風險最小化準則在中文大詞彙連續語音辨識之研究

郭人瑋　劉士弘　陳柏琳
國立台灣師範大學資訊工程研究所
{rogerkuo, g93470185, berlin}@csie.ntnu.edu.tw

摘要

本論文探討風險最小化(Risk Minimization)準則在中文大詞彙連續語音辨識(Large Vocabulary Continuous Speech Recognition, LVCSR)之初步研究，內容包括了聲學模型訓練、非監督式聲學模型調適與搜尋演算法等方面。本論文以公視電視新聞語料庫作爲中文廣播新聞實驗題材。在聲學模模型訓練方面，我們使用了*最小化音素錯誤*(Minimum Phone Error, MPE)鑑別式訓練方法；實驗結果顯示，*最小化音素錯誤*訓練能較傳統最大相似度訓練相對地降低約 12%的字錯誤率。另一方面，在聲學模型調適上，我們則探討*最小化音素錯誤線性迴歸*(Minimum Phone Error Linear Regression, MPELR)調適法在非監督式聲學模型調適的使用；實驗結果顯示，*最小化音素錯誤線性迴歸*可以再進一步相對地降低約 5%的字錯誤率。最後，在搜尋演算法方面，本文探討詞錯誤最小化(Word Error Minimization, WEM)搜尋方法；實驗結果初步顯示，詞錯誤最小化搜尋方法較傳統最大化事後機率解碼方法來的稍佳。

1. 序論

隨著科技的快速演進，電腦早已融入每個家庭日常生活之中，而消費性的電子產品，更是改變了傳統的生活方式。然而，爲了攜帶上的便利，科技產品也愈做愈小，卻換來了輸入上的不便，人與機器的溝通，需要更簡便的方式才行。語音是人跟人之間最自然的溝通方式，自然地，我們也會希望人與機器之間最自然的溝通就是透過語音交談，因此自動語音辨識的研究也變得更加重要，特別是針對中文的輸入不便。另一方面，由於多媒體影音資訊迅速累積，例如廣播電視節目、語音信件、演講錄影和數位典藏等，這些多媒體資訊可以從網路上大量地取得，已成爲傳統文字資訊外社會大眾廣泛使用的資訊來源。是顯而易見的是，在上述的絕大部分多媒體資訊中，語音可以說是最具語意的主要內涵之一，當播出放多媒體的語音資訊或是顯示出對應的正確轉寫資訊，我們就可以大概地瞭解其中的主題或概念。因此，語音辨識技術對多媒體資訊的處理也扮演著相當重要的角色。

語音辨識可視爲一個分類的過程，在[1]中介紹了以最小化(分類)錯誤率(Minimum Error Rate, MER)來作爲分類的法則。在傳統的架構中，均以零壹函數(Zero-One Function)作爲減損函數(Loss Function)，藉此冀求分類過程能達到最小化分類錯誤。但在語音辨識中，每一文句代表一個類別，使用零壹函數作爲減損函數，可以最小化句錯誤率(Sentence Error Rate, SER)。但語音辨識在進行效能評估時，通常會以詞錯誤率(Word Error Rate, WER)，或在中文以字錯誤率(Character Error Rate, CER)作爲評估標準，使得最小化錯誤率法則與語音辨識實際上的評估方式產生了相當大的差異。換句話說，句錯誤愈小，不一定帶來較少的詞錯誤；而詞錯誤愈少，也不一定會有最少的

句錯誤。爲了克服此問題，近年來最常見的作法是以編輯距離(Levenshtein Edit Distance)[2]來取代零壹函數作爲減損函數，不論是在模型的訓練或是在搜尋演算法上，均有不錯的成果。本論文將在此架構下，探討風險最小化(Risk Minimization)準則在中文大詞彙連續語音辨識(Large Vocabulary Continuous Speech Recognition, LVCSR)之初步研究，內容包括了聲學模型訓練、非監督式聲學模型調適與搜尋演算法等方面。在聲學模模型訓練方面，我們使用了*最小化音素錯誤*(Minimum Phone Error, MPE)鑑別式訓練方法；在聲學模型調適上，我們則探討*最小化音素錯誤線性迴歸*(Minimum Phone Error Linear Regression, MPELR)調適法在非監督式聲學模型調適的使用；最後，在搜尋演算法方面，我們探討詞錯誤最小化(Word Error Minimization, WEM)搜尋方法。在以公視電視新聞語料庫作爲中文廣播新聞實驗題材下，初步驗證了上述這些方法在中文大詞彙連續語音辨識上均有不錯的成效。

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*爲基礎的線性轉換調適技術；第五章則探討詞錯誤最小化搜尋演算法；第六章爲實驗與討論；第七章爲結論與未來展望。

2. 貝氏風險與全面風險

若 O_r 爲一語句的聲學特徵向量序列，將 O_r 歸類至文句 s 時，可以用函數 $R(s \mid O_r)$ 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文句。將 O_r 歸類至 s 的風險 $R(s \mid O_r)$ 可定義如下[1]：

$$R(s \mid O_r) = \sum_{u \in \mathbf{W}_h} P(u \mid O_r) L(s,u), \tag{1}$$

其中 $\mathbf{W}_h$ 爲聲學特徵向量序列 O_r 所有可能對應的文句所成之集合； $P(u \mid O_r)$ 表示給定 O_r 時，文句 u 的事後機率(Posterior Probability)； $L(s,u)$ 爲一減損函數(Loss Function)，用以表示文句 s 與 u 之間差異所造成的損失(Loss)， $R(s \mid O_r)$ 爲將 O_r 歸類至 s 時的期望損失(Expected Loss)，又稱爲條件風險(Conditional Risk)。在語音辨識解碼上，可以最小化此條件風險來尋找最佳的文句 s^* ：

$$s^* = \arg\min_s R(s \mid O_r) = \arg\min_s \sum_{u \in \mathbf{W}_h} P(u \mid O_r) L(s,u), \tag{2}$$

