基於參考簡化模式之聲音識別

Audio Identification Based on Reference-Reduced Model

林敬舜[1]、王大任[2]、趙永誠[2]、王俊祥[2]

摘要

聲音識別首重特徵選取，篩選的標準一般取決於分類訊號是否正確地被表示，以及之後的識別是否較為容易。在本文中，我們採取可減低對資料庫依賴的訊號表示法，同時也簡化了在識別過程中比對需要的編碼方法。本文提出的方法主要分為三部分。在第一個步驟中，我們採用了相關頻譜轉換-感知線性預測法(RASTA-PLP)，也就是利用基於心理聲學的線性預測法所獲得的資訊來表示數位訊號的頻譜封包。其次，為了保留音色變化時的自然逼真性，輸入訊號與已知資料的特徵配對與對齊就變得不可或缺。傳統的聲音辨識方式使用基於頻譜震幅與調和音域攫取出的特徵向量來達到這個目的，但對於需要在時間與頻率同時精確定位的頻譜圖來說，使用餘弦相似量測法(CSM)在資料庫裏找尋與輸入訊號最相似的頻譜圖將是更好的方法。在最後的步驟中，我們利用間歇性的碎形空間填充曲線(虛擬Hilbert曲線)對頻譜圖上的資訊進行區域編碼，因該輸出序列較短，使得利用高斯混合模型與存放在資料庫中的序列做比較時更加有效率。與完全參考模型不同的是，參考簡化模式只使用了少量的比對資料，這使得此法更適合用在手持式的設備上。

關鍵字：聲音識別、線性預測編碼、餘弦相似量測法、虛擬 Hilbert 曲線、高斯混合模型

Abstract

One of the first decisions in any audio identification system is to choose appropriate features. The criteria may count on how exactly the classifying signals are represented, and how easily the following identification can be performed. In this paper, we not only use a sophisticated representation to reduce the dependence on the database, but also provide a simplified encoding to process the matching in the identification. The proposed method is composed of three stages. In the first step, Relative Spectral Transform-Perceptual Linear Prediction (RASTA-PLP), an extension of Linear Predictive Coding (LPC), is used to represent the spectral envelope of a digital signal by using the information of a psychoacoustics-based linear predictive method. Second, to preserve a perceived naturalness in the timbral transitions, it is also necessary to match and align input features with those stored in the database. Conventional approach to audio identification uses feature vectors

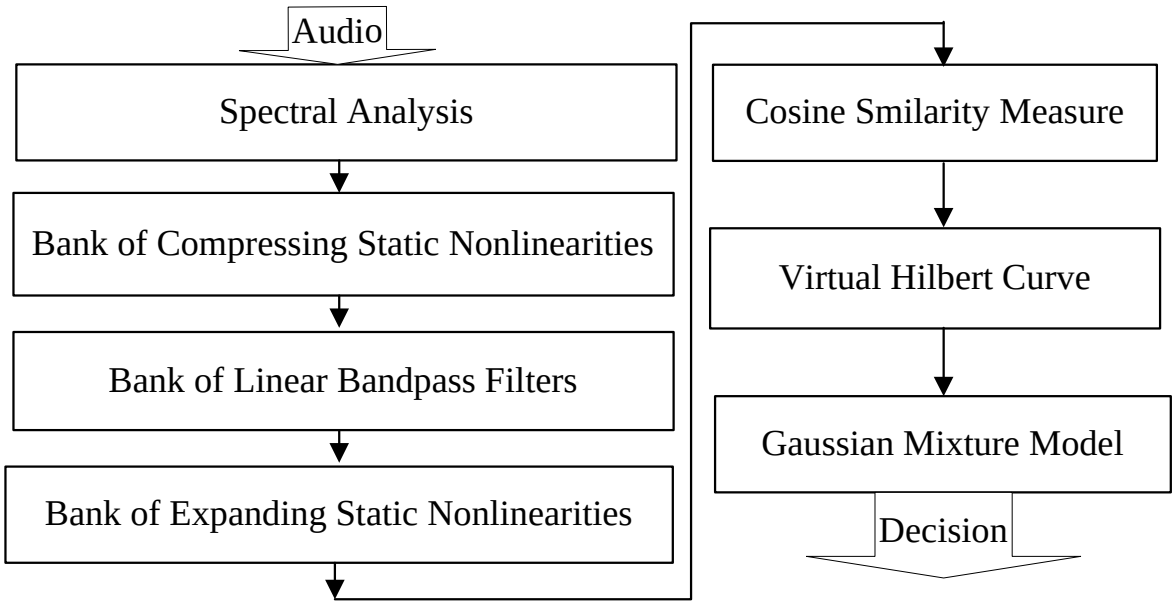
[1]國立台灣科技大學電子系助理教授(簡報作者)
[2]國立台灣科技大學電子系研究生

