行政院國家科學委員會專題研究計畫 成果報告

聲音訊號分類之研究：應用於鳥類鳴聲之辨識與音樂曲風之分類

研究成果報告(精簡版)

計 畫 類 別 ：個別型
計 畫 編 號 ：NSC 97-2221-E-216-037-
執 行 期 間 ：97 年 08 月 01 日至 98 年 07 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
計畫參與人員：碩士班研究生-兼任助理人員：林懷三
碩士班研究生-兼任助理人員：方仁政
碩士班研究生-兼任助理人員：李筱萱
碩士班研究生-兼任助理人員：張翔淵

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 ：本計畫涉及專利或其他智慧財產權，2 年後可公開查詢

中 華 民 國 98 年 10 月 28 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

聲音訊號分類之研究：應用於鳥類鳴聲之辨識與音樂曲風之分類

計畫類別：☑ 個別型計畫　　□ 整合型計畫
計畫編號：NSC 97-2221-E-216 -037-
執行期間：2008 年 08 月 01 日 至 2009 年 07 月 31 日

計畫主持人：李建興
共同主持人：連振昌
計畫參與人員： 林懷三、方仁政、張翔淵、李筱萱

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　98　年　10　月　30　日

摘要

隨著網際網路的普遍化，每天都有大量的多媒體音訊資料透過網路來傳送，如果我們將網際網路看成一巨大之多媒體資料庫 ，則此一多媒體資料庫可能有億萬首音訊檔案，包含大自然環境聲音、機械聲音、音樂檔案、動物叫聲、人類語音、以及其它各種聲音。對於數量如此龐大之音訊檔案，如何自動辨識音訊資料之類別，使得音訊資料之有效分類與整理成為重要的研究方向。在過去幾十年來，大部分之研究焦點集中於人類語音之自動辨識，對於其它音訊之辨識或分類相對較少，因此我們的研究計畫聚焦於鳥類鳴唱聲音之自動辨識與音樂曲風之自動分類等方向。

生態保育問題越來越受到全世界各國之重視，而評估一地區生態保育環境之好壞，一項重要之指標是調查此一地區之動植物分佈群相及其變化狀況，因此全球生物多樣性資訊機構(GBIF)在 2001 年正式成立，其目的在配合生物多樣性國際公約之要求，推動全球各國成立生物多樣性資訊交換中心，以進行生物多樣性資料之蒐集、整理與保存，使各國生物多樣性之資訊可與全球其他國家分享，促進生物多樣性之保育、利用、管理、研究及教育。國科會於 2001 年起推動建立「台灣生物多樣性國家資訊網」計畫 (Taiwan Biodiversity National Information Network, TaiBNET)，目前已完成本地生物多樣性專家名錄及台灣物種名錄兩個資料庫。此外國科會正在推動之「數位典藏國家型科技計畫」，其重點在將典藏之生物標本或文獻解說等基本資料予以數位化，農委會多年來也持續在推動全省分年分區生態之調查與時間空間分佈資料之數位化工作。從事生態調查之工作通常是由專家或具備豐富野外調查經驗之人員來執行 ，一般而言，都是依據視覺上(外形、顏色等)及聽覺上(聲紋)之特徵來辨識動物種類，但是對於某些動物，其習性隱匿不易觀察，若想見其行蹤，則是難上加難，例如八色鳥是國際鳥類聯盟公佈的全球遭受威脅鳥種之一，估計全球數量可能不到一萬隻，在台灣也十分稀少，而鳴叫聲是其互相溝通聯繫的重要工具，通常我們在野外比較容易聽到其叫聲而不易見其形體，此外生物的叫聲早已進化成與特定之物種相關(species-specific)，也就是不同之物種之聲音會有所不同，因此利用生物的鳴叫聲音來辨識生物種類是相當自然且有效可行的方法，可以幫助生態調查者確認生物之種類及其分佈定位。

在動物叫聲的辨識中，研究最多的是鳥類鳴叫聲音之辨識。目前全世界的鳥類約有 9,200 多種，臺灣已列入正式記錄的鳥類約有 456 種(台灣生物多樣性國家資訊網)，在分類學上分別隸屬於 18 目 68 科，而且台灣的特有鳥種有十九種，包括帝雉、冠羽畫眉、藍腹鷴、紫嘯鶇、烏頭翁、黃山雀、台灣藍鵲、火冠戴菊鳥、金翼白眉、栗背林鴝、紋

翼畫眉、深山竹雞、黃胸藪眉、白耳畫眉、台灣欒樹鶯、蘭嶼角鴞、白頭鶇、小翼鶇及鱗胸鷦鶇等。由於鳥類種類相當多，不同物種間的棲息環境及生活方式也都有所差異，因此研究人員投入研究鳥類叫聲的差異性，希望依此發現新的物種，然而目前所使用生物聲音的辨識方法，多採用人工至野外錄音，再回實驗室做人工的識別。因此若能利用生物的鳴叫聲來自動辨識生物種類，可以節省相當多的人力與時間，因此本計劃之一部分是對鳥類鳴叫聲音之自動辨識做一深入之研究。

由於寬頻網際網路的發展，透過網際網路來購買或下載數位音樂越來越普及，每天都有大量的數位音樂透過網路來傳送或下載，尤其是隨著蘋果電腦 i-Pod 音樂播放機在全球造成風潮，各式線上音樂網站一個個成立，每個線上音樂網站提供下載購買之音樂檔案可高達億萬首歌曲，就音樂網站管理者而言，如何有效率的管理如此龐大的音樂資料庫便是一個很重要的課題；就使用者角度而言，如何從此一巨大之音樂資料庫中搜尋感興趣之音樂資料是一項艱難之工作。因此，如果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，對於資料庫的管理有很大的幫助，對使用者也能依據音樂之曲風類型來瀏覽搜尋感興趣之音樂曲目。通常音樂曲風之分類是由具有經驗之管理者以人工的方式來分門別類，然而此種方式相當費時費力，如果能對音樂內涵加以分析，自動對每一首數位音樂之曲風類型加以分門別類，可以大大地減少搜尋之時間而且可對龐大的音樂資料庫做有效的分類與管理。因此，音樂曲風之自動分類對於音樂資料檢索系統扮演一重要之角色，此外也可做為音樂推薦(music recommendation)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。所以本研究計畫之另一重點與音樂曲風之自動分類之研究相關。

一. 報告內容

1. 前言

1.1 鳥類鳴唱聲音自動分類辨識之研究

鳥類的鳴叫聲依其功能性可大致分成至少　10 個不同的種類 [1]:一般的警戒聲(general alarm calls)、特殊的警戒聲(specialized alarm calls)、求救聲(distress calls)、遭到侵犯時的攻擊叫聲(aggressive calls)、捍衛領土的叫聲(territorial defence calls)、飛行時的叫聲(flight calls)、築巢時的叫聲(nest calls)、結伴成群的叫聲(flock calls)、餵食時的叫聲

(feeding calls)、及愉快的叫聲(pleasure calls)。這些鳥類的鳴叫聲就像簡單的語言可讓鳥類間互相溝通以及和其他種鳥類溝通，在許多情境下，鳥類的警戒聲是在警告其他的動物以脫離險境。

針對鳥類歌聲而言，其聲音結構是較複雜的，通常是將鳥類歌聲表示成一階層式之結構[2]，其中最簡單的一個鳥類聲音單元稱為音素(element)或是音調(note)，一系列連續出現且具規律模式的音素稱為音節(syllable)，而一連串的音節又組成了樂旨(motif)或是樂句(phrase)，一些重覆出現的樂旨的組合，就構成了歌聲的樂型(type)，最後，由一個或多個靜音區段隔開之樂旨則組成所謂的樂曲(bout)。

Kogan 和 Margoliash [3]比較動態時間較正(dynamic time warpping, DTW)和隱藏馬可夫模型(hidden Markov model, HMM)在辨識鳥類聲音上的效能，在使用 DTW 分類時，所採 用的特徵是對快速傳立葉轉換 (FFT) 後的振幅取對數值 ，而頻率範圍為 0.5~10KHz，在使用 HMM 做分類時，取六類不同的特徵及參數來當做 HMM 的輸入，這六個特徵分別為 linear predictive coding (LPC)、LPC cepstral coefficients (LPCC)、LPC reflection、mel-frequency cepstral coefficients (MFCC)、log mel-filter bank channel 和 linear mel-filter bank channel，其中以 MFCC 效能最好，而其選用之 MFCC 參數包括其能量值 energy)以及一次和二次導函數值。實驗結果顯示 DTW 之辨識效能還算不錯，但是對於雜訊干擾較嚴重的輸入聲音訊號或易混淆之短叫聲，使用 DTW 時還需要一些專業的背景知識來選擇恰當且具代表性之聲音訊號樣本(template)以輔助 DTW 之運作。對 HMM 而言，若想達到更好的辨識結果，就需要對聲音訊號做更好的切割(segmentation)和歸類(labeling)的動作，但 HMM 卻有一個缺點，就是對於發聲時間較短和結構複雜之鳥鳴聲常會有判斷錯誤的情形發生。

McIlraith 和 Card [4-7] 提出利用類神經網路和統計方法來分辨六種鳥鳴聲 (song sparrow、fox sparrow、marsh wren、sedge wren、yellow warbler 和 red-winged blackbird)，其擷取之特徵包括時域及頻譜上之資訊，與時域相關之資訊包含 song element 之個數，song element 長度之平均值及標準差，靜音時段之長度平均值及標準差等;而頻譜資訊包含 LPC 倒頻譜係數[8]，或將訊號分成九個次頻帶(subband)後，每個次頻帶之頻譜能量 (power spectral density) 之平均值與標準差 [5-7] ， 再利用倒傳遞類神經網路 (backpropagation neural networks)來做分類，最後的準確率為 82%，而利用二次區別分析演算法(quadratic discriminant analysis)，可以將準確率提升到 93%。

東華大學張勇富在其碩士論文中提出以語料分析為主的鳥音辨識系統[8]，其以能量

資訊來切割出鳥鳴聲中的音節(syllable)，而一個音節中會包含好幾個音框。其取每個音框的頻譜中發生振幅最大值的頻率當做基頻，最後以所有音框的基頻之中間值來辨識鳥類種類。

Anderson 等人[9]利用 DTW 來分析連續錄音中鳥類歌聲中的每一音節，他們直接比較這些聲音訊號的聲譜圖(spectrogram)，找出聲譜圖上的詞組單元(constituent)和其邊界，對聲譜圖上的振幅取對數當做特徵向量，而頻率範圍是 0.5~10KHz，他們試著用這套方法來辨識靛青鵐(indigo bunting)和錦花雀 (zebra finch)這兩種鳥類。採用的測識聲音檔是在一個低噪音的環境下收集而來的，而且用人工的方式來切出所每一鳥種具代表性之音節為其聲音訊號樣本，由實驗結果顯示，當音節變化不大時辨識結果的準確率可以達到 97%，但是，當音節結構變化大時，準確率會下降到 84%。

