以高斯混合模型表徵器與語言模型為基礎之語言辨認

Language Identification based on Gaussian Mixture Model Tokenizer and Language Model

張智傑、王小川
Zhi-Jie Chang and Hsiao-Chuan Wang

國立清華大學電機工程學系
Department of Electrical Engineering, National Tsing Hua University
E-mail : piscesboy@micro.ee.nthu.edu.tw hcwang@ee.nthu.edu.tw

摘要

本論文探討不需要標注資料的自動化語言辨認方法，基本觀念是建立高斯混合模型之表徵器，以表徵器輸出建立語言模型，加上切割處理與後端處理，提升語音資料的語言辨認正確率。所建議的系統架構，分別是串聯高斯混合模型表徵器和語言模型的“高斯混合模型表徵器-語言模型法”，以及將語言模型融合在表徵器裡面的“連結聲學-語言模型法”兩種型式。由實驗結果觀察，加入切割處理的幫助，的確能夠提升系統的辨認正確率。

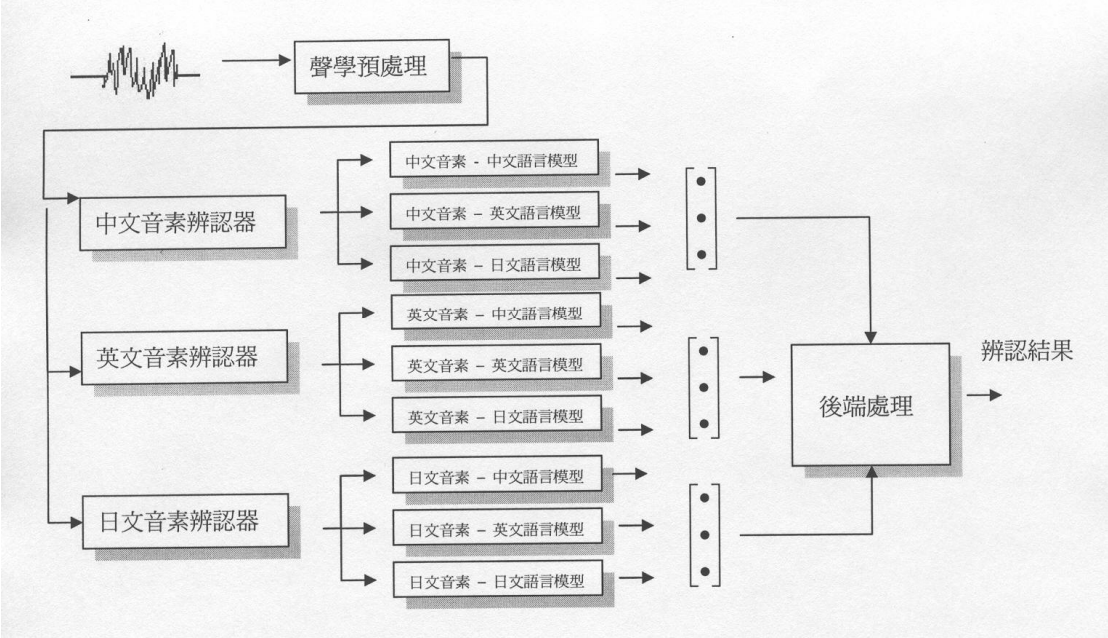
關鍵詞：語言辨認、高斯混合模型、表徵器、語言模型

一、緒論

近代語言辨認的方式，主要是對訓練語音資料，轉換成類音素（phone-like）序列，以類音素序列建立 N-連文模型作為語言模型。在做語言辨認時，計算測試語音之類音素序列與語言模型之間的相似度，經過後端處理做出語言辨認的判斷。所建議的系統有連結語言模型的音素辨認法（PRLM, Phone Recognition Language Model)[1][2]、連結語言模型的平行音素辨認法(PPRLM, Parallel-language PRLM)[1][2]、高斯混合模型表徵器-語言模型法（GMM-tokenizer-language model) [2][3] 、以及連結聲學-語言模型法（Joint-Acoustic Language Model) [4] [5] 等方式。

連結語言模型的音素辨認法[1]是將輸入語料經過預先訓練好的音素辨認器（phone recognizer），得出輸入語料的音素序列（phone sequence），再由音素序列統計產生語言模型(language model)。在辨認過程中，則是計算測試語音的音素序列與 N-連文法（N-gram)語言模型的相似度（likelihood），對應相似度最高的語言模型，就是辨認結果。圖一是以中英日三個語言的辨認為例，展示語言辨認系統之示意圖。輸入的測試語音，分別經由中英日三個語言的音素辨認器，產生三個不同的音素序列，將這三個不同的音素序列分別輸入到三個語言所建立的語

言模型，得出九個相似度值，後端處理器對這九個相似度值做運算，產生最後的辨認結果。



圖一、 連結語言模型的音素辨認法

表徵器-語言模型法的系統，需要有標註好的訓練語料做為音素辨認器訓練之用，要人工的介入才能完成系統建構；因此有研究者提出基本概念相似，但不需人工幫助的高斯混合模型表徵器-語言模型系統。其作法是將高斯混合模型的各個高斯機率密度函式(Gaussian probability density function)視為一個量化單位，給予模型中的每個高斯分布固定的表徵（token）值，將一個音框在各個高斯分布的機率值計算出來後，選擇機率最大的高斯分布作為表徵，視為此音框的代表值。對於輸入的測試語料，以高斯混合模型的表徵值序列（token sequence）取代連結語言模型的音素辨認法的音素序列（phone sequence），再做語言模型的處理。高斯混合模型表徵器-語言模型法在訓練語料充足時，實驗結果顯示有不錯的表現。

連結聲學-語言模型法是將語言模型合併到次字元的辨認器中，在決定音框的次字元時，加入語言模型的考量，以語言模型的機率值做為次字元間的轉移機率。使用連結聲學-語言模型法，可以降低次字元模型的數量 [4]，在次字元模型數更少的情況下，達到接近平行音素辨認法的效果。

