國立臺灣師範大學

資訊工程研究所碩士論文

指導教授：陳柏琳 博士

改善鑑別式聲學模型訓練於中文連續語音辨識之研究

Improved Discriminative Training of Acoustic Models for Mandarin Continuous Speech Recognition

研究生： 劉士弘 撰

中華民國 九十六 年 七 月

摘要

本論文探討改善鑑別式聲學模型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時間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，本論文提出一個新的以時間音框層次正規化熵值為基礎的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。此資料選取方法更進一步地應用到非監督鑑別式聲學模型訓練上。最後，本論文也嘗試修改鑑別式訓練的目標函數，以收集不同的統計值來改進最小化音素錯誤鑑別式訓練。所使用的實驗題材是公視新聞語料。由初步的實驗結果來看，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。

Abstract

This thesis considers improved discriminative training of acoustic models for Mandarin large vocabulary continuous speech recognition (LVCSR). First, we presented a new phone accuracy function based on the frame-level accuracy of hypothesized phone arcs instead of using the raw phone accuracy function of minimum phone error (MPE) training, which to some extent can sufficiently penalize deletion errors of speech recognition. Second, a novel data selection approach based on the normalized frame-level entropy of Gaussian posterior probabilities obtained from the word lattice of the training utterance was explored for discriminative training. It has the merit of making the training algorithm focus much more on the training statistics of those frame samples that center nearly around the decision boundary for better discrimination. The proposed data selection approach was further applied to unsupervised discriminative training of acoustic models. Finally, a few other modifications of the training objective functions, as well as the lattice structures, for the accumulation of MPE training statistics were investigated. Experiments conducted on the Mandarin broadcast news corpus (MATBN) collected in Taiwan showed that the integration of the frame-level data selection and new phone accuracy function could achieve slight but consistent improvements over the conventional MPE training at lower training iterations.

誌謝

感謝父母及家人對我經濟上及精神上的支持，使我能夠專心一致研究我有興趣的主題。

非常感謝指導教授陳柏琳老師三年來對我辛苦的教導，使我從一顆小豆苗慢慢的變成現在的一顆小樹，老師對研究的熱誠讓我非常感動，這也是我還要繼續努力的地方。老師的聰明及先知卓見一直都是學生望塵莫及的，我一直謹記老師說的”勤能補拙”，希望有朝一日也能跟老師一樣在學術領域上有成就。

感謝口試委員王新民博士、劉昭麟博士及洪志偉博士對我論文的指正與建議，使得我的論文更加完善。

感謝實驗室的人瑋學長，您一直都是我的榜樣，也是我要追求的標的，謝謝您給我研究上的數學知識，同時也謝謝您不辭辛苦地幫我程式除錯。感謝文鴻學長，謝謝您有問必答。感謝耀民學長、成韋學長、志豪學長及惠銘學長，謝謝你們讓我學習到不同領域的知識。感謝士傑學長，謝謝您每次都帶來不同的好壞消息。感謝實驗室的同學們，燦輝、怡婷、炫盛及士翔，有你們的陪伴，使我在研究室的生活不無聊，一同共進退。感謝實驗室的後進們，芳輝、鴻彬、庭瑋、斯涵及鴻欣，有你們的參與，實驗室的歡樂時光真是不少。

三年的研究生活一轉眼就過了，時光匆匆，歲月不饒人，希望大家都能快快樂樂地過生活。

謝謝大家 士弘謹誌

目錄

圖目錄

表目錄

第1章　序論

1.1 研究背景

隨著科技的高度發展，資訊工業革命的快速來臨，個人電腦(Personal Computer, PC)、智慧型個人數位助理(Personal Digital Assistant, PDA)和多功能手機已經是現在這個社會所不能缺少的資訊產品；尤其是多功能手機，在台灣社會中幾乎是每個人手上都有的必需品。消費性電子產品為了攜帶上的便利，勢必要將產品的體積縮得越小越輕便才會帶來可攜性(Portable)的便利，那麼傳統的輸入方式(鍵盤、滑鼠、觸控式螢幕等)便會造成輸入上的不方便；再者，由於全世界的數位化，電腦已經深植到各大公司行號及家庭之中，人們長時間使用電腦工作，日積月累下來，常會造成肩背痠痛、電腦失寫症(Dysgraphia)等之類的文明病產生，此時，便要想辦法去改變輸入的形式才行。語音，是人與人之間最自然的溝通橋樑，若輸入的形式是語音，那麼人與科技產品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)的研究就變得更加重要，這也是目前世界上熱門的研究議題之一。

語音辨識的技術目前在大詞彙連續語音辨識(Large Vocabulary Continuous Speech Recognition, LVCSR)還無法達到百分之百的辨識率，但在小字彙(Small Vocabulary)的辨識，例如數字辨識，已經有不錯的成效，在安靜的環境下，幾乎已達百分之百的辨識率。目前語音辨識已初步應用在通訊領域上，如可應用在手機內建的語音撥號，或用來取代電話服務人員等，由於自動語音辨識系統在大詞彙辨識上還無法完全辨識正確，所以跟安全性有關的語音辨識就還不能使用，例如門禁系統。因此，未來真要到全面實用性的階段，仍有許多的問題有待語音科技研究人員一同來克服。

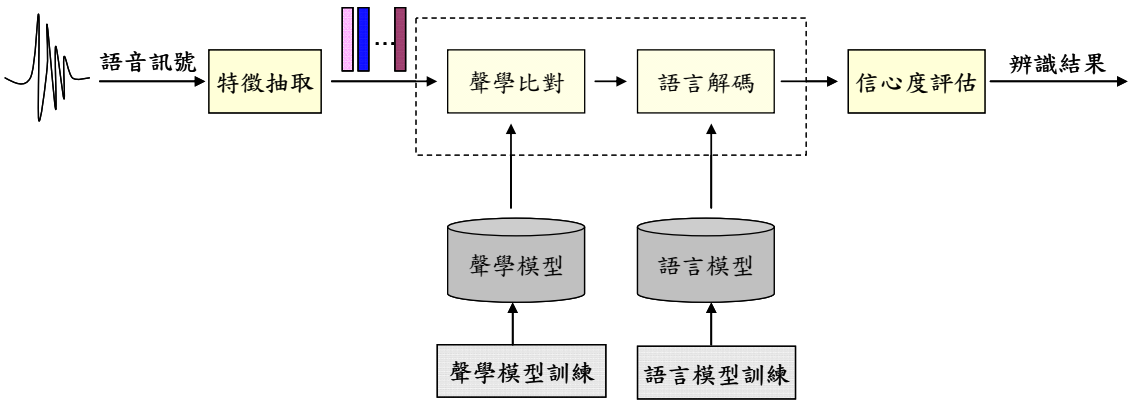


圖 1-1 統計式語音辨識系統簡易架構圖

1.2 統計式語音辨識

目前流行的語音辨識技術是以統計為基礎而建構出自動語音辨識(Automatic Speech Recognition, ASR)系統。使用統計式架構的方法可以有效率地去表達語音訊號(Speech Signals or Observations)和人工所建立的數學模型(Models)之間的關係，而人工建立的數學模型通常都是建立在一些機率分布(Probabilistic Distributions)，例如高斯(Gaussian)分布，因為使用機率分布有下列三項優點：

(1) 模型產生出的機率值可以直接被當成分數(Score)來使用。

(2) 根據不同的應用情境，機率值可以直接用來相乘或相加，也就是說可以很容易地結合分數而做出不同的運算。

(3) 機率可以很容易地表達模糊的關係，因為機率值是介於0到1之間的值，例如1代表一定有關係，0代表絕對沒有關係，介於0到1之間的值就代表模糊的關係。

在統計式語音辨識的處理過程中，一般是採用最大化事後機率(Maximum a Posteriori, MAP)來搜尋最有可能的辨識結果，如下所示:

$$\hat{W} = \arg\max_{W \in \mathbf{W}_h} P(W \mid O) \tag{1.1}$$

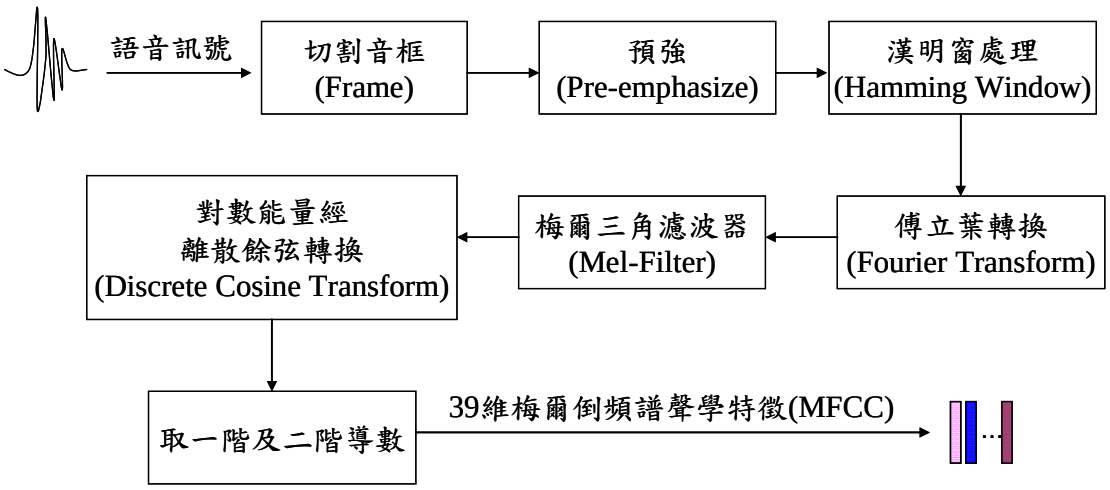


圖 1-2 梅爾倒頻譜係數特徵擷取步驟

其中 $\mathbf{W}_h$ 代表所有可能的詞序列，$P(W|O)$ 為給定語音特徵向量序列 O，產生某詞序列 W 的事後機率，$\hat{W}$ 為辨識結果(詞序列)。通常我們會用貝氏定理將式(1.1)中的事後機率展開:

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

其中 $p(O|W)$ 為給定某詞序列 W 所產生語音特徵向量序列 O 的機率或相似度(Likelihood)，一般我們會用模型來表示這個機率或相似度，由於此模型被用來決定語音特徵的機率，所以可稱之為聲學模型(Acoustic Model)。$P(W)$ 為某語言中詞序列 W 的事前機率，用來輔助聲學上的混淆，又稱為語言模型(Language Model)，$p(O)$ 為語音特徵向量序列 O 的事前機率，可視為一正規化因子(Normalization Factor)，不影響搜尋最大事後機率的詞序列，故式(1.1)可表示為:

$$\hat{W}=\arg\max_{W\in\mathbf{W}_h} p(O|W)P(W) \tag{1.3}$$

目前流行的自動語音辨識(ASR)系統大致有特徵擷取(Feature Extraction)、聲學比對(Acoustic Matching)與語言解碼(Linguistic Decoding)、聲學模型訓練、語言模型訓練以及信心度評估(Confidence Measure)五個部份，如圖 1-1 所示，下面將會

對這五個部份做一個簡略的介紹，以便了解整個語音辨識的基礎架構。

1.2.1 特徵擷取(Feature Extraction)

特徵擷取是要將人類說話時所產生類比訊號轉成自動語音辨識(ASR)系統可以處理的語音特徵向量序列(Speech Feature Vector Sequence)，也就是將類比訊號參數化(Parameteriation)。這個部份通常會透過類比數位轉換(Analog-to-digital Convert)、傅立葉轉換(Fourier Transform)及倒頻譜分析(Cepstral Analysis)擷取語音訊號中比較重要的參數。目前最具代表性的語音特徵參數為梅爾倒頻譜係數(Mel-frequency Cepstral Coefficients, MFCC)[Davis and Mermelstein 1980]，其擷取步驟如圖1-2所示。在擷取此特徵的時候，我們會將語音資料切割成一連串部份重疊的音框(Frames)，每一個音框最後表示成由12維梅爾倒頻譜係數和1維的能量特徵再加上其一階與二階的時間軸導數(Time Derivatives)所組成的39維特徵向量。其中取一階與二階時間軸導數的原因主要是為了能獲得語音特徵在時間上(Temporal)的相關資訊。因為訓練語料通常會與測試語料有不匹配(Mismatch)的問題存在，而我們也會希望所擷取的特徵係數具有強健性(Robust)，所以便有一些技術是以擷取語音訊號中較具有強健性的特徵為主要目的，使得擷取出來的特徵可以抵抗週遭的環境變化。常見的技術有倒頻譜平均消去法(Cepstral Mean Subtraction, CMS)[Atal 1974]、倒頻譜正規化法(Cepstral Normalization, CN)[Viikki and Laurila 1998]、統計圖等化法(Histogram Equalization, HEQ)[Korkmazsky *et al.* 2004; Lin *et al.* 2006]等。

除了以上強健性技術之外，還可以利用鑑別性分析(Discriminant Analysis)來計算原始語音資料的一些相關統計資訊，將原本的語音特徵投影到新的特徵空間，以得到較具有鑑別性的特徵。較常見的方法有線性鑑別分析(Linear Discriminant Analysis, LDA)[Duda *et al.* 2000]、異質性線性鑑別分析(Heteroscedastic Linear Discriminant Analysis, HLDA)[Kumar 1997]等。

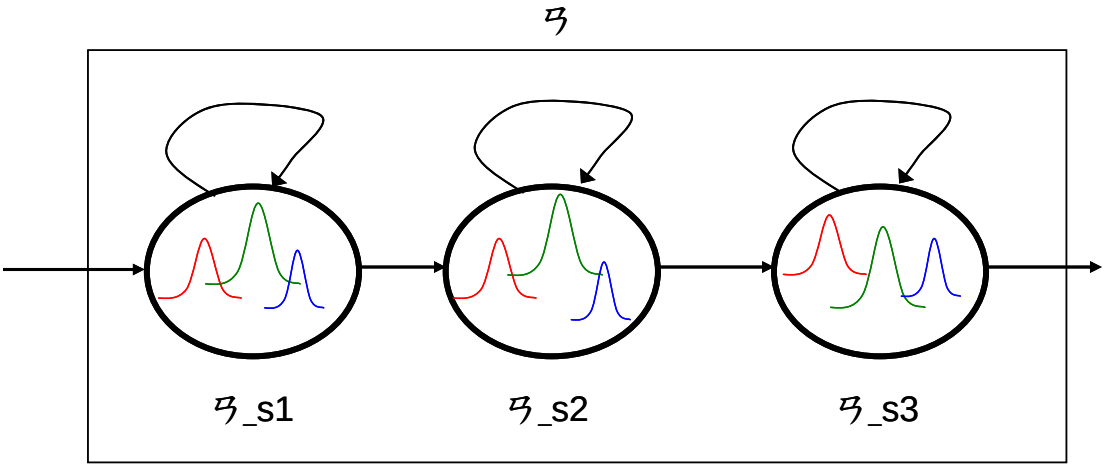


圖 1-3 連續密度隱藏式馬可夫模型示意圖

1.2.2 聲學模型(Acoustic Model)

為了處理語音訊號在時域上的變化，一般而言都是使用由左至右(Left-to-right)的連續密度隱藏式馬可夫模型(Continuous Density Hidden Markov Model, CDHMM)來作為聲學模型。如圖1-3便是一個具有三個狀態(State)的HMM模型，每個狀態中都具有高斯混合模型(Gaussian Mixture Model, GMM)分布，圖1-3中便是以三個高斯分布為例。另外，每個狀態也有相對應的狀態轉移機率(State Transition Probability)，用來控制下一個時間點要停留在自己或是轉移到下一個狀態。根據語音特徵參數是連續或非連續的值，HMM每個狀態中的觀測機率(Observation Probability)估測方式可分為離散型(Discrete)、半連續型(Semi-continuous)及連續型(Continuous)三種[Huang *et al.* 2001]，而目前的自動語音辨識系統主要都是連續型或半連續型為主。就連續型而言，為了減少估算觀測機率的參數量，也因為任何機率分布理論上皆可以由多個高斯分布(Gaussian Distributions)來逼近的特性，一般都是使用高斯混合分布(Gaussian Mixture Distributions)來近似此機率分布。而連續型與半連續型主要的差別在於在連續型中每個狀態擁有自己的高斯分布，而半連續型則會有共用高斯分布的情況。本論文是採用連續型的隱藏式馬可夫模型(HMM)，其中每個狀態有2到128個不等的高斯分布。

聲學模型在小詞彙上(如:數字辨識)，常以全詞模型(Whole-word Model)為單

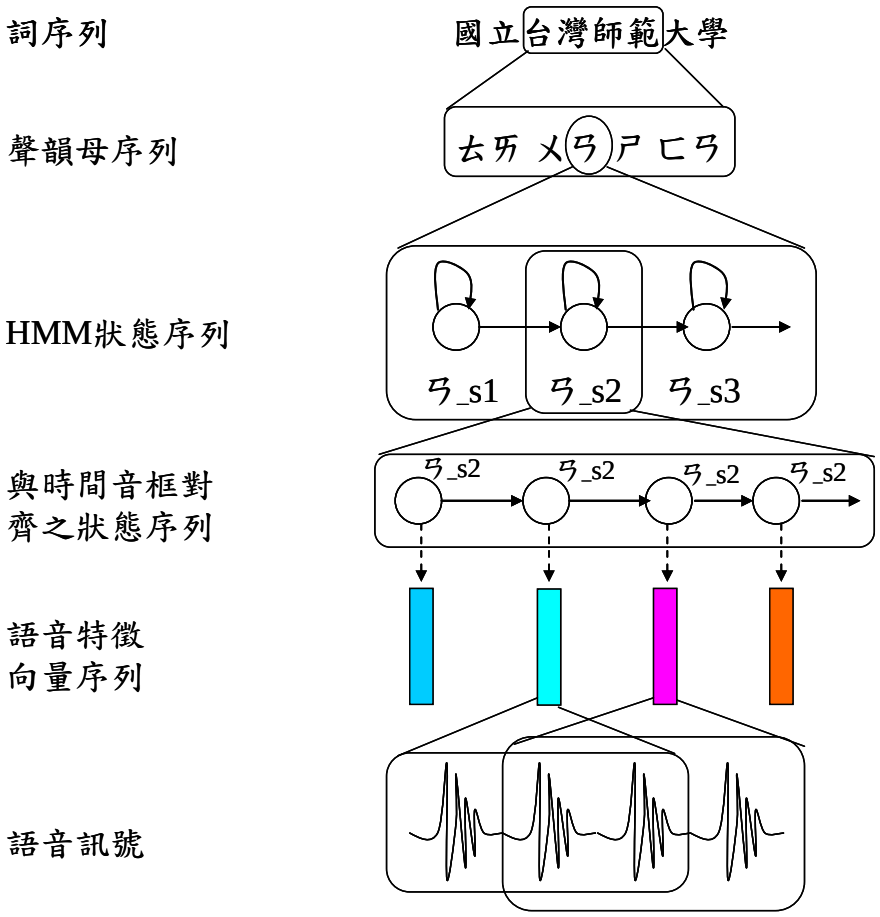


圖 1-4 聲學模型與語音訊號之階層性示意圖

位，但在中大詞彙上，因為訓練上的考量，不會對每個詞建一個聲學模型，而是以較小的單位來建模型，如:次詞單位(Sub-word Unit)、音節(Syllable)或音素(Phone)，再利用發音辭典(Pronunciation Lexicon)來串接每個聲學模型。

由於一個中文音節(Syllable)是由一個聲母(INITIAL)及一個韻母(FINAL)組成，22 個聲母及 38 個韻母構成約 400 個音節。基本上，我們只要為每個聲母及韻母建立屬於它的聲學模型，便可以辨識所有的中文音節。本論文共使用了 38 個韻母模型，但在聲母模型的部份，因為考慮到不同的右邊相連韻母對其聲母發音特性所造成不同的影響，而將 22 種聲母再細分成 112 種聲母模型，亦即聲母部份採用右相關聯模型(Right-context-dependent Model, RCD)[Chen *et al.* 2002]。另外，我們加入一個靜音(Silence)模型來估測語音訊號中靜音部份。圖 1-4 表示聲學模型與語音訊號的階層性關係。

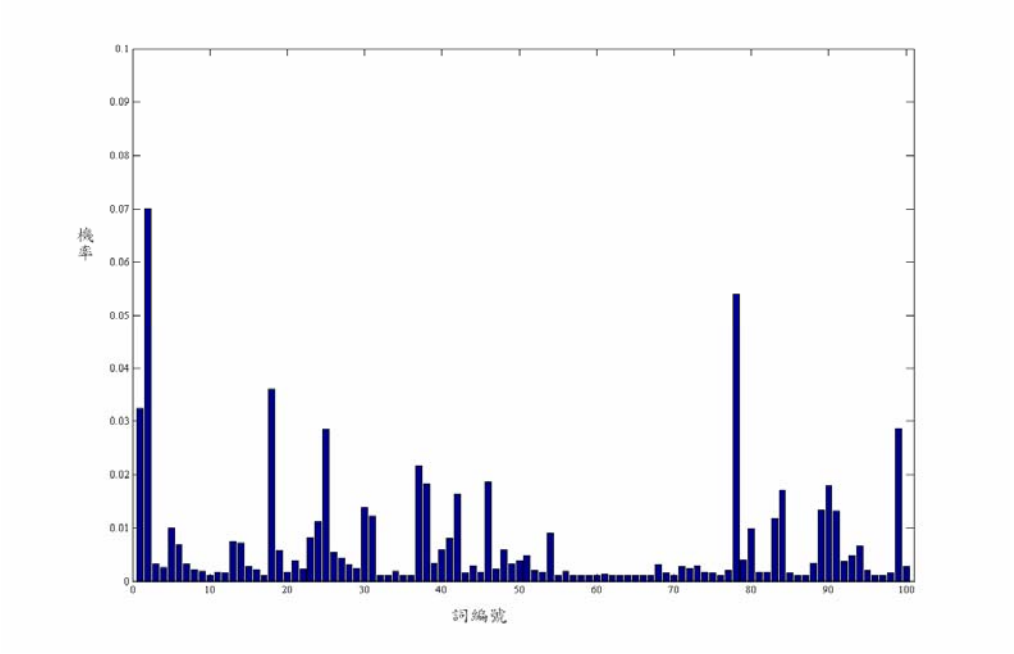


圖 1-5 語言模型(多項式分布)示意圖

1.2.3 語言模型(Language Model)

由於聲學模型本身只能辨識某一段語音訊號代表的是何種音素或音節序列，而無法確認其對應的詞(中文有許多同音詞)，且句子中詞跟詞的連接其實存在某種規則性，因此便需要有語言模型的存在。由於語言模型的機率分布是離散型的(多項式分布)，以詞單連(Unigram)語言模型為例，每一個詞編號都會有其對應的機率，如圖1-5所示。在估計語言模型的機率時，並不使用機率密度分布函數，而是直接估測個別詞序列的機率質量函數(Probability Mass Function, PMF) $P(w_1, w_2, \ldots w_N)$，其中 $w_1, w_2, \ldots, w_N$ 為此詞序列所包含的詞。但對整個詞序列的估測參數會隨著詞數量成指數成長，因此會遭遇資料稀疏(Data Sparseness)的問題。為了解決此問題，我們會先將語言模型的式子展開成條件機率的連乘積，再利用n-1階的馬可夫假設(n-1 Order Markovian Assumption)簡化，如下所示:

$$P(W) = P(w_1, w_2, \ldots, w_N) \approx \prod_{k=1}^{N} P(w_k \mid w_{k-n+1}, \ldots, w_{k-2}, w_{k-1}) \tag{1.4}$$

其中 N 為詞的個數，$w_{k-n+1},...,w_{k-2},w_{k-1}$ 則是 w_k 的歷史詞序列，式(1.4)便是常見的 n-連(n-gram)語言模型表示法。一般為了方便起見，以及減少參數量的複雜度，常使用詞二連(Bigram)及詞三連(Trigram)兩種模型(也就是分別使用一階及二階的馬可夫假設)。如同聲學模型，語言模型也需要有大量的文字語料來做為訓練之用。n-連語言模型的訓練方法有最大化相似度估測法(Maximum Likelihood Estimation, MLE)、最大熵值法(Maximum Entropy, ME)[Rosenfeld 1996]或是最小詞錯誤(Minimum Word Error, MWE)[Kuo *et al.* 2005]等，另外為了處理某些詞可能在訓練語料沒有出現的問題，通常會搭配如 Katz Smoothing[Katz 1987]及 Kneser-Ney Smoothing[Ney *et al*. 1994]等語言模型平滑技術，對這些估測機率原本為零的部份($P(w_k \mid w_{k-n+1},...,w_{k-2},w_{k-1}) = 0$)加以平滑化處理。

1.2.4 聲學比對與語言解碼(Linguistic Decoding)

在依式(1.2)尋找最佳詞序列時，由於分母的部份並不會影響最後詞序列排名的結果，因此實作上常將分母的部份省略。有了這項前提之後，就可以利用式(1.3)中的聲學模型與語言模型作聲學比對及語言解碼，聲學比對是負責將音素及詞序列中每一個可能的語音段落做比對，計算其相似度；語言解碼一般是使用維特比動態規劃搜尋(Viterbi Dynamic Programming Search)[Viterbi 1967]，結合聲學相似度及語言模型機率去找出一條最佳的詞序列。此外，由於搜尋空間會隨著詞典大小成指數成長，因此，在搜尋時，通常會透過搜尋路徑裁減(Pruning)技術來停止繼續尋找一些機率較低的詞序列，以減低其計算複雜度及記憶體使用量。搜尋時隨著語言模型愈複雜，搜尋空間也呈指數成長，為了降低搜尋時的複雜度，通常會透過兩階段的搜尋來完成：第一階段進行聲學比對並使用較低階的語言模型來搜尋，保留機率較高的候選文句；第二階段則使用較高階的語言模型再進行重新搜尋(Rescoring)[Ortmanns *et al.* 1997]。

1.2.5 信心度評估(Confidence Measure)

信心度評估的基本應用是給定語音辨識系統的輸出結果一個分數(輸出結果可以是針對整句詞序列，詞序列中的某個詞，或是音節等其它更小的單位，而給定的分數通常是介於0~1之間)，來判斷這個辨識結果的可靠度。舉例來說，信心度評估可以辨別每個辨識出來的詞它被辨識正確的機率有多高。如果依方法來分的話，大致上可分為三大類[Jiang 2005; 陳燦輝 2006]:

(i) 以特徵為基礎(Feature-based)之信心度評估

此種方法通常都是利用在進行語音辨識的過程中可獲得的一些所謂的預估特徵(Predictor Features，包含聲學及語言等資訊)。而一個特徵要能被稱為是預估特徵，其特徵值對正確辨認詞及錯誤辨認詞所建立的機率密度函式(Probability Density Function, PDF)必須具有很大的鑑別性。另一方面，每個預估特徵之間可利用某種方式結合，再配合不同的分類器，如有限向量機(Support Vector Machine, SVM)[Zhang and Rudnicky 2001]、天真貝氏分類器(Naïve Bayes Classifier)[Sanchis *et al*. 2004]等來決定辨識結果的正確性。

(ii) 發音確認(Utterance Verification)

此種方法則是將信心度評估視為統計式的假設檢定(Hypothesis Testing)的一種問題[Rose *et al*. 1995]。在這個架構之下，通常會提出兩個互斥的假設:

$$\begin{aligned} &H_0 \text{ (虛無假設，Null Hypothesis)}: O \text{之辨認的結果為正確} \\ &H_1 \text{ (對立假設，Alternative Hypothesis)}: O \text{之辨認的結果為錯誤} \end{aligned} \tag{1.5}$$

其中O 代表一段語音特徵向量序列。然後，我們測試虛無假設及對立假設，以決定辨識結果之正確與否。而測試之方法則是使用相似度比例檢測(Likelihood Ratio Testing, LRT) :

$$\frac{p(O|H_0)}{p(O|H_1)} \underset{H_1}{\overset{H_0}{\gtrless}} \tau \tag{1.6}$$

τ 為事先設定的門檻值(Threshold)，而 $p(O|H_0)$ 及 $p(O|H_1)$ 一般來說可使用隱藏式馬可夫模型來做估算。如果計算出來的值大於門檻值，我們便相信辨識結果的正確性，否則，便認為辨識結果為錯誤的。

(iii) 事後機率(Posterior Probability)

在傳統的最大事後機率(MAP)語音辨識方法中，式(1.2)的事後機率 $P(W|O)$ 對詞序列而言其實可以算是一種很好的信心度評估準則。但我們通常會省略分母項 $p(O)$ ，造成語音辨識系統輸出的分數不再是介於0到1的值。即使不省略分母項，但由於語音訊號有無窮多種，所以要如何估測出 $p(O)$ 便變成了一個癥結所在。為了解決這個問題，先前學者的研究曾提出了下列兩種方式來求得近似解:

(1)填充化基礎(Filler-based)法:此類方法主要是需要另外一組填充模型(Filler Model)或背景模型(Background Model)，如全音素辨識(All-phone Recognition)[Young 1994]、全包式模型(Catch-all Model)[Kamppari *et al.* 2000]。

(2)圖形化基礎(Graph-based)法:這類的方法主要是根據前向後向演算法(Forward-backward Algorithm)以詞圖上(Word Graph)資訊來估算 $p(O)$ [Wessel *et al* 2001]。

1.3 傳統聲學模型參數估測

1.3.1 最大化相似度(Maximum Likelihood)聲學模型估測

在本論文中，聲學模型是使用連續密度隱藏式馬可夫模型(CDHMM)，給定某詞序列 W 及其對應的語音特徵向量序列 $O=\{o_1,...,o_T\}$，其時間長度為 T。則 O 在聲學模型 W 的相似度可表示為:

$$p(O|W)=\sum_{s_1^T\in W} p(O,s_1^T) \tag{1.7}$$

其中 s_1^T 為時間 1 到 T，某詞序列所對應的一種可能的狀態序列，可視為隱藏變數(Latent Variable)。利用貝氏定理，可將式(1.7)展開為:

$$p(O|W) = \sum_{s_1^T \in W} \prod_{t=1}^{T} p(o_t | O_1^{t-1}, s_1^T) \cdot P(s_t | O_1^{t-1}, s_1^{t-1}) \tag{1.8}$$

再利用一階馬可夫假設(First-order Markovian Assumption)及觀測(或語音特徵向量間)獨立假設(Observation Independent Assumption)，式(1.8)可表示為:

$$p(O|W) \approx \sum_{s_1^T \in W} \prod_{t=1}^{T} p(o_t | s_t) \cdot P(s_t | s_{t-1}) \tag{1.9}$$

其中 $P(s_t | s_{t-1})$ 為從時間 $t-1$ 到 t 的狀態轉移機率，$P(o_t | s_t)$ 為狀態 s 在時間點 t 產生語音特徵向量 o_t 的機率，稱為觀測機率(Observation Probability)，在本論文中，使用高斯混合模型(GMM)來表示此觀測機率，可表示如下:

$$p(o_t | s_t) = \sum_m c_{s_t m} \cdot N(o_t; \mu_{s_t m}, \Sigma_{s_t m}) \tag{1.10}$$

其中 $c_{s_t m}$ 為狀態 s_t 中某個高斯分布 m 的權重(Mixture Weight)，$N(\cdot)$ 表示為高斯分布(Gaussian Distribution)，$\mu_{s_t m}$ 為狀態 s_t 中某個高斯分布 m 的平均值向量(Mean Vector)，$\Sigma_{s_t m}$ 為狀態 s_t 中某個高斯分布 m 的共變異矩陣(Covariance Matrix)。在連續密度隱藏式馬可夫模型中，模型參數包含狀態轉移機率、高斯分布之平均值向量、共變異矩陣及權重。這些參數均是使用機器學習(Machine Learning)的方法來加以訓練，由於每個人的語音特性不盡相同，所以需要訓練語料以估測模型參數，來代表大部份人類語音的統計分布。

傳統聲學模型參數的估測是使用最大化相似度估測法(Maximum Likelihood Estimation, MLE)[Bahl *et al.* 1983]。在收集到訓練語料之後，利用最大化相似度使其聲學模型與其對應的語音特徵量序列越像越好，即語音特徵向量序列落在其對應的聲學模型之相似度會最大，可以利用波氏重估(Baum-Welch Re-estimation, BW)演算法(又稱前向-後向演算法, Forward-Backward Algorithm, FBA)[Baum 1972]來訓練聲學模型中的參數。以平均值向量與共變異矩陣為例，其中平均值向量(Mean Vector)參數的調整可表示為[Rabiner 1989]:

$$\hat{\mu}_{sm} = \frac{\sum_{z=1}^{Z}\sum_{t=1}^{T}\gamma_{sm}^{z}(t)\cdot o_t}{\sum_{z=1}^{Z}\sum_{t=1}^{T}\gamma_{sm}^{z}(t)} \tag{1.11}$$

o_t 為在時間點 t 時的語音特徵向量序列，$\gamma_{sm}^{z}(t)$ 為在第 z 句訓練語料在時間點 t 時狀態 s 中某個高斯分布 m 的事後機率，可由有效率的前向-後向演算法(FBA)求得。而共變異矩陣(Covariance Matrix)參數的調整可表示為:

$$\begin{aligned}
\hat{\Sigma}_{sm} &= \frac{\sum_{z=1}^{Z}\sum_{t=1}^{T}\gamma_{sm}^{z}(t)\cdot(o_t-\hat{\mu}_{sm})(o_t-\hat{\mu}_{sm})^{Tr}}{\sum_{z=1}^{Z}\sum_{t=1}^{T}\gamma_{sm}^{z}(t)} \\
&= \frac{\sum_{z=1}^{Z}\sum_{t=1}^{T}\gamma_{sm}^{z}(t)\cdot o_t o_t^{Tr}}{\sum_{z=1}^{Z}\sum_{t=1}^{T}\gamma_{sm}^{z}(t)} - \frac{\sum_{z=1}^{Z}\sum_{t=1}^{T}\gamma_{sm}^{z}(t)\cdot o_t \hat{\mu}_{sm}^{Tr}}{\sum_{z=1}^{Z}\sum_{t=1}^{T}\gamma_{sm}^{z}(t)} - \frac{\sum_{z=1}^{Z}\sum_{t=1}^{T}\gamma_{sm}^{z}(t)\cdot \hat{\mu}_{sm} o_t^{Tr}}{\sum_{z=1}^{Z}\sum_{t=1}^{T}\gamma_{sm}^{z}(t)} + \frac{\sum_{z=1}^{Z}\sum_{t=1}^{T}\gamma_{sm}^{z}(t)\cdot \hat{\mu}_{sm}\hat{\mu}_{sm}^{Tr}}{\sum_{z=1}^{Z}\sum_{t=1}^{T}\gamma_{sm}^{z}(t)} \\
&= \frac{\sum_{z=1}^{Z}\sum_{t=1}^{T}\gamma_{sm}^{z}(t)\cdot o_t o_t^{Tr.}}{\sum_{z=1}^{Z}\sum_{t=1}^{T}\gamma_{sm}^{z}(t)} - \hat{\mu}_{sm}\hat{\mu}_{sm}^{Tr} - \hat{\mu}_{sm}\hat{\mu}_{sm}^{Tr} + \hat{\mu}_{sm}\hat{\mu}_{sm}^{Tr} \\
&= \frac{\sum_{z=1}^{Z}\sum_{t=1}^{T}\gamma_{sm}^{z}(t)\cdot o_t o_t^{Tr.}}{\sum_{z=1}^{Z}\sum_{t=1}^{T}\gamma_{sm}^{z}(t)} - \hat{\mu}_{sm}\hat{\mu}_{sm}^{Tr}
\end{aligned} \tag{1.12}$$

其中 Tr 代表向量或矩陣的轉置(Transposition)。

1.3.2 傳統聲學模型參數估測工具

在目前聲學模型訓練的階段，通常需使用大量的語料為模型作訓練，大致可分為三個階段，目前在國際上語音辨識的研究已廣泛地使用 HTK Toolkit 來完成此項工作[Young *et al.* 2006]：

第一階段：使用分段 *K*-平均值法(Segmental *K*-means Algorithm)，訓練語料需事先經過對齊(Alignment)決定音素的邊界。第一次的迭代，將語音片段等切對齊至狀態，每個狀態則利用對齊到的語音特徵向量，透過 *K*-中心法分成指定的群數，進而估測所屬的高斯分布。第二次之後的迭代，不再對語音片段做等切對

