國　立　交　通　大　學

電信工程學系碩士班

碩士論文

使用特徵參轉換之語音辨認與語者調適研究

A Study on Feature Transformation for Speaker-Adapted Speech Recognition

研 究 生：唐嘉俊

指導教授：王逸如　博士

中華民國九十三年七月

使用特徵參數轉換之語音辨識與語者調適研究

A Study on Feature Transformation for Speaker-Adapted Speech Recognition

研 究 生：唐嘉俊　　Student： Chia-Chun Tang

指導教授：王逸如 博士　　Advisor： Dr. Yih-Ru Wang

國立交通大學

電信工程學系碩士班

碩士論文

A Thesis
Submitted to Department of Communication Engineering
College of Electrical Engineering and Computer Science
National Chiao Tung University
In Partial Fulfillment of the Requirements
For the Degree of
Master of Science
In
Electrical Engineering
June, 2004
Hsinchu, Taiwan, Republic of China

中華民國九十三年七月

使用特徵參數轉換之語音辨認與語者調適研究

研究生：唐嘉俊　　指導教授：王逸如 博士

國立交通大學電信工程學系碩士班

中文摘要

本篇論文主要是探討特徵參數轉換方法對語者調適語音辨認的影響，我們以最小平均誤差(Mean Square Error)及最大相似度(Maximum Likelihood)為原則(Criteria)推導公式，以 MAT4500 語料庫 9:1 的比例為實驗的訓練及測試語料，並且使用測試語料中的長句為調適語料；實驗中觀察辨識率上限與調適語料為一句(4 秒)到八句時(約 37 秒)的辨識率；最後並分析在分群求取轉移函數的情形下，上限辨識率的改變和調適語料長短對分群的影響。歸納看來，特徵參數轉換方法可以有效去除語者/通道的差異而獲得較精準的 HMM 模型；轉移函數參數量越多時，上限辨識率越高，但在少量調適語料的情況下則越不理想，顯示出在調適語料有限的情形時，參數的估測有可能失去準確度而無法達到調適的效果。

關鍵詞：特徵參數轉換、語者調適、最小平均誤差、最大相似度

A Study on Feature Transformation for Speaker-Adapted Speech Recognition

Student： Chia-Chun Tang　　Advisor： Dr. Yih-Ru Wang

Department of Communication Engineering

National Chiao Tung University

Abstract

In this thesis, the effect of feature transformation in speaker-adapted speech recognition is exploited. Two criteria, minimum mean-squared error and maximum likelihood, are employed to formulate the feature transformation algorithm. Besides, the approach of using different transformation for three broad speech classes of initial, final, and silence is also studied. Effectiveness of the proposed method was examined by simulations using MAT4500 telephone speech database with 9/10 data for training and 1/10 for testing. Sentential utterances were used in the speaker adaptation test. The amount of adaptation date ranged from one utterance (4 seconds) to eight utterances (37 seconds). Experimental results showed that the proposed feature transformation method can eliminate the speaker/channel effect so as to make the HMM models more compact. We also found that, as more transformation parameters were used, the upper bound of recognition rate was better while the adaptation effect became worse for small adaptation data. This mainly resulted from the inaccuracy of parameter estimation when insufficient adaptation data were used.

Keywords: Feature transformation, Speaker adaptation, Minimum mean-squared error, Maximum likelihood

誌謝

兩年的研究生生活一轉眼就結束了，想當初大夥還在淡水忙著 ISCSLP 的籌備與服務，現在卻已經完成碩士研究並且即將踏上人生另一個里程。這雖然只是短短的兩年但是收穫卻是非常豐碩，首先要感謝我的指導教授王逸如老師及陳信宏老師讓我有機會進入語音處理實驗室。在陳老師及王老師的指導下，從原本對語音信號處理的陌生到熟悉，從做事的沒有條理到按部就班，從不熟悉軟體語言到三天完成一個複雜程式…等，真的都要感謝老師您們。然而不只於學術上的進步，在待人接物與處世上，你們也淺移莫化的給了我很多良好的典範。另外廖元甫老師在研究上亦給予我相當大的幫助，使得我的工作能更加順利。

兩年的時間幾乎都是在實驗室中度過，同學大家一起熬夜、一起打拼、一起打球、一起忙裡偷閒，使得研究這段路一點都不寂寞並讓我充滿感激。特別感謝郭威志學長、羅文輝學長及傅振宏學長，總是能在適當時候替我們打氣加油。還有我的鄰居王文德、小 z 及實驗室冷氣機江振宇，沒有你們的幫忙，我的程式技巧不會進步如此快速。另外實驗室模範生俊良、祺翰、智合也每每在我研究上有問題時給予我解答；睡神阿樹、苦瓜小孫更讓實驗室生活多采多姿，一點都不單調。謝謝你們，我只能說:有你們真好!還有學弟學妹，要加油喔，明年的語音處理實驗室就靠你們了，你們一定可以做的更好!

最後，要感謝我的父母及 little，沒有你們背後默默的支持打氣及讓我沒有後顧之憂的研究生活，我不會有今天如此的成就，更不會有這篇論文的產生。僅將這本論文獻給我最愛的家人及幫助過我的人。

目 錄

表 目 錄

圖 目 錄

第一章 緒論

1.1 研究動機

科技日新月益，越來越多的高科技產品如雨後春筍般的問世，然而這些產品的推出，無外乎是希望人們的生活更加的便利與豐富，因此具備隨身攜帶、隨時方便操作的功能便成為這些科技產品中重要的努力方向之一。若是在如此縮小及完備產品上裝置鍵盤或是厚重的輸入裝置，的確不是一個理想的辦法，而語言也正是我們人類最原始、最簡單、最自然、也是最方便的溝通工具。因此如何發展良好的語音辨識系統來作為人類與機器溝通的橋樑儼然成為非常重要的研究工作。

良好的語音辨識系統，不應該因為語者或是輸入通道的不同而影響辨識能力，也就是說，一個優良的語音辨識系統應該在任何一位使用者透過任何一種通道使用時，有能力去學習與適應這位使用者的語音特性並且消除背景雜訊干擾以提升此語者的語音辨識能力。因此語者調適(speaker adaptation)便成為我們研究的重要課題，希望藉由此方法來增加系統的強健性(robustness)。

1.2 研究方向

目前的語音辨識系統幾乎都是使用大量的語料且利用統計法來建立聲學模型，並藉由此聲學模型來模擬各種的語音特性。我們之所以需要大量的語料來訓練聲學模型，是希望各種情況下的聲音(例如:背景雜訊的不同，語者種族的不同…等)都能被考慮進入語者不特定模型(speaker independent model)中，藉以

盡量降低訓練與使用時不匹配的狀況發生。然而，就算訓練語料大到足以包含所有種族、所有發聲通道，我們還是沒辦法面面俱到地完全符合每個使用者的語音特性，這些語音的特性來自於語者間獨特的口腔構造、發聲習慣、講話腔調等先天上的差異。因此，我們可以利用使用者所提供的少數語料來修正、調整辨識系統，使得此辨識系統好像就是為此使用者所量身打造。我們稱此少數語料為調適語料(adaptation data)，這也就是語者調適的工作。

語者調適的技術，就調適基礎立論而言，可以分成兩大類，一種是以聲學模型為基礎之調適法(model-based adaptation)，另一種是以特徵向量為基礎之調適法(feature-based adaptation)。前者是目前較多人使用的調適技術，藉由新使用者的調適語料來將語者不特定模型調整修正成為語者特定模型(speaker dependent model)，這會使得調適過後的語音模型跟新使用者的語音特性相近，因此使用此調適過後的語音模型來辨識這位使用者，一定比使用眾人訓練出的語者不特定模型效果來的出色。貝氏調適法(Bayesian adaptation) 【1】【3】與最大相似線性回歸(Maximum Likelihood Linear Regression ，MLLR) 【2】【3】均是屬於此類調適法。以聲學模型為調適基礎之流程如圖 1.1 所示：

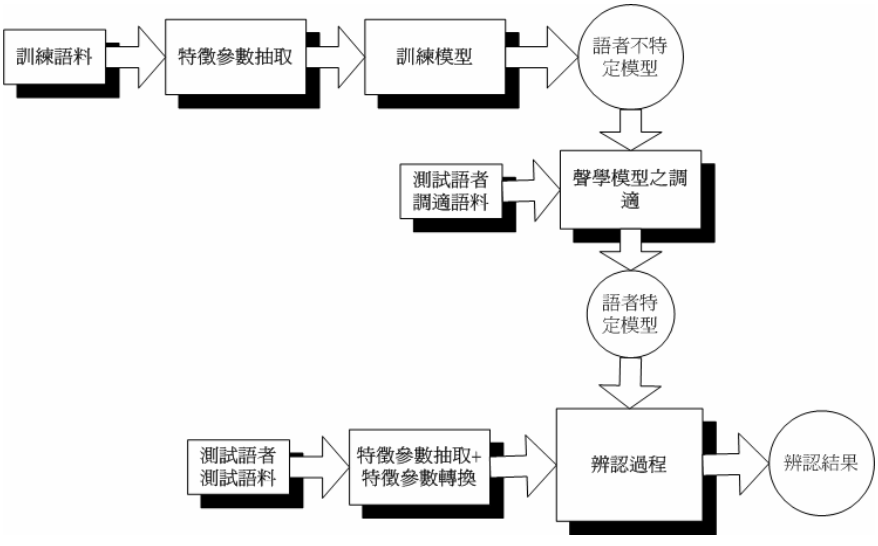

圖 1.1 以聲學模型為基礎之語者調適系統

以特徵向量為基礎的調適方法則是在處理特徵參數的時候，就把語者的聲學特性消除，這個消除語者特性的過程，我們稱為正規化(normalization)。接下來再把所有正規化後的特徵參數來訓練聲學模型，辨識的時候一樣需要把每位語者的特徵參數正規化，如此一來聲學模型與測試語者的語音性都去除了，辨識系統已經沒有語者與語者間差異的因子，辨識效果也因此會提升。整個系統流程圖如圖 1.2 所示。語音信號的頻譜特性對語音辨識而言是一個很重要的依據。舉例來說，對於相同語者所發出相同的音節在頻譜上的共振峰(format)位置應當非常類似。然而若因為不同語者的關係導致發出的相同音節在頻譜上共陣峰位置差距甚遠，再用這些差異性甚大的特徵參數訓練一個聲學模型，很容易與其他音節的聲學模型在頻譜上發生重疊的現象(overlap)，辨識系統因此容易發生誤判的現象。

