基於 GMM 及 PPM 模型的國、閩南、客語之語言辨識

Language Identification of Madarin, Holo, and Hakka Based on GMM and PPM Models

古鴻炎　　蔡仲明

Hung-Yan Gu　and　Zhong-Ming Cai

國立台灣科技大學資訊工程系

台北市基隆路四段 43 號

E-mail: {guhy, M9415007}@mail.ntust.edu.tw

摘要

本論文研究國語、閩南語、客語三種語言作語言辨識的方法與模型。我們採用梅爾倒頻譜係數及其差分係數來掌握聲學(acoustic)特性，並且用以建立聲道(vocal track)有關的高斯混合模型(GMM)；此外，也對各個音框的特徵係數作表徵化(tokenize)，再拿去建立基於部分匹配之預測模型(PPM)，以模式化(modeling)連續的表徵之間的語言相關的特性。在訓練完成 GMM 和 PMM 模型之後，我們使用兩者算出的機率值作組合，以進行語言辨識的測試實驗，在內部測試的實驗裡，三種語言的平均辨識率可達到 95.9% (3 秒發音)、96.3% (10 秒發音)；而在外部測試的實驗裡，平均辨識率則可達到 83.7% (3 秒發音)、79.6% (10 秒發音)。

關鍵詞：語言辨識、GMM 模型、PPM 模型、MFCC 特徵參數。

1. 前言

在台灣三種主要的方言是: 國語(Mandarin)、閩南語(Min-nan或稱為Holo)、及客語(Hakka)。由於能瞭解各種方言的人並不常見，而透過語言辨識，就可將撥入的電話自動轉接至熟悉該方言的人。所以在許多需要和人應答的自動化服務中，語言辨識亦為重要的前置處理。並且自動化的語言辨識處理，也可以省略掉互動系統中擾人的「國語請按1、台語請按2...」之類的開場說明，如此就可以節省時間，而用以服務更多的人。另外，關於語音辨識系統的製作上，若要同時處理多種語言，則其辨識效果通常會低於專門對一種語言作語音辨識者。因此，先由語言辨識系統作語言辨識，再交給該語言的單一語音辨識系統作處理，辨識效果應會較好。

語言辨識(Language Identification, LID)的問題已被研究了很久，約在70年代興起，但多數研究者都是針對異國間的語言作辨認，而非對一個國家內的方言進行辨識。在1980年，Li和Edwares就提出了以統計方法進行自動語言辨識的成果[1]。Zissman於1996年整理並比較了當時常見的四種電話語音之語言辨識方法 [2]，使用OGI-TS語料庫來進行各項實驗，此論文整理出的方法--以GMM(Gaussian Model Mixture)模式化(modeling)音素特性，和以n-gram模式化語言內語音發音之順序特性--仍是現在許多語言辨識系統的根基。此外Navrátil提出了一種不同的架構[3]，先用一個複雜的bigram結合動態規劃的機制去偵測出訊號代表的phone，再以phone為參數交給後端模型作處理。至於國內方面，林俊青[4]、張智傑[5]等人，也使用OGI-TS語料庫，來進行多國語言辨識的研究及系統實作。

漢語由於方言眾多，且彼此間的差異性不是很大，如何針對漢語方言研發一套效能好的語言辨識系統，還需要更多的努力。過去，針對台灣三種方言的研究，已有蔡偉和等人提出了Gaussian Mixture Bigram Model之方法[6]，此方法結合了GMM和bigram模型，而得到不錯的辨識效果。林奇嶽等人則對於音調軌跡提出了新的特徵抽取方法[7]。本論文的目標，則在於發展可分辨國語、閩南語、客語發音之自動語言辨識系統。

依據前人的經驗，在此我們選擇採用高斯混合模型來掌握各語言的聲學特性，並且選擇以PPM (Prediction by Partial Matching)模型來掌握各語言內語音發音的表徵時序關係。發展語言辨識系統會有兩個階段：訓練階段及辨識階段。我們在訓練階段的工作流程如圖1所示，原始wav檔語料先經過

