的曲風類型，有效率的管理龐大的音樂資料庫，此外也可做為音樂推薦系統使用，當使用者在選取一首喜愛的音樂時，可以將曲風相似之音樂曲目推薦給使用者，減少使用者搜尋性質相似之音樂所花的時間。當初提計劃書時預計以三年期之計劃來完成鳥類鳴唱聲音自動辨識與音樂曲風自動分類等兩系統，但是計劃只通過一年期，因此我們先完成音樂曲風自動分類系統，有關鳥類鳴唱聲音自動辨識系統則預計在後續之計劃中執行完成。目前我們已發表之相關論文如下：

期刊論文 (Journal Papers)：

[1] **C. H. Lee**, J. L. Shih, K. M. Yu, and H. S. Lin, "Automatic Music Genre Classification Based on Modulation Spectral Analysis of Spectral and Cepstral Features", *IEEE Trans. on Multimedia*, Vol. 11, No. 4, June 2009, pp. 670-682. (SCI, EI)

[2] **C. H. Lee**, C. C. Han, and C. C. Chuang, "Automatic Classification of Bird Species by Their Sounds Using Two Dimensional Cepstral Coefficients", *IEEE Trans. on Audio, Speech, and Language Processing*, Vol. 16, No. 8, Nov. 2008, pp. 1541-1550. (SCI, EI)

[3] **C. H. Lee**, C. H. Chou, C. H. Han, and R. Z. Huang, "Automatic Recognition of Animal Vocalizations Using Averaged MFCC and Linear Discriminant Analysis", *Pattern Recognition Letters*, Vol. 27, Issue 2, Jan. 2006, pp. 93-101. (SCI, EI)

[4] **C. H. Lee**, Y. K. Lee and R. Z. Huang, "Automatic recognition of bird songs using cepstral coefficients", *Journal of Information Technology and Applications,* Vol. 1, No. 1, May 2006, pp. 17-23.

[5] J. L. Shih, **C. H. Lee**, and S. W. Lin, "Automatic classification of musical audio signals", *Journal of Information Technology and Applications*, Vol. 1, No. 2, Sep. 2006, pp. 95-105.

研討會論文 (Conference Papers)：

[1] **C. H. Lee**, H. S. Lin, C. H. Chou, and J. L. Shih, “Modulation Spectral Analysis of Static and Transitional Information of Cepstral and Spectral Features for Music Genre Classification”, in *Proc. of the 5th International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP’2009)*, Sep. 12-14, 2009, Kyoto, Japan, pp. 1030-1033. (EI)

[2] **C. H. Lee**, J. L. Shih, K. M. Yu, H. S. Lin, and M. H. Wei, “Fusion of Static and Transitional Information of Cepstral and Spectral Features for Music Genre Classification”, in *Proc. of the IEEE Asia-Pacific Services Computing Conference*, Dec. 9-12, 2008, Yilan, Taiwan. (EI)

[3] **C. H. Lee**, J. L. Shih, K. M. Yu and H. S. Lin, “Modulation Spectral Analysis of Audio Features for Music Genre Classification”, in *Proc. of the 21th IPPR Conference on Computer Vision, Graphics, and Image Processing*, Yilan, Aug. 24-26, 2008.

[4] C. H. Chou, **C. H. Lee** and H. W. Ni, “Bird Species Recognition by Comparing the HMMs of the Syllables”, in *Proceedings of Second International Conference on Innovative Computing, Information and Control*, Kumamoto, Japan, Sep. 5-7, 2007. (EI)

[5] **C. H. Lee**, J. L. Shih, K. M. Yu and J. M. Su, “Automatic Music Genre Classification Using Modulation Spectral Contrast Feature”, in *Proceedings of IEEE International Conference on Multimedia and Expo*, Beijing China, July 2007, pp. 204-207. (EI)

[6] **C. H. Lee**, C. C. Lien and R. Z. Huang, “Automatic Recognition of Birdsongs Using Mel-frequency Cepstral Coefficients and Vector Quantization”, in *Proceedings of International MultiConference of Engineering and Computer Scientists*, Hong Kong, 2006, pp. 331-335.

[7] **C. H. Lee**, J. L. Shih, and S. W. Lin, “A novel approach to music genre classification”, in *Proceedings of the 18th IPPR Conference on Computer Vision, Graphics, and Image Processing*, Taipei, Aug. 20-22, 2005.