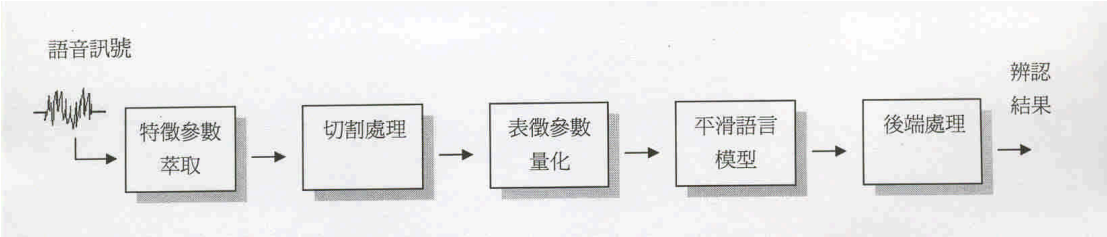
本篇論文的內容如下：第二節介紹本論文所使用的系統架構、流程，以及原理，第三節介紹所使用的程式工具，以及實驗的語料庫，第四節將實驗結果以圖表的方式呈現，並對其所顯示的現象加以討論。

二、 語言辨認之模型與方法

2.1 基本流程

語言辨認的基本流程如圖二所示，從語音訊號萃取出特徵參數後，經過切割成為類音素(phone-like)的聲音單位，經過表徵參數量化器（tokenizer)，將每個聲音單位轉換成一個音素代碼。對於一段輸入語音，即轉換成一序列的音素代碼，用以計算該段輸入語音與每一個語言模型

的對數相似度 (log-likelihood)。最後由後端處理器，對多個表徵器-語言模型(tokenizer –language model) 的相似度做處理後，得出最後的辨認結果。



圖二、 語言辨認的基本流程

2.2 位移差分化倒頻譜參數 (Shifted Delta Cepstrum)

倒頻譜參數由單一音框計算而得，並不包含發音腔道模型隨時間的變化特性。差分化 (delta) 便是將時間特性考慮進來，透過計算連續幾個相連音框的參數差值，表現出特徵參數的變動情況。本篇論文使用 HTK 所採用的回歸方程式 (regression formula) 來處理參數的差分化 ：

$$d_t = \frac{\sum_{k=1}^{K} k \cdot (c_{t+k} - c_{t-k})}{2\sum_{k=1}^{K} k^2} \qquad (1)$$

d_t 為差分化的特徵參數，k 為周圍音框和目前處理音框的距離，c_t 為音框 t 的特徵參數向量，K 為差分化音框 (delta window) 的大小。

位移差分化倒頻譜參數是將距離相同的多個音框的差分化倒頻譜參數結合起來，成為一個維度更大的向量，當做新的特徵參數來使用。一般使用四個關鍵值 (N, d, p, k) 來加以描述；

N : 單一音框計算出的倒頻譜參數的維度。

d : 差分化音框的大小

p : 串接差分向量的音框距離

k: 串接差分向量的個數

本論文中，關鍵值選用 (N, d, p, k) = (10, 1, 3, 3) ; 表示單一音框計算出 10 維倒頻譜參數，以其前後相連的兩音框計算出 10 維的差分化倒頻譜參數後，將間隔 3 音框的 3 組差分化倒頻譜參數串接成維度 30 的新向量，做為該音框的特徵參數向量。

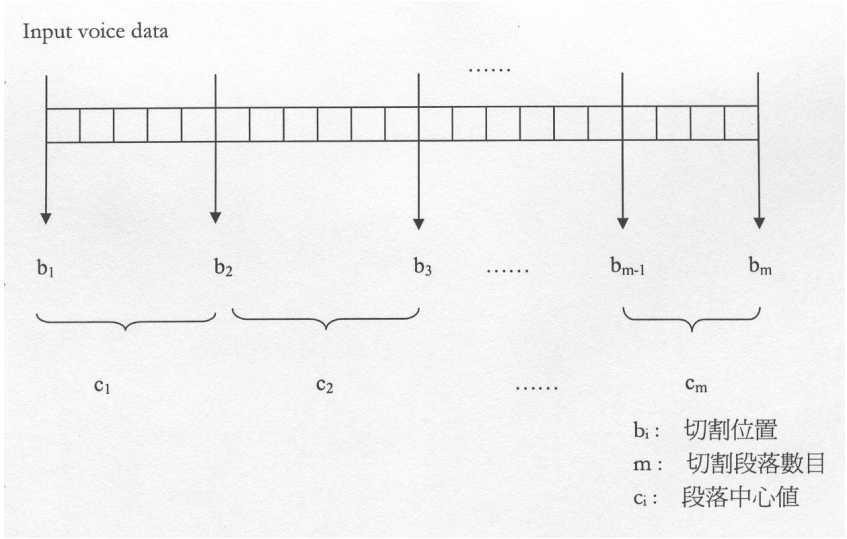
2.3 切割處理 (Segmentation)

將輸入的語音資料特徵參數序列，分割為預先指定的音段數目，稱做切割處理。音段的數目可以事先指定，也可以採用設定限制(constrained) 的方式讓系統自行決定。加入切割處理的目的在於取得類音素 (phone-like) 的單位作為語音段 (speech segment)。根據統計，各國語言音素出現的平均頻率為每秒 10 個音素。因此在本論文中，指定每秒切割出 10 個音段的方式。分割位置的決定，係以最小音段內差異(intra-segment distortion)的準則，讓每個分割出來的音段能夠有最大的聲學一致性 (acoustic homogeneity)。如圖三所示，輸入的語音特徵參數序列為 $X_1^T = (X_1, X_2, ...X_T)$ ， 若得出的分割點為 $B_0^m = (b_0, b_1, ..., b_m)$ ，計算每個音段的中心值 (centroid) $C_1^m = (c_0, c_1, ..., c_m)$ ，以及音段內各音框特徵參數和其中心值的距離，總和起來就