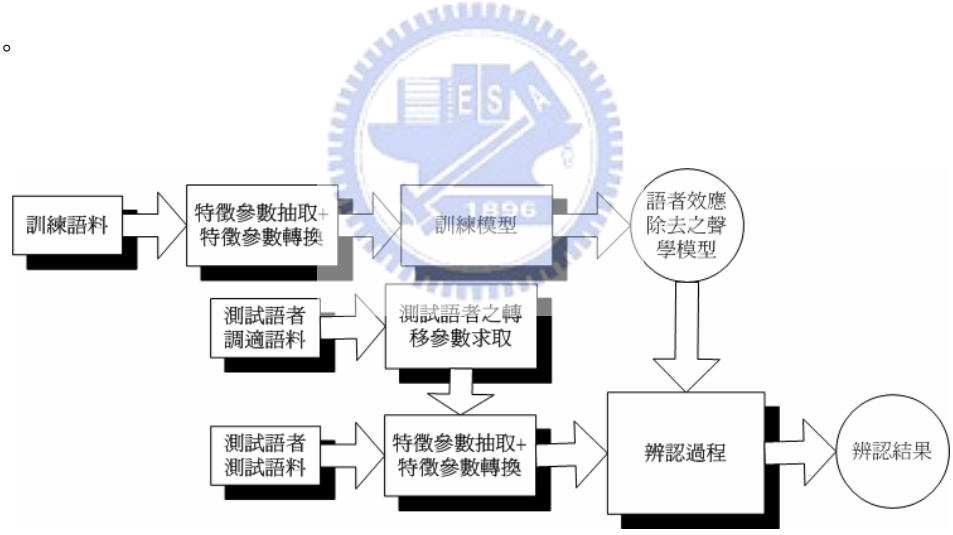

圖 1.2 以特徵向量為基礎之語者調適系統

若以調適時的做法而言，調適技術又可以分類如下。系統在做語調適時，必須先獲得使用者的語料，此語料稱為調適語料(adaptation data)。調適語料如果每收集一句即進行一次調適，稱為循序調適(sequential adaptation)、遞增調適法(incremental)或是線上調適法(on-line adaptation);調適語料如果收集到指定的句數之後才進行一次調適，稱為批次調適法、靜態調適法(static adaptation)或離線調

適法(off-line adaptation)。

另一種依調適做法的分類為：如果調適語料內容是事先已知的，將某特定語音對模型進行切割(force alignment)及可得到正確的切割資訊並對測試語料進行調整，此法稱為監督式調適法(supervised adaptation)；若事先不知道調適語料內容，須先對調適語料進行辨認，以辨認結果來當做調適語料內容，此方法稱為非監督式調適(unsupervised adaptation)。

本篇論文著重在研究以特徵向量為基礎的調適法，也稱這種調適法為語者正規化(speaker normalization)調適法。並且依照上述的分類方法，本論文所描述 HMM-based 語者偏移量除去法由於需要正確的切割資訊，所以使用的調適方法為批次和監督式調適法。希望每一位語者的特徵參數藉由一組轉移函數轉換去除語者在頻譜特性上的差異。研究方向只在於語音信號前處理與聲學層次的範疇，並不包含語言層次的處理。在聲學層次處理中僅探討中文連續音節的辨認，並未做聲調辨認方面的研究。

1.3 章節概要

本論文共分成四章，各章節編排如下：

第一章 緒論：說明研究動機、研究方向與章節概要。

第二章 語者通道效應之分析與移除：詳細述續如何求取特徵參數之轉移函數並且對使用此組轉換後的初步效能分析。之後再說明使用此轉移函數在語者調適系統的詳細流程。

第三章 實驗背景與實驗結果及分析：使用 MAT4500 語料庫以十分之九與十分之一的比例為訓練及測試語料，並抽取測試語料中之長句做為調適語料，以特徵參數轉換的方法來補償語者/通道效應，並觀察其在語者調適系統上的表現。

第四章 結論與未來展望：說明本論文之結論與未來研究方向。

第二章 語者通道效應之分析與移除

語音辨識系統常常因為訓練條件與測試條件的不匹配而影響其辨識結果。不匹配的原因通常都來自於語者間嘴型、聲道構造、腔調的不同。如果我們能在求取特徵參數時就去除影響語者間不同的因子訓練較精準的聲學模型，以提高模型得準確性，而測試者在使用系統時也以同樣的方法消除自己的語音效應，如此一來在訓練與測試都已經正規化的情況下，語音辨識效能應當有所提升。第一章研究方向中已經簡單介紹以特徵向量為基礎之調適系統，接下來我們將詳細介紹如何求取每一位語者轉移參數的方法。

2.1 語者偏移量除去法

本論文所述敘述之語者偏移量移除法【4】是先將整個訓練語料庫以 HMM 模型切割，找一組參數轉換函數 $\mathbf{Ax}+\mathbf{b}$，然後盡可能的把每個特徵參數往所屬 HMM 模型的平均值(one mixture mean)移動，而後再以移動過後的特徵參數重新訓練一個較精準的 HMM 模型；辨認時，在有適當切割資訊的情況下，也可以以相同於訓練語料的方法移除語者的偏移量，然後再進行辨認。

假設頻譜特性的差異可以利用一個線性轉換的關係來描述。希望對所有語料找到一個轉換關係，如下式：

$$\mathbf{y} = \mathbf{Ax} + \mathbf{b} \tag{2.1}$$

其中 **y** 和 **x** 分為轉移後和原始的語音特徵向量，**A**、**b** 為線性轉換關係的參數。以下兩個小節我們將詳細推導如何以兩種不一樣的原則得到 **A**、**b**。

2.1.1 使用 MSE 原則之轉換參數求取

基本的想法是欲使轉換過後的特徵向量 $\mathbf{y}_t$ (i.e. $\mathbf{y}_t = \mathbf{A}_k\mathbf{x}_t + \mathbf{b}_k$)和 $\mu_{s_t m_t}$ (第 m 個 model 的第 s 個 state)在頻譜上的距離最為相近，並且以語者 k 為單位，在最小化模型平均變異數向量的原則(Mean Square Error criterion)下求出矩陣 $\mathbf{A}_k$ 及向量 $\mathbf{b}_k$ ，同時對這一位語者所有語句中的特徵向量做轉換得到轉換後的特徵參數，使得特徵參數更為強健。

在詳細的推導之前，首先定義參數符號。假設 x_t 為第 t 個音框的 12 維或是 38 維梅爾倒頻譜參數(MFCC)，而整個訓練語料庫已經使用隱藏式馬可夫模型切割至音節的狀態序列，我們定義一個三組（3-tuple）的參數：(s_t, m_t, k_t)，來標示第 t 個音框的屬性，其中符號定義如下所示：

s_t ：隱藏式馬可夫模型狀態(HMM state)， $s_t \in \{0,1,2\}$ if m_t is an initial model（聲母模型）and $s_t \in \{0,1,2,3,4\}$ if m_t is a final model（韻母模型）；

m_t ：隱藏式馬可夫模型， $m_t \in$ {100 initials, 40 finals, silence}；

k_t ：語者編號(speaker index)。

定義一個客觀存在的函數如下式所示：

$$Q_k = \sum_{t=1}^{T_k} (\mathbf{A}_k\mathbf{x}_t + \mathbf{b}_k - \mu_{s_t m_t})^T (\mathbf{A}_k\mathbf{x}_t + \mathbf{b}_k - \mu_{s_t m_t}) \tag{2.2}$$

由 Q_k 最小化的觀點來求取 $\mathbf{A}_k$ 以及 $\mathbf{b}_k$ ，也就是將 Q_k 對 $\mathbf{A}_k$ 、$\mathbf{b}_k$ 做偏微分，設其值等於 0，可以求得 $\mathbf{A}_k$ 、$\mathbf{b}_k$ 之最佳解。

$$\frac{\partial Q_k}{\partial \mathbf{b}_k} = \frac{\partial(\sum_{t=1}^{T_k}(\mathbf{A}_k\mathbf{x}_t + \mathbf{b}_k - \mu_{s_t m_t})^T(\mathbf{A}_k\mathbf{x}_t + \mathbf{b}_k - \mu_{s_t m_t}))}{\partial \mathbf{b}_k}$$
$$= \sum_{t=1}^{T_k} \frac{\partial(\mathbf{A}_k\mathbf{x}_t + \mathbf{b}_k - \mu_{s_t m_t})^T(\mathbf{A}_k\mathbf{x}_t + \mathbf{b}_k - \mu_{s_t m_t})}{\partial \mathbf{b}_k}$$
$$= -2\sum_{t=1}^{T_k}(\mathbf{A}_k\mathbf{x}_t + \mathbf{b}_k - \mu_{s_t m_t})$$
$$= -T_k\mathbf{b}_k - \sum_{t=1}^{T_k}(\mathbf{A}_k\mathbf{x}_t - \mu_{s_t m_t})$$
$$= \mathbf{0}$$

$$\frac{\partial Q_k}{\partial \mathbf{A}_k} = \frac{\partial(\sum_{t=1}^{T_k}(\mathbf{A}_k\mathbf{x}_t + \mathbf{b}_k - \mu_{s_t m_t})^T(\mathbf{A}_k\mathbf{x}_t + \mathbf{b}_k - \mu_{s_t m_t}))}{\partial \mathbf{A}_k}$$
$$= \sum_{t=1}^{T_k} \frac{\partial(\mathbf{A}_k\mathbf{x}_t + \mathbf{b}_k - \mu_{s_t m_t})^T(\mathbf{A}_k\mathbf{x}_t + \mathbf{b}_k - \mu_{s_t m_t})}{\partial \mathbf{A}_k}$$
$$= \sum_{t=1}^{T_k} \frac{\partial tr[(\mathbf{A}_k\mathbf{x}_t + \mathbf{b}_k - \mu_{s_t m_t})^T(\mathbf{A}_k\mathbf{x}_t + \mathbf{b}_k - \mu_{s_t m_t})]}{\partial \mathbf{A}_k}$$
$$= \sum_{t=1}^{T_k} \frac{\partial tr[\mathbf{x}_t^T\mathbf{A}_k^T\mathbf{A}_k\mathbf{x}_t + (\mathbf{b}_k - \mu_{s_t m_t})^T\mathbf{A}_k\mathbf{x}_t + \mathbf{x}_t^T\mathbf{A}_k^T(\mathbf{b}_k - \mu_{s_t m_t})]}{\partial \mathbf{A}_k}$$
$$= \sum_{t=1}^{T_k}\left\{\frac{\partial tr[\mathbf{A}_k\mathbf{x}_t\mathbf{x}_t^T\mathbf{A}_k^T]}{\partial \mathbf{A}_k} + 2\frac{\partial tr[(\mathbf{b}_k - \mu_{s_t m_t})^T\mathbf{A}_k\mathbf{x}_t]}{\partial \mathbf{A}_k}\right\}$$
$$= \sum_{t=1}^{T_k}\left\{[\mathbf{x}_t\mathbf{x}_t^T\mathbf{A}_k^T + (\mathbf{x}_t\mathbf{x}_t^T)^T\mathbf{A}_k^T]^T + 2(\mathbf{b}_k - \mu_{s_t m_t})\mathbf{x}_t^T\right\}$$
$$= 2\sum_{t=1}^{T_k}\left\{\mathbf{A}_k\mathbf{x}_t\mathbf{x}_t^T + (\mathbf{b}_k - \mu_{s_t m_t})\mathbf{x}_t^T\right\}$$
$$= 2(\mathbf{A}_k\sum_{t=1}^{T_k}\mathbf{x}_t\mathbf{x}_t^T + \mathbf{b}_k\sum_{t=1}^{T_k}\mathbf{x}_t^T - \sum_{t=1}^{T_k}\mu_{s_t m_t}\mathbf{x}_t^T)$$
$$= \mathbf{0}$$

因此得到下式：

$$\mathbf{A}_k(\frac{1}{T_k}\sum_{t=1}^{T_k}\mathbf{x}_t) + \mathbf{b}_k - \frac{1}{T_k}\sum_{t=1}^{T_k}\mu_{s_t m_t} = \mathbf{0} \qquad (2.3)$$

$$\mathbf{A}_k \sum_{t=1}^{T_k} \mathbf{x}_t \mathbf{x}_t^T + \mathbf{b}_k \sum_{t=1}^{T_k} \mathbf{x}_t^T - \sum_{t=1}^{T_k} \mu_{s_t m_t} \mathbf{x}_t^T = \mathbf{0} \tag{2.4}$$