Harma[10]提出了一個演算法來把鳥類鳴叫聲音切割成一組音節的集合 ，所產生的每個音節以正弦波模型來表示，音節的正弦波之變化可分成振幅變化和頻率變化兩種情形，因此可以此正弦波之頻率和振幅會隨著時間變化之而軌跡來辨識不同品種的鳥類聲音，作者計算頻率軌跡和振幅軌跡之平均誤差的權重和(weighted sum)來辨識不同鳥類鳴叫聲音。其實驗對象為燕雀目的鳥類，由其實驗結果顯示，在有限的鳥類集合中所表現出來的辨識結果是不錯的，但對某些鳥種之辨識率相當低，主要原因是以頻率軌跡來表示鳥類鳴叫聲只能呈現曲調(melody)之變化而缺乏音色特徵之描述，另外對於有些鳥類所發出之喀嗒聲(clicks)或嘎嘎聲(rattles)，這類聲音就無法以正弦波模型來表示 。因此 Harma 等人提出一個方法把鳥類鳴叫聲音依其泛音結構分成四類[11]，第一類為不具泛音特性之鳥類聲音，第二類為聲音的基頻是主要部份，並有完整的泛音結構，第三類鳥類聲音在基頻部份較為微弱，反而在泛音序列裡的第一個泛音擁有最大的強度值，第四類鳥類聲音在泛音序列裡的第二個泛音擁有最大的強度值。實驗部份，作者對 150 種鳥類聲音從 2000 次的錄音檔案中切割出超過 30000 個音節來做實驗，發現大概有 60%的音節歸類為第一類，其次為第四類，大概有 14%，另外有 7%歸類為雜訊。當加入泛音結構類別來輔助辨識不同鳥類鳴叫聲音，約可提高 5-20%之正確率，但對某些鳥種其正確率之提升卻極微小，因此其認為以單一音節為辨識單元並不足夠，還必需考慮歌聲之結構(song structure)。所以他們又提出以一段鳥聲中相鄰的每對音 節所建立之直方圖(syllable pair histogram)為特徵，以此來辨識鳥類的種類[12]。建立直方圖前，先用 *k*-mean 演算法將訓練資料裡的所有音節分成 *k* 群，每一音節以其振幅軌跡為特徵且以 DTW(dynamic time warping)來計算兩個音節的差異性 。每一群則以高斯分佈模型來表

示，因此任一音節 x 和第 i 群的事後機率(posterior probability)之計算公式如下：

$$p(i \mid x) = \frac{1/(\sqrt{2\pi}\sigma_i)\exp(-d_{ix}^2/2\sigma_i^2)}{\sum_{j=1}^{k} 1/(\sqrt{2\pi}\sigma_j)\exp(-d_{jx}^2/2\sigma_j^2)}$$

其中 d_{ix} 為音節 x 和第幾 i 個群中心的 DTW 距離，而 σ_i^2 為第 i 群之變異數，其值由計算第 i 群內所有音節與其群中心的 DTW 距離平方之平均值得到。接下來對聲音中的所有成對相鄰的音節建立直方圖，假設 x_{t-1} 和 x_t 為連續兩個音節而 $\mathbf{P}_{t-1}$ 和 $\mathbf{P}_t$ 分別為其高斯模型之事後機率向量，對此兩個連續音節之二元成對機率值(bigram value)計算公式如下:

$$h_{i,j}(t) = \frac{p_{t-1,i}\, p_{t,j}}{\sum_{i',j'} p_{t-1,i'}\, p_{t,j'}}$$

對所有的 i 和 j 而言，$h_{i,j}(t)$ 可以表示為 $\mathbf{P}_{t-1}$ 和 $\mathbf{P}_t^{\mathrm{T}}$ 之乘積(product)，所以對於一段聲音的直方圖可用下列式子來表示：

$$\mathbf{H} = \sum_{t=2}^{L} \mathbf{P}_{t-1}\mathbf{P}_t^{\mathrm{T}} / \left| \mathbf{P}_{t-1}\mathbf{P}_t^{\mathrm{T}} \right|$$

其中 $\left|\mathbf{P}_{t-1}\mathbf{P}_t^{\mathrm{T}}\right|$ 為矩陣內所有元素的和。而計算兩個直方圖 $\mathbf{H}_1$ 和 $\mathbf{H}_2$ 之相似度則求其相關係數值，今 $\mathbf{h}_1$ 和 $\mathbf{h}_2$ 分別表示將 $\mathbf{H}_1$ 和 $\mathbf{H}_2$ 之每一行串接成起來之向量，如果將所有音節分為 k 群，則 $\mathbf{h}_1$ 和 $\mathbf{h}_2$ 的維度為 k^2 ，計算 $\mathbf{h}_1$ 和 $\mathbf{h}_2$ 之相關係數之公式如下：

$$c(\mathbf{h}_1, \mathbf{h}_2) = \frac{\mathbf{h}_1^{\mathrm{T}}\mathbf{h}_2}{\sqrt{\mathbf{h}_1^{\mathrm{T}}\mathbf{h}_1}\sqrt{\mathbf{h}_2^{\mathrm{T}}\mathbf{h}_2}}$$

所以 $c(\mathbf{h}_1, \mathbf{h}_2)$愈小表示 $\mathbf{h}_1$ 和 $\mathbf{h}_2$ 之間的差異性愈大。作者從 257 個鳥類歌聲檔案中切割音節，如果只有一個音節的話就將之移除，最後剩下 235 個檔案，並從這 235 個檔案中將所有音節分為 10 群、30 群和 50 群為實驗，正確率分別為 76%、79%和 80%。

此外，Fagerlund 和 Harma 提出兩種特徵參數來描述非泛音(inharmonic)結構之鳥類聲音[13]，第一種特徵參數為 10 種低階之描述參數(Low-level descriptive parameters)，其中分為頻譜特徵、時域特徵兩大類，其中頻譜特徵有頻譜質心(spectral centroid)、頻寬(signal bandwidth)、頻譜滑動頻率(spectral roll-off frequency)、頻譜變遷度(spectral flux)、頻譜平滑度 (spectral flatness)、頻譜範圍 (frequency range)，時域特徵有越零率 (zero crossing rate)、短時距能量(short time energy)、音節長度(syllable duration)和調變頻譜值(modulation spectrum)。其中頻譜質心、頻寬、頻譜滑動頻率、頻譜變遷度、頻譜平滑度、

越零率和短時距能量這 7 種是以音框為基礎的，所以真正用來辨識時，取所有音框之平均值和變異數為其真正的特徵。調變頻譜值則是對訊號先做 Hilbert 轉換，然後對訊號強度封套 (amplitude envelope)做調變 ，取調變索引值 (modulation index)及主要頻率 (dominating frequency)為特徵值。第二種特徵參數為梅爾倒頻譜係數(MFCC)，在辨識系統還利用了 LDA 來降低特徵向量之維度 。實驗部份則比較了用尤拉距離公式和墨氏(Mahalanobis)距離公式求最小距離之辨識率。當利用第一種低階之描述參數為特徵向量辨識時，利用尤拉距離公式計算最小距離之辨識率為 49%，利用墨氏距離公式的話，則辨識率升為 79%；利用第二種梅爾倒頻譜係數為特徵向量時，利用尤拉距離公式之辨識率為 73%，利用墨氏距離公式的辨識率則為 74%。

Somervuo 等人[14]更進一步針對 14 種北歐常見之燕雀目鳥種做辨識，並且提出一個新的切割音節之演算法，此一演算法是使用短時距訊號能量來把鳥類鳴叫聲切成一組音節的集合，首先是將聲音樣本取音框，每個音框大小設定為 128 個聲音樣本約 3 ms，而相鄰的音框取 50%的重疊，使用漢尼視窗(Hanning window)計算每一個音框的對數能量總和，之後將對數能量總和的最大值正規化為 0，即每個音框所得到的對數能量總和的範圍介於[0 , –∞]分貝(decibel)，接下來取全部音框的對數能量總和最小值作為背景環境噪音(noise, N_{dB})的初始值，並設定一個切割音節的臨界值：$T_{dB} = N_{dB}/2$，將大於此臨界值的音框視為存在鳥類鳴叫聲之音框，先大略切割出初步的音節，再把真正的鳥類音節切割出來。音節切割出來後，每一個音節以三種不同特徵參數模型來表示，分別為正弦波模型參數、梅爾倒頻譜係數、和一般常見的 12 種描述參數 (頻率特徵參數有頻譜質心、頻寬、頻譜滑動頻率、頻譜變遷度、頻譜平滑度、最大頻率和最小頻率；在時間域上的特徵 ：過零率 、短時距能量 、音節長度 、調變頻譜強度 (modulation spectrum magnitude)、調變頻譜頻率(modulation spectrum frequency)。首先利用動態時間校正比較此三種特徵的辨識率，就前三種特徵可觀察出 MFCC 在 12 種鳥類中有六種鳥類聲音的辨識率為最高，但就平均表現來看正弦波模型的表現為最好；之後並實驗梅爾倒頻譜係數在高斯混合模型(Gaussian mixture model, GMM)與隱藏式馬可夫模型(HMM)在不同數目的高斯函數下其辨識率的變化。

Trawicki 等人則以梅爾倒頻譜係數和 HMM 判別 Norwegian Ortolan Bunting 的鳴聲類型(song-type)[15]，其鳴聲種類可依英文字母順序分為 20 類：a，b，c，d，e，f，g，h，i，j，k，l，m，n，o，p，r，s，t 和 u。而這 20 種鳴聲還會相互組合成不同的鳴聲類型，比如 ab，cb，huf 等等，當然也有比較特別的鳥類鳴聲類型，比如 aaaabb，ccccbbb，

hhhuff，經過統計發現，ab，cb，cd，eb，f，gb，guf，h，huf 和 jufb 為 Ortolan Bunting 最常見的 10 種鳴聲類型。在其實驗部份，每種鳥類鳴聲類型各有 100 個測試檔案，只取前五種常見鳴聲類型辨識時，正確率為 92.4%，當十種常見之鳴聲類型全取時，其正確率則降為 63.6%。

由於在鳥類的鳴叫聲中其泛音有可能出現在非基頻整數倍的地方　，即非和諧音(inharmonic)的現象，所以 Selin 等人[16]使用小波分析對 8 種不同種類之鳥類鳴聲做辨識，其中有 5 種鳥類之鳴聲有非和諧音，其他 3 種鳥類之鳴聲為和諧音的鳴聲。首先針對鳥類鳴聲做前處理以去除雜訊並切割出音節，之後針對每個音節使用 Daubechies 小波轉換執行六次的小波封包分解(wavelet packet decomposition, WPD)，然後自分解後第 2-32 個頻帶之係數中擷取特徵，包括最大能量(maximum energy, E_m)、位置(position, P)、開展度(spread, S)與寬度(width, W)，並針對此四個特徵做正規化。最後使用兩種類神經網路做辨識：(1)監督式之多層認知網路(supervised multilayer perceptron, MLP)以及(2)非監督式之自我組織網路(unsupervised self-organizing map, SOM)去做訓練與測試，其使用 SOM 辨識率可達 78%，而 MLP 其辨識率則高達 96%。

