第5章　改進英文辨識之各項實驗

5.1　鑑別性特徵擷取

資料相關線性特徵轉換(Data-Driven Linear Feature Transform)近幾年在語音特徵擷取的研究上佔有相當重要的地位，因為資料相關線性特徵轉換可以藉由統計訓練資料來自動地找出特徵空間中重要的基底向量(Basis Vectors)，使得經轉換後的特徵能保有重要的成份或具有較高的鑑別力，並且可以進一步去除多餘的維度，由於基底向量是根據訓練資料而來，所以找出的基底向量將較能代表語音訊號的特徵[Hung *et al.* 2001]。

本論文嘗試使用資料相關線性特徵轉換中的線性鑑別分析(LDA)、異質性線性鑑別分析(HLDA)，並結合最大相似度線性轉換(MLLT)方法於語音特徵擷取。

線性鑑別分析也可以以最大相似度(Maximum-likelihood)估測法[Campbell 1984]來解釋，作法為使用訓練資料的類別資訊來統計各類別的分布，求取一個轉換矩陣再藉此作線性轉換與特徵降維，其目的在於使得轉換後特徵之間可以保有最大的分類鑑別資訊。期望轉換後類別內(Within-Class)的分布越凝聚越好，而類別間(Between-Class)的分布距離越遠越好；也就是說轉換後，類別內的變異量越小越好，而類別間的變異量越大越好。

線性鑑別分析假設各類別分布的變異量相同[Campbell 1984]，然而現實上大多的訊號特徵分布的變異量皆為異質性。異質性線性鑑別分析[Kumar 1997][Kumar and Andreou 1998]假設各類別分布的變異量為異質性(Heteroscedastic)，去除鑑別性特徵係數法中各類別分布變異量相同的限制，同樣再以最大相似度估

測目標函式(Objective Function)，以求得較具鑑別性的語音特徵向量，詳細異質性線性鑑別分析推導可參見[張志豪 2005]。

本論文使用線性鑑別分析或異質性線性鑑別分析之特徵擷取，再加入最大化相似度線性轉換[Gopinath 1998；Saon *et al.* 2000]，因為目前我們使用的隱藏式馬可夫模型為對角化(Diagonal)之共變異矩陣，最大相似度線性轉換目的為保留矩陣維度，並使轉換後類別的共變異矩陣對角化。

本論文使用不同特徵擷取方式與語音強健性技術。VOA 語料的辨識結果如表 5-1-1 所示。觀察可知，VOA 語料之前端擷取利用 LDA 配合 MLLT 與 CMVN、高斯混合數目依規則分配、語言模型使用 VOA 訓練語料訓練，可得詞正確率(57.21%)。

表 5-1-1　VOA 不同特徵擷取法之辨識結果

實驗	語音特徵	混合數	詞正確率(%)	
			TC	WG
1	MFCC	78,412	46.95	48.75
2	MFCC_CMS	76,073	47.95	49.92
3	MFCC_CMVN	73,083	47.12	49.10
4	LDA+MLLT_ CMVN	70,672	51.82	57.21
5	HLDA+MLLT_ CMVN	71,627	49.72	52.95

另一方面，EAT 語料之語音特徵利用基礎與線性鑑別式擷取，結果如表 5-1-2 所示，觀察可知，EAT 語料之語音擷取利用 HLDA 配合 MLLT 與 CMVN、高斯混合數目依規則分配、語言模型僅為 EAT 訓練語料訓練，可得最佳詞正確率(59.71%)。

表 5-1-2　EAT 不同特徵擷取法之辨識結果

實驗	語音特徵	混合數	詞正確率(%)	
			TC	WG
1	MFCC	145,319	29.69	40.04
2	MFCC_CMS	143,735	36.41	49.53
3	MFCC_CMVN	138,713	33.93	47.02
4	LDA+MLLT_CMVN	138,289	47.30	59.53
5	HLDA+MLLT_CMVN	141,333	46.48	59.71

5.2　語言模型調適

統計式語言模型調適技術，為利用大量語料訓練背景(Background)語言模型，這些語料包含許多領域和主題，可以從中求得一般性(Genaral)的自然語言規則。另外再準備調適語料，此為包含較少量的語料，並與欲辨識語料相關，利用調適語料中所取得的資訊來調適背景語言模型。本論文利用 BNC 文字語料當成背景語言模型之訓練語料，並以訓練聲學模型之正確人工轉寫當作調適語料。

