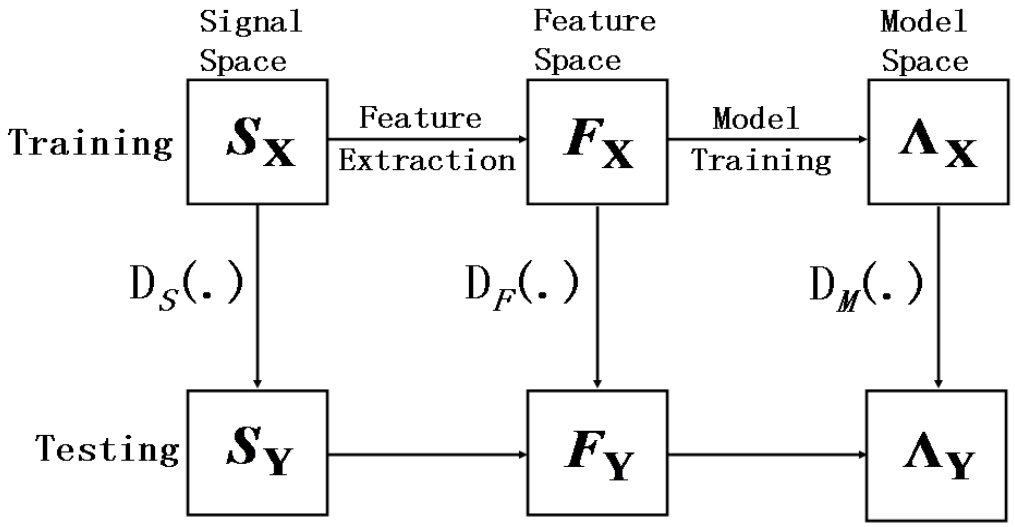
線性映射轉換函數於聲學模型調適之強健式語音辨識

曹昱，蘇煜程，王緒翔
中央研究院資訊科技創新研究中心

1. 簡介

隨著科技的發展與網路的普及，語音識別系統(automatic speech recognition，ASR)已應用於多項可攜式裝置。然而多項研究成果顯示，語音辨識技術的應用仍然受限於一項存在已久卻未完全解決的問題，就是聲學環境不匹配所造成語音辨識效能不佳的問題。聲學環境不匹配表示在語音辨識系統中，由於語者本身的發聲狀況、背景雜訊、通道及麥克風特性，造成語料的訓練環境以及測試環境有所差異，這樣的差異會嚴重影響語音辨識系統的效能。聲學環境的不匹配可以從語音訊號、特徵向量以及聲學模型等三個象限來觀察，如圖一所示[1]。假設訓練環境的語音訊號、特徵向量以及聲學模型象限的表現形式分別是 S_X、F_X、Λ_X，相對應測試環境的象限為 S_Y、F_Y、Λ_Y，則我們可以在這三個象限觀察到三種不匹配值，分別是 $D_S(.)$，$D_F(.)$、$D_M(.)$。對於改善聲學環境不匹配的研究，主要可以分成三大類。第一大類試圖降低 $D_S(.)$對原始語音訊號，S_X，的影響。主要的方法為語音增強技術(speech enhancement)[2]，包含了濾波器技術(filtering techniques)[2, 3]、頻譜回復技術(spectral restoration techniques)[2, 4]、以及模型技術(speech model based techniques)[2]。



圖一：訓練與測試環境不匹配示意圖

第二大類的目的是改善特徵向量，此類方法又可以再細分成兩種：如何求得對環境影響較不敏感的語音特徵向量以及語音特徵向量對雜訊的補償方式。第一種方法中知名的有 SS (spectral subtraction)及其衍伸的方法[5]、CMS(cepstral mean subtraction)[6]、更高維的倒頻譜平均消去法與 HEQ(histogram equalization)[7-9]。近年來歐洲電信標準協會(European telecommunications standards institute, ETSI)推出一套新的特徵向量提取方法—AFE(advanced front-end)[10]，並證明利用 AFE 取得的特徵函數，能顯著地提升語音辨識系統在雜訊環境中的效能。第二種方法，先求得轉換函數來模擬測試語料，F_Y，與利用訓練語料求得的聲學模型，Λ_X，之間的不匹配量。之後再由求出來的轉換函數調整 F_Y，使其匹配 Λ_X。主要的方法有 CDCN(codeword dependent cepstral normalization)[11]、SPLICE(stereo-based piecewise linear compensation environments)[12]。此外，fMLLR(feature-space maximum likelihood linear regression)[13, 14]、Feature-space Eigen-MLLR[15]、ML-based SFM(ML-based stochastic feature matching)[16]以及 MAP-based SFM[17]等等，都是利用轉換函數來補償測試語料的特徵向量，也是被視為有效而快速的特徵向量補償演算法。

第三大類則是在聲學模型象限下降低環境不匹配的影響。此類方法亦可再分為兩種：強化聲學模型對環境的適應性以及調整聲學模型參數來匹配測試環境。對於第一種方法而言，我們可以從兩個方面進行：第一，是收集不同語者在不同聲學環境下的語音資料；或者使用人工合成的方法模擬出不同聲學環境下的語音資料。收集到的語料稱為複合環境訓練語料(multi-condition training data)。利用複合環境訓練語料，我們可以訓練出一套聲學模型。在聲學環境不匹配的情況之下，研究指出這種複合環境訓練(multi-style training)產生出的聲學模型比單一環境訓練(single-style training)而得的聲學模型，在環境的不匹配的情況下，能提供較佳的抗噪能力[18, 19]。第二，利用鑑別式訓練方法(discriminative training，DT)，改善傳統的最大相似度估測法(maximum likelihood，ML)[20]。有效的鑑別式訓練方法包括最小化分類錯誤(minimum classification error, MCE) [21]、最大化交互資訊(maximum mutual information, MMI)[22]、最小化音素錯誤(minimum phone error, MPE)[23]、與柔性邊際估測法則(soft margin estimation，SME)[24]等等。第二種方法是模型調適，也是本專文討論的重點，我們將在下一節中詳述。

2. 聲學模型調適

聲學模型調適法，會先準備一個聲學模型結構，Ω，這個聲學模型結構可以是單一聲學模型($\Omega = \{\Lambda^X\}$)或是多套聲學模型($\Omega = \{\Lambda^1, \Lambda^2, \ldots \Lambda^P\}$)，其中$\Lambda^X$以及 P 套聲學模型Λ^p(p=1,2…P)都是由訓練語料求得。對於聲學模型調適而言，我們利用(式