而因此產生的風險即爲貝氏風險(Bayes Risk) R_{Bayes} [1]：

$$R_{Bayes} = \min_s R(s \mid O_r) = \min_s \sum_{u \in \mathbf{W}_h} P(u \mid O_r) L(s,u). \tag{3}$$

目前有許多語音辨識器根據貝氏決策定理(Bayesian Decision Rule)，即最小化此條件風險前提下來設計其搜尋演算法，如傳統的最大化事後機率(Maximum a Posteriori, MAP)解碼方法[3]、ROVER(Recognizer Output Voting Error Reduction)[4]、最小化貝氏風險(Minimum Bayes Risk, MBR)[5]及詞錯誤最小化(Word Error Minimization, WEM)[6] 等。然而，就模型訓練而言，則需

要最小化全面風險(Overall Risk) $R_{Overall}$ [1]：

$$R_{Overall} = \int R(s_r \mid O_r) P(O_r) dO_r, \tag{4}$$

其中 s_r 爲 O_r 對應之正確轉譯文句(Correct Transcription)， $P(O_r)$ 爲 O_r 的事前機率(Prior Probability)；全面風險 $R_{Overall}$ 是在語句空間上作積分，爲所有訓練語句的期望條件風險(Expected Conditional Risk)。由於訓練語料有限，故全面風險可簡化：

$$R_{Overall} = \sum_r R(s_r \mid O_r) P(O_r) = \sum_r \sum_{u \in \mathbf{W}_h} P(u \mid O_r) L(s_r, u) P(O_r). \tag{5}$$

若事後機率分佈 $P(u \mid O_r)$ 由聲學模型(Acoustic Model) λ 及語言模型(Language Model) Γ 所決定，記作 $P_{\lambda,\Gamma}(u \mid O_r)$ ，則全面風險可改寫成：

$$R_{Overall} = \sum_r \sum_{u \in \mathbf{W}_h} P_{\lambda,\Gamma}(u \mid O_r) L(s_r, u) P(O_r). \tag{6}$$

若再假設 $P(O_r)$ 對所有聲學特徵向量序列 O_r 均有一致(Uniform)的機率，且此項與模型參數 λ 及 Γ 的訓練無關，則可將此項省略：

$$R_{Overall} \approx \sum_r R(s_r \mid O_r) = \sum_r \sum_{u \in \mathbf{W}_h} P_{\lambda,\Gamma}(u \mid O_r) L(s_r, u). \tag{7}$$

在估測模型時，希望估測之模型 (λ, Γ) 能將全面風險降至最低：

$$(\lambda, \Gamma) = \arg\min_{\lambda', \Gamma'} \sum_r \sum_{u \in \mathbf{W}_h} P_{\lambda', \Gamma'}(u \mid O_r) L(s_r, u). \tag{8}$$

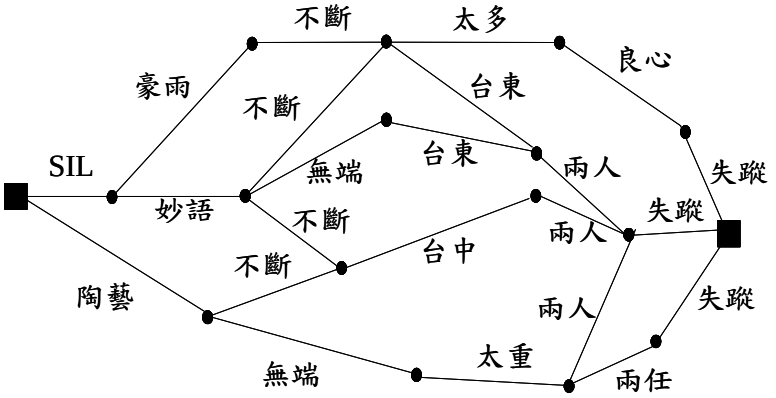
3. 最小化音素錯誤訓練

在估測模型時，我們希望估測之模型 (λ, Γ) 能將全面風險降至最低，式(8)因此可進一步表示成：

$$(\lambda, \Gamma) = \arg\min_{\lambda', \Gamma'} \sum_r \sum_{u \in \mathbf{W}_h} \frac{p_{\lambda'}(O_r \mid u) P_{\Gamma'}(u)}{\sum_{v \in \mathbf{W}_h} p_{\lambda'}(O_r \mid v) P_{\Gamma'}(v)} L(u, s_r), \tag{9}$$

其中 $p_{\lambda'}(O_r \mid u)$ 爲文句 u 對應的聲學模型產生聲學特徵向量序列 O_r 的機率分佈，如連續密度隱藏式馬可夫模型(Continuous Density HMM, CDHMM)； $P_{\Gamma'}(u)$ 爲文句 u 對應的語言模型機率分佈，如詞 *n*-連(詞雙連、詞三連)語言模型。

*全面風險法則估測*首先由 Na 等人所提出[7]，初步地使用零壹減損函數，來最小化訓練語料中的貝氏風險(Bayes Risk)，在獨立數字辨識(Isolated Digit Recognition)上可降低不少的錯誤率。Kaiser 等人在 2000 年時則以 Levenshtein 距離來取代零壹函數，以 *N*-最佳序列(*N*-Best List)作爲所有可能文句之近似，並使用延伸波氏重估(Extended Baum-Welch Re-estimation, EBW)演算法來爲模型參數進行最佳化[8][9]。



圖一、詞圖爲所有可能文句 $\mathbf{W}_h$ 的近似。

*最小化音素錯誤*訓練法則[10][11]類似全面風險法則的概念，同樣也是以最大化所有文句的期望辨識率爲目標。但*最小化音素錯誤*與*全面風險法則估測*有下列的差異：

