國立臺灣大學電機資訊學院電信工程學研究所

碩士論文

Graduate Institute of Communication Engineering

College of Electrical Enginnering and Computer Science

National Taiwan University

Master Thesis

# 基於變數選取聲學模型調適法之強健式語音辨識

# Acoustic Model Adaptation with Variable Selection for Robust Speech Recognition

胡庭曜

Ting-Yao Hu

指導教授：李琳山 教授

Advisor: Lin-Shan Lee, Ph.D.

中華民國一百零一年六月

June, 2013

# 摘要

聲學模型調適是改善聲學環境不匹配問題，增進語音辨識系統效能的一個重要方向。 仿射轉換方法 (affine transformation) ，如最大相似度線性回歸 (Maximum Likelihood Linear Regression, MLLR) ，在一般的聲學模型調適任務中有很好的效果。 然而在缺乏調適語料及正確轉寫 (transcription) 的自我調適情境 (self adaptation) 下，一般的仿射轉換調適方法容易造成過度貼合 (over-fitting) 問題。 為此，本論文利用變數選取 (variable selection) 的技術，提出兩種方法: 變數選取-最大相似度線性回歸 (Variable Selection MLLR, VSMLLR) 以及變數選取-特徵最大相似度線性回歸 (Variable Selection feature MLLR, VSfMLLR) 。 這兩種方法先以一些事前知識建構變數子集，再基於各變數子集以最大相似度準則求出對應的仿射轉換矩陣，最後利用正則化訓練準則 (regularization criterion) 當作子集與其對應仿射轉換的評量分數。 利用此方法，我們可以在系統上線時，即時為每句測試語料找出適合的變數子集、有效控制調適參數的複雜度、克服過度貼合問題並使辨識率進步。 當這些方法實做在Aurora-4語料庫上時，可發現測試集的辨識率較基本系統有顯著的進步。並勝過了一般常見的仿射轉換調適法，以及各種正則化訓練準則的延伸。 相較於基本系統的77.47%字正確率，在沒有額外的調適語料的條件下， 變數選取-最大相似度線性回歸以及變數選取-特徵最大相似度線性回歸分別將字正確率提升至80.10%與81.15%，相對進步率分別為11.67%以及16.33%。

# Contents

# 圖目錄

# 表目錄

# 第一章　緒論

## 1.1　研究動機

語音辨識 (speech recognition) 技術利用數位訊號處理、統計及機器學習方法，達到以電腦自動辨識人口說內容的功能，是一門具有前瞻性、應用價值的學問。 許多大型研究計畫，如美國國防部高級研究計畫局 (DARPA) 主導的全球自動化語言開發計畫 (Global Autonomous Language Exploitation, GALE) [1]，希望能建立高效率、高正確率的大字彙語音辨識(Large Vocabulary Continous Speech Recognition, LVCSR)系統，及其衍伸應用如翻譯 (translation) 及知識擷取 (distillation) 。 這項技術經過長時間的發展，今日在已被實際應用在各種科技產品中。 Google的聲音搜尋功能，以及Apple公司安裝於行動裝置上的虛擬個人秘書Siri，都是很好的例子。 這些產品利用語音辨識作為輸入的前端，將使用者說的聲音轉換成文字後，再做後續處裡。 因此使語音辨識的效能提升能夠幫助這些產品的主要功能，如文字搜尋以及語言理解 (speech understanding) ，跟著有所進步。 反之，過高的辨識錯誤 (recognition error) 也會使搜尋及語言理解的結果出現嚴重的問題。 由此可知，語音辨識在現代科技中佔有舉足輕重的地位。 如上一段所說，作為一個實用的語音辨識系統，最重要的條件就是良好的辨識正確率 [2]。 然而，語音辨識系統的效能會因訓練及測試時的聲學條件 (acoustic condition) 而有所不同。聲學條件由許多要素構成，例如:背景雜訊 (background noise) 、麥克風通道效應 (microphone channel effect) 、語者性別及聲道長度等等。 當語音辨識系統的訓練及測試聲學條件相匹配 (matched) ，通常能得到令人滿意的辨識結果。 反之，如果訓練及測試聲學條件不匹配 (mismatched) ，正確率將會大受影響。 因此，如何提升語音辨識系統的強健性 (robustness) ，使其不受聲學條件變化的影響，成為

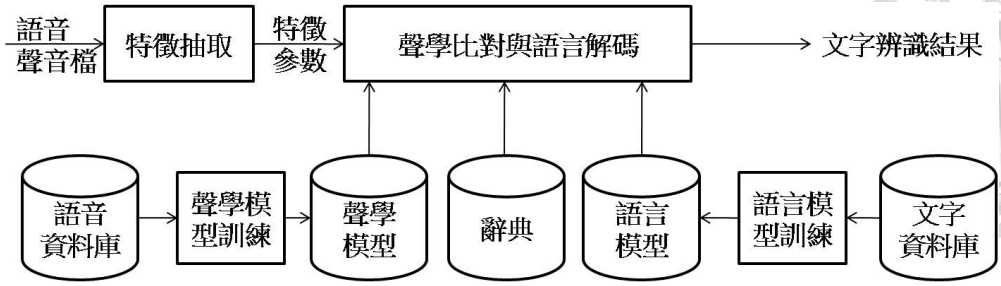



圖 1.1: 語音辨識基本架構圖

了一個重要的研究課題。 本論文發展了一系列基於變數選取方法的聲學模型調適技術，補償語音辨識系統測試時的聲學條件不匹配， 進而使得辨識正確率有效的被提升。

## 1.2 語音辨識原理簡介

本節簡介基於統計機器學習方法的語音辨識系統，其基本架構如圖1.1所示。我們會先從語音聲音檔中抽取出特徵參數， 再經過聲學比對 (acoustic matching) 及語言解碼 (linguistic decoding) 之後，得到文字辨識結果。特徵抽取步驟首先將語音訊號通過一組經過特殊設計的濾波器， 再於時間上有重疊 (overlap) 的音框 (frames) 上產生一組特徵參數序列。好的特徵參數對於語音資料的內容會有好的代表性， 梅爾倒頻譜係數 (Mel-frequency Cepstral Coefficients, MFCC) 和感知線性預測係數 (Perceptual Linear Predictive, PLP) 都是很好的例子。 在實做辨識的階段，我們使用聲學模型 (acoustic model) 的資訊來實做聲學比對，考慮特徵參數與模型的相似度；使用語言模型 (Language Model) 獲得相連詞出現的機率； 使用辭典 (Lexicon) 找出詞與音素的關係。 基於這些資訊，執行維特比動態規劃搜尋 (viterbi dynamic programming) 便能夠得到最合適的文字辨識結果。

統計式機器學習方法以機率的角度去解釋何謂合適的文字辨識結果。假

設$O$是輸入的觀察語音 (observation) ，給定此條件下，出現機率最大的詞串 (word sequence) $w^*$就是最適合的文字辨識結果。這樣的過程可以寫成:

$$w^* = \arg\max_{w \in W} p(w|O) \tag{1.1}$$

其中$W$為所有可能出現詞串的集合，$w$為$W$中的任一元素。$p(w|O)$被稱為詞串w的事後機率， 可進一步利用貝式定理 (Baye's Theorem) 拆解成:

$$p(w|O) = \frac{p(O|w)p(w)}{p(O)} \tag{1.2}$$

其中$p(O|w)$代表給定詞串$w$產生特徵向量觀測$O$的機率，為此我們建立了聲學模型來計算。$p(w)$ 代表詞串$w$的事前機率，須建立語言模型來計算。$p(O)$代表特徵向量觀測$O$的事前機率，因為和最適文字辨識結果無關， 在實作上並不考慮。最後，式1.1可被寫為:

$$w^* = \arg\max_{w \in W} p(O|w)p(w) \tag{1.3}$$

接下來將簡介語音辨識中會用到的聲學模型以及語言模型。

## 1.3 聲學模型

使用聲學模型的目的，在於模擬聲學基本單位 (basic unit) 在特徵空間 (feature space) 中產生特徵向量觀測$O$的機率。聲學的基本單位，可以是音素 (Phone) ， 中文的聲母/韻母 (Initial/Final) 或是詞 (Word) 。為了建立聲學基本單位和特徵向量之間的關係，我們需要蒐集語音資料庫，其中包含人說話的聲音檔， 以及對應的正確轉寫 (transcription) 。利用此資料庫進行聲學模型訓練，提高正確轉寫出現的機率，即可獲得令人滿意的聲學模型。

實作語音辨識系統時，常使用隱藏式馬可夫模型 (Hidden Markov Model, HMM) 來描述每個聲學基本單位。模型結構如圖1.2所示， 每個聲學基本單

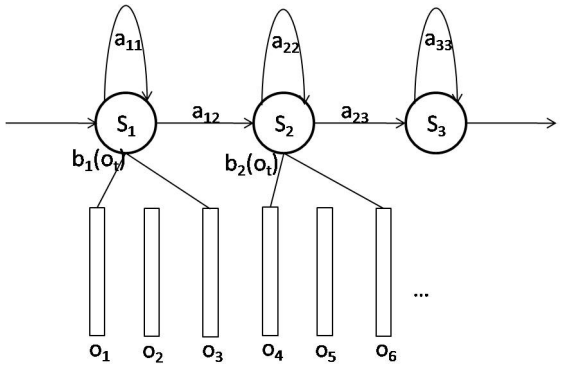



圖 1.2: 隱藏式馬可夫模型示意圖

位可被拆解成數個隱藏狀態 (State) ， 對該基本單位生成的語音訊號來說，時間軸上每一音框都有其對應的隱藏狀態。因此，每個隱藏式馬可夫模型需要由模型參數$\lambda = (A, B, \pi)$來描述。其中$\pi = \{\pi_i\}$，代表第i個 狀態的初始機率，也就是第一個音框屬於第i個狀態的機率；$A = \{a_{ij}\}$用以描述狀態間的轉移 (transition) ，$a_ij = p(s_t = j|s_{t-1} = i)$表示相鄰音框從狀態i轉移到狀態j的機率，一般我們會限制狀態轉移只能停留在同一個狀態，或是跳至鄰接的下一個狀態；$B = b_j(o_t)$中，$b_j(o_t) = p(o_t|s_t = j)$代表給定第j的狀態的條件下，產生特徵向量觀測$o_t$的機率。 最常見的聲學模型使用高斯混合模型 (Gaussian Mixture Model, GMM) ，將觀測機率寫為下式:

$$b_j(o_t) = \sum_{l=1}^{L} \omega_l g(o_t, \mu_l, \Sigma_l) \tag{1.4}$$

$$g(o_t, \mu_l, \Sigma_l) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_l|}} exp\left(-\frac{1}{2}(o_t - \mu)^T \Sigma_l^{-1} (o_t - \mu)\right) \tag{1.5}$$

其中$\omega_l$分別為L個高斯分布組成的權重；$\mu_l$與$\Sigma_l$分別為第l個高斯分布的平均 (mean) 和共變異矩陣 (covariance matrix) 。在一般的聲學模型中，我們假設共變異矩陣是對角矩陣 (diagonal matrix) 。

## 1.4 語言模型

使用語言模型的目的在於，提供一特定詞串$w$出現的機率$p(w)$。假定詞串$w$總共由K個詞， $w_1, w_2...w_k...w_K$，所組成，我們能將該詞串出現的機率拆解成下式:

$$p(w) = p(w_1, w_2...w_k...w_K) = \prod_k p(w_k|w_1, w_2...w_{k-1}) \tag{1.6}$$

在此式中，$w_1, w_2...w_{k-1}$被稱為歷史詞串 (history word sequence) 。因此上式的意義在於，給定到目前為止的歷史詞串，我們可以估測下一個詞出現的機率，由此方法估測出每個詞的機率後將其連乘，就可得到$p(w)$。

實作上為了得到適合的詞串機率分佈，我們同樣需要訓練語料來實做模型訓練。和聲學模型不同，語言模型訓練所需要的是大量文字資料，一般會使用比語音資料庫還要多的語料。然而，即使使用大量的語料，訓練語言模型仍然會碰到資料稀疏 (data sparsity) 的問題。這是因為，語言模型的機率分佈不適合用特定的密度函數 (density function) 來描述，我們建表的方式來儲存每一個詞在給予特定歷史詞串下的離散機率值，又因歷史詞串的可能性太多，該表將會大到無法儲存，我們能蒐集到的資料量也無法將這麼多的參數準確估計。為了解決這個問題，我們將1.6加上了n階馬可夫假設的限制，也就是假設第k個詞的出現機率只跟前n-1個詞構成的歷史詞串有關:

$$p(w_k|w_1, w_2...w_{k-1}) = p(w_k|w_{k-n+1}...w_{k-1}) \tag{1.7}$$

根據上式，需統計的參數將大大的減少，同時1.6，可被改寫為:

$$p(w) = \prod_k p(w_k|w_{k-n+1}w_{k-1}) \tag{1.8}$$

這就是常見的n連文法 (n-gram) 模型。本論文實作的語音辨識系統採用這種模型來模擬詞串的機率。

## 1.5 基於聲學模型調適的强健式語音辨識

如第一段提到，聲學條件的不匹配會對語音辨識系統的效能造成很大的傷害。為了解決這個問題、提高辨識率，許多方法相繼被提出。 著名的例子包括:特徵向量正規化 (feature normalization) [3] [4] [5]、聲學模型調適 (acoustic model adaptation) [6] [7]、以及調適訓練 (adaptive training) [8] [9]。 特徵向量正規化目的在於將原本的語音訊號轉換到新的特徵空間中， 以期待在這個空間中的特徵向量有强健性、不受聲學條件改變的影響。有名的例子包括梅爾倒頻譜平均消去 (Ceptral Mean Subtraction, CMS) 及梅爾倒頻譜平均變異度正規化 (Ceptral Mean Variance Normalization, CMVN) 。聲學模型調適利用少量測試聲學條件的語料，基於統計方法調整聲學模型或是特徵參數。最常見的方法有最大相似度線性迴歸 (Maximum Likelihood Linear Regression, MLLR) 、限制型最大相似度線性迴歸 (Constraint Maximum Likelihood Linear Regression, CMLLR) 以及聲道長度正規化 (Vocal Tract Length Normalization, VTLN) 。 最後，調適訓練結合了模型調適以及模型訓練方法，利用來自多樣化聲學條件的語料同時對模型參數及調適參數作統計上的最佳化， 常見的方法有語者調適訓練 (Speaker Adaptive Training, SAT) 和分群調適訓練 (Cluster Adaptive Training, CAT) 。

## 1.6 本論文研究貢獻

本論文的主要貢獻有如下兩點：

1. 本論文將變數選取方法 (Variable Selection) 應用在模型調適上、設計了一系列演算法、並實際在大字彙語音辨識的實驗上測試， 證實此方法能改善系統效能，使辨識率顯著的提升。

2. 本論文顯示變數選取方法能夠解決在實作模型調適時常碰到的問題：過度貼合 (over-fitting) 。 同時，本論文也實作了數個先前被提出，用以解決此問題的方法。實驗結果證明，變數選取方法更能有效的改善系統效能。

# 第 二 章　基於模型調適之强健型語音辨識

## 2.1　聲學模型調適摘要

如同1.1節中提到的，訓練與測試聲學條件的不匹配會嚴重影響語音辨識系統的效能。 聲學模型調適法藉由一些在測試條件下的語料，利用統計方法補償上述的不匹配。 以語者調適 (Speaker Adaptation) 為例；一般用來訓練聲學模型的語音資料庫會包含多個語者的聲音檔， 由於該模型並未特地適應某一語者，而被稱為語者不特定模型 (Speaker Independent Model) 。這樣的聲學模型對於任一語者通常能提供尚可的辨識率。 如果我們能為一特定語者蒐集到夠大的語音資料庫，建立語者特定模型 (Speaker Dependent Model) ，則理論上我們能獲得最佳的辨識率。然而，這樣的假設是不實際的， 因為在正常的情況下單一語者不願意錄製如此大量的資料庫。因此，想提升特定目標語者的辨識率，合理的做法是蒐集少量的語音資料，再使用模型調適技術稍微調整聲學模型， 如此即可獲得超過語者不特定模型、甚至逼近語者特定模型的效能。一般的調適技術也可以應用於處理其他種類的聲學條件不匹配。

由於模型調適方法可以有效地改善語音辨識系統的效能，各式各樣的研究相繼被提出來解決不同情境 (scenario) 下的模型調適任務 (task) 。 這些模型調適任務依據測試條件語料的多寡，以及正確轉寫的有無，可被區分成以下三類:

1. 監督式調適 (supervised adaptation) ：當有一定量符合測試聲學環境的調適語料 (adaptation data) 以及正確轉寫可利用，我們稱這樣的情境為監督式調適。 在此情境下，一般方法以最大化調適語料相似度為目標，調整聲學模型。跟另外兩種情境相比較，監督式調適是相對容易的。

