國立臺灣師範大學

資訊工程研究所碩士論文

指導教授：王新民博士

陳柏琳博士

信心度評估於中文大詞彙連續語音辨識之研究

Exploring the Use of Confidence Measures for Mandarin Large Vocabulary Continuous Speech Recognition

研究生：陳燦輝 撰

中華民國 九十五年 七月

摘要

本論文初步地探討信心度評估(Confidence Measures)於中文大詞彙連續語音辨識上之研究。除了討論原本一般信心度評估應用於判斷語音辨識結果(例如候選詞)是否正確之外，也嘗試將信心度評估應用在詞圖搜尋(Word Graph Rescoring)或*N*-最佳詞序列(*N*-best List)重新排序(Reranking)的研究。而實驗語料則是使用公視新聞語料庫(MATBN)中的外場記者(Field Reporters)跟受訪者(Interviewees)語句，以分別探討信心度評估在偏朗讀語料(Read Speech)或偏即性口語(Spontaneous Speech)等兩種不同性質的語句上是否能有不同的效能。首先，本論文嘗試使用熵值(Entropy)資訊並結合以事後機率為基礎之信心度評估方法，在MATBN外場記者(Read Speech)及外場受訪者(Spontaneous Speech)測試語料所得到的最佳實驗結果，可較傳統僅使用以事後機率為基礎之信心度評估可以分別有16.37%及12.00%的信心度錯誤率相對減少(Relative Reduction)。另一方面，在以最小化音框錯誤率(Time Frame Error)搜尋法來增進詞圖搜尋的正確率之實驗中，本論文嘗試結合以梅爾倒頻譜係數(Mel-frequency Cepstral Coefficients, MFCC)，以及以異質性線性鑑別分析(Heteroscedastic Linear Discriminant Analysis, HLDA)搭配最大相似度線性轉換(Maximum Likelihood Linear Transformation, MLLT)兩種不同語音特徵參數所形成的詞圖資訊，並以最小化音框錯誤率搜尋法來降低語音辨識系統的字錯誤率，經由實驗顯示在外場記者測試語料能有4.6%的字錯誤率相對減少，而在外場受訪者測試語料的部份則有4.8%的字錯誤率相對減少，相較於僅使用異質性線性鑑別分析及最大相似度線性轉換求得語音特徵參數的詞圖並配合最小化音框錯誤率法有較佳的結果。最後，本論文嘗試在傳統以Levenshtein距離為成本函式(Cost Function)的最小化貝氏風險(Minimum Bayes Risk)辨識法則中，適當的加入以特徵為基礎的信心度評估。雖然經由實驗得知，在外場記者以及外場受訪者的語料中，對於辨識錯誤率並沒有很明顯的進步或退步，但相

較於傳統利用Levenshtein距離為成本函式的最小化貝氏風險辨識法則而言，卻有較佳的結果。

Abstract

This thesis investigated the use of various kinds of confidence measures for Mandarin large vocabulary continuous speech recognition (LVCSR). These confidence measures were not only used as a post processor to justify the correctness of final recognition hypotheses, but also directly integrated into the word graph rescoring and *N*-best list reranking procedures for the generation of better recognition hypotheses. All experiments were carried out on the Mandarin broadcast news corpus (MATBN), including the speech utterances of field reporters and interviewees which also respectively belong to the read speech style and the spontanesous speech one. Several approaches to utilizing confidence measures for Mandarin LVCSR were presented and extensively studied in this thesis. First, the entropy information and the posterior probability based confidence measure were tightly combined, and the experimental results showed that such an approach could give relative confidence error rate reductions of 16.37% and 12.00%, respectively, for the field reporters' speech and the interviewees' speech, compared to those obtained by using the posterior probability based confidence measure alone. On the other hand, we attempted to jointly consider the information inherent in the word graph constructed by using the Mel-frequency cepstral coefficients (MFCC), and the word graph constructed by using the discriminant acoustic features resulting form the heteroscedastic linear discriminant analysis and maximum likelihood linear transformation (HLDA+MLLT). The minimum time frame error decoding was conducted on these two word graphs simultanesously to find the best word sequence among them. The experimental results showed that such an approach could achieve character error rate reductions of 4.6% and 4.8%, respectively, for the field reporters' speech and the interviewees' speech, which were better than the results obtained by conducting the minimum time frame error decoding on the word graph of

HLDA+MLLT alone. Finally, we incorporated the feature-based confidence measure with the minimum Bayes risk decoding. Compared to the conventional minimum Bayes risk decoding, the proposed approach demonstrates slight but consistent performance gains.

