Using Speaker and Glossary Identification for Access Control

應用語者及詞彙辨識於門禁控制

Fu-Hua Chou(周復華)
Ching Yun University, Taiwan, ROC
Email:fhchou@cyu.edu.tw

Yu-Cyuan Ciou(邱昱銓)
Ching Yun University, Taiwan, ROC
Email: B9511155@cyu.edu.tw

摘要

科技日新月異，而許多嵌入式系統也漸漸進入人們的生活之中，與它們之間的溝通也顯得重要，因此若是能直接以語音和這些系統進行溝通，那將會為我們帶來不小的方便。那首先我們必須使系統能認出使用者以及所下達的指令，而這便運用到所謂的語者和詞彙的辨識。本文以高斯混合模型(Gaussian Mixture Model, GMM) 為基礎，建立一個不限語言與說話內容的語者辨識系統，並進一步以隱藏式馬可夫模型(Hidden Markov Model, HMM) 用於詞彙辨識上。系統以軟體配合硬體架構，當軟體部分做出識別動作於使用者以及詞彙之後，系統便送出觸發訊號於相對應的硬體。測試環境設定為一般環境並有些許雜音，使辨識時更能模擬出如同正常使用環境下的各種情形，亦可藉以測試背景雜訊對於系統的影響。

關鍵詞:語者識別、高斯混合模型、詞彙識別、隱藏式馬可夫模型

Abstract

Many embedded technique systems are rapidly development and integrated into people's lives, and a convenient way, voice, for human to communication with their digital embedded systems is more important today. The first thing we need to make the system recognize the user as well as the instructions issued, this will used the so-called language and vocabulary identification technology. This paper uses Gaussian Mixture Models to build a speaker recognition system with language-free and text-free features, and further utilities the Hidden Markov Models to vocabulary identification. The another feature of this system is software combined with hardware schemas, when the recognizing work of the user and vocabulary is finish, this system will sends a trigger signal to a mechanical hardware to complete the specified work. The testing environment is to simulate the general environment with some noises, so it can imitate some voice commands utility environment under various situations, and emerge the influence of the background noise for this recognition system.

Keywords：Speaker recognition, Gaussian Mixture Model, Vocabulary identification, Hidden Markov Model

1. 前言

目前語音辨識，在於錄音室環境中，已可達到100%的辨識率，若是在於實際環境之下，或多或少會產生不同的外在因素導致辨識率下降，如：背景雜訊。在此將探討語者及詞彙辨識於實際環境的應用和影響，系統上主要採用高斯混合模型(Gaussian Mixture Model, GMM) 以及隱藏式馬可夫模型(Hidden Markov Model, HMM) ，兩種不同的模型比較方法，做為識別與統計其平均機率之依據。一般測試個別語者發聲差異方面，主要可分為語者發聲器官、共振腔型的不同，或者外界異常背景雜訊之干擾，以及語者內心情緒起伏的差異，這些條件同時也相對影響著辨識率的高低。

在於語者特徵參數建立方面，本系統採用長程統計時之作法，因此將不會受限於語者之說話內容，也就達到與文句不相干連的效果。待收音完成後利用梅爾頻率倒頻譜係數(Mel-Frequency Cepstral Coefficients, MFCC) 之特徵參數建立語者的GMM，不同特徵所建立的模型也皆有所不同，面對受測者本身所收錄之音訊特徵，也將會對其自己所建立之模型有著最大的符合度，而該模型便為此語者之語者模型(Speaker Model) 。

輸出部分則是模擬實際應用環境，製作正反雙向橫移式自動門，藉以提供更多樣的種類變化，而控制部份應用8051單晶片分別連接系統與驅動馬達，當系統辨別出語者及下達之指令種類後，送出相對應之觸發信號，再透過繼電器電路以及微動開關做為控制驅動馬達不同正反轉向和停止機制之用。在使用者介面(User Interface)方面，主要設計四種功能包含：加入使用者、訓練語料庫、語者辨識及詞彙辨識，並引導使用者以分階段式輸入資料，讓使用者不會感到系統使用上的困難。

2. 語音識別技術

語音識別技術，一般又可稱為(Automatic Speech Recognition, ASR) ，其主要目的是將人類所發出語音中的類比訊號，透過類比到數位轉換器(Analog-To-Digital Converter, ADC)切換成可經由計算機所解析、運算的數位訊號，而後建立語音模型以供辨識。這部份的處理主要可分為三個階段，如圖 1 所示。第一階段是語音訊號處理(Speech Signal Processing) 進行的前置處理，主要是當音訊收錄後將其精簡化的以致方便後續解析的一個過程，第二階段是語音訊號之特徵萃取(Feature Extraction) ，進一步將精簡化後的訊息位於時間上及頻率上的特徵加以存取。第三階段是語者識別(Speaker Recognition) ，主要使用機率的模型，將不同的聲音特徵加以分門別類以供識別。以下將先針對前兩部份做介紹。