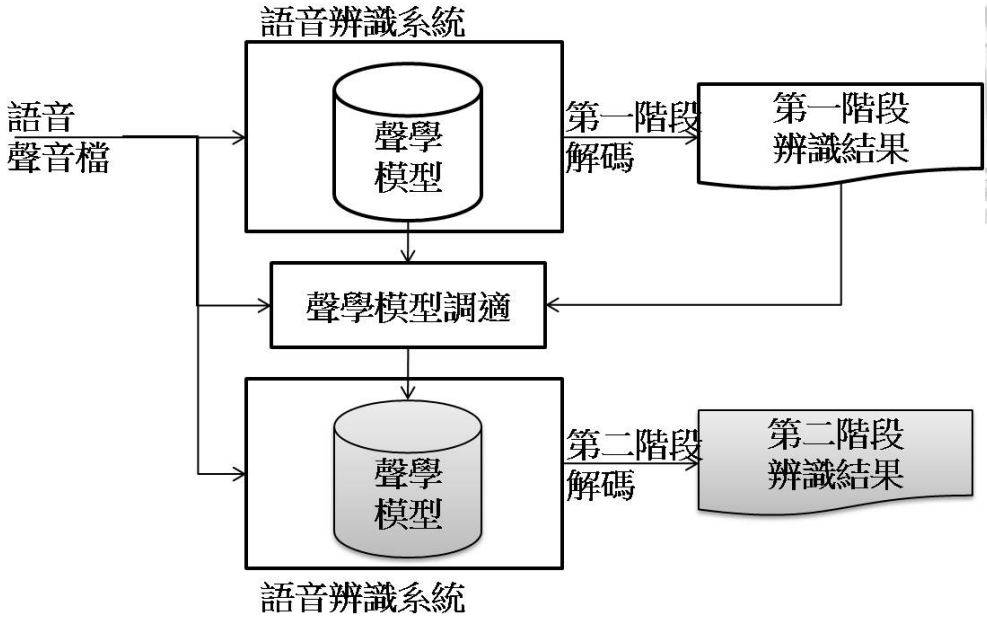



圖 2.1: 自我調適方法流程圖

2. 非監督式調適 (unsupervised adaptation) ：在非監督式調適情境下，雖然有一些調適語料可利用，但是缺少這些語料的正確轉寫。 我們通常先對調適語料做第一階段解碼 (first-path decoding) ，並將解碼出來的辨識結果當做正確轉寫，如此便能利用各種監督式調適的方法來調適模型。 然而，第一階段解碼有一個顯而易見的問題: 辨識結果免不了有錯誤，把有錯誤的轉寫當做正確答案去調適模型，有降低系統效能的風險。 有些非監督式調適的任務假設沒有額外的調適語料能取得，但是系統能夠一次獲得整批的測試語料。這種情境被稱為非監督式批次調適 (unsupervised batch adaptation) ，演講或是課程錄音的自動辨識都是常見的例子。

3. 自我調適 (self adaptation) ：自我調適比非監督式調適條件更為嚴苛。在此情境下，除了假設沒有任何額外的調適語料能使用外， 系統一次只能看到單獨一句，沒有轉寫資訊的測試語料 [10]。本論文做的實驗，皆是假設在這樣的的情境之下。 我們使用和非監督式調適情境相似的方式，先對測試語料

做第一階段解碼，並將辨識結果當做正確轉寫，作模型調適。 完成調適之後，再對原測試語料第二階段解碼 (second-pass decoding) ，流程如圖2.1所示。圖中標深色的聲學模型已依據測試語料被調適過。 在資料量極少、且缺乏正確轉寫的情況下，過度貼合問題非常嚴重，如直接套用和上面兩種情境相同的方法，通常不會獲得令人滿意的辨識結果。

## 2.2 過度貼合問題

過度貼合是在實作統計機器學習方法時常需要面對的問題 [11]，意指使用太過複雜的函數去貼合訓練資料 (training data) ， 如圖2.2所示。圖中給定訓練資料的特徵向量X以及標記Y， 我們利用一個函數來描述X和Y之間的關係，並希望當新的測試語料的特徵向量出現時，其標記可被預測。當我們使用太過複雜的函數，就如圖中紅線所示， 雖然在訓練資料上可以獲得極小的誤差，但是對測試資料來說卻不是一個好的預測。反之，如果選擇複雜度適中的函數，表現就相對穩定，如藍線所示， 在訓練及測試資料上都有不錯的效果。 過度貼合問題的發生，通常需滿足兩個條件: 其一是訓練資料量太少，以至於無法表達資料的正確分佈;再者是，我們使用的方法太過相信已知的訓練資料， 導致當測試語料和訓練語料稍微有不同時，系統效能會下降。

為了要解決過度貼合問題，可以從上面提到的兩個條件著手，把解決方法也分成兩種。首先，我們可以搜集更多的訓練資料，使之更逼近真正的資料分佈。在資料充足的情況下，使用複雜的模型也可以模擬出正確的分佈，系統效能自然也會很好。 然而，在沒有辦法取得更多訓練資料的情境下，我們必需使用更穩定的機器學習方法來改善系統效能。

聲學模型調適的任務本身特別容易遇到過度貼合問題。主要原因在於，調適

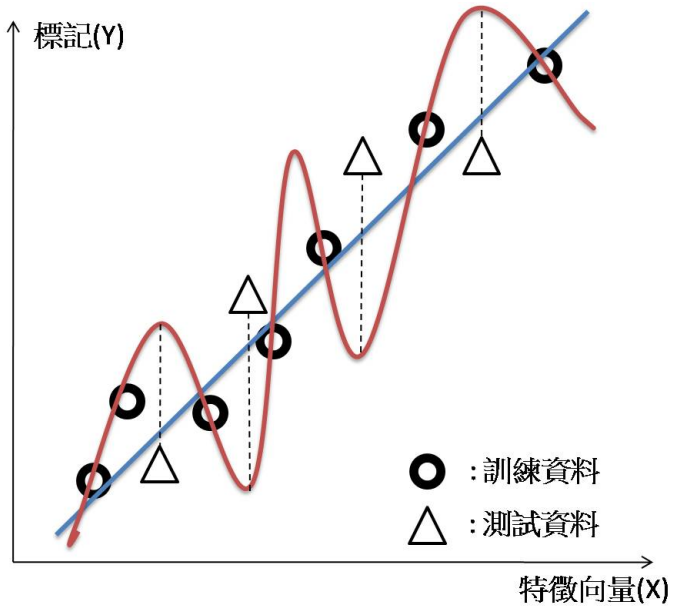



圖 2.2: 過度貼合示意圖

語料一般來說非常稀少，而聲學模型非常的複雜，參數非常多， 用正常模型訓練的方法來貼合調適語料不是一個好方法。考慮上述兩類解決過度貼合的方法，增加資料的方式不符合情境假設， 因此提出更穩定的機器學習方法來做模型調適。仿射轉換 (affine transformation) 是近年來最成功、最常被使用的模型調適法。 此方法讓聲學模型參數共用一組仿射轉換矩陣，有效減少了需要被估計的參數，成功的改善系統辨識率。此方法將會在下面的段落中詳細介紹。

## 2.3 仿射轉換模型調適方法

本節將介紹仿射轉換模型調適方法的理論以及實作方式。給定一個向量$x$，仿射轉換的通式如下:

$$x' = Ax + b \tag{2.1}$$

其中，$x'$為轉換後的向量；$A$和$b$分別為旋轉矩陣 (rotation matrix) 和偏差向量 (bias vector) 。此類調適方法便是利用仿射轉換， 將聲學模型參數轉換到另一個參數空

間中，得到新的模型，並期望新的模型會適合特定的測試聲學環境，增進語音辨識系統效能。 在調適語料量少的時候，仿射轉換會比直接重估測 (re-estimate) 模型參數要來的穩定。這是因為參數量從原本的數十到數百萬個模型參數，減少到數千個旋轉矩陣及偏差向量元素。 以常見的39維梅爾倒頻譜係數為例，旋轉矩陣及偏差向量元素總數為$39 \times 39 + 39 = 1560$。

仿射轉換可以應用到不同類型的模型參數上；同時，為了決定出適合的旋轉矩陣及偏差向量元素，我們需要定義出和調適語料有關的目標函數 (objective function) 。 根據不同的應用法及目標函數，可以得到各種仿射轉換模型調適方法。下面我們將介紹最常見的兩種方法: 最大相似度線性回歸 (Maximum Likelihood Linear Regression, MLLR) 以及限制型最大相似度線性回歸 (Constrained Maximum Likelihood Linear Regression, CMLLR) [12] [13]。

### 2.3.1 最大相似度線性回歸

最大相似度線性回歸以調適語料的的相似度為目標函數。首先我們介紹基於聲學模型，聲音特徵向量序列相似度的求法。給定一聲學模型，其內部的高斯混合模型參數集 (parameter set) 可被寫為$\Omega = \{\mu_i, \Sigma_i, \omega_i\}_1^I$，其中$i$是高斯分布的標籤，此聲學模型總共由$I$個高斯組成。$\mu_i, \Sigma_i, \omega_i$分別為第$i$個高斯的平均向量、共變異矩陣及權重。一段特徵向量序列$O$的相似度可被寫成:

$$p(O|\Omega) = \sum_Z p(O, Z|\Omega) \tag{2.2}$$

其中Z代表一些無法觀測到的變數，在語音辨識中，這些變數是在每個時間點的隱藏狀態。為了算出式2.2，理論上必須窮舉所有可能的$Z$，但在實作上無法達到。 因此我們利用期望值最大化演算法 (Expectation Maximization algorithm, EM algorithm) [14]，改計算一個最佳化條件與式2.2近似的輔助函數 (auxiliary

function) :

$$F(\Omega,\Omega')=E_{\Omega}(logp(O|\Omega'))=-\sum_{i}\sum_{t}\gamma_i(t)[(o_t-\mu_i')^T(\Sigma_i')^{-1}(o_t-\mu_i')-log|\Sigma_i'|]+C \tag{2.3}$$

這個輔助函數要被最大化的對象是對數相似度 (log-likelihood) 的期望值，並需要迭代以求得最大值。其中Ω、Ω′分別為原本模型參數及更新後的模型參數，$\gamma_i(t)$是在時間t的音框被第$i$個高斯產生的機率，$C$則是其他和模型參數無關的項。

最大相似度線性回歸利用仿射轉換調整聲學模型參數，包括平均向量和共變異矩陣。在這段中主要介紹調整平均向量的方法，利用式2.1，調適方程式可被寫成:

$$\mu_i'=A\mu_i+b=W\xi_i \tag{2.4}$$

其中$W=[A\ b]$，$\xi=[\mu_i\ 1]^T$。將式2.4代入式2.3，並將與$W$無關的項整理至$C$，即可得到最大相似度線性回歸的目標函數:

$$F(W)=-\sum_{i}\sum_{t}\gamma_i(t)(o_t-W\xi_i)^T(\Sigma_i)^{-1}(o_t-W\xi_i)+C \tag{2.5}$$

給定此函數，我們需要解出一個最適合的$W$，使之最大化。由於式2.5存在直接公式解 (close form solution) ，此問題並不困難。首先我們將一些較複雜的符號經過整理， 代換成以下變數:

$$\begin{aligned} G_j&=\sum_{i}\sum_{t}\frac{\gamma_i(t)}{\sigma_{ij}}\xi_i\xi_i^T \\ k_j&=\sum_{i}\sum_{t}\frac{\gamma_i(t)}{\sigma_{ij}}o_{tj}\xi_i^T \end{aligned} \tag{2.6}$$

其中$\sigma_{ij}$、$o_{tj}$分別為共變異矩陣的第 ($j$,$j$) 項元素及在時間t音框的第$j$維觀測特徵向量。將式2.6代入式2.5可得:

$$F(W)=-\sum_{j}\sum_{i}\sum_{t}w_jG_jw_j^T+2k_jw_j^T+C \tag{2.7}$$

其中$w_j$為仿射轉換矩陣$W$的第$j$列，另外與$W$無關的項在上式中被集中到$C$。我們很容易能看出，這是$w_j$的二次式，只要對$w_j$偏微分並令其為零，就可以求出最佳解。 以下即為最大相似度線性回歸法計算出仿射轉換矩陣的每一列:

$$w_j = k_j(G_j)^{-1} \tag{2.8}$$

最大相似度線性回歸在實作上有許多好處。其中最重要的一項就是有封閉型式解，這使得計算量比起其他方法要少得多、執行速度快，實做起來也容易。另外，當調適語料增加， 此方法還可以藉由將模型參數分群，每一群使用自己的仿射轉換，以達成更為細緻的模型調整。

### 2.3.2 限制型最大相似度線性回歸

限制型最大相似度線性回歸同樣也以調適語料對於模型的相似度作為目標函數，但是使用了不一樣的調適方程式:

$$\begin{aligned} \mu_i' &= A'\mu_i + b' \\ \Sigma_i' &= A'\Sigma_i(A')^T \end{aligned} \tag{2.9}$$

由式2.9可看出，聲學模型的平均向量也是經過仿射轉換得到，而共變異矩陣則是經過同樣大小的旋轉矩陣轉換，同時，此矩陣被限制成與用來轉換平均向量的旋轉矩陣相同。 將式2.9代入式2.3， 並經過適當整理可得到:

$$\begin{aligned} F(W) &= -\sum_i \sum_t \gamma_i(t)[(o_t' - \mu_i)^T(\Sigma_i)^{-1}(o_t' - \mu_i) + log|A|^2] + C \\ o_t' &= (A')^{-1}o_t + (A')^{-1}b' = Ao_t + b = W\zeta_t \end{aligned} \tag{2.10}$$

其中$\zeta_t = [o_t \ 1]^T$。由此可看出這個調適方程式，實質上等價於直接以仿射轉換調整聲學特徵向量。因此，此方法又常被稱為特徵空間最大相似度線性回歸(feature-space Maximum Likelihood Linear Regression, fMLLR) 。

限制型最大相似度線性回歸的目標函數中多了一項$log|A|^2$，因此在解其最佳化問題時，沒有辦法找出封閉形式解。利用一種較為特殊的迭代法， 求得最佳的仿射轉換矩陣的每一列$w_j$，如下式:

$$w_j = (k_j + \alpha p_j)(G_j)^{-1} \tag{2.11}$$

其中$p_j = [c_{j1}\ c_{j2}...c_{jD}\ 0], c_{jk} = cofactor(A_{jk})$。

$$\begin{aligned} G_j &= \sum_i \sum_t \frac{\gamma_i(t)}{\sigma_{ij}} \zeta_t \zeta_t^T \\ k_j &= \sum_i \sum_t \frac{\gamma_i(t)}{\sigma_{ij}} \mu_{ij} \zeta_t^T \end{aligned} \tag{2.12}$$

對照2.12，可發現2.11等式左右兩邊都和$W$有關，$\alpha$也是一待估參數。詳細的解法在附錄A中說明。

截至目前為止我們介紹了兩種以調適語料相似度為目標函數的仿射轉換方法。最大相似度線性回歸轉換的是模型參數， 而限制型最大相似度線性回歸則是轉換特徵向量。針對同一組測試語料， 哪一種方法會有比較好的表現，目前並沒有定論。另外，這兩種方法雖然都有效的減少待估參數量，但在遇到調適語料非常少、 過度貼合問題非常嚴重的情境時，仍然得不到更好的辨識結果。因此，我們需要更好更穩定的調適方法。

## 2.4 正則化仿射轉換

正則化 (regularization) 是一種相當汎用的機器學習方法 [11]，其主要概念在於以修改目標函數的方式，使待估參數變得更穩定。在仿射轉換模型調適法的例子中，我們將目標函數改成下列形式， 來達到正則化的效果:

$$F(W) = logp(O|W, \Omega) - \lambda R(W) \tag{2.13}$$

等式右邊第一項同樣是調適語料的對數相似度，而第二項中$R(W)$被稱為正則化項 (regularization term) ，其目的為量化參數$W$的變異程度。 當$R(W)$的值增加通常代表仿射轉換的穩定度下降， 因此將這項放入目標函數更能克服過度貼合的問題。式中$\lambda$為代表正則化項權重 (weight)的超參數 (hyper-parameter) ，越大代表越重視參數穩定度， 反之設定為零就等同於一般的仿射轉換方法。

將式2.13中的$R(W)$以不同的計算式帶入，即可得到不同的正則化方法。這些方法有各自的物理意義以及優缺點。下面將介紹三種不同的正則化仿射轉換方法，分別為: L2規範 (L2 norm) 正則化 [15]、L1規範 (L1 norm) 正則化 [16]以及最大事後機率線性回歸 (Maximum a Posterior Linear Regression, MAPLR) [17] [18]。

### 2.4.1 L2規範正則化

L2規範正則化以下式作為正則化項:

$$R(W) = ||A||^2 = \sum_{i,j} a_{ij}^2 \tag{2.14}$$

等號最右邊為旋轉矩陣$A$每項元素的平方和。利用此式來作正則化，能使得$A$中元素的大小被限制，進而使回歸值較穩定、不會隨特徵向量的擾動而改變太多。 L2規範正則化可以被應用在基於模型或是特徵向量的模型調適法中 (MLLR及CMLLR) ，作法在以下介紹。