Selouani 等人提出以自動回歸之時間延遲類神經網路 (Autoregressive time delay neural networks, AR-TDNN)來辨識 16 種位於加拿大新布倫茲維克省(New Brunswick)之鳥類鳴叫聲[17]。在切割音節的部分是採用 Harma[10]所提出的切割音節方法，將 16 種鳥類聲音切割成 482 個音節，其中 290 個音節用來訓練，292 個音節用來測試，其擷取的特徵是使用線性預測編碼(Linear Predictive Coding, LPC)分析，針對每一個音節使用 LPC 分析取得 20 個係數作為特徵，然後再以 AR-TDNN 來訓練測試，其辨識率可達 83%，和傳統的類神經網路比較，其辨識錯誤率降低了 16%。

1.2 音樂曲風自動分類方法之研究

一個音樂曲風分類系統所面臨的最基本問題是如何決定其分類架構，一般而言，如何明確地定義廣為眾人接受之分類架構是一艱難的問題。通常階層式的分類架構的有三點優點：1)使用者通常習慣於以瀏覽階層式目錄之方式來搜尋音樂；2)曲風之間在分類上的關聯性能夠明確的定義出來，因此階層式架構提供了一種由粗糙到精細的分類方法，可提升分類效率與正確性；3)階層式的分類法也可以清楚的知道每一階層發生分類錯誤的機率。因此，有許多人提出各種階層式音樂分類架構來解決分類之問題[18-21]。

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：低能量特徵 (low-energy feature, LEF)、越零率 (zero-crossing rate)、頻譜質心 (spectral centroid)、頻譜頻寬(spectral bandwidth)、頻譜滑動率(spectral rolloff)、頻譜變遷度(spectral flux)、梅爾倒頻譜係數(Mel-frequency cepstral coefficients, MFCC)及八度音程頻譜對比值 octave-based spectral contrast, OSC) [22]等。

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[23, 24]。

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[25]，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[26, 27]，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[28]，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[29]或調變頻譜[30-32]來分析。

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，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。Meng 等人以 AR 模型來分析音樂訊號隨著時間變化的特性[29]，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。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[30]，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[31]，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[32]，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。

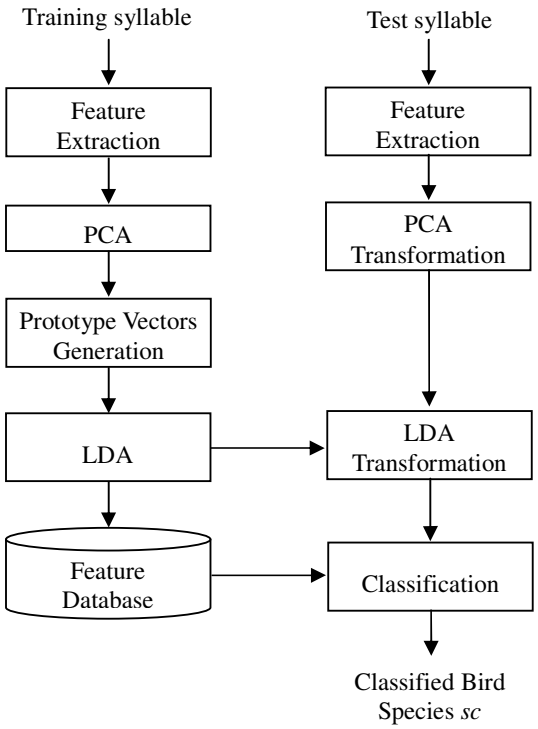
Tzanetakis 及 Cook [33]在其提出之音樂曲風分類系統中，使用音色、節奏及音高等音樂特徵，並且以高斯混合模型來分類。在其分類系統中，包含以下幾種音樂曲風：古典音樂、鄉村音樂、嘻哈音樂(hip-hop)、爵士樂、搖滾樂、藍調音樂、雷鬼音樂(reggae)、流行音樂及金屬音樂(metal)等，並且將音樂曲風之分類建構成階層式架構，因此古典音樂可再細分為聖歌(choir)、管弦樂(orchestra)、鋼琴音樂及四重奏等，而爵士樂則細分為爵士樂團(bigband)、冷峻爵士樂(cool)、融合爵士樂(fusion)、鋼琴爵士樂、爵士樂四重奏及搖擺爵士樂(swing)，其實驗結果顯示以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[34]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[35]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[36]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 [37]建構一個新的以音框為辨識單元之音樂曲風分類系統，在其系統中，一些無效之音框將先偵測出來

並且將其排除，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高斯混合模型，然後以此高斯混合模型來辨識時，那些辨識錯誤之音框將被濾除，而且辨識正確之音框將用以更新高斯混合模型之參數，此外對於那些無效之音框也以一個高斯混合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[38]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)。

2. 研究目的與研究方法

2.1 鳥類鳴唱聲音之自動分類辨識

本計畫提出之自動化鳥類鳴叫聲音辨識系統包含兩個階段，分別為訓練階段和辨識階段 ，訓練階段是由 四個主要模組所組成 :特徵擷取 、主軸分析演算法 (Principal Component Analysis, PCA)、高斯混合模型之分群演算法 、線性區別分析演算法(linear discriminant analysis, LDA)。辨識階段是由四個主要模組所組成:特徵擷取、主軸分析轉換(PCA transformation)、線性區別分析轉換(LDA transformation)和分類(classification)。圖一為本計劃之鳥類鳴唱聲音之自動分類辨識系統架構圖。



圖一、 鳥類鳴唱聲音之自動分類辨識系統架構圖。

2.1.1 特徵擷取

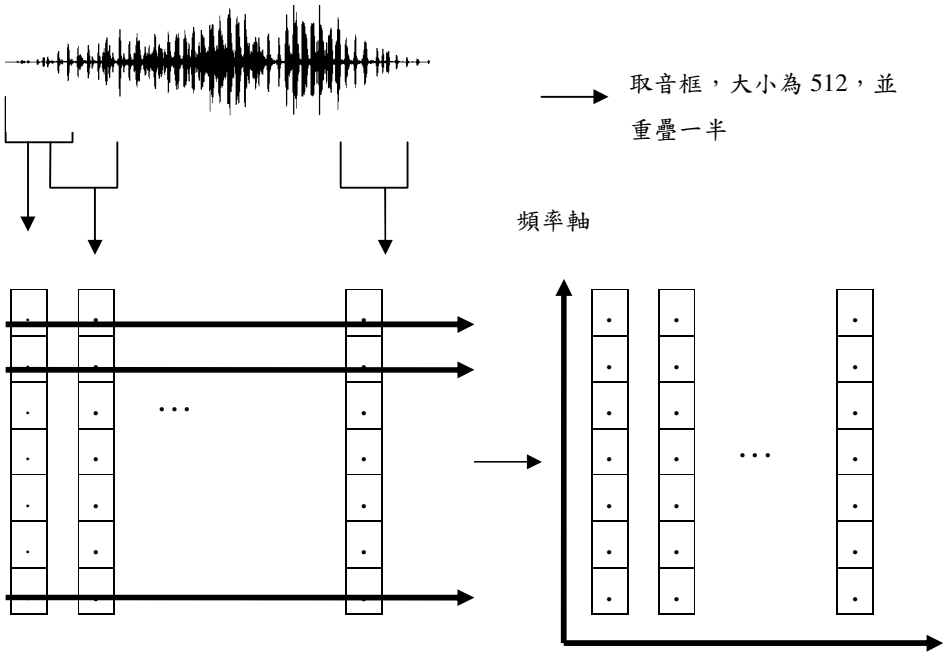
梅爾倒頻譜係數是最廣泛利用於語音辨識之特徵 [39-41]，而二維倒頻譜係數(Two-dimensional cepstrum, TDC)已被用於語音辨識上[42-44]，主要原因是二維倒頻譜係數能夠表現出倒頻譜係數隨著時間的變化，對於描述相鄰音框特徵的關聯性是一個不錯的方法，另外也能表現一個音節裡聲音頻譜圖裡之靜態和動態特性，也就是說可以表現出一個音節整體的頻率變化和細微的頻率變化；另外，二維倒頻譜係數還能夠同時解決音節長度不同的問題，因為在二維倒頻譜係數中真正有意義的是分佈於低頻的係數，所以對語音辨識真正有幫助的是分佈在低頻的係數，而分佈在高頻的係數在語音辨識上是比較沒有意義的。因此我們擬採用二維梅爾倒頻譜係數以及動態二維梅爾倒頻譜係數來表示每一個隨時間改變其特性之 鳥類鳴叫聲音 ，不只提供了梅爾倒頻譜係數的基本特性，也描述了梅爾倒頻譜係數隨著時間改變的性質。其做法是對各個音框之每一頻帶的對數能量頻譜值(logrithmic spectra)做二維離散餘弦轉換，由於二維離散餘弦轉換具有可分離特性(separability)，因此我們可以先對一音節內之每一音框計算其梅爾倒頻譜係數為此音框之特徵向量，再將這些梅爾倒頻譜係數依時間排成一矩陣之方式，針對同參數的梅爾倒頻譜係數做離散餘弦轉換，即可得到二維梅爾倒頻譜係數矩陣，其示意圖如圖

二所示。計算每一個音節之二維梅爾倒頻譜係數之步驟如下：

步驟 1: 取音框 (Framing)
將每一個聲音檔案切割成一個一個的音框，大小為 512，而且為了讓每個音框的差異性不大，我們又讓每個音框重疊一半。
步驟 2: 求出梅爾倒頻譜係數
求出梅爾倒頻譜係數 $C_i(m)$，$0 \le m \le L-1$，其中 $C_i(m)$表第 i 個音框之第 m 個梅爾倒頻譜係數，L 表梅爾倒頻譜係數的個數，在這裡 L=15。
步驟 3: 離散餘弦轉換(DCT)
令 $CC_q(m)$為對所有 $C_i(m)$沿著時間軸做離散餘弦轉換得到的二維梅爾倒頻譜係數:

$$CC_q(m) = \frac{1}{M}\sum_{i=0}^{M-1} C_i(m)\cos(2\pi iq/M),\ 0 \le q \le M-1,\ 0 \le m \le L-1,$$

其中 q 表時間軸，M 為音節音框總數，另外，在選取 $CC_q(m)$參數當作特徵時，本計劃擬取時間軸的前五個索引值，也就是二維梅爾倒頻譜係數區塊大小為 15×5。



圖二、 計算二維梅爾倒頻譜係數矩陣之流程圖

Furui 提出以動態特徵來辨識語音之方法[45]，其動態特徵是以迴歸係數(regression coefficient)來表現頻譜上的瞬間變化，應用在語者辨識中有著不錯的效果。對一段聲音切出數個音框，並對每個音框求出線性預估編碼(LPC)之後，將每個音框所求出線性預估編碼依時間排列 ，求出迴歸係數當做特徵並使用動態規畫比對演算法來辨識單詞語音，可以得到不錯的效果。令 $a_i(j)$表示在第 i 個音框之第 j 個迴歸係數，其計算方程式

如下:

$$a_i(j)=\frac{\sum_{n=1}^{n_0} n(\left|E_{i+n}(j)-E_{i-n}(j)\right|)}{\sum_{n=-n_0}^{n_0} n^2},$$