經由調適語料中取得的 N 連詞頻(N-gram Count)：N 連詞出現於訓練語料的次數，可透過不同方式來調適背景語言模型，以求得語言模型分數 $P(w_i \mid h_i)$ 之值，其中 h_i 是詞 w_i 的歷史詞序列。常見的語言模型調適法有詞頻數混合法(Count Merging)與模型插補法(Interpolation)。此兩種方法可視為最大事後機率(Maximun A Posteriori, MAP)[Bacchiani *et al.* 003]調適法的一種。

詞頻數混合法作用在詞頻數階級(Frequency Count Level)，算法如式(5-1)所示，其中 $C_B(\bullet)$ 表某個詞或詞序列在背景語料中出現的次數、$C_A(\bullet)$ 表在調適語料中出現的次數，α 與 β 可藉由期望值最大化(Eapectation-Maximization, EM)[Dempster *et al.* 1977]演算法求得。

$$P(w_i \mid h_k) = \frac{\alpha \cdot C_B(h_k w_i) + \beta \cdot C_A(h_k w_i)}{\alpha \cdot C_B(h_k) + \beta \cdot C_A(h_k)} \tag{5-1}$$

線性插補法作用在模型階段(Model Level)，算法如式(5-2)所示，其中λ代表線性插補法的係數，也可視為各個機率分布的權重。

$$P(w_i \mid h_k) = \lambda P_A(w_i \mid h_k) + (1-\lambda) P_B(w_i \mid h_k) \tag{5-2}$$

5.2.1 詞頻數混合法

表 5-2-1 為 VOA 詞頻數混合法之語言模型之辨識結果，所用之語音特徵為 LDA 加上 MLLT 配合 CMVN，高斯混合數目為 76,672 個，而語言模型調適法中之 α 與 β 值設定如表所示，觀察背景語料與調適語料混合，能比原先單獨使用調適語料的辨識率高，因 BNC 語料不僅包含與 VOA 統計特性較相關的會議或廣播新聞等文字語料，且 BNC 語料的量較大。1:100 59.14

表 5-2-1　VOA 詞頻數混合法之辨識結果

實驗	α	β	語言模型	詞正確率(%)	
				TC	WG
1	1	0	BNC	47.68	56.60
2	0	1	VOA	51.82	57.21
3	1	1	BNC+VOA	49.51	59.04
4	1	50	BNC+VOA*50	48.96	59.07
5	1	100	BNC+VOA*100	48.91	59.14

表 5-2-2 為 EAT 詞頻數混合法之語言模型對之辨識結果，所用之語音特徵為 HLDA 配合 CMVN，高斯混合數為依規則分配，共 141,333 個，語言模型調適法中之α與β值設定如表所示，觀察實驗結果發現，BNC 語料對 EAT 之辨識率提

升效果非常小，可能原因為 EAT 與 BNC 語料的統計特性差異很大，EAT 語料中大多為英文單字、片語或數字連續語音，而 BNC 為開會或是廣播新聞等對話資料。

表 5-2-2　EAT 不同語言模型之辨識結果

實驗	α	β	語言模型	詞正確率(%)	
				TC	WG
1	1	0	BNC	37.46	34.72
2	0	1	EAT	46.48	59.71
3	1	1	BNC+EAT	37.20	38.50
4	1	100	BNC+EAT*100	44.28	48.24

5.2.2　線性插補法

表 5-2-3 與圖 5-2-1 代表利用表 5-2-1 實驗 1 之以 BNC 為背景語料實驗之第一階段產生之詞圖資訊，代入詞三連機率之線性插補法，其中 λ 值為調適語料所佔比重。當調適模型與背景模型比重為 0.15 與 0.85，可得較佳的詞正確率 59.04%。

表 5-2-3　VOA 語言模型線性插補法之辨識結果

調適模型比重(%)	詞正確率(%)	調適模型比重(%)	詞正確率(%)
0.00	56.60	0.55	58.72
0.05	58.70	0.60	58.61
0.10	59.07	0.65	58.36
0.15	59.04	0.70	58.40
0.20	59.04	0.75	58.47
0.25	59.18	0.80	58.22
0.30	58.95	0.85	57.92
0.35	58.70	0.90	57.58
0.40	58.66	0.95	56.92
0.45	58.56	1.00	51.82