1)求得目標Λ^{Y}的聲學模型：

$$\Lambda^{\mathrm{Y}} = \mathrm{T}_{\varphi}(\Omega)，\qquad (式\ 1)$$

其中T_{φ}是轉換函數，φ是轉換函數中的參數。我們利用T_{φ}描述訓練與測試語料之間聲學環境的差別。最後利用轉換函數，T_{φ}，轉換聲學模型結構，Ω，而求得一套匹配測試環境的聲學模型，Λ^{Y}[1, 25]。為了求取轉換函數T_{φ}中的參數，φ，通常需要一段測試環境的語音訊號，稱為調適語料(adaptation data)。由(式 1)，我們利用(式 2)估算測試環境模型，Λ^{Y}，中的第 m 個 Gaussian 的 mean vector，μ_m^{Y}：

$$\mu_m^{\mathrm{Y}}=\mathrm{T}_{\varphi}(\xi_m)，\qquad (式\ 2)$$

其中ξ_m是一個延伸向量，可以來自單一聲學模型，$\xi_m = \{\mu_m^{\mathrm{X}}\}$，或是多套聲學模型，$\xi_m = \{\mu_m^1, \mu_m^2 \ldots \mu_m^P\}$，其中$\mu_m^{\mathrm{X}}$與$\mu_m^P$是$\Lambda^{\mathrm{X}}$與$\Lambda^P$的第 m 個 Gaussian 的 mean vector。之前的研究顯示，決定最佳調適效能有兩項關鍵性的因素：(1)轉換函數；(2)目標函數。接下來，本文將分別介紹這兩項關鍵性的因素。

2.1 轉換函數

所用的聲學模型結構，Ω，將決定轉換函數的種類。由於來自語者及聲學環境所造成的總體影響可能很複雜，如果只用一個聲學模型來做調適($\Omega = \{\Lambda^{\mathrm{X}}\}$)，會需要複雜的轉換函數來準確地描述這個總體影響。一項先前的研究利用非線性轉換(non-linear transformation)來模擬這個總體影響，此非線性轉換隨後被用於調整模型參數，實驗結果證明這種非線性轉換之模型調適方法可以有效地改進辨識效能[26]。然而，在可獲得調適語料是有限的情況下，複雜的轉換函數可能導致過適現象(over-fitting)，反而降低辨識率。因此在有限的調適語料情況下，我們較常使用簡單的函數，如補償向量(bias compensation，BC)[1]以及線性迴歸(linear regression，LR) [27]作為轉換函數。然而，若調適語料逐漸增加，卻仍使用過於簡單的函數，將使得聲學模型調適的進步曲線快速達到飽和。

另一方面，如果我們使用多套聲學模型($\Omega = \{\Lambda^1, \Lambda^2, \ldots \Lambda^P\}$)，即可用比較簡單的轉換函數完成調適，比如：BF(best first)、線性組合(linear combination，LC)、線性組合加上補償向量(linear combination with correction bias，LCB)。這些轉換函數被應用在 Eigenvoice[28]、CAT(cluster adaptive training)[29]、ESSEM(ensemble speaker and speaking environment modeling) [30, 31]。在 Eigenvoice 方法中，首先在離線運算時利用主成分分析(principal component analysis，PCA)[32]，對多套聲學環境模型建立起一個特徵空間(Eigenspace)，在線上運算時利用 LC 轉換函數來求得新的聲學模型。另外，CAT 在離線時把整個訓練語料集依照聲學特歸類成若干群，每群內部的訓練語料都有較為相似的聲學特性；接著，利用每一群訓練語料求出一套聲學模型。在線上運算時利用 LC 轉換函數求得新的聲學模型。而 ESSEM 則在離線時準備多套單一環境訓練的聲學模型，每一套模型表示某一種

特殊的聲學環境;然後利用所有的聲學模型集做環境群組(environment clustering，EC)及環境切割(environment partitioning，EP)；最後在線上運作時利用一個轉換函數得到新的聲學模型[30, 33]。

2.2 目標函數

在模型調適中，我們定義目標函數來求取轉換函數，T_{φ}，的參數，φ。研究指出，選用適當的目標函數，能有效地提升聲學模型調適的效能。一般而言，最大相似度估測法(ML)能利用少量調適語料適當地估算轉換函數參數，因此被廣泛的使用。ML-based stochastic matching[1]、MLLR(maximum likelihood linear regression)[27]都是基於 ML 估測法提出的演算法。然而，以 ML 為目標函數時，在調適語料量較少的情形下，常常會出現over-fitting的問題。為了改善這個問題，最常見的方法是改用 MAP 為目標函數。基於 MAP 為目標函數的方法包括 SMAP(structural maximum a posteriori)[34]、MAPLR(maximum a posteriori linear regression)[35]等等。使用 MAP 的另一項好處是可以設計適當的先驗知識，如基於語言或是統計學理論而設計的，來幫助轉換參數的估算。另外，我們也可以引入鑑別式函數為目標函數，調適聲學模型，知名的例子有 MCELR(MCE linear regression)[36]、SMELR(SME-based linear regression)[37]以及 MPELR(MPE-based linear regression)[38]等等。

在下一節中，我們將探討一項最近提出的轉換函數，線性映射(linear projection, LP)。我們首先討論利用 ML 以及 MAP 為目標函數來求取 LP 轉換函數的參數。接著，我們將探討線性映射與現有知名的轉換函數之間的關係。

3. 線性映射轉換函數之聲學模型調適

在本節中，我們將介紹本專文的主題：線性映射轉換函數於聲學模型調適。我們將介紹以 ML 以及 MAP 為目標函數來求取線性映射轉換函數中的參數，並且比較線性映射與其他幾種知名的轉換函數之間的關係。

3.1 線性映射函數

對於線性映射函數，我們把(式 2)寫成：

$$\mu_m^{\mathrm{Y}}=\mathrm{T}_{\varphi}(\xi_m)=\mathrm{A}^1\mu_m^1+\mathrm{A}^2\mu_m^2+\cdots\mathrm{A}^P\mu_m^P+b\,, \qquad (式\ 3)$$

其中b是D維的補償向量($b=[b_{(1)}\ b_{(2)}\ \ldots b_{(D)}]$)，$\mathrm{A}^p(p=1,2\ldots P)$為$(D\times D)$矩陣。對於$\mu_m$，我們準備一個先驗機率函數(prior density)：

$$p\ \{\mathrm{F}_{\varphi}(\xi_m)\}\sim\left\{\prod_{i=1}^{D}\exp\left[-\frac{1}{2\,\mathrm{V}_{m_{(i)}}}\left(\mathrm{F}_{\varphi}(\xi_m)_{(i)}-\eta_{m_{(i)}}\right)^2\right]\right\}, \qquad (式\ 4)$$

其中V_m、η_m是先驗機率函數的超參數(hyper-parameters)。

當利用 MAP 求取 (式 3)中的$\{\mathrm{A}^1,\mathrm{A}^2,\ldots\mathrm{A}^P\}$ 跟b，可以得到：

$$[\mathrm{A}^1\ \mathrm{A}^2 \dots \mathrm{A}^P\ b]'_{(i)} = (G_{(i)})^{-1} k_{(i)}\ ,\qquad (式\ 5)$$

