行政院國家科學委員會專題研究計畫　成果報告

應用生物聲紋特徵於自動化物種辨識之研究

計畫類別：個別型計畫
計畫編號：NSC94-2213-E-216-022-
執行期間：94 年 08 月 01 日至 95 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：李遠坤
計畫參與人員：蘇忠茂、林炳佑

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 95 年 10 月 27 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

應用生物聲紋特徵於自動化物種辨識之研究

計畫類別：☑ 個別型計畫　　□ 整合型計畫
計畫編號：NSC 94-2213-E-216-022-
執行期間：　　2005 年 08 月 01 日 至 2006 年 07 月 31 日

計畫主持人：李建興
共同主持人：李遠坤
計畫參與人員： 蘇忠茂、林炳佑

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　95　年　10　月　31　日

摘要

許多生物發出聲音之目的，主要是為了彼此之溝通或是某些特定行為所發出的聲音，像進食、移動、飛行、求偶和警戒等等，而生物聲紋自動辨識之研究對於生物學、生態學和環境監控應用上是相當重要的，特別是應用在偵測生物之種類和生物之分佈定位上。本計畫提出一套自動辨識生物聲紋特徵之系統，此系統可用以辨識蛙聲和蟋蟀聲。對於輸入一生物聲音，首先我們將此聲音之每一音節切取出來，然後對每一個音節分析其音色特徵，我們是以整個音節之平均線性倒頻譜係數(ALPCC)及梅爾倒頻譜係數(AMFCC)當做此音節之音色特徵向量。然後以線性區別分析演算法來提升辨識之正確率，此一演算法可以縮小同類特徵向量之距離而且加大不同種類特徵向量之距離，因此可以減少特徵向量維度之情況下提高辨識率。

一. 報告內容

1. 前言

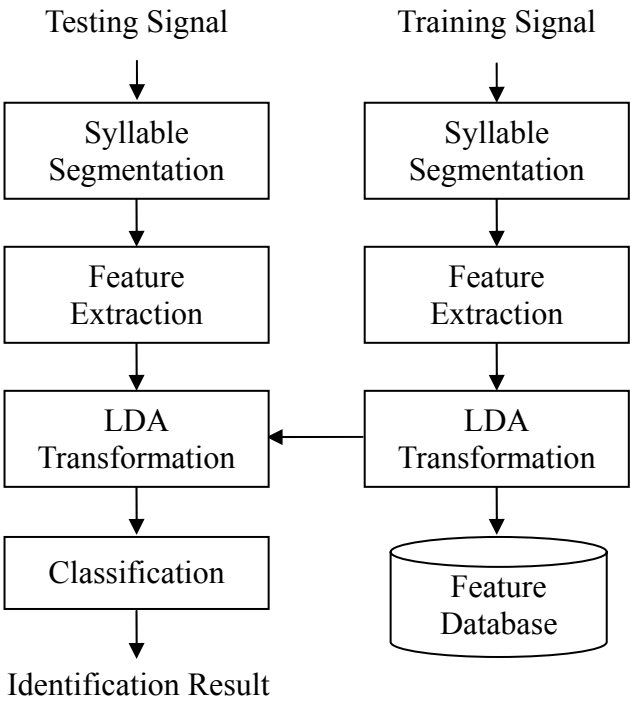
在日常生活中，我們能夠聽到許多生物的聲音，像是人類講話的聲音、狗吠聲、鳥鳴聲、蛙叫聲、蟬聲和蟋蟀聲等等。許多生物發出聲音之目的，主要是為了彼此之溝通或是某些特定行為所發出的聲音，像進食、移動、飛行、求偶和警戒等等，而生物聲紋自動辨識之研究對於生物學、生態學和環境監控應用上是相當重要的，特別是應用在偵測生物之種類和生物之分佈定位上。一般而言，人們只會聽到生物發出的聲音，較少看到生物本身，除此之外，生物的發聲(animal vocalizations)早已進化成與特定之物種相關(species-specific)，也就是不同之物種之聲音會有所不同，因此利用生物的聲音來辨識生物種類是非常自然不過的，另外還能夠應用在生物種類統計(ecological censusing)、環境監控(environment monitoring)和生物多樣性評估(biodiversity assessment)等方面。

通常在做生態調查時都是靠人力在白天用拍攝方式去做生態記錄，雖然現今科技進步，即使在晚上也可以利用紅外線拍攝，但是拍攝者必須花費很大的耐心和時間，且必須小心翼翼在不驚動到動物的情況下去做拍攝，但如果是記錄動物的聲音則不同，通常我們在野外容易聽得到動物的聲音，若想見其行蹤，則是難上加難，如果我們可以利用動物聲音之記錄來做生態評估，應該能省下許多的時間精力，且可以更有效的從事生態物種記錄，因此利用聲音去辨識物種做生態評估是近來常用的方法。由於生物種類繁多，不同物種間的棲息環境及生活方式也都有所差異，因此眾多研究人員投入研究生物叫聲的差異性，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，如此相當沒效率且不方便做大規模之調查，因此我們提出一套自

動辨識生物聲紋特徵之系統，用以辨識青蛙及蟋蟀之叫聲。

2. 研究目的與研究方法

本計劃之生物聲紋自動辨識系統包含兩個階段，分別為訓練階段(training phase)和辨識階段(recognition phase)，訓練階段是由三個主要模組所組成:音節切割(syllable segmentation)、特徵擷取(feature extraction)和線性區別分析(linear discriminant analysis, LDA)。辨識階段是由四個主要模組所組成: 音節分割、特徵擷取、線性區別分析轉換(LDA transformation)和分類(classification)。圖一為本計劃之系統架構圖。



圖一 生物聲紋自動辨織系統的架構圖

2.1 音節切割

當輸入一段生物聲音訊號時，首先將一個個音節切割出來，每一個音節視為辨識系統之基本的辨識單元。而選擇以音節當成辨識單元，是因為當所輸入聲音訊號同時有許多不同種類的動物聲音時，要切出一個個的音節是比較簡單的，除此之外，利用音節所擷取出來的特徵值較為穩定。在這裡我們是依據 Harma 之方法以頻率上的資訊來完成音節切割，其詳細步驟如下:

Step 1. 利用 short-time Fourier transform(STFT)建立輸入生物訊號的頻譜，我們用一個矩陣來表示此頻譜 $M(f, t)$，f 和 t 分別表示頻率值和時間的索引值。

Step 2. 設 n =0。

Step 3. 針對所有的(f, t)，找出振幅強度之最大值，即$|M(f_n, t_n)| \geq |M(f, t)|$，且將第 n 個音節的位置記錄為(f_n, t_n)。

Step 4. 求得$A_n(0) = 20\log_{10}|M(f_n, t_n)|$ dB，當$A_n(0) < A_0(0) - \beta$ dB，停止切割動作，這代表第 n 個音節的強度太小了因此就不需再切割出任何音節了，β為一門檻值(threshold)且其值設為 20。

Step 5. 從(f_n, t_n)的t_n位置找出$t < t_n$的$|M(f, t)|$最大值，直到$A_n(t - t_n) < A_n(0) - \beta$ dB，同時也尋找出$t > t_n$的$|M(f, t)|$最大值，直到$A_n(t - t_n) < A_n(0) - \beta$ dB。這個步驟是要找出第 n 個音節的起始時間$(t_n - t_s)$和結束時間$(t_n + t_e)$。

Step 6. 儲存第 n 個音節的時間點軌跡$A_n(\tau)$, 其中$\tau = t_n - t_s, ..., t_n + t_e$。

Step 7. 令$M(f, [t_n - t_s, ..., t_n + t_s]) = 0$，目的是要將第 n 個音節的部份給清空，並令 n=n+1 且回到 **Step 3** 尋找下一個音節。

2.2 特徵擷取

當每一動物聲音之音節已切割出來後，我們先將每一音節分為一個個部分重疊之音框(frame)，對每一個音框，我們以 LPCCs/MFCCs 為其特徵向量，而所有屬於同一音節中音框之 LPCCs/MFCCs 的平均值則視為此一音節之特徵向量，因此所有音節之特徵向量長度是固定的，與音節之長短無關。

(a)線性倒頻譜係數(Linear Predictive Cepstral Coefficients, LPCCs)

線性預測編碼 (linear predictive coding, LPC) 已經被廣泛的應用在語音辨識的技術上，可以說是語音分析技術上最重要的一項表示法，線性預測編碼的基本觀念，在於一個語音的訊號可以由前面 p 個語音訊號的線性組合來預測，而其做法是將預測的訊號值與實際的訊號值的誤差值減至最小，如此就可以得到最佳的預測器，而預測器內的係數即為線性組合所需的係數稱為線性預測編碼係數。