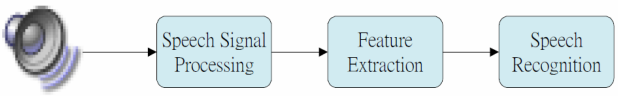

圖 1 語音識別技術

2.1. 語音訊號前處理

當我們將人們所發出之語音訊號做解析時，可以發現到此訊號為一線性且時變(Time-Variant) 的信號，而它有著非常迅速的變化且很難加以預測。但若從頻率區間來做觀察的話，將可發現頻譜(Spectrum) 會隨著時間而有著緩慢的變化，因此在處理上可以使用短時距(Short Time) 的方法來擷取並處理訊號，一般又將其方法稱為短時段語音處理，也就是收錄的語音訊號分割成固定時間長度進一步將其細分，區間段的單位則稱為音框(Frame) ，對於每一音框內的訊號，抽取出代表語音的特徵參數。一般將語音訊號由收錄音訊直至到語音之特徵萃取的前置處理過程當中，必須經由下列各個步驟：音框取樣(Sample) ，預強調(Pre-emphasis) ，端點

偵測(End Point Detection) ，漢明視窗(Hamming Window) 等。以下將介紹語音訊號前置處理的各個步驟。

2.1.1. 音框取樣

人類所發出的語音訊息，在於空氣之中是以波的形式經由介質所傳輸，統稱為聲波(Acoustic Wave) ，而若是要透過計算機來處理的話則必須先將接收到的類比訊號(Analog Signal) 透過轉換器(Transducer) 轉換成可以處理的數位訊號(Digital Signal) ，這種轉換稱之為取樣。

一般人類能夠聽見的聲音頻率範圍大約在 20 到 20000 Hz 之間，其中人類語音頻率主要分布在 300 到 4000 Hz 之間。根據取樣定理，取樣頻率必須是最高語音訊號的兩倍以上，才不會造成失真，故通常取樣頻率為 8000 Hz。這頻率值表示每秒的語音信號，有 8000 個取樣點。換算起來，則每 0.00125 秒就會取一個值。由於語音訊號在於時域的變化非常大，且語音在頻域上每一小段都有著不同特性。故需將整段語音訊號分成數個音框，且假設每個音框內的語音特性都是線性的。音框的長度大小一般可由使用者所決定，其中包含多少個取樣點並沒有所規定，但因後續特徵萃取過程中需將訊號做快速傅立葉轉換(Fast Fourier Transform, FFT) ，而其轉換限制必須以 2 為底之數值，故通常取 256 個取樣點為一個音框，換算成時間就是 16 ms，這也就是一個音框的長度。為了使語音特徵改變具有延續性，故需將前後音框重疊，至多可重疊達二分之一個音框長度，則每次移動視窗的距離就是 8 ms，也就是 128 個取樣點。

2.1.2. 預強調

當人類發聲使聲帶產生振動時，聲帶端之結構可以看成如同一串脈衝訊號通過一個由聲帶所形成的濾波器(Glottal Shaping Filter) ，其中產生的空氣流量速度波形即有-12dB/oct 的高頻衰減。而嘴唇部分對於氣流產生的阻擋效果則可視為一個輻射阻抗，若是當聲音訊號通過時則會產生一個高通濾波的效果，有著+6dB/oct 的高頻增強。若發聲腔道模型為0dB/oct 的高頻衰減，則最後所產生的訊號有-6dB/oct 高頻衰減。因此若是要補償此衰減，則必須將語音信號通過一階高頻濾波器來做處理，而此處理過程即稱為預強調。補償公式如下。

$$y(n)=x(n)-ax(n-1) \qquad 0\le n\le N-1 \ldots\ldots..(2.0)$$

其中 $y(n)$ 即代表經過預強調之取樣值。而 a 為0.9~1 之間，通常採用計算發元音時的第一個正規化自相關函數作為 a 的值。

2.1.3. 端點偵測

通常在於龐大的音訊處理量中，本身包含著許多不帶音訊或是次要音訊的內容在其中，這將會大大拖慢計算機的處理時間，因此為了更進一步精簡音訊處理量，則有了端點偵測這項技術的產生，在端點偵測技術上，主要使用越零率(Zero-Crossing Rate) 、能量曲線(Energy Contour) 、作為偵測端點的特徵。越零率是指在於每一個音框中取樣值所通過零點的次數。一般分析人類發聲情形時，語音訊號於區段時間內的越零率會較低。相反的，若是無語音訊號，或是僅發出摩擦音或收錄到規律雜訊時的訊號，所獲得的越零率值則會較高。但這也只是較為基本的判斷條件，故單單只靠越零率是不夠的，還必須搭配能量(Energy) 來加以判斷。因為語音訊號在於有聲部分的波形振幅較大，因此能定出一個數值作為門檻值，若是當音框能量超過了門檻值，則可視為有聲。實際上，越零率計算公式如下：

$$Z_x(m)=\frac{1}{N}\sum_{n=m-N+1}^{m}\frac{1}{2}\left|\operatorname{sgn}\left[x(n)\right]-\operatorname{sgn}\left[x(n-1)\right]\right| \ldots(2.1)$$