其中：

$$G_{(i)} = \sum_{t=1}^{T} \sum_{s \in W_c} r_s(t) \frac{1}{\Sigma_{s_{(i)}}} \rho_s \rho_s{}' + \sum_{s \in S} \frac{\epsilon_s}{\mathrm{V}_{s_{(i)}}} \rho_s \rho_s{}'\ ,\qquad (式\ 6)$$

$$k_{(i)} = \sum_{t=1}^{T} \sum_{s \in W_c} r_s(t) \frac{1}{\Sigma_{s_{(i)}}} o_{t_{(i)}} \rho_s + \sum_{s \in S} \frac{\epsilon_s}{\mathrm{V}_{s_{(i)}}} \eta_{s_{(i)}} \rho_s\ ,\qquad (式\ 7)$$

在(式 6)及(式 7)中o_t是第 t 個時間點的觀測值，$r_s(t)$是第 t 個時間點在第 s 個 Gaussian 的後驗機率；$\rho_s = [\mu_s^{1\prime}\mu_s^{2\prime} \dots \mu_s^{P\prime} 1]'$，$\mu_s^p$是第 p 個聲學環境模型Λ^p的第 s 個 mean vector；$\Sigma_{s_{(i)}}$是第 s 個 Gaussian 的 covariance matrix 的第 i 個參數值。$s \in W_c$ 是指第 s 個 Gaussian 有出現在參考文檔(reference transcription)，Wc 中。(式 5)–(式 7)是以 MAP 為目標函數所求的解，我們稱之為最大後驗概率線性映射(maximum a posteriori linear projection，MAPLP)。當我們設定$\epsilon = 0$，(式 5)–(式 7)成為以 ML 為目標函數所求的解，我們將之稱為最大相似度線性映射(maximum likelihood linear projection，MLLP)。

我們可以簡化(式 5)中的$\mathrm{A}^p, p = 1,2 \dots P$。利用對角矩陣取代全矩陣，我們可以得到，$\mathrm{A}^p = \mathrm{diag}[a_{(1)}^p, a_{(2)}^p, \dots a_{(D)}^p]$。在這個設定下，我們利用(式 8)–(式 10)求$[\mathrm{A}^1\ \mathrm{A}^2 \dots \mathrm{A}^P\ b]$：

$$[a_{(i)}^1\ a_{(i)}^2 \dots\ a_{(i)}^P\ b_{(i)}]' = \left(G_{(i)}\right)^{-1} k_{(i)}\ ,\qquad (式\ 8)$$

其中

$$G_{(i)} = \sum_{t=1}^{T} \sum_{s \in W_c} r_s(t) \frac{1}{\Sigma_{s_{(i)}}} \rho_s \rho_s{}' + \sum_{s \in S} \frac{\epsilon_s}{\mathrm{V}_{s_{(i)}}} \rho_s \rho_s{}'\ ,\qquad (式\ 9)$$

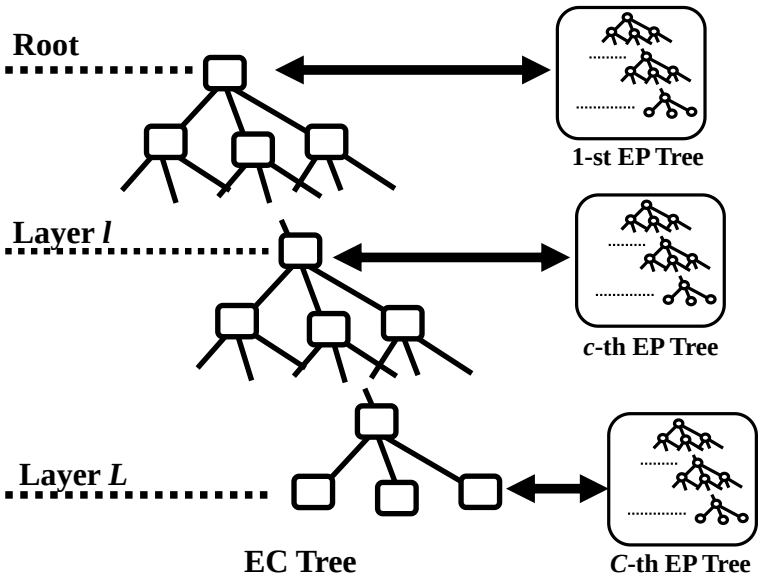
$$k_{(i)} = \sum_{t=1}^{T} \sum_{s \in W_c} r_s(t) \frac{1}{\Sigma_{s_{(i)}}} o_{t_{(i)}} \rho_s + \sum_{s \in S} \frac{\epsilon_s}{\mathrm{V}_{s_{(i)}}} \eta_{s_{(i)}} \rho_s\ ,\qquad (式\ 10)$$

其中 $\rho_s = [\mu_{s_{(i)}}^1 \mu_{s_{(i)}}^2 \dots \mu_{s_{(i)}}^P\ 1]'$。

3.2 準備先驗知識

先驗知識會影響 MAP 方法的效能。這篇專文中我們介紹四種先驗知識，分別是 CP(clustered prior)、SP(sequential prior)、HP(hierarchical prior)以及 IP(integrated prior)。我們可以由一個兩層的樹狀結構，如圖二，求得這些先驗知識。第一層對所有訓練語料分類，我們稱之為 EC 樹狀結構；每一個 EC 樹狀結構的節點代表一群訓練語料，這些訓練語料有較為接近的聲學特性。利用同一節點中的訓練語料，我們可以準備一套或是多套聲學模型。第二層則是對聲學模型來分類，我

們稱之為 EP 樹狀結構。每一個 EP 樹狀結構是基於 EC 樹狀結構裡一個節點中的訓練語料所建立的，建立的方法是依據此節點中的聲學模型參數以 data driven 的方式得到的。因此，在兩層樹狀結構中，總共會有一個 EC 樹狀結構以及 C 個 EP 樹狀結構，我們用一個 EP 樹狀結構描述某一群聲學環境的特性，也用來準備先驗知識。



圖二：兩層式樹狀結構

在準備 CP 時，我們先將訓練語料分成 K 群$\{\mathrm{T}_\varphi(\xi_m)^{(1)}, \mathrm{T}_\varphi(\xi_m)^{(2)} \ldots \mathrm{T}_\varphi(\xi_m)^{(K)}\}$基於這 K 群的訓練語料，基於這些分群過後的訓練語料，我們可以利用(式 11)、(式 12)求得先驗機率函數的 hyper-parameters，$\{\eta_m^{CP}, \mathrm{V}_m^{CP}\}$：

$$\eta_{m_{(i)}}^{CP} = \frac{1}{K}\sum_{k=1}^{K} \mathrm{T}_\varphi(\xi_m)_{(i)}^{(k)} \ , \qquad \text{(式 11)}$$

$$\mathrm{V}_{m_{(i)}}^{CP} = \frac{1}{K}\sum_{k=1}^{K} \left[\mathrm{T}_\varphi(\xi_m)_{(i)}^{(k)} - \eta_{m_{(i)}}^{CP}\right]^2 \ , \qquad \text{(式 12)}$$