線性倒頻譜係數(Linear prediction cepstral coefficients, LPCCs)是一種較可靠之特徵而且已被證明比起線性預測編碼係數更能應用在語音辨識上，並且更能表現出語音訊號中頻譜的波峰與細部的變化。計算線性倒頻譜係數之詳細步驟如下:

Step 1. Preemphasis

使用一階的 FIR 濾波器:

$$\tilde{s}(n) = s(n) - \tilde{a}s(n-1)$$

其中$s(n)$為輸入的聲音訊號，$\tilde{a}$=0.95。

Step 2. Framing

將聲音訊號切割為一個個音框，其區段大小可為 256、512 或是 1024 個 sample，每次重疊一半的 sample 數

$$x_\ell(n) = \tilde{s}(M\ell + n), \qquad n = 0,\ 1,\ ...,\ N-1,\ \ \ell = 0,\ 1,\ ...,\ L-1$$

其中 ℓ 表示第 ℓ 個區段。

Step 3. Windowing

使用漢明視窗(Hamming window)消除每個區段中起始與終點訊號的不連續性

$$\tilde{x}_\ell(n) = x_\ell(n)w(n), \qquad 0 \le n \le N-1$$

漢明視窗公式如下：

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \qquad 0 \le n \le N-1$$

Step 4. Autocorrelation analysis

計算每個區段的自相關函數 (autocorrelation)，以便求取線性預測編碼的係數，求取自相關函數的公式如下:

$$r_\ell(m) = \sum_{n=0}^{N-1-m} \tilde{x}_\ell(n)\tilde{x}_\ell(n+m), \qquad m = 0,\ 1,\ ...,\ p \text{ a}$$

Step 5. LPC analysis

利用 Durbin Recursive Method 求得線性預測編碼係數，演算法如下:

$$E^{(0)} = r(0)$$

$$k_i = \left\{ r(i) - \sum_{j=1}^{i-1} \alpha_j^{(i-1)} r(|i-j|) \right\} \Big/ E^{(i-1)}, \qquad 0 \le i \le p$$

$$\alpha_i^{(i)} = k_i$$

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)}$$

$$E^{(i)} = (1 - k_i^2)E^{(i-1)}$$

$$a_m = \text{LPC coefficient} = \alpha_m^{(p)}, \qquad 1 \le m \le p$$

其中 p 為 LPC 係數的階數。

Step 6. LPC coefficients conversion to cepstral coefficients (LPCCs)

將線性預測編碼係數轉換為倒頻譜係數

$$c_0 = \ln \sigma^2$$

$$\sigma^2 = G^2 = r(0) - \sum_{k=1}^{p} \alpha_k r(k)$$

$$c_m = a_m + \sum_{k=1}^{m-1}\left(\frac{k}{m}\right)c_k a_{m-k}\ , \qquad 1 \le m \le p$$

其中 σ^2 為與語音發聲模型之增益值(gain term):

$$\sigma^2 = r[0] - \sum_{k=1}^{p} a_k r[k]$$

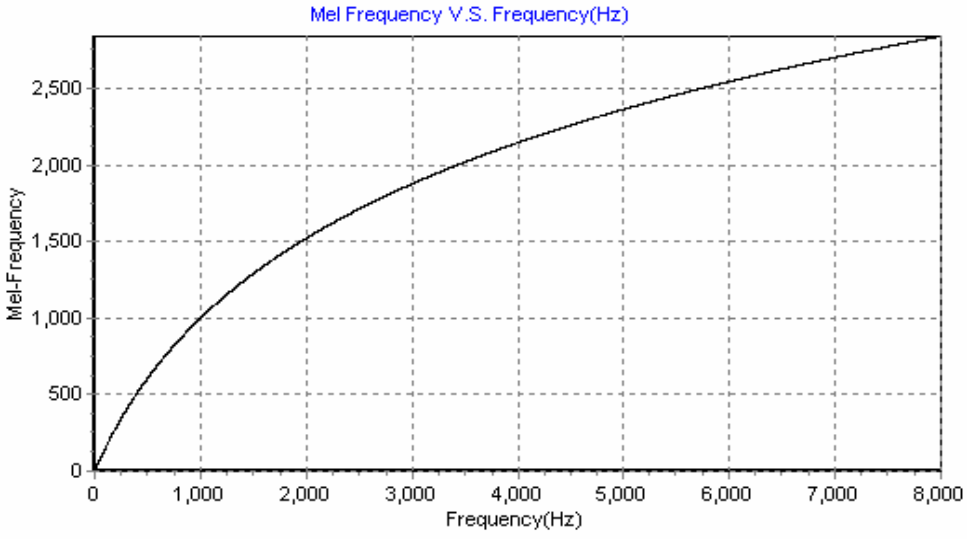
(b)梅爾倒頻譜係數（Mel-scale Frequency Cepstral Coefficients, MFCCs）

梅爾倒頻譜係數已被廣泛的利用在語音辨識上，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。mel 來是用以表示聽覺上對一個 tone 感覺上的的音高(pitch)或者是頻率的計算單位，在人類聽覺系統中，人類對於一個 tone 的實際頻率(physical frequency)之反應並不是呈現線性變化，而實際頻率和梅爾頻率之間的對應關係在頻率低於 1 KHz 時呈現線性變化，但在高頻的部份是呈現對數變化的。而實際頻率和梅爾頻率之間的關係圖顯示於圖二，其數學式如下:

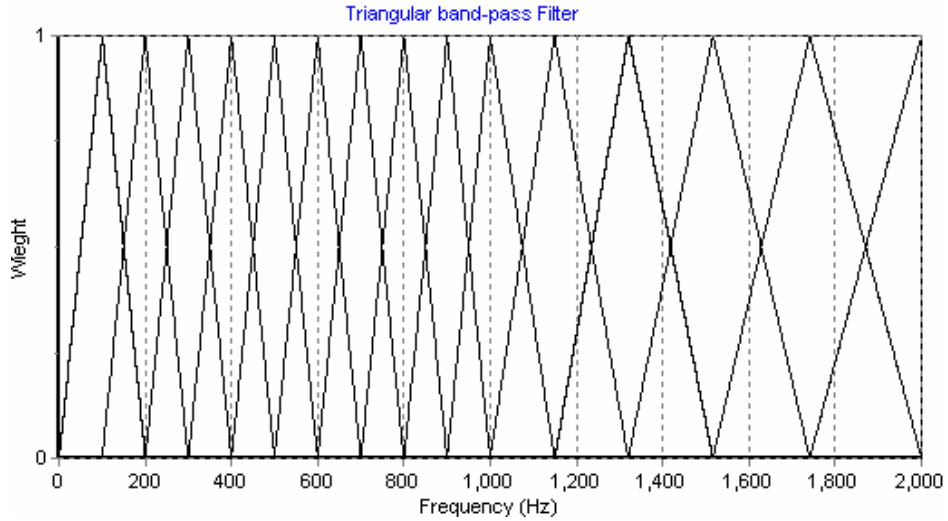
$$mel = 2595\log_{10}(1+\frac{f}{700}),$$

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

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聲音訊號，另外每一個臨界頻帶的頻寬會隨著頻率值而改變，圖三顯示一組三角臨界頻帶濾波器之形狀及頻寬，表一是每個臨界頻帶濾波器的頻寬範圍。



圖二 實際頻率和梅爾頻率之間的關係圖



圖三 一組三角臨界頻帶濾波器

表一 每個臨界頻帶濾波器的頻寬範圍

Index	Low Freg.(Hz)	Center Freg. (Hz)	High Freg. (Hz)
Filter 1	0	100	200
Filter 2	100	200	300
Filter 3	200	300	400
Filter 4	300	400	500
Filter 5	400	500	600
Filter 6	500	600	700
Filter 7	600	700	800
Filter 8	700	800	900
Filter 9	800	900	1000
Filter 10	900	1000	1149
Filter 11	1000	1149	1320
Filter 12	1149	1320	1516
Filter 13	1320	1516	1741
Filter 14	1516	1741	2000
Filter 15	1741	2000	2297
Filter 16	2000	2297	2639
Filter 17	2297	2639	3031
Filter 18	2639	3031	3482
Filter 19	3031	3482	4000
Filter 20	3482	4000	4595
Filter 21	4000	4595	5278
Filter 22	4595	5278	6063
Filter 23	5278	6063	6964
Filter 24	6063	6964	8000
Filter 25	6964	8000	9190

求取 MFCC 之步驟如下：

Step 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。

Step 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。

Step 3: 乘上漢明視窗(Hamming Windowing)

為了來消除每個音框與開始與結束的不連續性，每個音框都乘上一個漢明視窗，漢明視窗式子如下。

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

Step 4: 快速傅立葉轉換(FFT)