將式2.14代入式2.13，並將對數相似度以與式2.3相同的輔助函數代入，可得基於模型的L2規範仿射轉換法的目標函數:

$$F(W) = -\sum_i \sum_t \gamma_i(t)(o_t - W\xi_i)^T(\Sigma_i)^{-1}(o_t - W\xi_i) - \lambda||A||^2 \tag{2.15}$$

此式的最佳化同樣有直接公式解。對$W$的每一列偏微分並令其為零，我們便能以式2.8求得最適的仿射轉換，其中$G_j, k_j$要被改寫為:

$$G_j = \sum_i \sum_t \frac{\gamma_i(t)}{\sigma_{ij}} \xi_i \xi_i^T + [\lambda,\ \lambda,\ ..., 0] I$$
$$k_j = \sum_i \sum_t \frac{\gamma_i(t)}{\sigma_{ij}} o_{tj} \xi_i^T \tag{2.16}$$

其中$[\lambda,\ \lambda,\ ..., 0] I$為一方陣，其非對角線元素及最後一個對角線元素0。

以相同的方法改寫式2.10，即可得到基於特徵向量的L2規範仿射轉換法之目標函數:

$$F(W) = -\sum_i \sum_t \gamma_i(t) [(W\zeta_i - \mu_i)^T (\Sigma_i)^{-1} (W\zeta_i - \mu_i) + log|A| \tag{2.17}$$

此式的最佳化步驟與限制型最大相似度線性回歸相同，需要以迭代法來求。我們同樣利用式2.11來更新仿射轉換舉陣$W$，其中$G_j, k_j$要被改寫為:

$$G_j = \sum_i \sum_t \frac{\gamma_i(t)}{\sigma_{ij}} \zeta_t \zeta_t^T + [\lambda,\ \lambda,\ ..., 0] I$$
$$k_j = \sum_i \sum_t \frac{\gamma_i(t)}{\sigma_{ij}} \mu_{ij} \zeta_t^T \tag{2.18}$$

在實作上，L2規範正則化無論是被應用在基於模型或是特徵向量的仿射轉換法，都幾乎不需要增加任何的計算量， 並且都能有效的克服過度貼合問題，使語音辨識系統效能進步。

### 2.4.2 L1規範正則化

L1規範正則化以參數的L1規範作為正則化項:

$$R(W) = ||A|| = \sum_{i,j} |a_{ij}| \tag{2.19}$$

將此項加入目標函數中，同樣也會有限制$A$中元素的大小，使回歸值穩定的功能。然而與L2規範不同的是，L1規範有使待估參數稀疏 (sparse) 的特性，也就是

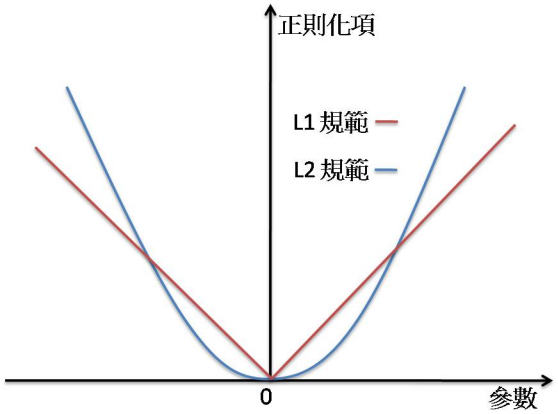



圖 2.3: L1與L2規範正則化項

非零的參數數目傾向減少。 此特性的完整證明十分困難，在本論文中不會提及，但此特性可用圖2.3來作粗略的解釋。圖中座標橫軸為某一待估參數，縱軸為其對正則化項增加造成的貢獻， 紅線藍線分別代表L1、L2規範。在靠近座標原點處可以發現L2規範的斜率接近零，代表在參數很小的時候，繼續減小參數對減小正則化項的貢獻很少。而L1規範則是有固定的斜率， 即使在參數已經很小的時候，仍然有往原點靠近的趨勢。參數的稀疏性 (sparsity) 的好處目前沒有一個定論。一個常見的解釋是，特徵向量中與標記無關的元素， 對回歸之類的方法來說如同雜訊，而具稀疏性 (sparsity) 的回歸方法可以用將參數設為零的方式，自動的去除這些雜訊。

仿射轉換目標函數在加上L1規範後，最佳化變得比較困難。截至目前為止的研究，只將此正則化的方法應用在基於模型的仿射轉換法上，下面將介紹詳細的作法。 把式2.19代入式2.13，並用和L2規範正則化法相同的方式代入輔助函數，可以得到L1規範仿射轉換法的目標函數:

$$F(W) = -\sum_{i}\sum_{t}\gamma_i(t)(o_t - W\xi_i)^T(\Sigma_i)^{-1}(o_t - W\xi_i) - \lambda||A|| \qquad (2.20)$$

由於此函數的直接公式解並不存在，我們採用的作法是給定初始的仿射轉換矩陣 (通常為最大相似度的解) ，在單獨更新旋轉矩陣$A$中的每個元素$a_{jk}$。 在假設$A$中$a_{jk}$以外的元素、偏差向量$b$都固定，式2.20可被改寫成:

$$F(A_{jk}) = -\sum_i \sum_t \frac{\gamma_i(t)}{\sigma_{ij}}(o_{tj} - b_j - (A_{-jk}\mu_i)_j)^2 - \lambda|a_{jk}| \tag{2.21}$$

其中$A_{-jk}$為目前估測的旋轉矩陣，除了第$(j,k)$元素被取代為零。將式2.21對$a_{jk}$微分並令其為零，整理後可得$A_{jk}$的更新式:

$$a_{jk} = (|\frac{c}{d}| - \frac{\lambda}{d})_+ sign(\frac{c}{d}) \tag{2.22}$$

其中

$$\begin{aligned} c &= \sum_i \sum_t \frac{\gamma_i(t)}{\sigma_{ij}} \mu_{ik}(o_{tj} - b_j - (A_{-jk}\mu_i)_j) \\ d &= \sum_i \sum_t \frac{\gamma_i(t)}{\sigma_{ij}} \mu_{ik}^2 \\ (x)_+ &= max(0, x) \end{aligned} \tag{2.23}$$

利用式2.22輪流更新$A$中每一個元素，待其收斂後即可得到目標函數2.20的近似解。

L1規範正則化法的實作較為困難，並且會消耗更多的計算資源，然而通常會帶來比L2規範更多的進步，在各個機器學習的應用領域中都有相同的結論。 本論文將這兩種方法都實作在模型調適問題上，比較兩方法的效能。

### 2.4.3 最大事後機率線性回歸

此方法從機率的角度來考慮正則化。以仿射轉換$W$的事前機率 (prior) 負對數值$p(W)$作為正則化項，並將之代入2.13，可得到最大事後機率線性回歸的目標函數:

$$F(W) = logp(O|W,\Omega) - \lambda R(W) = logP(O|W) + \lambda logp(W) \tag{2.24}$$

此式正好符合$W$的事後機率$p(W|O,\Omega)$以貝式定理展開。事前機率項在此方法中同樣發揮了限制參數變異程度的作用，使調適方法穩定性增加。

如何估測合理的事前機率分布式此方法的重要議題，有許多相關研究提出各種不同的估測方法。這些方法大多都需要利用額外的、處於不同聲學條件下的語料， 求得多個$W$的樣本後，再利用這些樣本求得事前機率分布。在本論文中，我們手動設計了一個合理的機率分布:

$$p(W) = p(\Omega') = \prod_i p(\mu_i') = \prod_i \mathcal{N}(\eta_i, \Phi_i) = \prod_i \mathcal{N}(\mu_i, a\Sigma_i) \tag{2.25}$$

在此式中，我們作了兩個假設:

1. 仿射轉換矩陣$W$的事前機率可以寫成被調適聲學模型 (adapted acoustic model) $\Omega'$的事前機率。而此機率又只跟被調適過的參數有關，此例中是每個高斯分布組成的平均向量。

2. 每個高斯平均向量$\mu_i'$的事前機率可被寫成另一個高斯分布，平均向量以及共變異矩陣分別為$\eta_i$、$\Phi_i$。以此為基礎，進一步假設$\eta_i = \mu_i$、$\Phi_i = a\Sigma_i$。

根據此事前機率的設定以及式2.24，同樣可以利用式2.8來求的最佳的仿射轉換矩陣的每一列$W_j$，其中:

$$\begin{aligned} G_j &= \sum_i \sum_t (\frac{\gamma_i(t)}{\sigma_{ij}} \xi_i \xi_i^T) + \lambda \sum_i \frac{1}{\sigma_{ij}} \xi_i \xi_i^T \\ k_j &= \sum_i \sum_t \frac{\gamma_i(t)}{\sigma_{ij}} o_{tj} \xi_i^T + \lambda \sum_i \frac{1}{\sigma_{ij}} \mu_{ij} \xi_i^T \end{aligned} \tag{2.26}$$

式2.25中的$a$以被整合進$\lambda$中故並不需要寫出，調整$\lambda$同時也相當於對事前機率的變異度作調整。

本論文所提出的最大事後機率線性回歸，在實作上不需增加太多的計算量。因為算式中關係到事前機率的部分與調適語料無關，可以先計算好。另外此方法調整超參數$\lambda$也比較容易，通常$\lambda$為一接近零的正數。

## 2.5 本章結論

本章前半段簡介聲學模型調適方法，包含其基本概念以及分類介紹，接著介紹了最常見的一種模型調適法: 仿射轉換法。後半段提出在調適語料極少的情況下，會碰到的過度貼合問題，並介紹如何將前人提出的正則化法，應用在仿射轉換模型調適法中，以解決過貼合問題。接下來的章節中將透過實驗比較這些方法的效能。

# 第 三 章　基礎實驗

## 3.1　Aurora-4基本設定

本論文中，我們使用Aurora-4語料庫 [19]來對各種模型調適方法作效能評估。Aurora-4是一個世界知名的基準 (benchmark) 語料庫，主要被用來評估語音辨識系統在不同雜訊以及通道效應下的效能。 Aurora-4的文字內容來自Wall Street Journal 0，在錄製語音資料時，使用Sennhieser麥克風錄製一份，多種第二麥克風 (如:Sony ECM-55, Sony ECM-50ps, Crown PZM-6FS, Crown PCC-160, RAdioShack omni-electret, Nakamichi CM100, AT&T 5400cordless phone, Panasonic KXT2365 speaker phone...) 錄製了另一份。另外，除了乾淨語料外，Aurora-4另外以人工方式加上了六種不同的加成性雜訊，其干擾程度為訊噪比 (signal noise ratio, SNR) 10dB至20dB。 製作成聲音檔時有8kHz和16kHz兩種不同取樣頻率，本論文中的實驗僅採用8kHz取樣頻率的聲音檔。

Aurora-4本身提供了兩種聲學模型訓練模式，分別為:乾淨條件訓練 (clean condition training) ，其訓練語料由7138句乾淨語音檔構成; 複合條件訓練 (multi-condition training) 同樣也有7138句訓練語料， 其中四分之一沒有雜訊，另外四分之三加有人工雜訊，詳細配置如圖3.1。本論文以複合條件訓練來得到聲學模型，並基於此架構基礎的語音辨識系統。 我們將訓練語料中每一語句轉換成梅爾倒頻譜係數及其一次二次微分，共39維特徵向量序列，並利用Hidden Markov ToolKit (HTK) 實做跨詞三連音素 (cross-word triphone) 聲學模型的訓練。語言模型則是利用7138句訓練語料的正確轉寫做出三連文法模型。

測試語料部份，Aurora-4提供了14組測試集，每組有166句。第01～07組是Sennhieser麥克風錄製，其中第01組為乾淨語料，第02～07組分別以人工的方

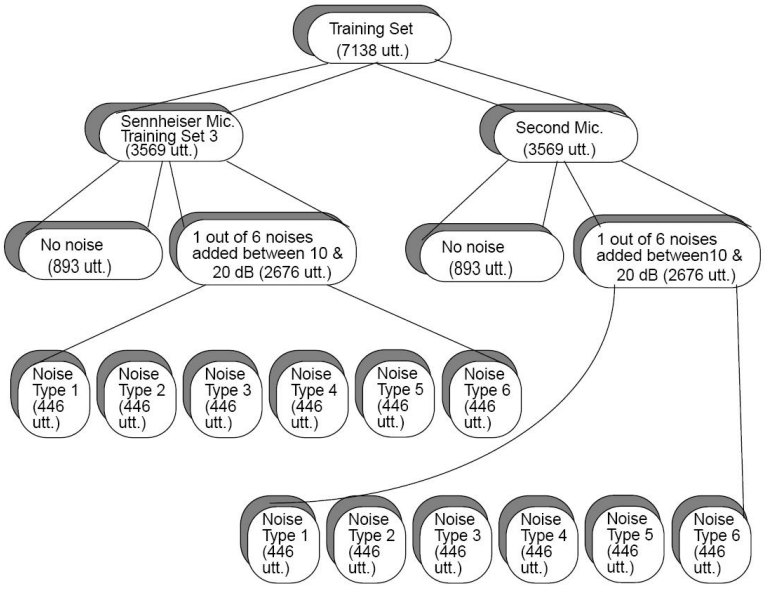



圖 3.1: Aurora-4 訓練集語料詳細設定

式加入以下類型的雜訊: 汽車雜訊、人聲雜訊 (babble noise) 、餐廳雜訊、街道雜訊、機場雜訊、火車雜訊。第08～14組語料是第二支麥克風錄製，第08組是乾淨語料，第09～14組的雜訊設定順序和第02～07組相同。 詳細的測試集語料設定如圖3.2所示。

本論文以Aurora-4為基礎，完成了一系列自我調適情境 (self adaptation) 的實驗，意即針對每一句測試語料，先做第一道解碼辨識，以辨識結果當做正確答案來做聲學模型調適，接著做第二道解碼當作最後答案。以下將列出各種方法在14組測試集下的分別以及平均辨識結果，以字正確率 (word accuracy) 表示。

## 3.2 基礎實驗結果

本節列舉Aurora-4複合條件訓練的基本系統 (baseline system) 以及一般仿射變換調適法 (最大相似度線性回歸、限制型最大相似度線性回歸) 在測試集上的字正確率。根據實驗結果，提出相關的比較和討論。

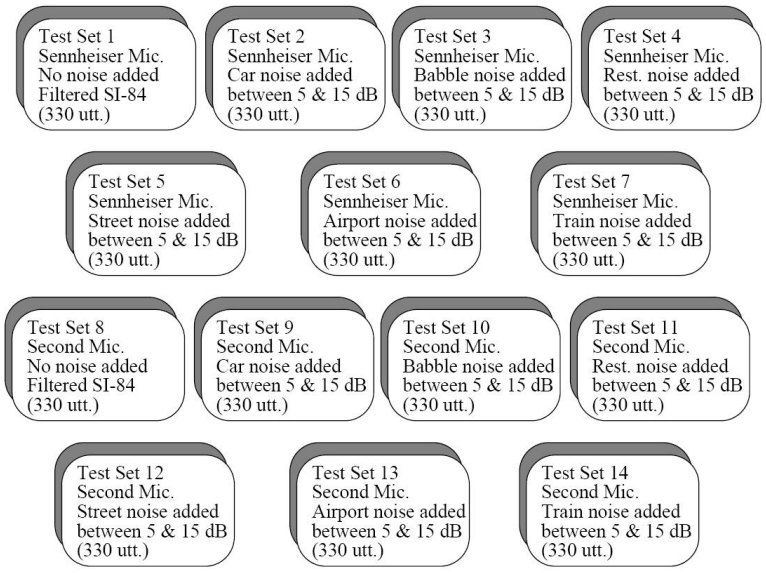



圖 3.2: Aurora-4 測試集語料詳細設定

表3.1列舉了Aurora-4複合條件訓練的測試集字正確率。由此表我們可以觀察到，在相同麥克風條件下， 乾淨語料測試集 (組合01、組合08) 的辨識率都超過有雜訊的測試集；另外，給定相同雜訊條件， 第一麥克風的測試集的辨識率都超過第二麥克風。這是因為在複合條件訓練的測試集中，乾淨語料佔大多數，第一麥克風錄製的語料也佔大多數。 最後，複合條件訓練的基本系統，對於測試集組合01～14平均辨識率為77.47%。

我們將Aurora-4複合條件訓練下的聲學模型，在自我調適的情境下分別用最大相似度線性回歸 (MLLR) 、限制型最大相似度線性回歸調適 (CMLLR) ， 在各測試集下得到的辨識率列舉於圖3.3，並將測試集平均的辨識率列於表3.2。 由圖3.3我們可以觀察到，各個測試集中，兩種調適方法跟基本系統 (baseline) 互有勝負。這是因為在自我調適的情境下， 調適方法只有一句測試語料可利用，同時也缺少正確轉寫，特別容易使得過度貼合問題發生，使得系統效能進步不明顯甚至退步。 由表3.2中可看出兩種仿射變換調適法，在測試集平均辨識率上都比基本系統差， 由此結果我們再次證實自我調適是一個非常困難的任務。

