行政院國家科學委員會補助專題研究計畫成果報告

※※※※※※※※※※※※※※※※※※※※※※※※
※　　　　　　　　　　　　　　　　　　　　　　※
※　　　　利用樹狀決策架構製作關鍵詞辨認系統　　　※
※　　　　　　　　　　　　　　　　　　　　　　※
※※※※※※※※※※※※※※※※※※※※※※※※

計畫類別：☑個別型計畫　　□整合型計畫
計畫編號：NSC89－2213－E－009－119－
執行期間：88 年 8 月 1 日至 89 年 7 月 31 日

計畫主持人：王逸如

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

執行單位：國立交通大學電信工程學系

中　華　民　國　89 年 9 月 10 日

行政院國家科學委員會專題研究計畫成果報告
利用樹狀決策架構製作關鍵詞辨認系統
The Keyword spotting system using decision tree clustering

計畫編號：NSC89－2213－E－009－119－
執行期限：88 年 8 月 1 日至 89 年 7 月 31 日

主持人：王逸如　國立交通大學電信工程學系
計畫參與人員 : 游山瑞、謝寶華、周樂生

一、中文摘要

本計畫使用決策樹方式建立國語 411 音節之右文相關之聲韻母 HMM 模式作為一個關鍵詞辨認系統中之關鍵詞模式，並利用決策樹方式產生一套較粗糙之右文相關模式來做為填充模式，以期獲得較佳之關鍵詞辨認系統。
在計畫中，首先使用 MAT-2000(電話語音)作為訓練語料，利用決策樹建立右文相關之韻母辨認模式，對 500 句測試語料可獲得 67.3%的辨認率。在關鍵詞辨認系統之製作，關鍵詞辨認使用上述辨認模式，填充模式則使用一組較粗糙之右文相關模式，對一個以 1013 個人名為關鍵詞的電話號碼查詢系統上，可得到 86.43%的關鍵詞辨認率(FOM 為 0.38FA/KW/hr)。

關鍵詞：決策樹分類、關鍵詞辨認、相似度量測、關鍵詞確認

Abstract
In this project, we improved our previous HMM models, containing 100 RFD initial and 40 CI final models, by using a decision-tree clustering method to refine these 40 final models into a set of RCD models. Performance of the method was examined by experiments using the MAT-2000 telephone-speech database as the training data and a 500-utterance set as the testing data. A recognition rate of 67.3% was achieved. A keyword-spotting system was then constructed by using these RCD HMM models in keyword recognition with filler models formed from coarser CD models found by the decision-tree clustering method. A keyword recognition rate of 86.43% with FOM of 0.38FA/KW/hr was achieved in a telephone number inquiry system with 1013 keywords.

Keywords: Decision tree clustering、Keyword spotting、Similarity measure、Keyword Verification

二、緣由與目的

在經電話網路語音辨認應用系統，常是目的明確的簡單對話系統。因為在許多電話語音辨認的應用中，只要關鍵詞辨認正確即可。一個成功的關鍵詞辨認系統可以讓許多語音辨認之應用成真。

三、研究報告內容

本計畫首先利用決策樹方法來建立 HMM 國語語音辨認模型。HMM 辨認系統中採用『梅爾倒頻譜參數』（Mel-ceptrum feature），並且是混合高斯的由左至右隱藏式馬可夫模型（Mixture Gaussian HMM），加入通道和語者效應的消除（signal bias removal），而以加瑪（Gamma）分布模擬模型中之狀態長度分布。

1. 決策樹之原理與製作步驟

利用決策樹的分裂產生右文相關的韻母模型。這種結合語音學和語言學的方法使得相關模型的產生較為客觀，以下是整個決策樹分裂的流程：

(A)根節點的選取

針對我們所要處理的聲母或韻母，把其特徵參數集合起來成為一個根節點(root node)。所以如果我們有 40 種不相關韻母模型，則有 40 個根節點，每個節點皆可自行長出一棵樹，所以共可長出 40 棵決策樹。

