國立臺灣大學電機資訊學院電信工程學研究所

碩士論文

Graduate Institute of Communication Engineering
College of Electrical Engineering and Computer Science
National Taiwan University

Master Thesis

最小音素錯誤訓練法及其改進方法在國語大字彙辨識上之評估與分析

Evaluation and Analysis of Minimum Phone Error Training and Its Modified Versions for Large Vocabulary Mandarin Speech Recognition

程永任

Cheng Yung-Jen

指導教授：李琳山 博士

Advisor: Lee Lin-Shan, Ph.D.

中華民國 97 年 6 月

June, 2008


國立臺灣大學
人格品德・愛國愛學

中文摘要

傳統的語音模型訓練以最大相似度(Maximum Likelihood, ML)來訓練聲學模型，雖然可以使正確的轉寫在訓練語料中有最大的事後機率，卻無法保證錯誤的聲學特徵(feature)不會產生更大的事後機率。鑑別式訓練(discriminative training)同時將可能的辨識結果與正確轉寫納入訓練，設法避免不正確的聲學特徵產生高於正確轉寫的事後機率。

本論文以最小音素錯誤訓練法(Minimum Phone Error, MPE)以及其改進方法為主軸，詳細介紹鑑別式訓練法的背景知識、理論基礎以及實驗結果。本論文可分為五個部份：

第一部份為鑑別式訓練的基礎理論，從貝氏風險(Bayes Risk)開始，介紹目前廣泛研究的若干種模型訓練法，包括最大相似度估測法、最大相互資訊(Maximum Mutual Information, MMI)估測法、全面風險法則估測(Overall Risk Criterion Estimation, ORCE)、最小分類錯誤(Minimum Classification Error, MCE)訓練法以及最小音素錯誤(Minimum Phone Error, MPE)訓練法，這些訓練法的目標函數都可以視為貝氏風險的延伸。

第二部份為本論文的實驗架構：包括師大的新聞語料庫；實驗的前端處理方式，梅爾倒頻譜係數(Mel-Frequency Cepstrum Coefficient, MFCC)；初始聲學模型的訓練，由 HTK 以最大相似度估測法訓練而成；詞典及語言模型的建立，以中央通訊社收集的文字語料由 SRILM 訓練而成；以及語音辨識工具，為台大語音實驗室的 TTK。基礎實驗為初始聲學模型的辨識結果。

第三部份為最小音素錯誤訓練法，先介紹目標函數最佳化的理論推導過程，求得模型參數的更新公式。再介紹模型參數的更新公式中，各項統計值在實作上的計算方法，其中包含正確度的定義，以及詞弧正確度和詞圖期望正確度的算法。實驗結果最小音素錯誤訓練法有約 2.4%字正確率的進步。

第四部份介紹最小音素錯誤訓練法的改進方法，包括最小音素音框錯誤

(Minimum Phone Frame Error, MPFE)訓練法、狀態層級最小貝氏風險(physical state level Minimum Bayes Risk, sMBR)訓練法和最小歧異度(Minimum Divergence, MD)訓練法，這些方法主要差異在於目標函數中正確度的定義。實驗結果包括最小音素錯誤訓練法的四種方法之中，除了最小歧異度訓練法之外的三種方法都可以在詞正確率以及字正確率上進步，其中又以最小音素錯誤訓練法在字正確率的表現最好，而詞正確率則是以最小音素音框錯誤訓練法表現最好。此外，本論文也在目標函數中正確度的定義做了更進一步的改進：在正確度中加入了錯誤處罰以及音素長度正規化，實驗結果這個正確度的改進版本會產生字正確率進步，而在詞正確率上退步的情形。

第五部份介紹基於詞弧期望正確度的資料選取方法，目標是篩選出較具有鑑別力的詞弧納入訓練，實驗在最小音素錯誤訓練法和最小音素音框錯誤訓練法的其中一種修改版本上，實驗結果顯示資料選取對於正確率的變化並沒有很大的影響，不過可以加快訓練的收斂速度。

目錄

圖目錄

表目錄

第1章　緒論

1.1　研究動機

在傳統的語音模型訓練中，模型參數的估測是由最大相似度估測法(Maximum Likelihood Estimation, MLE)求得，此方法的目標是讓正確轉寫(transcription)在訓練語料中產生最大的事後機率(posterior probability)，然而最大相似度估測法並未考慮到競爭字串(competing word sequence)，以致於在辨識的語料時，正確轉寫的聲學模型相似度(likelihood)未必高於競爭字串的聲學模型相似度，而造成辨識的錯誤。鑑別式訓練(discriminative training)的目的在於訓練過程中，加入對於競爭字串的考慮，目標是使正確轉寫的聲學模型相似度高於競爭字串的聲學模型相似度，將混淆的模型有效地分開，以達成提高辨識率的效果。

鑑別式訓練法在約二十年前首先由 IBM 提出的最大相互資訊 (Maximum Mutual Information, MMI) 估測法【1】開始，之後亦有最小分類錯誤(Minimum Classification Error, MCE)估測法【2】提出，都表現出比最大相似度估測法更好的成效，到了 2002 年劍橋大學又更進一步提出了最小音素錯誤(Minimum Phone Error, MPE)模型訓練法【3】，以降低音素錯誤率為目標，充份利用詞圖(word graph)資訊，並且找到了更有效率的參數最佳化方法，讓鑑別式訓練法在大字彙辨識上也有顯著的成效，因而最小音素錯誤模型訓練法成為目前鑑別式聲學模型訓練法中最具代表性的方法之一。

在最小音素錯誤模型訓練法之後，又提出了許多根據此方法改進而來的鑑別式聲學模型訓練法，如最小音素音框錯誤(Minimum Phone Frame Error, MPFE)模型訓練法【4】，是在錯誤率的計算上，使用比音素(phone)更小的音框(frame)為單位。以及最小歧異度(Minimum Divergence, MD)模型訓練法【5】，是在計算錯誤率時針對不同的比對錯誤給與不同的扣分因素(penalty)，這些方法都能讓鑑別式聲學模型訓練法的辨識率有更近一步的提升。

1.2 統計式語音辨識

語音辨識的直覺上的做法可以理解成：「找出聽起來最像、最可能的句子」，而相像、可能概念的量化，可以用機率來表示，這就是統計式語音辨識的基本概念。因此，「找出聽起來最像、最可能的句子」就可理解成「找出機率最高的句子」。若 O 是給定的觀測語句(observation)，要從所有文句 W_h 中找出機率最大的文句 s 可表示成：

$$s = \arg\max_{u \in W_h} P(u \mid O) \tag{1.1}$$

其中 u 為所有文句 W_h 中的某一句，$P(u \mid O)$ 代表在 O 發生時，文句 u 的事後機率。進一步使用貝氏定理(Bayes' Theorem)將 $P(u \mid O)$ 展開可以得到：

$$P(u \mid O) = \frac{P(O \mid u)P(u)}{P(O)} \tag{1.2}$$

$P(O \mid u)$ 表示給定文句 u 其聲音是語句 O 的相似度或機率，通常使用機率分佈(probability distribution)來呈現，由於這個機率分佈主要用來決定聲學特徵的機率，故稱為聲學模型(acoustic model)，而此機率分佈中的參數便稱為聲學模型參數；$P(u)$ 則是文句 u 的事前機率，表示語言中出現 u 的機率，同樣使用機率分佈來呈現，由於這個機率用來決定語言機率，故稱為語言模型(language model)。$P(O)$ 則是指觀測語句 O 的出現機率，由於在(1.2)中 $P(O)$ 與 u 無關，因此拿掉此項對於尋找機率最大的文句 u 並無影響，因此(1.1)可以簡化為：

$$s = \arg\max_{u \in W_h} P(O \mid u)P(u) \tag{1.3}$$

1.2.1 聲學模型

聲學模型的主要功能，便是對於觀測語句，能夠針對不同的發音可能，給與相對應的機率或相似度，即(1.3)中的 $P(O \mid u)$，一般使用機率密度函數(probability density function)來近似。而聲學模型訓練，就是在訓練語料中給定的觀測語句，以

及其對應的正確轉寫，在訓練過程中調整聲學模型參數，使得正確轉寫和其對應的發音產生最大的事後機率，簡易流程如圖 1.1。

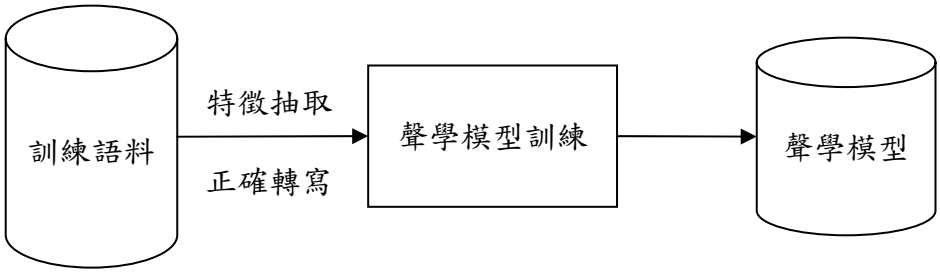


圖 1.1　聲學模型訓練流程

本論文中，使用連續密度隱藏式馬可夫模型(Continuous Density Hidden Markov Models, CDHMM)【6】做為聲學模型，模型的結構如圖 1.2 所示，每一個模型都由連續的數個狀態(state)，以及狀態間的轉移(transition)構成，每一個轉移均有其轉移機率(transition probability)，一般語音的聲學模型，狀態轉移只允許停留在原狀態或跳至鄰接的下一狀態，而其中每一個狀態對一音框的聲學特徵觀測機率(observation probability)，則使用連續的高斯混合模型(Gaussian Mixture Model, GMM)來決定。

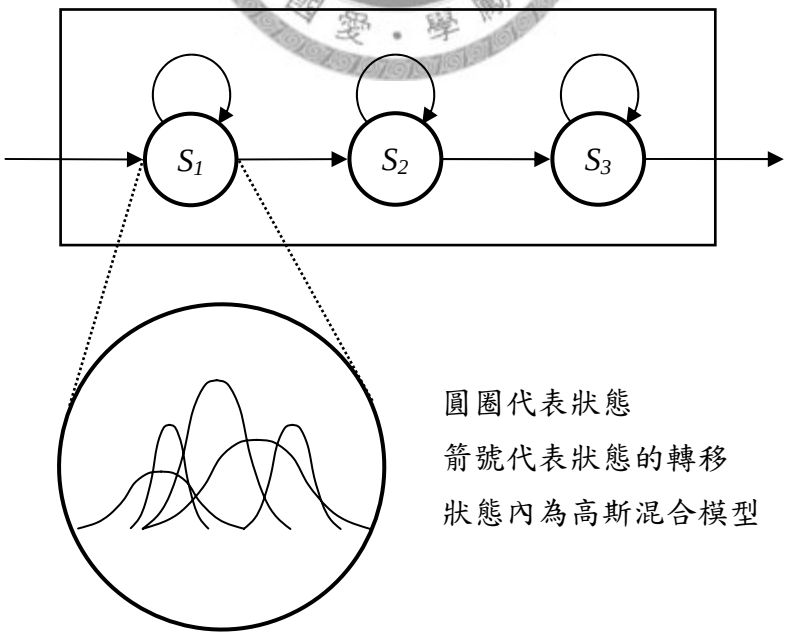


圖 1.2　連續密度隱藏式馬可夫模型示意圖

1.2.2 語言模型

語言模型的主要功能，便是針對不同的文句，給與一個該文句在語言中的使用機率，即(1.3)中的 $P(u)$，若文句 u 由 N 個詞 $w_1, w_2, ..., w_N$ 組成，則 $P(u) = P(w_1, w_2, ..., w_N)$，為 $w_1, w_2, ..., w_N$ 的聯合機率(joint probability)。由於語言機率是離散的分佈，故語言模型的建立不使用機率密度函式來近似，而是對個別的機率作直接估測。由於需要估測的參數量很大，存在資料稀疏的問題，故將聯合機率 $P(w_1, w_2, ..., w_N)$ 展開成條件機率的連乘 $\prod_{k=1}^{N} P(w_k \mid w_1, w_2, ..., w_{k-1})$，再使用 $n-1$ 階馬可夫假設(n-1 order Markov Assumption)來簡化，稱為 n 連(n-gram)語言模型，可表示為：

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

$w_1, w_2, ..., w_N$ 為歷史詞序列(history word sequences)，條件機率 $P(w_k \mid w_1, w_2, ..., w_{k-1})$ 可以解釋為根據歷史詞預測下一個詞為 w_k 的機率，故建立 n 連語言模型即是為每一種詞序列建立各自的條件機率分佈。實作上，常見的有使用一階馬可夫假設的詞雙連(bigram)語言模型，可表示為：

$$P(w_k \mid w_{k-n+1}, w_{k-n+2}, ..., w_{k-1}) \approx P(w_k \mid w_{k-1}) \tag{1.5}$$

以及使用二階馬可夫假設的詞三連(trigram)語言模型，可表示為：

$$P(w_k \mid w_{k-n+1}, w_{k-n+2}, ..., w_{k-1}) \approx P(w_k \mid w_{k-2}, w_{k-1}) \tag{1.6}$$

n 連語言模型的機率常以最大相似度估測法來做測，配合使用語言模型平滑技術，對無法由訓練語料估測的詞序機率加以平滑化。

1.3 研究主題與主要成果

本論文主要探討最小音素錯誤訓練法、最小音素音框錯誤訓練法、狀態層級最小貝氏風險訓練法，以及最小歧異度訓練法，這四種方法在中文大字彙辨識上的成效。以及基於這四種方法，針對詞弧正確度做進一步改進的方法之比較。實

驗結果發現這些方法在詞正確率與字正確率會有不一致的變化，其中最小音素錯誤訓練法是偏好字正確率的方法，字正確率的表現較其它方法好；最小音素音框錯誤訓練法和狀態層級最小貝氏風險訓練法則是偏好詞正確率的方法，詞正確率的表現較其它方法好。而將詞弧正確度做進一步改進的方法實行在這兩種偏好詞正確率的方法上，會產生詞正確率下降而字正確率上升的變化，表示將詞弧正確度做進一步改進之後，這兩種原本偏好詞正確率的方法會轉變成偏好字正確率的方法。

另外，本論文也實驗詞弧篩選的資料選取方法在最小音素錯誤訓練和將詞弧正確度改進後的最小音素音框錯誤上，實驗結果顯示資料選取對於正確率的變化並沒有很大的影響，不過可以加快訓練的收斂速度。

1.4 論文架構

本論文第二章將介紹鑑別式訓練法則，從貝氏風險出發，回顧並介紹鑑別式訓練法的發展流程。

第三章將介紹本論文的實驗系統以及基礎實驗設定和實驗結果。

第四章將介紹最小音素錯誤訓練法的理論基礎以及實作的方法。

第五章將介紹最小音素音框錯誤訓練法、狀態層級最小貝氏風險訓練法、最小歧異度訓練法，以及這三種方法進一步的修改版本。

第六章將介紹基於詞弧期望正確度的資料選取方法，實驗在最小音素錯誤訓練法和第五章中最小音素音框錯誤訓練法的其中一種修改版本上。

第七章會提出總結以及未來展望。

第2章　背景知識

2.1　鑑別式訓練法則

鑑別式訓練法的主要概念，在於訓練模型時，不以訓練語料相似度的最大化為目標，而是以分類錯誤的最小化為目標，進而增進辨識率。傳統的聲學模型訓練，以最大相似度估測法為原則，在訓練時調整模型參數的目標是使得正確的語音聲學特徵在此聲學模型的相似度變大，但是這種訓練方式沒有考慮到模型間彼此的關係，所以在使正確的語音聲學特徵在對應的模型上的相似度增加時，可能同時使不正確的語音聲學特徵在此聲學模型的相似度也變大，造成辨識上的混淆，舉例如圖 2.1，(a)表示一個正確轉寫為 u 的觀測語句 O_u，在模型 M 上可以得到一相似度 $P(O_u|\mathrm{M})$，在訓練時以相似度的最大化為目標的過程就如(b)所示，訓練時會調整模型 A 使得正確轉寫為 A 的觀測語句 O_A 落在模型 A 上的相似度 $P(O_A|\mathrm{A})$ 增加，

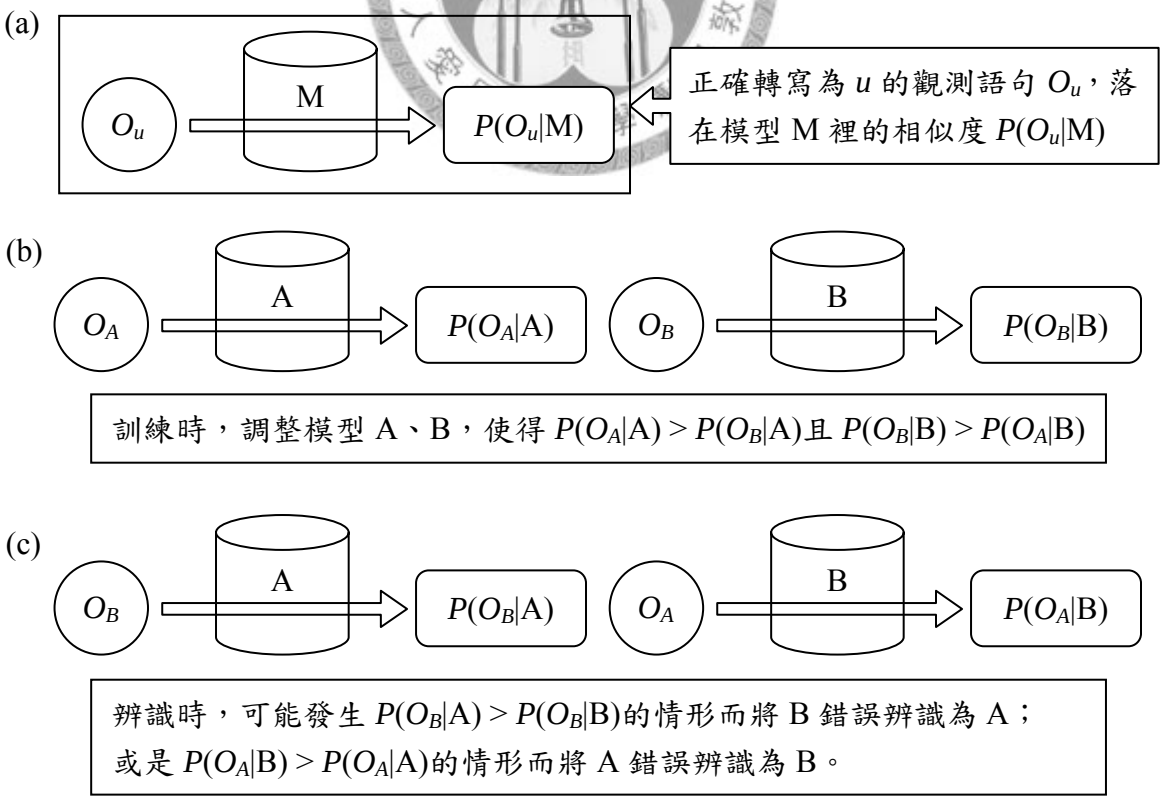


圖 2.1　最大相似度估測法造成混淆的情形

大於其它正確轉寫不為 A 的觀測語句落在模型 A 上的相似度，同樣的也調整模型 B 使得 O_B 落在模型 B 上的相似度 $P(O_A|\mathrm{B})$增加，大於其它正確轉寫不為 B 的觀測語句落在模型 B 上的相似度，這樣的訓練自然會有 $P(O_A|\mathrm{A}) > P(O_B|\mathrm{A})$且 $P(O_B|\mathrm{B}) > P(O_A|\mathrm{B})$的結果，然而在(c)的辨識時，辨識的準則是測試觀測語句落在每個模型裡的相似度，再挑選出落在哪個模型裡的相似度最大，而採定為辨識結果，這與相似度最大化的訓練原則並不一致，在(b)中雖然訓練結果會使正確轉譯為 A 的觀測語句 O_A 落在模型 A 的相似度一定大於 O_B 落在模型 A 的相似度，卻無法確定確 O_A 落在模型 B 的相似度是否會更大，意即發生 $P(O_B|\mathrm{B}) > P(O_A|\mathrm{B}) > P(O_A|\mathrm{A}) > P(O_B|\mathrm{A})$的情況，如此一來雖然 B 可以正確辨識為 B，A 卻會被錯誤辨識為 B；反之發生 $P(O_A|\mathrm{A}) > P(O_B|\mathrm{A}) > P(O_B|\mathrm{B}) > P(O_A|\mathrm{B})$的情況一樣會造成辨識錯誤，然而這兩種情況皆不違反在(b)中的訓練原則。

鑑別式訓練法便是針對這個缺點改進，企圖在訓練模型時同時考慮正確與不正確的語音聲學特徵，使得正確的語音聲學特徵在其聲學模型上的相似度可以大於不正確的語音聲學特徵在此聲學模型上的相似度，意即在圖 2.1 中的(b)訓練目標是 $P(O_A|\mathrm{A}) > P(O_B|\mathrm{A})$且 $P(O_B|\mathrm{B}) > P(O_A|\mathrm{B})$。

以下幾節將從貝氏風險開始，以鑑別式訓練方法的演進，介紹數個廣泛研究過的模型訓練法，包括其目標函數(objective function)及物理意義。

2.2 貝氏風險(Bayes Risk)

如果將語音辨識視為一個分類的行為，即對一語句 O_r 分類至一文句 s ，而辨識所做的分類未必正確，因此存在一個分類錯誤的風險，用一個函數 $R(s|O_r)$ 代表將語句 O_r 分類至文句 s 的風險，這個風險函數可以定義如下：

$$R(s|O_r) = \sum_{u \in W_h} P(u|O_r) L(s,u) \tag{2.1}$$

O_r 表示觀測語句的特徵向量，W_h 表示所有可能文句之集合，$P(u|O_r)$ 表示給定觀測語句的特徵向量 O_r 時，文句 u 的事後機率，$L(s,u)$ 為一減損函數(loss function)，

表示文句 s 和 u 之問的差異造成的損失；因此 $R(s|O_r)$ 就是將 O_r 辨識為 s 的損失期望值。在實作語音辨識的解碼上，會以風險值最小化的 s^* 作為辨識結果：

$$s^* = \arg\min_s R(s|O_r) = \arg\min_s \sum_{u \in W_h} P(u|O_r) L(s,u) \tag{2.2}$$

這個風險的最小值即為貝氏風險 R_{Bayes} ：

$$R_{Bayes} = \min_s R(s|O_r) = \min_s \sum_{u \in W_h} P(u|O_r) L(s,u) \tag{2.3}$$

目前許多辨識的解碼方法都是以最小貝氏風險為原則，如最大事後機率解碼(Maximum a posterior decoding, MAP decoding)【7】、最小貝氏風險(Minimum Bayesian Risk decoding, MBR decoding)【8】，以及最小詞錯誤解碼(word error minimization decoding)【9】，都是這個方法的應用。

至於將貝氏風險運用在模型訓練上，則是把風險函數作為目標函數：

$$\begin{aligned}(\lambda^*, \Gamma^*) &= \arg\min_{\lambda,\Gamma} \sum_r R(s_r|O_r) \\ &= \arg\min_{\lambda,\Gamma} \sum_r \sum_u P_{\lambda,\Gamma}(u|O_r) L(s_r,u)\end{aligned} \tag{2.4}$$

其中 λ, Γ 分別是聲學模型與語言模型的參數集，s_r 是 O_r 的正確轉寫，r 是訓練語句(utterance)的索引，$P_{\lambda,\Gamma}(u|O_r)$ 表示 u 基於聲學模型與語言模型的事後機率。模型訓練的目標可以視為將訓練語句的整體風險總合最小化。

以下將介紹數種聲學模型訓練法，其目標函數皆由貝氏風險而來，但也各有所不同，以下會介紹其目標函數與貝氏風險之間的關係。

2.3 最大相似度(Maximum Likelihood, ML)

最大相似度估測法的目標函數是將貝氏風險中的減損函數定義為零壹函數(zero-one function)，即是將(2.4)中的 $L(s_r,u)$ 定義如下：

$$L(s_r,u) = \begin{cases} 0 & ,u = s_r \\ 1 & ,u \neq s_r \end{cases} \tag{2.5}$$

也就是當文句 s_r 與 u 相同時損失為 0，否則損失為 1。則(2.4)可推導為：

$$\begin{aligned}\left(\lambda^*,\Gamma^*\right) &= \arg\min_{\lambda,\Gamma}\sum_r\sum_u P_{\lambda,\Gamma}\left(u\,|\,O_r\right)L\left(s_r,u\right)\\ &= \arg\min_{\lambda,\Gamma}\sum_r\sum_{u\neq s_r} P_{\lambda,\Gamma}\left(u\,|\,O_r\right)\\ &= \arg\min_{\lambda,\Gamma}\sum_r\left(1-P_{\lambda,\Gamma}\left(s_r\,|\,O_r\right)\right)\end{aligned} \tag{2.6}$$

倘若將(2.6)繼續代換如下：

$$\begin{aligned}\left(\lambda^*,\Gamma^*\right) &= \arg\min_{\lambda,\Gamma}\sum_r\left(1-P_{\lambda,\Gamma}\left(s_r\,|\,O_r\right)\right)\\ &= \arg\max_{\lambda,\Gamma}\sum_r P_{\lambda,\Gamma}\left(s_r\,|\,O_r\right)\end{aligned} \tag{2.7}$$

(2.7)的結果就是所謂的最大事後機率(maximum a posterior, MAP)。

如果再將(2.6)的結果套用詹氏不等式(Jensen's inequality)：

$$1-P_{\lambda,\Gamma}\left(s_r\,|\,O_r\right)\le -\log P_{\lambda,\Gamma}\left(s_r\,|\,O_r\right) \tag{2.8}$$

(2.6)就會進一步變成：

$$\begin{aligned}\left(\lambda^*,\Gamma^*\right) &= \arg\min_{\lambda,\Gamma}\sum_r\left(1-P_{\lambda,\Gamma}\left(s_r\,|\,O_r\right)\right)\\ &= \arg\min_{\lambda,\Gamma}\sum_r\left(-\log P_{\lambda,\Gamma}\left(s_r\,|\,O_r\right)\right)\\ &= \arg\max_{\lambda,\Gamma}\sum_r\log P_{\lambda,\Gamma}\left(s_r\,|\,O_r\right)\end{aligned} \tag{2.9}$$

之後再使用貝氏定理(Bayes' theorem)推導如下：

$$\begin{aligned}\left(\lambda^*,\Gamma^*\right) &= \arg\max_{\lambda,\Gamma}\sum_r\log P_{\lambda,\Gamma}\left(s_r\,|\,O_r\right)\\ &= \arg\max_{\lambda,\Gamma}\sum_r\log\frac{P_\lambda\left(O_r\,|\,s_r\right)P_\Gamma\left(s_r\right)}{P\left(O_r\right)}\end{aligned} \tag{2.10}$$

由於假設所有的O_r為均勻分布(uniform distribution)，因此在(2.10)中又可以省略此項：

$$\left(\lambda^*,\Gamma^*\right) = \arg\max_{\lambda,\Gamma}\sum_r\log P_\lambda\left(O_r\,|\,s_r\right)P_\Gamma\left(O_r\right) \tag{2.11}$$

最後因為最大相似度估測法只訓練聲學模型，因此只保留與聲學模型λ有關的項目，於是就成為最大相似度估測法的目標函數$F_{ML}\left(\lambda\right)$：

$$F_{ML}(\lambda)=\sum_{r}\log P_{\lambda}(O_r\mid s_r) \tag{2.12}$$

而最大相似度聲學模型參數的估測就是其目標函數的最大化：

$$\lambda_{ML}=\arg\max_{\lambda}F_{ML}(\lambda)=\arg\max_{\lambda}\sum_{r}\log P_{\lambda}(O_r\mid s_r) \tag{2.13}$$

2.4 最大相互資訊(Maximum Mutual Information, MMI)

對於(2.10)的結果，如果對分母項$P(O_r)$使用貝氏定理展開：

$$P(O_r)=\sum_{u\in W_h}P_{\lambda}(O_r\mid u)P_{\Gamma}(u) \tag{2.14}$$

其中W_h表示所有可能的辨識結果，u為可能辨識結果的其中一句，則(2.10)可以進一步推導為：

$$(\lambda^*,\Gamma^*)=\arg\max_{\lambda,\Gamma}\sum_{r}\log\frac{P_{\lambda}(O_r\mid s_r)P_{\Gamma}(s_r)}{\sum_{u\in W_h}P_{\lambda}(O_r\mid u)P_{\Gamma}(u)} \tag{2.15}$$

這裡如(2.12)同樣的省略掉$P_{\Gamma}(s_r)$後，就成為最大相互資訊估測法的目標函數$F_{MMI}(\lambda)$：

$$F_{MMI}(\lambda)=\sum_{r}\log\frac{P_{\lambda}(O_r\mid s_r)}{\sum_{u\in W_h}P_{\lambda}(O_r\mid u)P_{\Gamma}(u)} \tag{2.16}$$

而最大相互資訊聲學模型參數的估測就是其目標函數的最大化：

$$\lambda_{ML}=\arg\max_{\lambda}F_{MMI}(\lambda)=\arg\max_{\lambda}\sum_{r}\log\frac{P_{\lambda}(O_r\mid s_r)}{\sum_{u\in W_h}P_{\lambda}(O_r\mid u)P_{\Gamma}(u)} \tag{2.17}$$

而在(2.17)中由於：

$$\log\frac{P_{\lambda}(O_r\mid s_r)}{\sum_{u\in W_h}P_{\lambda}(O_r\mid u)P_{\Gamma}(u)}=\log\frac{P_{\lambda}(O_r\mid s_r)}{P(O_r)}=\log\frac{P_{\lambda}(O_r,s_r)}{P(O_r)P(s_r)} \tag{2.18}$$

表示觀測語句O_r與正確轉寫s_r的相互資訊(mutual information)，是故最大相互資訊估測法就是在最大化觀測語句O_r與正確轉寫s_r的相互資訊。

將最大相互資訊用於聲學模型的訓練，最早是由 IBM 在 1986 年提出【1】，在辨識 2000 個獨立的詞的辨識實驗中，比最大相似度估測法降低了 18%的詞錯誤率。布氏(Brown)在 1987 年時使用最大相互資訊估測法訓練連續隱藏式馬可夫模型【10】，可以產生 18%的相對進步率，由於最佳化的過程十分複雜，所以使用了斜率遞減法(gradient descent)來求解。之後在 1995 年，諾氏(Normandin)更將延伸式波氏重估(extended Baum-Welch re-estimation, EBW)【11】演算法用於連續隱藏式馬可夫模型的參數最佳化上【12】。之後范氏(Valtchev)等人則將最大相互資訊估測法應用到大字彙連續語音辨識(Large Vocabulary Continuous Speech Recognition, LVCSR)上【13】，在 64000 個詞彙的實驗中，可以產生約 5~10%的相對進步率，此時的實驗已經使用語音辨識產生的詞圖作為可能的辨識結果((2.14)中的W_h)的近似，進而利用(2.14)近似出觀測語句的事前機率$P(O_r)$。

2002 年由伍氏(Woodland)等人於劍橋大學提出 I 平滑(I-Smoothing)技術【14】，由訓練語料中使用最大相似度估測之統計資訊，作為待測模型的事前機率分佈，來加強最大相互資訊估測模型的強健性。

2.5 全面風險法則估測(Overall Risk Criterion Estimation, ORCE)

在上述的最大相似度及最大相互資訊中，皆使用零壹函數作為減損函數，零壹函數可以視為句錯誤率的計算，但在語音辨識結果的評量上，英文習慣使用詞錯誤率(word error rate, WER)，中文則使用字錯誤率(character error rate, CER)較為合理，兩種評量方式皆與零壹函數的錯誤率計算方式相左，因此，以零壹函數的減損函數為最小化的目標並不一定帶來較低的辨識錯誤率，為了克服這個問題，在全面風險法則估測中便提出了使用編輯距離(Levenshtein distance)取代零壹函數為減損函數的作法，編輯距離的定義如圖 2.2 及(2.19)式所示【15】：

```
int LevenshteinDistance(char s_r[1..m], char u[1..n])
  // d is a table with m+1 rows and n+1 columns
  declare int d[0..m, 0..n]

  for i from 0 to m
    d[i, 0] := i
  for j from 0 to n
    d[0, j] := j
  for i from 1 to m
    for j from 1 to n
      if s_r[i] = u[j] then cost := 0
      else cost := 1
      d[i, j] := minimum(
                          d[i-1, j] + 1,        // deletion
                          d[i, j-1] + 1,        // insertion
                          d[i-1, j-1] + cost    // substitution
                         )
  return d[m, n]
```

圖 2.2 編輯距離之計算方式

$$L(s_r,u)=L_{Edit}(s_r,u)=d[m,n] \tag{2.19}$$

編輯距離為計算文句 u 與正確轉寫 s_r 之間的距離，將文句中的字詞一一比對(圖 2.2 中之 u 的 1 到 m 個字詞及 s_r 的 1 到 n 個字詞)，過程中可以統計出字詞的遺失(deletion)、插入(insertion)、取代(substitution)，用來計算辨識正確率。將(2.4)中的 $P_{\lambda,\Gamma}(u|O_r)$ 同(2.10)般展開，再將 $P(O_r)$ 如同(2.14)般展開，並將編輯距離作為減損函數，便可推導出全面風險法則估測：

$$(\lambda^*,\Gamma^*)=\arg\min_{\lambda,\Gamma}\sum_r\frac{\sum_{u\in W_h}P_\lambda(O_r|s_r)P_\Gamma(s_r)L_{Edit}(s_r,u)}{\sum_{u\in W_h}P_\lambda(O_r|u)P_\Gamma(u)} \tag{2.20}$$

全面風險法則估測在 2002 由凱氏(Kaiser)等人提出以編輯距離取代零壹函數作為減損函數，以 N 最佳路徑(N-best)作為所有辨識可能的近似，並使用延伸式波氏重估演算法來實行模型最佳化【16】，在 TIMIT 的音素辨識實驗中，約可降低 20.8%的詞錯誤率。

2.6 最小分類錯誤(Minimum Classification Error, MCE)

在(2.1)中，一般均使用零壹函數作為減損函數來計算貝氏風險，然而最小分類錯誤訓練法則重新定義風險函數，並套用至 S 形函數(Sigmoid function)：

$$R(s_r \mid O_r) = \frac{1}{1+\exp\{-\gamma\, d_{\lambda,\Gamma}(O_r \mid s_r) + \theta\}} \tag{2.21}$$

其中 $d_{\lambda,\Gamma}(O_r \mid s_r)$ 為錯誤分類計量(misclassification measure)，定義為：

$$d_{\lambda,\Gamma}(O_r \mid s_r) = \log\left(\frac{\frac{1}{M-1}\sum_{u\in W_h,\, u\neq s_r} P_\lambda(O_r \mid u)^\eta\, P_\Gamma(u)^\eta}{P_\lambda(O_r \mid s_r)^\eta\, P_\Gamma(s_r)^\eta}\right)^{1/\eta} \tag{2.22}$$

其中 M 表示所有辨識可能 W_h 的文句數目，γ、θ 為 S 形函數的控制參數，$u \in W_h, u \neq s_r$ 表示 s_r 的競爭文句，η 為比例控制參數，可以決定競爭文句中不同機率文句的影響程度差異，η 越大會使得機率高的文句具有越大的影響，而 S 形函數的功能是將錯誤分類計量的數值投影到 0 與 1 之間。如果定義分類條件相似度函數(class conditional likelihood functions) $g_v(O;\lambda,\Gamma)$ 如下：

$$g_v(O;\lambda,\Gamma) = \log P_\lambda(O \mid v)\, P_\Gamma(v) \tag{2.23}$$

則(2.22)可以代換如下：

$$d_{\lambda,\Gamma}(O_r \mid s_r) = -g_{s_r}(O_r;\lambda,\Gamma) + \log\left(\frac{1}{M-1}\sum_{u\in W_h,\, u\neq s_r} \exp\{g_u(O_r;\lambda,\Gamma)\eta\}\right)^{1/\eta} \tag{2.24}$$

由(2.24)可以將錯誤分類計量理解為競爭文句的平均相似度與正確轉寫的相似度的比值，越小表示錯誤越低，會計算出越小的錯誤分類計量。因此最小分類錯誤的目標函數 $F_{MCE}(\lambda,\Gamma)$ 可以表示如下：

$$F_{MCE}(\lambda,\Gamma) = \sum_r \left(1 + e^{\theta}\left(\frac{P_\lambda(O_r \mid s_r)^\eta\, P_\Gamma(s_r)^\eta}{\frac{1}{M-1}\sum_{u\in W_h,\, u\neq s_r} P_\lambda(O_r \mid u)^\eta\, P_\Gamma(u)^\eta}\right)^{\gamma/\eta}\right)^{-1} \tag{2.25}$$

而使用最小分類錯誤訓練聲學模型時，就是在最大化其目標函數，因為目標只有訓練聲學模型，因此將語言模型視為已知：

$$\begin{aligned}\lambda_{MCE} &= \arg\max_{\lambda} F_{MCE}(\lambda) \\ &= \arg\max_{\lambda} \sum_{r} \left(1 + e^{\theta \left(\frac{P_{\lambda}(O_r \mid s_r)^{\eta} P_{\Gamma}(s_r)^{\eta}}{\frac{1}{M-1} \sum_{u \in W_h,\, u \neq s_r} P_{\lambda}(O_r \mid u)^{\eta} P_{\Gamma}(u)^{\eta}} \right)^{\gamma/\eta} } \right)^{-1}\end{aligned} \tag{2.26}$$

最小分類錯誤在 1992 年由莊氏(Juang)等人提出【17】【2】，使用一般化機率遞減法(generalized probability descent, GPD)來進行最佳化。有許多成果在小字彙訓練上提出，如周氏(Chou)在 TI Digit String 的實驗可以降低25%的字串錯誤率【18】，薛氏(Sual)等人在小型及中型詞彙辨識上也能產生 10%的相對進步率【19】。而在2000 年，舒氏(Schlüter)則將最小分類錯誤訓練法應用到大字彙連續音辨識上【20】，利用詞圖作為競爭文句的近似。

2.7 最小音素錯誤(Minimum Phone Error, MPE)

最小音素錯誤模型訓練法在 2002 年由劍橋大學的波氏(Povey)等人提出，相較於最大相互資訊估測法是將正確轉寫的事後機率最大化，最小音素錯誤模型訓練法是將訓練語料的音素正確度(Raw Pone Accuracy)期望值最大化【3】，將(2.4)中的減損函數改為音素的正確率，將原本對減損函數的最小化改為對音素正確率的極大化，可得如下：

$$\begin{aligned}(\lambda^{*}, \Gamma^{*}) &= \arg\max_{\lambda,\Gamma} \sum_{r} \sum_{u} P_{\lambda,\Gamma}(u \mid O_r)\, Acc(s_r, u) \\ &= \arg\max_{\lambda,\Gamma} \sum_{r} \frac{\sum_{u \in W_h} P_{\lambda}(O_r \mid u) P_{\Gamma}(u)\, Acc(s_r, u)}{\sum_{u \in W_h} P_{\lambda}(O_r \mid u) P_{\Gamma}(u)}\end{aligned} \tag{2.27}$$

是故最小音素錯誤訓練法的目標函數 $F_{MPE}(\lambda)$ 為：

$$F_{MPE}(\lambda)=\sum_{r}\frac{\sum_{u\in W_h}P_{\lambda}(O_r\mid u)P_{\Gamma}(u)Acc(s_r,u)}{\sum_{u\in W_h}P_{\lambda}(O_r\mid u)P_{\Gamma}(u)} \tag{2.28}$$

此目標函數可視為音素正確度的期望值，也就是將辨識可能 u 的音素正確度 $Acc(s_r,u)$ 乘上 u 的事後機率 $P_{\lambda,\Gamma}(u\mid O_r)$ 做為權重再累加起來。實作上以詞圖來近似所有的辨識可能，如(2.14)。

在數學式上，最小音素錯誤模型訓練法類似於全面風險法則估測的一種變化形，然而在實作上，全面風險法則估測一般使用 N 最佳路徑作為辨識可能的近似，並以詞錯誤率的最小化為目標；而最小音素錯誤模型訓練法則使用詞圖作為辨識可能的近似，並以音素正確率的最大化為目標，此外還有引進待測模型的事前分布(prior distribution)來增加模型的強健性。

而在實作上，最小音素錯誤模型訓練法類似於最大相互資訊估測法的一種變化形，其中差別在於最大相互資訊估測法只將正確轉寫當成分子詞圖(numerator lattices)，將整個詞圖當成分母詞圖(denominator lattices)；最小音素錯誤模型訓練法則會統計整個詞圖的音素正確度平均值，將正確度高於平均值的辨識可能當成分子詞圖，將正確度低於平均值的辨識可能當成分母詞圖。

2.8 綜合各種訓練法之目標函數推導流程

綜合本章節之前的討論及推導，可知各種模型訓練的目標函數概念皆源自於貝氏風險的最小化，經由使用不同的減損函數定義、事前機率的假設以及函式推導的近似方式，發展出各式的模型估測法，如圖 2.3 所示。

2.9 本章結論

本章回顧了鑑別式訓練法發展的歷程，從貝氏風險的觀念開始，介紹了最大相似度估測法，和諸多著名的鑑別式訓練法：最大相互資訊、全面風險法則估測、最小分類錯誤，以及最小音素錯誤，最後再以這些訓練法目標函數的差異流程做為結束。

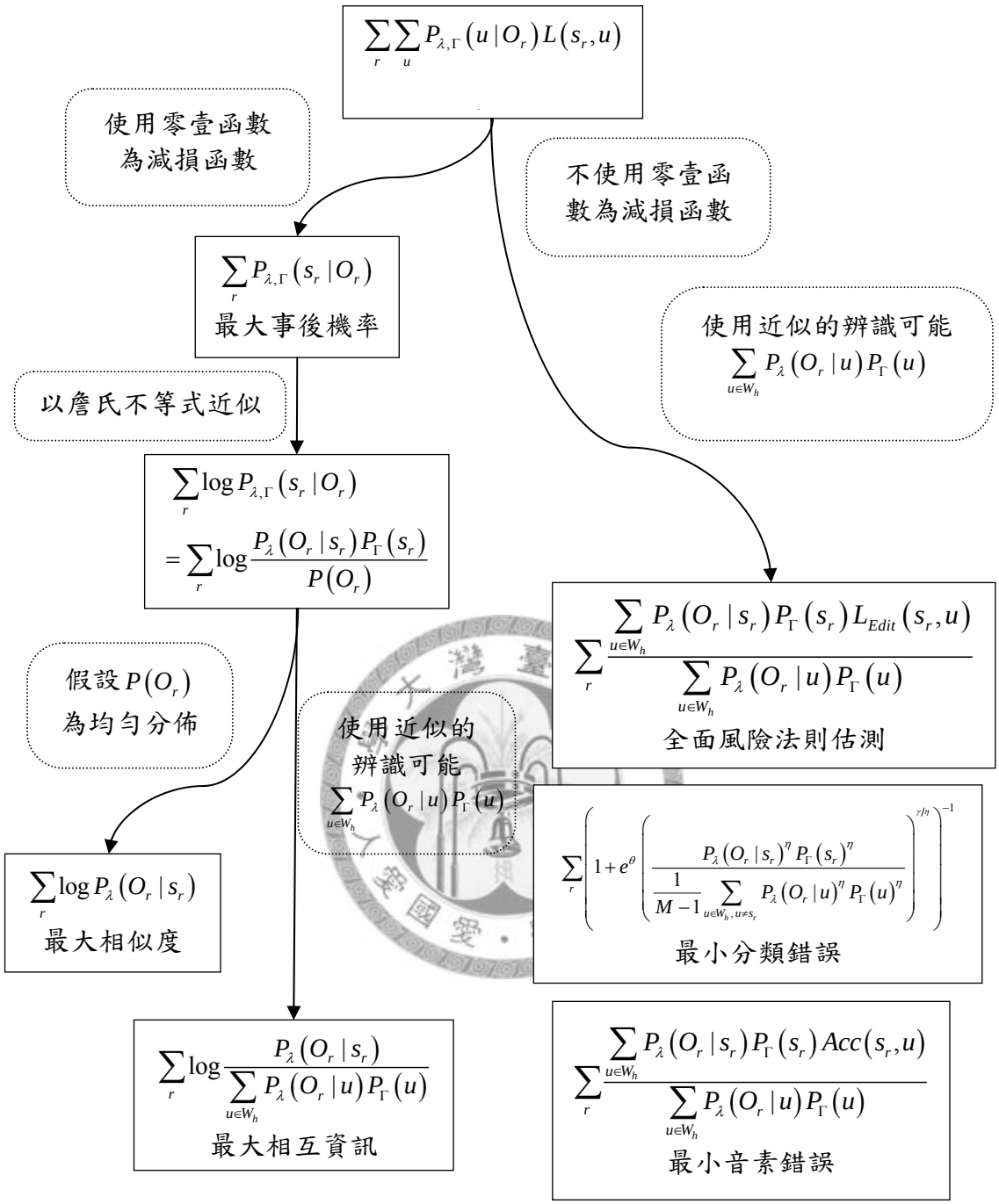


圖 2.3　各種目標函數之推導流程

第3章　實驗基礎架構及語料庫

本章將介紹本論文的實驗架構，包括公視新聞語料庫、語音特徵前端處理、詞典和語言模型的建立、初始聲學模型的訓練、以及語音辨識工具。論文中以最大相似度估測的結果作為基礎實驗。

3.1　實驗語料

本論文使用的語料為 MATBN 電視新聞語料【21】，由中央研究院資訊所口語小組耗時三年與公共電視台合作錄製完成。每天一個小時的公視晚間新聞深度報導，收了 200 天的電視新聞語料，其中包含 2001 年的新聞 30 小時、2002 年的 146 小時及 2003 年的 24 小時。本論文選擇採訪記者語料作為實驗語料，其中包含 25.5 小時的訓練集(training set)共 5774 則，用來訓練聲學模型；1.5 小時的評估集(test set)共 292 則，作辨識評估之用，如表 3.1。

其中訓練集的語料都經由人工切割為一句一句的語音檔，共 34672 句，每一句都有詞和音素的時間標記(alignment)。

語者性別	訓練集		評估集		語料重疊
	時間(秒)	人數(人)	時間(秒)	人數(人)	人數(人)
男性	46001	≤66	1301	9	9
女性	46007	≤111	3914	≤23	≥13

表 3.1　訓練集與評估集的語料資訊

3.2　訓練與辨識系統

本論文的基礎實驗系統可分為三個部份，各由不同工具完成：聲學模型訓練工具為劍橋大學的 HTK【22】；語言模型訓練的工具為史丹佛大學的 SRILM【23】；語音辨識工具則為台大語音實驗室的 TTK【24】。