0.50	58.66	-	-

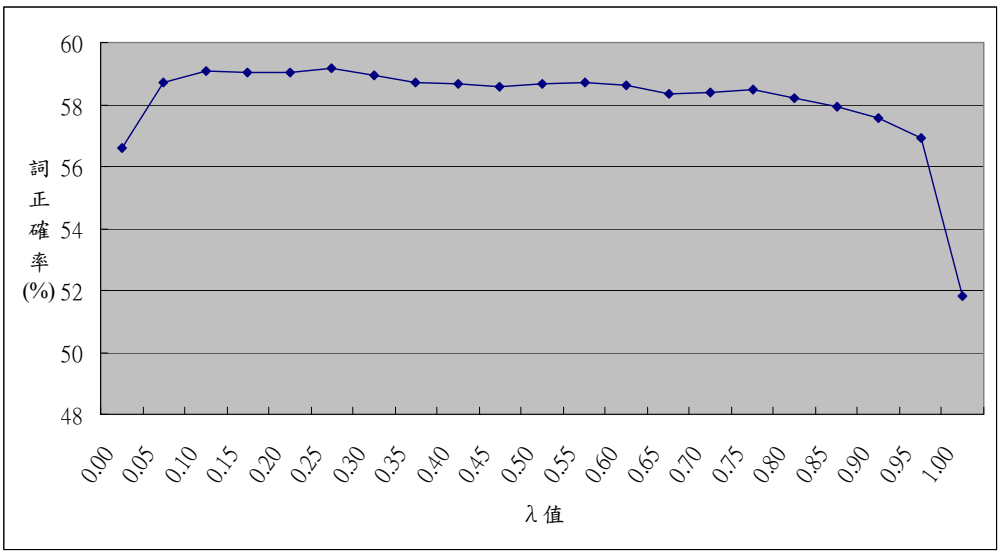

圖 5-2-1　VOA 語言模型線性插補法辨識結果示意圖

5.3　模糊矩陣之使用

本論文利用英文辨識器之第二階段辨識結果，與正確轉寫文字做單連音素、三連音素與詞之比對，統計發生「取代」的個數，利用模糊矩陣(Confusion Matrix)法統計並正規化(Normalized)容易辨識錯誤的個數。取出大於門檻值之辨識錯誤情況，將單連音素辨識錯誤統計結果，加入訓練聲學模型階段之決策樹問題條件，觀測辨識率之變化。以及將三連音素辨識錯誤統計結果，應用於辨識器搜尋階段，觀測辨識率之變化。

5.3.1　聲學模型訓練階段使用

此節實驗利用表 5-2-2 實驗 4 之 VOA 內測試(Inside test)語料，詞正確率為 78.05%之實驗做模糊矩陣之實驗，門檻值設定為 0.5，取出大於門檻值之辨識錯誤單連

音素統計結果，結果如表 5-3-1 所示：

表 5-3-1　VOA 內測試語料之模糊矩陣之單連音素辨識錯誤統計表

正確音素	辨識音素	次數正規化	正確音素	辨識音素	次數正規化
ng	ae	1.00	ih	ae	0.50
oy	ay	1.00	jh	b	0.50
s	z	0.67	jh	uw	0.50
z	s	0.60	l	w	0.50
er	r	0.57	p	t	0.50
ey	l	0.50	p	uw	0.50
ey	t	0.50	-	-	-

將表 5-3-1 之辨識音素變化應用於訓練聲學模型階段分裂決策樹之問題條件，重新訓練聲學模型與辨識，當門檻值設為 0.5，得到新的詞正確率為 78.10%；門檻值降低為 0 時，新的詞正確率為 78.07%，此規則對模型辨識率之提升影響力較小，可能原因為額外加入的問題條件占訓練語料之機率值過小。

5.3.2　辨識器搜尋階段使用

本論文統計語料之正確人工轉寫與辨識結果，找出編輯距離(Levenshtein Distance)中每個三連音素 M 「取代」(Substitution)成 $N_1...N_k$ 的次數正規化值，以 A_{MN_i} 表示，其中 $i=1...k$ 且 $\sum_{i=1}^{k} A_{MN_i}=1$。再將此模糊矩陣挑選門檻值大於 α 以上的結果，代入英文辨識器，重新計算每個時間點每個狀態的機率值，以 $\widetilde{B}_M$ 表示，$\widetilde{B}_M$ 計算方式如式(5-3)所示，其中 λ 代表原本三連音素 M 之狀態機率值所佔比列：

$$\widetilde{B}_M = \lambda B_M + (1-\lambda)\sum_{i=1}^{k}\left(A_{MN_i} \times B_{N_i}\right) \tag{5-3}$$

圖 5-3-1 第一列代表編號 10 號的三連音素容易被取代成編號 12 號，正規化次數為 0.5，其他變異如表所示。當 α 值設定為 0.4，則取出編號 10 的所有變異狀況於辨識器搜尋階段。

M	N	A_{MN}
10	12	0.5
10	15	0.5
12	16	0.4
102	140	0.4
:	:	:

$\alpha = *$

圖 5-3-1　模糊矩陣示意圖

本節實驗利用論文 5.4.2 之實驗設定，訓練聲學模型，利用辨識結果建立模糊矩陣。表 5-3-2 為 EAT 測試語料之單連音素辨識錯誤統計表，取出 0.2 以上的單連音素變化。

表 5-3-2　EAT 測試語料之單連音素辨識錯誤統計

正確音素	辨識音素	次數正規化	正確音素	辨識音素	次數正規化
z	s	0.38	ay	ax	0.25
sh	s	0.38	ay	t	0.25
jh	r	0.33	k	t	0.23
jh	t	0.33	uh	ax	0.23
zh	ax	0.33	m	n	0.23
zh	l	0.33	ao	ow	0.23
zh	sh	0.33	ch	n	0.22
aw	l	0.30	th	s	0.22
ng	n	0.29	b	f	0.21
d	t	0.27	l	r	0.20
aw	aa	0.25	iy	ih	0.20

表 5-3-3 代表將 EAT 測試語料之三連音素模糊矩陣變化，應用於辨識器階段

之詞正確率，α 全為 0，代表模糊矩陣找出之變異全部加入辨識器搜尋階段，λ 為 0.97 時有最佳詞正確率 62.98%。

表 5-3-3　EAT 測試語料之模糊矩陣應用於辨識器階段之詞正確率

		詞正確率(%)	
實驗	λ	TC	WG
1	-	50.14	57.84
2	0.20	39.25	48.15
3	0.50	40.84	52.36
4	0.80	52.14	62.85
5	0.97	52.87	62.98

如果使用大量 EAT 語料進行辨識，將其辨識結果與正確轉寫文字比對，建立一般化(General)模糊矩陣，再將此矩陣應用於辨識階段，詞正確率如表 5-3-4 所示。

表 5-3-4　EAT 一般化模糊矩陣應用於辨識器階段之詞正確率

			詞正確率(%)	
實驗	λ	α	TC	WG
	-		50.14	57.84
1	0.80	0.0	48.13	55.07
2	0.97	0.0	50.18	57.40
3	0.97	0.1	50.32	57.40
4	0.97	0.3	50.54	57.94

觀察實驗結果，當 α 為 0.3、λ 為 0.97 時有最佳詞正確率 57.94%，比原本 57.84 之辨識率提高 0.17%。

如果利用信心度評估(詳見論文 5.4.1、5.4.2)，選出辨識結果中的詞信心度為 1 的轉寫文字(26,810 句)，由此建立一般化模糊矩陣，再將矩陣加入辨識階段，詞正確率如表 5-3-5 所示。觀測當 α 為 0.3、λ 為 0.97 時有最佳詞正確率 58.41%，比原本 57.84 之辨識率提高 0.98%。並與表 5-3-4 比較，使用信心度挑選出的語句對詞正確率有提升效果。可知使用信心度評估法可挑選出適當語句且其變異狀況較有代表性。

表 5-3-5　EAT 一般化模糊矩陣應用於辨識器階段之詞正確率

			詞正確率(%)	
實驗	λ	α	TC	WG
	-		50.14	57.84
1	0.80	0.0	49.09	56.06
2	0.97	0.0	50.19	57.34
3	0.97	0.1	51.44	57.98
4	0.97	0.3	51.10	58.41

5.4　非監督式聲學模型訓練

過去研究語音辨識，語料的來源取得較為困難，因需人工特別去錄製，且需大量人力利用聽寫的方式，找出訓練語料對應的正確轉寫文字(True Transcription)及詞與音素之邊界。如圖 5-4-1(a)所示，稱其為監督式(Supervised)聲學模型訓練，此法須人工的介入，相當費力耗時。