1. 使用詞圖(Word Graph)來取代 *N*-最佳序列(*N*-best List)作爲所有可能文句之近似，如圖一所示。
2. 引入模型參數事前機率，來增加估測值的強健性。
3. 對於延伸波氏重估演算法中的控制參數，提出更佳的設定方式。
4. 強調音素層次的正確率而非詞正確率。

*最小化音素錯誤*訓練法則聲學模型訓練的目標函數 $F_{MPE}(\lambda)$ 爲：

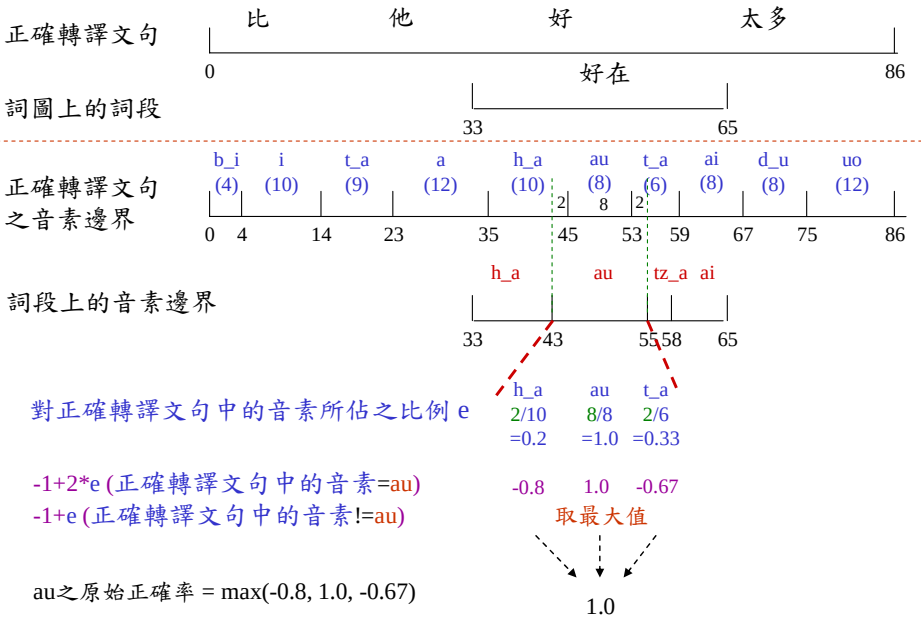
$$F_{MPE}(\lambda)=\sum_r \sum_{u\in \mathbf{W}_h} \frac{p_\lambda(O_r\mid u)P(s)}{\sum_{v\in \mathbf{W}_h} p_\lambda(O_r\mid v)P(v)} A(u,s_r). \tag{10}$$

在實作上，由於不可能對聲學特徵向量序列 O_r 所有可能對應的文句 $\mathbf{W}_h$ 作窮舉，與*全面風險法則估測*以 *N*-最佳路徑作爲所有可能文句之近似不同的是，*最小化音素錯誤*是以 $\mathbf{W}_{lat}^r$ 來近似，它是以第 r 句訓練語句辨識過後所產生的詞圖並加入正確轉譯文句 s_r 的詞分枝所形成之可能文句集合。另一方面， $A(v,s_r)$ 爲文句 v 相對於正確轉譯文句 s_r 的正確率，由於傳統以全域比對(Global Matching)結合編輯距離(Levenshtein Edit Distance)計算正確率並沒有考慮到時間上的相關性，無法提供充分正確的資訊供聲學模型訓練使用。因此 Povey 等人於 2002 年時對*最小音素錯誤訓練*提出一套音素之間計算正確率的方式[10]。如圖二所示，正確轉譯文句 s_r 爲「比-他-好-太多」，而以詞圖中某一詞段「好在」爲例，要計算詞圖上某一音素的正確率有三個步驟(在此以辨識文句 *v* 中的音素 au 爲例)：

Step 1. 在正確轉譯文句中找出與 au 有時間重疊之音素 h_a、au 與 t_a，分別的重疊長度爲 2、8、2 個音框(Frame)。

Step 2. 計算辨識文句中 au 對此三音素所重疊比例，如對 h_a 重疊 2 個音框，而 h_a 在正確轉譯文句中實際長度爲 10 個音框，所以所重疊的比例爲 0.2。同理可求得對轉譯文句中 au 的重疊比例爲 1.0、對 t_a 的重疊比例爲 0.33。

Step 3. 再來先計算辨識文句中 au 對此三音素的正確率，若音素相同，則計算方式爲-1+2*重疊比例，否則爲-1+重疊比例。對 h_a 來說，因爲 h_a ≠ au，所以對 h_a 之正確率爲-0.8，同理可得對 au 的正確率爲 1.0、對 t_a 的正確率爲-0.67。而最後 au 之正確率取上述三個值中的最大值 1.0。



圖二、音素原始正確率計算方式[10]。

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[11]：

$$g_{MPE}(\lambda,\bar{\lambda})=\sum_r \sum_{q\in \mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \sum_m \gamma_q^{r\,MPE}\gamma_{qm}^r(t)\log N(o_r(t);\mu_m,\Sigma_m) \tag{11}$$

其中 s_q 與 e_q 分別代表音素 q 的起始時間(Start Time)與結束時間(End Time)；$o_r(t)$ 是 O_r 的第 t 個語音特徵向量；$N(;\mu_m,\Sigma_m)$ 是音素 q 的第 m 個高斯分佈，μ_m 與 Σ_m 分別是它的平均值向量與共變異矩陣；$\gamma_{qm}^r(t)$ 則是第 r 句訓練語句中在時間 t 時音素上 q 的高斯分佈 m 的佔有機率；$\gamma_q^{r\,MPE}$ 可進一步表示成[11]：

$$\gamma_q^{r\,MPE}=\left.\frac{\partial F_{MPE}(\lambda)}{\partial \log p_\lambda(O_r\mid q)}\right|_{\lambda=\bar{\lambda}}=\gamma_q^r\left(c_r(q)-c_{avg}^r\right) \tag{12}$$