是該語音特徵參數序列的距離總和，

$$D(m,T)=\sum_{i=1}^{m}\sum_{n=b_{i-1}+1}^{b_i} d(X_n,\mu_i) \qquad (2)$$

設定分割點的目標，就是要使得距離總和為最小。



圖三、 特徵參數序列的分割

根據特徵參數選用的不同，各音段中心點以及最後距離總合的計算方式也應該跟著改變[6]。選用梅爾刻度倒頻譜參數時，中心點即為各音段中所有音框特徵參數的算數平均值，距離的計算則是採用歐氏幾何距離。分割位置採用動態程式規劃（Dynamic Programming）的方式，

$$D(i,b_i)=\min_{b_{i-1}}\{D(i-1,b_{i-1})+\Delta(b_{i-1},b_i)\} \qquad (3)$$

將所有可能的分割組合都加以嘗試。

2.4 語言模型（Language Model）

將量化單位給予一個類音素的代碼，於是一段語音就被轉換成一序列的類音素代碼。從各個語言的訓練語料，可以計算出每個類音素代碼量出現的機率。在做語言辨認測試時，即依據訓練語料統計出的機率，計算出測試語料的對數相似度。若一次選取連續 N 個語音段當作統計標準，則稱作 N-連文語言模型（N-gram language model）。以二連文法為例，根據各個類音素代碼的出現次數計算出狀態機率（conditional probability）：

$$P(w_t \mid w_{t-1})=\frac{C(w_{t-1},w_t)}{C(w_{t-1})} \qquad (4)$$

其中，w_{t-1},w_t 表示時間上相連的兩個類音素代碼，$C(w_{t-1},w_t)$ 為兩者共同出現的次數，$C(w_{t-1})$ 則為 w_{t-1} 出現的總次數。測試語料經由前端處理器轉換成類音素代碼序列

$W=\{w_0,w_1,...,w_T\}$ 後，輸入到各語言的 N-連文法語言模型 λ_l^{NG}，計算出其對數相似度

$$L(W \mid \lambda_l^{NG}) = \sum_{t=1}^{T} \log P(w_t \mid w_{t-1}, \ldots, w_{t-(N-1)}, \lambda_l^{NG}) \qquad (5)$$

做為語言模型的輸出結果。

如果其中某一個類音素代碼在訓練語料中沒有出現，可能會發生 $P(w_t \mid w_{t-1}) = 0$ 的狀況，使得測試時出現 log 0 這樣無意義的成份。解決的方法是以平滑化的 *N* 連文語言模型（smoothed N-gram language model）來取代原本的 *N* 連文語言模型。以二連文法為例，平滑化的二連模型為二連及一連模型的線性組合

$$\tilde{P}(w_t \mid w_{t-1}) = \alpha_2 P(w_t \mid w_{t-1}) + \alpha_1 P(w_t) + \alpha_0 \qquad (6)$$

其中，$P(w_t \mid w_{t-1})$ 為二連文法語言模型，$P(w_t)$ 為一連文法語言模型，α_2, α_1 為其對應的權重值，α_0 則為防止兩者皆為 0 的偏差值。α 的大小是由評估最大值演算法迭代求得。根據前人研究的結果 [1]，$0.3 < \alpha_1, \alpha_2 < 0.7$ 時系統會有最佳的表現。

在本論文中，選取的語言模型為平滑化的二連語言模型，對應的參數則選取 $\alpha_2 = 0.666, \alpha_1 = 0.333, \alpha_0 = 0.001$。另外為了避免 $P(w_t \mid w_{t-1})$ 分母出現 0 的情況，每個語音段量化單位出現次數至少為 1。

2.5 常用的系統組合與方法

A. 高斯混合模型表徵法（GMM Tokenization）

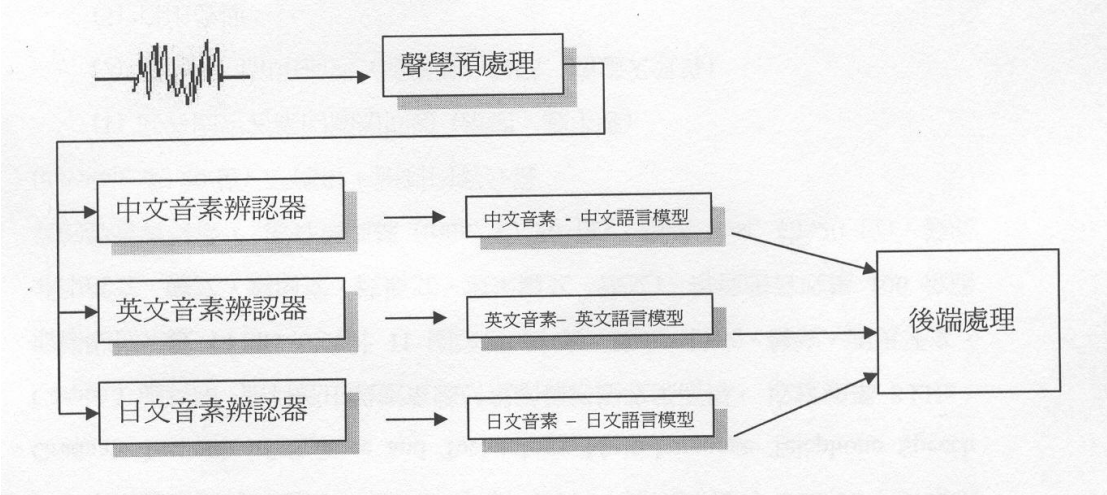
高斯混合模型表徵法是在高斯混合模型中每個高斯分布有一個固定的表徵（token）值，計算出單一音框在高斯混合模型中各個高斯分布發生的機率值，選擇機率最大的高斯分布的表徵做為此音框的代表值。其概念和向量量化（Vector Quantization, VQ）近似，不同的是，在向量量化中，我們取用和輸入音框距離最接近的中心值（centroid）的代號當做輸入資料的量化值。在高斯混合模型表徵法中，則是採用音框機率值成份最大的高斯模型的代號當作量化結果。

