改善以最小化音素錯誤爲基礎的鑑別式聲學模型訓練於中文連續語音辨識之研究

劉士弘, 朱芳輝, 陳柏琳
國立臺灣師範大學資訊工程學系
{ g93470185, g94470144, berlin}@ntnu.edu.tw

摘要

本論文探討改善最小化音素錯誤爲基礎的鑑別式聲學模型訓練於中文大詞彙連續語音辨識之研究。首先，本論文提出一個新的音框層次音素正確率函數來取代最小化音素錯誤訓練的原始音素正確率函數，此新的音素正確率函數在某種程度上能充分地懲罰刪除錯誤。其次，本論文提出一個以音框層次正規化熵值爲基礎的嶄新資料選取方法來改進鑑別式訓練，其正規化熵值是由訓練語料所產生之詞圖中高斯分布之事後機率所求得。此資料選取方法可以讓鑑別式訓練更集中在那些離決定邊界較近的訓練樣本所收集的統計值，以達到較佳的鑑別力。所使用的實驗題材是公視新聞外場記者語料。初步的實驗結果顯示，結合時間音框層次的資料選取方法和新的音素正確率函數在前幾次的迭代訓練中確實有些微且一致的進步。

關鍵詞：最小化音素錯誤訓練，鑑別式訓練，資料選取方法，大詞彙連續語音辨識

一、緒論

語音，是人與人之間最自然的溝通橋樑，倘若語音能夠成爲資訊產品的主要輸入形式，那麼人與機器之間的溝通就會變得簡單許多，並且可以盡量避免文明病的產生。因此自動語音辨識(Automatic Speech Recognition, ASR)的研究已變得非常重要，這也是目前語音與語言處理領域中熱門的研究議題之一。

大詞彙連續語音辨識(Large Vocabulary Continuous Speech Recognition, LVCSR)所使用的鑑別式訓練法則是不以最大化訓練語料的相似度爲目標，而以最小分類錯誤爲目標，進而增進辨識率。傳統在聲學模型之訓練上，大都使用最大化相似度(Maximum Likelihood, ML)法則，配合波氏重估演算法(Baum-Welch algorithm)來進行聲學模型的訓練，但此種訓練方法並沒有考慮語音辨識時聲學模型彼此間的關係，在調整聲學模型參數之後，可以使得相關的語音特徵落在此聲學模型的相似度(Likelihood)變大，卻也可能同時讓非相關的語音特徵落在此聲學模型的相似度更大，造成辨識上的混淆。因此，近來有不少研究針對此項缺點，提出鑑別式訓練(Discriminative Training)法則來加以改進。故本文著重於探討以最小化音素錯誤(Minimum Phone Error Training, MPE)爲基礎的鑑別式訓練法則，藉著提出一個新的音框層次音素正確率函數來改善原始音素正確率函數之缺點，同時也提出一個以音框層次正規化熵值爲基礎的嶄新資料選取方法來改進最小化音素錯誤訓練。

本論文接下來的安排如下：第二章將介紹貝氏風險與全面風險；第三章則介紹最小化音素錯誤聲學模型訓練；第四章探討最小化音素錯誤訓練之改進；第五章則探討資

料選取方法於改進最小化音素錯誤聲學模型訓練；第六章爲實驗與討論；第七章爲結論與未來展望。

二、貝氏風險與全面風險

語音辨識的過程可視爲一個分類的動作，將每句可能的詞序列都視爲一類，語音辨識即是要從所有可能類別(詞序列)中找出最佳的一類(一句)。若 O_z 爲一語句的語音特徵向量序列，將 O_z 歸類至詞序列 W 時，可以用函數 $R(W \mid O_z)$ 代表此歸類行爲的風險(Risk)；而語音辨識則可視爲找出此風險最低的詞序列。 $R(W \mid O_z)$ 可定義如下[1]：

$$R(W \mid O_z) = \sum_{W' \in \mathbf{W}} l(W, W') P(W' \mid O_z) \tag{1}$$

其中 $\mathbf{W}$ 爲所有可能詞序列所成的集合； $P(W' \mid O_z)$ 表示給定語音特徵向量序列 O_z 時，詞序列 W' 的事後機率(Posterior Probability)； $l(W, W')$ 爲一減損函數(Loss Function)，用以表示詞序列 W 與 W' 之間差異所造成的損失(Loss)， $R(W \mid O_z)$ 爲將 O_z 歸類至 W 時的期望損失(Expected Loss)，又稱爲貝氏風險(Bayes Risk)或條件風險(Conditional Risk)。在語音辨識或解碼上，需要最小化此貝氏風險來找最佳的詞序列 $\hat{W}$ ，即：

$$\hat{W} = \arg\min_{W \in \mathbf{W}} R(W \mid O_z) = \arg\min_{W \in \mathbf{W}} \sum_{W' \in \mathbf{W}} l(W, W') P(W' \mid O_z) \tag{2}$$

目前有許多辨識器根據貝氏決策定理(Bayesian Decision Theorem)，即最小化此貝氏風險(式(2))來設計其搜尋演算法，如標準最大化事後機率解碼方法(Maximum a Posteriori Decoding, MAP)[2]、ROVER(Recognizer Output Voting Error Reduction)[3]、最小化貝氏風險(Minimum Bayes Risk, MBR)[4]、最小化時間音框錯誤搜尋(Minimum Time Frame Error Search)[5]及詞錯誤最小化(Word Error Minimization) [6]等。

然而，若在聲學模型和語言模型的訓練上，則需要計算全面風險(Overall Risk)，並且最小化此全面風險 R_{all} [1]：

$$R_{all} = \int R(W \mid O) P(O) dO \tag{3}$$

其中 W 爲語音特徵向量序列 O 對應之正確轉譯詞序列， $P(O)$ 爲 O 的事前機率(Prior Probability)；全面風險 R_{all} 是在語句空間(語音特徵向量序列空間)上作積分，爲所有訓練語句(語音特徵向量序列)的期望條件風險(Expected Conditional Risk)。由於訓練語料有限，故全面風險可簡化爲 Z 個訓練語句的條件風險總和：

$$R_{all} = \sum_{z=1}^{Z} R(W_z \mid O_z) P(O_z) = \sum_{z=1}^{Z} \sum_{W' \in \mathbf{W}} l(W_z, W') P(W' \mid O_z) P(O_z) \tag{4}$$

若事後機率分布 $P(W' \mid O_z)$ 由聲學模型 λ 及語言模型 Γ 所決定，令 $\theta = \{\lambda, \Gamma\}$ ，所以事後機率我們將之表示爲 $P(W' \mid O_z; \theta)$ ，則全面風險可改寫成：

$$R_{all} = \sum_{z=1}^{Z} R(W_z \mid O_z) P(O_z) = \sum_{z=1}^{Z} \sum_{W' \in \mathbf{W}} l(W_z, W') P(W' \mid O_z; \theta) P(O_z) \tag{5}$$

若假設 $P(O_z)$ 對所有 O_z 均有一致(Uniform)的機率，且此項與模型參數 λ 及 Γ 無關，則可將此項省略：

$$R_{all} = \sum_{z=1}^{Z} \sum_{W' \in \mathbf{W}} l(W_z, W') P(W' \mid O_z; \theta) \tag{6}$$

