整合邊際資訊於鑑別式聲學模型訓練方法之比較研究

A Comparative Study on Margin-Based Discriminative Training of Acoustic Models

羅永典　Yueng-Tien Lo
國立臺灣師範大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan Normal University
g96470198@csie.ntnu.edu.tw

陳柏琳　Berlin Chen
國立臺灣師範大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan Normal University
berlin@csie.ntnu.edu.tw

摘要

鑑別式聲學模型訓練在近代自動語音辨識(Automatic Speech Recognition, ASR)中扮演重要的角色。在許多基於不同思維且能有效地提昇辨識率的鑑別式聲學模型訓練方法陸續被提出後，對於訓練方法的相關推廣與改進便如雨後春筍般地興起；而這些方法在本質上，皆是在描述訓練語句與語音辨識器所產生對應詞圖(Word Graph)之間的關係。本論文首先將統整與歸納近年來所發展的多種鑑別式聲學模型訓練方法，並以三種最具代表性鑑別式訓練方法：最小化分類錯誤(Minimum Classification Error, MCE)、最大化交互資訊(Maximum Mutual Information, MMI)、最小化音素錯誤(Minimum Phone Error, MPE)為範例，透過有系統地轉換與化解方程式，得到聲學模型訓練準則的共通表示函數型態。我們可以發現到，對於上述鑑別式訓練方法，此共通表示函數背後物理意義之差別乃是在於欲觀察訓練語料不同層級的鑑別資訊，如音素(Phone)、語句(Utterance)等，以及共通表示函數之參數設定。其次，本論文針對語音辨識結果所形成的假設空間上所觀察到錯誤(或正確)率的不同細緻層度，在模型訓練時引入了機器學習領域中的邊際概念；其背後的物理意義，事實上就是從不同層級的訓練語料中選取適合的資訊供聲學模型訓練所使用。本論文的目的在於分析近代對於以隱藏式馬可夫模型為聲學模型之模型訓練方法與邊際概念在演進上的一致性；從琳瑯滿目的訓練方法之中，闡述近年來鑑別式聲學模型訓練發展演進之中心思想。最後，我們實作於中文大詞彙連續語音辨識系統，驗證了多種鑑別式聲學模型訓練方法以及我們所提出方法之效能。

關鍵詞：語音辨識、聲學模型鑑別式訓練、邊際資訊、資料選取

Proceedings of the 22nd Conference on Computational Linguistics and Speech Processing (ROCLING 2010), Pages 65-85, Puli, Nantou,Taiwan, September 2010.

一、緒論

以最大化相似度估測(Maximum Likelihood Estimation, MLE)來訓練聲學模型(Acoustic Models)，在過去數十年廣為語音辨識領域所採用，它主要是考量如何能從訓練語料中獲得統計資訊，以讓聲學模型可以代表訓練語料(也就讓聲學模型產生對應的訓練語料之相似度最大)。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，雖可使相關的語音特徵落在某一個聲學模型的相似度變大，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。鑑別式訓練不僅是考慮了訓練語句的正確(參考)轉寫(Correct or Reference Transcription)，同時也考慮由語音辨識器對語句進行辨識後產生的、與正確轉寫不同的候選詞序列假設(Candidate Word Sequence Hypotheses)，以增進訓練後聲學模型的鑑別性。

長久以來，鑑別式訓練方法為語音辨識中聲學模型辨識能力提昇中的重要一環，相關研究與延伸族繁不及備載，以下列三種方法較具代表性：(一)最小化分類錯誤(Minimum Classification Error, MCE)[1]考慮到訓練語句的正確轉寫與不正確轉寫在假設空間上的分離程度；(二)最大交互資訊法則(Maximum Mutual Information, MMI)[2]以最大化訓練語句與其對應詞正確轉寫的交互資訊；(三)最小化音素錯誤(Minimum Phone Error, MPE)[3]目的為最小化語音辨識器輸出(亦即所有候選詞序列)的音素期望錯誤率。這三種以不同思維出發的方法，它們背後涵義皆是描述訓練語句的正確轉寫與其它候選詞序列之間的關係，日後許多被提出的聲學模型訓練方法也都是架構於這樣的關係上。

通常在執行鑑別式訓練時，每一句訓練語句所對應的詞圖(Word Lattice)扮演的角色不僅僅為所有候選詞序列(可能包括正確轉寫)之假設空間(Hypothesis Space)，更是提供聲學模型參數估測過程中鑑別資訊的重要來源之依據。對於鑑別資訊的需求因而有了所謂的資料選取(Data Selection)概念：在機器學習中，支撐向量機邊際資訊概念在分類問題上有著非常成功的成效，邊際資訊所闡述的理念是決策邊際與訓練資料分布對分類問題所產生影響；相同的概念被推廣到在語音辨識的聲學模型訓練中使用[4][5][6][8]，並且證明了其效用[7]，而相同有效果的資料選取方法如[9][10][11][12]同樣在對於語音辨識率的提昇有著正面影響。此外，基於對訓練語料的分布以及鑑別式訓練所需的權重的分析，吾人結合40 delta 調整讓兩類邊際資訊方法於鑑別式聲學模型訓練上：一方面，除了在辨識率上可以有效提昇之外；另一方面，對於聲學模型訓練時因增進權重因子所帶來過度訓練(Overfitting)的問題，亦可以有效的獲得解決。

本論文對於近年鑑別式訓練及其延伸方法，做出統整歸納之研究，並從代表性的資料選取方法中結合其優點，獲得同時兼具各項優點的聲學模型訓練目標函數。吾人將各類訓練準則與本論文所提及改善方法實作於中文大詞彙連續語音辨識實驗中，同時從實驗結果中觀察資料選取方法的優缺點。本論文接下的安排如下：第二章以最具代表性三種鑑別式聲學模型訓練方法為例，整合從各種不同角度思考的目標函數，藉由數學推導詮釋其目的的一致性；第三章為歸納各種邊際資訊概念的延伸，分析各種方法所專注的訓練資料，並提出以柔性邊際法則與增進式因子為基礎的目標函數；第四章為各種方法

的實驗比較與討論；第五章為結論及未來研究展望。

二、 鑑別式訓練法則及其一致性

本章節首先將從語音辨識原則為出發點，闡述大多數方法所依循準則為何，據此將輸入的語音訊號(或語音特徵向量序列)辨識成對應的自動轉寫做為輸出。因此，在聲學模型訓練時，我們自然地就可以利用此原則做為鑑別式訓練當中定義目標函數的主要根據。

一般來說，在執行鑑別式訓練前，我們通常先使用最大化相似度估測(MLE)做為基礎聲學模型的訓練方法，其目的是最大化聲學模型產生對應的訓練語料之相似度。爾後在鑑別式訓練上，透過不同的思維對於訓練語句的語音辨識混淆資訊產生需求，也就是以正確轉寫與語音辨識器所產生的詞圖(內含許多候選詞序列)形成所謂的假設空間；以時下最流行的三種最具代表性之鑑別式訓練方法為例，說明其聲學模型訓練目標函數的最終目的都是在於使用鑑別式函數描述正確轉寫與其它在詞圖上的候選詞序列在假設空間所形成的關係。我們可以歸納出這其中的差別主要在於不同層級(如語句層級、音素層級)或不同細緻程度的訓練資料選取。美國微軟公司(Microsoft)的學者針對這點在近期提出了完整的證明[13]；同樣地，日本電報電信公司(NTT)[14]也抱持相同看法並且引入增進權重，其所隱含的即是邊際資訊概念，並且說明了基於相同型式的不同目標函數皆可使用同樣的最佳化方法來求得模型參數；其它如[15][16]也都對目標函數與最佳化方法做深入的解析來探討鑑別式訓練方法的一致性。

（一）語音解碼原則與最大化相似度法則：

在語音辨識的解碼原則上，大多數的作法通常是採用貝式決策定理(Bayesian Decision Theorem)：即是在給定一句語句的語音特徵向量序列$O=\{o_1,...,o_t\}$已知的情況下，要找一段文字$\hat{W}$的相似度(發生的機率)最大：

$$\hat{W}=\arg\max_{W_i} P(W_i|O) \tag{1}$$

其中W_i代表所有某一條候選詞序列，$P(W_i|O)$為給定語音特徵向量序列後O，W_i發生的事後機率(Posterior Probability)。若使用貝式定理(Bayes' Theorem)將式(1)中的事後機率項展開成

$$P(W_i|O)=\frac{p(O|W_i)P(W_i)}{p(O)} \tag{2}$$

其中$p(O)$為產生語句O的事前機率並不影響所有候選詞序列之排序，故式(1)可簡化為：

$$\hat{W}=\arg\max_{W_i} p(O|W_i)P(W_i)=\arg\max_{W_i} p(O,W_i) \tag{3}$$

以式(3)為目標函數的解碼(搜尋)演算法即是所謂的最大化事後機率(Maximum a Posteriori Probability, MAP)解碼方法。

鑒於語音辨識解碼原則為評估詞序列是否可能為較正確輸出的依據，自然的以式(3)為鑑別式函數，做為每一句訓練語句的正確轉寫與詞圖上其它不同於正確轉寫的選詞序列在假設空間上之間比較的依據。鑑別式訓練設法將訓練語句的正確轉寫與其它候選詞序列在假設空間上的關係透過目標函數來描述，使得在函數最佳化過程中，得以透過鑑別式訓練讓模型更能分辯訓練語句的正確轉寫與其它候選詞序列的差異來達到增進辨識率的目的。接下來，對於每一句訓練語句，我們將以 O_z 代表第 z 句語句特徵序列，W_{zR} 代表第 z 句語句特徵序列對應的正確轉寫，W_{zi} 代表第 z 句語句特徵序列在詞圖上的候選詞序列之一。