對於取樣之語音訊號訂定為 $x(n)$ ，其中 N 表示音框的大小， m 為音框編號。sgn [.]為符號函數，倘若相鄰之兩個訊號不同號，即可得到2，同號就得到0。而能量曲線計算公式如下：

$$E_x(m)=\log[|\sum_{n=m-N+1}^{m}|x(n)|^2|] \text{.....(2.2)}$$

其中 m 為音框編號， $x(n)$ 表示整段語音之第 n 個取樣值， N 則為音框大小。

2.1.4. 視窗化

通常在求取音框內數值的同時，音框兩邊界所產生的不連續現象，在聽覺感受上會有著明顯的落差，若觀察在於頻域上的語音頻譜也會發現其延續性受到破壞。因此必須使用可以將邊界誤差降低的方法，也就是將音框做視窗化的步驟，在於邊界上的抑制效果會讓其變化更具有延續性的結果產生。

一般常見的種類有(Hamming Window) 、漢尼視窗(Hanning Window) 、矩形窗(Rectangular Window) ，在此採用漢明視窗，其特性為兩邊緩慢減小，如圖2 所示。

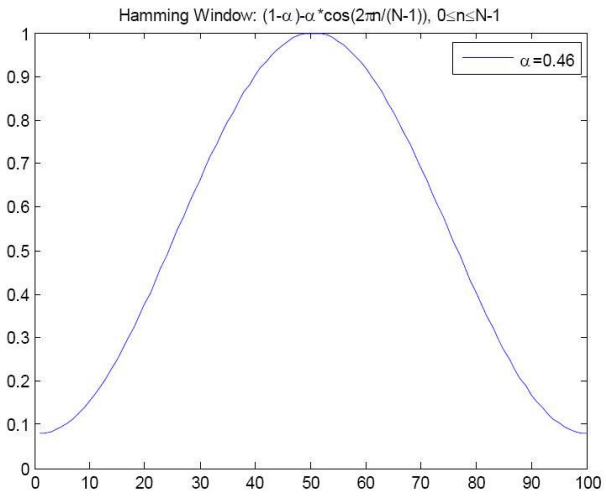

圖2 漢明視窗

視窗公式如下：

$$w(n)=\begin{Bmatrix}0.54-0.46\cos\left(\frac{2n\pi}{(N-1)}\right),0\le n\le N-1\\ 0,otherwise\end{Bmatrix} \text{...(2.3)}$$

音框乘上漢明視窗公式如下：

$$s(n)=y(n)\times w(n) \text{.....(2.4)}$$

$s(n)$ 代表經過漢明視窗後之取樣值。經過漢明視窗後的取樣值，便可算是完成語音訊號前置處理的程序，接下來即是特徵萃取，在於萃取過程中全都將以音框為單位。

2.1.5. 特徵萃取

當完成語音前置處理之後的資料量，雖說以省去不少次要資訊，但整體來說，剩餘的精簡化資料仍是非常之龐大，若是直接交由計算機來做交互比對，那將會非常沒有效率，因此還必須更進一步的將語音資料，尋其特性來取出當中適合的特徵參數來代表語音訊號，而此一過程即稱為特徵萃取，目前已有多種特徵萃取之方法，如：線性預測倒頻譜參數(Linear Prediction Cepstrum Coefficient, LPCC) 、梅爾頻率倒頻譜參數(Mel Frequency Cepstrum Coefficient, MFCC) 、感覺加權線性預測(Perceptual Linear Predictive, PLP) 與口音敏感參數(Accent Sensitive Cepstrum Coefficient, ASCC) 等，本文主要採用MFCC ，因此以下將對於MFCC 做介紹。

語音訊號當中，各頻帶區間之能量分佈，稱之為語音訊號之頻譜特徵，而在於語音辨識當中，最廣為被運用的方法則是梅爾頻率刻度。梅爾頻率刻度是根據一般人耳在於接收不同基頻頻率的狀態表現時，模擬其聽覺特性而發展出的