其中 γ_q^r 是音素 q 在已知 O_r 情況下的事後機率；$c_r(q)$ 爲在詞圖 $\mathbf{W}_{lat}^r$ 中所有經過音素分枝 q 的文句對於 s_r 之期望正確率; c_{avg}^r 爲在詞圖 $\mathbf{W}_{lat}^r$ 中所有文句相對於 s_r 之期望正確率。γ_q^r 、$c_r(q)$ 與 c_{avg}^r 的統計量可在詞圖上使用波氏重估演算法來求得[11]。爲了增加模型估測的強健性，防止*最小化音素錯誤*過度的訓練並增進聲學模型的一般性(Generalization)，克服訓練與測試環境的不匹配，故在此引入以舊有模型參數爲超參數的平滑函數 $g_{EBW}^{smooth}(\lambda)$ 來加以輔助，由於弱性輔助函數加上平滑函數仍會滿足弱性輔助函數的性質，可提供較佳的參數估測。平滑函數 $g_{EBW}^{smooth}(\lambda)$ 則定義爲：

$$g_{EBW}^{smooth}(\lambda,\bar{\lambda})=\sum_m -\frac{D_m}{2}\left[\log(|\Sigma_m|)+(\mu_m-\bar{\mu}_m)^T\Sigma_m^{-1}(\mu_m-\bar{\mu}_m)+tr(\bar{\Sigma}_m\Sigma_m^{-1})\right] \tag{13}$$

其中 D_m 爲高斯分佈層次的平滑係數。此平滑函數以舊有之模型參數，如平均值向量 $\bar{\mu}_m$ 與共變

異矩陣 $\overline{\Sigma}_m$ 作爲超參數(Hyper-parameters)，使新估測之模型參數不致改變太大。加入此平滑函數後，輔助函數即成爲[11]：

$$g_{MPE}(\lambda,\overline{\lambda}) = \sum_r \sum_{q\in\mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \sum_m \gamma_q^{r\,MPE}\gamma_{qm}^r(t)\log N(o_r(t);\mu_m,\Sigma_m) - \sum_m \frac{D_m}{2}\left[\log(|\Sigma_m|) + (\mu_m-\overline{\mu}_m)^T\Sigma_m^{-1}(\mu_m-\overline{\mu}_m) + tr(\overline{\Sigma}_m\Sigma_m^{-1})\right] \tag{14}$$

分別對平均值向量與共變異矩陣作偏微分並使式爲 $\vec{0}$ 向量及 **0** 矩陣，則可推導出用於平均值向量與共變異矩陣的延伸波氏重估公式[12]：

$$\mu_m = \frac{\sum_r \sum_{q\in\mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \gamma_q^{r\,MPE}\gamma_{qm}^r(t)o_r(t) + D_m\overline{\mu}_m}{\sum_r \sum_{q\in\mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \gamma_q^{r\,MPE}\gamma_{qm}^r(t) + D_m}, \tag{15}$$

$$\Sigma_m = \frac{\sum_r \sum_{q\in\mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \gamma_q^{r\,MPE}\gamma_{qm}^r(t)o_r(t)o_r(t)^T + D_m\left[\overline{\Sigma}_m + \overline{\mu}_m\overline{\mu}_m^T\right]}{\sum_r \sum_{q\in\mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \gamma_q^{r\,MPE}\gamma_{qm}^r(t) + D_m} - \mu_m\mu_m^T, \tag{16}$$

其中 D_m 必須確保估測值共變異矩陣(Σ_m)爲正定矩陣。I-平滑(I-smoothing)技術[10]同樣是爲輔助函數引入平滑函數 $g^{I-smooth}(\lambda)$，但此平滑函數是以*最大化相似度*(Maximum Likelihood, ML)估測之統計資訊作爲超參數(Hyper-parameters)，故此函數 $g^{I-smooth}(\lambda)$ 可定義爲[11]：

$$g^{I-smooth}(\lambda) = \sum_m -\frac{\tau_m}{2}\left[\log(|\Sigma_m|) + (\mu_m - \frac{\theta_m^{ML}(O)}{\gamma_m^{ML}})^T\Sigma_m^{-1}(\mu_m - \frac{\theta_m^{ML}(O)}{\gamma_m^{ML}}) + tr\left(\left(\frac{\theta_m^{ML}(O^2)}{\gamma_m^{ML}} - \left(\frac{\theta_m^{ML}(O)}{\gamma_m^{ML}}\right)\left(\frac{\theta_m^{ML}(O)}{\gamma_m^{ML}}\right)^T\right)\Sigma_m^{-1}\right)\right], \tag{17}$$

其中 τ_m 是一個高斯分佈層次的平滑係數，表示要由*最大化相似度*統計資訊加入的資料點數；$\gamma_m^{r\,ML}(t)$ 表示在第 r 句訓練語句中，時間 *t* 時，以*最大化相似度*法則所估測之高斯分佈 *m* 的佔有機率，可表示成 $\gamma_m^{ML} = \sum_r \sum_t \gamma_m^{r\,ML}(t)$；而 $\theta_m^{ML}(O)$ 與 $\theta_m^{ML}(O^2)$ 分別可表示成 $\theta_m^{ML}(O) = \sum_r \sum_t \gamma_m^{r\,ML}(t)o_r(t)$ 與 $\theta_m^{ML}(O^2) = \sum_r \sum_t \gamma_m^{r\,ML}(t)o_r(t)o_r(t)^T$。因此重估公式要修改爲[10]：