（二）鑑別式聲學模型訓練法則：

我們可以依循[13]來分析上述三種代表性鑑別式法則之目標函數的一致性，透過它們的一致性來說明鑑別式聲學模型訓練法則主要在於正確轉寫與其它候選詞序列的關係上的最佳化：

（I）最小化分類錯誤法則：以最小化分類錯誤為基礎設計的語音辨識器(分類器)，其決策的法則可以透過錯誤分類評估(Misclassification Measure)函數來表示，錯誤分類評估函數代表的是語音辨識器對於其它非正確轉寫之候選詞序列產生的分數減去對於正確轉寫產生的分數。若錯誤分類評估函數輸出大於零則表示語音辨識(分類)錯誤，反之則表示語音辨識(分類)正確。若聲學模型的訓練若能依循最小化錯誤分類評估函數的輸出來設計，則預期將會有較少的語音辨識錯誤。在實作時，語音辨識器給予每一句訓練語句 z 對應的正確轉寫 W_z 的分數(即鑑別函數輸出)可定義為

$$g_\Lambda(O_z) = \log p_\Lambda(O_z, W_{zR}) \tag{4}$$

其中 Λ 是語音辨識器模型(聲學模型及語言模型)參數，而語音辨識器給予其它非正確轉寫之候選詞序列的分數可表示成

$$G_\Lambda(O_z) = \log\left\{\frac{1}{N}\sum_{W_{zi},\, W_{zi}\neq W_{zR}} p_\Lambda^\eta(O_z, W_{zi})\right\}^{\frac{1}{\eta}} \tag{5}$$

其中 N 代表所有非正確轉寫之候選詞序列的個數，而 w_{zi} 是其中的一條候選詞序列；η 為調整聯合機率分布的範圍因子，為簡化計算，在本論文我們令 $\eta = 1$，關於 η 的進一步討論可參閱[13]。式(5)對於所有非正確轉寫之候選詞序列分數的總和在一正規化後再取對數，可視為於是一種柔性最大臨界值，代表的是與正確轉寫分數相近的錯誤語音辨識結果。於是，分類錯誤評估函數 $d_\Lambda(O_z)$ 可以定義為

$$d_\Lambda(O_z) = -g_\Lambda(O_z) + G_\Lambda(O_z) \tag{6}$$

值得注意的是(6)式中呼應了鑑別式訓練的精神，描述訓練語句之正確轉寫與語音辨識結果(詞圖中候選詞序列)之間的關係。藉由分類錯誤評估 $d_\Lambda(O_z)$ 來描述語音辨識器對每一訓練語句的決策結果：當 $d_\Lambda(O_z) < 0$，時表示訓練語句 z 被模型 Λ 分類正確；$d_\Lambda(O_z) \geq 0$ 則表示 z 被錯誤分類。而一般在給定分類錯誤評估函數後，對於每一句訓練語句的損失函數經常可透過使用 s 型函數(Sigmoid Function)而得：

$$l\left(d_{\Lambda}(O_z)\right)=\frac{1}{1+e^{-\alpha\cdot d_{\Lambda}(O_z)}} \tag{7}$$

其中$\alpha>0$，可調整s型函數傾斜程度，在本論文為我們設定$\alpha=1$，關於α討論亦可參閱[13]。由s型平滑函數我們可以注意到損失函數的值落於 0 至 1 區間內，我們將鑑別函數(式(4)與(5))套用於式(7)，令$\alpha=1$與$\eta=1$，並重新整理後可得到：

$$l\left(d_{\Lambda}(O_z)\right)=\frac{\sum_{W_{zi},W_{zi}\neq W_{zR}} p_{\Lambda}(O_z,W_{zi})}{\sum_{W_{zi},W_{zi}\neq W_{zR}} p_{\Lambda}(O_z,W_{zi})+p_{\Lambda}(O_z,W_{zR})}=\frac{\sum_{W_{zi},W_{zi}\neq W_{zR}} p_{\Lambda}(O_z,W_{zi})}{\sum_{W_{zi}} p_{\Lambda}(O_z,W_{zi})} \tag{8}$$

對於訓練語料，以最小分類錯誤為準則的鑑別式訓練，其聲學模型訓練目標函數$F_{MCE}(\Lambda)$可表示成最小化所有訓練語句的期望錯誤率

$$F_{MCE}(\Lambda)=\frac{1}{Z}\sum_{z=1}^{Z} l\left(d_{\Lambda}(O_z)\right) \tag{9}$$

我們若進一步定義一個為“1 減去損失函數”的功用函數：

$$u\left(d_{\Lambda}(O_z)\right)=1-l\left(d_{\Lambda}(O_z)\right) \tag{10}$$

則可看出最小化損失函數(式(9))等同於最大化下列最小分類錯誤目標函數$\hat{F}_{MCE}(\Lambda)$：

$$\hat{F}_{MCE}(\Lambda)=\sum_{z=1}^{Z} u\left(d_{\Lambda}(O_z)\right)=\sum_{z=1}^{Z}\frac{p_{\Lambda}(O_z,W_{zR})}{\sum_{W_{zi}} p_{\Lambda}(O_z,W_{zi})} \tag{11}$$

（II）最大化交互資訊估測法則：以最大化交互資訊估測法為準則的鑑別式聲學模型訓練，主要目的是最大化所有訓練語句z的語音特徵向量序列O_z與其對應轉寫W_{zR}的交互資訊$F_{MMI}(\Lambda)$，定義如下：

$$\begin{aligned} F_{MMI}(\Lambda) &= \sum_{O_z,W_{zR}} p(O_z,W_{zR})\log\frac{p_{\Lambda}(O_z,W_{zR})}{p_{\Lambda}(O_z)P_{\Lambda}(W_{zR})} \\ &= \sum_{O_z,W_{zR}} p(O_z,W_{zR})\log\frac{p_{\Lambda}(W_{zR}\mid O_z)}{P_{\Lambda}(W_{zR})}=H(W_{zR})-H(W_{zR}\mid O_z) \end{aligned} \tag{12}$$

其中$H(W_{zR})=-\sum_{W_{zR}} P_{\Lambda}(W_{zR})\log P_{\Lambda}(W_{zR})$為$W_{zR}$的熵值；而$H(W_{zR}\mid O_z)$為給定條件$O_z$時$W_{zR}$的熵值，可表示為$H(W_{zR}\mid O_z)=-\sum_{W_{zR},O_z} p(W_{zR},O_z)\log p_{\Lambda}(W_{zR}\mid O_z)$。假定在模型訓練時，語言模型$P_{\Lambda}(W_{zR})$不做調整(亦即$H(W_{zR})$不改變)的情況下，最大化式(12)等同於最小化$H(W_{zR}\mid O_z)$。若再假設訓練語句為有相同機率分布值(Uniformly Distributed)下，則$H(W_z\mid O_z)$可表示近似地成

$$H(W_{zR}\mid O_z)=-\frac{1}{Z}\sum_{z=1}^{Z}\log p_{\Lambda}(W_{zR}\mid O_z)=-\frac{1}{Z}\sum_{z=1}^{Z}\log\frac{p_{\Lambda}(W_{zR},O_z)}{p_{\Lambda}(O_z)} \tag{13}$$

因此，我們最小化式(12)即是最大化下列函數：

$$\hat{F}_{MMI}(\Lambda)=\sum_{z=1}^{Z}\log\frac{p_{\Lambda}(O_z,W_{zR})}{p_{\Lambda}(O_z)}=\sum_{z=1}^{Z}\log\frac{p_{\Lambda}(O_z,W_{zR})}{\sum_{W_{zi}}p_{\Lambda}(O_z,W_{zi})} \tag{14}$$

接下來，我們將探討式(14)與其它基於不同準則鑑別式函數的比較。若對式(14)中目標函數 $\hat{F}_{MMI}(\Lambda)$ 取指數函數，則目標函數變成：

$$\tilde{F}_{MMI}(\Lambda)=\exp\left[\hat{F}_{MMI}(\Lambda)\right]=\prod_{z=1}^{Z}\frac{p_{\Lambda}(O_z,W_{zR})}{\sum_{W_{zi}}p_{\Lambda}(O_z,W_{zi})} \tag{15}$$