致謝

首先，最要感謝的就是我的父母。是他們辛苦地將我養大，教我學會做人做事的態度，也一直勉勵我努力唸書。也要感謝我的奶奶以及伯父、姊姊和弟弟，從小到大他們一直陪在我身邊，在背後一直默默地支持我。因為有這些親愛的家人陪伴與鼓勵，我才能順利完成學業。

感謝我的兩位指導教授-王新民博士及陳柏琳博士。兩位教授在我碩士求學過程中，其認真的研究態度以及對學生的敦敦教誨，一直深刻地記在我的腦海中。覺得自己非常的幸運，能獲得兩位如此認真及和善的好老師之教導。在此再次謝謝兩位老師，您們辛苦了。此外，也感謝口試委員林順喜博士及洪志偉博士在口試時的指導，讓學生的論文能更臻完善。

也要謝謝實驗室已畢業的學長們。謝謝人瑋，不論在中研院或是師大，你總是很樂意地灌輸你的知識，在你的身邊，我學到了很多，也謝謝你一直陪我打桌球。謝謝文鴻、耀民、志豪、成韋及惠銘，不論你們畢業與否，還是一直很關心我們，對於我們課業或是生活的疑問，總是很熱心地幫我解答。也謝謝士翔、士弘、炫盛及怡婷，在一起學習的這二年，我們一起渡過了許多寫作業、哈啦聊天、打球以及互相討論的日子，與你們在一起的生活很快樂。我想我永遠會記得這段的時光。芳輝、鴻彬、家豪及庭瑋，因為有你們的加入，讓實驗室的氣氛變得更加活潑，也祝你們未來的研究順利。也要特別感謝斯涵，謝謝妳跟人瑋的"榮華富貴"車票，妳以後也要加油囉。

謝謝偉和、中研院實驗室的士賢、怡翔、弘明、永承、宏毅及御仁學長，由你們身上我學到了許多做研究的態度。不論我有什麼困難，你們也都很樂意幫忙。

最後要感謝思吟。謝謝妳陪我走過這段時間，在我最累的這段時間，因為有妳的支持與鼓勵，才能讓我一直走下去，也祝妳明年順利畢業囉。

謹將此論文獻給所有曾經幫助我的人。

燦輝謹誌

目錄

圖目錄

表目錄

第1章 序論

「霹靂遊俠」是70年代一部非常熱門的電視影集。其中最令大家印象深刻的是主角李麥克的伙計-霹靂車，它不但擁有極高的人工智慧，並且還能直接接受主人的語音指令。的確，語音是人類最自然的溝通方式之一，人們也一直希望電腦能聽得懂人類的語言。過去由於電腦軟硬體技術的限制，加上對個人電腦使用者來說，利用鍵盤及滑鼠來當作輸入裝置也不會有太多不便，因此，電腦距離達到許多科幻電影或小說裡面的語音辨識功能，似乎相當遙遠。

如今隨著科技的發展，電腦技術的發展也跟著突飛猛進，消費性電子產品及許多可攜性裝置跟著融入了我們的生活中。為了節省體積及方便使用者攜帶，電子產品的體積越來越小是一定的趨勢。但鍵盤及滑鼠卻不可能跟著變小，否則會造成輸入不便，也不可能隨身攜帶它們，否則便失去了可攜性的便利。因此，過去利用鍵盤或滑鼠的輸入方式在可攜性裝置已經不太可行。許多研究學者便開始試著用語音來當作輸入裝置的一部份。過去數十年來，由於眾多學者專家的努力以及隱藏式馬可夫模型(Hidden Markov Model, HMM)[Rabiner 1989]的發展，自動語音辨識(Automatic Speech Recognition, ASR)的技術也獲得了很大的提昇。在今天，只要提供足夠的訓練資料，再利用一些標準的程序(如語音特徵參數擷取、聲學與語言模型訓練及建立辨識器等等)，便可以為某個特定的領域建立一套語音辨識系統，達到人類與電腦對話的初步夢想。但是，當我們將系統從實驗室環境移到真實世界應用時，由於環境的噪音、語者之間的差異以及通道效應等在真實環境才會遇到的問題，或是訓練資料與測試資料環境不匹配的情況，即使是目前最好的語音辨識系統，其效能依舊會大幅地下降。因此，要如何使語音辨識系統在真實環境使用，仍然有最好的效果，便成為了語音辨識研究裡一項極為重要的課題。此外，由於目前的語音辨識系統依然無法達到百分之百的正確，要如何讓系統能自動的判斷辨識結果的可靠性，對許多自動語音辨識相關應用，例如口語對話系統(Spoken Dialog System)[Hazen *et al*. 2002]、關鍵詞擷取(Keyword Spotting)[Wilpon *et al*. 1990]等，也很重要。除了