將(2.3)、(2.4)式改寫為下式：

$$\mathbf{A}_k B + T_k \mathbf{b}_k - C = \mathbf{0} \tag{2.5}$$

$$\mathbf{A}_k D + \mathbf{b}_k B^T - E = \mathbf{0} \tag{2.6}$$

其中 B、C、D、E 矩陣的定義如下所示：

$$B = \sum_{t=1}^{T_k} \mathbf{x}_t$$

$$C = \sum_{t=1}^{T_k} \mu_{s_t m_t}$$

$$D = \sum_{t=1}^{T_k} \mathbf{x}_t \mathbf{x}_t^T$$

$$E = \sum_{t=1}^{T_k} \mu_{s_t m_t} \mathbf{x}_t^T$$

由(2.5)式可得：

$$\mathbf{b}_k = \frac{1}{T_k}(C - \mathbf{A}_k B) \tag{2.7}$$

將(2.7)式帶入(2.6)式得：

$$\begin{aligned} &\mathbf{A}_k D + \frac{1}{T_k}(C - \mathbf{A}_k B) B^T - E = \mathbf{0} \\ &\mathbf{A}_k (D - \frac{1}{T_k} B B^T) = E - \frac{1}{T_k} C B^T \end{aligned} \tag{2.8}$$

最後 $\mathbf{A}_k$ ，$\mathbf{b}_k$ 可由下列式子求得：

$$\mathbf{A}_k = (D - \frac{1}{T_k} B B^T)^{-1} (E - \frac{1}{T_k} C B^T) \tag{2.9}$$

$$\mathbf{b}_k = \frac{1}{T_k}(C - (D - \frac{1}{T_k}BB^T)^{-1}(E - \frac{1}{T_k}CB^T)B) \tag{2.10}$$

從式(2.9)，(2.10)可以使用簡單的矩陣運算找到轉換後的特徵參數$\mathbf{y}_t$ ($\mathbf{y}_t = \mathbf{A}_k\mathbf{x}_t + \mathbf{b}_k$)。

2.1.2 使用 ML 原則之轉換參數求取

上述方法是很直覺的使轉換過後的特徵向量$\mathbf{y}_t$和$\mu_{s_t m_t}$在頻譜上的距離最為相近，所利用的準則是最小化模型平均變異數向量；現在我們則是希望轉換過後的特徵向量，在所屬的聲學模型狀態下得到最大的機率值(Maximum Likelihood)，如式(2.11)所示：

$$Q = \underset{\mathbf{A}_k, \mathbf{b}_k}{Max}\ \mathrm{P}(\mathbf{A}_k\mathbf{x}_t + \mathbf{b}_k | s_t, \Lambda_{s_t m_t}) \tag{2.11}$$

式(2.11)中的$\Lambda_{s_t m_t}$為特徵參數在t時間時所屬的 one mixture 聲學模型狀態。並且經由式(2.12) 利用 viterbi algorithm【5】得到最好的狀態序列(state sequence)資訊

$$\underset{s_1, s_2, \ldots, s_{t-1}}{Max}\ \mathrm{P}(s_1 s_2 \ldots s_{t-1}, s_t, \mathbf{x}_1\mathbf{x}_2\ldots\mathbf{x}_t | \Lambda) \tag{2.12}$$

利用此方法主要是希望藉由更多的正確參數(多考慮了HMM狀態(state)內的共變異矩陣(covariance matrix))可以預估更準確的轉移矩陣**A**以及**b**。接下來我們將詳細推導求取轉移參數的過程。在此，我們先假設共變異矩陣為對角矩陣(diagonal)。

由$\frac{\partial Q}{\partial \mathbf{A}} = 0$ ，$\frac{\partial Q}{\partial \mathbf{b}} = 0$ 可得到

$$\sum_t \mathbf{R}_{S_t m_t}^{-1}(\mathbf{A}\mathbf{x}_t + \mathbf{b} - \boldsymbol{\mu}_{S_t m_t})\mathbf{x}_t^{\mathrm{T}} = 0 \tag{2.13}$$

$$\sum_t \mathbf{R}_{S_t m_t}^{-1}(\mathbf{A}\mathbf{X}_t + \mathbf{b} - \boldsymbol{\mu}_{S_t m_t}) = 0 \tag{2.14}$$

將上式矩陣的每一項分別展開可得

$$\begin{aligned}
&\sum_t \left(\sum_j w_{ii}^{S_t} a_{ij} x_j(t) x_l(t) \right) + \sum_t \left(w_{ii}^{S_t} b_i x_l(t) \right) = \sum_t \left(w_{ii}^{S_t} \mu_i^{S_t} x_l(t) \right), \quad \forall i, l \\
&\sum_t \left(\sum_j w_{ii}^{S_t} a_{ij} x_j(t) \right) + \sum_t \left(w_{ii}^{S_t} b_i \right) = \sum_t \left(w_{ii}^{S_t} \mu_i^{S_t} \right), \quad \forall i
\end{aligned} \tag{2.15}$$

上述的線性方程式可以分成 i 個獨立方程組，令

$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 \\ \vdots \\ \mathbf{A}_n \end{bmatrix}$，$\mathbf{A}_i$ 是列向量。 上式經過重新排列可以整理為

$$\begin{bmatrix} \sum_t w_{ii}^{S_t} \mathbf{X}(t)\mathbf{X}^{\mathrm{T}}(t) & \sum_t w_{ii}^{S_t} \mathbf{X}(t) \\ \sum_t w_{ii}^{S_t} \mathbf{X}^T(t) & \sum_t w_{ii}^{S_t} \end{bmatrix} \begin{bmatrix} \mathbf{A}_i^{\mathrm{T}} \\ b_i \end{bmatrix} = \begin{bmatrix} \sum_t w_{ii}^{S_t} \mu_i^{S_t} \mathbf{X}(t) \\ \sum_t w_{ii}^{S_t} \mu_i^{S_t} \end{bmatrix}, \quad \forall\, i \tag{2.16}$$

最後可得

$$\begin{bmatrix} \mathbf{A}_i^{\mathrm{T}} \\ b_i \end{bmatrix} = \begin{bmatrix} \sum_t w_{ii}^{S_t} \mathbf{X}(t)\mathbf{X}^{\mathrm{T}}(t) & \sum_t w_{ii}^{S_t} \mathbf{X}(t) \\ \sum_t w_{ii}^{S_t} \mathbf{X}^T(t) & \sum_t w_{ii}^{S_t} \end{bmatrix}^{-1} \begin{bmatrix} \sum_t w_{ii}^{S_t} \mu_i^{S_t} \mathbf{X}(t) \\ \sum_t w_{ii}^{S_t} \mu_i^{S_t} \end{bmatrix}, \quad \forall\, i \tag{2.17}$$

比較式(2.16)與 P.C. Woodland 在推導 MLLR 裡的式子，式(2.16)中 $\mathbf{X}(t)$ 變成模型的平均值 $\boldsymbol{\mu}_t$，原本 $\boldsymbol{\mu}_t$ 的部分則變成調適語料的特徵參數。這可以解讀成，MLLR 是希望聲學模型的特性盡量接近調適語料，而我們則是希望每個特徵參數盡量接近所屬聲學模型的平均值。

2.2 特徵參數轉換後之效能評估

利用上述幾種轉換函數求取方法並重新訓練聲學模型且應用在辨識系統求得其辨認效能前，我們可以藉由測量經轉換後之特徵參數在所屬模型與狀態時的轉移程度來評鑑該轉換在去除語者效應的能力。在此我們可以使用 F-ratio 觀察、評估此偏移量移除後的效能。

F-ratio 定義如下：

$$F = \frac{\text{variance of mean}}{\text{mean of variance}} = \frac{\sum_{s,m} f(s,m)(\mu_{s,m} - \bar{\mu}_{s,m})^2}{\bar{\sigma}^2} \tag{2.18}$$

由上式可知，F-ratio 的數值大小表示每個 HMM 模型狀態的相對距離變化，若 F-ratio 變大，則表示狀態的相對距離變遠，反之，則變近。通道效應的移除，基本上會使得聲學模型狀態本身更為緊密，而模型狀態與狀態之間的距離幾乎不變，因此 F-ratio 提升越多表示效果變得越好。

首先我們先來介紹特徵向量每一個維度所代表的意義，前 1 到 12 維是 12 維的梅爾頻率倒頻係數(MFCC)，第 13 維到 24 維是 12 維的回歸倒頻係數(△MFCC)，第 25 維是 1 維的能量回歸頻譜係數(△energy)，第 26 到 37 維是 12 維的回歸回歸倒頻係數(△△MFCC)，第 38 維則是能量回歸回歸係數(△△energy)，總共 38 維。

一開始我們只對特徵參數的前十二維，由最小化模型平均變異數向量的原則(MSE criterion)求取的矩陣 **A** 及 **b** 並將特徵參數轉換後來觀察其前十二維的 F-ratio。之後則是對特徵參數的所有維度(三十八維)，由最大化模型平均機率(ML criterion)求取矩陣 **A** 及 **b** 並觀察移除偏移量後 HMM 模型參數(三十八維)的

F-ratio。圖 2.1、2.2 中，十字曲線為原始特正參數每一個維度的 F-ratio 值變化趨勢，圓形曲線為使用 MSE criterion 並轉移特徵向量後之 F-ratio，星形曲線則是使用 ML criterion 的 F-ratio 變化趨勢。可以觀察出 F-ratio 在兩種轉換準則下的提升程度；使用 ML criterion 的 F-ratio 幾乎所有的特徵向量維度都大於使用 MSE criterion 的 F-ratio，更明顯大於沒有經過轉換之特徵參數的 F-ratio。如此特性可以解釋，在有更多正確參數參與估算轉移參數 **A** 及 **b** 的情況下，語者效應更可以適當的被消除，聲學模型本身也因此更為緊密。