行政院國家科學委員會補助國內專家學者出席國際學術會議報告

98 年 10 月 22 日

附件三

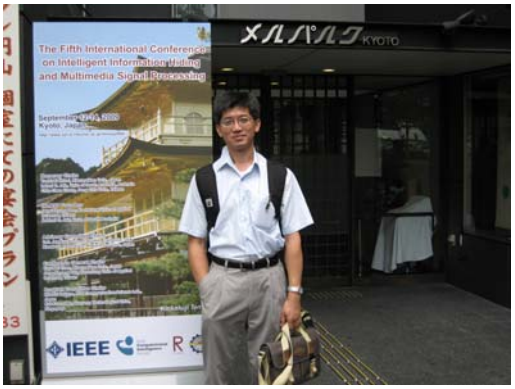
報告人姓名	李建興	服務機構及職稱	中華大學資訊工程學系 副教授
時間 會議 地點	Sep. 12-14, 2009 Kyoto Japan	本會核定 補助文號	NSC-98-2221-E-216-028-
會議名稱	(中文) (英文) The *5th International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIHMSP'2009)*		
發表論文題目	(中文) (英文) Modulation Spectral Analysis of Static and Transitional Information of Cepstral and Spectral Features for Music Genre Classification		

一、參加會議經過

本人於2009年9月11-14日赴日本京都參加 "The *5th International Conference on Intelligent Information Hiding and Multimedia Signal Processing*" 國際會議，會中發表論文一篇，如下所示：

C. H. Lee, H. S. Lin, C. H. Chou, and J. L. Shih, "Modulation Spectral Analysis of Static and Transitional Information of Cepstral and Spectral Features for Music Genre Classification", in *Proc. of the 5th International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIHMSP'2009)*, Sep. 12-14, 2009, Kyoto, Japan, pp. 1030-1033.

本人於9月11日會議前一天由台灣出發前往日本京都，我們的論文是安排在第一天9月12日上午10:30至12:30，Session A01: *Multimedia signal Processing for Intelligent Applications*，從台灣來參加研討會的學者人數相當多，幾乎與日本的學者人數相當，顯現國內在多媒體訊號分析及資料隱藏等領域之研發能量相當強勁。



二、與會心得

會議中與各國學者作深切的學術交流，而且認識許多國內相關學術領域之學者，獲益良多。

三、考察參觀活動(無是項活動者省略)

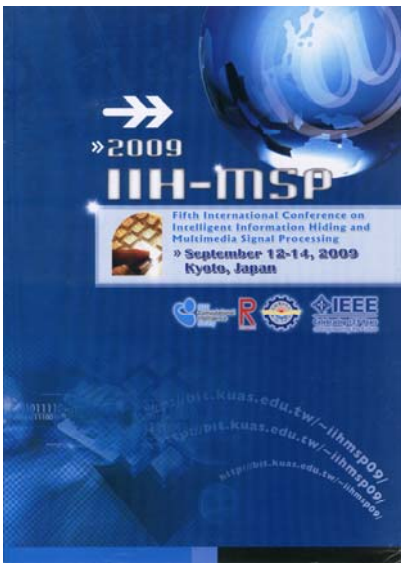
無

四、建議

建議台灣多爭取舉辦國際學術研討會，除了可以和各國學者作廣泛之學術交流，並能促進觀光產業之發展。

五、攜回資料名稱及內容

IIHMSP’2009 議程手冊及論文光碟



六、其他

非常感謝國科會之補助得以參加該研討會。

行政院國家科學委員會補助國內專家學者出席國際學術會議報告

100 年 1 月 12 日

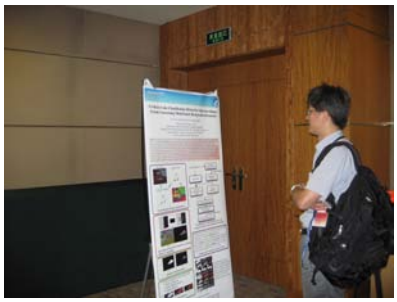
附件三

報告人姓名	李建興	服務機構及職稱	中華大學資訊工程學系 副教授
時間 會議 地點	Sep. 21-25, 2010 Shanghai, China	本會核定 補助文號	NSC 98-2221-E-216-028-
會議名稱	(中文) (英文) The 2010 Pacific-Rim Conference on Multimedia (PCM 2010)		
發表論文題目	(中文) (英文)		