3.2.1 前端處理

本論文使用梅爾倒頻譜係數(Mel-Frequency Cepstrum Coefficient, MFCC)【25】作為語音訊號的特徵參數。特徵抽取流程如圖 3.1 所示，將語音資料切割成一連串部份重疊的音框，每個音框中抽出 13 維的梅爾倒頻譜係數特徵，再加上其一階與二階的時間軸導數(time derivatives)所形成的 39 維聲學特徵向量所組成。其中 13 維的梅爾倒頻譜係數是由 18 個梅爾頻譜濾波器組(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，在實驗系統中，亦使用倒頻譜平均消去法(Ceptral Mean Subtraction, CMS)【26】。

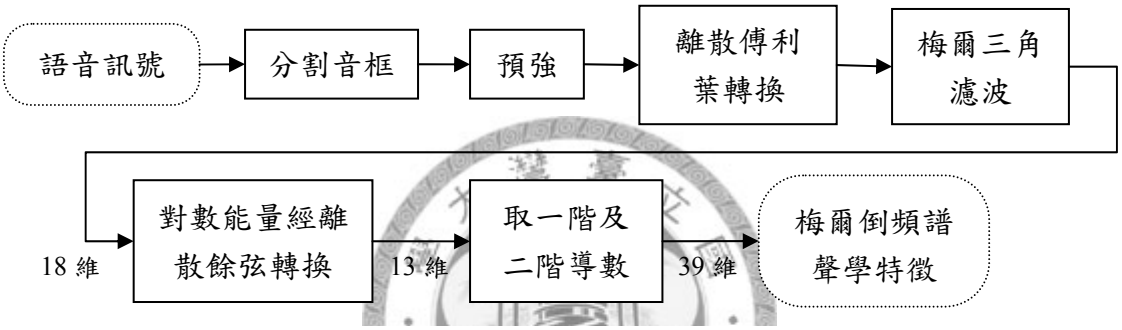


圖 3.1 MFCC 特徵抽取流程

3.2.2 聲學模型設定

本論文聲學模型的音素設定使用右相關聲韻母(Right-Context Dependent INITIAL/FINAL, RCDIF)模型，共包括 22 個聲母(INITAIL)隨右接韻母的起始音素不同而展開成的 112 個右相關聲母模型、38 個韻母(FIANL)模型與一個靜音(Silence)模型，總共 151 個右相關聲韻母音素。詳細模型列表見附錄 A。

聲學模型採用連續密度隱藏式馬可夫模型，共 151 個聲學模型，每個模型包含進入與離開的 2 個狀態共有 5 個狀態，每個狀態中高斯混合模型則依照該音素在訓練語料中的多寡，分別由 1 至 64 個高斯混合模型構成。詳細音素出現次數及高斯混合數見附錄 A。

3.2.3 詞典建立與語言模型設定

詞彙抽取使用中央通訊社(Central News Agency, CAN)在 2001 年與 2002 年所收集到約 1 億 7000 萬字語料作為文字語料。在中文裡約有 7000 個常用單字詞，新詞可由此 7000 個單字詞合併產生。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方式如下所述：對於語料中任意相鄰的兩個詞$\left(w_i, w_j\right)$，可以分別計算它們的前雙連文法(forward bigram)機率$P_f\left(w_j \mid w_i\right)$，與後雙連文法(backward bigram)機率$P_b\left(w_i \mid w_j\right)$，並以前後雙連文法(forward and backward bigrams)的機率幾何平均作為$\left(w_i, w_j\right)$是否合併的依據。機率幾何平均$FB\left(w_i, w_j\right)$為：

$$FB\left(w_i, w_j\right) = \sqrt{P_f\left(w_j \mid w_i\right) P_b\left(w_i \mid w_j\right)} \tag{3.1}$$

抽取過程中先經由一個含有 1~4 字詞約 68000 個詞的詞典對文字語料斷詞，然後利用(3.1)的公式，經數次的疊代以及不同的閥值(threshold)設定，產生約 5000 個由 2~10 個詞組成的複合詞，使得最後的語音辨識詞典裡總共有約 72000 個詞。

語言模型使用與詞典建立相同的文字語料，訓練得到詞三連語言模型，訓練過程中使用了凱氏語言模型平滑技術(Katz language model smoothing)【27】。

3.2.4 語音辨識工具

大字彙連續語音辨識可視為樣式辨認(pattern recognition)與搜尋演算法的結合，其中樣式辨認的部份是使用訓練語料訓練出的聲學及語言模型做為統計式樣式辨認的基礎，再經由這些模型去做後端的解碼(decoding)。解碼的程序主要是根據聲學及語言學的模型，找出一個最符合輸入信號的詞串(word sequence)，由此可以發現這樣的解碼的流程其實就是一個搜尋的過程。

在語音辨識時，首先需要依照聲學模型中的狀態轉移展開出可能的狀態樹狀圖，同時利用聲學及語言模型計算出樹狀圖中由樹根(root)到節點(node)的機率，之後再從中搜尋出機率最高的路徑作為辨識結果，但在搜尋的過程中，由於狀態樹的增長會隨著語音訊號的音框數量增加成指數成長，因此會加以使用光束搜尋法

(beam search)來避免狀態樹增長過大使得搜尋時間過長。光束搜尋法的主要作法是對同一階層(level)的節點依照機率由高至低排序，挑出同一階層中排名前 w 的節點，下一層便只有這些節點才會繼續增長，其中 w 便稱為光束的寬度；排在 w 名之後的節點則視為機率低於最佳路徑太多而忽略不計，如此一來，無論原本的搜尋空間多大，在光束搜尋演算法下的搜尋空間都可以限制在一定的範圍內。

此外，一般的語音辨識只會找出機率最高的路徑作為辨識結果，然而在狀態樹中，每個未被光束搜尋刪去的節點也會記錄著其歷史資訊，如語言模型歷史、對應的候選詞的首尾音框，以及搜尋時至此節點的機率分數等等，依照這些資訊可以建立詞圖。由於詞圖可以將語言模型與聲學模型的分數計算作出某種程度的分解，因此可以在詞圖上使用更高階的語言模型，重新進行一次詞圖重計分(rescoring)，而不至於使得第一階段的搜尋複雜度增加過多。另外在鑑別式訓練法中，也經常使用詞圖作為所有辨識可能的近似，在本論文中，語音辨識與詞圖的產生都是使用同一套工具。

3.3 基礎實驗(baseline)

本實驗依照 3.2.2 的初始聲學模型設定，語音訊號使用 3.2.1 的方式抽取梅爾倒頻譜係數特徵，使用 HTK 抽取語音訊號特徵及訓練聲學模型。而訓練過程使用最大相似度估測法。訓練過程首先必須給每個模型一個原型(prototype)，之後再經過最大相似度估測法的反覆疊代(iteration)訓練出最終模型。

本實驗的聲學模型原型為 5 個狀態，每個狀態有一個高斯混合數。在一個高斯混合數的情況下，經過 50 次的疊代之後，開始增加高斯混合數，高斯混合數的增加順序為：1→2→3→4→5→6→7→8→16→24→32→64。每次增加混合數之後，都會再做 4 次的疊代，才繼續增加至下一個數量的混合數，每個聲韻母模型的高斯混合數依照表 A.3 決定。在增加高斯混合數的過程中，每個聲韻母模型在增加到目標數量之後，該聲韻母模型在之後的增加過程中便不再增加混合數。

語音辨識是使用 TTK 完成，詞典與語言模型是依照 3.2.3 的方式產生，辨識

時使用詞三連語言模型，基礎實驗結果如表 3.2：

level	Corr(%)	Acc(%)	H	D	S	I	N
word	71.04	57.99	11424	411	4246	2099	16081
char	76.43	75.17	19997	444	5723	330	26164
syl	82.72	81.42	21689	466	4064	342	26219
I/F	86.29	84.76	45251	1054	6133	806	52438

表 3.2　基礎實驗結果

實驗結果以 4 個不同的層級表示，分別為詞(word)、字(character, 表中簡寫為 char)、音節(syllable, 表中簡寫為 syl)、聲韻母(Initial/Final, 表中簡寫為 I/F)，而這 4 個層級的規則中，詞是以詞典中列出的詞為準，在實驗中原本的辨識結果就是詞；字則是將詞全部拆開成一個一個的字，在實驗中是將辨識結果的詞斷開而來；音節則是將同音字視為相同的單位，在實驗中是將辨識結果的詞，依照詞典中的發音對應轉換而來；聲韻母則是聲學模型的單位，在實驗中也是將辨識結果的詞，依照詞典中的發音對應轉換而來，而在本實驗中因為是使用右相關聲韻母模型，每一個音節都是由一個聲母加一個韻母構成，所以音節跟聲韻母的差異就是兩個聲韻母組成一個音節。

而表中的 H(hit, 命中)是代表辨識結果與正確答案相同的部份，D(deletion, 遺失)是正確答案中有出現但辨識結果沒出現的部份，S(substation, 取代)是辨識結果中與正確答案中相異的部份，I(insertion, 插入)是正確答案中沒出現但辨識結果中有出現的部份，N(number)是正確答案的總數(如在 word 中就是指總詞數，在 char 中就是總字數)。D、S、I 是由計算編輯距離而來，計算方式如圖 2.2，最後 H 的計算方式是：

$$H = N - D - S \tag{3.2}$$

而在表中的 Corr(correct)命中率的計算方式是：

$$Corr = \frac{H}{N} \times 100\% \tag{3.3}$$

表中的 Acc(accuracy)正確率的計算方式是：

$$Acc = \frac{H - I}{N} \times 100\% \tag{3.4}$$

一般的語音辨識結果的評估，都是以詞正確率為標準，而在中文的語音辨識中，一般又以字正確率為標準。唯本論文為了深入分析不同方法的表現，之後的實驗結果依然會列出詳細各項數據。

3.4 本章結論

本章介紹了本論文使用的中文大字彙辨識系統，以及基礎實驗的訓練方式，包括前端處理、聲學模型訓練、詞典與語言模型的建立，以及辨識工具的解碼方式。在公視新聞語料上，經由最大相似度估測法，可以得到約 75.17%字正確率的結果，這個結果將作為本論文之後鑑別式訓練法的基礎實驗。

第4章　最小音素錯誤訓練

本章將介紹最小音素錯誤訓練的理論基礎以及實作上的方法，包含理論的推導過程和實作上的各項設定【28】【29】。

4.1　目標函數

最小音素錯誤聲學模型訓練的目標函數 $F_{MPE}(\lambda)$ 為音素正確度的期望值【30】:

$$\begin{aligned} F_{MPE}(\lambda) &= \sum_r \sum_u P(u|O_r)\,Acc(s_r,u) \\ &= \sum_r \frac{\sum_{u\in W_h} P_\lambda(O_r|u)\,P_\Gamma(u)\,Acc(s_r,u)}{\sum_{u\in W_h} P_\lambda(O_r|u)\,P_\Gamma(u)} \end{aligned} \tag{4.1}$$

其中 λ 為模型參數，O_r 是第 r 句語句，s_r 是 O_r 的正確轉寫，$Acc(s_r,u)$ 為其中一句可能辨識 u 的正確度，W_h 為所有的可能辨識結果，$P(u|O_r)$ 為辨識結果 u 的事後機率。而在實作中由於不可能列舉出所有的可能辨識結果來計算事後機率，因此通常使用詞圖做為所有可能辨識結果的近似。

4.1.1　目標函數之最佳化

由於(4.1)中的 $Acc(s_r,u)$ 未必是正數，且目標函數整體是由數個分數項目相加而不同於最大相互資訊是相乘而來，因此目標函數在使用延伸式波氏重估最佳化上會有困難。在這裡使用弱性輔助函數(weak-sense auxiliary function)處理這個問題(見附錄 B)，令一弱性輔助函數 $H_{MPE}(\lambda,\bar{\lambda})$ 為：

$$H_{MPE}(\lambda,\bar{\lambda}) = \sum_r \sum_{q\in W_{lat}^r} \left.\frac{\partial F_{MPE}(\lambda)}{\partial \log P_\lambda(O_r|q)}\right|_{\lambda=\bar{\lambda}} \log P(O_r|q) \tag{4.2}$$

其中 q 是第 r 句訓練語句的詞圖 W_{lat}^r 上的一個詞弧(arc)音素，$P_\lambda(O_r|q)$ 則表示音素 q 的相似度。$H_{MPE}(\lambda,\bar{\lambda})$ 是 $F_{MPE}(\lambda)$ 在 $\lambda=\bar{\lambda}$ 附近的弱性輔助函數，由於在(4.1)中 $F_{MPE}(\lambda)$ 會隨著 λ 變動的項目只有 $P_\lambda(O_r|s_r)$，而 $P_\lambda(O_r|s_r)$ 又是由 $P_\lambda(O_r|q)$ 的連乘

組成，因此$F_{MPE}(\lambda)$與$\log P_{\lambda}(O_r|q)$在$\lambda=\overline{\lambda}$時對$\lambda$的偏微分結果會相同，所以可以使用$H_{MPE}(\lambda,\overline{\lambda})$做為$F_{MPE}(\lambda)$的弱性輔助函數。而針對$H_{MPE}(\lambda,\overline{\lambda})$又可以再使用一弱性輔助函數$G_{MPE}(\lambda,\overline{\lambda})$以方便對模型參數的微分：

$$G_{MPE}(\lambda,\overline{\lambda})=\sum_{r}\sum_{q\in W_{lat}^{r}}\left.\frac{\partial F_{MPE}(\lambda)}{\partial\log P_{\lambda}(O_r|q)}\right|_{\lambda=\overline{\lambda}}F_{MLE}(\lambda,\overline{\lambda},r,q) \tag{4.3}$$

其中$F_{MLE}(\lambda,\overline{\lambda},r,q)$可以理解為在詞圖$r$上的音素$q$的相似度$P_{\lambda}(O_r|q)$的輔助函數。

在兩次的弱性輔助函數轉換後，$G_{MPE}(\lambda,\overline{\lambda})$依然是$F_{MPE}(\lambda)$的弱性輔助函數。而在(4.3)中再定義：

$$\gamma_{q}^{MPE}=\frac{\partial F_{MPE}(\lambda)}{\partial\log P_{\lambda}(O_r|q)} \tag{4.4}$$

(4.4)即是目標函數$F_{MPE}(\lambda)$對對數相似度$P_{\lambda}(O_r|q)$的偏微分，經過一連串的輔助函數代換後，目標函數的最佳化就會在(4.4)式中的偏微分完成。

另外將(4.3)式中的$F_{MLE}(\lambda,\overline{\lambda},r,q)$進一步代換之後可以得到：

$$G_{MPE}(\lambda,\overline{\lambda})=\sum_{r}\sum_{q\in W_{lat}^{r}}\sum_{t=s_q}^{t=e_q}\sum_{m}\gamma_{qr}^{MPE}\gamma_{qm}^{r}(t)\log N(O_r(t),\mu_m,\Sigma_m) \tag{4.5}$$

其中s_q與e_q分別代表詞弧q的起始與結束時間；$O_r(t)$代表O_r在時間t的語音特徵向量；$\gamma_{qm}^{r}(t)$表示在第r句訓練語句中的詞弧q，在時間t時，在第m個高斯混合的佔有機率(occupation probability)；γ_{qr}^{MPE}表示第r句訓練語句的微分結果。

4.1.2 目標函數之微分

對於(4.4)的目標函數$F_{MPE}(\lambda)$對對數相似度$P_{\lambda}(O_r|q)$的偏微分，首先把目標函數中對整個詞圖的詞弧之加總$u\in W_h$，針對特定的詞弧$\hat{v}$，分成包含$\hat{v}$以及不包含$\hat{v}$的文句：

$$
\begin{aligned}
F_{MPE}(\lambda) &= \sum_r \frac{\sum_{u\in W_{lath}^r} P_\lambda(O_r\,|\,u)P_\Gamma(u)Acc(s_r,u)}{\sum_{u\in W_{lat}^r} P_\lambda(O_r\,|\,u)P_\Gamma(u)} \\
&= \sum_r \frac{\sum_{u\in W_{lat}^r,\,\hat{v}\in u} P_\lambda(O_r\,|\,u)P_\Gamma(u)Acc(s_r,u) + \sum_{u\in W_{lat}^r,\,\hat{v}\notin u} P_\lambda(O_r\,|\,u)P_\Gamma(u)Acc(s_r,u)}{\sum_{u\in W_{lat}^r,\,\hat{v}\in u} P_\lambda(O_r\,|\,u)P_\Gamma(u) + \sum_{u\in W_{lat}^r,\,\hat{v}\notin u} P_\lambda(O_r\,|\,u)P_\Gamma(u)}
\end{aligned}
\tag{4.6}
$$

並且已知：

$$
\frac{\partial P_\lambda(O_r\,|\,\hat{v})}{\partial \log P_\lambda(O_r\,|\,q)} = \begin{cases} P_\lambda(O_r\,|\,\hat{v}) & ,\hat{v}\in q \\ 0 & ,\hat{v}\notin q \end{cases}
\tag{4.7}
$$

另外再設：

$$
\begin{aligned}
a_r(\lambda) &= \sum_{u\in W_{lat}^r} P_\lambda(O_r\,|\,u)P_\Gamma(u)Acc(s_r,u) \\
b_r(\lambda) &= \sum_{u\in W_{lat}^r} P_\lambda(O_r\,|\,u)P_\Gamma(u)
\end{aligned}
\tag{4.8}
$$

於是：

$$
F_{MPE}(\lambda) = \sum_r \frac{\sum_{u\in W_{lat}^r} P_\lambda(O_r\,|\,u)P_\Gamma(u)Acc(s_r,u)}{\sum_{u\in W_{lat}^r} P_\lambda(O_r\,|\,u)P_\Gamma(u)} = \sum_r \frac{a_r(\lambda)}{b_r(\lambda)}
\tag{4.9}
$$

因此 $F_{MPE}(\lambda)$ 對 $\log P_\lambda(O_r\,|\,q)$ 的微分可得：

$$
\begin{aligned}
\frac{\partial F_{MPE}(\lambda)}{\partial \log P_\lambda(O_r\,|\,q)} &= \sum_r \frac{\partial \dfrac{a_r(\lambda)}{b_r(\lambda)}}{\partial \log P_\lambda(O_r\,|\,q)} \\
&= \sum_r \frac{\dfrac{\partial a_r(\lambda)}{\partial \log P_\lambda(O_r\,|\,q)} b_r(\lambda) - \dfrac{\partial b_r(\lambda)}{\partial \log P_\lambda(O_r\,|\,q)} a_r(\lambda)}{b_r(\lambda)^2} \\
&= \sum_r \frac{\dfrac{\partial a_r(\lambda)}{\partial \log P_\lambda(O_r\,|\,q)}}{b_r(\lambda)} - \frac{a_r(\lambda)}{b_r(\lambda)} \frac{\dfrac{\partial b_r(\lambda)}{\partial \log P_\lambda(O_r\,|\,q)}}{b_r(\lambda)}
\end{aligned}
\tag{4.10}
$$

根據(4.7)，可以計算出：

$$\frac{\partial a_r(\lambda)}{\partial \log P_\lambda(O_r \mid q)} = \frac{\partial \sum_{u \in W_{lat}^r,\, \hat{v} \in u} P_\lambda(O_r \mid u) P_\Gamma(u) Acc(s_r, u)}{\partial \log P_\lambda(O_r \mid q)} = \sum_{u \in W_{lat}^r,\, \hat{v} \in u} P_\lambda(O_r \mid u) P_\Gamma(u) Acc(s_r, u) \tag{4.11}$$

$$\frac{\partial b_r(\lambda)}{\partial \log P_\lambda(O_r \mid q)} = \frac{\partial \sum_{u \in W_{lat}^r,\, \hat{v} \in u} P_\lambda(O_r \mid u) P_\Gamma(u)}{\partial \log P_\lambda(O_r \mid q)} = \sum_{u \in W_{lat}^r,\, \hat{v} \in u} P_\lambda(O_r \mid u) P_\Gamma(u) \tag{4.12}$$

將(4.11)與(4.12)代入(4.10)可以得到：

$$\begin{aligned}\frac{\partial F_{MPE}(\lambda)}{\partial \log P_\lambda(O_r \mid q)} &= \sum_r \frac{\sum_{u \in W_{lat}^r,\, \hat{v} \in u} P_\lambda(O_r \mid u) P_\Gamma(u) Acc(s_r, u)}{\sum_{u \in W_{lat}^r} P_\lambda(O_r \mid u) P_\Gamma(u)} \\ &\quad - \frac{\sum_{u \in W_{lat}^r} P_\lambda(O_r \mid u) P_\Gamma(u) Acc(s_r, u)}{\sum_{u \in W_{lat}^r} P_\lambda(O_r \mid u) P_\Gamma(u)} \frac{\sum_{u \in W_{lat}^r,\, \hat{v} \in u} P_\lambda(O_r \mid u) P_\Gamma(u)}{\sum_{u \in W_{lat}^r} P_\lambda(O_r \mid u) P_\Gamma(u)} \\ &= \sum_r \frac{\sum_{u \in W_{lat}^r,\, \hat{v} \in u} P_\lambda(O_r \mid u) P_\Gamma(u) Acc(s_r, u)}{\sum_{u \in W_{lat}^r,\, \hat{v} \in u} P_\lambda(O_r \mid u) P_\Gamma(u)} \frac{\sum_{u \in W_{lat}^r,\, \hat{v} \in u} P_\lambda(O_r \mid u) P_\Gamma(u)}{\sum_{u \in W_{lat}^r} P_\lambda(O_r \mid u) P_\Gamma(u)} \\ &\quad - \frac{\sum_{u \in W_{lat}^r} P_\lambda(O_r \mid u) P_\Gamma(u) Acc(s_r, u)}{\sum_{u \in W_{lat}^r} P_\lambda(O_r \mid u) P_\Gamma(u)} \frac{\sum_{u \in W_{lat}^r,\, \hat{v} \in u} P_\lambda(O_r \mid u) P_\Gamma(u)}{\sum_{u \in W_{lat}^r} P_\lambda(O_r \mid u) P_\Gamma(u)}\end{aligned} \tag{4.13}$$

其中令

$$\gamma_q^r = \frac{\sum_{u \in W_{lat}^r,\, \hat{v} \in u} P_\lambda(O_r \mid u) P_\Gamma(u)}{\sum_{u \in W_{lat}^r} P_\lambda(O_r \mid u) P_\Gamma(u)} \tag{4.14}$$

代表在O_r的詞圖上，詞弧q的佔有機率。於是(4.13)就可簡寫為：

$$\frac{\partial F_{MPE}(\lambda)}{\partial \log P_\lambda(O_r \mid q)} = \sum_r \gamma_q^r \left(\frac{\sum_{u \in W_{lat}^r,\, \hat{v} \in u} P_\lambda(O_r \mid u) P_\Gamma(u) Acc(s_r, u)}{\sum_{u \in W_{lat}^r,\, \hat{v} \in u} P_\lambda(O_r \mid u) P_\Gamma(u)} - \frac{\sum_{u \in W_{lat}^r} P_\lambda(O_r \mid u) P_\Gamma(u) Acc(s_r, u)}{\sum_{u \in W_{lat}^r} P_\lambda(O_r \mid u) P_\Gamma(u)} \right) \tag{4.15}$$

然後再令$C_r(q)$為在語句O_r的詞圖W_{lat}^r中所有有經過詞弧$\hat{v}$的文句對於正確轉寫s_r的期望正確度：

$$C_r(q)=\frac{\sum_{u\in W_{lat}^r,\hat{v}\in u}P_\lambda(O_r|u)P_\Gamma(u)Acc(s_r,u)}{\sum_{u\in W_{lat}^r,\hat{v}\in u}P_\lambda(O_r|u)P_\Gamma(u)} \tag{4.16}$$

令C_{avg}^r為在語句O_r的詞圖W_{lat}^r中所有文句對於正確轉寫s_r的期望正確度：

$$C_{avg}^r=\frac{\sum_{u\in W_{lat}^r}P_\lambda(O_r|u)P_\Gamma(u)Acc(s_r,u)}{\sum_{u\in W_{lat}^r}P_\lambda(O_r|u)P_\Gamma(u)} \tag{4.17}$$

因此整個(4.4)的微分推導結果成為：

$$\gamma_q^{MPE}=\frac{\partial F_{MPE}(\lambda)}{\partial\log P_\lambda(O_r|q)}=\sum_r\gamma_q^r\left(C_r(q)-C_{avg}^r\right) \tag{4.18}$$

而對第r句訓練語句的微分結果就是：

$$\gamma_{qr}^{MPE}=\gamma_q^r\left(C_r(q)-C_{avg}^r\right) \tag{4.19}$$

最小音素錯誤模型訓練法的主要概念是，對於一個詞弧q，若其$C_r(q)>C_{avg}^r$，則認為通過詞弧q的文句平均正確度高於所有文句的平均正確度，因此將詞弧q視為正確的詞弧，而對詞弧q中的音素聲學模型做正向訓練；反之若$C_r(q)<C_{avg}^r$，則將詞弧q視為錯誤的詞弧，而對詞弧q中的音素聲學模型做負向訓練。若$C_r(q)=C_{avg}^r$，則表示詞弧q沒有相異的競爭詞弧(可能沒有別的競爭詞弧，或競爭詞弧中的音素聲學模型相同)，在這個情況下，詞弧q不會用來對音素聲學模型訓練。

4.1.3 聲學模型參數更新

對於聲學模型參數的估測，基本上就是將輔助函數(4.5)對聲學模型參數λ偏微分，再設其微分結果為 0 並求解，就可以得到模型參數λ的波氏重估公式。但是因為$G_{MPE}(\lambda,\overline{\lambda})$只是一弱性輔助函數，所以原目標函數$F_{MPE}(\lambda)$不一定與$G_{MPE}(\lambda,\overline{\lambda})$有相同的遞增或遞減的趨勢。為了增加參數估測的一般性，故在此引入

一個以舊有模型參數為超參數(hyperparameter)的平滑函數$G_{EBW}^{smooth}\left(\lambda,\overline{\lambda}\right)$：

$$G_{EBW}^{smooth}\left(\lambda,\overline{\lambda}\right)=\sum_{m}-\frac{D_m}{2}\left(\log\left(\left|\Sigma_m\right|\right)+\left(\mu_m-\overline{\mu}_m\right)^T\Sigma_m^{-1}\left(\mu_m-\overline{\mu}_m\right)+tr\left(\overline{\Sigma}_m\Sigma_m^{-1}\right)\right) \tag{4.20}$$

其中μ_m、Σ_m與$\overline{\mu}_m$、$\overline{\Sigma}_m$分別代表新模型與舊模型的平均值向量與共變異矩陣；D_m為高斯分佈層次的平滑係數，其主要目是為了保證更新的模型參數共變異矩陣Σ_m為正定(positive definite)。平滑函數$G_{EBW}^{smooth}\left(\lambda,\overline{\lambda}\right)$的選擇是在舊模型$\lambda=\overline{\lambda}$時$G_{EBW}^{smooth}\left(\lambda,\overline{\lambda}\right)$會有全域最大值，藉此來改善弱性輔助函數的收斂速度與效果。原輔助函數(4.5)加上平滑函數(4.20)後仍然符合輔助函數的條件：

$$\begin{aligned}G_{MPE}^{smooth}\left(\lambda,\overline{\lambda}\right)&=\sum_{r}\sum_{q\in W_{lat}^r}\sum_{t=s_q}^{t=e_q}\sum_{m}\gamma_{qr}^{MPE}\gamma_{qm}^{r}\left(t\right)\log N\left(O_r\left(t\right),\mu_m,\Sigma_m\right)+G_{EBW}^{smooth}\left(\lambda,\overline{\lambda}\right)\\&=\sum_{r}\sum_{q\in W_{lat}^r}\sum_{t=s_q}^{t=e_q}\sum_{m}\gamma_{qr}^{MPE}\gamma_{qm}^{r}\left(t\right)\log N\left(O_r\left(t\right),\mu_m,\Sigma_m\right)\\&\quad-\sum_{m}\frac{D_m}{2}\left(\log\left(\left|\Sigma_m\right|\right)+\left(\mu_m-\overline{\mu}_m\right)^T\Sigma_m^{-1}\left(\mu_m-\overline{\mu}_m\right)+tr\left(\overline{\Sigma}_m\Sigma_m^{-1}\right)\right)\end{aligned} \tag{4.21}$$

最後模型參數的估測就是將(4.21)式對參數μ_m與Σ_m偏微分：

$$\begin{aligned}\frac{\partial G_{MPE}^{smooth}\left(\lambda,\overline{\lambda}\right)}{\partial\mu_m}&=\sum_{r}\sum_{q\in W_{lat}^r}\sum_{t=s_q}^{t=e_q}\gamma_{qr}^{MPE}\gamma_{qm}^{r}\left(t\right)\Sigma_m^{-1}\left(O_r\left(t\right)-\mu_m\right)\\&\quad-D_m\left(\Sigma_m^{-1}\left(\mu_m-\overline{\mu}_m\right)\right)\end{aligned} \tag{4.22}$$

$$\begin{aligned}\frac{\partial G_{MPE}^{smooth}\left(\lambda,\overline{\lambda}\right)}{\partial\Sigma_m^{-1}}&=\sum_{r}\sum_{q\in W_{lat}^r}\sum_{t=s_q}^{t=e_q}\gamma_{qr}^{MPE}\gamma_{qm}^{r}\left(t\right)\frac{1}{2}\left(\Sigma_m^T-\left(O_r\left(t\right)-\mu_m\right)\left(O_r\left(t\right)-\mu_m\right)^T\right)\\&\quad+\frac{D_m}{2}\left(\Sigma_m^T-\left(\mu_m-\overline{\mu}_m\right)\left(\mu_m-\overline{\mu}_m\right)^T-\overline{\Sigma}_m^T\right)\end{aligned} \tag{4.23}$$

然後設(4.22)與(4.23)等於 0(0 向量或 0 矩陣)，就可以推導出用於平均值向量與共變異矩陣的延伸式波氏重估公式：

$$\begin{aligned}\mu_m&=\frac{\theta_m^{MPE}\left(O\right)+D_m\overline{\mu}_m}{\gamma_m^{MPE}+D_m}\\\Sigma_m&=\frac{\theta_m^{MPE}\left(O^2\right)+D_m\left(\overline{\Sigma}_m+\overline{\mu}_m\overline{\mu}_m^T\right)}{\gamma_m^{MPE}+D_m}-\overline{\mu}_m\overline{\mu}_m^T\end{aligned} \tag{4.24}$$

其中γ_m^{MPE}、$\theta_m^{MPE}\left(O\right)$和$\theta_m^{MPE}\left(O^2\right)$分別是：

$$\gamma_m^{MPE} = \sum_r \sum_{q \in W_{lat}^r} \sum_{t=s_q}^{t=e_q} \gamma_{qr}^{MPE} \gamma_{qm}^r(t)$$

$$\theta_m^{MPE}(O) = \sum_r \sum_{q \in W_{lat}^r} \sum_{t=s_q}^{t=e_q} \gamma_{qr}^{MPE} \gamma_{qm}^r(t) O_r(t) \tag{4.25}$$

$$\theta_m^{MPE}(O^2) = \sum_r \sum_{q \in W_{lat}^r} \sum_{t=s_q}^{t=e_q} \gamma_{qr}^{MPE} \gamma_{qm}^r(t) O_r(t) O_r(t)^T$$

承 4.1.2 節對於(4.19)式的說明，若 $\gamma_{qr}^{MPE} > 0$ 則視 q 為正確詞弧；若 $\gamma_{qr}^{MPE} < 0$ 則視 q 為錯誤詞弧。為了統計的方便及呈現模型更新的意義，可以將 γ_{qr}^{MPE} 拆解成正負兩個部份，正的部份為分子詞圖(numerator lattice)，負的部份為分母詞圖(denominator lattice)，於是(4.25)就可以改寫為：

$$\gamma_m^{num} = \sum_r \sum_{q \in W_{lat}^r} \sum_{t=s_q}^{t=e_q} \max\left(0, \gamma_{qr}^{MPE}\right) \gamma_{qm}^r(t)$$

$$\theta_m^{num}(O) = \sum_r \sum_{q \in W_{lat}^r} \sum_{t=s_q}^{t=e_q} \max\left(0, \gamma_{qr}^{MPE}\right) \gamma_{qm}^r(t) O_r(t) \tag{4.26}$$

$$\theta_m^{num}(O^2) = \sum_r \sum_{q \in W_{lat}^r} \sum_{t=s_q}^{t=e_q} \max\left(0, \gamma_{qr}^{MPE}\right) \gamma_{qm}^r(t) O_r(t) O_r(t)^T$$

$$\gamma_m^{den} = \sum_r \sum_{q \in W_{lat}^r} \sum_{t=s_q}^{t=e_q} \max\left(0, -\gamma_{qr}^{MPE}\right) \gamma_{qm}^r(t)$$

$$\theta_m^{den}(O) = \sum_r \sum_{q \in W_{lat}^r} \sum_{t=s_q}^{t=e_q} \max\left(0, -\gamma_{qr}^{MPE}\right) \gamma_{qm}^r(t) O_r(t) \tag{4.27}$$

$$\theta_m^{den}(O^2) = \sum_r \sum_{q \in W_{lat}^r} \sum_{t=s_q}^{t=e_q} \max\left(0, -\gamma_{qr}^{MPE}\right) \gamma_{qm}^r(t) O_r(t) O_r(t)^T$$

依照(4.26)與(4.27)的方式改寫(4.24)可以得到：

$$\begin{aligned} \mu_m &= \frac{\theta_m^{num}(O) - \theta_m^{den}(O) + D_m \overline{\mu}_m}{\gamma_m^{num} - \gamma_m^{den} + D_m} \\ \Sigma_m &= \frac{\theta_m^{num}(O^2) - \theta_m^{den}(O^2) + D_m\left(\overline{\Sigma}_m + \overline{\mu}_m \overline{\mu}_m^T\right)}{\gamma_m^{num} - \gamma_m^{den} + D_m} - \overline{\mu}_m \overline{\mu}_m^T \end{aligned} \tag{4.28}$$

由(4.28)可以看出，平均值向量 μ_m 的更新會使 μ_m 逐漸靠近屬於分子詞圖的特徵 $\theta_m^{num}(O)$，遠離屬於分母詞圖的特徵 $\theta_m^{den}(O)$；同理，共變異矩陣 Σ_m 的調整也有一樣的效果。而因為有這樣的正向及負向訓練，新的模型可以更具鑑別性。

4.1.4 I 平滑

I 平滑技術同樣是為輔助函數引入平滑函數，方法是在估測模型參數的過程中，引入待測模型的事前分佈。估測事前機率的方法有很多種，I 平滑是以最大相似度估測的統計資訊做為最小音素錯誤訓練法之估測模型的事前分佈，此平滑函數 $G_{I-smooth}(\lambda)$ 為：

$$\begin{aligned} G_{I-smooth}(\lambda) = -\frac{\tau}{2}\sum_m \log\left(|\Sigma_m|\right) + \left(\mu_m - \frac{\theta_m^{ML}(O)}{\gamma_m^{ML}}\right)^T \Sigma_m^{-1}\left(\mu_m - \frac{\theta_m^{ML}(O)}{\gamma_m^{ML}}\right) \\ + tr\left(\left(\frac{\theta_m^{ML}(O^2)}{\gamma_m^{ML}} - \left(\frac{\theta_m^{ML}(O)}{\gamma_m^{ML}}\right)^2\right)\Sigma_m^{-1}\right) \end{aligned} \tag{4.29}$$

其中 τ 為一常數，γ_m^{ML} 、$\theta_m^{ML}(O)$ 和 $\theta_m^{ML}(O^2)$ 為最大相似度估測的統計值：

$$\begin{aligned} \gamma_m^{ML} &= \sum_r \sum_t \gamma_{mr}^{ML}(t) \\ \theta_m^{ML}(O) &= \sum_r \sum_t \gamma_{mr}^{ML}(t) O_r(t) \\ \theta_m^{ML}(O^2) &= \sum_r \sum_t \gamma_{mr}^{ML}(t) O_r(t) O_r(t)^T \end{aligned} \tag{4.30}$$

$\gamma_{mr}^{ML}(t)$ 表示在時間 t 時第 r 句訓練語句以最大相似度所估測之高斯混合 m 的佔有機率。之後再將(4.30)統計的結果加入分子詞圖(4.26)的統計中：

$$\begin{aligned} \gamma_{m,I-smooth}^{num} &= \gamma_m^{num} + \tau \\ \theta_{m,I-smooth}^{num}(O) &= \theta_m^{num}(O) + \frac{\tau}{\gamma_m^{ML}}\theta_m^{ML}(O) \\ \theta_{m,I-smooth}^{num}(O^2) &= \theta_m^{num}(O^2) + \frac{\tau}{\gamma_m^{ML}}\theta_m^{ML}(O^2) \end{aligned} \tag{4.31}$$

由(4.31)式可以看出，I 平滑在某種程度上，可以視為將最小音素錯誤的估測結果與最大相似度的估測做內插(interpolation)，而 τ 就是在內插中最大相似度的估測值所佔的權重。

4.2 實作流程

由 4.1 節的理論推導，可以大略看出最小音素錯誤訓練的概念，然而這些統計資訊的實際計算方式並不直接，諸如用於近似辨識可能的詞圖，詞圖中詞弧的正確度以及通過機率之計算，以下會對這些實作流程作詳細的介紹。

4.2.1 詞圖

詞圖的產生，是在語音辨識的過程中，相對於使用詞彙樹複製(tree copy)搜尋【31】只找出最高分的文句，詞圖的產生是在搜尋時不只保留最高分的選項，而將分數雖不是最高，但也離最高分不遠的詞串留下，做為辨識結果的候選空間。產生的候選空間便是詞圖。

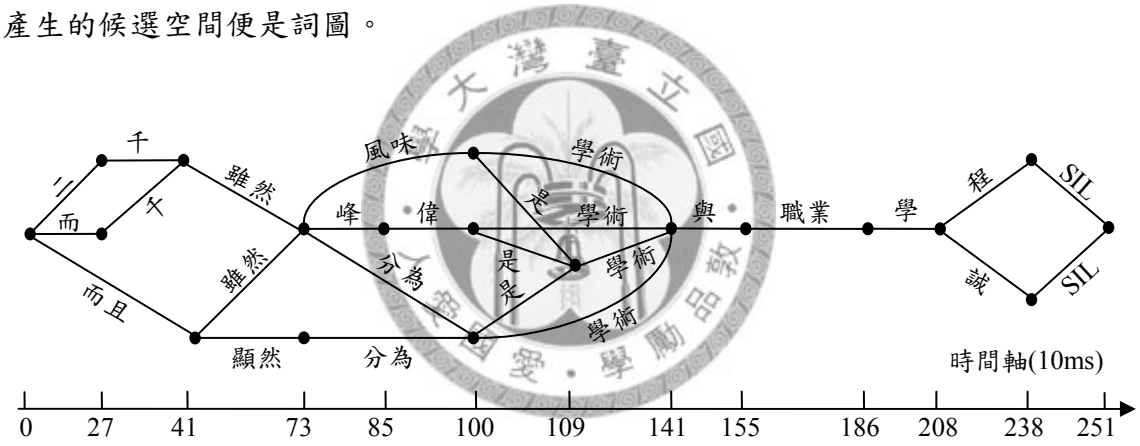


圖 4.1 詞圖範例

圖 4.1 是一個中文詞圖範例。詞圖是一單向無迴圈(acyclic directed graph)的網路，對一句時間長度為 T 的語音，網路起點位於語句時間的起始處，終點位於語句時間的結尾處。從起點出發到終點之間路徑的選擇，只能往時間較大的方向前進，路徑上經過的每一段詞弧代表此段時間語音可能對應的詞。每一個詞弧上的資訊包括此段詞弧的開始與結束時間，以及詞弧在詞圖上連接的前一個詞和詞弧本身代表的詞。另外，為了便於搜尋，每段詞弧也可以記錄在以詞彙樹複製搜尋時產生的聲學模型分數或其它信心分數等等，省去之後使用詞圖時重新計算的時間。

詞圖的產生可以跟語音辨識時一並完成,在語音辨識如 3.2.4 節的搜尋方法時,光束搜尋法會設定一個閥值,只保留目前的狀態中分數高於基準的繼續往下搜尋,直到語句結尾,才選出最高分的一條路徑做為辨識結果。而若要產生詞圖,則需再另設一詞圖閥值,決定在搜尋過程中與最高分接近的程度,要到多接近才認定為與最高分分數足夠接近,可以在產生詞圖時將詞弧放進候選空間。因此詞圖閥值最大不能超過光束閥值,當詞圖閥值設到跟光束閥值一樣大時,所有在光束搜尋法的過程中,有保留下來的狀態最後都會納入候選空間裡。

詞圖閥值的設定方法有許多種,本論文用於鑑別式訓練法的可能辨識近似的詞圖,建立詞圖時的詞圖閥值使用了兩種不同的方式共產生了兩份詞圖:一是給定一個與最高分分數的比值,表示要保留與最高分分數相差在幾倍以內的詞弧要納入候選空間;二是給定一個數量,表示要保留從最高分開始依照分數排序下來,在這個數量之內的詞弧要納入候選空間。這兩種方式分別由兩種不同的語音辨識工具完成,分別是 TTK 以及 NDecoder。

另外在為鑑別式訓練法的可能辨識近似產生詞圖時,通常會使用較低連的語言模型,在本論文中辨識時使用詞三連語言模型,產生詞圖時則使用詞雙連語言模型,目的是因為鑑別式訓練法目標是在改進聲學模型,所以為了使聲學模型的錯誤不要被語言模型修復太多,而在訓練時可以出現在詞圖裡而被改進到,所以僅使用詞雙連語音模型來產生詞圖。

4.2.2 詞弧正確度

在(4.1)式中的音素正確度 $Acc(s_r, u)$,理論上應該使用如圖 2.2 編輯距離的計算方式而來,然而編輯距離的計算需要在整個句子都有辨識結果的情況下才能計算,因此若要使用編輯距離計算詞圖的正確度,等於是要把詞圖經過每一段詞弧的路徑組合都窮舉出來,再一一計算,這種方式顯然計算過於龐大而不切實際,因此波氏(Povey)提出的最小音素錯誤訓練法【3】使用了一個編輯距離的近似方法來計算正確度。近似正確度首先定義每個音素個別的正確度:

$$PhoneAcc(q)=\max_{z}\begin{cases}-1+2e(q,z) & \text{if } q=z\\ -1+e(q,z) & \text{if } q\neq z\end{cases} \tag{4.32}$$

其中音素 z 是正確轉寫，音素 q 是辨識結果，而 $e(q,z)$ 則是：

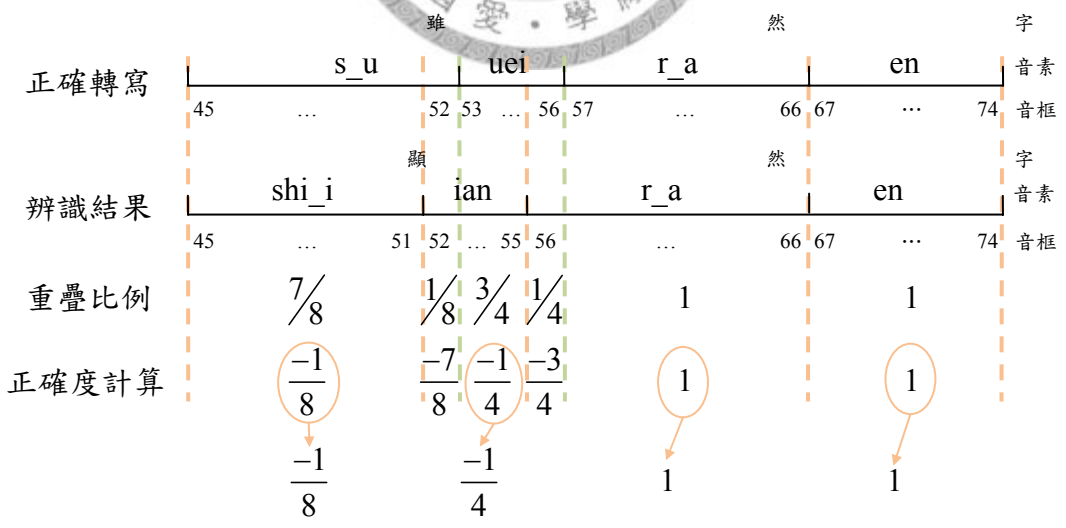
$$e(q,z)=\frac{\text{音素}q\text{與}z\text{之重疊長度}}{\text{音素}z\text{之長度}} \tag{4.33}$$

PhoneAcc 的值會落在-1 和+1 之間，這樣的近似方式是為了在正確音素和內插音素的逼近之間求取平衡。而整句文筍的近似正確度就是把一句當中的每一個音素正確度加總起來：

$$Acc(s_r,u)=\sum_{q\in u}PhoneAcc(q) \tag{4.34}$$

其中 q 代表組成文句 u 的音素，s_r 為正確轉寫。

近似正確度的計算方式是，首先分別對每個辨識出來的音素比對同時間範圍內的正確轉寫，與每個在正確轉寫中有時間重疊的音素做一次正確度的計算，從中選出最大值當成該音素的正確度，然後再把整句的每個音素正確度加起來，就成為整句的近似正確度。精確正確度就是由計算編輯距離而來，編輯距離只能在



近似正確度 $=\frac{-1}{8}+\frac{-1}{4}+1+1=\frac{13}{8}$

精確正確度 $=2\ \text{hit}(\text{r\_a \& en})=2$

圖 4.2　音素正確度的近似及精確計算範例

整句都已知的情況下計算，因此沒有每個音素的正確度存在。

圖 4.2 是一個近似正確度與精確正確度計算的例子，在圖中 shi_i 對應到重疊的音素為 s_u，重疊比例為 $7/8$，因為是不同的音素，故正確度為 $-1+7/8=-1/8$；ian 同時與 s_u 和 uei 都有重疊，重疊比例分別為 $1/8$ 和 $3/4$，同樣都是不同的音素，故正確度分別為 $-1+1/8=-7/8$ 和 $-1+3/4=-1/4$，最後取較大者為 $-1/4$；r_a 與 uei 和 r_a 都有重疊，重疊比例分別為 $1/4$ 和 1，其中一個為不同的音素，另一個為相同的音素，故正確度分別為 $-1+1/4=-3/4$ 和 $-1+2\times1=1$，最後取較大者為 1；en 只與 en 重疊，重疊比例為 1，為相同的音素，故正確度為 $-1+2\times1=1$，而整句的近似正確度就是各音素正確度的總合 $13/8$。另外，由於靜音在辨識的結果中不列入正確率的計算，所以在詞弧正確度的計算中遇到靜音時也一律跳過不計算。

在最小音素錯誤訓練法中，每個詞弧的正確度計算方法，就是將詞弧中的各音素正確度計算出來後再加總，因此每個詞弧的正確度都可以個別計算出。

4.2.3 詞圖期望正確度

在(4.18)中的 C_{avg}^{r} 以及 $C_r(q)$ 分別是指整個詞圖的期望正確度以及詞圖中所有的路徑中，有通過詞弧 q 的路徑的期望正確度。以 4.2.2 節的方式可以計算出詞弧的正確度，再配合詞圖中每個詞弧通過的機率，就可以計算出期望正確度。