將音訊訊號從時域轉換成頻率域

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

N：音框大小，$\tilde{s}[n]$：離散訊號

Step 5: 三角帶通濾波器(Triangular band-pass filter)

由於人耳對聲音的頻率的解析度不是呈線性關係，而是呈現對數(logarithm)變化，利用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:

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

J 為三角帶通濾波器之個數，A_k 為 $X[k]$ 的振幅:

$$A_k = |X[k]|^2,\ 0 \le k < N/2$$

而 ϕ_j 為第 j 個濾波器:

$$\phi_j[k] = \begin{cases} 0 & \text{if } k \le I_l^j \text{ or } k \ge I_h^j \\ (k - I_l^j)/(I_c^j - I_l^j) & I_l^j \le k \le I_c^j \\ (I_h^j - k)/(I_h^j - I_c^j) & I_c^j \le k \le I_h^j \end{cases}$$

在這裡，I_l^j，I_c^j 和 I_h^j 分別代表第 j 個濾波器之低頻索引值，中間頻率索引值，和高頻索引值:

$$I_l^j = \frac{f_l^j}{(f_s/N)},$$

$$I_c^j = \frac{f_c^j}{(f_s/N)},$$

$$I_h^j = \frac{f_h^j}{(f_s/N)},$$

f_s 為取樣頻率，f_l^j，f_c^j，f_h^j 為第 j 個濾波器的低頻、中頻和高頻值，而每個濾波器的低頻、中頻和高頻值可參考表二。

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

最後將這些不同頻帶的能量乘上不同的 cosine 值，求出梅爾倒頻譜係數:

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

L 代表的是梅爾倒頻譜係數的個數。

我們共用了 25 個三角濾波器，所以 J=25，而梅爾倒頻譜係數的長度為 15(L=15)。

(c)Averaged LPCCs/MFCCs

如前所述，我們先把每一個音節切成一個個的音框，接著計算每個音框的 LPCCs/MFCCs 為其特徵值。可想而知，針對不同長度之音節所切割出來的音框個數會不盡相同，而為了解決這個問題，我們取這些音框的 LPCCs/MFCCs 的平均值來當做此一音節之特徵值，因此對所有的音節所取出的特徵值之長度是固定的，不會隨著音節的長度而變化。計算平均 LPCCs/MFCCs 的計算公式如下:

$$f_m = \frac{1}{K}\sum_{i=1}^{K} C_m^i, \qquad 0 \le m \le L-1$$

f_m 為第 m 個特徵值，K 是一個音節中的音框數量，C_m^i 為第 i 個音框的第 m 個特徵值，L 為特徵向量的長度。

在系統之訓練過程中，我們把屬於同一種動物聲音的所有訓練音節的特徵平均值來當做此一物種之特徵，計算公式如下:

$$F_m = E(f_m), \quad 0 \le m \le L-1$$

$E(\cdot)$ 指的是期望值。因為不同的特徵值，其範圍大小也不一樣，所以我們利用正規化來求出 $\hat{F}_m$ ，計算公式如下:

$$\hat{F}_m = \frac{F_m - f_m^{\min}}{f_m^{\max} - f_m^{\min}},$$

其中 $f_m^{\max}$ 和 $f_m^{\min}$ 為第 m 個特徵的最大值和最小值。

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

線性區別分析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，線性區別分析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。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:

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

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:

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

$\boldsymbol{\mu}$ 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 eigenvectors 來得到，而 A_{opt} 之 d 個行向量為前 d 個最大 eigenvalue 值所對應之 eigenvector。假設 eigenvalues 的順序為非遞增的，則所保留之 eigenvector 個數可由以下公式求得:

$$\sum_{i=1}^{d}\lambda_i \ge 0.95\sum_{i=1}^{n}\lambda_i,$$

λ_i 為第 i 個 eigenvalue。

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:

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

2.4 辨識階段

在辨識的部份中，在輸入每個聲音檔後，首先將每一個音節切割出來，並求出每個音節的 LPCCs/MFCCs 平均值。我們同時利用轉換矩陣 A_{opt} 來將經過正規化的 LPCCs/MFCCs 轉換成較低維度的特徵向量，接著，計算此一特徵向量和代表每一個生物種類之特徵向量之間的距離，在這裡的距離公式是歐基里德距離(Euclidean distance)，令 r 代表所辨識出來之生物種類:

$$r = \arg\min_{1 \le k \le C} \sum_{m=1}^{d} \left\| x_m - x_m^k \right\|,$$

C 為類別的個數，d 為特徵向量之維度，x_m 為輸入音節的第 m 個特徵值，x_m^k 為種類 k 的第 m 個特徵值。

3. 實驗結果與討論

在實驗所用動物聲音資料中，包含了兩個資料庫，分別有 30 種青蛙聲紋與 19 種蟋蟀聲紋，取樣頻率為 44100 Hz，音訊範圍大小為 16 bits。在對資料庫中所有動物聲音擷取特徵進行辨識前，我們以頻譜能量之資訊對每個動物聲音檔案切出音節，一半之聲音音節做為訓練音節，而另一半之聲音音節為測試音節，表二及表三分別列出各種青蛙與蟋蟀聲音檔案之訓練音節及測試音節數目。

表四及表五分別比較HMM, ALPCC,及AMFCC辨識30種青蛙聲音及19種蟋蟀聲音之正確率，由這兩個表中，我們可看出AMFCC之正確率比HMM及ALPCC高，而其對青蛙聲音及蟋蟀聲音之辨識正確率分別為96.2%及98.9%。

表二 青蛙聲音資料庫 (SC 為代碼)

SC	Scientific name (Popular name)	Ns
1	Bufo bankorensis (Central Formosan Toad)	38
2	Bufo melanosticus (Spectacled Toad)	48
3	Hyla chinensis (Chinese Tree Frog)	46
4	Microhyla butleri (Butler's Narrow-Mouthed Toad)	15
5	Microhyla heymonsi (Heymonsi's Narrow-Mouthed Toad)	235
6	Microhyla ornata (Ornate Narrow-Mouthed Toad)	193
7	Rana adenopleura (Olive Frog)	36
8	Rana catesbiana (American Bull Frog)	13
9	Rana guentheri (Guenther's Amoy Frog)	15
10	Rana kuhlii (Kuhli's Wart Frog)	52
11	Rana latouchii (Brown Wood Frog)	95
12	Rana limmocharis (Indian Rice Frog)	38
13	Rana rugulosa (Chinese Bull Frog)	98
14	Rana swinhoana (Swinhoe's Frog)	13
15	Rana sauteri (Sauter's Frog)	132
16	Rana taipehensis (Taipei Grass Frog)	27
17	Buergeria japonica (Japanese Tree Frog)	60
18	Buergeri robusta (Brown Tree frog)	67
19	Chirixalus eiffingeri (Eiffinger's Tree Frog)	10
20	Chirixalus idiootocus (Meintein Tree Frog)	112
21	Polypedates megacephalus (White lipped Tree Frog)	23
22	Rhacophorus arvalis (Farmland Tree Frog)	46
23	Rhacophorus Aurantiventris (Orange-Belly Tree Frog)	38
24	Rhacophorus moltrechti (Moltrecht's Tree Frog)	191
25	Rhacophorus prasinatus (Emerald Tree Frog)	52
26	Rhacophorus taipeianus (Taipei Green Tree Frog)	49
27	Microhyla steinegeri (Steinger's Narrow-Mouthed Toad)	3
28	Kaloula pulchra (Malaysian Narrow-Mouthed Toad)	6
29	Rana longicrus (Long-Legged Frog)	9
30	Rana psaltes (Harpist Frog)	65

表三 蟋蟀聲音資料庫 (SC 為代碼)

SC	Scientific name (Popular name)	Ns
1	Gryllotalpa fossor (Mole Cricket)	80
2	Teleogryllus occipitalis (Oil guord)	24
3	Teleogryllus mitratus (Oil guord)	3
4	Teleogryllus emma (Oil guord)	10
5	Gryllus bimaculatus (Painted Mirror)	6
6	Brachytrupes portentosus (Formosan Giant Crickets)	67
7	Loxoblemmus equestris (Coffin-headed cricket)	9
8	Dianemobius flavoantennalis (Flowered Bell)	22
9	Homoeogryllus japonicus (Horse bell)	3
10	Sclеropterus punctatus (Rocky bell)	7
11	Oecanthus longicaudus (Bamboo bell)	7
12	Xenogryllus marmoratus (Pagoda Bell)	6
13	Anaxipha pallidula (Yellow Bell)	58
14	Svistella bifasciatata (Golden Bell)	15
15	Homoeoxipha lycoides (Inky bell)	4
16	Mecopoda elongta L. (Weaving Lady)	74
17	Gryllotalpa fossor (Mole Cricket)	135
18	Teleogryllus occipitalis (Oil guord)	2
19	Teleogryllus mitratus (Oil guord)	5