在估測聲學模型和語言模型時，希望估測之模型θ能將全面風險降至最低：

$$\hat{\theta} = \arg\min_{\theta} \sum_{z=1}^{Z} \sum_{W' \in \mathbf{W}} l(W_z, W') P(W' \mid O_z; \theta) \tag{7}$$

在此所表示的減損函數是一般化減損函數(Generalized Loss Function)，並沒有明確定義要如何計算，這也因此成爲一個開放的研究議題(亦即要如何去設計一個減損函數以期望訓練出較佳的模型θ，進而提高辨識率)。目前有許多的模型訓練的方法都是以風險最小化(Risk Minimization)爲基礎，並搭配其設計的減損函數來達成鑑別式之模型訓練，如最大化交互資訊估測(Maximum Mutual Information Estimation, MMIE) [7]、全面風險估測法則(Overall Risk Criterion Estimation, ORCE) [8]、最小化貝氏風險鑑別式訓練(Minimum Bayes Risk Discriminative Training, MBRDT) [9]、最小化音素錯誤訓練(Minimum Phone Error Training, MPE) [10]等。

三、最小化音素錯誤之聲學模型訓練

新近劍橋大學提出的最小化音素錯誤(Minimum Phone Error, MPE)聲學模型訓練，是以全面風險爲出發，以辨識出詞序列的原始音素正確率(Raw Phone Accuracy)函數$A(W_i, W_z)$來取代其中減損函數$l(W_i, W_z)$。因此，它的目標函數變成是最大化語音辨識器對所有訓練語句(語音特徵向量序列)O_z的可能辨識出候選詞序列W_i($W_i \in \mathbf{W}_z = \{W_1, W_2, W_3, \cdots\}$)的期望音素正確率(也就是最小化語音辨識器對所有訓練語句可能辨識出候選詞序列W_i的期望錯誤率)，最小化音素錯誤的目標函數可表示如下：

$$F_{MPE}(\lambda) = \sum_{z=1}^{Z} \sum_{W_i \in \mathbf{W}_z} p(W_i \mid O_z) A(W_i, W_z) = \sum_{z=1}^{Z} \sum_{W_i \in \mathbf{W}_z} \frac{p_\lambda(O_z \mid W_i) P(W_i)}{p(O_z)} A(W_i, W_z) \tag{8}$$

其中$p(O_z)$可用語音辨識器產生的詞圖$\mathbf{W}_{z,lattice}$來近似[11]，因此目標函數可進一步表示成：

$$F_{MPE}(\lambda) \approx \sum_{z} \sum_{W_i \in \mathbf{W}_{z,lattice}} \frac{p_\lambda(O_z \mid W_i) P(W_i)}{\sum_{W_k \in \mathbf{W}_{z,lattice}} p_\lambda(O_z \mid W_k) P(W_k)} A(W_i, W_z) \tag{9}$$

其中W_i與W_k分別表示詞圖$\mathbf{W}_{z,lattice}$上任兩條候選詞序列(假設O_z對應的正確詞序列W_z亦包含在詞圖裡)。

爲了對目標函數$F_{MPE}(\lambda)$進行最佳化，Povey 等人提出最小化音素錯誤的弱性(Weak-sense)輔助函數$H_{MPE}(\lambda, \bar{\lambda})$爲[12]：

$$H_{MPE}(\lambda, \bar{\lambda}) = \sum_{z=1}^{Z} \sum_{q \in \mathbf{W}_{z,lattice}} \left[\frac{\partial F_{MPE}(\lambda)}{\partial \log p_\lambda(O_z \mid q)} \Big|_{\lambda=\bar{\lambda}} \right] \log p_\lambda(O_z \mid q) \tag{10}$$

其中$\frac{\partial F_{MPE}(\lambda)}{\partial \log p_\lambda(O_z \mid q)} \Big|_{\lambda=\bar{\lambda}}$的值可爲正或負，取決於詞圖上通過此音素的候選詞序列的期望正確率$c_z(q)$是否大於詞圖上所有候選詞序列的期望正確率c_{avg}^z。也就是：

$$\frac{\partial F_{MPE}(\lambda)}{\partial \log p_{\lambda}(O_z \mid q)}\Big|_{\lambda=\bar{\lambda}} = \gamma_q^z\left(c_z(q) - c_{avg}^z\right) \tag{11}$$

其中：

$$\gamma_q^z = \frac{\sum_{W_i \in \mathbf{W}_{z,lattice}: q \in W_k} p_{\bar{\lambda}}(O_z \mid W_k) P(W_k)}{\sum_{W_k \in \mathbf{W}_{z,lattice}} p_{\bar{\lambda}}(O_z \mid W_k) P(W_k)} \tag{12}$$

爲詞圖上通過音素段落 q 的候選詞序列的事後機率和，而

$$c_z(q) = \frac{\sum_{W_i \in \mathbf{W}_{z,lattice}, q \in W_i} p_{\bar{\lambda}}(O_z \mid W_i) P(W_i) A(W_i, W_z)}{\sum_{W_k \in \mathbf{W}_{z,lattice}: q \in W_k} p_{\bar{\lambda}}(O_z \mid W_k) P(W_k)} \tag{13}$$

爲詞圖上通過此音素段落的候選詞序列的期望正確率，而

$$c_{avg}^z = \frac{\sum_{W_i \in \mathbf{W}_{z,lattice}} p_{\bar{\lambda}}(O_z \mid W_i) P(W_i) A(W_i, W_z)}{\sum_{W_k \in \mathbf{W}_{z,lattice}} p_{\bar{\lambda}}(O_z \mid W_k) P(W_k)} \tag{14}$$

爲詞圖上所有候選詞序列的期望正確率。 γ_q^z 、 $c_z(q)$與 c_{avg}^z 的統計量可在詞圖上使用波氏重估演算法來求得[12]。

另一方面，針對對數機率函數 $\log p_{\lambda}(O_z \mid q)$ ，必需透過一個強性輔助函數 $Q_{ML}(\lambda, \bar{\lambda}, z, q)$ 來估測新的模型參數值，因此弱性輔助函數 $H_{MPE}(\lambda, \bar{\lambda})$ 可表示成：

$$H'_{MPE}(\lambda, \bar{\lambda}) = \sum_{z=1}^{Z} \sum_{q \in \mathbf{W}_{z,lattice}} \left[\frac{\partial F_{MPE}(\lambda)}{\partial \log p_{\lambda}(O_z \mid q)}\Big|_{\lambda=\bar{\lambda}} \right] Q_{ML}(\lambda, \bar{\lambda}, z, q) \tag{15}$$

若以 $\gamma_q^{z\,MPE}$ 來表示 $\frac{\partial F_{MPE}(\lambda)}{\partial \log p(O_z \mid q)}\Big|_{\lambda=\bar{\lambda}}$ ，且 $Q_{ML}(\lambda, \bar{\lambda}, z, q)$ 可表示如下:

$$Q_{ML}(\lambda, \bar{\lambda}, z, q) = \sum_{t=s_q}^{e_q} \sum_{m} \gamma_q^z(t) \log N(o_z(t); \mu_{qm}, \Sigma_{qm}) \tag{16}$$