值得注意的是，因為在指數函數是單調遞增轉換，經此函數轉換後的新目標函數 $\tilde{F}_{MMI}(\Lambda)$ 在極大值時對應的參數是不變的，亦即 $\tilde{F}_{MMI}(\Lambda)$ 與 $\hat{F}_{MMI}(\Lambda)$ 在各自極大值點有相同參數解集合。我們亦可將 $p_{\Lambda}(W_{zR}\mid O_z)$ 作改寫，以便於與其他鑑別式訓練準則做比較：

$$\begin{aligned}p_{\Lambda}(W_{zR}\mid O_z)&=\frac{p_{\Lambda}(O_z,W_{zR})}{\sum_{W_{zi}}p_{\Lambda}(O_z,W_{zi})}=1-\sum_{W_{zi}\neq W_{zR}}p_{\Lambda}(W_{zi}\mid O_z)\\&=1-\overbrace{\sum_{W_{zi}}\left(1-\delta(W_{zi},W_{zR})\right)p_{\Lambda}(W_{zi}\mid O_z)}^{\text{期望錯誤}}\end{aligned} \tag{16}$$

最後，我們可以將式(16)視為訓練語句 z 的期望正確率，也就是等於 1 減去訓練語句 z 的期望錯誤率。值得注意是在 MCE 中為所有訓練語句期望正確率的總和而 MMI 為所有訓練語句期望正確率的連乘積，背後的意義在某種程度上皆代表最大化語音辨識器對於所有訓練語句的期望正確率。

（III）最小化音素錯誤訓練(Minimum Phone Error Training, MPE)法則：相較於最大交互資訊法則與最小化分類錯誤法則是著重於訓練語句整體(String Level)的正確率提昇，最小化音素錯誤訓練法則著重於對訓練語句較細微層級(如音素、字或詞等)的正確率提昇。例如，使用最大化音素的期望正確率作為最小化音素錯誤的目標函數，可以定義為[3]：

$$F_{MPE}(\Lambda)=\sum_{z=1}^{Z}\frac{\sum_{W_{zi}}p_{\Lambda}(O_z,W_{zi})A_{Phone}(W_{zi},W_{zR})}{\sum_{W_{zi}}p_{\Lambda}(O_z,W_{zi})} \tag{17}$$

其中 $A(W_{zi},W_{zR})$ 為訓練語句 z 的候選詞序列 W_{zi} 之原始音素正確個數，可以定義為在正確轉寫 W_{zR} 上所有音素個素減去 W_{zi} 產生插入、刪除、替換等錯誤個數，MPE 目標函數希望能增進語音辨識器對於訓練語料所有辨識輸出(也就是所有候選詞序列，可包括正確轉寫)的期望音素正確率

$$F_{MPE}(\Lambda)=\sum_{z=1}^{Z}\sum_{W_{zi}}P_{\Lambda}(W_{zi}\mid O_z)A_{Phone}(W_{zi},W_{zR}) \tag{18}$$

其中

$$P_{\Lambda}(W_{zi}\mid O_z)=\frac{p_{\Lambda}(O_z,W_{zi})}{p_{\Lambda}(O_z)}=\frac{p_{\Lambda}(O_z,W_{zi})}{\sum_{W_{zi}}p_{\Lambda}(O_z,W_{zi})} \tag{19}$$

是候選詞序列W_{zi}的事後機率。$A_{Phone}(W_{zi},W_{zR})$可以定義為正確轉寫上音素的個數減去插入、刪除和替換錯誤；同樣地，藉由修改$A_{Phone}(W_{zi},W_{zR})$，我們可以定義最小化詞錯誤訓練(Minimum Word Error Training, MWE)。

以上三種鑑別式訓練方法的目標函數，如式(11)、(15)與(18)所示，在經過適當地推導(與化簡)後可以得到相同型式的目標函數[13]，而我們主要可以從兩種不同角度來分析其中差異；首先在訓練語句只有一句時，MMI 是一種希望透過提昇給定語句與聲學模型對假設空間上詞序列的期望正確率來增進辨識率，如式(16)所示；MCE 是以平滑錯誤分類評估函數來評估辨識器的辨識結果，目的希望藉由評估結果來減少辨識器的分類錯誤來降低錯誤率，如式(6)、(7)與(10)所示，來最小化辨識器的分類錯誤；MPE/MWE 概念上相似於 MMI，也是尋求提昇訓練語料的期望正確率，但架構於較細微層次如音素(Phone)與詞(Word)等。其次、當訓練語句大於一句時，MMI 是最大化所有訓練語句在語句層次上(Sentence or String Level)期望正確率的乘積，如式(15)所示；MCE 為最大化平滑後所有語句功用函數的總和，如式(11)所示；而 MPE/MWE 是最大化所有訓練語句在音素(詞)層次上期望正確率的總和。

三、基於不同法則之資料選取方法及其一致性

（一）邊際方法於鑑別式聲學模型訓練

（1）最大邊際估測法則(Large-Margin Estimation)：啟蒙於最大邊際分類器，在語音辨識議題中，最大邊際估測法則的主旨在於藉由調整隱藏式馬可夫模型的參數來最大化語音辨識器的邊際，使得在訓練集中屬於正邊際的訓練語句，藉此能讓語音辨識器的一般化能力(Generalization Capability)可以獲得增進。給定某一句訓練語句z的語音特徵向量序列O_z與其正確轉寫(正確的候選詞序列)W_{zR}，最大邊際估測法的分離邊際(Separation Margin)定義為[4]：

$$\begin{aligned} d(O_z) &= p_{\Lambda}(O_z\mid W_{zR})-\max_{W_{zi}\in\mathbf{W},W_{zi}\neq W_{zR}}p_{\Lambda}(O_z\mid W_{zi}) \\ &= \min_{W_{zi}\in\mathbf{W},W_{zi}\neq W_{zR}}\left[p_{\Lambda}(O_z\mid W_{zR})-p_{\Lambda}(O_z\mid W_{zi})\right] \end{aligned} \tag{20}$$

其中 **W** 為語音辨識器產生的所有可能候選詞序列所成集合，$p_{\Lambda}(O_z\mid W_{zR})$為給定正確詞序列$W_{zR}$產生語音特徵向量序列$O_z$的相似度，$W_{zi}$為語音特徵向量序列$O_z$的可能辨識結果(候選詞序列)，$p_{\Lambda}(O_z\mid W_{zi})$為候選詞序列$W_{zi}$產生語音特徵向量序列$O_z$的相似度。由式(20)可以知道分離邊際的計算就是正確轉寫與最有可能(機率最大)的候選詞序列的相似度之差。而當語音辨識器乃是建構在最大化事後機率解碼方法上時(這裡假設$P_{\Lambda}(W_{zR})$與每一個$P_{\Lambda}(W_{zi})$都是相等的)，若$d(O_z)>0$，則表示語音特徵向量序列O_z被目前的辨識器正確地辨識；反之，若$d(O_z)<0$，則此語音特徵向量序列O_z會被此辨識器錯誤地辨識。然而，當$d(O_z)=0$，則表示語音特徵向量序列O_z既可能會被辨識正確，亦可能被

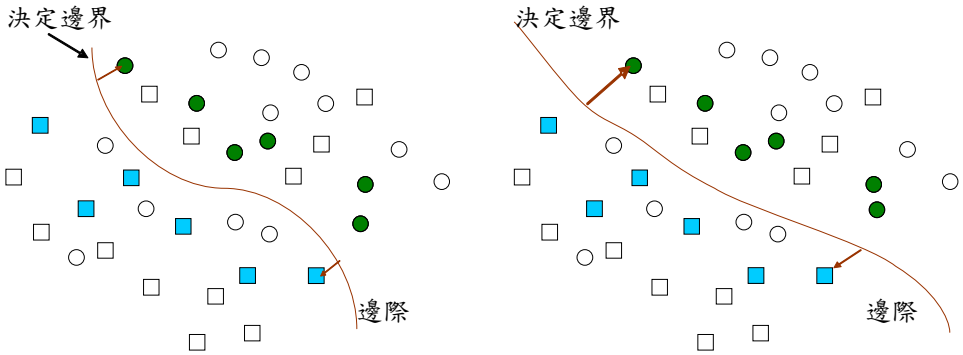

圖 1. 最大邊際估測法(左)與調整後最大邊際估測法(右)示意圖

辨識錯誤，全仰賴辨識器如何實作。

因此分離邊際 $d(O_z)$ 可以視為在相似度定義域中的一項決策要素，用以決定 O_z 是否被正確辨識($d(O_z)>0$)抑或被錯誤辨識($d(O_z)<0$)，而決定邊界自然就是當 $d(O_z)=0$ 首先，必須先找出在目前辨識器上的最小邊際，為此，可先定義一個子集合 $\mathbf{S}$ ：

$$\mathbf{S}=\{O_z \mid O_z \in R, 0 \le d(O_z) \le \gamma\} \tag{21}$$

其中 R 代表所有的語音特徵向量序列(訓練語句)； γ 為事先定義的門檻值，為一個大於零的正實數。此子集合稱為支持向量集合(Support Vector Set)，在此集合裡的語音特徵向量序列 O_z (或訓練語句 z)都是離決定邊界較近(小於 γ)且可以被正確地辨識出的語音特徵向量序列(訓練語句)，可稱之為支持樣本(Support Tokens)，如圖 1 中實心圖案所示。決定目前的最小邊際(定義於支持向量集合中)之後，最大邊際估測法則即以最大化此最小邊際為目標來進行聲學模型訓練，如：

$$\overline{\Lambda}=\arg\max_{\Lambda}\min_{O_z\in\mathbf{S}} d(O_z) \tag{22}$$

其中 Λ 與 $\overline{\Lambda}$ 分別為連續密度隱藏式馬可夫模型訓練前與訓練後的參數，將式(20)帶入式(22)，則最大邊際估測的目標即為：

$$\overline{\Lambda}=\arg\max_{\Lambda}\min_{O_z\in\mathbf{S},W_{zi}\in\mathbf{W},W_{zi}\neq W_{zR}}\left[p(O_z \mid W_{zR})-p(O_z \mid W_{zi})\right] \tag{23}$$