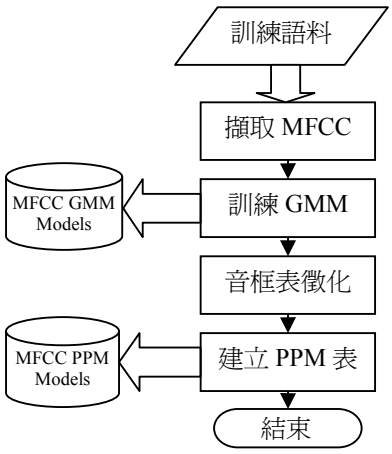

圖1 訓練階段的工作流程

特徵擷取步驟以轉換成MFCC(mel frequency cepstrum coefficient)[8]特徵向量的序列，然後作K-Means分群[9]以建立初步的GMM，再接著以MLE (maximum likelihood estimation)方法加以訓練[4,5]。完成後再用GMM對各音框進行表徵化(tokenize)處理，來轉換成一序列的表徵代碼，然後使用此表徵序列去訓練出一個PPM模型。

辨識階段的處理流程如圖2所示，輸入的語音信號先經過MFCC擷取而得到特徵向量序列，然後此序列再經各語言的表徵化處理而得到分別的表徵碼序列，用以計算各語言的PPM模型的機率值；此外，MFCC特徵向量序列也用以計算各語言的GMM模型的機率值。然後各語言分別將將兩種機率值作加權和(weighted sum)，而得到該語言的分數，之後就依分數來作語言的辨識及輸出辨識結果。

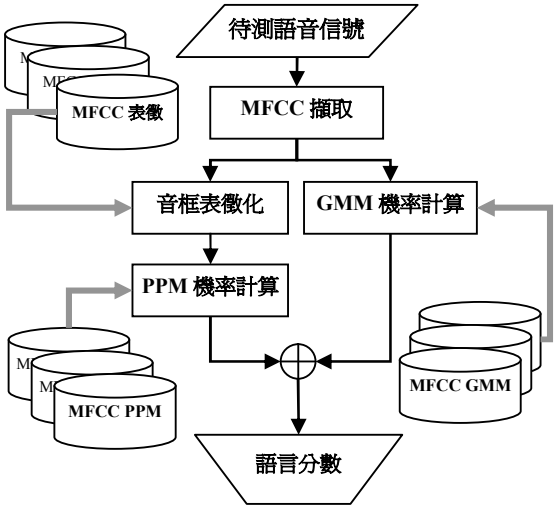

圖2 辨識階段的處理流程

2. 語料收集與特徵擷取

2.1 語料收集

我們用數位電視棒和電視卡來錄製國語、閩南語、客語的新聞播報的語音，音檔格式都轉換成16-bits/sample，取樣率 22050Hz 的 PCM 檔案。由於新聞播報當中，除了主播的語音之外，還會穿插外場記者的訪問聲音和環境噪音，因此我們花了一些人力，以手工方式來把主播之外的聲音片段刪除掉。如此所收集到的語料，其來源電台、總語者數及時間長度，就如表 1 所示，其中"語者數"欄裡的(m/f)，代表模型訓練時使用的男女語者的人數。

2.2 特徵擷取

我們設定語音音框的長度為 20ms，且相鄰音框之間重疊 10ms。對於各個音框，必須先乘上漢明窗(Hamming window)[8]，以避免音框兩側邊界

表 1 語料來源及時間長度

	來源電台	語者數	時間長
國語	台視、中視、華視、民視、公視	15 (6/4)	0.51hr
閩南語	民視、公視、大愛	19 (6/4)	0.63hr
客語	客家電視台、哈客廣播網	13 (5/5)	0.46hr

的波形不連續而造成退化的頻譜，然後經由 FFT (fast Fourier transform)轉換成頻域的頻譜振幅。

由於人耳聽聲音的知覺頻率和物理上的頻率不是線性關係，因此擷取 MFCC 時，要先使用一串頻帶有重疊的三角濾波器，來把傅立葉頻譜之振幅對應到符合人類聽覺的梅爾(mel)頻率尺度上，然後再對其作 DCT (Descrete Cosine Transform)來得到梅爾頻率的倒頻譜係數，即 MFCC [8]。