其中$\eta_{m_{(i)}}^{CP}$及$\mathrm{V}_{m_{(i)}}^{CP}$分別是$\eta_m^{CP}$與$\mathrm{V}_m^{CP}$的第 i 個參數值。

在準備 SP 先驗知識時，我們利用之前聲學模型調適的結果當作目前聲學模型調適步驟的參考。為了減低運算量，我們在使用 SP 的時候，僅更新η_m^{SP}而保持V_m^{SP}不變。由於第一句並無先驗知識，因此我們設定$\epsilon_s = 0, \forall s \in S$來求解(式 6)、(式 7)以及(式 9)、(式 10)。而從第二句開始，前一句所求得的$\mathrm{T}_\varphi(\xi_m)$將被利用為超參數來完成當下的聲學模型調適。對於完成第 u 句聲學模型調適時，超參數

$\eta_m^{SP(u)}$來自於$\mathrm{T}_\varphi(\xi_m)^{(u-1)}$：

$$\eta_m^{SP(u)} = \mathrm{T}_\varphi(\xi_m)^{(u-1)} \ , \qquad \text{(式 13)}$$

其中$\mathrm{T}_\varphi(\xi_m)^{(u-1)}$是第$(u\text{-}1)$句求得的模型參數。

在利用 HP 先驗知識時，我們先對 EP 樹狀結構的根結點(root node)，求取一個轉換函數$\mathrm{T}_\varphi^{(0)}$，基於這個$\mathrm{T}_\varphi^{(0)}$我們先轉換所有的聲學模型參數。接著，我們利用根結點中轉換好的聲學模型參數當作 HP 的超參數，持續求取 EP 樹狀結構下一層的轉換函數。對於第 n 層中第 q 個節點，其超參數，$\eta_m^{HP(n)}$，為

$$\eta_m^{HP(n)} = \mathrm{T}_\varphi(\xi_m)^{(n-1)} , \qquad \text{(式 14)}$$

其中$\mathrm{T}_\varphi(\xi_m)^{(n-1)}$是來自於第 q 個節點的父節點(parent node)。

本研究中，我們設計出一個函數，$\Gamma(.)$，來結合 CP、SP 與 HP 先驗知識：

$$\eta_n^{IP} = \Gamma(\eta_n^{CP}, \eta_n^{SP}, \eta_n^{HP}) \ , \qquad \text{(式 15)}$$

其中η_n^{CP}、η_n^{SP}與η_n^{HP}分別為 CP、SP 與 HP 的超參數，η_n^{IP}則為組合的 IP 超參數。而$\Gamma(.)$則被設計成η_n^{CP}、η_n^{SP}與η_n^{HP}的線性組合，因此從(式 15)得到：

$$\eta_n^{IP} = w_{CP}\eta_n^{CP} + w_{SP}\eta_n^{SP} + w_{HP}\eta_n^{HP} \ , \qquad \text{(式 16)}$$

其中w_{CP}、w_{SP}與w_{HP}分別是給予 CP、SP 與 HP 結合的權重值，而且滿足：

$$w_{CP} + w_{SP} + w_{HP} = 1 , \qquad \text{(式 17)}$$

在實作上，我們可以利用訓練語料來求w_{CP}、w_{SP}以及w_{HP}。

3.3 線性映射轉換函數與其他轉換函數的關係

在此節中，我們將討論線性映射轉換函數與其他知名函數之間的關係。我們將由使用單一聲學模型($\Omega = \{\Lambda^{\mathrm{X}}\}$)來做調適，以及使用多套聲學模型($\Omega = \{\Lambda^1, \Lambda^2, \ldots \Lambda^P\}$)來做調適來分別討論。

3.3.1 單一聲學模型($\Omega = \{\Lambda^X\}$)

在此類方法中，最有名的轉換函數為線性迴歸(LR)以及補償向量(BC)。對於 LR 轉換函數而言，我們設定(式 2)中的$\xi_m = \{\mu_m^{\mathrm{X}}\}$，$\mu_m^{\mathrm{Y}}$可由(式 18)求得：

$$\mu_m^{\mathrm{Y}} = \mathrm{T}_\varphi(\xi_m) = \mathrm{A}\mu_m^{\mathrm{X}} + b 。 \qquad \text{(式 18)}$$

(式 18)中的A以及b可由(式 19)、(式 20)與(式 21)求得：

$$[\mathrm{A}\ b]'_{(i)} = (G_{(i)})^{-1}k_{(i)} , \qquad \text{(式 19)}$$

其中：

$$G_{(i)} = \sum_{t=1}^{T}\sum_{s\in W_c} r_s(t)\frac{1}{\Sigma_{s_{(i)}}}\rho_s{\rho_s}' + \sum_{s\in S}\frac{\epsilon_s}{\mathrm{V}_{s_{(i)}}}\rho_s{\rho_s}' \ , \qquad \text{(式 20)}$$

$$k_{(i)} = \sum_{t=1}^{T} \sum_{s \in W_c} r_s(t) \frac{1}{\Sigma_{s_{(i)}}} o_{t_{(i)}} \rho_s + \sum_{s \in S} \frac{\epsilon_s}{\mathrm{V}_{s_{(i)}}} \eta_{s_{(i)}} \rho_s \ ,$$ (式 21)

其中$\rho_s = [\mu_s^{\mathrm{X}\prime}\ 1]'$。

同樣地，我們可以將(式 18)中的A設計為對角矩陣的形式，$\mathrm{A} = \mathrm{diag}[a_{(1)}, a_{(2)}, \ldots a_{(D)}]$，並由(式 22)-(式 24)來求取A以及$b$：

$$[a_{(i)}\ b_{(i)}]' = (G_{(i)})^{-1} k_{(i)} \ ,$$ (式 22)

其中：

$$G_{(i)} = \sum_{t=1}^{T} \sum_{s \in W_c} r_s(t) \frac{1}{\Sigma_{s_{(i)}}} \rho_s \rho_s{}' + \sum_{s \in S} \frac{\epsilon_s}{\mathrm{V}_{s_{(i)}}} \rho_s \rho_s{}' \ ,$$ (式 23)

$$k_{(i)} = \sum_{t=1}^{T} \sum_{s \in W_c} r_s(t) \frac{1}{\Sigma_{s_{(i)}}} o_{t_{(i)}} \rho_s + \sum_{s \in S} \frac{\epsilon_s}{\mathrm{V}_{s_{(i)}}} \eta_{s_{(i)}} \rho_s \ ,$$ (式 24)