一種頻率刻度轉換。梅爾頻率倒頻譜參數主要擷取的是在於頻率域中的特徵參數，因此必須先將語音訊號從時域狀態轉換至頻域狀態，此一轉換過程通常使用快速傅立葉轉換(Fast Fourier Transform, FFT) 。轉換至頻率域後之音框再通過一組梅爾頻率刻度(Mel-scale Frequency) 所設計之三角數位帶通濾波器。通過濾波器後之參數的數量龐大，因此再使用離散餘旋轉換(Discrete Cosine Transform, DCT) 將其減量，由於先前透過FFT 轉換至頻率域狀態，所以採用DCT 轉換是期望能轉回類似時域之情況來觀察，也就是倒頻譜(Cepstrum) 動作。又因為之前所採用的是梅爾刻度轉換至梅爾頻率(Mel-frequency) ，故稱之為梅爾倒頻譜(Mel-frequency Cepstrum) 。

*梅爾頻率刻度(Mel-scale Frequency) 其公式如下：

$$Mel = 2595 \times \log\left(1 + \frac{f}{700}\right) \text{.....(2.5)}$$

故

$$f = 700 \times \left(10^{\frac{Mel}{2959}} - 1\right) \text{.....(2.6)}$$

其中 Mel 代表經由頻率刻度轉為梅爾刻度， f 代表透過梅爾刻度轉為頻率刻度。梅爾三角濾波器主要將4KHz 分割為20個頻帶，如圖3 。當頻率小於1kHz 時，梅爾刻度具有線性關係，在大於1kHz 時， Mel 刻度則呈現對數關係。

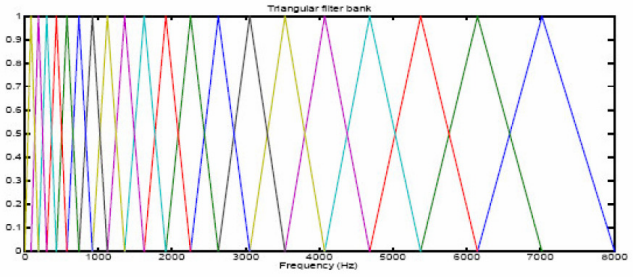

圖3 梅爾三角濾波器

這些三角形濾波器的中心頻率定為臨界頻帶的中心頻率，而兩邊的截止頻率就是兩個相鄰臨界頻帶的中心頻率。以數學表示，其公式如下：

$$B_m(k) = \begin{cases} 0, k < f_{m-1} \\ \dfrac{k - f_{m-1}}{f_m - f_{m-1}}, f_{m-1} \le k \le f_m \\ \dfrac{f_{m+1} - k}{f_{m+1} - f_m}, f_m \le k \le f_{m+1} \\ 0, f_{m+1} < k \end{cases} \text{.....(2.7)}$$

$B_m(k)$ 表示第 m 個頻帶的三角濾波器，為第 m 頻帶的中心頻率，與即是相鄰的前後兩個頻帶的中心頻率，m 為全部的頻帶數目。計算梅爾倒頻譜參數前，必須先計算每一個梅爾刻度內的之能量總和，公式如下：

$$Y(m) = \log\left\{\sum_{k=f_{m-1}}^{f_{m+1}} |E(k)|^2 B_m(k)\right\} \text{.....(2.8)}$$

其中 $E(k)$ 為音框內各頻率的能量。

*離散餘弦轉換(Discrete Cosine Transform, DCT) 公式如下:

$$c_x(n) = \frac{1}{M}\sum_{m=1}^{M} Y(m)\cos\left(\frac{\pi n\left(m - \frac{1}{2}\right)}{M}\right) \text{.....(2.9)}$$

其中 M 為全部的頻帶數目。

若只取梅爾頻率倒頻譜係數是無法準確的進行辨識的，因此必須加入其他參數。所以將12階的梅爾頻率倒頻譜係數加上音框之對數能量，組成13階特徵參數，再以此13階特徵參數，

取其一階差量倒頻譜參數與二階倒頻譜參數，全部共39階係數來代表一個音框的MFCC。其中差量的意義為參數相對於時間上的斜率，也代表參數在時間上的變化程度。差量公式如下：

$$\Delta C_n(t)=\frac{\sum_{\tau-n}^{n}\tau\bullet C_n(t+\tau)}{\sum_{\tau=-n}^{n}\tau^2},n=1,2,...,L\ \text{.....(2.10)}$$

2.2. 高斯語者辨識模型

對於語者辨識模型，本文主要採用高斯混合模型(Gaussian Mixture Model, GMM) 來建立語者辨識模型，高斯混合模型屬於一種高維機率密度函數，經由高維機率密度函數可以使用機率統計的方式來表示這種可變性，高斯混合模型作為高斯機率密度函數的一個線性組合，是語音訊號處理上常用的統計模型，其基本理論之前提是當只要有足夠多數目的混合分量，則就可以逼近任意一種密度函數，而語音特徵通常有著平滑的機率密度函數，因此在於有限數目的高斯密度函數就足以對語音特徵的密度函數形成平滑逼近。以下將介紹高斯混合模型建立方式，並簡單地描述高斯混合模型的架構，以及在語者辨識實驗中使用高斯混合模型來代表每位語者特性、模型參數的估算及語者辨認的公式。