其中 $o_z(t)$爲 O_z 的第 t 個語音特徵向量； $N(\cdot; \mu_{qm}, \Sigma_{qm})$ 是音素段落 q 的第 m 個高斯分布， μ_{qm} 與 Σ_{qm} 分別是它的平均值向量與共變異矩陣。因此弱性輔助函數 $H_{MPE}(\lambda, \bar{\lambda})$ 可進一步表示成：

$$H'_{MPE}(\lambda, \bar{\lambda}) = \sum_{z} \sum_{q \in \mathbf{W}_{z,lattice}} \sum_{t=s_q}^{t=e_q} \sum_{m} \gamma_q^{z\,MPE} \gamma_{qm}^z(t) \log N(o_z(t); \mu_{qm}, \Sigma_{qm}) \tag{17}$$

其中 s_q 與 e_q 分別爲音素段落 q 的開始與結束時間， $\gamma_{qm}^z(t)$ 爲語音特徵向量 $o_z(t)$在音素段落 q 上的高斯分布 m 的佔有機率。若把平滑函數 $H_{SM}(\lambda, \bar{\lambda})$ 加入弱性輔助函數 $H'_{MPE}(\lambda, \bar{\lambda})$ ，則 $H'_{MPE}(\lambda, \bar{\lambda})$ 可進一步表示成[12]：

$$H''_{MPE}(\lambda,\overline{\lambda}) = \sum_z \sum_{q\in \mathbf{W}_{z,lattice}} \sum_{t=s_q}^{t=e_q} \sum_m \gamma_q^{z^{MPE}} \gamma_{qm}^z(t) \log N(o_z(t), \mu_{qm}, \Sigma_{qm}) - \sum_{q,m} \frac{D_{qm}}{2} \left[\log(|\Sigma_{qm}|) + (\mu_{qm} - \overline{\mu}_{qm})^T \Sigma_{qm}^{-1} (\mu_{qm} - \overline{\mu}_{qm}) + tr(\overline{\Sigma}_{qm} \Sigma_{qm}^{-1})\right] \quad (18)$$

而平滑函數 $H_{SM}(\lambda,\overline{\lambda})$ 表示爲：

$$H_{SM} = \sum_{q,m} \frac{D_{qm}}{2} \left[\log(|\Sigma_{qm}|) + (\mu_{qm} - \overline{\mu}_{qm})^T \Sigma_{qm}^{-1} (\mu_{qm} - \overline{\mu}_{qm}) + tr(\overline{\Sigma}_{qm} \Sigma_{qm}^{-1})\right] \quad (19)$$

其中 $\overline{\mu}_{qm}$ 與 $\overline{\Sigma}_{qm}$ 爲舊有模型的平均值向量與共變異矩陣。我們可以對 $H'_{MPE}(\lambda,\overline{\lambda})$ 使用延伸波式(Extended Baum-Welch ，EBW)演算法得到聲學模型參數估測更新公式(當假設語音特徵向量維度間爲無關時，亦即共變異矩陣爲對角矩陣)[12]：

$$\mu_{qmd} = \frac{\{\theta_{qmd}^{num}(O) - \theta_{qmd}^{den}(O)\} + D_{qmd}\overline{\mu}_{qmd}}{\{\gamma_{qm}^{num} - \gamma_{qm}^{den}\} + D_{qmd}}$$
$$\sigma_{qmd}^2 = \frac{\{\theta_{qmd}^{num}(O^2) - \theta_{qmd}^{den}(O^2)\} + D_{qmd}(\overline{\sigma}_{qmd}^2 + \overline{\mu}_{qmd}^2)}{\{\gamma_{qm}^{num} - \gamma_{qm}^{den}\} + D_{qmd}} - \mu_{qmd}^2 \quad (20)$$

其中統計值資訊可分爲兩類，亦即 *num*(numerator)與 *den*(denominator)兩類，*num* 代表 $\gamma_q^{z^{MPE}}$ 爲正時的統計值資訊，而 *den* 則代表 $\gamma_q^{z^{MPE}}$ 爲負時的統計值資訊，詳細統計資訊可分別表示如下：

$$\gamma_{qm}^{num} = \sum_{z=1}^{Z} \sum_{q\in \mathbf{W}_{z,lattice}} \sum_{t=s_q}^{e_q} \gamma_{qm}^z(t) \max(0, \gamma_q^{z^{MPE}}) \quad (21)$$

$$\theta_{qmd}^{num}(O) = \sum_{z=1}^{Z} \sum_{q\in \mathbf{W}_{z,lattice}} \sum_{t=s_q}^{e_q} \gamma_{qm}^z(t) \max(0, \gamma_q^{z^{MPE}}) o_z(t) \quad (22)$$

$$\theta_{qmd}^{num}(O^2) = \sum_{z=1}^{Z} \sum_{q\in \mathbf{W}_{z,lattice}} \sum_{t=s_q}^{e_q} \gamma_{qm}^z(t) \max(0, \gamma_q^{z^{MPE}}) o_z(t)^2 \quad (23)$$

$$\gamma_{qmd}^{den} = \sum_{z=1}^{Z} \sum_{q\in \mathbf{W}_{z,lattice}} \sum_{t=s_q}^{e_q} \gamma_{qm}^z(t) \max(0, -\gamma_q^{z^{MPE}}) \quad (24)$$

$$\theta_{qmd}^{den}(O) = \sum_{z=1}^{Z} \sum_{q\in \mathbf{W}_{z,lattice}} \sum_{t=s_q}^{e_q} \gamma_{qm}^z(t) \max(0, -\gamma_q^{z^{MPE}}) o_z(t) \quad (25)$$

$$\theta_{qmd}^{den}(O^2) = \sum_{z=1}^{Z} \sum_{q\in \mathbf{W}_{z,lattice}} \sum_{t=s_q}^{e_q} \gamma_{qm}^z(t) \max(0, -\gamma_q^{z^{MPE}}) o_z(t)^2 \quad (26)$$

其中式(20)中的 D_{qmd} 是一個常數，需要用來確保每一維度的變異數必須要是正數，同時它也會影響收斂速度。一般而言， D_{qmd} 的值都設爲最小確保變異數爲正數的兩倍。另外，爲了要增加正確詞序列的對於模型參數訓練時的貢獻，可以引入所謂的 I-Smoothing 技術[12]，其公式如下：

$$\theta'^{num}_{qmd}(O) = \theta^{num}_{qmd}(O) + \frac{\tau_{qm}}{\gamma^{ML}_{qm}}\theta^{ML}_{qmd}(O)$$

$$\theta'^{num}_{qmd}(O^2) = \theta^{num}_{qmd}(O^2) + \frac{\tau_{qm}}{\gamma^{ML}_{qm}}\theta^{ML}_{qm}(O^2) \quad (27)$$

$$\gamma'^{num}_{qm} = \gamma^{num}_{qm} + \tau_{qm}$$

其中 γ^{ML}_{qmd} 、 $\theta^{ML}_{qmd}(O)$ 及 $\theta^{ML}_{qmd}(O^2)$ 爲使用傳統最大化相似度訓練法所求得的統計值資訊，τ_{qm} 爲訓練時設定之常數。最後，對於詞圖 $\mathbf{w}_{z,lattice}$ 上候選詞序列正確率可以以每一候選詞序列所組成音素的正確率加總來代表，原始最小化音素錯誤(MPE)聲學模型訓練的計算候選詞序列中每一個組成音素 q 正確率的公式是：

$$A(q) = \max_u \begin{Bmatrix} -1+2e(q,u) & \text{if } u \text{ and } q \text{ are same phone} \\ -1+e(q,u) & \text{if } u \text{ and } q \text{ are different phones} \end{Bmatrix} \quad (28)$$