表 3.1: Aurora-4 複合條件訓練 實驗結果

| 測試集 | 第一麥克風 (Sennhieser) | | | | | | |
|---|---|---|---|---|---|---|---|
| | 組合01 | 組合02 | 組合03 | 組合04 | 組合05 | 組合06 | 組合07 |
| 雜訊型態 | 乾淨 | 汽車 | 人聲 | 餐廳 | 街道 | 飛機場 | 火車 |
| 字正確率(%) | 89.02 | 88.29 | 81.55 | 78.45 | 77.75 | 80.33 | 76.24 |

| 測試集 | 第二麥克風 | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 組合08 | 組合09 | 組合10 | 組合11 | 組合12 | 組合13 | 組合14 | 平均 |
| 雜訊型態 | 乾淨 | 汽車 | 人聲 | 餐廳 | 街道 | 飛機場 | 火車 | 組合01 14 |
| 字正確率(%) | 82.95 | 80.22 | 73.52 | 69.54 | 66.52 | 72.56 | 67.66 | 77.47 |

表 3.2: Aurora-4 仿射變換調適法 實驗結果

| 調適方法 | baseline | MLLR | CMLLR |
|---|---|---|---|
| 字正確率(%) | 77.47 | 77.24 | 77.38 |

## 3.3 正則化仿射變換實驗結果

我們將2.4節中提出的各種正則化仿射變換調適方法實做在Aurora-4上，其中基於模型的方法包含L2、L1規範正規化、 最大事後機率線性回歸；基於特徵向量的方法僅有L2規範正規化。列出這些方法在測試集上的辨識率、並提出相關比較與討論。

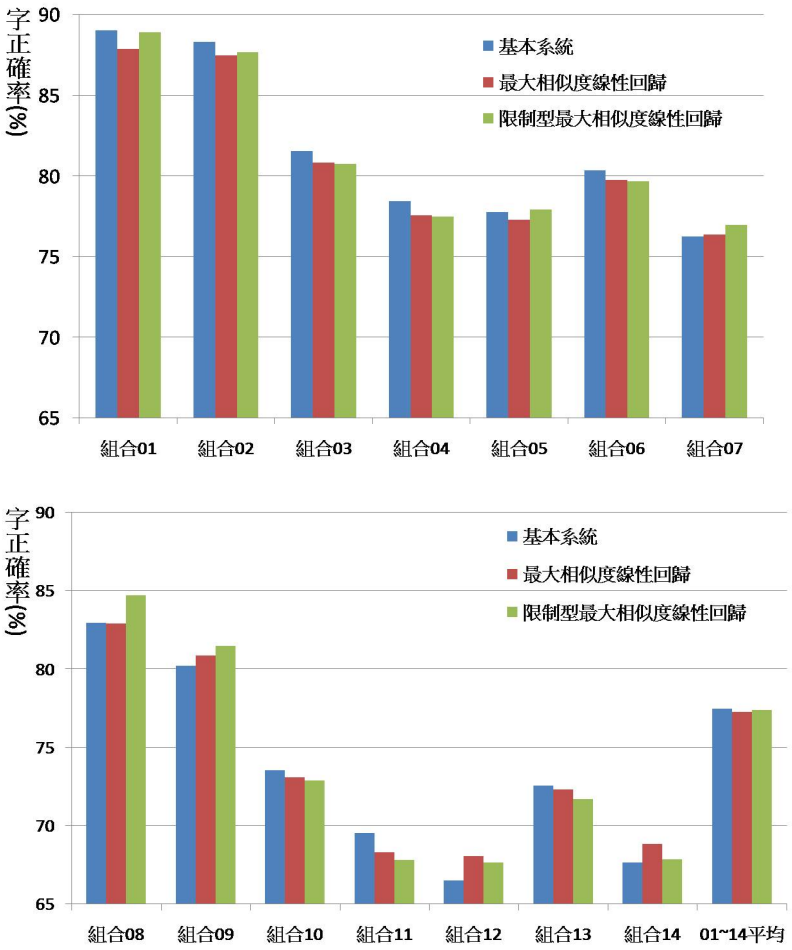



圖 3.3: Aurora-4 一般仿射變換調適法 實驗結果

### 3.3.1 基於模型的正則化仿射變換

各種基於模型的正則化仿射變換方法在Aurora-4各測試集上的辨識結果列於圖3.4，基本系統的辨識率也被列出以供比較。由於每個方法都有一個超參數， 正則化項權重λ可供調整，圖3.4中列出的是各方法效能最佳的超參數，我們另外將每個方法的測試集平均辨識率和超參數關係圖畫出。

首先，觀察圖3.4可發現，與基本系統相比較，各種正則化仿射變換對測試集組合01～07效能差別不明顯，對組合08～14則是都有辨識率進步。 比較全部測試集01～14的平均可看出，L1規範正則化和最大事後機率線性回歸得到較好的結

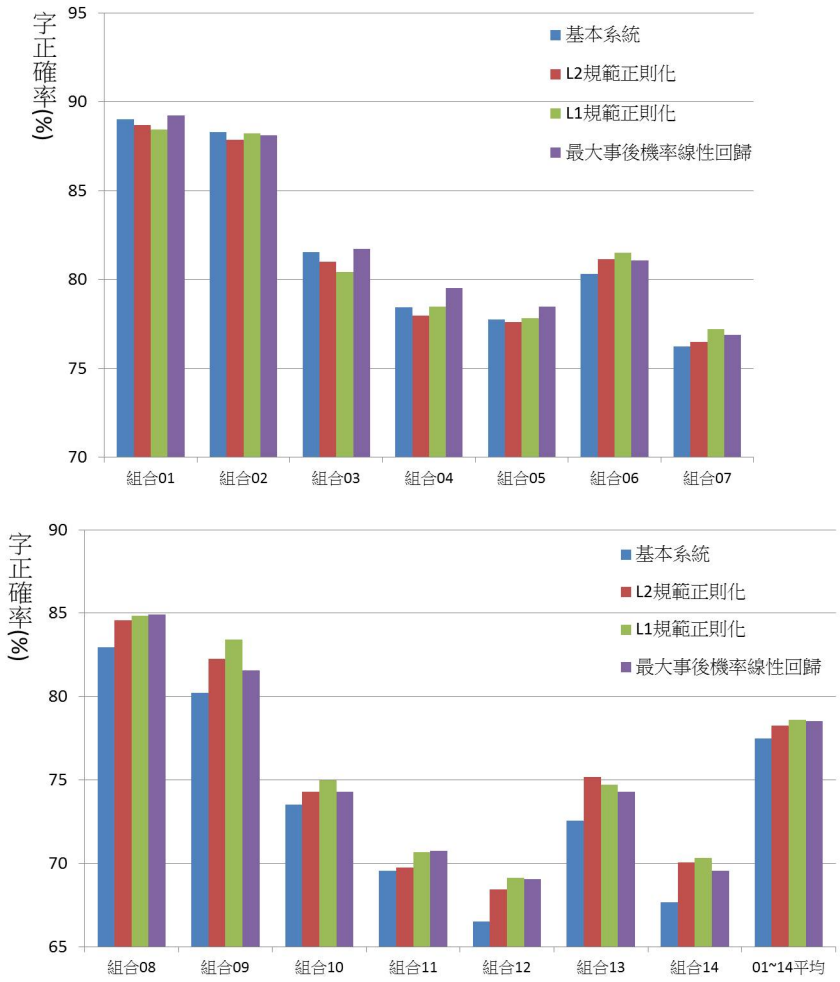
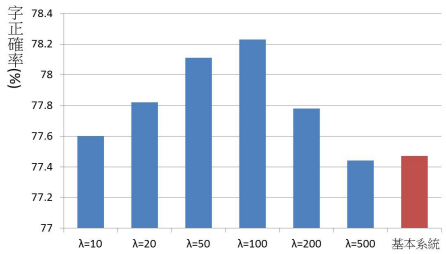
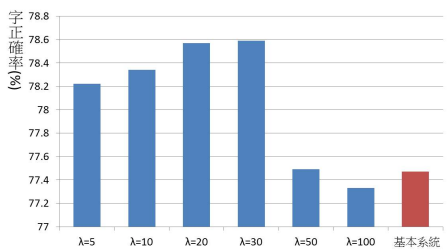


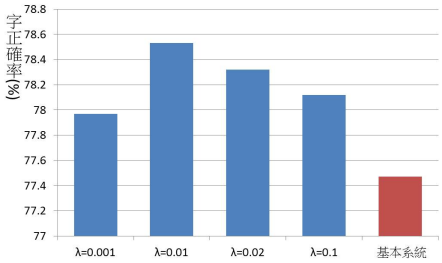

圖 3.4: Aurora-4 基於模型之正則化仿射變換調適法 實驗結果




(a) L2規範正則化




(b) L1規範正則化




(c) 最大事後機率線性回歸

圖 3.5: 不同正則化權重λ的實驗結果

果。

圖3.5a、3.5b、3.5c分別畫出L2規範正則化、L1規範正則化和最大事後機率線性回歸， 正則化權重λ與測試集平均辨識率的關係。從圖中可看出， 各種正則化方法都有一個適合的超參數範圍， 在這個範圍內可以使得辨識率超過基本系統。當超參數太大，會給待估參數太多限制；太小則是沒有達到正則化的效果，趨近於一般仿射變換方法的結果。

### 3.3.2 基於特徵向量的正則化仿射變換

基於特徵向量的正則化仿射變換法，在本論文中只實做了L2規範這一種。主要是因為L1規範的方法較困難，目前還沒有適當解法被發表；而最大相似度線性回歸實做在特徵向量空間， 通常需要額外的資料估計事前機率，要求比其他方法多，因此不放在一起比較。圖3.6列出基於特徵向量的L2規範正則化在各測試集上的辨

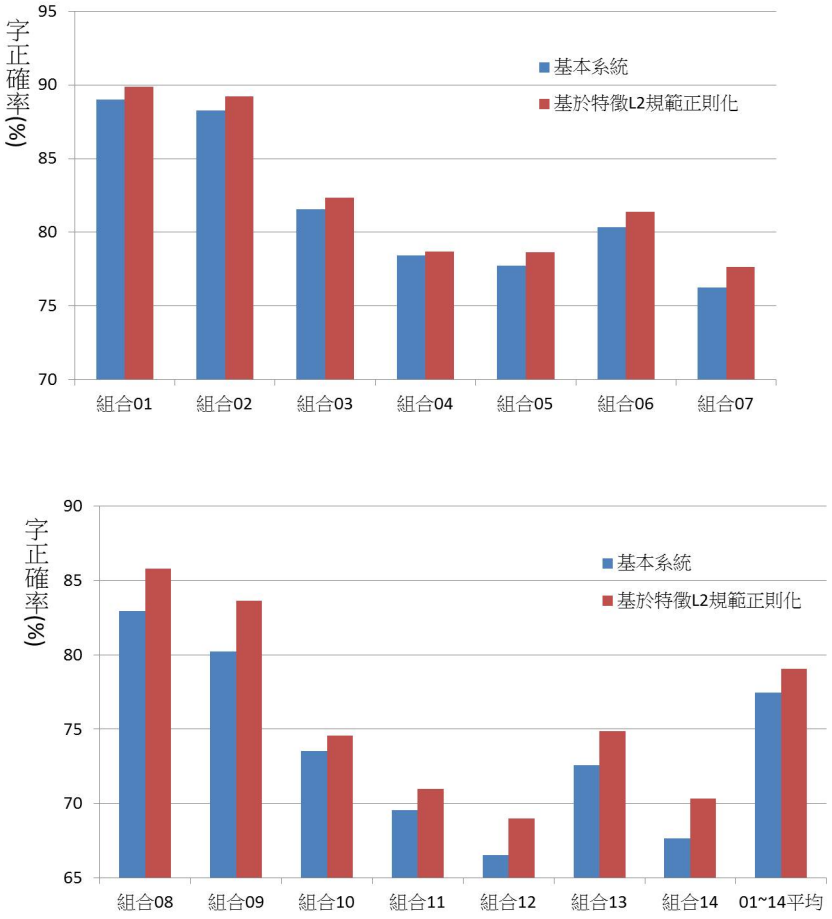



圖 3.6: Aurora-4 基於特徵向量之正則化仿射變換調適法 實驗結果

識結果， 觀察可發現在測試集中每個組合；以及平均值都明顯勝過基本系統。圖3.7同樣列出超參數λ與測試集平均辨識率的關係， 可以看出與基於模型的各種方法有一樣的趨勢。 最後表3.3列出截至目前為止所有調適方法在所有測試集上的平均辨識率，包含基本系統、最大相似度線性回歸 (MLLR) 、限制型最大相似度線性回歸 (CMLLR) 、 基於模型L2規範正則化 (L2-norm) 、基於模型L1規範正則化 (L1-norm) 、最大事後機率線性回歸 (MAPLR) 以及基於特徵向量L2規範正則化 (feature L2-norm) 。我們發現從平均上來看每種正則化方法都達到克服過度貼合問題的效果， 其中基於特徵向量L2規範正則化效果最好。

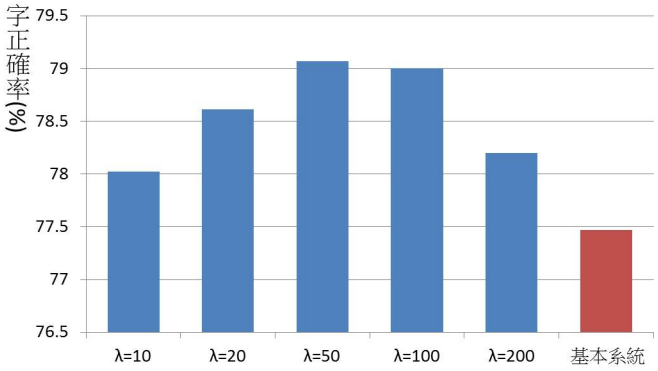



圖 3.7: L2規範正則化 不同正則化權重λ的實驗結果

表 3.3: Aurora-4 仿射變換調適法 實驗結果

| 調適方法 | baseline | MLLR | CMLLR | L2-norm |
|---|---|---|---|---|
| 字正確率(%) | 77.47 | 77.24 | 77.38 | 78.24 |
| 調適方法 | L1-norm | MAPLR | feature L2-norm | |
| 字正確率(%) | 78.57 | 78.53 | 79.07 | |

## 3.4 本章結論

本章列舉並比較了基本系統以及各種調適方法的效能。從中我們能觀察到數點：

1. 在許多調適任務中獲得成功的一般仿射變換調適法，在自我調適上沒有辦法得到令人滿意的結果。原因是由於過度貼合的發生
2. 各種正則化法可以幫助我們解決過度貼合問題，改善系統效能。
3. 對Aurora-4測試語料來說，基於特徵向量L2規範正則化可得到最好的辨識率。可能是因為在特徵向量空間上做調整比較符合背景雜訊、通道效應這類的聲學條件不匹配。

截至目前為止，我們證明了正則化法對於自我調適任務的幫助。文獻中指出 [11]，正則化法之所以能發揮作用，是因為此方法控制了模型參數的複雜度，例如L1正則化會使得不重要的參數自動被設定為零。如果我們能以更有效的方式來控制參數，也許能夠使系統效能更進一步提升。因此，本論文接下來的章節提出了基於變數選取方法的模型調適法。

# 第 四 章　變數選取方法

## 4.1 原理簡介

本章介紹一種在統計機器學習、資料探勘 (data mining) 領域常用到的一門技術: 變數選取 (variable selection) [20] [21]。此技術可以幫助機器提升處理預測問題 (prediction problem) 的效能，包含預測準確性和計算資源的消耗量。

我們首先需要敘述預測問題。假設有N筆觀測資料$\{\boldsymbol{x_k}, y_k\}(k = 1, ..., N)$，其中$y_k$代表第$k$筆資料的標記， 可以是一實數或屬於一個有限集合，$\boldsymbol{x_k}$則是第$k$筆資料的變數向量。所謂的預測問題，就是利用給定的N筆資料，求得一個函數$y = f(\boldsymbol{x})$， 當我們收到一筆僅有變數向量資料，可以利用函數來預測標記。舉例來說，在自然語言處理 (natural language processing, NLP) 領域中常見的文章分類 (document classification) 任務就是一個典型的預測問題。在這個例子中文章的類別就是待預測的標記$y$， 如廣告、科技、遊記...等等；而變數向量$x$可以包含任何跟一篇文章有關的資訊: 例如文章中出現的詞、n連文法。 給定已知類別的文章N篇，我們希望找出一個函數能夠分類未知類別的新文章。

在實做預測問題時，我們會盡可能的多蒐集各式各樣的變數來幫助預測，因為蒐集的越多，越有可能得到能夠區別不同標記的變數。例如在文章分類任務中， 作者的資訊顯然對文章類別的區分很有幫助。但是蒐集太多的變數也存在一些缺點。比如說耗去較多的儲存空間、在使用統計機器學習方法時增加訓練及測試所需要的計算量、 還有太多雜亂的變數會影響預測準確性。前兩個缺點是顯而易見的，關於第三點，理論上只要使用夠穩定的統計機器學習方法，任何和標記無關的雜亂變數， 在最後得出的預測函數中會被自動忽略，不影響預測準確性。然而，在觀測資料的數目不足的時候，我們很難去觀測到這些雜亂變數的真正統