高斯混合模型是由 M 個基本密度的加權總合。其公式表示如下：

$$p(\vec{x}\mid\lambda)=\sum_{i=1}^{M}p_i b_i(\vec{x})\ \text{.....(2.11)}$$

其中 $\vec{x}$ 是維度為 D 的特徵向量， M 是高斯混合密度的混合數(mixture)， $b_i(\vec{x}),i=1,...,M$ 是基本密度而 $P_{i,i}=1,...,M$ 是混合權數，其中必須滿足 $\sum_{i=1}^{M}P_i=1$ 的限制。每一個基本密度是 D 維的高斯函數，如下所示：

$$b_i(\vec{x})=\frac{1}{(2\pi)^{D/2}\mid\sum\mid^{1/2}}\times\exp\left[\frac{(\vec{x}-\overline{u_i})}{2}\right]\text{.....(2.12)}$$

其中 $\overline{\mu_i}$ 是平均向量(mean vector)， $\sum_i$ 是共變異矩陣(covariance matrix)。一個完整的高斯混合密度可以用平均值向量，共變異矩陣和混合權值來表示，我們可以用來表示這些參數的集合。

$$\lambda=\left\{P_i,\bar{u}_i,\sum_i\right\}i=1,...,M\ \text{.....(2.13)}$$

會使用高斯混合模型來代表語者模型主要有兩個原因，第一個是因為高斯混合模型的每一個基本密度皆可以模擬出一些發聲狀態的特徵。因此我們使用高斯混合模型中第 i 個平均值來代表第 i 個聲音特徵的頻譜形狀，並且使用共變異矩陣來代表頻譜形狀的變化。

第二個原因是因為利用高斯混合模型便能很平滑地近似任意形狀的密度。單一型態的高斯混合語者模型是利用一個平均值向量和共變異矩陣來代表語者特徵參數的分佈情形。而向量量化模型則是利用一組離散的特徵樣板來代表語者的分佈。因此高斯混合模型整體上來說算是結合了上述兩種模型的優點，並且它利用了一組離散的高斯函數，加上高斯函數具有的平均值向量和共變異矩陣使得它擁有更好的模型能力。而且當我們比較使用單一型態高斯模型，高斯混合模型和向量量化模型三者之間的差異時，可以發現到使用高斯混合模型確實可以達到跟原來的語

者資料有最近似的結果。並且，當高斯混合模型的混合數越多，它就越能達到近似原來資料分佈的目的，但是相對所需的訓練及辨識時間也會跟著增加。

2.3. 語者識別

在於語者辨認上的過程主要又分為，語者識別以及語者確認兩部份，首先對於語者識別當中，若是訂定有一群語者 $S=\{1,2,...,S\}$，我們則可利用一群高斯混合模型來做表示。而對於其中一段測試語句 $X=\{x_1,x_2,...,x_t\}$，其中語者識別的目的就是在這一群語者模型當中，找出一組事後符合機率最大的模型，如下所表示：

$$\hat{S}=\arg\max_{1\le k\le s} p(X\mid\lambda_k) \quad \text{.....(2.14)}$$

對上式取對數可改寫成

$$\hat{S}=\arg\max_{1\le k\le s}\sum_{t=1}^{T}\log p(X\mid\lambda_k) \quad \text{.....(2.15)}$$

2.4. 語者確認

語者確認部分，主要則是對於一段測試的語句Y，去計算它與宣稱之語者模型的對數近似值，與此模型的門檻值之間的關係，如果對數近似值大於門檻值則將會判定此人為宣告者而接受，若是對數近似值小於門檻值的話，則判定此人為假冒者而拒絕。演算法公式如下：

$$S(X\mid k)=\log p(X\mid\lambda_k)\begin{cases}\ge\ \theta & \text{接受}\\ <\ \theta & \text{拒絕}\end{cases} \quad \text{.....(2.16)}$$

門檻值所設定的大小將會造成系統判斷的兩難，對於語者確認系統的建立當中，除了要建立效能不錯的語者模型之外，門檻值的設定也相對影響著整個系統效能。若是將門檻值設定的過高，會容易導致真實語者被系統拒絕，此即False Reject (FR) 之情況；而門檻值設定的過低，則會使假冒者被系統誤判為宣告語者，即False Acceptance (FA) 之情況，使得系統的效能降低。在本文中，門檻值的大小設定在於FR 與FA 兩條曲線的交會點，選擇此點將可使整體的等錯誤率(EER ：Equal Error Rate，即FA = FR)達到最低。其中FA 與FR 定義如下：

$$\text{FA}=\frac{\text{假冒者被判定爲真實語句的句數}}{\text{假冒者測試總句數}}$$

在於本論文所做的實驗中，求取 EER 的計算方式如下所示:

$$\text{ERR}=\frac{1}{2}(\text{FA+FR}) \quad \text{.....(2.17)}$$

2.5. 詞彙識別

透過語者辨識系統分辨出個別使用者後，而進一步希望能藉由語音直接去做指令之下達，如此便可省去許多控制器操作之過程，並可延伸至各類嵌入式系統當中，使人們的生活更加便利。當中談到經由語音指令下達使計算機能做出判斷，這部份的識別則稱為詞彙辨識，通常會依照詞彙數量大小以及發音方式的不同而有所區分，而常用於聲學比對之方法有，動態時間校準(Dynamic Time Warping, DTW) 與隱藏式馬可夫模型(Hidden Markov Model, HMM) ，本文主要採用 HMM，以下將針對 HMM 做簡單介紹。

2.6. 隱藏式馬可夫模型

本文中主要使用隱藏式馬可夫模型(Hidden

Markov Model, HMM) 做為詞彙辨識之模型，早先關於馬可夫鏈(Markov Chains) 的基本理論在二十世紀初期就為廣為數理學家和工程師所熟知，然而直到波氏(Baum) 提出將馬可夫模型之參數最佳化的方法之後，才將它轉用於語音識別上，而且有著相當的成效。隱藏式馬可夫模型屬於一種雙重的隨機程序(Double Stochastic Process) ，是以一種無法觀察(Hidden) 且為有限可能值(Finite Number) 的隨機程序做為基礎，在透過另一個隨機程序，可從中觀察到隱藏式馬可夫模型所產生的一連串觀測值(Observation) 。其隱藏式馬可夫模型定義公式如下：

$$\lambda = \{A, B, \pi, S, V\} \quad \text{.....(2.18)}$$

模型 λ 當中包含三個機率集合，以及觀察結果與狀態數，其中，$S = \{s_1, s_2, ..., s_N\}$ 代表著每一個隱藏式馬可夫模型的若干種狀態，數目量為 N ，在於每一個狀態當中包含一組狀態轉移機率(State Transition Probability) ，$A = \{a_{ij}\}$ ，用以決定狀態 i 轉移至狀態 j 的機率，而當中觀察出的M 個結果則以$V = \{v_1, v_2, ..., v_M\}$ 來表示，同時在$N \times M$ 的狀態觀測機率分佈則(Observation Probability Distribution) 做為，$B = \{b_j(k)\}$ ，用來決定觀測對象k 出現在狀態 j 的機率值。而初始狀態分佈$\pi = \{\pi_i\}$ ，用來表示模型是從狀態 i 開始的機率大小。整體模型示意則如下圖4 所示。

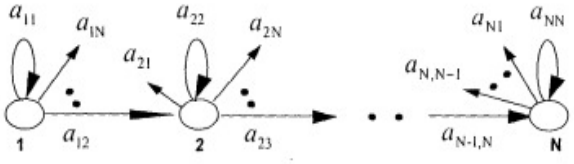

圖4 一個狀態數為N的HMM示意圖

3. 語音雙模識別系統

將前文所敘述之語音識別技術加以整合應用，便可將其建立為一套有規模架構之系統，而本文所使用之系統主要由本實驗室所開發完成，當中將詞彙以及語者辨識技術整合於其中，系統之特色，即在特徵萃取時分別同時建立語者聲頻及語音詞彙雙模型。在辨識上，由於本系統設有語者模型與詞彙模型，所以先識別出語者身分，再經存有語者個人之特色詞彙庫中找出詞彙模型以進行辨識出正確詞彙，故稱之為語音雙模識別系統。

4. 硬體架構

硬體架構上主要是與軟體系統相互搭配，將其運用至模擬實際環境之模型作品上，藉以發揮實物設計開發以及製作操作之精神。本文主要設計為一正反雙向橫移式自動門，欲從其中獲得更多樣之種類變化，可供模擬條件上的不同應用，而將其各區塊細分則可分為：實體之做動結構設計、8051 單晶片與計算機之傳輸、Relay 控制驅動電路。圖5 所示為基本硬體架構方塊圖，硬體實體成品則如下圖6：

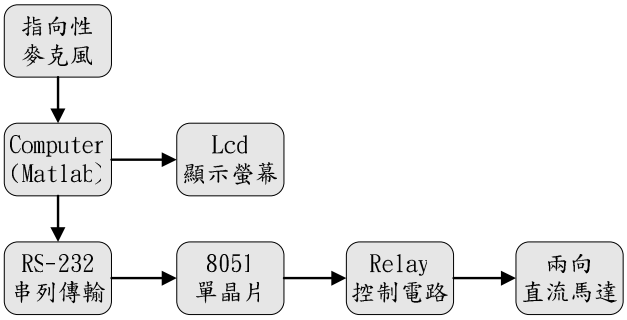

圖5 硬體架構方塊圖

其中透過指向性麥克風進行收音動作，將音訊資料交由計算機做處理，運算演算法主要架構於Matlab程式之上，處理結果將會顯示於螢幕，並且送出輸出訊號透過RS-232串列傳輸至8051單晶片作判斷，最後將會由8051單晶片送出相對應的觸發信號給Relay 控制電路部份，進行兩向直流馬達電源的切換，當中使用微動開關做為直流馬達停止機制之用。