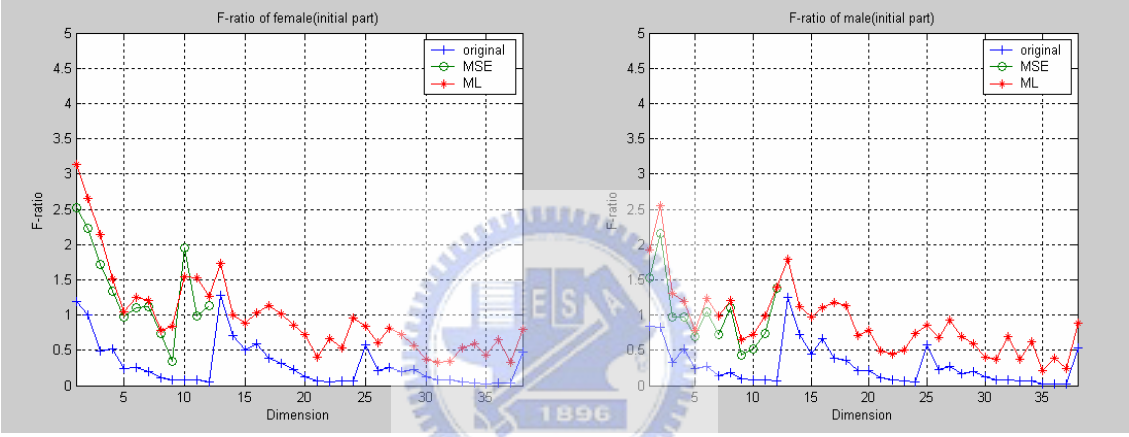

圖 2.1：原始特徵參數與經過轉移參數後的特徵參數之 F-ratio 聲母部分比較圖

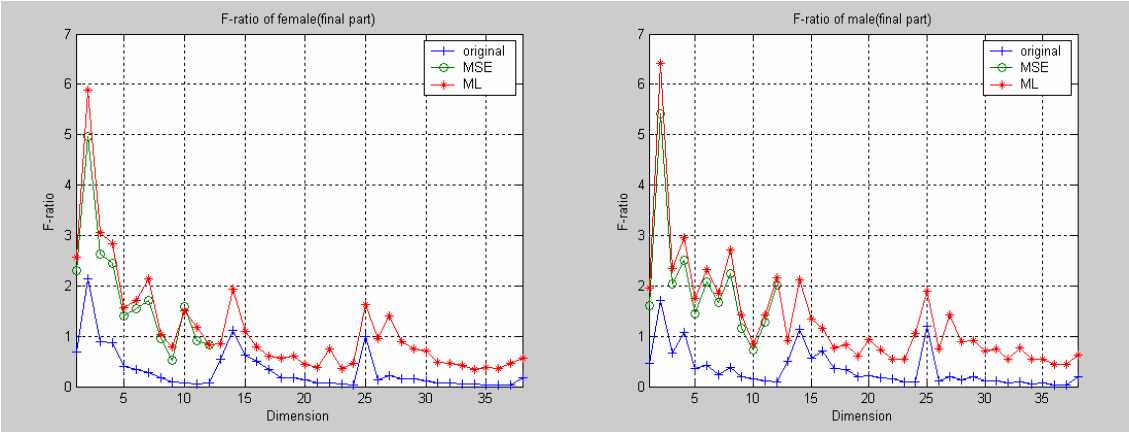

圖 2.2：原始特徵參數與經過轉移參數後的特徵參數之 F-ratio 韻母部分比較圖

另外可以觀察語者/通道偏移量是否去除的方法就是繪出數個 HMM state 第一維及第二維的機率統計分佈 $N(x_t;\mu_{s,m},\sigma^2_{s,m})$，觀察移除語者/通道效應後的分佈，是否較移除前緊密。圖 2.3 與圖 2.4 分別為聲母第二個狀態(ㄚ、ㄧ、ㄨ、ㄝ、

ㄛ)與韻母第三個狀態模型的 31~37 類(分別為ㄨㄣ、ㄨㄤ、ㄨㄥ、ㄩㄝ、ㄩㄢ、ㄩㄣ)機率分佈，其中較大的實線橢圓形分佈為原始的特徵參數的模型狀態統計分佈，較小的虛線橢圓分佈為特徵參數轉換後(ML criterion)的模型狀態統計分佈，菱形、正方形、圓形等幾何形狀代表著某個聲學模型在二維頻譜空間中的平均值。箭頭的方向則是從原始特徵參數的平均值指向轉移後的特徵參數平均值。由圖中觀察得知，相對於原始的特徵參數的模型統計分佈，特徵參數轉移後的統計分佈確實變得更緊密，因此模型參數修正得更加精確，但是模型的平均值則會有些許的移動。

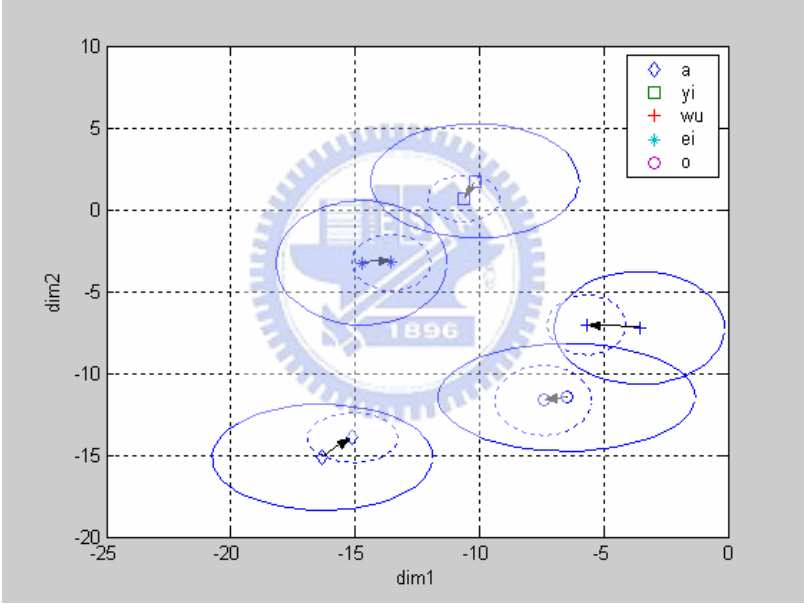

圖 2.3　特徵參數轉換 v.s. 原始特徵參數的特徵參數分布

(ㄚ、ㄧ、ㄨ、ㄝ、ㄛ)

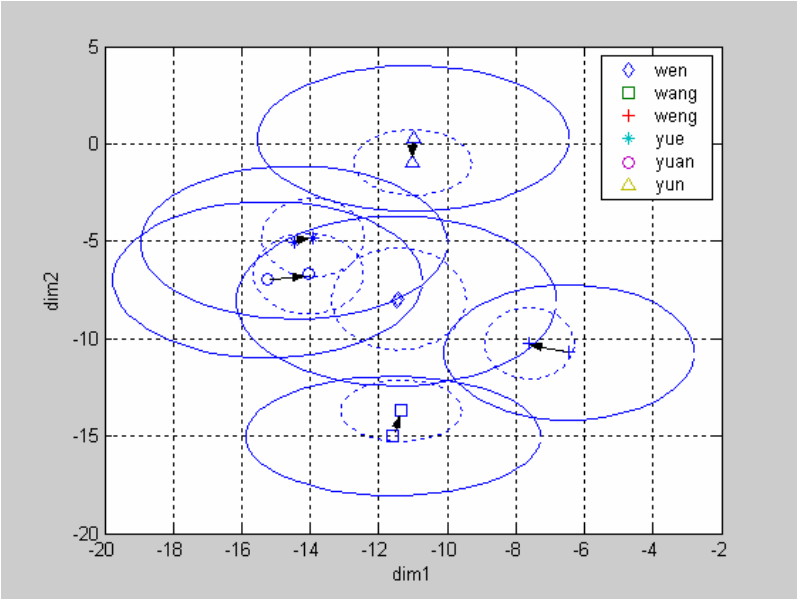

圖 2.4　特徵參數轉換 v.s. 原始特徵參數的特徵參數分布

(ㄨㄣ、ㄨㄤ、ㄨㄥ、ㄩㄝ、ㄩㄢ、ㄩㄣ)

章節 2.2 以前主要介紹在兩種不同原則下，如何對每一位語者求得矩陣 **A** 及向量 **b** 的方法，並且透過 $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$ 之轉換函式得到轉換後的特徵參數並觀察其 F-ratio 值。然而我們之所以可以得到如此精確的模型是因為訓練語料有正確切割位置的資訊，有了切割資訊才可以精確的得到每位語者的轉移矩陣及偏移向量進而消除語者/通道效應。接下來則要詳細介紹利用參數轉換之整個語者調適系統流程；詳細說明訓練聲學模型步驟及如何得到每一位測試語者些許的語音及其正確切割資訊來轉換測試語料並進行辨認。

2.3　語者偏移量移除之語者調適系統

2.3.1 模型的訓練

語者效應補償由特徵參數轉移法取代原本的 SBR(Signal Bias Remove) 【6】【7】法，以原始的特徵參數及切割資訊求取轉移參數矩陣 A 及向量 b，之後得

到訓練語料的轉移特徵參數、建立初始模型並對其進行訓練後產生更精準的聲學模型，系統模型參數訓練流程如圖 2.5 所示：

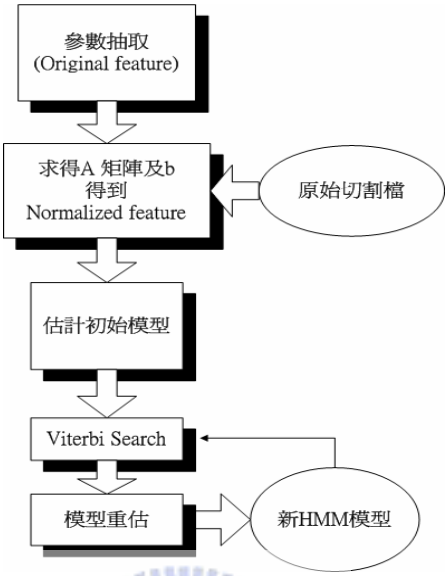

圖 2.5 以轉換後之特徵參數訓練聲學模型流程圖

2.3.2 語者調適--從調適語料中求取轉移參數

由於調適方法為監督式調適法，所以可以用原始的語者不特定模型對調適語料加以切割得到切割資訊，然後以調適語料的原始的特徵參數及此切割資訊求取轉移參數矩陣 A 及向量 b 做為接下來此位語者辨識語料的轉移參數。求得測試語者轉移參數矩陣 A 及向量 b 之流程如圖 2.6 所示

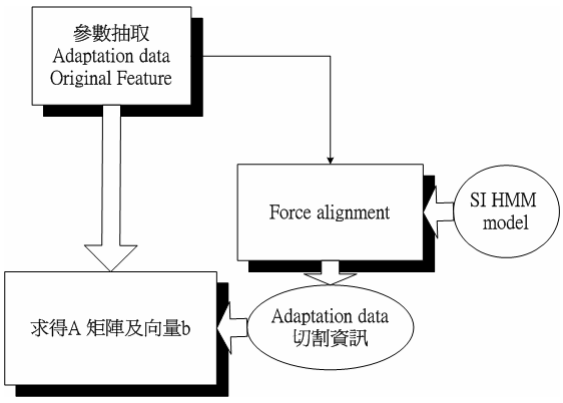

圖 2.6 從調適語料求得轉移參數流程圖

2.3.3 辨認流程

使用語者調適時所得到之矩陣 A 及向量 b 用來求取測試語料轉移後的特徵參數，最後再使用轉換後的特徵參數所訓練之 HMM 模型進行維特比搜尋(Viterbi search) 【8】，進而得到連續單音節辨認結果，辨認之流程如圖 2.7 所示

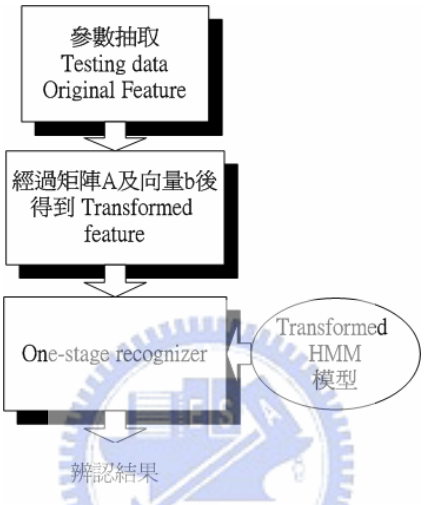

圖 2.7 移除語者偏移量測試流程圖

第三章 實驗背景與實驗結果及分析

3.1 語音資料庫

3.1.1 訓練語料 (train data)