其中：$\rho_s = [\mu_{s_{(i)}}^{\mathrm{X}}\ 1]'$。比較 LP 與 LR，我們可以看出不同點在於 LP 使用了 P 套矩陣，$\mathrm{A}^P, p = 1 \ldots P$，而 LR 只用一個矩陣，A。

當使用 BC 轉換函數，我們設定(式 2)中的$\xi_m = \{\mu_m^{\mathrm{X}}\}$，$\mu_m^{\mathrm{Y}}$可由(式 25)求得：

$$\mu_m^{\mathrm{Y}} = \mathrm{T}_\varphi(\xi_m) = \mu_m^{\mathrm{X}} + b \text{。}$$ (式 25)

(式 25)中的補償向量b可由(式 26)-(式 28)來求取

$$b_{(i)} = (G_{(i)})^{-1} k_{(i)} \ ,$$ (式 26)

其中：

$$G_{(i)} = \sum_{t=1}^{T} \sum_{s \in W_c} r_s(t) \frac{1}{\Sigma_{s_{(i)}}} + \sum_{s \in S} \frac{\epsilon_s}{\mathrm{V}_{s_{(i)}}} \ ,$$ (式 27)

$$k_{(i)} = \sum_{t=1}^{T} \sum_{s \in W_c} r_s(t) \frac{(o_{t_{(i)}} - \rho_{s_{(i)}})}{\Sigma_{s_{(i)}}} + \sum_{s \in S} \frac{\epsilon_s (\eta_{s_{(i)}} - \rho_{s_{(i)}})}{\mathrm{V}_{s_{(i)}}} \ ,$$ (式 28)

其中$\rho_{s_{(i)}} = \mu_{s_{(i)}}^{\mathrm{X}}$。比較 LP 與 BC，LP 使用了 P 套矩陣，而 BC 使用一個 D×D 單位矩陣(identity matrix)，I，作為矩陣A。

3.3.2 多套聲學模型($\Omega = \{\Lambda^1, \Lambda^2, \ldots \Lambda^P\}$)

在這部分，我們討論三種轉換函數，線性組合加上補償向量(LCB)、線性組合(LC)、BF(best first)。與 LP 不同的是，這些轉換函數利用不同形式的$\{\mathrm{A}^1, \mathrm{A}^2, \ldots \mathrm{A}^P\}$與 b，以下我們將分別討論這三種方法。

使用 LCB 轉換函數時，我們設定$\xi_m = \{\mu_m^1, \mu_m^2 \ldots \mu_m^P\}$，$\mu_m^{\mathrm{Y}}$可由(式 29)求得：

$$\mu_m^{\mathrm{Y}} = \mathrm{T}_\varphi(\xi_m) = \mathrm{A}^1 \mu_m^1 + \mathrm{A}^2 \mu_m^2 + \cdots \mathrm{A}^P \mu_m^P + b \ ,$$ (式 29)

其中μ_m^p是第 p 個聲學模型，Λ^p，第 m 個 Gaussian 的 mean vector。與線性映射

轉換函數(LP)不同的是，LCB 中的$\mathrm{A}^p = \mathrm{diag}[\omega^p, \omega^p, \dots \omega^p]\,, p = 1,2 \dots P$，我們可由(式 30)- (式 32)求得$\{\mathrm{A}^1, \mathrm{A}^2, \dots \mathrm{A}^P\}$：

$$[\omega^1\ \omega^2 \dots \omega^P\ b']' = G^{-1}k\,, \qquad \text{(式 30)}$$

其中：

$$G = \sum_{t=1}^{T}\sum_{s\in W_c} r_s(t)\ \mathrm{H}_s'\Sigma_s^{-1}\mathrm{H}_s + \sum_{s\in S}\epsilon_s\ \mathrm{H}_s'\mathrm{V}_s^{-1}\mathrm{H}_s\ , \qquad \text{(式 31)}$$

$$k = \sum_{t=1}^{T}\sum_{s\in W_c} r_s(t)\mathrm{H}_s'\Sigma_s^{-1}o_t + \sum_{s\in S}\epsilon_s\ \mathrm{H}_s'\mathrm{V}_s^{-1}\eta_s\ , \qquad \text{(式 32)}$$

其中$\mathrm{H}_s = [\mu_s^1\ \mu_s^2 \dots \mu_s^P\ I]$，$I$是一個 D×D 的單位矩陣。

使用 LC 轉換函數時，我們設定 $\xi_m = \{\mu_m^1, \mu_m^2 \dots \mu_m^P\}$；$\mu_m^{\mathrm{Y}}$可由(式 33)求得：

$$\mu_m^{\mathrm{Y}} = \mathrm{T}_\varphi(\xi_m) = \mathrm{A}^1\mu_m^1 + \mathrm{A}^2\mu_m^2 + \cdots \mathrm{A}^P\mu_m^P\,, \qquad \text{(式 33)}$$

LC 中的$\mathrm{A}^p = \mathrm{diag}[\omega^p, \omega^p, \dots \omega^p]\,, p = 1,2 \dots P$，我們可由(式 34)-(式 36)求得$\{\mathrm{A}^1, \mathrm{A}^2, \dots \mathrm{A}^P\}$：

$$[\omega^1\ \omega^2\ \dots \omega^P]' = G^{-1}k\,, \qquad \text{(式 34)}$$

其中：

$$G = \sum_{t=1}^{T}\sum_{s\in W_c} r_s(t)\mathrm{H}_s'\Sigma_s^{-1}\mathrm{H}_s + \sum_{s\in S}\epsilon_s\ \mathrm{H}_s'\mathrm{V}_s^{-1}\mathrm{H}_s\ , \qquad \text{(式 35)}$$

$$k = \sum_{t=1}^{T}\sum_{s\in W_c} r_s(t)\mathrm{H}_s'\Sigma_s^{-1}o_t + \sum_{s\in S}\epsilon_s\ \mathrm{H}_s'\mathrm{V}_s^{-1}\eta_s\ , \qquad \text{(式 36)}$$