圖 4.3 是一個 C_q 與 C_{avg} 的計算範例，在每個詞弧的正確度以及通過機率都已經求出的情況下，列舉出每一條路徑就可以完成 C_q 與 C_{avg} 的計算。圖中的路徑總共有「二千雖然」、「而千雖然」、「而且雖然」三條路徑，其中「二」跟「千 1」的通過路徑是完全相同的「二千雖然」，因此

$$C(\text{二})=C(\text{千1})=Acc(\text{二})+Acc(\text{千1})+Acc(\text{雖然1})=5.54$$

另外「而」跟「千 2」的通過路徑是完全相同的「而千雖然」，因此

$$C(\text{而})=C(\text{千2})=Acc(\text{而})+Acc(\text{千2})+Acc(\text{雖然1})=5.26$$

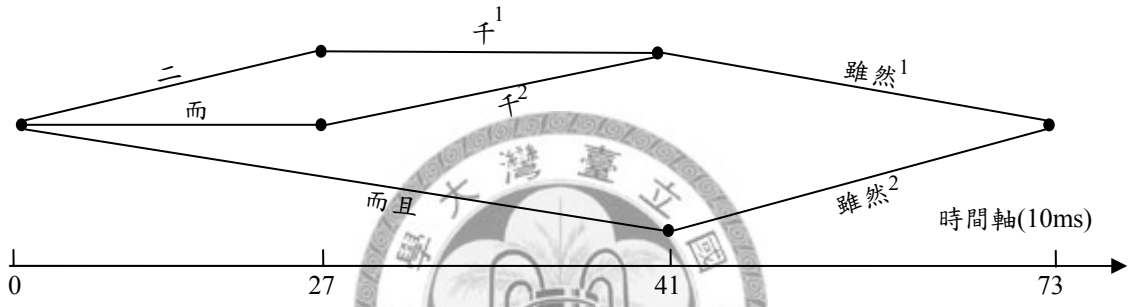
「雖然 1」則有「二千雖然」、「而千雖然」兩條路徑，因此依照兩條路徑的機率加權後相加：

$$\begin{aligned} C(\text{雖然1}) &= \left[Acc(\text{二}) + Acc(\text{千1})\right] \times \left[\Pr(\text{二}) \div \left(\Pr(\text{二}) + \Pr(\text{而})\right)\right] \\ &\quad + \left[Acc(\text{而}) + Acc(\text{千2})\right] \times \left[\Pr(\text{而}) \div \left(\Pr(\text{二}) + \Pr(\text{而})\right)\right] \\ &\quad + Acc(\text{雖然1}) \\ &= 5.40 \end{aligned}$$

其中 $\Pr(\text{二}) + \Pr(\text{而}) = \Pr(\text{雖然1})$ 是必然的結果，因為在雖然 1 的時候兩條路徑是合併的。另外「而且」、「雖然 2」則只有「而且雖然」這條路徑，因此

$$C(\text{而且}) = C(\text{雖然2}) = Acc(\text{而且}) + Acc(\text{雖然2}) = 6.31$$

最後 C_{avg} 則是計算整個詞圖的期望正確率，因此是把每個詞弧的正確率乘上通過



詞弧 Arc	正確率 Acc	通過機率 Prob	$C(q)$ 期望正確率
而	1.28	2.05E-03	Acc(而)+Acc(千 2)+Acc(雖然 1)=5.26
二	1.56	2.28E-03	Acc(二)+Acc(千 1)+Acc(雖然 1)=5.54
而且	3.06	9.96E-01	Acc(而且)+Acc(雖然 2)=6.31
千 2	0.73	2.05E-03	Acc(而)+Acc(千 2)+Acc(雖然 1)=5.26
千 1	0.73	2.28E-03	Acc(二)+Acc(千 1)+Acc(雖然 1)=5.54
雖然 1	3.25	4.33E-03	[Acc(二)+Acc(千 1)]×[Prob(二)÷Prob(雖然 1)] +[Acc(而)+Acc(千 2)]×[Prob(而)÷Prob(雖然 1)] +Acc(雖然 1)=5.40
雖然 2	3.25	9.96E-01	Acc(而且)+Acc(雖然 2)=6.31

C_{avg} = [Acc(二)+Acc(千 1)]×[Prob(二)] + [Acc(而)+Acc(千 2)]×[Prob(而)]
+ Acc(雖然 1)×Prob(雖然 1)
+ Acc(而且)×Prob(而且) + Acc(雖然 2)×Prob(雖然 2)
= 6.31

圖 4.3　C_q 與 C_{avg} 計算範例

機率後加總：

$$\begin{aligned} C_{avg} &= \left[Acc(二)+Acc(千1)\right]\times \Pr(二)+\left[Acc(而)+Acc(千2)\right]\times \Pr(而) \\ &\quad +Acc(雖然1)\times \Pr(雖然1) \\ &\quad +Acc(而且)\times \Pr(而且)+Acc(雖然2)\times \Pr(雖然2) \\ &= 6.31 \end{aligned}$$

然而在實作上，窮舉出所有的路徑來計算是個不切實際的作法，因此通常會使用詞圖前向後向演算法(forward-backward algorithm)，分別對每個詞弧計算其前接路徑及後接路徑的平均正確率，再以此計算出 C_q 和 C_{avg}。

4.2.4 詞圖前向後向演算法

在(4.18)中的 γ_q^{MPE} 可以利用詞圖的前向後向演算法求得，計算方式如圖 4.4、圖 4.5 和圖 4.6。圖中 $P_\lambda(O|q)$ 為詞弧 q 的聲學相似度，$P_\Gamma(q|r)$ 是詞弧 r 連接到詞弧 q 的機率；α_q 為前向相似度，β_q 為後向相似度，$\bar{\alpha}_q$ 為詞弧 q 的所有前接路徑連接到 q 的期望正確度，$\bar{\beta}_q$ 為詞弧 q 的所有後續路徑由 q 之後連接過去的期望正確度；$\sum_{r\in q\text{的前接詞弧}}$ 代表對每個可以連接到詞弧 q 的詞弧累加，$\sum_{r\in q\text{的後續詞弧}}$ 代

for 詞圖中的每一個詞弧 q，依照詞弧開始時間由小到大的順序

  if q 為起始詞弧 (q 沒有前接的詞弧)

$$\alpha_q = PhoneAcc(q)$$

$$\bar{\alpha}_q = P_\lambda(O|q)$$

  else

$$\alpha_q = \sum_{r\in q\text{的前接詞弧}} \alpha_r P_\Gamma(q|r) P_\lambda(O|q)$$

$$\bar{\alpha}_q = \frac{\sum_{r\in q\text{的前接詞弧}} \bar{\alpha}_r \alpha_r P_\Gamma(q|r)}{\sum_{r\in q\text{的前接詞弧}} \alpha_r P_\Gamma(q|r)} + PhoneAcc(q)$$

  end

end

圖 4.4　詞圖前向演算法

表對每個可以由詞弧 q 連接過去的詞弧累加。

for 詞圖中的每一個詞弧 q ，依照詞弧開始時間由大到小的順序

if q 為結尾詞弧 (q 沒有後續的詞弧)

$$\beta_q = 1$$

$$\bar{\beta}_q = 0$$

else

$$\beta_q = \sum_{r \in q\text{的後續詞弧}} P_\Gamma(r|q) P_\lambda(O|r) \beta_r$$

$$\bar{\beta}_q = \frac{\sum_{r \in q\text{的後續詞弧}} P_\Gamma(r|q) P_\lambda(O|r) \beta_r \left(\bar{\beta}_r + PhoneAcc(r)\right)}{\sum_{r \in q\text{的後續詞弧}} P_\Gamma(r|q) P_\lambda(O|r) \beta_r}$$

end

end

圖 4.5 詞圖後向演算法

$$x = \sum_{q\text{為結尾的詞弧}} \alpha_q$$

$$C_{avg} = \frac{\sum_{q\text{為結尾的詞弧}} \bar{\alpha}_q \alpha_q}{x}$$

for 詞圖中的每一個詞弧 q

$$\gamma_q = \frac{\alpha_q \beta_q}{x}$$

$$C_q = \bar{\alpha}_q + \bar{\beta}_q$$

$$\gamma_q^{MPE} = \gamma_q \left(C_q - C_{avg}\right)$$

end

圖 4.6 根據詞圖前向後向演法計算 C_q 與 C_{avg}

對照 4.2.3 節的說明來看，α_q 和 β_q 可以視為在計算每個詞弧通過的機率，分別由頭尾開始計算，每一步的詞弧通過機率都利用在上一步已經算出的前接或後續詞弧的機率計算出來；$\bar{\alpha}_q$ 和 $\bar{\beta}_q$ 則可視為在計算機率時同時也乘上詞弧正確度，

成為詞弧正確度的期望值，同樣也在每一步利用上一步已經算出的前接或後續詞弧的詞弧正確度的期望值來算出目前詞弧正確度的期望值。在整個詞圖的詞弧都計算完成之後，就可以依照圖 4.6 的方式計算出C_{avg}、γ_q、C_q和γ_q^{MPE}。

另外，在圖 4.6 中的x代表的意義為該觀測語句O的事前機率，也就是目標函數(4.1)式中的$\sum_{u\in W_h} P_\lambda(O_r \mid u) P_\Gamma(u)$；而$\gamma_q = \alpha_q \beta_q / x$則是詞弧$q$在整句語句中的事後機率。$C_{avg}$為所有結尾詞弧的前向相似度$\overline{\alpha}_q$的期望正確度，由於$\overline{\alpha}_q$本來就是代表詞弧$q$的所有前接路徑的期望正確度，因此$C_{avg}$等於是由結尾詞弧把全部的期望正確度都加總起來，於是C_{avg}就是整句觀測語句O_r的期望正確度。

4.3 實驗結果

本實驗使用 3.3 節基礎實驗的聲學模型作為初使模型進行最小音素錯誤聲學模型訓練，由 4.1.1 節的推導可知最小音素錯誤訓練法是經由反覆疊代來調整聲學模型的，因此本實驗會經過多次訓練的疊代，每一次疊代都是把上一次疊代的訓練結果做為新的初始模型。實驗中也使用了這兩種不同的詞圖，分別為 4.2.1 節說明的兩種詞圖產生方式；另外在 4.1.4 節說明的 I 平滑，需要一個平滑係數τ，本實驗也測試了許多不同的 τ 值。

實驗結果如表 4.1～表 4.8 及圖 4.7～圖 4.14 所示：實驗使用的詞圖分為詞圖 N 以及詞圖 T 兩種，詞圖 N 代表由 NDecoder 產生，為詞弧保留原則為分數排名的詞圖；詞圖 t 代表由 TTK 產生，為詞弧保留原則為分數比值的詞圖。兩種詞圖都測試了多個平滑係數 τ 值：5、10、25、100、200、400，表中的 itr 代表疊代的次數。

實驗結果使用詞、字、音節、聲韻母四種層級來呈現，在正確率的表現上，最小音素錯誤訓練法在四種層級上都有進步，而且在疊代初期時的進步量較多，到接近最後幾次的疊代時，進步就會趨緩，之後正確率便不再提升，甚至有出現下降的情形。正確率在疊代數次達到最大值後下降，在鑑別式訓練法中是十分普遍的現象，一般認為是因過度訓練(over training)所造成的，過度訓練是指模型訓練

得與訓練集太過匹配而失去了一般性，導致對評估集的辨識率下降。

平滑係數在 25 時的詞正確率與字正確率為實驗中的最大值，平滑係數為 10 時的音節與聲韻母正確率為實驗中的最大值。在詞圖 N 的實驗中，平滑係數 25 時，字正確率會進步 2.46%(相對 9.91%)；在詞圖 T 的實驗中，平滑係數 25 時，字正確率會進步 2.31%(相對 9.30%)。

實驗中發現不同的平滑係數對於正確率的影響，若使用的平滑係數的數值較接近得到正確率最高時的平滑係數的數值時，就會有較高的正確率，反之在使用的平滑係數的數值較遠離得到正確率最高時的平滑係數的數值時，就會有較差的正確率，因此在經驗上可以假設在正確率最高時的平滑數值十分接近最佳值。波式(Povey)在【32】中提出關於平滑係數的看法，認為平滑係數的最佳值與分子詞圖和分母詞圖的統計值(即(4.26)式中的$\theta_m^{num}(O)$與(4.27)式中的$\theta_m^{den}(O)$)成正比，因此在本論文之後的實驗中，會基於這個假設選擇平滑係數的數值，減少調整參數的時間。

4.4 本章結論

本章詳細介紹了最小音素錯誤訓練法，包括目標函數的最佳化過程到模型參數的更新，音素正確度的估算到詞弧正確度的計算方式，使用詞圖做為辨識可能近似的演算法。由實驗結果可以看出，在公視新聞的語料上，最小音素錯誤訓練法在詞、字、音節、聲韻母四種層級上的正確率都有進步。

MPE			詞圖 N		詞正確率(%)	
itr	τ=5	τ=10	τ=25	τ=100	τ=200	τ=400
0	57.99	57.99	57.99	57.99	57.99	57.99
1	58.56	58.68	58.63	58.66	58.75	58.58
2	59.31	59.38	59.48	59.13	59.09	58.97
3	59.87	59.97	59.88	59.47	59.18	58.89
4	59.88	60.17	60.29	59.77	59.18	59.08
5	59.98	60.29	60.38	59.82	59.39	59.06
6	60.06	60.43	60.73	59.92	59.59	59.09
7	60.20	60.50	60.59	60.10	59.58	59.08
8	60.26	60.67	60.79	60.18	59.51	59.16
9	60.13	60.74	60.78	60.15	59.66	59.07
10	59.92	60.60	**60.87**	60.19	59.56	59.09

表 4.1　最小音素錯誤訓練法－詞圖 N－詞正確率

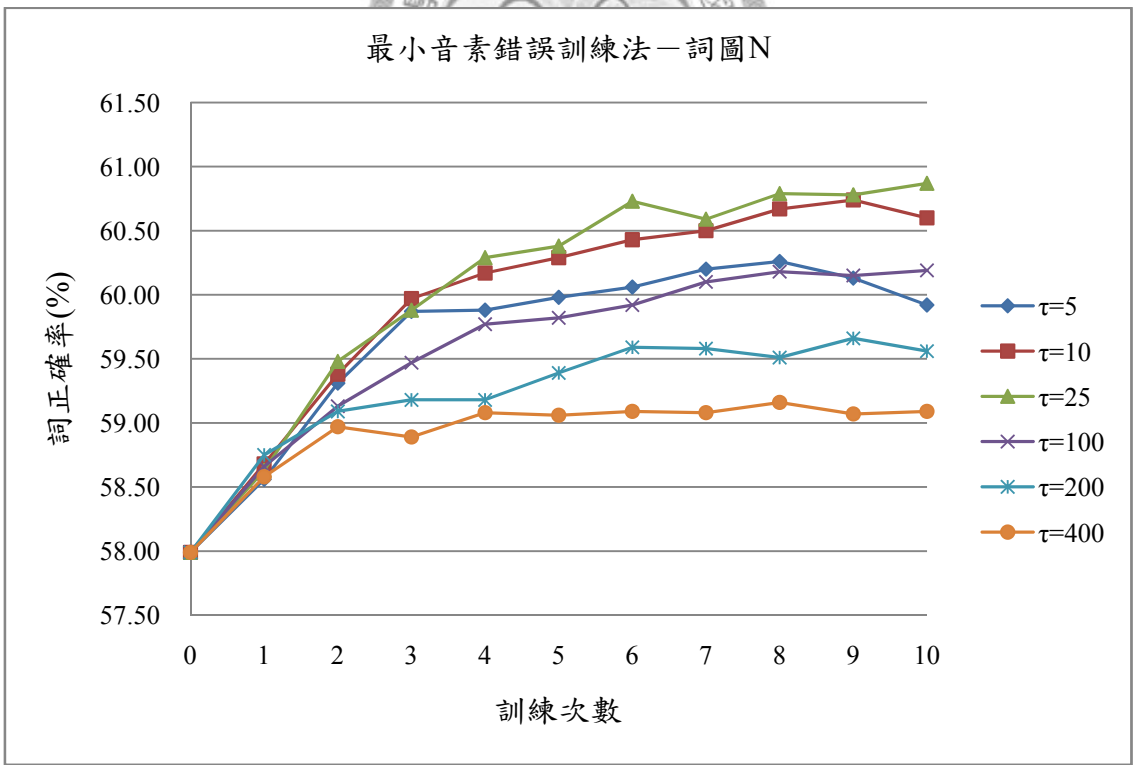


圖 4.7　最小音素錯誤訓練法－詞圖 N－詞正確率

MPE			詞圖 N		字正確率(%)	
itr	τ=5	τ=10	τ=25	τ=100	τ=200	τ=400
0	75.17	75.17	75.17	75.17	75.17	75.17
1	75.79	75.89	75.84	75.84	75.86	75.64
2	76.52	76.57	76.58	76.17	76.12	75.97
3	77.01	77.10	77.03	76.49	76.21	75.94
4	77.02	77.26	77.39	76.69	76.23	76.08
5	77.09	77.32	77.41	76.70	76.35	76.07
6	77.13	77.34	77.60	76.83	76.52	76.10
7	77.18	77.34	77.46	77.00	76.52	76.19
8	77.10	77.46	77.62	77.10	76.59	76.24
9	76.81	77.45	**77.63**	77.13	76.74	76.25
10	76.60	77.19	77.61	77.18	76.63	76.21

表 4.2　最小音素錯誤訓練法－詞圖 N－音節正確率

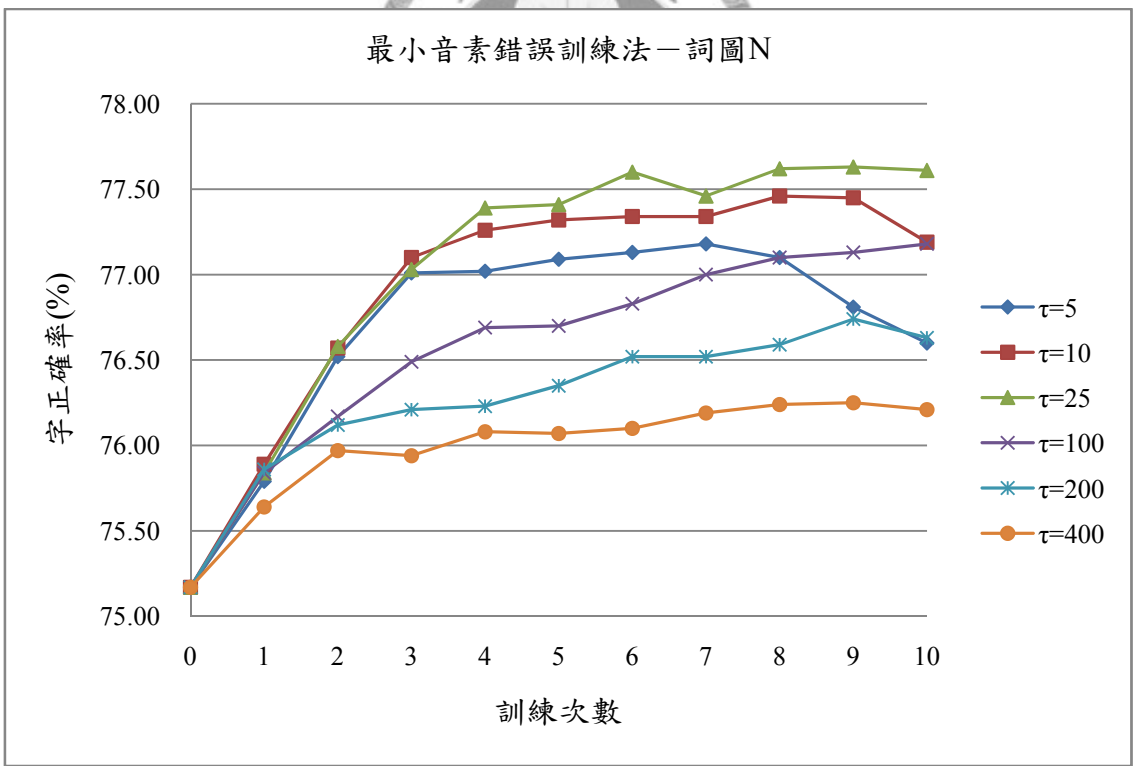


圖 4.8　最小音素錯誤訓練法－詞圖 N－字正確率

MPE			詞圖 N		音節正確率(%)	
itr	τ=5	τ=10	τ=25	τ=100	τ=200	τ=400
0	81.42	81.42	81.42	81.42	81.42	81.42
1	82.06	82.17	82.15	82.09	82.04	81.82
2	82.76	82.81	82.80	82.34	82.27	82.15
3	83.22	83.30	83.23	82.64	82.39	82.11
4	83.38	83.54	83.58	82.87	82.40	82.20
5	83.46	83.63	83.65	82.89	82.52	82.22
6	83.60	83.65	83.82	83.04	82.69	82.23
7	83.68	83.68	83.77	83.19	82.70	82.30
8	83.68	**83.90**	83.85	83.33	82.75	82.34
9	83.32	83.87	83.84	83.32	82.85	82.33
10	83.04	83.56	83.86	83.33	82.76	82.28

表 4.3　最小音素錯誤訓練法－詞圖 N－音節正確率

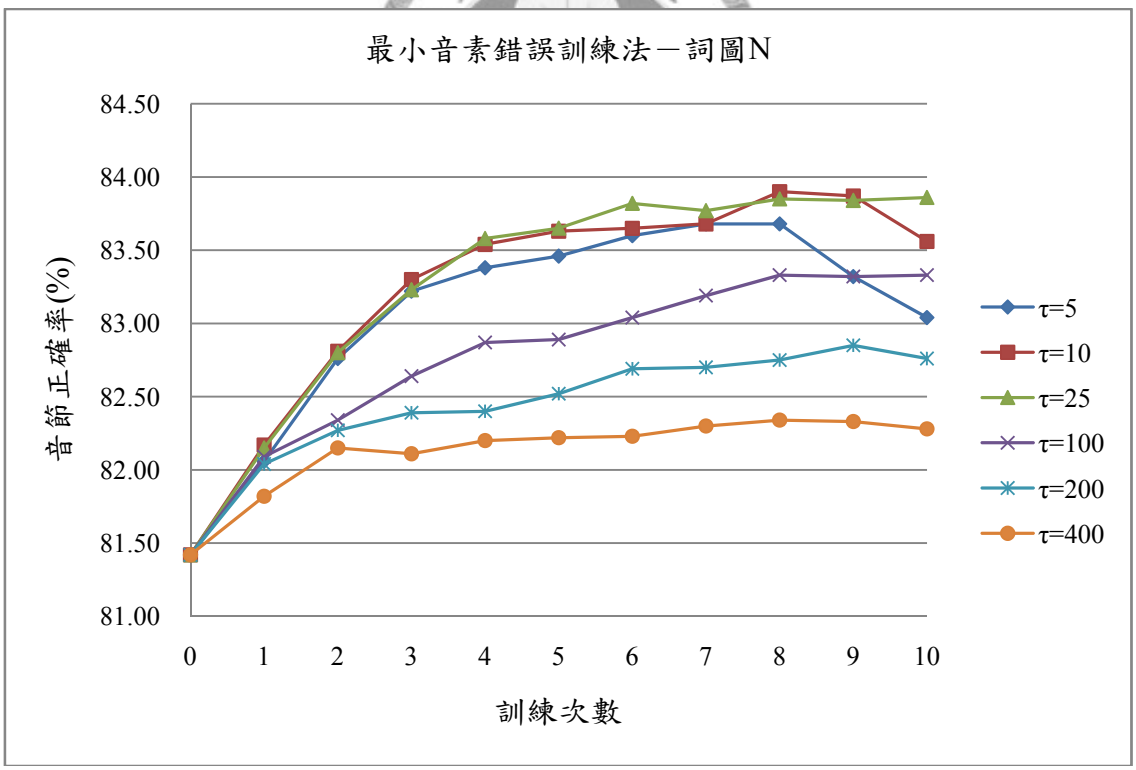


圖 4.9　最小音素錯誤訓練法－詞圖 N－音節正確率

MPE			詞圖 N		聲韻母正確率(%)	
itr	τ=5	τ=10	τ=25	τ=100	τ=200	τ=400
0	84.76	84.76	84.76	84.76	84.76	84.76
1	85.40	85.48	85.47	85.36	85.29	85.11
2	85.93	85.97	86.00	85.62	85.51	85.41
3	86.37	86.44	86.38	85.89	85.66	85.42
4	86.51	86.66	86.66	86.07	85.68	85.50
5	86.58	86.73	86.72	86.09	85.80	85.52
6	86.70	86.78	86.88	86.21	85.94	85.54
7	86.72	86.73	86.83	86.33	85.94	85.60
8	86.74	**86.92**	86.87	86.42	86.01	85.64
9	86.37	86.90	86.86	86.39	86.06	85.60
10	86.19	86.56	86.88	86.41	86.00	85.55

表 4.4　最小音素錯誤訓練法－詞圖 N－聲韻母正確率

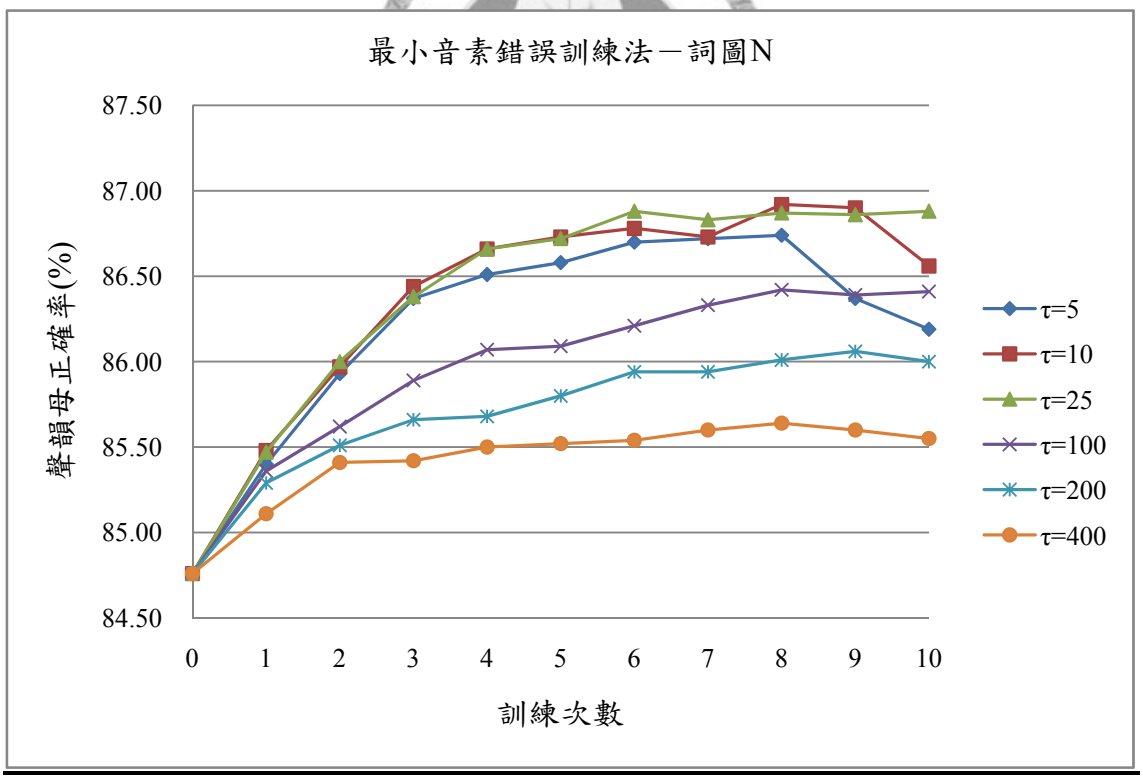


圖 4.10　最小音素錯誤訓練法－詞圖 N－聲韻母正確率

MPE			詞圖 T		詞正確率(%)	
itr	τ=5	τ=10	τ=25	τ=100	τ=200	τ=400
0	57.99	57.99	57.99	57.99	57.99	57.99
1	58.46	58.57	58.68	58.83	58.65	58.57
2	59.12	59.24	59.26	59.26	59.14	58.88
3	59.70	59.93	59.93	59.49	59.13	58.95
4	60.03	60.38	60.29	59.75	59.26	58.88
5	59.95	60.27	60.49	60.00	59.41	58.97
6	60.24	60.43	60.66	60.11	59.40	59.05
7	60.26	60.54	60.90	60.28	59.42	59.18
8	60.27	60.61	60.85	60.23	59.46	59.18
9	60.49	60.84	60.97	60.19	59.75	59.33
10	60.58	60.84	**61.03**	60.22	59.68	59.33

表 4.5　最小音素錯誤訓練法－詞圖 T－詞正確率

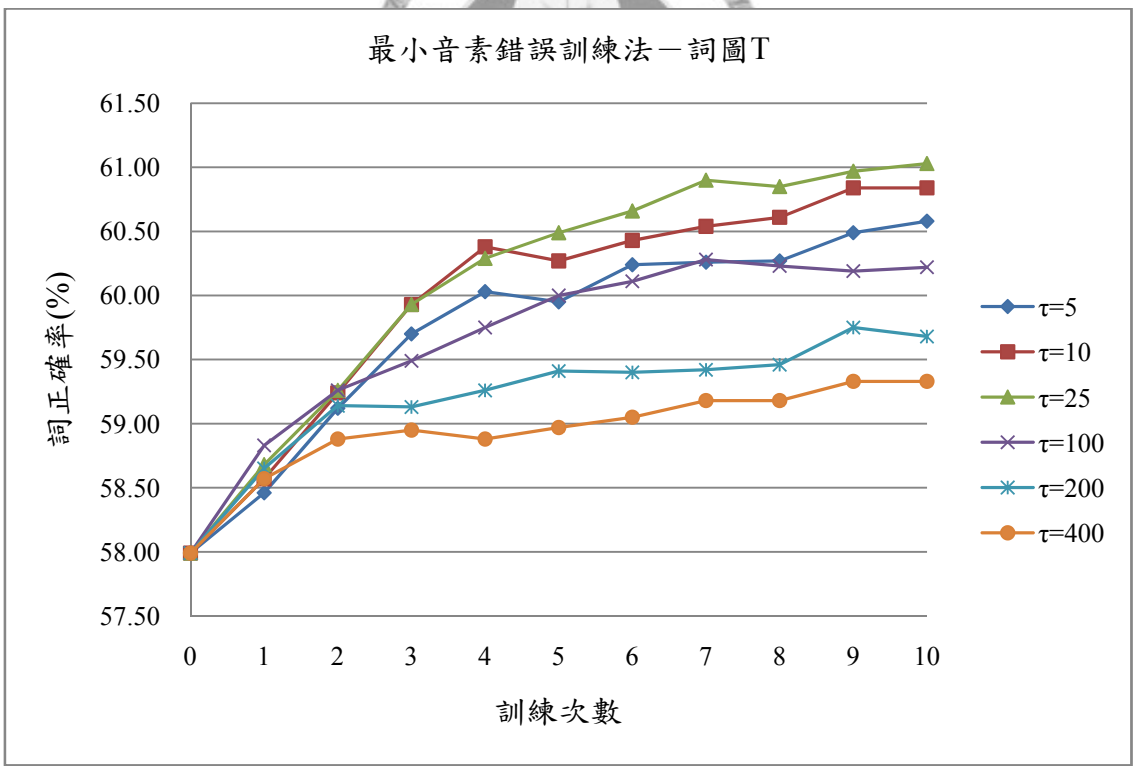


圖 4.11　最小音素錯誤訓練法－詞圖 T－詞正確率

MPE			詞圖 T		字正確率(%)	
itr	τ=5	τ=10	τ=25	τ=100	τ=200	τ=400
0	75.17	75.17	75.17	75.17	75.17	75.17
1	75.62	75.73	75.87	75.91	75.72	75.61
2	76.13	76.28	76.31	76.26	76.12	75.92
3	76.76	76.91	76.85	76.46	76.18	75.96
4	76.99	77.24	77.21	76.63	76.29	75.92
5	76.87	77.16	77.22	76.87	76.36	75.94
6	77.12	77.18	77.32	76.97	76.36	76.02
7	77.04	77.22	**77.48**	76.99	76.35	76.13
8	76.97	77.21	77.37	76.89	76.43	76.15
9	76.98	77.31	77.45	76.93	76.57	76.27
10	76.84	77.23	77.47	76.92	76.50	76.29

表 4.6　最小音素錯誤訓練法－詞圖 T－字正確率

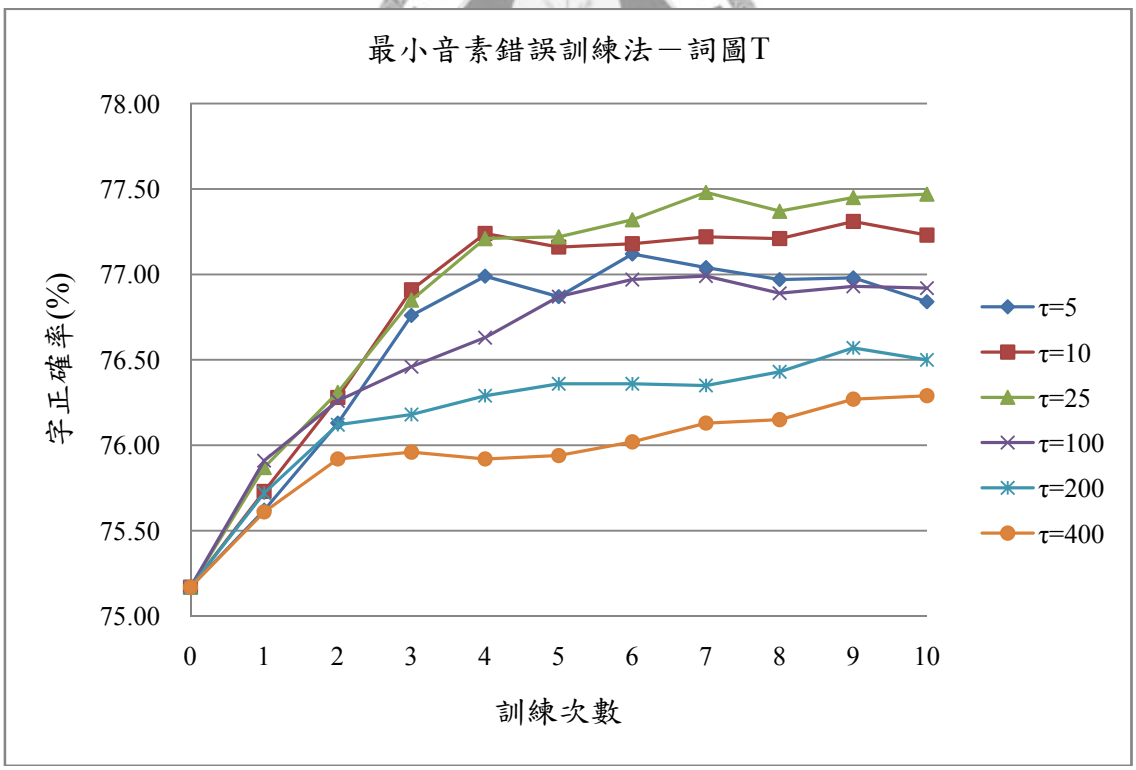


圖 4.12　最小音素錯誤訓練法－詞圖 T－字正確率

MPE			詞圖 T		音節正確率(%)	
itr	τ=5	τ=10	τ=25	τ=100	τ=200	τ=400
0	81.42	81.42	81.42	81.42	81.42	81.42
1	81.97	82.06	82.14	82.11	81.95	81.81
2	82.48	82.56	82.59	82.45	82.26	82.05
3	83.08	83.17	82.99	82.63	82.34	82.10
4	83.31	83.43	83.35	82.77	82.42	82.10
5	83.33	83.52	83.35	82.96	82.51	82.14
6	83.56	83.62	83.49	83.03	82.54	82.20
7	83.54	83.65	83.58	83.05	82.52	82.28
8	83.50	83.68	83.54	83.01	82.59	82.26
9	83.50	**83.77**	83.65	83.05	82.68	82.33
10	83.41	83.70	83.71	83.06	82.61	82.34

表 4.7　最小音素錯誤訓練法－詞圖 T－音節正確率

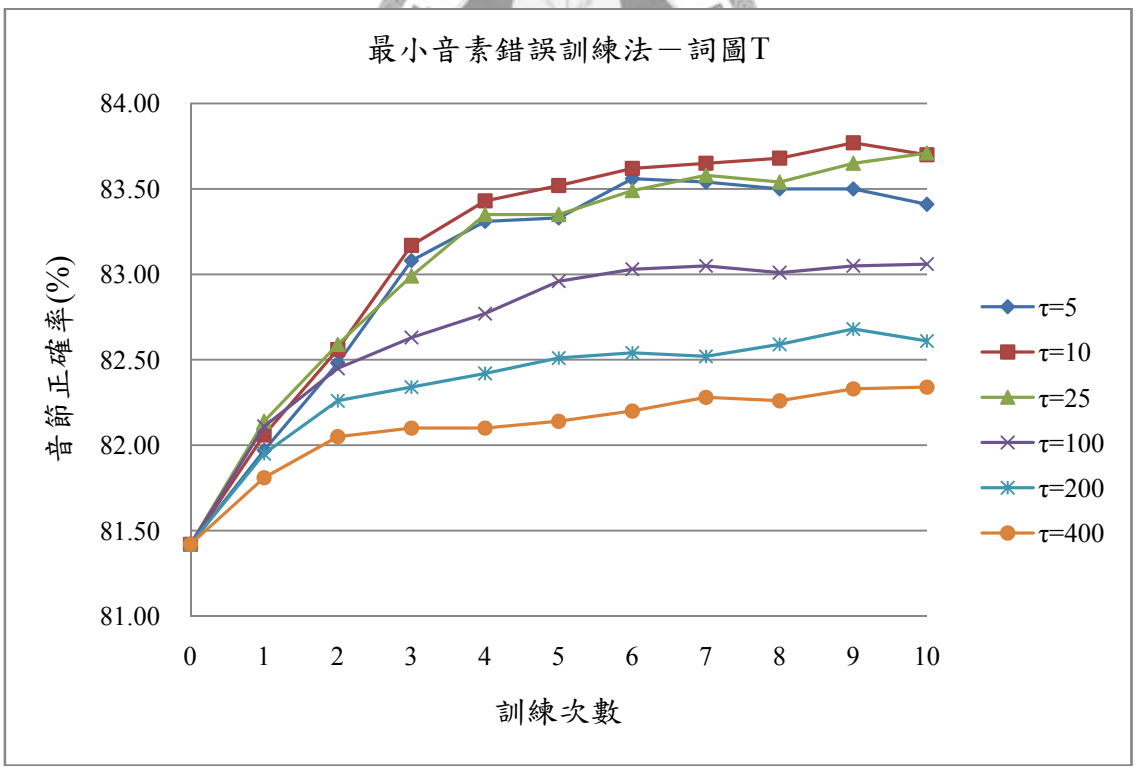


圖 4.13　最小音素錯誤訓練法－詞圖 T－音節正確率

MPE			詞圖 T		聲韻母正確率(%)	
itr	τ=5	τ=10	τ=25	τ=100	τ=200	τ=400
0	84.76	84.76	84.76	84.76	84.76	84.76
1	85.27	85.34	85.42	85.38	85.25	85.13
2	85.69	85.72	85.78	85.69	85.51	85.33
3	86.23	86.27	86.12	85.83	85.62	85.41
4	86.48	86.52	86.45	85.96	85.70	85.41
5	86.49	86.65	86.45	86.12	85.77	85.45
6	86.69	86.73	86.58	86.16	85.77	85.50
7	86.69	86.74	86.66	86.17	85.78	85.56
8	86.68	86.77	86.62	86.15	85.82	85.54
9	86.68	**86.84**	86.71	86.18	85.91	85.60
10	86.63	86.80	86.74	86.21	85.87	85.61

表 4.8　最小音素錯誤訓練法－詞圖 T－聲韻母正確率

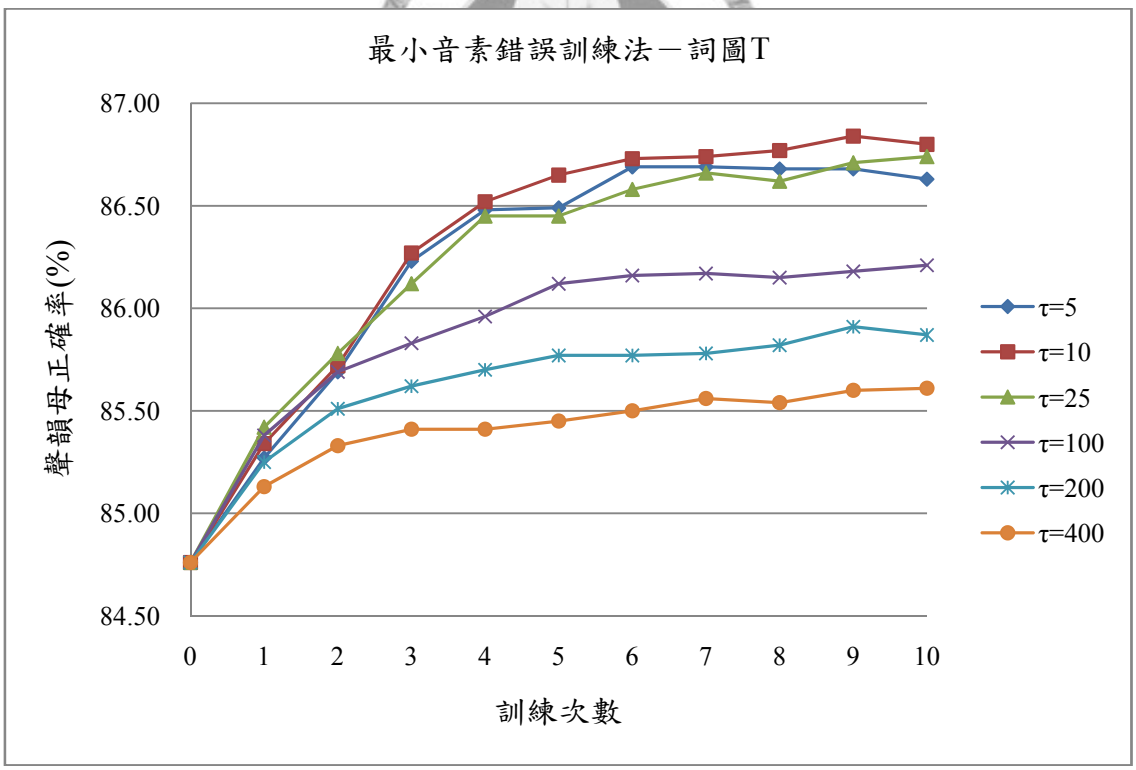


圖 4.14　最小音素錯誤訓練法－詞圖 T－聲韻母正確率

第5章　基於最小音素錯誤改進之鑑別式訓練法

本章將介紹一些基於最小音素錯誤修改而來的鑑別式訓練法，包括最小音素音框錯誤(Minimum Phone Frame Error, MPFE)、狀態層級最小貝氏風險(physical state level Minimum Bayes Risk, sMBR)、最小歧異度(Minimum Divergence)，以及這三種方法再進一步的修改版本。

5.1　最小音素音框錯誤訓練

5.1.1　目標函數

最小音素音框錯誤是由鄭氏(Zheng)等人在 2005 年提出的方法【4】，基本上的目標函數與最小音素訓練法(4.1)式相同：

$$F_{MPFE}(\lambda)=\sum_{r}\sum_{u}P(u|O_r)\,Acc(s_r,u) \tag{5.1}$$

與最小音素訓練法的差異在於正確度 $Acc(s_r,u)$ 的計算方式有所不同，最小音素音框錯誤計算的正確度為音素音框正確度，首先定義每個音素個別的正確度：

$$PhoneFrameAcc(q)=\sum_{t=start(q)}^{end(q)}\delta\left(q,s_{r,phone}(t)\right) \tag{5.2}$$

$$\delta\left(q,s_{r,phone}(t)\right)=\begin{cases}1 & \text{if } q=s_{r,phone}(t)\\ 0 & \text{if } q\neq s_{r,phone}(t)\end{cases} \tag{5.3}$$

其中音素 q 是辨識結果，$start(q)$ 和 $end(q)$ 分別表示音素 q 以音框為單位的開始與結束時間；$s_{r,phone}(t)$ 表示正確轉寫在時間 t 時的音素。而整個文句的音素音框正確度，就是將每個音素個別的正確度加總：

$$Acc(s_r,u)=\sum_{q\in u}PhoneFrameAcc(q) \tag{5.4}$$

音素音框正確度的計算，就是在比對文句 u 與正確轉寫 s_r 的每個音框，計算音框所屬的音素相同的數量，因此在實作上，最小音素音框錯誤訓練的詞弧正確度計算方式，就是比對詞弧區間的正確轉寫，然後累加辨識音素正確的音框數量。

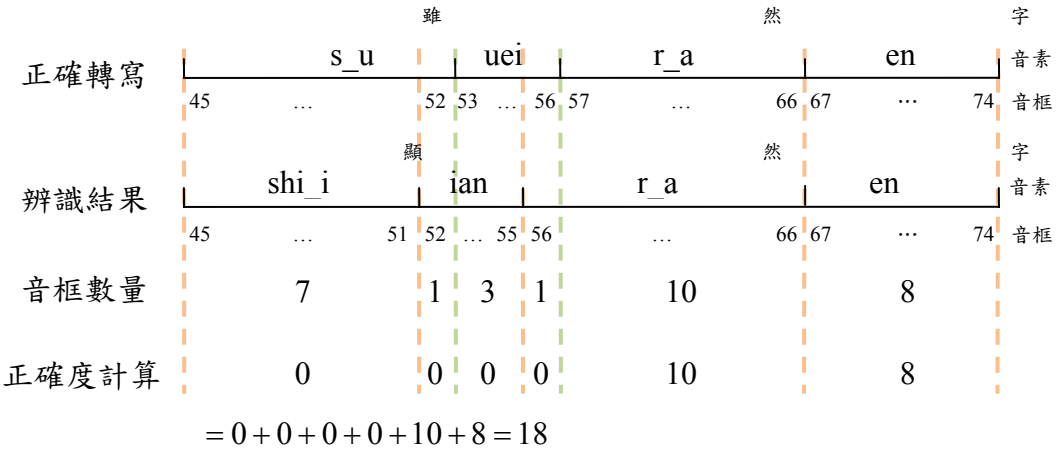


圖 5.1　音素音框正確度的計算範例

圖 5.1 是一個音素音框正確度的計算範例，圖中的詞弧長度是 45～74 共 30 個音框，其中 45 至 55 的音框為 shi_i 與 ian，辨識的音素皆與正確轉寫不同，所以正確度為 0；576 至 74 的音框為 r_a 與 en，辨識的音素有 28 個音框與正確轉寫相同，每個音框記正確度 1，總共 28 個音框，所以正確度為 28。整段詞弧的正確音框數量 28 個，故詞弧正確度為 28。