但必需滿足以下條件：

$$p(O_z \mid W_{zR})-p(O_z \mid W_{zi})>0 \tag{24}$$

式(23)可以轉換為標準的約束型『最小最大』最佳化問題(Constrained Minimax Optimization Problem)：

$$\overline{\Lambda}=\arg\min_{\Lambda}\max_{O_z\in\mathbf{S},W_{zi}\in\mathbf{W},W_{zi}\neq W_{zR}}\left[p(O_z \mid W_{zi})-p(O_z \mid W_{zR})\right] \tag{25}$$

則約束條件為

$$p(O_z \mid W_{zi}) - p(O_z \mid W_{zR}) < 0 \tag{26}$$

因此，最大邊際估測法則的目標函數可表示為

$$Q(\Lambda) = \max_{O_z \in \mathbf{S}, W_{zi} \in \mathbf{W}, W_{zi} \neq W_{zR}} \left[p(O_z \mid W_{zi}) - p(O_z \mid W_{zR}) \right] \tag{27}$$

而過去從事此種以最大化邊際估測為主題的研究，主要是在琢磨於目標函式的最佳化方法。相關的最佳化方法[15]在近年陸續提出，如在鑑別式訓練中廣泛的被應用至求取目標函式解的一般化機率遞減(Generalized Probabilistic Descent, GPD)，一般化機率遞減被用來估測模型參數 $\overline{\Lambda}$ ；接著，約束聯合優化方法(Constrained Joint Optimization)提出；或者，藉由一些逼近方法，將最大化邊際估測的目標函式轉換為凸函數(Convex Function)最佳化的問題，使用半正定(Semi-Define Programming, SDP)來求得模型參數 $\overline{\Lambda}$ ，並且有鑒於改善半正定規劃法訓練速度，二次圓錐規劃模式(Second Order Cone Programming, SOCP)被用來改善最大邊際估測法則訓練速度。

而針對一個分類器來說，最大邊際估測法則所強調的訓練重點為那些被認為可以正確的訓練語句(或樣本)。而那些不滿足最大邊際估測法則限制條件(即可能是辨識或分類錯誤)的訓練語句將被最大邊際估測法則排除在訓練之外。但是，這樣一來所衍生的問題是，這些訓練語句對語音辨識器(分類器)來說也是關鍵的，可以提供聲學模型訓練所需的鑑別資訊。其次，在經最大邊際估測法則的模型訓練後，訓練語句的分離邊際變大且可容錯的能力也變大，代表聲學模型更具有一般化能力。然而，實際上的語音辨識器多半是無法完全正確分類訓練語句的(也就是說辨識器對於訓練語料的辨識率尚未臻完美)，尤其是當最大邊際估測法則應用於大詞彙連續語音辨識任務上時，這會造成存在於支持向量集合內的樣本個數非常有限，使得於聲學模型調整後對整體辨識率提昇的影響不大。於是，有所謂柔性之最大邊際估測法則(Soft-Large Margin Estimation, Soft-LME)被提出以改善此一問題[17]，同時將所有分類正確與錯誤的訓練語句全部納入考量，並使用半正定規劃法加以最佳化。

（2）柔性邊際估測法則(Soft Margin Estimation)：有鑑於前述最大邊際估測法則忽略訓練樣本所造成決定邊際附近資訊的遺失，一味的最大化分類邊際可能使得所訓練出來的語音辨識器(分類器)一般化的能力不足，因此有以柔性邊際為訓練法則被提出。它的觀點是，語音辨識器對測試語料的辨識(分類)錯誤已被證明存在一定的機率之上，會受限於一個上界，而此上界為分類器之所謂的經驗風險(Empirical Risk)加上另一項一般化量值(Generalization or Regularization Term)[18]。因此，欲加強語音辨識器在測試語料上的辨識率，除了要最大化此語音辨識器的邊際外，更要同時降低其在訓練語料上的經驗風險。將這兩個面向的資訊整合在一起，則柔性邊際估測法則的目標函數可表示為[5]：

$$L^{SME}(\Lambda) = \frac{\eta}{\rho} + \frac{1}{Z}\sum_{z=1}^{Z} l(O_z, \Lambda) \tag{28}$$

其中 Λ 為隱藏式馬可夫模型的參數； ρ 為柔性邊際(Soft Margin)； η 為一常數，用來平

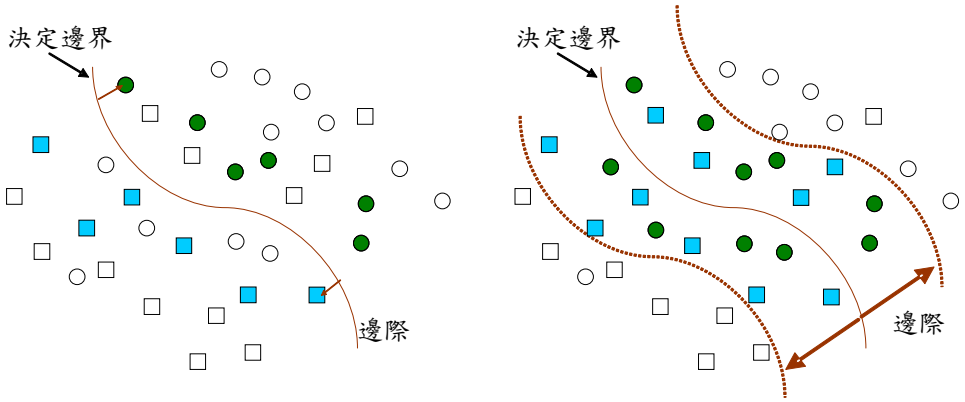

圖 2. 最大邊際估測法(左)與柔性邊際估測法(右)選取資料示意圖

衡柔性邊際的最大化與經驗風險的最小化，顯而易見地，當η越小則越強調此分類器的經驗風險；Z為所有的訓練語句個數；$l(O_z,\Lambda)$為一語句O_z的減損函數。因此，柔性邊際估測法則便藉由最小化此目標函數來降低辨識器於測試語料上的分類錯誤。與傳統鑑別式聲學模型訓練最大的不同處是，傳統鑑別式訓練只專注於經驗風險(即為$1/Z\sum_{z=1}^{Z} l(O_z,\Lambda)$)的最小化，而忽略一般化量值(以$\eta/\rho$來近似)。然而，在最大邊際估測法則卻只把重點放在加大邊際以降低一般化量值，而忽略了經驗風險的影響。

柔性邊際估測法則中的減損函數$l(O,\Lambda)$是一項影響整體訓練的重要元件，其設計必須跟其目標(即增進分類器之邊際)有一定的關連性。故定義為：

$$l(O_z,\Lambda)=\left(\rho-d_{\Lambda}^{SME}(O_z)\right)I(O_z\in U) \tag{29}$$

其中$I(\bullet)$為一個指示函數(Indicator Function)；U為所有訓練語句的子集合，其定義為

$$U=\left\{O_i\middle|\rho-d_{\Lambda}^{SME}(O_i)>0\right\} \tag{30}$$

而$d^{SME}(O_z,\Lambda)$為O_z的分離估量(Separation Measure)，用以衡量其正確轉寫與對應之最為競爭(Most Competiting)候選詞序列在聲學分數上的差距，定義為：

$$d_{\Lambda}^{SME}(O_z)=\frac{1}{n_z}\sum_{t}\log\left[\frac{p_{\Lambda}(o_{zt}\mid W_{zR})}{p_{\Lambda}(o_{zt}\mid W_{z,c})}\right]I(o_{zt}\in F_z) \tag{31}$$

其中W_{zR}為O_z的正確轉寫、$W_{z,c}$為所有候選詞序列中相似度最大的一條，就以最大事後機率解碼原則而言，其對於正確轉寫(或詞序列)擁有最大的競爭力；F_z是W_{zR}與$W_{z,c}$對應(Aligned)到訓練語句z後，兩者(指W_{zR}與$W_{z,c}$)含有不同音素類別(Phone Label)的時間音框t所構成之集合；n_z為F_z的元素個數；$p_{\Lambda}(o_{zt}\mid W_{zR})$與$p_{\Lambda}(o_{zt}\mid W_{z,c})$分別為給定$W_{zR}$和$W_{z,c}$後語音特徵向量$o_{zt}$的相似度(於目前的聲學模型$\Lambda$上所估測求得)。因此，只有其分離估量小於$\rho$的訓練語句才會被納入於目標函數內(其意謂著那些辨識錯誤處在一定範圍的訓練語句也會被納入考量)，而對於那些分離估量遠大於ρ的訓練語句(換句話說，即是很容易被辨識或分類正確的)則被視為不存在分類風險，對於聲學模型參數的估測