其中$\mathrm{H}_s = [\mu_s^1\ \mu_s^2 \dots \mu_s^P]$。

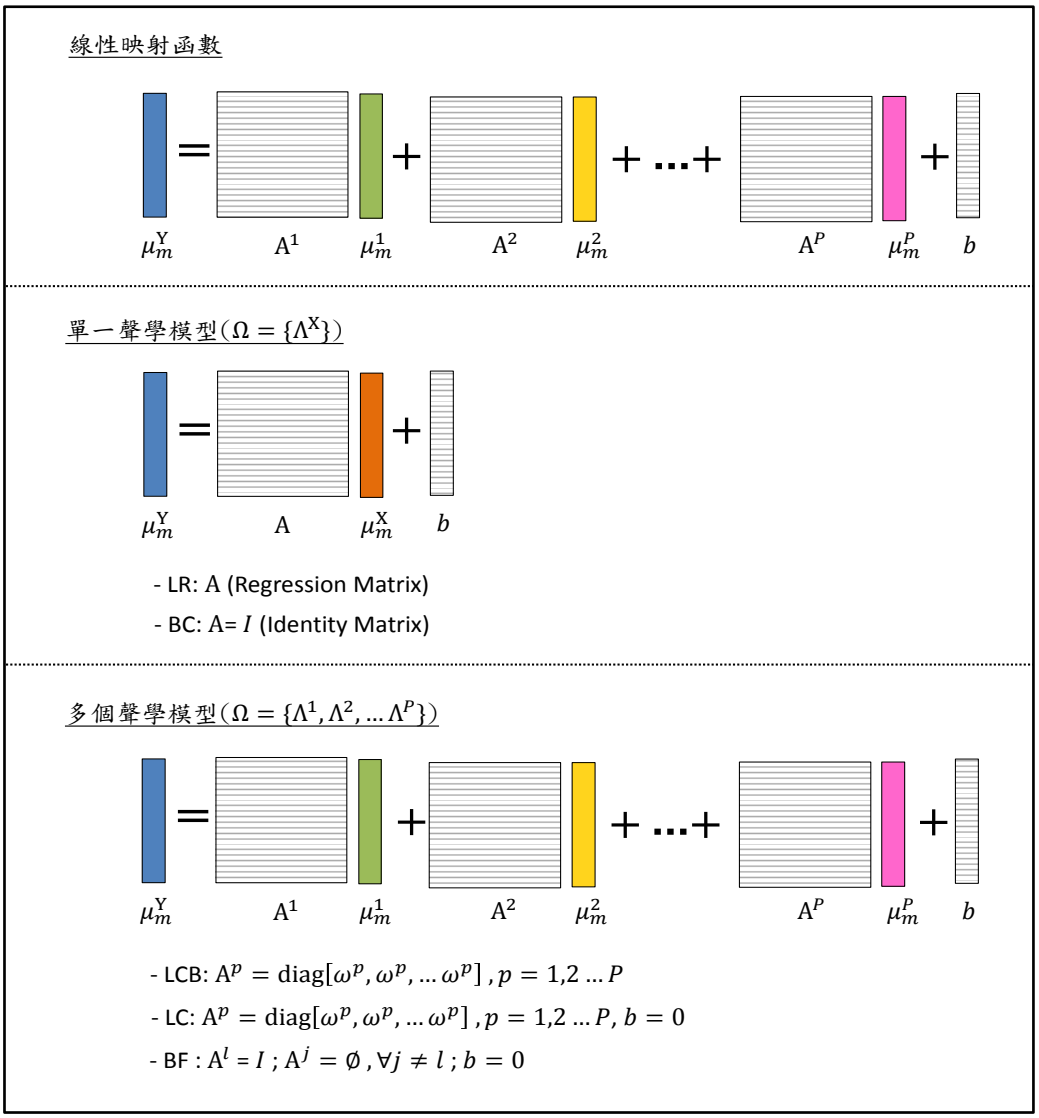
使用 BF 時，我們設定(式 2)中的$\xi_m = \{\mu_m^1, \mu_m^2 \dots \mu_m^P\}$，$\mu_m^{\mathrm{Y}}$可由(式 37)求得：

$$\mu_m^{\mathrm{Y}} = \mathrm{T}_\varphi(\xi_m) = \mathrm{A}^1\mu_m^1 + \mathrm{A}^2\mu_m^2 + \cdots \mathrm{A}^P\mu_m^P\,, \qquad \text{(式 37)}$$

其中只有第 l 個矩陣A^l為一個 D×D 單位矩陣，I，其他的矩陣皆為零矩陣。我們利用(式 38)來求取l：

$$l = \underset{p}{argmin}\sum_{t=1}^{T}\sum_{s\in W_c} r_s(t)\left[\left(o_t - \mu_s^p\right)'\Sigma_s^{-1}\left(o_t - \mu_s^p\right)\right]\,,\quad p = 1, 2, \dots P。 \qquad \text{(式 38)}$$

圖三中我們整理線性映射函數與其他轉換函數的比較，包括使用單一聲學模型($\Omega = \{\Lambda^{\mathrm{X}}\}$)的兩種方法，LR 與 BC；以及多套聲學模型($\Omega = \{\Lambda^1, \Lambda^2, \dots \Lambda^P\}$)的三種方法，LCB、LC 與 BF。



圖三：線性映射函數與其他轉換函數的比較

4. 實驗環境設置與實驗結果討論

在本節中我們介紹實驗環境以及討論實驗結果。實驗中我們比較各種轉換函數及ML、MAP 兩種目標函數。其中在使用 MAP 目標函數時，我們使用 IP 先驗知識。

4.1. 實驗設置

我們利用 Aurora2 語料庫[39]來做語音辨識實驗，並利用其中的 multi-condition 訓練語料集來訓練聲學模型。在這個 multi-condition 訓練語料集中共有 8440 句語料，這些語料包含四種加成式雜訊，以及五種訊雜比(signal-to-noise ration，SNR)，訊雜比為 5dB 到 20B SNRs 以及 clean condition。我們將整個訓練語料集依照語者的性別分成各有 4220 句的男性語料以及女性語料。利用整個 multi-condition 訓練語料集以及兩個 GD(gender dependent)訓練語料子集，我們訓練出三套聲學模型，包含一個 GI(gender independent)以及兩個 GD 聲學模型。在本實驗中，以

上的三個聲學模型，是利用複雜型的隱藏式馬可夫模型來求得的[40]，每套模型包含十一個 digit、一個 silence 以及一個 short pause 模型。每個 digit 模型由 16 個狀態組成，每個狀態利用 20 個高斯混和模型來描述。silence 模型由三個狀態組成，short pause 模型由一個狀態組成，每個狀態包括 36 個高斯混和模型。我們用 AFE 技術取得語音特徵參數[10]。除了 13 維的 static 特徵參數，我們加上 26 維的一、二階導數(first- and second-order time derivative)特徵參數。

我們利用 Aurora-2 測試集中的 50 個測試環境(10 種雜訊，訊雜比為 0dB 到 20B SNRs)來比較辨識結果，每一個測試環境有 1,001 句測試語料。這 50 個測試環境分成 SetA、SetB、以及 SetC 等三個子集。其中 SetA 包含四類雜訊，這四類雜訊與 multi-condition 訓練語料集中的四類雜訊相同，SetB 則包含了另外四類加成式雜訊、SetC 除了加成式雜訊之外，還有通道雜訊。我們利用字錯誤率(word error rate，WER)來比較辨識結果。在接下來的實驗結果中，除了 Baseline 結果之外，所有的實驗結果都是依單句非監督式(per-utterance unsupervised)調適的模式完成。