計特性，進而得出"與標記無關"的結論。 因此，一系列有關變數選取的方法相繼被提出，希望能選出真正有用的變數，進而改善系統效能。

變數選取方法主要目的在於，從所有能從蒐集到的變數中選出真正對預測問題有幫助的部分。一般來說，變數選取方法可以被區分為兩大類： 打包法 (wrapper method) 、以及濾波器法 (filter method) 。本章接下來的部份將介紹這兩類方法，並討論它們與模型調適任務間的關係。

## 4.2 打包法

打包法的基本實作方式非常直觀。在預測問題中，如果將能蒐集到的所有變數種類視為一個集合，打包法的作用在於找出一個變數子集 (subset) 使得系統對此預測問題的效能達到最好。基本架構圖如4.1所示。由圖中可看出，打包法需要定義在一個用來解決預測問題的目標演算法上 (target algorithm) ， 在整個實作過程中，該演算法會被視為一個黑盒子，打包法會將每個變數子集帶入該演算法中，並在一個額外的發展資料集 (development data set) 中評估使用該變數子集的預測準確率。得到最高預測準確率的變數子集被選取，作為後續正式執行目標演算法使用的變數集。 打包法基本上適用於任何的預測問題以及演算法，並且通常提供令人滿意的結果。在一般的情況中，打包法找出的子集，其元素數量大多會遠小於整個變數集， 因此目標演算法的計算速度一定能被大幅提高。另外，只要發展資料集夠具代表性，我們可以預期打包法選出來的變數子集被應用在目標演算法後， 同樣對真正測試集能提供較高的預測準確率。

雖然打包法可以有效減少實際執行目標演算法的時間，但是執行打包法本身所需要的時間非常驚人，這是因為打包法需要窮舉所有可能的變數子集。 假設我們蒐集到$M$種變數，所有變數子集的數量為$2^M - 1$，也就是說，目標演算法必須

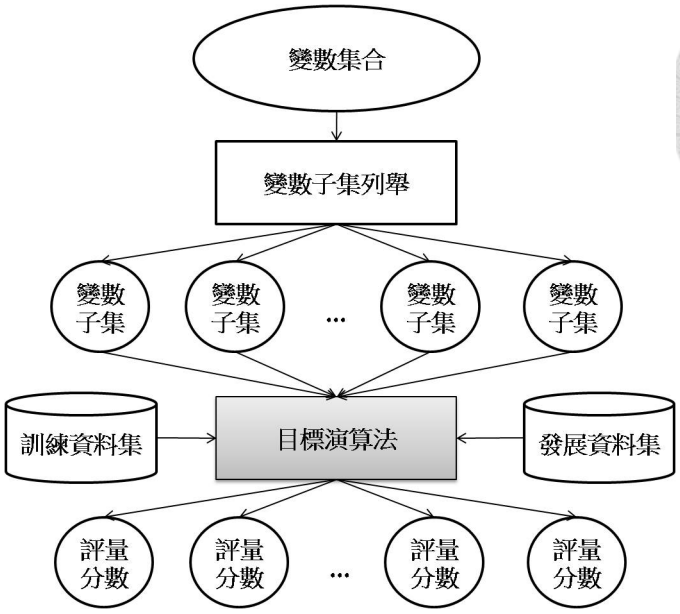



圖 4.1: 打包法 基本架構圖

被執行$2^M - 1$次。在實際的例子中，$M$可以等於數千數萬， 即使把執行打包法的工作視為離線程序 (offline process) ，執行時間還是太久。因此，我們需要一些近似 (approximation) 的方法來解決這個問題。

打包法中最常見的近似法是利用貪婪搜尋 (greedy search) 來取代窮舉法，實作變數子集列舉。正向選取 (forward selection) 和反向消去 (backward elimination) 都是很有效的貪婪搜尋法。向前選取演算法，由空集合開始建構變數子集，將變數一個個加入子集中。第一個變數選取單獨使用使目標演算法效能最好的變數，從第二個變數開始， 選取和上一步變數子集合併後，使效能最好的變數。如此重複至加入任何一變數都不能使效能上升。向後消去演算法則是從完整的變數集開始，每次消去一個變數，使剩下的變數子集效能提昇最多。 使用這兩種方法，我們最多只需要執行目標演算法$M(M+1)/2$次，大幅縮減了計算時間，然而不保證能找到最佳的變數子集。另外，比較兩種貪婪搜尋法，一般來說向前選取的執行速度會比向後消去要快， 因為在多次執行目標演算法時使用的變數子集較小；然

而，向後消去法通常能提供較好的效能，這是因為有些變數需要和其他變數一起存在，才能使得目標演算法進步，向前選取法會忽略這一點。

## 4.3 濾波器法

濾波器法和打包法比起來輕巧簡便、通常也有不錯的效果。這種方法不用多次執行目標演算法，只需要對預測問題的訓練集作一些統計前處裡，就可以得到合適的變數子集。 濾波器法首先會定義一個評量變數的分數，利用這個分數對所有的變數排名 (ranking) ，接著根據這個排名表，從第一名的變數開始往下選取，直到符合一個特定的停止標準 (stop criterion) 才停下來。 在 [20]中提到了介紹三種類見的評量分數: 相關標準 (Correlation criterion) 單一變數分類器 (single variable classifier) 、以及交互資訊 (mutual information) ，下面將分別介紹。

以$y$代表預測問題的標記，$x^{(i)}$代表第$i$種變數。如果標記為連續的實數，可以計算$y$與$x^{(i)}$的皮氏相關係數 (pearson correlation coefficient) :

$$\mathcal{R}(i) = \frac{cov(X^{(i)}, Y)}{\sqrt{var(X^{(i)})var(Y)}} \tag{4.1}$$

其中，$cov(X^{(i)}, Y)$代表標記和第$i$種變數的共變異係數；$var(X^{(i)})$、$var(Y)$分別為標記和第$i$種變數的變異係數。皮氏相關係數介於$\pm 1$之間， 我們通常利用$\mathcal{R}^2(i)$作為評量分數，其意義在於第$i$種變數針對標記能做多好的線性預測。此類方法計算容易， 然而缺點是只適用於連續實數標記，且無法考慮變數和標記間的非線性關係。

統計機器學習領域中，有各式各樣線性或非線性的分類方法。這些分類方法，大部份都是基於多個變數。然而，將這些方法用單一變數代入，實做在發展資料集上，就可以利用預測準確率來當作變數的評量分數。 這類方法對不同的預

測問題可以選擇不同的分類器，和相關標準比較起來有比較大的自由度。另外，雖然實做特定的分類方法有一定的計算量，但是僅只有一個變數使得需要的計算資源會在合理的範圍內， 在一般離線的使用上是沒有問題的。

基於消息理論 (information theory) ，我們可以藉由兩隨機變數間的交互資訊來做為評量分數:

$$\mathcal{I}(i) = \int_{x_i} \int_{y} p(x_i, y) log \frac{p(x_i, y)}{p(x_i)p(y)} \tag{4.2}$$

這個評量分數指出了第$i$種變數與標記機率分佈的相關性，越大代表兩個分佈能傳遞的資訊越相似。使用此方法最大的問題在於$p(x_i, y)$、$p(x_i)$以及$p(y)$通常是未知， 通常我們利用統計訓練資料的方式來估計這些分佈。另外，當遇到連續變數時連統計訓練資料都會遇到困難，需要更多近似的技巧。

## 4.4 變數選取與模型調適

如同4.1節中提到，變數選取方法一個重要目的是，解決雜亂變數太多的問題。這個問題和前面提到的過度貼合問題其實是一體兩面的。過度貼合造成的問題在於訓練資料太少時， 我們無法了解每一種變數的統計特性。從變數選取的角度來看，因為變數種類的數目相對於訓練資料量來說太多，才需要選取變數，這同樣也是資料不足的問題。另外，在2.4.2節中提到的L1正則化法， 在對目標函數作最佳化的同時，會使得一些參數被設定為零，與其對應的變數不會被考慮到。這其實也包含了變數選取的概念。然而解目標函數的方法是較為間接的方式，本論文接下來的部份希望以本章介紹的方法為基礎， 以直接選取變數的方式改善模型調適方法。

# 第 五 章　基於模型變數選取方法之强健型語音辨識

從預測問題的角度來看，聲學模型的調適任務中新模型的參數相當於標記$y$， 初始模型的參數 (高斯平均向量) 相當於特徵參數向量$x$。 由於本論文專注於解決自我調適的任務，希望在僅有一句測試語料的條件下即時做調適來提升辨識率，在上一章中提到的各種變數選取方法在使用上會遇到一些限制。以打包法來說， 執行時間是最大的問題，即使使用了上一章提到的各種貪婪搜尋法，仍然會花太多時間；濾波器法雖然可以在短時間內完成，但是僅對特徵向量作前處理，對調適的幫助較有限， 這一點後面實驗會多做說明。因此，我們在本論文中提出適用於調適任務的變數選取方法。

我們提出了變數選取-最大相似度線性回歸 (Variable Selection MLLR, VSM-LLR) 的方法，其基本架構如圖5.1所示。此方法共可分為三部份：變數子集建立、子集選取以及模型調適。 第一部分在系統離線時完成，稱為離線程序；剩餘兩部分則是在系統上線時完成，稱為線上程序。這個方法能夠針對每一句測試語料都找出合適的變數子集，使辨識結果最佳化。

## 5.1　離線程序

變數選取-最大相似度線性回歸在系統離線時需準備好一些候選的變數子集。與一般打包法不同，我們利用一些事前知識，只建立出少數資訊豐富的變數子集。 如此有助於後續的變數選取步驟能夠在可接受的時間內完成。後續實驗也證明列舉這些少量的變數子集就足以幫助辨識效能的提昇。接下來兩小節會介紹兩種建立

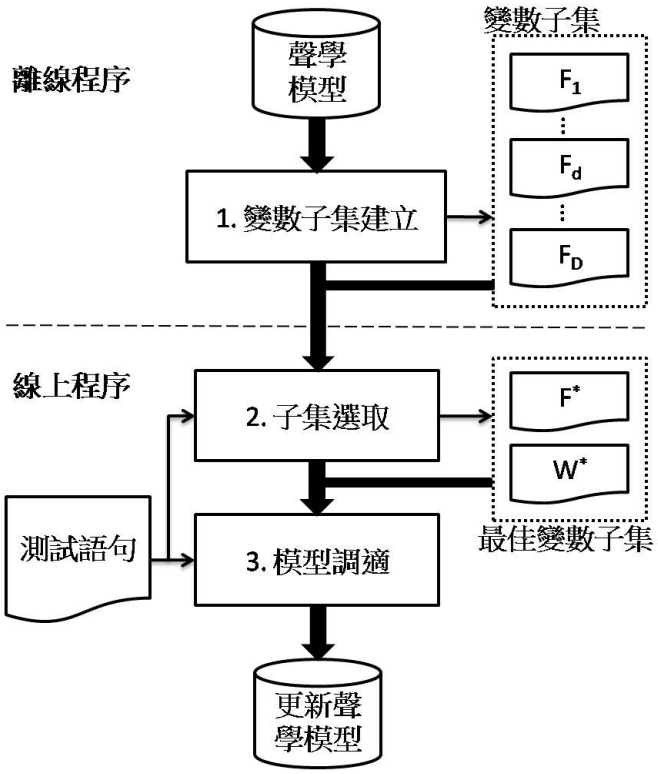



圖 5.1: 變數選取-最大相似度線性回歸 基本架構圖

變數子集的方法： 主成分分析 (principal component analysis, PCA) 以及窗型變數集。

### 5.1.1 變數子集建立-主成份分析

在模型調適任務中，原本使用的變數是初始聲學模型的高斯平均向量。為了使變數資訊更加集中，我們將每個高斯平均向量投影到平均向量集的主成分上 [22]。假設Γ為初始聲學模型的高斯平均向量所組成，為$D \times I$矩陣，其中$D$為高斯平均向量的維度，同時也是聲學特徵向量的維度，I為聲學模型的高斯分布數目。將Γ的共變異矩陣$\Gamma\Gamma^T$做特性分析 (eigen analysis) ，即可得到由平均向量集主成分組成的空間基底：

$$X = [e^{(1)}, e^{(2)}, ..., e^{(d)}, ..., e^{(D)}]^T \tag{5.1}$$

其中$e^{(1)}, e^{(2)}..., e^{(d)}, ..., e^{(D)}$為共變異矩陣$\Gamma\Gamma^T$的特性向量 (eigenvector)，按照特性值 (eigenvalue) 由大到小排列。 將高斯平均向量$\mu_i$投影到空間基底$X$上，我們可以得到新的變數向量$\epsilon_i$：

$$\epsilon_i = X\mu_i \tag{5.2}$$

利用此轉換式，我們按照空間基底特性值的大小將新變數選進子集，建構出$D$個變數子集:$F^{(d)}, d = 1, 2..., D$，其中 $F^{(1)} = \{f^{(1)}\}, F^{(2)} = \{f^{(1)}, f^{(2)}\}, ..., F^{(D)} = \{f^{(1)}, f^{(2)}, ..., f^{(D)}\}$，而變數$f^{(d)}$是由原本的高斯平均向量投影到$e^{(d)}$上獲得。 使用$F^{(d)}$變數子集，第$i$個高斯分佈組成的變數向量可被寫成：

$$\epsilon_i^{(d)} = X^{(d)}\mu_i^{(d)} \tag{5.3}$$

其中$X^{(d)} = [e^{(1)}, e^{(2)}, ..., e^{(d)}]^T$。

主成份分析有一個重要的特性，是能夠將資料分佈大部分的變異度保留在前幾個特性向量所構成的子空間中。 因此，藉由此方法我們可以將原本變數的重複性 (redundency) 消除，建構出好的變數子集。 雖然在此方法中，我們僅列舉$D$個變數子集，但是每一個變數子集和其他有相同變數數目的子集比較起來，保留了較多資訊。

### 5.1.2 變數子集建立-窗型變數集

窗型變數集的建構是基於以下的假設：在使用梅爾倒頻譜係數作為聲學特徵向量時，高斯平均向量的每一維度變數與自己以及梅爾倒頻譜上相鄰近的變數相關。舉例來說，預測更新過後的高斯平均向量第五維，我們應優先使用初始高斯平均向量同樣第五維作為變數。除此之外高斯平均向量的第四、第六維也應該納入考慮，因為在抽取聲學特徵的過程中， 第四、五、六維梅爾倒頻譜特徵對應到相近的聲音特性。

根據上述的假設，我們將聲學模型高斯平均向量的調適任務視為$D$個預測問題。對於每個預測問題，分別建立數個變數子集供後續選擇。建立變數子集的過程，首先我們將所有的變數$\mu^{(1)}, \mu^{(2)}, ..., \mu^{(D)}$，按照上述的假設排序，再按此順序將變數依序選入子集。以高斯平均向量的第五維預測問題來說，變數排列順序為$\mu^{(5)}, \mu^{(4)}, \mu^{(6)}, \mu^{(3)}, \mu^{(7)}...$，建構出的變數子集包括$\{\mu^{(5)}\}, \{\mu^{(4)}, \mu^{(5)}\}, \{\mu^{(4)}, \mu^{(5)}, \mu^{(6)}\}...$。考慮到39維梅爾倒頻譜聲學特徵可分成原係數、一階導數及二階導數三部份，針對高斯平均向量每維的預測問題，我們只會排序同性質的變數，建構13個變數子集。另外，當考慮$D$個預測問題時，理論上可以各自選擇不同變數數目的子集。但在實做上為了進一步簡化計算，我們限制$D$個預測問題必須選取同樣大小的變數子集。這使得變數子集的選取像是在頻率軸上取一個固定大小的窗，因此稱這套變數子集的建立為窗型變數集。

在這兩小節中提到的變數子集建立方法，都將原有變數做了排序，這正是濾波器法的特徵。但與一般濾波器法不同的是，在此步驟結束後，真正在做模型調適時會使用的子集仍未決定，後面介紹的線上程序會根據每一句測試語料，分別選出適合的子集。

## 5.2 線上程序

系統的線上程序負責為每一句測試語料選出適合的子集，並完成模型調適以供第二階段解碼使用。由於子集選取及模型調適的理論與實做方法類似，在此節中一併介紹。變數子集的選取，我們採取和一般打包法相同的策略：給定一目標演算法，評估基於各個變數子集該演算法的表現。針對模型調適任務，我們選擇仿射變換調適法作為目標演算法。然而，與一般打包法不同，在自我調適的情境下我們沒有發展資料集可供評估各變數子集的表現。因此，本論文使用2.3節提到的各