本論文使用的語音資料是由「中華民國計算語言學會」(ROC Computational Linguistic Society)所提供的「台灣之國語語音資料庫」(Mandarin speech database Across Taiwan,簡稱 MAT),是透過電話網路經由個人電腦錄製,以 8KHz 的取樣頻率及 16 位元的取樣位元數來進行語音的收集,再經處理而成。 實驗中取 MAT 語料庫中的十分之九作為訓練語料。經過整理,參與訓練語料的人數為男生 2057 人,女生 2263 人,共 4320 人。總共訓練音節男生為 244463 個,女生為 263683 個,共 508146 個。

3.1.2 辨認語料 (test data)

以語者為單位,扣除調適語料後的十分之一 MAT 4500 為辨認語料。其中 MAT2000 的測試人數為 222 人,MAT2500 測試人數為 250 人,總共測試語者有 472 人。MAT4500 測試語料詳細的分類及統計如表 3.1:

MAT2000 測試語料分類	內容	備註
第一部分	short spontaneous answers to questions	(沒有使用)
第二部分	numbers	(沒有使用)

第三部分	isolated Mandarin syllables	(沒有使用)
第四部分	國語 2-4 字詞	
第五部分	國語平衡長句	(平均每句話有 14 個音節)

MAT2500 測試語料分類	內容	備註
第六部分	國語數字串	
第七部分	國語專有名詞	
第八部分	國語 2-4 字詞	
第九部分	國語平衡長句	(平均每句話有 13 個音節)

表 3.1 MAT4500 測試語料詳細的分類及統計

3.1.3 調適語料

從十分之一 MAT 4500 測試語料中，以語者為單位，取出國語平衡長句(參考表 3.1 ；MAT2000 為第五部分、MAT2500 為第九部分)為調適語料，句數從一句開始增加。如果調適語料不夠時，則整個語者的語料皆不作辨認。可做調適語料之句子的平均音節數為 14 個，平均時間為 4 秒。

3.2 語音訊號的特徵參數擷取

每一個音框(frame)的聲學特性經過數學的運算可以用一組特徵向量來表示，表 3.2 詳細敘述參數擷取的詳細情形

取樣頻率	8K
快速傅利葉(FFT)轉換點數	256
音框長度(frame size)	30ms
音框平移	10ms
濾波器(filter bank)與濾波器個數	30 個梅爾刻度三角濾波器

表 3.2 本論文實驗所使用之特徵參數擷取

3.3 聲學模型架構

本論文所用的聲學模型皆為連續密度隱藏式馬可夫模型(continues density HMM)，每一個中文音節可分成聲母與韻母兩部分；聲母模型為右相關模型(right dependent)或稱為 final-dependent，共有 100 個，由三個狀態(state)組成，韻母模型總數為 40 個是由五個狀態組成。每個模型的高斯混合數（mixture）以 50 個音框取一個混合數，最大混合數不超過 32 個。除此之外還有一個靜音模型只由一個狀態組成，靜音模型的混合數取 64 個。

3.4 辨認率之計算

對連續音而言，由於辨認結果所得的音節總數，未必等於正確的音節總數，因此辨認的結果除了「替代型」(Substitution)錯誤以外，還包含「插入型」(Insertion)以及「刪除型」(Deletion) 錯誤。在此對於替代型、插入型、刪除型錯誤的認定方式，是以得到最佳辨認率為原則，其具體做法是利用動態規劃法，將正確音節字串與辨認結果做一對應，進行錯誤類型的認定，找出一條可得最佳辨認率的路徑，再以下式計算辨認率：

辨認率 = (正確音節數 - (Sub + Ins + Del)) / 正確音節數 (3.1)

3.5 基本系統之測試比較

首先定義基本系統(baseline system)為沒有利用經過轉換的特徵參數所訓練出來的聲學模型來對所有的測試語者進行辨認，訓練語料為十分之九 MAT4500 語料，測試語料為十分之一 MAT4500 語料。辨認結果可以當作語者調適系統時的一個參考。如果沒有考慮 CMN(Cepstral Mean Normalization)時，基本辨識系統辨識率為 58.59%；有考慮 CMN 的狀況時，基本辨識系統辨識率為 61.96%，兩者相差大約 3 個百分點。詳細資訊如表 3.3 所示

	Ins.	Del.	Sub.	Total test syllable	Syllable correct rate
Baseline no CMN	2.69%	1.43%	34.58%	56691	58.59%
Baseline with CMN	2.26%	1.25%	34.54%	56691	61.96%

表 3.3 基本測試系統辨識率

3.6 偏移量移除之語者調適結果比較

3.6.1 使用 MSE 原則之轉換參數之語者調適系統

首先我們先來觀察以最小化模型平均變異數向量為原則(MSE criterion)，也就是利用第二章(章節 2.1.1)敘述特徵向量轉換函數之求取方法。這個部分我們只考慮特徵參數前十二維對求取轉參數矩陣 **A** 及向量 **b** 的影響，並且在轉換前十二

維的特徵參數後再求取 12 維的回歸倒頻係數(△MFCC)、12 維的回歸回歸倒頻係數(△△MFCC)、1 維的能量回歸頻譜係數(△energy)與 1 維的能量回歸回歸係數(△△energy)。詳細參數求取流程如圖 3.1 所示：

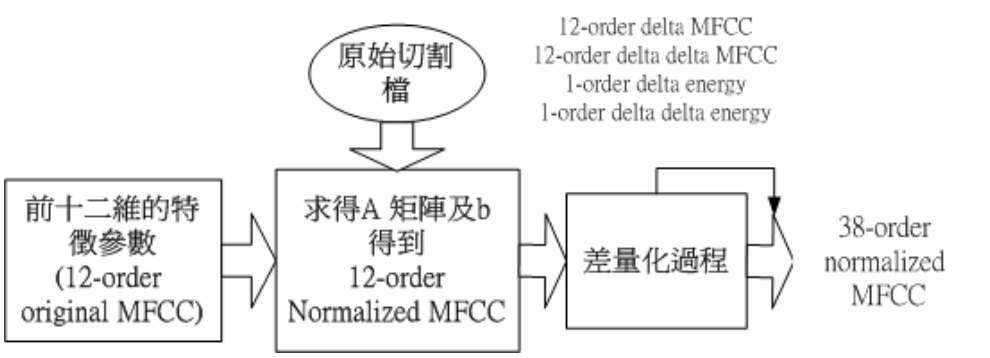

圖 3.1 特徵參數處理流程

我們在此先定義辨識率上限(Upper bond)，做法是以正確的測試語料 HMM 切割位置為輸入，求取測試語料特徵參數轉移函數矩陣 **A** 及向量 **b**，再用此組轉移函數轉移此語者所有的測試語料並且進行辨認。詳細的調適結果如表 3.4 與圖 3.2 所示：

Utterance number of adaptation data	Length of adaptation data (sec)	Ins.	Del.	Sub.	Total test syllable	Syllable correct rate
1	4.096	8.55%	1.31%	52.71%	50972	37.42%
2	8.426	5.60%	1.05%	40.70%	45158	52.66%
3	12.894	4.69%	0.97%	36.69%	39433	57.65%
4	17.430	4.36%	0.98%	34.91%	33740	59.75%
5	22.042	4.27%	0.88%	33.56%	27957	61.29%
6	26.775	4.42%	0.75%	32.50%	22280	62.32%
7	31.512	4.79%	0.67%	31.34%	16609	63.20%
8	37.053	5.16%	0.53%	30.75%	10958	63.56%
Upper bond		3.12%	0.99%	31.32%	56691	64.57%

表 3.4：調適語料的句數對辨識率之影響(語者數目隨著調適語料變多而減少)

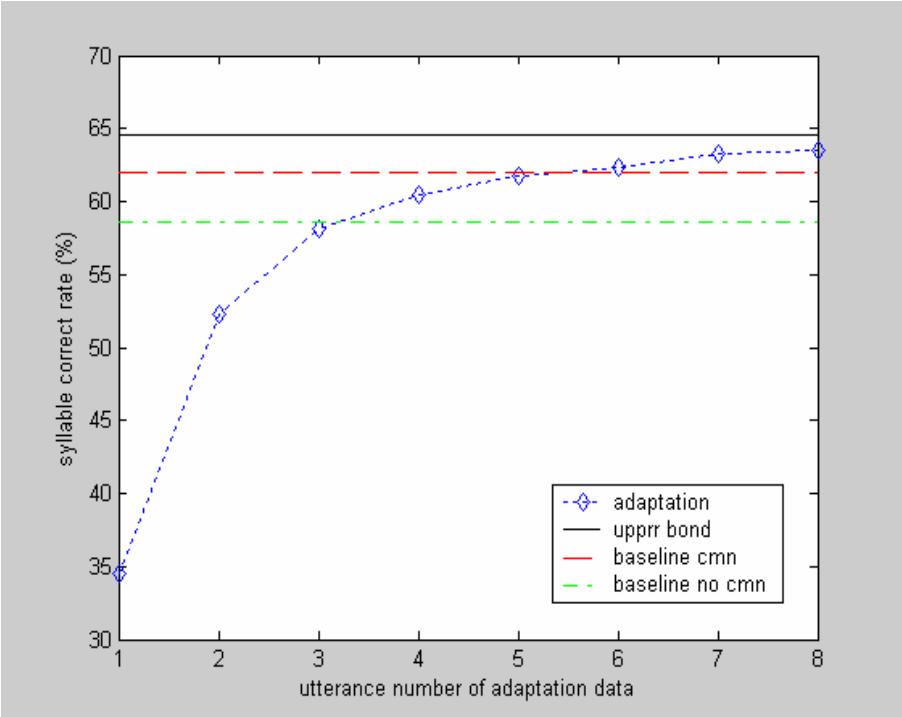

圖 3.2 辨識率 v.s. 調適語料的句數

調適語料從一句到八句增加，由於大部分測試語者的國語平衡常句部分都不超過九句，所以我們不做調適語料九句以後的測試。由表格中可以看出，在調適語料不參語辨認且不足調適語料的句數時整個語者不參語辨認的情形下總共測試音節隨著調適語句的增加而減少。進一步可以觀察出，當調適語料很少的時候調適效果非常的差，在調適語料增加到約 15 秒(約 60 個音節)時達到沒有使用 CMN 之基本系統的辨識水準並且在調適語料增加到 22 秒(大約 90 個音節)時才能達到使用 CMN 之基本系統的辨識水準。到有八句調適語料(約 37 秒)時，可以比使用 CMN 基本辨識系統高出快兩個百分點並且離辨識率上限只差一個百分點。

MAT4500 十分之一測試語料是經過品質挑選的語料，如果長句太少的語者表示此語者有太多語句是被標示為不能使用的句子，造成原因可能是因為此語者錄音時的狀況及環境太差的緣故。為了避免上述調適結果是因為隨著調適語料需求越多而使得錄音狀況不好的語者越不能參與辨識進而造成辨識率隨著調適語料增多而變高的現象，我們現在挑出調適語料足夠八句的 153 位語者來做相同語

者調適實驗，結果如表 3.5 所示

Utterance number of adaptation data	Speaker number(人)	Ins.	Del.	Sub.	Total test syllable	Syllable correct rate
1	153	2616	254	13791	25455	34.55%
2	153	1580	156	9524	23585	52.26%
3	153	1134	140	7758	21578	58.14%
4	153	952	132	6653	19552	60.42%
5	153	816	117	5729	17439	61.80%
6	153	749	102	4907	15308	62.39%
7	153	651	85	4086	13114	63.23%
8	153	565	58	3370	10958	63.56%

表 3.5 調適語料的句數對辨識率之影響(語者數目固定為 153 人)