$E_i(j)$表示在第 i 個音框之第 j 個線性預估編碼。在動態二維梅爾倒頻譜係數中，我們利用迴歸係數求出在頻譜上的瞬間變化，而頻譜上的瞬間變化就像是在一張圖片中的邊線部份，也就是說如果把每一種類之鳥類鳴聲當成是一張特定的圖片，而這些圖片各自擁有獨特的邊線部份，這樣我們便能利用邊線部份進行辨識，所以我們便能利用迴歸係數來表示梅爾倒頻譜係數隨著時間變化之特性。動態二維梅爾倒頻譜係數的做法是利用迴歸係數來當做一個高通濾波器求出頻譜中變化較大的部份，也就是說，對三角帶通濾波器之輸出值計算其迴歸係數，再去做二維離散餘弦轉換後便求得動態二維梅爾倒頻譜係數。計算動態二維梅爾倒頻譜係數之詳細步驟如下:

步驟 1: 預強調 (Pre-emphasis)

$$\hat{s}[n]=s[n]-\hat{a}s[n-1],$$

$s[n]$ 為我們輸入訊號，$\hat{a}$ 的預設值為 0.95。

步驟 2: 取音框 (Framing)

將每一個音節切割成一個一個的音框，大小為 512，而且為了讓每個音框的差異性不大，我們又讓每個音框重疊一半。

步驟 3: 傅立葉轉換(DFT)

$$X_q[k]=\sum_{n=0}^{N-1} x_q[n]w[n]e^{-j2\pi\frac{k}{N}n},\ 0\le k<N$$

其中 N 為音框大小，令 $x_q[n]$表示第 q 個音框之第 n 個訊號值，$X_q[k]$為第 q 個音框之第 k 個傅立葉係數，$w[n]$為漢明視窗(Hamming window)之第 n 個係數值：

$$w[n]=0.54-0.46\cos(\frac{2\pi n}{N-1}),\ 0\le n<N$$

步驟 4: 三角帶通濾波器(Triangular band-pass filter)

以三角帶通濾波器將聲音訊號分成一個個頻帶，並算出每個頻帶的能量:

$$E_j=\sum_{k=0}^{K-1}\phi_j(k)A_k\ ,\ 0\le j\le J\ 。$$

步驟 5: 計算迴歸係數

令 $E_i(j)$表示第 i 個音框之第 j 個三角帶通濾波器輸出值，將所有音框之三角帶通濾波器之輸出值依時間順序排列，計算其迴歸係數 $a_i(j)$:

$$a_i(j)=\frac{\sum_{n=1}^{n_0} n(\left|E_{i+n}(j)-E_{i-n}(j)\right|)}{\sum_{n=-n_0}^{n_0} n^2},0\le j\le J\ 。$$

步驟 6: 離散餘弦轉換(Discrete Cosine Transform)

對這些迴歸係數乘上不同的餘弦值，求出動態梅爾倒頻譜係數 $C_i'(m)$:

$$C_i'(m) = \sum_{j=0}^{J-1} \cos(m\frac{\pi}{J}(j+0.5))\log_{10}(a_i(j)), \quad 0 \le m \le L-1$$

步驟 7: 對同索引值係數沿時間軸做離散餘弦轉換

令 $CC_q'(m)$ 為對所有 $C_i'(m)$ 沿著時間軸做離散餘弦轉換得到的動態二維梅爾倒頻譜係數，式子如下:

$$CC_q'(m) = \frac{1}{M-2}\sum_{i=1}^{M-2} C_i'(m)\cos(2\pi q / M),$$

其中 q 表時間軸，$1 \le q \le M-2$ ，M 為音節音框總數。另外，在選取 $C_i'(m)$ 參數當作特徵時，本計劃只要取時間軸的前五個索引值，也就是動態二維梅爾倒頻譜係數區塊大小為 15×5。

對於二維梅爾倒頻譜係數或動態二維梅爾倒頻譜係數也有特徵值範圍大小不同之問題，所以我們利用正規化來解決這個問題，令 $F(n)$為由二維梅爾倒頻譜係數或者是動態二維梅爾倒頻譜係數組成之特徵向量，其正規化計算公式如下:

$$\hat{F}(n) = \frac{F(n) - F_{\min}(n)}{F_{\max}(n) - F_{\min}(n)},$$

其中，$\hat{F}(n)$ 為正規化後之特徵向量，$F_{\max}(n)$和 $F_{\min}(n)$ 為第 n 個特徵值之最大值和最小值。

2.1.2 主軸分析演算法(Principal Component Analysis, PCA)

主軸分析演算法 [46]之主要目的是降低特徵向量之維度，但是降低特徵向量之維度會損失部分資訊，所以我們要如何在降低維度後還能保持最大之資訊量，因而不影響辨識之結果，甚至是刪除那些降低辨識率的特徵，而使得辨識率上升，這個問題是 PCA 所要解決的主要課題。

PCA 是先計算所有訓練資料之特徵向量的平均變異數矩陣之 eigenvalue 及 eigenvector，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 T(表示所要保留的資訊量程度)，以計算轉換後維度 d

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

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

$$\mathbf{y} = \mathbf{E}^{\mathrm{T}}\mathbf{x}_i^{'}$$

其中 **E** 為此 d 個較大 eigenvector 構成之轉換矩陣。

2.1.3 高斯混合模型之分群演算法

由於鳥類鳴叫聲音相當豐富多變化，因此就算有兩個音節是從同一種鳥類聲音中所切割出來的，所擷取出來的特徵向量也可能會有明顯的不同 ，所以對於每一種鳥類聲音，我們將使用高斯混合模型(Gaussian mixture model, GMM)來描述，因此屬於同一種鳥類聲音之不同音節可以分成幾個小群(高斯分佈)，而屬於同一小群之不同音節其特徵向量會較相似。在一般聲音辨識系統中，特徵向量維度太大時往往會降低辨識率，其原因在於有許多特徵之重要性並不高，當這些不重要的特徵過多時往往無法顯示目標聲音之獨特性，所以特徵之選擇就顯得非常重要，在本計劃中我們將利用貝氏訊息準則作為判定特徵重要性之依據。

2.1.3.1 高斯混合模型

傳統上，使用高斯混合模型於分類辨識時，對於不同類別之資料需分別建立其高斯混合模型，一般傳統上高斯混合模型之參數預測是使用 EM (Expectationn Maximization) 演算法，此演算法主要是用在預測高斯混合模型中多變數機率分佈函數之參數值，其目的是要找到最佳之參數 Θ 使得 $p(\mathbf{X}|\Theta)$最大，其中 $\mathbf{X} = \{\mathbf{x}(t), t = 1, 2, \ldots, N\}$為訓練資料之集合，$N$ 為訓練資料之數目；$\Theta \equiv \{ p(\theta_r), \boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r \mid r = 1, 2, \ldots, M\}$，$p(\theta_r)$為在高斯混合模型

中第 r 個高斯分佈之事前機率(prior probability)，$\boldsymbol{\mu}_r$ 為平均值向量，$\boldsymbol{\Sigma}_r$ 為共變異數矩陣(covariance matrix)，M 為高斯混合模型中高斯分佈之群數。EM 演算法之詳細步驟如下：

步驟 1: 執行 *k*-means 演算法

首先依據高斯混合模型中所指定之高斯分佈群數執行 *k*-means 演算法分群，以每群之平均值向量作為每個高斯分佈之平均值向量之初始值，且將共變異數矩陣之初始值設為單位矩陣。

步驟 2: Expectation-step

計算所有資料屬於高斯混合模型中每一高斯分佈之機率比值作為預測值：

$$p(\theta_r \mid \mathbf{x}(t)) = \frac{p(\theta_r)\, p(\mathbf{x}(t) \mid \theta_r)}{\sum_{r=1}^{M} p(\theta_r)\, p(\mathbf{x}(t) \mid \theta_r)}$$

其中

$$p(\mathbf{x}(t) \mid \theta_r) = \frac{1}{\sqrt{(2\pi)\boldsymbol{\Sigma}_r}} \exp\left(-\frac{(\mathbf{x}(t) - \boldsymbol{\mu}_r)^T \boldsymbol{\Sigma}_r^{-1} (\mathbf{x}(t) - \boldsymbol{\mu}_r)}{2} \right)$$

步驟 3: Maximization-step

利用**步驟 2** 所計算之預測值，更新預測之參數值：

權重值：
$$p(\overline{\theta}_r) = \frac{1}{N} \sum_{t=1}^{N} p(\theta_r \mid \mathbf{x}(t))$$

平均值向量：
$$\overline{\boldsymbol{\mu}}_r = \frac{\sum_{t=1}^{N} p(\theta_r \mid \mathbf{x}(t))\mathbf{x}(t)}{\sum_{t=1}^{N} p(\theta_r \mid \mathbf{x}(t))}$$

共變異數矩陣：
$$\overline{\boldsymbol{\Sigma}}_r = \frac{\sum_{t=1}^{N} p(\theta_r \mid \mathbf{x}(t))(\mathbf{x}(t) - \overline{\boldsymbol{\mu}}_r)(\mathbf{x}(t) - \overline{\boldsymbol{\mu}}_r)^T}{\sum_{t=1}^{N} p(\theta_r \mid \mathbf{x}(t))}$$

步驟 4: 重覆執行**步驟 2～3**，直到收斂為止。

2.1.3.2 自動決定高斯混合模型之高斯分佈群數

通常每一 GMM 模型中高斯分佈之群數是固定的，此做法其實並不適當，因為對不同類別之資料分佈皆不相同，若是使用相同群數之高斯分佈來建立各自的 GMM 模型，可能造成有些類別之資料使用太少之高斯分佈群數來描述，無法呈現資料之真正分佈狀況，另外有些類別之資料卻使用太多之高斯分佈群數來描述，造成過度調適(overfitting)之現象，使得辨識效果變差，所以較好的方式是依據不同類別之資料分佈狀況來調整其 GMM 模型中高斯分佈之群數。

Cheng 等人提出基於自我分裂法則之高斯混合模型學習演算法 (Self-Splitting Gaussian Mixture Learning, SGML)以自動決定 GMM 模型中高斯分佈之群數[47]，其中高

斯分佈之群數是由演算法中自我分裂法則之結束與否來決定，主要是以貝氏訊息準則為依據。然而在傳統 GMM 模型中當訓練資料數目過於太少時，其 GMM 模型中高斯分佈之參數會變得不可靠，這是因為 GMM 模型為一統計學上之模型，需要大量的訓練資料才能預測出可靠的高斯分佈之參數值；當訓練資料非常少之狀況下，使用簡單的向量量化法(vector quantization, VQ)其產生之辨識效果比傳統 GMM 模型要來的好，但我們無法事前去預期訓練資料之分佈情形，故 Nishida 及 Kawahara[48]提出以貝氏訊息準則為依據之模型選擇方法，討論如何根據訓練資料且固定模型架構下選擇一最佳化之分佈模型(GMM 模型或向量量化法)，使得辨識之效果最佳。

A. 貝氏訊息準則(Bayesian Information Criteria, BIC)

貝氏訊息準則是一個模型選擇之統計標準，以避免模型產生過度調適之方法。貝氏訊息準則是計算以一模型來表示一組資料時之最大相似度(maximum likelihood)，再減去模型複雜度作為懲罰(penalty)之相似度準則(likelihood criterion)，使得模型被控制在一個範圍下不會太過於複雜，其計算公式之定義如下：

$$BIC(\Theta, \mathbf{X}) \equiv \log p(\mathbf{X} \mid p(\Theta), \Theta) - \alpha \frac{1}{2} d(\Theta) \cdot \log N ,$$