表四 青蛙聲音之辨識正確率

SC	HMM	ALPCC	AMFCC
1	66%	89%	97%
2	98%	91%	100%
3	89%	97%	100%
4	40%	100%	100%
5	76%	71%	95%
6	82%	80%	97%
7	53%	58%	100%
8	54%	100%	100%
9	93%	100%	100%
10	50%	76%	100%
11	80%	66%	100%
12	68%	100%	100%
13	81%	97%	100%
14	62%	46%	38%
15	77%	87%	88%
16	63%	85%	92%
17	75%	100%	100%
18	96%	100%	97%
19	100%	100%	100%
20	99%	100%	100%
21	43%	60%	56%
22	15%	30%	73%
23	79%	68%	100%
24	94%	96%	99%
25	67%	96%	96%
26	94%	85%	100%
27	0%	100%	100%
28	0%	66%	100%
29	89%	88%	100%
30	98%	69%	100%
Average	78.9%	83.9%	96.2%

表五 青蛙聲音及蟋蟀聲音之辨識正確率

SC	HMM	ALPCC	AMFCC
1	90%	92%	98%
2	83%	100%	100%
3	100%	100%	100%
4	50%	20%	100%
5	50%	100%	100%
6	99%	97%	100%
7	89%	100%	88%
8	68%	86%	100%
9	33%	100%	100%
10	86%	85%	85%
11	71%	100%	100%
12	100%	100%	100%
13	93%	94%	100%
14	87%	93%	93%
15	100%	100%	100%
16	99%	98%	98%
17	100%	97%	100%
18	0%	100%	100%
19	20%	80%	80%
Average	91.2%	94.6%	98.9%

二. 參考文獻

[1] E. D. Chesmore, “Application of time domain signal coding and artificial neural networks to passive acoustical identification of animals,” *Applied Acoustics*, vol: 62, Issue 12, December, 2001.

[2] A. Harma,.” Automatic identification of bird species based on sinusoidal modeling of syllables,” *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, pp.v_545-v_548, 2003.

[3] J. A. Kogan and D. Margoliash, “Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models:A Comparative study,” *J. Acoust. Soc. Am.*, vol. 103, No. 4, April 1998.

[4] A. L. McIlraith and H. C. Card, “A comparison of backpropagation and statistical classifiers for bird identification,” *Proceedings of the International Conference on Neural Networks*, vol.1, pp. 100-104, vol. 1, June 1997.

[5] A. L. McIlraith and H. C. Card, "Birdsong recognition with DSP and neural networks," *Proceedings of the IEEE International Conference on Communications, Power, and Computing*, vol. 2, pp. 409-414, May 1995.

[6] A. L. McIlraith and H. C. Card, "Birdsong recognition using backpropagation and multivariate statistics," *IEEE Transactions on Signal Processing*, vol. 45, Issue 11, pp. 2740-2748, Nov. 1997.

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

[8] K. Sasaki and M. Yamazaki, "Vector compression of bird songs spectra in water sites by using the linear prediction method and its application to an automated Bayesian species classification," *Proceedings of the 38th Annual Conference on SICE Annual*, pp. 1083-1088, July 1999.

[9] C. Rogers, "High resolution analysis of bird sounds," *Proceedings of the 1995 International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, pp. 3011-3014, May 1995.

[10] A. Harma and M. Juntunen, "A method for parametrization of time-varying sounds," *IEEE Signal Processing Letters*, vol. 9, Issue 5, pp. 151-153, May 2002.

[11] 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, 1999, pp. 525 –532.

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

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

[14] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," *IEEE Trans. Speech Audio Processing*, Vol. 8, pp. 708–716, Nov. 2000.

[15] Cristianini, *An Introduction to Support Vector Machine*, Cambridge, 2000.

[16] L. Lu, S.Z. Li and H.J. Zhang, "Content-based audio segmentation using support vector machines," *Proceedings of the IEEE International Conference on Multimedia and Expo*, pp. 749-752, Aug. 2001.

[17] Hung Wei Ng, Y. Sawahata and K. Aizawa, "Summarization of wearable videos using support vector machine," *Proceedings of the IEEE International Conference on Multimedia and Expo*, pp. 325-328, Aug. 2002.

[18] S. I. Hill, P. J. Wolfe and P. J. W. Rayner, "Nonlinear perceptual audio filtering using support vector machines," *Proceedings of the 11th IEEE Signal Processing Workshop on Statistical*

Signal Processing, pp. 488-491, Aug. 2001.

[19] M. Davy and S. Godsill, "Detection of abrupt spectral changes using support vector machines: an application to audio signal segmentation," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 1313-1316, 2002.

[20] G. D. Guo and S. Z. Li, "Content-based audio classification and retrieval by support vector machines," *IEEE Transactions on Neural Networks*, vol. 14, Issue 1, pp. 209-215, Jan. 2003.

[21] Y. Liu, P. Ding and B. Xu, "Using nonstandard SVM for combination of speaker verification and verbal information verification in speaker authentication system," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. I-673-I-676, May 2002.

三. 計畫成果自評

建立生物多樣性資料庫是推動生物保育、教育及研究的重要基礎工作。於「生物多樣性公約」之第十七條即要求各國需成立生物多樣性資訊之交換中心，積極蒐集整理本土生物多樣性之資料，並與其他國家分享，以促進生物多樣性之保育、利用、管理、研究及教育，同時也可提振各國分類學的能力建設。

台灣的土地面積雖不大，卻擁有異常豐富的生物多樣性資源，特有生物種類繁多，臺灣已列入正式紀錄的鳥類約有 450 種，青蛙種類約有 31 種，蟬類有 59 種(4 種新種)，台灣蟋蟀種類約有八十幾種，以聲音自動辨識系統來記錄生物的棲息環境，不僅有助於了解這些生物的生態變化，並能減少對生態的影響。生物聲音辨識的研究，尤其是國內，進行的仍舊很少，希望透過此自動辨識系統配合適當的硬體設備，能發現更多未曾記載的生物物種，建立更完善的台灣生物聲音資料庫。本計畫已完成可自動辨識青蛙及蟋蟀聲紋之辨識系統，未來希望能進一步辨識較複雜多變化之鳥類鳴聲。

目前我們已發表三篇相關論文，包括一篇期刊論文及兩篇研討會論文：

期刊論文 (Journal Papers) ：

[1] 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)

研討會論文 (Conference Papers) ：

[1] 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.

[2] 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.

以下為發表之期刊論文的全文

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Pattern Recognition Letters 27 (2006) 93–101
Pattern Recognition Letters
www.elsevier.com/locate/patrec

Automatic recognition of animal vocalizations using averaged MFCC and linear discriminant analysis

Chang-Hsing Lee [a,*], Chih-Hsun Chou [a], Chin-Chuan Han [b], Ren-Zhuang Huang [a]

[a] Department of Computer Science and Information Engineering, Chung Hua University, Hsinchu 300, Taiwan, ROC
[b] Department of Computer Science and Information Engineering, National United University, Miao-Li 360, Taiwan, ROC

Received 13 August 2004; received in revised form 30 June 2005
Available online 31 August 2005

Communicated by O. Siohan

Abstract

In this paper we propose a method that uses the averaged Mel-frequency cepstral coefficients (MFCCs) and linear discriminant analysis (LDA) to automatically identify animals from their sounds. First, each syllable corresponding to a piece of vocalization is segmented. The averaged MFCCs over all frames in a syllable are calculated as the vocalization features. Linear discriminant analysis (LDA), which finds out a transformation matrix that minimizes the within-class distance and maximizes the between-class distance, is utilized to increase the classification accuracy while to reduce the dimensionality of the feature vectors. In our experiment, the average classification accuracy is 96.8% and 98.1% for 30 kinds of frog calls and 19 kinds of cricket calls, respectively.

MSC: 140.000

Keywords: Linear discriminant analysis; Mel-frequency cepstral coefficients

1. Introduction

Many animals generate sounds either for communication or as a by-product of their living activities such as eating, moving, or flying. Automatic recognition of bioacoustic sounds is valuable for applications such as biological research and environmental monitoring; this is particularly true for detecting and locating animals. In our daily life, we often hear the animal vocalizations rather than see the animals. In general, the animals generate sounds to communicate with members of the same species and thus the animal vocalizations have evolved to be species-specific. Therefore, identifying animal species from their vocalizations is valuable to ecological censusing.