沒有影響力，所以不納入考量。以一個非線性且不可分割的二元分類器來加以說明，如圖 2 (右)所示。在決定邊界兩側的虛線到決定邊界的距離即為邊際 ρ ，針對每一個訓練樣本(其中□代表"+"類別，○代表"-"類別)，其與決定邊界的距離即為分離估量，當此分離估量小於邊際 ρ 時(實心的樣本圖例)，便會被選入於集合 U ，不管其是否被分類器給正確分類。

故將式(30)、(31)整合於式(29)，則柔性邊際估測法則的目標函數可表示為：

$$\begin{aligned} L^{SME}(\Lambda) &= \frac{\eta}{\rho} + \frac{1}{Z}\sum_{z=1}^{Z}\left(\rho - d_z^{SME}(O_z)\right) I(O_z \in U) \\ &= \frac{\eta}{\rho} + \frac{1}{Z}\sum_{z=1}^{Z}\left(\rho - \frac{1}{n_z}\sum_{t}\log\left[\frac{p_\Lambda(o_{zt}\,|\,W_{zR})}{p_\Lambda(o_{zt}\,|\,W_{z,c})}\right] I(o_{zt} \in F_z)\right) I(O_z \in U) \end{aligned} \tag{32}$$

由式(32)可知柔性邊際估測法則的目標函數有兩層選取資料(Data Selection)的涵義，其一為語句的選取(Utterance Selection)，即 $I(O_z \in U)$ ；另一為時間音框的選取(Frame Selection)，即 $I(o_{zt} \in F)$ 。

然而，在本小節上述所介紹的減損函數，皆一致地認為當訓練語句的分離估量小於預先定義的柔性邊際 ρ ，即對於模型參數調整能有貢獻。但當某一句訓練語句其分離估量非常小(遠小於零)時，它其實可能是一個異常訓練樣本(Outlier)，對於整個訓練過程無法提供任何幫助。因此，可藉由引進一個其值為負的參數 τ 將之過濾，則語句層次的減損函數可定義為[5]：

$$l(O_z, \Lambda) = \begin{cases} \rho - d_\Lambda^{SME\_U}(O_z), & \text{if } \rho > d^{SME\_U}(O_z, \lambda) > \tau \\ 0, & \text{otherwise} \end{cases} \tag{33}$$

而音框層次減損函數可定義為：

$$l(o_{zt}, \Lambda) == \begin{cases} \rho' - d_\Lambda^{SME\_F}(o_{zt}), & \text{if } \rho' > p(W_{zi}|o_{zt}) > \tau' \\ 0, & \text{otherwise} \end{cases} \tag{34}$$

（二）以強化混淆資訊為出發的邊際因子

（1）增進式最大交互資訊法則(Boosted MMI)

同樣受到最大邊際技術[6]影響，對於每一句訓練語句而言，給定假設空間(詞圖)上的候選詞序列與正確轉寫差別越大(錯誤越多)時，則該候選詞序列機率值得到權重越大在訓練時邊際得以最大化。增進給定假設空間(詞圖)上辨識錯誤較多候選詞序列的機率值來強化資料上的混淆程度，可以被詮釋為一種柔性邊際方法：在每一句訓練語句所對應給定假設空間(詞圖)上，硬性的將對應辨識結果錯誤的候選詞序列的機率值增進使其靠近決策邊際附近而被再重新考慮，這些辨識錯誤詞序列於是構成了一個柔性邊際空間。此層柔性邊際空間在意義上為藉由強化混淆資訊，使得鑑別式訓練過程中模型更新的目的除了正確文句與易混淆詞序列間鑑別式函數的分離外，同時藉由增進被錯誤辨識候選詞序列的機率值，來強調更新後模型的辨識結果不可與錯誤過多辨識候選詞序列太相近，

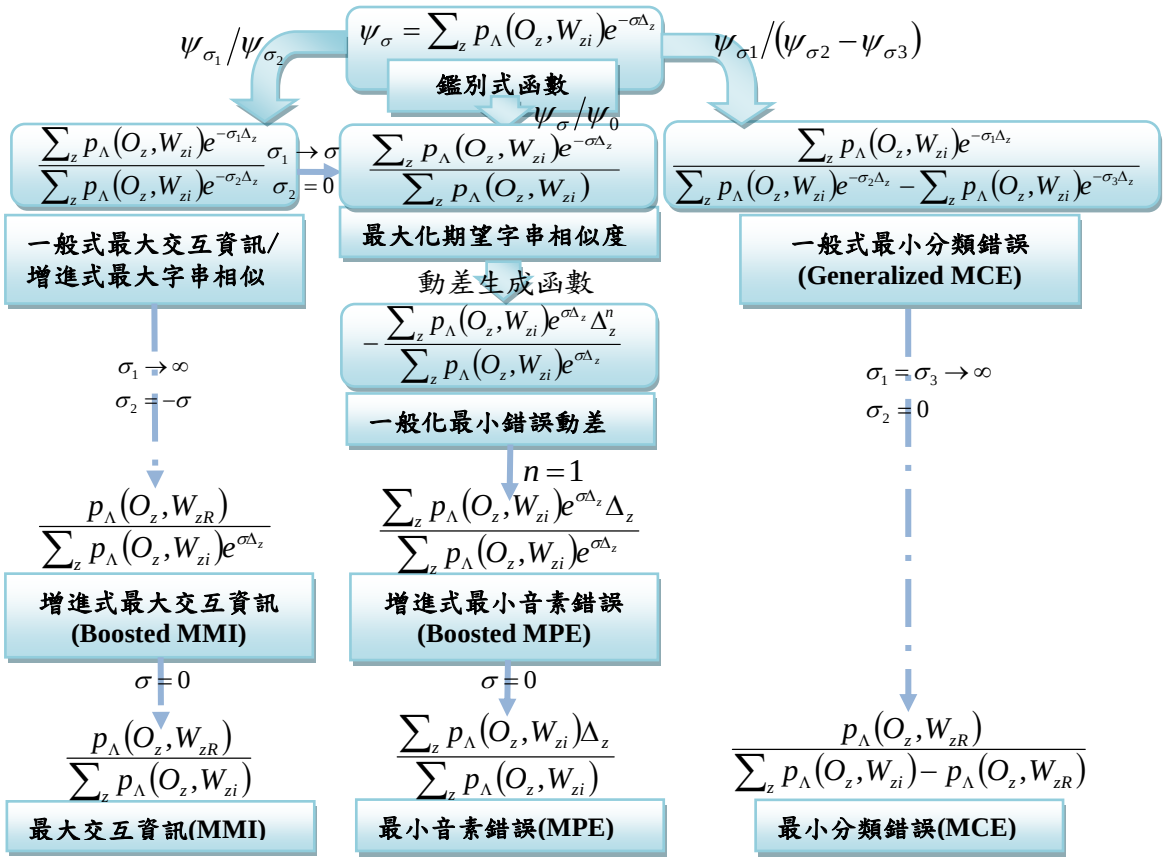

圖 3. 整合邊際鑑別式訓練方法一致性[14]

達成鑑別式訓練資料上正確與易混淆資訊邊際的最大化。以增進式最大化交互資訊法則[11]為例，其目標函數定義如下：

$$F_{\Lambda,\sigma}^{BMMI} = \sum_{z}^{Z} \log \frac{P(W_{zR})^{\eta} p_{\Lambda}(O_z|W_{zR})}{\sum_{W_{zi}} P(W_{zi})^{\eta} p_{\Lambda}(O_z|W_{zi}) e^{\sigma\varepsilon(W_{zR},W_{zi})}} \tag{35}$$

其中，$\varepsilon(W_{zR},W_{zi})$ 為給定正確轉寫 W_{zR} 反應相對應給定假設空間上候選詞序列 W_{zi} 的錯誤個數(在[11]中以正確率的負值來表示)，σ 為衰退因子(decay factor)，目的為減緩或提昇 $\varepsilon(W_{zR},W_{zi})$ 所產生影響，σ 越大越能反映正確轉寫與辨識產生之候選詞序列間的差異；η 為控制語言模型範圍因子。