表 1 列出各種轉換函數的複雜度，其中 LR 的A以及 LP 中的A^{p}，我們都使用簡化的對角矩陣。另外，在本實驗中，我們設定 D=39，P=3。

表 1：各種轉換函數的複雜度

Function	LR	BC	LCB	LC	LP
Complexity	$D+D$	D	$P+D$	P	$D\times P+D$

4.2. 實驗結果及討論

在本節中，我們將呈現在 Aurora-2 測試集 SetA、SetB、SetC、以及 50 個測試環境的平均值(All)的實驗結果。

4.2.1 Baseline 以及 BF

表 2 列出 Baseline 以及 BF 的辨識結果。我們使用的 Baseline，是直接利用 GI 聲學模型而得，不對 GI 模型做任何調適。而 BF 則是針對每一句話，直接從三個聲學模型(一個 GI 以及兩個 GD)中，選擇一個最佳的聲學模型來辨識。選擇模型的方法如(式 38)。

表 2：BASELINE 以及 BF 的辨識結果(WER %)

Test Condition	SetA	SetB	SetC	All
Baseline	5.92	6.69	7.11	6.46
BF	5.68	6.48	6.90	6.24

4.2.2 以 ML 為目標函數，幾種不同轉換函數的比較

在表 3 中，我們列出以 ML 為目標函數來求取上述五種轉換函數參數在 SetA、SetB、SetC 以及 All 所得到的辨識結果(WER %)。第一列列出 MLLP 的結果；第二、第三列列出 MLLR 以及 MLBC 的結果；第四、第五列則列出 MLLCB 以及 MLLC 的結果。請注意如同之前討論到的，LP、LCB、以及 LC 轉換函數，使用三個聲學模型(一個 GI 以及兩個 GD 聲學模型)完成模型調適，而 LR 以及 BC 轉換函數，只使用單一聲學模型(一個 GI 聲學模型)完成模型調適。

我們首先比較表 3 中第一、第二以及第三列，在 SetA、SetC 以及 All 的字錯誤率上，MLLR 提供了比 MLBC 更好的結果，這驗證了線性迴歸矩陣對於模型調適的功效。此外，我們也觀察到在所有測試環境中，MLLP 都提供了比 MLLR 更佳的效能；由此，我們得知有效地利用訓練語料中的 local information，並利用多個線性迴歸矩陣，可以提升模型調適的效能。

比較表 3 中第一、第四以及第五列的結果，可以發現使用 MLLCB 調適模型的方式比起 MLLP 與 MLLC 有較優良的效能；相較於 LCB 轉換函數，LP 轉換函數有較多的參數量，所以 MLLP 在實驗中會發生 over-fitting 的現象，關於這點，我們接下來將利用 MAP 為目標函數，搭配有效的先驗機率來解決。

表 3：以 ML 為目標函數的辨識結果(WER %)

Test Condition	SetA	SetB	SetC	All
MLLP	5.58	6.14	6.32	5.95
MLLR	5.65	6.20	6.33	6.01
MLBC	5.89	**6.10**	6.62	6.12
MLLCB	5.52	6.26	**6.01**	**5.91**
MLLC	**5.51**	6.52	6.50	6.11

4.2.3 以 MAP 為目標函數，幾種不同轉換函數的比較

表 4 中，我們列出以 MAP 為目標函數來求取上述五種轉換函數參數在 SetA、SetB、SetC 以及 All 所得到的辨識結果(WER %)。第一列列出 MAPLP 的結果；第二、第三列列出 MAPLR 以及 MAPBC 的結果；第四、第五列則列出 MAPLCB 以及 MAPLC 的結果。請同樣地注意到 LP、LCB、以及 LC 轉換函數，使用三個聲學模型(一個 GI 以及兩個 GD 聲學模型)完成模型調適，而 LR 以及 BC 轉換函數，僅使用了單一聲學模型(一個 GI 聲學模型)完成模型調適。

首先比較表 4 中，第一、二、三列在 SetA、SetB、SetC 以及 All 的辨識結果，MAPLP 轉換函數都提供了比 MAPLR 以及 MAPBC 更好的辨識結果。此外，比較表 4 中第一、第四以及第五列的結果，可以發現，不同於表 3 的結果， MAPLP

調適模型的方式比起 MAPLCB 與 MAPLC 都有較優良的效能。最後，比較表 3 以及表 4，我們發現以 MAP 為目標函數能明顯地提升 LP 轉換函數的效能，證明了在調適語料不足的情況下，利用 MAP 為目標函數，能夠有效地減輕 over-fitting 的問題，進而提升辨識率。

表 4：以 MAP 為目標函數的辨識結果(WER %)

Test Condition	SetA	SetB	SetC	All
MAPLP	**5.34**	**6.07**	5.94	**5.75**
MAPLR	5.59	6.28	6.14	5.98
MAPBC	5.86	6.08	6.60	6.10
MAPLCB	5.46	6.22	**5.93**	5.86
MAPLC	5.51	6.44	6.55	6.09

5. 結論

此專文討論最近一項關於 LP 轉換函數用於聲學模型調適的研究成果。首先，我們討論 LP、LR 與 BC 轉換函數之間的關聯，這三種轉換函數的差別，在於 LP 使用了多套聲學模型，而 LR 與 BC 轉換函數則使用單一聲學模型來完成調適。接著討論了 LP、LCB 與 LC 轉換函數的關聯，相較於 LCB 與 LC 轉換函數，LP 轉換函數顯得較為複雜。就求取轉換函數的參數部分，我們探討了 ML 以及 MAP 兩種目標函數。對於 MAP 目標函數而言，我們事先提供一個 IP 先驗函數，此 IP 包含了來自於不同方法得到的先驗知識，如之前的聲學模型轉換、樹狀結構下得到的結果等。在實驗設定上，我們使用 Aurora-2 語料庫，以單句非監督式(per-utterance unsupervised)調適的模式來完成聲學模型調適，最後獲得實驗成果。我們發現在以 ML 為目標函數的結果中，比起 LR、BC 與 LC，LP 有較佳的效能，惟較 LCB 差，其原因在於 LP 所需估測的參數量較多，所以當調適語料量不足的情形時，會有 over-fitting 的問題。因此，當我們使用 MAP 為目標函數來估測時，LP 就能提供相對於其他四種轉換函數更佳的辨識效能。在我們的實驗結果中，MAPLP 得到最佳的效能，相較於 Baseline 的辨識結果，MAPLP 的 WER 在 Aurora-2 的 50 個測試環境中，平均值降低了 10.99% (從 6.46%到 5.75%)。

6. 參考文獻

[1] A. Sankar and C.-H. Lee, "A maximum-likelihood approach to stochastic matching for robust speech recognition," *IEEE Trans. Speech Audio Proc.*, vol. 4, pp. 190-202, 1996.

[2] J. Chen, Fundamentals of Noise Reduction in Spring Handbook of Speech Processing, Chapter 43, Springer, 2008.