一、參加會議經過

本人於 2010 年 9 月 21-25 日赴中國上海參加 "**The 2010 Pacific-Rim Conference on Multimedia (PCM 2010)**" 國際會議，研討會會場位於復旦大學校內，該研討會主要是與多媒體系統相關之專業研討會，主要分為五個 tracks: Multimedia analysis and retrieval, Multimedia security rights and management, Multimedia compression and optimization, Multimedia communication and networking, 及 Multimedia systems and applications 等。由於本人之研究領域偏重多媒體訊號分析，因此本人主要參加之 sections 是 Multimedia Analysis and Retrieval 及 Multimedia Systems and Applications 等。

本人於9月21日會議當天由台灣出發前往中國上海，由於會議期間適逢上海世博會及中秋假期，因此機票及飯店不容易預定，雖然上海市區到處都是人擠人，但是來參加研討會的學者人數卻不多。



二、與會心得

會議中與各國學者作深切的學術交流，獲益良多。

三、考察參觀活動(無是項活動者省略)

無

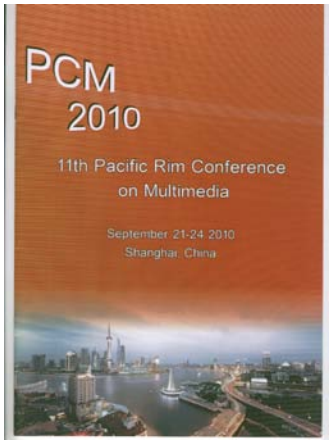
四、建議

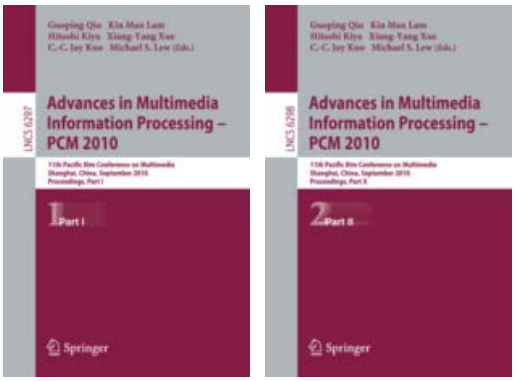
建議台灣多爭取舉辦國際學術研討會，除了可以和各國學者作廣泛之學術交流，並能促進觀

光產業之發展。

五、攜回資料名稱及內容

PCM'2010 論文手冊及論文集





六、其他

非常感謝國科會之補助得以參加該研討會。

國科會補助計畫衍生研發成果推廣資料表

日期:2011/01/19

國科會補助計畫	計畫名稱: 鳥類鳴唱聲音自動辨識與音樂曲風自動分類之研究		
	計畫主持人: 李建興		
	計畫編號: 98-2221-E-216-028- 學門領域: 自然語言處理與語音處理		
研發成果名稱	(中文) 應用靜態及動態調變頻譜分析於音樂曲風之自動分類		
	(英文) Modulation Spectral Analysis of Static and Transitional Information of Cepstral and Spectral Features for Music Genre Classification		
成果歸屬機構	中華大學	發明人(創作人)	李建興, 林懷三, 周智勳, 石昭玲
技術說明	(中文) 本技術提出以調變頻譜分析來自動分類音樂之曲風。首先將輸入之音樂聲音切割為一些固定長度之分析視窗，對於每一分析視窗分別對其八度音程頻譜對比值(OSC)、MPEG-7之正規化聲音頻譜封包(NASE)及MFCC倒頻譜等之靜態及動態特徵做調變頻譜分析，分別產生靜態OSC調變聲譜圖、動態OSC調變聲譜圖、靜態NASE調變聲譜圖、動態NASE調變聲譜圖、靜態MFCC倒頻譜調變聲譜圖及動態MFCC倒頻譜調變聲譜圖，然後對每一個調變頻譜分解成對數間距之調變頻帶，自每一調變頻帶中我們擷取調變頻譜能量值、調變頻譜波谷值及調變頻譜對比值，然後以主軸向量分析(PCA)來選取適當之調變頻譜特徵值並降低特徵向量維度，最後再以LDA來辨識決定每一分析視窗屬於每一特定類別聲音之相似度，以辨識此一輸入之音樂檔案是屬於何種類別之音樂曲風。實驗結果比較我們所提出之方法及2004年音樂曲風分類競賽之前五名參賽者，還有其他具備相同實驗設定之論文，我們所提出之方法得到最佳之辨識率(87.79%)，比2004年音樂曲風分類競賽之優勝者(84.07%)還高3.72%。		
	(英文) In this research, we will propose an automatic music genre classification approach based on long-term modulation spectral analysis on the static and dynamic information of spectral (OSC and MPEG-7 NASE) as well as cepstral (MFCC) features. First, modulation spectral analysis is performed respectively on the static and dynamic OSC spectrogram, NASE spectrogram, and MFCC spectrogram to obtain the corresponding modulation spectrograms. Second, the modulation subband energy, modulation spectral valley, and modulation spectral contrast are computed as the modulation features from each decomposed modulation subband. Principal component analysis (PCA) and linear discriminant analysis are then employed to reduce the feature dimension and improve the classification accuracy. An information fusion approach is finally employed to further improve the classification accuracy.		
產業別	出版事業；其他專業、科學及技術服務業		
技術/產品應用範圍	電子資訊產業、線上網站、音樂產業		
技術移轉可行性及預期效益	技術可移轉線上網站及音樂產業		