考量了每一句訓練語句 z 的正確轉寫(分子項)與給定假設空間(分母項)關係。以最大化交互資訊目標函數(14)為例，最大化交互資訊所隱藏另外一層意義為 $p_{\Lambda}(W_{zR}|O_z)$ 事後機率[3]，由給定訓練語句之正確轉寫的鑑別式函數(分子)與辨識產生假設空間(詞圖)上候選詞序列的鑑別式函數(分母)所構成。緊接著我們討論增進了可能錯誤辨識的詞序列機率值在訓練上帶來的影響：在給定假設空間詞圖上，錯誤越多的詞序列將使得分母項隨著變大(因為和正確轉寫差異越多而經由 $e^{\sigma\varepsilon(W_{zR},W_{zi})}$ 增進使機率值變大)，造成目標函數值(參照式(35))變小。在訓練集中單一語句正確轉寫與其對應假設空間詞圖的關係上，詞圖上的混淆程度提昇，意味著該語句在辨識產生的詞圖詞序列上表現是較多錯誤的；由整個訓練集來看，在原本的目標函數(14)並沒有強調辨識錯誤較多詞序列，而以目標

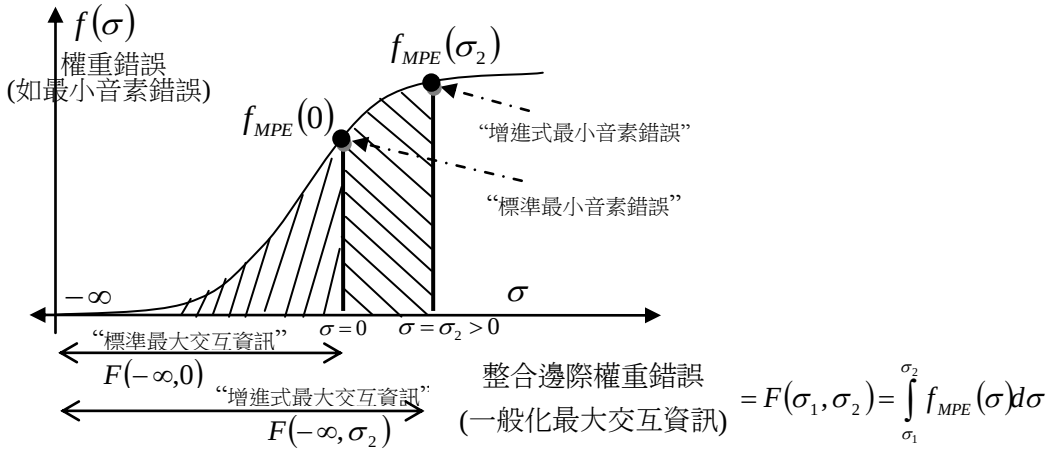

圖 4. 最大交互資訊與最小音素錯誤關係圖[12]

函數(35)增進式最大化交互資訊(也就是事後機率)作為分類標準，所想要的結果是將正確文句與分類錯誤較多(因為強化了混淆資訊使分母項變大)的詞序列分開，使得在模型的訓練上以最大化此目標函數出發做調整，但過於強化混淆資訊則反而會造成資料上容易被過度的訓練(Overtraining)。

在新近的研究上，日本電報電信公司學者同時亦將衰退因子σ詮釋為邊際參數[12]。所代表的是在固定一句語句z中，當σ越大，則分母項表示給定假設空間(詞圖)上含有錯誤越多詞序列機率值將得到越大的增進，使得至正確轉寫與其它在詞圖上辨識相較正確的分數差距縮小，產生混淆資訊；當σ越小，則反之。故邊際參數σ越大，使得詞圖上辨識錯誤較多的詞序列與其他較正確的詞序列在機率值差距變得更小(更靠近決策邊際)，邊際參數的意義正是說明了σ越大，則考慮到詞圖上辨識較錯誤的詞序列越多(因為增進的權重使得機率值提昇)。在下一節，我們將針對衰退因子(邊際參數)σ作統整性討論。

（三）以邊際資訊與錯誤為基礎的統一觀點

此一觀點以論述鑑別式訓練演進之間共通性[14]為主，如同文獻[13]，同樣以鑑別式函數為基礎，額外乘上了指數函數做為增進式權重因子，其目的如同前小節所述為了強化假設假設空間(詞圖)上混淆資訊。以帶有增進指數權重型態的鑑別式函數為出發，根據不同準則所定義目標函數，亦可以將三種方法演進整理歸納如圖 3，並且引入了增進因子權重，清楚的明白透過了調整邊際參數σ可以達到不同的目標函數。

在文獻[12]發現了最小化音素錯誤為增進式最大化交互資訊目標函數的微分，並由[7]驗證於支撐向量機中的邊際關係，有別於式(35)，增進式最大化交互資訊目標函數可改定義如下(此時為最小化整個目標函數$F_{\Lambda,\sigma}^{BMMI}$)：

$$F_{\Lambda,\sigma}^{BMMI}=-\frac{1}{\psi}\sum_{z=1}^{Z}\log\frac{P(W_{zR})^{\psi\eta}\,p_{\Lambda}(O_z|W_{zR})^{\psi}}{\sum_{W_{zi}}P(W_{zi})^{\psi\eta}\,p_{\Lambda}(O_z|W_{zi})^{\psi}\,e^{\psi\sigma\varepsilon(W_{zR},W_{zi})}} \tag{36}$$

其中ψ為一種近似程度控制目標函數的平滑[7]，對增進式最大交互資訊中邊際參數σ微分如下式：

$$\frac{\partial}{\partial\sigma}F_{\Lambda,\sigma}^{MMI}=\frac{1}{\psi}\sum_{z=1}^{Z}\frac{\partial}{\partial\sigma}\log\left(\sum_{W_{zi}}P(W_{zR})^{\psi\eta}\,p_{\Lambda}(O_z|W_{zR})^{\psi}\,e^{\psi\sigma\varepsilon(W_{zR},W_{zi})}\right)$$
$$=\sum_{z=1}^{Z}\frac{\sum_{W_{zi}}P(W_{zi})^{\psi\eta}\,p_{\Lambda}(O_z|W_{zi})^{\psi}\,e^{\psi\sigma\varepsilon(W_{zR},W_{zi})}\varepsilon(W_{zR},W_{zi})}{\sum_{W_{zi}}P(W_{zi})^{\psi\eta}\,p_{\Lambda}(O_z|W_{zi})^{\psi}\,e^{\psi\sigma\varepsilon(W_{zR},W_{zi})}} \tag{37}$$
$$=f_{\Lambda,\sigma}^{MPE}$$

關係描述如圖 4.，整合邊際錯誤(Margin-Integrated Weighted Error)是一種在最小音素錯誤所形成的空間上，對於邊際(橫軸)的範圍不同而求得的錯誤估測。

受到這層關係的啟發，日本學者定義了兩種在邊際空間上以微積分基本定理為架構的訓練目標函數。第一種作法為給定一段邊際段落區間，在此段落上求取最小化錯誤音素的積分故稱為集成式最小音素錯誤訓練(Integrated MPE, iMPE) 定義如下式：

$$F_{\Lambda,\sigma_1,\sigma_2}^{iMPE}=\int_{\sigma_1}^{\sigma_2}f_{\Lambda,\sigma}^{MPE}d_{\sigma}=F_{\Lambda,\sigma_2}^{MMI}-F_{\Lambda,\sigma_1}^{MMI} \tag{38}$$

第二種方法為集成式最小音素錯誤訓練一般化該段邊際區間稱為微分式最大交互資訊訓練(Differentiated MMI, dMMI)如下式：

$$F_{\Lambda,\sigma_1,\sigma_2}^{dMMI}=\frac{F_{\Lambda,\sigma_2}^{MMI}-F_{\Lambda,\sigma_1}^{MMI}}{\sigma_2-\sigma_1} \tag{39}$$

可以明顯的觀察到兩個目標函式皆是非常類似，兩者的差別僅在於式(39)多針對該段邊際區間一般化處理，因此集成式最小音素錯誤(iMPE)訓練較適合探討於區間較大的邊際資訊段落；微分式最大交互資訊訓練適合較小的邊際資訊區間。在這兩種以不同邊際資訊區間的方法上，集成式最小音素錯誤(iMPE)式(39)其實就是一般式最大交互資訊(可參考圖 3.)，增進式最大化交互資訊(BMMI)(35)也是其中的一個特例。而在與一般最小錯誤音素訓練的比較上，從過去試驗可以發現：某些議題上[10][11]，增進式為基礎的最大交互資訊結果是優於傳統最小音素錯誤訓練。對此種情況，可以解讀為有時只針對單一邊際頂點(最小化音素錯誤訓練)，有時需要一整段的邊際資訊(增進式的最大化交互資訊訓練)。而式(38)與(39)皆可對邊際資訊段落作可以掌握的調整，利用調整邊際資訊，來求得到最好的答案。同時，參照圖 4 我們可以察覺在最佳化過程中，σ_1 控制了對錯誤較少詞序列的權重(圖 4 左半部)、σ_2 則相反地對錯誤較多詞序列掌握權重，而兩種權重即是代表了對資料鑑別性的描述。