齊，而是使用前一次所估測的模型做分段對齊(Segmental Alignment)。此階段訓練可使用 HTK Toolkit 內的 HInit 函數來實作。

第二階段：對每個音素及其對應的語料，使用最大化相似度訓練法則配合使用波氏重估(BW)演算法(前向-後向演算法(FBA))來訓練聲學模型，期望訓練語料在對應的聲學模型上有較高的相似度。此階段訓練可使用 HTK Toolkit 內的 HRest 函數來實作。

第三階段：與第二階段同樣使用波氏重估演算法來訓練聲學模型，差別在於此階段移除了音素邊界，將整段語句的音素都考慮進來，使跨音素的統計資訊能加以利用。此階段訓練可使用 HTK Toolkit 內的 HERest 函數來實作。

1.4 本論文研究內容與貢獻

影響一個語音辨識器效能的因素有許多種，例如：語音特徵擷取的方法、聲學模型與語言模型的結構、聲學模型與語言模型的訓練方法以及搜尋演算法等。近年來的研究大致可分為四個部份，語音特徵擷取技術、聲學模型訓練、語言模型訓練和搜尋演算法的改進，本論文主要專注在聲學模型訓練的改進。

傳統聲學模型的訓練是以最大化相似度估測法(MLE)，配合波氏重估演算法(BW)來做聲學模型的訓練，但此種訓練方法沒有考慮到聲學模型間彼此的關係，在調整模型參數之後，使得相關的語音特徵向量序列落在此聲學模型的相似度(Likelihood)變大，卻可能同時使非相關的語音特徵向量序列落在此聲學模型的相似度更大，因而產生辨識上的混淆。過去十多年來，有不少研究針對此項缺點，提出鑑別式(Discriminative)的訓練法則來加以改進，其中又以最小化音素錯誤(Minimum Phone Error, MPE)[Povey 2004; 郭人瑋 2005]之鑑別式訓練為最好的方法之一。本論文首先以最小化音素錯誤(MPE)訓練為基礎，深入探討其原理並改進其缺點。接著本論文提出以熵值(Entropy)為基礎的資料選取(Data Selection)方法來改進所有的鑑別式聲學模型訓練，再將其資料選取方法應用到非監督(Unsupervised)鑑別式聲學模型訓練。以下為本論文之研究成果與貢獻。

(1) 提出時間音框正確率函數:最小化音素錯誤訓練的原始音素正確率(Raw Phone Accuracy)函數在模型參數訓練時有給予插入錯誤(Insertion Errors)及取代錯誤(Substitution Errors)適當的懲罰，但其訓練方法並沒有考慮到刪除錯誤(Deletion Errors)的影響，本論文便是旨在改善其缺點，因而提出時間音框正確率(Time Frame Accuracy, TFA)函數來取代原本的原始音素正確率函數。

(2) 考慮事前機率:目前以全面風險為基礎(將在2.2小節介紹)的鑑別式訓練都是假設事前機率(Prior Probability)是一致的(Uniform)，本論文提出以統計式的方法來計算事前機率，使得每個訓練語料具有不同的事前機率，進而改善鑑別式訓練。

(3) 提出以熵值為基礎的資料選取方法:本論文的主要貢獻是提出以熵值(Entropy)為基礎的資料選取(Data Selection)方法來改進所有的鑑別式訓練。過去幾年來，以邊際(Margin)為基礎的模型訓練在語音辨識的小詞彙(Small Vocabulary)應用情境中有不錯的成效[Hui *et al.* 2006; Jinyu Li *et al.* 2006]。從某種角度看，其以邊際為基礎的模型訓練可以視為是資料選取的方法，再以不同的目標函數來最佳化模型參數。其邊際是定義在給定某語音特徵向量序列時，正確詞序列與辨識詞序列的相似度(Likelihood)之差，所以可以視為在相似度定義域中(Likelihood Domain)來選取資料[Hui *et al.* 2006]。在本論文中是以給定在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(Posterior Probability，此事後機率是有考慮到詞與詞之間的轉移機率，即語言模型)來求得熵值，再經由事先所設定的門檻值來選取資料，所以可以視為在事後機率定義域(Posterior Probability Domain)中(不同於以往的相似度定義域)來取選資料。本論文所提出以熵值為基礎的資料選取方法著重在時間音框(Frame)的選取，因為現階段的聲學模型訓練是以時間音框為最小的統計值收集單位。目前所有的鑑別式訓練都是以所有的時間音框(Frame)所收集到的統計值來調整模型參

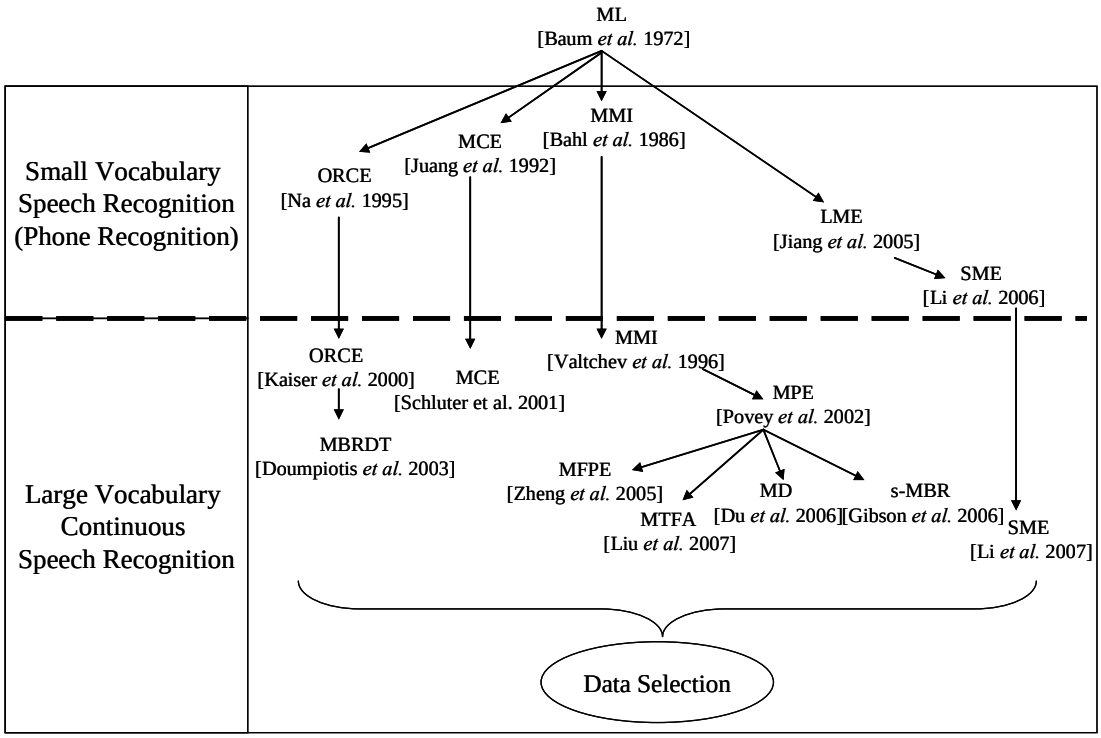


圖 1-6 現階段鑑別式聲學模型訓練

數，事實上有些時間音框的統計值是沒有貢獻的，所以本論文提出之資料選取方法可以濾掉那些沒有貢獻的統計值，只採用重要、具鑑別性的音框所提供的統計值來調整模型參數，以改進所有的鑑別式聲學模型訓練。圖1-6簡略說明現階段鑑別式聲學模型訓練之演進，其中本論文所提出的以熵值為基礎之資料選取方法可以適用於所有的鑑別式訓練。

(4) 應用資料選取方法於非監督式訓練:由於網路上可以取得的新聞語料越來越多，但都大部分都缺乏人工所轉譯的正確詞序列，而且人工轉寫需要的成本非常高，所以本論文利用以熵值為基礎的資料選取方法來幫助最小化音素錯誤(MPE)於非監督鑑別式聲學模型訓練，以求得較佳的辨識結果。

1.5 論文架構

本論文的章節概要如下:

第二章將回顧貝氏風險(2.1小節)、全面風險(2.2小節)以及過去有關鑑別式聲學模型訓練(2.3節)的研究；

第三章則是描述過去針對最小化音素錯誤訓練之改進(3.1小節)以及本論文提出之方法於改進最小化音素錯誤訓練，如時間音框正確率函數(3.2小節)和考慮事前機率(3.3小節)；

第四章首先敘述以邊際為基礎的模型訓練(4.1小節)，事實上此方法從某種角度來看，我們視為是資料選取的方法。本論文提出以熵值為基礎的資料選取方法來挑選重要的音框，以改進所有的鑑別式聲學模型訓練(4.2小節)；

第五章將介紹非監督式之聲學模型訓練，並利用4.2小節所提出之方法來改進非監督之鑑別式聲學模型訓練；

第六章則描述臺灣師大之大詞彙連續語音辨識系統的基本架構、實驗語料、實驗設定、基礎實驗及相關改進鑑別式聲學模型訓練之實驗結果；

第七章則是結論與未來展望，討論未來可能的研究方向。

第2章　鑑別式模型訓練

本章將在 2.1 小節介紹貝氏風險，2.2 小節介紹全面風險，在 2.3 小節回顧過去學者所提出的鑑別式聲學模型訓練法則。

2.1 貝氏風險(Bayes Risk)

語音辨識的過程可視為一個分類的動作，將每句可能的詞序列都視為一類，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 為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 $R(W|O_z)$ 代表此歸類行為的風險(Risk)；而語音辨識則可視為找出此風險最低的詞序列。將 O_z 歸類至 W 的風險 $R(W|O_z)$ 可定義如下[Duda *et al.* 2000]：

$$R(W|O_z)=\sum_{W'\in\mathbf{W}} l(W,W')P(W'|O_z) \tag{2.1}$$

其中 **W** 為所有可能詞序列所成的集合；$P(W'|O_z)$ 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)；$l(W,W')$ 為一減損函數(Loss Function)，用以表示詞序列 W 與 W' 之間差異所造成的損失(Loss)，$R(W|O_z)$ 為將 O_z 歸類至 W 時的期望損失(Expected Loss)，又稱為貝氏風險(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：

$$\hat{W}=\arg\min_{W\in\mathbf{W}} R(W|O_z)=\arg\min_{W\in\mathbf{W}}\sum_{W'\in\mathbf{W}} l(W,W')P(W'|O_z) \tag{2.2}$$

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(2.2))來設計其搜尋演算法，如標準最大化事後機率解碼方法(Standard Maximum a Posterior Decoding, MAP)[Bahl *et al.* 1983]、ROVER(Recognizer Output Voting Error Reduction)[Fiscus 1997]、最小化貝氏風險(Minimum Bayes Risk, MBR)[Goel & Byrne 2000]、最小化時間音框錯誤搜尋(Minimum Time Frame

Error Search)[Wessel *et al* 2001]及詞錯誤最小化(Word Error Minimization) [Mangu *et al.* 2000]等。

2.2 全面風險(Overall Risk)

前2.1小節提到貝氏風險的計算可以用在搜尋上，但若用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} [Duda *et al.* 2000]：

$$R_{all} = \int R(W \mid O)P(O)dO \tag{2.3}$$

其中 W 為語音特徵向量序列 O 對應之正確轉譯詞序列，$P(O)$ 為 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，為所有訓練語句(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，故全面風險可簡化為 Z 個訓練語句的條件風險總和：

$$R_{all} = \sum_{z=1}^{Z} R(W_z \mid O_z)P(O_z) = \sum_{z=1}^{Z} \sum_{W' \in \mathbf{W}} l(W_z, W')P(W' \mid O_z)P(O_z) \tag{2.4}$$

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為 $P(W' \mid O_z;\theta)$，則全面風險可改寫成：

$$R_{all} = \sum_{z=1}^{Z} R(W_z \mid O_z)P(O_z) = \sum_{z=1}^{Z} \sum_{W' \in \mathbf{W}} l(W_z, W')P(W' \mid O_z;\theta)P(O_z) \tag{2.5}$$

若 $P(O_z)$ 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：

$$R_{all} = \sum_{z=1}^{Z} \sum_{W' \in \mathbf{W}} l(W_z, W')P(W' \mid O_z;\theta) \tag{2.6}$$

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：

$$\hat{\theta} = \arg\min_{\theta} \sum_{z=1}^{Z} \sum_{W' \in \mathbf{W}} l(W_z, W')P(W' \mid O_z;\theta) \tag{2.7}$$

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，這也因此成為一個開放式的議題，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)為基礎，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) [Normandin 1991]、全面風險估測法則(Overall Risk Criterion Estimation, ORCE) [Kaiser *et al.* 2002]、最小化貝氏風險鑑別式訓練(Minimum Bayes Risk Discriminative Training, MBRDT) [Doumpiotis *et al.* 2004]、最小化音素錯誤訓練(Minimum Phone Error Training, MPE) [Povey & Woodland 2002]等。

2.3 鑑別式聲學模型訓練

鑑別式訓練法則是不以最大化訓練語料的相似度為目標，而以最小分類錯誤為目標，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)來做聲學模型的訓練，但此種訓練方法沒有考慮到語音辨識時聲學模型間彼此的關係，在調整聲學模型參數之後，使得相關的語音特徵落在此聲學模型的相似度(Likelihood)變大，卻可能同時使非相關的語音特徵落在此聲學模型的相似度更大，產生辨識上的混淆。近來，有不少研究針對此項缺點，提出鑑別式的訓練(Discriminative Training)法則來加以改進。

2.3.1 最大化交互資訊估測(Maximum Mutual Information)

從風險最小化的角度出發，經由減損函數的設計以及數學的推導，我們就可以得到最大化交互資訊估測的目標函數(Objective Function)，說明如下，若式(2.7)中的減損函數定義為零一函數(Zero-One Function)，即：

$$l_{0-1}(W_z, W') = \begin{cases} 0 & , W_z = W' \\ 1 & , W_z \neq W' \end{cases} \tag{2.8}$$

當詞序列W_z與W'相同時損失為 0，否則損失為 1。在此令$\theta = \{\lambda\}$，即只考慮聲學模型，則全面風險可簡化為：

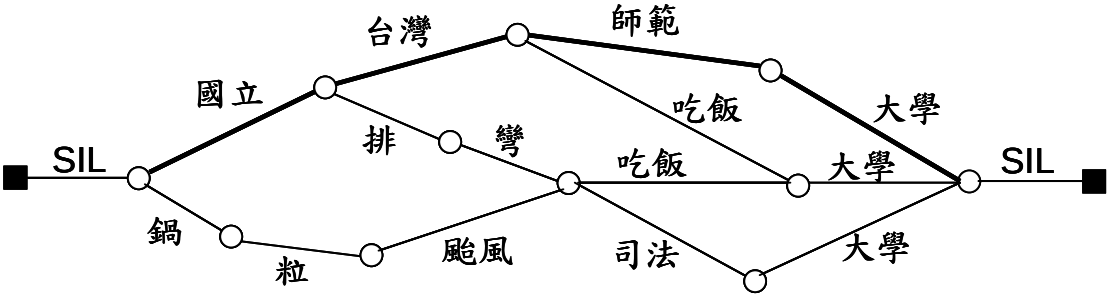


圖 2-1 詞圖為所有可能詞序列 **W** 的近似

$$R_{all}=\sum_{z=1}^{Z}\sum_{\{W'\in\mathbf{W}|W'\neq W_z\}}P(W'|O_z;\theta)=\sum_{z=1}^{Z}\big(1-P(W_z|O_z;\theta)\big) \tag{2.9}$$

為了方便處理，在此利用 Jensen's 不等式為此風險找一上界(Upper Bound)：

$$\sum_{z=1}^{Z}\big(1-P(W_z|O_z;\theta)\big)\leq-\sum_{z=1}^{Z}\log P(W_z|O_z;\theta) \tag{2.10}$$

故全面風險可寫成：

$$R_{all}\leq-\sum_{z=1}^{Z}\log P(W_z|O_z;\theta) \tag{2.11}$$

在模型 θ 的估測上，需要最小化全面風險的上界，以間接地方式最小化實際的全面風險：

$$\begin{aligned}\hat{\theta}&=\arg\min_{\theta}-\sum_{z=1}^{Z}\log P(W_z|O_z;\theta)\\&=\arg\max_{\theta}\sum_{z=1}^{Z}\log P(W_z|O_z;\theta)\\&=\arg\max_{\theta}F_{MMI}(\theta)\end{aligned} \tag{2.12}$$

式(2.12)中的對數事後機率之總和即為最大化交互資訊估測(Maximum Mutual Information Estimation, MMIE)的目標函數。利用貝氏定理將事後機率展開便可得到如下式子：

$$F_{MMI}(\theta)=\sum_{z=1}^{Z}\log\frac{P(O_z|W_z;\theta)P(W_z)}{\sum_{W'\in\mathbf{W}_z}P(O_z|W';\theta)P(W')} \tag{2.13}$$

其中 $\mathbf{W}_z$ 為 O_z 所可能產生的詞序列，在實作上是以詞圖(Word Graph or Lattice，如圖2-1)來表示。由式(2.13)可知其目標函數為一有理函數(Rational Function)，傳統的波氏重估演算法便不再適用，取而代之的是延伸波氏重估演算法(Extended Baum-Welch, EBW) [Gopalakrishnan *et al.* 1991]。最大化交互資訊估測的目的是要使語音特徵向量序列 O_z 與其對應的正確轉譯詞序列 W_z 之事後機率要越大越好。然而以中文語音辨識而言，其最終目的就是要提高字辨識率或降低字錯誤率，所以在模型訓練上之減損函數最好能與語音辨識評估準則(Evaluation Criterion)一致，在最大化交互資訊估測所使用的零一函數似乎與評估準則有很大的差異，在下一小節中，將介紹使用與評估標準一致的減損函數之鑑別式聲學模型訓練。

2.3.2 全面風險估測準則(Overall Risk Criterion Estimation)

在語音辨識中，每一詞序列代表每一個類別，使用零一函數作為減損函數，雖然可以最小化語句錯誤率(Sentence Error Rate, SER)，但在對語音辨識進行評估時，卻是以詞錯誤率(Word Error Rate, WER)，或在中文上以字錯誤率(Character Error Rate, CER)作為評估標準，使得以零一函數為基礎的鑑別式訓練法則與評估之間產生了相當大的差異。換句話說，句子錯誤愈小，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子錯誤。為了克服此問題，近年來陸續有人提出以Levenshtein距離[Levenshtein 1966]或稱為編輯距離(Edit Distance)取代零一函數作為減損函數，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。全面風險法則估測則是因應此概念所提出的鑑別式訓練方式。由式(2.7)中，可以看出有兩個重要的值需要被計算出來，一為事後機率 $P(W'|O_z;\theta)$，另一為編輯距離 $L(W_z,W')$。然而實際應用到語音辨識中，我們將會遇到一些問題，即如何有效率地在大量的其他可能詞序列中去計算編輯距離，詞圖可以簡潔地表示所有可能詞序列的近似，因其結構的關係，利用動態程式規劃(Dynamic Programming)和前向後向演算法(Forward-Backward Algorithm)就可以快速的算

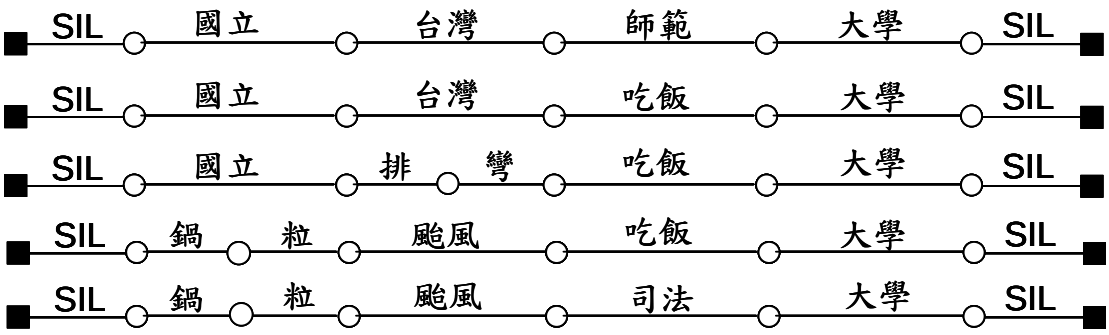


圖 2-2 *N*-最佳序列

出詞圖中每一個詞段(Word Arc)的事後機率，但卻無法有效率地計算任意兩個詞序列間的編輯距離，所以我們只好退而求其次，使用*N*-最佳序列(*N*-Best List，如圖2-2所示)去計算詞序列的編輯距離。使用*N*-最佳序列 $\mathbf{W}_{z,N\text{-Best}}$ 去近似所有可能的詞序列 $\mathbf{W}_z$ ，以及使用編輯距離 $L(\cdot)$ 做為減損函數，我們便可以將式(2.7)改寫如下：

$$\hat{\theta} = \arg\min_{\theta} \sum_{z=1}^{Z} \sum_{W' \in \mathbf{W}_{z,N-\text{Best}}} L(W_z, W') P(W' \mid O_z; \theta) \tag{2.14}$$

所以全面風險法則估測(Overall Risk Criterion Estimation, ORCE)的目標函數變為：

$$f_{ORCE}(\theta) = R_{all} = \sum_{z=1}^{Z} \sum_{W' \in \mathbf{W}_{z,N-\text{Best}}} L(W_z, W') P(W' \mid O_z; \theta) \tag{2.15}$$

因全面風險的目標函數也是含有事後機率，故為一有理函數，則需要使用延伸波式重估演算法(EBW)來進行模型參數的調整，延伸波氏重估演算法是最大化目標函數，而全面風險是要最小化，所以我們必須將全面風險的目標函數乘上負號，延伸波氏重估演算法的一般式為[Gopalakrishnan *et al.* 1991]：

$$\hat{\theta}_a = \frac{\theta_a \left(\dfrac{\partial(-f_{ORCE}(\theta))}{\partial \theta_a} + D \right)}{\sum_{e=1}^{E} \theta_e \dfrac{\partial(-f_{ORCE}(\theta))}{\partial \theta_e} + D} \tag{2.16}$$

其中，θ_a 可為隱藏式馬可夫模型(Hidden Markov Model, HMM)參數 θ 中某一個參數，如從狀態 i 移轉至狀態 j 的轉移機率 a_{ij} (Transition Probability)、狀態 j 中第 m 個高斯分布的混合權重 c_{jm} (Mixture Weight)和在時間點 t 時狀態 j 的生成機率 $b_j(o_t)$ (Output Probability)。E 為 θ_a 所屬同一類的隱藏式馬可夫模型中的所有參數個數，例如 θ_a 屬於具有五個狀態的隱藏式馬可夫模型，則 $E=5$。D 為一常數，常用來控制收斂速度。要求得聲學模型參數的更新式子就要先求得全面風險目標函數的斜率值：

$$\frac{\partial(-f_{ORCE}(\theta))}{\partial\theta_a}=\frac{\partial}{\partial\theta_a}-\sum_{z=1}^{Z}\sum_{W'\in\mathbf{W}_{z,N-Best}}L(W_z,W')P(W'|O_z;\theta) \tag{2.17}$$

接著用貝氏定理將事後機率展開：

$$\frac{\partial(-f_{ORCE}(\theta))}{\partial\theta_a}=\frac{\partial}{\partial\theta_a}-\sum_{z=1}^{Z}\sum_{W'\in\mathbf{W}_{z,N-Best}}L(W_z,W')\frac{P(O_z|W';\theta)P(W')}{P(O_z;\theta)} \tag{2.18}$$

其中 $P(O_z;\theta)=\sum_{W''\in\mathbf{W}_{z,N-Best}}P(O_z|W'';\theta)P(W'')$，式(2.18)可改寫如下：

$$\frac{\partial(-f_{ORCE}(\theta))}{\partial\theta_a}=\frac{\partial}{\partial\theta_a}-\sum_{z=1}^{Z}\frac{\sum_{W'\in\mathbf{W}_{z,N-Best}}L(W_z,W')P(O_z|W';\theta)P(W')}{\sum_{W''\in\mathbf{W}_{z,N-Best}}P(O_z|W'';\theta)P(W'')} \tag{2.19}$$

接著利用簡單的微分公式進行有理多項式(Rational Polynomial)的微分：

$$\frac{\partial(-f_{ORCE}(\theta))}{\partial\theta_a}=-\sum_{z=1}^{Z}\left(\frac{\left[\sum_{W''\in\mathbf{W}_{z,N-Best}}P(O_z|W'';\theta)P(W'')\right]\left(\sum_{W'\in\mathbf{W}_{z,N-Best}}L(W_z,W')P(W')\frac{\partial}{\partial\theta_a}P(O_z|W';\theta)\right)}{\left[\sum_{W''\in\mathbf{W}_{z,N-Best}}P(O_z|W'';\theta)P(W'')\right]^2}\right.$$
$$\left.-\frac{\left[\sum_{W'\in\mathbf{W}_{z,N-Best}}L(W_z,W')P(O_z|W';\theta)P(W')\right]\left(\sum_{W''\in\mathbf{W}_{z,N-Best}}P(W'')\frac{\partial}{\partial\theta_a}P(O_z|W'';\theta)\right)}{\left[\sum_{W''\in\mathbf{W}_{z,N-Best}}P(O_z|W'';\theta)P(W'')\right]^2}\right) \tag{2.20}$$

我們利用變數互換，讓後面的分子項將W'換成W''，且將W''換成W'，目的則是為方便整理聲學模型的微分項，所以式(2.20)可寫為：

$$\frac{\partial(-f_{ORCE}(\theta))}{\partial\theta_a}=-\sum_{z=1}^{Z}\left(\frac{\left[\sum_{W''\in\mathbf{W}_{z,N-Best}}P(O_z|W'';\theta)P(W'')\right]\left(\sum_{W'\in\mathbf{W}_{z,N-Best}}L(W_z,W')P(W')\frac{\partial}{\partial\theta_a}P(O_z|W';\theta)\right)}{\left[\sum_{W''\in\mathbf{W}_{z,N-Best}}P(O_z|W'';\theta)P(W'')\right]^2}\right.$$
$$\left.-\frac{\left[\sum_{W''\in\mathbf{W}_{z,N-Best}}L(W_z,W'')P(O_z|W'';\theta)P(W'')\right]\left(\sum_{W'\in\mathbf{W}_{z,N-Best}}P(W')\frac{\partial}{\partial\theta_a}P(O_z|W';\theta)\right)}{\left[\sum_{W''\in\mathbf{W}_{z,N-Best}}P(O_z|W'';\theta)P(W'')\right]^2}\right) \tag{2.21}$$

將聲學模型微分項前面的加總符號往前提，不影響結果可得：

$$\frac{\partial(-f_{ORCE}(\theta))}{\partial\theta_a}=-\sum_{z=1}^{Z}\sum_{w'\in\mathbf{W}_{z,N-Best}}\left(\frac{\left[\sum_{W''\in\mathbf{W}_{z,N-Best}}P(O_z|W'';\theta)P(W'')\right]L(W_z,W')P(W')\frac{\partial}{\partial\theta_a}P(O_z|W';\theta)}{\left[\sum_{W''\in\mathbf{W}_{z,N-Best}}P(O_z|W'';\theta)P(W'')\right]^2}\right.$$
$$\left.-\frac{\left[\sum_{W''\in\mathbf{W}_{z,N-Best}}L(W_z,W'')P(O_z|W'';\theta)P(W'')\right]P(W')\frac{\partial}{\partial\theta_a}P(O_z|W';\theta)}{\left[\sum_{W''\in\mathbf{W}_{z,N-Best}}P(O_z|W'';\theta)P(W'')\right]^2}\right) \tag{2.22}$$

將聲學模型微分項獨立出來可得：

$$\frac{\partial(-f_{ORCE}(\theta))}{\partial\theta_a}=-\sum_{z=1}^{Z}\sum_{W'\in\mathbf{W}_{z,N-Best}}\left(\frac{\left[\sum_{W''\in\mathbf{W}_{z,N-Best}}P(O_z|W'';\theta)P(W'')\right]L(W_z,W')}{\left[\sum_{W''\in\mathbf{W}_{z,N-Best}}P(O_z|W'';\theta)P(W'')\right]^2}\right.$$
$$\left.-\frac{\left[\sum_{W''\in\mathbf{W}_{z,N-Best}}L(W_z,W'')P(O_z|W'';\theta)P(W'')\right]}{\left[\sum_{W''\in\mathbf{W}_{z,N-Best}}P(O_z|W'';\theta)P(W'')\right]^2}\right)P(W')\frac{\partial}{\partial\theta_a}P(O_z|W';\theta) \tag{2.23}$$

在此利用一個數學轉換，如下所示:

$$\begin{aligned}\frac{\partial}{\partial\theta_a}P(O_z\,|\,W';\theta) &= P(O_z\,|\,W';\theta)\frac{\partial}{\partial\theta_a}\log P(O_z\,|\,W';\theta)\\ &= P(O_z\,|\,W';\theta)\frac{1}{P(O_z\,|\,W';\theta)}\frac{\partial}{\partial\theta_a}P(O_z\,|\,W';\theta)\\ &= \frac{\partial}{\partial\theta_a}P(O_z\,|\,W';\theta)\end{aligned} \qquad (2.23^*)$$

故可將式(2.23)改寫成：

$$\frac{\partial(-f_{ORCE}(\theta))}{\partial\theta_a} = -\sum_{z=1}^{Z}\sum_{W'\in\mathbf{W}_{z,N-Best}}\left(\frac{\left[\sum_{W''\in\mathbf{W}_{z,N-Best}}P(O_z\,|\,W'';\theta)P(W'')\right]L(W_z,W')}{\left[\sum_{W''\in\mathbf{W}_{z,N-Best}}P(O_z\,|\,W'';\theta)P(W'')\right]^2} - \frac{\left[\sum_{W''\in\mathbf{W}_{z,N-Best}}L(W_z,W'')P(O_z\,|\,W'';\theta)P(W'')\right]}{\left[\sum_{W''\in\mathbf{W}_{z,N-Best}}P(O_z\,|\,W'';\theta)P(W'')\right]^2}\right)P(W')P(O_z\,|\,W';\theta)\frac{\partial}{\partial\theta_a}\log P(O_z\,|\,W';\theta)$$

(2.24)

假設$K(W',\mathbf{W}_{z,N-Best})$定義為:

$$K(W',\mathbf{W}_{z,N-Best}) = -\frac{\left[\sum_{W''\in\mathbf{W}_{z,N-Best}}P(O_z\,|\,W'';\theta)P(W'')\right]L(W_z,W') - \left[\sum_{W''\in\mathbf{W}_{z,N-Best}}L(W_z,W'')P(O_z\,|\,W'';\theta)P(W'')\right]}{\left[\sum_{W''\in\mathbf{W}_{z,N-Best}}P(O_z\,|\,W'';\theta)P(W'')\right]^2}P(W')P(O_z\,|\,W';\theta)$$

(2.25)

我們就可以將全面風險的微分簡潔表示如下：

$$\frac{\partial(-f_{ORCE}(\theta))}{\partial\theta_a} = \sum_{z=1}^{Z}\sum_{W'\in\mathbf{W}_{z,N-Best}}K(W',\mathbf{W}_{z,N-Best})\frac{\partial}{\partial\theta_a}\log P(O_z\,|\,W';\theta) \qquad (2.26)$$

其中$K(W',\mathbf{W}_{z,N-Best})$這個權重值可再經由化簡便可得：

$$\begin{aligned}K(W',\mathbf{W}_{z,N-Best}) &= \left[\sum_{W''\in\mathbf{W}_{z,N-Best}}L(W_z,W'')P(W''\,|\,O_z;\theta) - L(W_z,W')\right]P(W'\,|\,O_z;\theta)\\ &= \left[R(W_z\,|\,O_z) - L(W_z,W')\right]P(W'\,|\,O_z;\theta)\end{aligned} \qquad (2.27)$$

其中 $R(W_z \mid O_z) = \sum_{W'' \in \mathbf{W}_{z,N-Best}} L(W_z, W'') P(W'' \mid O_z; \theta)$ ，為正確詞序列 W_z 的貝氏風險(Bayes Risk)。$\log P(O_z \mid W'; \theta)$ 是最大化相似度(Maximum Likelihood, ML)的目標函數，在處理這項微分值時，需要引用一個輔助函數(Auxiliary Function)來幫助求取統計值，此輔助函數通常稱為Q函數(Q Function)。由式(2.28)以及式(2.29)，我們可以求得隱藏式馬可夫模型中某個狀態 s 中的某個高斯分布 m 之平均值向量 $\overline{\mu}_{sm}$ (Mean Vector)和共變異矩陣 $\overline{\Sigma}_{sm}$ (Covariance Matrix)的更新式子：

$$\overline{\mu}_{sm} = \frac{\sum_{z=1}^{Z} \sum_{W' \in \mathbf{W}_{z,N-Best}} K(W', \mathbf{W}_{z,N-Best}) \sum_{t} \gamma_{sm}^{z,W'}(t) o_t + D_{sm} \mu_{sm}}{\sum_{z=1}^{Z} \sum_{W' \in \mathbf{W}_{z,N-Best}} K(W', \mathbf{W}_{z,N-Best}) \sum_{t} \gamma_{sm}^{z,W'}(t) + D_{sm}} \tag{2.28}$$

$$\overline{\Sigma}_{sm} = \frac{\sum_{z=1}^{Z} \sum_{W' \in \mathbf{W}_{z,N-Best}} K(W', \mathbf{W}_{z,N-Best}) \sum_{t} \gamma_{sm}^{z,W'}(t) o_t^2 + D_{sm} (\Sigma_{sm} + \mu_{sm} \mu_{sm}^T)}{\sum_{z=1}^{Z} \sum_{W' \in \mathbf{W}_{z,N-Best}} K(W', \mathbf{W}_{z,N-Best}) \sum_{t} \gamma_{sm}^{z,W'}(t) + D_{sm}} - \overline{\mu}_{sm} \overline{\mu}_{sm}^T \tag{2.29}$$

其中 $\gamma_{sm}^{z,W'}(t)$ 為在第 z 句訓練語料中，某個詞序列 W' 上，在時間 t 時狀態 s 中的高斯分布 m 之佔有機率。$K(W', \mathbf{W}_{z,N-Best})$ 可視為每一詞序列 W' 對於平均值向量和共變異矩陣的調整量大小有多少的貢獻。

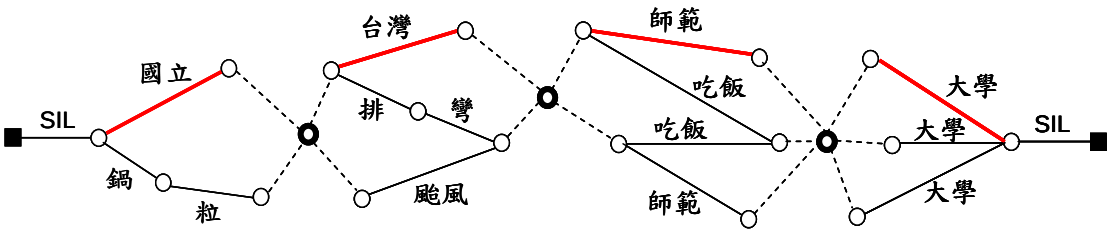


圖 2-3 對齊至正確轉譯詞段的詞圖(具有四的次詞圖)，
粗線(紅線)為正確轉譯詞段

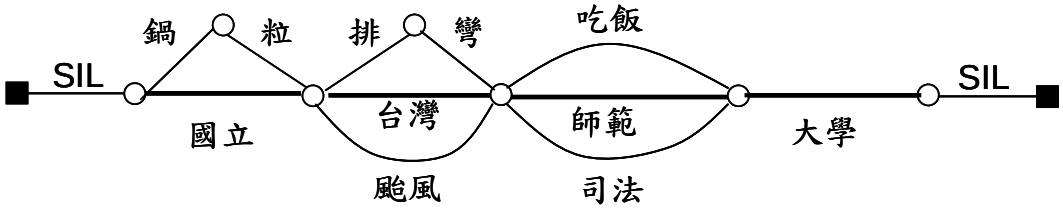


圖 2-4 將次詞圖整合而形成狹縮詞圖(未經刪除)

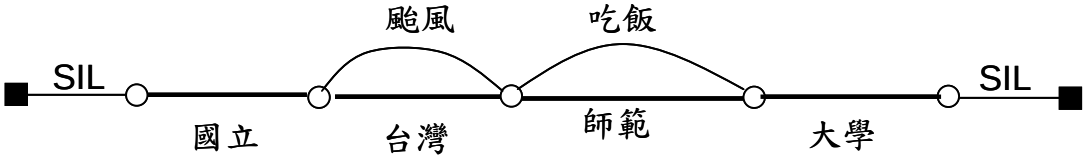


圖 2-5 具混淆對之狹縮詞圖(如吃飯與師範是混淆對)