(B)問題集的選取

根據韻母右邊所接的聲母，利用問題集裏的問題，把根節點分成符合此問題跟不符合此問題兩群資料，求出分成兩群資料與原來一群資料間所相對應之相似度改變量（ΔL）。以ㄥ＋ㄅ及ㄥ＋ㄍ為例，見圖 1。

對於問題集的選取，是採用中文的發音特性【1】，來訂出合適的問題【2,3】，並配合中文之聲母、韻母特性把問題集做一個整理。

(C)分裂的標準

一個母節點利用一個問題可分裂成兩個的子節點時，其相似度（maximum likelihood）的變化以ΔL 代表。對不同問題如果ΔL 大，表示

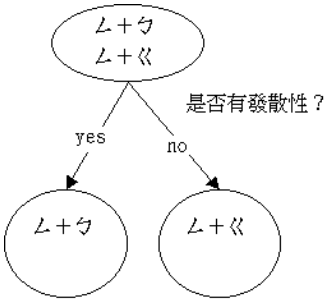


圖 1 決策樹利用問題做資料分裂示意圖。

此問題可以把母節點內語料分成兩個差異性較大的子節點。每次分裂都是以擁有最大的相似度變化的母節點來分裂，則可以想見，這棵決策樹最後剩下的末節點（left node）**彼此間的相似度變化應是最大的。不同模型間彼此差異性大，也表示所得的每一個模型其所含特徵參數的分布是最相似的【**4**】。**

在此計畫中採用的相似度改變量有兩種：

(1)**假設特徵向量呈高斯函數分布，則相似度的比值**(likelihood ratio) a**，表示當一個高斯分布分成兩個高斯分布時相似度的變化。而且** a **是對平均的相似度比值和變異量的相似度比值的乘積【**5**】：**

$$\Delta L = -(\log a_{\text{cov}} + \log a_{mean}) \qquad (1)$$

$$a_{mean} = (1+\frac{n_1 \times n_2}{n^2}(u_1-u_2)^t W^{-1}(u_1-u_2))^{\frac{-n}{2}};$$

$$a_{\text{cov}} = \left(\frac{|\Sigma_1|^{\sim}\cdot|\Sigma_2|^{(1-\sim)}}{|W|}\right)^{n/2};$$

$$W = \frac{n_1}{n}\Sigma_1 + \frac{n_2}{n}\Sigma_2, \quad \sim = \frac{n_1}{n};$$

其中 u_i**、**Σ_i **及** n_i **是兩個子節點資料特徵參數的平均向量、變異量矩陣及資料數目;其中變異量矩陣假設為** diagonal**。**

(2)**如果根據一些假設條件經過適當的簡化，可以得到下面公式：**

$$\Delta L = \frac{(n_1+n_2)\times\log|\Sigma| - n_1\times\log|\Sigma_1| - n_2\times\log|\Sigma_2|}{(n_1+n_2)} \qquad (2)$$

(D)**停止分裂的條件**

停止分裂的條件有二【6**】：一為節點內的音框數太少;一為最大相似度變化(**ΔL**)過小。**

(E)**合併**

如果最後所得的模型數太多，則對所有葉節點(leaf nodes)**找出其分裂時之**ΔL **小的兩個節點合併回去，直到模型數符合所需。**

2. 利用決策樹製作右文相關模式之辨認效果

訓練語料是使用「台灣之國語語音資料庫」（MAT**）中的第四（短句詞彙）與第五（平衡長句）部份，總共約** 48 **萬個音節。測試語料是** TEST-500**，是** 1998 **年語音辨認評比時採用之自行測試語料，其中包含** 50 **句單音節、**150 **句短句詞彙及** 300 **句平衡長句，總共有** 4731 **個音節。**

基本系統是以100**個韻母相關的聲母模型，**40**個不相關韻母模型（**independent final models**）做辨認，基本系統的辨認率為**63.94%**(插入及刪除型錯誤率個為**1.84%**及**1.14%**)。其餘兩組實驗則是把韻母利用決策樹方法做聲母相關的韻母模型。在表**1**中使用相似度改變量量測**(1)**及不同模型數目，可以發現模型數量越多，辨認率越高。在表**2**中則比較使用相似度改變量量測之辨認結果。**