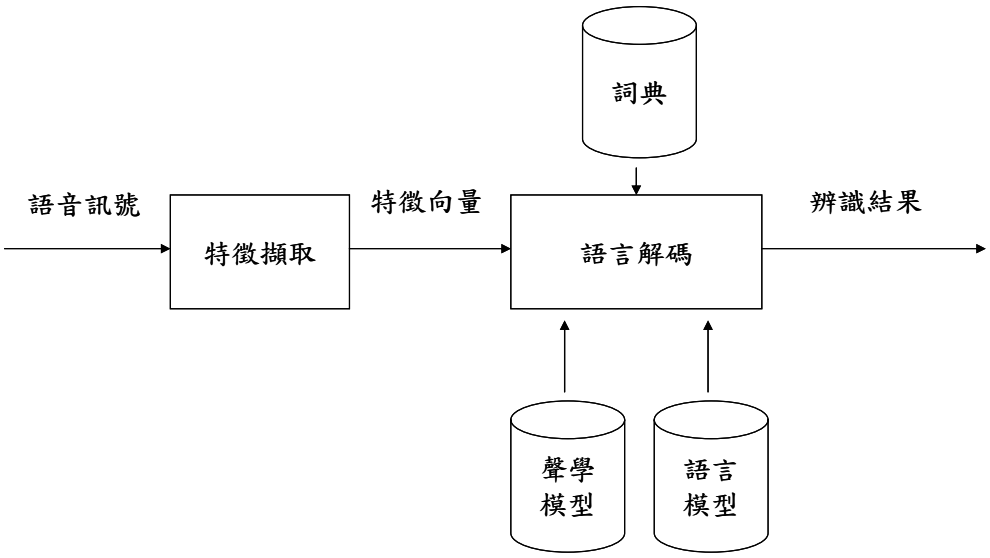

圖 1-1 語音辨識流程圖

利用信心度評估(Confidence Measures)找出辨識錯誤，並進一步決定是否要請使用者重新輸入外。更甚者，可以利用信心度評估修正辨識結果，提高系統的辨識正確率。而要如何達到上述之目標，也是目前亟待研究的方向。

1.1 語音辨識之流程

一般我們是以統計式的方法來實行自語音辨識[Jelinek 1999]。其數學式子可以寫成

$$W^* = \arg\max_{W} P(W \mid X) \tag{1-1}$$

其中 W 為一個語言 $\overline{\mathbf{W}}_{\Sigma}$ 中可能形成的任一詞序列(Word Sequence)，而 X 則代表輸入的語音訊號。統計式的語音辨識其方法很直覺，它希望在給定一段語音訊號後，從許多可能的詞序列中，找出一條詞序列，使得其在 X 出現時的機率是最大的。以人類的聽覺來說，就好像是要我們的大腦找出一條聽起來最有可能的詞序列。目前語音辨識系統的流程大致如圖 1-1所示，當一段語音進來時，語音辨識系統會先將語音訊號做特徵擷取的動作，接著針對所形成的語音特徵向量序列(Feature Vector Sequence)找出一條最符合的詞序列。而在這個步驟中，通常會將式(1-1)利用貝式定理展開

$$P(W \mid X) = \frac{p(X \mid W)P(W)}{p(X)} \tag{1-2}$$

其中，$p(X \mid W)$及$P(W)$便是分別代表經由圖 1-1中的聲學模型(Acoustic Model)及語言模型(Language Model)產生的分數。而這兩個模型一般是經由統計求得，也就是分別假設一組機率模型，利用訓練語料來估測其機率分佈(Probability Distribution)。以下分別簡介圖 1-1的每個模組運作方式。

1.1.1 特徵擷取(Feature Extraction)

這個部份主要是擷取語音訊號中比較重要的參數，一般較常用的語音特徵參數為梅爾倒頻譜係數(Mel-frequency Cepstral Coefficients, MFCC)[Davis and Mermelstein 1980]，其擷取步驟如圖 1-2所示。在取此特徵的時候，我們會將語音資料切割成一連串部份重疊的音框(Frames)，每一個音框最後表示成由13維的梅爾倒頻譜係數加上其一階與二階的時間軸導數(Time Derivatives)所組成的特徵向量。其中取一階與二階時間軸導數的原因主要是為了能獲得語音特徵在時間上(Temporal)的相關資訊。