2.3.3 最小化貝氏風險為基礎的鑑別式聲學模型訓練(MBRDT)

全面風險估測法則使用*N*-最佳序列來近似所有可能的詞序列，雖然方便計算編輯距離，但若*N*值不夠大時，*N*-最佳序列彼此間就會很相像，所以調整模型的資訊就會相對減少很多，若*N*值很大時，那麼計算編輯距離就會相當地耗時。在實作上我們也不可能將*N*值設得很大，而且詞圖又不方便計算編輯距離，若想要以編輯距離當減損函數，還是要將重心放在*N*-最佳序列上。想要讓*N*-最佳序列不要太相像，又不能使用詞圖來近似所有可能詞序列，目前只有一個辦法可以暫時解決這個問題，那就是從詞圖中找出具有鑑別力的*N*-最佳序列，因此就有學者提出

狹縮詞圖(Pinched Lattice(如圖2-4所示))的概念[Doumpiotis & Byrne 2004]來近似原本的詞圖，而且又可從中取出具有鑑別力的*N*-最佳序列，因而提出最小化貝氏風險漸別式訓練(Minimum Bayes Risk Discriminative Training, MBRDT)。狹縮詞圖是由語音辨識器所產生出來的詞圖強迫對齊(Forced Alignment)至正確轉譯文句所形成的次詞圖(Sub-Lattice)之聯結(如圖2-3)，再使用原始詞圖中詞段的事後機率和次數(Count)來做刪除(Pruning)的動作，使得每個次詞圖只留下一個正確轉譯詞段和一個最具競爭力的詞段(如圖2-5)，通常稱為混淆對(Confusion Pair)，再將所形成具混淆對之狹縮詞圖展開成*N*-最佳序列，於是此*N*-最佳序列便會是具鑑別力的。若我們使用具混淆對之狹縮詞圖來取代*N*-最佳序列，則式(2.28)與式(2.29)可改寫為[Doumpiotis & Byrne 2004]：

$$\overline{\mu}_{sm} = \frac{\sum_{z=1}^{Z} \sum_{W' \in \mathbf{W}_{z,Pinched}} K(W', \mathbf{W}_{z,Pinched}) \sum_{t} \gamma_{sm}^{z,W'}(t) o_t + D_{sm} \mu_{sm}}{\sum_{z=1}^{Z} \sum_{W' \in \mathbf{W}_{z,Pinched}} K(W', \mathbf{W}_{z,Pinched}) \sum_{t} \gamma_{sm}^{z,W'}(t) + D_{sm}} \tag{2.30}$$

$$\overline{\Sigma}_{sm} = \frac{\sum_{z=1}^{Z} \sum_{W' \in \mathbf{W}_{z,Pinched}} K(W', \mathbf{W}_{z,Pinched}) \sum_{t} \gamma_{sm}^{z,W'}(t) o_t^2 + D_{sm}(\Sigma_{sm} + \mu_{sm}\mu_{sm}^T)}{\sum_{z=1}^{Z} \sum_{W' \in \mathbf{W}_{z,Pinched}} K(W', \mathbf{W}_{z,Pinched}) \sum_{t} \gamma_{sm}^{z,W'}(t) + D_{sm}} - \overline{\mu}_{sm}\overline{\mu}_{sm}^T \tag{2.31}$$

此即為最小化貝氏風險為基礎的鑑別式聲學模型訓練。

2.3.3.1 最差式狹縮詞圖最小化貝氏風險之模型訓練

為了加快計算編輯距離以利聲學模型估測，有學者提出最差式狹縮詞圖最小化貝氏風險之鑑別式訓練(One Worst Pinched Lattice MBRDT) [Doumpiotis & Byrne 2004]，所謂最差式狹縮詞圖就是要在狹縮詞圖中，找到一條與正確轉譯詞序列最不相像的詞序列，亦即找出與正確轉譯詞序列W_z之間有最大編輯距離的詞序列W_z^*：

$$W_z^* = \arg\max_{W' \in \mathbf{W}_{z,Pinched}} L(W_z, W') \tag{2.32}$$

這也就意味著這條詞序列包含最多可以用來調整模型的資訊。因此使用一個強性假設(Strong Assumption)，假設所有可能詞序列的集合 $\mathbf{W}_z$ 只包含兩條詞序列，記做 $\mathbf{W}_{z,worst}=\{W_z,W_z^*\}$，可將式(2.7)改寫如下：

$$\begin{aligned}\hat{\theta}&=\arg\min_{\theta}\sum_{z=1}^{Z}\sum_{W'\in\mathbf{W}_{z,worst}}l(W_z,W')P(W'\mid O_z;\theta)\\&=\arg\min_{\theta}\sum_{z=1}^{Z}L(W_z,W_z^*)P(W_z^*\mid O_z;\theta)\end{aligned}\tag{2.33}$$

在此強性假設下，事後機率總和只由 W_z 以及 W_z^* 所構成：

$$1=P(W_z\mid O_z;\theta)+P(W_z^*\mid O_z;\theta)\tag{2.34}$$

且正確轉譯詞序列 W_z 的貝氏風險為：

$$R(W_z\mid O_z)=L(W_z,W_z^*)P(W_z^*\mid O_z;\theta)\tag{2.35}$$

則相關的權重值 $K(W_z,\mathbf{W}_{z,worst})$ 及 $K(W_z^*,\mathbf{W}_{z,worst})$ 可寫為：

$$K(W_z,\mathbf{W}_{z,worst})=\left[R(W_z\mid O_z)-L(W_z,W_z)\right]P(W_z\mid O_z;\theta)=L(W_z,W_z^*)P(W_z^*\mid O_z;\theta)P(W_z\mid O_z;\theta)\tag{2.36}$$

$$K(W_z^*,\mathbf{W}_{z,worst})=\left[R(W_z\mid O_z)-L(W_z,W_z^*)\right]P(W_z^*\mid O_z;\theta)=-L(W_z,W_z^*)P(W_z^*\mid O_z;\theta)P(W_z\mid O_z;\theta)\tag{2.36*}$$

事實上權重值 $K(W_z,\mathbf{W}_{z,worst})$ 和 $K(W_z^*,\mathbf{W}_{z,worst})$ 只差一個負號。因此隱藏式馬可夫模型的平均值向量及共變異矩陣的更新式子可寫為[Doumpiotis & Byrne 2004]：

$$\overline{\mu}_{sm}=\frac{\sum_{z=1}^{Z}K(W_z,\mathbf{W}_{z,worst})\left(\sum_t\gamma_{sm}^{z,W_z}(t)o_t-\sum_t\gamma_{sm}^{z,W_z^*}(t)o_t\right)+D_{sm}\mu_{sm}}{\sum_{z=1}^{Z}K(W_z,\mathbf{W}_{z,worst})\left(\sum_t\gamma_{sm}^{z,W_z}(t)-\sum_t\gamma_{sm}^{z,W_z^*}(t)\right)+D_{sm}}\tag{2.37}$$

$$\overline{\Sigma}_{sm}=\frac{\sum_{z=1}^{Z}K(W_z,\mathbf{W}_{z,worst})\left(\sum_t\gamma_{sm}^{z,W_z}(t)o_t^2-\sum_t\gamma_{sm}^{z,W_z^*}(t)o_t^2\right)+D_{sm}(\Sigma_{sm}+\mu_{sm}\mu_{sm}^T)}{\sum_{z=1}^{Z}K(W_z,\mathbf{W}_{z,worst})\left(\sum_t\gamma_{sm}^{z,W_z}(t)-\sum_t\gamma_{sm}^{z,W_z^*}(t)\right)+D_{sm}}-\overline{\mu}_{sm}\overline{\mu}_{sm}^T\tag{2.38}$$

Step 1. 從 Z 句訓練語料中，經由辨識器產生出 Z 個詞圖。
Step 2. 將 Z 個詞圖強迫對齊至所屬的正確轉譯詞序列 W_z。
Step 3. 使對齊好的詞圖分割成次詞圖，並產生狹縮詞圖。
Step 4. 利用詞段之事後機率，將狹縮詞圖中混淆的地方強迫變成混淆對。
Step 5. 在混淆對裡，刪除出現次數較少的詞段。
Step 6. 從已經過刪除的狹縮詞圖中，擷取出最差之詞序列 W_z^*。
Step 7. 利用前向-後向演算法，對正確轉譯詞詞序列 W_z 和最差之詞序列 W_z^*，分別收集統計值 $\gamma_s^{z,W_z}(t)$ 和 $\gamma_s^{z,W^*}(t)$。
Step 8. 使用式(2.29)和式(2.30)更新隱藏式馬可夫模型的平均值向量和共變異矩陣。

圖 2-6 最差式最小化貝氏風險鑑別式聲學模型訓練之流程

因為所有可能詞序列的集合大大地減少，所以計算編輯距離變得有效率，在聲學模型訓練上，時間也相對地減少很多。若直接忽略權重值 $K(W_z, \mathbf{W}_{z,worst})$ 的計算，亦即使用零一函數當減損函數，則隱藏式馬可夫模型的平均值向量及共變異矩陣的更新式子可寫為：

$$\overline{\mu}_{sm} = \frac{\sum_{z=1}^{Z}\left(\sum_t \gamma_{sm}^{z,W_z}(t)o_t - \sum_t \gamma_{sm}^{z,W_z^*}(t)o_t\right) + D_{sm}\mu_{sm}}{\sum_{z=1}^{Z}\left(\sum_t \gamma_{sm}^{z,W_z}(t) - \sum_t \gamma_{sm}^{z,W_z^*}(t)\right) + D_{sm}} \tag{2.39}$$

$$\overline{\Sigma}_{sm} = \frac{\sum_{z=1}^{Z}\left(\sum_t \gamma_{sm}^{z,W_z}(t)o_t^2 - \sum_t \gamma_{sm}^{z,W_z^*}(t)o_t^2\right) + D_{sm}(\Sigma_{sm} + \mu_{sm}\mu_{sm}^T)}{\sum_{z=1}^{Z}\left(\sum_t \gamma_{sm}^{z,W_z}(t) - \sum_t \gamma_{sm}^{z,W_z^*}(t)\right) + D_{sm}} - \overline{\mu}_{sm}\overline{\mu}_{sm}^T \tag{2.30}$$

此即為最差式最小化貝氏風險鑑別式訓練的方法，此方法與最小化分類錯誤(Minimum Classification Error, MCE)[B. H. Juang *et al.* 1992]非常相似(將在2.4.5小節介紹)，差別在於最差式最小化貝氏風險鑑別式訓練只考慮了正確轉譯詞序列與最差的詞序列，而最小化分類錯誤則是考慮到所有辨識的詞序列。在圖2-6中，總結最差式最小化貝氏風險鑑別式訓練之流程。

2.3.4 最小化音素錯誤之模型訓練(Minimum Phone Error)

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，是以式(2.7)為出發，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：

$$\begin{aligned} F_{MPE}(\lambda) &= \sum_{z=1}^{Z} \sum_{W_i \in \mathbf{W}_z} p(W_i \mid O_z) A(W_i,W_z) \\ &= \sum_{z=1}^{Z} \sum_{W_i \in \mathbf{W}_z} \frac{p_\lambda(O_z \mid W_i) P(W_i)}{p(O_z)} A(W_i,W_z) \end{aligned} \tag{2.31}$$

其中 $p(O_z)$ 可用語音辨識器產生的詞圖 $\mathbf{W}_{z,lattice}$ (如圖2-1所示)來近似，因此目標函數可進一步表示成：

$$F_{MPE}(\lambda) \approx \sum_z \sum_{W_i \in \mathbf{W}_{z,lattice}} \frac{p_\lambda(O_z \mid W_i) P(W_i)}{\sum_{W_k \in \mathbf{W}_{z,lattice}} p_\lambda(O_z \mid W_k) P(W_k)} A(W_i,W_z) \tag{2.32}$$

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)。由於式(2.32)中的聲學模型為隱藏式馬可夫模型，有潛藏變數(Latent Variables，也就指隱藏式馬可夫模型的狀態序列)問題，同樣也需透過輔助函數來推導出模型參數訓練式。但由於式(2.32)是有理函數(Rational Function)，並沒有所謂的「強性」輔助函數，僅能透過有所謂的「弱性」輔助函數來求取新的模型參數。而當對弱性輔助函數求極值時模型參數設定的並不能保證能讓原目標函數(式(2.32))值增大，因此通常會再加入所謂「平滑」函數於弱性

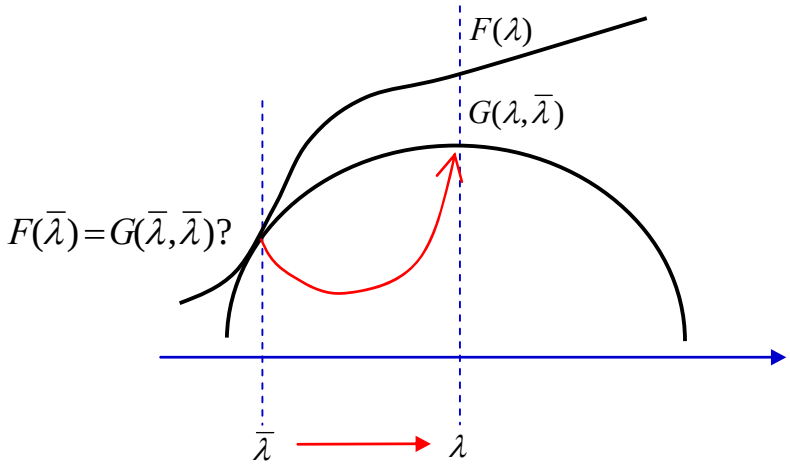


圖2-7 強性輔助函數G與目標函數F關係之示意圖。其中G與F要在舊有模型參數$\bar{\lambda}$設定時有相同的斜率，且滿足G為F的下界。

輔助函數，來限制模型參數的調整量。以下將對這三種不同性質的輔助函數作簡要的介紹：

(a)強性輔助函數(Strong-sense Auxiliary Function)

假設$\bar{\lambda}$為舊有已知的模型參數，λ為欲求取的新模型參數(為一變數，可視為隱藏式馬可夫模型中某個狀態的高斯混合模型的係數)，F為目標函數，G為強性輔助函數，則F與G需滿足：

$$\left.\frac{\partial}{\partial\lambda}G(\lambda,\bar{\lambda})\right|_{\lambda=\bar{\lambda}}=\left.\frac{\partial}{\partial\lambda}F(\lambda)\right|_{\lambda=\bar{\lambda}} \tag{2.33}$$

即F與G在舊有模型參數$\bar{\lambda}$設定時有相同的斜率(相同的變化趨勢)，同時也必需滿足不等式：

$$G(\lambda,\bar{\lambda})-G(\bar{\lambda},\bar{\lambda})\le F(\lambda)-F(\bar{\lambda}) \tag{2.34}$$

或者是：

$$F(\bar{\lambda})-G(\bar{\lambda},\bar{\lambda})\le F(\lambda)-G(\lambda,\bar{\lambda}) \tag{2.35}$$

式(2.34)與式(2.35)指在新模型參數λ設定下，強性輔助函數G的值增加量恆小於等於目標函數F的增加量(可視強性輔助函數G為目標函數F的下界)；也就是只

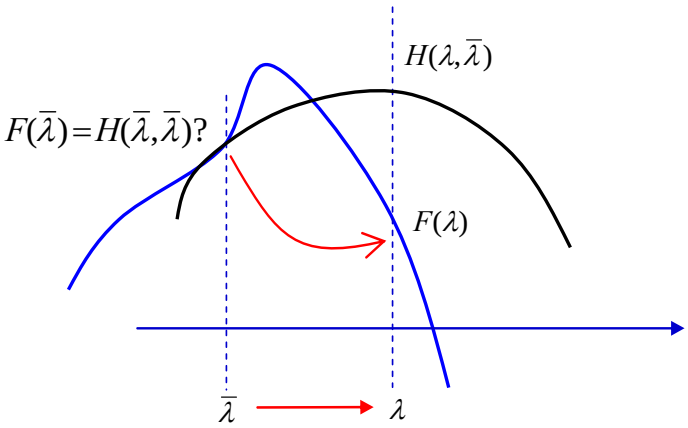


圖2-8 弱性輔助函數H與目標函數F關係之示意圖。其中H與F並不一定在舊有模型參數$\bar{\lambda}$設定時相切於一點，僅需要在舊有模型參數$\bar{\lambda}$設定時有相同的斜率。

要輔助函數G在新的模型參數λ設定下有較大的函數值，那麼目標函數也同樣地會有較大的新函數值，如圖2-7所意示。

(b)弱性輔助函數(Weak-sense Auxiliary Function)

弱性輔助函數H僅需滿足：

$$\left.\frac{\partial}{\partial\lambda}H(\lambda,\bar{\lambda})\right|_{\lambda=\bar{\lambda}}=\left.\frac{\partial}{\partial\lambda}F(\lambda)\right|_{\lambda=\bar{\lambda}} \tag{2.36}$$

即F與H在舊有模型參數$\bar{\lambda}$設定時有相同的斜率(相同的變化趨勢)，但不保證在以H所求得的新參數設定λ下能讓F的函數值增大，如圖2-8所意示。

(c)平滑函數(Smoothing Function)

平滑函數H_{SM}通常配合弱性輔助函數一起使用，滿足在舊有參數值$\bar{\lambda}$時有極值：

$$\left[\left.\frac{\partial H_{SM}(\lambda,\bar{\lambda})}{\partial\lambda}\right|_{\lambda=\bar{\lambda}}\right]=0 \tag{2.37}$$

而新的弱性輔助函數H'可表示成：

$$H'(\lambda,\bar{\lambda})=H(\lambda,\bar{\lambda})+H_{SM}(\lambda,\bar{\lambda}) \tag{2.38}$$

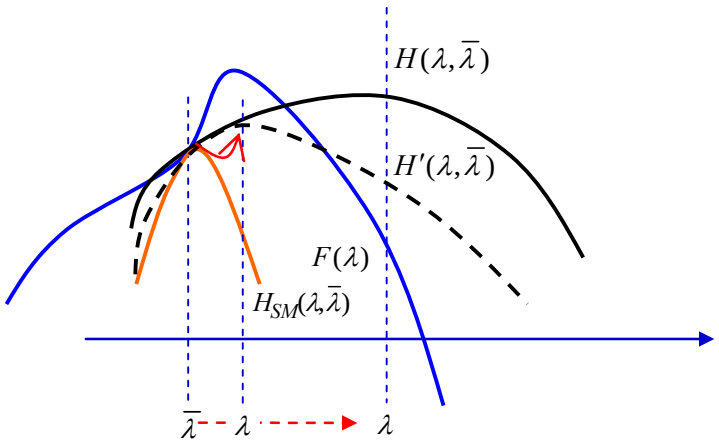


圖2-9 平滑函數H_{SM}與弱性輔助函數H、目標函數F關係之示意圖。其中H_{SM}在舊有模型參數值$\bar{\lambda}$時有極值。

如圖2-9所示，平滑函數H_{SM}可以限制模型參數的調整量，儘量使得目標函數F與弱性輔助函數H'在新模型參數λ設定下能保持函數值增加的趨勢。

因此，對於最小化音素錯誤鑑別式聲學訓練而言，我們需找到一個弱性輔助函數H來作為模型參數估測使用，它必須滿足與目標函數F在舊有模型參數值$\bar{\lambda}$下有相同的斜率。由微積分中的鏈鎖法則得知：

$$\frac{\partial F(\lambda)}{\partial \lambda}=\frac{\partial g(\lambda)}{\partial \lambda}\frac{\partial F(\lambda)}{\partial g(\lambda)},$$
$$\frac{\partial H(\lambda)}{\partial \lambda}=\frac{\partial g(\lambda)}{\partial \lambda}\frac{\partial H(\lambda)}{\partial g(\lambda)} \qquad (2.39)$$

因此，若將弱性輔助函數H設為$H(\lambda,\bar{\lambda})=\left[\frac{\partial F(\lambda)}{\partial g(\lambda)}\Big|_{\lambda=\bar{\lambda}}\right]g(\lambda)$的形式，其中$\frac{\partial F(\lambda)}{\partial g(\lambda)}\Big|_{\lambda=\bar{\lambda}}$為常數值，則可保證$H$與$F$在舊有模型參數值$\bar{\lambda}$下有相同的斜率。當$g(\lambda)$表示成$\log p_{\lambda}(O_z|q)$，也就是詞圖上某一個候選詞中的音素段落$q$對應之隱藏式馬可夫模型產生訓練語句(語音特徵向量序列)O_z的對數機率函數，則最小化音素錯誤的弱性輔助函數可表示成：

$$H_{MPE}(\lambda,\bar{\lambda})=\sum_{z=1}^{Z}\sum_{q\in\mathbf{W}_{z,lattice}}\left[\frac{\partial F_{MPE}(\lambda)}{\partial\log p_{\lambda}(O_z\mid q)}\Big|_{\lambda=\bar{\lambda}}\right]\log p_{\lambda}(O_z\mid q) \tag{2.40}$$

其中 $\frac{\partial F_{MPE}(\lambda)}{\partial\log p_{\lambda}(O_z\mid q)}\Big|_{\lambda=\bar{\lambda}}$ 的值可為正或負，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音素的平均期望正確率 c_{avg}^{z}。我們可先對目標函數作整理，將詞圖上的所有可能候選詞序列分成兩類，即包含音素段落 q 與否：

$$\begin{aligned}F_{MPE}(\lambda)&=\sum_{z=1}^{Z}\frac{\sum_{W_i\in\mathbf{W}_{z,lattice}}p_{\lambda}(O_z\mid W_i)P(W_i)A(W_i,W_z)}{\sum_{W_i\in\mathbf{W}_{z,lattice}}p_{\lambda}(O_z\mid W_k)P(W_k)}\\&=\sum_{z=1}^{Z}\frac{\sum_{W_i\in\mathbf{W}_{z,lattice},q\in W_i}p_{\lambda}(O_z\mid W_i)P(W_i)A(W_i,W_z)+\sum_{W_i\in\mathbf{W}_{z,lattice},q\notin W_i}p_{\lambda}(O_z\mid W_i)P(W_i)A(W_i,W_z)}{\sum_{W_k\in\mathbf{W}_{z,lattice},q\in W_k}p_{\lambda}(O_z\mid W_k)P(W_k)+\sum_{W_k\in\mathbf{W}_{z,lattice},q\notin W_k}p_{\lambda}(O_z\mid W_k)P(W_k)}\\&=\sum_{z=1}^{Z}\frac{a_z(\lambda)}{b_z(\lambda)}\end{aligned} \tag{2.41}$$

又在音素 q 的段落已知的情況下，會有以下的關係式：

$$\frac{\partial p_{\lambda}(O_z\mid W_i)}{\partial\log p_{\lambda}(O_z\mid q)}=p_{\lambda}(O_z\mid W_i)\ \ if\ q\in W_i \tag{2.42}$$

所以 $\frac{\partial F_{MPE}(\lambda)}{\partial\log p_{\lambda}(O_z\mid q)}\Big|_{\lambda=\bar{\lambda}}$ 可表示成：

$$\frac{\partial F_{MPE}(\lambda)}{\partial\log p_{\lambda}(O_z\mid q)}\Big|_{\lambda=\bar{\lambda}}=\frac{\partial\frac{a_z(\lambda)}{b_z(\lambda)}}{\partial\log p_{\lambda}(O_z\mid q)}\Big|_{\lambda=\bar{\lambda}}=\frac{\frac{\partial a_z(\lambda)}{\partial\log p_{\lambda}(O_z\mid q)}b_z(\lambda)-a_z(\lambda)\frac{\partial b_z(\lambda)}{\partial\log p_{\lambda}(O_z\mid q)}}{b_z(\lambda)^2}\Big|_{\lambda=\bar{\lambda}} \tag{2.43}$$

可再進一步化簡成：

$$\frac{\partial F_{MPE}(\lambda)}{\partial \log p_\lambda(O_z \mid q)}\Big|_{\lambda=\bar{\lambda}} = \left[\frac{\frac{\partial a_z(\lambda)}{\partial \log p_\lambda(O_z \mid q)}}{b_z(\lambda)} - \frac{a_z(\lambda)}{b_z(\lambda)}\frac{\frac{\partial b_z(\lambda)}{\partial \log p_\lambda(O_z \mid q)}}{b_z(\lambda)}\right]_{\lambda=\bar{\lambda}}$$

$$= \left[\begin{array}{l} \dfrac{\partial \sum_{W_i \in \mathbf{W}_{z,lattice}, q \in W_i} p_\lambda(O_z \mid W_i) P(W_i) A(W_i, W_z)}{\partial \log p_\lambda(O_z \mid q)} \cdot \dfrac{1}{b_z(\lambda)} \\ -\dfrac{\partial \sum_{W_k \in \mathbf{W}_{z,lattice}, q \in W_k} p_\lambda(O_z \mid W_k) P(W_k)}{\partial \log p_\lambda(O_z \mid q)} \cdot \dfrac{a_z(\lambda)}{b_z(\lambda)^2} \end{array}\right]_{\lambda=\bar{\lambda}}$$

$$= \left[\begin{array}{l} \dfrac{\sum_{W_i \in \mathbf{W}_{z,lattice}, q \in W_i} p_{\bar{\lambda}}(O_z \mid W_i) P(W_i) A(W_i, W_z)}{\sum_{W_k \in \mathbf{W}_{z,lattice}: q \in W_k} p_{\bar{\lambda}}(O_z \mid W_k) P(W_k)} \cdot \dfrac{\sum_{W_k \in \mathbf{W}_{z,lattice}: q \in W_k} p_{\bar{\lambda}}(O_z \mid W_k) P(W_k)}{\sum_{W_k \in \mathbf{W}_{z,lattice}} p_{\bar{\lambda}}(O_z \mid W_k) P(W_k)} \\ -\dfrac{\sum_{W_k \in \mathbf{W}_{z,lattice}, q \in W_k} p_{\bar{\lambda}}(O_z \mid W_k) P(W_k)}{\sum_{W_k \in \mathbf{W}_{z,lattice}} p_{\bar{\lambda}}(O_z \mid W_k) P(W_k)} \cdot \dfrac{\sum_{W_i \in \mathbf{W}_{z,lattice}} p_{\bar{\lambda}}(O_z \mid W_i) P(W_i) A(W_i, W_z)}{\sum_{W_k \in \mathbf{W}_{z,lattice}} p_{\bar{\lambda}}(O_z \mid W_k) P(W_k)} \end{array}\right]$$

$$= \gamma_q^z \left(c_z(q) - c_{avg}^z\right) \tag{2.44}$$

其中：

$$\gamma_q^z = \frac{\sum_{W_i \in \mathbf{W}_{z,lattice}: q \in W_k} p_{\bar{\lambda}}(O_z \mid W_k) P(W_k)}{\sum_{W_k \in \mathbf{W}_{z,lattice}} p_{\bar{\lambda}}(O_z \mid W_k) P(W_k)} \tag{2.45}$$

為詞圖上通過音素段落 q 的候選詞序列的事後機率和，而

$$c_z(q) = \frac{\sum_{W_i \in \mathbf{W}_{z,lattice}, q \in W_i} p_{\bar{\lambda}}(O_z \mid W_i) P(W_i) A(W_i, W_z)}{\sum_{W_k \in \mathbf{W}_{z,lattice}: q \in W_k} p_{\bar{\lambda}}(O_z \mid W_k) P(W_k)} \tag{2.46}$$

為詞圖上通過此音素段落的候選詞序列的期望正確率，而

$$c_{avg}^z = \frac{\sum_{W_i \in \mathbf{W}_{z,lattice}} p_{\bar{\lambda}}(O_z \mid W_i) P(W_i) A(W_i, W_z)}{\sum_{W_k \in \mathbf{W}_{z,lattice}} p_{\bar{\lambda}}(O_z \mid W_k) P(W_k)} \tag{2.47}$$

為詞圖上所有候選詞序列的期望正確率。

另一方面，由於對數機率函數 $\log p_\lambda(O_z \mid q)$ 還是有有潛藏變數(Latent Variables，也就是指我們無法直接觀測隱藏式馬可夫模型所有狀態序列個別產生訓練語句 O_z 的機率)問題，不能直接對其最佳化，必需透過一個強性輔助函數

$Q_{ML}(\lambda,\bar{\lambda},z,q)$(與傳統最大化相似度訓練法的輔助函數相同)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：

$$H'_{MPE}(\lambda,\bar{\lambda})=\sum_{z=1}^{Z}\sum_{q\in\mathbf{W}_{z,lattice}}\left[\frac{\partial F_{MPE}(\lambda)}{\partial\log p_{\lambda}(O_z\mid q)}\Big|_{\lambda=\bar{\lambda}}\right]Q_{ML}(\lambda,\bar{\lambda},z,q) \tag{2.48}$$

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:

$$Q_{ML}(\lambda,\bar{\lambda},z,q)=\sum_{t=s_q}^{e_q}\sum_{m}\gamma_q^z(t)\log N(o_z(t),\mu_{qm},\Sigma_{qm}) \tag{2.48*}$$

則弱性輔助函數 $H_{MPE}(\lambda,\bar{\lambda})$ 可表示成：

$$H'_{MPE}(\lambda,\bar{\lambda})=\sum_{z}\sum_{q\in\mathbf{W}_{z,lattice}}\sum_{t=s_q}^{t=e_q}\sum_{m}\gamma_q^{z\,MPE}\,\gamma_{qm}^z(t)\log N(o_z(t),\mu_{qm},\Sigma_{qm}) \tag{2.49}$$

其中 s_q 與 e_q 分別為音素段落 q 的開始與結束時間，$\gamma_{qm}^z(t)$ 為時間 t 時語音特徵向量落在隱藏式馬可夫模型 q 的第 m 個高斯分布的機率， $N(o_z(t),\mu_{qm},\Sigma_{qm})$ 第 m 個高斯分布。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：

$$\begin{aligned}H''_{MPE}(\lambda,\bar{\lambda})=&\sum_{z}\sum_{q\in\mathbf{W}_{z,lattice}}\sum_{t=s_q}^{t=e_q}\sum_{m}\gamma_q^{z\,MPE}\gamma_{qm}^z(t)\log N(o_z(t),\mu_{qm},\Sigma_{qm})\\&-\sum_{q,m}\frac{D_{qm}}{2}\left[\log(|\Sigma_{qm}|)+(\mu_{qm}-\bar{\mu}_{qm})^T\Sigma_{qm}^{-1}(\mu_{qm}-\bar{\mu}_{qm})+tr(\bar{\Sigma}_{qm}\Sigma_{qm}^{-1})\right]\end{aligned} \tag{2.50}$$

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

$$H_{SM}=\sum_{q,m}\frac{D_{qm}}{2}\left[\log(|\Sigma_{qm}|)+(\mu_{qm}-\bar{\mu}_{qm})^T\Sigma_{qm}^{-1}(\mu_{qm}-\bar{\mu}_{qm})+tr(\bar{\Sigma}_{qm}\Sigma_{qm}^{-1})\right] \tag{2.51}$$

其中 $\bar{\mu}_{qm}$ 與 $\bar{\Sigma}_{qm}$ 舊有模型的平均值向量與共變異矩陣。我們可以對 $H'_{MPE}(\lambda,\bar{\lambda})$ 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為無關時，亦即共變異矩陣為對角矩陣)[Povey 2004]：

$$\mu_{qmd} = \frac{\{\theta^{num}_{qmd}(O) - \theta^{den}_{qmd}(O)\} + D_{qmd}\overline{\mu}_{qmd}}{\{\gamma^{num}_{qm} - \gamma^{den}_{qm}\} + D_{qmd}}$$
$$\sigma^2_{qmd} = \frac{\{\theta^{num}_{qmd}(O^2) - \theta^{den}_{qmd}(O^2)\} + D_{qmd}(\overline{\sigma}^2_{qmd} + \overline{\mu}^2_{qmd})}{\{\gamma^{num}_{qm} - \gamma^{den}_{qm}\} + D_{qmd}} - \mu^2_{qmd} \quad (2.52)$$

其中統計值資訊可分為兩類，亦即 *num* (numerator)與 *den* (denominator)兩類，*num* 代表 $\gamma_q^{z^{MPE}}$ 為正時的統計值資訊(代表通過此音素段落 q 的所有候選詞序列期望正確率大於詞圖所有候選詞序列平均正確率，亦意謂此音素段落 q 較為可能為正確聲學模型)，*den* 代表 $\gamma_q^{z^{MPE}}$ 為負時的統計值資訊(代表通過此音素段落 q 的所有候選詞序列期望正確率小於詞圖所有候選詞序列平均正確率，亦意謂此音素段落 q 較為可能為錯誤聲學模型)，詳細統計資訊可分別表示如下：

$$\gamma^{num}_{qm} = \sum_{z=1}^{Z} \sum_{q \in \mathbf{W}_{z,lattice}} \sum_{t=s_q}^{e_q} \gamma^z_{qm}(t) \max(0, \gamma_q^{z^{MPE}}) \quad (2.53)$$

$$\theta^{num}_{qmd}(O) = \sum_{z=1}^{Z} \sum_{q \in \mathbf{W}_{z,lattice}} \sum_{t=s_q}^{e_q} \gamma^z_{qm}(t) \max(0, \gamma_q^{z^{MPE}}) o_z(t) \quad (2.54)$$

$$\theta^{num}_{qmd}(O^2) = \sum_{z=1}^{Z} \sum_{q \in \mathbf{W}_{z,lattice}} \sum_{t=s_q}^{e_q} \gamma^z_{qm}(t) \max(0, \gamma_q^{z^{MPE}}) o_z(t)^2 \quad (2.55)$$

$$\gamma^{den}_{qmd} = \sum_{z=1}^{Z} \sum_{q \in \mathbf{W}_{z,lattice}} \sum_{t=s_q}^{e_q} \gamma^z_{qm}(t) \max(0, -\gamma_q^{z^{MPE}}) \quad (2.56)$$

$$\theta^{den}_{qmd}(O) = \sum_{z=1}^{Z} \sum_{q \in \mathbf{W}_{z,lattice}} \sum_{t=s_q}^{e_q} \gamma^z_{qm}(t) \max(0, -\gamma_q^{z^{MPE}}) o_z(t) \quad (2.57)$$

$$\theta^{den}_{qmd}(O^2) = \sum_{z=1}^{Z} \sum_{q \in \mathbf{W}_{z,lattice}} \sum_{t=s_q}^{e_q} \gamma^z_{qm}(t) \max(0, -\gamma_q^{z^{MPE}}) o_z(t)^2 \quad (2.58)$$

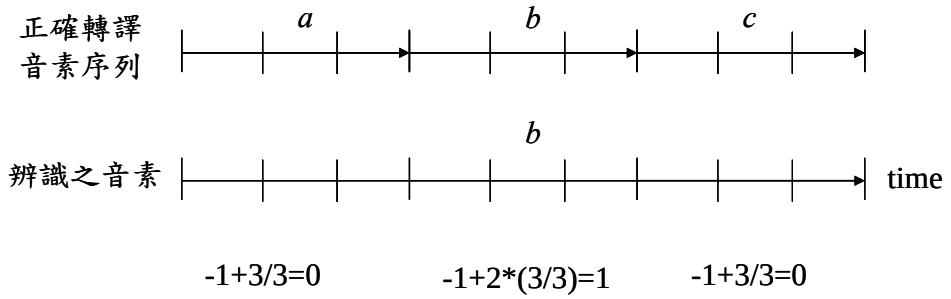


圖 2-10 原始音素正確率函數的範例

其中式(2.52)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，若 D_{qmd} 太大，則收斂的速度會非常的慢；若 D_{qmd} 太小，則會讓新的參數估測值不穩定，一般而言，D_{qmd} 的值都設為最小確保變異數為正數的兩倍。由式(2.52)到(2.58)可得知最小化音素錯誤聲學訓練為一種鑑別式訓練，也就它不但最大化正確聲學模型產生語音特徵向量的機率，也會同時嘗試最小化錯誤聲學模型產生語音特徵向量的機率。另外，為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[Povey 2004]，其公式如下：

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

$$\theta'^{num}_{qmd}(O^2) = \theta^{num}_{qmd}(O^2) + \frac{\tau_{qm}}{\gamma^{ML}_{qm}}\theta^{ML}_{qm}(O^2) \qquad (2.59)$$

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

其中 γ^{ML}_{qmd} 、 $\theta^{ML}_{qmd}(O)$ 及 $\theta^{ML}_{qmd}(O^2)$ 為使用傳統最大化相似度訓練法(使用僅使用正確文句資訊)求得的統計值資訊，τ_{qm} 為訓練時設定之常數。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：

$$A(q)=\max_{u}\begin{Bmatrix}-1+2e(q,u)\text{ if }u\text{ and }q\text{ are same phone}\\-1+e(q,u)\ \text{ if different phones}\end{Bmatrix} \tag{2.60}$$