$$\mu_m = \frac{\sum_r \sum_{q\in\mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \gamma_q^{r\,MPE}\gamma_{qm}^r(t)o_r(t) + \frac{\tau_m}{\gamma_m^{ML}}\theta_m^{ML}(O) + D_m\overline{\mu}_m}{\sum_r \sum_{q\in\mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \gamma_q^{r\,MPE}\gamma_{qm}^r(t) + \tau_m + D_m}, \tag{18}$$

$$\Sigma_m = \frac{\sum_r \sum_{q\in \mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \gamma_q^{r\,MPE} \gamma_{qm}^r(t) o_r(t) o_r(t)^T + \frac{\tau_m}{\gamma_m^{ML}} \theta_m^{ML}(O^2) + D_m \left[\Sigma_m + \mu_m \mu_m^T\right]}{\sum_r \sum_{q\in \mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \gamma_q^{r\,MPE} \gamma_{qm}^r(t) + \tau_m + D_m} - \mu_m \mu_m^T, \tag{19}$$

其中 $\gamma_m^{r\,ML}(t)$ 表示在第 *r* 句訓練語句中，時間 *t* 時，以*最大化相似度*法則所估測之音素 *q* 的第 *m* 個高斯分佈的佔有機率。

4. 最小化音素錯誤爲基礎的線性轉換調適技術

在模型空間上的調適方法中，連續密度隱藏式馬可夫模型中的平均值向量與共變異矩陣分別使用不同的線性迴歸矩陣，本文的研究只針對平均值向量的調適。在平均值向量的調適中，假設調適前的高斯分佈 *m* 的平均值向量爲 $\overline{\mu}_m$ ，調適後的平均值向量爲 μ_m ，並希望透過一線性迴歸矩陣來調適此平均值向量：

$$\mu_m = A\overline{\mu}_m + b = W\overline{\xi}_m, \tag{20}$$

其中 $W = \left[b\ A\right]$ 爲調整平均值向量的線性迴歸矩陣，其中 A 爲旋轉矩陣，而 b 爲偏移向量；$\overline{\xi}_m = \left[1\ \overline{\mu}_m^T\right]^T$ 是調適前的延伸平均值向量(Extended Mean Vector)。由式(11)可得到輔助函數 $g_{MPE}(W, \overline{W})$ 爲[13]：

$$g_{MPE}(W, \overline{W}) = \sum_m \sum_{q\in \mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \gamma_q^{MPE} \gamma_{qm}(t) \log N(o(t); W\xi_m, \Sigma_m), \tag{21}$$

其中 $\overline{W}$ 爲舊有的轉換矩陣，γ_q^{MPE} 同樣可由式(12)求得，而估測 $\gamma_{qm}(t)$ 與 γ_q^{MPE} 所需的聲學模型，由舊有的轉換矩陣 $\overline{W}$ 所轉換。爲了增加線性迴歸矩陣估測的強健性，同式(14)，在此也加入了以舊有線性迴歸矩陣 $\overline{W}$ 爲超參數的平滑函數 $g_{EBW}^{smooth}(W, \overline{W})$ 來輔助最佳化，$g_{EBW}^{smooth}(W, \overline{W})$ 則定義爲[13]：

$$g_{EBW}^{smooth}(W, \overline{W}) = \sum_m -\frac{D_m}{2} (W\xi_m - \overline{W}\xi_m)^T \Sigma_m^{-1} (\overline{W}\xi_m - W\xi_m), \tag{22}$$

其中 D_m 爲高斯分佈層次的平滑係數，所以新的輔助函數爲[13]：

$$\begin{aligned} g_{MPE}(W, \overline{W}) = &\sum_m \sum_{q\in \mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \gamma_q^{MPE} \gamma_{qm}(t) \log N(o(t); W\xi_m, \Sigma_m) \\ &- \frac{1}{2} \sum_m D_m \left[(\overline{W}\xi_m - W\xi_m)^T \Sigma_m^{-1} (\overline{W}\xi_m - W\xi_m)\right], \end{aligned} \tag{23}$$

我們若將式(23)對 W 作偏微分，並設之等於零，可得：

$$\mathbf{G}^{-1}\mathbf{w} = \mathbf{z}, \tag{24}$$

其中 $\mathbf{z} = vec(Z)$ ， $\mathbf{w} = vec(W)$ ，*vec*(・)是一個將矩陣轉換爲向量的函數，以列優先的排序方式；而 Z 表示成：

$$Z = \sum_m \Sigma_m^{-1} \left(\sum_{q\in \mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \gamma_q^{MPE} \gamma_{qm}(t) o(t) + D_m \overline{W}\xi_m \right) \xi_m^T, \tag{25}$$

另一方面，$\mathbf{G}$ 可進一步表示成：

$$\mathbf{G} = \sum_{m} kron(V_m, R_m), \tag{26}$$

而其中 *kron*(・)則爲 kronecker 矩陣乘法[14]，V_m 與 R_m 分別爲：

$$V_m = \left(\sum_{q \in \mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \gamma_q^{MPE} \gamma_{qm}(t) + D_m \right) \Sigma_m^{-1}, \tag{27}$$

$$R_m = \xi_m \xi_m^T. \tag{28}$$

W 可由線性方程組(Systems of Linear Equations)式(24)求出。爲了避免過度訓練，I-平滑技術同樣也可以使用在此。以*最大化相似度*之統計資訊爲超參數(Hyper-parameters)的平滑函數 $g^{I-smooth}(W)$ 可定義爲[13]：

$$g^{I-smooth}(W) = \sum_{m} -\frac{\tau_m}{2} \left[(o(t) - W\xi_m)^T \Sigma_m^{-1} (o(t) - W\xi_m) \right], \tag{29}$$