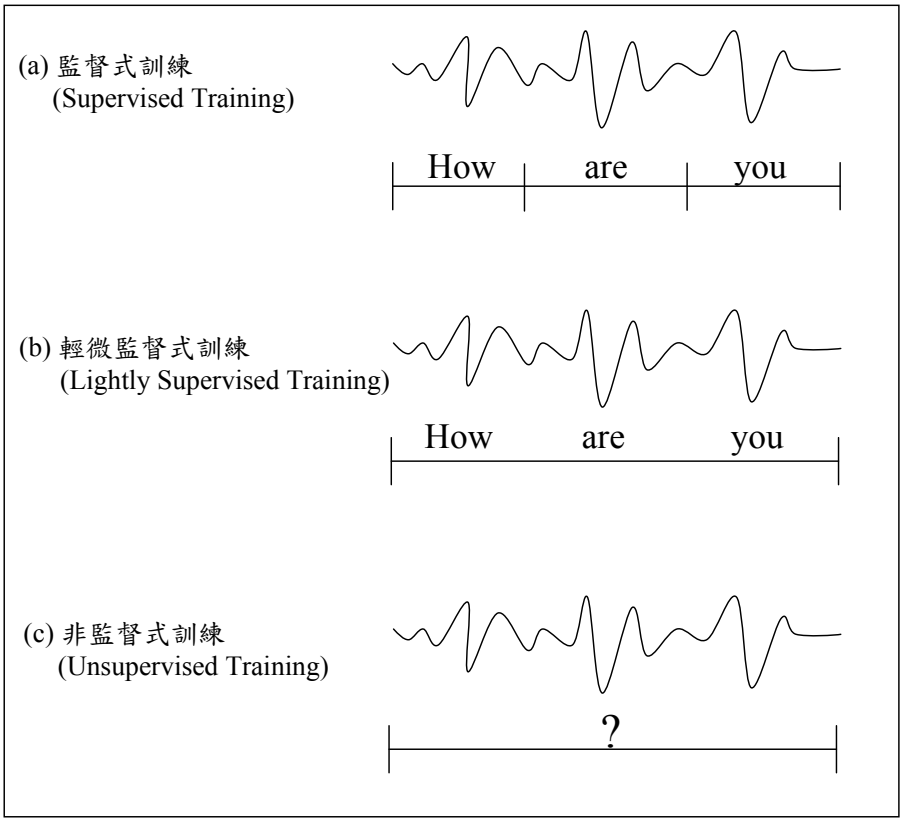

圖 5-4-1　三種訓練聲學模型方式示意圖

隨著科技發達，語料的來源可以有很多，使得大量的多媒體資料存在於網路中，例如：電視新聞、廣播新聞、多媒體影音等，使得收集語料不再是困難的工作。但是收集到的語料通常沒有正確的轉寫文字(如廣播新聞)，或者僅有近似的正確文字，如現場直播(Live)的電視新聞有人工即時(Real-time)將字幕(Closed-caption)打出，然而這些字幕可能並不視為完整的人工轉寫。利用字幕或是近似且沒有正確詞或音素邊界之語料來訓練聲學模型的方法，稱為輕微監督式(Lightly Supervised)聲學模型訓練[Lamel *et al.* 2002]，如圖 5-4-1(b)所示。如果只有語料而沒有正確轉寫文字和詞或音素邊界，稱其為非監督式(Unsupervised)聲學模型訓練[Wessel *et al.* 2001, Chen *et al.* 2004b]，如圖 5-4-1(c)所示。

對於大詞彙連續語音辨識研究中，訓練語料的多寡對模型的強健度佔了很重要的因素，例如三連音素聲學模型訓練實驗中，如果訓練語料量夠多，三連音素

平均出現的機會也增多，辨識率相對有提升之效果，但是在語料隨手可得的今日，我們仍無法有效地提升自動語音辨識器(ASR)效能，這是因為取得的大量語料可能不具有正確轉寫文字，且字幕取得也不容易，故如果想要有正確的轉寫文字或詞與音素的邊界，必須投入大量人力去標註，非常耗時。如果不想花費大量人力去轉寫正確的文字，可用現有的自動語音辨識器去辨識大量未轉寫之語料，再進行非監督式模型訓練，在大詞彙連續語音辨識研究中，非監督式模型訓練為一重要議題。

非監督式之聲學模型訓練通常利用最大化相似度(Maximum Likelihood)估測法來達成模型參數最佳化，作法為利用現有的人工轉寫過少量的語料，訓練初始聲學模型，利用此聲學模型對大量未經人工轉寫的語料做辨識，利用辨識後第一名(Top 1)辨識結果當成正確轉寫文字，再利用第一名之辨識結果的大量語料與現有之人工語料重新利用最大化相似度訓練法訓練聲學模型。

5.4.1 信心度評估法

非監督式聲學模型訓練中，利用初始之聲學模型對大量未經人工轉寫的語料做一次辨識，利用辨識後之詞圖產生第一名之辨識結果當成正確轉寫文字，再重新訓練初始之聲學模型。但是語音辨識總有辨識錯誤產生，如果拿錯誤的轉寫文字去訓練模型，會使訓練出的聲學模型不正確，降低辨識效能。信心度評估[Wessel *et al* 2001]法為判斷辨識結果的可靠度，給定辨識結果一個分數，如 0 至 1 之間的實數值，再設定門檻值，選出大於門檻值之語料與原本語料重新訓練模型。

本論文利用的信心度評估為事後機率法中的圖形化基礎(Graph-based)法來求詞圖中每個詞段(Word Arc)的信心度值。假設某一段語音特徵序列為O，詞圖Ψ^O中每個節點代表一個時間點，每個詞段(Word Arc)a由三個變數組成，