音素音框正確度不同於音素正確度，並不是編輯距離計算的精確正確度的近似，在 4.2 節的說明可以看出，詞弧正確度的功能是為了將詞弧區分成分子詞圖與分母詞圖，意即要分辨每個詞弧是較接近正確而要做正向訓練，還是較遠離正確而要做負向訓練，因此詞弧正確度的計算，不一定需要跟精確正確度的計算結果相似，只要區分出的較正確詞弧與較不正確詞弧跟精確正確度的區分結果相似即可。

音素音框正確度可以改善音素正確度對於遺失(deletion)處罰不夠的問題。如圖 5.2 是一個音素正確率與音素音框正確率比較的一個例子，圖中的(a)是有一個插入錯誤的例子，(b)是有一個遺失錯誤的例子。精確正確度的算法，在兩種錯誤的情況下正確度都是 2；音素音框正確度的算法，在兩種錯誤的情況下也會相同，正確度為 18；然而在音素正確度的算法下，插入錯誤的正確度為 1，刪除錯誤的正確度為 2，相差 2 倍。由此可以發現音素正確度對於插入錯誤的處罰較遺失錯誤的處罰為重，音素音框正確度則無此問題。探討原因會注意到，音素正確度的計算方

式(4.32)式，無論辨識的音素對或錯，正確度都會先加上-1，於是辨識結果中每有一個音素正確度都會-1，自然辨識結果的音素數量越多，正確度就會有越低的傾向，所以會有偏向處罰插入錯誤較遺失錯誤重的情況，單純是因為插入錯誤發生時的音素數量較多的緣故。如此可能造成最小音素錯誤訓練偏向增加遺失錯誤以減少插入錯誤的現象。

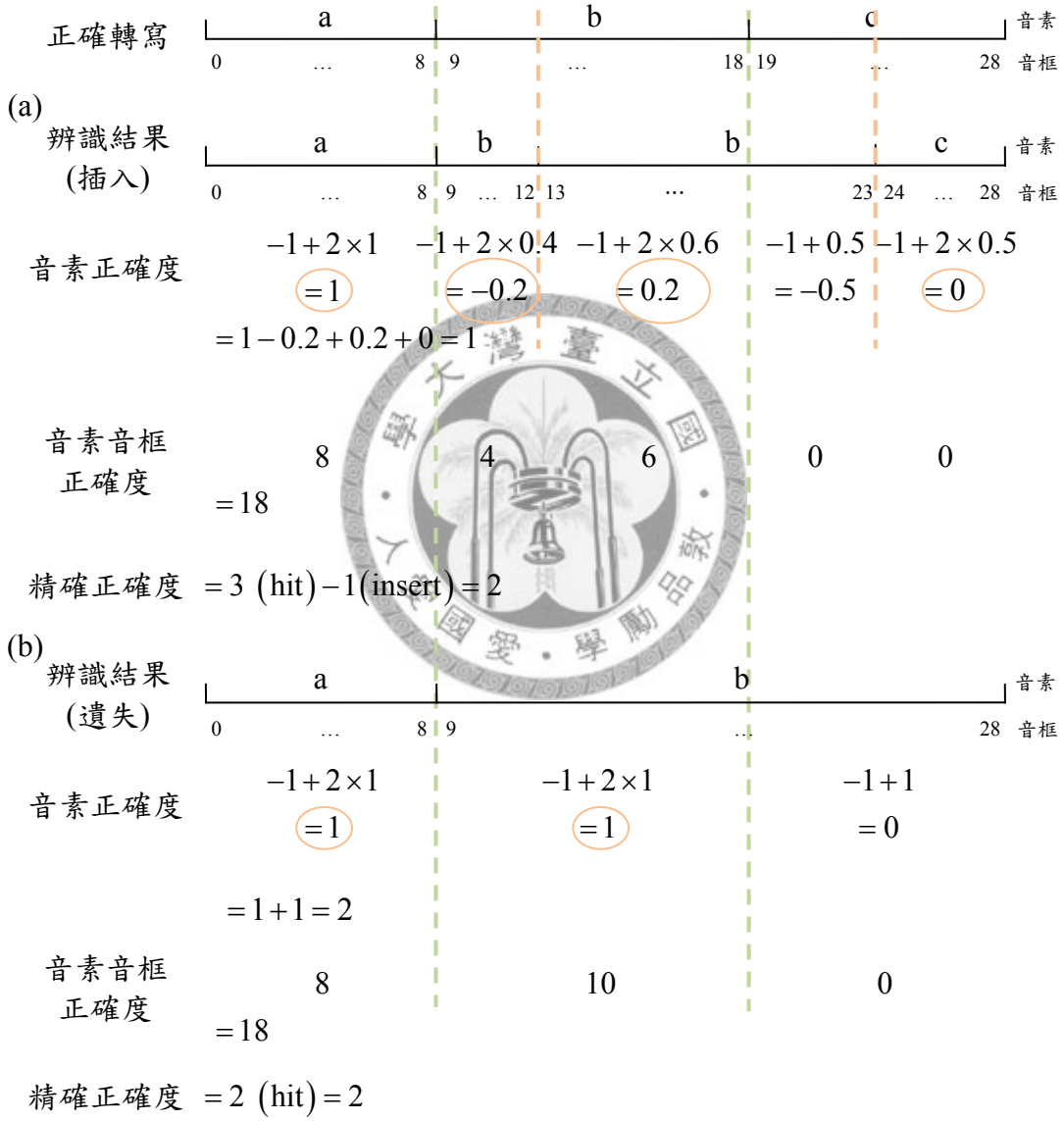


圖 5.2　音素正確度與音素音框正度之比較範例

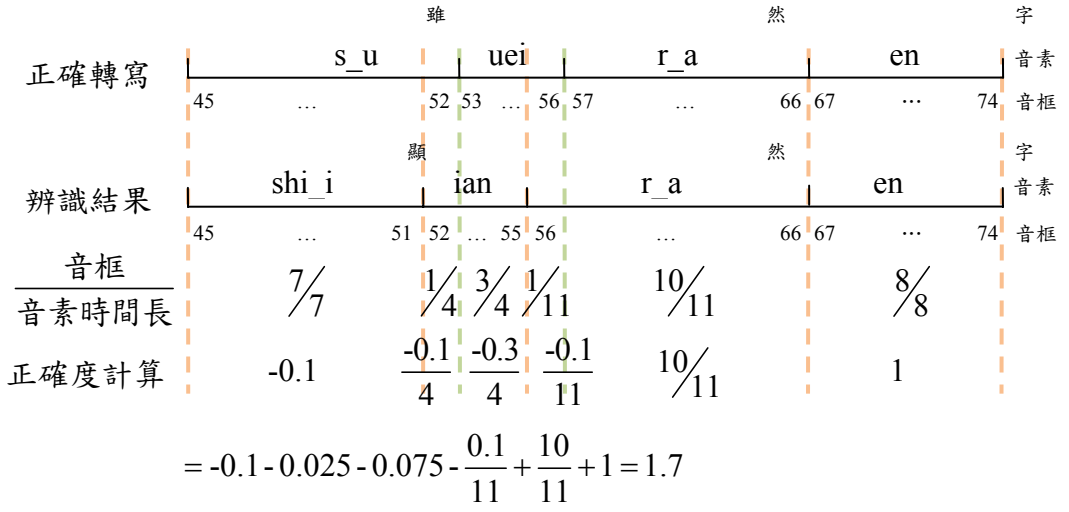
5.1.2 加入錯誤處罰與音素長度正規化的詞弧正確度

5.1.1 節所討論的音素音框正確度，為計算辨識之音素正確的音框數，對於辨識錯誤的情況並未施與處罰，並且音框正確數量的計算，會使得時間較長的音素(即音框數量較多的音素)，對於正確度有較大的影響。因此將原本的音素音框正確度加以修改，加入錯誤處罰與音素長度的正規化【33】，原本(5.2)與(5.3)式各別音素正確度的計算方式就改為：

$$PhoneFrameAcc(q)=\sum_{t=start(q)}^{end(q)}\frac{\delta\left(q,s_{r,phone}(t)\right)}{end(q)-start(q)+1} \tag{5.5}$$

$$\delta\left(q,s_{r,phone}(t)\right)=\begin{cases}1 & \text{if } q=s_{r,phone}(t)\\ -\rho & \text{if } q\neq s_{r,phone}(t)\end{cases} \tag{5.6}$$

其中 $s_{r,phone}(t)$ 代表正確轉寫 s_r 在時間 t 時的音素；ρ 是錯誤處罰的權重，ρ 的選擇必須介於 0 到 1 之間，在本論文中依照【33】選擇 $\rho=0.1$。修改後的音素音框正確度在分別計算每一個音素的時候，除了加總辨識正確的音框總數之外，辨識錯誤的音框每一個也要給與 $-\rho$ 的錯誤處罰，最後的總和再除以辨識音素的時間長度(即(5.5)式中的 $end(q)-start(q)+1$)，即音素長度正規化。如此一來，每個音素的正確度就會落在 $-\rho$ 到 1 的範圍之間。



$$=-0.1-0.025-0.075-\frac{0.1}{11}+\frac{10}{11}+1=1.7$$

圖 5.3 加入錯誤處罰與音素長度正規化的音素音框正確度的計算範例

圖 5.3 是一個加入錯誤處罰與音素長度正規化的音素音框正確度的計算範例，

圖中 shi_i 的長度為 7 個音框，皆與正確轉寫不同，因此正確度為 7/7×-0.1=-0.1；ian 長度為 4 個音框，皆與正確轉寫不同，因此正確度為 4/4×-0.1=-0.1；r_a 長度為 11 個音框，其中一個音框與正確轉寫不同，剩下 10 個音框則相同，因此正確率為 1/11×-0.1+10/11=0.9；en 長度為 8 個音框，皆與正確轉寫相同，因此正確率為 8/8=1；整段詞弧的正確率就是這 4 個音素正確度的加總等於 1.7。

5.1.3 實驗結果

本實驗為最小音素音框錯誤訓練法，以及加入錯誤處罰與音素長度正規化的實驗，實驗的設定基本上與 4.3 節的最小音素錯誤訓練的實驗相同，不過因為最小音素音框錯誤訓練的詞弧正確度範圍跟最小音素錯誤訓練的有一段差距，因此本實驗測式的平滑係數 τ 與最小音素錯誤訓練所使用的數值不能相同，這裡測試的 τ 值為：10、100、150、200、250、400，測試這些數值的一方面也是為了驗證【32】中平滑係數的選擇方式，經由測試不同的平滑係數尋找到最佳值，比較與從最小音素訓練找到的最佳值推測來的最佳值有何差異。

最小音素音框錯誤的實驗結果如表 5.2～表 5.9 及圖 5.4～圖 5.11 所呈現，由於 τ=10 的結果明顯較差，所以不畫入圖中。實驗結果使用詞、字、音節、聲韻母四種層級來呈現，在正確率的表現上，詞正確率的進步較明顯，而且直到最後幾次疊代時才有進步漸緩的趨勢，然而字、音節、聲韻母層級的正確率，則在疊代 4 次左右就已經飽和，之後便不再進步。換言之，字、音節、聲韻母層級的過度訓練發生得比詞層級早得多，由此可以推論，最小音素音框訓練對於詞正確率

訓練方法	詞圖	分子詞圖統計	分母詞圖統計	平滑係數 τ
MPE	N	1.005032E+06	1.005032E+06	25.00
MPE	T	8.953452E+05	8.953498E+05	22.27
MPFE	N	8.559231E+06	8.559231E+06	212.91
MPFE	T	7.646058E+06	7.646075E+06	190.19
MPFE_pen	N	9.895510E+05	9.895510E+05	24.61
MPFE_pen	T	8.859587E+05	8.859611E+05	22.04

表 5.1　平滑係數最佳值之估測

的效果較佳，字正確率的提升則效果有限。在詞圖 N 的實驗中，τ=200 時有最高的字正確率 76.37%，進步 1.20%(相對 4.83%)；而在詞圖 T 的情況下，τ=250 時有最高的字正確率 76.22%，進步 1.05%(相對 4.23%)。

表 5.1 是關於平滑係數τ最佳值的估測，表中 MPE 代表最小音素錯誤訓練法，MPFE 代表最小音素音框錯誤訓練法，MPFE+pen+len 代表最小音素音框錯誤訓練法正確度計算加入錯誤處罰與音素正規化的版本。這個表是建立在假設τ的最佳值與分子分母的詞圖統計值成正比的關係下所作的估測【32】，在 4.3 節的實驗中，最小音素錯誤訓練法在詞圖 N 找到的τ最佳值為 25，表中的該值由此而來，之下的估測則是假設此值為最佳值下所做的預估。如表中所估測的，詞圖 T 的最小音素錯誤訓練法估測τ最佳值為 22.27，實驗中τ最佳值為 25，算是十分接近；另外在最小音素音框錯誤訓練法中，詞圖 N 的估測τ最佳值為 212.91，與實驗中τ最佳值為 200 亦相去不遠，而在詞圖 T 的估測τ最佳值為 190.19，與實驗中τ的最佳值 250 相差有一段距離，不過在 MPFE 的詞圖 T 實驗中，觀察在不同τ值時的字正確率，會發現雖然在 τ=250 時最高，但次高者為 τ=150 與 τ=100 的時候，τ=200 時反而是位居第三，τ對於正確率的影響出現不連續的情況，這裡本論文假設是一個例外情況，因為 τ=150 與 τ=100 時的最高正確率相差僅 0.01%，在語音辨識裡可視為誤差範圍，以此推論在最小音素音框錯誤訓練法詞圖 T 的實驗中，τ最佳值為介於 150 至 250 之間亦算合理，因此估測τ最佳值為 190.19 也算是有效的估測。

另外分子詞圖與分母詞圖的統計值，理論上會是相同的值，因為分子與分母詞圖的區分方式，是把詞圖中全部的詞弧路徑的正確度期望值計算出來後，把詞弧正確度高於平均的歸為分子詞圖，低於平均的歸為分母詞圖，而分子詞圖與分母詞圖的統計值就是把各詞弧的正確度與詞圖正確度的期望值之差加總，因此兩者的值應該相等，在實作上兩者的值也會非常接近，因此τ最佳值的估測使用分子或分母詞圖統計值皆可，估測出來的值不會有明顯差別。

經過 MPE 與 MPFE 對於τ最佳值估測的驗證，基本上可以證明這個估測方式有一定的準確度，由於測試不用τ值相當耗時，之後的實驗τ值都直使用此估測法

決定，不再嘗試不同值的結果。至於 MPFE+pen+len 經過分子或分母詞圖統計值的計算後，詞圖 N 的 τ 最佳值估測為 24.61，詞圖 N 的 τ 最佳值估測為 22.04，實驗中兩者皆使用 25。

表 5.10、表 5.11 及圖 5.12～圖 5.19 為正確度計算加入錯誤處罰與音素正規化的最小音素音框錯誤訓練法的實驗結果，在圖中以+pen+len 代表，本實驗同樣以詞、字、音節、聲韻母四種層級來呈現，在正確率的表現上，詞圖 N 除了詞正確率在第 8 次疊代達到最高正確率外，其餘皆在疊代第 10 次時達到最高正確率；詞圖 N 則一致都在第 10 次疊代時達到最高正確率。在詞圖 N 的實驗中，字正確率達到 77.67%，進步 2.5%(相對 10.07%)；而在詞圖 T 的實驗中，字正確率達到 77.36%，進步 2.19%(相對 8.82%)。在圖中與 MPFE 的比較可以發現，MPFE+pen+len 雖然在詞正確率表現較差，但在其餘 3 種正確率上都有較佳的表現，由實驗結果可以推測，加入錯誤處罰與音素正規化的正確度可能跟字正確率較相似，而 MPFE 的正確度則跟詞正確率較相關，推測可能原因是加入錯誤處罰與音素長度正規化之後，一來正規化使得每個聲韻母對於正確度的影響程度是等價的，且右相關聲韻母的結構每個字都恰好由一組聲韻母組成，進而造成每個字的效果也等價，所以對字正確率的進步較有效果；二來錯誤處罰使得錯誤的長詞與短詞會有不一樣的分數，在錯誤詞中長詞的聲韻母也較短詞的聲韻母多，錯誤處罰便給錯誤長詞更低的正確度，而不像原本的正確度計算法不論錯誤詞長短都是 0 分；在原本的正確度計算法，會使得越長的正確聲韻母影響越大，導致由較多字組成的長詞，在正確時長詞通常會有比短詞高的正確度，但在錯誤時長詞卻不會有較短詞低的正確度，進而影響訓練方向，使得正向訓練偏好正確長詞勝於正確短詞，負向訓練不偏好錯誤長詞與短詞的程度則相同，然而在中文的語音辨識中，將 1 個長詞辨識成 2 個以上的短詞是不算錯誤的，偏好長詞的特性可能使得模型訓練得過份避免長詞拆成短詞的情形，使得原本正確的多個短詞也訓練成一個不正確的長詞，進而造成詞正確率提高，但字正確率卻下降的情形。

MPFE			詞圖 N		詞正確率(%)	
itr	τ=10	τ=100	τ=150	τ=200	τ=250	τ=400
0	57.99	57.99	57.99	57.99	57.99	57.99
1	58.78	58.89	58.85	58.93	58.90	58.85
2	59.21	59.77	59.78	59.74	59.60	59.61
3	59.64	60.05	60.20	60.20	60.19	60.05
4	60.06	60.75	60.61	60.59	60.63	60.38
5	59.92	60.99	60.89	60.65	60.78	60.66
6	60.00	60.93	61.13	60.99	60.88	60.67
7	60.21	61.24	61.26	61.15	61.15	60.89
8	60.41	61.20	61.36	61.32	61.35	60.87
9	60.15	61.56	61.56	61.46	61.57	61.20
10	60.05	61.49	61.52	**61.69**	61.62	61.36

表 5.2　最小音素音框錯誤訓練法－詞圖 N－詞正確率

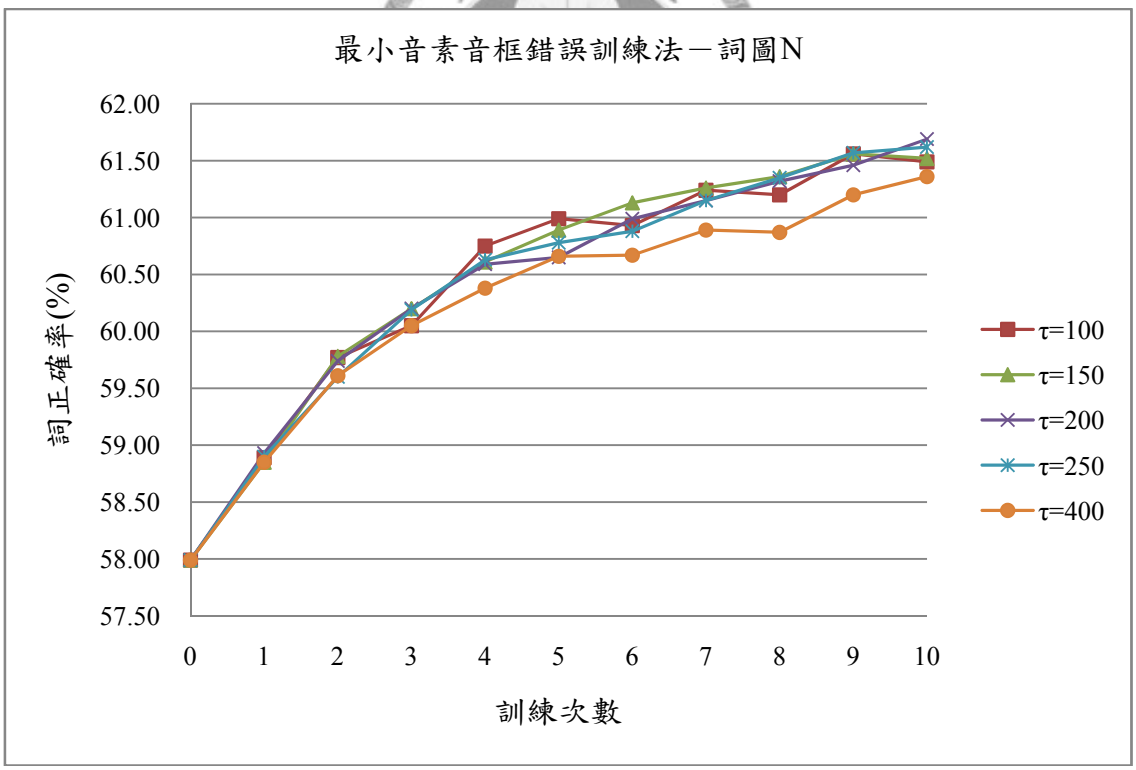


圖 5.4　最小音素音框錯誤訓練法－詞圖 N－詞正確率

MPFE			詞圖 N		字正確率(%)	
itr	τ=10	τ=100	τ=150	τ=200	τ=250	τ=400
0	75.17	75.17	75.17	75.17	75.17	75.17
1	75.67	75.72	75.69	75.77	75.73	75.73
2	75.88	76.21	76.19	76.18	76.17	76.16
3	75.84	76.15	76.31	76.31	76.30	76.27
4	75.67	76.28	76.30	**76.37**	76.36	76.27
5	75.12	76.13	76.13	76.07	76.20	76.23
6	74.63	75.75	76.00	75.99	76.02	76.11
7	74.35	75.66	75.87	75.87	75.96	76.04
8	74.02	75.49	75.81	75.83	75.92	75.76
9	73.73	75.54	75.70	75.84	76.10	76.19
10	73.35	75.60	75.79	76.01	76.09	76.21

表 5.3　最小音素音框錯誤訓練法－詞圖 N－字正確率

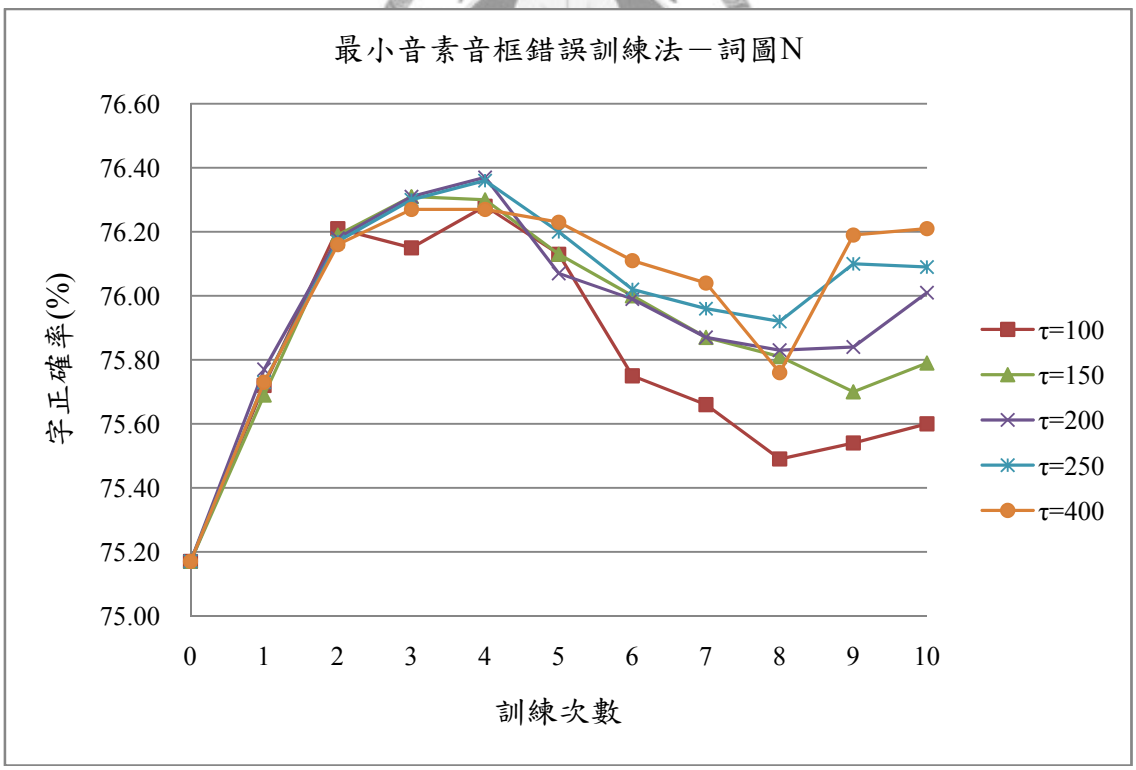


圖 5.5　最小音素音框錯誤訓練法－詞圖 N－字正確率

MPFE			詞圖 N		音節正確率(%)	
itr	τ=10	τ=100	τ=150	τ=200	τ=250	τ=400
0	81.42	81.42	81.42	81.42	81.42	81.42
1	81.97	82.02	81.99	82.06	82.02	81.99
2	82.26	82.48	82.42	82.41	82.38	82.38
3	82.25	82.49	82.60	82.60	82.57	82.50
4	82.12	82.60	82.60	82.62	**82.63**	82.52
5	81.77	82.46	82.43	82.36	82.47	82.48
6	81.44	82.28	82.46	82.39	82.35	82.34
7	81.24	82.26	82.36	82.30	82.32	82.33
8	80.98	82.12	82.34	82.28	82.28	81.97
9	80.77	82.17	82.28	82.30	82.42	82.44
10	80.47	82.18	82.34	82.42	82.40	82.45

表 5.4　最小音素音框錯誤訓練法－詞圖 N－音節正確率

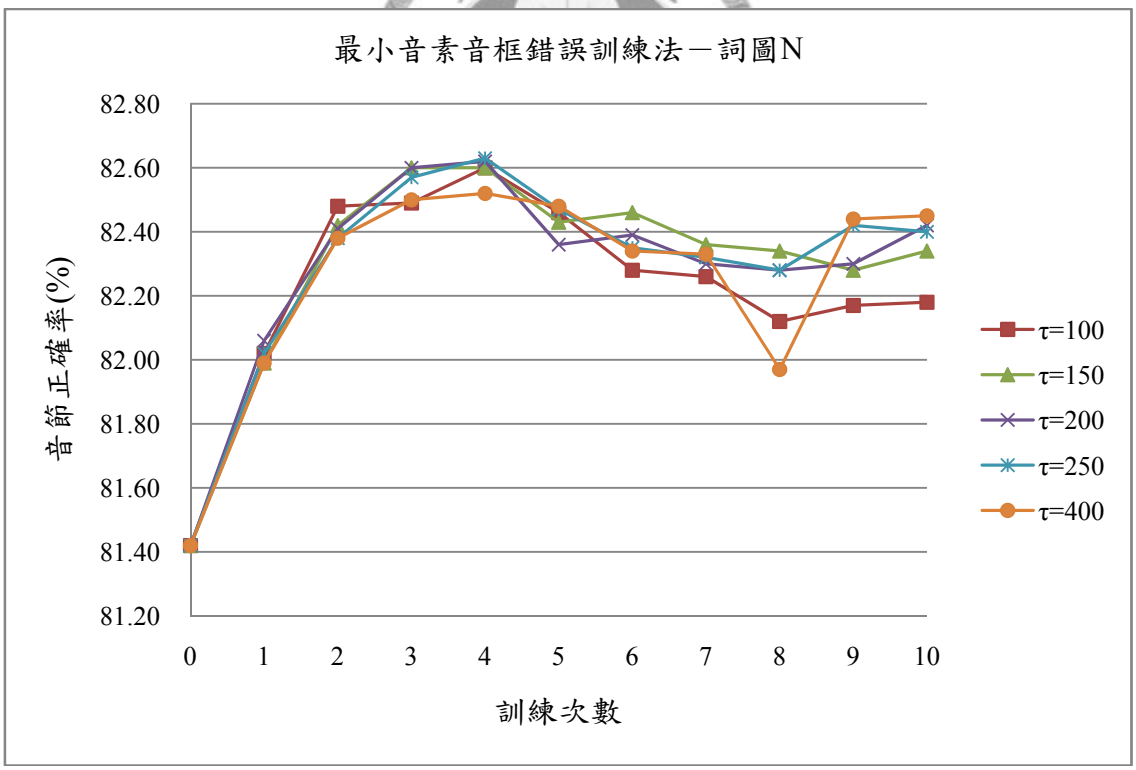


圖 5.6　最小音素音框錯誤訓練法－詞圖 N－音節正確率

MPFE			詞圖 N		聲韻母正確率(%)	
itr	τ=10	τ=100	τ=150	τ=200	τ=250	τ=400
0	84.76	84.76	84.76	84.76	84.76	84.76
1	85.22	85.26	85.24	85.28	85.24	85.21
2	85.41	85.60	85.57	85.56	85.54	85.53
3	85.40	85.63	85.72	**85.74**	85.72	85.64
4	85.28	85.69	85.70	85.72	**85.74**	85.66
5	84.97	85.57	85.56	85.49	85.59	85.59
6	84.64	85.39	85.53	85.47	85.47	85.49
7	84.44	85.34	85.42	85.37	85.40	85.43
8	84.25	85.25	85.43	85.38	85.38	85.03
9	84.10	85.28	85.38	85.42	85.52	85.51
10	83.82	85.29	85.44	85.51	85.51	85.53

表 5.5　最小音素音框錯誤訓練法－詞圖 N－聲韻母正確率

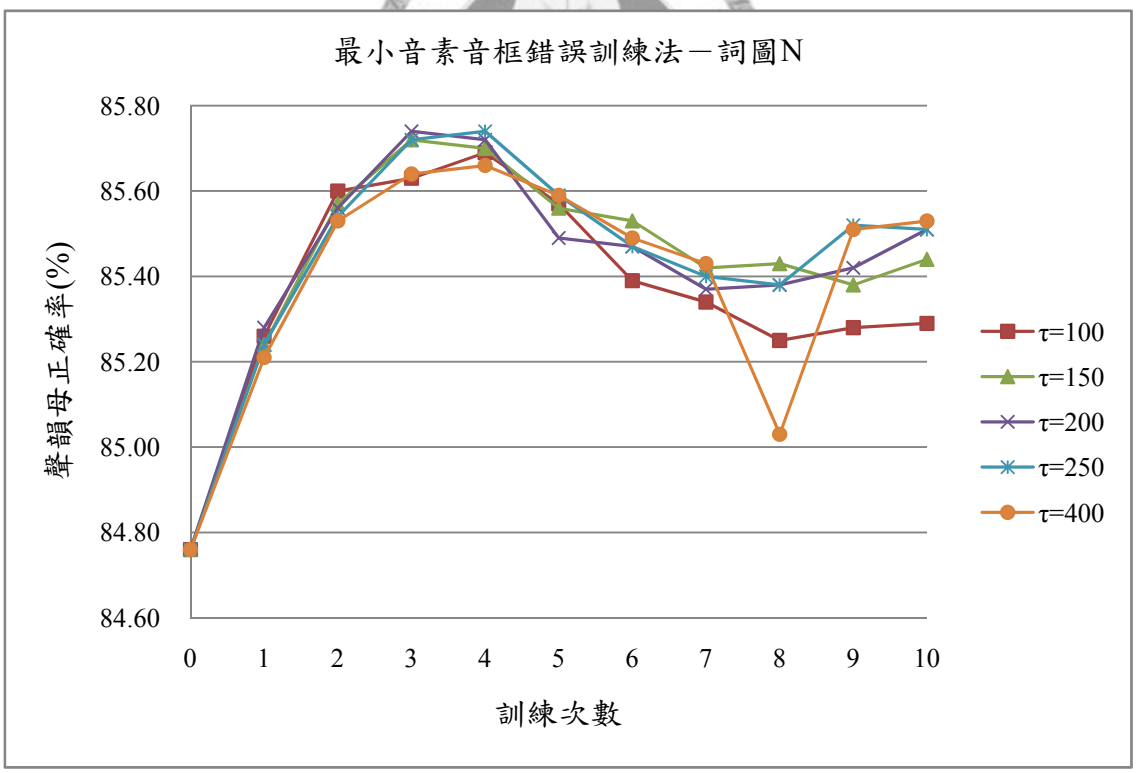


圖 5.7　最小音素音框錯誤訓練法－詞圖 N－聲韻母正確率

MPFE			詞圖 T		詞正確率(%)	
itr	τ=10	τ=100	τ=150	τ=200	τ=250	τ=400
0	57.99	57.99	57.99	57.99	57.99	57.99
1	58.60	58.81	58.85	58.88	58.85	58.76
2	59.35	59.63	59.59	59.48	59.44	59.46
3	59.72	60.00	60.12	60.11	60.08	59.85
4	59.79	60.62	60.55	60.55	60.36	60.25
5	59.59	60.82	60.95	60.76	60.69	60.59
6	59.49	60.75	61.01	61.19	60.99	60.88
7	59.64	61.61	61.30	61.30	61.45	61.07
8	59.88	61.89	61.71	61.56	61.49	61.24
9	59.65	61.86	61.89	61.77	61.66	61.21
10	59.67	61.76	**62.03**	61.95	61.85	61.38

表 5.6　最小音素音框錯誤訓練法－詞圖 T－詞正確率

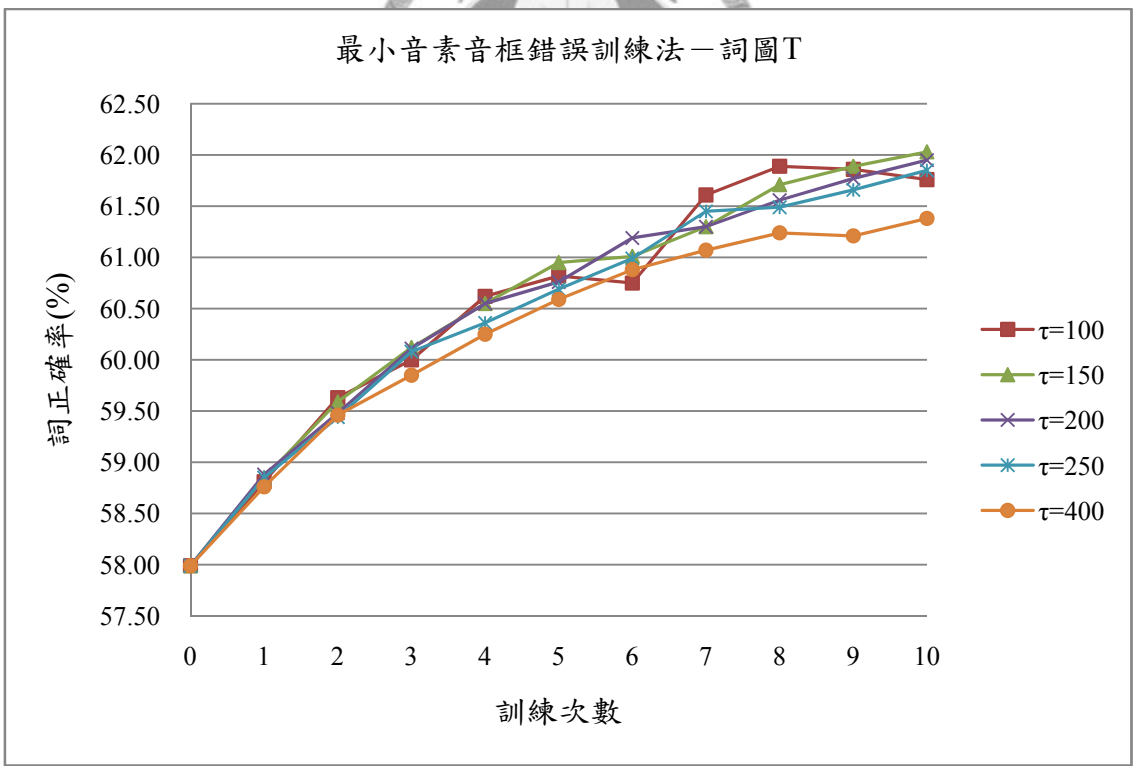


圖 5.8　最小音素音框錯誤訓練法－詞圖 T－詞正確率

MPFE			詞圖 T		字正確率(%)	
itr	τ=10	τ=100	τ=150	τ=200	τ=250	τ=400
0	75.17	75.17	75.17	75.17	75.17	75.17
1	75.50	75.67	75.64	75.68	75.70	75.68
2	75.93	76.19	76.18	76.06	76.01	75.99
3	75.88	76.11	76.19	76.18	**76.22**	76.09
4	75.42	76.11	76.16	76.18	76.13	76.06
5	74.78	75.81	75.99	76.01	76.02	75.98
6	73.85	75.24	75.59	75.76	75.76	75.99
7	73.13	75.24	75.28	75.54	75.76	75.83
8	73.04	75.13	75.28	75.43	75.46	75.76
9	72.48	75.05	75.32	75.40	75.52	75.36
10	72.41	75.01	75.16	75.47	75.58	75.33

表 5.7　最小音素音框錯誤訓練法－詞圖 T－字正確率

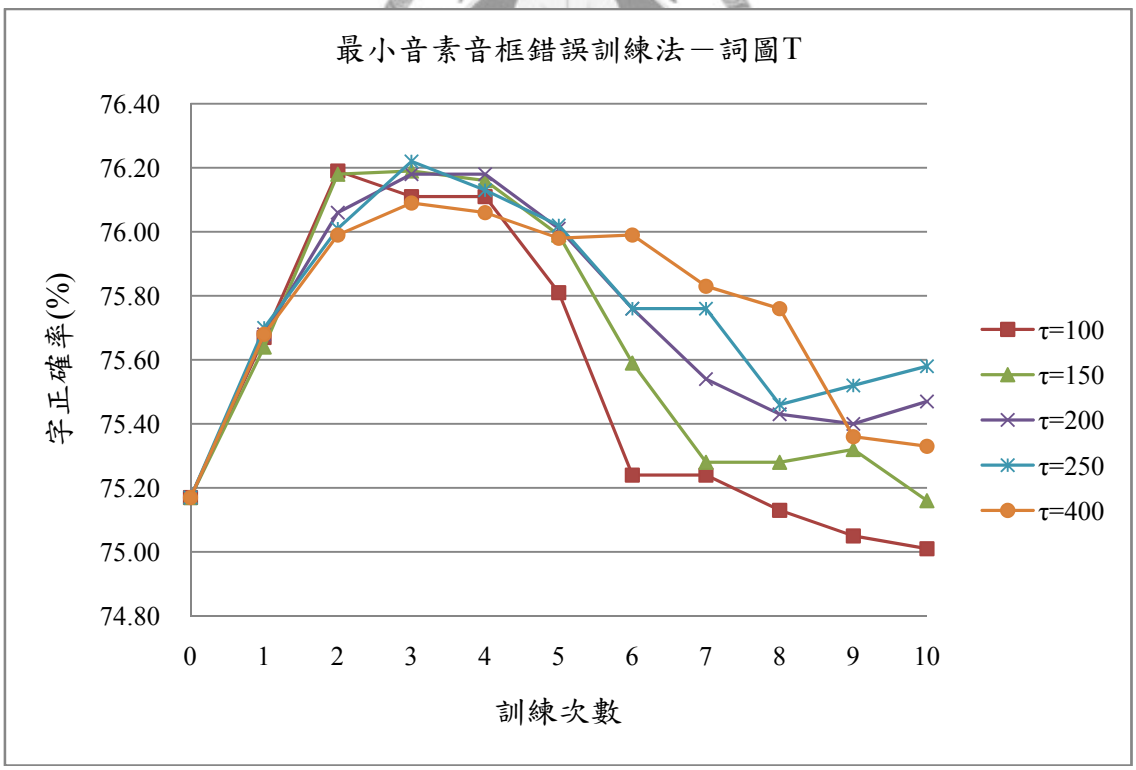


圖 5.9　最小音素音框錯誤訓練法－詞圖 T－字正確率

MPFE			詞圖 T		音節正確率(%)	
itr	τ=10	τ=100	τ=150	τ=200	τ=250	τ=400
0	81.42	81.42	81.42	81.42	81.42	81.42
1	81.90	82.00	81.99	82.00	82.02	81.95
2	82.36	82.48	82.47	82.31	82.29	82.23
3	82.30	82.43	**82.50**	82.46	82.47	82.34
4	81.93	82.42	82.44	82.41	82.39	82.32
5	81.60	82.26	82.39	82.33	82.29	82.26
6	80.90	81.82	82.10	82.17	82.19	82.22
7	80.31	81.81	81.81	82.00	82.16	82.13
8	80.26	81.81	81.83	81.93	81.91	82.06
9	79.83	81.80	81.91	81.87	81.99	81.66
10	79.94	81.81	81.88	82.03	82.09	81.69

表 5.8　最小音素音框錯誤訓練法－詞圖 T－音節正確率

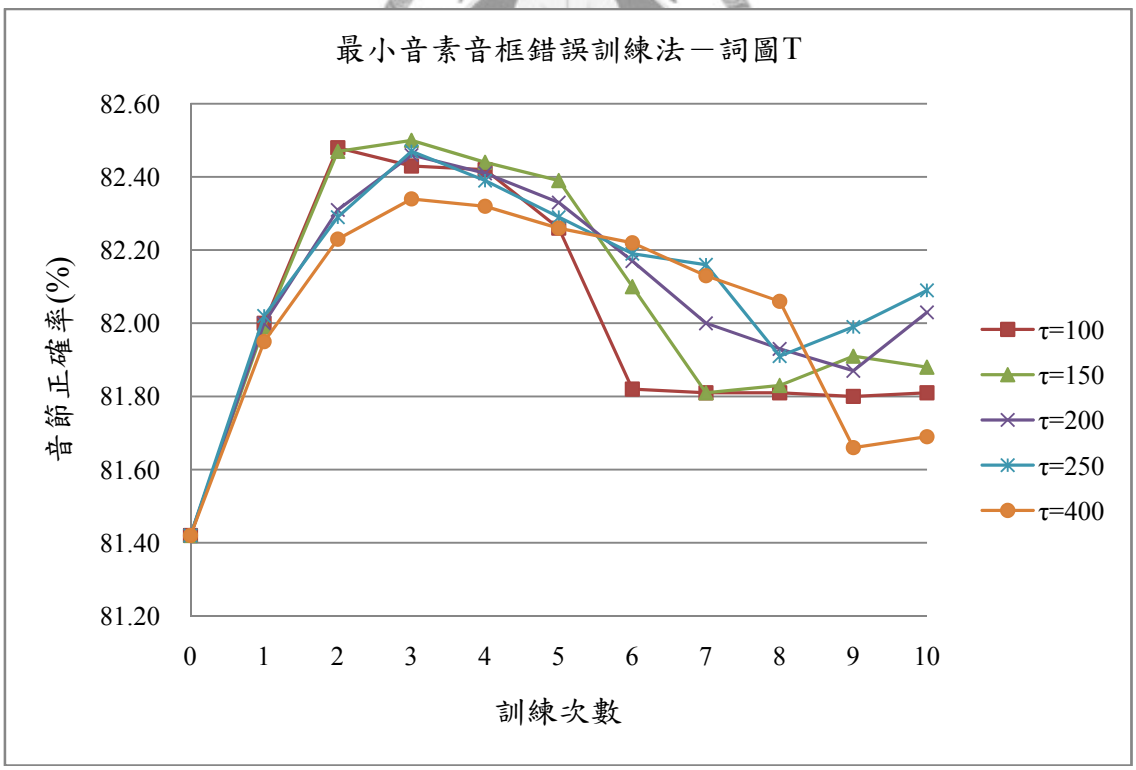


圖 5.10　最小音素音框錯誤訓練法－詞圖 T－音節正確率

MPFE			詞圖 T		聲韻母正確率(%)	
itr	τ=10	τ=100	τ=150	τ=200	τ=250	τ=400
0	84.76	84.76	84.76	84.76	84.76	84.76
1	85.19	85.28	85.26	85.26	85.29	85.22
2	85.51	85.60	85.58	85.48	85.46	85.45
3	85.47	85.60	**85.64**	85.61	**85.64**	85.55
4	85.15	85.55	85.57	85.56	85.56	85.50
5	84.83	85.37	85.50	85.48	85.47	85.45
6	84.25	85.02	85.22	85.33	85.34	85.37
7	83.74	84.94	84.99	85.14	85.28	85.28
8	83.72	84.95	85.00	85.08	85.08	85.18
9	83.41	84.99	85.09	85.02	85.14	84.78
10	83.60	85.02	85.07	85.21	85.26	84.80

表 5.9　最小音素音框錯誤訓練法－詞圖 T－聲韻母正確率

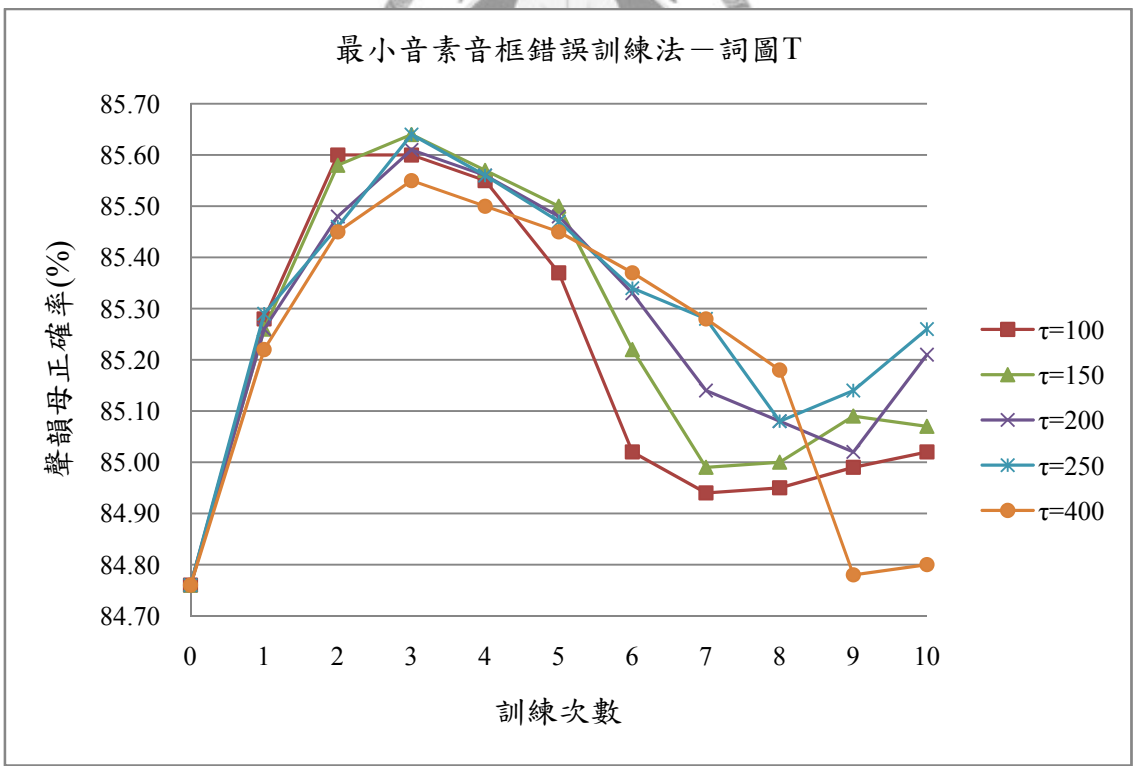


圖 5.11　最小音素音框錯誤訓練法－詞圖 T－聲韻母正確率

MPFE+pen+len		詞圖 N		τ =25
itr	詞正確率(%)	字正確率(%)	音節正確率(%)	聲韻母正確率(%)
0	57.99	75.17	81.42	84.76
1	58.76	75.90	82.18	85.43
2	59.27	76.54	82.83	85.98
3	59.48	76.90	83.15	86.32
4	59.57	77.05	83.39	86.52
5	59.80	77.36	83.65	86.75
6	59.92	77.54	83.80	86.89
7	59.86	77.47	83.78	86.82
8	59.80	77.59	83.91	86.92
9	59.83	77.65	83.96	86.93
10	59.85	77.67	84.00	86.95