其中 τ_m 是一個係數，要表示要由*最大化相似度*統計資訊加入的資料點數，因此可將式(25)與式(27)修改爲[13]：

$$Z = \sum_{m} \Sigma_m^{-1} \left(\left(\sum_{q \in \mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \gamma_q^{MPE} \gamma_{qm}(t) + \tau_m \right) o(t) + D_m \overline{W} \xi_m \right) \xi_m^T, \tag{30}$$

$$V_m = \left(\sum_{q \in \mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \gamma_q^{MPE} \gamma_{qm}(t) + \tau_m + D_m \right) \Sigma_m^{-1}, \tag{31}$$

5. 詞錯誤最小化搜尋演算法

本論文以詞錯誤最小化(Word Error Minimization, WEM)[6]搜尋演算法(或稱爲最小化貝氏風險搜尋(Minimum Bayes Risk, MBR)[5]作爲搜尋法則，在式(2)中，由於實際上不可能對所有可能的文句 $\mathbf{W}_h$ 作窮舉，所以我們根據[6]先對測試聲學特徵向量序列 O_l 先進行語音辨識產生詞圖 $\mathbf{W}_{lat}^l$，並從詞圖 $\mathbf{W}_{lat}^l$ 上找出 *N*-最佳序列(*N*-best List) $\mathbf{N}^l$，再接著估測 $\mathbf{N}^l$ 中每一路徑(或文句)的條件風險，進而選出擁有最小條件風險的文句。其詞錯誤最小化搜尋的標準法則爲：

$$s^* = \arg\min_{s} \sum_{u \in \mathbf{N}^l} P(u \mid O_l) L(s,u), \tag{32}$$

在上式中，每一條路徑 s 都要與其它所有的路徑 u 算編輯距離 $L(s,u)$ 並乘上 u 的事後機率，將之加總起來即爲一條路徑 s 的條件風險。我們可以基於此一條件風險從 $\mathbf{N}^l$ 找出最小條件風險的文句 s^* 作爲最後的語音辨識結果。由於在中文的斷詞上會有混淆的問題，詞錯誤率通常不是一個評估語音辨識效能的很好標準；因此，在本研究我們在計算編輯距離 $L(s,u)$ 時，實作上是以字作爲比對與計算的單位。

N-最佳路徑	正確答案：”昔日-執政黨-大-掌櫃”		
u (sentence)	$P(u \mid O_l)$	$\sum_{u \in \mathbf{N}^l} P(u \mid O_l) L(s,u)$	CER
“昔日-行政-長-大-掌櫃”	0.23	85	2
“七-日-執政黨-把-掌櫃”	0.21	90	2
“昔日-執政黨-把-掌櫃”	0.19	73	1
“昔日-指-政黨-大-掌櫃”	0.18	87	1
“昔日-行政黨-把-掌櫃”	0.17	101	2

表一、最小條件風險之例子，$\mathbf{N}^l$ 大小設爲 50。

性別	訓練語料		評估語料			語料重疊人數
	總長(分鐘)	人數(人)	總長(分鐘)	人數(人)	平均長度(字)	(人)
男生	766.69	≤66	21.69	9	89.8	9
女生	766.79	≤111	65.23	≤23		≥13

表二、語音實驗語料統計資訊。

如表一所示，正確答案爲“昔日-執政黨-大-掌櫃”，用傳統最大事後機率所找出的文句爲“昔日-行政-長-大-掌櫃”其 CER(字錯誤率)=2，若用詞錯誤最小搜尋法則，找到的最小條件風險文句爲“昔日-執政黨-把-掌櫃”其 CER=1。在此一例子可明顯地看出，利用詞(字)錯誤最小化搜尋法則可以降低字錯誤率，進而提高語音辨識效能。

6. 實驗與討論

6.1 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台灣師範大學資工所目前所發展的新聞語音辨識系統[15]，它基本上是一套大詞彙連續語音辨識系統，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)[16]等部分。

在前端處理方面，本文使用梅爾倒頻譜特徵作爲語音訊號的特徵參數，並且使用倒頻譜平均消去法(Ceptral Mean Subtraction, CMS)以移除錄音時通道效應所造成的影響。在聲學模型我們總共使用了 151 個隱藏式馬可夫模型來作爲中文 INITIAL-FINAL 的統計模型，其中隱藏式馬可夫模型的每個狀態會依據其對應到的訓練語料多寡，以 2 到 128 個高斯分佈來表示，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[17]。在本文中的語言模型使用了 Katz 語言模型平滑技術[18]，在訓練時是採用 SRL Language Modeling Toolkit (SRILM)[19]。在詞彙樹搜尋時，本系統採用詞雙連語言