其中$e(q,u)$為音素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為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計算式值最大者為音素q正確率(介於-1與1之間)，圖2-10為計算原始音素正確率的一個範例。而$\gamma_q^{z^{MPE}}$則可以由前向後向演算法(Forward-Backward Algorithm)求得。有關更詳細的最小化音素錯誤訓練的說明及在中文上的實驗結果，可以參考[郭人瑋 2005; Kuo *et al.* 2006]。

2.3.5 最小化分類錯誤之模型訓練(Minimum Classification Error)

前面 2.4.1 到 2.4.4 小節中，介紹了以全面風險為出發，定義了不同的減損函數及使用不同的所有詞序列可能之集合(詞圖、狹縮詞圖或 *N*-最佳序列)來設計模型訓練的目標函數。事實上，以最小化全面風險來設計分類器(如: 隱藏式馬可夫模型(HMM))，就是將分類器的設計問題轉化為模型分布的估測問題[B.H. Junag *et al.* 1997]，因為全面風險中有事後機率項，而其實我們根本不知道此事後機率的分布為何，所以我們會先假設此事後機率會遵循某種分布(如:高斯(Gaussian)分布)，藉由事先定義好的目標函數最大化過程中來調整模型的參數，所以使用最小化全面風險的觀念來設計分類器就是模型分布的估測問題。過去有學者基於以分類器的設計而提出最小化分類錯誤(MCE)來進行模型參數的調整[B.H. Junag *et al.* 1992]，在語音辨識的任務中，我們通常會使用隱藏式馬可夫模型(HMM)來當成分類器，儘管 HMM 有一些假設[Rabiner 1989]，但在目前語音辨識的效能表現上有還算不差的效果。若我們現在以 HMM 當成分類器，接著說明如何利用最小化分類錯誤的概念來調整模型的參數，最小化分類錯誤的實現有三個步驟，且

假設所有可能詞序列的集合是 *N*-最佳序列(*N*-Best List)所組成，茲分述如下:

(1)首先要先定義鑑別函數(Discriminant Function):

$$g_i(O_z;\Lambda) = P(O_z \mid W_i) \tag{2.61}$$

Λ 為隱藏式馬可夫模型的參數，$P(O_z \mid W_i)$ 為給定某詞序列 W_i 產生語音特徵向量序列 O_z 的相似度。當定義了鑑別函數之後，就可以此鑑別函數當成分類的準則，則分類器會有以下的決定規則(Decision Rule):

$$F(O) = W_i \quad if\ g_i(O;\Lambda) = \max_j g_j(O;\Lambda) \tag{2.62}$$

其中 $F(\cdot)$ 為分類器，O 為未知的語音特徵向量序列。

(2)接著要定義分類錯誤估量(Misclassification Measure):

$$d_z(O_z) = -g_z(O_z;\Lambda) + \log\left[\frac{1}{N-1}\sum_{j,j\neq z}\exp\left(g_j(O_z;\Lambda)\eta\right)\right]^{1/\eta} \tag{2.63}$$

其中 $g_z(O_z;\Lambda)$ 為正確詞序列 W_z 的鑑別函數，$g_j(O_z;\Lambda)$ 為其他詞序列 W_j 的鑑別函數，N 為所有可能最佳序列的個數，η 是一個正數，用來控制其他詞序列與正確詞序列的相似度之差異程度，若 η 接近無限大($\eta \to \infty$)，$[\cdot]$ 的其他詞序列會被具有相似度最大的詞序列所取代，即:

$$d_z(O_z) = -g_z(O_z;\Lambda) + \max_{j,j\neq z} g_j(O_z;\Lambda) \quad when\ \eta \to \infty \tag{2.64}$$

由分類錯誤估量的定義(式(2.63))可知，當 $d_z(O_z) \le 0$，代表第 z 個語音特徵向量序列 O_z 被分類正確；若 $d_z(O_z) > 0$，則代表第 z 個語音特徵向量序列 O_z 分類錯誤。

(3)最後，定義一個減損函數能使分類錯誤估量的值域為 0 到 1 之間的實數，任何能使值域為 0 到 1 之間的實數之函數都可以使用，最常使用的就是 S 型函數(Sigmoid Function):

$$l(d_z(O_z)) = \frac{2}{1+\exp(-\alpha d_z(O_z)+\beta)} \tag{2.65}$$

其中 α 及 β 為 S 型函數中可調整的參數，α 控制 S 型函數的曲度(Slope)，β 則

控制 S 型函數的平移(Shift)，在最小化分類錯誤的模型訓練上，α 的設定為大於等於 1 的實數，即 $\alpha \geq 1$，β 通常設為 0($\beta = 0$)。使用 S 型函數有個好處，就是分類錯誤估量值域範圍縮到 0 至 1 之間，可以代表為分類錯誤的機率，即 $l(\cdot) = 0$ 為分類正確，分類錯誤的機率為 0，若 $l(\cdot) = 1$ 為完全分類錯誤，分類錯誤的機率為 1。

按照上面所描述的三個步驟，我們可以得知某個語音特徵向量序列在目前辨識器中的分類錯誤機率，則所有語音特徵向量序列 O 的分類錯誤率可以表示成:

$$R = \int l(d(O))1(d(O) > 0)dO \tag{2.66}$$

其中 $1(\cdot)$ 為指示函數，表示為:

$$1(\cdot) = \begin{cases} 1 \ if \ d(O) > 0 \\ 0 \ if \ d(O) \leq 0 \end{cases} \tag{2.67}$$

假設我們可以收集到有限的訓練語料(語音特徵向量序列)，則式(2.66)可以近似為:

$$R = \sum_{z=1}^{Z} l(d_z(O_z))1(d_z(O_z) > 0) \tag{2.68}$$

式(2.68)即為最小分類錯誤訓練想要最小化的目標函數，即最小化分類錯誤率，在實作上通常都會做正規化的動作，所以最小化分類錯誤的目標函數可寫成:

$$F_{MCE}(\Lambda) = \frac{1}{Z}\sum_{z=1}^{Z} l(d_z(O_z;\Lambda))1(d_z(O_z;\Lambda) > 0) \tag{2.69}$$

上述的目標函數一般都是用一般化機率遞減(Generalized Probabilistic Descent, GPD)[B.H. Junag *et al.* 1992]來調整模型的參數:

$$\Lambda_{t+1} = \Lambda_t - \varepsilon_t U_t \nabla F_{MCE}(\Lambda)|_{\Lambda=\Lambda_t} \tag{2.70}$$

其中 U_t 為一正定矩陣(Positive Definite Matrix)，t 為迭代子(Iterator)，ε_t 為一調整量(Step Size)，用來決定這次迭代應該要調整多少。

最小化分類錯誤最早是在 1992 被提出，已有十多年的發展歷史，相關的文獻也非常豐富，可以參考[B.H. Junag *et al.* 1992; B.H. Juang *et al.* 1997; McDermott *et al.* 1997; Katagiri *et al.* 1998; Schlüter *et al.* 2001]。

第3章　最小化音素錯誤訓練之改進

本章首先在 3.1 小節回顧過去學者針對最小化音素錯誤訓練之改進方法，在 3.2 小節提出時間音框正確率函數來改進最小化音素錯誤訓練。最後在 3.3 小節，提出使用統計式的方法來近似每一個訓練語句的事前機率，以改進以全面風險為基礎的鑑別式訓練。

3.1 最小化音素錯誤訓練之變形

本小節探討近幾年來針對最小化音素錯誤(MPE)鑑別式聲學模型訓練的改進方法，分別為 3.1.1 小節的最小化音素音框錯誤之模型訓練、3.1.2 小節的以狀態為基礎的最小化貝氏風險之模型訓練以及 3.1.3 小節的最小化散度錯誤之模型訓練，茲介紹如下:

3.1.1　最小化音素音框錯誤之模型訓練

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，而對於插入錯誤(Insertion Errors)和取代錯誤(Substitution Errors)有給予適當的懲罰。

(2) 原始音素正確率函數是以音素為單位(Phone-by-phone)來做計算，如式(2.60)所示，且其值域範圍為-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為 1，因此遇到訓練語料不足時，模型訓練所收集到的統計值會不夠強健。

為了克服上述之問題，因此有學者提出了以時間為單位(Frame-by-frame)的音素音框正確率函數(Phone Frame Accuracy, PFA)[J. Zheng *et al.* 2005]:

$$PhoneFrameAccuracy(q) = \sum_{t=s_q}^{e_q} P(q_t \mid O) \tag{3.1}$$

其中O為語音特徵向量序列，q為詞圖中某個音素段落，s_q為音素q的開始時間，e_q為音素q的結束時間，$P(q_t|O)$為給定O在某個時間點t音素段落q的事後機率(Posterior Probability)。式(3.1)之音素音框正確率函數的算法是累加某個音素在其時間段落中的事後機率值。在詞圖中一整條路徑(詞序列)W_i的音素音框正確率函數為:

$$PhoneFrameAcc(W_i) = \sum_{q \in W_i} PhoneFrameAccuracy(q) \tag{3.2}$$

以式(3.2)取代式(2.60)，則最小化音素音框錯誤訓練(Minimum Phone Frame Error, MPFE)的目標函數為[J. Zheng *et al.* 2005]:

$$\begin{aligned} F_{MPFE}(\lambda) &= \sum_{z=1}^{Z} \sum_{W_i \in \mathbf{W}_{z,lattice}} p(W_i|O_z) PhoneFrameAcc(W_i) \\ &= \sum_{z=1}^{Z} \sum_{W_i \in \mathbf{W}_{z,lattice}} \frac{p_\lambda(O_z|W_i)P(W_i)}{p(O_z)} PhoneFrameAcc(W_i) \end{aligned} \tag{3.3}$$

另一方面，有學者提出以標記比對(Label Matching)為基礎的音框音素正確率函數來取代音素音框正確率函數，其音框音素正確率(Frame Phone Accuracy, FPA)函數可表示如下[Povey *et al.* 2007]:

$$FramePhoneAccuracy(q) = \sum_{t=s_q}^{e_q} \left\{ \begin{array}{l} 1 \text{ if } u \text{ and } q \text{ are same phone at time } t \\ 0 \text{ if different phones} \end{array} \right\} \tag{3.4}$$

其中q為詞圖中某個音素段落，u為正確音素段落。式(3.4)是以時間音框為單位的音素標記比對(Phone Label Matching)，若某個音素段落q與正確音素段落u在時間點t時的標記一樣，則音框音素正確率為1，反之為0。詞圖中一整條路徑(詞序列)W_i的音框音素正確率函數為:

$$FramePhoneAcc(W_i) = \sum_{q \in W_i} FramePhoneAccuracy(q) \tag{3.5}$$

若以式(3.5)取代式(2.60)，則最小化音框音素錯誤訓練(Minimum Frame Phone Error, MFPE)的目標函數為[Povey *et al.* 2007]:

$$F_{MFPE}(\lambda)=\sum_{z=1}^{Z}\sum_{W_i\in\mathbf{W}_{z,lattice}} p(W_i\mid O_z)FramePhoneAcc(W_i)$$
$$=\sum_{z=1}^{Z}\sum_{W_i\in\mathbf{W}_{z,lattice}} \frac{p_\lambda(O_z\mid W_i)P(W_i)}{p(O_z)}FramePhoneAcc(W_i) \tag{3.6}$$

3.1.2 以狀態為基礎的最小化貝氏風險之模型訓練(s-MBR)

最小化貝氏風險在模型訓練上是一種一般化(General)的方法，而最小化音素錯誤只是它的一個特例，由 2.4.3 小節所述，最小化貝氏風險在模型訓練上是使用編輯距離作為減損函數(Loss Function)，在詞圖的實作上有其困難，所以才會使用最小化音素錯誤(MPE)訓練的原始音素正確率函數來逼近編輯距離的算法[Povey 2004]。由 3.1.1 小節所述，最小化音框音素錯誤(MFPE)的音框音素正確率函數是在某時間點 t 時以音素為單位來做標記比對，其時可以比對更細的單位(如狀態)，因此有學者提出以狀態(State)為單位來做類別的標記比對，其狀態正確率(State Accuracy)函數可表示如下[Gibson *et al.* 2006]:

$$stateAccuracy(q)=\sum_{t=s_q}^{e_q}\left\{\begin{array}{l}1 \text{ if } \mathrm{s}(u_t) \text{ and } \mathrm{s}(q_t) \text{ are same state at time } t\\ 0 \text{ otherwise}\end{array}\right\} \tag{3.7}$$

其中 $s(u_t)$ 為正確音素段落 u 在時間點 t 的狀態類別，$s(q_t)$ 為詞圖中某音素段落 q 在時間點 t 的狀態類別。詞圖中一整條路徑(詞序列) W_i 的狀態正確率函數為:

$$stateAcc(W_i)=\sum_{q\in W_i} stateAccuracy(q) \tag{3.8}$$

式(3.8)取代式(2.60)，則以狀態為基礎的最小化貝氏風險之模型訓練(State Based Minimum Bayes Risk, s-MBR)的目標函數為[Gibson *et al.* 2006]:

$$F_{s-MBR}(\lambda)=\sum_{z=1}^{Z}\sum_{W_i\in\mathbf{W}_{z,lattice}} p(W_i\mid O_z)stateAcc(W_i)$$
$$=\sum_{z=1}^{Z}\sum_{W_i\in\mathbf{W}_{z,lattice}} \frac{p_\lambda(O_z\mid W_i)P(W_i)}{p(O_z)}stateAcc(W_i) \tag{3.9}$$

3.1.3 最小化散度(Minimum Divergence)之模型訓練

最小化音素錯誤(MPE)訓練中的原始音素正確率函數、3.1.1 小節中的音框音素正確率函數與 3.1.2 小節中的狀態正確率函數都是以類別比對為基礎的，這樣子的類別比對很容易受到類別定義(如:狀態、音素、詞...等)及語言模型的影響，而且如果使用全詞模型(Whole Word Model)來作為聲學模型的話，這樣便沒有清楚的音素類別定義，因此沒有辦法計算音素的正確率，為了改進此一缺失，因此有學者提出以聲學模型之間的相似程度來取代以類別比對為基礎的正確率函數[Jun Du *et al.* 2006]，稱為最小化散度鑑別式訓練(Minimum Divergence Discriminative Training, MD)。在統計的方法上，要計算兩個模型分布的相似程度，可以使用 KL 距離(KL Distance)來計算其相似程度:

$$\begin{aligned} D(W_i \| W_z) &= D(s_i^{1:T} \| s_z^{1:T}) \\ &= \int p(o^{1:T} \mid s_i^{1:T}) \log \frac{p(o^{1:T} \mid s_i^{1:T})}{p(o^{1:T} \mid s_z^{1:T})} do^{1:T} \end{aligned} \tag{3.10}$$

其中 $o^{1:T}$ 為時間點 1 到 T 的語音特徵向量序列。$s_i^{1:T}$ 為詞圖中某一個路徑 W_i 從時間點 1 到時間點 T 的狀態序列，$s_z^{1:T}$ 為正確詞序列 W_z 從時間點 1 到時間點 T 的狀態序列，由式(3.10)可以知道此 KL 距離是非對稱的。由於我們會假設每一個時間點都是獨立的(Independent)，所以我們可以將式(3.10)進一步寫成:

$$D(s_i^{1:T} \| s_z^{1:T}) \approx \sum_{t=1}^{T} D(s_i^t \| s_z^t) \tag{3.11}$$

在語音辨識的應用情境中，我們通常都會假設狀態是由高斯混合模型(GMM)所組成，所以在每個時間點 t，某個狀態 s 的相似度(Likelihood)為:

$$p(o_t \mid s) = \sum_{m=1}^{M_s} w_{sm} N(o_t; \mu_{sm}, \Sigma_{sm}) \tag{3.12}$$

其中 M_s 為狀態 s 中所有的高斯個數，w_{sm} 為在狀態 s 中某個高斯 m 的權重。如此一來，要計算聲學模型的相似度便可簡化為計算兩個狀態中的高斯混合模型的 KL 距離，但由於計算 KL 距離需要作積分，而且沒有閉析解(Closed-form

Solution)，這使得實作上有困難，因此有學者提出近似解[Jun Du *et al.* 2006]:

$$D(s_i \| s_z) \approx \frac{1}{2N}\sum_{m=1}^{M_s} w_{s_i m} \sum_{k=1}^{2N} \log \frac{p(o_{m,k} \mid s_i)}{p(o_{m,k} \mid s_z)} \tag{3.13}$$

其中 N 為語音特徵的維度，$o_{m,k}$ 為第 m 個的高斯的第 k 個人造特徵值，又稱為 Sigma Point，其算法如下:

$$o_{m,k} = \mu_m + \left(\sqrt{N \cdot \lambda_{m,k}}\right) \cdot u_{m,k} \tag{3.14}$$

$$o_{m,k+N} = \mu_m - \left(\sqrt{N \cdot \lambda_{m,k}}\right) \cdot u_{m,k} \tag{3.15}$$

其中 $\lambda_{m,k}$ 為第 m 個的高斯的共變異矩陣(Covariance Matrix)中第 k 個特徵值(Eigenvalue)，$u_{m,k}$ 為第 m 個的高斯的共變異矩陣中第 k 個特徵值向量(Eigenvector)。由於我們是要將目標函數最大化，所以必須在算完 KL 距離之後，在其前面加上一個負號，將式(3.11)取代式(2.60)，則最小化散度(Minimum Divergence, MD)訓練的目標函數便可寫為[Jun Du *et al.* 2006]:

$$\begin{aligned} F_{MD}(\lambda) &= \sum_{z=1}^{Z} \sum_{W_i \in \mathbf{W}_{z,lattice}} p(W_i \mid O_z)\left[-D(W_i \| W_z)\right] \\ &= \sum_{z=1}^{Z} \sum_{W_i \in \mathbf{W}_{z,lattice}} \frac{p_\lambda(O_z \mid W_i) P(W_i)}{p(O_z)}\left[-D(W_i \| W_z)\right] \end{aligned} \tag{3.16}$$

事實上，最小化散度可能不會被最佳化(Optimized)，因為當聲學模型的參數變動會使得 KL 距離也跟著變動，亦即散度會隨著模型的調整而有所更動，所以散度是跟模型參數有絕對的關係。在最小化音素錯誤的原始音素正確率函數及前兩小節的改進方法中，其正確率都是固定的(Fixed)，故只要專心調整模型的參數即可。但是最小化散度為求實作上的方便，通常都會假設任兩個狀態(State)間的 KL 距離不會隨著模型參數最佳化的過程而有所更動，所以可以事先算好任兩個狀態間的 KL 距離，也因此最小化散度的訓練方法有可能不會被最佳化。有關此一缺失之改進，讀者可參考[Jun Du *et al.* 2007]。

在介紹了有關最小化音素錯誤的變形，可以發現事實上所有的變形都是在改

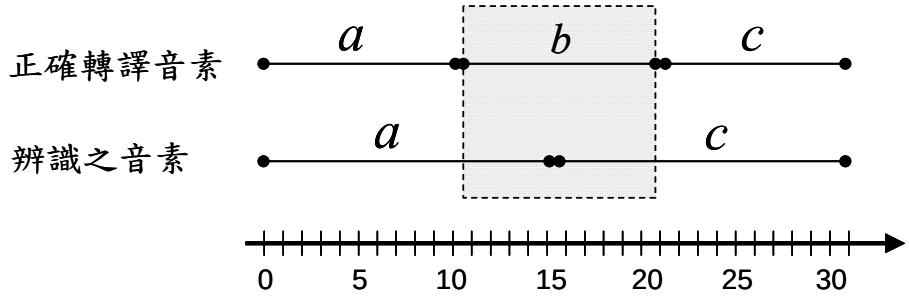
進減損函數以求得更好的模型訓練方式，而此減損函數在最小化貝氏風險的模型訓練中一直都是一個開放的議題，且是跟任務有關的(Task-dependent)，在表 3-1 中整理了最小化音素錯誤(MPE)的減損函數及一些減損函數的改進方法。

表 3-1 最小化音素錯誤訓練的減損函數與其他變形的減損函數之比較

目標函數	減損函數的表示式
MPE	$A(q)=\max_{u}\begin{Bmatrix}-1+2e(q,u)\text{ if } u \text{ and } q \text{ are same phone}\\ -1+e(q,u)\quad \text{if different phones}\end{Bmatrix}$
MPFE	$PhoneFrameAccuracy(q)=\sum_{t=s_q}^{e_q}P(q_t \mid O)$
MFPE	$FramePhoneAccuracy(q)=\sum_{t=s_q}^{e_q}\begin{Bmatrix}1 \text{ if } u \text{ and } q \text{ are same phone at time } t\\ 0 \text{ if different phones}\end{Bmatrix}$
s-MBR	$stateAccuracy(q)=\sum_{t=s_q}^{e_q}\begin{Bmatrix}1 \text{ if } \mathrm{s}(u_t) \text{ and } \mathrm{s}(q_t) \text{ are same state at time } t\\ 0 \text{ otherwise}\end{Bmatrix}$
MD	$-D(W_i \parallel W_z)=-D(s_i^{1:T} \parallel s_z^{1:T})$

3.2 時間音框音素正確率函數(Time Frame Accuracy Function)

如同3.1.1小節所描述，最小化音素錯誤訓練有兩個主要的缺點，第一個缺點是原始音素正確率函數並沒有懲罰刪除錯誤(Deletion Errors)，第二個缺點是原始音素正確率函數的值域範圍(介於-1到+1)過於狹窄。所以有學者提出最小化音素音框錯誤(MPFE)或最小化音框音素錯誤(MFPE)來改進，但事實上他們提出的方法只解決了遇到訓練語料不足時，模型訓練所收集到的統計值不夠強健的問題，仍然沒有給刪除錯誤一個適當的懲罰。舉一個例子來說明，如圖3-1所示，假設正確轉譯詞序列中有三個音素，即a、b和c，辨識之詞序列中有兩個，即a和c，則b即為刪除錯誤，那麼依最小化音素錯誤(MPE)訓練的原始音素正確率函數做計算



MPE之原始音素正確率= 2

MPFE之音素音框正確率= 介於0~30

MFPE之音框音素正確率= 20

MTFA之時間音框音素正確率= $\frac{2\cdot(10+5\cdot(-\rho))}{30}\approx 1.27\ (\rho=0.1)$

圖 3-1 最小化音素錯誤訓練及其變形對於刪除錯誤的影響與時間音框音素正確率計算示意圖

會得到2；因最小化音素音框錯誤(MPFE)的音素音框正確率函數是每個時間點音素的事後機率做累加，所以會得到介於0到30得實數值；若依最小化音框音素錯誤(MFPE)的音框音素正確率函數做計算會得到20。顯然地，上述三個方法都沒有適當地給予刪除錯誤懲罰。本論文針對此一缺點加以改善，因而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[Liu *et al.* 2007]：

$$TimeFrameAccuracy(q)=\frac{\sum_{t=s_q}^{e_q}\delta(q,u(t))}{e_q-s_q+1} \tag{3.17}$$

$$\delta(q,u(t))=\begin{Bmatrix}1 & ,if\ q=u(t)\\ -\rho & ,if\ q\neq u(t),0<\rho<1\end{Bmatrix} \tag{3.18}$$

其中 q 為詞圖中某一音素段落，s_q 和 e_q 分別為音素段落 q 的開始時間及結束時間，$u(t)$ 為正確音素段落 u 在時間 t 時的音素標記(Phone Label)，ρ 為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 的正確率值域範圍為介於 $-\rho$ 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[Wessel *et al.* 2001]，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3-1即為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有三個，即a、b和c，此語句的辨識音素有兩個，即a和c，那麼b就是刪除錯誤，在圖3-1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，我們應該要給予這些錯誤的時間音框一些刪除錯誤的懲罰。在詞圖中一整條路徑(詞序列)W_i的時間音框音素正確率為:

$$TimeFrameAcc(W_i) = \sum_{q \in W_i} TimeFrameAccuracy(q) \tag{3.19}$$

將式(3.19)取代式(2.60)，則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)的目標函數為:

$$\begin{aligned} F_{MTFA}(\lambda) &= \sum_{z=1}^{Z} \sum_{W_i \in \mathbf{W}_{z,lattice}} p(W_i \mid O_z) TimeFrameAcc(W_i) \\ &= \sum_{z=1}^{Z} \sum_{W_i \in \mathbf{W}_{z,lattice}} \frac{p_\lambda(O_z \mid W_i) P(W_i)}{p(O_z)} TimeFrameAcc(W_i) \end{aligned} \tag{3.20}$$

另外，為了更充分地懲罰刪除錯誤，本論文使用了 S 型函數(Sigmoid Function)來正規化式(3.17)的時間音框音素正確率函數的分子項，稱之為 S 型時間音框音素正確率函數(Sigmoid Time Frame Phone Accuracy, 記作 STFA):

$$SigTimeFrameAccuracy(q) = \frac{2}{1+\exp(-\alpha \cdot net + \beta)} - 1 \tag{3.21}$$

其中

$$net = \sum_{t=s_q}^{e_q} \delta(q, u(t)) \tag{3.22}$$

其$\delta(\cdot)$的定義同式(3.18)，α及β為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。式(3.21)的值域範圍為-1 到+1 之間。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為:

$$SigTimeFrameAcc(W_i) = \sum_{q \in W_i} SigTimeFrameAccuracy(q) \tag{3.23}$$

將式(3.23)取代式(2.60)，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)的目標函數為:

$$\begin{aligned} F_{MTFA}(\lambda) &= \sum_{z=1}^{Z} \sum_{W_i \in \mathbf{W}_{z,lattice}} p(W_i \mid O_z) SigTimeFrameAcc(W_i) \\ &= \sum_{z=1}^{Z} \sum_{W_i \in \mathbf{W}_{z,lattice}} \frac{p_\lambda(O_z \mid W_i) P(W_i)}{p(O_z)} SigTimeFrameAcc(W_i) \end{aligned} \tag{3.24}$$

本論文所提出的時間音框音素正確率函數主要不是去逼近編輯距離，而只是有考量到刪除錯誤的適當懲罰，以改進最小化音素錯誤(MPE)鑑別式聲學模型訓練。有關如何在詞圖中正確地計算編輯距離，可以參考[G. Heigold *et al.* 2005]。

3.3 考慮事前機率

在 2.2 小節中介紹了全面風險之於模型的估測，目前大部分的鑑別式訓練都是由全面風險出發，並假設訓練語句的事前機率(Prior Probability)均為一致(Uniform)，進而推導出實作上的目標函數。事實上，每一個訓練語句應該是要有不同的機率，所以本論文嘗試使用統計式的方式來近似訓練語句的事前機率。在鑑別式聲學模型訓練的應用情境中，事前機率可以視為是一個權重值(Weighting)，用來加權模型參數調整時所收集的統計值。因為語音特徵向量序列(訓練語句)O_z 的事前機率 $p(O_z)$ 無法直接求得，所以我們先假設訓練語句 O_z 布於一個給定的假設空間(Hypothesis Space) $\mathbf{W}$，通常此假設空間在語音辨識的應用情境中就是以詞圖(Word Graph or Lattice)來表示，記作 $\mathbf{W}_{lattice}$。現在我們便可以使用總體機率定律(Law of Total Probability)將事前機率展開:

$$p(O_z) = \sum_{W \in \mathbf{W}_{z,lattice}} p(O_z, W) \tag{3.25}$$

使用貝氏定理將聯合機率(Joint Probability)拆為聲學模型(Acoustic Model)與語言模型(Language Model)的相乘積:

$$p(O_z) = \sum_{W \in \mathbf{W}_{z,lattice}} p(O_z \mid W) P(W) \tag{3.26}$$

從式(3.26)中可以看出若語音特徵向量序列(訓練語句)所對應的詞序列越長，則此句的事前機率會越小，因為機率會越乘越小。為了克服此一問題，我們使用時間音框長度來作正規化(Normalization)，假設語音特徵向量序列 O_z 的時間音框長度為 T_{O_z} ，則其數學式可表示如下:

$$p(O_z) = \sqrt[T_{O_z}]{\sum_{W \in \mathbf{W}_{z,lattice}} p(O_z \mid W) P(W)} \tag{3.27}$$

在實作上，再對所有的訓練語句做正規化(Normalization):

$$\tilde{p}(O_z) = \frac{\sqrt[T_{O_z}]{\sum_{W \in \mathbf{W}_{z,lattice}} p(O_z \mid W) P(W)}}{\sum_{r=1}^{Z} \left[\sqrt[T_{O_r}]{\sum_{W \in \mathbf{W}_{z,lattice}} p(O_r \mid W) P(W)} \right]} \tag{3.28}$$

其中 Z 為所有可收集到的訓練語句。另外，考量實作上的效率，我們可以假設聲學模型的相似度(Likelihood)會被正確轉譯詞序列所支配(Domination)，亦即詞圖中只含有正確轉譯詞序列，故式(3.28)可表示成:

$$\tilde{p}(O_z) = \frac{\sqrt[T_{O_z}]{p(O_z \mid W_z) P(W_z)}}{\sum_{r=1}^{Z} \left[\sqrt[T_{O_r}]{p(O_r \mid W_r) P(W_r)} \right]} \tag{3.29}$$

其中 W_z 為語音特徵向量序列 O_z 所對應的正確轉譯詞序列，W_r 為語音特徵向量序列 O_r 所對應的正確轉譯詞序列。

由統計式方法所計算而得的事前機率 $\tilde{p}(O_z)$ 值越大，代表語音特徵向量序列(訓練語句) O_z 平均相似度越大，所以在收集統計值時，此訓練語句就會有比較大的權重。以最小化音素錯誤(MPE)訓練統計值的收集為例，每個訓練語句在收集統計值時要乘上一個事前機率(式(3.28)或式(3.29))，則其數學式可表示為:

$$\gamma_{qm}^{num} = \sum_{z=1}^{Z} \left[\sum_{q \in \mathbf{W}_{z.lattice}} \sum_{t=s_q}^{e_q} \gamma_{qm}^{z}(t) \max(0, \gamma_q^{z^{MPE}}) \right] \cdot \tilde{P}(O_z)$$

$$\theta_{qmd}^{num}(O) = \sum_{z=1}^{Z} \left[\sum_{q \in \mathbf{W}_{z.lattice}} \sum_{t=s_q}^{e_q} \gamma_{qm}^{z}(t) \max(0, \gamma_q^{z^{MPE}}) o_z(t) \right] \cdot \tilde{P}(O_z) \tag{3.30}$$

$$\theta_{qmd}^{num}(O^2) = \sum_{z=1}^{Z} \left[\sum_{q \in \mathbf{W}_{z.lattice}} \sum_{t=s_q}^{e_q} \gamma_{qm}^{z}(t) \max(0, \gamma_q^{z^{MPE}}) o_z(t)^2 \right] \cdot \tilde{P}(O_z)$$

$$\gamma_{qmd}^{den} = \sum_{z=1}^{Z} \left[\sum_{q \in \mathbf{W}_{z.lattice}} \sum_{t=s_q}^{e_q} \gamma_{qm}^{z}(t) \max(0, -\gamma_q^{z^{MPE}}) \right] \cdot \tilde{P}(O_z)$$

$$\theta_{qmd}^{den}(O) = \sum_{z=1}^{Z} \left[\sum_{q \in \mathbf{W}_{z.lattice}} \sum_{t=s_q}^{e_q} \gamma_{qm}^{z}(t) \max(0, -\gamma_q^{z^{MPE}}) o_z(t) \right] \cdot \tilde{P}(O_z) \tag{3.31}$$

$$\theta_{qmd}^{den}(O^2) = \sum_{z=1}^{Z} \left[\sum_{q \in \mathbf{W}_{z.lattice}} \sum_{t=s_q}^{e_q} \gamma_{qm}^{z}(t) \max(0, -\gamma_q^{z^{MPE}}) o_z(t)^2 \right] \cdot \tilde{P}(O_z)$$

第4章　資料選取方法於改進鑑別式聲學模型訓練

本章首先在 4.1 小節介紹以邊際為基礎的模型訓練方法，在 4.2 小節介紹本論文所提出的以正規化熵值為基礎的資料選取方法，用於改進鑑別式聲學模型訓練。

4.1 以邊際為基礎(Margin-based)的模型訓練

近年來，最大邊際分類器(Large Margin Classifier)在機器學習(Machine Learning)的領域中已有高度的發展。例如，支持向量機(Support Vector Machine, SVM)、布斯丁演算法(Boosting)等[A. Smola *et al.*]，在二元類別(Binary class)或多元類別(Multi-class)的分類(Classification)任務中，都可以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，而這是傳統鑑別式訓練所無法達到的，因為鑑別式訓練旨在最小化訓練樣本的分類錯誤，而並不是直接提升分類器的一般化能力。在觀念上，我們以二元類別且可分離的(Separable)訓練樣本為例，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，是一個超表面(Hypersurface)，能明確地將訓練樣本分為兩個類別)越遠越好，訓練樣本到決定邊界的距離我們一般會稱為邊際(Margin)，而邊際越大且邊際內沒有其它的訓練樣本代表其一般化能力及容錯能力會越好[Vapnik 1995]。在語音辨識的應用情境中，我們通常都是使用連續密度隱藏式馬可夫模型(CDHMM)來當成分類器，因此有學者利用最大邊際的概念來訓練連續密度隱藏式馬可夫模型，將在下一小節 4.1.1 節介紹，但此最大邊際訓練法則只有考慮模型的一般化能力，並沒有考慮到錯誤分類的樣本，因此有學者提出柔性邊際(Soft Margin)的概念來改善此一問題，將在 4.1.2 節介紹。

4.1.1 最大邊際估測法(Large Margin Estimation, LME)

要使用最大邊際的概念來最佳化連續密度隱藏式馬可夫模型則要先定義樣本分離估量(Separation Measure)。在語音辨識的應用情境中，通常都是使用語音特徵向量序列的相似度(Likelihood)來當作分離估量。給定某個語音特徵向量序列O_z及其正確對應的詞序列W_z，則其分離邊際(Separation Margin)可定義為[Li *et al.* 2005]:

$$\begin{aligned} d(O_z) &= p(O_z \mid W_z) - \max_{W_i \in \mathbf{W}, W_i \neq W_z} p(O_z \mid W_i) \\ &= \min_{W_i \in \mathbf{W}, W_i \neq W_z} \left[p(O_z \mid W_z) - p(O_z \mid W_i) \right] \end{aligned} \tag{4.1}$$

其中 **W** 為所有可能詞序列所成的集合，$P(O_z \mid W_z)$ 為給定正確詞序列W_z產生語音特徵向量序列O_z的相似度，W_i為語音特徵向量序列O_z可能的辨識結果(詞序列)，$P(O_z \mid W_i)$為給定可能的詞序列W_i產生語音特徵向量序列O_z的相似度。

由式(4.1)可以知道分離邊際的計算就是正確詞序列與最有可能的辨識詞序列的相似度之差，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;若$d(O_z) < 0$，則表示語音特徵向量序列O_z被目前的辨識器錯誤地辨識；若$d(O_z) = 0$，則表示語音特徵向量序列O_z剛好落在決定邊界上。在相似度定義域中，分離邊際為 0 即為決定邊界。最大邊際估測法就是要找出離決定邊界較近的語音特徵向量序列(亦即較有可能會被別的辨識器所辨識錯誤的語音特徵向量序列)，以進行隱藏式馬可夫模型參數的調整。首先要找出離決定邊界較近的語句，所以我們會先定義一個子集合S:

$$S = \{O_z \mid O_z \in R, 0 \le d(O_z) \le \gamma\} \tag{4.2}$$

其中R為所有的語音特徵向量序列(訓練語句)，γ為事先定義的門檻，為一個大於 0 的正實數，此子集合又可稱為支持向量集合(Support Vector Set)，在此集合裡的語音特徵向量序列O_z都是離決定邊界較近且可以被正確辨識的語音特徵向量序列，又可稱為支持標誌(Support Tokens)。在定義了支持向量集合後，最大邊際估測便是要使支持向量集合裡的支持標誌離決定邊界越遠越好，可以用下式來

進行最大邊際估測:

$$\overline{\Lambda} = \arg\max_{\Lambda} \min_{O_z \in S} d(O_z) \tag{4.3}$$

其中 Λ 為連續密度隱藏式馬可夫模型的參數，將式(4.1)帶入式(4.3)，則最大邊際估測的目標函數便為:

$$\overline{\Lambda} = \arg\max_{\Lambda} \min_{O_z \in S, W_i \in \mathbf{W}, W_i \neq W_z} \left[p(O_z \mid W_z) - p(O_z \mid W_i) \right] \tag{4.4}$$