經過高斯混合模型表徵器 **（GMM Tokenizer）** 的處理後，輸入的特徵參數序列轉變成表徵序列 **（token sequence）**，於是就可以進行下個階段的語言模型處理。語言模型的大小決定於前端高斯混合模型表徵器混合數 **（mixture）**。混合數越大，表示原資料在量化時的單位越細，語言模型在統計時就必須記錄下越多種可能發生的語音段排列。以二連文法為例，混合數為 N 時，對應的二連文法語言模型大小為 N^2。

當訓練語料不夠多時，會使得許多成份出現機率太小，在測試端使用時就會讓相似度的變動範圍變小而難以做比較。在使用高斯混合模型表徵法時，混合數和語言模型的大小必須跟著訓練語料的大小、分布做調整。

B. 平行模式法（Parallel Model）

如圖四所示，平行模式是每個音素辨認器只連接和其語言相同的語言模型，目的在於彌補音素辨認器未考慮到的時間性質及順序問題。如果是採用高斯混合模型表徵法，則是將前端的音素辨認器改為高斯混合模型表徵器。



圖四、 平行模式法

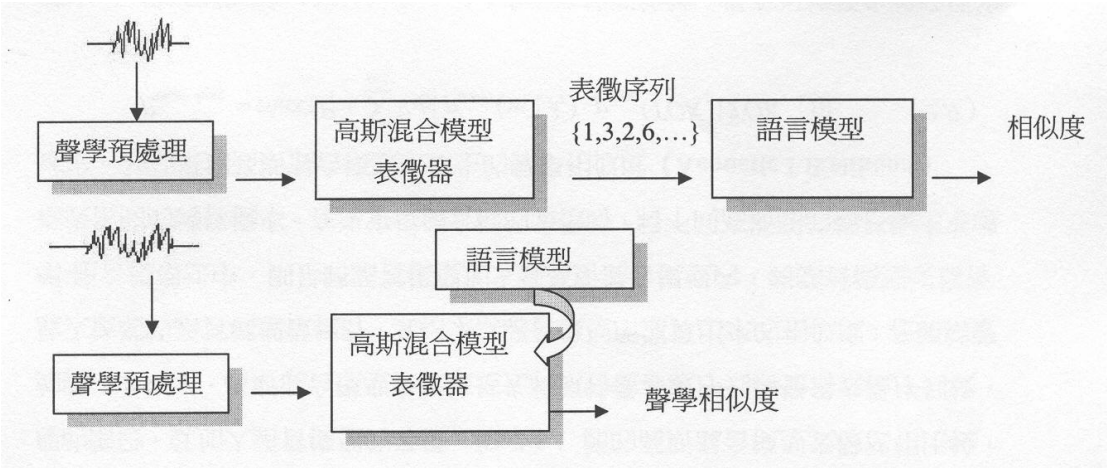
C. 連結聲學-語言模型法 (Joint-Acoustic-Language Model)

連結聲學-語言模型法的概念是，在做表徵序列 **(tokenization sequence)** 處理的階段，就加入語言模型的考量。高斯混合模型表徵法是先將特徵參數序列轉變為表徵序列後，作為語言模型輸入，得出對於該語言模型的相似度。在連結聲學-語言模型法中，則是將語言模型加入到高斯混合模型內，將語言模型視為音框表徵間的轉移機率。在決定每個音框的表徵時，將轉移機率考慮進去。其結果為高斯混合模型所產生的聲學相似度 **(Acoustic Likelihood)**

$$P_{token-lang} = \max_{\lambda}\{P_1 + \sum_{t=2}^{T} \log[\, p_{token}(w_t \mid \lambda_t) \cdot p_{lang}(TOK_t \mid TOK_{t-1})]\} \qquad (7)$$

其中，T 為音框總數，$\lambda = (\lambda_1, \lambda_2, ..., \lambda_T)$ 表示最有可能產生該音框的高斯成份序列，

$p_{token}(w_t \mid \lambda_t)$ 為音框 w_t 在 λ_t 的機率值， $p_{lang}(TOK_t \mid TOK_{t-1})$ 為語言模型，表示第 t-1 個音框表徵連接第 t 個音框表徵的發生機率，P_1 為第一個音框的最大高斯成份機率。



圖五、連結聲學-語言模型法

2.6 後端處理 (Back-end Process)

A. 專家投票法 (Voting)

專家投票法就是將所有要辨認的語言視為候選者, 每個語言模型的輸出視對候選者投下的選票。統計每個語言模型的輸出，以得票數最高的語言模型當做辨識結果。

B. 幾何平均法

使用平行音素辨認器時，某一個語言音素辨認器的輸出，只經過該語言音素所訓練的語言辨認器，產生一序列的相似度值。將各個語言所產生的相似度值做相乘後開次方，得出的結果即為測試語料和該國語言的相似度。

C. 算術平均法

和幾何平均法相同，得出一序列的相似度值，而後端處理改為對相似度值做相加後平均。由於語言模型的輸出結果為機率值乘積，取對數的結果

$$L_k = \log(\prod_{i=1}^{T} p_k(x_i)) \qquad (8)$$

將不同音素辨認器中同語言模型的部份相加，

$$L_1 + L_2 + L_3 = \log(\prod_{i=1}^{T} p_1(x_i)\, p_2(x_i)\, p_3(x_i)) \qquad (9)$$

在計算輸入音素的相似度時, 將不同的音素-語言模型組合視為彼此獨立的事件。

三、 實驗語料庫及使用工具

3.1 OGI-TS 語料庫

本篇論文使用的語料庫是 1992 年完成錄製的語料庫 OGI-TS (Oregon Graduate Institute of Science and Technology Multi-language Telephone Speech Corpus)[7]。取樣頻率為 8 kHz，取樣點位元數為 14 bits，共計有 11 國語言(中文、英文、日文、德文、西班牙文、北印度文、韓文、越南文、波斯文、坦米爾文、法文)。每種語言收集大約 100 位語者的聲音資料，這些資料分為訓練 (train, 約 50 位)、測試 (test, 約 20 位)、發展 (develop, 約 20 位) 三部份。語料內容包含