$a:[w_a;s_a,e_a]$，其中 w_a 代表為詞編號、s_a 代表詞段開始時間、e_a 代表詞段結束時間，每個詞段產生這個語音段落的聲學分數 $P(O_{s_a}^{e_a} \mid w_a)$，且每個詞圖有兩個特殊節點，分別為詞圖的開始與結束(如圖 3-13 之方形節點)，只要從開始節點到結束節點的任何路徑都可視為一條完整路徑(Complete Path)，而任一條完整路徑代表某一條聲學觀測值之辨識詞序列。在詞圖 Ψ^O 上利用前向後向演算法計算某詞段 $a:[w_a;s_a,e_a]$ 的事後機率如式(5-4)所示：

$$P(a:[w_a;s_a,e_a] \mid \Psi^O) = \frac{\sum_{\{\overline{W}:[w^n;s^n,e^n]_{n=1}^N\}\in\Psi^O, a\subset\overline{W}} \left\{\prod_{n=1}^{N} p\left(O_{s^n}^{e^n} \mid w^n\right)\cdot P\left(w^n \mid h^n\right)\right\}}{\sum_{\{\overline{W}:[w^m;s^m,e^m]_{m=1}^M\}\in\Psi^O} \left\{\prod_{n=1}^{M} p\left(O_{s^m}^{e^m} \mid w^m\right)\cdot P\left(w^m \mid h^m\right)\right\}} \tag{5-4}$$

其中，$\overline{W}$ 表在詞圖的一條完整路徑，共有 N 個詞段，$a\subset\overline{W}$ 代表包含詞段 a 的完整路徑 $\overline{W}$，h^n 為 w^n 的詞歷史(Word History)，$p\left(O_{s^n}^{e^n} \mid w^n\right)$ 代表開始時間 s^n 至結束時間 e^n 此段語音特徵序列的聲學相似度、$P\left(w^n \mid h^n\right)$ 代表語言模型分數。於實作時，求得每個詞段的信心度，再利用維特比(Viterbi)動態搜尋解碼得第一名之詞序列，此詞序列中的每個詞都有一個信心度值，再利用事先訂好的門檻值(Threshold)，來決定第一名詞序列中的某個詞是否拿來作聲學模型訓練，本論文挑選詞序列其信心度值全為 1 之句子。

研究指出[Wessel *et al* 2001b]，非監督式之模型訓練應以迭代方式實現，即以現有人工轉寫語料訓練之聲學模型，對未轉寫的語料做一次辨識，再將第一名辨識結果與現有人工轉寫語料再次訓練聲學模型，迭代(Iterative)過程可以有多次，且信心度評估於迭代過程中，應隨迭代次數而門檻值降低，因一開始之聲學模型之辨識率較低，所以門檻值設高，可過濾許多辨識錯誤語句，當幾次迭代後，可得到較佳的聲學模型。

5.4.2 實驗設定與結果

本論文利用 EAT 語料做非監督式最大化相似度聲學模型訓練，所使用的語音特徵、語料、詞典個數、語言模型設定如下表 5-4-1 所示。首先監督式訓練語料與測試語料同 4.2.1 節之實驗設定，另外再使用其他更多的英語系男生、女生，與非英語系男生、女生之語料混合，共 33.4 小時當非監督式訓練語料。而詞典個數為出現在監督式與非監督式訓練語料與測試語料中所有相異詞，共 4,229 個詞。

表 5-4-1　EAT 語料之非監督式最大化相似度聲學模型訓練實驗設定

語音特徵	HLDA+MLLT+CMVN			
實驗語料	種類	句數	時間(hr)	詞彙數
	監督式訓練語料	20,000	7.02	53,922
	非監督式訓練語料	42,960	33.4	108,323
	測試語料	1,000	0.65	2,781
詞典個數	4,229 個			
語言模型	監督式訓練語料之轉寫文字			

首先尋找非監督式之聲學模型上界，實驗流程如圖 5-4-2 所示。首先針對 7.02 小時之監督式語料訓練三連音素狀態分享模型、高斯混合數依據規則分配(表 4-1-2)，訓練之聲學模型為 HMM(1)(圖 5-4-2 中)。再將 7.02 與 33.4 小時監督式語料混合，依據訓練語料量重新分配並依規則分配高斯混合數目，與聲學模型 HMM(1)重新訓練成聲學模型 HMM(2)(圖 5-4-2 中)。HMM(2)為本節實驗詞正確率上界之聲學模型。詞正確率為 64.74%，如表 5-4-2 所示。