其中 $e(q,u)$ 爲音素 q 與正確詞序列中音素 u 的重疊比例(根據正確音素 u 的長度)，如果 q 與 u 爲同一音素則套用計算式 $-1+2e(q,u)$ ，反之套用計算式 $-1+e(q,u)$ ，最後 $A(q)$ 取與所有重疊的正確詞序列中音素 u 計算式值最大者爲音素 q 正確率(介於-1 與 1 之間)，圖 1 爲計算原始音素正確率的一個範例。而 $\gamma^{z\,MPE}_{q}$ 則可以由前向後向演算法(Forward-Backward Algorithm)求得[12]。

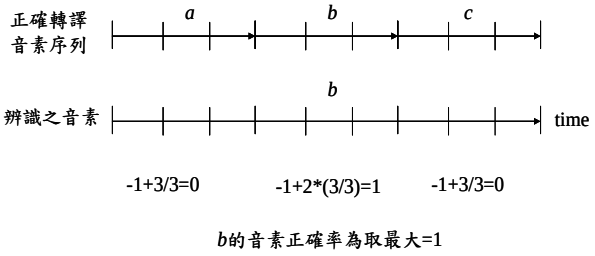

圖 1 最小化音素錯誤訓練之原始音素正確率函數的範例

四、最小化音素錯誤訓練之改進

最小化音素錯誤(MPE)訓練主要有兩個缺點:

1. 最小化音素錯誤中的原始音素正確率函數(Raw Phone Accuracy Function)並沒有給予刪除錯誤(Deletion Errors)適當的懲罰，只有對於插入錯誤(Insertion Errors)和取代錯誤(Substitution Errors)給予適當的懲罰。
2. 原始音素正確率函數是以音素爲單位(Phone-by-Phone)來做計算，如式(28)所示，且其值域範圍爲-1 到+1，這樣的範圍可能過於狹窄。每個音素段落(Phone Arc)所收集到的正確率統計值最大爲 1，因此遇到訓練語料不足的任務時，模型訓練所收集到的統計值會不夠強健。

爲了克服上述之問題，吾人提出了時間音框音素正確率(Time Frame Phone Accuracy Function, 記作 TFA)函數來取代原始音素正確率函數[13]：

$$TimeFrameAccuracy(q) = \frac{\sum_{t=s_q}^{e_q}\delta(q,u(t))}{e_q - s_q + 1} \quad (29)$$

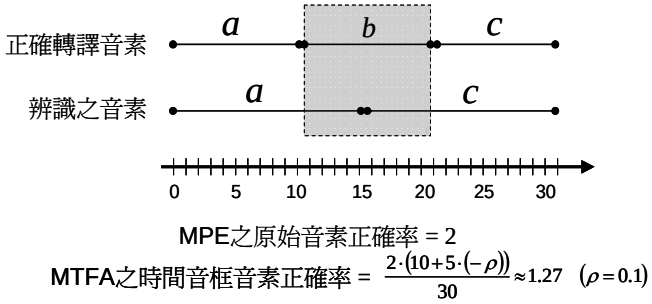

圖 2 最小化音素錯誤訓練之原始音素正確率及時間音框音素正確率對於刪除錯誤的影響

$$\delta(q,u(t))=\begin{Bmatrix}1 & ,if\ q=u(t)\\ -\rho & ,if\ q\neq u(t),0<\rho<1\end{Bmatrix} \tag{30}$$

其中 q 爲詞圖中某一音素段落，s_q 和 e_q 分別爲音素段落 q 的開始時間及結束時間，$u(t)$ 爲正確音素段落 u 在時間 t 時的音素標記(Phone Label)，ρ 爲刪除錯誤的懲罰權重(Deletion Penalty Weight)，用來懲罰某不完全正確音素段落 q 的正確率，因此某一音素段落在某個時間點 t 的正確率值域範圍爲介於 $-\rho$ 到 1 之間。時間音框音素正確率公式是看每一個音框的音素標記是否與正確音素標記一致來計算音素段落的正確率，因此對於一個完整的語句所對應的詞序列，就只要計算是否擊中(Hit)或取代(Substitution)，而不用考慮插入(Insertion)或刪除(Deletion)，因此在音素段落比對時比計算編輯距離(Edit or Levenshtein Distance)有效率，且時間音框音素正確率與我們要做評估的音素正確率有很大的正相關[5]，所以使用時間音框音素正確率的確可以去近似某個音素段落的音素正確率。圖 2 即爲計算時間音框音素正確率(TFA)的一個例子，假設某個語句有 30 個音框，此語句的正確轉譯音素共有三個，即 a、b 和 c；而此語句的辨識音素只有兩個，即 a 和 c，那麼 b 就是刪除錯誤。在圖 2 中灰色部份代表出現刪除錯誤，此刪除錯誤發生在第 11 個到第 20 個時間音框，我們理當給予這些錯誤的時間音框一些刪除錯誤的懲罰。而在詞圖中一整條路徑(詞序列) W_i 的時間音框音素正確率爲:

$$TimeFrameAcc(W_i)=\sum_{q\in W_i} TimeFrameAccuracy(q) \tag{31}$$

將式(31)取代式(28)，即本論文所提出的最大化時間音框音素正確率(Maximum Time Frame Phone Accuracy, 記作 MTFA)的目標函數:

$$\begin{aligned}F_{MTFA}(\lambda)&=\sum_{z=1}^{Z}\sum_{W_i\in\mathbf{W}_{z,lattice}} p(W_i\mid O_z)TimeFrameAcc(W_i)\\ &=\sum_{z=1}^{Z}\sum_{W_i\in\mathbf{W}_{z,lattice}}\frac{p_\lambda(O_z\mid W_i)P(W_i)}{p(O_z)}TimeFrameAcc(W_i)\end{aligned} \tag{32}$$

另外，爲了更充分地懲罰刪除錯誤且使其值域與原始音素正確率同爲介於-1 到 1 之間，本論文使用了 S 型函數(Sigmoid Function)來正規化時間音框音素正確率函數(式(29))的分子項，稱之爲 S 型時間音框音素正確率函數(Sigmoid Time Frame Phone Accuracy, 記作 STFA):

$$SigTimeFrameAccuracy(q)=\frac{2}{1+\exp(-\alpha\cdot net+\beta)}-1 \tag{33}$$

其中

$$net = \sum_{t=s_q}^{e_q} \delta(q, u(t)) \tag{34}$$

其$\delta(\cdot)$的定義同式(30)，α及β爲*S*型函數中可調整的參數，α控制*S*型函數的曲度，β則控制*S*型函數的平移。故式(33)的值域範圍介於-1 到+1 之間。而在詞圖中一整條路徑(詞序列)W_i的*S*型時間音框音素正確率爲:

$$SigTimeFrameAcc(W_i) = \sum_{q \in W_i} SigTimeFrameAccuracy(q) \tag{35}$$

將式(35)取代式(28)，則本論文所提出的最大化 *S* 型時間音框音素正確率(Maximum Sigmoid Time Frame Phone Accuracy, 記作 MSTFA)的目標函數爲:

$$\begin{aligned} F_{MTFA}(\lambda) &= \sum_{z=1}^{Z} \sum_{W_i \in \mathbf{W}_{z,lattice}} p(W_i \mid O_z) SigTimeFrameAcc(W_i) \\ &= \sum_{z=1}^{Z} \sum_{W_i \in \mathbf{W}_{z,lattice}} \frac{p_\lambda(O_z \mid W_i) P(W_i)}{p(O_z)} SigTimeFrameAcc(W_i) \end{aligned} \tag{36}$$