註：本項研發成果若尚未申請專利，請勿揭露可申請專利之主要內容。

98 年度專題研究計畫研究成果彙整表

計畫主持人：李建興	計畫編號：98-2221-E-216-028-
計畫名稱：鳥類鳴唱聲音自動辨識與音樂曲風自動分類之研究	

成果項目			量化			單位	備註（質化說明：如數個計畫共同成果、成果列為該期刊之封面故事...等）
			實際已達成數（被接受或已發表）	預期總達成數(含實際已達成數)	本計畫實際貢獻百分比		
國內	論文著作	期刊論文	0	0	100%	篇	
		研究報告/技術報告	1	1	100%		
		研討會論文	0	0	100%		
		專書	0	0	100%		
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（本國籍）	碩士生	3	3	100%	人次	
		博士生	0	0	100%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		
國外	論文著作	期刊論文	0	0	100%	篇	
		研究報告/技術報告	0	0	100%		
		研討會論文	1	1	100%		
		專書	0	0	100%	章/本	
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（外國籍）	碩士生	0	0	100%	人次	
		博士生	0	0	100%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		

其他成果 (無法以量化表達之成果如辦理學術活動、獲得獎項、重要國際合作、研究成果國際影響力及其他協助產業技術發展之具體效益事項等，請以文字敘述填列。)	無

	成果項目	量化	名稱或內容性質簡述
科教處計畫加填項目	測驗工具(含質性與量性)	0	
	課程/模組	0	
	電腦及網路系統或工具	0	
	教材	0	
	舉辦之活動/競賽	0	
	研討會/工作坊	0	
	電子報、網站	0	
	計畫成果推廣之參與（閱聽）人數	0	

國科會補助專題研究計畫成果報告自評表

請就研究內容與原計畫相符程度、達成預期目標情況、研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）、是否適合在學術期刊發表或申請專利、主要發現或其他有關價值等，作一綜合評估。

1. 請就研究內容與原計畫相符程度、達成預期目標情況作一綜合評估

■達成目標

□未達成目標（請說明，以 100 字為限）

□實驗失敗

□因故實驗中斷

□其他原因

說明：

2. 研究成果在學術期刊發表或申請專利等情形：

論文：■已發表 □未發表之文稿 □撰寫中 □無

專利：□已獲得 □申請中 ■無

技轉：□已技轉 □洽談中 ■無

其他：(以 100 字為限)

已發表於 IIHMSP'2009 研討會

3. 請依學術成就、技術創新、社會影響等方面，評估研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）(以 500 字為限）

本計劃提出以調變頻譜分析來自動分類音樂之曲風。實驗結果比較我們所提出之方法及 2004 年音樂曲風分類競賽之前五名參賽者，還有其他具備相同實驗設定之論文，我們所提出之方法得到最佳之辨識率(87.79%)，比 2004 年音樂曲風分類競賽之優勝者(84.07%)還高 3.72%。本計劃之成果可應用於各搜尋網站或數位音樂出版業。