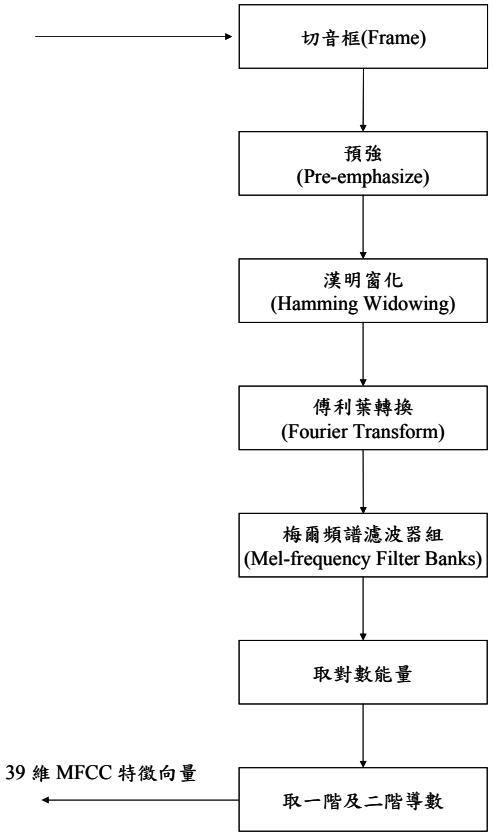

圖 1-2 梅爾倒頻譜係數特徵擷取步驟

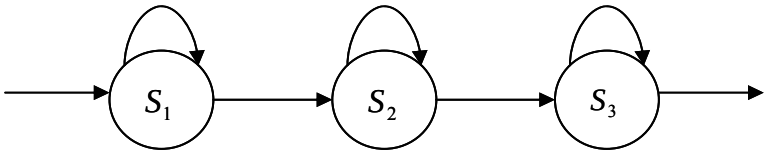

圖 1-3 隱藏式馬可夫模型範例

1.1.2 聲學模型(Acoustic Model)

為了處理語音訊號在時域上的變化，一般而言都是使用由左至右(Left-to-right)的隱藏式馬可夫模型(Hidden Markov Model, HMM)[Rabiner 1989]來作為聲學模型。圖 1-3 便是一個具有三個狀態(State)的HMM模型，每個狀態中都會有對每個音框所形成的語音特徵參數向量之觀測機率分佈(Observation Probability Distribution)。另外，每個狀態也有相對應的狀態轉移機率(State Transition Probability)，用來控制下一個時間點要停留在自己或是轉移到下一個狀態。根據語音特徵參數是連續或非連續的值，HMM每個狀態中的觀測機率估測方式可分為離散型(Discrete)、半連續型(Semi-continuous)及連續型(Continuous)三種[Huang *et al*. 2001]，目前的語音辨識系統主要都是連續型或半連續型為主。就連續型而言，為了減少估算觀測機率的參數量，以及因為任何機率分佈理論上皆可以由多個高斯分佈(Gaussian Distributions)來逼近的特性，一般而言都是使用高斯混合分佈(Gaussian Mixture Distribution)來近似此機率分佈。而連續型與半連續型主要的差別在於連續型每個狀態擁有自己的高斯分佈，半連續型則會有共用高斯分佈的情況。

由於一個中文音節(Syllable)是由一個聲母(INITIAL)及一個韻母(FINAL)組成，22 個聲母及 38 個韻母構成約 400 個音節。基本上，我們只要為每個聲母及韻母建立屬於它的聲學模型，便可以辨識所有的中文音節。本論文共使用了 38 個韻母模型，但在聲母模型的部份則是考慮不同種類的右邊相連韻母對其發音特性會造成不同的影響，而將 22 種聲母再細分成 112 種聲母模型，亦即聲母部份採用右相關聯模型(Right-context-dependent Model, RCD);另外，我們加入一個靜音(Silence)模型來估測語音訊號中靜音部份。為了讓聲學模型有更精確的估算能力，除了要有足夠的訓

練資料之外，還需要有好的訓練方法，較常見的有最大化相似度訓練法則[Bahl *et al.* 1983] 配合使用波氏重估(Baum-Welch Re-estimation)演算法(又稱前向-後向演算法，Forward-backward Algorithm)[Baum 1972]、最大化交互資訊(Maximum Mutual Information, MMI)[Bahl *et al.* 1986]、最小化分類錯誤(Minimum Classification Error, MCE)[Juang and Katagiri 1992]以及新近被提出的最小化音素錯誤(Minimum Phone Error, MPE)[Povey 2004]訓練法則等。

1.1.3 語言模型(Language Model)

由於聲學模型本身只能辨識某一段語音訊號發的是何種音節(Syllable)序列，無法確認其對應的詞(中文有許多同音詞)，且句子中詞跟詞的連接其實存在概略的規則，因此便需要有語言模型的存在。由於語言模型的機率分佈是離散型的，在估計語言模型的機率時，並不使用機率密度分佈函數，而是直接估測個別詞序列的機率資量函數 $P(w_1, w_2, \ldots w_N)$，其中 $w_1, w_2, \ldots, w_N$ 為此詞序列所包含的詞。但對整個詞序列的估測參數會隨著詞數量成指數成長，因此會遭遇資料稀疏(Data Sparseness)的問題。為了解決此問題，我們會先將語言模型的式子展開成機率的連乘積，再利用 n-1階的馬可夫假設(n-1 Order Markovian Assumption)做簡化的動作，如式(1-3)所示:

$$P(W) = P(w_1, w_2, \ldots, w_N) = \prod_{k=1}^{N} P(w_k \mid w_{k-1}, w_{k-2}, \ldots, w_{k-n+1}) \qquad (1\text{-}3)$$