表 5.10 最小音素音框錯誤－加入錯誤處罰與音素正規化－詞圖 N

MPFE+pen+len		詞圖 T		τ =25
itr	詞正確率(%)	字正確率(%)	音節正確率(%)	聲韻母正確率(%)
0	57.99	75.17	81.42	84.76
1	58.71	75.87	82.17	85.43
2	59.08	76.49	82.78	85.92
3	59.42	76.91	83.16	86.28
4	59.30	77.01	83.31	86.41
5	59.51	77.19	83.53	86.58
6	59.46	77.18	83.58	86.64
7	59.37	77.17	83.55	86.64
8	59.54	77.36	83.75	86.81
9	59.41	77.34	83.69	86.75
10	59.48	77.36	83.71	86.77

表 5.11 最小音素音框錯誤－加入錯誤處罰與音素正規化－詞圖 T

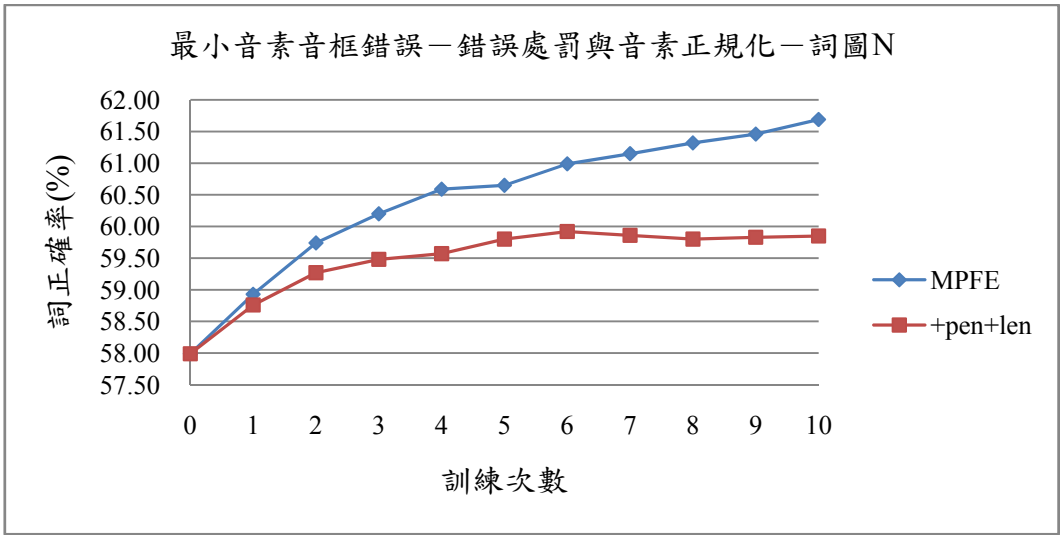


圖 5.12 最小音素音框錯誤－錯誤處罰音素正規化－詞圖 N－詞正確率

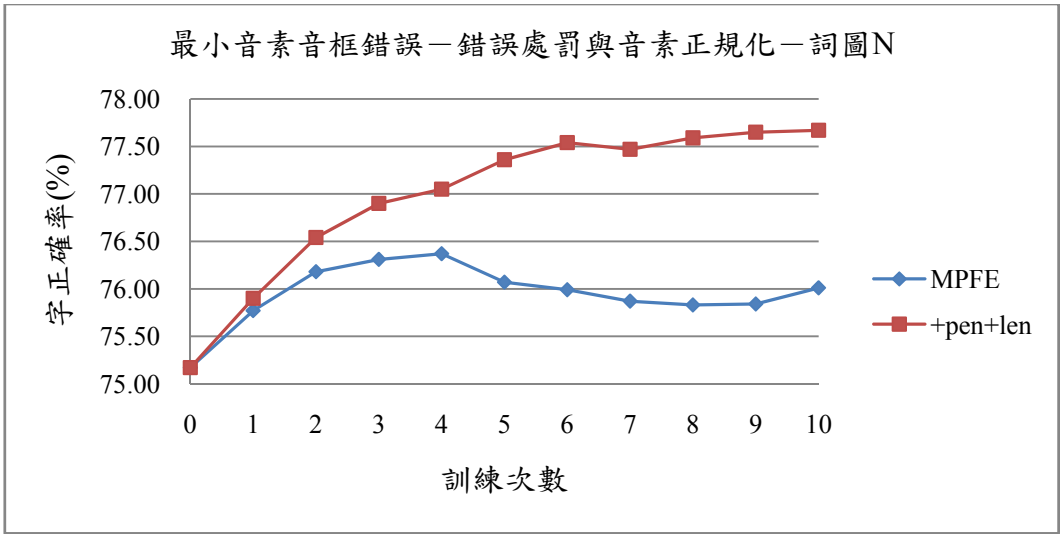


圖 5.13 最小音素音框錯誤－錯誤處罰音素正規化－詞圖 N－字正確率

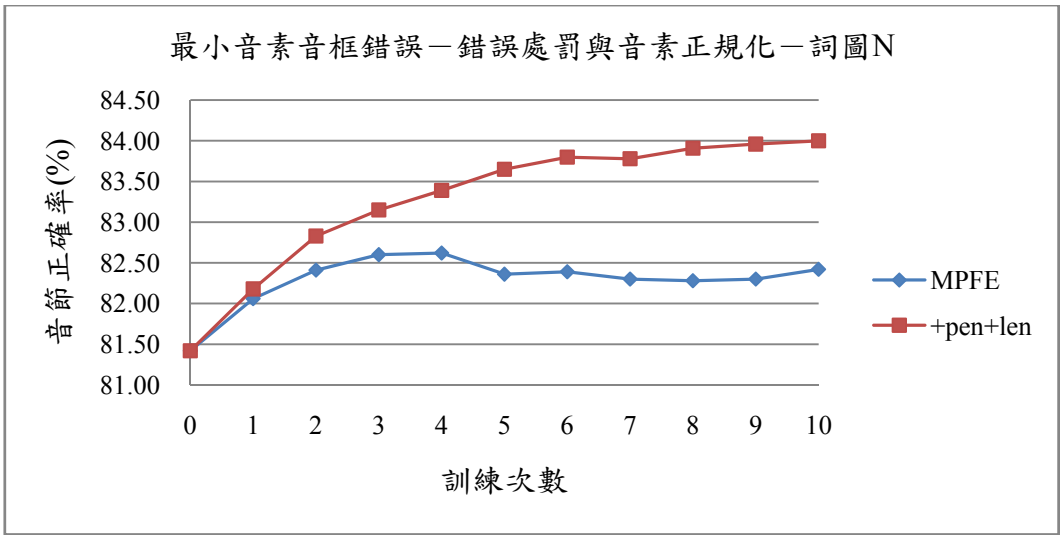


圖 5.14 最小音素音框錯誤－錯誤處罰音素正規化－詞圖 N－音節正確率

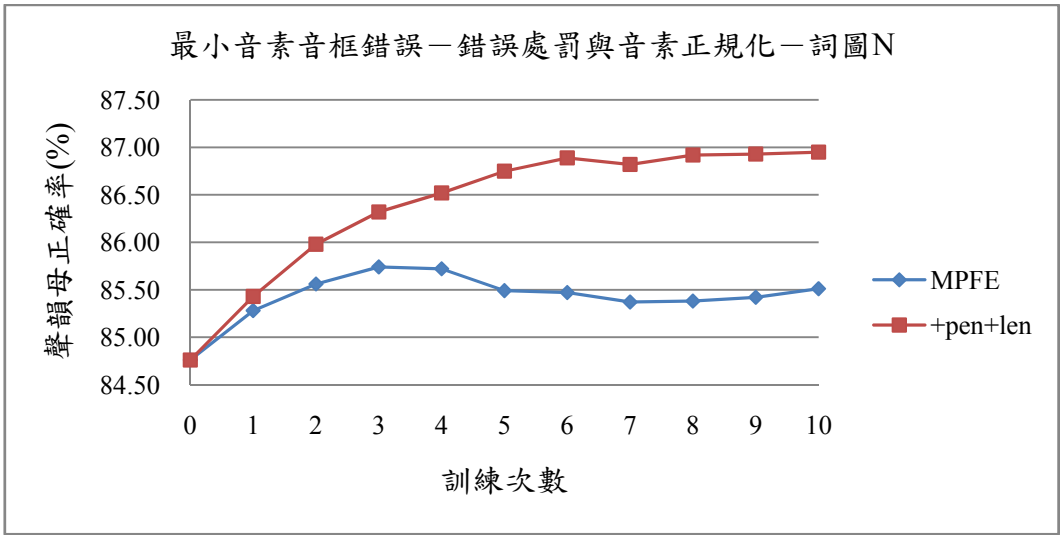


圖 5.15 最小音素音框錯誤－錯誤處罰音素正規化－詞圖 N－聲韻母正確率

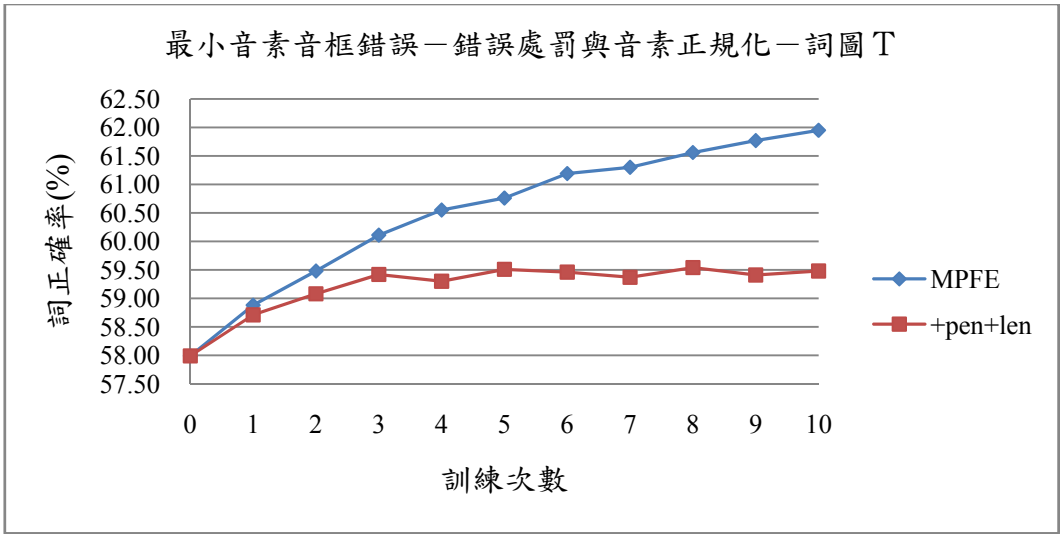


圖 5.16 最小音素音框錯誤－錯誤處罰音素正規化－詞圖 T－詞正確率

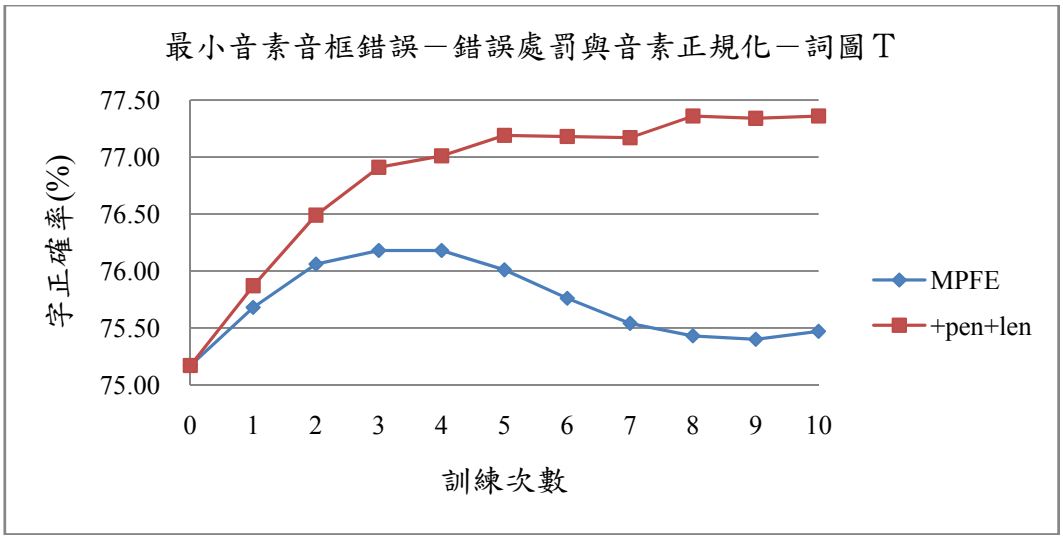


圖 5.17 最小音素音框錯誤－錯誤處罰音素正規化－詞圖 T－字正確率

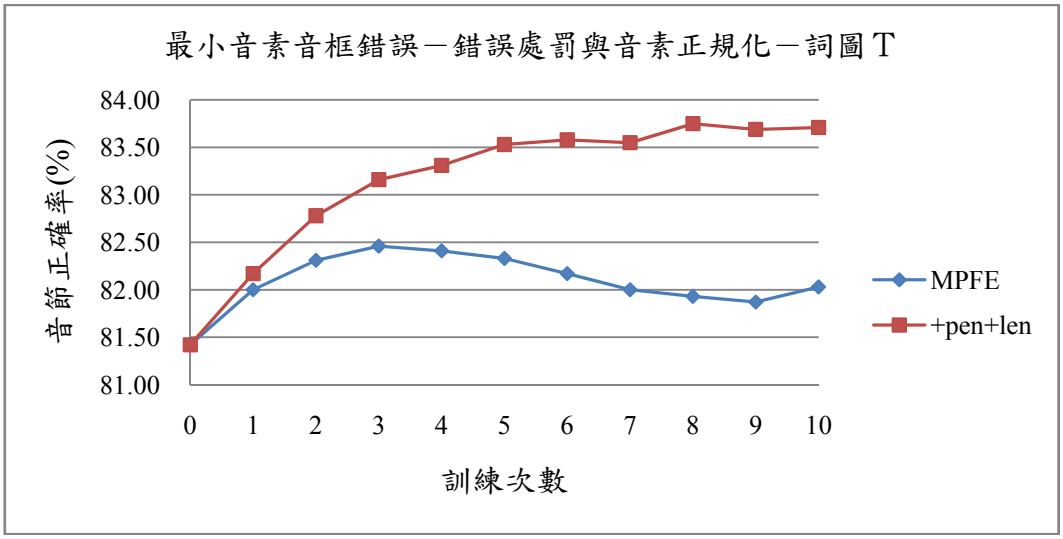


圖 5.18 最小音素音框錯誤－錯誤處罰音素正規化－詞圖 T－音素正確率

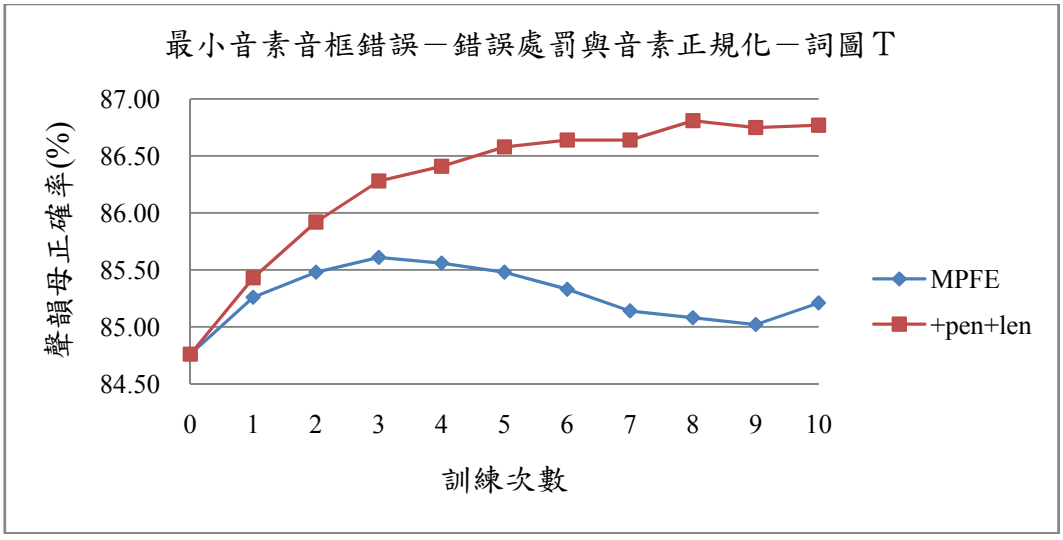


圖 5.19 最小音素音框錯誤－錯誤處罰音素正規化－詞圖 T－聲韻母正確率

5.2 狀態層級最小貝氏風險訓練

5.2.1 目標函數

在 2006 年吉氏(Gibson)提出基於最小貝氏風險，以不同層級計算音框正確度做為減損函數依據的方法【34】，其中狀態(state)就是這些層級中的一種，本節的狀態層級最小貝氏風險(state level Minimum Bayes Risk, sMBR)就是以狀態層級計算音框正確度的方法。目標函數基本上與最小音素訓練法(4.1)式相同：

$$F_{sMBR}(\lambda)=\sum_{r}\sum_{u}P(u\mid O_r)\,Acc(s_r,u) \tag{5.7}$$

與最小音素訓練法的差異在於正確度$Acc(s_r,u)$的計算方式有所不同，狀態層級最小貝氏風險計算的正確度為狀態音框正確度。定義每個音素個別的正確度：

$$StateFrameAcc(q)=\sum_{t=start(q)}^{end(q)}\delta\left(q_{state}(t),s_{r,state}(t)\right) \tag{5.8}$$

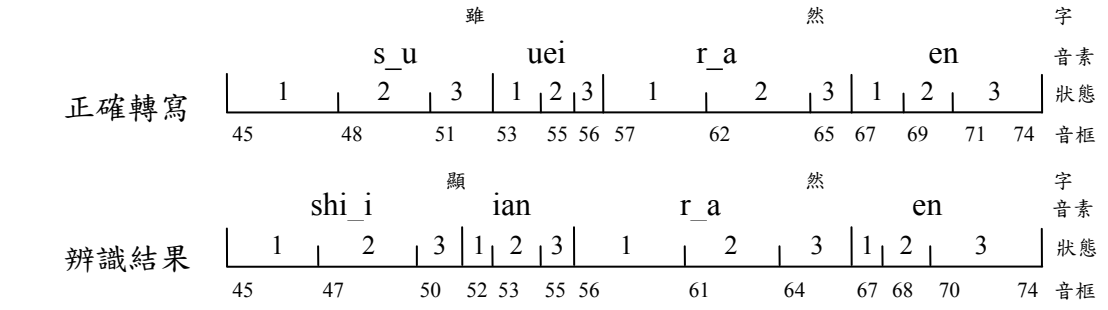
$$\delta\left(q_{state}(t),s_{r,state}(t)\right)=\begin{cases}1 & \text{if } q_{state}(t)=s_{r,phone}(t)\\ 0 & \text{if } q_{state}(t)\neq s_{r,phone}(t)\end{cases} \tag{5.9}$$

其中音素q是辨識結果，$start(q)$和$end(q)$分別表示音素q以音框為單位的開始與結束時間；$q_{state}(t)$表示音素q在時間t時的狀態，$s_{r,state}(t)$表示正確轉寫在時間t時的狀態。而整個文句的狀態音框正確度，就是將每個音素個別的正確度加總：

$$Acc(s_r,u)=\sum_{q\in u}StateFrameAcc(q) \tag{5.10}$$

狀態音框正確度的計算，是在比對文句u與正確轉寫s_r的每個音框所屬的狀態，然後計算相同狀態的數量，因此在實作上，狀態層級最小貝氏風險的詞弧正確度計算方式，就是比對詞弧與其區間的正確轉寫，然後計算辨識狀態正確的音框數量。

圖 5.20 是一個狀態音框正確度的計算範例，圖中時間 45～55 的音框，由於



狀態音框數	2	3	2	1	2	1	5	3	3	1	2	5	
相同狀態音框數	0	0	0	0	0	0	4	2	2	1	1	4	

正確度計算 $=0+0+0+0+0+0+4+2+2+1+1+4=14$

圖 5.20 狀態音框正確度的計算範例

音素不同，狀態當然也不同，所以正確度皆為 0；音框 56～60 為 r_a 的狀態 1，其中音框 57～60 的 4 個音框同為 r_a 的狀態 1，故記正確度 5；音框 62～63 同為 r_a 的狀態 2，共 2 個音框，故記正確度 2；音框 65～66 同為 r_a 的狀態 3，共 2 個音框，故記正確度 2；音框 67 同為 en 的狀態 1，1 個音框故記正確度 1；音框 69 同為 en 的狀態 2，1 音框故記正確度 1；音框 71～74 同為 en 的狀態 3，共 4 個音框，故記正確度 4。以上正確度的總合就是整段詞弧的正確度 14。

另外，由於正確轉寫通常是不會標記到狀態的等級的，一般最小單位也只會記到音素的等級，所以在狀態音框正確度的計算中，正確轉寫的狀態標記是經由強制對齊(forced alignment)而來。

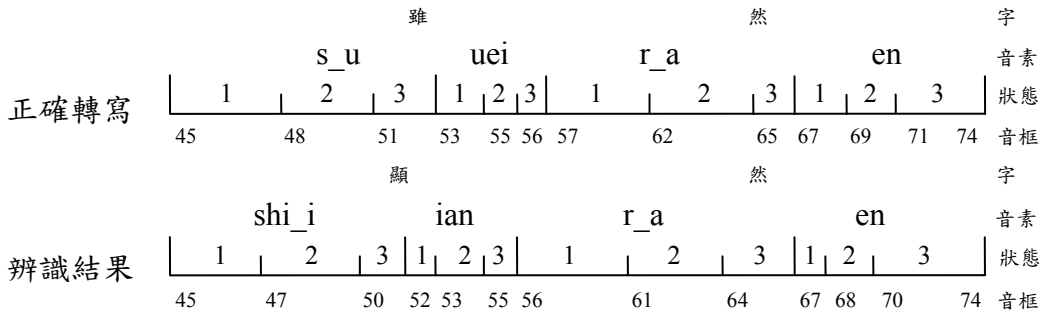
5.2.2 加入錯誤處罰的詞弧正確度

5.2.1 節所討論的狀態音框正確度，為計算辨識之狀態正確的音框數，本節仿照 5.1.2 節的方式，也對狀態音框正確度做類似的修改。本節在原本的狀態音框正確度加入對於錯誤的處罰，原本(5.8)與(5.9)式各別音素正確度的計算方式就改為：

$$StateFrameAcc(q)=\sum_{t=start(q)}^{end(q)} U\left(q_{state}(t),s_{r,state}(t)\right) \tag{5.11}$$

$$U\left(q_{state}(t),s_{r,state}(t)\right)=\begin{cases}1 & \text{if } q_{state}(t)=s_{r,phone}(t)\\ 0 & \text{if } q_{state}(t)\neq s_{r,phone}(t), \text{ but } q=s_{r,phone(t)}\\ -\rho & \text{if } q\neq s_{r,phone(t)}\end{cases} \tag{5.12}$$

其中 $s_{r,state}(t)$ 表示正確轉寫在時間 t 時的狀態，$s_{r,phone}(t)$ 代表在時間 t 時的音素；ρ 是錯誤處罰的權重，ρ 的選擇與 5.1.2 節一樣使用 $\rho=0.1$。修改後的狀態音框正確度在對每一個音素計算正確度的時候，不再只有累加辨識狀態正確的音框數而已，對於辨識音素不正確的音框，每一個會給與 $-\rho$ 的處罰，在這裡對於辨識狀態錯誤但音素卻正確的音框仍然維持不處罰，主要的考量是對於辨識的正確率而言，狀態錯誤但音素卻正確的情況實際上是視為正確的，因此這裡對這種情況不給與處罰，以期增加狀態音框正確度的訓練目標，同時也可以增加辨識率。



相同狀態音框數	0	0	0	0	0	0	4	2	2	1	1	4	
不同音素音框數	2	3	2	1	2	1	1	0	0	0	0	0	
相同狀態數－處罰權重	-0.2	-0.3	-0.2	-0.1	-0.2	-0.1	4.9	2	2	1	1	4	

正確度計算 $= -0.2 - 0.3 - 0.2 - 0.1 - 0.2 - 0.1 + 4.9 + 2 + 2 + 1 + 1 + 4 = 13.8$

圖 5.21 加入錯誤處罰的狀態音框正確度的計算範例

圖 5.21 是一個加入錯誤處罰的狀態音框正確度的計算範例，圖中時間 45～56 的音框，因為音素不同因此每一個音框都給與-0.1 的處罰；時間 57～74 的音框由於音素皆與正確轉寫相同，因此不給與處罰。音框 56～61 同為 r_a 的狀態 1，共 5 個音框，扣除音框 56 為一個音素不同之音框的處罰-0.1，故正確度為 $5 - 0.1 = 4.9$；剩下後面的音框 61～74，由於並無出現不同音素的情況需要處罰，因此正確度的計算與圖 5.20 未加入錯誤處罰的正確度相同。以上正確度的總合即整段詞弧的正確率 13.8。

5.2.3 加入錯誤處罰與音素長度正規化的詞弧正確度

本節承 5.2.2 節在狀態音框正確度加入錯誤處罰之後，更進一步加入音素長度的正規化，於是，原本(5.11)與(5.12)式各別音素正確度的計算方式就變更為：

$$StateFrameAcc(q) = \sum_{t=start(q)}^{end(q)} \frac{U\left(q_{state}(t), s_{r,state}(t)\right)}{len\left(s_{r,phone}(t)\right)} \tag{5.13}$$

$$U\left(q_{state}(t), s_{r,state}(t)\right) = \begin{cases} 1 & \text{if } q_{state}(t) = s_{r,phone}(t) \\ 0 & \text{if } q_{state}(t) \neq s_{r,phone}(t), \text{ but } q = s_{r,phone(t)} \\ -\rho & \text{if } q \neq s_{r,phone(t)} \end{cases} \tag{5.14}$$

基本上與(5.11)和(5.12)式大致相同，ρ 也一樣使用 0.1，差別只有在於(5.13)多除了一個 $len\left(s_{r,phone}(t)\right)$，而 $len\left(s_{r,phone}(t)\right)$ 代表的是正確轉寫 s_r 在時間 t 的音素的長度。也就是說，每個音框計算出來的正確度會再對正確轉寫在該時間的音素長度做正規化。這裡的音素長度正規化不同於 5.1.2 節使用的是辨識結果的音素長度，而是正確轉寫的音素長度，考慮因素是在第 4 章的最小音素錯誤訓練法中，重疊比例的計算亦是以正確轉寫的音素長度為分母，就某種程度上而言，也算是對正確轉寫的音素長度正規化，因此這裡也使用正確轉寫的音素長度做正規化。另外也是考慮到正確度的計算，盡量以正確轉寫為標準也較為合理。

圖 5.22 是一個加入錯誤處罰與音素長度正規化的狀態音框正確度的計算範例，基本上每個音框的正確度計算都與圖 5.21 大致上是相同的，主要的差別就只是需要再除以正確轉寫的音素長度。在時間 45～52 的音框，正確轉寫的音素 s_u 的長度為 8 個音框，所以這個區間的音框正確度要除以 8；在時間 53～56 的音框正確轉寫的音素 uei 長度為 4，因此這個區間的音框正確度要除以 4；在時間 57～66 的音框正確轉寫的音素 r_a 長度為 10，因此這個區間的音框正確度要除以 10；在時間 67～74 的音框正確轉寫的音素 en 長度為 8，因此這個區間的音框正確度要除以 8。如音框 56～60 的音素 r_a 狀態 1，就有 1 個音框為錯誤音素，該音框的正確轉

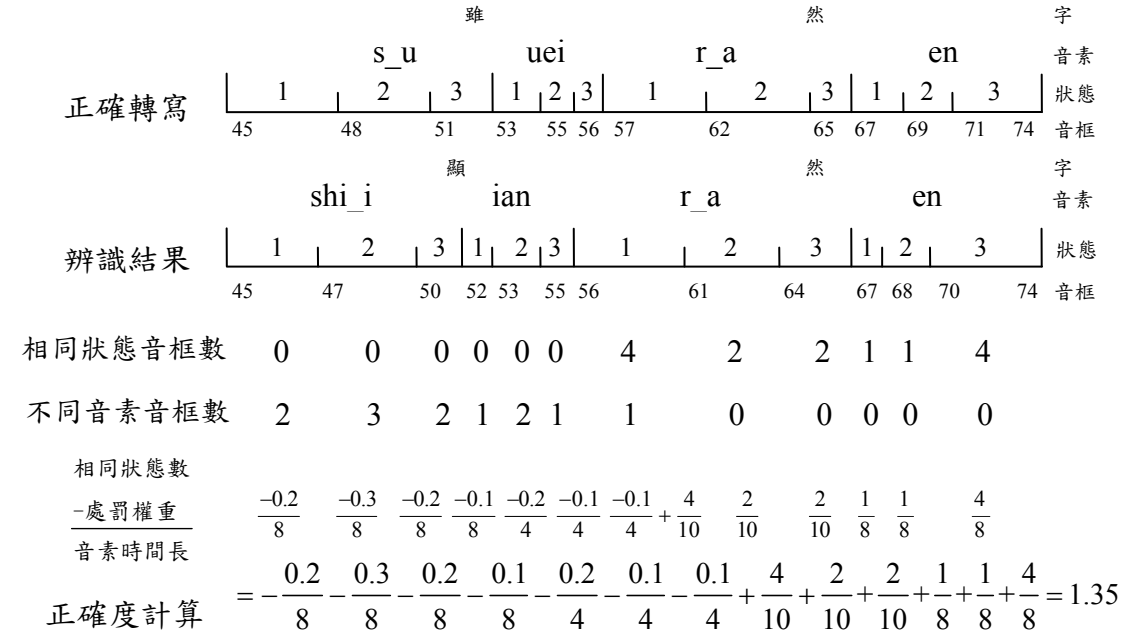


圖 5.22 加入錯誤處罰與音素長度正規化的狀態音框正確度的計算範例

寫音素長度為 4 個音框，故正確度罰處 $-0.1\times1/4=-0.1/4$；另外有 4 個音框為正確狀態，該音框的正確轉寫音素長度為 10，故正確度加上 $4/10=0.4$。以上正確度的總合即整段詞弧的正確率 1.35。

5.2.4 實驗結果

本實驗為狀態層級最小貝氏風險訓練法，以及加入錯誤處罰和再加入音素長度正規化的實驗。在本實驗中狀態音框正確度的計算需要正確轉寫的狀態對齊，而訓練語料的正確轉寫只有標計到音素對齊，因此正確轉寫的狀態對齊是由強制對齊而來(forced alignment)，本實驗中強制對齊時使用的聲學模型是上一次疊代結束產生的模型，也就是說，每次強制對齊所使用的模型並不相同，所以對齊的結果也不會一樣。第一次疊代時的強制對齊則是使用 3.3 節的初使模型。

實驗設定基本上與 4.3 節的最小音素錯誤訓練的實驗相同，而平滑係數 τ 的設定則是根據 5.1.3 節的方法推估而來，估測如表 5.12 所示：

訓練方法	詞圖	分子詞圖統計	分母詞圖統計	估測 τ	τ
MPE	N	1.005032E+06	1.005032E+06	25.00	25
sMBR	N	5.267830E+06	5.267830E+06	131.04	130
sMBR	T	4.684503E+06	4.684513E+06	116.53	115
sMBR+pen	N	5.956818E+06	5.956818E+06	148.17	150
sMBR+pen	T	5.298603E+06	5.298614E+06	131.80	130
sMBR+pen+len	N	6.580281E+05	6.580281E+05	16.37	16
sMBR+pen+len	T	5.882187E+05	5.882202E+05	14.63	15

表 5.12 狀態層級最小貝氏風險－平滑係數最佳值之估測

表中 MPE 代表最小音素錯誤訓練法，sMBR 代表狀態層級最小貝氏風險，sMBR+pen 代表正確度加入錯誤處罰的版本，sMBR+pen+len 代表正確度加入錯誤處罰與音素長度正規化。同樣的表中是在假設最小音素錯誤訓練法的最佳 τ 值為 25 之下所做的估測。估測 τ 欄內為估測的結果，由於此估測的精準度並沒有很高，因此最後實驗使用的 τ 值大概取在估測值的附近即可，在【32】中亦為如此之作法，最右欄 τ 內為實驗採用的值。

實驗結果如表 5.13～表 5.20 及圖 5.23～圖 5.30 所呈現。圖中 sMBR 代表狀態層級最小貝氏風險，+pen 代表正確度加入錯誤處罰的版本，+pen+len 代表正確度加入錯誤處罰與音素長度正規化，圖表同時呈現這 3 種方法。在字正確率上，詞圖 N 的實驗中 sMBR 最高達到 76.78%的正確率，進步 1.61%(相對 6.48%)，+pen 最高達到 77.40%的正確率，進步 2.23%(相對 8.98%)，+pen+len 最高達到 77.42%的正確率，進步 2.25%(相對 9.06%)；而在詞圖 T 的實驗中，sMBR 最高達到 76.54%的正確率，進步 1.37%(相對 5.52%)，+pen 最高達到 77.17%的正確率，進步 2.00%(相對 8.05%)，+pen+len 最高達到 77.15%的正確率，進步 1.98%(相對 7.97%)。而就 4 種層級的正確率來看，兩種詞圖的實驗，sMBR 都有最佳的詞正確率，其它則大部份是+pen+len 具有最佳的正確率，除了在詞圖 T 的實驗中的字正確率是+pen 有最佳的正確率。整體實驗的疊代大部份都在 5～7 次之間達到最大正確率，其中 sMBR 通常較早達到最大正確率。

+pen 與+pen+len 在整體的表現上相當類似，不過與 sMBR 的表現上則有明顯差異。與 5.1.3 節最小音素音框錯誤的實驗結果有類似的情況，加入錯誤處罰的版本會有較好的字正確率，但是較差的詞正確率。此外，這裡又多比較了在加入錯誤處罰時是否也加入音素長度正規化的情形。由實驗結果發現，加入錯誤處罰後對於詞、字正確率的改變就會出現，再加入音素正規化之後進一步則無明顯變化，依照 5.1.3 節的推測，原因為原始的版本在正確的詞中，因長詞的聲韻母多，所以會得到比短詞更高的正確度，但是在錯誤的詞中，正確度皆為 0，所以長詞會得到跟短詞相同的正確度，然而在加入錯誤處罰之後，使得錯誤的詞中，聲韻母較多的長詞具有比短詞更低的正確度，進而使訓練不會有偏向長詞勝於短詞的情形；至於音素長度正規化之所以沒有明顯影響，可能因為聲韻母的時間長短相差並沒有大到造成時間較長的聲韻母控制了對正確度的影響，使得訓練偏向時間較長的聲韻母；另外也有可能是因為中文其實並沒有特定的聲韻母時間一定較其它聲韻母長，時間較長的聲韻母為語者隨機發生，沒有特定發生在特定的聲韻母上，而使得訓練偏好某些特定的聲韻母。

sMBR		詞正確率(%)	
詞圖 N itr	sMBR τ=130	+pen τ=150	+pen+len τ=16
0	57.99	57.99	57.99
1	58.74	58.65	58.76
2	59.65	59.56	59.55
3	60.03	60.01	59.82
4	60.57	60.43	60.20
5	**61.00**	60.61	60.55
6	60.68	60.67	60.65
7	60.85	60.86	60.66
8	60.67	60.80	60.50
9	60.64	60.78	60.60
10	60.60	60.71	60.59

表 5.13 狀態層級最小貝氏風險－詞圖 N－詞正確率

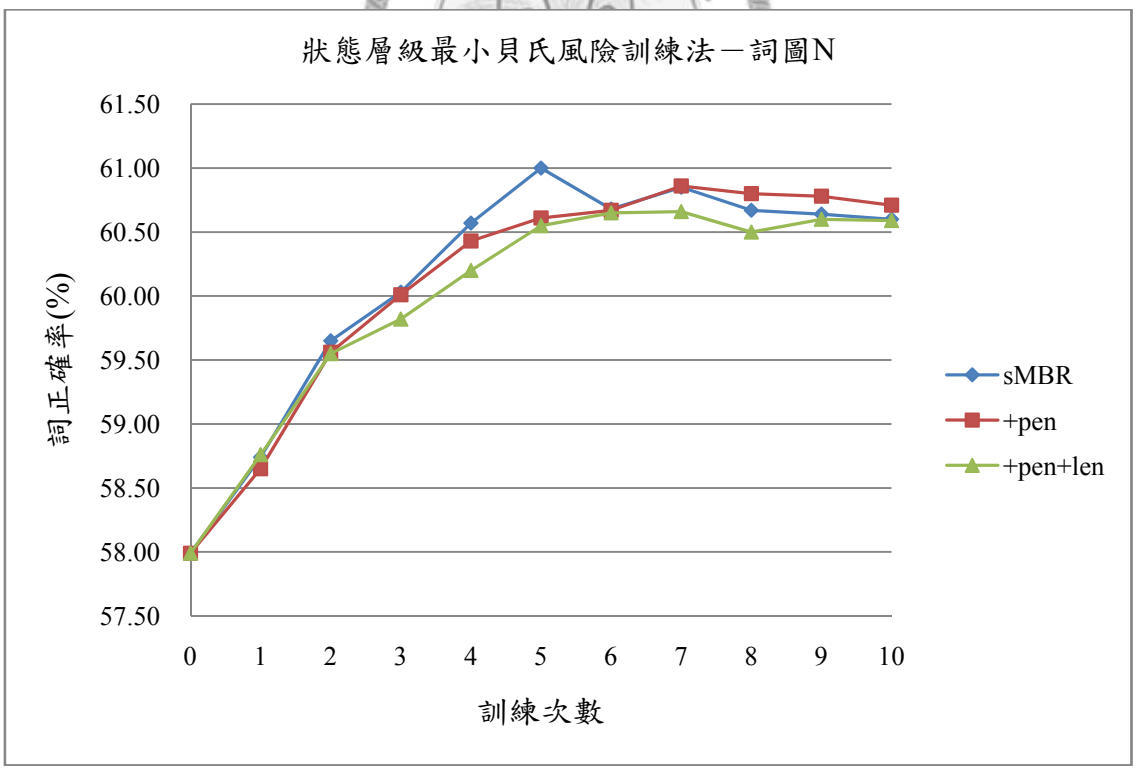


圖 5.23 狀態層級最小貝氏風險－詞圖 N－詞正確率

sMBR		字正確率(%)	
詞圖 N itr	sMBR τ=130	+pen τ=150	+pen+len τ=16
0	75.17	75.17	75.17
1	75.82	75.82	75.86
2	76.35	76.52	76.54
3	76.68	77.02	76.87
4	76.68	77.14	77.15
5	76.78	77.34	77.32
6	76.45	77.31	**77.42**
7	76.61	77.40	77.39
8	76.18	77.32	77.21
9	76.20	77.32	77.24
10	76.04	77.26	77.17

表 5.14 狀態層級最小貝氏風險－詞圖 N－字正確率

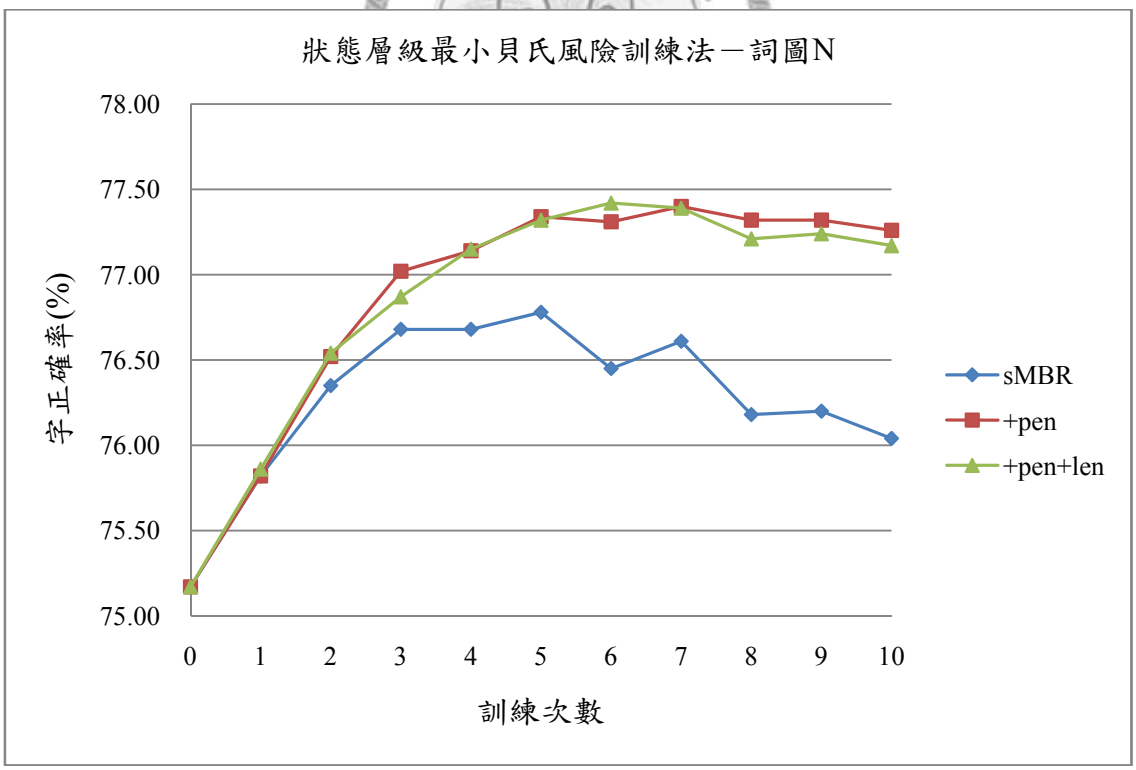


圖 5.24 狀態層級最小貝氏風險－詞圖 N－字正確率

sMBR		音節正確率(%)	
詞圖 N itr	sMBR τ=130	+pen τ=150	+pen+len τ=16
0	81.42	81.42	81.42
1	82.10	82.13	82.18
2	82.63	82.79	82.76
3	82.91	83.25	83.16
4	82.97	83.42	83.45
5	83.02	83.56	83.58
6	82.80	83.59	83.66
7	82.98	83.66	**83.72**
8	82.50	83.63	83.61
9	82.54	83.63	83.66
10	82.40	83.58	83.63

表 5.15 狀態層級最小貝氏風險－詞圖 N－音節正確率

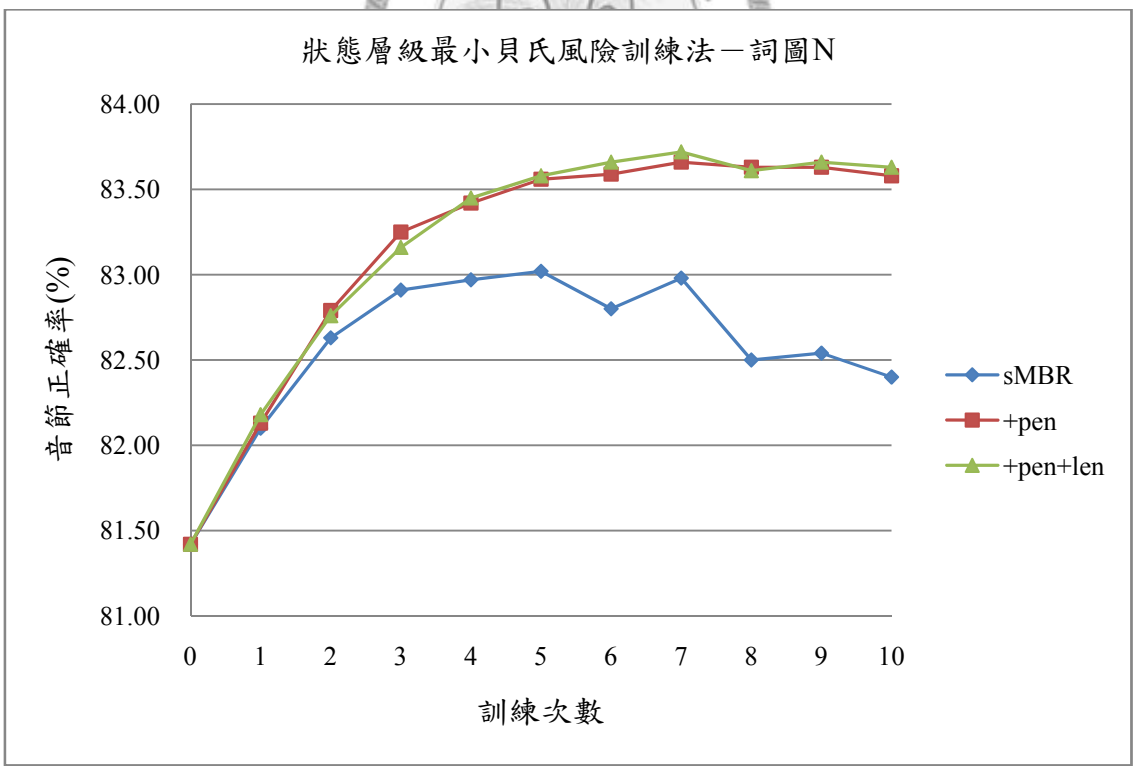


圖 5.25 狀態層級最小貝氏風險－詞圖 N－音節正確率

sMBR		聲韻母正確率(%)	
詞圖 N itr	sMBR τ=130	+pen τ=150	+pen+len τ=16
0	84.76	84.76	84.76
1	85.35	85.38	85.44
2	85.77	85.94	85.90
3	86.01	86.33	86.26
4	86.07	86.53	86.52
5	86.09	86.66	86.63
6	85.90	86.63	86.69
7	85.99	86.66	**86.72**
8	85.48	86.61	86.60
9	85.56	86.61	86.63
10	85.45	86.55	86.58