based on spectral amplitude and harmonic location for this purpose. However, for such an application as spectrogram analysis that requires accurate locations both in time and in frequency, we use the cosine similarity measure (CSM) as an alternative to search the database for the spectrograms that are most similar to the input spectrogram. Finally, virtual Hilbert curve, an intermittent fractal space-filling curve, is introduced to preserve and encode the locality behavior of spectrogram. The result is represented as a size-reduced sequence to facilitate the comparison with the pre-transformed sequences in the database by the Gaussian mixture model (GMM). Unlike the full reference model that makes comparisons among a series of audio samples and the classifying audio, the reference-reduced approach uses only a handful of database, which makes this algorithm more suitable for being implemented on a handheld device.

Keywords: Audio Identification, Linear Predictive Coding, Cosine Similarity Measure, Virtual Hilbert Curve, Gaussian Mixture Model

壹、前言

聲音識別的主要步驟有特徵表示與分類法則的選取，特徵選取的目的在於從原有的特徵點中挑選出最佳的特徵點，使其辨識率能夠達到最高。這些代表性較佳的特徵點，不但能夠簡化分類的計算，而且也有助於了解分類問題的因果關係[1]。一般說來，頻譜分析提供了聲音識別上較佳的資訊，而相關頻譜轉換-感知線性預測法(RASTA-PLP)[2, 3]利用心理聲學的線性預測法則，針對頻譜進一步的做非線性轉換，透過這個技術，我們更加容易瞭解訊號的特性，在保留重要資訊的同時亦可減低相似訊號間的差異性。相關頻譜轉換-感知線性預測法的輸出比起其他方法在人耳聽覺上更具一致性，這使得之後的辨識過程更顯效率。而餘弦相似量測法(CSM)依據時間與頻率框架間的頻譜結構來建構相似矩陣，而經過平移、遮罩與核心函數加總的音色變異量則可用來切割頻譜圖。此法依據區域自我相似度自動地標示出音訊的明顯變化處，這也使得整個處理過程對外在資訊的依賴可減至最低。而分類法則的選擇一般可依資料的分佈狀況、維度、非線性程度與其應用來決定。為了在編碼時同時提高資料間的相似度，我們提出了間歇性的虛擬Hilbert曲線法來對頻譜圖上的RGB資訊進行區域編碼，因該輸出序列較短，使得不論在建立比對資料庫與利用高斯混合模型進行比較時都顯得較具效率。而這樣的技術可廣泛地應用在各種音訊上，特別是高度自我重覆的聲音。而潛在性的應用包含聲音資料庫的建立、專業的聲音編輯、以及如野雁、海豚、鯨魚等生物聲的辨識。針對圖一所列的訊號處理流程，其相對應的方法敘述如下。