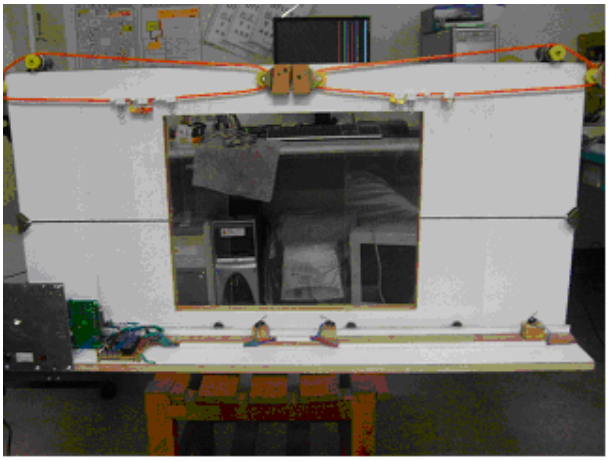

圖6 硬體實體成品圖

5. 實驗結果

5.1 語者辨識

系統內設有436位語者模型，其中正確需開門者語料有5位，31位為不可開門之測試語料，另400非使用語者使用MAT-400語料庫內之語料所建立，藉此提升辨識的難度。每位可開門者語料需訓練兩次，每次60秒，以100次語音輸入進行門禁系統之開門測試，其正確識別率如表1所示。測試中語者語氣平緩不含情緒化，辨識率如表1。另外，安排5位不可開門者，測試系統的排除率，每位輸入100次語料以統計其排除率，所得資料如表2。根據本文之實驗結果，證實本文設計確可應用於門禁開啟系統，其中排除率可達93%。排除率較辨識率為高正是本系統之特色。

表1.需開啟者語者辨識正確率

語者	A	B	C	D	E	平均
辨識率	92%	90%	85%	94%	87%	89%

表2.不可開啟者之語者辨識排除率

語者	A	B	C	D	E	平均
排除率	87%	100%	98%	96%	85%	93%

系統當中設定之門檻值為-3.4564e+003 到-4.7063e+003 這區間範圍，加入之後排除率大幅提升。除此之外，訓練正確語者的說話方式，雖然犧牲些許系統便利性，卻可換來更高的安全性。圖7 為系統門檻值之示意圖。

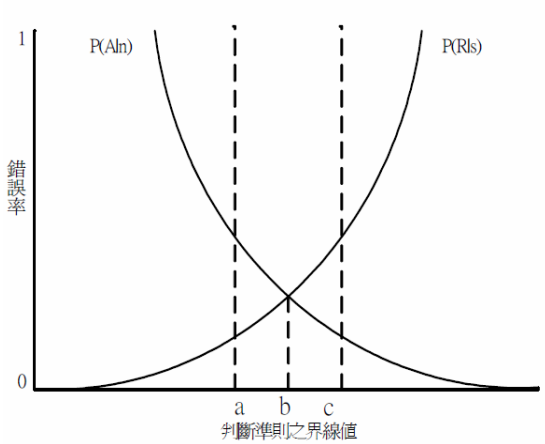

圖7 兩種錯誤機率的曲線

5.2 詞彙辨識

本論文於詞彙辨識實驗上主要使用英國劍

橋大學所開發之軟體工具(Hidden Markov Model Toolkit, HTK) 加以進行隱藏式馬可夫模型之相關演算以及模型建立，當中包含錄製音訊、標音、分析、到隱藏式馬可夫模型的訓練與辨識比對及結果分析，皆可透過HTK內建相對應的函式進行操作，以HTK建立之語音辨識系統流程示意圖如下圖8 。目前經由單一字詞測試之詞彙辨識結果如下表3，整體求得之平均識別率將近八成，當中包含系統本身建立之詞彙模型多寡因素所在，也存有外在背景雜訊等錄音品質之影響，若是能將其內外在影響因素做進一步改進，將可再使辨識率有所提升。