由表 3.4 與表 3.5 可以觀察出上述所擔心的狀況幾乎沒有對調適結果造成影響，之後的實驗我們也將不考慮此種狀況的影響。

3.6.2 使用 ML 原則之轉換參數之語者調適系統

最大化模型平均機率原則(ML criterion)也就是利用章節 2.1.2 敘述特徵向量轉換函數之求取方，目的是希望藉由更多正確的參數估計轉換函數，可以使得每位語者的轉移矩陣 **A** 及向量 **b** 被更準確的計算出來以提升辨識效果。我們已經於章節 2.2 看到用此種轉移方法的 F-ratio 值比用 MSE criterion 的 F-ratio 值大，這可以顯示出使用 ML criterion 轉移後的特徵參數所訓練出的聲學模型比使用 MSE criterion 的聲學模型精確。現在我們來看使用此種轉移方法在語者調適系統上的表現。有別於第一種使用 MSE 原則之轉換調適系統，這個部分我們直接取三十八維的特徵參數來求取轉移參數，並直接對三十八維的特徵參數做轉換。詳細的調適結果如表 3.6 與圖 3.3 所示：

Utterance of adaptation data (句)	Length of adaptation data (sec)	Ins.	Del.	Sub.	Total test syllable	Syllable correct rate
1	4.096	7.41%	9.23%	65.66%	50972	17.70%
2	8.426	8.04%	1.07%	49.92%	45158	40.97%
3	12.894	6.54%	0.85%	42.07%	39433	50.54%
4	17.430	5.90%	0.82%	38.59%	33740	54.70%
5	22.042	5.71%	0.78%	36.28%	27957	57.24%
6	26.775	5.79%	0.68%	34.82%	22280	58.70%
7	31.512	6.24%	0.58%	33.22%	16609	59.96%
8	37.053	6.82%	0.51%	32.57%	10958	60.10%
Upper bond		2.91%	0.86%	29.07%	56691	67.16%

表 3.6：調適語料的句數對辨識率之影響(以 ML criterion 求取轉移參數)

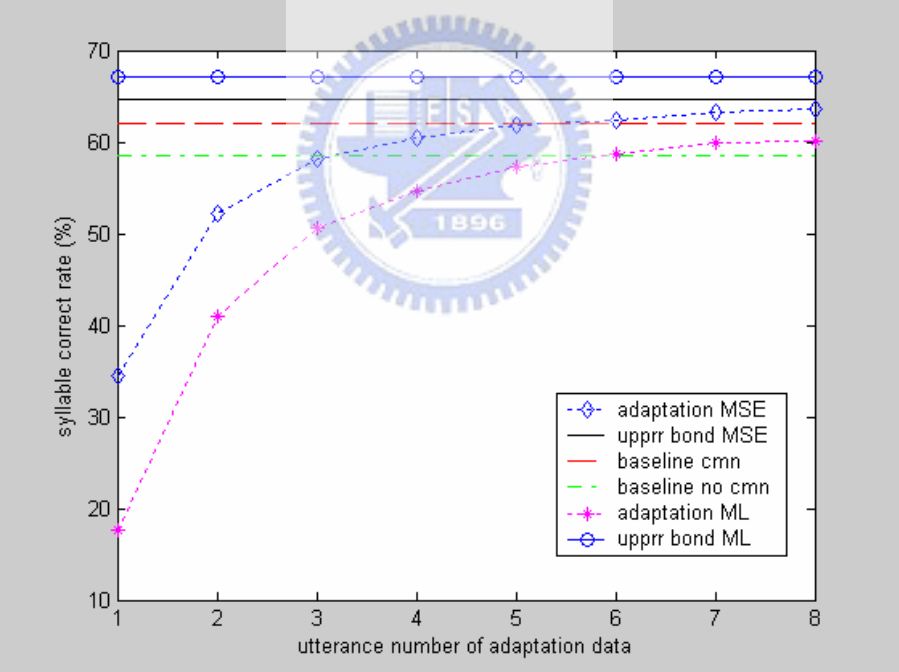

圖 3.3 辨識率 v.s. 調適語料的句數

由以上結果可以觀察到以 ML criterion 求取轉移後特徵參數的方法比用 MSE criterion 的方法在辨識率上限(upper bond)表現的好約 2~3 個百分點，但是在語者調適的系統上，似乎是因為需要預估的參數變多而使得調適語料有不足的現象產生，表現反而沒有比 MSE criterion 求取轉移後參數的方法來的優秀，但

兩者之間的辨識率差別也從一句調適語料的相差 20 個百分點到八句調適語料時只差 4~5 個百分點，有漸漸拉進的趨勢。

3.7 偏移量移除之語者調適系統改進與修正

3.7.1 使用新的切割資訊來求轉移參數

我們使用之偏移量移除方法來求取轉移函數所需要最重要的資訊就是正確的語音切割位置，之前的系統在訓練聲學模型時並沒有把新模型對訓練與料所產生的切割位置再拿來求取轉移函數。我們認為轉移後的特徵參數所訓練出的模型應該更為精確，而此模型對訓練語料做切割的切割位置也應該更為精準。以下為詳細執行步驟整理：

1. 首先使用原始的三十八維特徵參數(original MFCC)，利用原始切割資訊將靜音(silence)的部分去除。
2. 利用 HMM 模型對語料做切割，並利用每個模型之 one mixture 平均值得到 $\mathbf{A}_k$ 、 $\mathbf{b}_k$ 。
3. 得到 $\mathbf{A}_k$ 、後 $\mathbf{b}_k$ ，再用此參數求轉換後之特徵向量。
4. 利用轉換後之特徵向量重新訓練以求得新的 HMM 模型。
5. 重複步驟 2-4

我們先利用上述疊代法重複兩次以比較是否可獲得較佳之結果。圖 3.4 與圖 3.5 比較一次疊代與與兩次疊代時 F-ratio 的差異，十字曲線為一次疊代的每一個維度 F-ratio 變化趨勢，圓形曲線兩次疊代後之 F-ratio (兩者都以 ML criterion 求取特徵向量轉移函數)：

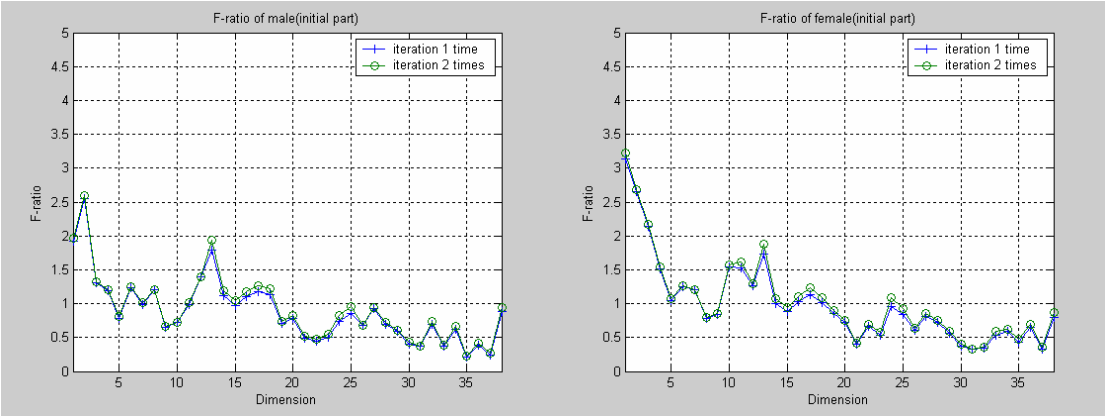

圖 3.4：一次疊代與兩次疊代後之聲母部分特徵參數 F-ratio 比較圖

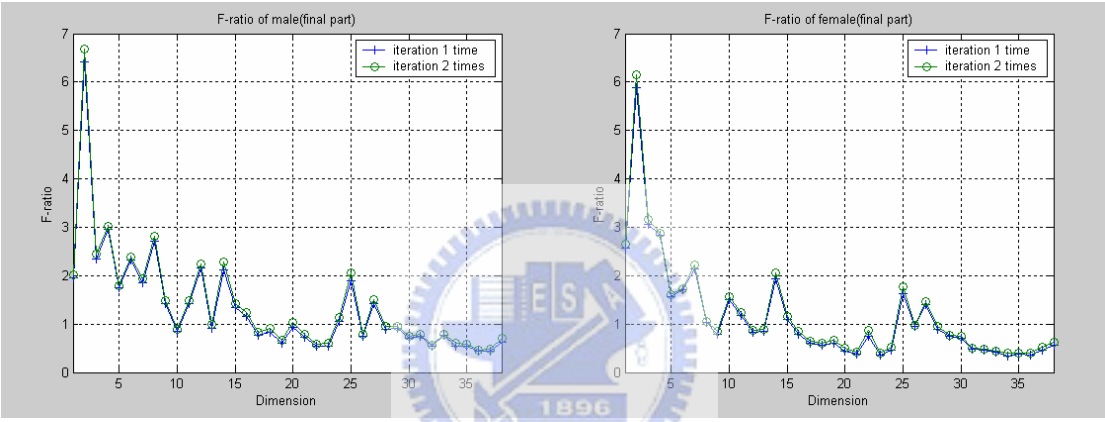

圖 3.5：一次疊代與兩次疊代後之韻母部分特徵參數 F-ratio 比較圖

Initial Part (ML criterion)				
	normalize feature 一次疊代		normalize feature 兩次疊代	
Dimension	SEX=F	SEX=M	SEX=F	SEX=M
C1	3.138970	1.928944	3.221669	1.964144
C2	2.660529	2.556048	2.678310	2.590629
C3	2.147169	1.310638	2.164336	1.328672
Final Part				
	normalize feature 一次疊代		normalize feature 兩次疊代	
Dimension	SEX=F	SEX=M	SEX=F	SEX=M
C1	2.575548	1.954941	2.648397	2.009606

C2	5.875198	6.417751	6.140684	6.692082
C3	3.056217	2.342376	3.171889	2.447390

表 3.7：一次疊代與兩次疊代後特徵參數前三維之 F-ratio 詳細數值

由圖 3.4 與圖 3.5 可以觀察出兩次疊代後聲母部分與韻母部分的 F-ratio 雖然有所提升，但是很有限，由於差距很小，我們特別列出特徵參數前三維的 F-ratio 詳細數值供參考，如表 3.7 所示。兩次疊代後之聲學模型的精確度可能只比一次疊代好一點點，甚至差不多。辨認時，我們以相同於上述的方法對調適語料做兩次疊代，並且先觀察查利用所有的測試語料當作調適語料時的上限(upper-bond)辨識率。表 3.8 是兩次疊代後的辨識率上限詳細資訊：

	Ins.	Del.	Sub.	Total syllable	Syllable correct rate
Iteration one time (ML criterion) Upper bond	2.95%	0.85%	29.03%	56691	67.17%

表 3.8 以 ML criterion 並且兩次疊代後之辨識率上限

由表 3.8 觀察出經過兩次疊代後的辨識率上限與一次疊代時的辨識率上限(表 3.6， 67.16%)幾乎一樣，並沒有顯著的提升。因為第一次估計參數轉換函式時所用的切割位置資訊已經十分準確了，經過疊代後的切割位置幾乎沒有改變。所以疊代的過程對辨識率的影響並不大，之後的實驗我們也將不考慮疊代的流程。