表 1. **使用相似度改變量量測**(1)**及不同模型數目之辨認率。**

Number of RCD final models	Ins(%)	Del(%)	Sub(%)	Syllable correct rate (%)
Female:386 Male :293	4.0	0.7	28.7	66.7
Female : 293 Male : 293	3.7	0.7	29.4	66.3
Female : 240 Male : 240	3.6	0.7	29.6	66.1

表 2.**使用不同相似度改變量量測之辨認結果之比較。**

Measures and Number of RCD final models	Ins(%)	Del(%)	Sub(%)	Syllable correct rate (%)
Measure (1) Female:293 Male :293	3.7	0.7	29.4	66.3
Measure (2) Female : 290 Male : 290	3.8	0.6	28.3	67.3

3. 關鍵詞辨認系統

本系統是多關鍵詞系統。測試語料則從工研院電話分機查詢系統挑出 500 **句查詢人名的語料。共** 1144 **秒，包含** 239 **個音節。系統是以** 1013 **個人名為關鍵詞，加入光速搜尋法。在搜尋多條最佳路徑時採用** A* algorithm**，並以詞為單位【**7**】，做最佳路徑的選擇，辨認結果為詞組序列**(word sequence)**，而我們僅留下關鍵詞之資料，每一個詞組序列可能包含多個關鍵詞。**

在本計畫中之關鍵詞辨認器中，對關鍵詞部份使用 gender-dependent **的** context-dependent model **(**100 **韻母相關的聲母模型及** 250 **個聲母**

相關的韻母模型，40 個不相關韻母模型）。並且對辨認結果認為是關鍵詞的音框其辨認分數加一個正的嘉獎量。所使用之填充模式則為下列幾種精細程度不同的聲韻母模式：

(A)填充模型 A：100 個韻母相關的聲母模型，100 個聲母相關的韻母模型及 40 個韻母不相關模型組成 411 個音節填充模型。

(B)填充模型 B：100 個韻母相關的聲母模型，39 個不相關的韻母模型，組成 411 個音節填充模型。

(C)填充模型 C：29 個不相關的聲母模型，39 個不相關的韻母模型，組成 411 個音節填充模型。

對使用不同填充模式之關鍵詞辨認系統進行測試。在此光束搜尋法中的光束寬度設為 3000。

表 3. 利用不同填充模型所得之關鍵詞辨認結果。

	辨認率 %	假警報 (FOM)
填充模型 A top1	85.2	0.63
填充模型 A top10	92.4	0.34
填充模型 B top1	85.2	0.53
填充模型 B top10	93.4	0.44
填充模型 C top1	84.6	0.88
填充模型 C top10	89.2	0.63

由表 3 發現，模型數量的增加，可以得到一定的好處，使用較粗略的填充模型辨認效果不佳（填充模型 B、C）。因為用粗略的填充模型，在不能明確的描述下，非關鍵詞會吃掉一些關鍵詞的部份聲母，這都會造成關鍵詞辨認的困擾。不過模型數持續的增加，並不能帶來更多的好處，原因在於利用精細填充模型的辨認分數跟關鍵詞非常相近且關鍵詞的位置較為準確，此時嘉獎量只用來凸顯關鍵詞，嘉獎量太大則會使長詞優先，並且產生過多的假警報；太小則關鍵詞的分數會輸給由填充模型組合而成的辨認分數。粗略的填充模型辨認的分數低，而且不夠精確，所以嘉獎量大，可以使關鍵詞容易凸顯，並彌補關鍵詞前後因填充模型粗略所喪失的分數。