其中 $\mathbf{X}$ 表示所有訓練資料之集合，$\mathbf{X} = \{\mathbf{x}(t) ; t = 1, 2, \ldots, N\}$；$N$ 表示資料數目；Θ 表示 GMM 模型中所有高斯分佈之參數，包含平均值向量與共變異數矩陣；$p(\Theta)$表示此一 GMM 模型之權重值；α 為懲罰權重(penalty weight)；$d(\Theta)$表示此 GMM 模型之自由度，即其所有參數之數目，其值之計算方式會依分佈模型之選擇(GMM 模型或向量量化法)而有所不同。

B. 利用貝氏訊息準則選擇最佳模型

Nishida 及 Kawahara[48]將 GMM 模型中高斯分佈之參數值分為兩種方法去預估，一為傳統 GMM 模型預估參數方式，即使用 EM 演算法去估算每一高斯分佈之參數值；另一種方法則是使用一種稱為 "延伸向量量化法"(extended VQ, EVQ)，其作法只是將 GMM 模型中所有高斯分佈參數的權重值與共變異數矩陣取平均，讓 GMM 模型中所有高斯分佈之權重值與共變異數矩陣皆相同，便可以與傳統 GMM 模型做比較。如此一來便可以經由計算貝氏訊息準則之 值大小來決定最佳之 GMM 模型及其高斯分佈參數值，其中計算傳統 GMM 模型之 BIC 值公式如下：

$$BIC_{GMM} = \log p(\mathbf{X} \mid \Theta_{GMM}) - \frac{1}{2} M (2d+1) \log N,$$

其中 M 為 GMM 模型中高斯分佈之群數，d 為特徵之維度，懲罰權重設為 1。而 EVQ 之 BIC 值計算公式則是：

$$BIC_{EVQ} = \log p(\mathbf{X} \mid \Theta_{EVQ}) - \frac{1}{2}(M+1) d \log N,$$

$$p(\theta_{EVQ}) = \frac{1}{M},$$

$$\Sigma_{EVQ} = \frac{1}{M} \sum_{i=1}^{M} \Sigma_{GMM_i},$$

其中 Σ_{GMM_i} 為 GMM 模型中第 i 個高斯分佈之共變異數矩陣。

C. 基於自我分裂法則之高　斯混合模型學習演算法　(Self-Splitting Gaussian Mixture Learning, SGML)

SGML 演算法主要用於自動決定 GMM 模型中高斯分佈之群數，首先對每一個高斯分佈計算其分裂之後所增加之 BIC 值，並選取 BIC 值增加最大之高斯分佈，將其分裂為兩個，並且在分裂之後計算新的 GMM 模型之 BIC 值，如此重覆執行分裂之動作，直到找到具有最大 BIC 值之 GMM 模型出現為止，即代表此一組訓練資料以此一 GMM 模型來描述有最大之相似度。首先，假設 $\mathbf{X} = \{\mathbf{x}(t); t = 1, \ldots, N\}$ 為訓練資料之集合；$\mathbf{w} = \{ p(\theta_r),\ \theta_r \mid r = 1, 2, \ldots, \text{bestNum}\}$ 為輸出參數之集合，其中 $p(\theta_r)$ 為第 r 個高斯分佈之事前機率，$\theta_r = \{\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r\}$ 為第 r 個高斯分佈之參數集合，包含平均值向量與共變異數矩陣。SGML 演算法之詳細步驟如下：

步驟 1: 初始設定 (Initialization)

首先將所有訓練資料視為一群，計算全部訓練資料之平均值向量與共變異數矩陣作為初始值。令

SRange = 5;
compoNum = 1;
$\mathbf{w} = \{ p(\theta_1),\ \theta_1 = \{\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1\}\}$;
$\text{BIC\_set}(1) = \text{BIC}(\text{GMM}_1, \mathbf{X})$;
$\text{GMM\_set}(1) = \mathbf{w}$;

其中 SRange 此變數為確認目前之 BIC 為最大時還需繼續參照之高斯分佈群數範圍；compoNum 此變數為目前之 GMM 模型中高斯分佈之群數；$p(\theta_1) = 1$, $\boldsymbol{\mu}_1$ 與 $\boldsymbol{\Sigma}_1$ 分別為全部訓練資料之平均值向量與共變異數矩陣；BIC_set(compoNum) 為計算以 compoNum 個高斯分佈之 GMM 模型來描述全部訓練資料之 BIC 值；GMM_set(compoNum)為 GMM 模型中所有參數之集合。

步驟 2: 資料分群 (Data Clustering)

將目前之訓練資料利用 $GMM_{compoNum}$ 之參數計算其最大相似度值，以決定每筆訓練資料是屬於哪一高斯分佈群集。

for i = 1, 2, ..., compoNum

$EM\_cluster_r = \phi$ ；

for t = 1, 2, ..., N

$j = \arg\max_{1 \le r \le compoNum} \{p(\theta_r \mid \mathbf{x}(t)\}$;

$EM\_cluster_j = EM\_cluster_j \cup \mathbf{x}(t)$; //將 $\mathbf{x}(t)$ 資料加入 $EM\_cluster_j$ 中

步驟 3: 分裂 (Splitting)

假設目前所有高斯分佈之間是互相獨立，計算每一高斯分佈分裂前後 BIC 值之差值，即 $\Delta BIC_{21} = BIC(GMM_2, EM\_cluster_r) - BIC(GMM_1, EM\_cluster_r)$，$1 \le r \le compoNum$，然後對具有最大 ΔBIC_{21} 之高斯分佈群集進行分裂之動作，其餘高斯分佈群集皆保留原狀：

$$whichSplit = \arg\max_{1 \le r \le compoNum} \{\Delta BIC_{21}(EM\_cluster_r)\};$$

假設 $EM\_cluster_{whichSplit}$ 為具有最大 ΔBIC_{21} 之高斯分佈群集，其分裂後之參數為

$\bar{\lambda}_1 = \{p(\bar{\theta}_1), \bar{\theta}_1\}, \bar{\lambda}_2 = \{p(\bar{\theta}_2), \bar{\theta}_2\}$

$\bar{\Theta} = \{\bar{\boldsymbol{\mu}}_1, \bar{\boldsymbol{\Sigma}}_1, \bar{\boldsymbol{\mu}}_2, \bar{\boldsymbol{\Sigma}}_2\}$

令

$p(\bar{\theta}_1) = p(\bar{\theta}_2) = p(\theta_{whichSplit}) / 2$;

$\mathbf{w} = \mathbf{w} \setminus \{ p(\theta_{whichSplit}), \theta_{whichSplit} \}$; //從參數集合 $\mathbf{w}$ 中移除 $\theta_{whichSplit}$ 參數

$\mathbf{w} = \mathbf{w} \cup \{ \lambda_1 , \lambda_2 \}$; //將分裂後之參數加入參數集合 $\mathbf{w}$ 中

compoNum = compoNum + 1; // GMM 模型中高斯分佈之群數增加 1

步驟 4: 整體 EM 學習 (Global EM learning)

以目前之 GMM 模型參數 $\mathbf{w}$ 做為初始值執行 EM 演算法，收斂後更新 GMM 模型之參數：

GMM_set(compoNum) = $\mathbf{w}$;

計算更新後之 GMM 模型之 BIC 值：

BIC_set(compoNum) = BIC($\mathbf{w}$, $\mathbf{X}$);

假如(compoNum 大於 SRange)並且(BIC_set(compoNum-SRange)在整個學習過程之曲線上是最大值)兩者同時成立之時，則：

bestNum = compoNum – Srange;

$\mathbf{w}$ = GMM_set(bestNum);

同時結束 SGML 演算法;反之，重複執行**步驟 2-4**。

執行上述之演算法後，對於每一種聲音，我們可以得到其 GMM 模型及其參數集合，因此我們將以所有高斯分佈群集之平均值向量為此一種類聲音之特徵向量集合，而且以所選定之特徵值向量來計算此待辨識之聲音訊號之特徵向量與資料庫中所有種類聲音 之每一特徵向量之距離。

2.1.4 線性區別分析演算法(Linear Discriminant Analysis, LDA)

線性區別分析演算法[46]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，線性區別分析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。線性區別分析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 (transformation matrix)來將維度 n 的特徵向量轉換成維度 d 的向量，在這裡 $d \le n$，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:

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

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:

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

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:

$$S_B = \sum_{j=1}^{C} n_j(\mu_j - \boldsymbol{\mu})(\mu_j - \boldsymbol{\mu})^T$$

$\boldsymbol{\mu}$ 為所有類別的平均向量。線性區別分析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(transformation matrix) A_{opt}，而其維度大小為 $n \times d$:

$$A_{opt} = \arg\max_{A} \frac{tr(A^T S_B A)}{tr(A^T S_W A)}.$$

此一轉換矩陣，可經由求出 $S_W^{-1}S_B$ 的特徵向量(eigenvectors)來得到，而 A_{opt} 之 d 個行向量為前 d 個最大特徵值(eigenvalue)值所對應之特徵向量。在此我們是取(鳥鳴聲種類數目-1)作為 d 值。

在我們決定出最佳的轉換矩陣 A_{opt} 後，我們以 A_{opt} 將每一正規化(normalized)後之 n 維的特徵向量轉換為 d 維之向量。令 $\mathbf{f}_j$ 為類別 j 裡維度為 n 的特徵向量，轉換成維度為 d 的向量之公式如下:

$$\mathbf{x}_j = A_{opt}^T \mathbf{f}_j$$

2.1.5 辨識階段

在辨識的部份中，在輸入每個鳥類鳴叫聲音後，首先將每一個音節切割出來，並求出每個音節之特徵向量 **x**，然後對各個特徵值作正規化，接著以主軸分析演算法來降低降低特徵向量維度，最後利用線性區別分析演算法再進一步降低特徵向量維度且提高不同類別間特徵向量之距離，辨識時以歐基里德距離(Euclidean distance)來計算測試特徵向量和每一鳥類鳴叫聲所建立之高斯混合模型中高斯分佈的向量平均值之間的距離，取最小距離作為代表距離該種鳥類之距離，最後我們取測試特徵向量與不同鳥類之高斯混合模型所計算出之距離最小者，即代表辨識結果為該鳥種，令 C 為鳥種數目，b 代表辨識出來之鳥類種類，式子如下：

$$b = \arg\min_{1\le k\le C}\left(\min_{1\le i\le M_k} d(\mathbf{x}, \boldsymbol{\mu}_{k,i})\right)$$

其中 **x** 為測試特徵向量；M_k 為第 k 個高斯混合模型中高斯分佈群數；$\boldsymbol{\mu}_{k,i}$ 為第 k 個高斯混合模型中第 i 個高斯分佈的向量平均值。

2.2 音樂曲風之自動分類

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:特徵擷取、線性區別分析演算法 、以及分類器之學習 。辨識階段是由三個主要模組所組成:特徵擷取、線性區別分析轉換、和分類。