(1) 回答固定答案的簡短問題 (星期、數字等)

(2) 對問題之簡單描述 (最喜歡的餐點、描述天氣等)

(3) 自由發揮

3.2 語料庫之修改

在後半段的實驗中，我們從發展語料 (developing data) 取出部份檔案，將各語言的訓練語料填補至 85 分鐘左右的長度，如表一所示。

訓練語料只選取 45 秒的長句，每種語言 20 筆資料，測試語料長句數未滿 20 者，從發展語料填補。在 11 國語言辨認涉及偏差值 (bias) 的實驗中， 使用訓練語料內約 50 筆 45 秒的長句來做偏差值的計算。

表一、填補後的訓練語料

語言	訓練語料大小 (Byte)	語者數
中文	80,301,372	62
英文	81,481,458	56
日文	81,089,452	54

德文	80,817,724	52
西班牙文	80,724,618	51
印度文	81,031,972	107
韓文	80,559,128	59
越南文	80,481,152	61
波斯文	80,937,144	57
坦米爾文	81,040,500	61
法文	80,479,746	52

3.3 使用工具

特徵參數之萃取，採用劍橋大學提供的 Hidden Markov Model Toolkit (HTK)，版本為 3.2.1。輸入語料的取樣頻率是 8 kHz，經過 $(1- 0.97Z^{-1})$ 的高通濾波器做預處理。音框大小為 32 ms，音框間距為 16 ms，每個音框乘上漢明窗 (Hamming Window)。倒頻譜參數的臨界頻帶 (critical band) 數設定為 20, 並作倒頻譜參數均值刪除法 (Cepstrum Mean Subtraction) 處理。

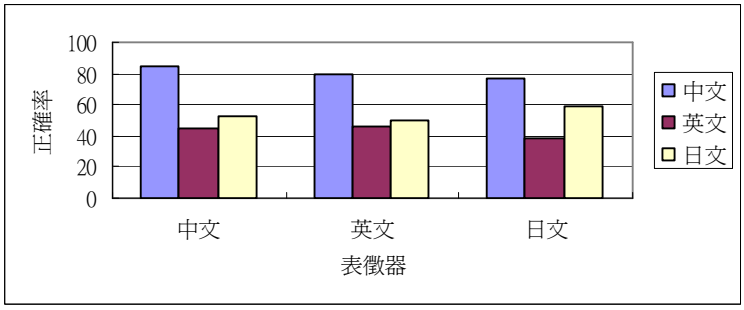
採用程式是建立在 C 語言上的 mat 2D 函式庫，版本為 1.8.1，作者為 Mike Schuster。該函式庫主要功能在處理向量及矩陣的基本運算、分類處理、基本統計模型建立、以及傅立葉轉換等語音處理經常使用到的演算。

四、 實驗結果及討論

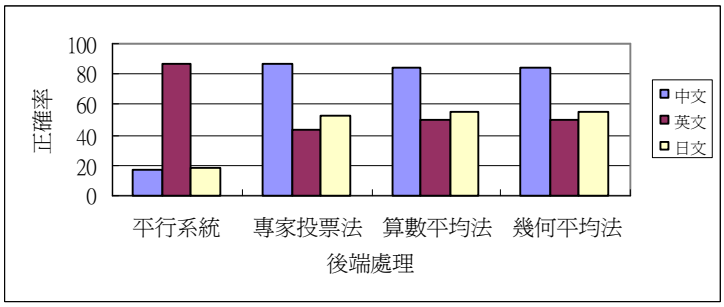
實驗的安排如下; 4.1 節與 4.2 節針對 OGI-TS 的三國語言 (中、英、日) 原本設定的訓練及測試語料做模擬實驗。對兩種不同的特徵參數 (30 階位移差分化倒頻譜參數、38 階梅爾刻度式倒頻譜參數)，作高斯混合模型表徵器-語言模型 (GMM-tokenizer language model)的語言辨認。在前端採用了單一表徵器及多表徵器兩種形式，也試驗了多種後端處理 (專家投票法、幾何平均法、算數平均法) 的方法。 4.3 節以連結聲學語言模型做為辨認系統，4.4 節則是將語料庫重新整理分類後，讓訓練語料大小相似，測試語料統一為 20 句 45 秒的長句。

4.1 中英日三國語言辨認 - 30 階位移式差分化倒頻譜參數

本節使用 30 階位移式差分化倒頻譜參數做為特徵參數，使用在單一表徵器及多表徵器兩種系統上。



(a) 單一表徵器



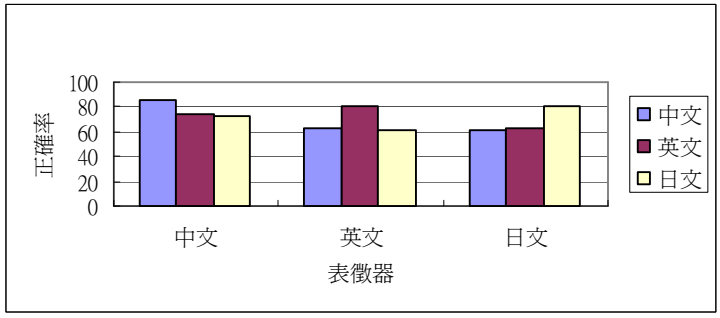
(b) 多表徵器

圖六、 使用 30 階位移式差分化倒頻譜參數之辨認結果

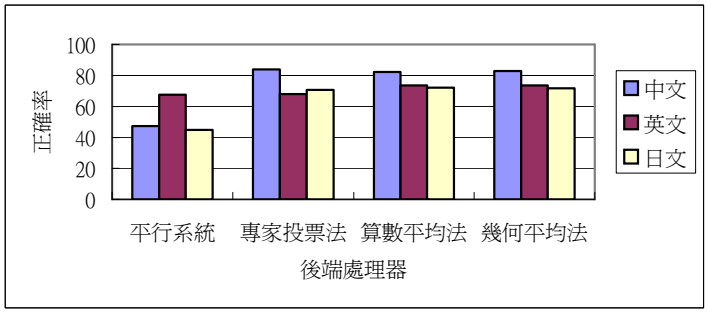
由圖六中發現， 使用位移式差分化倒頻譜參數時， 用單一表徵器及多表徵器的效果都不太理想，正確率大多不到五成。可能造成的原因是，位移式差分化倒頻譜參數將更多隨時間變化的因素加入參數萃取的步驟內，因此和語言模型相結合後，訓練語料必須要夠充足才能提供變化特性。參考文獻 [8] 使用的是比 OGI-TS 語料庫 (1.5 hr) 大 6 倍的 CallFriend 語料庫 (10 hr)，所以得到較高的辨認率。