在此，我們對各音框擷取出 12 維的 MFCC 係數，並且訂定 3 個音框的距離，來計算另外 12 維的 MFCC 差分(delta)係數。不採取一般的 1 個音框的距離，是因為我們所做的辨識實驗顯示，3 個音框比起 1 個音框的距離，可以讓辨識正確率大幅度提升。

3. 高斯混合模型(GMM)訓練

在擷取出 MFCC 特徵向量後，如果將特徵向量適當地分成一定數量的群，則每一群大致會代表一種語音特性，例如音素(phoneme)。由於各種語音特性在三種語言中的出現頻率和次序皆不同，故可以使用統計模型來辨認待測的語音最接近何種語言。在此我們採用高斯混合模型(Gaussian Mixture Model，GMM)來模式化各種語音特性出現的機率，並以經過表徵器後得到的表徵(token)的n-gram 來模式化各種語音特性出現的前後關係。

高斯混合模型的概念是，利用多個高斯機率分布的組合，表達出各種語音特性的分布狀況。在GMM 中，第 j 個高斯機率分布的公式如下：

$$D_j(x) = \frac{1}{(2\pi)^{\frac{D}{2}}\left|C_j\right|^{\frac{1}{2}}} \exp\left(-\frac{(x-\mu_j)^T C_j^{-1}(x-\mu_j)}{2}\right) \quad (1)$$

其中 Cj 為共變異矩陣(covariance matrix)，μj 為平均向量，Dj(x)表示向量 x 由此機率分布取得的機率密度。進一步把數個高斯機率分布用加權和的方式加總，就可得到高斯混合模型：

$$P(x \mid \lambda) = \sum_{i=0}^{K-1} w_j D_j(x) \quad (2)$$

其中 λ 為模型中所有參數的集合。我們採用一個高斯混合模型來模式化一種欲辨識的語言。訓練的目的是，使一種語言的所有特徵向量在該語言的模型

中的機率總和為最大：

$$\lambda = \arg\max_{\lambda d} P(x_1, x_{2,} \cdots x_n \mid \lambda_d) \tag{3}$$

其中 $x_1,x_2,\ldots, x_n$ 是表示所有的特徵向量，λ_d 表示所有有可能的參數。由於實作上不可能窮舉出所有的參數值來計算，所以我們採用 MLE(Maximum Likelihood Estimate)的迭代估計方法來尋找最佳的參數值 [4]。

K-Mean 為一種相當廣泛使用的分群演算法[9]，可以把訓練用的特徵向量分成 K 群，其優點是簡單、速度快，但略嫌粗糙。在此我們用它來計算出 GMM 的初始參數值，得到 K-Mean 分群的中心向量後，就可以令一群內的特徵向量屬於一個高斯分布，如此第 j 個高斯分布的初始參數值就可以計算如下：

$$\mu_j = \frac{1}{N_j}\sum_n (x_n \mid g(x_n) = j) \tag{4}$$

$$C_j^{m,n} = \begin{cases} \dfrac{1}{N_j}\sum_n ((x_n - \mu_n)^2 \mid g(x_n) = j), \text{ if } m = n \\ 0, \text{ if } m \neq n \end{cases} \tag{5}$$

$$w_j = N_j \Big/ \sum_{k=0}^{K-1} N_k \tag{6}$$

其中 $g(x_n)$表示 x_n 在執行 K-means 分群後所屬的群編號，N_j 表示第 j 群的成員個數，μ_j 表示平均向量，C_j 表示共變異矩陣，w_j 表示加權值。由公式(5)可看出，我們假設特徵向量各維之間互相無關，所以共變異矩陣其實是一個對角矩陣。

得到初始值後，就可用以計算訓練語料對現在模型的 Log Likelihood，Lg(λ)：

$$L_g(\lambda) = \log \prod_{i=0}^{N-1} P(x_i \mid \lambda) = \sum_{i=0}^{N-1} \log P(x_i \mid \lambda) \tag{7}$$