2.2.1 特徵擷取

在本計劃中，我們用來自動分類音樂曲風之特徵可分為以下幾類：梅爾倒頻譜係數之調變頻譜(MMFCC)、八度音程頻譜對比值之調變頻譜(MOSC)、及 MPEG-7 正規化聲音頻譜封包之調變頻譜(MNASE)，由於梅爾倒頻譜係數已於鳥類鳴叫聲音辨識系統中描述，因此不再重覆敘述，以下我們僅描述八度音程頻譜對比值(OSC)及 MPEG-7 正規化聲音頻譜封包(NASE)等。

2.2.1.1 八度音程頻譜對比值(OSC)

八度音程頻譜對比值是用來描述一音樂訊號之頻譜特性，首先將音樂訊號依據八度音程之觀念將其分解為一些次頻帶，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，頻譜波峰大略是反映聲音訊號之泛

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。

對於一音樂訊號，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一些次頻帶，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：

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

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

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峰值($Peak_b$, $1 \leq b \leq 8$)及頻譜對比值(SC_b, $1 \leq b \leq 8$)為此一音框之特徵向量，然後我們就可以對每一音框之特徵向量做調變頻譜分析，得到八度音程頻譜對比值之調變頻譜特徵。

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

2.2.1.2 MPEG-7 正規化聲音頻譜封包(NASE)

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

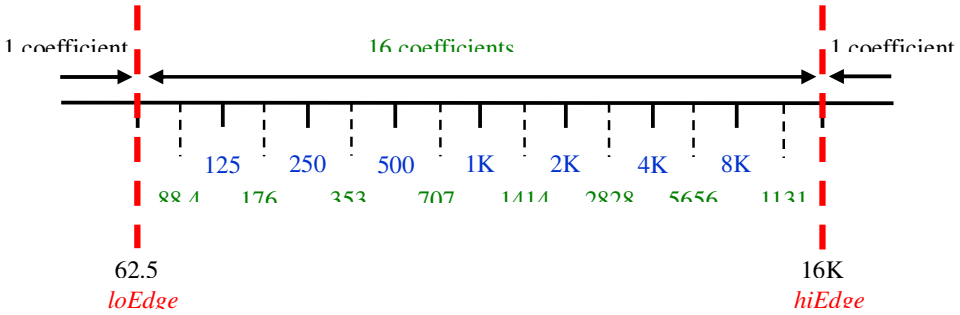
兼顧描述性與簡潔性。聲音頻譜封包(audio spectrum envelope, ASE)在MPEG-7標準裡普遍用於表示原始聲音訊號中每一頻帶之功率頻譜[49, 50]，主要是描述介於*loEdge* (預設62.5Hz)與*hiEdge* (預設為16000Hz)間的頻譜資訊，將介於*loEdge*與*hiEdge*間的頻率再分解為B個頻帶，而每一頻帶的頻寬解析度是以八度音 (octave)解析度為基準，以1000Hz為中心上下區分，總計介於[*loEdge*, *hiEdge*]間之頻帶數目為 $B = 8/r$，其中r是八度音的解析度，其範圍是介於1/16倍八度音至8倍八度音之間：

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

在本計劃中，我們擬採用之r值為1/2，因此$B = 16$，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+2) 個頻帶，圖三為一個八度音解析度之邊界頻率f_{edge}分隔圖。



圖三、 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，擷取每一音框之NASE係數之詳細步驟如下(其流程圖請參考圖四)。

步驟 1: 將輸入音框乘上漢明視窗(Hamming windowing)：

$$w[n] = 0.54 - 0.46\cos(\frac{2\pi n}{N-1}), \ 0 \le n \le N_w - 1,$$

N_w 為漢明視窗之大小

步驟 2: 快速傅立葉轉換(FFT)

將聲音訊訊號從時間域轉換至頻率域：

$$X[k]=\sum_{n=0}^{N-1}s[n]e^{-j2\pi\frac{k}{N}n},\ 0\le k<N,$$

N：音框大小，$s[n]$：聲音訊號

步驟 3: 計算 ASE

計算功率頻譜之係數值並做正規化：

$$P(k)=\frac{1}{N\times E_w}\left|X[k]\right|^2,\ k=0,N/2$$

$$P(k)=\frac{2}{N\times E_w}\left|X[k]\right|^2,\ 0<k<N/2$$

其中

$$E_w=\sum_{n=0}^{N_w-1}\left|w[n]\right|^2$$

計算出每個頻帶的能量總合：

$$ASE(b)=\sum_{k=loK_b}^{hiK_b}P(k),\ 0\le b\le B+1$$

其中 loK_b 及 hiK_b 為第 b 個頻帶之下方邊界頻率及上方邊界頻率之索引值，最後將其轉換至分貝單位：

$$ASE_{dB}(b)=10\log_{10}(ASE(b)),\ 0\le b\le B+1$$

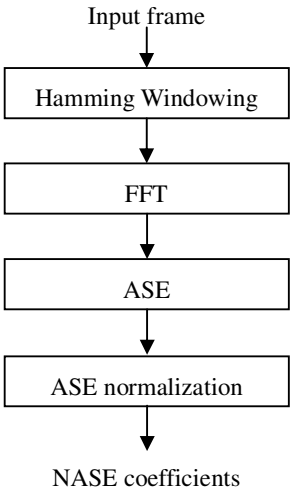
步驟 4: 計算 NASE

首先，先計算每一個音框之 *RMS* 值，R：

$$R(b)=\sqrt{\sum_{b=0}^{B+1}((ASE_{dB}(b))^2},\ 0\le b\le B+1$$

接著對每一個 ASE_{dB} 做正規化之動作以計算 NASE 值：

$$NASE(b)=\frac{ASE_{dB}(b)}{R(b)},\ 0\le b\le B+1$$



圖四、 計算一音框之 NASE 係數之流程圖

2.2.1.3 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。

傳統的方法是在流動頻譜上使用帶通濾波器來擷取此一頻帶，但這方法不能正確地獲得調變頻譜，因此，Wada 等人提出調變頻譜控制(modulation spectrum control, MSC)方法[51]，對每一個頻率中的調變頻譜乘上一個權重值，以消除不需要的調變頻譜成份。此一權重值法可以避免帶通濾波器之一些限制，如濾波器係數的數目、延遲的時間、穩定度、和相位的誤差等，因此更適合用在聲音辨識系統。

相對於頻譜消去法(spectrum subtraction, SS)，以調變頻譜方法來強化語音辨識在抗雜訊干擾之主要目的是其與雜訊之構成要件(例如雜訊的類型、變化、和 SNR 等)是無關的。然而，此一方法可以和 SS 共用於聲音辨識系統中。當將輸入之含雜訊之聲音訊號 $y(t)$做傅立葉轉換到頻率域，可以表示成:

$$y(t) = h(t) \otimes (x(t) + n(t))$$
$$Y(f) = H(f)X(f) + H(f)N(f)$$

其中 $x(t)$表示原始聲音訊號，$h(t)$表示系統雜訊，而 $n(t)$表示環境雜訊，$H(f)N(f)$則表示外加之雜訊，此一雜訊可以用 SS 來移除，或是對流動頻譜以帶通濾波器來移除。然而，過度消除較低頻的調變頻率帶可能會造成負的頻譜能量值。假設附加之雜訊以 SS 做適當消除後，則此訊號之對數能量頻譜可以表示成：

$$\log|Y(f)| = \log|X(f)H(f)| = \log|X(f)| + \log|H(f)|$$

因此，對於對數能量頻譜的時間軌跡做帶通濾波可以將系統雜訊 $H(f)$移除。一般而言，SS 是在能量頻譜(power spectrum)域中去除調變頻譜之 DC 值，而 CMS 則在對數能量頻譜(logarithmic power spectrum)域中消除調變頻譜之 DC 值。然而，MSC 可以消除更多不需要的組成元素，且若同時使用 SS 和 MSC 會比單獨使用任一種得到更好的辨識效果。

Kanedera 等人[52, 53]將調變頻譜中 0 至 40 赫茲之頻率範圍分成若干頻帶，實驗在不同調變頻帶之係數對於聲音辨識上之重要性，此一結果顯示在調變頻率 1 至 16 赫茲之範圍內所擷取之特徵有較好之辨識效果。Vuuren 與 Hermansky[54]同樣也將調變頻率分成若干頻帶，且對調變頻率較低頻介於 0 至 1 赫茲部份更細分成三個頻帶，實驗結果

顯示在低頻 0.5 至 1 赫茲之頻帶範圍內其重要性亦不差，其實驗結果顯示取調變頻譜之頻率介於 0.5 和 16 赫茲範圍內對語音辨識之重要性較大。

在本計劃中，我們擬將此一調變頻譜之觀念應用於不同之特徵向量，包括梅爾倒頻譜係數、OSC 係數、及 NASE 係數及梅爾倒頻譜係數等，計算各種特徵向量之調變頻譜，然後將調變頻譜介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。

表二、 調變頻帶之頻率範圍

Modulation Frequency Band	Modulation frequency interval (Hz)
0	[0, 0.5)
1	[0.5, 1)
2	[1, 2)
3	[2, 4)
4	[4, 8)
5	[8, 16)
6	[16, 32)
7	[32, 64)

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：

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

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五(J = 8)，每個調變頻帶我們擷取 二種調變特徵值：調變頻譜波谷值(modulation spectral valley, MSV)及調變頻譜對比值(modulation spectral contrast, MSC)。首先定義調變頻譜波谷值及調變頻譜波峰值(modulation spectral peak, MSP)如下：

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

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

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。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及調變頻譜波谷值可構成二個 $D \times J$ 之矩陣，每一矩陣可視為二維之影像圖，統稱為調變聲譜圖(實際上含調變頻譜對比值聲譜圖及調變頻譜波谷值聲譜圖)。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。

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

此一步驟與前述之鳥類鳴叫聲音辨識系統相同 (請參考 2.1.4)。

2.2.3 分類

此步驟與前述之鳥類鳴叫聲音辨識系統相同 (請參考 2.1.5)。

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

事實上，沒有一個分類器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，某些分類器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分類器可以得到更好的辨識率。因此，對於一輸入之音樂，在一組分類器之不同辨識結果中，該選擇那一個分類器之辨識結果才能得到最佳之辨識效能，或是如何整合數個分類器之辨識結果以提高辨識率，是我們的研究方向，因此我們以資訊融合方法來整合各種分類器之結果以提高辨識率(如圖五所示)。