其本論文所提出的時間音框音素正確率函數主要並非去逼近編輯距離，而只是有考量給予刪除錯誤一些適當懲罰，以改進最小化音素錯誤(MPE)鑑別式聲學模型訓練。至於有關如何在詞圖中正確地計算編輯距離，可以參考[14]。

五、資料選取方法於改進最小化音素錯誤聲學模型訓練

近年來，由於最大邊際分類器(Large Margin Classifier)[15]在機器學習(Machine Learning)的領域中已有高度的發展，且在分類(Classification)任務中都已達到非常不錯的分類效果。其設計理念就在於提升分類器的一般化能力(Generalization Ability)，以致能夠在未知的測試樣本中達到較好的分類效果。在觀念上，我們以二元類別且可分離的(Separable)訓練樣本爲例，因訓練樣本通常與測試樣本會有不一致(Mismatch)的現象，要提升分類器的一般化能力，就要使得訓練樣本在某種定義域中(如相似度定義域(Likelihood Domain))離此定義域的決定邊界(Decision Boundary)越遠越好，訓練樣本到決定邊界的最近距離我們一般會稱之爲邊際(Margin)，而此邊際越大且邊際內沒有其他的訓練樣本代表一般化能力及容錯能力會越好[16]。

最大邊際估測法(Large Margin Estimation, LME)[17]是以相似度(Likelihood)爲基礎的分離邊際(Separation Margin)來選取距離決定邊界(Decision Boundary)較近的語音特徵向量序列，依其選取門檻(Threshold)，可以定義出支持向量集合(Support Vector Set)，再利用最大邊際估測法則進而調整聲學模型。對於那些不在支持向量集合裡的訓練樣本(訓練語句)，因爲距離決定邊界較遠，所以較不具鑑別力，因此就沒有拿來調整聲學模型的參數。所以我們可以視最大邊際估測法爲以相似度作爲選取準則的資料選取方法，選出較爲重要的語音特徵向量序列(訓練語句)。在柔性邊際估測法(Soft Margin Estimation, SME)[18]中，也是以相似度作爲選取準則，藉由定義不同的門檻值，進而選取出較有影響力的訓練語句，而且從選取出來的訓練語句中，更進一步地用類別比對(Label Matching)的方式選取出重要的時間音框(Frame)。所以我們也可以視柔性邊際估測法(SME)爲以相似度和類別比對爲基礎的進階資料選取方法。

最大邊際估測法與柔性邊際估測法所使用的資料選取方法都是在相似度定義域

(Likelihood Domain)中來執行資料的選取。在本論文中，吾人提出以熵值(Entropy)爲基礎的時間音框資料選取(Data Selection)方法來改進最小化音素錯誤聲學模型訓練。其中是以給定在某語音特徵向量序列(訓練語句)O_z，某個狀態中的某個高斯分布出現的事後機率(Posterior Probability，此事後機率有考慮到詞與詞之間的轉移機率，即語言模型機率)來求得熵值，再利用事先所設定的門檻值來選取資料，故可視爲在事後機率定義域中來取選資料。因其熵值的計算是在事後機率定義域(Posterior Domain)中，故有別於以相似度定義域爲基礎的傳統資料選取方法。然而傳統熵值的值域爲 0 到 $\log_2 N$，其中 N 爲參與熵值計算的樣本個數，但爲了方便決定門檻值進而選取時間音框，故在此我們使用正規化熵值(Normalized Entropy)來使其值域介於 0 到 1 之間，其公式如下：

$$E_z(t) = \frac{1}{\log_2 N} \sum_{q=1}^{Q} \sum_{m \in q} \gamma_{qm}^{z}(t) \cdot \log_2 \frac{1}{\gamma_{qm}^{z}(t)} \tag{37}$$

其中 $E_z(t)$ 爲在第 z 句訓練語句時間 t 時的正規化熵值，$\gamma_{qm}^{z}(t)$ 爲在第 z 句訓練語句時間 t 時，在音素段落 q 中之高斯模型 m 的事後機率，Q 爲在時間 t 時所有的音素段落個數，N 爲在時間 t 中所有事後機率不爲零的高斯模型 m 。

然而在資料選取方法中，資料(或樣本)可以定義在不同的單位上，以語音辨識爲例，訓練樣本(Training Sample)可以定義在語音特徵向量序列(訓練語句(Sentence or Utterance))、詞圖中的某詞段(Word Arc)、音素段落(Phone Arc)或時間音框(Frame)等。在語音辨識的任務中，鑑別式訓練收集統計值時是以時間音框(Frame)爲最小單位，所以本論文將著重在時間音框之選取(Frame Selection)，並將每一個時間音框視爲一個訓練樣本(Training Sample)。鑑別式訓練時是將所有的時間音框所收集到的統計值都用來調整模型的參數，事實上有些時間音框對於鑑別式訓練是沒有幫助的，例如那些已經可以被分類器(在語音辨識中，通常使用連續密度隱藏式馬可夫模型(CDHMM)來當成分類器)很正確分類或很錯誤分類的時間音框，故本論文提出的以熵值(Entropy)爲基礎的時間音框選取方法(Frame Selection)旨在找出哪些時間音框是會被很正確或很錯誤地分類，哪些是不容易被分類正確，進而丟棄那些被很正確分類和被很錯誤分類之時間音框所收集到的統計值，且只利用這些被收集到的統計值來調整模型參數，以幫助鑑別式聲學模型訓練。因此使用此資料選取方法可以適用於所有的鑑別式聲學模型訓練，不僅能夠保持鑑別式訓練最小化訓練樣本分類錯誤率，還可以增進分類器的一般化能力。以最小化音素錯誤(MPE)訓練的統計值收集爲例，每個時間音框要先算正規化熵值，再由其門檻值決定是否累加統計值，則其數學式可表示爲(以 *num* 類爲例):

$$\begin{aligned}
\gamma_{qm}^{num} &= \sum_{z=1}^{Z} \sum_{q \in \mathbf{W}_{z,lattice}} \sum_{t=s_q}^{e_q} \left[\gamma_{qm}^{z}(t) \max(0, \gamma_{q}^{z\,MPE}) \right] \cdot I(E_z(t) > \rho) \\
\theta_{qmd}^{num}(O) &= \sum_{z=1}^{Z} \sum_{q \in \mathbf{W}_{z,lattice}} \sum_{t=s_q}^{e_q} \left[\gamma_{qm}^{z}(t) \max(0, \gamma_{q}^{z\,MPE}) o_z(t) \right] \cdot I(E_z(t) > \rho) \\
\theta_{qmd}^{num}(O^2) &= \sum_{z=1}^{Z} \sum_{q \in \mathbf{W}_{z,lattice}} \sum_{t=s_q}^{e_q} \left[\gamma_{qm}^{z}(t) \max(0, \gamma_{q}^{z\,MPE}) o_z(t)^2 \right] \cdot I(E_z(t) > \rho)
\end{aligned} \tag{38}$$