表 5.16 狀態層級最小貝氏風險－詞圖 N－聲韻母正確率

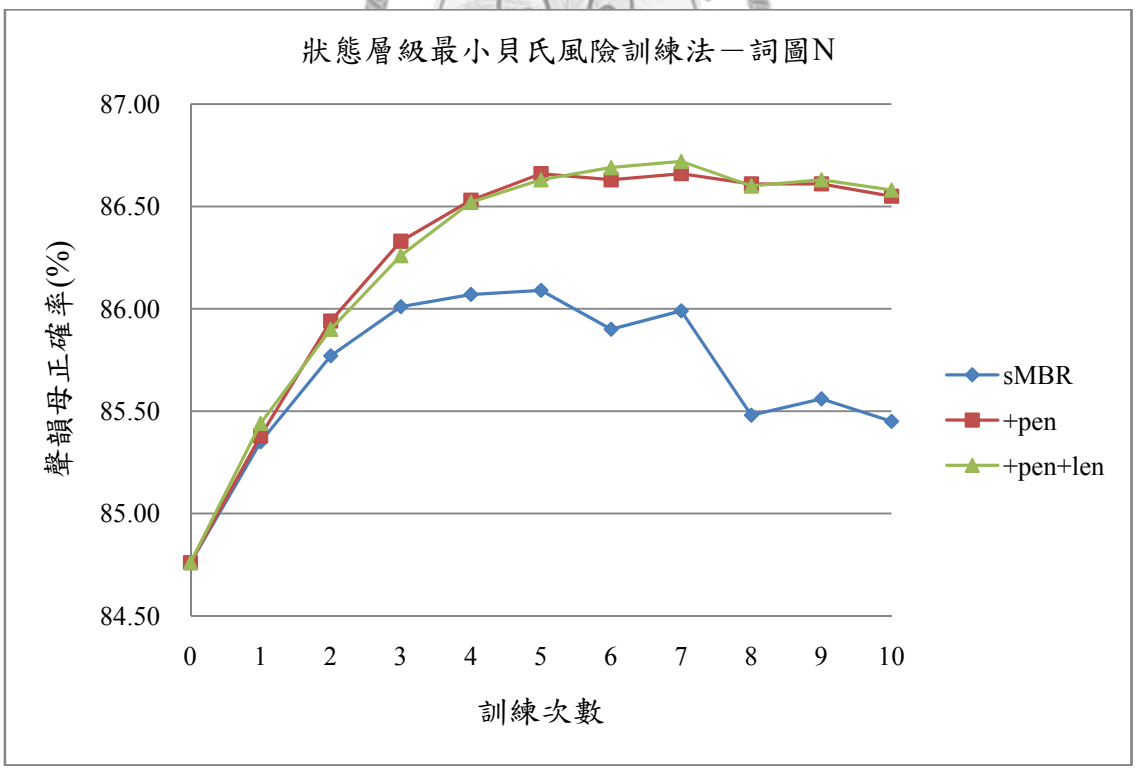


圖 5.26 狀態層級最小貝氏風險－詞圖 N－聲韻母正確率

sMBR		詞正確率(%)	
詞圖 T itr	sMBR τ=115	+pen τ=130	+pen+len τ=15
0	57.99	57.99	57.99
1	58.64	58.65	58.80
2	59.31	59.12	59.18
3	60.03	59.81	59.51
4	60.42	60.13	59.77
5	60.61	60.35	59.84
6	60.64	60.37	60.05
7	**60.75**	60.26	60.15
8	60.44	60.31	60.04
9	60.61	60.38	60.06
10	60.49	60.39	60.04

表 5.17 狀態層級最小貝氏風險－詞圖 T－詞正確率

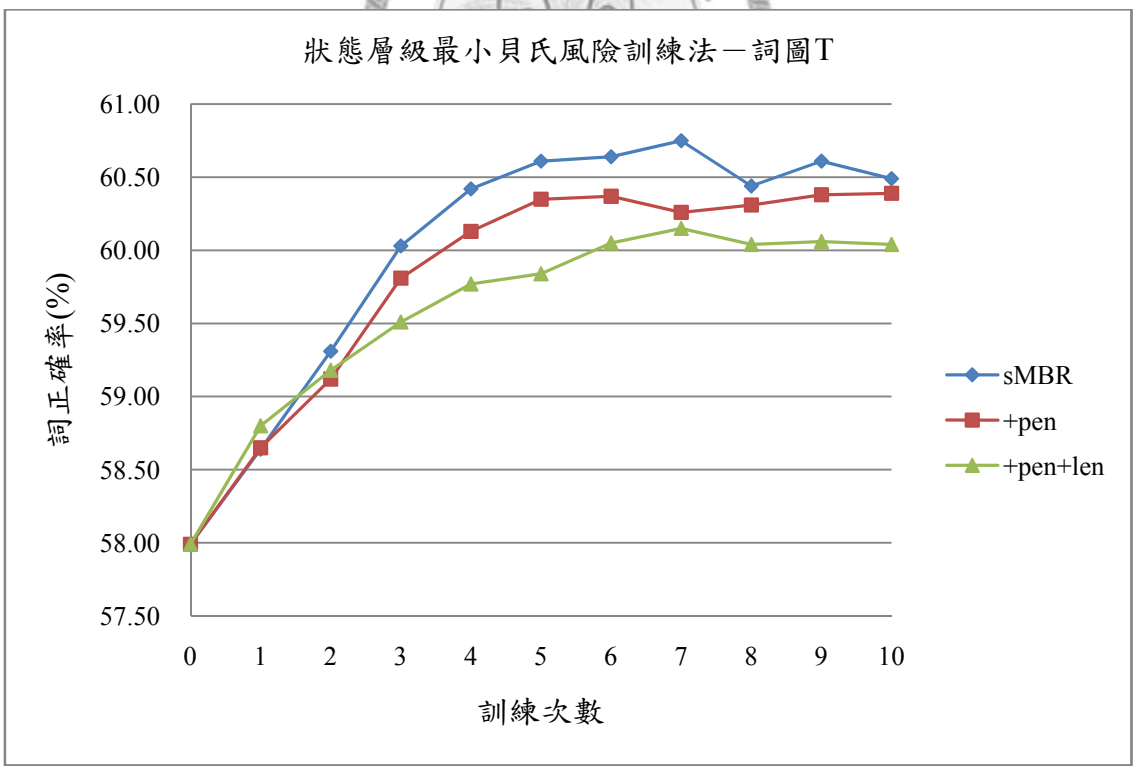


圖 5.27 狀態層級最小貝氏風險－詞圖 T－詞正確率

sMBR		字正確率(%)	
詞圖 T itr	sMBR τ=115	+pen τ=130	+pen+len τ=15
0	75.17	75.17	75.17
1	75.67	75.79	75.92
2	76.15	76.31	76.45
3	76.46	76.91	76.72
4	76.54	77.06	77.00
5	76.46	**77.17**	77.06
6	76.13	77.11	77.15
7	76.07	77.03	77.09
8	75.54	76.91	76.96
9	75.48	76.82	76.85
10	75.32	76.79	76.82

表 5.18 狀態層級最小貝氏風險－詞圖 T－字正確率

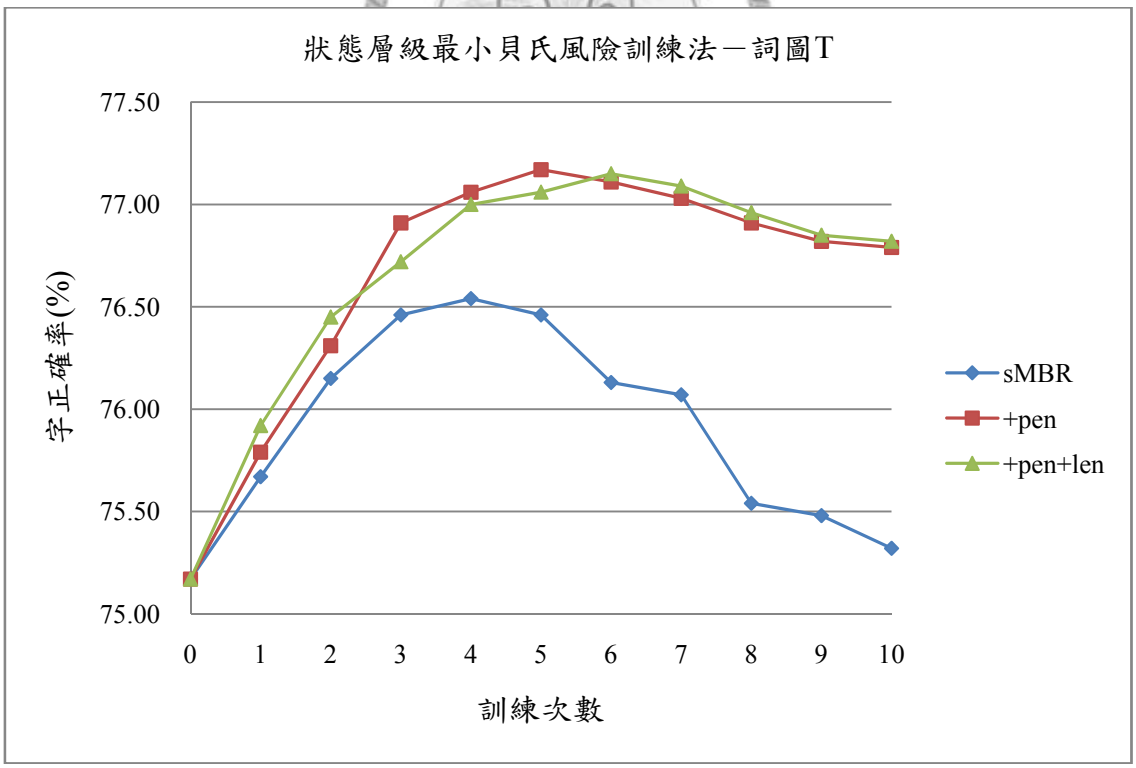


圖 5.28 狀態層級最小貝氏風險－詞圖 T－字正確率

sMBR		音節正確率(%)	
詞圖 T itr	sMBR τ=115	+pen τ=130	+pen+len τ=15
0	81.42	81.42	81.42
1	82.02	82.14	82.23
2	82.49	82.64	82.73
3	82.68	83.13	83.02
4	82.81	83.29	83.30
5	82.78	83.43	83.37
6	82.52	83.45	83.47
7	82.45	83.41	**83.50**
8	81.96	83.42	83.42
9	81.93	83.31	83.38
10	81.85	83.30	83.38

表 5.19 狀態層級最小貝氏風險－詞圖 T－音節正確率

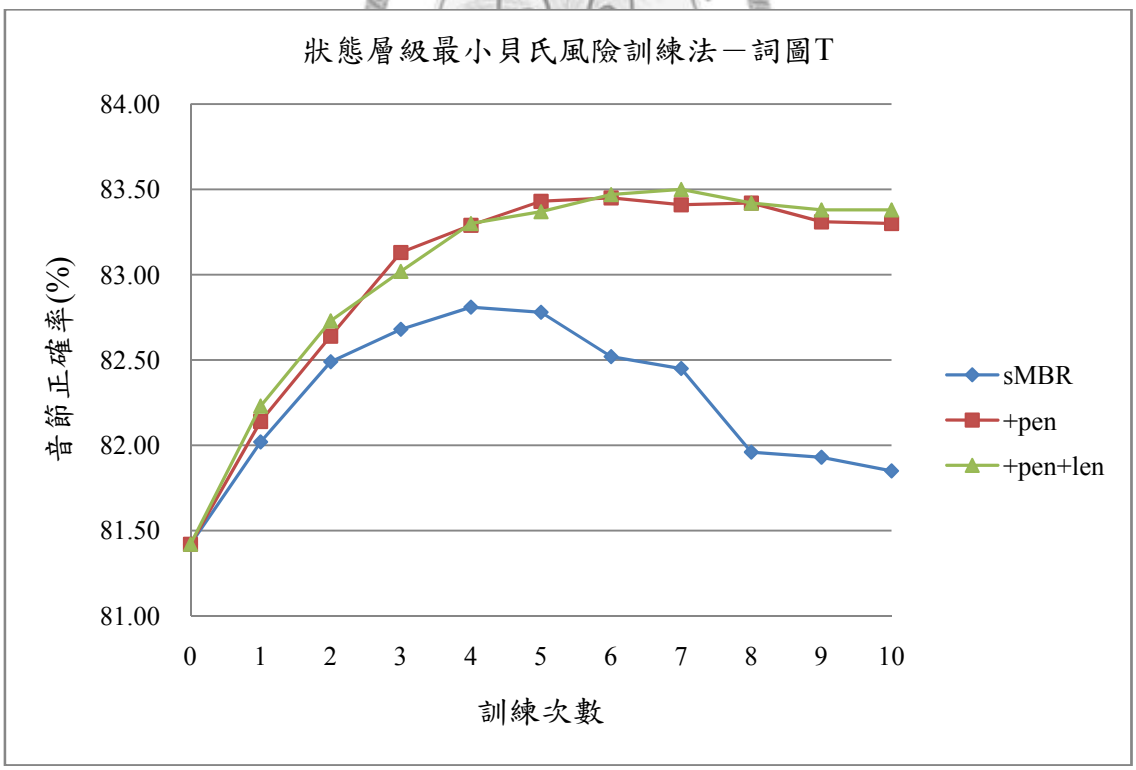


圖 5.29 狀態層級最小貝氏風險－詞圖 T－音節正確率

sMBR		聲韻母正確率(%)	
詞圖 T itr	sMBR τ=115	+pen τ=130	+pen+len τ=15
0	84.76	84.76	84.76
1	85.28	85.38	85.47
2	85.63	85.77	85.85
3	85.81	86.23	86.14
4	85.88	86.39	86.38
5	85.86	86.50	86.45
6	85.66	**86.52**	**86.52**
7	85.57	86.46	**86.52**
8	85.10	86.49	86.45
9	85.04	86.38	86.41
10	84.97	86.36	86.40

表 5.20 狀態層級最小貝氏風險－詞圖 T－聲韻母正確率

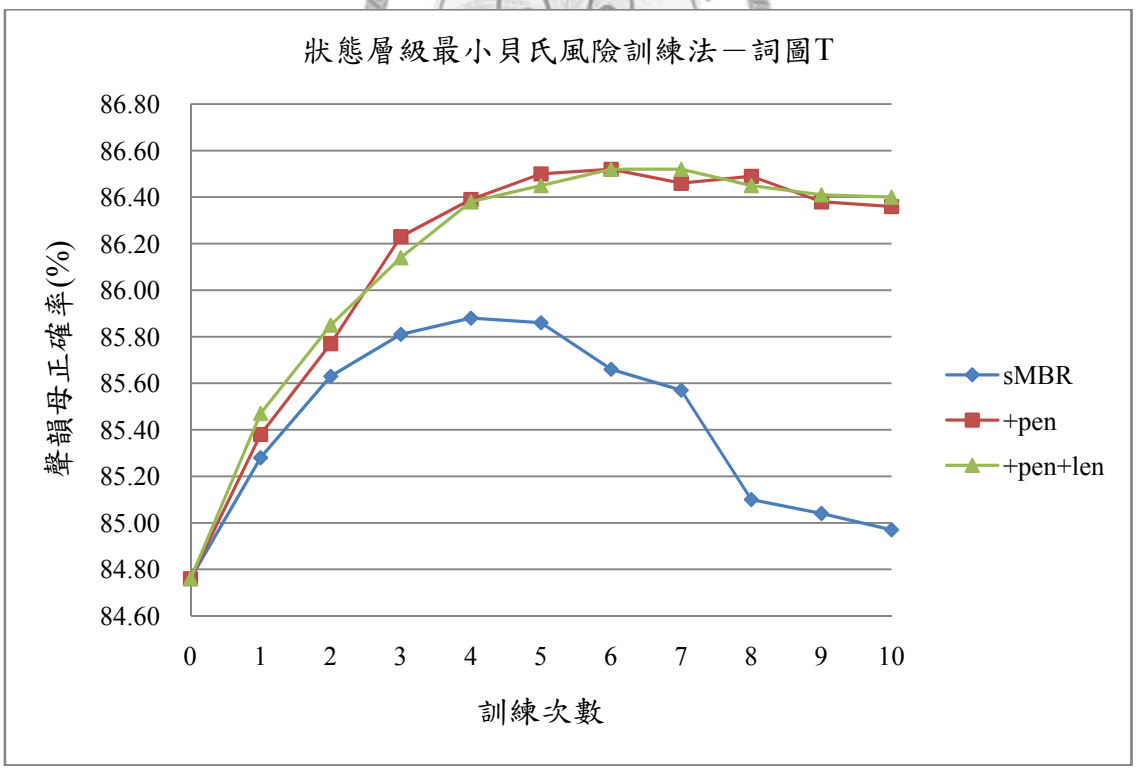


圖 5.30 狀態層級最小貝氏風險－詞圖 T－聲韻母正確率

5.3 最小歧異度訓練

5.3.1 目標函數

最小歧異度在 2006 年由杜氏(Du)等人提出，其正確度的計算方式類似 5.2 節的狀態層級最小貝氏風險，差異在於最小歧異度的正確度不是單純以狀態是否相同為準，而是另外計算辨識狀態與正確狀態兩者的歧異度，歧異度越大則正確度越低。目標函數基本上與狀態層級最小貝氏風險(5.7)式相同，差別在於正確度 $Acc(s_r,u)$ 的計算方式有所不同，最小歧異度計算定義每個音素個別的正確度為：

$$DivergenceAcc(q) = -\sum_{t=start(q)}^{end(q)} D_{KL}\left(q_{state}(t), s_{r,state}(t)\right) \tag{5.15}$$

其中 $D_{KL}\left(q_{state}(t), s_{r,state}(t)\right)$ 代表音素 q 在時間 t 時的狀態 $q_{state}(t)$，與正確轉寫 s_r 在時間 t 時的狀態 $s_{r,state}(t)$ 的 KL 距離(Kullback-Leibler distance)。因為 KL 距離為越相像態差距越小，所以再加上負號以做為正確度。

KL 距離對於兩個維度相同為 d 的多元高斯分佈(multivariate Gaussian distribution) $G_1 \sim N_d(\mu_1,\Sigma_1)$ 和 $G_2 \sim N_d(\mu_2,\Sigma_2)$ 的定義為【35】【36】：

$$D_{KL}(G_1,G_2) = \frac{1}{2}\left[\ln\frac{\det(\Sigma_2)}{\det(\Sigma_1)} + \mathrm{tr}\left(\Sigma_2^{-1}\Sigma_1\right) + (\mu_2-\mu_1)^T\left(\Sigma_2^{-1}\right)(\mu_2-\mu_1) - d\right] \tag{5.16}$$

其中 $\det(\Sigma)$ 代表矩陣 Σ 的行列式(determinant); $\mathrm{tr}(\Sigma)$ 代表矩陣 Σ 的跡數(trace)。

而兩個狀態 S_1, S_2 之間的 KL 距離則定義為：

$$D_{KL}(S_1,S_2) = \sum_{i=1}^{M_1} w\left(m_{S_1,i}\right)\sum_{j=1}^{M_2} w\left(m_{S_2,j}\right) D_{KL}\left(m_{S_1,i}, m_{S_2,j}\right) \tag{5.17}$$

其中 $m_{S_1,i}$ 表示狀態 S_1 的第 i 個高斯混合模型； $w\left(m_{S_1,i}\right)$ 表示高斯混合模型 $m_{S_1,i}$ 的權重(weight)； M_1 表示狀態 S_1 的高斯混合總數。簡而言之，(5.17)式的意義就是將兩個狀態中的高斯混合個取一個兩兩計算 KL 距離，再乘上各自的權重，然後將所有算出來的結果加總，就成為狀態的 KL 距離。而整個文句的狀態音框正確度，就是將每個音素個別的正確度加總：

$$Acc(s_r,u) = -\sum_{q\in u} DivergenceAcc(q) \tag{5.18}$$

最小歧異度正確度的計算方試與 5.2.1 節狀態音框正確度的計算方式類似，差別只在於比對狀態的時候，累加的值是依據(5.15)式算出來的為準。

5.3.2 實驗結果

本實驗同樣以強制對齊得到正確轉寫的狀態對齊。而在每次訓練開始前，必需先計算出所有聲學模型中任意兩個狀態的 KL 距離。另外實驗的狀態的 KL 距離有分為每次疊代皆不更新(static)、每次疊代皆更新(renew)、以及將 KL 距離利用 S 形函數對應到 $-1\sim1$ 範圍中(range)的三個版本，實驗結果如圖 5.31～圖 5.34 所呈現。由圖中可以發現，這三個版本無論在哪一種情況下正確率都不會進步，至於退步的原因則有很多可能，因為最小歧異度訓練法遠不如之前所提的方法來得單純，僅僅距離的計算就有很多變化，例如某些聲韻母由於在訓練語句中出現次數太少，導致該模型的共變異矩陣皆近 0 矩陣，於是算出的距離就接近無限大；以及部份同一狀態內的高斯混合彼此差異就很大，造成這些狀態與自己的距離還大於跟其它某些狀態的距離；還有把距離對應該需求範圍內的 S 形函數，也有參數需要調整以控制飽和數值的界限，這種種許多因素的存在，使得最小歧異度訓練法企圖找到有效的正確度計算方式變得十分困難。因為可調整選擇太多，因此最後沒有嘗試出有效的方法。

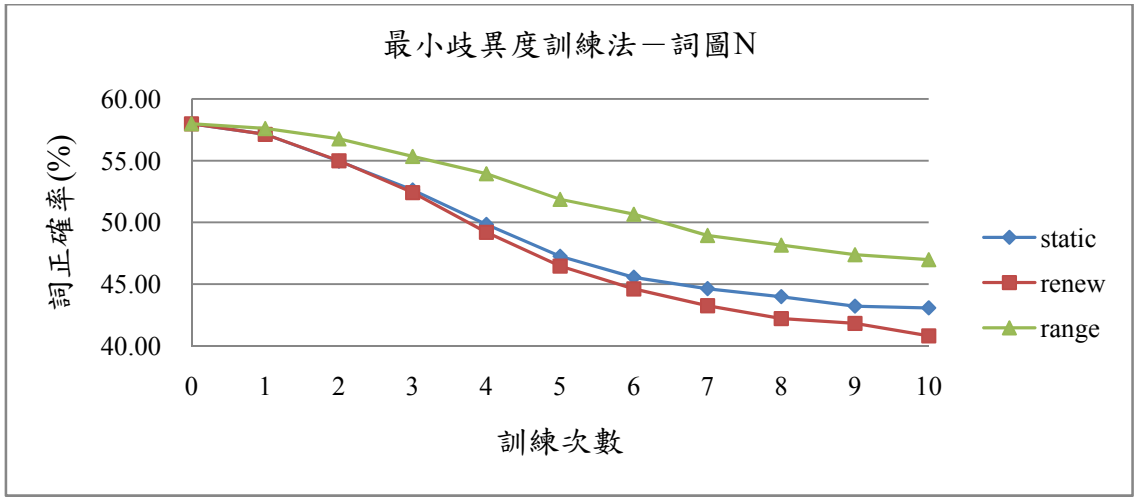


圖 5.31 最小歧異度訓練法－詞圖 N－詞正確率

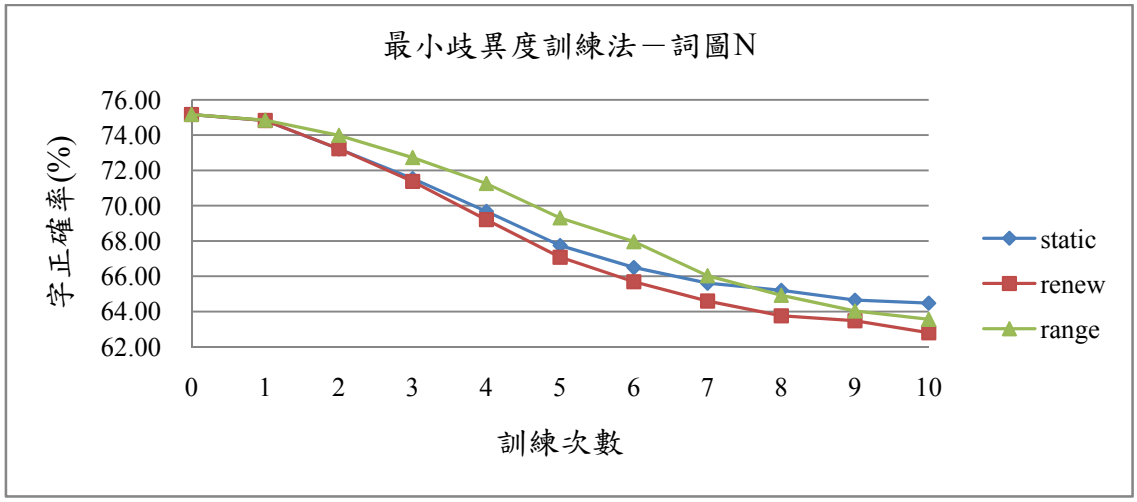


圖 5.32 最小歧異度訓練法－詞圖 N－字正確率

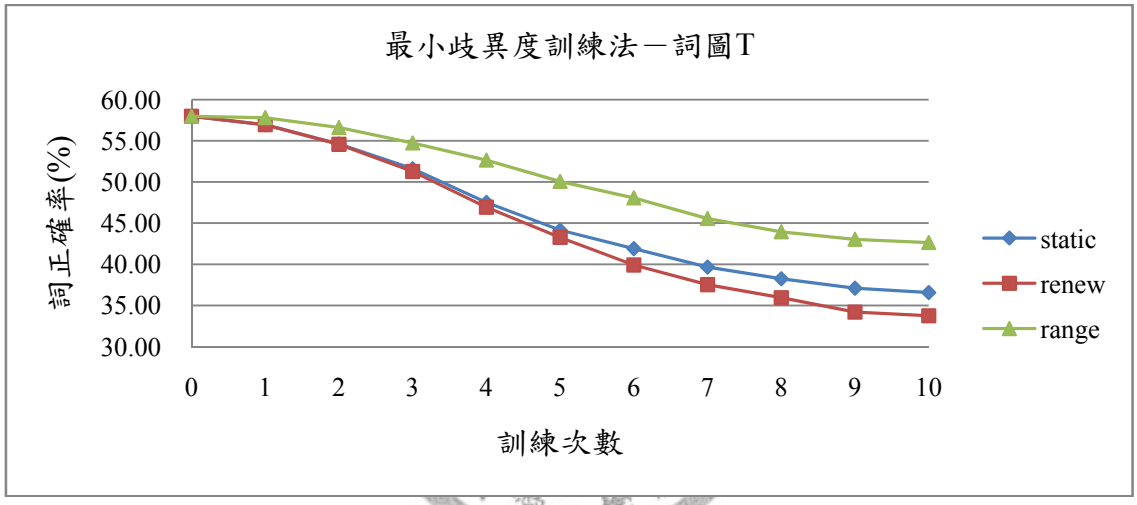


圖 5.33 最小歧異度訓練法－詞圖 T－詞正確率

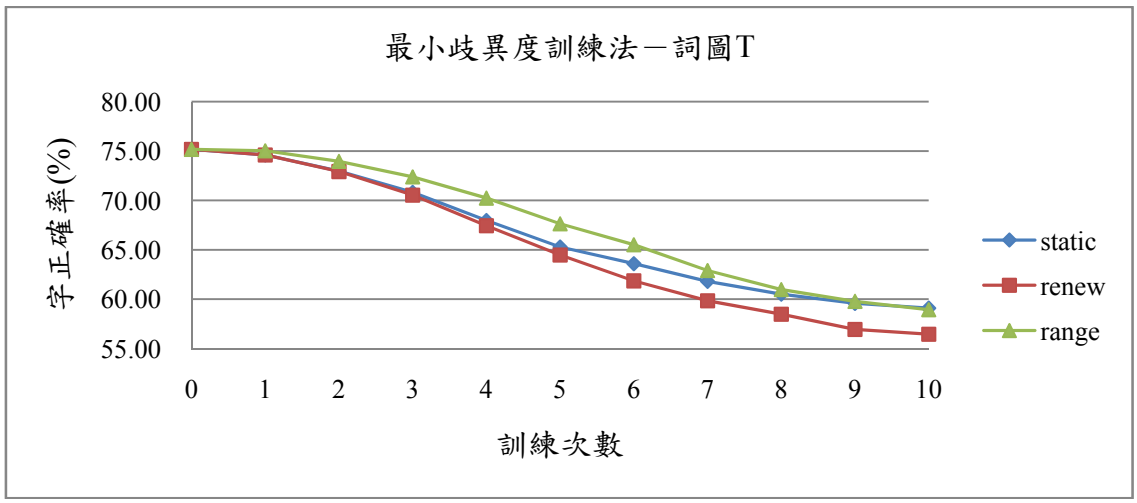


圖 5.34 最小歧異度訓練法－詞圖 T－字正確率

5.4 本章結論

本章介紹了最小音素音框錯誤、狀態層級最小貝氏風險、最小歧異度，三種最小音素錯誤的改進方法，與最小音素錯誤的比較如圖 5.35～圖 5.38 所示，這個比較與【32】中的類似，然而結果不大相同。本論文中最小歧異度並沒有進步的效果，故不畫入圖中；在詞正確率上最小音素音框錯誤有最好的表現，而在字正確率上則依然是最小音素錯誤的結果最佳，狀態層級最小貝氏風險的表現則多為居中。

另外本章也在最小音素音框錯誤及狀態層級最小貝氏風險上修改正確度計算方法，有加入錯誤處罰以及音素長度正規化的版本，實驗結果顯示加入錯誤處罰之後，字正確率會有明顯提升，詞正確率則反而會下降；而在加入錯誤處罰之後，再加入音素長度正規化則不會使正確率出現明顯改變。

而在這些方法之中，字正確率表現最好的方法，在詞圖 N 上是加入錯誤處罰的最小音素音框錯誤訓練法 77.67%，較最小音素錯誤訓練法進步 0.04%(相對 0.18%)；而在詞圖 T 上仍是最小音素錯誤訓練法最佳 77.48%。而詞正確率表現最好的方法，兩種詞圖都是最小音素音框錯誤訓練法最好，在詞圖 N 上正確率為 61.69%，較最小音素錯誤訓練法進步 0.82%(相對 2.10%)；在詞圖 T 上正確率為 61.95%，較最小音素錯誤訓練法進步 1.08%(相對 2.76%)。

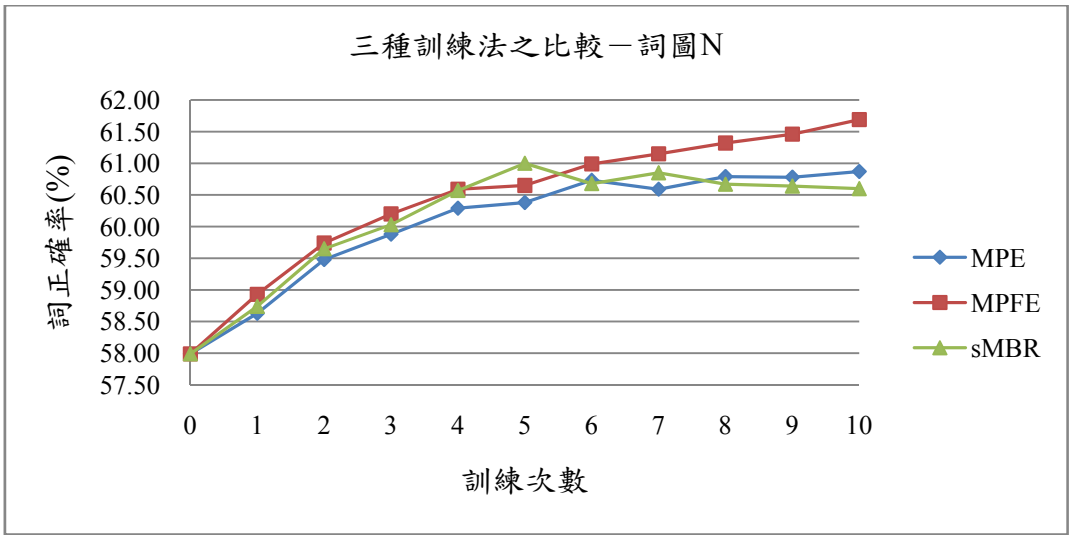


圖 5.35 三種訓練法之比較－詞圖 N－詞正確率

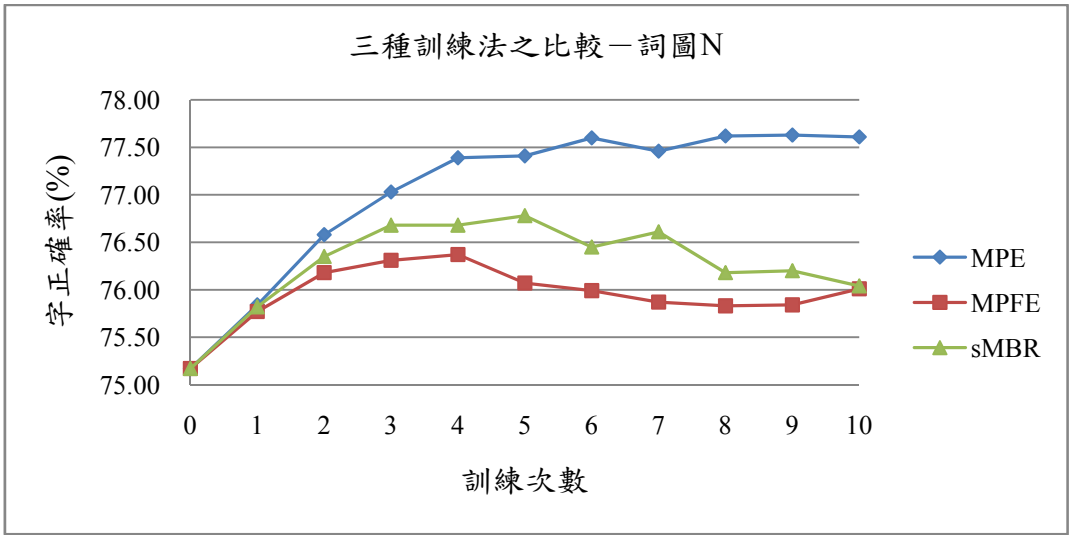


圖 5.36 三種訓練法之比較－詞圖 N－字正確率

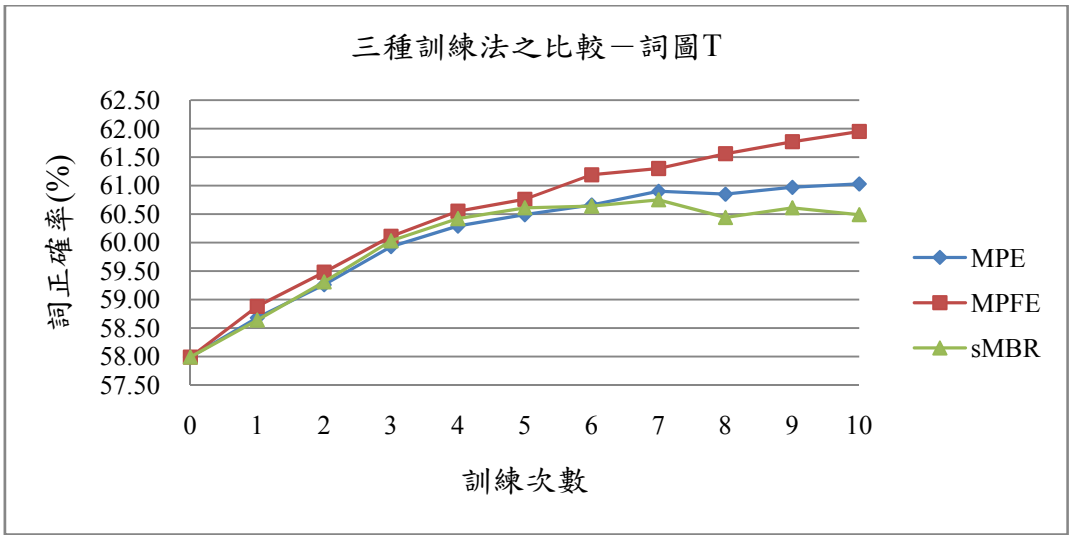


圖 5.37 三種訓練法之比較－詞圖 T－詞正確率

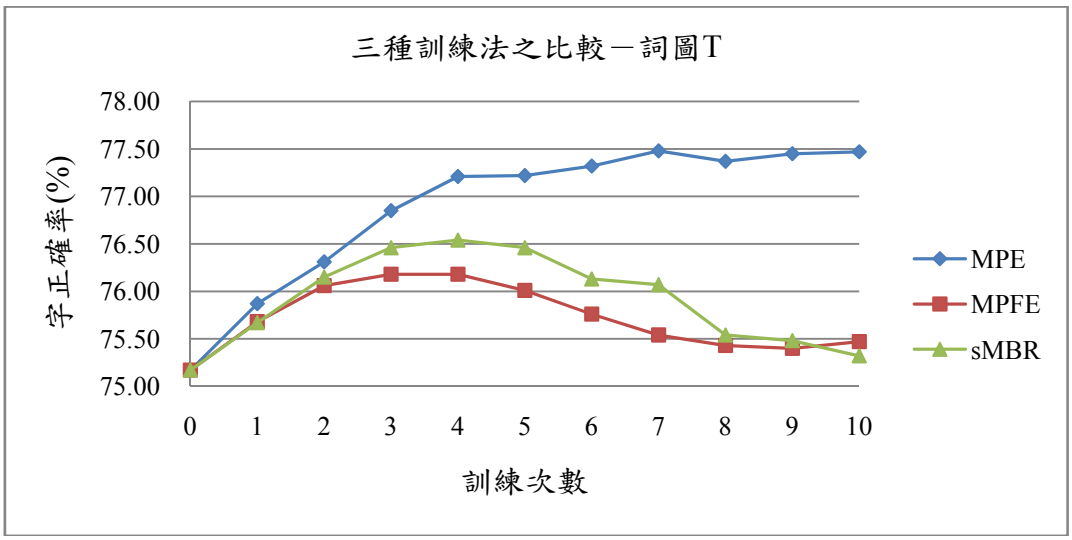


圖 5.38 三種訓練法之比較－詞圖 T－詞正確率

第6章　最小音素錯誤與最小音素音框錯誤的資料選取

本章將介紹一種資料選取的方式，為基於詞弧期望正確度來選取詞弧的方法。本論文將此方法實行於最小音素錯誤與加入錯誤處罰與音素長度正規化的最小音素音框錯誤訓練法上。本章最後會對本論文提出之方法做一個綜合整理。

6.1　基於詞弧期望正確度的資料選取

使用訓練語料訓練模型參數的方法，在訓練語料增加的情況下，往往可以讓訓練出來的模型更為強健。然而，模型的辨識率並不會隨著訓練語料的增加一起增加下去，通常會在到達某個程度之後就已經飽和，不再增加辨識率。資料選取的意義就在於在訓練語料中挑選出對模型參數估測最有效果的部份，讓模型的訓練更有效率。

本章資料選取方式的概念是由寬邊界隱藏式馬可夫模型(Large Margin HMMs)【37】而來。寬邊界隱藏式馬可夫模型的想法是參考支撐向量機(Support Vector Machine, SVM)【38】的分類方式引入隱藏式馬可夫模型，訓練目標不是事後機率的最大化，而是分類邊際的最大化，如圖 6.1【39】中，將圖(a)的分類邊際調整到圖(b)的位置，就是寬邊界隱藏式馬可夫模型的訓練目標。觀察這種類型的訓練方式可以發現，在尋找最大邊界時，每筆資料對訓練的影響程度是不同的，接近邊界部份的資料會對訓練結果影響較為重要，而不同於事後機率的最大化中，所有資料都是同等重要的。

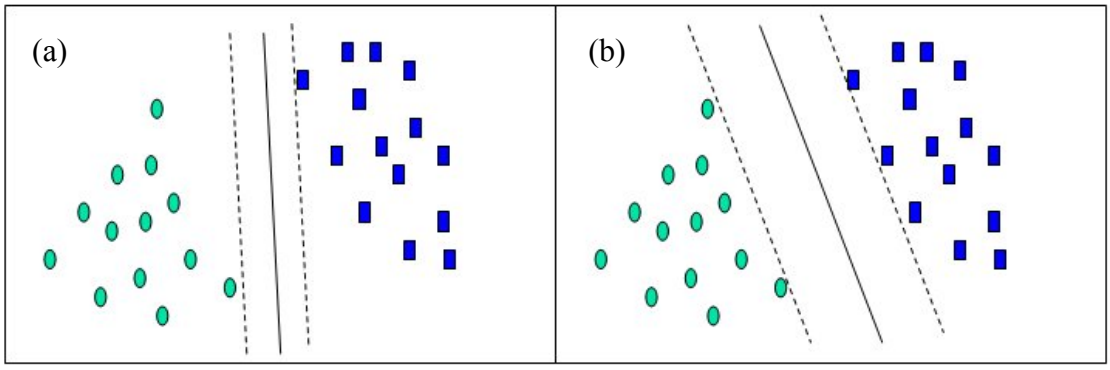


圖 6.1　分類邊際的最大化

將這種訓練目標引用到最小音素錯誤訓練法，在最小音素錯誤訓練法中，認為所謂邊界的資料就是指臨界於辨識成正確與錯誤的資料。套用到詞弧的統計上，就是在詞圖中的每一段詞弧都會分別計算出詞弧的期望正確度，依據詞弧期望正確度是大於或小於整個詞圖的期望正確度，當做是辨識為正確或錯誤的資料，以此將詞弧分為分子詞圖與分母詞圖兩個部份。另外再將詞弧的期望正確度與整個詞圖的期望正確度的差值視為距離邊界的大小，因此基於邊界最大化的原則，對於距離邊界較近的詞弧應該做為主要的訓練資料，於是在詞弧分至分子或分母詞圖前，先依照該詞弧的期望正確度C_q與整個詞圖的期望正確度C_{avg}的差距，決定該詞弧是否為邊界的資料。而決定是否要加入訓練的原則為，該詞弧的期望正確率越接近整個詞圖的期望正確度，就越傾向將該段詞弧加入訓練；反之若該詞弧的期望正確率越遠離整個詞圖的期望正確度，就越傾向不將該段詞弧加入訓練。

本論文中詞弧的選取方式是先對詞弧期望正確度及整個詞圖的期望正確度正規化，由於詞弧正確度的計算方式為計算每個音素的音素正確度之累加，因此音素數量會跟詞弧正確度有關，詞弧正確度除以音素數量後，成為每個音素的平均正確度，詞弧正確度就不會跟音素數量有關，使期望正確度落在一定的範圍內。之後再對詞弧期望正確度與整個詞圖的期望正確度的差值設定一個閥值(threshold)，然後只挑選差值落在閥值範圍內的詞弧加入訓練，超出閥值範圍的詞弧則忽略不計【40】【41】。於是最小音素錯誤就是將(4.18)式改為：

$$\gamma_q^{MPE} = \sum_r \gamma_q^r \left[\left(C_r(q) - C_{avg}^r \right) I\left(C_{avg}^r \in \Omega_r \right) \right]$$

$$I\left(C_{avg}^r \in \Omega_r \right) = \begin{cases} 1 & \text{if } -\alpha \le \dfrac{\left(C_r(q) - C_{avg}^r \right)}{len(r)} \le \beta \\ 0 & \text{otherwise} \end{cases} \tag{6.1}$$

其中$len(r)$是指第r句語句中音素的數量；$\left(C_r(q) - C_{avg}^r \right) / len(r)$就是期望正確度的正規化；$\alpha$與$\beta$是決定詞弧是否加入訓練的閥值。因此只有正規化的詞弧期望正確度落在閥值範圍內的詞弧會加入最小音素錯誤訓練法的詞圖統計之中。

6.2 實驗結果

本章詞弧選擇的方法分別實行在最小音素錯誤訓練法以及加入錯誤處罰和音素長度正規化的最小音素音框錯誤訓練法上。由於資料選取的閥值會對結果有很大的影響，因此首先必需觀察訓練語料中平均正確度的分佈情形，分佈情形如圖 6.3～圖 6.6，由於當 $C_r(q)-C_{avg}^r=0$ 時該段詞弧不會加入訓練，因此圖中的統計資訊不包含 $C_r(q)-C_{avg}^r=0$ 的次數。

由圖中可以看出，絕大多數的詞弧平均正確度差值都很靠近 0，而且小於 0 的數量多於大於 0 的數量。此外，由於在最大相似度估測法的初始模型中，已經針對正確轉寫訓練過，在之後的鑑別式訓練法中，應該加強的是對錯誤競爭字串的鑑別能力，因此在最小音錯誤訓練法中使用的詞弧選擇閥值如表 6.1 所示：

疊代次數	α	β
1、2	1.00	0.03
3、4	1.00	0.08
5、6	1.00	0.18
8、7	1.00	0.30
9、10	1.00	0.38

表 6.1　最小音素錯誤訓練－詞弧選擇閥值

α=1.00 即表示所有分母詞圖的統計值都選入訓練，而使用 β=0.03 時，會在疊代次數 5 次之後發生嚴重的過度訓練，如圖 6.2 之 init 即為使用 α=1.00 與 β=0.03 疊代 10 次的結果，因此選擇 β 值隨著疊代次數逐漸放大選取範圍，又因為兩種詞圖裡

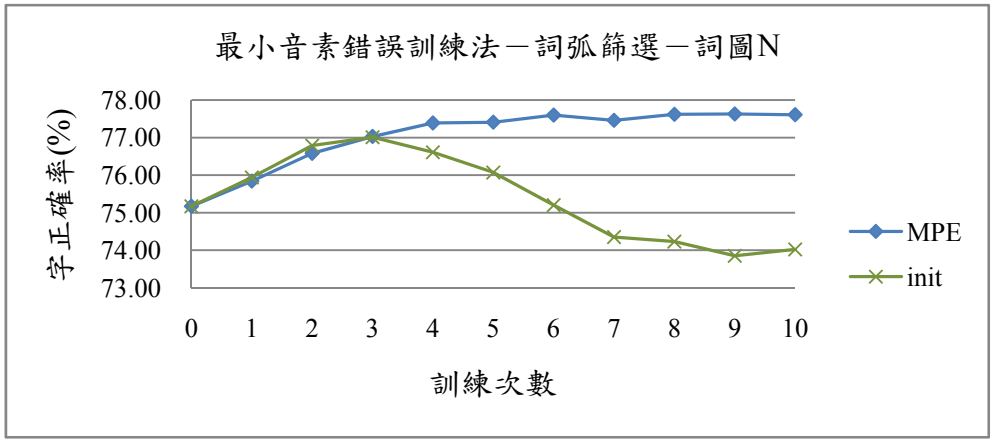


圖 6.2　最小音素錯誤－詞弧篩選－詞圖 N－過度訓練情形

的正確度分佈十分相似，因此兩種詞圖使用的閥值都相同；而在加入錯誤處罰和音素長度正規化的最小音素音框錯誤訓練法上，詞弧選擇閥值則是根據最小音錯誤訓練法中的選擇，選擇到同等比例的詞弧時的值(如原本選擇的值會選到後 60%的詞弧，對應過來的閥值就是也會選到後 60%詞弧的值)，因此在加入錯誤處罰音素長度正規化的最小音素音框錯誤訓練法中對照過來使用的詞弧選擇閥值就如表 6.4 和表 6.5 所示：

疊代次數	α	β
1、2	1.00	0.03302280
3、4	1.00	0.08186985
5、6	1.00	0.18090070
8、7	1.00	0.28907700
9、10	1.00	0.36248670

表 6.2　MPFE－加入錯誤處罰音素長度正規化－詞弧選擇閥值－詞圖 N

疊代次數	α	β
1、2	1.00	0.02723290
3、4	1.00	0.07051789
5、6	1.00	0.15896940
8、7	1.00	0.27925270
9、10	1.00	0.36150360