In general, the acoustic signal representing animal vocalizations can be regarded as a sequence of syllables. Thus, a better way to identify animals from their vocalizations is to use a syllable as the acoustic component. It is necessary to segment the syllables of animal vocalizations before the recognition process. Segmentation of speech or audio signals is often based on energy (Lamel et al., 1981; Li et al., 2001; Lu, 2001; Wold et al., 1996; Zhang and Kuo, 2001) and/or zero-crossing rate (Li et al., 2001; Lu, 2001; Tian et al., 2002; Wold et al., 1996; Zhang and Kuo, 2001). A disadvantage of using these segmentation methods to extract syllables from animal vocalizations is that the full syllable cannot be extracted exactly. To overcome this problem, we exploit the frequency information to segment the syllables of animal vocalizations (Harma, 2003).

* Corresponding author. Tel.: +886 3 5186406; fax: +886 3 5186416.
E-mail addresses: chlee@chu.edu.tw (C.-H. Lee), chc@chu.edu.tw (C.-H. Chou), cchan@nuu.edu.tw (C.-C. Han).

0167-8655/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2005.07.004

Once the syllables have been properly segmented, a set of features will be calculated to represent each syllable. The most well-known features for speech/speaker recognition are linear predictive coefficients (LPCs) (Rabiner and Juang, 1993) or Mel-frequency cepstral coefficients (MFCCs) (Picone, 1993; Rabiner and Juang, 1993; Vergin et al., 1999). In this paper, we use the averaged MFCCs in a syllable to identify animals from their sounds due to the fact that MFCCs can represent the spectrum of animal sounds in a compact form. In the next section, we will describe the proposed recognition method for animal vocalizations.

2. The proposed recognition method for animal vocalizations

The recognition system consists of two parts: the training part and the recognition part. The training part is composed of three main modules: syllable segmentation, averaged MFCCs extraction, and linear discriminant analysis (LDA). The recognition part consists of four modules: syllable segmentation, averaged MFCCs extraction, LDA transformation, and classification. A detailed description of each module will be described below.

2.1. Syllable segmentation

The input acoustic signal is first segmented into a set of syllables (Harma, 2003). Each syllable is regarded as the basic acoustic unit for recognition. The syllable segmentation method based on the frequency information is described as follows:

Step 1. Compute the spectrogram of the input bioacoustic signal using short-time Fourier transform (STFT). We denote the spectrogram a matrix $S(f,t)$, where f represents frequency index and t is the frame index.

Step 2. Set $n = 0$.

Step 3. Find f_n and t_n, such that $|S(f_n,t_n)| \geqslant |S(f,t)|$, for every pair of (f,t). Set the position of the nth syllable to be (f_n,t_n).

Step 4. Compute the amplitude $A_n(0) = 20\log_{10}|S(f_n,t_n)|$ dB and set the frequency parameter $W_n(0) = f_n$. If $A_n(0) < A_0(0) - 20$ dB, stop the segmentation process. This means that the amplitude of the nth syllable is too small and hence no more syllables need to be extracted.

Step 5. Starting from (f_n,t_n), trace the maximal peak of $|S(f,t)|$ for $t < t_n$ until $A_n(t) < A_n(0) - ß$dB, where ß is the stopping criteria and its default value is 20. Next, trace the maximal peak of $|S(f,t)|$ for $t > t_n$ until $A_n(t) < A_n(0) - ß$dB. The step is to determine the starting time $(t_n - t_s)$ and the ending time $(t_n + t_e)$ of the nth syllable around t_n.

Step 6. Store the amplitude trajectories corresponding to the nth syllable in function $A_n(\tau)$, where $\tau = t_n - t_s, \ldots, t_n + t_e$.

Step 7. Set $S(f,[t_n - t_s, \ldots, t_n + t_e]) = 0$ to delete the area of nth syllable. Set $n = n + 1$ and goto Step 3 to find the next syllable.

Fig. 1 shows the waveform of Olive Frog (*Rana adenopleur*) as well as the segmentation results by using the energy information and the spectrogram frequency information. It is evident that a better result can be obtained. After segmenting each syllable, the averaged MFCCs are extracted to represent the syllable.

2.2. Averaged MFCCs extraction

MFCCs have been the most widely used features for speech recognition (Picone, 1993; Rabiner and Juang, 1993; Vergin et al., 1999), bird song recognition (Kogan and Margoliash, 1998), and audio retrieval (Slaney, 2002) due to their ability to represent the signal spectrum in a com-

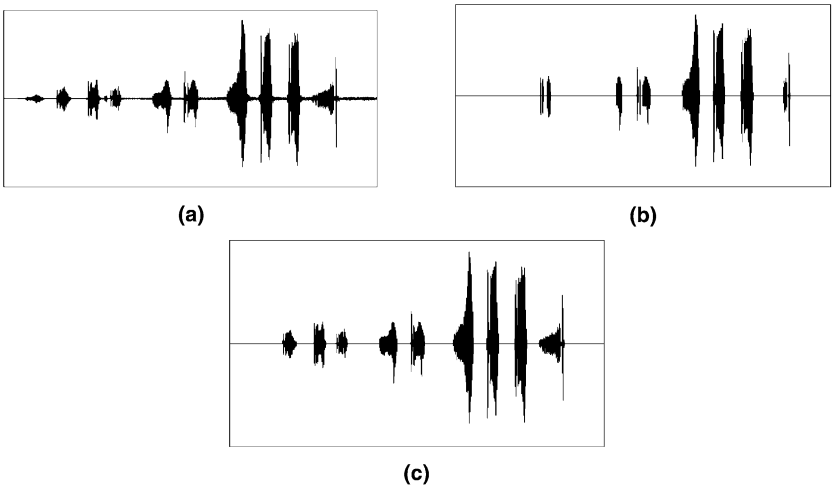


Fig. 1. (a) The waveform of Olive Frog (*Rana adenopleur*). (b) The segmentation result by using the energy information. (c) The segmentation result by using the spectrogram frequency information.

pact form. In fact, the MFCCs have been proven to be very effective in automatic speech recognition or in modeling the subjective frequency content of audio signals. In general, an input signal is first divided into a set of frames. The MFCCs for each frame are then computed and are regarded as the features of this frame. However, the number of frames varies for different syllables. To deal with this problem, the averaged MFCCs of all the frames in a syllable are computed and used as features to represent the syllable. Therefore, the number of features is fixed regardless of the length of the acoustic syllable. A detailed description for deriving the MFCCs of an acoustic signal is given as follows:

Step 1. *Pre-emphasis*.

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

where $s[n]$ is the signal denoting the input syllable, a typical value for $\hat{a}$ is 0.95.

Step 2. *Framing*.Each syllable is divided into a set of overlapped frames with frame size of N samples, and the overlapping size is M samples for each pair of successive frames. Therefore, consecutive frames will never change too much. In our experiments, N is 512 and M is 256.

Step 3. *Windowing*.To reduce the discontinuity on both ends of a frame, each frame is multiplied by a Hamming window

$$\tilde{s}[n] = \hat{s}[n]w[n], \quad 0 \leqslant n \leqslant N-1, \tag{2}$$

where $w[n]$ is the Hamming window function

$$w[n] = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \leqslant n \leqslant N-1. \tag{3}$$

Step 4. *Spectral analysis*.Take the discrete Fourier transform of each frame using FFT

$$X[k] = \sum_{n=0}^{N-1} \tilde{s}[n]\mathrm{e}^{-j2\pi\frac{k}{N}n}, \quad 0 \leqslant k \leqslant N-1. \tag{4}$$

Step 5. *Band-pass filtering*.The amplitude spectrum is then filtered using a set of triangular band-pass filters

$$E_j = \sum_{k=0}^{N/2-1} \phi_j(k)A_k, \quad 0 \leqslant j \leqslant J-1, \tag{5}$$

where J is the number of filters, ϕ_j is the jth filter, and A_k is the amplitude of $X[k]$

$$A_k = |X[k]|^2, \quad 0 \leqslant k < N/2. \tag{6}$$

Step 6. *DCT*.The MFCCs for the ith frame are computed by performing DCT on the logarithm of E_j

$$C_m^i = \sum_{j=0}^{J-1} \cos\left(m\frac{\pi}{J}(j+0.5)\right)\log_{10}(E_j), \quad 0 \leqslant m \leqslant L-1, \tag{7}$$

where L is the number of MFCCs.