要求極大值，可將 L(λ)分別以 μ_j、C_j 和 w_j 作微分，再分別令微分值為 0 後，而推得如下公式 [4, 5]：

$$\beta_j(x_i) = \frac{w_j D_j(x_i)}{P(x_i)} \tag{8}$$

$$\hat{\mu}_j = \frac{\sum_{i=0}^{N-1} \beta_j(x_i) \cdot x_i}{\sum_{i=0}^{N-1} \beta_j(x_i)} \tag{9}$$

$$\hat{C}_j = \frac{\sum_{i=0}^{N-1} \beta_j(x_i) \cdot x_i x_i^T}{D\sum_{i=0}^{N-1} \beta_j(x_i)} - \hat{\mu}_j \hat{\mu}_j^T \tag{10}$$

$$\hat{w}_j = \frac{\sum_{i=0}^{N-1} \beta_j(x_i)}{N} \tag{11}$$

MLE 就是利用上述公式作迭代，來修正 λ，以找出讓 $L_g(\lambda)$最大的參數集合。每迭代一次，就去計算現在的 log likelihood 值，當該值超過一定門檻(如 100)，就結束訓練。

4. 基於 GMM 距離之 K-means 分群

前一節是對一個 GMM 模型求取最大 Likelihood 值的目標下，去作模型參數的估計，但是考慮下一節所要建造的 PPM 模型時，我們只會用一個高斯機率分布來表徵化一個音框的 MFCC 特徵向量，所謂"表徵化"，類似於向量量化，就是要以一個高斯機率分布的編號來代表一個特徵向量。所以當使用 GMM 來作為 PPM 模型的表徵器時，求取一個音框對於 GMM 模型整體的最大 Likelihood 其實是缺乏意義的，因此我們提出另一種方法，就是求取一個音框對於個別高斯機率分布的最大 Likelihood，也就是把公式(7)改成：

$$L_p(\lambda) = \log \prod_{n=0}^{N-1} \max_j D_j(x_n) \tag{12}$$

其中 Dj(xn)的定義如公式(1)。

由於一個音框只屬於一個高斯機率分布，故要求最大的 Lp(λ)，其實等價於將所有音框作最好的分群。因此我們再度使用 K-Mean，不過把距離公式改成以 GMM 機率來定義：

$$d(x_n, G_k) = D_k(x_n) \tag{13}$$

其中 G_k 代表第 k 個高斯機率分布。在 K-means 分群時，一個特徵向量要找的是離它距離最近的的群中心，但是使用公式(13)時，xn 要找的是，使得 xn 的機率達到最大的 G_k。

每進行一次分群之後，都可再用公式(4)(5)(6)來重新計算出 GMM 的參數值。我們就以如此方式來進行迭代訓練，直到相鄰兩次迭代的平均量化誤差比值大於門檻值(如 0.95)。

由前一節和本節的說明可知，我們會對一個語言的 MFCC 特徵向量訓練出兩個 GMM 模型，一個拿來測量 GMM 為基礎的聲學分數，一個拿來作 PPM 模型的表徵器。

5. PPM 模型訓練

GMM 模型可用來模式化各種語言特性單獨的出現機率，可是無法模式化它們之間的時間次序關係。因此我們採取 PPM 模型[9]來實作 n-gram 模型，以模式化語音特性的時間次序關係。由於

n-gram 模型的建立，需要使用一個音框序列所對應的表徵值序列，因此我們要先將各個音框的特徵向量表徵化(tokenize)成一個整數值，表徵化的公式為：

$$T_t = \arg\max_j D_j(x_t) \tag{14}$$

其中 x_t 表示時間 t 時的特徵向量，$Dj(x_t)$的定義如公式(1)，T_t 表示 x_t 的表徵值。也就是要將一個特徵向量轉換成一個編號，而那一個編號的高斯機率分布可以讓 x_t 的機率達到最大。

當 n-gram 的 n 為 2 時，也就是 bigram 模型，它計算機率的公式如下：

$$P(T_t \mid T_{t-1}) = \frac{\text{count}(T_{t-1}, T_t)}{\text{count}(T_{t-1})} \tag{15}$$