圖一、參考簡化模式聲音識別系統流程圖

貳、方法

一、相關頻譜轉換-感知線性預測

人類聽覺對於變化緩慢的刺激訊息較不敏感，這可以解釋為何聆聽者不需花太多的注意力在頻率特徵變化緩慢的溝通環境上，也就是為什麼統計量不變的背景噪音不會劇烈地影響人們的言語溝通[3]。即便如此，在聲音訊號中對於變化緩慢部份的抑制仍具有其物理意義與工程價值。因此，為了能分析聲音中對於變化緩慢較不敏感的因素亦或是分析聲音中的穩態因子，我們利用另一種頻譜評估來替代一般感知線性預測法(PLP)中的臨界頻帶短期頻譜，也就是每個頻率通道在低頻時的訊號將被帶有頻譜零點的帶通濾波器給濾除。因每個頻率通道中的任何常數或變化緩慢的元素被此運算抑制，所以此頻譜評估方法對短期頻譜中緩慢的頻譜變化將變得較不敏銳，其音框的分析步驟如下[2, 3]：

1. 計算臨界頻帶的功率頻譜
2. 利用壓縮靜態的非線性轉換來改變頻譜振幅
3. 對每個轉換過的頻譜元素的時間軌跡做濾波
4. 藉由延展靜態的非線性轉換來改變濾波後的訊號
5. 藉由等響曲線乘以0.33次方來模擬聽覺的冪次法則
6. 利用PLP技術來計算頻譜的全極點模型