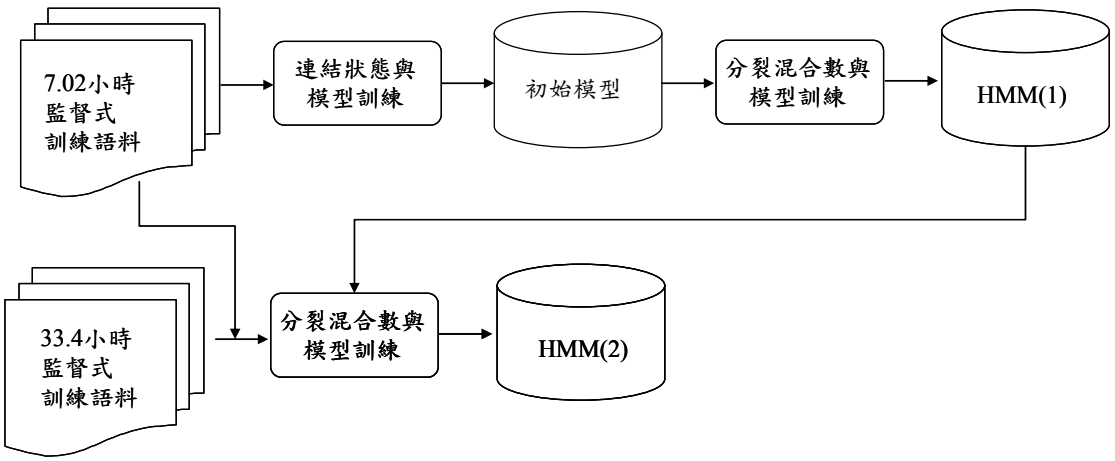

圖 5-4-2　非監督式之聲學模型上界

表 5-4-2　非監督式之聲學模型上界之詞正確率

實驗	聲學模型	混合數	詞正確率(%)	
-	-	-	TC	WG
1	HMM(1)	141,333	50.14	57.84
2	HMM(2)	216,318	56.29	64.74

非監督式模型訓練實驗流程如圖 5-4-3 所示。首先對 7.02 小時之監督式語料訓練三連音素狀態分享模型、高斯混合數依據規則分配，訓練聲學模型(圖 5-4-3 中 HMM(1))。再對 33.4 小時非監督式訓練語料做辨識，將辨識結果當成標記文字，與原本 7.02 小時訓練語料混合，重新分配高斯混合數並訓練聲學模型(如圖 5-4-3 中 HMM(3))，詞正確率為 51.73%(表 5-4-3)。如利用信心度評估法，尋找 33.4 小時辨識結果中詞序列之信心度全為 1 的轉寫文字，與原本 7.02 小時訓練語料結合，重新分配高斯混合數並訓練聲學模型(如圖 5-4-3 中 HMM(4))，詞正確率為 58.20%(表 5-4-3)。