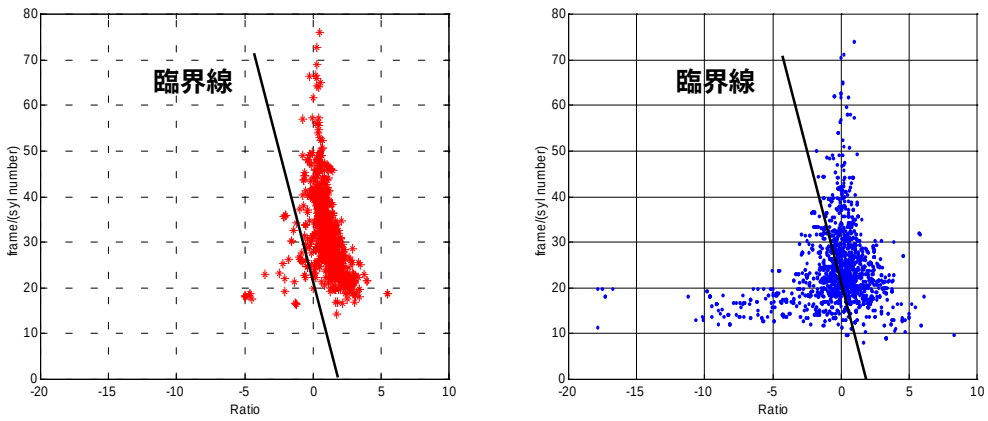
4. 關鍵詞的確認

在前節關鍵詞 Top n 之辨認，找出多條最佳路徑後，必須在做確認的工作(verification)。在此必須定義新的關鍵詞辨認分數，再重新挑出最佳的關鍵詞。

假警報的產生，往往都是因為短詞效應，原因在於系統對辨認成關鍵詞的部份加上一個嘉獎量，且輸入的語音要和一個短詞相似較為容易。結果造成原本沒有關鍵詞的地方，多辨認出一個關鍵詞，即假警報。所以假設真正關鍵詞每個音節的平均音框長度會比假警報來的長。另外，由於每個單音辨認的難易度不同，有些單音本身就很容易和別的音混淆，辨認分數低，因此絕對的分數沒有太大意義，應該要和此時其他填充模型的分數做比較，求出相對的分數，才能去除語者效應和音節辨認分數上的差異。

在關鍵詞辨認時，我們留下與關鍵詞切割位置最接近的填充模型字串之辨認分數，並考慮音框數不同的情況，關鍵詞與填充模式辨認分數均除以音框長度作為關鍵詞確認時之輸入參數。此外，因為假警報發生不但此短詞發生機率較高，而且常常音節長度也十分的短(會得到較高的 per frame 分數)。所以在此，定義了兩個參數作為確認步驟的輸入參數：

Score_Ratio = (關鍵詞分數/關鍵詞音框數) (非



(a) 關鍵詞特性圖　　(b) 假警報特性圖

圖 2. 關鍵詞與假警報的確認參數之分佈情形。

關鍵詞數/非關鍵詞音框數)

Lengthen = 關鍵詞音框數/關鍵詞音節數

在圖 2 中，標示了關鍵詞與假警報這兩個確認參數之分佈圖。我們可以訂定一個確認函式(臨界線)來參除部分假警報。從表 4 中也看到在設定臨界線後，Top-10 之辨認率雖然有些許下降，不過卻能刪除大量假警報。

	辨認率	假警報 (FOM)	拒絕率
原本 Top10	93.41%	0.44	0.002
經確認後	90.62%	0.25	0.004

表 4．經確認後假警報與辨認率關係圖。

接著我們要從 Top 10 關鍵詞組序列中找出最佳詞組序列，由於在每個詞組序列中關鍵詞之總長度不同，且每個音節的辨認分數都不一樣，其長度也有所不同，所以我們從訓練語料裏對每個音節求出每個音框之平均辨認分數及平均音框數，進而得知其平均長度的辨認分數。有了此項依據，就可以把各關鍵詞組的辨認分數正規化（normalization）再做比較。挑選一句裡面含有比平均長度辨認分數高最多的關鍵詞為最後結果。除了針對關鍵詞辨認分數，還可以加入關鍵詞長度跟一般平均長度的考量。首先，我們定義：*Score* 為關鍵詞的辨認分數；*Frame* 為關鍵詞的音框長度；關鍵詞之平均音框分數 $\bar{S} = Score / Frame$，$Score_{est}$ 為從訓練語料估計出的關鍵詞平均辨認分數；$Frame_{est}$ 為從訓練語料估計出的關鍵詞平均音框長度；關鍵詞之估計平均音框分數 $\bar{S}_{est} = Score_{est} / Frame_{est}$。由於辨認的分數是取對數後的結果，所以比較時是採用相減運算。實驗結果於表 5 中：