3.7.2 使用多組(聲母、韻母、靜音)特徵向量轉換

這個方法是參考 MLLR 的調適方法，當語料多的時候，為了將調適語料有效的利用，可以把所有的模型分成多群，以每一群模型所共有的語料去求出該群

的 **A** 、**b** 參數。類似於 MLLR 的方法，由於不同種類的語音彼此之間特徵參數差異性頗大，例如聲母和韻母就是屬於完全不同的兩大類語音。因之前的作法將所有的語音共享同一個線性轉換關係並不恰當。所以可以嘗試將所有的語音分類成多種，讓屬於同一個種類的語音信號都能共用同一個轉換矩陣 $\mathbf{A}_k$ 及向量 $\mathbf{b}_k$ ，如此一來才能確保屬於同一種類的每一個音素都能夠得到最佳的轉換。在本篇論文中，我們僅簡單的考慮三種類別，分別為聲母(initial)、韻母(final)與靜音(silence)部分三類。

分成三組來求取個別轉換函數 $\mathbf{A}_{k,(u,v,s)}$ ， $\mathbf{b}_{k,(u,v,s)}$ $(u,v,s$,分別代表 unvoice，voice，silence)的時候，聲母與韻母的轉換函數求法跟前面章節敘述求一組轉換函數時一模一樣，最大的差別就是靜音部分，稍後我們在來探討靜音部分的處理問題，首先我們先來比較沒有轉換問題的聲母、韻母部分在用分組轉移參數轉換過後 F-ratio 值的表現，如圖 3.6 與圖 3.7 所示，十字曲線為一組轉移函數轉移後的特徵向量每一個維度 F-ratio 變化趨勢，圓形曲線為分兩組轉換後之 F-ratio：

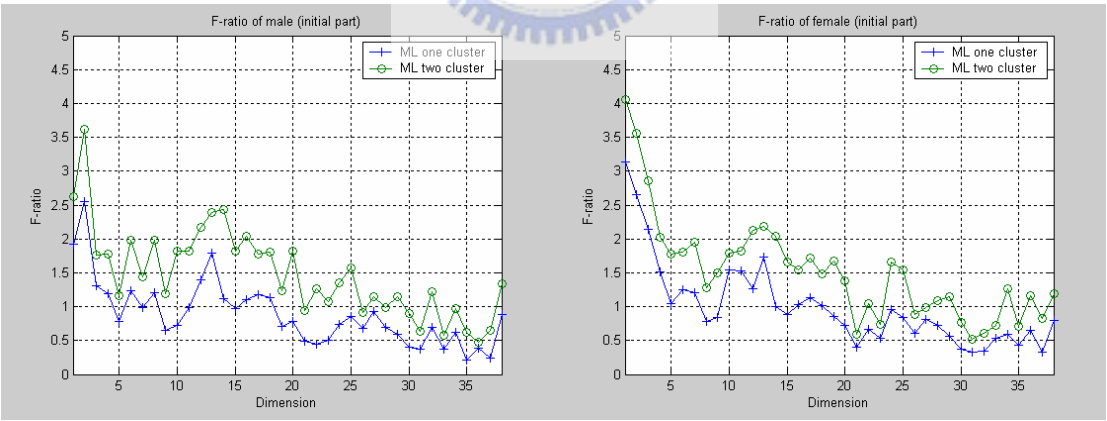

圖 3.6：只有一組轉移函數與分組求取轉移函數轉移後聲母部分之 F-ratio 比較圖

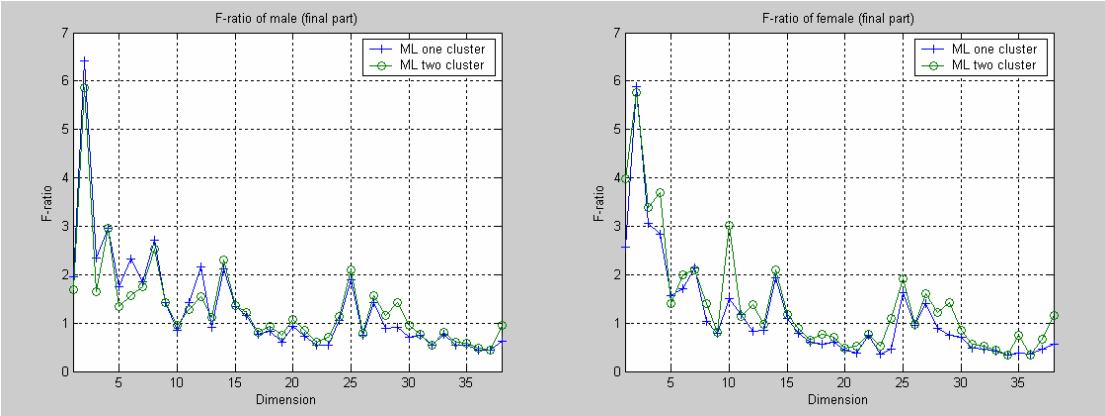

圖 3.7：只有一組轉移函數與分組求取轉移函數轉移後韻母部分之 F-ratio 比較圖

由圖 3.6 與圖 3.7 可以觀察出，用多組轉移函數並且在切割位置正確時普遍可以使得非靜音之聲學模型更加精密，F-ratio 比只使用一組轉移函數轉換時在聲母部分增加不少。比較特別的地方是男生韻母模型的部分，分組的結果比沒有分組的 F-ratio 值在不同維度時互有領先。以訓練過程總體看來，分組過後的成效還是比沒有分組時更優良一些。

現在我們來敘述靜音單獨求取轉移參數時所會遇到的問題，我們可以回顧到第二章(2.11)式 $Q=\underset{\mathbf{A}_k,\mathbf{b}_k}{Max}\ \mathrm{P}(\mathbf{A}_k\mathbf{x}_t+\mathbf{b}_k,s_t=sil|\Lambda_{sil})$，因為每個靜音的特徵參數都想盡量靠近靜音模型的平均值 $\boldsymbol{\mu}_{sil}$，在 ML 的原則下可以得最好的解就是 $\mathbf{A}_s$ 等於 Null matrix，$\mathbf{b}_s$ 等於 $\boldsymbol{\mu}_{sil}$。如此的轉換函數在辨認時不只會把原本屬於靜音部分的特徵參數轉移到靜音模型的平均值，同樣也會把原本不屬於靜音的特徵參數也全部轉移到靜音的平均值，這絕對不是正確的做法。接下來我們必須面對之前所提到的靜音部分處理過程，現在嘗試用三種不一樣的方法來處理靜音轉換的問題：

方法一、對靜音取一組轉換：

我們將所有訓練語料的靜音特徵參數分群分成 16 群或是 64 群，我們之所以用所有訓練語料的靜音特徵參數來求取靜音的轉移函數是因為每位訓練語者

或是測試語者所屬的語料靜音部分長短不一，如果每一位語者自己靜音的部分求取一組轉移函數時會造成靜音部分特徵參數較少的語者求不出自己的轉移矩陣 $\mathbf{A}_s$ 及向量 $\mathbf{b}_s$，而且發生此種問題的語者比例還不少(尤其是屬於 MAT2500 的部分)。分群的方法採取 Partition VQ(Vector Quantization) 的原則，再針對每一個靜音特徵參數是屬於哪一群去對所屬那一群的平均值與共變異矩陣來求得矩陣 $\mathbf{A}_s$ 及向量 $\mathbf{b}_s$ (如式(2.11)所示)。辨識時，每位語者靜音部份的轉換 $\mathbf{A}_s$、$\mathbf{b}_s$ 與訓練語料靜音用的轉換參數是同一組。圖 3.8 可以看出訓練語料靜音特徵參數在頻譜上第一維和第二維的分布圖，實線為原始特徵參數的靜音分布，圓形為其平均值，點虛線是靜音部分分成 16 群並且求出轉移參數 $\mathbf{A}_s$、$\mathbf{b}_s$ 並對靜音轉移後的頻譜分布圖，叉號為其平均值，虛線則是分成 64 群並轉移特徵參數後的結果，菱形為其平均值。靜音分的群數越少，則轉換後的特徵參數分布的越密集，這個結果看起來靜音模型似乎比不轉換時更緊密許多。

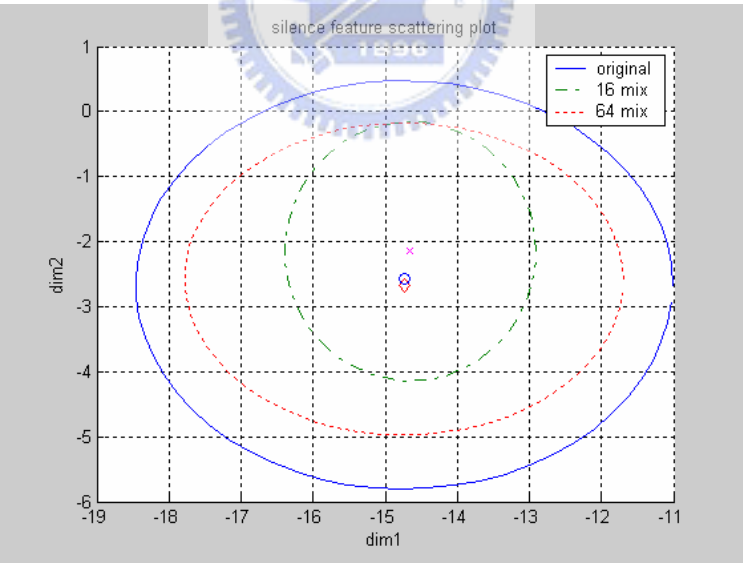

圖 3.8 靜音特徵參數在頻譜上的分布

但是在辨識的時候，由於參與辨識的每一個特徵參數並不知道它是屬於聲母、韻母或是靜音部分。所以必須使每個特徵參數都經過三組參數 $\mathbf{A}_{k,(u,v,s)}$、

$\mathbf{b}_{k,(u,v,s)}$ 的轉換，並且在辨認時互相競爭比較分數以得到辨認結果。在此時問題就產生了，靜音部分的轉換不只會把屬於靜音的特徵參數盡量移到平均值，也會把不屬於靜音的所有特徵參數也移到靠近靜音的平均值，如圖 3.9 所示，左邊的圖可以看出聲母部分的音節在經過靜音轉換參數後被移至靠近靜音模型附近，右邊的圖則是韻母部分被移動的情形，這樣看來，所有的非靜音特徵參數都有想往靜音平均值移動的趨勢。兩圖中，實線橢圓為原始特徵參數的靜音分布，點虛線橢圓則為經過靜音轉移參數轉換後之靜音分布。

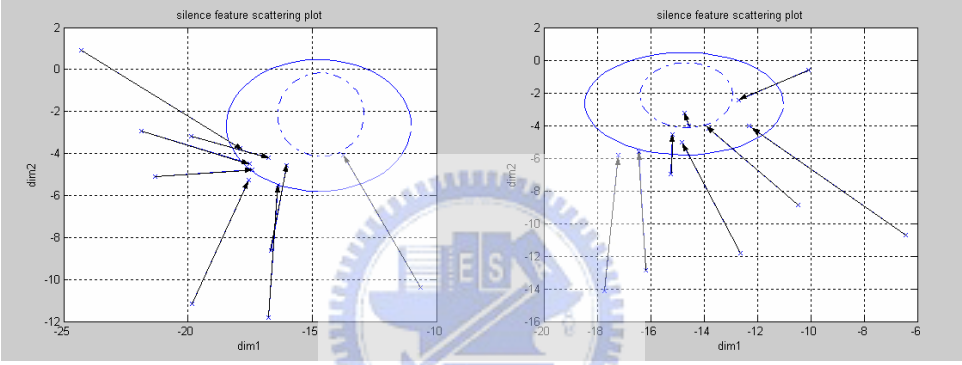

圖 3.9 部分聲母(左圖)及部分韻母(右圖)經過靜音的轉換函數轉移後之移動趨勢

我們先觀察查利用所有的測試語料求取三組轉移函數並對所有測試語料辨識的上限(upper-bond)辨識率，詳細資訊如表 3.9 所示：

	Ins.	Del.	Sub.	Total test syllable	Syllable Correct rate
Sil_16mixture	0.84%	26.16%	32.80%	56487	40.20%
Sil_64mixture	1.77%	8.13%	30.41%	56487	59.69%