且式(4.4)被限制在:

$$p(O_z \mid W_z) - p(O_z \mid W_i) > 0 \tag{4.5}$$

式(4.4)可以轉換為標準的最小最大最佳化問題(Minimax Optimization Problem):

$$\overline{\Lambda} = \arg\min_{\Lambda} \max_{O_z \in S, W_i \in \mathbf{W}, W_i \neq W_z} \left[p(O_z \mid W_i) - p(O_z \mid W_z) \right] \tag{4.6}$$

則式(4.6)被限制在:

$$p(O_z \mid W_i) - p(O_z \mid W_z) < 0 \tag{4.7}$$

則最大邊際的目標函數可表示為:

$$Q(\Lambda) = \max_{O_z \in S, W_i \in \mathbf{W}, W_i \neq W_z} \left[p(O_z \mid W_i) - p(O_z \mid W_z) \right] \tag{4.8}$$

要最佳化式(4.8)可以使用一般化機率遞減(Generalized Probabilistic Descent, GPD) [B.H. Junag *et al.* 1992]來求解，則連續密度隱藏式馬可夫模型(CDHMM)的參數最佳化如下所示(以平均值向量(Mean Vector)為例):

$$\mu_{sm}^{(n+1)} = \mu_{sm}^{(n)} - \varepsilon \frac{\partial Q(\Lambda)}{\partial \mu_{sm}} \bigg|_{\Lambda = \Lambda^{(n)}} \tag{4.9}$$

其中 n 為迭代子(Iterator)，s 、m 分別代表某個隱藏式馬可夫模型的狀態及高斯分布，ε 為一調整量(Step Size)，用來決定這次迭代應該要調整多少。以線性可分割的(Linear Separable)樣本為例，如圖 4-1 所示，其中□代表"+"類別，○代表"-"類別。在經最大邊際估測法的模型訓練後，如圖 4-2 所示，訓練樣本的分離邊際變大且可容錯的能力也變大，代表模型更具有一般化能力。有關更多最大邊際估測在語音辨識上的相關文獻，可參考[X. Li *et al.* 2005; X. Li *et al.* 2006; Jiang *et al.*

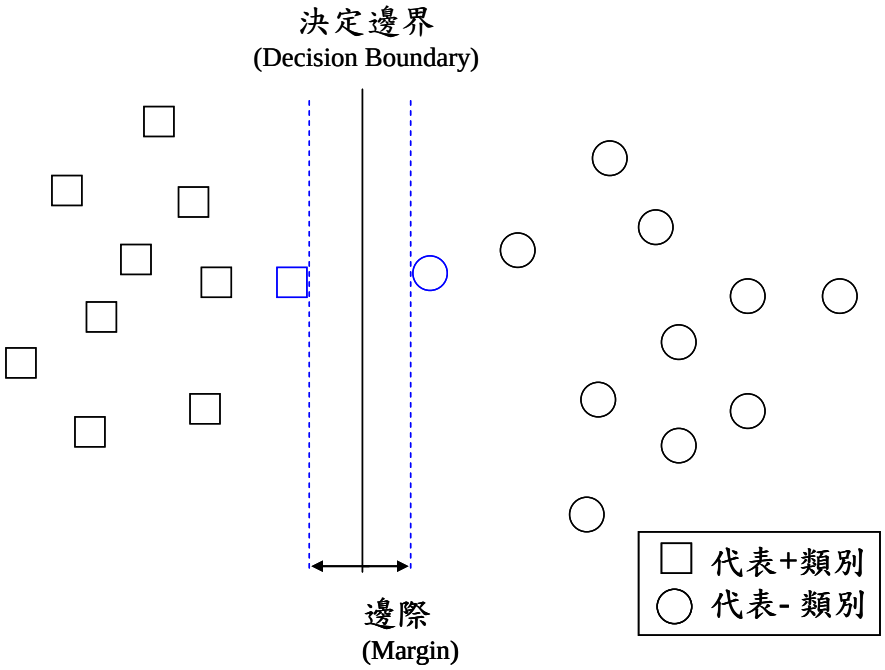


圖 4-1 最大邊際估測的模型訓練示意圖(未調整)

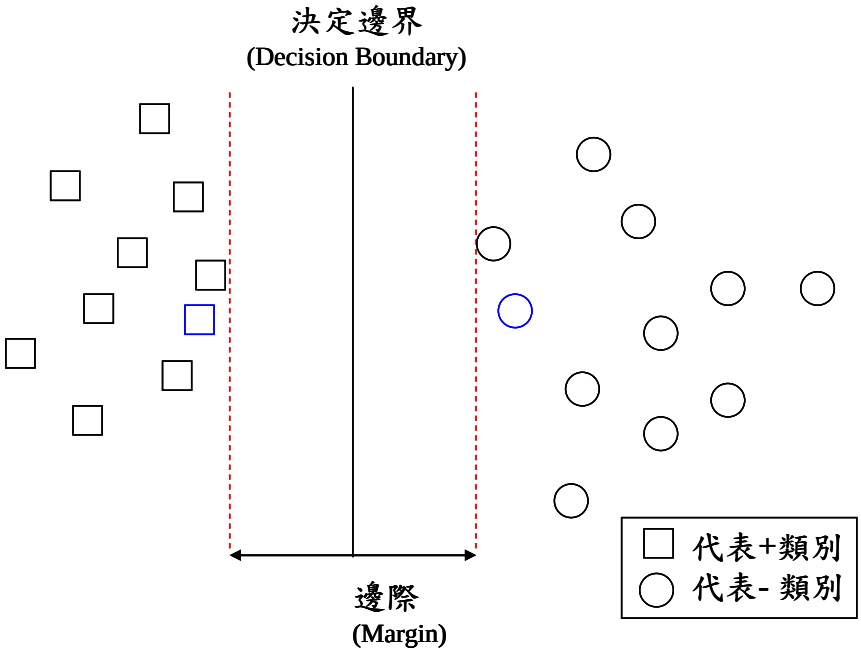


圖 4-2 最大邊際估測的模型訓練示意圖(調整後)

2006]。

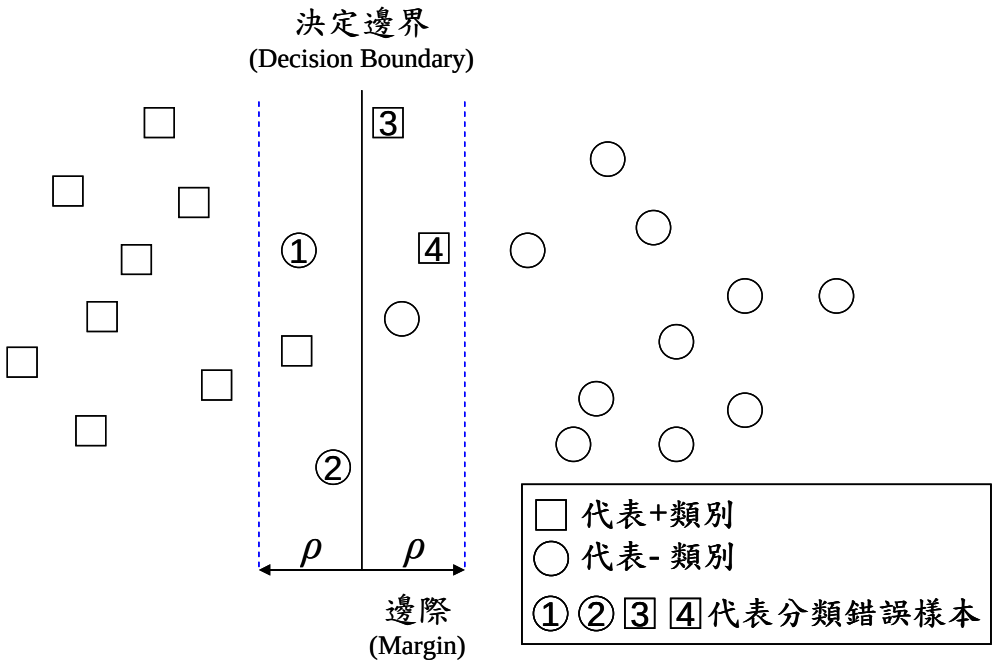


圖 4-3 柔性邊際之模型訓練示意圖

4.1.2 柔性邊際估測法(Soft Margin Estimation, SME)

前一小節 4.1.1 所描述的最大邊際估測法旨在直接提升模型的一般化能力，並沒有考慮到最小化訓練語句的錯誤分類，並且也沒有考慮到被錯誤分類的語句。以線性不可分割的樣本為例，如圖 4-3 所示，標號 1、2、3、4 為被錯誤分類的訓練語句，在使用最大邊際估測法(LME)時是忽略這些錯誤分類的語句，事實上，那些分類錯誤的語句對於鑑別式訓練有重要的資訊。為了改進最大邊際估測法的缺失，於是有學者提出柔性邊際估測法來訓練隱藏式馬可夫模型[Jinyu Li *et al.* 2006]，其分離邊際定義為:

$$d^{SME}(O_z) = \frac{1}{n_z}\sum_t \log\left[\frac{p(o_{zt}\mid W_z)}{p(o_{zt}\mid W_{z,c})}\right] I(o_{zt}\in F_z) \tag{4.10}$$

其中 $W_{z,c}$ 為語音特徵向量序列 O_z 所辨識的詞序列中相似度最大的詞序列，是與正確詞序列最為競爭(Most Competitive)的詞序列，F_z 為正確詞序列 W_z 與最競爭

詞序列$W_{z,c}$在每個時間點上，音素類別比對(Phone Label Matching)不同所形成的集合，n_z為F_z的個數，$p(o_{zt}|W_z)$為在時間點t給定正確詞序列W_z產生語音特徵向量o_{zt}的相似度，$p(o_{zt}|W_{z,c})$為在時間點t給定最競爭序列$W_{z,c}$產生語音特徵向量o_{zt}的相似度。$I(o_{zt}\in F_z)$為指示函數(Indicator Function)，可表示如下:

$$I(o_{zt}\in F_z)=\begin{Bmatrix}1\ if\ \ W_z(t)=W_{z,c}(t)\\0\ if\ \ W_z(t)\neq W_{z,c}(t)\end{Bmatrix} \tag{4.10*}$$

其中$W_z(t)$和$W_{z,c}(t)$分別為在時間點t時的正確詞序列之音素類別和最競爭詞序列之音素類別。所以式(4.10)為正確詞序列與最競爭詞序列的正規化對數相似度差之和。在定義了分離邊際之後，就可以定義柔性邊際估測(Soft Margin Estimation, SME)的目標函數為[Jinyu Li *et al.* 2006]:

$$L^{SME}(\Lambda)=\frac{\lambda}{\rho}+\frac{1}{N}\sum_{z=1}^{N}l(O_z,\Lambda) \tag{4.11}$$

其中Λ為隱藏式馬可夫模型參數，ρ為柔性邊際(Soft Margin)，λ為一常數，用來平衡柔性邊際的最大化與訓練語句分類錯誤最小化，N為所有收集到的語音特徵向量序列(訓練語句)，$l(O_z,\Lambda)$為一減損函數，即為柔性邊際估測法的分離邊際。通常語音特徵向量序列會發生分類錯誤都是因為減損函數所計算的分離邊際小於柔性邊際，因此我們重新定義柔性邊際的目標函數:

$$L^{SME}(\Lambda)=\frac{\lambda}{\rho}+\frac{1}{N}\sum_{z=1}^{N}\left(\rho-d_z^{SME}(O_z)\right)_+ \tag{4.12}$$

其中

$$\left(\rho-d_z^{SME}(O_z)\right)_+=\begin{Bmatrix}\rho-d_z^{SME}(O_z),\ if\ \ \rho-d_z^{SME}(O_z)>0\\0,\ otherwise\end{Bmatrix} \tag{4.13}$$

整合式(4.11)與式(4.12)則柔性邊際的目標函數為:

$$L^{SME}(\Lambda)=\frac{\lambda}{\rho}+\frac{1}{N}\sum_{z=1}^{N}\left(\rho-d_z^{SME}(O_z)\right)_+$$

$$=\frac{\lambda}{\rho}+\frac{1}{N}\sum_{z=1}^{N}\left(\rho-d_z^{SME}(O_z)\right)I(O_z\in U) \tag{4.14}$$

$$=\frac{\lambda}{\rho}+\frac{1}{N}\sum_{z=1}^{N}\left(\rho-\frac{1}{n_z}\sum_{t}\log\left[\frac{p(o_{zt}\mid W_z)}{p(o_{zt}\mid W_{z,c})}\right]I(o_{zt}\in F_z)\right)I(O_z\in U)$$

其中U為語音特徵向量序列的分離估量小於柔性邊際所形成的集合，F_z為正確詞序列W_z與最競爭詞序列$W_{z,c}$在每個時間點上音素類別比對不同所形成的集合。$I(O_z\in U)$為指示函數可表示如下:

$$I(O_z\in U)=\begin{cases}1,\ if\ \ \rho-d_z^{SME}(O_z)>0\\ 0,\ otherwise\end{cases} \tag{4.14*}$$

由式(4.14)可知柔性邊際的目標函數有兩層的選取資料(Data Selection)之動作，其一為語句(Utterance)的選取，即$I(O_z\in U)$；另一為時間音框(Frame)的選取，即$I(o_{zt}\in F)$。使用柔性邊際估測法可以考慮到錯誤分類的語句，如圖 4-3 標號為 1、2、3、4 之語句。事實上柔性邊際估測法中的柔性邊際ρ要當成一個變數，在進行隱藏式馬可夫模型最佳化的同時，也要考量到柔性邊際的最佳化，但此目標函數沒有閉析解，所以在實作上，要事先決定柔性邊際，然後使用一般化機率遞減(GPD)來進行模型參數的最佳化:

$$\overline{\Lambda}^{(n+1)}=\Lambda^{(n)}+\varepsilon\frac{\partial L^{SME}(\Lambda)}{\partial\Lambda}\bigg|_{\Lambda=\Lambda^{(n)}} \tag{4.15}$$

其中n為迭代子(Iterator)，ε為調整量(Step Size)，用來決定這次迭代應該要調整多少，有關柔性邊際估測法的進一步延伸，可參考[Jinyu Li *et al.* 2007]。

4.2 以熵值為基礎(Entropy-based)之資料選取

在4.1.1小節中，最大邊際估測法(LME)是以相似度(Likelihood)為基礎的分離邊際(Separation Margin)來選取離決定邊界(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)，因為離決定邊界較遠，所以較不具鑑別力，因此就沒有拿來調整聲學模型的參數。所以最大邊際估測法我們可以視為是以相似度為選取準則的資料選取方法，選出比較重要的語音特徵向量序列(訓練語句)。在4.1.2小節中，也是以相似度為選取準則，定義不同的門檻，進而選取出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)的方式選取出有重要性的時間音框(Frame)。所以柔性邊際估測法(SME)我們也可以視為是以相似度和類別比對為基礎的進階資料選取方法。

在資料選取的方法中，資料(或樣本)可以定義在不同的單位上，以語音辨識為例，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 Arcs)、音素段落(Phone Arcs)或時間音框(Frames)等。

最大邊際估測與柔性邊際估測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。本論文提出以熵值(Entropy)為基礎的時間音框資料選取(Data Selection)方法來改進所有的鑑別式聲學模型訓練。在本論文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)來求得熵值，再經由事先所設定的門檻值來選取資料，所以可以視為在事後機率定義域中來取選資料。其熵值的計算是在事後機率定義域(Posterior Domain)中，有別於傳統的相似度定義域。傳統熵值的值域為0到$\log_2 N$，其中N為參與熵值計算的樣本個數，為了方便決定門檻值而來選取時間音框，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：

$$E_z(t) = \frac{1}{\log_2 N} \sum_{q=1}^{Q} \sum_{m \in q} \gamma_{qm}^{z}(t) \cdot \log_2 \frac{1}{\gamma_{qm}^{z}(t)} \tag{4.16}$$

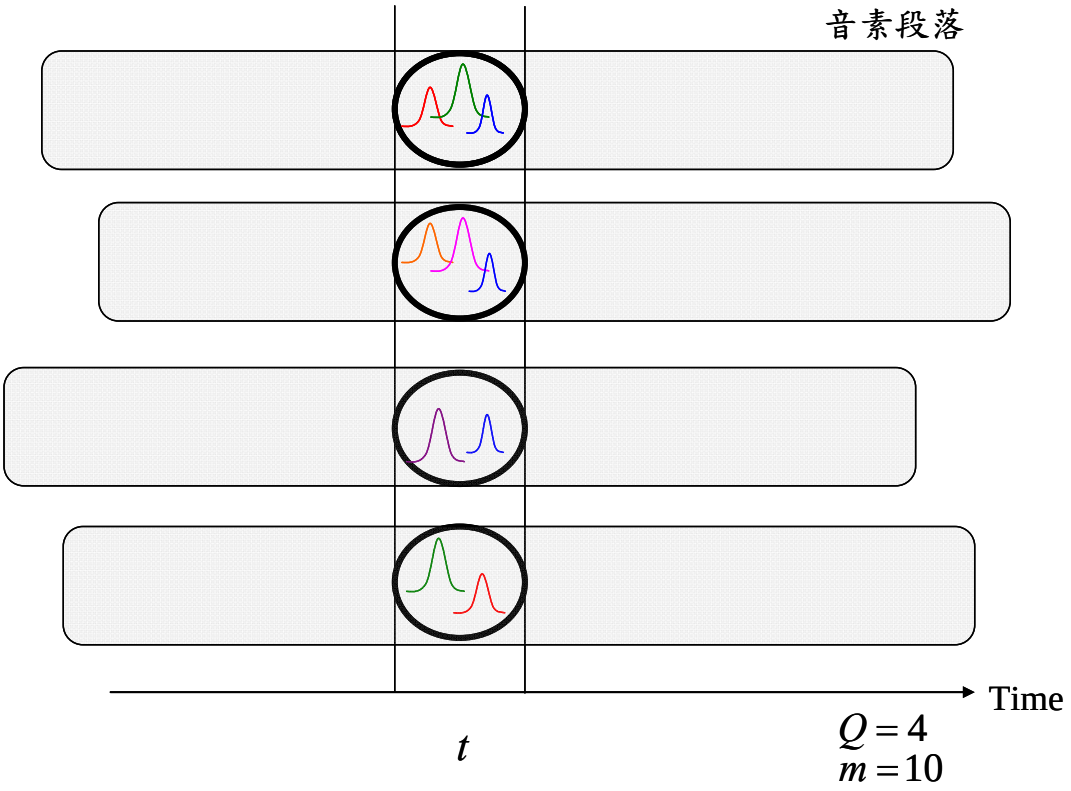


圖 4-4 詞圖中之音素段落及高斯模型在時間 t 時之示意圖

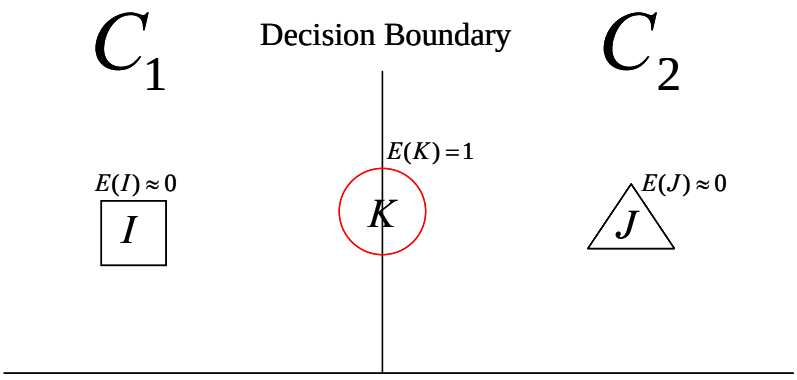


圖 4-5 正規化熵值圖例

其中 $E_z(t)$ 為在第 z 句訓練語句時間 t 時的正規化熵值，$\gamma_{qm}^{z}(t)$ 為在第 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 為在時間 t 時所有的音素段落個數，N 為在時間 t 中所有事後機率不為零的高斯模型 m。其範例如圖4-4所示。

為了理解正規化熵值在此所代表的意義，我們舉一個例子來說明。假設現在有兩個類別和三個樣本，其所對應的分類器分別記作C_1與C_2，其樣本記作I、J和K。現在我們要做分類的動作，即判斷某樣本(假設為K)是屬於C_1或C_2，於是我們會將樣本K分別帶到分類器C_1與C_2去算一個分數，若分類器C_1算出來的分數比較大，則我們會認為樣本K是屬於C_1這個類別，反之則屬於C_2這個類別。對於分類器C_1與C_2對樣本K所產生的分數，我們可以算一個正規化熵值，現在考慮兩種極端的情形，若正規化熵值等於1，則代表分類器C_1與C_2算出來的分數一樣，那麼此樣本K就不能被判別屬於C_1或C_2，所以我們會認為樣本K是非常混淆的(Confused);若正規化熵值等於0，則代表某分類器(C_1或C_2)算出來的分數會非常大，我們可以非常確定樣本K是屬於哪一個類別。如圖4-5所示，因為可以非常確定樣本I和J的類別，所以樣本I和J的正規化熵值會很接近0；樣本K剛好在兩個分類器C_1與C_2的決定邊界上，不能判斷其類別，所以樣本K的正規化熵值等於1。

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)為最小單位，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為一個訓練樣本(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，所以本論文提出的以熵值(Entropy)為基礎的時間音框選取方法(Frame Selection)就是要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，進而丟棄很正確分類和很錯誤分類之時間音框所收集到的統計值，且利用未被正確分類之時間音框所收集的統計值來調整模型參數，以幫助鑑別式聲學模型訓練。

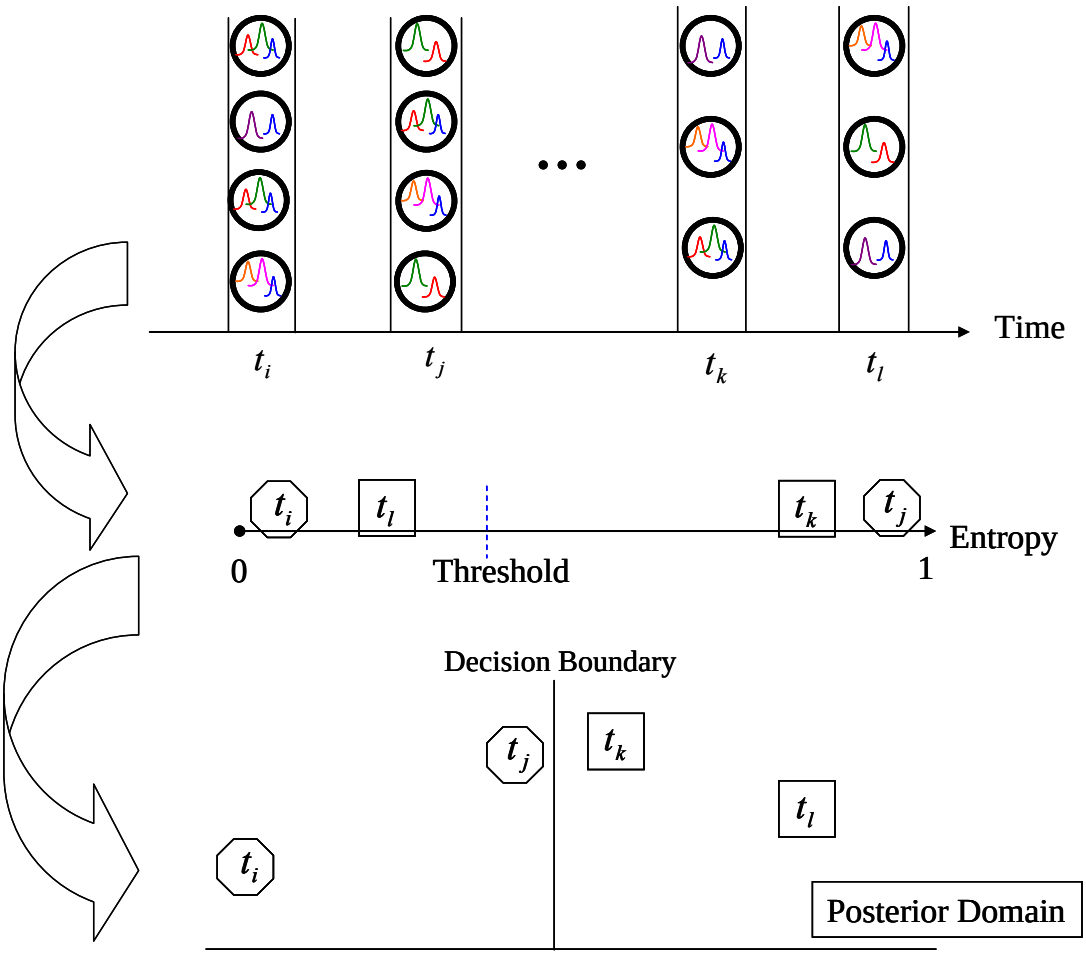


圖 4-6 正規化熵值圖例(在語音辨識應用情境中)

下面舉一個在語音辨識上的例子來說明，如圖4-6所示，以時間音框當做資料選取的單位，考慮某個語音特徵向量序列中時間音框t_i、t_j、t_k和t_l(t_i和t_j屬於某個音素類別，以八邊形表示；t_k和t_l屬於另一個音素類別，以正方形表示)，我們分別計算每一個時間音框的正規化熵值，若正規化熵值近似於1，如t_j或t_k，則代表此時間音框很混淆(Confused)，因為此時間音框落在分類器(連續密度隱藏式馬可夫模型)的事後機率都差不多，所以不容易辨識其所屬的類別。在事後機率定義域中，此時間音框會離決定邊界比較近，我們直覺地認為此時間音框對鑑別式訓練會有重要的影響力，所以此音框要被選取出來當做訓練樣本(Training Sample)；反之，若正規化熵值近似於0，如t_i或t_l，則代表此時間音框可以很容

易地被分類器完全辨識正確或完全辨識錯誤。在事後機率定義域中，此時間音框離決定邊界比較遠，對鑑別式訓練沒有什麼幫助，因此就會被捨棄。

目前的聲學模型訓練都是在最大化相似度估測(MLE)之後才開始應用鑑別式訓練，語音特徵向量序列(訓練語句)在最大化相似度估測時，大部分的語音特徵向量或時間音框就會遠離分類器(連續密度隱藏式馬可夫模型)所形成的決定邊界。少部分的語音特徵向量或時間音框還留在決定邊界附近，若能只將那些少部份的語音特徵向量或時間音框拉離決定邊界，事實上就隱含著分類器(連續密度隱藏式馬可夫模型)具有一般化(Generalization)的能力。但傳統的鑑別式聲學模型訓練將所有的訓練樣本(不論在決定邊界附近或遠離決定邊界)都考慮進來一同調整模型的參數。本論文旨在改進此缺點，因而利用正規化熵值(Normalized Entropy)來只選取決定邊界附近的時間音框，再來做鑑別式訓練，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)統計值收集為例，每個時間音框要先算正規化熵值，再由門檻值決定是否累加統計值，則其數學式可表示為:

$$\gamma_{qm}^{num} = \sum_{z=1}^{Z} \sum_{q \in \mathbf{W}_{z,lattice}} \sum_{t=s_q}^{e_q} \left[\gamma_{qm}^{z}(t) \max(0, \gamma_q^{z^{MPE}}) \right] \cdot I(E_z(t) > \rho)$$

$$\theta_{qmd}^{num}(O) = \sum_{z=1}^{Z} \sum_{q \in \mathbf{W}_{z,lattice}} \sum_{t=s_q}^{e_q} \left[\gamma_{qm}^{z}(t) \max(0, \gamma_q^{z^{MPE}}) o_z(t) \right] \cdot I(E_z(t) > \rho) \tag{4.17}$$

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

$$\gamma_{qmd}^{den} = \sum_{z=1}^{Z} \sum_{q \in \mathbf{W}_{z,lattice}} \sum_{t=s_q}^{e_q} \left[\gamma_{qm}^{z}(t) \max(0, -\gamma_q^{z^{MPE}}) \right] \cdot I(E_z(t) > \rho)$$

$$\theta_{qmd}^{den}(O) = \sum_{z=1}^{Z} \sum_{q \in \mathbf{W}_{z,lattice}} \sum_{t=s_q}^{e_q} \left[\gamma_{qm}^{z}(t) \max(0, -\gamma_q^{z^{MPE}}) o_z(t) \right] \cdot I(E_z(t) > \rho) \tag{4.18}$$

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

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

$$I(E_z(t) > \rho) = \begin{cases} 1, \ if \ E_z(t) > \rho \\ 0, \ if \ E_z(t) \le \rho \end{cases} \tag{4.19}$$

式(4.19)使用的是指示函數，其值不是 0 就是 1，我們可將它視為是一種硬性選取(Hard Selection)的資料選取方法。另一方面，在實作上，我們可以將每個時間音框所計算出來的正規化熵值當作是權重(Weight)，用來強調(Emphasized)或非強調(Deemphasized)此時間音框的重要性，我們可以將此方法視為是一種柔性選取(Soft Selection)的資料選取方法，其數學式如下所示:

$$\gamma_{qm}^{z}(t) = \gamma_{qm}^{z}(t)\left(1 + \omega \cdot E_z(t)\right) \tag{4.20}$$

其中 ω 為一比例控制參數。

第5章　非監督式模型訓練

5.1 非監督式最大化相似度聲學模型訓練

在早期的語音辨識應用情境中，需要一些語料來訓練聲學模型，而這些訓練語料通常都是經由大量的人工轉譯成文字，即正確轉譯文字(True Transcription)，並且要用聽的來找出正確的詞和音素的邊界。在找出詞和音素的邊界時需要有人為全程地介入，這種情形我們稱之為監督式(Supervised)模型訓練。收集語料在早期是一項困難的工作，因為需要利用人工特別去錄製。隨著網路的發達，多媒體的資料透過網路可以很容易地取得，如:電視新聞、無線電廣播等，使得收集語料不再是困難的工作。其所收集的語料，往往都沒有正確的轉譯文字(如無線電廣播)，不過有些語料或許有近似正確的文字，例如字幕(Closed-caption)。現場直播(Live)的電視新聞語料會有工作人員用聽的，即時地(Real-time)打出字幕，這些字幕可能會有打字錯誤的發生，而且這些字幕也沒有正確的詞或音素的邊界。所以若是有利用字幕或近似正確的文字來訓練模型，通常稱之為輕微監督式(Lightly Supervised)模型訓練[Lamel *et al.* 2002]。若是只有語料而沒有正確的轉譯文字及詞和音素邊界，通常稱之為非監督式(Unsupervised)模型訓練[Wessel *et al.* 2001b; Chen *et al.* 2004]。

在大詞彙連續語音辨識(LVCSR)的架構中，訓練語料量的多寡，對模型訓練來說是一個重要的因子。直覺地，我們會希望訓練語料的量越多，對聲學模型的訓練會越有幫助，因為可以看到更多以前所沒有看過的語音特徵。但在語料隨手可得的今天，我們卻沒有辦法很容易地提升自動語音辨識器的效能，因為通常我們所收集到的大量語料是不具有正確轉譯文字，且字幕的取得也不是非常容易。想要有正確的轉譯文字及詞和音素的邊界，是需要大量人工去標註，這是非常耗時耗力的工作。另一方面，我們若想將自動語音辨識器應用到不同的語言或不同的領域當中，可是又不想花費大量的人力去轉譯正確的文字，這時便可以利用現

有的自動語音辨識器去辨識大量未轉譯的語料，省去大量人工轉譯的力氣，以達成非監督式模型訓練。所以在大詞彙連續語音辨識系統中，非監督式模型訓練變成一項重要的研究。

通常非監督式聲學模型訓練是使用最大化相似度估測法(MLE)來達成模型參數的最佳化。其作法是利用現有人工轉譯的語料訓練出一個初始的自動語音辨識器，再去對大量未經人工轉譯的語料做一次辨識，然後利用辨識後所產生之第一名(Top 1)的辨識結果當成正確轉譯文字。接著利用含第一名辨識結果的大量語料和現有人工轉譯的語料去訓練初始的自動語音辨識器中的聲學模型。

5.1.1 信心度評估於非監督式最大化相似度聲學模型訓練

非監督式聲學模型訓練的作法是先將大量未轉譯的語料先作辨識，再重新訓練初始的聲學模型。但辨識時總是會有辨識錯誤的產生，若是拿錯誤的轉譯文字去訓練模型，反而會使模型不強健，可能會降低辨識效能。此時，信心度評估或許是個解決辦法，利用信心度評估來判斷哪些未經人工轉譯的訓練語句可能是辨識錯誤的，哪些可能是辨識正確的；更精細的說，我們可以利用信心度評估來判斷一個經辨識後的訓練語句中的哪一個詞可能是辨識錯誤的，哪一個詞可能是辨識正確的。經由信心度評估來選擇辨識後的訓練語句，很自然地會過濾掉許多有可能辨識錯誤的詞或詞序列，在過去文獻的實驗報告中，說明信心度評估確實能幫助非監督式聲學模型訓練[Wessel *et al.* 2005]。

信心度評估在 1.2.5 小節中有做了簡略的介紹，本論文所使用的信心度評估是以事後機率法中的圖形化基礎(Graph-based)法來求得詞圖中每一個詞段(Word Arc)的信心度值。在實作上，先求得每一個詞段的信心度，接著在詞圖中經由維特比(Viterbi)解碼可以得到第一名(Top 1)詞序列，此第一名詞序列中的每個詞都含有信心度，再利用事先定好的門檻值(Threshold)，來決定此第一名詞序列中的某個詞要不要拿來作聲學模型的訓練。如圖 5-1 所示，簡略說明含有信心度評估的非監督式聲學模型訓練之流程圖。

Step1. 利用現有(少量或大量)含人工轉譯文字的語料訓練出一個初始的聲學模型。

Step2. 用初始聲學模型所構成的語音辨識器去辨識大量未轉譯的語料，產生出詞圖(Lattice)。

Step3. 在詞圖中使用前向-後向演算法(FBA)求得每一個詞段的信心度，利用維特比解碼產生出含信心度的第一名詞序列(Top 1 Word Sequence)。

Step4. 利用含信心度的第一名詞序列和現有人工轉譯文字來訓練初始的聲學模型(使用最大化相似度估測法(MLE))。

圖 5-1 非監督式聲學模型訓練步驟

5.1.2 迭代方法(Iterative Approach)

為了使人工監督的力氣花得越少，非監督式的模型訓練應該要使用迭代方法(Iterative Approach)來實現[Wessel *et al.* 2005]。其迭代的作法是將重複非監督式聲學模型訊練，即利用訓練好的聲學模型再對未轉譯的語料作一次辨識，將所產生的第一名辨識結果及現有人工轉譯文字拿去重新訓練之前已經訓練好的聲學模型。在具有大量人工轉譯的語料所訓練出來的初使聲學模型，其辨識效能以達不錯的效果，所以迭代方法在具有大量人工轉譯的語料上會有比較不明顯的效果。若在只有少量的人工轉譯語料上，會有比較明顯的效果，因為只有少量人工轉譯的語料所訓練出來的初始模型，對大量未人工轉譯的語料之辨識率會比較差，經由迭代方法的模型訓練，可以使模型越來越好，辨識率相對提升許多，所以會有比較明顯的效果。

信心度評估在迭代方法中，應該要隨著迭代次數的增加，而將信心度門檻值降低，因為一開始的聲學模型可能會造成比較多的辨識錯誤，所以信心度門檻值要設高一點，過濾掉比較多的辨識錯誤文字。當做了幾次迭代方法後，所訓練的聲學模型會比較強健，對大量未人工轉譯的語料會有比較少的辨識錯誤文字，所