此處濾波器的低截止頻率取決於對數頻譜中變化最快的部分，低於這個頻率的頻譜將被濾除，相對地，高截止頻率所標示出的最快頻譜變化則會在輸出時被保留。

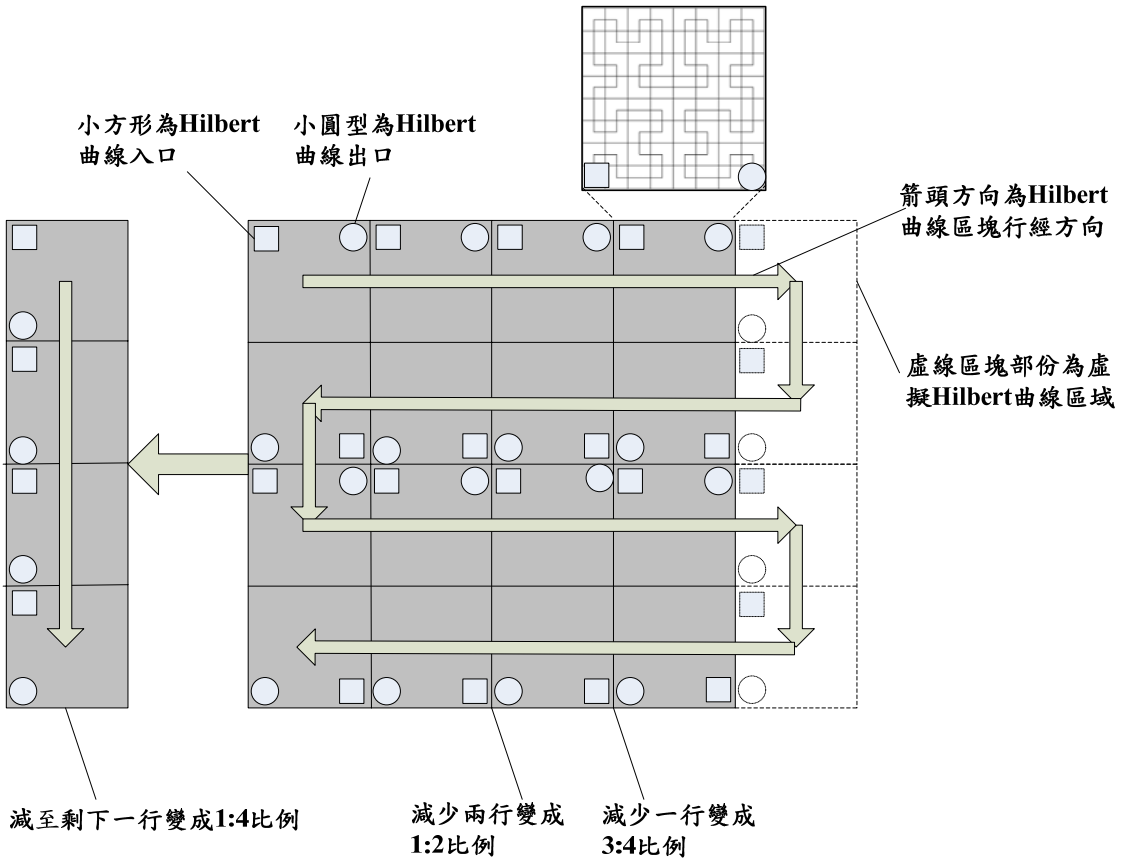
二、餘弦相似量測

聲音識別中很重要的其中一部分便是訊號的切割，常見的作法有越零法(Zero Crossing Measure)與短期能量量測法(Short-Term Energy Measure)等[4]。由於輸入訊號的比對率高低與否跟切割區間的大小與區域極其相關，使用上述的方法雖可快速的找出音框中變化與能量相對較大的部分，但為確定所擷取出來的音訊具代表性與逼真性，並在時間與頻率上對頻譜圖同時做精確的比對，使用餘弦相似量測法(CSM)將是更好的作法[5, 6]。首先，我們依據時間與頻率框架間的頻譜結構來建構相似矩陣，此相似矩陣的每個元素分別由頻譜向量間取餘弦運算得出，再經由一個主對角區均為 1 與非對角區均為 -1 的區域矩陣沿著相似矩陣的對角線平移、遮罩與加總，因頻譜相異的轉換區間的主對角區多半被乘以 1，而非對角區均多半被乘以-1，相較於相似的區域(頻譜圖 RGB 值差異不大)的正負相消，最後可分隔出差異較大的音色，這也是我們用來切割頻譜圖的斷點。此法依據區域自我相似度標示出頻譜明顯變化處，並使得訊號切割過程對外在資訊的依賴降低。

三、虛擬 Hilbert 曲線

Hilbert曲線一種能遞迴地填充滿一個平面正方形的空間填充曲線[7]，此曲線目的是把一組2維資料轉換成1維資料，且其填充空間的順序是以目前所在位置的鄰近資料作為

優先編碼的依據。基於已有的頻譜圖，其需求是必須由上向下依循偵測，因Hilbert曲線受限於必須是一個邊長為2的N次方的正方形，所以我們把一個2維平面分割成多個小的正方形2維平面，而每個小正方形再做Hilbert曲線編碼。另外，由於頻譜圖的2維資料不會是剛好等長，所以我們設計了一個可以調整長寬的模型。這個模型一樣以Hilbert曲線為基準，但曲線繞法可依輸入資料的長寬比來改變。同時為了不失去Hilbert曲線本身的特性，資料長寬不足的部分我們以補"0"代替，其示意圖如下所示：