表 3.9 將靜音分成 16 群與 64 群之兩種靜音轉移函數對辨識率的影響

就如以上我們所擔憂的情形，辨識結果刪除型(deletion)的錯誤大量增加，這種情況就是原本應該是非靜音的特徵函數都被辨識成靜音，但是交替型(substitution)的錯誤還是維持在 30%左右，顯示音節與音節間辨識結果還算良

好。使用此方法的辨識率上限沒有比基本系統(baseline system)優越，所以沒有繼續觀察此方法在語者調適時的表現。

方法二、靜音部分不做參數轉換

我們只對韻母(final)、聲母(initial)作參數轉換，靜音(silence)部分則是使用原始的特徵參數訓練靜音模型。

Ins.	Del.	Sub.	Total test syllable	Syllable Correct rate
24.56%	0.52%	34.62%	56487	40.30%

表 3.10 只對非靜音的特徵參數做轉換之辨識結果

使用方法二辨識出來的結果則是插入型(Insertion)錯誤數量大量增加達到25%，也就是大部份原本是靜音的特徵參數被辨識成非靜音的音節，由辨識結果觀察插入型錯誤幾乎都是被辨識成ㄗ與ㄙ的音節。

方法三、使用聲母之參數轉換函數來做靜音之轉換

此時回到以語者為單位，我們用較相近於靜音部分的聲母(initial)轉換函數 $A_{k,(u)}$ 、$b_{k,(u)}$ 來轉移靜音(silence)的特徵參數。我們在此觀察兩種情況：

狀況三.1

所有的辨識語料拿來求取矩陣 $A_{k,(v,u)}$、$b_{k,(v,u)}$，並且分別對兩組轉換 $A_{k,(v,u)}$、$b_{k,(v,u)}$ 所分轉移過後的特徵參數求取個別的 frame observation probability,然後再帶入 viterbi algorithm 辨識。辨識率請見表 3.11 第一列。

狀況三.2

所有的辨識語料拿來求取 $A_{k,(v,u)}$ 、$b_{k,(v,u)}$ 並且知道參與辨識的特徵參數是屬於聲母、韻母或是靜音來決定此特徵參數是用哪一組轉換來轉換。詳細辨識率請見表 3.11 第二列：

	Ins.	Del.	Sub.	Total test syllable	Syllable Correct rate
狀況三.1	2.86%	0.88%	27.08%	56487	69.18%
狀況三.2	0.02%	0.01%	22.36%	56487	77.61%

表 3.11 使用聲母的轉移函數對靜音做轉換之辨識結果上限

表 3.11 第一列的辨識率上限達到 69.18%，比使用 MSE criterion 及 ML criterion 求取一組轉移函數的辨識率上限 64.57%、67.16%都還要高 2~4 個百分點，接下來我們一樣觀察上述兩種狀況在語者調適統的表現，請見表 3.12 與表 3.13：

Utterance number of adaptation data (句)	Length of adaptation data (sec)	Ins.	Del.	Sub.	Total test syllable	Syllable correct rate
1	4.096	6.45%	27.00%	63.05%	50680	3.50%
2	8.426	11.30%	2.59%	67.80%	45005	18.31%
3	12.894	10.03%	1.27%	56.72%	39296	31.98%
4	17.430	8.54%	1.09%	49.64%	33613	40.72%
5	22.042	8.04%	0.92%	45.31%	27846	45.73%
6	26.775	7.80%	0.78%	42.87%	22186	48.54%
7	31.512	7.94%	0.70%	41.15%	16537	50.21%
8	37.053	8.40%	0.57%	38.97%	10902	52.06%
Upper bond		2.86%	0.88%	27.08%	59487	69.18%

表 3.12　不知道辨識特徵參數屬於聲母、韻母或靜音之語者調適辨識結果

Utterance number of adaptation data (句)	Length of adaptation data (sec)	Ins.	Del.	Sub.	Total test syllable	Syllable correct rate
2	8.426	0.01%	0.44%	61.73%	45005	37.82%

3	12.894	0.02%	0.02%	49.70%	39296	50.25%
4	17.430	0.01%	0.01%	42.61%	33613	57.37%
5	22.042	0.01%	0.01%	38.50%	27846	61.47%
6	26.775	0.01%	0.01%	35.92%	22186	64.05%
7	31.512	0.01%	0.01%	33.79%	16537	66.19%
8	37.053	0.03%	0.00%	31.65%	10902	68.33%
Upper bond		0.02%	0.01%	22.36%	56487	77.61%

表 3.13　知道辨識特徵參數屬於聲母、韻母或靜音之語者調適辨識結果

3.8　本章結論與實驗結果分析

比較本篇論文所使用的調適方法，由圖 3.10 可以看出上限辨識率(upper bond) 越高的調適方法，在真正做語者調適時反而越不理想。可能的原因是因為上限辨識率越高的方法在求取轉移參數時所要估計的參數越多，如果在調適語料還不足夠的情形下，反而轉移函數預估的越不精準。這個情況猶如使用 MLLR 語者調適系統的情況【9】，在分群越多但是少量語料語者調適的情形下調適效果反而越差；但在有充足調適語料時，因為可以更精準的預估每一組的轉移函數，把屬於某一組的聲學模型調到更正確的位置，調適效果可以變的更好。

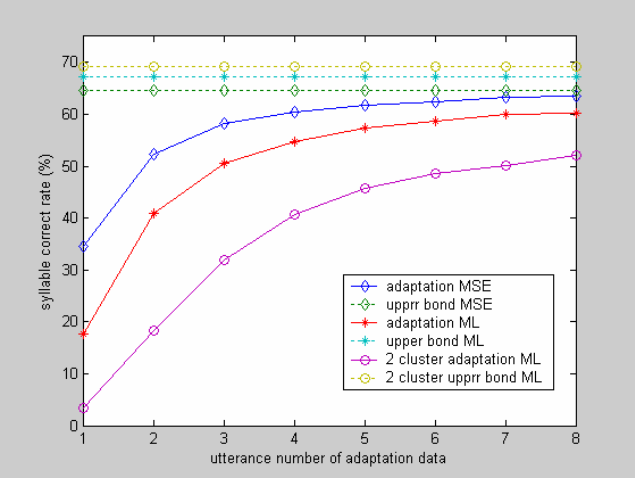

圖 3.10 所有調適方法之調適結果與辨識率上限

第四章 結論與未來展望

4.1 結論

本篇論文分別以兩種原則來求取特徵參數之轉移函數，並透過此組轉移函數使得語者效應被去除。尤其是在訓練聲學模型的過程中，由於可以得到正確可靠的切割訊息，每位語者的轉移函數都精確的被預估進而使得聲學模型更加緊密。由 F-ratio 值我們觀察出，使用 MSE criterion 求取轉移函數與使用 ML criterion 求取轉移函數的情形下，轉移過後的特徵參數 F-ratio 都比沒有轉移時提升許多，並且後者也比前者的 F-ratio 值大一些，這可以顯示有更多正確的參數參與估計轉移函數確實可以使得模型更準確。實驗中，我們取 MAT4500 語料庫 9:1 的比例為訓練及測試語料做外部測試，並且先以測試語料也有正確的切割資訊時的辨識率作為辨識率上限。以 MSE criterion 求取轉移函數時，辨識率上限為 64.57%，以 ML criterion 求取轉移函數時，辨識率上限可以達到 67.16%，分別比基本系統辨識率 61.96%提升 2.61%與 5.2%。但是這兩種原則求取的轉移函數應用在語者調適系統並且在我們調適語料上限八句時，MSE criterion 的辨識率 63.56%反而比使用 ML criterion 的辨識率 60.10%高，並且接近辨識率上限，顯示八句的調適語料已經足夠於使用 MSE criterion 求取測試語者的轉移函數，反而是對 ML criterion 而言，由於需要預估的參數較多(多約九倍)，調適語料反而顯的不足。

這種情形在對轉移參數分群時看的更明顯，如果聲母、韻母分 別求一組轉移函數 $\mathbf{A}_{k,(u,v)}$、$\mathbf{b}_{k,(u,v)}$，且靜音的部分使用聲母的轉移函數來轉換。辨識率上限為 69.18%，比基本辨識率高 7.22%，但是在八句調適語料時卻只剩 52.06%，這顯示出以八句調適語料估計聲母與韻母的轉移函數是非常不足的，這樣一來參數的估測值上有可能失去了準度而無法達到調適的效果。這也可以顯示出，本論文

的轉移函數求取方法使用在少量語料語者調適的情形下，表現並不出色。但相信在調適語料充足時，調適結果都會慢慢趨於辨識率上限(upper bond)。

4.2 未來展望

1. 更換語音資料庫，使用有充足調適語料之語料庫可以充分觀察語者調適的過程與最後辨識率收斂結果。TCC300 可以是下一個分析的語音資料庫。

2. 現在求取轉移參數時所需的切割資訊是由 Viterbi search 所得到的最佳狀態序列(hard decision)，以後我們可以使用 Baum-Welch forward-backward formula 求出特徵參數在每一個狀態之 state occupation probability，作為轉換之依據(soft decision)，取代現有的方法。

3. 每位語者的轉移參數 $\mathbf{A}_k$ 、$\mathbf{b}_k$ 皆有代表語者的特性，我們可以進一步的分析每位語者的轉移矩陣 $\mathbf{A}_k$ 行向量與列向量，希望可以獲得有用的資訊，並且可以朝向 eigenvoice 的概念來使用此矩陣與向量資訊。

參考文獻

【1】Jean-Luc Gauvain, Chin-Hui Lee, “Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains”, *IEEE Transactions on Speech and Audio Processing*, 1994.

【2】C.J Leggetter, P.C. Woodland, “Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models”, *Computer speech and Language*, 1995.

【3】曾國裕,“國語語音辨認之快速與者調適技術之研究”, 國立台灣大學電機工程研究所碩士論文, 中華民國八十七年六月.

【4】葉人鳳,“國語連續音節辨認系統之電話通道語者效應偏移量移除與分析”, 國立交通大學電信工程研究所碩士論文, 中華民國九十二年六月.

【5】Lawrence Rabiner and Bing-Hwang Juang, “Fundamentals of speech reognition”, Prentice Hall, 1993.

【6】Mazin G. Rahim and Bing-Hwang Juang, “Signal Bias Removal by Maximum Likelihood Estimation for Robust Telephone Speech Recognition”, *IEEE Transactions on Speech and Audio Processing*, January 1996.

【7】廖于棻,“通道偏移量分析以及不匹配環境下的電話語音辨認”, 國立交通大學電信工程學系碩士論文, 中華民國九十一年六月.

【8】H. Ney, “The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition”, *IEEE Trans. Acoustics, Speech, Signal Processing*, vol.32, no.2, pp.263-271, April 1984.

【9】曹昱, “國語音節與聲調辨識之少量語料語者調適”, 國立台灣大學電信工程研究所碩士論文, 中華民國九十年六月.