種正則化訓練準則 (regularization criterion) 作為變數子集的評量標準。

### 5.2.1 子集選取及模型調適-主成分分析

基於主成份分析找出的變數子集，我們可以用下式來實做仿射變換調適法：

$$\mu_i' - \mu_i = A^{(d)}\epsilon_i^{(d)} + b^{(d)} \tag{5.4}$$

其中$A^{(d)}, b^{(d)}$分別為$D \times d$矩陣及$d$維向量。與式2.4相比較可發現，式5.4以新舊聲學模型高斯平均向量的差$\mu_i' - \mu_i$作為轉換的標記， 而非新模型的平均向量$\mu_i$。這是因為，基於主成分分析找出的變數向量維度較高斯平均向量低 ($d < D$) ，雖然式5.4將變數投影至高維， 投影過後的標記仍然被限制在$d$維的子空間中，這會使得標記的鑑別度 (discrimination) 喪失。舉例來說當$d = 2$，如果使用新模型的平均向量$\mu_i'$作為標記， 因為旋轉矩陣$A^{(d)}$只有兩行，$A^{(d)}\epsilon_i^{(d)} + b^{(d)}$被限制在$D$維空間中的一個二維平面上，使得$\mu_i'$的維度只有2， 明顯不可能表達原本散布在39維空間中的高斯平均向量結構。如改以$\mu_i' - \mu_i$作為標記，平均向量間的結構及鑑別度仍會被保留。

基於式5.4，我們利用最大相似度訓練準則來求出最佳的$A^{(d)}, b^{(d)}$。首先將式5.4改寫為：

$$\mu_i' - \mu_i = W^{(d)}\eta_i^{(d)} \tag{5.5}$$

其中$W^{(d)} = [A^{(d)}b^{(d)}]$，$\eta_i = [\epsilon_i^T 1]^T$。將式5.5中的$\mu_i'$移項後代入式2.3之後，可得到最大相似度作為準則的目標函數。最佳化此目標函數，$W^{(d)}$的每一列$W_j^{(d)}$有封閉

形式解：

$$W_j^{(d)} = k_j^{(d)}(G_j^{(d)})^{-1}$$
$$G_j^{(d)} = \sum_i \sum_t \frac{\gamma_i(t)}{\sigma_{ij}} \eta_i^{(d)}(\eta_i^{(d)})^T \tag{5.6}$$
$$k_j^{(d)} = \sum_i \sum_t \frac{\gamma_i(t)}{\sigma_{ij}}(o_{tj} - \mu_{ij})(\eta_i^{(d)})^T$$

其中$\gamma_i(t)$代表第$i$個高斯分布組成在時間t的占據機率；$\sigma_{ij}$、$o_{tj}$分別為共變異矩陣的第 $(j,j)$ 項元素及在時間t音框的第$j$維觀測特徵向量。 利用式5.6，我們可以得到基於$D$個不同變數子集的仿射變換矩陣。另外，觀察可發現$G_j^{(d)}$、$k_j^{(d)}$的每個元素都在$G_j^{(D)}$、$k_j^{(D)}$中出現。 以$(G_j^{(d)})_{lm}$代表$G_j^{(d)}$的第$(l,m)$個元素，$(k_j^{(d)})_l$代表$k_j^{(d)}$的第$l$個元素， 當$l < d+1, m < d+1$，$(G_j^{(d)})_{lm} = (G_j^{(D)})_{lm}$，$(k_j^{(d)})_l$=$(k_j^{(D)})_l$；當$l < d+1, m = d+1$，$(G_j^{(d)})_{lm} = (G_j^{(D)})_{l(D+1)}$； 當$l = d+1, m < d+1$，$(G_j^{(d)})_{lm} = (G_j^{(D)})_{(D+1)m}$，$(k_j^{(d)})_l$=$(k_j^{(D)})_{D+1}$；最後$(G_j^{(d)})_{(d+1)(d+1)} = (G_j^{(D)})_{(D+1)(D+1)}$。 因此，我們只需要統計出$G_j^{(D)}$、$k_j^{(D)}$，就可以求得每個變數子集所對應到的仿射變換矩陣。這個特性使得變數選取-最大相似度線性回歸具備及時完成的條件。

求得$D$個仿射變換矩陣後，下一步是要選出其中最佳的矩陣及其對應的變數子集。本論文使用2.3節提到的各種正則化訓練準則 (regularization criterion) 作為變數子集的評量標準。 首先，計算個變數子集的評量分數:

$$S^{(d)} = logP(O|\Omega, W^{(d)}) - \lambda||A^{(d)}||^2 \tag{5.7}$$

等式右邊與L2正則化仿射變換的目標函數幾乎相同，然而在這邊我們並非要求得能夠最佳化式5.7的變換矩陣，而是先以最大相似度的目標函數找到$D$個候選矩陣， 再以式5.7做評估。評量分數最大的變數子集會被選擇，利用其對應個仿射變換矩陣，以式5.5調適高斯平均向量，即可獲得新的聲學模型。

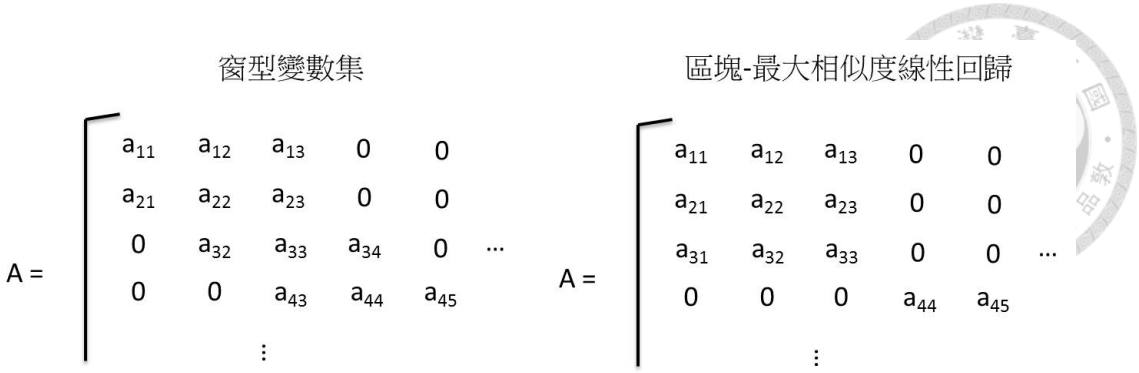



圖 5.2: 窗型變數集/區塊最大相似度線性回歸 旋轉矩陣比較圖

### 5.2.2 子集選取及模型調適-窗型變數集

列舉出窗型變數集後，我們可以利用下式來實做仿射變換調適法：

$$\mu_i' = A^{(d)}\mu_i + b^{(d)} \tag{5.8}$$

其中上標$d$代表各預測問題都採用元素數目為$d$的變數子集。$A^{(d)}$、$b^{(d)}$為$D \times D$的旋轉矩陣以及$D$維偏差向量。變數子集的選擇會反映在旋轉矩陣$A^{(d)}$上。 $A^{(d)}$的每一列，只會在對應到已選變數的元素被設為非零。根據最大相似度訓練準則，我們可以利用與式2.8相似的方法去求得$A^{(d)}$、$b^{(d)}$。

基於窗型變數集所求得的旋轉矩陣$A^{(d)}$，與方塊最大相似度線性回歸 (block MLLR) 很相似。兩者的相異之處可以從圖5.2看出，以$d = 3$為例， 方塊最大相似度線性回歸的旋轉矩陣，非零元素可組成$3 \times 3$的區塊；而基於窗型變數集的旋轉矩陣，除了$13 \times 13$區塊的最前最後列， 每一列的非零元素都以對角線元素為中心。這兩個方法有相同的假設:高斯平均向量的每一維度變數與自己以及頻譜上相鄰近的變數相關。理論上兩方法會有相似的效果， 在這邊選擇窗型變數集，是因為列舉子集比較直觀且方便。

與主成分分析相同，求得$D$個仿射變換矩陣後，正則化訓練準則被用來當作變數子集的評量標準。同樣以5.7計算評量分數後，選取最高分的變數子集和對應

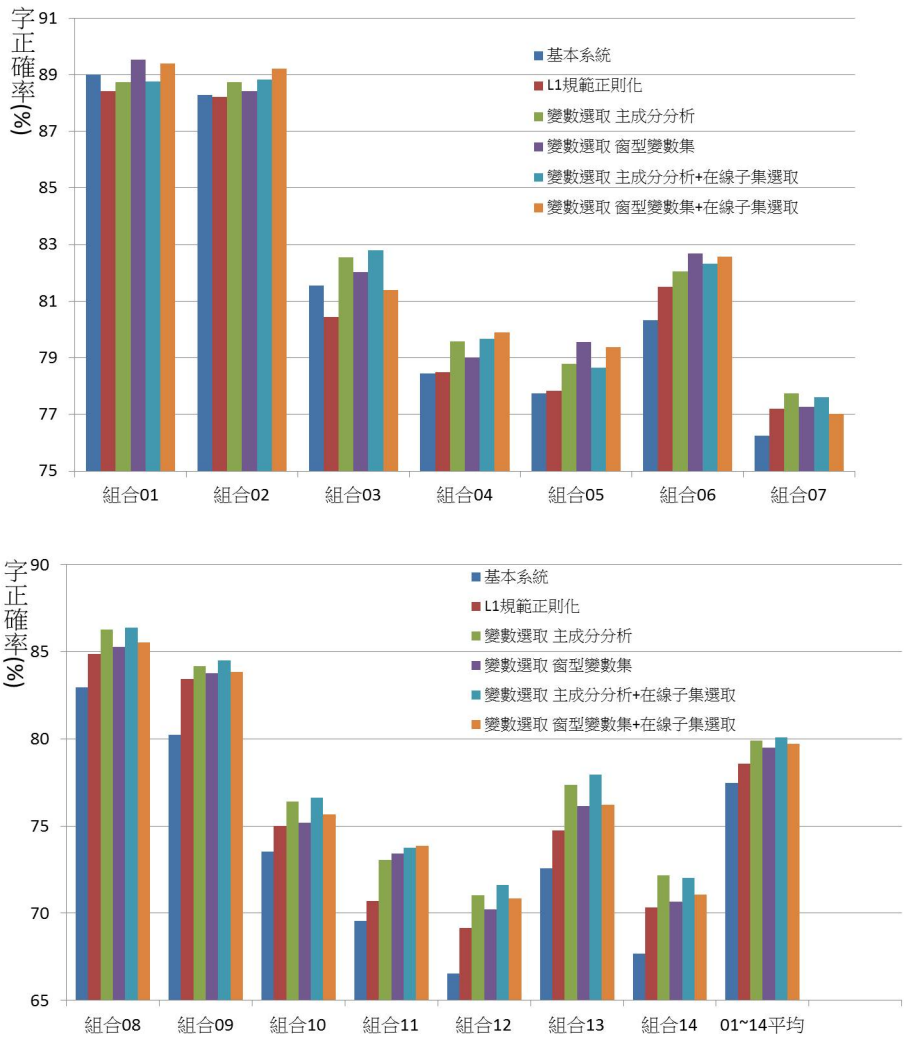
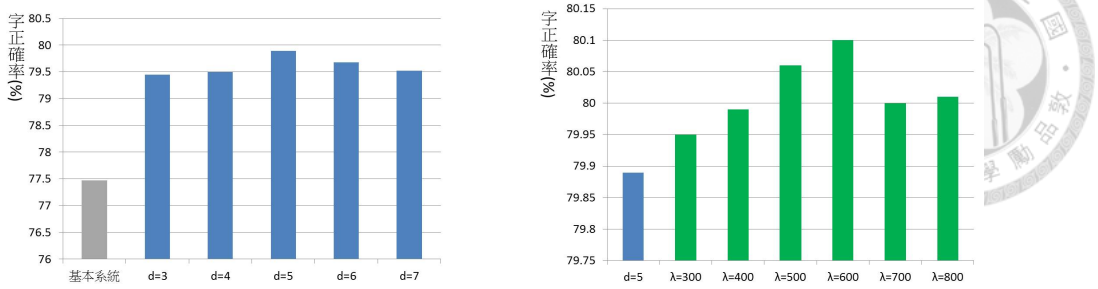



圖 5.3: Aurora-4 變數選取-最大相似度線性回歸 實驗結果

的仿射轉換矩陣，實做模型調適。

## 5.3 Aurora-4實驗結果

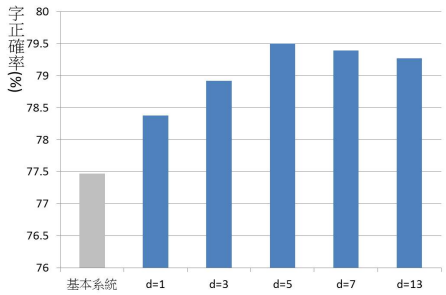
本節列舉了變數選取-最大相似度線性回歸實做在Aurora-4上得到的實驗結果。由於變數選取-最大相似度線性回歸包括離線及線上程序，我們可以在離線變數子集建立的步驟完成後， 手動選擇最佳變數子集 (意即每一測試語句所採納的子集相同，相當於使用濾波器法) ，並同樣實做在Aurora-4上。



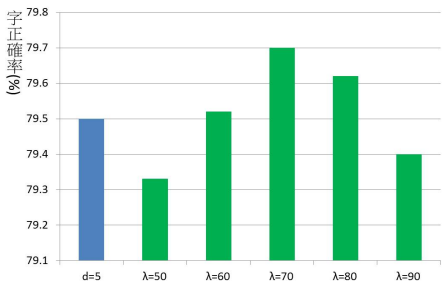

(a) 主成分分析




(b) 主成分分析-線上子集選取




(c) 窗型變數集




(d) 窗型變數集-線上子集選取

圖 5.4: 變數選取-最大相似度線性回歸 不同超參數設定的實驗結果

圖5.3列出了變數選取-最大相似度線性回歸在Aurora-4所有測試集上的辨識結果，包含兩種變數子集建立方法、只做離線程序和完整做完離線線上程序。基本系統以及基於模型的L1正則化法也同樣被列舉出來作比較。 此圖中除了基本系統之外，所有方法都是列出在所有測試集平均上表現最好的超參數 (hyper-parameter) 。由數據可看出所有變數選取方法，在大多數的測試集中都勝過基本系統以及基於模型的L1正則化法。另外， 在使用同樣變數子集建立方法的條件下，本論文提出以正則化訓練準則評量變數子集的策略 (即線上子集選取) 可以獲得更好的效能。最後，使用主成份分析作為變數子集建立法，其效能又勝過窗型變數集。

## 5.4 本章結論

本章介紹了如何將變數選取方法運用在基於模型的仿射變換調適法上。為了配合自我調適的情境，本論文改進了一般方法，設計出變數選取-最大相似度線性回歸，使得系統能夠線上為 每句測試語料分別決定最適合的變數子集。我們使用兩種變數子集列舉的方法: 主成分分析以及窗型變數集，並在系統線上時使用正則化訓練準則做為評量子集的方法。

實驗結果顯示以下三點:

1. 將變數選取方法應用於基於模型的仿射變換模型調適法，可以克服過度貼合問題，可使辨識系統效能進步。

2. 本論文提出的線上變數選取方法，用於主成分分析法所列舉的變數子集，可獲得令人滿意的效果，而對窗型變數集來說則效果不明顯。

3. 同樣考慮正則化訓練準則，一般正則化法是直接將其作為目標函數對參數最佳化，本論文則是用此函數做為評量分數。實驗結果證明，後者可以使辨識率進步更多。

# 第 六 章　基於聲學特徵變數選取方法之強健型語音辨識

本章介紹如何將變數選取方法應用在基於聲學特徵的仿射變換調適法上。如將基於聲學特徵的仿射變換調適法考慮為預測問題，初始以及更新過的聲學特徵向量分別為參數向量$x$以及標記$y$。同樣為了配合自我調適的情境，我們採用和上一章相似的系統架構，如圖6.1所示，首先建構變數子集，接著為每一句測試語料找出適合的變數子集並完成模型調適。然而兩章方法間最大的不同在於，此章的方法中所有步驟，包括變數子集建立都是在系統上線時完成，以及圖6.1中的模型調適是作用在聲學特徵向量上。我們將其本章介紹的方法稱為變數選取-特徵最大相似度線性回歸 (Variable Selection feature-based MLLR, VSfMLLR) 。

## 6.1　變數子集-建立與選取

本章所提出的變數選取-特徵最大相似度線性回歸，同樣可以選擇主成分分析、窗型變數集兩種子集建立方法。變數子集的建立完成後，我們再一次使用正則化訓練準則，作為變數子集作用於目標演算法 (基於特徵的仿射變換調適法) 的評量分數。各步驟的詳細做法詳述如以下小節。

### 6.1.1　變數子集建立-主成份分析

一般基於特徵的仿射變換調適法，所採用的變數是梅爾倒頻譜特徵向量。對每一句測試語料，我們將這些特徵向量投影到主成分基底構成的空間中，形成新的變數向量。令$O$為某一句測試語料所有梅爾倒頻譜特徵向量構成的矩陣，維度