圖二、虛擬Hilbert曲線的繞行方式

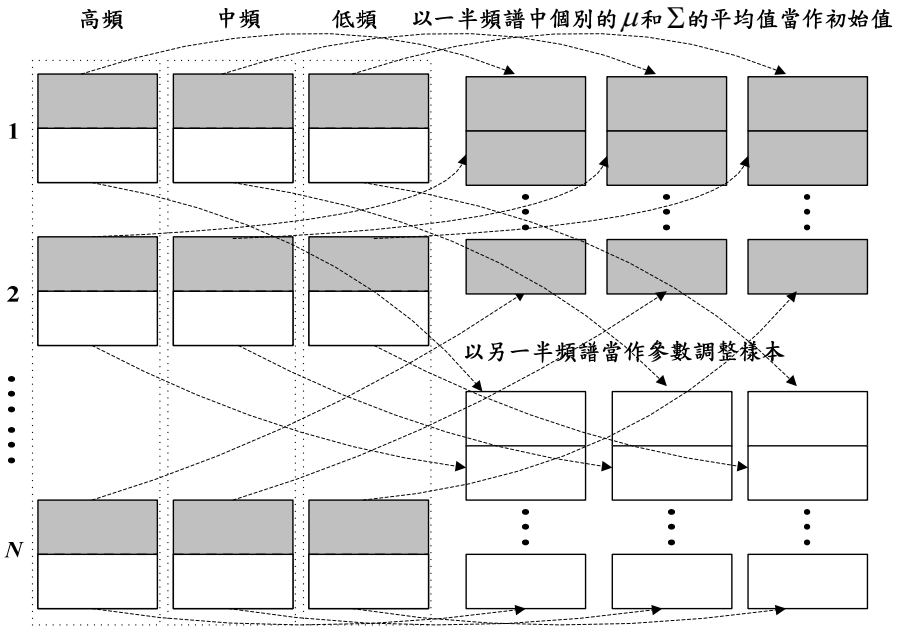
四、高斯混合模型

高斯混合模型(Gaussian mixture model, GMM)是單一高斯機率密度函數的延伸，由於GMM能夠平滑地近似任意形狀的密度分佈，因此近年來常被用在語音辨識上。假設存在一組多維空間的點x_i，$i=1,\cdots,n$，若這些點的分佈近似橢球狀，則我們可利用高斯密度函數$g(x;\mu,\Sigma)$來描述產生這些點的機率密度函數[8]：

$$g(x_i;\mu,\Sigma)=\frac{1}{\sqrt{(2\pi)^d\,|\Sigma|}}\exp\left(-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right)$$

其中d為維度，μ為此密度函數的均值，Σ則是此密度函數的共變異矩陣，這些參數決定了函數形狀的中心點及分佈情況。若$X=\{x_1,\cdots,x_n\}$彼此間互為獨立事件，則發生X的機率密度可表示為$p(X;\mu,\Sigma)=\Pi_{i=1}^{n}g(x_i;\mu,\Sigma)$。我們希望藉由已知的$X$找出使

$p(X;\mu,\Sigma)$ 為最大值之均值與共變異矩陣，經過推導，其最佳值分別是 $\hat{\mu} = \left(\Sigma_{i=1}^{n} x_i\right)/n$ 與 $\hat{\Sigma} = \left(\Sigma_{i=1}^{n}(x_i - \hat{\mu})(x_i - \hat{\mu})^T\right)/(n-1)$ [8]。由於我們將此模型運用在頻譜圖的分類上，其頻譜大小分佈情況不同，此單一高斯機率密度函數需改以 $p(x_i) = \Sigma_{j=1}^{3}\lambda_j g(x_i;\mu_j,\Sigma_j)$ 來表示，其中 $g(x_i;\mu_j,\Sigma_j)$ 分別代表高中低頻的高斯密度函數，配合權重和 $\Sigma_{j=1}^{3}\lambda_j = 1$ 的限制可進一步利用Lagrange Multiplier定義目標函數來進行求解。我們將此高斯混合模型運用在聲音識別上，假設有*N*組取樣音框，將每個音框依不同頻率區間等分成高、中、低頻三段，以一半頻譜中個別的 μ 和 Σ 的平均值當作是3組個別單一高斯機率密度函數的初始值，之後再以另一半頻譜當作是參數調整樣本來疊代求得 $\lambda_i, \mu_i, \Sigma_i$ 等參數[9]。此法基本上會讓所定義目標函數逐步遞增，但並無法保證此局部最大值就是全域最大值。