Step1. 利用現有(少量)含人工轉譯文字的語料訓練出一個初始的聲學模型。

Step2. 用目前訓練好的聲學模型所構成的語音辨識器去辨識大量未轉譯的語料，產生出詞圖(Lattice)。

Step3. 在詞圖中使用前向-後向演算法(FBA)求得每一個詞段的信心度，利用維特比解碼產生出含信心度的第一名詞序列(Top1 Word Sequence)。

Step4. 利用含信心度的第一名詞序列和現有人工轉譯文字來重新訓練目前的聲學模型(使用最大化相似度估測法(MLE))。

重複 Step2-Step4 並且降低信心度門檻值，直到收斂或達到可以滿足的效能。

圖 5-2 迭代方法之非監督式聲學模型訓練流程圖

以信心度門檻值就可以設低一點，以考慮更多可能辨識正確的訓練語料。圖 5-2 簡略介紹迭代方法的流程。

5.2 非監督鑑別式聲學模型訓練

目前鑑別式訓練在監督式(Supervised)聲學模型的訓練上已有不錯的成效，所以我們期望利用現階段的鑑別式訓練在非監督式(Unsupervised)聲學模型訓練上也能達到不錯的效果。近三、四年來，有學者嘗試使用鑑別式訓練在非監督式聲學模型訓練上，確實可以達到一定的效果[Mathias *et al.* 2005; Ma *et al.* 2006; Wang *et al.* 2007]。但非監督鑑別式聲學模型訓練與非監督最大化相似度聲學模型都會遇到同樣的難題，就是會有遇到辨識錯誤文字，以致效能會打折扣。如何產生出精確的轉譯文字及詞和音素邊界是很重要的一個議題，如 5.1.1 小節所描述的信心度評估，是一個可行的解決方式。事實上，利用信心度評估來選取可能正確的辨識詞或詞序列，可以視為是一種資料選取的方法。如第四章所述，訓練資料樣本可以定義在不同的單位上，若現在以詞(Word)當成資料樣本單位，則信心度評估可以視為一種選取的工具(如同本論文所提出的正規化熵值)，用來選取可能正

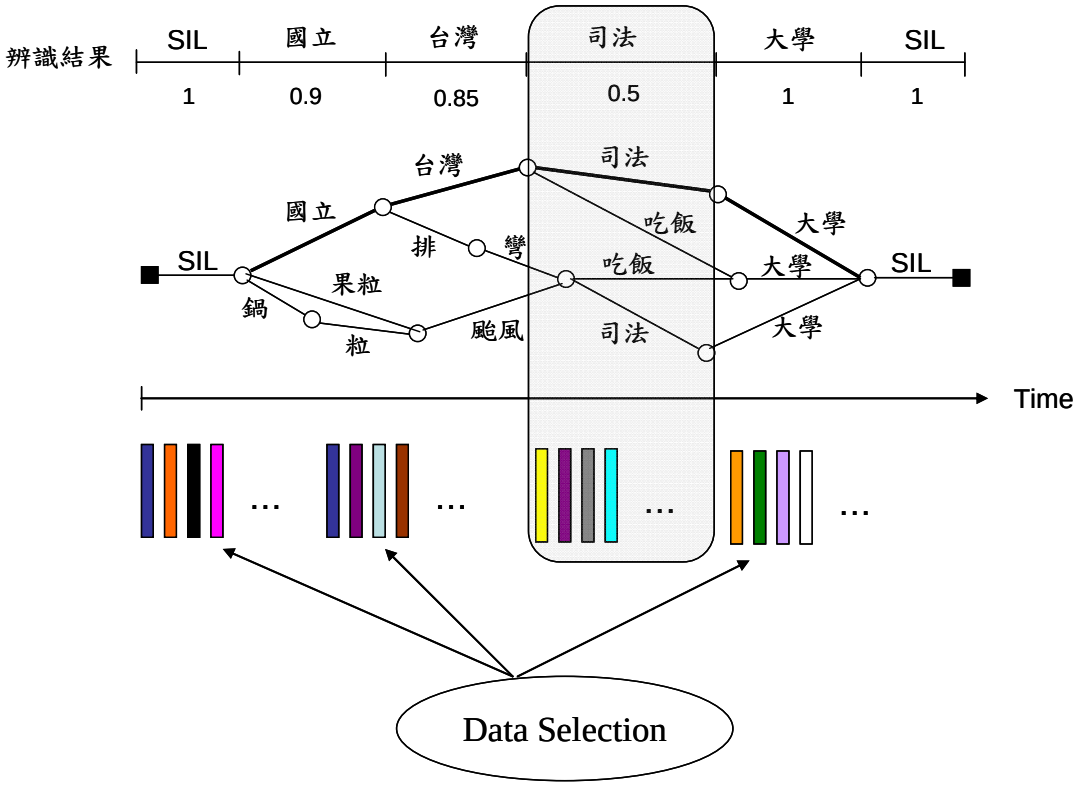


圖 5-3 資料選取於非監督鑑別式聲學模型訓練之示意圖

確的詞，以利非監督式最大化相似度和鑑別式聲學模型訓練。

因過去的非監督鑑別式聲學模型訓練只有使用信心度評估來做訓練樣本(詞)的選取，本論文嘗試使用 4.2 小節所提出的正規化熵值進一步地做訓練樣本(時間音框)的選取，如圖 5-3 所示，辨識結果(詞序列"SIL 國立台灣司法大學 SIL")下方的數字代表每一個詞的信心度值(Confidence Score)，假設信心度門檻值為 0.8，那麼"司法"這個詞就會被過濾掉，剩下的詞所對應的時間音框(語音特徵向量序列)再由正規化熵值來選取。非監督鑑別式聲學模型訓練在少量的人工轉譯語料上應該也要使用迭代方法來實現，圖 5-4 簡略介紹迭代方法於非監督鑑別式聲學模型訓練。

Step1. 利用現有(少量)含人工轉譯文字的語料訓練出一個初始的聲學模型。

Step2. 用目前訓練好的聲學模型所構成的語音辨識器去辨識大量未轉譯的語料，產生出詞圖(Lattice)。

Step3. 在詞圖中使用前向-後向演算法(FBA)求得每一個詞段的信心度，利用維特比解碼產生出含信心度的第一名詞序列(Top 1 Word Sequence)。

Step4. 利用含信心度的第一名詞序列和現有人工轉譯文字來重新訓練目前的聲學模型(使用最大化相似度估測法(MLE))。

重複 Step2-Step4 並且降低信心度門檻值，直到收斂或達到可以滿足的效能。

Step5. 利用目前最好的聲學模型所構成的語音辨識器去辨識大量未轉譯的語料，產生出詞圖。

Step6. 與 Step3 相同。

Step7. 使用目前流行的鑑別式訓練(如 MPE)、信心度評估(詞的選取)及正規化熵值(時間音框的選取)來做聲學模型參數的調整。

重複 Step5-Step7 並且降低信心度門檻值，直到收斂。

圖 5-4 迭代方法於非監督鑑別式聲學模型訓練步驟

第6章　實驗架構與實驗結果

在 6.1 小節吾人將先介紹臺灣師大的中文大詞彙連續語音辨識系統[Chen *et al.* 2004; 2005]，接著在 6.2 小節介紹及分析本論文所使用的公視晚間新聞(MATBN)外場記者語料[Wang *et al.* 2005]和實驗的評估方式。在 6.3、6.4、6.5 及 6.6 小節說明基礎實驗結果、改進以最小化音素錯誤鑑別式聲學模型訓練之結果、資料選取方法於鑑別式訓練之實驗結果、非監督式訓練的實驗結果和實驗分析。

6.1 臺灣師大之中文大詞彙連續語音辨識系統

以下將分別介紹臺灣師大的中文大詞彙連續語音辨識系統統採用的前端處理(Front-end Processing)、聲學模型(Acoustic Model)、詞典建立(Lexicon Construction)、語言模型(Language Model)以及詞彙樹複製搜尋(Tree-copy Search)等部份。

6.1.1 前端處理

在本論文中使用梅爾倒頻譜係數特徵(MFCC)作為語音訊號的特徵參數。在求取梅爾倒頻譜係數特徵時，將語音資料切割成一連串部分重疊的音框，每一個音框由13維的梅爾倒頻譜係數特徵加上其一階與二階的時間軸導數(Time Derivatives)所形成的39維語音特徵向量所組成。其中13維的梅爾倒頻譜係數特徵是由18個梅爾頻譜上濾波器組(Filter Banks)的輸出經餘弦轉換求得。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)。

另外本論文還使用鑑別性特徵，所採用的是目前流行的線性鑑別分析(Linear Discriminant Analysis, LDA)[Duda *et al.* 1973]和異質性線性鑑別分析(Heteroscedastic Linear Discriminant Analysis, HLDA)[Kumar 1997]，異質性線性鑑別分析(HLDA)是線性鑑別分析的一般化與去除多餘限制後的另一種線性特徵轉換作法。線性鑑別分析假設各類別分布的變異量皆相同，因此在所有類別分布的

變異量為相同的情況下有最好的解。可是現實上大多的訊號特徵分布的變異量皆為異質性，線性鑑別分析求出的基底矩陣就可能不是最佳的解，而異質性線性鑑別分析就是拿掉這個假設，以求得較具鑑別性的語音特徵向量。本論文在做完異質性線性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)[Gopinath 1998]，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)之共變異矩陣。

6.1.2 聲學模型

聲學模型是採用傳統的連續密度隱藏式馬可夫模型(CDHMM)，模型內狀態的轉移情形只有兩種，一種是停留在原狀態，一種是由左至右跳到下一個相鄰的狀態。模型的總數量有151個，其中包含了1個靜音模型(Silence)，112個聲母模型(INITIAL)，以及38個韻母模型(FINAL)。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。此外，聲母和韻母共有403種不同的音節組合。

6.1.3 詞典建立與語言模型訓練

在中文裡約有 7,000 個單字詞，新詞可由此 7,000 個單字詞合併產生，則可根據字詞在語料中的統計特性，以自動化的方式產生新的複合詞(Compound Words)。新增複合詞的自動產生方式如下面所述：對於語料中任意相鄰的兩個詞 (w_i, w_j)，可以分別計算它們的前二連(Forward Bigram)機率 $P_f(w_j \mid w_i)$，與後二連(Backward Bigram)機率 $P_f(w_i \mid w_j)$，並以前後二連(Forward and Backward Bigrams) 的機率幾何平均(Geometric Average) $FB(w_i, w_j) = \sqrt{P_f(w_j \mid w_i) P_b(w_i \mid w_j)}$，作為 (w_i, w_j) 是否合併的依據。文字語料先經由一個含有一至四字詞約六萬八千個詞的詞典來斷詞，然後利用上述的公式，經數次的迭代以及不同的基準閥值

(Threshold)設定，產生約五千個二至十字詞的複合詞，使得最後的語音辨識詞典約含有七萬二千個一至十字詞。本系統使用詞二連以及詞三連語言模型(Word Bigram and Trigram Language Models)，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個中文字語料作為背景語言模型訓練時的訓練資料[LDC]。本論文中的語言模型使用 Katz 語言模型平滑技術，語言模型訓練工具採用 SRI Language Modeling Toolkit (SRILM) [SRILM 2002]。

6.1.4 詞彙樹複製搜尋

本系統是採用由左至右(Left-to-right)且音框同步(Frame Synchronous)的詞彙樹複製搜尋方式[Aubert 2002]。詞彙樹的架構如所示，樹中的每個分枝(Arc)代表一個聲母(INITIAL)、韻母(FINAL)或靜音(Silence)模型。由樹的根節點(Root Node，圖6-1的圓形實心點)走到樹的葉節點(Leaf Node，圖6-1的方形實心點)的某一條完整路徑代表走完一個或一組發音相同的詞。而路徑上的每一個分枝正好對應到這些詞的一組聲學模型。詞彙樹複製搜尋在執行時，每個音框會同時存在數棵詞彙樹複製(Tree Copies)，而每棵詞彙樹代表來自不同的語言歷史或限制(Language Model History or Constraint)。在同一棵詞彙樹裡，會進行隱藏式馬可夫模型狀態層次(State Level)維特比(Viterbi)動態規劃搜尋。在詞彙樹搜尋中，只有在走到葉節點時，才能確定所搜尋的一個完整詞為何。另外，當具有相同語言模型歷史之不同詞彙樹分別都已經走到自己所屬那棵樹的葉節點時，則會進行結合(Recombination)，只保留其中分數最大者，並針對留下來的詞彙樹繼續執行詞彙樹複製搜尋。然而，真正在實作時，並不需要產生如此多的詞彙樹，僅需建立一棵詞彙樹作為參考之用，並分別記錄搜尋時存活下來之隱藏式馬可夫模型狀態節點的相關資訊(如到目前為此所累積的分數及前一狀態為何)。另外一方面，由於存活的狀態節點通常會隨著音框數呈指數倍成長，因而必須以光束剪裁(Beam Pruning)技術將分數較低的狀態節點做剪裁的動作。在對每個狀態節點執行光束剪裁時，會依此節點所有可拜訪的葉節點之最大單連語言模型往前觀測分數

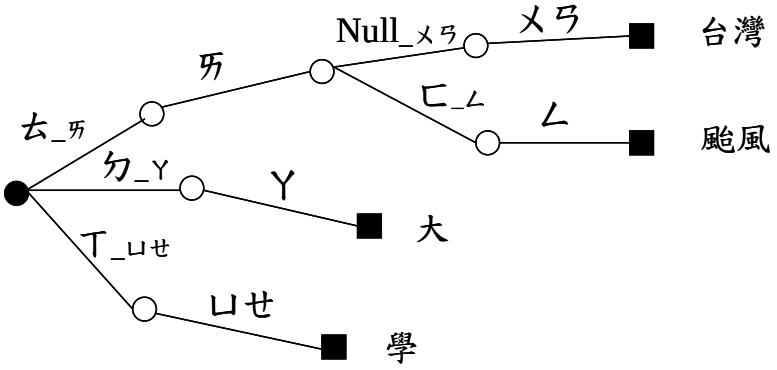


圖 6-1 詞彙樹範例

(Unigram Language Model Look-ahead Score)[Aubert 2002]及聲學往前觀測分數(Acoustic Look-ahead Score)[Chen *et al*. 2004; 2005]做為剪裁與否的依據。此外，在每個音框，利用存活的詞彙樹複製樹其葉節點(代表可能的候選詞)所儲存的語言模型歷史、開始音框、結束音框及其聲學解碼的分數等資訊，建立如第二章所提到的詞圖。而後使用更高階的語言模型，如詞三連(Trigram)或詞四連(Fourgram)語言模型，抑或採用更複雜的聲學模型，如三連音素(Triphone)，進行詞圖重評分(Word Graph Rescoring)搜尋[Ortmanns *et al*. 1997]，找出最佳的詞序列。在本論文中，詞彙樹複製搜尋階段是採用詞二連語言模型，詞圖搜尋階段則是使用詞三連語言模型。

6.2 實驗語料與評估方式

6.2.1 實驗語料之說明

本實驗主要使用的語料庫為 MATBN 新聞語料，為中央研究院資訊所口語小組(SLG)[SLG]耗時三年與公共電視台(PTS)[PTS]合作錄製完成。錄製的對象為公視晚間新聞，其每天的長度皆為一個小時，收錄了 200 天(約 200 小時)的新聞語料，其中包含 2001 年的新聞 30 小時、2002 年 146 小時及 2003 年 24 小時。所有的

表 6-1 主播語料分布表

語者姓名	性別	句數(句)	語音總長度(秒)	所含語音百分比(%)
余佳璋-主播	男	36	452.20	0.50
林建成-主播	男	427	5,298.10	5.70
某主播一_PTSND20020226	女	1	7.90	0.008
洪蕙竹-主播	女	89	1,407.40	1.50
洪蕙竹-氣象主播	女	155	1,443.60	1.50
徐惠玲-主播	女	225	3,208.20	3.40
馬紹-主播	男	35	465.60	0.50
黃明明-主播	女	175	2,932.60	3.10
葉明蘭-主播	女	5,101	78,584.70	83.60
蘇怡如-氣象主播	女	17	213.80	0.20

新聞語料都有正確的人工轉寫以及其它的標註資訊(如：音樂、背景雜訊、停頓、語助詞、呼吸、強調語氣、反覆、不適當的發音等)，所有的人工轉寫與標注均使用 DGA&LDC 的轉寫器(Transcriber)[Barras *et al.* 2001]來完成。

每天的新聞約含有二十多則報導，每則報導為一完整主題。除了語音資料，文字語料在其它應用上也有很大的價值(如資訊檢索、文件摘要等)。此新聞語料大致上可分內場及外場兩個部份，內場部分主要為主播(Studio Anchors)的語料，外場部分則可分為採訪記者(Field Reporters)與受訪者(Interviewees)的語料。在篩選實驗語料時，考量新聞的特性，主播多為同一人所擔任。如表 6-1 所示，葉明蘭主播的語料在本語料庫中約佔了所有主播語料的 85%，這將使得實驗偏向語者相依(Speaker-dependent)的環境，加上女性主播約佔了所有主播語料的 94%，也造成了性別相依(Gender-dependent)的問題。如果使用主播語料的話，缺少足夠的變異來提供良好的訓練與客觀的評估，故本實驗不採用主播語料。此系統可檢索語句的統計資訊，如語者資訊、語音長度、所含背景雜訊、說話速度及正確轉譯文句等資訊，適合用來分析且定義出實驗的訓練集(Training Set)與評估集(Evaluation Set)。目前本研究初步地只選擇外場採訪記者部份作為實驗語料，在

表 6-2　外場記者訓練與測試語料分布表

性別	訓練語料總長(分)	評估語料總長(分)
男生	766.69	21.68
女生	766.79	65.23

表 6-3 語助詞出現次數統計表

語者型別	所含語音百分比(%)	語助詞出現次數(句)	每句平均語助詞出現次數(次)
外場採訪記者	48.69	877	0.07
外場受訪者	29.33	18,991	2.03
內場主播	21.98	771	0.12

表 6-4 外場受訪者訓練與評估語料分布表

性別	訓練語料總長(分)	評估語料總長(分)
男生	269.03	25.91
女生	259.22	10.53

未來將會納入受訪者的部份。

本研究所使用外場記者的語料總共約 26 小時，其中 24.5 小時(5774 句，再切成 34,672 個短句供聲學模型訓練之用)做為聲學模型訓練的語料，1.5 小時(292 句)則為辨識評估的資料。訓練語料由 2001 及 2002 年的新聞中篩選出不含語助詞(Particle)的語料，為了建立性別平衡(Gender-balanced)的訓練環境，男、女語料分別篩取 12.2 小時。為了準確客觀地評估研究方法，我們採用由中研院從 2003 年語料庫中所選定的外場記者語料作為評估語料，部份語音片段含有語助詞，關於訓練語料及測試語料詳細的資訊可見表 6-2。

而在外場受訪者部份，由於有較多的語助詞出現，如表6-3所示，因此如果直接只採用不含語助詞的外場語料的話，測試資料(只從2003年的語料擷取)將會非常的稀少。而訓練資料(從2001及2002年的語料擷取)為了要顧及性別平衡的因素，也會有訓練資料量不足的問題，所以本論文就沒有使用外場受訪者的語料。另外所有語料的詳細統計資訊可經由師大資工語音實驗室[NTNU 2004]所開發的公視新聞語料檢索系統獲得。此系統可檢索語句的統計資訊，如語者資訊、語音長度、所含背景雜訊、說話速度及正確轉寫文等內容，極為適合用來篩選聲學模型訓練所需的語料。

6.2.2 實驗評估方式

此評估法則是採用美國國家標準與技術中心(National Institute of Standards and Technology, NIST)[NIST]所訂立的評估標準來進行正確答案的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(也就是第二章所提到的編輯距離(Edit or Levenshtein Distance))。由於在中文會有斷詞不一致的問題，因此在本論文的實驗中主要是以字為比對單位。令 H 為正確答案詞序列與辨識詞序列比對後相同(Match)的字的個數、I 為辨識詞序列多餘插入(Insertion)的字的個數、N 為正確答案詞序列的字的個數，則語音辨識系統的正確率(Accuracy)的計算方式為 $\frac{H-I}{N}\times 100\%$ ，而錯誤率(Error Rate)則為 1-正確率。在進行動態規畫比對時，替代(Substituion)錯誤的懲罰權重(Penalty Weight)為 10 分，插入及刪除的權重則皆為 7 分。因為中文是以字(Character)為單位，所以在以下的實驗數據中，都是以字錯誤率(Character Error Rate, CER)來呈現實驗結果。

6.3 基礎實驗結果

本小節主要是比較不同的語音特徵以及不同的聲學模型訓練方法對外場記者語料辨識字錯誤率的影響。所比較的語音特徵有三種，包含梅爾倒頻譜特徵加上倒頻譜正規化法(MFCC+CN)、線性鑑別分析加上最大化相似度線性轉換與倒頻譜正規化法(LDA+MLLT+CN)及異質性線性鑑別分析加上最大化相似度線性轉換與倒頻譜正規化法(HLDA+MLLT+CN)。其次，所比較不同的聲學模型訓練方法也是有三種，包含最大化相似度(ML)估測法、最大化交互資訊(MMI)估測法以及最小化音素錯誤(MPE)鑑別式訓練[Povey 2004]。本小節使用最小化音素錯誤(MPE)鑑別式訓練來比較三種不同的語音特徵，其實驗結果可參考圖6-2和表6-5，其中Baseline是利用最大化相似度(ML)估測法訓練10次所得到的字錯誤率(CER)。其次，使用異質性線性鑑別分析加上最大化相似度線性轉換與倒頻譜正規化法(HLDA+MLLT+CN)來比較三種不同的聲學模型訓練方法，其實驗結果可參考圖6-3和表6-6，Baseline是利用最大化相似度(ML)估測法訓練10次所得到的字錯誤率。由上述之比較可以得知在基礎實驗中，目前最好的語音特徵與模型訓練方法分別為為異質性線性鑑別分析加上最大化相似度線性轉換與倒頻譜正規化法(HLDA+MLLT+CN)與最小化音素錯誤(MPE)鑑別式訓練。所以接下來的實驗中，都是以最小化音素錯誤(MPE)與異質性線性鑑別分析加上最大化相似度線性轉換及倒頻譜正規化法(HLDA+MLLT+CN)訓練10次為比較對象，其字錯誤率(CER)為20.77%。

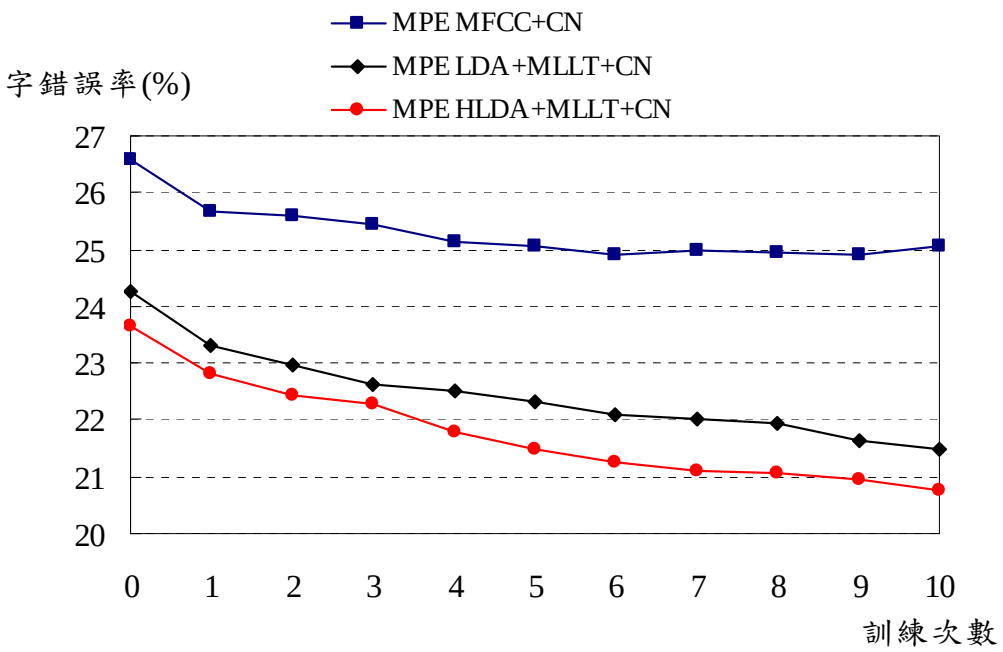


圖 6-2 比較不同的語音特徵(使用最小化音素錯誤訓練)

表 6-5 比較不同的語音特徵(使用最小化音素錯誤訓練)

CER(%)	MPE MFCC+CN	MPE LDA+MLLT+CN	MPE HLDA+MLLT+CN
Baseline	26.60	24.25	23.64
Itr01	25.68	23.32	22.82
Itr02	25.58	22.95	22.44
Itr03	25.43	22.64	22.28
Itr04	25.13	22.50	21.79
Itr05	25.06	22.31	21.48
Itr06	24.90	22.10	21.24
Itr07	24.97	22.01	21.10
Itr08	24.96	21.95	21.06
Itr09	24.89	21.65	20.97
Itr10	25.05	21.50	20.77

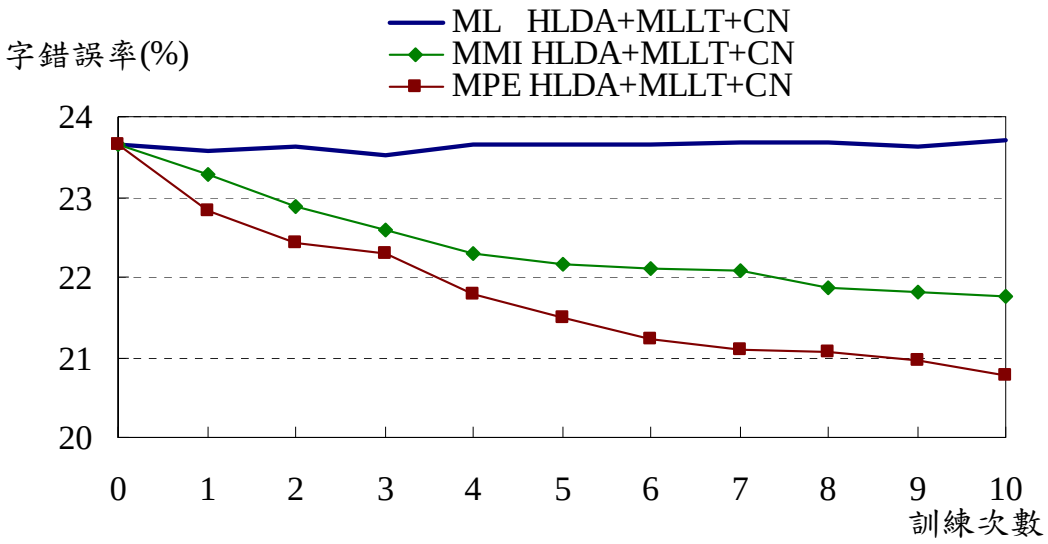


圖 6-3 比較不同的聲學模型訓練方法(使用異質性線性鑑別分析)

表 6-6 比較不同的聲學模型訓練方法(使用異質性線性鑑別分析)

CER(%)	HLDA+MLLT+CN ML	HLDA+MLLT+CN MMI	HLDA+MLLT+CN MPE
Baseline	23.64		
Itr01	23.58	23.28	22.82
Itr02	23.62	22.89	22.44
Itr03	23.52	22.58	22.28
Itr04	23.64	22.28	21.79
Itr05	23.64	22.16	21.48
Itr06	23.66	22.10	21.24
Itr07	23.68	22.08	21.10
Itr08	23.68	21.88	21.06
Itr09	23.62	21.81	20.97
Itr10	23.70	21.75	20.77

6.4 改進最小化音素錯誤之實驗結果

本小節主要是呈現第三章所描述過去學者對最小化音素錯誤(MPE)訓練的正確函數之改進以及吾人所提出之改進的訓練方法的實驗結果，分為三個子小節呈現實驗數據。6.4.1 小節呈現最小化音素錯誤正確率改進之實驗，6.4.2 小節為本論文提出之改進方法於最小化音素錯誤訓練的實驗結果，最後 6.4.3 小節則是本論文所提出之以統計式方法去近似訓練語句的事前機率於最小化音素錯誤訓練之實驗結果。

6.4.1 最小化音素錯誤訓練正確率函數改進之實驗

最小化音素錯誤鑑別式訓練之不同正確率函數有最小化音素音框錯誤(MPFE)訓練、最小化音框音素錯誤(MFPE)訓練、以狀態為基礎的最小化貝氏風險(s-MBR)以及最小化散度(MD)鑑別式聲學模型訓練。其中最小化音框音素錯誤又有一個改進方式(記作 MFPE_nosil)，就是在詞圖中收集統計值時不考慮靜音(Silence)的影響，那麼在詞圖中只要有遇到靜音詞段，則其音素音框正確率設為 0[Povey *et al.* 2007]。有關最小化音素錯誤鑑別式聲學模型訓練之不同減損函數公式可以參考表 3-1。

在使用最小化音素錯誤(MPE)訓練聲學模型時，一定要搭配使用重要的平滑技術，即 I-平滑技術(I-smoothing，詳見 2.3.4 小節之說明)，不然其效能便沒有辦法完全發揮，甚至會比最大化交互資訊(MMI)估測的效能還要差[Povey *et al.* 2007]。因為減損函數的不同，使得所收集的統計值之範圍也會不盡相同，所以 I-平滑技術的參數(記作 Tau)就需要重新調整。最小化音素錯誤(MPE)訓練的 I-平滑技術參數最佳化設定為 10[郭人瑋 2005]，其他不同減損函數的訓練也試了許多組實驗設定，在此只呈現最佳設定的實驗結果，其實驗結果可參考圖 6-4 和表 6-7。由實驗結果得知最小化音素錯誤(MPE)訓練與其他改進方法的結果差不多，只有最小化散度(MD)和最小化音素音框錯誤(MPFE)訓練的結果稍微差了一點。

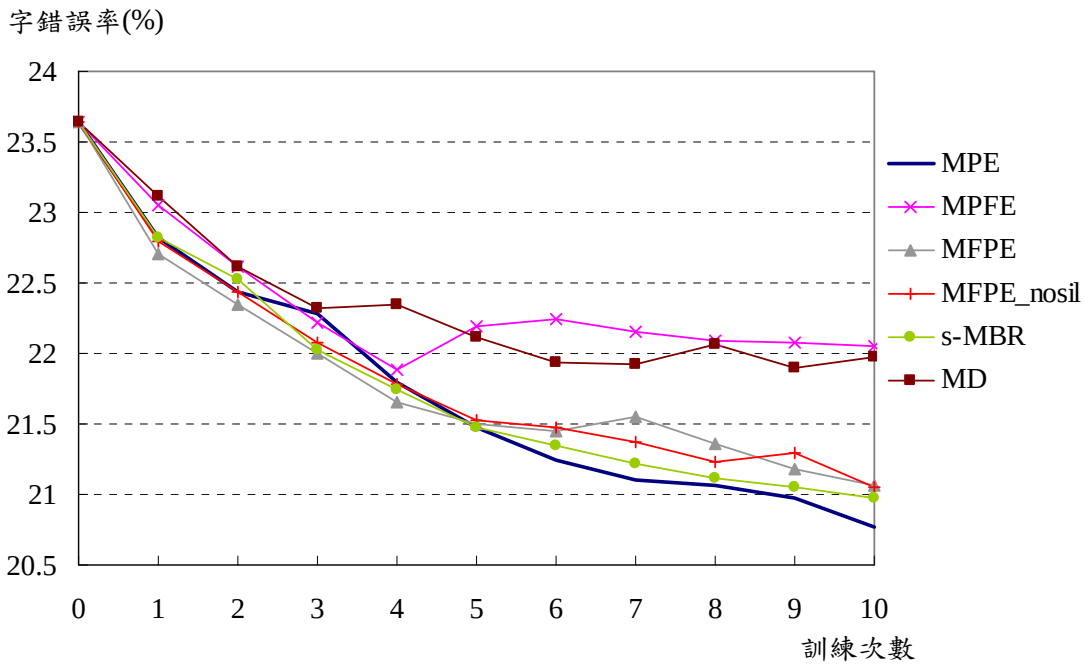


圖 6-4 最小化音素錯誤訓練正確率改進之實驗結果

表 6-7 最小化音素錯誤正確率改進之實驗結果

CER(%)	MPE Tau=10	MPFE Tau=10	MFPE Tau=100	MFPE_nosil Tau=100	s-MBR Tau=250	MD Tau=50
Baseline	23.64					
Itr01	22.82	23.05	22.70	22.80	22.82	23.11
Itr02	22.44	22.62	22.35	22.44	22.52	22.62
Itr03	22.28	22.22	22.00	22.08	22.03	22.32
Itr04	21.79	21.89	21.66	21.78	21.74	22.35
Itr05	21.48	22.19	21.50	21.52	21.48	22.12
Itr06	21.24	22.24	21.45	21.47	21.35	21.93
Itr07	21.10	22.16	21.55	21.37	21.22	21.92
Itr08	21.06	22.09	21.36	21.23	21.11	22.06
Itr09	20.97	22.08	21.18	21.29	21.05	21.90
Itr10	20.77	22.05	21.06	21.05	20.97	21.97

6.4.2 本論文提出的時間音框正確率函數之實驗

本子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。因最小化音素錯誤訓練沒有給予刪除錯誤(Deletion Error)適當的懲罰(Penalty)，而吾人所提出之時間音框正確率函數有考量到刪除錯誤的懲罰，所以在字辨識錯誤率上，確實比最小化音素錯誤來的好。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，而是要讓詞圖中某詞段因受到刪除錯誤的影響，而減少其收集的正確率統計值，以利聲學模型訓練的強健。實驗結果可參考表 6-8，其中刪除錯誤的懲罰權重(Penalty Weight)以 Lo(ρ)表示。由實驗數據顯示出在前幾次的迭代訓練中，有有考慮不同懲罰權重的刪除錯誤之時間音框正確率函數都會稍微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的效果，由表 6-8 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訓練上會有比較不明顯的效果，但都比最小化音素錯誤訓練來得佳，可是在第 10 次的訓練上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 6-5。

為了能使時間音框正確率函數能充分地懲罰刪除錯誤，且也是為了要和原始音素正確率的值域同為-1 到 1 之間，本論文使用一個常見的 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設為零($\beta=0$)。實驗結果可以參考圖 6-6，實驗數據顯示出使用 S 型函數來充分地懲罰刪除錯誤確實可以達到不錯的效果，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%，訓練次數 1 到 10 次的字錯誤率曲線圖請參考表 6-9。

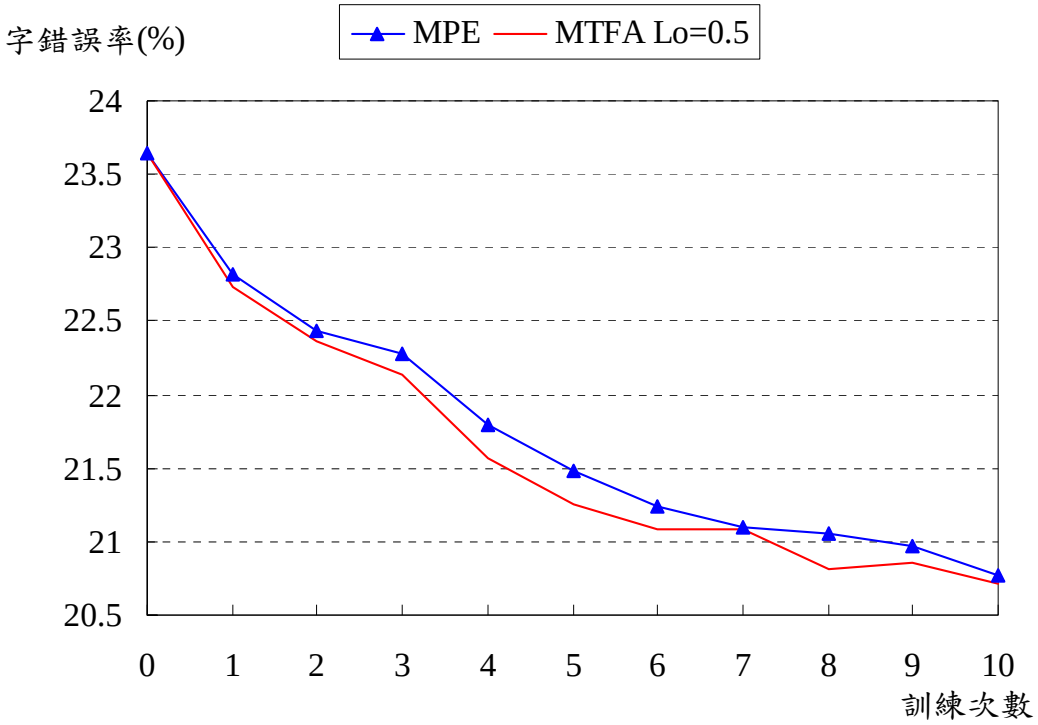


圖 6-5 最大化時間音框正確率函數(MTFA)最佳設定(ρ =0.5)