[3] E. Hänsler and G. Schmidt, Topic in Acoustic Echo and Noise Control, Chapter 9,

Springer, 2006.
[4] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator,” *IEEE Trans. Acoustics Speech and Signal Proc.*, vol. ASSP-32, pp. 1109–1121, 1984.
[5] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. Acoustics, Speech, Signal Proc.*, vol. 27, pp.113-120, 1979.
[6] H. Kim and R. C. Rose, “Cepstrum-domain acoustic feature compensation based on decomposition of speech and noise for ASR in noisy environments,” *IEEE Trans. Speech Audio Proc.*, vol. 11, pp. 435-446, 2003.
[7] C.-W. Hsu and L.-S. Lee, “Higher order cepstral moment normalization (HOCMN) for robust speech recognition,” in *Proc. ICASSP*, pp.197-200, 2004.
[8] Y. H. Suk, S. H. Choi, and H. S. Lee, “Cepstrum third-order normalization method for noisy speech recognition,” *Electronics Letters*, vol. 35, pp. 527-528, 1999.
[9] A. Torre, A. M. Peinado, J. C. Segura, J. L. Perez-Cordoba, M. C. Bentez, and A. J. Rubio, “Histogram equalization of speech representation for robust speech recognition,” *IEEE Trans. Speech and Audio Proc.*, vol. 13, pp. 355-366, 2005.
[10] ETSI (2002). Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithm. ETSI standard document ES 202 050.
[11] A. Acero, “Acoustical and environmental robustness in automatic speech recognition,” Ph.D. Dissertation, ECE, Department, CMU, 1990.
[12] L. Deng, J. Droppo, and A. Acero, “Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition,” *IEEE Trans. Speech Audio Proc.*, vol. 11, pp.568-580, 2003.
[13] M. J. F. Gales, “Maximum likelihood linear transformations for HMM-based speech recognition,” Technical Report, Cambridge University, 1997.
[14] Y. Li, H. Erdogan, Y. Gao, and E. Marcheret, “Incremental online feature space MLLR adaptation for telephony speech recognition,” in *Proc. ICSLP*, pp. 1417-1420, 2002.
[15] K. Visweswariah, V. Goel, and R. Gopinath, “Structuring linear transforms for adaptation using training time information,” in *Proc. ICASSP*, pp. 585-588, 2002.
[16] H. Jiang, F. Soong and C.-H. Lee, “Hierarchical stochastic matching for robust speech recognition,” in *Proc. ICASSP*, pp. 217 - 220, 2001.
[17] Y. Tsao, P. R. Dixon, C. Hori, and H. Kawai, “Incorporating regional information to enhance MAP-based stochastic feature compensation for robust speech recognition,” in *Interspeech*, pp. 2585-2588, 2011.
[18] R. P. Lippmann, E. A. Martin, and D. B. Paul, “Multi-style training for robust isolated-word speech recognition,” in *Proc. ICASSP*, pp. 705-708, 1987.

[19] X. Huang, A. Acero, and H. Hon, Spoken Language Processing: A Guide to Theory, Algorithm and System Development, Prentice Hall, 2001.
[20] A. P. Dempster, N. M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. B*, vol. 39, pp. 1-38, 1977.
[21] B.-H. Juang, W. Chou, and C.-H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Trans. Speech Audio Proc.*, vol. 5, pp. 257-265, 1997.
[22] V. Valtchev, J. Odell, P. C. Woodland and S. Young, "MMIE training of large vocabulary recognition systems," *Speech Communication*, vol. 22, pp. 303-314, 1997.
[23] D. Povey and P. C. Woodland, "Minimum phone error and i-smoothing for improved discriminative training," in *Proc. ICASSP*, pp. I105-I108, 2002.
[24] J. Li, "Soft margin estimation for automatic speech recognition," Ph.D. Dissertation, School of ECE, Georgia Institute of Technology, 2008.
[25] C.-H. Lee, "On stochastic feature and model compensation approaches to robust speech recognition," *Speech Communication*, pp.29-47, vol. 25, 1998.
[26] A. C. Suredran, C.-H. Lee, and M. Rahim, "Nonlinear compensation for stochastic matching," *IEEE Trans. Speech Audio Proc.*, vol. 7, pp.643-655, 1999.
[27] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech and Lang.*, vol. 9, pp.171-185, 1995.
[28] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in Eigenvoice space," *IEEE Trans. Speech Audio Proc.*, vol. 8, pp. 695–707, 2000.
[29] M. J. F. Gales, "Cluster adaptive training of hidden Markov models," *IEEE Trans. Speech Audio Proc.*, vol. 8, pp. 417–428, 2000.
[30] Y. Tsao and C.-H. Lee, "An ensemble speaker and speaking environment modeling approach to robust speech recognition," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 17, pp. 1025–1037, 2009.
[31] Y. Tsao, J. Li, and C.-H. Lee, "Ensemble speaker and speaking environment modeling approach with advanced online estimation process," in *Proc. ICASSP*, pp. 3833–3836, 2009.
[32] I. T. Jolliffe, Principal Component Analysis. Berlin, Germany: Springer-Verlag, 1986.
[33] Y. Tsao, R. Isotani, H. Kawai, and S. Nakamura, "An environment structuring framework to facilitating suitable prior density estimation for MAPLR on robust speech recognition," in *Proc. ISCSLP*, pp. 29-32, 2010.
[34] K. Shinoda and C.-H. Lee, "A structural Bayes approach to speaker adaptation," *IEEE Trans. Speech Audio Proc.*, vol. 9, pp. 276-287, 2001.
[35] O. Siohan, C. Chesta, and C.-H. Lee, "Hidden Markov model adaptation using

maximum a posteriori linear regression," in *Proc. Workshop Robust Methods for Speech Recognition in Adverse Conditions*, pp. 147-150, 1999.
[36] J. Wu and Q. Huo, "A study of minimum classification error (MCE) linear regression for supervised adaptation of MCE-trained continuous-density hidden Markov models," *IEEE Trans. Speech Audio Proc.*, vol. 15, pp. 478-488, 2007.
[37] S. Matsuda, Y. Tsao, J. Li, S. Nakamura, and C.-H. Lee, "A study on soft margin estimation of linear regression parameters for speaker adaptation," in *Proc. Interspeech*, pp. 1603-1606, 2009.
[38] Wang, L. and Woodland, P. C., "MPE-based discriminative linear transform for speaker adaptation," in *Proc. ICASSP*, pp. 321-324, 2004.
[39] D. Macho, L. Mauuary, B. Noe, Y. M. Cheng, D. Ealey, D. Jouver, H. Kelleher, D. Pearce, and F. Saadoun, "Evaluation of a noise-robust DSR front-end on Aurora databases," in *Proc. ICSLP*, pp. 17-20, 2002.
[40] J. Wu and Q. Huo, "Several HKU approaches for robust speech recognition and their evaluation on Aurora connected digit recognition tasks," in *Proc. Eurospeech*, pp. 21-24, 2003.