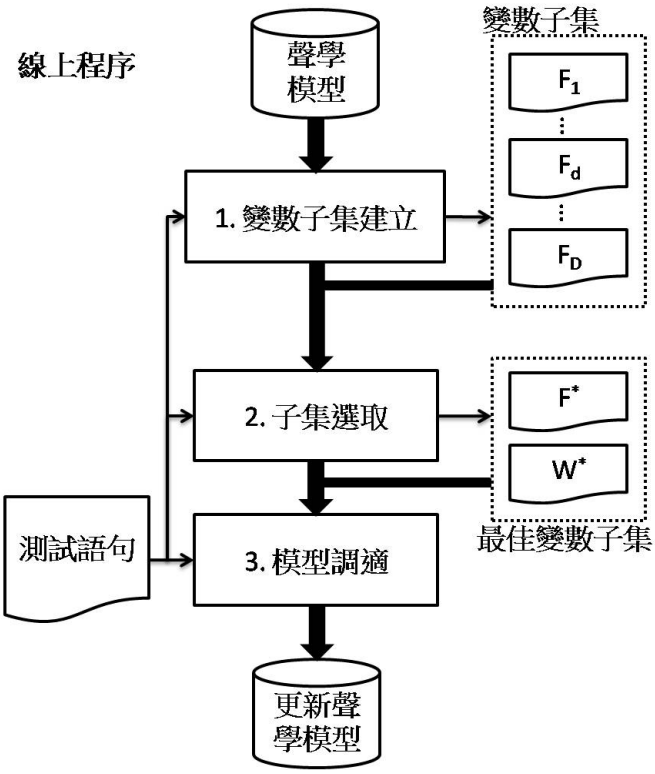



圖 6.1: 變數選取-特徵最大相似度線性回歸 基本架構圖

為$D \times T$。其中$D$代表特徵向量的維度 (通常為39) ，$T$代表這句測試語料的音框總數。 將$O$的共變異矩陣$OO^T$做特性分析 (eigen analysis) ，即可得到主成分組成的空間基底：

$$X = [e^{(1)}, e^{(2)}, ..., e^{(d)}, ..., e^{(D)}]^T \tag{6.1}$$

其中$e^{(1)}, e^{(2)}..., e^{(d)}, ..., e^{(D)}$為共變異矩陣$OO^T$的特性向量 (eigenvector) ，按照特性值 (eigenvalue) 由大到小排列。 將聲學特徵向量投影到空間基底$X$上，得到新的變數向量，如下式所示：

$$\epsilon_t = Xo_t \tag{6.2}$$

接著，按照空間基底特性值的大小將新變數選進子集，建構出$D$個變數子集:$F^{(d)}, d = 1, 2..., D$，其中$F^{(1)} = \{f^{(1)}\}, F^{(2)} = \{f^{(1)}, f^{(2)}\}, ..., F^{(D)} = \{f^{(1)}, f^{(2)}, ..., f^{(D)}\}$，而變數$f^{(d)}$是由原本的聲學特徵向量投影到$e^{(d)}$上獲得。

使用$F^{(d)}$變數子集，第$t$個音框的變數向量可被寫成：

$$\epsilon_t^{(d)} = X^{(d)} o_t^{(d)} \tag{6.3}$$

### 6.1.2 變數子集建立-窗型變數集

在聲學特徵向量中建立窗型變數集，其基本原則與5.1.2節敘述的相同。只要將5.1.2節中的高斯平均向量變數$\mu^{(1)}, \mu^{(2)}, ..., \mu^{(D)}$代換為聲學特徵向量的各維度$o^{(1)}, o^{(2)}, ..., o^{(D)}$，即可以相同方法得到聲學特徵的窗型變數集。

### 6.1.3 子集選取與模型調適

變數子集建立完成後，我們需要執行目標演算法才能完成子集選取。這邊以限制型最大相似度線性回歸作為目標演算法。這種演算法沒有直接公式解，實作到各種變數子集上較困難。

將主成份分析找出的變數子集，應用於基於特徵的仿射變換調適法，可利用下式：

$$o_t' - o_t = A^{(d)} \epsilon_t^{(d)} + b^{(d)} \tag{6.4}$$

在此仍須注意，仿射變換的目標為$o_t' - o_t$初始及更新聲學特徵向量的差。針對各變數子集，我們基於最大相似度訓練準則找出最佳的仿射轉換矩陣。與之前提到各種實作在聲學特徵向量上的仿射轉換法相同，這個問題並不存在直接公式解。令$W^{(d)} = [A^{(d)}\ b^{(d)}]$，根據附錄A介紹的低維度最佳化法，可以利用以下的更新式，以迭代法求得最佳解：

$$w_j^{(d)} = (k_j^{(d)} + \alpha p_j Q^T)(G_j^{(d)})^{-1} \tag{6.5}$$

其中

$$G_j^{(d)} = \sum_{i,t} \frac{\gamma_i(t)}{\sigma_{ij}} \eta_t^{(d)} (\eta_t^{(d)})^T$$
$$k_j^{(d)} = \sum_{i,t} \frac{\gamma_i(t)}{\sigma_{ij}} (\mu_{ij} - o_t j)(\eta_t^{(d)})^T \quad (6.6)$$
$$Q = [(X^{(d)})^T \ \vec{0}]^T$$

$\vec{0}$為$D$維全零向量、$w_j^{(d)}$是$W^{(d)}$的第$j$列、$\alpha$也為一待估參數。低維度最佳化法的詳細步驟將在附錄中介紹。 完成各變數子集的仿射轉換矩陣計算後，我們再次以正則化訓練準則當作評量分數：

$$S^{(d)} = logP(O|\Omega, W^{(d)}) - log|A| - \lambda||A^{(d)}||^2 \quad (6.7)$$

以窗型變數集作為候選的子集，可使用下式調適聲學特徵向量：

$$o_t' = A^{(d)} o_t + b^{(d)} \quad (6.8)$$

與基於模型的方法相同，變數子集的選擇會反映在旋轉矩陣$A^{(d)}$上。 $A^{(d)}$的每一列，只會在對應到已選變數的元素被設為非零，其非零元素分佈可參考圖5.2。根據最大相似度訓練準則，我們可以利用與式2.11相似的方法去求得$A^{(d)}$、$b^{(d)}$。 最後，同樣利用正則化訓練準則求得各窗型變數集的評量分數，為每句測試語料選出最佳變數子集和對應的仿射轉換矩陣。

## 6.2 Aurora-4 實驗結果

本節列舉了變數選取-特徵最大相似度線性回歸實做在Aurora-4上得到的實驗結果，其中包括手動選擇固定的變數子集，以及線上選取子集的結果。

首先圖6.2列出了變數選取-特徵最大相似度線性回歸在Aurora-4所有測試集上的辨識結果，包含兩種變數子集建立方法、手動及自動選擇最佳變數子集的結

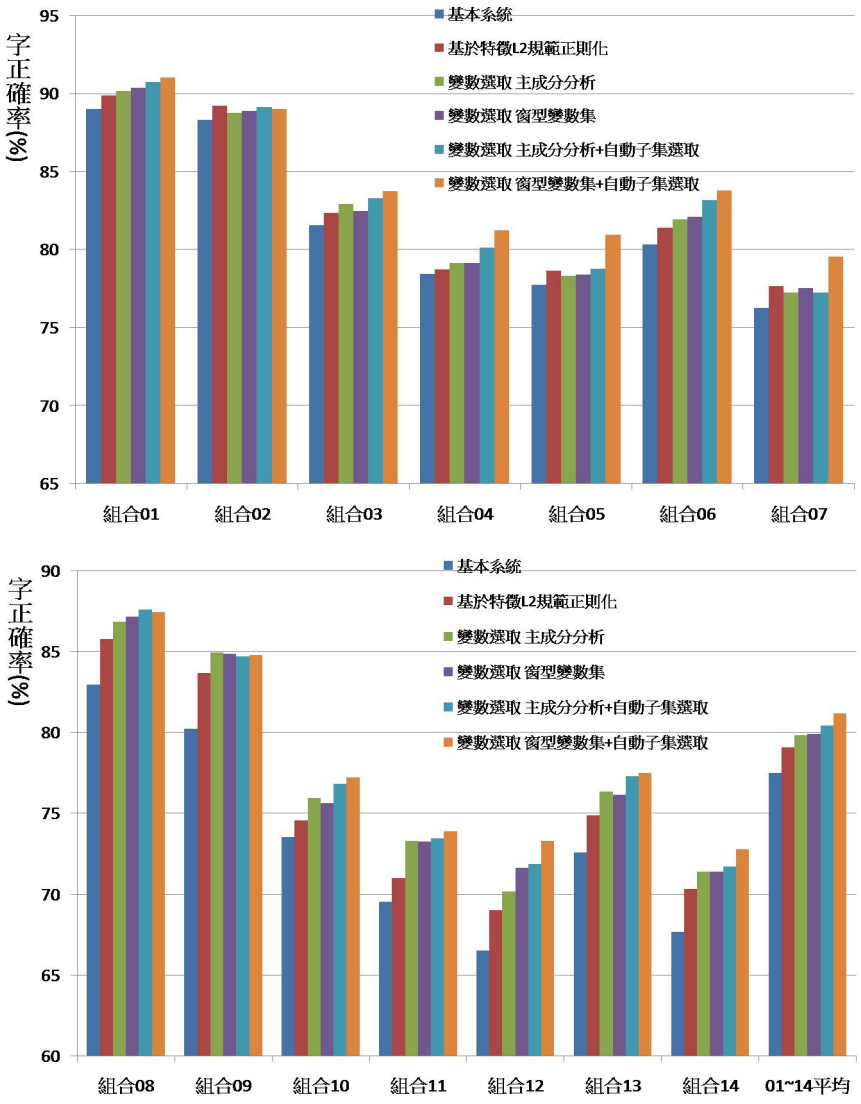
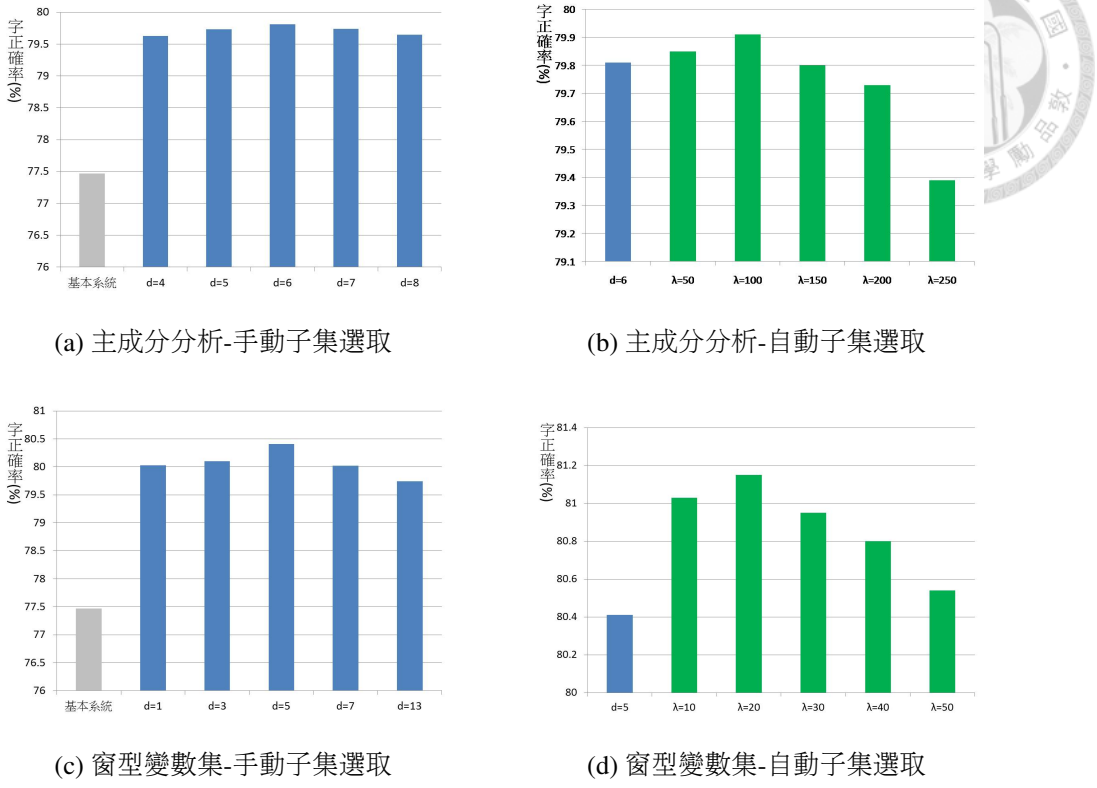



圖 6.2: Aurora-4 變數選取-特徵最大相似度線性回歸 實驗結果




(a) 主成分分析-手動子集選取

(b) 主成分分析-自動子集選取

(c) 窗型變數集-手動子集選取

(d) 窗型變數集-自動子集選取

圖 6.3: 變數選取-特徵最大相似度線性回歸 不同超參數設定的實驗結果

果。基本系統以及基於聲學特徵向量的L2正則化法也同樣被列舉出來作比較。 此圖中除了基本系統之外，所有方法都是列出在所有測試集平均上表現最好的超參數。由數據可看出，兩種變數子集建立的方法都能夠得到比基本系統和一般正則化方法更好的辨識結果。其中使用窗型變數集在大部分的大部分的大部分的測試集上的效果都更勝主成份分析，這與前一章得到的結果相反。最後，以正則化訓練準則自動選變數子集，所得到的辨識率都勝過手動選擇。

圖6.3a、6.3b、6.3c、6.3d分別畫出兩種變數子集選取策略，搭配手動選擇或是自動選取變數子集， 調整不同超參數在Aurora-4測試集上的平均辨識率。從圖6.3a、6.3c可看出，在採用變數選取方法後，即可克服過度貼合問題，使辨識效能勝過基本系統。在此必須提及，圖6.3c中$d = 1$、$d = 13$相當於一般常見的對角特徵最大相似度線性回歸 (diagnonal fMLLR) ，即將旋轉矩陣限制為對角矩陣、以

及13 13 13區塊特徵最大相似度線性回歸。也就是說對於自我調適情境， 簡單的控制仿射轉換矩陣複雜度 (選擇元素少的變數子集) ，比一般的正則化方法更有效。最後觀察圖6.3c、6.3d，本論文提出的在子集選取方法應用在聲學特徵向量，於窗型變數集上可取得很大的進步，在主成分分析上則進步的不明顯。

表 6.1: Aurora-4 所有方法 實驗結果

| 調適方法 | baseline | MLLR | CMLLR | L2-norm |
|---|---|---|---|---|
| 字正確率(%) | 77.47 | 77.24 | 77.38 | 78.24 |
| 調適方法 | L1-norm | MAPLR | feature L2-norm | |
| 字正確率(%) | 78.57 | 78.53 | 79.07 | |
| 調適方法 | VS-MLLR PCA-fix | VS-MLLR PCA | VS-MLLR slide-fix | VS-MLLR slide |
| 字正確率(%) | 79.89 | 80.10 | 79.50 | 79.71 |
| 調適方法 | VS-fMLLR PCA-fix | VS-fMLLR PCA | VS-fMLLR slide-fix | VS-fMLLR slide |
| 字正確率(%) | 79.81 | 79.91 | 80.41 | 81.15 |

最後，表6.1列出本論文實做所有方法，在Aurora-4測試集上的平均辨識率，包括了：基本系統 (baseline) 、最大相似度線性回歸 (MLLR) 、限制型最大相似度線性回歸 (CMLLR) 、 基於模型L2規範正則化 (L2-norm) 、基於模型L1規範正則化 (L1-norm) 、最大事後機率線性回歸 (MAPLR) 以及基於特徵向量L2規範正則化 (feature L2-norm) ，還有本論文所提出的各種變數選取方法： 基於模型參數的主成份分析-手動 (VS-MLLR PCA-fix) 、自動 (VS-MLLR PCA) 選取子集、窗型變數集-手動 (VS-MLLR slide-fix) 、自動 (VS-MLLR slide) 選取子集； 基於聲學特徵的主成份分析-手動 (VS-fMLLR PCA-fix) 、自動 (VS-fMLLR PCA) 選取子集、窗

型變數集-手動 (VS-fMLLR slide-fix) 、自動 (VS-fMLLR slide) 選取子集等等。

## 6.3 本章結論

本章介紹了如何將變數選取方法運用在基於聲學特徵向量的仿射變換調適法上。將上一章提出的兩種變數子集建立方法應用在聲學特徵向量，本章提出了變數選取-特徵最大相似度線性回歸。此方法也能夠讓系統線上為 每句測試語料分別決定最適合的變數子集，使調適的結果最佳化。

實驗結果顯示以下三點:

1. 在Aurora-4測試集上，我們發現基於特徵向量的仿射轉換法比基於模型參數能使辨識率進步更多。對於一般正則化方法以及變數選取方法，都有相同的趨勢。

2. 本章提出的變數選取-特徵最大相似度線性回歸，即使只實做第一階段的變數子集建立，也能使辨識效能勝過基於特徵向量的正則化方法。這顯示在自我調適任務中，相較於完全相信正則化訓練準則與系統效能成正相關，不如專注於控制仿射轉換矩陣的複雜度。