其中 N 為詞的個數，$w_{k-1}, w_{k-2}, \ldots, w_{k-n+1}$ 則是 w_k 的歷史詞序列，式(1-3)便是常見的 n-連(n-gram)語言模型表示法。一般為了方便起見，以及減少參數量的複雜度，常使用詞雙連(Bigram)及詞三連(Trigram)兩種模型(也就是分別使用一階及二階的馬可夫假設)。如同聲學模型，語言模型也需要有大量的文字語料來做為訓練之用。n-連語言模型的訓練方法有最大化相似度估測法(Maximum Likelihood Estimation, MLE)及最大熵值法(Maximum Entropy, ME)[Rosenfeld 1996]等，另外為了處理某些詞可能在訓練語料沒有出現的問題，通常會搭配如Katz Smoothing[Katz 1987]及Kneser-Ney

Smoothing[Ney *et al*. 1994]等語言模型平滑技術，對這些估測機率原本為零的部份加以平滑化處理。

1.1.4 語言解碼(Lingusitc Decoding)

在依式(1.2)尋找最佳詞序列時，由於分母的部份並不會影響最後詞序列排名的結果，因此實作上常將分母的部份省略。有了這項前提之後，再將每個音節比對語句中每一個音框，找出一條最佳的詞序列。而為了有效率的求解，一般是使用維特比動態規劃搜尋(Viterbi Dynamic Programming Search)[Viterbi 1967]。此外，由於搜尋空間會隨著詞典大小成指數成長，因此，在搜尋時，通常會透過搜尋路徑裁減(Pruning)技術來停止繼續尋找一些機率較低的詞序列，以減低其計算複雜度及記憶體使用量。

1.2 現階段語音辨識研究內容

目前在語音辨識的研究中，強健性語音辨識(Robust Speech Recognition)主要在探討如何使得語音辨識系統在真實充滿各種噪音的環境底下仍有一定的辨識效果，而除了維護語音辨識系統的效能之外，信心度評估則是應用於準確地判斷語音辨識系統的結果正確性，在1.2.1及1.2.2小節將分別簡介強健性語音辨識及信心度評估的研究。

1.2.1 強健性語音辨識

此方法主要為降低環境噪音或不同語者等因素對語音訊號的影響，或是找出比較具有鑑別性的語音特徵，使得語音辨識的正確率不會因環境的噪音因素而有所降低。大致來說，可再細分為兩個方向:

(i) 語音強化(Speech Enhancement)

這類方向主要是提昇語音訊號本身的品質，希望藉由乾淨語音及噪音不同的統計特性，想辦法將受噪音影響的聲音訊號還原成乾淨語音。常見的技術有頻譜消去法(Spectral Subtraction)[Boll 1979]及維爾濾波器(Wiener Filter)[Wiener 1949]等。

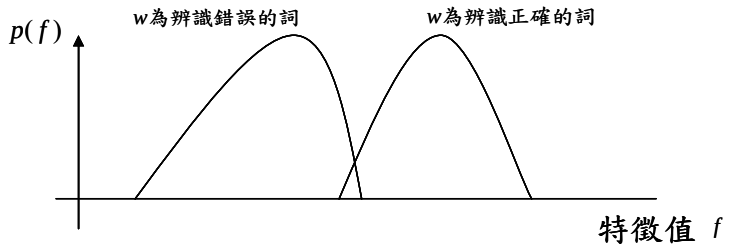

圖 1-4 一個預估參數範例

(ii) 強健性語音特徵(Robust Speech Features)

這類技術是以擷取語音訊號中較具有強健性的特徵為主要目的，使得擷取出來的特徵可以抵抗週遭的環境變化。常見的技術有倒頻譜平均消去法(Cepstral Mean Subtraction, CMS)[Atal 1974]、倒頻譜正規化法(Cepstral Normalization, CN)[Viikki and Laurila 1998]、統計圖等化法(Histogram Equalization, HEQ)[Korkmazsky 2004]等。