4.2 中英日三國語言辨認 – 38 階梅爾刻度式倒頻譜參數

圖七顯示，使用梅爾刻度式倒頻譜參數，用單一表徵器時相同語言的語言模型辨認率較高，各種後端處理器的表現則相差不大，效能最好的是算術平均法，平均正確率為 75.92 % 。



(a) 單一表徵器



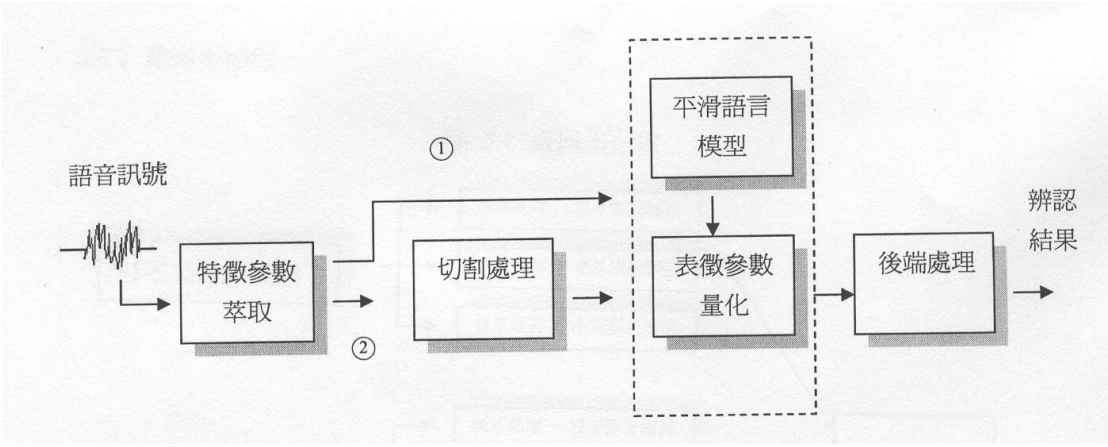
(b) 多表徵器

圖七、 使用 38 階梅爾刻度式倒頻譜參數之辨認結果

將圖六與圖七的實驗結果加以比較，38 階梅爾刻度式倒頻譜參數在後端處理為算術平均時，表現最好也最平均，平均正確率為 75.92 % 。因此我們選用該參數及算術平均之後端處理，應用在後面的實驗中。

4.3 中英日三國語言辨認 – 連結聲學語言模型

本實驗使用連結聲學語言模型做為辨認方法，由於使用這個方法可能會有偏差值的現象產生，即某些語言的相似度值會偏高，使得辨認結果偏向該國語言。使用的高斯混合模型之混合數為 32、64、128 和 256，參數採用的是 38 階梅爾刻度式倒頻譜參數。系統如圖八所示，實驗 A 未使用切割法，直接將特徵參數序列傳進表徵器內，實驗 B 則加入切割法，將特徵參數序列切割為事先指定的每秒 10 個單位，再傳入表徵器，表徵器的訓練語料也同樣經過切割處理。

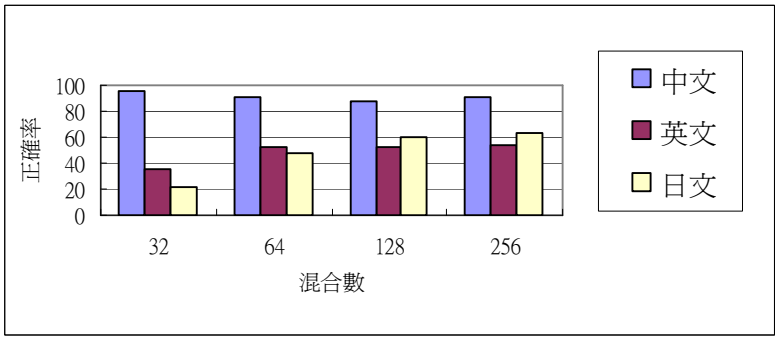


(1) 實驗 A　　(2) 實驗 B

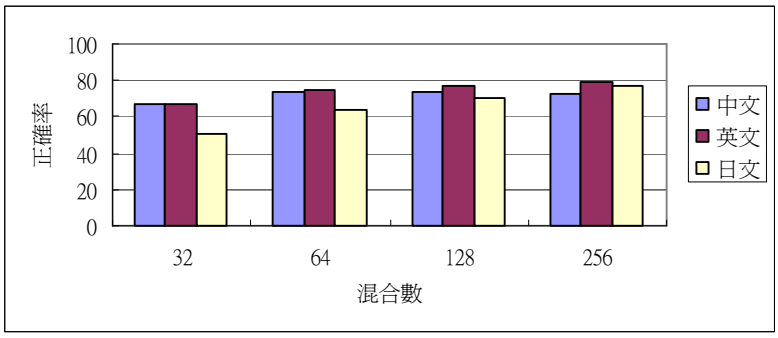
圖八、 切割處理與未切割處理之連結聲學語言模型法

(1) 實驗 A

圖九為針對測試語料做開放測試 (open test) 的實驗結果，以未經偏差值刪除的相似度做為辨認依據。



(a) 含偏差值



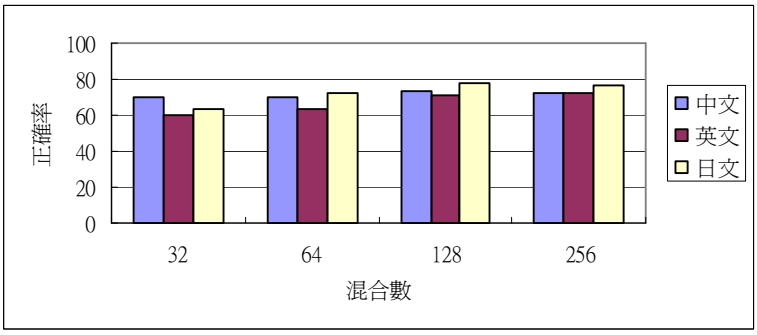
(b) 去除偏差值

圖九、 連結聲學語言模型之實驗結果(未使用切割法)