表 6.3　MPFE－加入錯誤處罰音素長度正規化－詞弧選擇閥值－詞圖 T

由於在兩種詞圖對應到的值並不同，因此使用了不同的閥值。

實驗結果如表 6.4～表 6.11 及圖 6.7～圖 6.14 所示，圖表中 MPE 表示最小音素錯誤訓練法，再加入詞弧篩選則是 MPE+sel；MPFE+pen+len 代表加入錯誤處罰音素長度正規化的最小音素音框錯誤訓練法，MPFE+p+l+sel 則是再加入詞弧篩選的方法。圖表中呈現加入詞弧篩選之前及之後的比較，在詞圖 N 上，MPE+sel 在詞、音節、聲韻母上都有正確率的提升，MPFE+p+l+sel 則在音節、聲韻母上有正確率的提升;在詞圖 T 上,MPE+sel 只有在詞正確率上有提升,MPFE+pen+len+sel 則在字、音節、聲韻母上有正確率的提升。MPFE+pen+len 在加入詞弧篩選後較有進步,在詞圖 T 時字正確率較 MPFE+pen+len 進步 0.09%(0.40%相對)。整體看來，加入詞弧篩選後即使最高正確率沒有提升，大都也有較快的收斂速度。

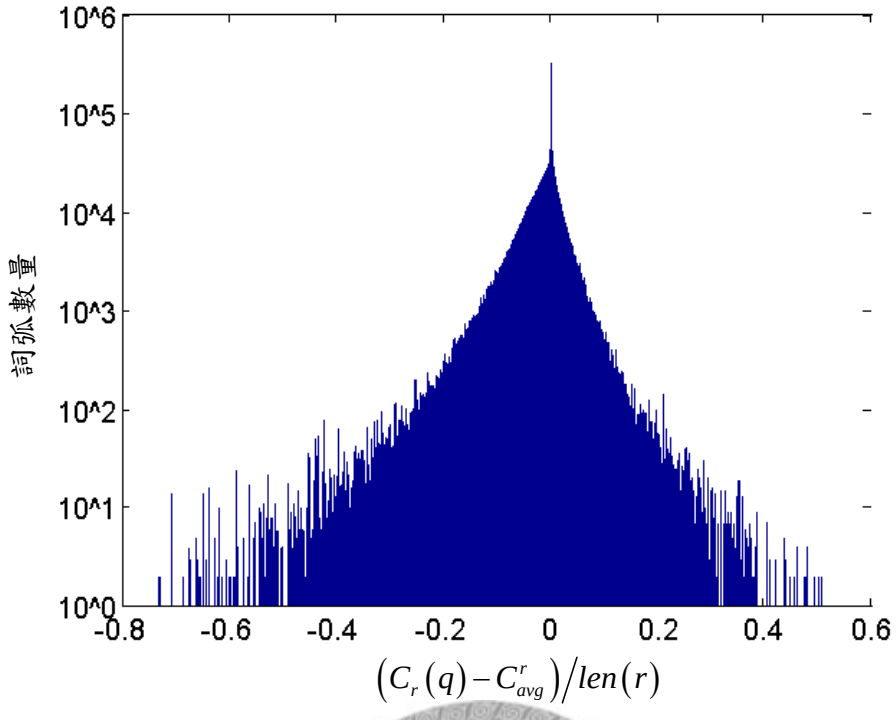


圖 6.3　最小音素錯誤－詞圖 N－詞弧正確度分佈

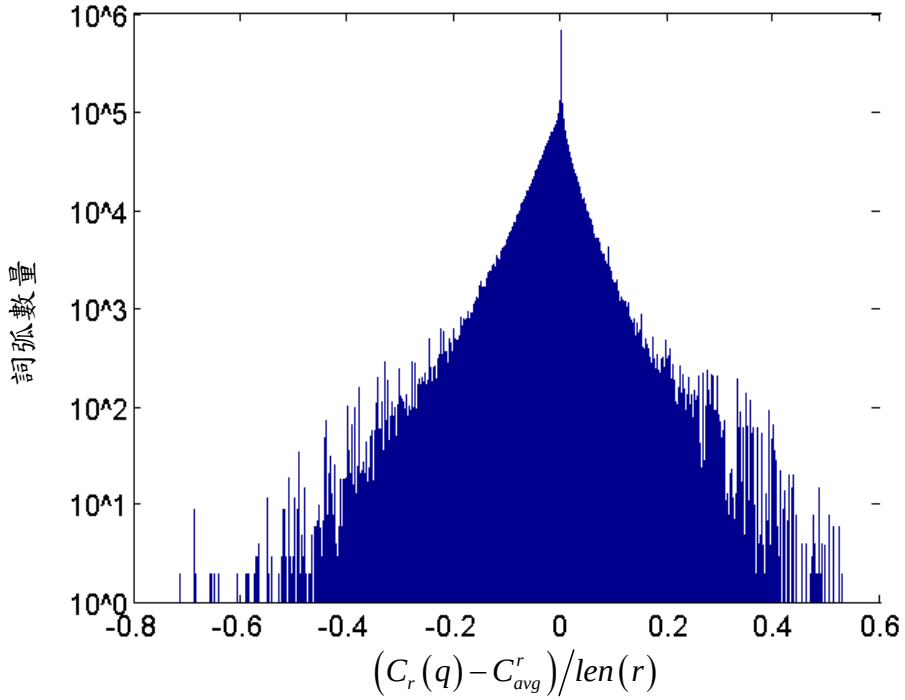


圖 6.4　最小音素錯誤－詞圖 T－詞弧正確度分佈

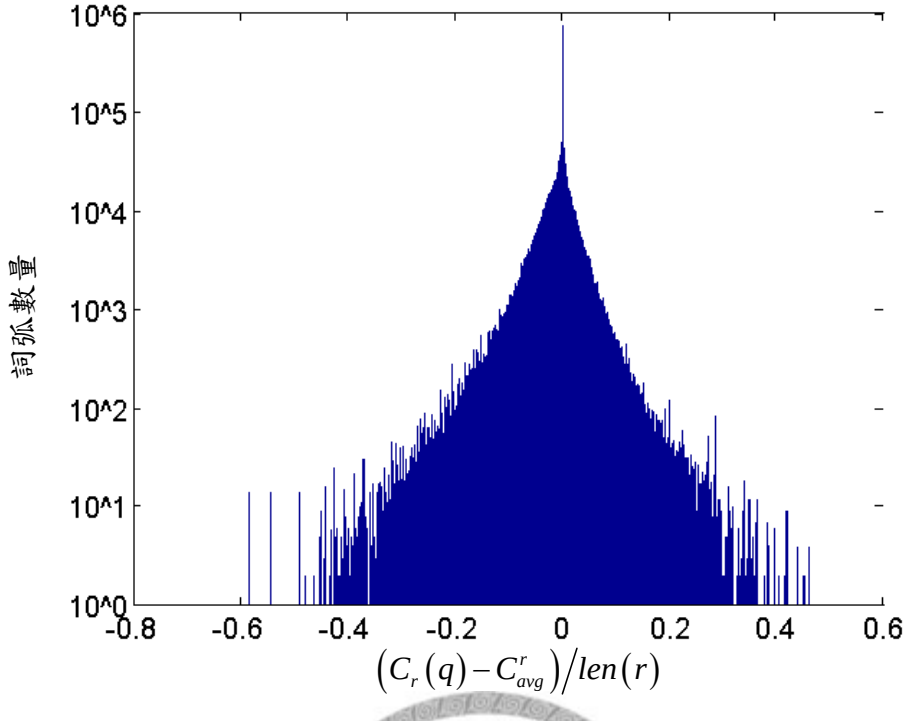


圖 6.5　MPFE+pen+len－詞圖 N－詞弧正確度分佈

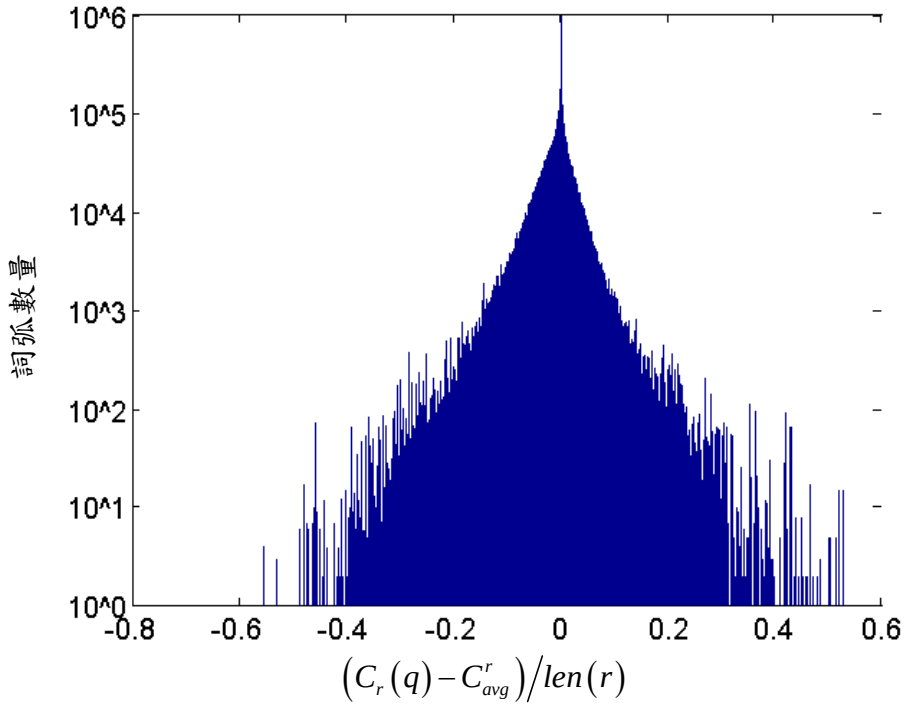


圖 6.6　MPFE+pen+len－詞圖 T－詞弧正確度分佈

詞圖 N		τ =25	詞正確率(%)	
itr	MPE	MPE+sel	MPFE+pen+len	MPFE+p+l+sel
0	57.99	57.99	57.99	57.99
1	58.63	58.74	58.76	58.55
2	59.48	59.67	59.27	59.23
3	59.88	60.03	59.48	59.32
4	60.29	60.25	59.57	59.44
5	60.38	60.52	59.80	59.58
6	60.73	60.81	**59.92**	59.87
7	60.59	60.80	59.86	59.90
8	60.79	**60.92**	59.80	59.90
9	60.78	60.77	59.83	59.73
10	60.87	60.90	59.85	59.74

表 6.4　詞弧篩選－詞圖 N－詞正確率

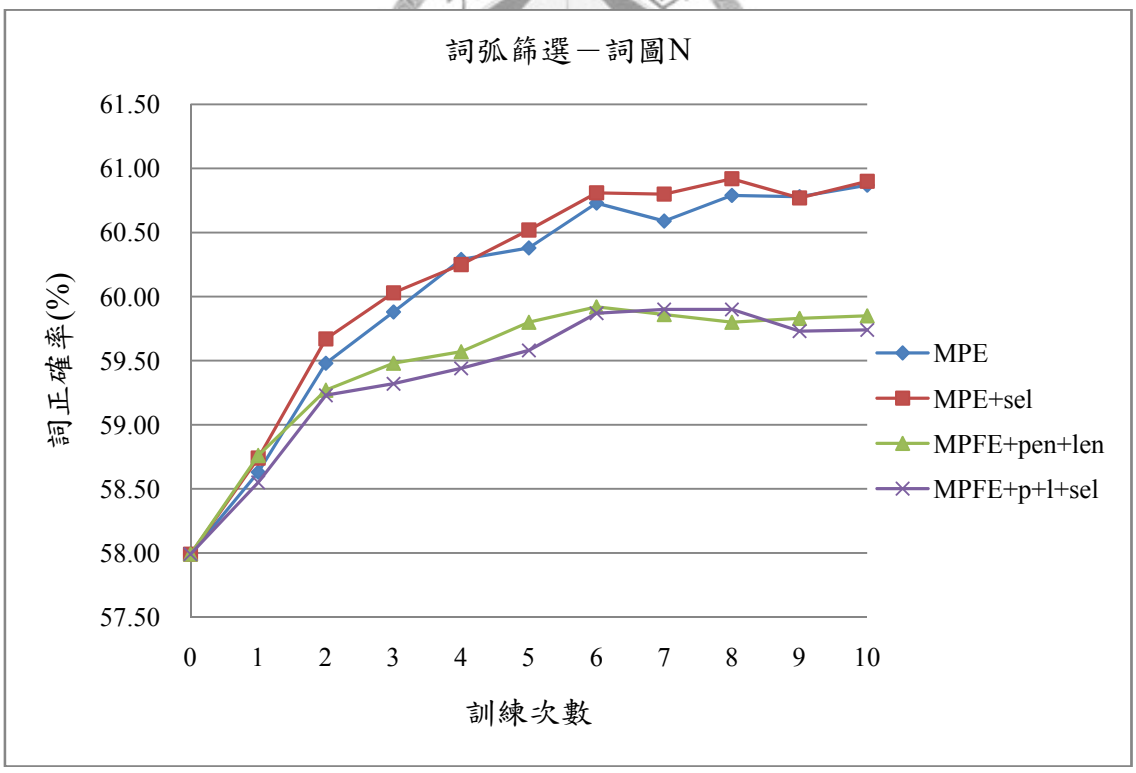


圖 6.7　詞弧篩選－詞圖 N－詞正確率

詞圖 N		τ =25	字正確率(%)	
itr	MPE	MPE+sel	MPFE+pen+len	MPFE+p+l+sel
0	75.17	75.17	75.17	75.17
1	75.84	75.94	75.90	75.83
2	76.58	76.79	76.54	76.69
3	77.03	77.11	76.90	76.86
4	77.39	77.21	77.05	77.06
5	77.41	77.35	77.36	77.23
6	77.60	77.53	77.54	77.49
7	77.46	77.56	77.47	77.56
8	77.62	77.62	77.59	**77.67**
9	**77.63**	77.47	77.65	77.55
10	77.61	77.53	**77.67**	77.62

表 6.5　詞弧篩選－詞圖 N－字正確率

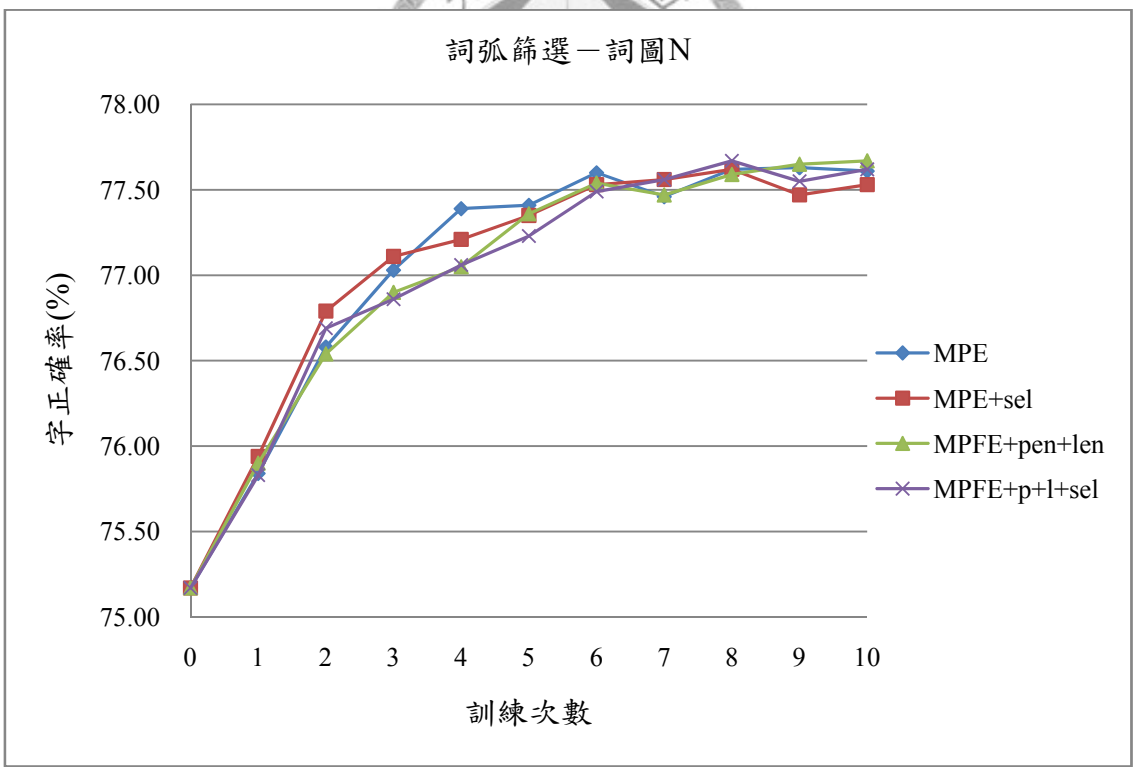


圖 6.8　詞弧篩選－詞圖 N－字正確率

詞圖 N		τ =25	音節正確率(%)	
itr	MPE	MPE+sel	MPFE+pen+len	MPFE+p+l+sel
0	81.42	81.42	81.42	81.42
1	82.15	82.24	82.18	82.17
2	82.80	83.06	82.83	82.99
3	83.23	83.27	83.15	83.20
4	83.58	83.46	83.39	83.46
5	83.65	83.58	83.65	83.61
6	83.82	83.81	83.80	83.89
7	83.77	83.83	83.78	83.91
8	83.85	**83.92**	83.91	83.97
9	83.84	83.70	83.96	83.92
10	83.86	83.76	84.00	**84.01**

表 6.6　詞弧篩選－詞圖 N－音節正確率

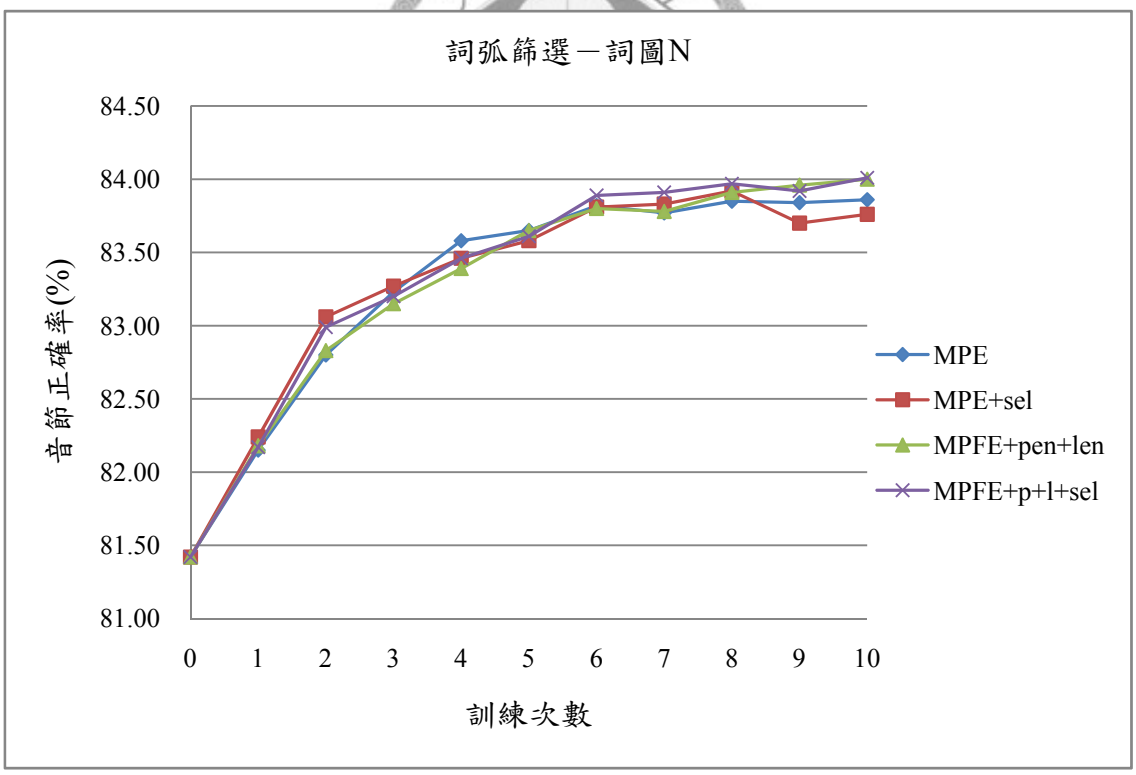


圖 6.9　詞弧篩選－詞圖 N－音節正確率

詞圖 N		τ =25	聲韻母正確率(%)	
itr	MPE	MPE+sel	MPFE+pen+len	MPFE+p+l+sel
0	84.76	84.76	84.76	84.76
1	85.47	85.54	85.43	85.43
2	86.00	86.20	85.98	86.13
3	86.38	86.42	86.32	86.38
4	86.66	86.62	86.52	86.61
5	86.72	86.70	86.75	86.71
6	86.88	86.89	86.89	86.90
7	86.83	86.88	86.82	86.90
8	86.87	**86.94**	86.92	86.96
9	86.86	86.76	86.93	86.92
10	86.88	86.79	86.95	**86.98**

表 6.7　詞弧篩選－詞圖 N－聲韻母正確率

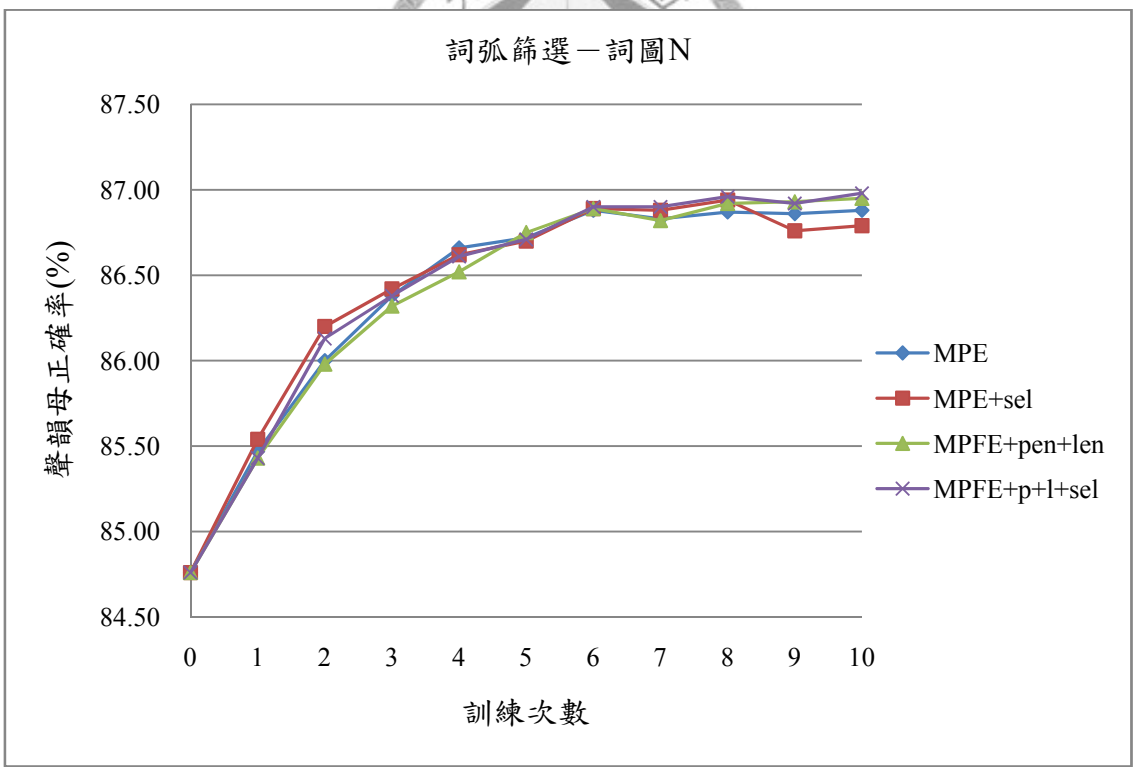


圖 6.10　詞弧篩選－詞圖 N－聲韻母正確率

詞圖 T		τ =25	詞正確率(%)	
itr	MPE	MPE+sel	MPFE+pen+len	MPFE+p+l+sel
0	57.99	57.99	57.99	57.99
1	58.68	58.78	58.71	58.67
2	59.26	59.66	59.08	59.04
3	59.93	60.21	59.42	59.39
4	60.29	60.19	59.30	59.33
5	60.49	60.49	59.51	59.41
6	60.66	60.71	59.46	59.44
7	60.90	60.84	59.37	59.45
8	60.85	60.89	**59.54**	59.32
9	60.97	61.07	59.41	59.46
10	61.03	**61.18**	59.48	59.48

表 6.8　詞弧篩選－詞圖 T－詞正確率

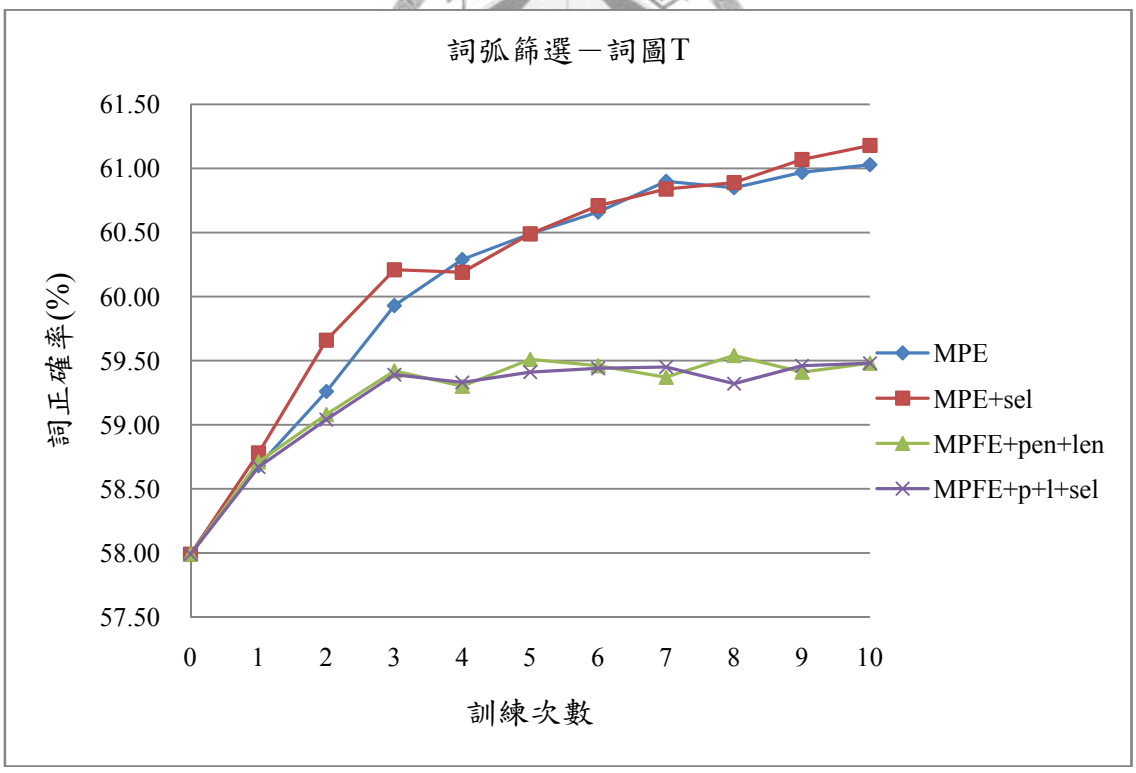


圖 6.11　詞弧篩選－詞圖 T－詞正確率

詞圖 T		τ =25	字正確率(%)	
itr	MPE	MPE+sel	MPFE+pen+len	MPFE+p+l+sel
0	75.17	75.17	75.17	75.17
1	75.87	75.94	75.87	75.95
2	76.31	76.63	76.49	76.52
3	76.85	76.99	76.91	77.04
4	77.21	76.97	77.01	77.21
5	77.22	77.07	77.19	77.36
6	77.32	77.20	77.18	77.39
7	**77.48**	77.33	77.17	77.37
8	77.37	77.30	77.36	77.30
9	77.45	77.43	77.34	**77.45**
10	77.47	77.37	77.36	77.44

表 6.9　詞弧篩選－詞圖 T－字正確率

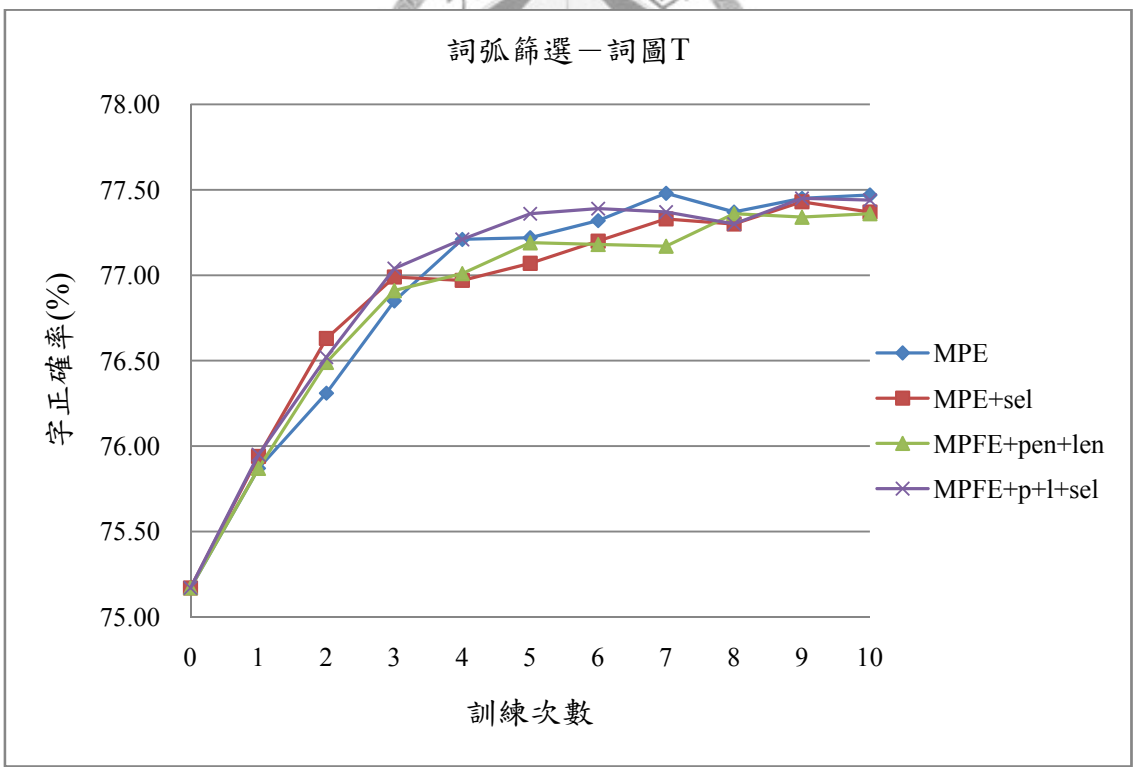


圖 6.12　詞弧篩選－詞圖 T－字正確率

詞圖 T		τ =25	音節正確率(%)	
itr	MPE	MPE+sel	MPFE+pen+len	MPFE+p+l+sel
0	81.42	81.42	81.42	81.42
1	82.14	82.30	82.17	82.28
2	82.59	82.97	82.78	82.96
3	82.99	83.31	83.16	83.41
4	83.35	83.30	83.31	83.63
5	83.35	83.37	83.53	83.77
6	83.49	83.46	83.58	83.84
7	83.58	83.55	83.55	83.78
8	83.54	83.50	83.75	83.76
9	83.65	83.63	83.69	83.82
10	**83.71**	83.64	83.71	**83.86**

表 6.10 詞弧篩選－詞圖 T－音節正確率

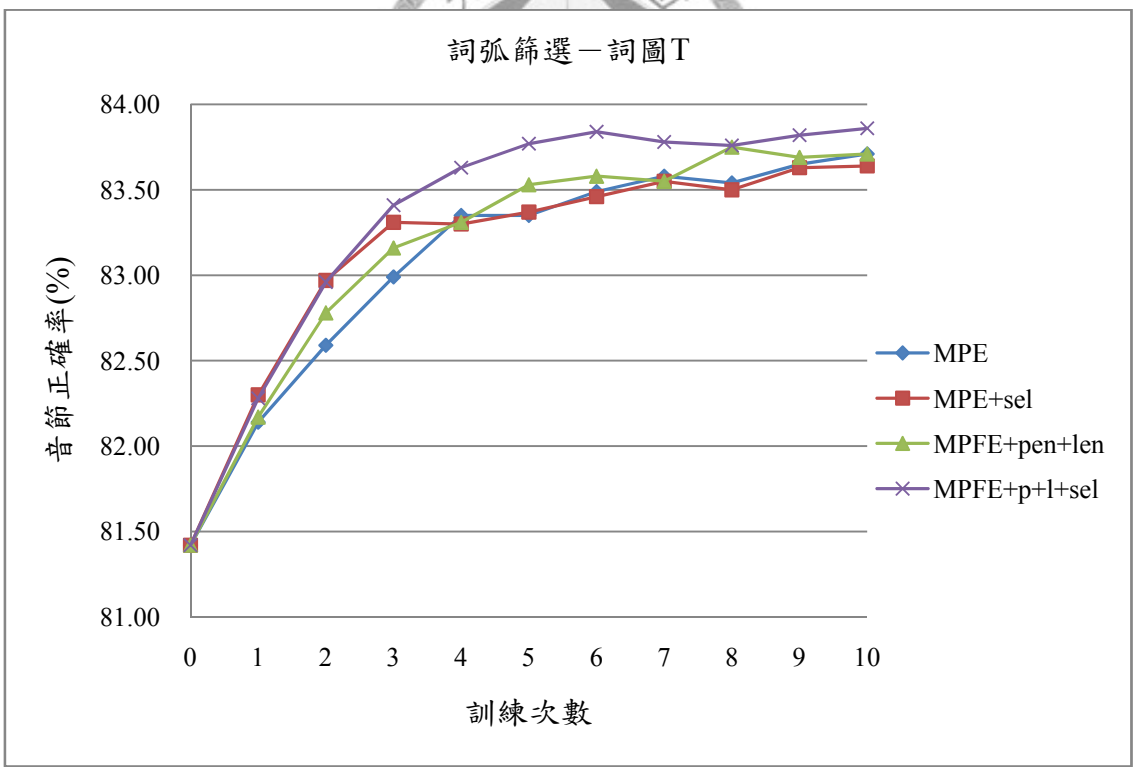


圖 6.13 詞弧篩選－詞圖 T－音節正確率

詞圖 T		τ =25	聲韻母正確率(%)	
itr	MPE	MPE+sel	MPFE+pen+len	MPFE+p+l+sel
0	84.76	84.76	84.76	84.76
1	85.42	85.60	85.43	85.54
2	85.78	86.15	85.92	86.11
3	86.12	86.45	86.28	86.45
4	86.45	86.41	86.41	86.67
5	86.45	86.47	86.58	86.77
6	86.58	86.56	86.64	**86.84**
7	86.66	86.64	86.64	86.79
8	86.62	86.62	86.81	86.79
9	86.71	86.69	86.75	**86.84**
10	**86.74**	86.69	86.77	86.82

表 6.11 詞弧篩選－詞圖 T－聲韻母正確率

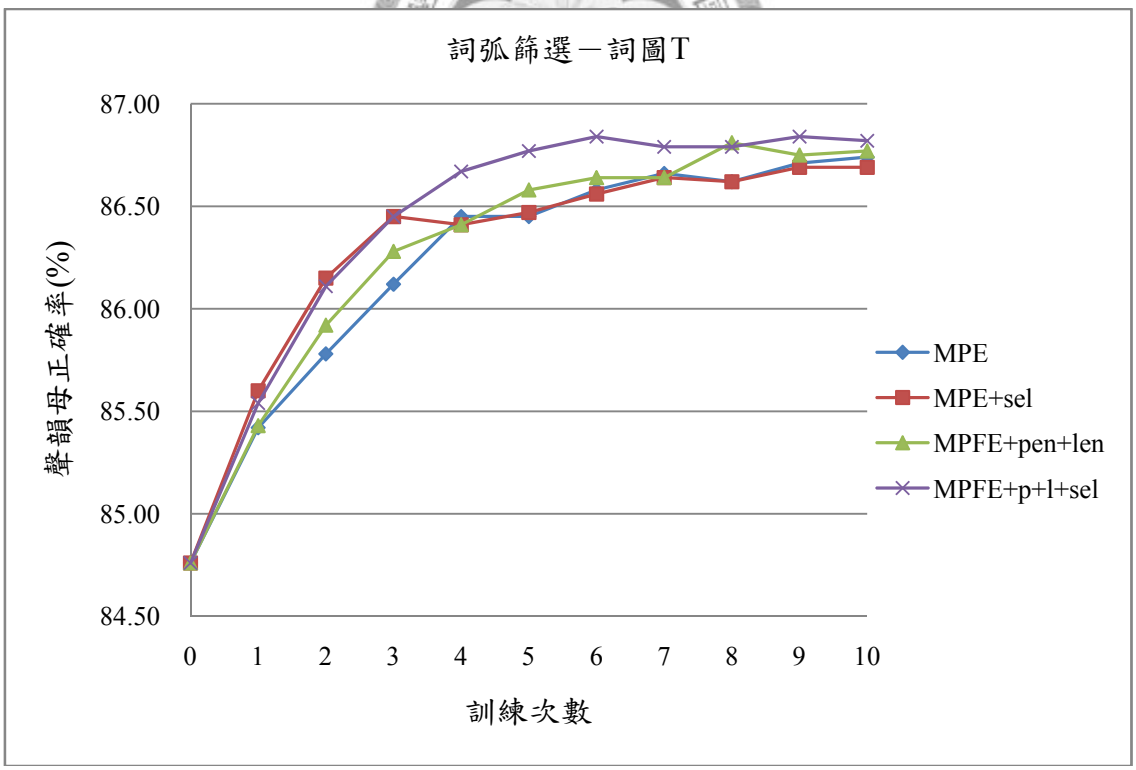


圖 6.14 詞弧篩選－詞圖 T－聲韻母正確率

6.3 各實驗綜合整理

本節列出本論文中所有實驗的結果，列出的數據為每個方法在以 5.1.3 節討論的平滑係數決定方式所選用的平滑係數下，在不同疊代次數中的詞、字、音節、聲韻母的最高正確率，唯最小歧異度訓練因為沒有嘗試出有效的方法，故不在此列出。

各實驗結果的最高正確率如表 6.13、表 6.14 及圖 6.15～圖 6.18 所示，其中各方法名稱的代表如下：

方法代稱	方法	對應章節
baseline	基礎實驗	第 3 章
MPE	最小音素錯誤訓練	第 4 章
MPE+sel	MPE 加上詞弧篩選	第 6 章
MPFE	最小音素音框錯誤訓練	5.1 節
MPFE+pen+len	MPFE 加上錯誤處罰和音素長度正規化	5.1.2 節
MPFE+p+l+sel	MPFE+pen+len 加上詞弧篩選	第 6 章
sMBR	狀態層級最小貝氏風險訓練法	5.2 節
sMBR+pen	sMBR 加上錯誤處罰	5.2.2 節
sMBR+pen+len	sMBR+pen 加上音素長度正規化	5.2.3 節

以及各目標函數之主要差異處的詞弧正確度計算方式呈現如表 6.12。

由表中看來，MPFE 在兩種詞圖上都有最好的詞正確率；MPFE+p+l+sel 則是有最好的音節和聲韻母正確率；字正確率上則是由 MPFE+pen+len 以及 MPFE+p+l+sel 在詞圖 N 上有最好的正確率，在詞圖 T 上則仍是 MPE 有最好的字正確率。而由圖中看來，會發現在字、音節、聲韻母正確率上，各方法的優劣關係相當類似；然而詞正確率卻跟其它的正確率有相當不同的優劣關係，這顯示了詞正確率與字正確率在根本上並沒有一致的變動關係，因此在詞弧正確度計算函數的選擇上，也會出現偏好詞正確率和偏好字正確率的差異。

整體而言，在中文以字確率為目標的標準上，MPFE+p+l+sel、MPFE+pen+len 的表現較佳，其次是 MPE、MPE+sel，然後是 sMBR+pen+len、sMBR+pen，而 MPFE 和 sMBR 的成效則遠不如 MPE。

訓練法	詞弧正確度函數
MPE	$PhoneAcc(q)=\max_{z}\begin{cases}-1+2e(q,z) & \text{if } q=z\\ -1+e(q,z) & \text{if } q\neq z\end{cases}$
MPFE	$PhoneFrameAcc(q)=\sum_{t=start(q)}^{end(q)}\delta\left(q,s_{r,phone}(t)\right)$ $\delta\left(q,s_{r,phone}(t)\right)=\begin{cases}1 & \text{if } q=s_{r,phone}(t)\\ 0 & \text{if } q\neq s_{r,phone}(t)\end{cases}$
MPFE+pen +len	$PhoneFrameAcc(q)=\sum_{t=start(q)}^{end(q)}\frac{\delta\left(q,s_{r,phone}(t)\right)}{end(q)-start(q)+1}$ $\delta\left(q,s_{r,phone}(t)\right)=\begin{cases}1 & \text{if } q=s_{r,phone}(t)\\ -\rho & \text{if } q\neq s_{r,phone}(t)\end{cases}$
sMBR	$StateFrameAcc(q)=\sum_{t=start(q)}^{end(q)}\delta\left(q_{state}(t),s_{r,state}(t)\right)$ $\delta\left(q_{state}(t),s_{r,state}(t)\right)=\begin{cases}1 & \text{if } q_{state}(t)=s_{r,phone}(t)\\ 0 & \text{if } q_{state}(t)\neq s_{r,phone}(t)\end{cases}$
sMBR+pen	$StateFrameAcc(q)=\sum_{t=start(q)}^{end(q)}U\left(q_{state}(t),s_{r,state}(t)\right)$ $U\left(q_{state}(t),s_{r,state}(t)\right)=\begin{cases}1 & \text{if } q_{state}(t)=s_{r,phone}(t)\\ 0 & \text{if } q_{state}(t)\neq s_{r,phone}(t),\text{ but } q=s_{r,phone(t)}\\ -\rho & \text{if } q\neq s_{r,phone(t)}\end{cases}$
sMBR+pen +len	$StateFrameAcc(q)=\sum_{t=start(q)}^{end(q)}\frac{U\left(q_{state}(t),s_{r,state}(t)\right)}{len\left(s_{r,phone}(t)\right)}$ $U\left(q_{state}(t),s_{r,state}(t)\right)=\begin{cases}1 & \text{if } q_{state}(t)=s_{r,phone}(t)\\ 0 & \text{if } q_{state}(t)\neq s_{r,phone}(t),\text{ but } q=s_{r,phone(t)}\\ -\rho & \text{if } q\neq s_{r,phone(t)}\end{cases}$

表 6.12 各方法目標函數之詞弧正確度計算方法

詞圖 N	τ	itr	word	itr	char	itr	syl	itr	I/F
baseline			57.99		75.17		81.42		84.76
MPE	25	10	60.87	9	77.63	10	83.86	6	86.88
MPE+sel	25	8	60.92	8	77.62	8	83.92	8	86.94
MPFE	200	10	61.69	4	76.37	4	82.62	3	85.74
MPFE+pen+len	25	6	59.92	10	77.67	10	84.00	10	86.95
MPFE+p+l+sel	25	7	59.90	8	77.67	10	84.01	10	86.98
sMBR	130	5	61.00	5	76.78	5	83.02	5	86.09
sMBR+pen	150	7	60.86	7	77.40	7	83.66	5	86.66
sMBR+pen+len	16	7	60.66	6	77.42	7	83.72	7	86.72

表 6.13 詞圖 N－最高正確率

詞圖 T	τ	itr	word	itr	char	itr	syl	itr	I/F
baseline			57.99		75.17		81.42		84.76
MPE	25	10	61.03	7	77.48	10	83.71	10	86.74
MPE+sel	25	10	61.18	9	77.43	10	83.64	9	86.69
MPFE	200	10	61.95	3	76.18	3	82.46	3	85.61
MPFE+pen	25	8	59.54	8	77.36	8	83.75	8	86.81
MPFE+p+l+sel	25	10	59.48	9	77.45	10	83.86	6	86.84
sMBR	115	7	60.75	4	76.54	4	82.81	4	85.88
sMBR+pen	130	10	60.39	5	77.17	6	83.45	6	86.52
sMBR+pen+len	15	7	60.15	6	77.15	7	83.50	6	86.52