其中 T_t 表示時間 t 時的表徵，count(y)表示訓練語料中表徵或表徵序列 y 的出現次數。

訓練 n-gram 模型，就是要將各音框的表徵和它的前置表徵序列(context)填入 PPM 表，而 n 階的 PPM 表會有 n+1 層，每層存放訓練語料裡各種遇見過的前置表徵序列(context)和表徵 T_t相連出現的次數，一個例子如表 2 所示，表 2 裡，前置表徵序列和目前表徵連接成一個表徵序列，後面的 count 代表這個序列在訓練語料的音框序列中出現的次數。

表 2　PPM 表的例子

order-1			order-2		
context	Token	count	Context	token	count
40	5	7	30, 30	30	55
22	5	18	30, 35	30	2
17	8	22	30, 33	30	6
4	12	15	22, 14	14	21
…	…	…	…	…	…
	<Esc>	85		<Esc>	227

n-gram 模型有一個問題，就是在辨識階段時可能發生某個表徵序列不存在於表中，也就是該表徵序列在訓練階段沒出現過，那麼要如何計算它的機率。PPM 驅動之 n-gram 模型(簡稱 PPM 模型)解決此問題的方法是，使用一個想像的<Esc>脫逃表徵，當遇到最高階序列(如 4 階序列，即 4 個相連表徵形成的前置表徵序列)在 PPM 表裡找不到時，就去掉最遠的一個表徵，而到下一階(第 3 階)去找這個較短序列的出現機率，並乘上前一階<Esc>表徵的出現機率。若在這階還是找不到，就再乘上這一階的<Esc>機率並再作降階，直到 0 階為止。

在此我們採用 PPMc 之逃脫機率估計方法[9]來設定<Esc>表徵的機率，PPMc 設定<Esc>的 count 值等於「此層所有表徵序列的種類數」。實作上，我們計算 PPM 模型機率的遞迴公式如下：

$$P\big(\text{Esc} \mid X(t,r)\big) = \frac{N\big(\text{Esc}|X(t,r)\big)}{N\big(X(t,r)\big)} \tag{16}$$

$$P(T_t \mid X(t,r)) = \begin{cases} [1 - P\big(\text{Esc} \mid X(t,r)\big)] \cdot \dfrac{\text{N}(T_t \mid X(t,r))}{\text{N}(X(t,r))}, \\ \qquad \text{if N}(T_t \mid X(t,r)) \neq 0 \\ P\big(\text{Esc} \mid X(t,r)\big) \cdot \dfrac{\text{N}(T_t \mid X(t,r-1))}{\text{N}(X(t,r-1))}, \\ \qquad \text{otherwise} \end{cases} \tag{17}$$

其中 X(t, r)代表時間點 t 時前 r 個表徵組成的前置表徵序列，N(Tt|X)代表在前置表徵序列 X 下表徵 Tt 的出現次數，N(X)代表前置表徵序列 X 的總出現次數。

6. 測試實驗

由表 1 可知我們所收集的語料的來源和語者人數，各語言中 10 個語者的發音被拿來訓練 GMM 和 PPM 模型，而剩下的語者的發音，則可被用來作外部測試。詳細來說，我們從模型訓練的語料中隨機抽出一定長度的語音作為 Inside 測試的語料，而 Outside 測試的語料，則另外找三個未參加模型訓練的語者，一樣隨機抽出一定長度的語音作為測試語料。測試語料分為 3 秒、10 秒、45 秒等三種發音長度，3 秒和 10 秒長的，三種語言各別有 49 個測試音檔，45 秒長的，各語言則分別有 29 個測試音檔。

6.1 GMM 模型之混合數測試

我們先對 GMM 模型的混合(mixture)數量進行實驗，即對於聲道 GMM 的混合數及 GMM 表徵器的混合數進行組合實驗，而兩種模型(聲道 GMM 模型和時序關聯性 PPM 模型)算出的機率作組合時，我們把加權值設為 β_{GMM}=0.3，β_{PPM}=0.18，此加權值是在觀察只有其中一個模型誤判時，需要怎樣的比例才能讓加權後分數的正確率達到最高的結果。至於 PPM 模型裡的階數，在此設定最高階數為 4。