表3.單一字詞彙識別率

詞彙	1	2	3	4	平均
辨識率	80%	80%	77%	80%	79%

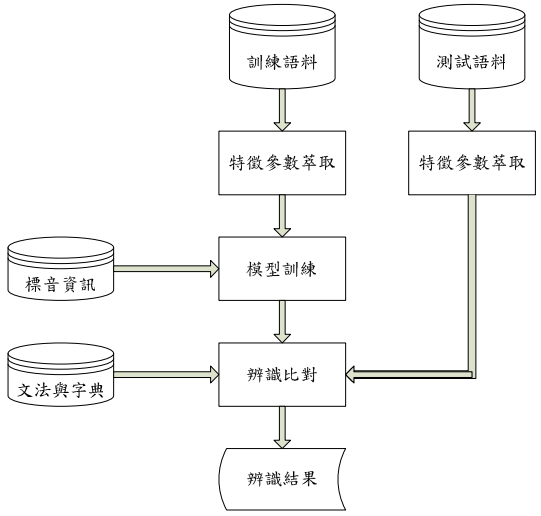

圖8 以HTK進行詞彙辨識流程

6. 相關分析比較

6.1 門禁系統

若以門禁管制方式來區別，排除傳統鎖匙之不同型式外，目前在於日常生活中已能看到 RFID 相關型式的應用，如：磁感應卡片、磁感應鑰匙圈等，其若與本系統設計以聲音做為識別依據來做比較，首先在於安全性來說，RFID 相關產品雖然不像傳統鎖匙那般容易被仿製，但由於屬個人隨身物品，一但掉落遺失，對於門禁管制來說一樣是認物不認人，而聲音則屬於個人與生俱來之特質，並且由於每個人發聲腔結構及發聲方式的不同，而產生屬於個人獨一無二的語音特色，以達到無法輕易的被他人模仿之要求，接著對於便利性來說，RFID 相關物件已能從卡片形式進一步縮小置入於個人相關隨身周邊物品之中，對於攜帶以及使用上並沒有太大的困難，但是仍有可能會發生遺忘或遺失等情形，聲音則不會，若討論識別所需時間，目前來說 RFID 這類以電磁感應通過系統做出判斷只需要極短的時間，而若以語音做為輸入源，由於其計算資料量的多寡相對影響著處理時間以及識別準確度，當中之取捨則必須多做考量。

6.2 系統環境因素

對於語音辨識系統而言，收錄音訊的品質將會對其識別結果有著極大的影響，因此一般會對於週遭環境所產生的異音有所要求，如同本文當中測試環境為一般實驗室內僅有些許正常交談音量，另外便是透過語音訊號前處理步驟，盡可能將收錄音訊做純化處理，如本文中介紹之視窗化步驟便可以達到邊界誤差降低的效果，使其變化更具有延續性，此外環境對於個人情緒之影響也必須多做考量，由於過多的情緒化語音對於辨識系統來說，會造成比對的難易度增加進而導致整體辨識率下降，因此對於環境變因的考量，目前也有其他針對於語音情緒部分以及語音雜訊排除的相關研究，若將這些技術加以運用，便能使整體系統對於環境因素方面做好的加強。

7. 結論

本論文主要希望能將語音辨識技術應用於日常生活當中，為人們帶來便利，為了展現其識

別效果，故採用門禁系統來做模擬，搭配詞彙辨識將控制技術導入其中。當中使用高斯混合模型進行語者識別亦有些許條件限制，如：情緒化語料無法準確辨認、麥克風之通道效應。而在實驗中以居家安全為出發點特別強調對於錯誤語者之排除率，同時也印證了文中的門檻值對於語音排除率能有所提升，但相對有可能會有錯誤排除之情形發生，這也影響著整體系統上的便利性。另外對於詞彙辨識的應用，也可將一一導入人們在於生活上的不同需求，若希望進一步提高辨識率，那對於辨識時間考量上也必須有所取捨，若以便利性為出發點，那其辨識反應時間以及辨識結果高低兩者更是相對重要。

未來研究發展，則希望系統更臻完美，希望在自發語料收音完成後再加入經驗模態分解，將整體的識別率再達到提升。本文證實語者辨識系統以及詞彙辨識兩者應用於一般環境中，的確可以增加未來語音在數位家庭生活的實用性。

8. 參考文獻

[1] 王小川，“語音訊號處理”，初版，全華科技圖書，台北，2004年。

[2] 劉于碩，“應用經驗模態分解技術於 情緒化自發語音之辨識”，清雲科技大學，碩士論文，2007年7月。

[3] 李政益，“特定語者特定中文語音指令雙模辨識技術”，清雲科技大學，碩士論文，2005年7月。

[4] 范世明，“高斯混合模型在語者辨識與國語語音辨認之應用”，國立交通大學，碩士論文，2002年。

[5] 周復華、黃捷群、許智源、涂世民、李玉翔、廖文翰、曾偉榮，“應用語者辨識於門禁開啟”，proc. of 2008 National Symposium on System Science and Engineering Conference，P0540，ILan，Taiwan，2008.6.6~7.

[6] D. Jurafsky, J. H.Martin, *Speech and Language Processing*, Prentice Hall PTR, Inc., 2000.

[7] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," Speech and Audio Processing, IEEE Transactions on, vol. 3, no. 1, pp.72-83, 1995.