（四）從邊際資訊方法到增進式權重因子

本論文在前面章節所提及的邊際資訊為考量，主要以邊際方法(LME and SME)與增進因子兩類為主。在 2006 年柔性邊際方法[19]首次被提出後，2008 年以增進權重方式在最大化交互資訊被實現[11]，我們對這兩種方法進行歸納探討，並嘗試結合兩種方法。首先在柔性邊際方法上，不同的評估方法可做為分離評估，為求與增進式最大交互資訊法則(BMMI)為比較，以最大交互資訊法則作為分離評估的依據。我們嘗試將柔性邊際方法與增進式因子方法做連結，得到整理過後的柔性邊際法則於最大交互資訊法則如下

式：

令 $\Delta = \frac{\lambda}{\sigma} + \sigma$ 為一常數，則

$$L^{SME}(\rho,\Lambda) = \frac{\lambda}{\rho} + \frac{1}{Z}\sum_{z=1}^{Z}\left(\rho - d_{\Lambda}(O_z)\right)_{+} = -\frac{1}{U}\sum_{z=1}^{U}\log\left(\frac{p_{\Lambda}(O_z \mid W_{zR})P(W_{zR})}{\sum_{W_{zi}} p_{\Lambda}(O_z \mid W_{zi})P(W_{zi})e^{\Delta}}\right) \tag{40}$$

從上式結論可知(推導過程請參見附錄)，柔性邊際法則為一個集合 U 經過錯誤函數檢查過濾而得，轉換後的關鍵不同在於柔性邊際方法 σ 為固定值使得 Δ 也是固定值，使得在不同的分離評估上，我們可以輕易的將邊際資訊整合入評估方法中，而這兩類方法雖皆以柔性邊際為考量，但在物理意義上有不同含意。柔性邊際法則[5]是架構於訓練語句資料集合上，每一句訓練語句的評估值(如本論文以最大交互資訊法則為評估)；然而，增進因子隱含的柔性邊際意義所代表的是在固定單一訓練語句與其相對應辨識所產生給定假設空間詞圖上。在詞圖上，增進那些錯誤較多詞序列的鑑別式函數，使得與辨識較正確詞序列的鑑別式函數值差距縮小，造成了詞圖上的混淆資訊。但是增進因子在整個訓練集合的架構上產生的影響，卻是將辨識錯誤較多之語句的事後機率降低(分母因為增進而變大)，因為在詞圖上增進辨識錯誤的詞序列鑑別式函數，造成該訓練語句在事後機率的評估上降低。故在以整個訓練語句集合來看，柔性邊際法則與增進式因子雖然皆標榜柔性邊際，但在意義上，是截然不同的。

吾人嘗試將上述兩類邊際方法做結合：在式(40)為基礎下，利用式(36)分母項中 $\sigma\varepsilon(W_{zR},W_{zi})$ 來動態地給予式(40)的分母項中每一條候選詞序列之不同貢獻權重

$$L^{HMMI}(\rho,\Lambda) = \frac{\lambda}{\rho} + \frac{1}{Z}\sum_{z=1}^{Z}\left(\rho - d_{\Lambda}(O_z)\right)_{+} = -\frac{1}{U}\sum_{z=1}^{U}\log\left(\frac{p_{\Lambda}(O_z \mid W_{zR})P(W_{zR})}{\sum_{W_{zi}} p_{\Lambda}(O_z \mid W_{zi})P(W_{zi})e^{\sigma\varepsilon(W_{zR},W_{zi})}}\right) \tag{41}$$

此方法(往後將被稱做 Hybrid MMI)優點在於：利用語句層級柔性邊際法則[5]，更有彈性的控制了所專注訓練的語料；加入增進式最大化交互資訊訓練精神來強化了辨識錯誤候選詞序列，將使得辨識正確與錯誤語句差距更明顯。這樣一來可以避免因為在過度強化訓練語料的貢獻而造成過度訓練的問題，而同時可以使用增進式因子來能提昇語音辨識率的效能，因而保留了上述兩種方法的優點。在下面的實驗中，我們將以最大化交互資訊方法(MMI)為例，分別討論增進式最大化交互資訊法則與柔性邊際資料選取方法，以及它們的結合之實驗結果。

四、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為臺灣師範大學語音實驗室所發展的新聞語音

CER(%)	MMI	SME-utter $\tau = -41$	SME-utter $\tau = -43$	SME-utter $\tau = -45$
Baseline	23.64			
Itr01	23.30	23.27	23.25	23.18
Itr02	22.95	22.94	22.95	22.92
Itr03	22.74	22.46	22.55	22.68
Itr04	22.40	22.37	22.41	22.31
Itr05	22.14	22.22	22.33	22.23
Itr06	22.12	22.19	22.13	22.19
Itr07	21.99	22.15	21.98	21.90
Itr08	21.89	21.94	22.00	21.88
Itr09	21.84	22.04	21.93	21.89
Itr10	21.80	22.00	21.92	21.80

表 1. 語句層級柔性邊際估測法實驗結果

辨識系統[20]，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)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)[21]。且在執行鑑別分析之後還額外使用最大化相似度線性轉換(Maximum Likelihood Linear Transform, MLLT)[22]，其目的是為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)共變異矩陣。同時，為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為中文INITIAL-FINAL的統計模型，而每個模型的狀態數分別為3至6個不等，每個狀態皆為高斯混合分布，其中每個高斯混合分布的分布個數分別為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)個中文字語料作為背景語言模型訓練時的訓練資料[23]。在本文中的語言模型使用了Katz語言模型平滑技術[24]，在訓練時是採用SRI Language Modeling Toolkit (SRILM)[25]。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[27]，則採用詞三連語言模型(Trigram Language Model)。

（二）實驗語料

本論文實驗使用的訓練與測試語料為MATBN電視新聞語料庫[26]，是由中央研究院資訊所口語小組耗時三年與公共電視台合作錄製完成。我們初步地選擇採訪記者語料作為實驗語材，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；評估集則均為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

CER(%)	MMI	BMMI	BMMI-d	MPE
Itr01	23.30	22.98	22.92	22.89
Itr02	22.95	22.54	22.53	22.45
Itr03	22.74	22.20	22.20	22.26
Itr04	22.40	21.97	21.85	21.85
Itr05	22.14	21.66	21.48	21.45
Itr06	22.12	21.36	21.39	21.15
Itr07	21.99	21.36	21.29	21.15
Itr08	21.89	21.37	21.16	20.98
Itr09	21.84	21.35	21.10	20.94
Itr10	21.80	21.35	20.98	20.83

表 2. 語句層級柔性邊際估測法實驗結果

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為比對單位。令H為正確轉譯詞序列與辨識詞序列比對後相同字的個數(Hit)、I為辨識詞序列多餘插入字元的個數(Insertion)、N為正確轉寫中詞序列的字總數，則語音辨識系統之正確率(Accuracy)的計算方式為$\frac{H-I}{N}\times 100\%$，錯誤率(Error Rate)則為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

在基礎實驗中，我們先使用最大化相似度(ML)估測法訓練10次，所得到的字錯誤率(CER)為23.64%(記作Baseline)。接著，分別進行最大化交互資訊(MMI)最小化音素錯誤(MPE)訓練10次，I-平滑的在MMI與BMMI設定為100[11]、MPE設定為10[3]，最後最大化交互資訊(MMI)所得到的字錯誤率(CER)為21.80%，最小化音素錯誤(MPE)所得到的字錯誤率(CER)為20.83%。故於接下來的實驗中，將以這組實驗為比較對象。

（五）柔性邊際方法於語句選取實驗結果

藉由調整柔性邊際參數對資料選取產生影響，在此實驗中我們特別觀察在可能辨識錯誤較多語句產生影響，配合式(35)，$\rho=-30$，而調整τ觀察辨識錯誤較多語句產生結果如表 1.，在本實驗中，可以從結果發現到在十次訓練中，在多數的情況下，當$\tau=-45$時，使用有較佳的結果；τ越來越小所代表的意義是納入了越多可能辨識錯誤的語句，因為較具鑑別資訊，可以觀察部分結果錯誤率有稍微的降低，語句層級分類略微粗淺，故是較難以反應實驗結果上。嚴格來看，在語句選取上的改善率並不大。但對於所要選取增進鑑別式訓練的資料是有正面的影響。

（六）增進式最大交互資訊法則(BMMI)實驗結果

CER(%)	BMMI	Hybrid-MMI	BMMI-d	Hybrid-MMI-d	MPE
Itr01	22.98	22.93	22.92	22.95	22.89
Itr02	22.54	22.69	22.53	22.51	22.45
Itr03	22.20	22.24	22.20	22.27	22.26
Itr04	21.97	21.97	21.85	21.88	21.85
Itr05	21.66	21.67	21.48	21.60	21.45
Itr06	21.36	21.51	21.39	21.40	21.15
Itr07	21.36	21.25	21.29	21.23	21.15
Itr08	21.37	21.11	21.16	21.14	20.98
Itr09	21.35	21.17	21.10	20.99	20.94
Itr10	21.35	21.21	20.98	21.03	20.83