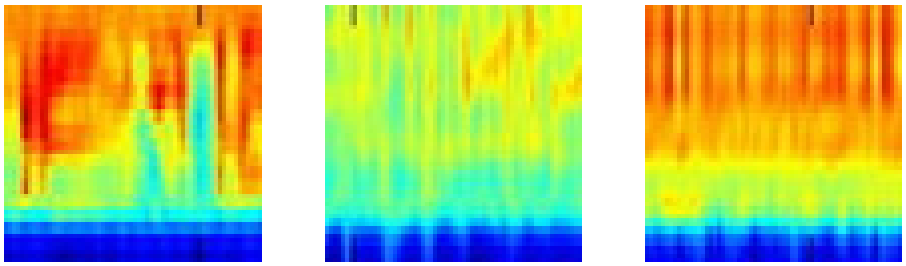


圖三、音框切割方式

叁、實驗結果

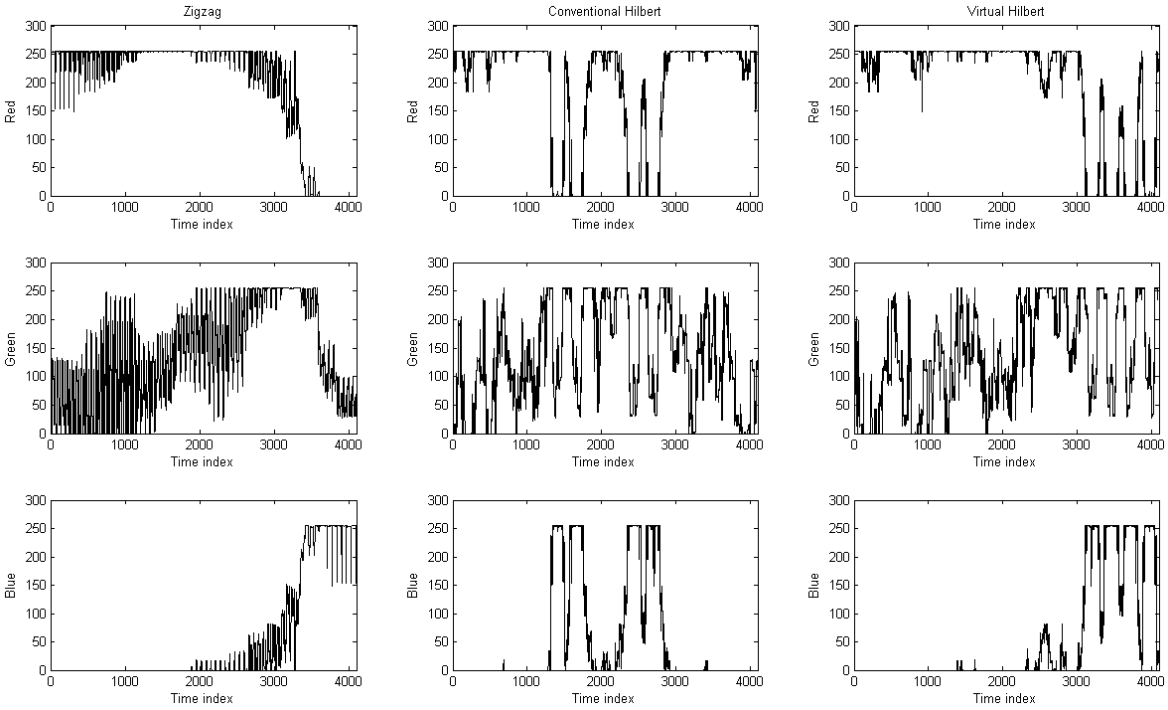
本實驗中，我們使用野雁聲、海豚聲與鯨魚聲三種不同的聲音來測試以上方法，圖四是這些測試訊號透過RASTA-PLP與CSM所得到的部分取樣音框。接下來我們使用三種偵測方法來驗證虛擬Hilbert曲線的效率，第一種為一般循序偵測法，作法為第一列由左至右依序掃描完後再換第二列，依此類推掃描完全圖。第二種方法為傳統的二維正方形Hilbert曲線繞法。第三種方式則是把整個正方形平均分割成16個小正方形，依頻譜特性依序繞完每個小正方形，其中每個小正方形則分別以Hilbert曲線法繞線。以上三種方