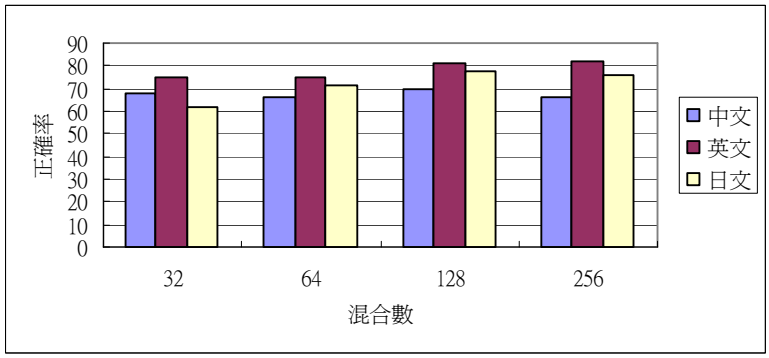
由圖中可看出，未除去偏差值時，中文語料的相似度有偏高的狀況，使得其它兩國語言的辨認效能很難提升，即使提高表徵器的混合數，也沒有顯著的效果。當除去偏差值的影響後，隨著表徵器混合數的增加，三國語言的辨認率都會提升，且會逐漸拉近，混合數 256 時有 76.27 % 的平均辨認率，和圖六的實驗結果 75.92 % 相比上升了一點。

(2) 實驗 B

本實驗對於做切割處理 **(segmentation)** 後的語料做語音辨認。



(a) 含偏差值



(b) 去除偏差值

圖十、 連結聲學語言模型之實驗結果(使用切割法)

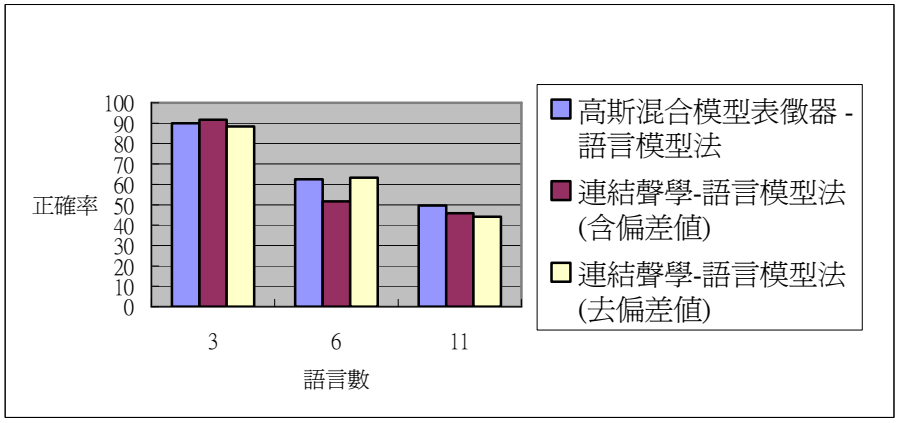
圖十顯示，混合數 128 時的表現和混合數 256 差不多，去除偏差值的情形下，在混合數 128 時平均正確率為 76.10 % 。對於切割語料，偏差值的考慮與否效果並不明顯。根據上述的觀察，

我們在接下來的實驗中保留偏差值的變因，並對切割過的資料採用混合數 128 的高斯混合模型。

4.4 使用修改語料庫之實驗

在 OGI-TS 語料庫中，日文比起中文和英文短少了約 15 分鐘的語料，這可能是日文辨認效能不高的原因。而測試語料也有長短不一的問題，較短的關鍵字詞語料大約 3 秒就結束，較長的自由發揮問題則有 45 秒的長度。為了做出較客觀的比較，訓練語料的大小和測試語料的長度就必須加以調整。

對於重新分類的語料庫，我們試驗了“高斯混合模型表徵器-語言模型法”、“連結聲學-語言模型法 (含偏差值)”、 以及 “連結聲學-語言模型法 (去除偏差值)” 三種方法。語料未經切割處理，傳入辨認系統的是 38 階梅爾刻度式倒頻譜的特徵參數序列。每種實驗分別測試三國語言 (中、英、日) 、六國語言 (中、英、日、德、西、印) 、 與 11 國語言 (中、英、日、德、西、印、韓、越、波斯、坦米爾、法)的辨認。

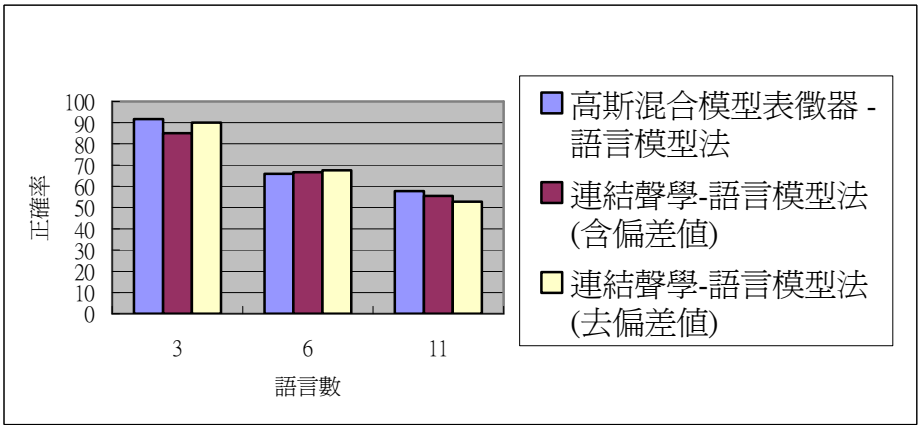


圖十一、 修改語料庫之實驗結果 (未使用切割法)