實驗後，我們得到表 3 及圖 3 的結果，其中 Mg 值代聲道 GMM 的混合數，Mp 值則代表 GMM 表徵器的混合數。由表 3 及圖 3 可看出，Mg 值對於辨識率不管在內部(inside)或外部測試都有明顯的影響，例如在 Mp=32 之 45 秒外部測試，當 Mg 值由 32、64 變化到 128，則辨識率會由 63.95%、70.16%，上升至 80.26%，也就是上升了 16.3%。但是，Mp 值對辨識率則無明顯的影響，例如在 Mg=128 之 45 秒外部測試，不同 Mp 值所得的辨識差異只有 2.7% (80.26% - 77.55%)。另外在語料的長度方面，一般來說時間越長則辨識率越高，例如在 Mp=32, Mg=128 之內部測試中，長度 3 秒、10 秒、45 秒的辨識率會由 91.25%上升至 94.28%及

表 3　GMM 模型混合數組合之實驗

Mp / Mg	Inside			Outside		
	3 秒	10 秒	45 秒	3 秒	10 秒	45 秒
32/32	80.8%	83.5%	84.4%	69.2%	64.3%	63.9%
32/64	83.5%	78.9%	93.2%	71.7%	71.1%	70.2%
32/128	91.3%	94.3%	95.2%	76.8%	77.5%	80.3%
32/256	95.9%	95.2%	96.6%	83.7%	79.6%	82.8%
64/32	79.8%	80.8%	91.8%	67.2%	58.3%	58.6%
64/64	80.4%	86.5%	93.2%	67.2%	63.0%	67.4%
64/128	90.2%	94.3%	95.2%	74.0%	75.4%	77.6%
128/32	82.8%	86.2%	94.6%	69.3%	67.3%	65.3%
128/64	86.5%	91.9%	95.2%	72.3%	73.7%	71.4%
128/128	90.6%	96.3%	95.9%	78.0%	78.5%	79.1%

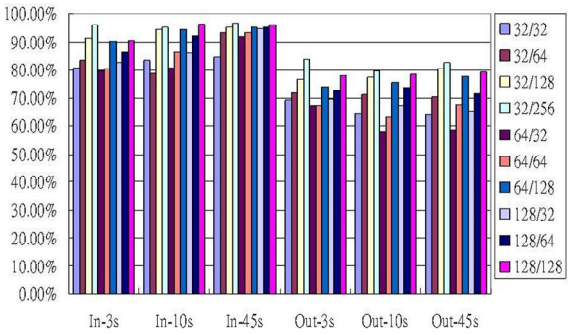

圖 3　GMM 模型混合數組合之實驗

95.24%，並且在聲道 GMM 的混合數足夠的情況下，10 秒和 45 秒的辨識率差異並不明顯。

一般來說，內部測試的辨識率可達 96.6% (Mp=32, Mg=256, 45 秒)，而外部測試的辨識率最高只可達到 83.7% (Mp=32, Mg=256, 3 秒)。

6.2 PPM 模型階數測試

在此我們只依據 PPM 模型算出的機率來作三種語言的辨識，並且設定 GMM 表徵器的混合數為 Mp=32。PPM 模型的階數代表條件機率 P(x|y)裡 y 序列所含的表徵個數，當階數越大，表徵 x 的機率值就會受到越多前置表徵的影響。實驗後，我們得到的辨識率如表 4 及圖 4 所示。

由表 4 及圖 4 可看出，PPM 模型的辨識率會隨著階數的變高而跟著上升。例如在 10 秒的內部測試中，PPM 模型由 1 階變成 5 階時，辨識率會上升 34%，而在 10 秒的外部測試中，PPM 模型由 1 階變成 5 階時，辨識率也相上升了 18.7%。此外，高階的 PPM 模型對於測試語料的長度也比較敏感，例如在內部測試中，1 階 PPM 模型對於 3 秒和 45 秒發音的辨識率幾乎沒有差別，但是，4 階 PPM 模型對於 45 秒發音的辨識率，則比 3 秒發音的高了 20.1%。

雖然實驗顯示 PPM 模型的階數越高，則辨識效果越好，但是每提高一階，辨識處理所需要的記憶體就會大幅增加，如在 Mp=32 的情況下，1 階 PPM 模型只需要約 50MB 的記憶體，但是 5 階 PPM 模型卻需要將近 1GB 的記憶體空間，且辨識速度也明顯下降。