表 3. 結合柔性邊際估測法與增進式因子實驗結果

本小節呈現增進式最大交互資訊訓練實驗，結果如表 2 所示。增進式最大交互資訊對於邊際參數σ固定值($\sigma=0.5$，以 BMMI 表示)[11]與漸進遞減的調整($\sigma=1\sim0.1$，以 BMMI-d 表示)，實驗的結果都比最大化交互資訊法則好。值得注意的是，BMMI 在第六次訓練後，錯誤率呈現了收斂的趨勢，所代表的意思是增進了錯誤多的語句造成詞圖上混淆，但這些錯誤語句若被過度的重視，反而容易被該群資料所支配，這個問題在 BMMI-d 就不會較不易發生，原因是 BMMI-d 採用可調整的邊際參數σ來動態增進可能辨識錯誤的訓練語句，給予的權重隨訓練次數漸進遞減；由訓練語句分布權重的角度來觀察，在訓練初期，對於錯誤較多的訓練語句給予較大權重，象徵著重視那些可能辨識錯誤的語句在模型更新帶來的影響，詞圖上錯誤較多的語句因為也詞圖混淆(分母項變大)導致降低了事後機率值。簡而言之，漸進遞減增進式最大交互資訊訓練(BMMI-d)在一開始所專注的是錯誤較大訓練語句(代表較靠近其他類別)。在訓練初期時，能提供較多混淆資訊的訓練語句在訓練產生較大的貢獻，確立了模型的調整趨勢；然後，再漸漸的降低邊際參數σ，使錯誤較少的語句(較不具鑑別資訊)開始產生影響，確保資訊的充分利用。我們可以從實驗結果的走勢看出，在一開始時改善率較大主因是對於決策邊際間資料透過邊際參數σ更明確的存在，後期訓練次數增加時，因為慢慢納入較能正確辨識的訓練語句，使得錯誤率有慢慢的降低。

（七）結合柔性邊際與增進因子於最大交互資訊法則實驗結果

結合柔性邊際與增進因子於最大交互資訊法則實驗結果如表 3 所示，$\tau=-43$，從 Hybrid-MMI 中發現不管是較原本柔性邊際或是增進因子方法，都有所改善。更使得原本 BMMI 中在固定邊際參數σ容易被支配(Overfitting)的問題得到解決，解決的原因是透過了柔性邊際方法資料選取對於辨識錯誤較多的語句有了限制，雖然它們可能是較具鑑別資訊的資料也有可能是離群值(Outlier)。但也因此，在實驗數據上前五次訓練中，Hybrid 的方法改善率稍低於 BMMI，但在整體看來，Hybrid-MMI 有著較好的結果。

而在 Hybrid-MMI-d 的漸進遞減方法上與 BMMI-d 結果差不多，但可以得到較快的收斂且使用資料量亦減少；然而在後續改善的方法上，可以搭配柔性邊際參數的設定，與增進因子中邊際參數搭配，使得訓練的資料更能有效的發揮。

五、結論

本論文從語音辨識解碼原則出發，論述目前主流鑑別式聲學模型訓練之目標函數設計基礎；說明它們雖由不同考量的出發而定義了不同的目標函數，但經過若干數學推導與假設可得到近似的共通函數格式。本論文針對語音辨識結果所形成的假設空間上所觀察到錯誤(或正確)率的不同細緻層度，在模型訓練時引入了機器學習領域中的邊際概念；其背後的物理意義，事實上就是從不同層級的訓練語料中選取適合的資訊供聲學模型訓練所使用。

參考文獻

[1] B. H. Juang, W. Chou, and C. H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Trans. Speech and Audio Processing*, Vol. 5, no. 3, pp. 257–265, 1997.

[2] Y. Normandin, *Hidden Markov models, maximum mutual information estimation, and the speech recognition problem*, Ph.D. Dissertation, McGill University, Montreal, 1991.

[3] D. Povey, *Discriminative training for large vocabulary speech recognition*. Ph.D Dissertation, Peterhouse, University of Cambridge, July 2004.

[4] H. Jiang, X. Li and C.-J. Liu, "Large margin hidden Markov models for speech recognition," *IEEE Trans. on Audio, Speech and Language Processing*, pp. 1584-1595, Vol. 14, No. 5, September 2006.

[5] J. Li, *Soft margin estimation for automatic speech recognition*. Ph.D. Dissertation, Electrical and Computer Engineering, Georgia Institute of Technology, July 2008.

[6] Fei Sha, *Large margin training of acoustic models for speech recognition*. Ph.D. Dissertation, University of Pennsylvania. 2007.

[7] G. Heigold, T. Deselaers, R. Schlüter, and H. Ney. "Modified MMI/MPE: A Direct Evaluation of the Margin in Speech Recognition," in *Proc. ICML*, pp. 384-391, 2008.

[8] D. Yu, L. Deng, X. He, and A. Acero, "Large-margin minimum classification error training: A theoretical risk minimization perspective," *Computer Speech and Language*, Vol. 22, No. 4 pp. 415-429, October 2008.

[9] B. Chen, S. -H. Liu, and F.- H. Chu, "Training data selection for improving discriminative training of acoustic models," *Pattern Recognition Letters*, Vol. 30, No. 13, pp. 1228-1235, October 2009.

[10] G. Saon and D. Povey, "Penalty function maximization for large margin HMM training," in *Proc. Interspeech*, pp. 920–923, 2008.

[11] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for model and feature space discriminative training," in *Proc. ICASSP*, pp. 4057-4060, 2008.

[12] E. McDermott, S. Watanabe, and A. Nakamura: "Margin-space integration of MPE loss via differencing of MMI functionals for generalized error-weighted discriminative training," in *Proc. Interspeech*, pp. 224–227, 2009.

[13] Xiaodong He, Li Deng, and Chou Wu, "Discriminative Learning in Sequential Pattern Recognition --- A Unifying Review for Optimization-Oriented Speech Recognition," in *IEEE Signal Processing Magazine*, vol. 25, No. 5, pp. 14-36, September, 2008.

[14] A. Nakamura, E. McDermott, S. Watanabe, and S. Katagiri, "A unified view for discriminative objective functions based on negative exponential of difference measure between strings," in *Proc. ICASSP* ,pp. 1633-1636 ,2009.

[15] H. Jiang, "Discriminative training for automatic speech recognition: A survey," *Computer and Speech, Language*, pp. 589-608, Vol. 24, No. 4, October 2010.

[16] R. Schlüter, W. Macherey, B. Müller, and H. Ney, "Comparison of discriminative training criteria and optimization methods for speech recognition," *Speech Communication*, Vol. 34, pp. 287-310, May 2001.

[17] H. Jiang and X. Li, "Incorporating training errors for large margin HMMs under semi-definite programming framework," in *Proc. ICASSP*, pp. 629-632, 2007.

[18] V. Vapnik, *The nature of statistical learning theory.* New York: Springer-Verlag, 1995.

[19] J. Li, M. Yuan, and C. -H. Lee, "Soft margin estimation of hidden Markov model parameters," in *Proc. Interspeech*, pp. 2422-2425, 2006.

[20] B. Chen, J. W. Kuo and W. H. Tsai, "Lightly Supervised and Data-Driven Approaches to Mandarin Broadcast News Transcription," in *Proc. ICASSP*, p777-780, 2004.

[21] 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.

[22] R. A. Gopinath, "Maximum likelihood modeling with Gaussian distributions," in *Proc. ICASSP*, pp. 661-664, 1998.

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

[24] 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.

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

[26] H.-M. Wang, B. Chen, J.-W. Kuo, and S.-S. Cheng, "MATBN: A Mandarin Chinese broadcast news corpus," *International Journal of Computational Linguistics & Chinese Language Processing*, Vol. 10, No.2, pp. 219-236, 2005.

[27] 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.

附錄　式(40)推導過程如下：

$$
\begin{aligned}
L^{SME}(\rho,\Lambda) &= \frac{\lambda}{\rho}+\frac{1}{Z}\sum_{z=1}^{Z}\left(\rho-d_{\Lambda}(O_z)\right)_{+} \\
&= \frac{\lambda}{\rho}+\frac{1}{Z}\sum_{z=1}^{Z}(\rho-\log\frac{p_{\Lambda}(O_z\mid W_{zR})P(W_{zR})}{\sum_{W_{zi}}p_{\Lambda}(O_z\mid W_{zi})P(W_{zi})})I(\mathrm{O}_z\in U) \\
&= \frac{1}{U}\sum_{z=1}^{U}(-\log p_{\Lambda}(O_z\mid W_{zR})P(W_{zR})+\log\sum_{W_{zi}}p_{\Lambda}(O_z\mid W_{zi})P(W_{zi})+\frac{\lambda}{\rho}+\rho) \\
&= \frac{1}{U}\sum_{z=1}^{U}(-\log p_{\Lambda}(O_z\mid W_{zR})P(W_{zR})+\log\sum_{W_{zi}}p_{\Lambda}(O_z\mid W_{zi})P(W_{zi})+\log e^{\Lambda}) \\
&= -\frac{1}{U}\sum_{z=1}^{U}\log(\frac{p_{\Lambda}(O_z\mid W_{zR})P(W_{zR})}{\sum_{W_{zi}}p_{\Lambda}(O_z\mid W_{zi})P(W_{zi})e^{\Lambda}})
\end{aligned}
$$