圖十一顯示，在未經切割處理的語料中，對於三國語言的辨認，效果最好的是連結聲學-語言模型法，平均正確率可以達到 91.67 % ，顯示出當訓練語料長度均衡，且測試語料長度夠的時候，對於這三國語言，可以有不錯的辨認效果。我們也發現，在三國語言中，將偏差值去掉後表現反而下降一點，表示加入偏差值的考量並不一定能提升系統效能。

對於六國語言的辨認，則有不同的現象發生，表現最好的是去掉偏差值的連結聲學語言模型法，正確率為 63.33 %，不考慮偏差值的系統則為 51.67 %， 相差 11 % 左右，顯示在六國語言辨認中，確實有幾個語言有相似度偏高的狀況。

在十一國語言辨認的實驗中顯示，單純採用特徵參數序列時，即使採用不同的系統處理，也很難提高正確率。三種方法的正確率都沒辦法超過五成，可能的原因是，所採用的混合數 128 高斯混合表徵器，將特徵參數間較細微的變化模糊掉。因此，接下來我們嘗試對特徵參數做切割處理，將相近的特徵參數先聚集起來，以其中心值為代表參數，再傳入表徵器做辨認。



圖十二、 修改語料庫之實驗結果 (使用切割法)

圖十二顯示，對於三國語言的辨認，經過切割處理的語料一樣有很高的辨認率，以高斯混合模型表徵器-語言模型法的 91.67 % 為最高。對於六國語言的辨認， 和未切割的實驗結果同樣是以去偏差值的連結聲學-語言模型法最高，有 67.5 % 的辨認效果，比起未切割的語料進步 4 % 左右。在十一國語言辨認的實驗中，加入切割處理能讓三種方法稍微提升正確率，有超過五成的正確率。效果最好的是高斯混合模型表徵器-語言模型法的 57.73 %。多數錯誤發生在英、德、西、法等四國同屬印歐語系的語言，以德文和法文之間的混淆最嚴重；中文和韓文也有蠻高的比例被辨認成越南文，顯示來源語系相近的語言會互相影響。

五、結論

本論文的主要目標在於尋找不需要標註資料的自動化語言辨認方法，所採用的系統是以高斯混合模型表徵器和語言模型為基礎，加上切割處理以及後端處理的輔助，處理語音資料的語言辨認工作。我們對串聯高斯混合模型表徵器和語言模型的“高斯混合模型表徵器-語言模型法”，以及將語言模型融合在表徵器裡面的“連結聲學-語言模型法”分別進行實驗。實驗語料採用 OGI-TS 語料庫，實驗之語言數分別是三國語言 (中、英、日) 、六國語言 (中、英、日、德、西、印) 、 與 11 國語言 (中、英、日、德、西、印、韓、越、波斯、坦米爾、法)。三國語言辨認實驗結果顯示，以 38 階梅爾刻度式倒頻譜參數，使用“高斯混合模型表徵器-語言模型法”，並加入切割處理，效果最好，有 91.67 % 的辨認率。在六國語言的辨認中，則是去偏差值的“連結聲學-語言模型法”表現最好，平均辨認率為 67.50 %。在 11 國語言的辨認上，則以 ”高斯混合模型表徵器-語言模型法”的 57.73 % 辨認率為最高。由實驗結果觀察，加入切割處理能夠使辨認效能稍微提升，表示在語言辨認的問題中，以音框為單位的特徵參數序列可能太過細微，如果能夠尋找聲學一致性更大、更為粗糙的單位來取代特徵參數，有機會能夠提升辨認的效能。

致謝

本研究受國科會專題研究計畫補助，計畫編號 NSC-93-2213-E-007-019。

參考文獻

[1] Marc A. Zissman, “Comparison of Four Approaches to Automatic Language Identification of Telephone Speech”, *IEEE Transactions on Speech and Audio Processing*, pp. 31-44, 1996

[2] Pedro A. Torres-Carrasquillo, Elliot Singer, T. P. Gleason, W. M. Campbell, D. A. Reynolds, "Acoustic, Phonetic, and Discriminative Approaches to Automatic Language Identification", *Proc. Eurospeech 2003*, pp. 1345-1348.

[3] Pedro A. Torres-Carrasquillo, Douglas A. Reynolds, J. R. Deller, Jr. , "Language Identification Using Gaussian Mixture Model Tokenization", *Proc. ICASSP 2002*, pp. I-757-760.

[4] A. K. V. Sai Jayram, V. Ramasubramanian, T. V. Sreenivas, "Language Identification Using Parallel Sub-word Recognition", *Proc. ICASSP 2003*, pp-81-84.

[5] Wuei-He Tsai, Wen-Whei Chang, "Discriminative training of Gaussian Mixture Bigram Models with Application to Chinese Dialect Identification", *Speech Communication*, vol. 36, pp. 317-326, 2002

[6] A. K. V. Sai Jayram, V. Ramasubramanian, T. V. Sreenivas, "Robust Parameters For Automatic Segmentation of Speech", *Proc. ICASSP 2002*, pp. I-513-516.

[7] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, " The OGI multi-language telephone speech corpus," *Proc. ICSLP 1992*, pp. II-895-898.

[8] Pedro A. Torres-Carrasquillo, Elliot Singer, Mary A. Kohler, Richard J. Greene, Douglas A. Reynolds, J. R. Deller, Jr. , "Approaches to Language Identification Using Gaussian Mixture Models and Shifted Cepstral Features", *Proc. ICSLP 2002*, pp. 89-92.