與最小化音素錯誤(MPE)之比較結果

表 6-8 最大化時間音框正確率函數(MTFA)之實驗結果

CER(%)	MTFA ρ =0.1	MTFA ρ =0.3	MTFA ρ =0.4	MTFA ρ =0.45	MTFA ρ =0.5	MTFA ρ =0.55	MTFA ρ =0.6	MTFA ρ =0.8
Baseline	23.64							
Itr01	22.85	22.73	22.71	22.73	22.74	22.73	22.75	22.80
Itr02	22.35	22.33	22.31	22.30	22.36	22.33	22.29	22.39
Itr03	22.07	22.13	22.11	22.12	22.14	22.15	22.16	22.19
Itr04	21.65	21.50	21.57	21.60	21.56	21.57	21.58	21.69
Itr05	21.26	21.14	21.25	21.21	21.26	21.28	21.24	21.34
Itr06	20.98	20.97	21.00	21.00	21.09	21.17	21.11	21.23
Itr07	20.91	20.87	20.94	20.99	21.09	21.13	21.16	21.19
Itr08	20.87	20.81	20.85	20.79	20.82	20.78	20.82	20.93
Itr09	20.84	20.74	20.81	20.78	20.85	20.86	20.92	20.90
Itr10	20.82	20.80	20.81	20.80	20.72	20.79	20.83	20.93

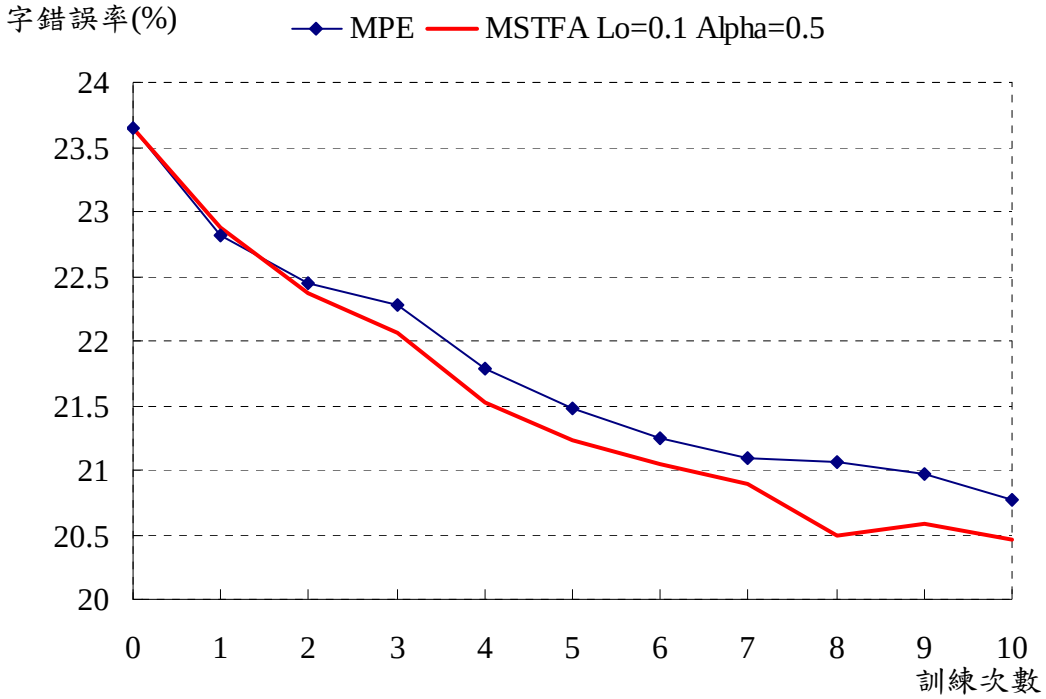


圖 6-6 最大化 S 型時間音框正確率函數(MSTFA)最佳設定(ρ =0.1, α = 0.5)與最小化音素錯誤(MPE)之比較結果

表 6-9 最大化 S 型時間音框正確率函數(MSTFA)之實驗結果

CER(%)	MSTFA ρ =0.1 α = 0.5	MSTFA ρ =0.2 α = 0.5	MSTFA ρ =0.5 α = 0.5	MSTFA ρ =0.1 α = 1	MSTFA ρ =0.2 α = 1	MSTFA ρ =0.5 α = 1
Baseline	23.64					
Itr01	22.88	22.82	22.82	22.83	22.82	22.77
Itr02	22.37	22.40	22.34	22.37	22.40	22.38
Itr03	22.06	22.10	22.10	22.02	22.09	22.05
Itr04	21.52	21.56	21.58	21.41	21.60	21.56
Itr05	21.23	21.29	21.47	21.30	21.39	21.52
Itr06	21.05	21.03	21.27	21.06	21.26	21.32
Itr07	20.89	20.90	21.11	20.80	20.91	21.19
Itr08	20.50	20.69	20.97	20.54	20.84	20.98
Itr09	20.58	20.69	20.82	20.57	20.63	21.03
Itr10	20.46	20.68	20.87	20.65	20.72	21.10

6.4.3 考慮事前機率之實驗

本子小節呈現吾人所提出之統計式方法來近似事前機率(Prior Probability)於最小化音素錯誤(MPE)訓練之實驗結果。在本實驗中，某語句之事前機率的算法是使用某詞圖中所有可能詞序列的前向分數(Forward Score)再正規化所有訓練語句的前向分數，如式(3.28)所示。為了不讓事前機率支配統計值的收集，在實作上使用刻度法(Scaling)來平滑事前機率值，其統計值的收集數學式可表示為(以正貢獻為例):

$$\gamma_{qm}^{num} = \sum_{z=1}^{Z} \left[\sum_{q \in \mathbf{W}_{z,lattice}} \sum_{t=s_q}^{e_q} \gamma_{qm}^{z}(t) \max(0, \gamma_q^{z^{MPE}}) \right] \cdot \tilde{P}(O_z)^{1/\kappa} \tag{6.1}$$

其中κ為刻度法的參數。

實驗結果可以參考表 6-10，由實驗數據顯示出考慮事前機率的影響，確實對辨識字錯誤率有效果，但其效果不是非常的明顯。在最佳設定($\kappa = 10$)的實驗中，可參考圖 6-7，前幾次的迭代訓練中都有比最小化音素錯誤(MPE)訓練好，在第 10 次訓練上只有 0.01%的進步。

表 6-10 考慮事前機率於最小化音素錯誤(MPE)訓練之實驗結果

CER(%)	Prior $\kappa = 3$	Prior $\kappa = 5$	Prior $\kappa = 8$	Prior $\kappa = 10$	Prior $\kappa = 12$	Prior $\kappa = 15$	Prior $\kappa = 20$
Baseline	23.64						
Itr01	22.56	22.80	22.78	22.80	22.74	22.79	22.79
Itr02	22.26	22.25	22.26	22.31	22.28	22.34	22.33
Itr03	21.67	21.83	21.95	21.98	22.01	22.01	22.10
Itr04	21.54	21.50	21.55	21.57	21.75	21.69	21.78
Itr05	21.44	21.32	21.27	21.36	21.43	21.41	21.53
Itr06	21.34	21.27	21.26	21.23	21.25	21.18	21.34
Itr07	21.42	21.10	21.01	21.01	21.18	21.10	21.37
Itr08	21.33	21.00	21.01	20.98	20.98	21.03	21.29
Itr09	21.22	20.91	20.91	20.95	20.95	20.99	21.25
Itr10	21.08	20.82	20.80	20.76	20.77	20.76	21.16

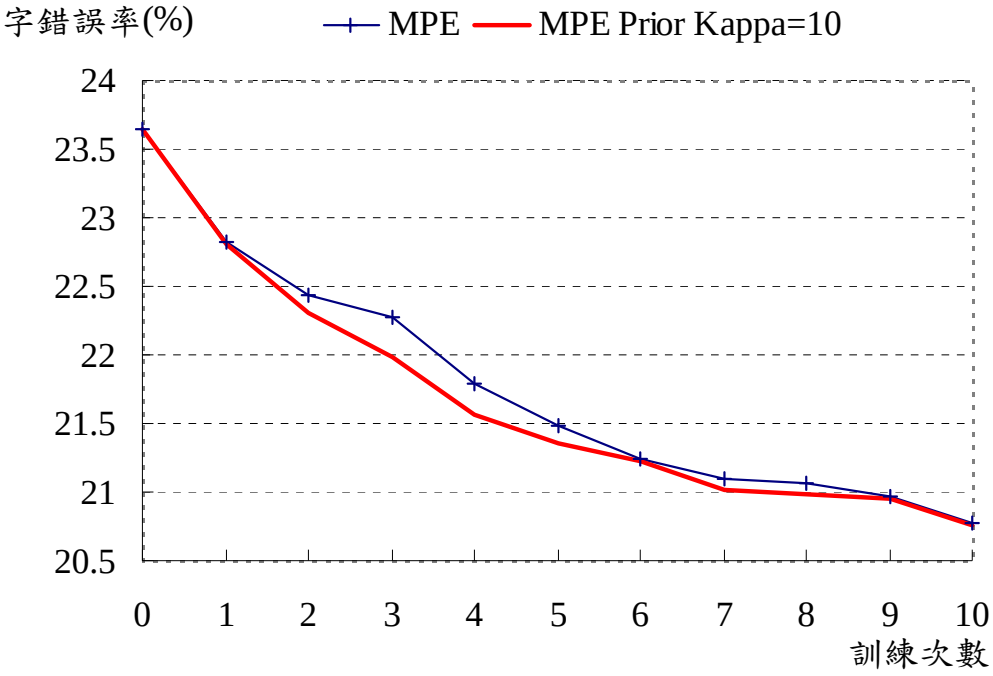


圖 6-7 考慮事前機率的最佳設定($\kappa=10$)
與最小化音素錯誤(MPE)訓練之比較

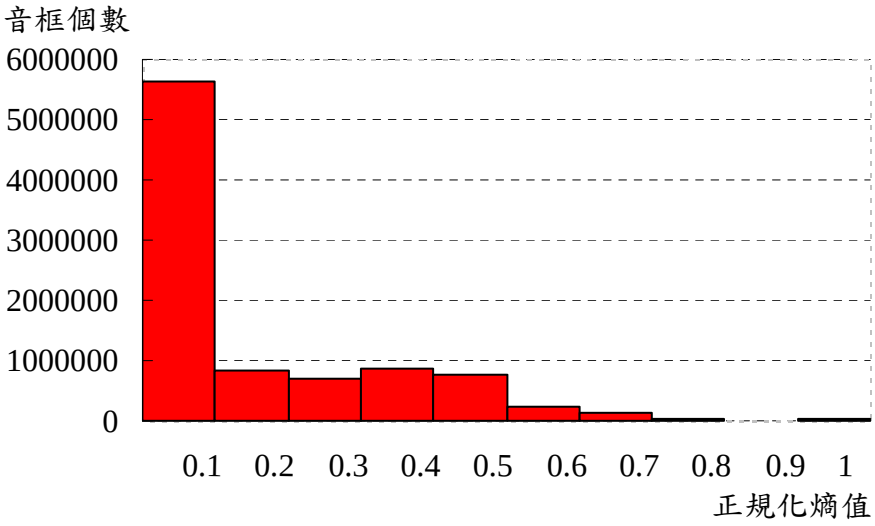


圖 6-8 所有訓練語句的正規化熵值分布圖

6.5 本論文提出的資料選取方法之實驗結果

本小節將呈現以正規化熵值為基礎的資料選取方法於鑑別式聲學模型訓練之實驗結果，包含最大化交互資訊(MMI)估測(6.5.1小節)、最小化音素錯誤(MPE)訓練(6.5.2小節)以及最大化*S*型時間音框正確率函數(MSTFA)(6.5.3小節)。首先，本論文分析所有訓練語料的正規化熵值分布，其分布圖如圖6-8所示，橫軸代表正規化熵值(Normalized Entropy)，縱軸代表時間音框的個數(Frame Number)，利用異質性線性鑑別分析(HLDA+MLLT+CN)之語音特徵與最大化相似度(ML)估測法訓練10次之後的聲學模型，對每一個音框算出其正規化熵值就可以求得此分布圖。在約25小時的訓練中，所有的時間音框總數為9183883個時間音框。從分析的數據中可以得知大部分時間音框的正規化熵值都集中在0.1以內(其時間音框總數為3561021個，佔所有時間音框總數的38.77%)，代表的意思就是大部份的時間音框都是可以完全辨識正確或完全辨識錯誤，只有少部分的時間音框是非常混淆的(Confused)。事實上，因為最大化相似度(ML)訓練10次之後所得到的字錯誤率為23.64%，其音素錯誤率會再更低一點，所以時間音框的正規化熵值集中在0.1以內，吾人認為大部份都是完全辨識正確，少部份則是完全辨識錯誤。完全辨識正確的時間音框在最大化相似度訓練中已貢獻非常多了，在鑑別式訓練中，其統計值就不再是那麼的重要，因此可以捨棄。由4.2小節的描述，以正規化熵值為基礎的時間音框資料選取方法有兩種，一種是硬性選取(Hard Selection，記作HS)，另一種為軟性選取(Soft Selection，記作SS)。本實驗嘗試將兩種選取方法結合(記作HS+SS)。

6.5.1 資料選取方法於最大化交互資訊估測

本子小節呈現資料選取方法於最大化交互資訊(MMI)估測之實驗結果。其最大化交互資訊的 I-平滑技術參數最佳化設定為 10[郭人瑋 2005]。所使用的時間音框資料選取方法為硬性選取(HS)，其實驗結果可以參考表 6-11。最佳門檻值(記作

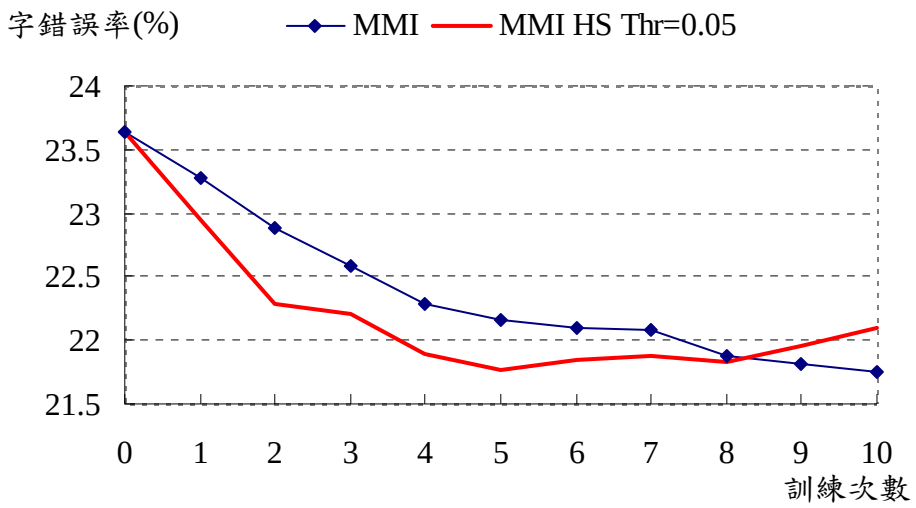


圖 6-9 硬性資料選取最佳設定(HS Thr=0.05)
與最大化交互資訊估測之比較

表 6-11 硬性資料選取方法(HS)於最大化交互資訊估測之實驗結果

CER(%)	MMI HS Thr=0.05	MMI HS Thr=0.06	MMI HS Thr=0.08	MMI HS Thr=0.1	MMI HS Thr=0.25
Baseline	23.64				
Itr01	22.95	22.96	22.91	22.84	22.89
Itr02	22.28	22.33	22.33	22.29	22.38
Itr03	22.21	22.18	22.06	22.11	22.33
Itr04	21.90	21.92	21.85	21.84	21.96
Itr05	21.77	21.86	21.88	21.94	22.08
Itr06	21.84	21.82	21.78	21.99	22.03
Itr07	21.88	21.90	21.92	21.92	21.90
Itr08	21.83	21.97	21.87	22.07	21.84
Itr09	21.95	22.20	22.27	22.46	21.76
Itr10	22.10	22.70	22.87	23.12	22.06

Thr)設定為 0.05(其時間音框總數為 4214360 個，佔所有時間音框總數的 45.88%)，其曲線圖可參考圖 6-9。資料選取方法可以加快收斂速度，但在第 10 次的訓練上卻沒有比最大化交互資訊估測的字錯誤率來得低，但其結果是差不多的。可能的原因將在 6.5.2 小節分析。

6.5.2 資料選取方法於最小化音素錯誤訓練

本子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。其最小化音素錯誤的 I-平滑技術參數最佳設定為 10[郭人瑋 2005]。所使用的時間音框資料選取方法為硬性選取(HS)及軟性選取(SS)，其實驗結果可以參考表 6-12 及表 6-13。其硬性選取固定門檻值(記作 Thr)最佳設定為 0.05(其時間音框總數為 4214360 個，佔所有時間音框總數的 45.88%)，如圖 6-10 所示，硬性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 次的訓練上卻沒有比最小化音素錯誤的字錯誤率來得低，其效果是差不多的。軟性資料選取的實驗結果曲線圖可以參考圖 6-11，其效果與最小化音素錯誤訓練差不多，沒有特別明顯的進步。針對硬性資料選取方法會加快收斂速度及在第 10 次的訓練上與最小化音素錯誤訓練之結果會差不多的原因，吾人猜想其可能的因素有兩個，茲分述如下:

1. 以下為針對加快收斂速度之猜想，鑑別式聲學模型訓練的目標函數(MMI or MPE)是使用延伸波氏重估演算法(Extended Baum-Welch, EBW)來作模型參數的最佳化，其中 EBW 有一個控制收斂速度的常數 D_s ，D_s 值越大則收斂速度越慢，D_s 值越小則收斂速度越快。在過去的文獻中，有學者在實作上是使用如下的公式可以得到最好的效果[Povey 2004]:

$$D_s = \max\{2D_s^{\min}, E \cdot \gamma_s^{den}\} \tag{6.2}$$

其中 $D_s^{\min}$ 為連續密度隱藏式馬可夫模型(CDHMM)中的某個狀態 s ，在模型參數調整時為確保共變異矩陣(Covariance Matrix)為正定(Positive Definition)

時的最小值。E 為可調整的常數，γ_s^{den} 為狀態 s 在所有訓練語料中所收集的統計值(負貢獻)。由式(6.2)可知控制收斂速度的常數 D_s 與訓練資料量的多寡有關，因此當我們使用資料選取方法時，勢必會減少訓練資料量，那麼統計值的收集便會減少，所以 D_s 的值就會變小，因此收斂速度會加快。

事實上，EBW 的模型參數調整公式如下(以平均值向量為例)，類似式(2.52):

$$\mu_s = \frac{\{\theta_s^{num}(O) - \theta_s^{den}(O)\} + D_s \overline{\mu}_s}{\{\gamma_s^{num} - \gamma_s^{den}\} + D_s} \tag{6.3}$$

在訓練資料量一樣的情形下，即沒有做資料選取，分子分母項的統計值沒有變動，那麼 D_s 的大小會影響收斂速度。但在使用資料選取方法調整時，控制收斂速度的常數 D_s 雖然變小，但其調整的響影力也隨著其分子分母項的統計值減少而變小。所以吾人的猜想是錯的，會加快收斂速度的原因應該是利用以正規化熵值為基礎的資料選取方法選到了具有鑑別力的時間音框樣本(Frame Samples)，即比較混淆的時間音框樣本，以幫助鑑別式模型訓練較快達到收斂。

2. 以下為在第 10 次會有差不多效果的猜想，在鑑別式訓練的目標函數一樣的情形下，不管有沒有使用資料選取方法，其目標函數的區域最佳值(Local Optimal)應該是一樣的。因此造成以正規化熵值的資料選取方法與最小化音速錯誤訓練在第 10 次的辨識率差不多。

事實上，目標函數的區域最佳值是與訓練資料量有關的，以最小化音素錯誤(MPE)訓練的目標函數為例:

$$F_{MPE}(\lambda) = \sum_{z=1}^{Z} \sum_{W_i \in \mathbf{W}_z} p(W_i \mid O_z) A(W_i, W_z) \tag{6.4}$$

當訓練資料量變動時，即 Z 變動(變大或變小)，那麼目標函數的值就會有所變動。所以此猜想也是錯的，那麼為什麼資料選取方法在第 10 次的訓練沒有比最小化音素錯誤訓練來的好呢，其原因就是使用以正規化熵值的資料選

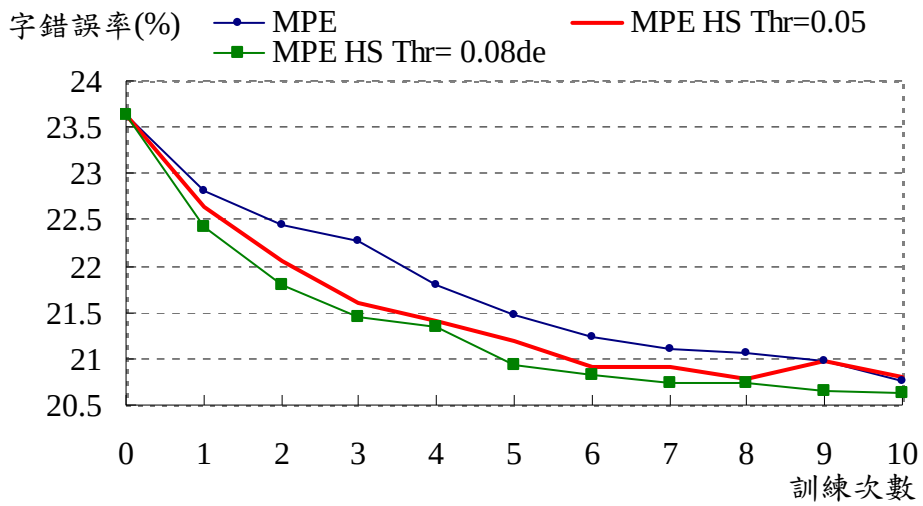


圖 6-10 硬性資料選取方法固定門檻值最佳設定(HS Thr=0.05)及動態最佳設定(HS Thr=0.08de)與最小化音素錯誤之比較

表 6-12 硬性資料選取方法於最小化音素錯誤訓練

CER(%)	MPE HS Thr=0.05	MPE HS Thr=0.06	MPE HS Thr=0.08de	MPE HS Thr=0.1	MPE HS Thr=0.15
Baseline	23.64				
Itr01	22.63	22.55	22.43	22.55	22.53
Itr02	22.05	22.02	21.80	21.94	21.88
Itr03	21.60	21.66	21.45	21.70	21.74
Itr04	21.40	21.44	21.34	21.53	21.54
Itr05	21.19	21.24	20.94	21.27	21.39
Itr06	20.92	20.95	20.82	21.28	21.41
Itr07	20.91	21.06	20.73	—	—
Itr08	21.22	21.17	20.74	—	—
Itr09	21.08	21.18	20.65	—	—
Itr10	21.29	21.02	20.63	—	—

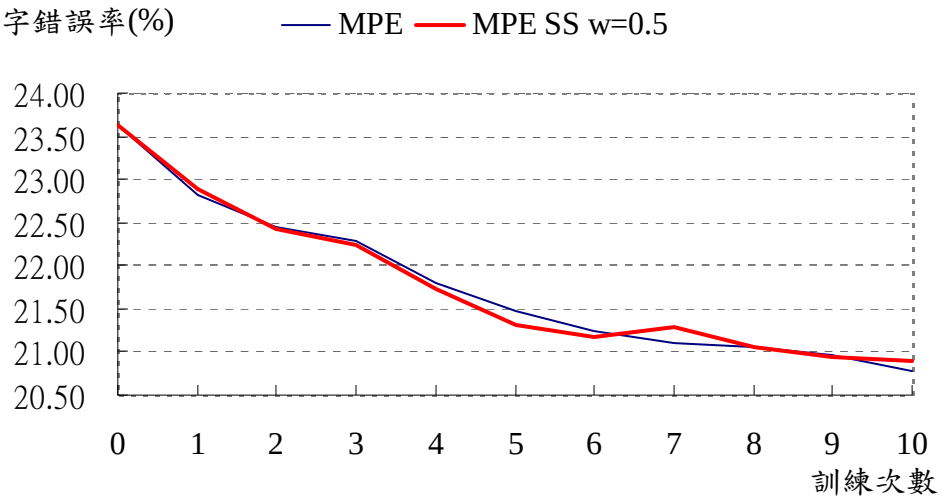


圖 6-11 軟性資料選取方法門檻值最佳設定(SS w=0.5)
與最小化音素錯誤之比較

表 6-13 軟性資料選取方法於最小化音素錯誤訓練

CER(%)	MPE	MPE SS w=1	MPE SS w=0.5	MPE SS w=2	MPE SS w=1.5
Baseline	23.64				
Itr01	22.82	22.84	22.88	22.91	22.86
Itr02	22.44	22.40	22.43	22.36	22.37
Itr03	22.28	22.21	22.25	22.11	22.12
Itr04	21.79	21.65	21.73	21.62	21.60
Itr05	21.48	21.34	21.31	21.34	21.31
Itr06	21.24	21.33	21.18	21.48	21.37
Itr07	21.10	21.29	21.29	21.18	21.24
Itr08	21.06	21.00	21.06	21.15	21.13
Itr09	20.97	21.02	20.93	20.97	21.02
Itr10	20.77	20.94	20.89	21.04	21.08

取方法只選出有鑑別力的時間音框資料樣本，訓練資料量減少，以致模型參數的訓練會遭遇到過度訓練(Over-training)的問題，因此資料選取方法在第10 次的訓練就沒有比最小化音素錯誤訓練來得好。

由以上兩點的分析，說明了以正規化熵值為基礎的資料選取方法確實是有用的，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，也因為這些時間音框樣本本身比較混淆，所以對鑑別式訓練會特別有幫助。

另一方面，吾人使用隨機選取(Random Selection)方法作為比較來驗證以正規化熵值為基礎的資料選取方法是有用的，而不是亂選的。實驗結果可以參考圖6-12 和表 6-14，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%。

為了想要加快收斂速度且要避免因使用資料選取方法而會遇到過度訓練的問題，本論文使用隨迭代次數的增加而減少門檻值設定的方法(記作 MPE HS 0.08de)來解決此問題，其實驗結果可以參考圖 6-10、圖 6-12、表 6-12 及表 6-14，由其曲線圖得知，使用門檻值遞減的方法確實能克服過度訓練的問題，而且在每一次的迭代訓練中都比最小化音素錯誤訓練來得好，在第 10 次的迭代中之辨識字錯誤率比 MPE 好 0.14%。

另外，本論文使用了另一個測試集(Testing Set，1.5hrs 包含 307 句，與 292 句為同一時期的資料)來驗證以熵值為基礎的資料選取方法的一般性，其實驗結果可以參考圖 6-13 及表 6-15，由曲線圖得知其資料選取方法確實能夠加快收斂速度。

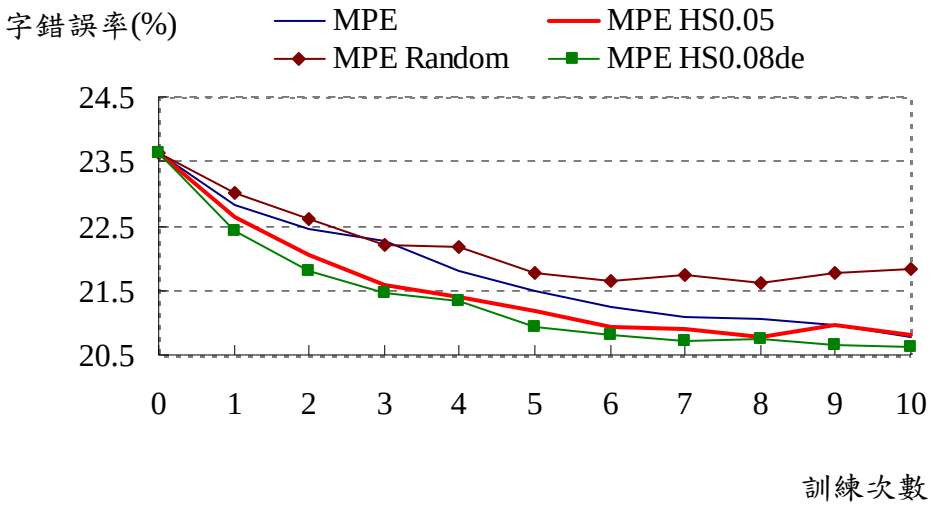


圖 6-12 以隨機選取作為比較對象

表 6-14 以隨機選取作為比較對象之實驗結果

CER(%)	MPE	MPE HS 0.05	MPE HS 0.08de	MPE Random
Baseline	23.64			
Itr01	22.82	22.63	22.43	23.02
Itr02	22.44	22.05	21.80	22.62
Itr03	22.28	21.60	21.45	22.22
Itr04	21.79	21.40	21.34	22.16
Itr05	21.48	21.19	20.94	21.76
Itr06	21.24	20.92	20.82	21.66
Itr07	21.10	20.90	20.73	21.74
Itr08	21.06	20.79	20.74	21.62
Itr09	20.97	20.97	20.65	21.78
Itr10	20.77	20.80	20.63	21.84

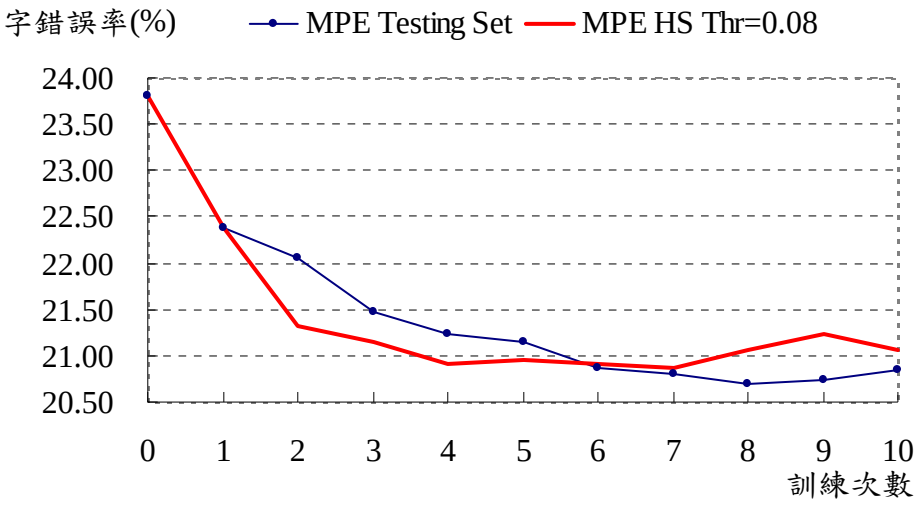


圖 6-13 硬性資料選取方法(固定門檻值 HS Thr=0.08)
於最小化音素錯誤訓練在另一測試集

表 6-15 硬性資料選取方法於最小化音素錯誤訓練在另一測試集

CER(%)	MPE Testing Set	MPE HS Thr=0.08
Baseline	23.80	
Itr01	22.38	22.37
Itr02	22.06	21.33
Itr03	21.48	21.14
Itr04	21.24	20.90
Itr05	21.15	20.95
Itr06	20.86	20.90
Itr07	20.80	20.86
Itr08	20.70	21.06
Itr09	20.74	21.24
Itr10	20.84	21.07

6.5.3 資料選取方法於最大化 *S* 型時間音框正確率函數

在 6.4.2 小節中，考慮刪除錯誤的最大化 *S* 型時間音框錯誤率函數(MSTFA)相對於最小化音素錯誤(MPE)訓練有明顯的效果。所以本子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為 10。所使用的時間音框資料選取方法為硬性選取(HS)、軟性選取(SS)以及結合硬性和軟性選取(HS+SS)，其實驗結果分別可以參考表 6-16、表 6-17 和表 6-18。其硬性選取最佳門檻值(記作 Thr)設定為 0.05，如圖 6-14、圖 6-15 和圖 6-16 所示，資料選取方法應用在最大化 *S* 型時間音框正確率函數確實可以加快收斂速度，但在第 10 次的訓練上卻沒有比最大化 *S* 型時間音框正確率函數的字錯誤率來得低。另外，如圖 6-15 所示，軟性資料選取方法比硬性選取效果來得好，在第 10 次的訓練上跟最大化 *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)的實驗結果似乎沒有比較好，沒有加成性的效果。

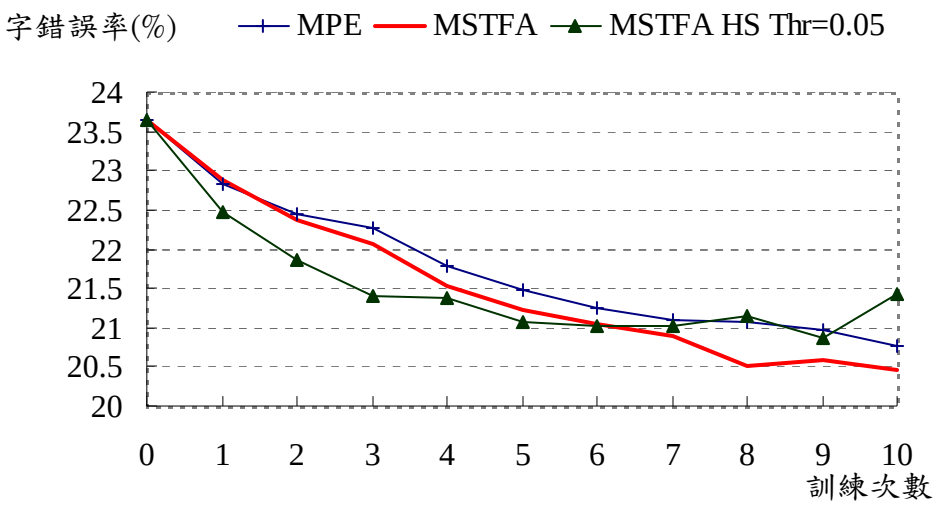


圖 6-14 硬性資料選取方法(HS Thr=0.05)與最大化 *S* 型時間音框正確率函數之比較

表 6-16 硬性資料選取方法(HS)於最大化 *S* 型時間音框正確率函數

CER(%)	MPE	MSTFA	MSTFA HS Thr=0.05
Baseline	23.64		
Itr01	22.82	22.88	22.46
Itr02	22.44	22.37	21.87
Itr03	22.28	22.06	21.40
Itr04	21.79	21.52	21.38
Itr05	21.48	21.23	21.08
Itr06	21.24	21.05	21.03
Itr07	21.10	20.89	21.02
Itr08	21.06	20.50	21.15
Itr09	20.97	20.58	20.86
Itr10	20.77	20.46	21.43

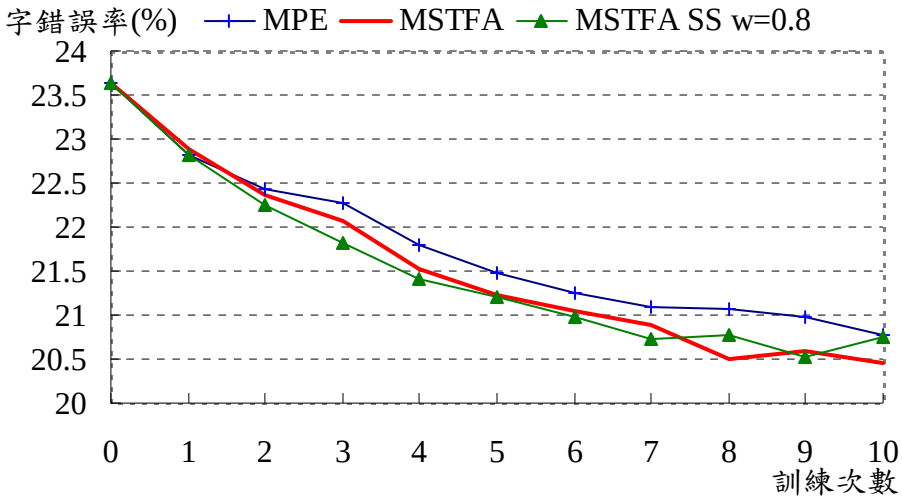


圖 6-15 軟性資料選取方法最佳化設定(SS w=0.8, Lo=0.1, Alpha=0.5)與最大化 *S* 型時間音框正確率函數之比較

表 6-17 軟性資料選取方法(SS)於最大化 *S* 型時間音框正確率函數

CER(%)	MPE	MSTFA Lo=0.1 Alpha=0.5	MSTFA SS w=0.1 Lo=0.1 Alpha=0.5	MSTFA SS w=0.5 Lo=0.1 Alpha=0.5	MSTFA SS w=0.8 Lo=0.1 Alpha=0.5	MSTFA SS w=1 Lo=0.1 Alpha=0.5
Baseline	23.64					
Itr01	22.82	22.88	22.88	22.88	22.81	22.75
Itr02	22.44	22.37	22.35	22.28	22.26	22.25
Itr03	22.28	22.06	21.98	21.88	21.81	21.83
Itr04	21.79	21.52	21.45	21.41	21.42	21.45
Itr05	21.48	21.23	21.21	21.16	21.21	21.27
Itr06	21.24	21.05	21.15	20.99	20.98	20.94
Itr07	21.10	20.89	—	—	20.73	20.65
Itr08	21.06	20.50	—	—	20.78	20.78
Itr09	20.97	20.58	—	—	20.53	20.56
Itr10	20.77	20.46	—	—	20.75	20.86