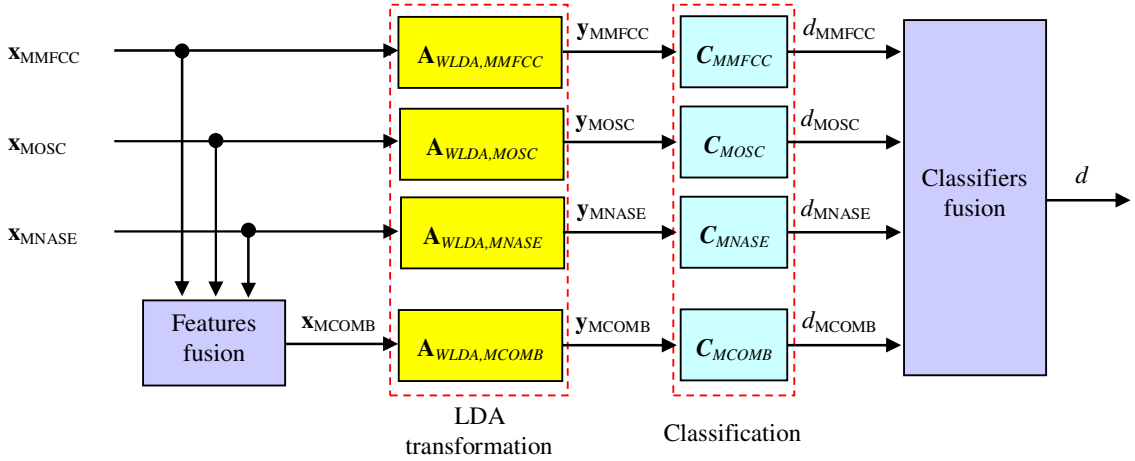
首先我們分別以 MMFCC、MOSC、及 MNASE 之特徵向量 $\mathbf{x}_{\mathrm{MMFCC}}$、$\mathbf{x}_{\mathrm{MOSC}}$、及 $\mathbf{x}_{\mathrm{MNASE}}$ 計算兩音樂檔案之特徵距離 d_{MMFCC}、d_{MOSC}、及 d_{MNASE}，此外我們還將此三個特徵向量

串接起來得到混合特徵向量 $\mathbf{x}_{COMB}$：

$$\mathbf{x}_{COMB} = [(\mathbf{x}_{MMFCC})^T, (\mathbf{x}_{MOSC})^T, (\mathbf{x}_{MNASE})^T]^T$$

再計算 $\mathbf{x}_{COMB}$ 特徵向量之距離 d_{MCOMB}，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：

$$d_{MMFCC} = d_{MMFCC} + d_{MOSC} + d_{MNASE} + d_{MCOMB}$$



圖五、 整合分類器之音樂曲風自動分類辨識系統之架構圖

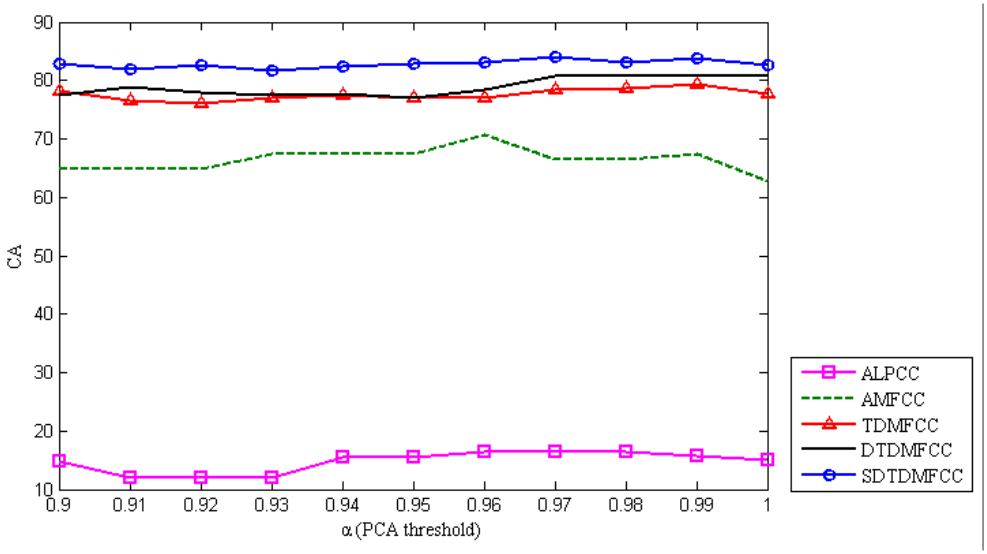
3. 實驗結果與討論

3.1 鳥類鳴唱聲音之自動分類辨識結果

在實驗中所使用之鳥類鳴聲資料，總共有 28 種台灣鳥類，其訓練資料之鳥類錄音與測試資料之鳥類錄音皆為在不同環境下使用不同錄音設備之錄音。首先，將所有鳥類錄音重新取樣調整為 44100 Hz，音訊範圍大小為 16 bits，對資料庫中所有鳥類擷取特徵進行辨識前，我們以人工的方式對每個聲音檔案切出鳥類音節，在 28 種鳥類共有 3143 個訓練音節和 646 個測試音節，在表三中顯示實驗之鳥類名稱及每種鳥類聲音所切割之音節數目。在此實驗中我們對每一音節取其二維梅爾倒頻譜係數(TDMFCC)、動態二維梅爾倒頻譜係數(DTDMFCC)和整合之二維梅爾倒頻譜係數(SDTDMFCC)作為特徵，圖六比較各種特徵向量對於 28 種鳥類於 PCA 門檻值介於 0.9 至 1.0 間之辨識率，我們比較之特徵向量有 ALPCC、AMFCC、TDMFCC、DTDMFCC、及 SDTDMFCC，其實驗結果顯示二維梅爾倒頻譜係數之辨識正確率較高，而且 SDTDMFCC 整合二維梅爾倒頻譜係數及動態二維梅爾倒頻譜係數有最佳之辨識正確率。表四則呈現每種鳥類聲音之個別辨識率、選取之辨識模型(EVQ 或 GMM)、分群數目(N_s)於 PCA 門檻值為 0.97 之數值。

表三、 28 種鳥類訓練音節與測試音節數目

Bird Name	Training Syllable	Test Syllable
大冠鷲	10	4
小卷尾	229	37
小啄木	17	25
小翼鶇	296	29
小彎嘴畫眉	120	22
火冠戴菊鳥	194	57
白耳畫眉	98	14
白喉笑鶇	100	37
白腹秧雞	172	15
灰鶯	70	8
竹鳥	31	31
岩鷚	122	53
青背山雀	140	14
冠羽畫眉	49	12
紅頭山雀	61	24
栗背林鴝	230	18
烏頭翁	131	30
深山竹雞	123	27
深山鶯	51	8
筒鳥	284	45
黃山雀	222	27
黃腹琉璃	76	12
煤山雀	149	34
鳳頭蒼鷹	32	16
頭烏線	32	18
鵂鶹	61	14
藍腹鷴	23	10
藪鳥	20	5



圖六、各種特徵向量對於 28 種鳥類之辨識率比較圖

表四、 當 PCA 門檻值為 0.97 時每種鳥類聲音之個別辨識率、選取之辨識模型(EVQ 或 GMM)及分群數目(N_s)

Subject Code	Bird Name	*CA* (%)	N_s	Selected Model
1	Crested Serpent Eagle	100.00	2	EVQ
2	Bronzed Drongo	86.49	5	EVQ
3	Gray-headed Pygmy Woodpecker	0.00	1	EVQ
4	Blue Shortwing	72.41	4	EVQ
5	Streak-breasted Scimitar Babbler	54.55	3	GMM
6	Taiwan Firecrest	100.00	3	EVQ
7	Taiwan Sibia	100.00	6	EVQ
8	White-throated Laughing Thrush	94.59	3	EVQ
9	White-breasted Water Hen	100.00	4	EVQ
10	Beavan's Bullfinch	100.00	3	EVQ
11	Gray-sided Laughing Thrush	100.00	3	EVQ
12	Alpine Accentor	71.70	1	EVQ
13	Green-backed Tit	7.14	5	EVQ
14	Taiwan Yuhina	100.00	3	EVQ
15	Red-headed Tit	100.00	2	EVQ
16	Collared Bush Robin	94.44	9	EVQ
17	Taiwan Bulbul	83.33	5	EVQ
18	Taiwan Hill Partridge	88.89	6	EVQ
19	Verreaux's Bush Warbler	100.00	4	EVQ
20	Oriental Cuckoo	95.56	3	GMM
21	Taiwan Tit	96.30	7	EVQ
22	Vivid Niltava	100.00	5	EVQ
23	Coal Tit	100.00	4	EVQ
24	Crested Goshawk	100.00	3	EVQ
25	Gould's Fulvetta	33.33	1	EVQ
26	Collared Pigmy Owlet	100.00	1	EVQ
27	Swinhoe's Pheasant	100.00	3	EVQ
28	Steere's Liocichla	80.00	3	EVQ

3.2 音樂曲風之自動分類結果

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[55]，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。

圖七顯示各種調變特徵向量及混合特徵向量之辨識混淆矩陣(confusion matrix)，由圖七(a)、(b)、(c)，我們可以發現沒有任一單獨之調變特徵向量對於每一音樂類別都能得到最佳之辨識率，因此將各種調變特徵向量混合後可能會有較佳之辨識率(如圖七(d)所示)，此外若我們將各種分類結果再整合後可進一步提高辨識率(如圖七(e)所示)。

	C	*E*	*J*	*M*	*R*	*W*
C	**297**	0	0	0	4	21
E	0	**90**	0	2	5	4
J	3	0	**18**	0	0	7
M	1	4	0	**35**	22	4
R	3	12	6	8	**64**	10
W	16	8	2	0	7	**76**

(a) MMFCC

	C	*E*	*J*	*M*	*R*	*W*
C	**300**	0	0	0	1	13
E	0	**90**	1	2	9	6
J	0	0	**21**	0	0	4
M	0	2	0	**31**	21	2
R	0	11	3	10	**64**	10
W	20	11	1	2	7	**87**

(b) MOSC

	C	*E*	*J*	*M*	*R*	*W*
C	**296**	2	1	0	0	17
E	1	**91**	0	1	4	3
J	0	2	**19**	0	0	5
M	0	2	1	**34**	20	8
R	2	13	4	8	**71**	8
W	21	4	1	2	7	**81**

(c) MNASE

	C	*E*	*J*	*M*	*R*	*W*
C	**298**	3	0	0	1	8
E	0	**98**	0	1	5	5
J	0	1	**20**	0	0	2
M	0	1	0	**31**	15	1
R	3	8	3	11	**75**	16
W	19	3	3	2	6	**90**

(d) MCOMB

	C	*E*	*J*	*M*	*R*	*W*
C	**311**	2	0	0	0	10
E	0	**98**	0	1	5	3
J	0	0	**20**	0	0	2
M	0	0	0	**33**	13	2
R	1	9	3	10	**78**	12
W	8	5	3	1	6	**93**

(e) MMFCC+MOSC+MNASE+MCOMB

圖七、 各種調變特徵向量及混合特徵向量之辨識混淆矩陣

表格五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(MMFCC+MOSC+MNASE+MCOMB)得到最佳之辨識率(86.83%)，比 2004 年音樂曲風分類競賽之優勝者(84.07%)還高

表五、 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 approach (MMFCC+MOSC+MNASE+MCOMB)	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] http://www.earthlife.net/birds/song.html

[2] E. A. Brenowitz, D. Margoliash, and K. M. Nordeen, “An introduction to birdsong and the avian song system”, *Journal of Neurobiology*, Vol. 33, Issue 5, pp. 495-500, Nov. 1997.

[3] J. Kogan and D. Margoliash, “Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: a comparative study”, *Journal of the Acoustical Society of America*, Vol. 103, No. 4, pp. 2187-2196, Apr. 1998.

[4] A. L. McIlraith and H. C. Card, “Birdsong recognition with DSP and neural networks”, in *Proceedings of IEEE Conference on Communications, Power, and Computing*, Vol. 2, pp. 409-414, May 1995.