除了以上兩項技術之外，還可以利用鑑別性分析(Discriminant Analysis)方法來計算原始語音資料的一些相關統計資訊，將原本的語音特徵投影到新的特徵空間，以得到較具有鑑別性的特徵。較常見的方法有線性鑑別分析(Linear Discriminant Analysis, LDA)[Duda and Hart 1973]、異質性線性鑑別分析(Heteroscedastic Linear Discriminant Analysis, HLDA)[Gales 1999]等。

1.2.2 信心度評估

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

(i) 以特徵為基礎(Feature-based)之信心度評估

此種方法通常都是利用在進行語音辨識的過程中可獲得的一些所謂的預估特徵(Predictor Features，包含聲學及語言等資訊)。而一個特徵要能被稱為是預估特徵，其特徵值對正確辨認詞及錯誤辨認詞所建立的機率密度函式(Probability Density Function, PDF)必須具有很大的鑑別性，如圖 1-4 所示。另一方面，每個預估特徵之間可利用某種方式結合，再配合不同的分類器，如有限向量機(Support Vector Machine, SVM)[Zhang and Rudnicky 2001]、自然貝氏分類器(Naïve Bayes Classifier)[Sanchis *et al*. 2004]或決策樹(Decision Tree)[Eide *et al*. 1995; Neti *et al*. 1997]等來決定辨識結果的正確性。

(ii) 發音確認 (Utterance Verification)

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

$$\begin{aligned} H_0 &\text{ (虛無假設，Null Hypothesis) : } X \text{ 之辨認的結果為正確} \\ H_1 &\text{ (對立假設，Alternative Hypothesis) : } X \text{ 之辨認的結果為錯誤} \end{aligned} \tag{1-4}$$

其中 X 代表一段聲學觀測序列。然後，我們測試虛無假設及對立假設，以決定辨識結果之正確與否。而測試之方法則是使用相似度比例檢測(Likelihood Ratio Testing, LRT) :

$$\frac{p(X \mid H_0)}{p(X \mid H_1)} \underset{H_1}{\overset{H_0}{\gtrless}} \tau \tag{1-5}$$

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

(iii) 事後機率 (Posterior Probability)

在傳統的最大事後機率(Maximum a Posterior, MAP)語音辨識方法中，式(1-2)的事後機率 $P(W \mid X)$ 對詞序列而言其實可以算是一種很好的信心度評估準則。但

是如1.1.4小節所提到的，我們通常會省略分母項，造成語音辨識系統輸出的分數不再是介於0到1的值。即使不省略分母項，但由於語音訊號有無窮多種，所以要如何估測出 $p(X)$ 便變成了一個癥結所在。為了解決這個問題，先前學者的研究曾提出了下列兩種方式來求得近似解:

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

B. 圖形化基礎(Graph-based)法:這類的方法主要是根據前向後向演算法(Forward-backward Algorithm)在詞圖上(Word Graph)上計算分母的值[Kemp and Schaaf 1997;Wessel *et al*. 2001]等，本論文將會於 2.2 小節更深入的探討此種方法。

1.3 本論文研究成果與貢獻

如之前所提到的，在過去的研究中，信心度評估主要是用來判斷語音辨識系統其辨識結果的正確性，再決定是否要接受其結果，應用範圍並不算廣泛。直至最近，開始有學者將信心度評估應用至別的研究領域，如使用事後機率增進語音辨識系統的正確率[Wessel *et al*. 2000]、非監督式(Unsupervised)聲學模型訓練[Wessel and Ney 2005; Chen *et al*. 2005]以及大詞彙連續語音辨識(Large Vocabulary Continuous Speech Recognition, LVCSR)系統的往前觀測(Look-ahead)法則[Afify *et al*. 2005]等。

在傳統判別辨識結果的正確性研究方面，本論文提出使用熵值資訊並結合以事後機率為基礎之信心度評估方法，在公視新聞外場記者及受訪者測試語料最佳實驗結果中，較傳統以事後機率為基礎之信心度評估有16.37%及12.00%的信心度錯誤率相對下降。而在探討信心度評估於提高語音辨識的正確率方面，吾人嘗試結合以梅爾倒頻譜係數及異質性線性鑑別分析搭配最大相似度線性轉換(Maximum Likelihood Linear Transformation, MLLT)兩種不同語音特徵參數所形成的詞圖資訊，並以最小化

音框錯誤率(Minimum Time Frame Error Decoding)[Wessel *et al*. 2001b]來降低辨識系統的字錯誤率，實驗結果顯示，在外場記者及受訪者測試語料中各獲得4.6%及4.8%的相對字錯誤率減少。最後，本論文嘗試在傳統以Levenshtein距離為成本函式(Cost Function)的最小化貝氏風險(Minimum Bayes Risk)[Goel and Byrne 2000]，適當的加入以特徵為基礎的信心度評估，經由*N*-最佳化詞序列重新排序(Reranking)實驗得知，相較於傳統單單使用Levenshtein距離為成本函式的最小化貝氏風險辨識法則而言，有少許的相對字錯誤率減少。