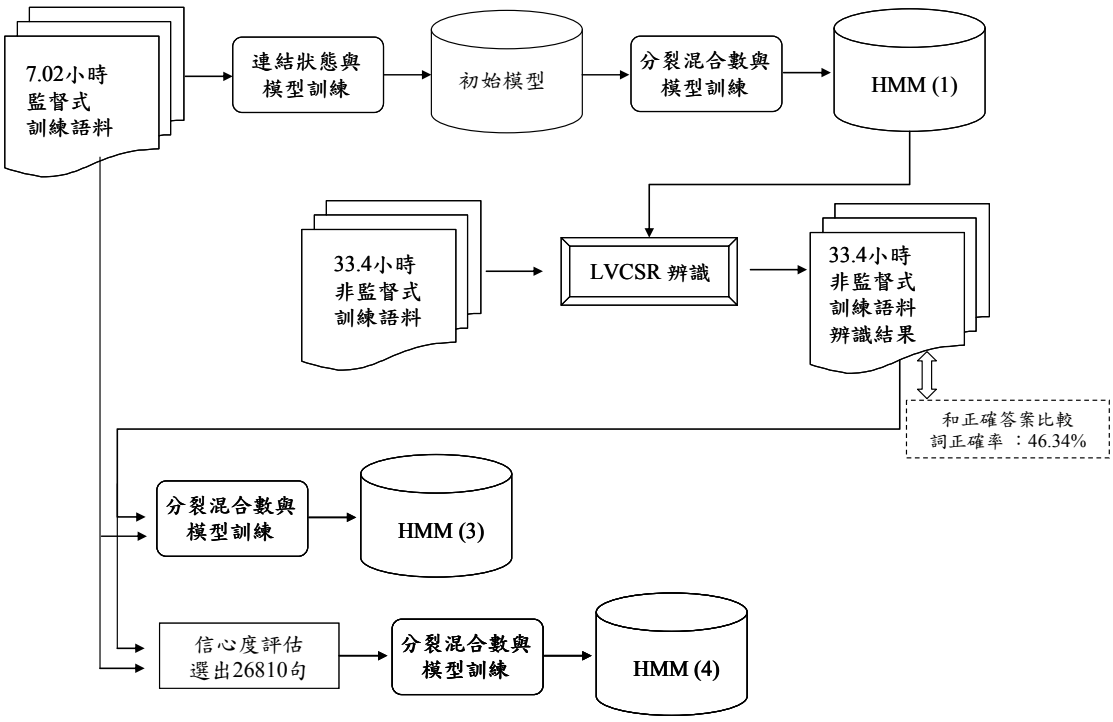

圖 5-4-3　非監督式之訓練示意圖

表 5-4-3　非監督式之訓練之詞正確率

實驗	聲學模型	混合數	詞正確率(%)	
-	-	-	TC	WG
1	HMM(1)	141,333	50.14	57.84
2	HMM(3)	221,820	49.78	51.73
3	HMM(4)	191,314	50.86	58.20

由實驗得知，33.4 小時之大量辨識結果與其正確轉寫文字比對，詞正確率為 46.34%，代表辨識結果含有 53.66%的錯誤比重，如果將全部辨識結果與 7.02 小時監督式語料之正確轉寫文字混合訓練聲學模型，將使詞正確率下降，因其中含有許多錯誤之轉寫文字，讓訓練的聲學模型不正確，降低辨識效能。如果以信心度評估法選出符合條件之語句，將使詞正確率提升 0.62%，代表信心度評估法能選出較正確之辨識語句，重新與原本訓練語料結合訓練模型，讓聲學模型更強健。

5.5 實驗討論

本節討論 5.1 至 5.4 實驗如下：

1. 鑑別性語音特徵擷取：

   A. VOA 語料之特徵擷取以 LDA 配合 MLLT 與 CMVN、高斯混合數依規則分配、語言模型使用 VOA 訓練語料訓練，可得較佳詞正確率(57.21%)。

   B. EAT 語料之特徵擷取以 HLDA 配合 MLLT 與 CMVN、高斯混合數依規則分配、語言模型為 EAT 訓練語料訓練，可得較佳詞正確率(59.71%)。

2. 語言模型調適：

   A. VOA 語料實驗，利用語言模型調適中的詞頻數混合法，以式(5-1)之α與β值，即 BNC 背景語料與 VOA 調適語料設定比重為 1 比 100 的情況下，語音特徵為 LDA 加上 MLLT 配合 CMVN，得到詞辨識率為 59.14%，觀察背景語料與調適語料混合，能比原先單獨使用調適語料的辨識率高，因 BNC 語料不僅包含與 VOA 統計特性較相關的會議或廣播新聞等文字語料，且 BNC 語料的量較大。

   B. VOA 語料之語言模型線性插補法實驗，將表 5-2-1 實驗 1 以 BNC 文字語料為背景語料，於第一階段辨識後詞圖產生的資訊，代入詞三連機率之線性插補，當調適模型與背景模型比重為 0.15 與 0.85 時，能讓詞正確率相對地提高 4.31%。

C. EAT 語料利用語言模型調適法中的詞頻數混合法，以式(5-1)之α與β值，即 BNC 背景語料與 EAT 調適語料設定比重為 0 與 1 的情況下，語音特徵為 HLDA 配合 MLLT 與 CMVN，得到詞辨識率為 59.71%，觀察 BNC 背景語料對 EAT 實驗影響非常小，因 EAT 語料中大多為英文單字、片語或數字連續語音，與 BNC 的統計特性差異很大。

3. 音素模糊矩陣之應用：

A. 尋找辨識結果與正確解答之音素取代情況，製作模糊矩陣並應用於訓練聲學模型階段之決策樹問題條件，觀測對辨識率之影響不大，可能原因為額外加入的問題條件占訓練語料之機率值過小，故不影響辨識率。

B. 應用音素模糊矩陣於辨識器搜尋階段，調整語音特徵向量之觀測機率，如果將辨識結果與正確解答比對建立後之模糊矩陣加入搜尋階段，可讓詞正確率提高。

4. 非監督式聲學模型訓練：

非監督式之聲學模型訓練結合信心度評估法，能選出非監督式訓練語料中較具正確性之語句，重新與原本監督式訓練語料混合訓練，讓聲學模型更強健。