法的RGB編碼結果如圖五所示。



野雁聲　　海豚聲　　鯨魚聲

圖四、經由RASTA-PLP與CSM所得到的音框



圖五、循序偵測RGB編碼、一般Hilbert曲線RGB編碼與虛擬Hilbert曲線RGB編碼

由以上編碼結果可看出，一般Hilbert曲線無法滿足由上往下的區塊編碼，綠色訊號集中在中央部份而紅色訊號集中在兩端。而循序偵測和虛擬Hilbert曲線偵測較接近我們想要的結果，但循序偵測訊號值變動過大，連續性差，而虛擬Hilbert曲線偵測的連續性明顯較好，提升了下一階段頻譜分類的一致性。而依此方法產生的樣本進一步的藉由高斯混合模型進行參數調整，之後並隨機使用與系統參數疊代時不同的音源進行測試，藉此得知此訊號與系統中既有的對應資料相符性有多高，只要模型輸出的結果高於門檻值，則就可判斷此測試音是屬於哪一類物種。下表為不同測試音源於聲音識別系統之交叉比較所得出的數據，其相同物種的比對結果明顯得要比不同物種所得到的數據要高，顯示此系統在相對比較上亦具鑑別力。

表一、不同測試音源於聲音識別系統之實驗數據

識別系統 / 測試音源	野雁聲	海豚聲	鯨魚聲
野雁聲	**1.8166e-005**	3.3705e-006	5.2336e-006
海豚聲	1.4469e-006	**1.7927e-005**	3.7471e-007
鯨魚聲	6.4008e-006	2.6224e-007	**4.8887e-005**

肆、結論

假設我們將音訊的各個特徵視為它的座標，那麼我們就可以將全部的資料視為一群分佈在此高維度空間中的點，而每筆資料的特徵數目便可視為該筆資料的維度。在本文中，我們結合了可減低對資料庫依賴的相關頻譜轉換-感知線性預測法來表示音訊，並利用虛擬 Hilbert 曲線法來簡化頻譜的編碼，其主要目的是希望用較少的變數去解釋原來資料中的大部份變異，更期望能將我們手中許多相關性很高的變數轉化成彼此相依性較低的變數。最後使用高斯混合模型進行各種音源的參數擷取與調整，期能選取較原始變數個數少，但仍能解釋大部份資料中變異的新參數，在不減低聲音識別度下，使得測試音源與原始存放在資料庫中的序列特徵做比較時的運算複雜度與儲存空間都能降低。

伍、參考文獻

[1] Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*, 2nd Ed., Wiley-Interscience, 2000.

[2] Hynek Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech," Journal of the Acoustical Society of America, vol. 87, no. 4, pp. 1738-1752, Apr. 1990.

[3] Hynek Hermansky and Nelson Morgan, "RASTA Processing of Speech," IEEE Trans. Speech and Audio Processing, vol. 2, no. 4, pp. 578-589, Oct. 1994.

[4] John. R. Deller Jr., John H. L. Hansen, and John G. Proakis, *Discrete-Time Processing of Speech Signals*, IEEE Press, 1993.

[5] J. Foote, "Automatic audio segmentation using a measure of audio novelty," Proc. IEEE Intl. Conf. on Multimedia and Expo, vol. 1, pp. 452-455, 2000.

[6] M. Cooper and J. Foote, "Summarizing popular music via structural similarity analysis," IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 127-130, 2003.

[7] Benoit B. Mandelbrot, *The Fractal Geometry of Nature*, W. H. Freeman & Co, Sep. 1982.

[8] G. J. McLachlan and D. Peel, *Finite Mixture Models*, Wiley, 2000.

[9] V. V. Digalakis, D. Rtischev, and L. G. Neumeyer, "Speaker Adaptation Using Constrained Estimation of Gaussian Mixtures," IEEE Trans. Speech and Audio Processing, vol. 3, no. 5, pp. 357-366, Sep. 1995.