關鍵詞重排名之量測值	辨認率	假警報 FOM	拒絕率
$\bar{S} - \bar{S}_{est}$	82.24%	0.44	0.02
$(\bar{S} - \bar{S}_{est})$+Weight* (Frame/Frame$_{est}$)	86.43%	0.38	0.004

表 5．關鍵詞組重排名後之結果。

從上面的結果可以看出考慮平均長度分數的因素，效果較好，如此便去除了不同音節辨認分數不同造成比較的不公平。而對於關鍵詞的長度部份，對辨認分數加入一個跟平均長度有關的分數後效果確有改善。

5. 結論

在關鍵詞辨認問題中，較粗糙的填充模式，雖可以將非關鍵詞的辨認分數降低並降低假警報，但是隨之而來的是切割位置較不正確，使得關鍵詞的辨認分數下降。若將填充模型變的較精細，雖然關鍵詞的切割位置可以較正確，不過非關鍵詞的辨認分數將會提高，連帶的假警報也會隨之提高。由於利用決策樹，我們可以隨意的製造出想要的模型數，可訓練精細的模型做關鍵詞辨認，較粗略的當填充模型，使得填充模型可以有彈性的變化。在關鍵詞的確認方面，我們首先僅用關鍵詞部分的辨認分數刪除了一半的假警報。在利用正規化後之關鍵詞辨認分數做關鍵詞確認工作。

四、計畫成果自評

利用決策樹方式建立國語 411 音節辨認器。(2)建立一套較佳的關鍵詞辨認填充模式。(3)利用(2)之填充模式建立一關鍵詞辨認系統。本計畫中利用決策樹方式建立國語 411 音節辨認器結果將發表於 2000 年國際中文語言處理研討會(ISCSLP-2000)中【8】。

五、參考文獻

【1】 國立台灣師範大學國音教材編輯委員編纂，“國音學”，中正書局。

【2】 梁伯宇，“國語連續語音辨識之聲學模型研究”，國立台灣大學碩士論文，民國八十七年六月。

【3】 Po-yu Liang, Jia-lin Shen, Lin-shan Lee, “Class-Triphone Acoustic Modeling Based On Decision Tree For Mandarin Continuous Speech Recognition “, ISCALP-98.

【4】 Julian James Odell, “The Use of Context in Large Vocabulary Speech Recognition”, Ph.D tesis, University of Cambridge, U.K., March 1995.

【5】 A. Kannan, M. Ostendorf, and J. R. Rohlicek, “Maximum Likelihood Clustering of Gaussians for Speech Recognition,” IEEE Trans. on Speech and Audio Processing. Vol. 2, NO. 3, pp.453-455, July 1994.

【6】 L. R. Bahl, P. V. de Souza, P.S. Gopalakrishnan, D. Nahamoo, M.A. Picheny, “Decision Trees for Phonological Rules in Continuous Speech,” IBM Research, T.J. Watson Research Center, IEEE 1991.

【7】 Mei-Yuh Huang, Xue-Dong Huang, “Dynamically Configurable Acoustic Models for Speech Recognition,” ICASSP 1998, Vol. 2.

【8】 Yih-Ru Wang, and Ke-Shiu Chen, ' RCD Sub-Syllable Hmm Modeling By Decision Tree Clustering Using Mat-2000 Database, ', to be appeared in International Conf. on Chinese Spoken Language Processing 2000, Beijing, China, Oct. 2000.