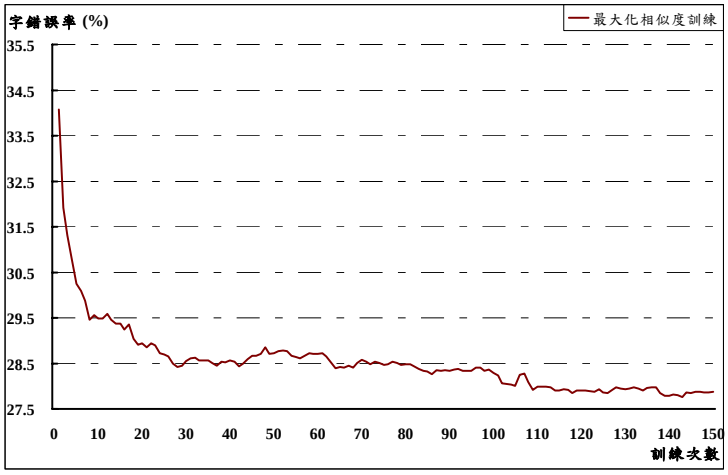
模型；在詞圖搜尋時，則採用詞三連語言模型。

6.2 實驗語料

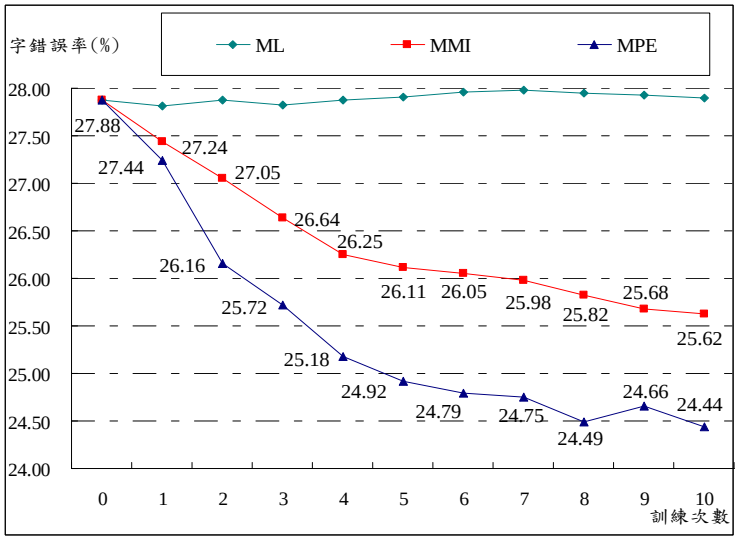
本論文實驗使用的訓練與測試語料爲 MATBN 電視新聞語料庫[20]，是由中央研究院資訊所口語小組[21]耗時三年與公共電視台[22]合作錄製完成。我們初步地選擇採訪記者語料作爲實驗語材，其中包含 25.5 小時的訓練集(5,774 句)，供聲學模型訓練之用，其中男女語料各半；1.5 小時的評估集(292 句)，供辨識評估之用。訓練集由 2001 及 2002 年的新聞語料所篩選出來的；評估集則均爲 2003 年的語料，由中研院的評估語料篩選出來，只選擇了採訪記者語料並濾掉了含有語助詞之語句。語料中語者資訊如表二所示。另外，評估集中每則新聞的平均長度爲 89.8 個字。

6.3 聲學模型訓練之實驗

本論文先進行 150 次迭代的*最大化相似度*訓練，字錯誤率曲線如圖三所示。然後分別再進行 10



圖三、150 次的最大化相似度聲學模型訓練之字錯誤率曲線。



圖四、10 次最小化音素錯誤訓練、最大化交互資訊訓練及最大化相似度訓練字錯誤率曲線。

	CER (%)
Baseline: MPE	24.44
MPE + MLLR	23.41
MPE + MMILR (τ_m=3)	23.52
MPE + MMILR (τ_m=10)	23.30
MPE + MMILR (τ_m=30)	23.20
MPE + MPELR (τ_m=3)	23.07
MPE + MPELR (τ_m=10)	23.11
MPE + MPELR (τ_m=30)	23.19

表三、鑑別式聲學模型調適之字錯誤率(%)。

	CER (%)
Baseline: MPE + MPELR	23.07
MPE + MPELR + WEM (N-best)	23.04
Word Graph Error Rate (GER)	12.28

表四、詞錯誤最小化搜尋之結果。

次迭代的鑑別式訓練(Large Scale Discriminative Training)，其中包含*最小化音素錯誤*訓練及*最大化交互資訊*訓練(Maximum Mutual Information, MMI)[23]。字錯誤率曲線如圖四所示，其中我們以經由額外 10 次迭代的*最大化相似度*訓練模型作爲對照組(所以總共經歷 160 次迭代的*最大化相似度*訓練)。在實驗的設定中，正確的轉譯文句要先經由強制對齊(Force Alignment)產生詞段，再加入詞圖中。I-平滑技術中的控制參數 $\tau_m = 10$ ，詞段的聲學分數使用維特比的分數再經過 1/12 次方來縮小和語言機率的比例，利用詞圖進行鑑別式聲學模型訓練時所用的語言限制則是使用詞單連(Unigram)語言模型。由實驗結果顯示，經過 150 次*最大化相似度*訓練之後，錯誤率的下降已趨於飽和，無法再藉由傳統*最大化相似度*訓練來降低錯誤率。需藉由鑑別式的方法做更進一步的訓練，如圖四，*最大化交互資訊*能相對降低 8.11%的字錯誤率，已有相當顯著的成效。而經過 10 次的*最小化音素錯誤*訓練之後，可相對降低 12.34%的字錯誤率，明顯優於*最大化交互資訊*訓練所帶來的成效，有兩個原因。第一、*最小化音素錯誤*使用 Levenshtein 函數的改良正確率計算方式作爲減損函數，較*最大化交互資訊*所使用的零壹函數更貼近於評估標準。第二、*最小化音素錯誤*直接最小化全面風險，*最大化交互資訊*卻是透過一個上界間接最小化全面風險。

6.4 非監督式聲學模型調適之實驗

在非監督式聲學模型調適的實驗中，以 10 次*最小化音素錯誤*訓練之模型作爲基礎(Baseline)，並將 14,396 個高斯分佈以聲母、韻母及靜音區分成 3 個迴歸群集(Regression Classes)對平均值向量做調適。實驗結果如表三所示，其中MLLR代表*最大化相似度線性迴歸*，MMILR代表*最大交互資訊線性迴歸*，MPELR代表*最小化音素錯誤線性迴歸*。在實驗設定方面，我們針對不同鑑別式聲學模型調適方法使用不同的τ_m最來作初步的探討，其他設定則與聲學模型訓練時維持一致。由

實驗結果可見，隨著τ_m 愈大，MMILR的表現愈來愈好，在τ_m=30 時有較佳的字錯誤率；而MPELR則剛好相反，在τ_m=3 時有較佳的字錯誤率。我們推測原因可能是MMILR的輔助函數與目標函數差異較大，需要引入較多的*最大化相似度*估測值，來得到較強健性的估測；而MPELR的輔助函數則較貼近於目標函數，使得愈多的*最大化相似度*估測值，反而會減緩估測的收斂速度。我們進一步分析可發現MPELR在τ_m=3 時能相對降低 5.61%的字錯誤率，相較於MMILR(τ_m=30)的 5.07%、MLLR的 4.21%，MPELR的確提供了較佳的聲學模型調適效能。