In the proposed method, the filter bank consists of 25 triangular filters, that is, $J = 25$. The length of MFCCs feature vector for each frame is 15 ($L = 15$). After deriving the MFCCs for each frame, we compute the averaged MFCCs of all frames within the syllable

$$f_m = \frac{\sum_{i=1}^{K} C_m^i}{K}, \quad 0 \leqslant m \leqslant L-1, \tag{8}$$

where f_m is the mth MFCC, K is the number of frames within the syllable, and C_m^i denotes the mth MFCC of the ith frame. In the training phase, the averaging of f_m over all training syllables for the acoustic vocalization of the same species is regarded as the mth feature value, F_m. Since the dynamic ranges of f_m's may be different, we perform a linear normalization process to get the final feature vector F'_m

$$F'_m = \frac{F_m - f_m^{\min}}{f_m^{\max} - f_m^{\min}}, \tag{9}$$

where $f_m^{\max}$ and $f_m^{\min}$ denote the maximum and minimum values of the mth MFCC of all f'_ms for the training syllables, respectively.

From Fig. 1, it seems that each syllable has different sound structure. Fig. 2 shows that the spectrograms of these syllables look similar except that of the last syllable. Table 1 shows the feature vectors extracted from these

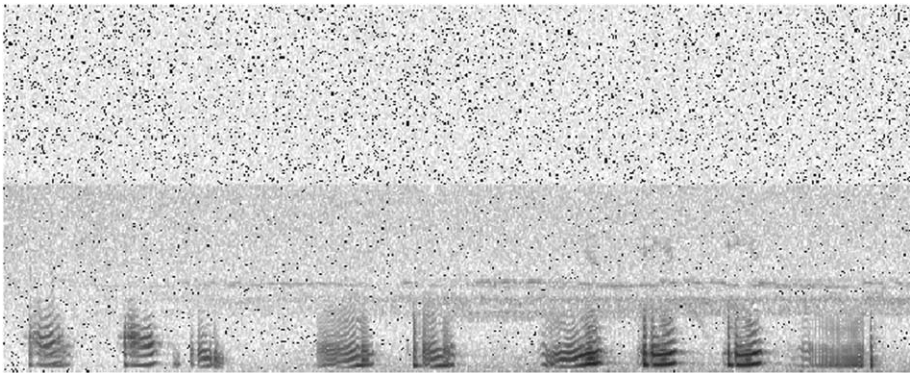


Fig. 2. Spectrogram for the example Olive Frog (*Rana adenopleur*).

Table 1
Feature vectors for the extracted syllables (SC denotes the subject code)

SC	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9
f_1	0.627	0.667	0.606	0.713	0.727	0.791	0.843	0.829	0.680
f_2	0.680	0.684	0.784	0.768	0.784	0.793	0.783	0.801	0.782
f_3	0.290	0.242	0.342	0.325	0.361	0.310	0.283	0.296	0.511
f_4	0.404	0.456	0.241	0.309	0.305	0.256	0.264	0.270	0.284
f_5	0.368	0.324	0.504	0.463	0.416	0.467	0.417	0.383	0.455
f_6	0.156	0.098	0.418	0.406	0.346	0.462	0.291	0.303	0.439
f_7	0.273	0.250	0.266	0.251	0.162	0.260	0.183	0.154	0.258
f_8	0.457	0.459	0.487	0.536	0.467	0.511	0.510	0.493	0.550
f_9	0.592	0.617	0.351	0.512	0.439	0.562	0.760	0.734	0.478
f_{10}	0.809	0.915	0.641	0.662	0.699	0.698	0.873	0.908	0.676
f_{11}	0.671	0.658	0.420	0.490	0.577	0.456	0.466	0.485	0.475
f_{12}	0.506	0.479	0.653	0.661	0.686	0.609	0.426	0.463	0.727
f_{13}	0.523	0.357	0.559	0.541	0.501	0.511	0.437	0.409	0.525
f_{14}	0.451	0.477	0.390	0.453	0.337	0.453	0.450	0.462	0.396
f_{15}	0.593	0.614	0.410	0.418	0.491	0.425	0.467	0.483	0.482

Table 2
Distance between each feature vector extracted from the example syllables and the 30 representative feature vectors with the minimum distance highlighted

SC	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9
1	0.7415	0.8188	0.6039	0.5936	0.6587	0.5493	0.7214	0.7292	0.6717
2	0.8217	0.9170	0.6944	0.6371	0.7520	0.6797	0.8420	0.8740	0.7215
3	0.7595	0.8941	0.8354	0.8194	0.8182	0.8604	0.9300	0.9503	0.8349
4	0.6782	0.7808	0.7048	0.6665	0.7097	0.6718	0.7571	0.7804	0.7471
5	0.6993	0.7981	0.6162	0.5661	0.6487	0.5693	0.7071	0.7289	0.5668
6	0.8232	0.8872	0.6899	0.6597	0.7372	0.6684	0.8217	0.8341	0.6709
7	0.3099	0.3917	0.4410	0.2813	0.2748	0.2685	0.2748	0.2622	0.3777
8	0.9130	1.0213	0.6975	0.6980	0.7500	0.6806	0.8556	0.8842	0.6811
9	0.8558	0.9303	0.6713	0.6936	0.7061	0.6909	0.8221	0.8495	0.7285
10	0.9542	1.0621	0.7159	0.7853	0.7150	0.7850	0.9780	0.9661	0.7123
11	0.7517	0.8106	0.5296	0.5299	0.6268	0.5263	0.7261	0.7387	0.6011
12	1.0211	1.1240	0.8332	0.8651	0.9174	0.8173	0.9882	1.0073	0.9110
13	0.6558	0.7210	0.6242	0.5835	0.6120	0.5959	0.6438	0.6716	0.6317
14	0.6947	0.8062	0.6483	0.6258	0.6534	0.6830	0.8248	0.8373	0.6077
15	0.5267	0.5939	0.6179	0.5592	0.5456	0.6100	0.7183	0.7158	0.5603
16	0.7915	0.8734	0.7215	0.6223	0.7144	0.6161	0.7501	0.7792	0.6388
17	0.6448	0.7294	0.6343	0.5876	0.6302	0.6157	0.7493	0.7423	0.5883
18	0.6568	0.7204	0.5167	0.5368	0.5833	0.5677	0.7192	0.7357	0.6165
19	1.2702	1.3241	1.3281	1.3180	1.3344	1.3571	1.4431	1.4364	1.2818
20	1.1940	1.2704	1.2935	1.3012	1.3162	1.3569	1.4272	1.4301	1.2355
21	0.6753	0.7434	0.4479	0.4495	0.5182	0.4171	0.5986	0.6129	0.5046
22	0.8649	0.9368	0.7933	0.7902	0.8599	0.8153	0.9374	0.9430	0.7289
23	0.7865	0.8738	0.7021	0.6580	0.6753	0.6612	0.7866	0.7947	0.5952
24	1.0395	1.1656	0.7796	0.8132	0.9295	0.7812	0.9507	0.9850	0.8778
25	0.7724	0.8922	0.7335	0.6541	0.7303	0.6930	0.8256	0.8609	0.6843
26	0.8508	0.9501	0.6928	0.6290	0.7261	0.6534	0.8123	0.8423	0.6436
27	1.0673	1.0742	0.9840	0.8746	1.0122	0.8696	1.0030	0.9894	0.9164
28	0.9436	1.0374	0.7056	0.7272	0.7848	0.7024	0.8491	0.8737	0.7178
29	0.7372	0.8596	0.5467	0.5054	0.6029	0.5086	0.7217	0.7424	0.4600
30	0.9759	1.0893	0.8906	0.8582	0.9431	0.8531	0.9991	1.0190	0.7676

syllables. From this table, we can see that these feature vectors are very close. Table 2 shows the distance between each feature vector extracted from these syllables and the representative feature vectors corresponding to the 30 kinds of frogs. From this table, we can see that the correct frogs (indexed with subject code 7) can be identified using Euclidean distance between two feature vectors.

2.3. Linear discriminant analysis (LDA)

LDA (Baker, 2004; Duda et al., 2000; Slaney, 2002) aims at improving the classification accuracy at a lower-dimensional feature vector space. LDA deals with discrimination between classes. The goal of LDA is to minimize the within-class distance while to maximize the between-class distance.

In LDA, an optimal transformation matrix from an n-dimensional feature space to a d-dimensional space is determined. The transformation matrix is a linear mapping that maximizes the so-called Fisher criterion J_F

$$J_F(A) = \mathrm{tr}((A^T S_W A)^{-1}(A^T S_B A)). \quad (10)$$

Here S_W and S_B are the within-class scatter matrix and between-class scatter matrix, respectively. The within-class scatter matrix is defined as

$$S_W = \sum_{j=1}^{C} \sum_{i=1}^{N_j} (X_i^j - \mu_j)(X_i^j - \mu_j)^T, \quad (11)$$

where X_i^j is the ith feature vector of class j, μ_j is the mean vector of class j, C is the number of classes, and N_j is the number of feature vectors in class j. The between-class scatter matrix is given by

$$S_B = \sum_{j=1}^{C} (\mu_j - \mu)(\mu_j - \mu)^T, \quad (12)$$

where μ is the mean vector of all classes.