表 6.14 詞圖 T－最高正確率

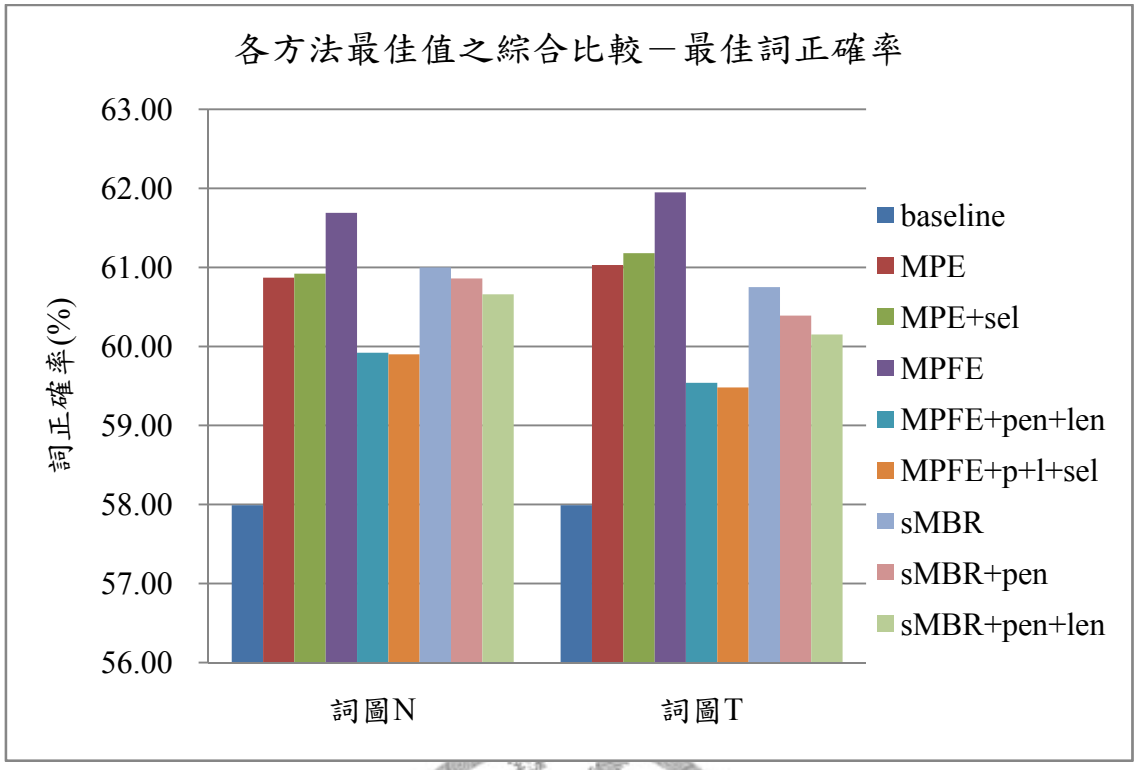


圖 6.15 各方法最佳值之綜合比較－最佳詞正確率

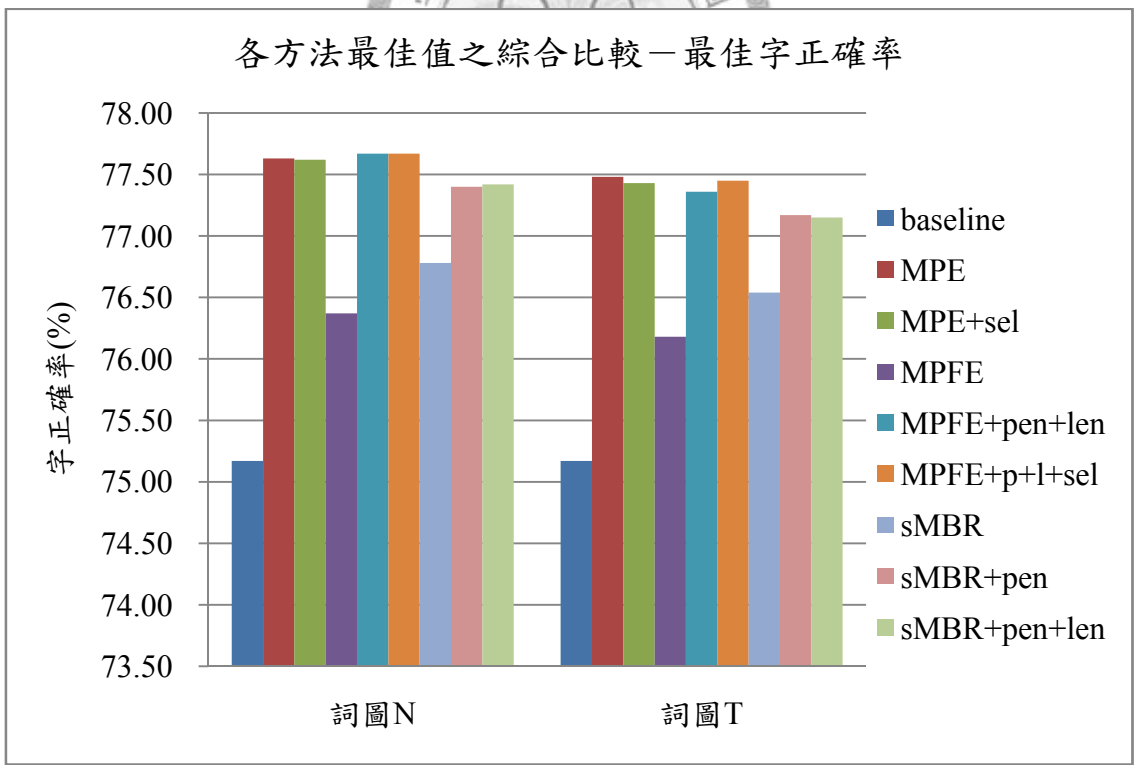


圖 6.16 各方法最佳值之綜合比較－最佳字正確率

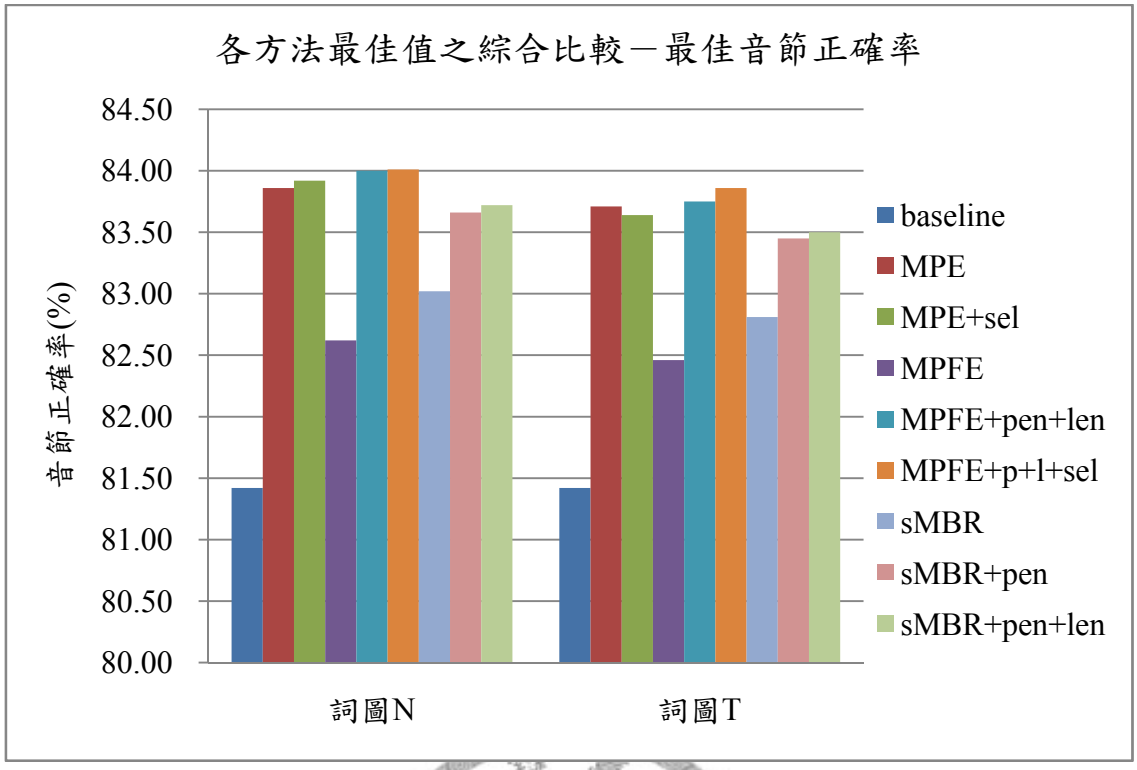


圖 6.17 各方法最佳值之綜合比較－最佳音節正確率

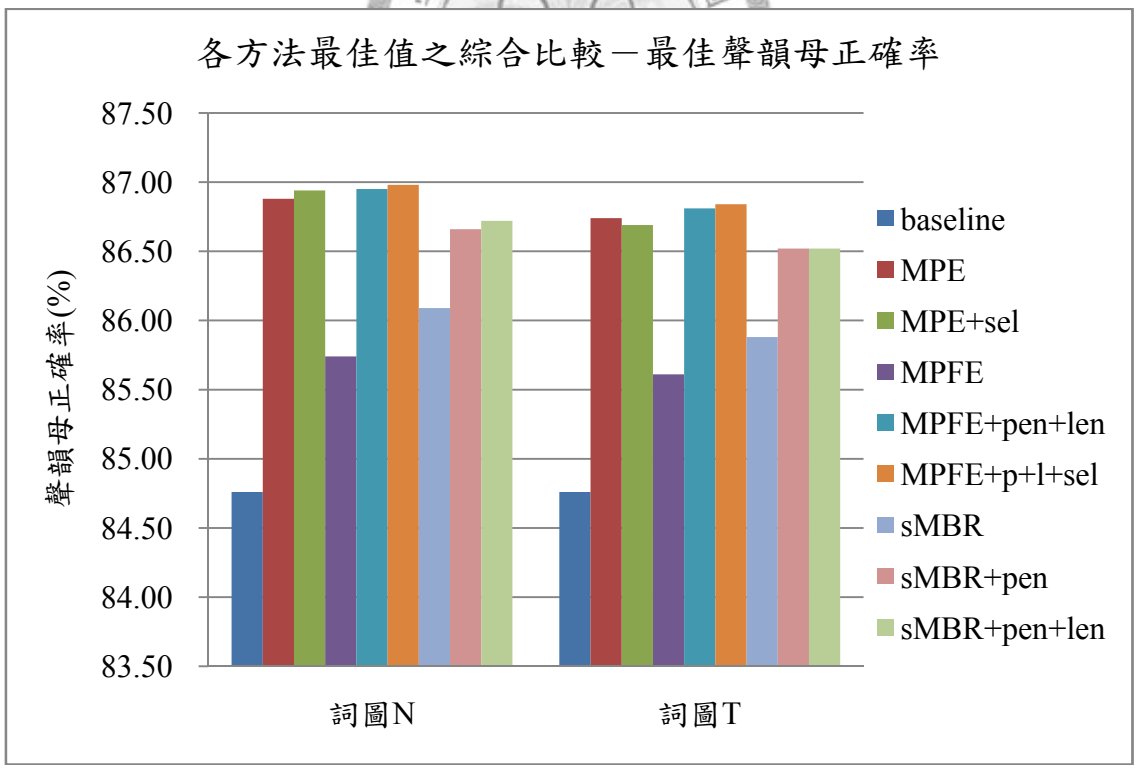


圖 6.18 各方法最佳值之綜合比較－最佳聲韻母正確率

第7章　結論與展望

7.1　總結

本論文深入探討了最小音素錯誤訓練法、最小音素音框錯誤訓練法、狀態層級最小貝氏風險訓練法，以及最小歧異度訓練法。在這四種方法之中，最小音素錯誤、最小音素音框錯誤以及狀態層級最小貝氏風險訓練法，都可以在詞正確率以及字正確率上進步，唯有最小歧異度訓練法，由於在實作上的細節設定及調整方式太複雜，所以沒有嘗試出會進步的設定值。而在有進步的三種方法之中，又以最小音素錯誤訓練法在字正確率的表現最好，而詞正確率則是以最小音素音框錯誤訓練法表現最好。

此外，本論文也在這些方法的詞弧正確度計算方式上做了更進一步的改進，有加入錯誤處罰以及音素長度正規化的方法，分別在最小音素音框錯誤訓練法上同時加入錯誤處罰與音素長度正規化，以及在狀態層級最小貝氏風險訓練法上加入錯誤處罰以及同時加入錯誤處罰與音素長度正規化，三種版本。這三種對詞弧正確度計算方式改進的版本，都會產生字正確率進步(相較於原本的正確度計算方式)，而在詞正確上退步的現象，由實驗結果可以推測在詞弧正確率加入錯誤處罰是產這個結果的主要原因。

由以上的結果可以推論，辨識結果的詞正確率與字正確率並沒有完全一致的變化，詞正確率的增加並不保證字正確率也會上升，反之亦然。因此在鑑別式訓練法中目標函數的最大化，目標函數的正確度定義也會有偏好詞正確率與字正確率的差異，偏好詞正確率的正確度定義方式並不能獲得字正確率的提升。這也解釋了何以相同的鑑別式訓練法在本論文的中文大字彙實驗結果與【32】中的英文大字彙實驗結果大相逕庭，由於英文是以詞正確率為評量標準，不同於中文是以字正確率為評量標準，因此正確度定義偏好詞正確率的目標函數，會在英文大字彙上有很好的表現。而同樣的方法實行在中文大字彙上，則會因為正確度定義偏好詞正確率，然而使詞正確率上升的同時卻無法使字正確率一併上升。因此將英

文大字彙上的鑑別式訓練法使用在中文大字彙上時，必須注意正確度的定義是否偏好詞正確率，並加以修改為偏好字正確率。在本論文中發現，在詞弧正確度的計算加入錯誤處罰可以將偏好詞正確率的詞弧正確度改變成偏好字正確率的詞弧正確度。

另外，本論文也在最小音素錯誤訓練和加入錯誤處罰和音素長度正規化的最小音素音框錯誤上實驗資料選取的方法，為基於詞弧期望正確度篩選詞弧的方法，實驗結果顯示最小音素錯誤訓練法會在詞正確率上進步，但在字正確率上退步；加入錯誤處罰和音素長度正規化的最小音素音框錯誤訓練則有恰好相反的結果，不過兩者正確率的變化都很小，所以基於詞弧期望正確度篩選詞弧的方法在正確率的效果上沒有一致的結論，不過在收斂速度上都有一定程度的提升。

7.2 未來展望

鑑別式訓練法主要的概念就是將錯誤資訊也加入訓練資料中，以期加強區分正確與錯誤資訊的能力。這種概念也可以套用在其它的方法上，如最大邊際估測法(Large Margin Estimation)【42】【43】，是將正確詞串與最可能的辨識詞串相似度的差值視為分離邊界，而將此邊界最大化的方法。而柔性邊際估測法，則是更進一步改進最大邊際估測法(Soft Margin Estimation)【44】，認為應該著重訓練分離邊界過小會辨識錯的資訊，因此只挑選分離邊界小於一定值的資料才納入訓練。而這種挑選容易辨識錯的資訊來訓練的概念，也有引入最大相互資訊訓練中，如增強式最大相互資訊法(Boosted MMI)【45】，是將近似所有辨識可能的詞圖，針對每個詞弧的正確度不同給與訓練不同的權重。這些以鑑別式訓練法的概念設計出來的訓練方式，都顯示出鑑別式訓練法是十分有效的方法，相信未來鑑別式訓練法的概念還可以推廣到更多不同的領域上。

附錄A　右相關聲韻母模型

起始音素	韻母									
empt	empt1 捲舌空韻母			empt2 不捲舌空韻母						
a	a ㄚ	ai ㄞ	au ㄠ	an ㄢ	ang ㄤ					
o	o ㄛ	ou ㄡ								
e	e ㄜ	en ㄣ	eng ㄥ	er ㄦ						
i	i ㄧ	ia ㄧㄚ	ie ㄧㄝ	iai ㄧㄞ	iau ㄧㄠ	ian ㄧㄢ	in ㄧㄣ	ing ㄧㄥ	iang ㄧㄤ	iou ㄧㄡ
u	u ㄨ	ua ㄨㄚ	uo ㄨㄛ	uai ㄨㄞ	uei ㄨㄟ	uan ㄨㄢ	uen ㄨㄣ	ueng ㄨㄥ	uang ㄨㄤ	
iu	iu ㄩ	iue ㄩㄝ	iuan ㄩㄢ	iun ㄩㄣ	iung ㄩㄥ					
E	ei ㄟ									

表 A.1　韻母模型列表

聲母		右相關聲母與對應的韻母起始音素							
		empt	a	o	e	i	u	iu	E
b	ㄅ		b_a	b_o	b_e	b_i	b_u		b_E
p	ㄆ		p_a	p_o	p_e	p_i	p_u		p_E
m	ㄇ		m_a	m_o	m_e	m_i	m_u		m_E
f	ㄈ		f_a	f_o	f_e		f_u		f_E
d	ㄉ		d_a	d_o	d_e	d_i	d_u		d_E
t	ㄊ		t_a	t_o	t_e	t_i	t_u		
n	ㄋ		n_a	n_o	n_e	n_i	n_u	n_iu	n_E
l	ㄌ		l_a	l_o	l_e	l_i	l_u	l_iu	l_E
g	ㄍ		g_a	g_o	g_e		g_u		g_E
k	ㄎ		k_a	k_o	k_e		k_u		
h	ㄏ		h_a	h_o	h_e		h_u		h_E
ji	ㄐ					ji_i		ji_iu	
chi	ㄑ					chi_i		chi_iu	
shi	ㄒ					shi_i		shi_iu	
j	ㄓ	j_empt	j_a	j_o	j_e		j_u		j_E
ch	ㄔ	ch_empt	ch_a	ch_o	ch_e		ch_u		
sh	ㄕ	sh_empt	sh_a	sh_o	sh_e		sh_u		sh_E
r	ㄖ	r_empt	r_a	r_o	r_e		r_u		
tz	ㄗ	tz_empt	tz_a	tz_o	tz_e		tz_u		tz_E
ts	ㄘ	ts_empt	ts_a	ts_o	ts_e		ts_u		
s	ㄙ	s_empt	s_a	s_o	s_e		s_u		
#	空聲母		#_a	#_o	#_e	#_i	#_u	#_iu	

表 A.2　右相關聲母模型列表

聲學模型	出現次數	高斯混合數	聲學模型	出現次數	高斯混合數	聲學模型	出現次數	高斯混合數
d_ee	0	1	l_iu	886	24	sic_e	2940	32
j_ee	0	1	ts_e	916	24	er	3032	32
n_o	0	1	t_o	947	24	f_u	3282	32
sh_ee	0	1	f_ee	951	24	l_e	3335	32
iai	1	1	p_a	952	24	g_a	3370	32
ts_o	3	1	m_u	1088	32	n_i	3404	32
tz_ee	3	1	tz_e	1231	32	ts_a	3420	32
r_o	52	8	ts_u	1382	32	s_u	3512	32
s_o	88	8	ts_empt	1469	32	uang	3561	32
h_ee	96	8	iung	1507	32	ji_iu	3725	32
l_o	105	16	m_e	1529	32	sh_a	3872	32
ch_o	137	16	n_a	1562	32	ch_a	3920	32
sic_o	147	16	o	1599	32	j_a	4195	32
p_u	162	16	s_a	1627	32	l_a	4384	32
f_o	220	24	m_a	1637	32	ch_e	4434	32
g_ee	252	24	sh_o	1679	32	ia	4507	32
p_o	285	24	iun	1731	32	t_u	4555	32
tz_o	301	24	h_o	1751	32	chi_iu	4578	32
p_e	308	24	s_empt	1767	32	h_a	4669	32
l_ee	320	24	n_e	1770	32	sh_u	4752	32
k_o	322	24	b_ee	1773	32	g_e	4940	32
d_o	348	24	uai	1793	32	ch_u	5152	32
j_o	354	24	r_a	1954	32	iue	5165	32
p_ee	360	24	sic_a	2098	32	sh_e	5201	32
n_iu	390	24	ua	2185	32	r_e	5247	32
s_e	402	24	p_i	2208	32	shi_iu	5506	32
t_e	467	24	f_e	2234	32	b_i	5556	32
b_o	489	24	r_u	2236	32	uen	5572	32
r_empt	502	24	k_u	2350	32	t_i	5705	32
n_u	533	24	m_ee	2358	32	tz_u	5732	32
ch_empt	660	24	l_u	2533	32	empt2	5773	32
g_o	721	24	tz_empt	2537	32	m_i	5794	32
b_e	850	24	k_a	2773	32	d_i	6028	32
m_o	857	24	k_e	2819	32	tz_a	6378	32
n_ee	869	24	h_e	2825	32	t_a	6644	32

聲學模型	出現次數	高斯混合數	聲學模型	出現次數	高斯混合數	聲學模型	出現次數	高斯混合數
ei	6982	32	iu	10908	64	uei	17459	64
j_empt	7186	32	h_u	10951	64	ueng	18729	64
ou	7211	32	iou	12048	64	sh_empt	19368	64
b_u	7223	32	l_i	12962	64	ian	20741	64
b_a	7365	32	a	13382	64	shi_i	21214	64
d_u	7871	32	an	13843	64	d_e	21862	64
iuan	8027	32	d_a	13921	64	ai	22455	64
f_a	8181	32	in	15082	64	sic_i	27125	64
iang	8451	32	uo	15155	64	empt1	27716	64
j_u	8528	32	eng	15317	64	ji_i	30712	64
uan	8529	32	ang	15340	64	u	32720	64
ie	9545	32	g_u	15404	64	i	39898	64
j_e	10287	64	au	15875	64	e	41752	64
chi_i	10314	64	en	15991	64	sil	64352	64
iau	10649	64	ing	16850	64			
sic_iu	10870	64	sic_u	16947	64			

表 A.3　聲韻母聲學模型在訓練語料的出現次數與狀態中的高斯混合數

其中在訓練語料中由於有四個聲韻母(d_ee、j_ee、n_o、sh_ee)未出現，因此分別以 d_ee → d_e、j_ee → j_e、n_o → n_u、sh_ee → sh_e 作為替代的模型，使用的替代模型為訓練過程中增加高斯混合數前只有一個混合數時的模型

附錄B　輔助函數(Auxiliary Function)

本附錄節錄自【28】。

當目標函數$F(\lambda)$(objective function)不易最佳化時，可以借助易處理的輔助函數$g(\lambda,\overline{\lambda})$，以藉著重覆對輔助函數進行最佳化，間接達成對目標函數的最佳化。由於輔助函數使用迭代方的式逼近最佳值，故此方法只能找到局部最佳值(local optimum)。在圖 B.1 中，$F(\lambda)$為進行最佳化的函數，假設初始的估測值為λ，則最佳化的演算法如下：

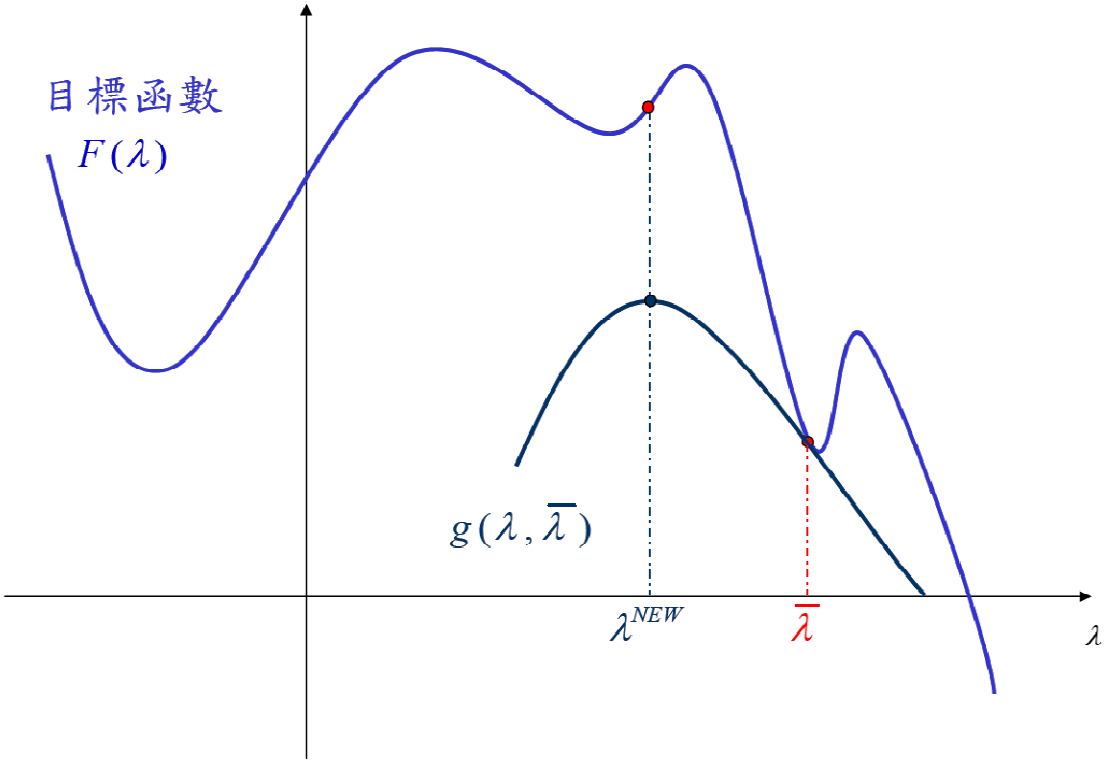


圖 B.1　輔助函數示意圖，橫軸代表λ值

Step 1.　在λ處先求出輔助函數$g(\lambda,\overline{\lambda})$，且在$\lambda$處與$F(\lambda)$相交

Step 2.　對$g(\lambda,\overline{\lambda})$求出最大值$\lambda^{NEW}$

Step 3.　將λ^{NEW}設為初始值λ

Step 4.　若滿足收斂條件則停止，否則回到 Step 1

可見利用輔助函數來進行最佳化，即是根據初始值找$g(\lambda,\overline{\lambda})$，然後求其最佳值$\lambda^{NEW}$，再以此點作為初始值，反覆這些步驟直至收斂為止。

輔助函數可分成兩類：強性輔助函數(Strong-Sense Auxiliary Function)與弱性輔

助函數(Weak-Sense Auxiliary Function)。

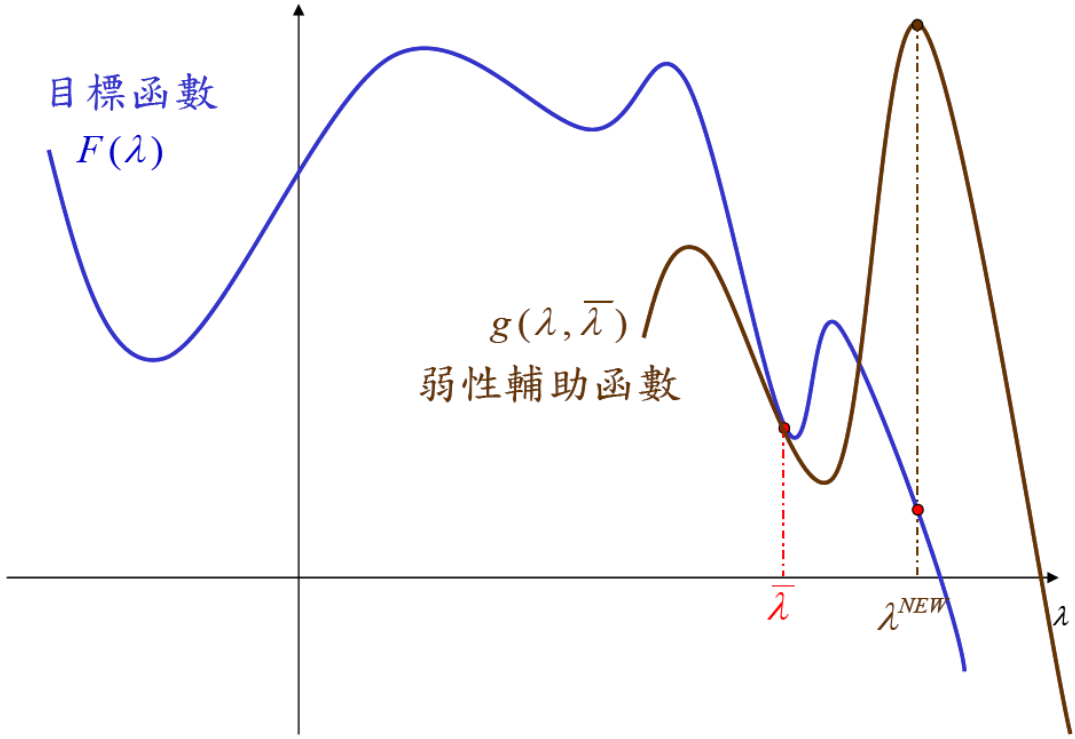


圖 B.2　弱性輔助函數示意圖，橫軸代表λ值

B.1　強性輔助函數(Strong-Sense Auxiliary Function)

強性輔助函數 $g\left(\lambda,\bar{\lambda}\right)$ 必須滿足

$$g\left(\lambda,\bar{\lambda}\right)-g\left(\bar{\lambda},\bar{\lambda}\right)\le F\left(\lambda\right)-F\left(\bar{\lambda}\right) \tag{B.1}$$

即 $g\left(\lambda,\bar{\lambda}\right)$ 為 $F\left(\lambda\right)$ 的下界(lower bound)，若更進一步要求 $F\left(\bar{\lambda}\right)=g\left(\bar{\lambda},\bar{\lambda}\right)$ 的話，則當增加 $g\left(\lambda,\bar{\lambda}\right)$ 時，也必定保證會增加 $F\left(\lambda\right)$。期望值最大化(Expectation Maximization)演算法即是使用強性輔助函數來最佳化的例子，本節將以最大化相似度法則訓練聲學模型為例，來說明期望值最大化演算法與強性輔助函數的關係。

以最大化相似度法則來訓練聲學模型可得到：

$$\lambda_{ML}=\max_{\hat{\lambda}}F\left(\hat{\lambda}\right)=\max_{\hat{\lambda}}\log P\left(O\mid\hat{\lambda}\right)=\max_{\hat{\lambda}}\log\sum_{S}P\left(O,S\mid\hat{\lambda}\right) \tag{B.2}$$

其中 O 為訓練語料，S 為 HMM 狀態序列(state sequence)。由於目標函數 $F\left(\lambda\right)$ 在

對數運算內包含了對所有狀態序列的加總，不易處理，故希望能藉由強性輔助函數 $g(\lambda,\bar{\lambda})$ 來最大化。此演算法可分為三個步驟：

1. 尋找下界：

由於強性輔助函數須為 $F(\lambda)$ 的下界，故在此引入詹氏不等式，尋找 $F(\lambda)$ 的下界：

$$\begin{aligned} F(\lambda) &= \log \sum_S P(O,S \mid \lambda) \\ &= \log \sum_S P(S) \frac{P(O,S \mid \lambda)}{P(S)} \\ &\geq \sum_S P(S) \log \frac{P(O,S \mid \lambda)}{P(S)} = g(\lambda, P(S)) \end{aligned} \tag{B.3}$$

其中 $P(S)$ 為任意 S 的機率分佈。

2. 取期望值(找最佳下界)：

為了使 $g(\lambda,P(S))$ 能在 λ 與 $F(\lambda)$ 相交，所以必須找出一組 $P(S)$ 使得 $g(\lambda,P(S))$ 在 $\lambda=\bar{\lambda}$ 處有最大值，即：

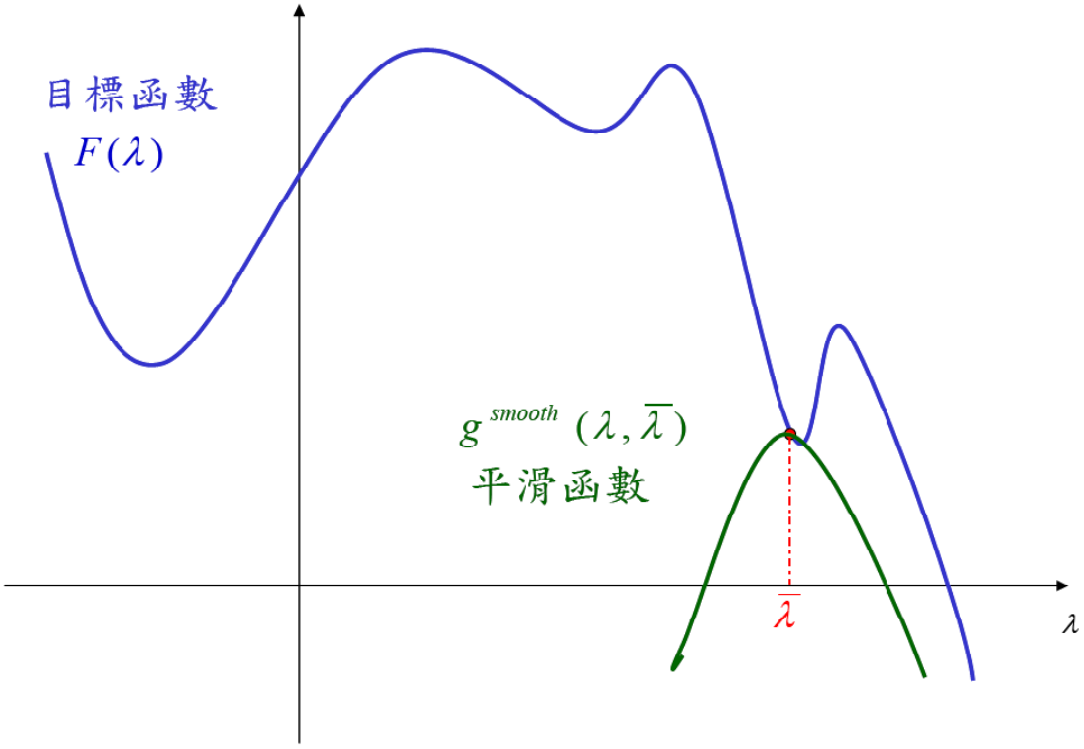


圖 B.3 平滑函數示意圖，橫軸代表λ值

$$P^{\lambda}(S)=\arg\max_{\hat{P}(S)} g\left(\bar{\lambda},\hat{P}(S)\right)=\arg\max_{\hat{P}(S)}\sum_{S}\hat{P}(S)\log\frac{P\left(O,S\mid\bar{\lambda}\right)}{\hat{P}(S)} \tag{B.4}$$

其中為確保$\sum_S \hat{P}(S)=1$，在此引入拉格朗日乘數 η (Lagrange Multiplier)，可以得到：

$$g\left(\bar{\lambda},P(S)\right)=\eta\left[1-\sum_{S}P(S)\right]+\sum_{S}P(S)\log P\left(O,S\mid\bar{\lambda}\right)-\sum_{S}P(S)\log P(S) \tag{B.5}$$

將$g\left(\bar{\lambda},P(S)\right)$對$P(S)$作微分並使之等於 0 可得：

$$\begin{aligned}
&-\eta+\log P\left(O,S\mid\bar{\lambda}\right)-\log P(S)-1=0\\
&\Rightarrow\log P(S)=-\eta+\log P\left(O,S\mid\bar{\lambda}\right)-1\\
&\Rightarrow P(S)=\frac{e^{-\eta}P\left(O,S\mid\bar{\lambda}\right)}{e}\Rightarrow\sum_{S}P(S)=e^{-\eta}\frac{\sum_{S}P\left(O,S\mid\bar{\lambda}\right)}{e}=1\\
&\Rightarrow e^{-\eta}=\frac{e}{\sum_{S}P\left(O,S\mid\bar{\lambda}\right)}\Rightarrow P(S)=\frac{P\left(O,S\mid\bar{\lambda}\right)}{\sum_{S}P\left(O,S\mid\bar{\lambda}\right)}=P\left(S\mid O,\bar{\lambda}\right)
\end{aligned} \tag{B.6}$$

由於$P(S)=P\left(S\mid O,\bar{\lambda}\right)$只跟$\bar{\lambda}$有關，故將$g\left(\lambda,P\left(S\mid O,\bar{\lambda}\right)\right)$記作$g\left(\lambda,\bar{\lambda}\right)$，所以

$$\begin{aligned}
g\left(\lambda,\bar{\lambda}\right)&=\sum_{S}P\left(S\mid O,\bar{\lambda}\right)\log\frac{P\left(O,S\mid\lambda\right)}{P\left(S\mid O,\bar{\lambda}\right)}\\
&=\sum_{S}P\left(S\mid O,\bar{\lambda}\right)\log P\left(O,S\mid\lambda\right)-\sum_{S}P\left(S\mid O,\bar{\lambda}\right)\log P\left(S\mid O,\bar{\lambda}\right)
\end{aligned} \tag{B.7}$$

此函數即為$F(\lambda)$的最佳下界。由於$\sum_S P\left(S\mid O,\bar{\lambda}\right)\log P\left(O,S\mid\lambda\right)$與待測模型的參數$\lambda$無關，所以$g\left(\lambda,\bar{\lambda}\right)$可以只取一部份

$$g\left(\lambda,\bar{\lambda}\right)=\sum_{S}P\left(S\mid O,\bar{\lambda}\right)\log P\left(O,S\mid\lambda\right) \tag{B.8}$$

做為輔助函數。

3. 最大化：

對$g\left(\lambda,\bar{\lambda}\right)$微分並使其等於 0，求得新的聲學模型參數。再以此參數重覆步驟 2、3，直至收斂為止。

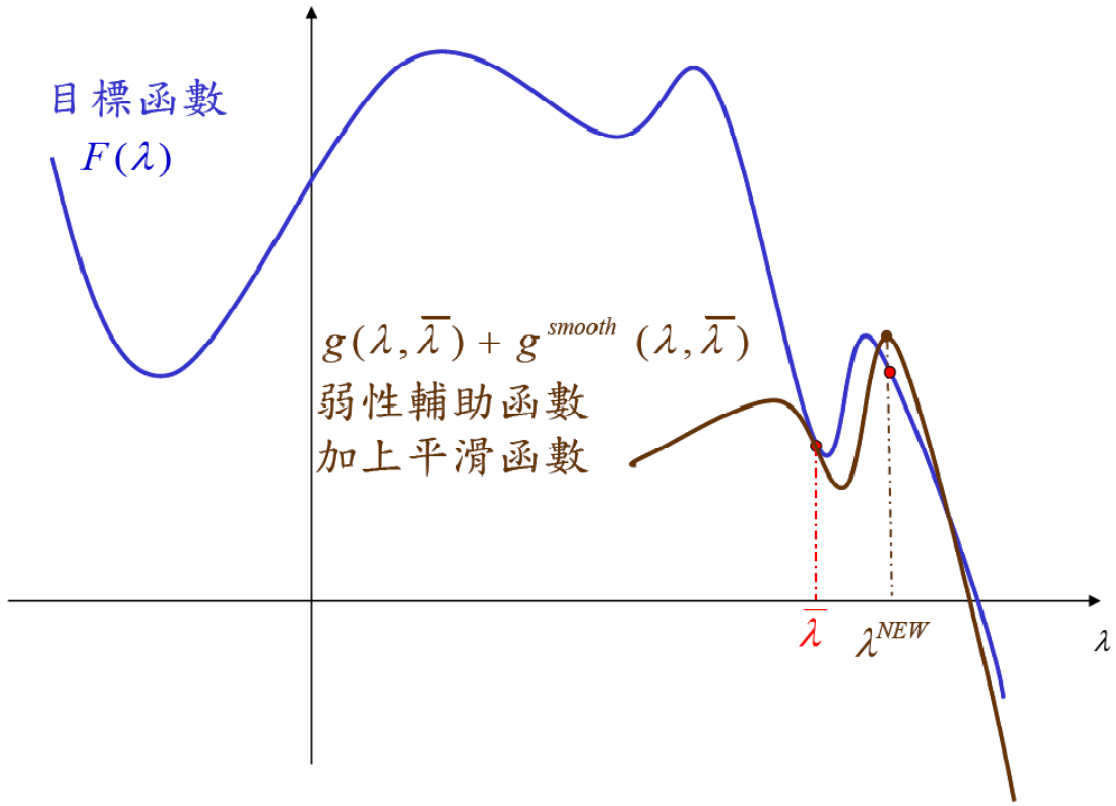


圖 B.4　弱性輔助函數加上平滑函數之示意圖，橫軸代表λ值

B.2 弱性輔助函數(Weak-Sense Auxiliary Function)

弱性輔助函數 $g(\lambda,\bar{\lambda})$ 只需滿足

$$\left.\frac{\partial}{\partial\lambda}g(\lambda,\bar{\lambda})\right|_{\lambda=\bar{\lambda}}=\left.\frac{\partial}{\partial\lambda}F(\lambda)\right|_{\lambda=\bar{\lambda}} \tag{B.9}$$

即要求 $g(\lambda,\bar{\lambda})$ 與 $F(\lambda)$ 在 $\lambda=\bar{\lambda}$ 有相同的斜率即可，如圖 B.2，由於增加 $g(\lambda,\bar{\lambda})$，不保證會增加 $F(\lambda)$，故弱性輔助函數不適合做為最佳化的輔助函數。但在鑑別式訓練中，由於強性函數不易求得，是故除使用弱性輔助函數之外也別無他法。由於弱性輔助函數不保證收斂，在使用上必須搭配平滑函數來加速其收斂速度。若 $g^{smooth}(\lambda,\bar{\lambda})$ 在 $\lambda=\bar{\lambda}$ 有全域最大值(global maximum)，則 $g^{smooth}(\lambda,\bar{\lambda})$ 可做為一平滑函數，如圖 B.3。雖然 $h(\lambda,\bar{\lambda})=g(\lambda,\bar{\lambda})+g^{smooth}(\lambda,\bar{\lambda})$ 仍為一弱性輔助函數，如圖 B.4，但加入平滑函數可加速其收斂速度。

參考文獻

【1】 L. Bahl, P. Brown, P de Souza, R. Merce, “Maximum Mutual Information Estimation Of Hidden Markov Model Parameters For Speech Recognition,” *Proc. ICASSP,* 1986.

【2】 B.-H. Juang, W. Chou, C.-H Lee, “Minimum Classification Error Rate Methods For Speech Recognition,” *IEEE Transactions on Speech and Audio Processing*, 1997.

【3】 D. Povey, P.C. Woodland, “Minimum Phone Error And I-smoothing For Improved Discriminative Training,” *Proc. ICASSP,* 2002.

【4】 J. Zheng, A. Stolcke, “Improved Discriminative Training Using Phone Lattices,” *Interspeech*, 2005.

【5】 J. Du, P. Liu, F. K. Soong, J.-L. Zhou, R.-H. Wang, “Minimum Divergence Based Discriminative Training,” *Interspeech,* 2006.

【6】 L.R. Rabiner, “A tutorial on Hidden Markov Models and selected applications in speech recognition,” *Proc. IEEE,* Vol.**77,** No.2, pp.257–285, 1989

【7】 L.R. Bahl, F. Jelinek, R. L. Mercer, “A Maximum Likelihood Approach to Continuous Speech Recongnition,” *IEEE Trans. Pattern Analysis and Machine Intelligence,* Vol. PAMI-5, No.2, pp.179–190, 1983

【8】 V. Goel, S. Kumar, W. Byrne, “Segmental Minimum Bayes-Risk Decoding for Automatic Speech Recognition,” *IEEE Trans. Speech and Audio Processing*, 2004

【9】 L. Mangu, E. Brill, A. Stolcke, “Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Nerworks,” *Computer Speech and Language*, 2004

【10】 P. F. Brown “The Acoustic-Modeling Problem in Automatic Speech Recognition,” *Ph.D Dissertation, Carnegie Mellon University, Pittsburg*, 1987.

【11】 P. S. Gopalakrishnan, D. Kanevsky, A. Nádas & D. Nahamoo “An Inequality for Rational Functions with Applications to Some Statistical Estimation Problems,” *IEEE Trans. Information Theory,* Vol. 37, pp.107-113, 1991.

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

【13】 V. Valtchev, J. J. Odell, P. C. Woodland, S. J. Young. (1997). “MMIE Training of Large Vocabulary Recognition Systems,” *Speech Communication,* Vol. 22, No. 4, pp.303-314, September 1997.

【14】 P. C. Woodland and D. Povey (2002). “Large Scale Discriminative Training of Hidden Markov Models for Speech Recognition,” *Computer Speech and Language,* Vol. 16, pp.25-47, 2002.

【15】 Wikipedia, Levenshtein distance, http://en.wikipedia.org/wiki/Levenshtein_distance

【16】 J. Kaiser, B. Horvat, Z. Kacic “Overall Risk Criterion Estimation of Hidden Markov Model Parameters,” *Speech Communication,* Vol. 38, pp.383-398, 2002.

【17】 B.-H. Juang and S. Katagiri “Discriminative Learning for Minimum Error Classification,” *IEEE Trans. Signal Processing,* Vol. 40, No. 12, pp. 3043-3054, 1992.

【18】 W. Chou, C.-H. Lee, B.-H. Juang (1993). “Minimum Error Rate Training based on N-Best String Models,” *Proc. ICASSP,* 1993.

【19】 L. K. Saul and M. G. Rahim “Maximum Likelihood and Minimum Classification Error Factor Analysis for Automatic Speech Recognition,” *IEEE Trans. Acoustics, Speech and Signal Processing,* Vol. 8, No. 2, pp. 115-125, March 2000.

【20】 R. Schlüter. “Investigations on Discriminative Training Criteria,” *Ph.D Dissertation, RWTH Aachen University of Technology,* 2000.

【21】 H.-M. Wang, B. Chen, J.-W. Kuo, and S.-S Cheng “MATBN: A Mandarin Chinese Broadcast News Corpus,” *Interational Journal of Computational Linguistics and Chinese Language Processing*, 2005

【22】 Cambridge University Engineering Dept. (CUED), Machine Intelligence Laboratory, “HTK,” http://htk.eng.cam.ac.uk/

【23】 SRI Speech Technology and Research Laboratory, “SRILM,” http://www.speech.sri.com/projects/srilm/

【24】 潘奕誠，『大字彙中文連續語音辨認之一段式及以詞圖為基礎之搜尋演算法』，*碩士論文，國立台灣大學資訊工程研究所*，2002

【25】 X. Huang, A. Acero, H.-W. Hon, “Spoken Language Processing,” *Pearson Education Taiwan Ltd.*, pp. 424-426, 2005

【26】 S. Furui, “Cepstral analysis technique for automatic speaker verification,” *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol.29, No.2, pp. 254-272, 1981.

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

【28】 郭人瑋，『最小化音素錯誤鑑別式聲學模型學習於中文大詞彙連續語音辨識之初步研究』，*碩士論文，國立台灣師範大學資訊工程研究所*，2005

【29】 陳佳妤，『最小化音素錯誤模型及特徵訓練法於中文大詞彙辨識上之初步研究』，*碩士論文，國立台灣大學電信工程研究所*，2006

【30】 D. Povey. “Discriminative Training for Large Vocabulary Speech Recognition,” *Ph.D Dissertation, Peterhouse, University of Cambridge*, 2004.

【31】 X. L. Aubert “An Overview of Decoding Techniques for Large Vocabulary Continuous Speech Recognition, “ *Computer Speech and Language,* 2002

【32】 D. Povey, B. Kingsbury, “Evaluation of Proposed Modifications to MPE for Large Scale Discriminative Training,” *Proc. ICASSP,* 2007.

【33】 S.-H. Liu, F.-H. Chu, S.-H. Lin, B. Chen, “Investigation Data Selection for Minimum Phone Error Training of Acoustic Models,” *Proc. ICME*, 2007.

【34】 M. Gibson, T. Hain, “Hypothesis Spaces for Minimum Bayes Risk Training in Large Vocabulary Speech Recognition,” *Interspeech*, 2006.

【35】 Wikipedia, Multivariate normal distribution, http://en.wikipedia.org/wiki/Multivariate_normal_distribution#Kullback-Leibler_divergence

【36】 蔡明怡，『國語語音之發音變異分析及提昇辨識效能之發音模型』，*博士論文，國立台灣大學電信工程研究所*，2006

【37】 X. Li, H. Jiang, C. Liu, “Larginn Margin HMMs for Speech Recognition,” *ICASSP,* 2005

【38】 Wikipedia, Support vector machine, http://en.wikipedia.org/wiki/Support_vector_machine

【39】 DTREG, Software For Predictive Modeling and Forecasting, SVM - Support Vector Machines, http://www.dtreg.com/svm.htm

【40】 S.-H. Liu, F.-H. Chu, S.-H. Lin, H.-S. Lee, B. Chen, “Training Data Selection for Improving Discriminative Training of Acoustic Models,” *ASRU*, 2007.

【41】 朱芳輝，『資料選取方法於鑑別式聲學模型訓練之研究』，*碩士論文，國立台灣師範大學資訊工程研究所*，2008

【42】 H. Jiang, X. Li, C. Liu, “Large Margin Hidden Markov Models for Speech Recognition,” *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol.14, No.5, pp. 1584-1595, 2006.

【43】 F. Sha, L. K. Saul, “Comparison of Large Margin Training to Other Discriminative Methods for Phonetic Recognition by Hidden Markov Models,” *ICASSP,* 2007

【44】 J. Li, M. Yuan, C.-H. Lee, “Soft Margin Estimation of Hidden Markov Model Parameters,” *Interspeech*, 2006.

【45】 D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Soan, K. Visweswariah, “Boosted MMI for Model and Feature- Space Discriminative Training,” *ICASSP,* 2008