其中 ρ 爲事先定義的門檻值(Threshold)，其值介於 0 到 1 之間， $I(E_z(t) > \rho)$ 可表示爲:

$$I(E_z(t) > \rho) = \begin{cases} 1, \ if \ E_z(t) > \rho \\ 0, \ if \ E_z(t) \le \rho \end{cases} \tag{39}$$

式(39)使用的是指示函數，其值非 0 即 1，故我們可將它視爲是一種硬性選取(Hard Selection)的資料選取方法。另一方面，我們亦可將每個時間音框所計算出的正規化熵值作爲權重(Weight)，用來強調(Emphasized)或非強調(Deemphasized)此時間音框的重要性，我們將此方法視爲另一種柔性選取(Soft Selection)的資料選取方法，其數學式如下所示:

$$\gamma_{qm}^{z}(t) = \gamma_{qm}^{z}(t)\left(1+\omega \cdot E_{z}(t)\right) \tag{40}$$

其中 ω 爲一比例控制參數。

六、實驗與討論

（一）實驗架構與設定

本論文所使用的大詞彙連續語音辨識器爲臺灣師大目前所發展的新聞語音辨識系統[19]，主要包括前端處理、詞彙樹複製搜尋(Tree-Copy Search)及詞圖搜尋(Word Graph Rescoring)[11]等部分。

在前端處理方面，本論文所採用的是異質性線性鑑別分析(Heteroscedastic Linear Discriminant Analysis, HLDA)[20]。且在做完鑑別分析之後還額外使用最大化相似度現性轉換(Maximum Likelihood Linear Transform, MLLT)[21]，其目的是爲了配合目前我們在連續密度隱藏式馬可夫模型所使用的對角化(Diagonal)之共變異矩陣。同時，爲了降低通道效應對語音辨識的影響，在此使用倒頻譜正規化法(Cepstral Normalization, CN)。

在聲學模型方面，我們採用 151 個連續密度隱藏式馬可夫模型作爲中文 INITIAL-FINAL 的統計模型，而每個模型的狀態數分別爲 3 至 6 個不等，每個狀態皆爲高斯混合分布，其中每個高斯混合分布的分布個數分別爲 1 至 128 個不等，本論文總共使用到約 14,396 個高斯分佈。另一方面，本論文所使用的詞典約含有七萬二千個一至十字詞，並以從中央通訊社(Central News Agency, CNA) 2001 與 2002 年所收集到的約一億七千萬(170M)個中文字語料作爲背景語言模型訓練時的訓練資料[22]。在本文中的語言模型使用了 Katz 語言模型平滑技術[23]，在訓練時是採用 SRL Language Modeling Toolkit (SRILM)[24]。在詞彙樹搜尋時，本系統採用詞雙連語言模型；在詞圖搜尋時，則採用詞三連語言模型。

（二）實驗語料

本論文實驗使用的訓練與測試語料爲 MATBN 電視新聞語料庫[25]，是由中央研究院資訊所口語小組[26]耗時三年與公共電視台[27]合作錄製完成。我們初步地選擇採訪記者語料作爲實驗語材，其中包含 25.5 小時的訓練集(5,774 句)，供聲學模型訓練之用，其中男女語料各半；1.5 小時的評估集(292 句，共 26,219 字)，供辨識評估之用。訓練集由 2001 及 2002 年的新聞語料所篩選出來的；評估集則均爲 2003 年的語料，由中研院的評估語料篩選出來，只選擇了採訪記者語料並濾掉了含有語助詞之語句。

（三）實驗評估方式

本論文採用美國國家標準與技術中心(National Institute of Standards and Technology, NIST)所訂立的評估標準來進行正確轉譯詞序列與辨識詞序列的比較。此評估標準需

要使用動態規畫(Dynamic Programming)來做詞序列比對。然而因在中文中存在著斷詞不一致的問題，故在本文的實驗中皆是以字爲比對單位。令 H 爲正確轉譯詞序列與辨識詞序列比對後相同(Match)的字元個數、I 爲辨識詞序列多餘插入(Insertion)的字元個數、 N 爲正確轉譯詞序列的字元總數，則語音辨識系統之正確率(Accuracy)的計算方式爲 $\frac{H-I}{N}\times 100\%$ ，錯誤率(Error Rate)則爲 1-正確率。本文的實驗數據中，皆是以字錯誤率(Character Error Rate, CER)來呈現實驗結果。

（四）基礎實驗結果

於基礎實驗中，先利用最大化相似度(ML)估測法訓練 10 次，所得到的字錯誤率(CER)爲 23.64% (記作 Baseline)。接著進行最小化音素錯誤(MPE)訓練 10 次，最後所得到的字錯誤率(CER)爲 20.77%。故於接下來的實驗中，皆以這組實驗爲比較對象。

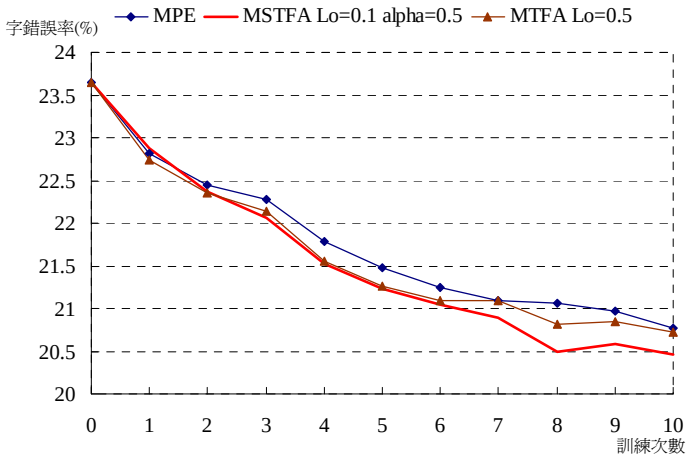

圖 3 時間音框正確率函數與最小化音素錯誤之比較結果

（五）改進最小化音素錯誤之實驗結果

本小節呈現本論文針對最小化音素錯誤訓練(MPE)的缺點而改進的最大化時間音框正確率函數(MTFA)訓練之實驗數據。事實上，本論文所提出的時間音框正確率函數並不是要減少刪除錯誤的個數(但實際上於共 26,219 字的評估集中，MPE 在第 10 次迭代訓練上有 359 個刪除錯誤，然而 MSTFA 則稍微減爲成 345 個)，而是去考量詞圖中某詞段受到刪除錯誤的影響，而減少其收集到的正確率統計值，以利聲學模型訓練之強健。實驗結果可參考表 1，其中刪除錯誤的懲罰權重(Penalty Weight)以 Lo(ρ)表示。由實驗數據顯示得知，在前幾次的迭代訓練中，時間音框正確率函數都會稍比最小化音素錯誤來得好。不同的刪除錯誤懲罰權重設定會有不同的刪除錯誤懲罰之效果。由表 1 的數據顯示，太大(ρ =0.8)或太小(ρ =0.1)的刪除錯誤懲罰權重設定，在一開始的迭代訓練上並無法得到明顯的效果，但都比最小化音素錯誤訓練來得佳，不過在第 10 次的迭代訓練上卻比最小化音素錯誤訓練的字錯誤率要高一點。最好的刪除錯誤懲罰權重設定(Lo=0.5)的時間音框正確率函數在第 10 次的迭代訓練上比最小化音素錯誤(MPE)訓練的辨識字錯誤率好 0.05%，相對字錯誤率將低約 0.1%，訓練次數 1 到 10 次的字錯誤率曲線圖請參考圖 3。