[5] A. L. McIlraith and H. C. Card, “A comparison of backpropagation and dtatistical classifiers for bird identification”, in *Proceedings of IEEE International Conference on Neural Networks* , Vol. 1, pp. 100-104, June 1997.

[6] A. L. McIlraith and H. C. Card, “Birdsong recognition using backpropagation and multivariate statistics”, *IEEE Trans. on Signal Processing*, Vol. 45, No. 11, pp. 2740-2748, Nov. 1997.

[7] A. L. McIlraith and H. C. Card, “Bird song identification using artificial neural networks and statistical analysis”, in *Proceedings of Canadian Conference on Electrical and Computer Engineering*, Vol. 1, pp. 63-66, May 1997.

[8] 張勇富, “以語料分析為主的鳥音辨識系統研究”, 國立東華大學碩士論文, 中華民國九十二年七月.

[9] S. E. Anderson, A. S. Dave, and D. Margoliash, “Template-based automatic recognition of birdsong syllables from continuous recordings”, *Journal of the Acoustical Society of America*, Vol. 100, No. 2, pp.1209-1219, Aug. 1996.

[10] A. Harma, “Automatic identification of bird species based on sinusoidal modeling of syllables”, in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 5, pp. 545-548, 2003.

[11] A. Harma and P. Somervuo, “Classification of the harmonic structure in bird vocalization”, in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 5, pp. 701-704, 2004.

[12] P. Somervuo and A. Harma, “Bird song recognition based on syllable pair histograms”, in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 5, pp. 825-828, 2004.

[13] S. Fagerlund and A. Harma, “Parametrization of inharmonic bird sounds for automatic recognition”, in Proceedings of the *13th European Signal Processing Conference* (*EUSIPCO 2005*), Antalya, Turkey, Sep. 2005.

[14] P. Somervuo, A. Harma, and S. Fagerlund, “Parametric representations of bird sounds for automatic species recognition”, *IEEE Trans. on Audio, Speech, and Language Processing*, Vol. 14, No. 6, pp. 2252-2263, Nov. 2006.

[15] M. B. Trawicki and M. T. Johnson, “Automatic song-type classification and speaker identification of Norwegian Ortolan Bunting (Emberiza Hortulana) vocalizations”, in *Proc. of IEEE Workshop on Machine Learning for Signal Processing*, pp. 277-282, Sep. 2005.

[16] A. Selin, J. Turunen, and J. T. Tanttu, “Wavelets in Recognition of Bird Sounds”, *EURASIP Journal on Advances in Signal Processing,* Vol. 2007, Article ID 51806, 9 pages.

[17] S. A. Selouani, M. Kardouchi, E. Hervet, and D. Roy, “Automatic Birdsong Recognition Based on Autoregressive Time-Delay Neural Networks”, in *Proc. of 2005 ICSC Congress on Computational Intelligence Methods and Applications*, Dec. 2005.

[18] J. Jose Burred and A. Lerch, “A hierarchical approach to automatic musical genre classification”, in *Proc. of the 6th Int. Conf. on Digital Audio Effects*, pp. 8-11, September 2003.

[19] J. G. A. Barbedo and A. Lopes, "Research article: automatic genre classification of musical signals", *EURASIP Journal on Advances in Signal Processing*, Vol. 2007, pp.1-12, June 2006.

[20] T. Li and M. Ogihara, “Music genre classification with taxonomy”, in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vol. 5, pp. 197-200, March 2005.

[21] J. J. Aucouturier and F. Pachet, “Representing musical genre: a state of the art”, *Journal of new musical research*, Vol. 32, No. 1, pp. 83-93, 2003.

[22] 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.*, Vol. 1, pp. 113-116, 2002.

[23] 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)*, 2005.

[24] W. A. Sethares, R. D. Robin, and J. C. Sethares, “Beat tracking of musical performance

using low-level audio feature", *IEEE Transactions on speech and audio processing,* Vol. 13, No. 12, March 2005.

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

[26] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model", *IEEE Transactions on Speech and Audio Processing*, Vol. 8, No. 6, pp. 708-716, November 2000.

[27] R. Meddis and L. O'Mard, "A unitary model of pitch perception", *Acoustical Society of America*, Vol. 102, No. 3, pp. 1811-1820, September 1997.

[28] 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.

[29] A. Meng, P. Ahrendt, J. Larsen, and L. K. Hansen, "Temporal feature integration for music genre classification", *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 15, Issue 5, pp.1654 – 1664, July 2007.

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

[31] S. Sukittanon, L. E. Atlas, and J. W. Pitton, "Modulation-scale analysis for content identification", *IEEE Transactions on signal processing,* Vol. 52, No. 10, pp.3023-3035, October 2004.

[32] 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 2006 *IEEE International Conference on Multimedia and Expo (ICME)*, pp.1085-1088, July 2006.

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

[34] 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.

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

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

[37] U. Bagci, and E. Erzin, "Automatic classification of musical genres using inter-genre

similarity", *IEEE Signal Processing Letters*, Vol. PP, Issue 99, pp.1-4, 2007.

[38] K. Umapathy, S. Krishnan, and R. K. Rao, "Audio signal feature extraction and classification using local discriminant bases", *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 15, Issue 4, pp.1236 – 1246, May 2007.

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

[40] 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.

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

[42] Y. Ariki, S. Mizuta, M. Magata, and T. Sakai, "Spoken-word recognition using dynamic features analysed by two-dimensional cepstrum", *IEE Proceedings, Pt. I*, No. 2, Apr. 1998, pp. 133-140.

[43] H. F. Pai and H. C. Wang, "A study of the two-dimensional cepstrum approach for speech recognition", *Computer Speech and Language*, No. 6, 1992, pp. 361-375.

[44] C. T. Lin, H. W. Nein, and J. Y. Hwu, "GA-based noisy speech recognition using two-dimensional cepstrum", *IEEE Trans. on Speech and Audio Processing*, Vol. 8, No. 6, Nov. 2000, pp. 664-675.

[45] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. ASSP-34, No. 1, February, 1986, pp. 52-59.

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

[47] S. S. Cheng, H. M. Wang and H. C. Fu, "A model-selection-based self-splitting Gaussian mixture learning with application to speaker identification", *EURASIP Journal on Applied Signal Processing*, Vol. 2004, Issue 17, 2004, pp. 2626-2639.

[48] M. Nishida and T. Kawahara, "Speaker Model Selection Based on the Bayesian Information Criterion Applied to Unsupervised Speaker Indexing", *IEEE Trans. on Speech and Audio Processing,* Vol. 13, No. 4, July 2005.

[49] 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.

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

[51] N. Wada, N. Hayasaka, S. Yoshizawa, and Y. Miyanaga, **"**Direct control on modulation spectrum for noise-robust speech recognition and spectral subtraction**",** in *Proceedings of IEEE International Symposium on Circuits and Systems*, 21-24 May, 2006.

[52] N. kanedera, T. Arai, H. Hermansky, and M. Pavel, "On the relative importance of various components of the modulation spectrum for automatic speech recognition", *Speech Communication*, Vol. 28, Issue 1, May 1999, pp.43-55.

[53] N. kanedera, T. Arai, H. Hermansky, and M. Pavel, "On the importance of various modulation frequencies for speech recognition", in Proc. of *ESCA*, 1997, pp. 1079-1082.

[54] S. V. Vuuren and H. Hermansky, "On the importance of components of the modulation spectrum for speaker verification", in Proc. of *ICSLP*, Nov 1998.

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

三. 計畫成果自評

本計劃完成可自動辨識鳥類鳴聲之辨識系統，可以應用於各種不同之錄音環境及錄音器材，最佳辨識率可達 84.06%。此一辨識系統可用以輔助調查鳥類族群之生態、棲地之變化，並能減少對生態的影響，建立更完善的台灣生物聲音資料庫。

此外我們也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。

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**, 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)

[2] **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.

[3] 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)

[4] **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)

[5] **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.

[6] **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.

[7] **C. H. Lee**, C. H. Chou, C. C. Han, and R. Z. Huang, "Automatic Recognition of Frog Calls Using Averaged MFCC and Linear Discriminant Analysis", in *Proceedings of the 9th Conference on Artificial Intelligence and Applications*, Taipei, Nov. 5-6, 2004.

[8] **C. H. Lee**, C. H. Chou, and R. Z. Huang, "Automatic Recognition of Bioacoustic Sounds: an Experiment on the Frog Vocalizations", in *Proceedings of the 17th IPPR Conference on Computer Vision, Graphics, and Image Processing*, Hualien, Aug. 15-17, 2004.

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

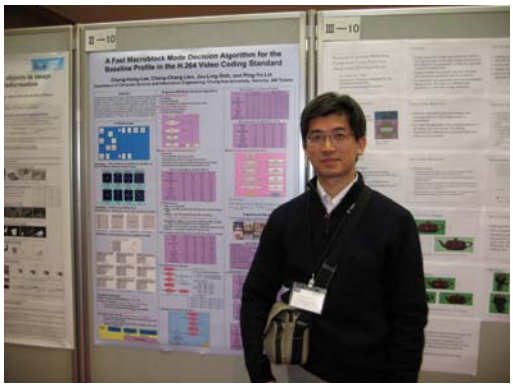
報告人姓名	李建興	服務機構及職稱	中華大學資訊工程學系 副教授
時間 會議 地點	Jan. 13-16, 2009 Tokyo Japan	本會核定 補助文號	NSC 97-2221-E-216 -037 -
會議名稱	(中文) (英文) The 3rd Pacific-Rim Symposium on Image and Video Technology		
發表論文題目	(中文) (英文) A fast macroblock mode decision algorithm for the baseline profile in the H.264 video coding standard		

一、參加會議經過

本人於2009年1月13-16日赴日本東京參加 "The 3rd Pacific-Rim Symposium on Image and Video Technology" 國際會議，會中發表論文一篇，如下所示：

C. H. Lee, C. C. Lien, J. L. Shih, and P. Y. Lin, "A Fast Macroblock Mode Decision Algorithm for the Baseline Profile in the H.264 Video Coding Standard", in *Proc. of the 3rd Pacific-Rim Symposium on Image and Video Technology (PSIVT'2009)*, Jan. 13-16, 2009, Tokyo, Japan, pp. 784-795.

本人於1月13日會議當天由台灣出發前往日本東京，我們的論文是安排在第三天1月15日下午poster session13:30至15:30，從台灣來參加研討會的學者人數不少，顯現國內對國際學術交流的重視。



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)

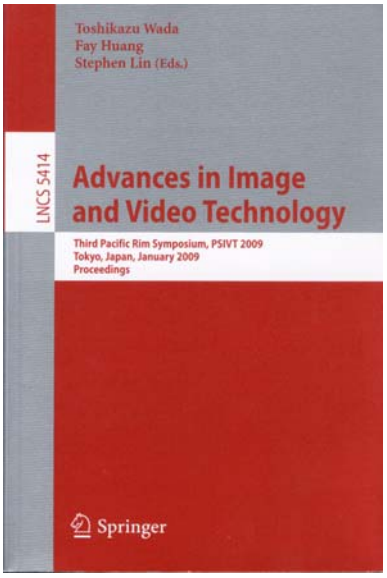
無

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

PSIVT'2009 論文集及論文光碟



六、其他

非常感謝國科會之補助得以參加該研討會。