From Eq. (10), we can see that LDA tries to find a transformation matrix that maximizes the ratio of between-class scatter to within-class scatter in a lower-dimensional space. The optimal solution of Eq. (10) is the transformation matrix, $\boldsymbol{A}_{opt}$, given by

$$\boldsymbol{A}_{opt} = \arg\max_{\boldsymbol{A}} \frac{\mathrm{tr}(A^T S_B A)}{\mathrm{tr}(A^T S_W A)}. \quad (13)$$

The transformation matrix $\boldsymbol{A}_{opt}$ can be determined by finding the eigenvectors of $S_W^{-1} S_B$.

2.4. Automatic recognition of animal vocalizations

At the recognition part, each input acoustic signal is first segmented into a set of syllables. The averaged MFCCs for each syllable are calculated. The same linear normalization process is applied to each MFCC. The normalized MFCCs are transformed to a lower-dimensional feature vector by the transformation matrix $\boldsymbol{A}_{opt}$ derived by LDA. For every species, the distance between the feature vector of the test syllable and the feature vector representing this species is calculated. The one with minimum distance is regarded as the identified species. In this paper, the distance between two feature vectors is the Euclidean distance. That is, the subject code that represents the identified species is determined by

$$r = \arg\min_{1 \leqslant k \leqslant C} \sum_{m=1}^{d} (F'_m - F_m^k)^2, \quad (14)$$

where C is the number of classes, d is the dimension of the feature vector, F'_m is the mth feature of the testing syllable,

Table 3
Frog call database (SC denotes the subject code)

SC	Scientific name (popular name)	Ns (Experiment 1)	Ns (Experiment 2)
1	*Bufo bankorensis* (Central Formosan Toad)	38	38
2	*Bufo melanosticus* (Spectacled Toad)	47	48
3	*Hyla chinensis* (Chinese Tree Frog)	45	46
4	*Microhyla butleri* (Butler's Narrow-Mouthed Toad)	15	15
5	*Microhyla heymonsi* (Heymonsi's Narrow-Mouthed Toad)	234	235
6	*Microhyla ornata* (Ornate Narrow-Mouthed Toad)	192	193
7	*Rana adenopleura* (Olive Frog)	35	36
8	*Rana catesbiana* (American Bull Frog)	12	13
9	*Rana guentheri* (Guenther's Amoy Frog)	15	15
10	*Rana kuhlii* (Kuhli's Wart Frog)	52	52
11	*Rana latouchii* (Brown Wood Frog)	95	95
12	*Rana limmocharis* (Indian Rice Frog)	38	38
13	*Rana rugulosa* (Chinese Bull Frog)	97	98
14	*Rana swinhoana* (Swinhoe's Frog)	13	13
15	*Rana sauteri* (Sauter's Frog)	131	132
16	*Rana taipehensis* (Taipei Grass Frog)	27	27
17	*Buergeria japonica* (Japanese Tree Frog)	60	60
18	*Buergeri robusta* (Brown Tree Frog)	66	67
19	*Chirixalus eiffingeri* (Eiffinger's Tree Frog)	10	10
20	*Chirixalus idiootocus* (Meintein Tree Frog)	112	112
21	*Polypedates megacephalus* (White lipped Tree Frog)	22	23
22	*Rhacophorus arvalis* (Farmland Tree Frog)	46	46
23	*Rhacophorus aurantiventris* (Orange-Belly Tree Frog)	38	38
24	*Rhacophorus moltrechti* (Moltrecht's Tree Frog)	190	191
25	*Rhacophorus prasinatus* (Emerald Tree Frog)	52	52
26	*Rhacophorus taipeianus* (Taipei Green Tree Frog)	49	49
27	*Microhyla steinegeri* (Steinger's Narrow-Mouthed Toad)	2	3
28	*Kaloula pulchra* (Malaysian Narrow-Mouthed Toad)	6	6
29	*Rana longicrus* (Long-Legged Frog)	9	9
30	*Rana psaltes* (Harpist Frog)	65	65

Table 4
Cricket call database (SC denotes the subject code)

SC	Scientific name (popular name)	Ns (Experiment 1)	Ns (Experiment 2)
1	*Gryllotalpa fossor* (Mole Cricket)	79	80
2	*Teleogryllus occipitalis* (Oil guord)	24	24
3	*Teleogryllus mitratus* (Oil guord)	3	3
4	*Teleogryllus emma* (Oil guord)	9	10
5	*Gryllus bimaculatus* (Painted Mirror)	5	6
6	*Brachytrupes portentosus* (Formosan Giant Crickets)	66	67
7	*Loxoblemmus equestris* (Coffin-headed cricket)	9	9
8	*Dianemobius flavoantennalis* (Flowered Bell)	22	22
9	*Homoeogryllus japonicus* (Horse bell)	2	3
10	*Scleropterus punctatus* (Rocky bell)	6	7
11	*Oecanthus longicaudus* (Bamboo bell)	6	7
12	*Xenogryllus marmoratus* (Pagoda Bell)	5	6
13	*Anaxipha pallidula* (Yellow Bell)	57	58
14	*Svistella bifasciatata* (Golden Bell)	14	15
15	*Homoeoxipha lycoides* (Inky bell)	4	4
16	*Mecopoda elongta* L. (Weaving Lady)	74	74
17	*Gryllotalpa fossor* (Mole Cricket)	135	135
18	*Teleogryllus occipitalis* (Oil guord)	1	2
19	*Teleogryllus mitratus* (Oil guord)	5	5

F_m^k is the mth feature of the kth species, and r is the subject code for the rth species as shown in Tables 3 and 4.

3. Experimental results

Two audio databases of 30 frog calls and 19 cricket calls derived from compact disk are used for the experiments (see Tables 3 and 4). The sampling frequency is 44,100 Hz and each sample is digitized in 16 bits. Most of the calls are field recordings with additional sounds in the background. Some of the calls are generated by multiple individuals vocalizing simultaneously. Each acoustic signal is first segmented into a set of syllables, in which half is used for training and half for testing. Two scenarios for dividing the syllables equally into training and testing sets are conducted: (1) the syllables are alternately divided into training and testing sets, that is, the odd-numbered syllables are used for training whereas the even-numbered syllables for testing; (2) the first-half of the syllables are used for training and the second half of the syllables are used for testing. In the second scenario, the training and testing syllables may be extracted from the calls of different frogs/crickets. The classification accuracy (CA) is defined as

$$\mathrm{CA} = \frac{N_{\mathrm{CA}}}{N_{\mathrm{S}}} \times 100, \tag{15}$$

where N_{CA} is the number of syllables which were recognized correctly and N_{S} is the total number of test syllables (shown in Tables 3 and 4).

3.1. Experiment 1—alternate training and testing

In this experiment, the syllables are alternately divided into training and testing sets. The odd-numbered syllables are used for training whereas the even-numbered syllables for testing. Tables 5 and 6 compare the recognition results of the proposed averaged MFCC (AMFCC) method with HMM and averaged LPC (ALPC) for 30 frog calls and 19 cricket calls, respectively. In addition, multiple feature vector templates were compared in these two tables, where LPC-i and MFCC-i denote that i ($i > 1$) vector templates are used to model the syllables derived from the same species. From these two tables, we can see that AMFCC greatly outperforms HMM and ALPC. HMM, which exploits the temporal information among the frames within a syllable, does not provide better performance than ALPC or AMFCC. Since variations between consecutive frames within a syllable are not regular and thus temporal information extracted for identification purpose is not very essential. In addition, most of the sounds are recorded in the field with additional sounds/noise in the background. Thus, the feature vector extracted from each syllable is not so stable. On the other hand, the averaged LPC/MFCC can attenuate the effect of background noise by averaging the feature vectors of all frames within the syllable. Therefore, the proposed AMFCC is adequate for the identification of frogs and crickets. From Tables 5 and 6, we can also see that if each species is modeled by more than one vector template, the classification accuracy will increase as well.