3. 本論文提出的線上變數選取方法，以正則化訓練準則作為評量分數，在基於特徵向量的仿射轉換法上也獲得成功。在Aurora-4測試集上比較所有方法，實做變數選取-特徵最大相似度線性回歸、使用窗型變數集、並完成線上變數選取可獲得最佳的辨識結果。

# 第 七 章　結論與展望

## 7.1　結論

近年來，語音辨識系統的效能逐步提升，使這項技術越來越被廣泛應用在日常生活中。然而，如何解決因為使用者或是錄音環境不同造成的聲學條件不匹配，仍然是一個重要的議題。 本論文研究的主要方向，是以聲學模型調適的方法建立強健的語音辨識系統，並針對自我調適情境深入探討。

由第二章的敘述、討論以及第三章的實驗結果，我們可以知道自我調適任務非常困難，因為完全沒有任何額外資訊可以使用。若是直接將最常見的仿射變換調適法，包括一般、限制型最大相似度線性回歸， 實做在自我調適任務上，會發生過度貼合問題，使得系統辨識率反而下降。為了嘗試解決這個問題，我們採用了統計機器學習領域中常用的正則化方法，來使得待估參數變得更穩定。實驗結果證明， 正則化方法在自我調適任務上也發揮了作用，克服過度貼合問題並改善系統效能。

本論文於第四章開始從不同的角度切入來解決過度貼合問題，並介紹了變數選取方法。變數選取主要的考量是，在訓練資料不足的條件下雜亂的變數會影響統計機器學習方法的效能。這個問題基本上和過度貼合類似， 因此我們將這個方法應用在聲學模型的調適任務上。然而，將一般變數選取方法直接實做於自我調適情境並不容易。主要原因是，我們通常希望應用在自我調適情境的方法能夠在系統上線時即時完成。一般最常見的打包法， 即使使用了各種貪婪演算法，要在系統上線時完成也有困難。因此本論文在第五、六章分別提出了變數選取-最大相似度線性回歸以及變數選取-特徵最大相似度線性回歸。這兩種方法的共同點在於， 先根據一些背景知識建立少量的變數子集，在利用正則化訓練準則當作評估

子集的方法。如此便能為每一句測試語料分別找到最適合的子集，最佳化調適效果。兩種方法相異之處則是， 前者屬於作用在聲學模型參數的仿射轉換法，而後者則是基於聲學特徵向量的仿射轉換法。兩種方法分別與基於模型參數、基於特徵向量的正則化方法比較，都能夠取得辨識率上的進步。

相對於一般的正則化方法，本論文提出的方法同樣使用了正則化訓練準則。兩類方法最大的不同在於，前者是將正則化訓練準則當作目標函數，直接將根據此函數的最佳化找出最適合的待估參數 (仿射轉換矩陣) ； 而後者則是將正則化訓練準則當作評估變數子集的標準，在計算每個變數子集所對應的待估參數時，仍是根據一般的最大相似度準則。本論文最重要的論點在於，證明在自我調適情境下， 將正則化訓練準則當作變數選取的評估標準更能有效的解決過度貼合問題，增加辨識率。

## 7.2 展望

語音辨識在現代科技中佔有舉足輕重的地位。這是因為語音辨識與各種語音應用高度相關。語音辨識率的進步將使這些應用的效能跟著提昇，或是將之前不可行的應用變為可行。因此，系統辨識率的提昇以及强健化， 在今後仍會是非常重要的研究方向。

我們提出的變數選取方法，可以與各種前端聲學特徵正規化疊加，進一步提昇辨識效能。另外，本論文嘗試以不一樣的方式使用正則化訓練準則，將其當作評估變數子集的標準，同樣的方法也能應用在鑑別式訓練準則 (discriminative training criterion) 上。一般的鑑別式訓練，在碰到訓練資料量少的調適問題時，很難能發揮效果。使用變數選取方法或許能改善這個問題。本論文提出的概念及方法，還有很大的發展空間。

另外，單純考慮到辨識率的提昇，近年來深度學習 (deep learning) 的技術快速發展，其中深度類神經網路 (deep neural network, DNN) [23]在語音辨識上獲得了巨大的成功。在這塊新興領域中還有許多不確定性， 如超參數的決定、適當的模型調適法等等，都是今後很好的發展方向。

# 附錄 A　基於聲學特徵的仿射轉換最佳化法

此附錄章節詳細介紹如何利用最大相似度訓練準則，求出用於聲學特徵向量的仿射變換矩陣。

## A.1　一般最佳化法

本節的內容主要來自參考資料 [13]，其目的在於如何以迭代法解出限制型最大相似度線性回歸的仿射變換矩陣。 2.3.2節提到的式2.10，限制型最大相似度線性回歸相當於對聲學特徵向量實做仿射變換，如下式所示：

$$o_t' = Ao_t + b = W\zeta_t \tag{A.1}$$

其中$o_t$、$o_t'$分別為音框$t$更新前、後的聲學特徵向量，$W = [A\ b]$，$\zeta_t = [o_t^T\ 1]^T$。 如根據最大相似度訓練準則來定義最佳的仿射變換矩陣，其目標函數即為:

$$F(W) = -\sum_i \sum_t \gamma_i(t)[(o_t' - \mu_i)^T(\Sigma_i)^{-1}(o_t' - \mu_i) + log|A|^2] + C \tag{A.2}$$

其中$i$為高斯分佈組成的標記。將一些較複雜的符號經過整理，代換成以下變數:

$$\begin{aligned} G_j &= \sum_i \sum_t \frac{\gamma_i(t)}{\sigma_{ij}} \zeta_t \zeta_t^T \\ k_j &= \sum_i \sum_t \frac{\gamma_i(t)}{\sigma_{ij}} \mu_{ij} \zeta_t^T \\ \beta &= \sum_i \sum_t \gamma_i(t) \end{aligned} \tag{A.3}$$

並將A.1式也代入，我們可將目標函數化簡如下：

$$F(W) = \beta log(p_j w_j^T) - \frac{1}{2}\sum_{j=1}^{D}(w_j G_j - 2w_j k_j^T) \tag{A.4}$$

其中$w_j$為$W$的第$j$列，$p_j = [c_{j1}\ c_{j2}...c_{jD}\ 0], c_{jk} = cofactor(A_{jk})$。對照目標函數的兩個形式可看出，$log|A| = log(p_j w_j^T)$，其中$j$可以代入$1, 2, ..., D$，(D為聲學特徵向量維度)。接著將A.4式對$W$的第$j$列$w_j$偏微分可得：

$$\frac{\partial F(W)}{\partial w_j} = \beta \frac{p_j}{p_j w_j^T} - w_j G_j w_j^T + k_j \tag{A.5}$$

令此式等於零，就是求得最佳$w_j$需要解的式子，整理過後可得：

$$p_j w_j^T k_j G_j^{-1} + \beta p_j G_j^{-1} = p_j w_j^T w_j \tag{A.6}$$

因為$p_j$與$w_j$相關，所以無法直接解出$w_j$。因此我們將解開此式的程序分為兩步：1.已知$p_j$的條件下解出$w_j$；2.已知$w_j$的條件下更新$p_j$ 對每個維度$j = 1, 2, ..., D$輪流執行這兩步驟，多次迭代之後，$W$會收斂到符合最大相似度的仿射轉換矩陣。 第二步十分容易，下面詳細介紹第一步的作法。首先已知$p_j w_j^T$是純量，令$p_j w_j^T = \beta/\alpha$。A.6式可被整理成下式：

$$w_j = (\alpha p_j + k_j) G_j^{-1} \tag{A.7}$$

這是$w_j$的更新式，式中還留著一個未知數$\alpha$。將此更新式以及$p_j w_j^T = \beta/\alpha$代回A.6，整理過後可得：

$$\alpha^2 (p_j G_j^{-1} p_j^T) + \alpha p_j G_j^{-1} k_j^T - \beta = 0 \tag{A.8}$$

此為以$\alpha$為變數的一元二次方程式，根據其公式解，可解出兩個$\alpha$。在實做上給定$p_j$，利用A.7、A.8式可以求出兩組$w_j$，將兩組解代回A.4式， 選擇能使目標函數值較大的解。

本方法在目標函數可被寫為式A.4、聲學特徵調適可直接被寫為式A.1時適用，包括本論文提到的限制型最大相似度線性回歸、基於特徵的L2規範正則化、使用窗型變數集的變數選取-特徵最大相似度線性回歸等等。

## A.2 低維度最佳化法

在實做聲學特徵向量的仿射轉換之前，若我們先以另一個$d \times D$線性轉換矩陣$W_L$將聲學特徵向量降維，則以下式調適聲學特徵向量:

$$o'_t - o_t = A^{(d)}\epsilon_t^{(d)} + b^{(d)} \tag{A.9}$$

其中$\epsilon_t^{(d)} = W_L o_t$，$A^{(d)}$、$b^{(d)}$分別為$D \times d$矩陣及$D$維向量。如此一來，我們便無法直接將最大相似度目標函數寫為式A.4。為了解決這樣的最佳化問題，本節提出低維度最佳化法。 首先，基於最大相似度訓練準則，我們先求低維度最佳化法的目標函數。為了把式A.9中的$A^{(d)}$、$b^{(d)}$合併， 可將其改寫為:

$$o'_t = W^{(d)}\eta_t^{(d)} + o_t \tag{A.10}$$

其中$\eta_t^{(d)} = [(\epsilon_t^{(d)})^T \ 1]^T$、$W^{(d)} = [A^{(d)} \ b^{(d)}]$。$W_d$即為我們要求得的目標。 為了能使用式A.2作為目標函數，令$Q_L = [W_L^T \ \vec{0}]^T$，$\vec{0}$為$D$維全零向量，式A.10可再被改寫為：

$$o'_t = Ao_t + b^{(d)}A = W^{(d)}Q_L + I \tag{A.11}$$

其中$I$為$D \times D$單位矩陣 (identity matrix) 將上式代入式A.2，即可得到低維度最佳化法的目標函數：

$$F(W^{(d)}) =$$

$$-\sum_{i,t}\gamma_i(t)[(W^{(d)}\eta_t^{(d)} + o_t - \mu_i)^T\Sigma_i^{-1}(W^{(d)}\eta_t^{(d)} + o_t - \mu_i) - log|W^{(d)}Q_L + I|^2] \tag{A.12}$$

接下來，我們將一些較複雜的符號經過整理，代換成以下變數:

$$
\begin{aligned}
G_j^{(d)} &= \sum_{i,t} \frac{\gamma_i(t)}{\sigma_{ij}} \eta_t^{(d)} (\eta_t^{(d)})^T \\
k_j^{(d)} &= \sum_{i,t} \frac{\gamma_i(t)}{\sigma_{ij}} (\mu_{ij} - o_t j)(\eta_t^{(d)})^T \\
\beta &= \sum_{i,t} \gamma_i(t) \\
|A| &= p_j A_j^T = p_j (Q_L^T w_j^{(d)} + I_j)
\end{aligned}
\tag{A.13}
$$

其中$w_j^{(d)}$、$A_j$分別為$W^{(d)}$、$A$的第$j$列；$I_j$為$I$的第$j$行；$p_j = [c_{j1} \;\; c_{j2} ...], c_{jk} = cofactor(A_{jk})$。代換後目標函數可以被寫為：

$$
F(W^{(d)}) = log(p_j(Q_L^T w_T^{(d)} + I_j)) - \sum_{j=1}^{D} [w_j^{(d)} G_j^{(d)} (w_j^{(d)})^T - 2 w_j^{(d)} k_j^T] \tag{A.14}
$$

為了將目標函數最佳化，我們將上式對$w_j^{(d)}$偏微分並令其為零，可得：

$$
w_j^{(d)} G_j^{(d)} - k_j^{(d)} = \beta \frac{p_j Q_L^T}{p_j (Q_L^T w_j^{(d)} + I_j)} \tag{A.15}
$$

觀察此式，我們再次發現$p_j$與$w_j^{(d)}$有關，因此$w_j^{(d)}$無法直接被解出。利用和上節相似的技巧，已知$p_j(Q_L^T w_j^{(d)} + I_j)$為純量，我們令$p_j(Q_L^T w_j^{(d)} + I_j) = \beta/\alpha$，代入上式可將其整理為：

$$
w_j^{(d)} = (k_j^{(d)} + \alpha p_j Q_L^T)(G_j^{(d)})^{-1} \tag{A.16}
$$

這是$w_j^{(d)}$的更新式，式中同樣還留著一個未知數$\alpha$。將此更新式以及$p_j(Q_L^T w_j^{(d)} + I_j) = \beta/\alpha$代回A.15，整理過後可得：

$$
\alpha^2 (p_j Q_L^T (G_j^{(d)})^{-1} Q_L p_j^T) + \alpha p_j (Q_L^T (G_j^{(d)})^{-1} k_j^T + I_j) - \beta = 0 \tag{A.17}
$$

此為以$\alpha$為變數的一元二次方程式，根據其公式解，可解出兩個$\alpha$。在實做上給定$p_j$，利用A.7、A.8式可以求出兩組$w_j^{(d)}$，將兩組解代回A.14式，選擇能使目標函數值較大的解。

本節介紹的方法主要功能在於，當使用式A.9實做仿射轉換調適法時，如何找出最佳解。其中$W_L$可以為任何有降維功能的矩陣，以本論文提出的變數選取-特徵最大相似度線性回歸來說， 若採用主成份分析建立變數子集，$W_L$即為聲學特徵向量主成分組成的空間基底。

# 參 考 文 獻


[1] "DSP history - understanding speech: An interview with john makhoul," *IEEE Signal Processing Magazine*, pp. 76–79.

[2] Douglas O'Shaughnessy, "Invited paper: Automatic speech recognition: History, methods and challenges," *Pattern Recognition*, vol. 41, no. 10, pp. 2965 – 2979, 2008.

[3] M. J. F. Gales and S. J. Young, "Cepstral parameter compensation for hmm recognition in noise," *Speech Commun.*, vol. 12, no. 3, pp. 231–239, July 1993.

[4] Liang-Che Sun and Lin-Shan Lee, "Modulation spectrum equalization for improved robust speech recognition," *Trans. Audio, Speech and Lang. Proc.*, vol. 20, no. 3, pp. 828–843, Mar. 2012.

[5] Yang Chang, "Robust speech recognition with two-dimensional frame-and-feature weighting and modulation spectrum normalization," M.S. thesis, National Taiwan University, Taiwan, 2012.

[6] Phil C. Woodland, "Speaker adaptation for continuous density HMMs: A review," in *ITRW on Adaptation Methods for Speech Recognition*, Aug. 2001, pp. 11–19.

[7] Yu Tsao, "Speaker adaptation for mandarin syllable/tone recognition with limited data," M.S. thesis, National Taiwan University, Taiwan, 2001.

[8] M.J.F. Gales, "Cluster adaptive training of hidden markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 8, pp. 417–428, 1999.

[9] Kai Yu and M. J.F. Gales, "Discriminative cluster adaptive training," *Trans. Audio, Speech and Lang. Proc.*, vol. 14, no. 5, pp. 1694–1703, Sept. 2006.

[10] Philip C. Woodland, D. Pye, and M. J. F. Gales, "Iterative unsupervised adaptation using maximum likelihood linear regression.," in *ICSLP*. 1996, ISCA.

[11] Christopher M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[12] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171 – 185, 1995.

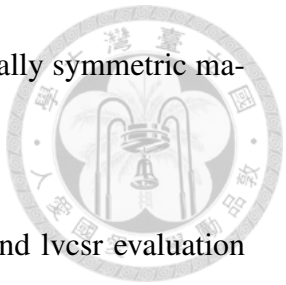
[13] M.J.F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[14] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, vol. 39, no. 1, pp. 1–38, 1977.

[15] C.-H. Lee J. Li, Y. Tsao, "Shrinkage model adaptation in automatic speech recognition.," in *InterSpeech*. 2010, pp. 1656–1659, ISCA.

[16] C.-H. Lee J. Li, M. Yuan, "Lasso model adaptation for automatic speech recognition.," in *ICML*, 2011.

[17] Olivier Siohan, Cristina Chesta, and Chin-Hui Lee, "Hidden markov model adaptation using maximum a posteriori linear regression," pp. 147–150, 1999.



[18] Wu Chou, "Maximum a posterior linear regression with elliptically symmetric matrix variate priors," in *EUROSPEECH'99*, 1999, pp. –1–1.

[19] N. Parihar and J. Picone, "Aurora working group: Dsr front end lvcsr evaluation au/384/02," *Institutefor Signal and Information Processing report*, 2002.

[20] Isabelle Guyon and André Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar. 2003.

[21] M. Dash and H. Liu, "Feature Selection for Classification," *Intelligent Data Analysis*, vol. 1, pp. 131–156, 1997.

[22] I. T. Jolliffe, *Principal Component Analysis*, Springer, second edition, Oct. 2002.

[23] George E. Dahl, Student Member, Dong Yu, Senior Member, Li Deng, and Alex Acero, "Context-dependent pre-trained deep neural networks for large vocabulary speech recognition," in *IEEE Transactions on Audio, Speech, and Language Processing*, 2012.