表 4　PPM 模型之階數實驗

階數	Inside			Outside		
	3 秒	10 秒	45 秒	3 秒	10 秒	45 秒
1	49.8%	53.8%	49.7%	46.3%	48.7%	50.2%
2	58.5%	66.7%	66.3%	52.5%	59.2%	58.6%
3	63.3%	75.5%	92.9%	57.5%	61.2%	62.1%
4	76.2%	81.0%	96.3%	59.5%	63.9%	65.5%
5	77.6%	87.8%	98.6%	60.4%	67.4%	67.8%

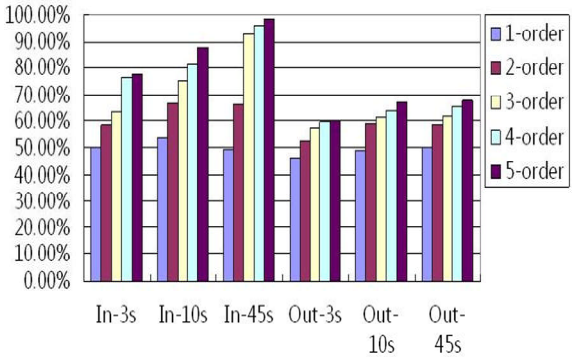

圖 4　PPM 模型之階數實驗

7. 結論

本論文針對國語、閩南語、及客語三者之間的語言辨識問題，研究了一種結合 GMM 和 PPM 模型的辨識方法。我們以 GMM 模型來掌握不同語言各自的聲學與聲道特性；此外，對各個音框的特徵係數作表徵化，再拿去建立 PPM 模型，用以掌握不同的語言裡，連續的表徵之間所存在的和語言相關的特性。

經由測試實驗的結果，我們發現聲道 GMM 模型裡，越多的混合數可以得到越好的辨識效果。但是，作為表徵器的 GMM，當混合數從 32 個增加到 128 個時，辨識率並沒有明顯的變化。另外，關於 PPM 模型的階數，實驗結果顯示，PPM 模型的辨識正確率會隨著階數的增加而有明顯的成長，但是所需的記憶體也是快速攀升。在內部測試的實驗裡，三種語言的平均辨識率可達到 95.9% (3 秒發音)和 96.3% (10 秒發音)；而在外部測試的實驗裡，平均辨識率則可達到 83.7% (3 秒發音)和 79.6% (10 秒發音)。

致謝

感謝國科會計畫 NSC 95-2218-E-011-006 的支援。

參考文獻

[1] K. Li and T. Edwards, “Statistical Models for Automatic Language Identification,” Proc. International Conference on Acoustics, Speech, and Signal Processing, 1980.

[2] M. A. Zissman, “Comparison of Four Approaches to Automatic Language Identification”, IEEE Trans. Speech and Audio Processing, Vol. 4, pp. 31-44, Jan. 1996.

[3] J. Navrátil, “Spoken Language Recognition—A Step Toward Multilinguality in Speech Processing”, IEEE Trans. Speech and Audio Processing, Vol. 9, Sep. 2001.

[4] 林俊青，多國語言辨識系統之特徵設計研究，碩士論文，國立中山大學電機工程研究所，2002。

[5] 張智傑，以高斯混合模型表徵器與語言模型爲基礎之語言辨認研究，碩士論文，國立清華大學電機工程研究所，2005。

[6] W. H. Tsai and W. W. Chang, “Discriminative Training of Gaussian Mixture Bigram Models with Application to Chinese Dialect Identification”, Speech Communication, Vol. 36, pp. 317-326, Mar. 2002.

[7] C. Y. Lin and H. C. Wang, “Language Identification Using Pitch Contour Information in the Ergodic Markov Model”, Int. Conf. on Acoustics, Speech, and Signal Processing, Jun. 2006.

[8] D. O’Shaughnessy, *Speech Communications*, 2nd ed., IEEE Press, 2000.

[9] K. Sayood, *Introduction to Data Compression*, 3rd ed., Morgan Kaufmann, 2006.