3.2. Experiment 2—progressive training and testing

In this experiment, the first-half of the syllables are used for training and the second half of the syllables are used for testing. Tables 7 and 8 compare the recognition results of AMFCC with HMM and ALPC for 30 frog calls and 19 cricket calls, respectively. In addition, multiple feature vector templates were compared in these two tables. From these two tables, we can see that AMFCC greatly outperforms HMM and ALPC. However, if each species is

Table 5
Classification accuracy of 30 frog calls for alternate training and testing

SC	HMM (%)	ALPC (%)	LPC-2 (%)	LPC-3 (%)	AMFCC (%)	MFCC-2 (%)	MFCC-3 (%)
1	96	81	97	97	92	92	92
2	75	93	98	100	100	100	100
3	67	100	100	100	100	100	100
4	89	100	100	100	100	100	100
5	80	88	87	87	96	99	98
6	85	89	88	85	93	91	91
7	78	82	94	100	100	100	100
8	77	100	100	100	100	100	100
9	100	100	100	100	100	100	100
10	100	84	98	100	100	100	100
11	83	68	81	77	96	99	96
12	60	89	100	100	100	100	100
13	93	100	97	97	100	100	100
14	100	53	46	62	61	77	77
15	100	83	85	90	96	95	97
16	95	85	89	89	88	89	85
17	99	100	100	100	100	100	100
18	100	98	100	100	98	100	100
19	40	100	100	100	100	100	100
20	91	100	100	100	100	100	100
21	96	63	95	95	95	100	100
22	75	82	85	80	91	93	96
23	67	76	97	97	97	100	100
24	89	95	97	98	100	100	100
25	80	96	98	96	80	90	88
26	85	59	76	80	100	100	100
27	78	100	100	100	100	100	100
28	77	66	67	67	83	100	100
29	100	100	89	100	100	100	100
30	100	86	89	91	100	100	100
Average	81.9	88.9	91.9	92.2	96.8	97.6	97.4

Table 6
Classification accuracy of 19 cricket calls for alternate training and testing

SC	HMM (%)	ALPC (%)	LPC-2 (%)	LPC-3 (%)	AMFCC (%)	MFCC-2 (%)	MFCC-3 (%)
1	65	97	97	97	100	100	100
2	92	100	96	100	100	100	100
3	100	100	100	100	100	100	100
4	33	22	89	89	100	100	100
5	100	100	100	100	100	100	100
6	86	84	85	88	100	98	98
7	89	100	89	89	88	89	89
8	100	90	91	91	100	100	100
9	0	0	50	50	0	50	50
10	17	83	100	100	33	100	100
11	83	100	100	100	100	100	100
12	100	100	100	100	100	100	100
13	79	98	100	98	100	100	100
14	64	100	100	100	100	100	100
15	100	100	100	100	100	100	100
16	92	91	84	88	95	96	96
17	99	91	95	94	100	100	100
18	100	100	100	0	100	100	100
19	20	80	80	80	100	100	100
Average	91.4	91.6	92.8	93.3	98.1	98.9	98.9

model by more than one vector template, the classification accuracy does not increase accordingly, especially for the recognition of cricket calls. The reason to explain the phenomenon is that in the case of progressive training

Table 7
Classification accuracy of 30 frog calls for progressive training and testing

SC	HMM (%)	ALPC (%)	LPC-2 (%)	LPC-3 (%)	AMFCC (%)	MFCC-2 (%)	MFCC-3 (%)
1	66	89	97	95	97	100	100
2	98	91	96	94	100	100	100
3	89	97	100	100	100	100	100
4	40	100	100	100	100	100	100
5	76	71	70	87	95	89	97
6	82	80	84	86	97	95	93
7	53	58	53	94	100	100	100
8	54	100	100	92	100	100	100
9	93	100	100	100	100	100	100
10	50	76	98	100	100	100	100
11	80	66	74	75	100	100	100
12	68	100	100	100	100	100	100
13	81	97	94	100	100	100	100
14	62	46	54	46	38	46	54
15	77	87	87	91	88	92	75
16	63	85	85	85	92	96	93
17	75	100	100	100	100	100	100
18	96	100	99	100	97	97	97
19	100	100	100	100	100	100	100
20	99	100	100	100	100	100	100
21	43	60	74	78	56	48	74
22	15	30	48	50	73	72	61
23	79	68	87	82	100	100	100
24	94	96	98	98	99	99	100
25	67	96	96	96	96	100	100
26	94	85	92	90	100	100	100
27	0	100	100	33	100	100	0
28	0	66	100	50	100	100	83
29	89	88	89	56	100	100	33
30	98	69	80	69	100	100	98
Average	78.9	83.9	86.8	89.8	96.2	95.5	94.6

Table 8
Classification accuracy of 19 cricket calls for progressive training and testing

SC	HMM (%)	ALPC (%)	LPC-2 (%)	LPC-3 (%)	AMFCC (%)	MFCC-2 (%)	MFCC-3 (%)
1	90	92	94	96	98	100	100
2	83	100	83	83	100	100	100
3	100	100	100	100	100	100	100
4	50	20	90	90	100	100	100
5	50	100	100	100	100	100	100
6	99	97	94	99	100	100	100
7	89	100	89	89	88	89	89
8	68	86	86	95	100	100	100
9	33	100	33	67	100	33	67
10	86	85	86	29	85	86	0
11	71	100	100	0	100	100	0
12	100	100	100	67	100	100	17
13	93	94	98	36	100	100	45
14	87	93	87	7	93	93	7
15	100	100	100	0	100	100	0
16	99	98	93	92	98	99	77
17	100	97	99	84	100	100	46
18	0	100	100	100	100	100	100
19	20	80	60	80	80	60	0
Average	91.2	94.6	94.0	79.5	98.9	98.5	69.1

and testing, the diversity of the training syllables is not large enough to learn all the possible calls, especially for the experiment on cricket calls.

Comparing Tables 3 and 5 (respectively, Tables 4 and 6), we can see that alternate training and testing performs better than progressive training and testing. This is due to the fact

that for progressive training and testing not all frog/cricket calls are well trained since the training and testing syllables may be extracted from the calls of different frogs/crickets.

4. Conclusions

In this paper we propose a method capable of identifying frogs/crickets automatically from the sounds they generate. Each syllable corresponding to a piece of vocalization is first segmented. The averaged MFCCs (AMFCC) over all frames within a syllable are used as vocalization features such that the effect of background noise can be attenuated. Linear discriminant analysis (LDA) is used to reduce the feature dimension and increase the classification accuracy. Experimental results have shown that AMFCC greatly outperforms HMM and ALPC. In the experiments, the average classification accuracy is up to 96.8% and 98.1% for 30 kinds of frog calls and 19 cricket calls, respectively. From the experimental results, we can see that the proposed simple approach is adequate for the identification of frogs and crickets since the sounds generated by them are simpler than other sounds. If the animal sounds are more complex like bird songs in which the types of sounds and the syntactical arrangements of the sounds change significantly, a more complicated system may be required to do the recognition process.

Acknowledgments

The authors would like to thank the anonymous referees for their valuable comments that improved the representation and quality of this paper. This research was supported in part by Chung Hua University under contract CHU-94-TR-02 and the National Science Council of ROC under contract NSC-92-2213-E-216-020.

References

Baker, M.C., 2004. The chorus song of cooperatively breeding laughing kookaburras: characterization and comparison among groups. Ethology 110 (1), 21–35.

Duda, R., Hart, P., Stork, D., 2000. Pattern Classification. Wiley, New York.

Harma, A., 2003. Automatic identification of bird species based on sinusoidal modeling of syllables. Internat. Conf. on Acoust. Speech Signal Process. 5, 545–548.

Kogan, J.A., Margoliash, D., 1998. Automated recognition of bird song elements from continuous recordings using DTW and HMMs. Journal of the Acoustical Society of America 103 (4), 2185–2196.

Lamel, L.F., Rabiner, L.R., Rosenberg, A.E., 1981. An improved endpoint detector for isolated word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing 29, 777–785.

Li, D., Sethi, I.K., Dimitrova, N., McGee, T., 2001. Classification of general audio data for content-based retrieval. Pattern Recognition Letters 22 (5), 533–544.

Lu, G.J., 2001. Indexing and retrieval of audio: A survey. Multimedia Tools and Applications 15, 269–290.

Picone, J.W., 1993. Signal modeling techniques in speech recognition. Proceedings of the IEEE 81, 1215–1247.

Rabiner, L., Juang, B.H., 1993. Fundamentals of Speech Recognition. Prentice-Hall, Englewood Cliffs, NJ.

Slaney, M., 2002. Mixture of probability experts for audio retrieval and indexing. IEEE Internat. Conf. on Multimedia Expo. 1, 345–348.

Tian, Y., Wang, Z., Lu, D., 2002. Nonspeech segment rejection based on prosodic information for robust speech recognition. IEEE Signal Processing Letters 9 (11), 364–367.

Vergin, R., O'Shaughnessy, D., Farhat, A., 1999. Generalized Mel-frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition. IEEE Transactions on Speech and Audio Processing 7 (5), 525–532.

Wold, E., Blum, T., Keislar, D., Wheaten, J., 1996. Content-based classification, search, and retrieval of audio. IEEE Multimedia Magazine 3 (3), 27–36.

Zhang, T., Kuo, C.-C.J., 2001. Audio content analysis for online audiovisual data segmentation and classification. IEEE Transactions on Speech and Audio Processing 9 (4), 441–457.