本論文亦使用一個常見的 S 型函數來平滑時間音框正確率函數，記作 MSTFA。其中 S 型函數有兩個參數可調整，在本實驗中只調整 α (alpha)，而 β 設爲零($\beta = 0$)。

實驗結果同樣參考表 1，實驗數據顯示出 MSTFA 在每次迭代訓練中都會比最小化音素錯誤(MPE)訓練來得好，最好的設定($\rho = 0.1$，$\alpha = 0.5$)在第 10 次的迭代訓練上可以比最小化音素錯誤的辨識字錯誤率好 0.31%，相對字錯誤率降低約 1.5%。

表 1 時間音框正確率函數之實驗結果

CER(%)	MPE	MTFA ρ =0.1	MTFA ρ =0.3	MTFA ρ =0.5	MTFA ρ =0.8	MSTFA ρ =0.1 $\alpha = 0.5$	MSTFA ρ =0.5 $\alpha = 0.5$	MSTFA ρ =0.1 $\alpha = 1$	MSTFA ρ =0.5 $\alpha = 1$
Baseline	23.64								
Itr01	22.82	22.85	22.73	22.74	22.80	22.88	22.82	22.83	22.77
Itr02	22.44	22.35	22.33	22.36	22.39	22.37	22.34	22.37	22.38
Itr03	22.28	22.07	22.13	22.14	22.19	22.06	22.10	22.02	22.05
Itr04	21.79	21.65	21.50	21.56	21.69	21.52	21.58	21.41	21.56
Itr05	21.48	21.26	21.14	21.26	21.34	21.23	21.47	21.3	21.52
Itr06	21.24	20.98	20.97	21.09	21.23	21.05	21.27	21.06	21.32
Itr07	21.10	20.91	20.87	21.09	21.19	20.89	21.11	20.80	21.19
Itr08	21.06	20.87	20.81	20.82	20.93	20.50	20.97	20.54	20.98
Itr09	20.97	20.84	20.74	20.85	20.90	20.58	20.82	20.57	21.03
Itr10	20.77	20.82	20.80	20.72	20.93	20.46	20.87	20.65	21.10

（六）資料選取方法之實驗結果

表 2 資料選取方法之實驗結果

CER(%)	MPE	MPE Random	MPE HS Thr=0.05	MPE HS Thr=0.08	MPE SS ω =1.0	MPE SS ω =0.5
Baseline	23.64					
Itr01	22.82	23.02	22.63	22.43	22.84	22.88
Itr02	22.44	22.62	22.05	21.80	22.40	22.43
Itr03	22.28	22.22	21.60	21.45	22.21	22.25
Itr04	21.79	22.16	21.40	21.34	21.65	21.73
Itr05	21.48	21.76	21.19	20.94	21.34	21.31
Itr06	21.24	21.66	20.92	20.82	21.33	21.18
Itr07	21.10	21.74	20.91	20.73	21.29	21.29
Itr08	21.06	21.62	21.22	20.74	21.00	21.06
Itr09	20.97	21.78	21.08	20.65	21.02	20.93
Itr10	20.77	21.84	21.29	20.63	20.94	20.89

本小節呈現資料選取方法於最小化音素錯誤(MPE)訓練之實驗結果。其中最小化音素錯誤的 I-平滑技術參數設定爲 10[28]。所使用的時間音框資料選取方法分爲硬性選取(HS)和軟性選取(SS)，其實驗結果皆可參考表 2。在軟性選取部分，所得到的結果跟基礎實驗結果不相上下，而其硬性選取最佳門檻值(記作 Thr)之設定爲 0.05(其時間音框總數爲 4,214,360 個，佔所有時間音框總數的 45.88%)。如實驗數據所顯示，資料選

取方法應用在最小化音素錯誤(MPE)訓練確實可以加快收斂速度，但在第 10 次的訓練上沒有明顯比最小化音素錯誤的字錯誤率來得低，其效果是差不多的。特別注意的是在 Thr=0.08 的這組實驗中，我們嘗試把門檻值隨著迭代訓練次數而遞減以企圖避免過度訓練之現象，所得結果亦符合我們所期望。故可以得知資料選取方法具有加快收斂速度之能力同時在第 10 次迭代訓練上與最小化音素錯誤訓練擁有差不多之結果。同時更說明了以正規化熵值爲基礎的資料選取方法確實能選出在事後機率定義域中離決定邊界較近的時間音框樣本，受惠於這些時間音框樣本本身比較具有鑑別力，故資料選取方法對於鑑別式訓練特別有幫助。

另一方面，吾人使用隨機選取(Random Selection)方法進行比較驗證以正規化熵值爲基礎的資料選取方法的確有效用的，而非亂選。實驗結果同樣參考表 2，其中隨機選取(記作 MPE Random)方法在每一次的迭代都隨機選取所有時間音框總數的 45.88%(與 MPE HS Thr=0.05 這組實驗的時間音框總數一致)。

（七）資料選取方法結合時間音框正確率函數之實驗結果

最後本小節將呈現資料選取方法於最大化 *S* 型時間音框正確率函數(MSTFA)之實驗結果。其最大化 *S* 型時間音框正確率函數的 I-平滑技術參數最佳化設定爲 10。所使用的時間音框資料選取方法爲硬性選取(HS)、軟性選取(SS)以及結合硬性和軟性選取(HS+SS)，實驗結果可參考表 3。其中最大化 *S* 型時間音框正確率函數的參數設定爲 $\rho = 0.1$、$\alpha = 0.5$；而各資料選取方法之參數設定皆列於表 3 中。由數據顯示得知，資料選取方法應用在最大化 *S* 型時間音框正確率函數依然保有加快收斂速度之成效，但在第 10 次迭代訓練上卻沒能比最大化 *S* 型時間音框正確率函數的字錯誤率來得低。此外，軟性資料選取方法比硬性選取效果來得好，在第 10 次迭代訓練上跟最大化 *S* 型時間音框正確率函數的字錯誤率差不多。最後，結合硬性與軟性選取(HS+SS)的實驗結果卻並沒有達到我們所預期的加成性效果。

表 3 資料選取方法結合時間音框正確率函數之實驗結果

CER(%)	MPE	MSTFA ρ =0.1 α = 0.5	MSTFA HS Thr=0.05	MSTFA SS ω =1.0	MSTFA HS+SS Thr=0.1 ω =0.5
Baseline	23.64				
Itr01	22.82	22.88	22.46	22.75	22.53
Itr02	22.44	22.37	21.87	22.25	21.72
Itr03	22.28	22.06	21.40	21.83	21.45
Itr04	21.79	21.52	21.38	21.45	21.38
Itr05	21.48	21.23	21.08	21.27	21.03
Itr06	21.24	21.05	21.03	20.94	20.90
Itr07	21.10	20.89	21.02	20.65	21.14
Itr08	21.06	20.50	21.15	20.78	21.14
Itr09	20.97	20.58	20.86	20.56	21.07
Itr10	20.77	20.46	21.43	20.86	21.37