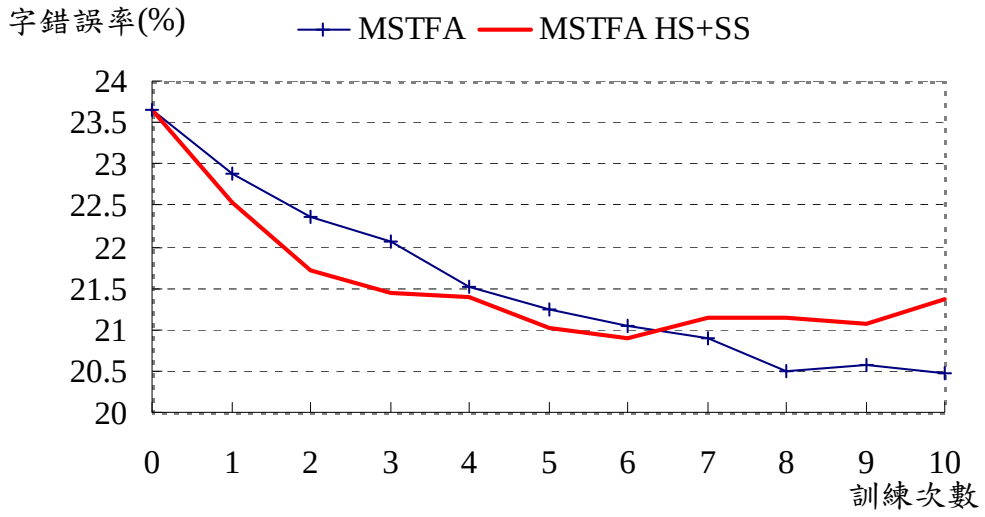


圖 6-16 結合硬性和軟性資料選取方法最佳化設定(HS Thr=0.1, SS w=0.5, Lo=0.1, Alpha=0.5)與最大化 *S* 型時間音框正確率函數之比較

表 6-18 結合硬性和軟性資料選取方法(HS+SS)於最大化 *S* 型時間音框正確率函數

CER(%)	MSTFA HS Thr=0.05 SS w=0.1 Lo=0.1 Alpha=0.5	MSTFA HS Thr=0.05 SS w=0.5 Lo=0.1 Alpha=0.5	MSTFA HS Thr=0.05 SS w=0.8 Lo=0.1 Alpha=0.5	MSTFA HS Thr=0.1 SS w=0.5 Lo=0.1 Alpha=0.5	MSTFA HS Thr=0.1 SS w=0.8 Lo=0.1 Alpha=0.5
Baseline	23.64				
Itr01	22.55	22.56	22.55	22.53	22.49
Itr02	21.79	21.77	21.73	21.72	21.85
Itr03	21.49	21.55	21.46	21.45	21.63
Itr04	21.28	21.61	21.31	21.38	21.56
Itr05	20.98	21.20	21.00	21.03	21.11
Itr06	20.96	21.26	20.98	20.90	21.28
Itr07	21.10	21.06	21.39	21.14	21.02
Itr08	21.18	21.28	21.49	21.14	21.34
Itr09	21.03	21.23	21.95	21.07	21.26
Itr10	21.39	21.73	21.97	21.37	21.21

6.6 非監督式之實驗結果

本小節將呈現非監督式最大化相似度、非監督鑑別式訓練以及資料選取方法於非監督鑑別式訓練之實驗結果。使用的實驗語料為公視新聞外場記者的部份(共34672 句，約 24.5 小時)，其中 200 句(約 11 分鐘)用來訓練初始模型(Initial Model)，其他的 34472 句(約 24 小時)當成大量未轉譯的語料。本實驗共使用兩種語音特徵，一為梅爾倒頻譜特徵加上倒頻譜正規化法(記作 MFCC+CN)，另一為異質性線性鑑別分析加上最大化相似度線性轉換與倒頻譜正規化法(記作 HLDA+MLLT+CN)。初始模型(Initial Model)使用較少的高斯模型個數(約 3100 個，依 11 分鐘的訓練量來決定其個數)，使用 HTK Toolkit 的 Hinit 和 HRest 函數來訓練初始的聲學模型。其初始模型的字錯誤率還蠻高的，實驗結果可參考表 6-19，表示用少量的訓練語料(如 11 分鐘)來訓練此初始的聲學模型會使模型不夠強健。當有大量未轉譯的語料時，我們應該提高模型的複雜度以訓練出較強健的模型，也就是增加高斯模型的個數，這個部份是使用 HHed 函數來實作，分裂後的高斯模型個數約 13500 個，其實驗結果可參考表 6-19。

在本實驗中迭代方法只用在最大化相似度(ML)的聲學模型訓練上，一共做了三次迭代，每一次的迭代中都作 10 次的最大化相似度聲學模型訓練，其實驗流程可參考圖 6-17，其實驗結果可以參考表 6-20、表 6-21、表 6-22、圖 6-19、圖 6-20 和圖 6-21。由實驗數據顯示出初始模型的字錯誤率在第一次的迭代中的第一次模型訓練就大幅下降，之後的訓練就很緩慢的下降，到第三次迭代時，就可以看出快要收斂了。

信心度評估的實驗數據顯示只有少量的進步，其門檻值(記作 Conf)隨著迭代次數的增加而遞減，實驗結果顯示出信心度評估有一點點的幫助，其信心度的分析圖可以參考圖 6-18，由圖中可以知道大部分的詞段之信心度都介於 0.9 到 1 之間，這也表示以事後機率為基礎的信心度評估相當不準，所以信心度評估在本實

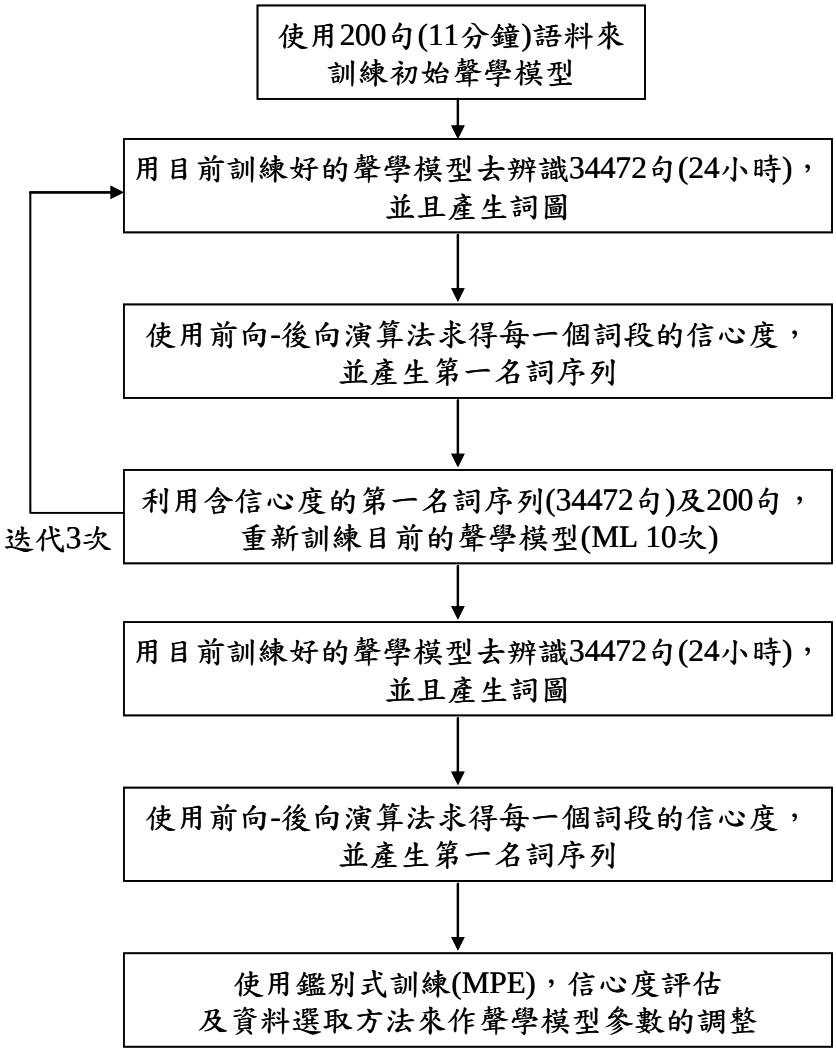


圖 6-17 非監督鑑別式訓練之流程

驗中才會幫助不大。接著就使用最小化音素錯誤(MPE)訓練法則來訓練聲學模型，其實驗結果可參考表 6-23 和圖 6-22。由實驗數據可以得知鑑別式訓練在非監督的情況下仍然是有幫助的。

最後，資料選取方法(記作 FS)應用在非監督鑑別式聲學模型訓練上的實驗可以參考表 6-23 和圖 6-22，其效果並不如預期。效果不好的可能原因是非監督的作法本身就會遇到辨識錯誤的問題，那麼應用資料選取方法在非監督式的訓練便可能沒有什麼幫助。

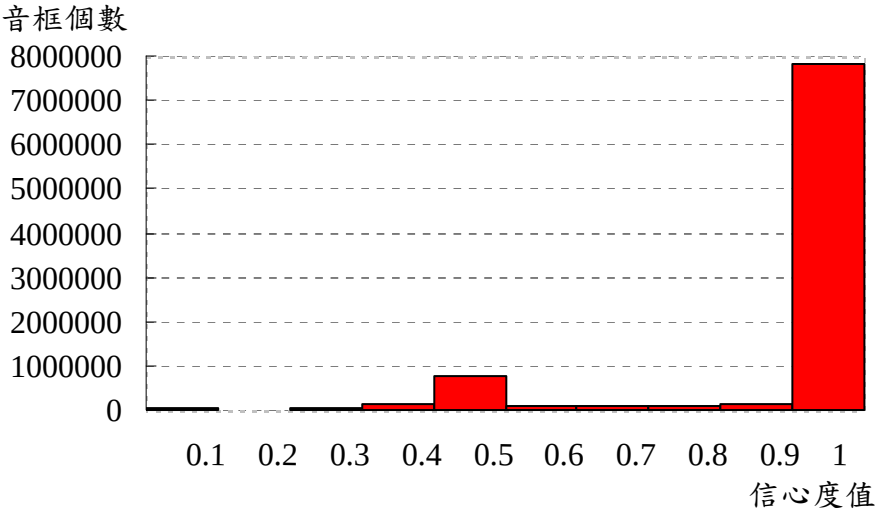


圖 6-18 信心度評估之分布圖

表 6-19 初始模型的實驗結果

CER(%)	MFCC+CN	HLDA+MLLT+CN
HRest	58.95	58.37
HHed	58.31	57.80

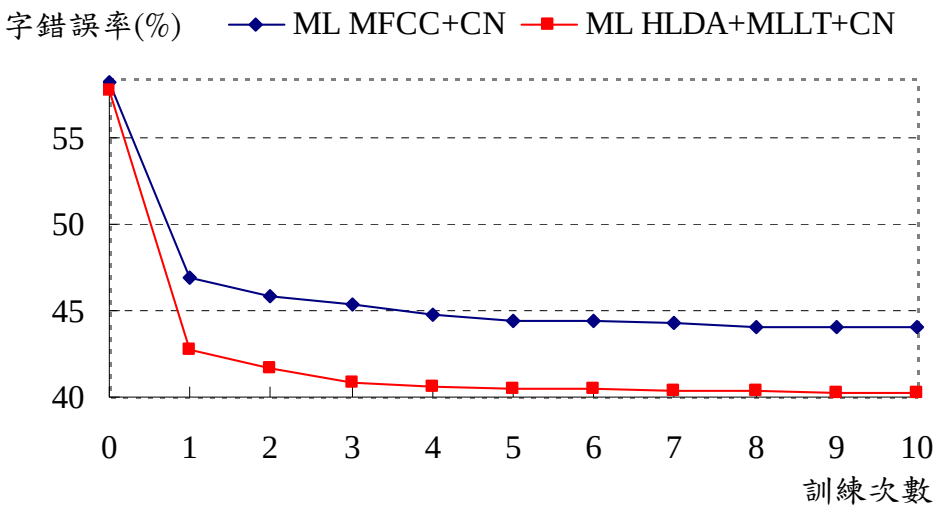


圖 6-19 最大化相似度模型訓練第一次迭代之實驗結果

表 6-20 最大化相似度模型訓練第一次迭代之實驗結果

CER(%)	MFCC+CN Conf=0	MFCC+CN Conf=0.9	HLDA+MLLT+CN Conf=0
HHed	58.31		57.80
Itr01	46.97	46.75	42.76
Itr02	45.87	45.62	41.66
Itr03	45.37	45.29	40.83
Itr04	44.73	44.78	40.56
Itr05	44.46	44.60	40.42
Itr06	44.41	44.38	40.51
Itr07	44.35	44.22	40.37
Itr08	44.07	44.17	40.35
Itr09	44.07	44.28	40.20
Itr10	44.04	44.11	40.20

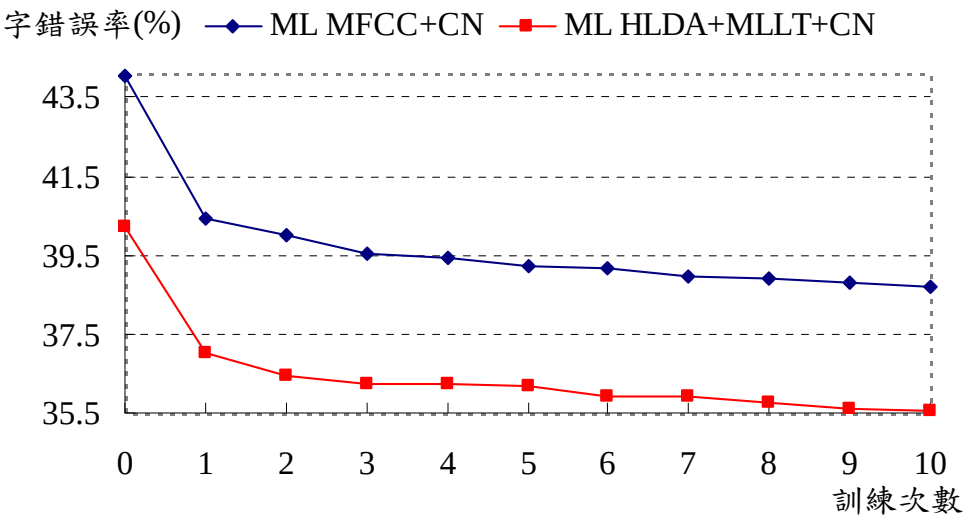


圖 6-20 最大化相似度模型訓練第二次迭代之實驗結果

表 6-21 最大化相似度模型訓練第二次迭代之實驗結果

CER(%)	MFCC+CN Conf=0	HLDA+MLLT+CN Conf=0
ML1_Itr10	44.04	40.20
Itr01	40.43	37.00
Itr02	40.00	36.47
Itr03	39.56	36.26
Itr04	39.45	36.24
Itr05	39.24	36.16
Itr06	39.16	35.93
Itr07	38.98	35.90
Itr08	38.93	35.75
Itr09	38.79	35.59
Itr10	38.68	35.56

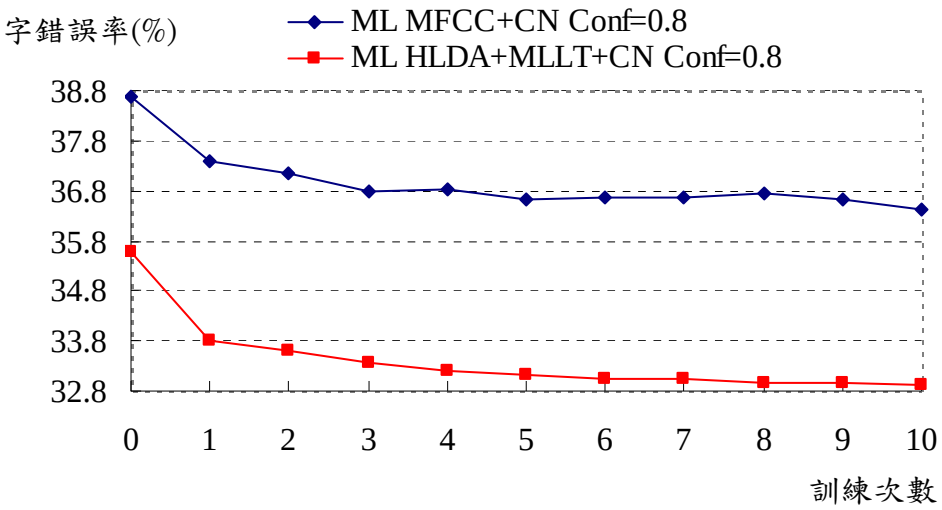


圖 6-21 最大化相似度模型訓練第三次迭代之實驗結果

表 6-22 最大化相似度模型訓練第三次迭代之實驗結果

CER(%)	MFCC+CN Conf=0	MFCC+CN Conf=0.8	HLDA+MLLT+CN Conf=0	HLDA+MLLT+CN Conf=0.8
ML2_itr10	38.68		35.56	
Itr01	37.31	37.40	33.90	33.80
Itr02	37.24	37.14	33.61	33.59
Itr03	37.14	36.79	33.47	33.35
Itr04	36.75	36.82	33.20	33.19
Itr05	36.63	36.64	33.16	33.14
Itr06	36.61	36.68	33.14	33.03
Itr07	36.64	36.66	33.19	33.06
Itr08	36.65	36.73	33.10	32.95
Itr09	36.65	36.62	32.93	32.97
Itr10	36.62	36.42	33.00	32.91

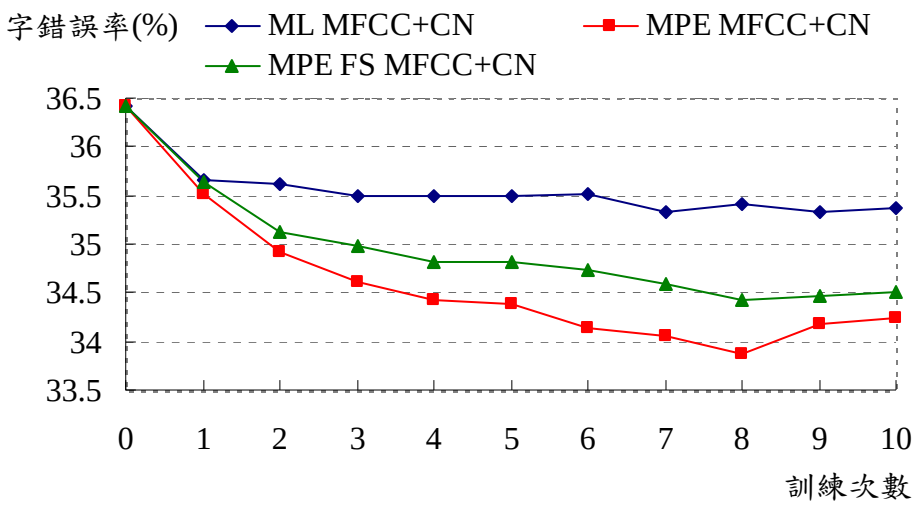


圖 6-22 非監督鑑別式聲學模型訓練之實驗結果(MFCC+CN)

表 6-23 非監督鑑別式聲學模型訓練之實驗結果(MFCC+CN)

CER(%)	ML MFCCCN Conf=0.5	MPE MFCC+CN Conf= 0	MPE MFCC+CN Conf= 0.5	MPE MFCC+CN Conf= 0.7	MPE FS MFCC+CN Conf= 0.5
ML3 itr10	36.42	36.42	36.42	36.42	36.42
Itr01	35.65	35.51	35.58	35.68	35.63
Itr02	35.61	34.91	34.86	34.95	35.13
Itr03	35.50	34.61	34.61	34.75	34.98
Itr04	35.49	34.43	34.42	34.63	34.82
Itr05	35.50	34.38	34.40	34.51	34.82
Itr06	35.52	34.14	34.23	34.33	34.73
Itr07	35.33	34.06	34.23	34.10	34.58
Itr08	35.41	33.87	34.08	34.31	34.43
Itr09	35.33	34.18	34.19	34.20	34.46
Itr10	35.38	34.24	34.14	34.39	34.51

第7章　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。本論文旨在改善鑑別式聲學模型訓練，相關研究內容與成果可從下面三個面向來作討論:

(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為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(此事後機率有考慮到詞與詞之間的轉移機率，即語言模型)來求得的，所以可以視為是在事後機率定義域中(有別於傳統是在相似度定義域中)來選取訓練樣本，且所選出來的訓練樣本是比較混淆的，那麼對鑑別式訓練來說，這些混淆的訓練樣本是比較具有鑑別力的。由大量的實驗結果顯示，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。本論文也企圖將此新的資料選取方法應用到非監督鑑別式聲學模型訓練上。但因非監督式訓練本身就會遭遇到辨識錯誤的問題，所以用以正規化熵值的資料選取方法選出來的訓練樣本即使很混淆，也沒有辦法用鑑別式訓練讓這些混淆的樣本離決定邊界較遠，因為沒有正確轉譯詞序列的資訊，所以在實驗結果上，並沒有比傳統的鑑別式訓練好。

(3) 最後，本論文考量以全面風險為出發的鑑別式聲學模型訓練之事前機率，提出以統計式的方法來近似每個訓練語句的事前機率。在實驗結果上，雖然進

步的幅度很小，但考慮事前機率在最小化音素錯誤訓練的影響確實對字錯誤率降低有幫助。

每個訓練語句本來就應該要有不同的分布，只是目前還不知道其分布為何，所以訓練語句之事前機率的估測或許是個研究的方向，未來可能嘗試使用音韻的資訊來估測語句的事前機率。

以全面風險為基礎的鑑別式聲學模型訓練中的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為基礎的原始音素正確率函數就還有改進的空間。有國外學者 Jun Du 等人提出以聲學模型間的關係來計算正確率以取代以類別為基礎的正確率函數。類別比對為基礎的和聲學模型間的關係的減損函數都各有其優缺點，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。

未來吾人也想要將以正規化熵值為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。除了正規化熵值是一個選取資料的工具，應該還有別的類似的方法(或許會更好)，可以應用在鑑別式聲學模型訓練上。事實上，由加快鑑別式訓練的收斂速度來看，此以正規化熵值為基礎之新的資料選取方法的確為鑑別式聲學模型訓練提供了一個新的視野。

參考文獻

[A. Smola *et al.*] A. J. Smola, P. Bartlett, B. Scholkopf, D. Schuurmans, “Advances in Large Margin Classifiers”, The *MIT* Press

[A. Stolcke *et al.* 1997] A. Stolcke, Y. Konig, M. Weintraub, “Explicit Word Error Minimization in N-Best List Rescoring”, *in Proc. ICASSP* 1997

[Atal 1974] B. S. Atal, “Effectiveness of Linear Prediction Characteristics of The Speech Wave for Automatic Speaker Identification and Verification,” *Journal of the Acoustical Society of America, Vol. 55, No. 6, pp.1304-1312*, 1974

[Aubert 2002] X. Aubert, “An Overview of Decoding Techniques for Large Vocabulary Continuous Speech Recognition,” *Computer Speech and Language, Vol. 16, pp. 89-114*, 2002

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

[Bahl *et al.* 1986] L. R. Bahl, P. F. Brown, P. V. de Souza and R. L. Mercer. “Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition,” *in Proc. ICASSP* 1986.

[Barras *et al.* 2001] C. Barras, E. Geoffrois, Z.B. Wu and M. Liberman, “Transcriber : Development and use of a tool for assisting speech corpora production,” Speech communication, 33 : 5-22, 2001.

[Baum 1972] L. E. Baum (1972). “An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes,” *Inequalities,* 3(1):1-8, 1972.

[B.H. Junag *et al.* 1992] B. H. Juang, S. Katagiri, “Discriminative Learning for Minimum Classification Error”, *IEEE Trans. Signal Processing, Vol.40, No.12*

1992

[B.H. Junag *et al.* 1997] B. H. Juang, Wu Chou, Chin-Hui Lee, “Minimum Classification Error Rate Methods for Speech Recognition”, *IEEE Trans. SAP, Vol.5, No.3* 1997

[Chen *et al.* 2002] B. Chen, H.-M. Wang , and L.-S. Lee, “Discriminating Capabilities of Syllable-based features and approaches of utilizing them for voice retrieval of speech information in Mandarin Chinese,” *IEEE Trans. Speech and Audio Processing , 10(5)* : 303-314, 2002.

[Chen *et al.* 2004] B. Chen, J.-W. Kuo, W.H. Tsai, “Lightly supervised ad data-driven approaches to Mandarin broadcast news transcription,” *in Proc. ICASSP*, 2004

[Chen *et al*. 2005] B. Chen, J.-W. Kuo and W.-H. Tsai, "Lightly Supervised and Data-driven Approaches to Mandarin Broadcast News Transcription," *International Journal of Computational Linguistics & Chinese Language Processing, Vol. 10, No. 1*, pp1-18,2005

[Davis and Mermelstein 1980] S. B. Davis and P. Mermelstein, “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences,” *IEEE Trans. Acoustic, Speech, and Signal Processing, Vol. 28, No. 4, pp.357-366*, 1980

[Doumpiotis *et al.* 2004] V. Doumpiotis, S. Tsakalidis, W. Byrne (2004). “Lattice Segmentation and Minimum Bayes Risk Discriminative Training,” in *Proc. Eurospeech’04.*

[Doumpiotis & Byrne 2004] V. Doumpiotis and W. Byrne (2004). “Pinched Lattice Minimum Bayes Risk Discriminative Traning for Large Vocabulary Continuous Speech Recognition,” in *Proc. ICSLP’04.*

[Duda *et al.* 1973] R. O. Duda, P. E. Hart and D. G. Stork (2000). *Pattern Classification, First Edition*. New York: John & Wiley, 2000.

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

[Fiscus 1997] J. Fiscus (1997). "A Post-processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER)," in *Proc. ASRU'97.*

[Gales 1999] M. J. F. Gales, "Semi-tied Covariance Matrices for Hidden Markov Models," *IEEE Trans. on Speech, Audio and Signal Processing, Vol. 7, No.3, pp. 272-281*, 1999

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

[Gibson *et al.* 2006] Gibson M. and Hain T., "Hypothesis Spaces for Minimum Bayes Risk Training in Large Vocabulary Speech Recognition", *in Proc. ICSLP 2006*

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

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

[Goel *et al*. 2004] V. Goel ,S. Kumar, W. Byrne (2004). "Segmental Minimum Bayes-Risk Decoding for Automatic Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, Vol. 12, No. 3, pp.234-249, 2004.

[Gopalakrishnan *et al.* 1991] P. S. Gopalakrishnan, D. Kanevsky, A. Nádas & D. Nahamoo (1991). "An Inequality for Rational Functions with Applications to Some Statistical Estimation Problems," *IEEE Trans. Information Theory,* Vol. 37, pp.107-113, 1991.

[Huang *et al*. 2001] X. Huang, A. Acero and H. Hon, "Spoken Language Processing," Prentice Hall, 2001

[Jiang *et al.* 2006] Hui Jiang, Xinwei Li, Chaojun Liu, “Large Margin Hidden Markov Models for Speech Recognition”, *IEEE Transaction on ASLP* 2006

[Jiang 2005] H. Jiang, “Confidence Measures for Speech Recognition: A Survey,” *Speech Communication, Vol. 45, pp. 455-470*, 2005.

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

[Jinyu Li *et al.* 2007] Jinyu Li, S. M. Siniscalchi, Chin-Hui Lee, “Appeoximate Test Risk Minimization Through Soft Margin Estimation”, *in Proc. ICASSP* 2007

[Juang & Katagiri 1992] B.-H. Juang and S. Katagiri (1992). “Discriminative Learning for Minimum Error Classification,” *IEEE Trans. Signal Processing*, Vol. 40, No. 12, pp. 3043-3054, 1992.

[J. Zheng *et al.* 2005] Jing Zheng and Andreas Stolcke (2005) “Improved Discriminative Training Using Phone Lattices”, *In Proc. Interspeech* 2005

[Jun Du *et al.* 2006] Jun Du, Peng Liu, F. K. Soong, J. L. Zhou, R. H. Wang, “Minimum Divergence Based Discriminative Training”, *in Proc. ICSLP* 2006

[Jun Du *et al.* 2007] J. Du, P. Liu, F. K. Soong, J. L. Zhou, R. H. Wang, “A New Minimum Divergence Approach to Discriminative Training”, *in Proc. ICASSP* 2007

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

[Kamppari *et al.* 2000] S. O. Kamppari and T. J. Hazen, “Word and Phone Level Acoustic Confidence Scoring,” *in Proc. of ICASSP* 2000

[Katagiri *et al.* 1998] S. Katagiri, B. H. Juang, Chih-Hui Lee, “Pattern Recognition Using a Family of Design Algorithms based upon the Generalized Probabilistic Descent Method”, *Proceeding of the IEEE, Vol. 86, No.11*, 1998

[Katz 1987] S. M. Katz, “Estimation of probabilities form sparse data for other language component of a speech recognizer,” *IEEE Trans. Acoustics, Speech and Signal Processing* , 35(5) : 300-401, 1987

[Korkmazsky *et al.* 2004] F. Korkmazsky, D. Fohr and I. Illina, “Using Linear Interpolation to Improve Histogram Equalization for Speech Recognition,” *in Proc. of ICSLP*, 2004

[Kumar 1997] 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

[Kuo *et al.* 2005] Jen-Wei Kuo, Berlin Chen, "Minimum Word Error Based Discriminative Training of Language Models," *in Proc. Eurospeech* 2005

[Kuo *et al.* 2006] Jen-Wei Kuo, Shih-Hung Liu, Hsin-min Wang, Berlin Chen, "An Empirical Study of Word Error Minimization Approaches for Mandarin Large Vocabulary Speech Recognition," *International Journal of Computational Linguistics & Chinese Language Processing, Vol. 11, No. 3*, 2006

[Lamel 2002] Lori Lamel, J. Gauvain, G.. Adda, “Lightly Supervised and Unsupervised Acoustic Model Training,” *Computer Speech and Language*, Vol.16, pp.115-129, 2002

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

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

[Li *et al.* 2005] Xinwei Li, Hui Jiang, Chaojun Liu, “Large Margin HMMs for Speech Recognition”, *in Proc. ICASSP* 2005

[Lin *et al.* 2006] Shih-Hsiang Lin, Yao-Ming Yeh, Berlin Chen, "Exploiting Polynomial-Fit Histogram Equalization and Temporal Average for Robust Speech

Recognition," *in Proc. ICSLP* 2006.

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

[Ma *et al.* 2006] J. Ma, S. Matsoukas, O. Kimball, R. Schwartz, "Unsupervised Training on Large Amounts of Broadcast News Data", *in Proc. ICASSP* 2006

[Matias *et al.* 2006] L. Mathias, G.. Y., J. Fritsch , "Discriminative Training of Acoustic Models Applied to Domains with Unreliable Transcripts", *in Proc. ICASSP* 2005

[Mangu *et al.* 2000] 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.

[McDermott *et al.* 1997] E. McDermott and S. Katagiri (1997). "String-Level MCE for Continuous Phoneme Recognition," in *Proc. Eurospeech* 1997

[Na *et al.* 1995] 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.

[Ney *et al*. 1994] H. Ney, U. Essen, and R. Kneser, "On Structuring Probabilistic Dependences in Stochastic Language Modeling," *Computer Speech and Language, Vol. 8, pp.1-38*, 1994

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

[NTNU 2004] Speech Lab, Graduate Institute of Computer Science and Information Engineering, Nation Taiwan Normal University. http://speech.csie.nctu.edu.tw/

[Ortmanns *et al.* 1997] 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

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

[Povey 2004] Daniel Povey , "Discriminative Training for Large Vocabulary Speech Recognition," Ph.D Dissertation, University of Cambridge, 2004.

[Povey *et al.* 2007] Daniel Povey and B. Kingsbury, "Evaluation of Proposed Modifications to MPE for Large Scale Discriminative Training", *in Proc. ICASSP* 2007.

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

[Rabiner 1989] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications inspeech recognition", *Proceedings of the IEEE* 1989

[Rose *et al*. 1995] R. C. Rose, B. H. Juang and C.-H. Lee, "A Training Procedure for Verifying String Hypothesis in Continuous Speech Recogniton," *in Proc. of ICASSP* 1995

[Rosenfeld 1996] R. Rosenfeld, "A Maximum Entropy Approach to Adaptive Statistical Language Modeling," *Computer Speech and Language, Vol. 10, No. 2, pp 187-228*, 1996

[Sanchis *et al*. 2004] A. Sanchis, A. Juan, and E. Vidal, "New Features Based on Multiple Word Graphs For Utterance Verification," *in Proc. of ICSLP*, 2004

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

[SLP NTNU] Speech Lab, Graduate Institute of Computer Science and Information Engineering, Nation Taiwan Normal University. http://speech.csie.nctu.edu.tw/

[SRILM 2002] A. Stolcke, “SRI language modeling toolkit,” Version 1.5.2, http://www.speech.sri.com/projects/srilm/ .

[Schwartz *et al.* 1990] R. Schwartz and Y. L. Chow, “The N-Best algorithm: an efficient

and exact procedure for finding the N most likely sentence hypotheses,”*in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol.1, pp. 81-84, 1990.

[Schlüter *et al.* 2001] R. Schlüter, W. Macherey, B. Muller, H. Ney (2001). “Comparison of Discriminative Training Criteria and Optimization Methods for Speech Recognition,” *Speech Communication*, Vol. 34, pp. 287-310, 2001

[Valtchev *et al.* 1996] V. Valtchev, J. J. Odell, P. C. Woodland, S. J. Young. (1996). “Lattice-Based Discriminative Training for Large Vocabulary Speech Recognition,” *in Proc. ICASSP* 1996.

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

[Viterbi 1967] A. J. Viterbi (1967). “Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm,” *IEEE Trans. Information Theory,* vol. 13, no. 2, April 1967.

[Viikki and Laurila 1998] O. Viikki, K. Laurila, “Cepstral Domain Segmental Feature Vector Normalization for Noise Robust Speech Recognition,” *Speech Communication, Vol. 25, pp. 133-147*, 1998

[Wang *et al*. 2005] Hsin-min Wang, Berlin Chen, Jen-Wei Kuo and Shih-Sian Cheng, "MATBN: A Mandarin Chinese Broadcast News Corpus," *International Journal of Computational Linguistics & Chinese Language Processing, Vol. 10, No. 2*, 2005

[Wang *et al.* 2007] L. Wang, M.J.F. Gales, P.C. Woodland, “Unsupervised Training for Mandarin Broadcast News and Conversation Transcrption”, *in Proc. ICASSP* 2007

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

[Wessel *et al* 2001b] Frank Wessel and Hermann Ney ,“Unsupervised Training of Acoustic Models for Large Vocabulary Continuous Speech Recognition”, *in Proc. ASRU* 2001

[Wessel *et al* 2005] Frank Wessel and Hermann Ney ,“Unsupervised Training of Acoustic Models for Large Vocabulary Continuous Speech Recognition”, *IEEE Trans. SAP, Vol.13, No.1* 2005

[X. Li *et al.* 2005] Xinwei Li, Hui Jiang, “A Constrained Joint Optimization Method for Large Margin HMM Estimation”, *in Proc. ASRU* 2005

[X. Li *et al.* 2006] Xinwei Li, Hui Jiang, “Solving Large Margin Estimation of HMMs via Semidefinite Programming”, *in Proc. ICSLP* 2006

[Young 1994] S. R. Young, “Detecting Misrecognition and Out-of-vocabulary Words,” *in Proc. of ICASSP* 1995

[Young *et al.* 2006] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. C. Woodland (2006). *The HTK Book.* Version 3.4, 2006. http://htk.eng.cam.ac.uk/

[Zhang and Rudnicky 2001] R. Zhang and A. I. Rundicky, “Apply N-Best List Re-ranking to Acoustic Model Combinations of Boosting Training,” *in Proc. of ICSLP* 2004

[郭人瑋 2005] 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*

[陳燦輝 2006] 陳燦輝, “信心度評估於中文大詞彙連續語音辨識之研究”, *Master Thesis, NTNU, 2006*