6.5 搜尋解碼之實驗

如表四的第三列所示，使用詞錯誤最小化(WEM)搜尋演算法後僅能將字錯誤率由 23.07%(表三最佳的字錯誤率)降低至 23.04%，效果並不如預期的好。經由觀察，我們發現在每一個詞圖所產生的前 1,000 名最佳序列中，即式(32)中的 $\mathbf{N}'$ 大小爲 1,000 時，大部分的序列都與第一名序列(即事後機率最大的序列)很相近，這樣我們便沒有辦法利用編輯距離的分數來降低某一條有可能是字錯誤最小序列的風險，進而將字錯誤最小序列取代第一名序列。另外，由於大部分序列都與第一名序列相近，若第一名序列的事後機率與其它序列的事後機率差不多的情況下時，利用詞錯誤最小化搜尋演算法有可能會將其他條序列取代第一名序列，造成辨識率下降；僅少部分的序列會與第一名序列較不相近而與其他前幾名的序列較相近，這樣情況下利用詞錯誤最小化搜尋演算法便可以找出字錯誤最小序列取代第一名序列(如表一的例子)，進而提高辨識率。

7. 結論與未來展望

由於傳統聲學模型訓練方式是以*最大化相似度法*則來訓練，使聲學模型在訓練語料上有最大的聲學相似度，但由於訓練環境與測試環境的差異(Mismatch)，過多的訓練次數，不僅無法帶來額外的辨識效能，反而會因此降低了辨識率，而*最小化音素錯誤*訓練法則，能在*最大化相似度*訓練之後，再對聲學模型施以鑑別式訓練及調適，使聲學模型更具鑑別力。本論文中，經由初步的實驗顯示，在聲學模型訓練上，*最小化音素錯誤*能有效的降低辨識錯誤率，相對於*最大化相似度*訓練能降低 15.52%的音節錯誤率、12.33%的字錯誤率及 10.02%的詞錯誤率。在聲學模型的調適上，也有小幅度的進步。在搜尋方面，由於詞圖的資訊提供了字正確率的上限爲 87.72%(如表四的第四列所示，詞圖的最佳字錯誤率爲 12.28%)，而在 *N*-最佳序列上的*詞錯誤最小化*搜尋僅能將字正確率提升至 76.95%(23.05%字錯誤率)，所以尚有 10.67%字錯誤率的空間可供改善。在寫作此論文的同時，我們正計畫使用 A*搜尋(A* Search)，以*詞錯誤最小化*準則直接在詞圖上找最佳的文句[5]，以冀求更佳的辨識率。

另外，我們也曾嘗試將*最小化音素錯誤*訓練法則應用在語言模型估測上，靠著估測語言模型之機率，來最大化訓練語料中詞圖的期望正確率，雖然在語音辨識率的提昇上不甚顯著，但是對語言模型估測卻提供了新的視野；而新近也有學者將*最小化音素錯誤*訓練法則應用在特徵空間轉換上[25]，用來求取特徵空間的轉換矩陣，研究結果顯示透過*最小化音素錯誤*訓練法則來求取轉換矩陣，除了能達到降低語音特徵向量維度與減少維度間相關性的功用外，對於語音辨識率的提昇也得到相當大的成效。

8. 參考文獻

[1] R. O. Duda, P. E. Hart and D. G. Stork. Pattern Classification, Second Edition. *New York: John & Wiley*, 2001.

[2] A. Levenshtein, “Binary Codes Capable of Correcting Deletions, Insertions and Reversals,” *Soviet Physics Doklady*, Vol. 10, No. 8, pp.707-710, 1966.

[3] L. R. Bahl, F. Jelinek and R. L. Mercer, “A Maximum Likelihood Approach to Continuous Speech Recognition,” *IEEE Trans. Pattern Analysis and Machine Intelligence,* Vol. PAMI-5, No.2, pp.179-190, March 1983.

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

[5] V. Goel and W. Byrne, “Minimum Bayes-risk Automatic Speech Recognition,” in Pattern Recognition in Speech and Language Processing, edited by W. Chou and B. H. Juang, *CRC Press*, 2003.

[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] K. Na, B. Jeon, D. Chang, S. Chae, and S. Ann, “Discriminative Training of Hidden Markov Models using Overall Risk Criterion and Reduced Gradient Method,” in *Proc. EUROSPEECH 1995.*

[8] J. Kaiser, B. Horvat, Z. Kacic, “A Novel Loss Function for the Overall Risk Criterion Based Discriminative Training of HMM Models,” in *Proc. ICSLP 2000.*

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

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

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

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

[13] L. Wang and P. C. Woodland, “MPE-Based Discriminative Linear Transform for Speaker Adaptation,” in *Proc. ICASSP 2004*.

[14] R. D. Schafer. An Introduction to Nonassociative Algebras. *New York: Dover*, 1996.

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

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

[17] LDC: Linguistic Data Consortium. http://www.ldc.upenn.edu

[18] 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.

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

[20] 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.

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

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

[23] Y. Normandin, R. Lacouture, R. Cardin, “MMIE Training for Large Vocabulary Continuous Speech Recognition,” in *Proc. ICSLP 1994*.

[24] J. W. Kuo and B. Chen, “Minimum Word Error Based Discriminative Training of Language Models,” in *Proc. EUROSPEECH 2005*.

[25] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, Geoffrey Zweig, “fMPE: Discriminatively Trained Features for Speech Recognition,” in *Proc. ICASSP 2005.*