七、結論與未來展望

鑑別式聲學模型訓練在大詞彙連續語音辨識的研究上一直扮演著重要的角色。本論文旨在改善最小化音素錯誤之聲學模型訓練，相關研究內容與成果可從下面兩個面向來作探討:

(1)首先，本論文提出了新的時間音框正確率函數來取代最小化音素錯誤訓練的原始音素正確率函數，進而充分地給予刪除錯誤適當的懲罰。在實驗結果上，最大化 *S* 型時間音框正確率函數(MSTFA)能達到比最小化音素錯誤(MPE)訓練約有1.5%的相對字錯誤率降低。

(2)其次，本論文提出以正規化熵值爲基礎之新的資料選取方法來改善鑑別式聲學模型訓練，由於正規化熵值是以給定某訓練語句的語音特徵向量序列中，某個狀態中的某個高斯分布出現的事後機率來求得的，所以可以視爲是在事後機率定義域中來選取訓練樣本，且所選出來的訓練樣本是比較混淆的，對鑑別式訓練來說，這些混淆的訓練樣本是較具有鑑別力的。根據初步的實驗結果顯示，此資料選取方法可以加快收斂速度，在前幾次的迭代訓練中，比最小化音素錯誤訓練有很大且一致的字錯誤率降低。最好的結果在第 6 次的迭代訓練上，比最小化音素錯誤訓練約有 1.5%的相對字錯誤率降低。

以全面風險爲基礎的鑑別式聲學模型訓練中，減損函數的設計一直都是一個重要的議題，如最流行的最小化音素錯誤訓練目標函數中，以類別比對爲基礎的原始音素正確率函數就還存在著改進的空間。已有學者提出以聲學模型間的關係來計算正確率以取代以類別爲基礎的正確率函數。類別比對爲基礎的音素正確率函數和聲學模型彼此間關係的減損函數皆各有其優缺點，未來吾人想要嘗試將這兩種不同的資訊結合，企圖改進鑑別式聲學模型訓練。

同時，吾人未來也想嘗試將以正規化熵值爲基礎的資料選取方法應用到其他的鑑別式訓練，如最小化分類錯誤、最小化貝氏風險鑑別式訓練等，以驗證此方法的一般性。事實上，由加最小化音素錯誤訓練的收斂速度來看，此以正規化熵值爲基礎之新的資料選取方法的確爲鑑別式聲學模型訓練提供了一個新的方向。

參考文獻

[1] R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification, Second Edition*. New York: John & Wiley, 2000.

[2] Lalit R. Bahl, F. Jelinek and Robert L. Mercer, *A Maximum Likelihood Approach to Continuous Speech Recognition*, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. PAMI-5, no.2, March 1983.

[3] J. Fiscus, *A Post-processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER)*, in Proc. ASRU 1997.

[4] V. Goel and W. Byrne, *Minimum Bayes-Risk Automatic Speech Recognition*, Computer Speech and Language, Vol. 14, pp.115-135, 2000.

[5] F. Wessel, R. Schluter, K. Macherey and H. Ney, *Explicit Word Error Minimization Using Word Hypothesis Posterior Probability*, in Proc. ICASSP 2001.

[6] L. Mangu, E. Brill and A. Stolcke, *Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Networks*, Computer Speech and Language, Vol. 14, pp.373-400, 2000.

[7] Y. Normandin, *Hidden Markov Models, Maximum Mutual Information Estimation, and the Speech Recognition Problem*, Ph.D Dissertation, McGill University, Montreal, 1991.

[8] J. Kaiser, B. Horvat and Z. Kacic, *Overall Risk Criterion Estimation of Hidden Markov Model Parameters*, Speech Communication, Vol. 38, pp.383-398, 2002.

[9] V. Doumpiotis, S. Tsakalidis and W. Byrne, *Lattice Segmentation and Minimum Bayes Risk Discriminative Training*, in Proc. Eurospeech 2004.

[10] D. Povey and P. C. Woodland, *Minimum Phone Error and I-smoothing for Improved Discriminative Training*, in Proc. ICASSP 2002.

[11] S. Ortmanns, H. Ney and X Aubert, *A Word Graph Algorithm for Large Vocabulary Continuous Speech Recognition*, Computer Speech and Language, Vol. 11, pp.11-72, 1997.

[12]D. Povey, *Discriminative Training for Large Vocabulary Speech Recognition*. Ph.D Dissertation, Peterhouse, University of Cambridge, July 2004.

[13]Shih-Hung Liu, Fang-Hui Chu, Shih-Hsiang Lin and Berlin Chen, *Investigating Data Selection for Minimum Phone Error Training of Acoustic Models*, in Proc. ICME 2007.

[14]G. Heigold *et al*, *Minimum Exact Word Error Training*, in Proc. ASRU 2005

[15]A. J. Smola, P. Bartlett, B. Scholkopf and D. Schuurmans, *Advances in Large Margin Classifiers*, The MIT Press.

[16]V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.

[17]Xinwei Li, Hui Jiang and Chaojun Liu, *Large Margin HMMs for Speech Recognition*, in Proc. ICASSP 2005.

[18] Jinyu Li, Ming Yuan and Chin-Hui Lee, *Soft Margin Estimation of Hidden Markov Model Parameters*, in Proc. ICSLP 2006.

[19]B. Chen, J. W. Kuo and W. H. Tsai, *Lightly Supervised and Data-Driven Approaches to Mandarin Broadcast News Transcription*, in Proc. ICASSP 2004.

[20] N. Kumar, *Investigation of Silicon-Auditory Models and Generalizaion of Linar Discriminant Analysis for Improved Speech Recognition*, Ph.D. Thesis, John Hopkins University, Baltimore, 1997.

[21] R. A. Gopinath, *Maximum Likelihood Modeling with Gaussian Distributions*, in Proc. of ICASSP 1998.

[22] LDC: Linguistic Data Consortium, *http://www.ldc.upenn.edu*

[23] S. M. Katz, *Estimation of Probabilities from Sparse Data for Other Language Component of a Speech Recognizer*, IEEE Trans. Acoustics, Speech and Signal Processing, Vol. 35, No.3, pp. 400-401, 1987.

[24]A. Stolcke, *SRI language Modeling Toolkit*, version 1.3.3, *http://www.speech.sri.com/projects/srilm/*

[25] H. M. Wang, B. Chen, J.-W. Kuo, and S.S. Cheng. *MATBN: A Mandarin Chinese Broadcast News Corpus*, International Journal of Computational Linguistics and Chinese Language Processing, Vol. 10, No. 2, pp. 219-236, 2005.

[26] SLG: Spoken Language Group at Chinese Information Processing Laboratory, Institute of Information Science, Academia Sinica. *http://sovideo.iis.sinica.edu.tw/SLG/index.htm*

[27] PTS: Public Television Service Foundation. *http://www.pts.org.tw*

[28]郭人瑋, *最小化音素錯誤鑑別式聲學模型學習於中文大詞彙連續語音辨識之初步研究*, Master Thesis, NTNU, 2005.

[29]劉士弘, *改善鑑別式聲學模型訓練於中文連續語音辨識之研究*, Master Thesis, NTNU, 2007.