1.4 論文架構

本論文第二章將回顧過去有關於信心度評估的研究，主要討論過去有關如何計算信心度評估的方法，以及一些應用信心度評估至其它領域的研究。第三章則是介紹大詞彙連續語音辨識系統的基本架構、實驗語料的設定以及實驗評估方法。第四章則是有關信心度評估、使用事後機率增進辨識系統正確率的基礎實驗。第五章的部份則是除了探討如何結合熵值(Entropy)資訊與傳統信心度評估外，還討論關於結合了不同語音特徵參數所形成的詞圖後之最小化音框錯誤率(Time Frame Error)解碼以及結合了信心度評估的最小化貝氏風險(Minumum Bayes Risk)解碼之實驗。第6章則是結論與未來之展望，探討未來可繼續研究之方向。

第2章 文獻回顧

2.1 以特徵為基礎之信心度評估

在過去文獻中，有許多信心度評估的相關研究都是為了找出可以有效判斷語音辨識結果正確性的預估特徵。通常這些預估特徵是在辨識過程中從聲學模型分數、語言模型分數、語法(Syntax)等三種不同的資訊收集得來的。一些聲學、語言模型或語法相關常見的特徵如下:

(i) 正規化聲學對數相似度(Normalized Acoustic Log Likelihood):聲學對數分數除以辨識結果所佔的音框個數。

(ii) *N*-最佳(*N*-best)詞序列相關特徵:先由詞圖(Word Graph，在 2.2.1 節會詳細的介紹)產生 N 條分數最高的詞序列。其相關的特徵有:某個候選詞在 N 條最佳詞序列中出現的次數，如式(2.1) :

$$\frac{\sum_{w \in Wn} 1}{N} \tag{2-1}$$

其中 w 代表一個詞，而 W_n 代表在*N*-最佳詞序列中分數排名第 n 高的詞序列;或是前*N*條中包含某個辨識結果的相似度權重比例(Weighted Ratio)，如式(2.2)所示:

$$\frac{\sum_{w \in Wn} f(n)P(W_n \mid X)}{\sum_{n=1}^{N} f(n)P(W_n \mid X)} \tag{2-2}$$

其中 $f(n)$ 為權重函式， $P(W_n \mid X)$ 代表聲學觀測序列 X 產生詞序列 W_n 的機率。

(iii) 聲學穩定度(Acoustic Stability):其主要概念為我們在尋找最佳詞序列時，可在語言模型分數的部份使用不同的權重β ，如式(2-3)所示:

$$p(X \mid W)P(W)^{\beta} \tag{2-3}$$

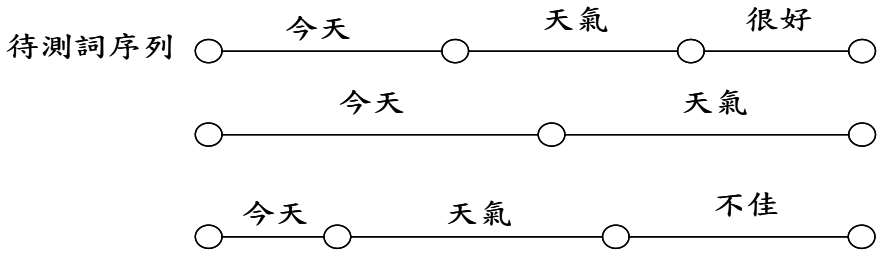

圖 2-1 三個不同語言模型權重所產生的詞序列

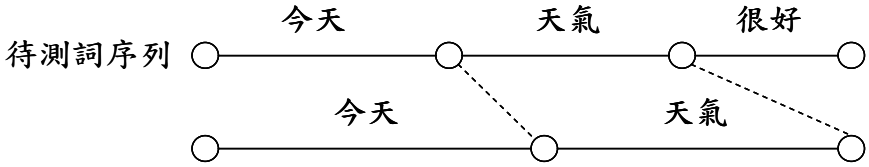

圖 2-2 第一條及第二條詞序列做完比對後之結果

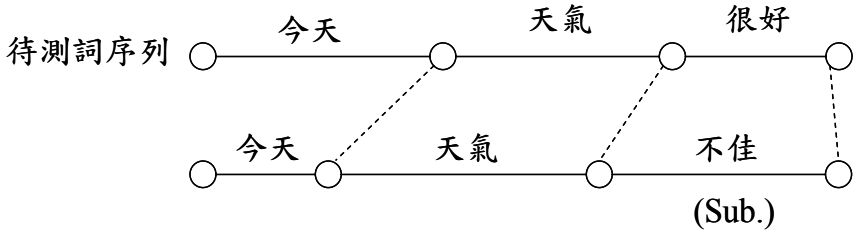

圖 2-3 第一條及第三條詞序列做完比對後之結果

因而得到不同的最佳詞序列。接下來針對某一個特定語言模型權重的詞序列(也就是實際上真正需做信心度評估的輸出結果，通常會事先由一組訓練資料中，找出能使訓練資料的辨識率為最高的權重值)，與其它不同語言模言權重所得到的詞序列做比對(利用Levenshtein Alignment)，計算某個辨識結果是否有出現在其它詞序列的一個位置。舉例來說，假設我們現在有三條不同語言模型權重所產生的詞序列，如圖 2-1所示。其中第一條是我們辨識器最後輸出的結果，也就是這裡所謂的待測詞序列。因此我們必須將其結果分別與第二條與第三條詞序列做比對，其比對後的結果如圖 2-2及圖 2-3所示。其中(Sub.)表示詞的替代(Substitution)。根據這三個圖，我們便可以知道在待測詞序列中，”今天”跟”天

氣”與其它兩條詞序列中都對齊至同一位置，其聲學穩定度的值便為2。而”很好”則沒有任何一條詞序列的詞對齊至同一位置，其聲學穩定度的值便是0。

(iv) 候選詞假設密度(Word Hypothesis Density):計算詞圖(Word Graph)中，在一個詞段(Word Arc)中，平均其它詞段出現的個數($HD(\bullet)$)，可以下列數學式表示

$$D(t') = \left|\{b : [w_b; s_b, e_b] \in WG \wedge (s_b \le t' \le e_b)\}\right| \tag{2-4}$$

$$HD(a : [w_a; s_a, e_a]) = \frac{1}{e_a - s_a + 1} \sum_{t=s_a}^{e_a} D(t) \tag{2-5}$$

其中 w_a 代表詞圖 WG 的一個詞， s_a 及 e_a 分別為 w_a 的開始及結束時間， $D(t')$ 則是代表在詞圖中有多少不同詞段其開始時間及結束時間有包含 t' 這個時間點。

(v) 持續時間(Duration)相關之特徵:一般而言，詞、音節或音素等辨識單位各自的持續時間差異性不太，因此，持續時間也可算是一個適合的預估特徵。

(vi) 語言模型:包括語言模型分數或語言模型回退(Back-off)行為[Chen and Goodman 1999]。假設我們現在有一個詞序列，包含 w_1 、 w_2 及 w_3 三個詞。在我們要估計這個詞的三連語言模型時，加上回退行為，其計算可能有下列四種情況:

A. w_1 、 w_2 及 w_3 此序列在訓練語料中的確存在，可由訓練語料直接估得此三連語言模型的機率 $P(w_3 \mid w_2, w_1)$ 。

B. 無法由訓練語料直接估得 $P(w_3 \mid w_2, w_1)$ 此三連機率，需由二連－二連語言模型回退行為所估得，如式(2-6)所示:

$$P(w_3 \mid w_2, w_1) \approx P(w_2 \mid w_1) + P(w_3 \mid w_2) \tag{2-6}$$

其中 $P(w_2 \mid w_1)$ 及 $P(w_3 \mid w_2)$ 可直接由訓練語料求得。

C. 無法由訓練語料直接估得 $P(w_3 \mid w_2, w_1)$ 此三連機率及 $P(w_2 \mid w_1)$ 二連機率，需由單連-二連語言模型回退行為所估得，如式(2-7)所示 :

$$P(w_3 \mid w_2, w_1) \approx P(w_1) + P(w_3 \mid w_2) \tag{2-7}$$

其中 $P(w_1)$ 及 $P(w_3 \mid w_2)$ 可直接於訓練語料估計。

D. 無法由訓練語料直接估得 $P(w_3 \mid w_2, w_1)$ 及 $P(w_3 \mid w_2)$ 此二項機率，需由二連－單連語言模型回退行為所估得，如式(2-8)所示:

$$P(w_3 \mid w_2, w_1) \approx P(w_3) + P(w_2 \mid w_1) \tag{2-8}$$

其中 $P(w_2 \mid w_1)$ 及 $P(w_3)$ 可直接由訓練語料求得。

而對上述四項情況，我們可以分別給定不同的信心度值。在[Uhrik and Ward 1997]中提到:能直接在訓練語料中求得的語言模型，其信心度會比較高，而如果需要回退行為才能求得的語言模型機率，其信心度較低。

(vii) 與語法剖析相關之特徵:一個詞是否能被文法(Grammar)正確地剖析[Sarikaya *et al*. 2005]。

有關預估特徵更詳細的資訊可以進一步參考[Cox and Rose 1996; Schaaf and Kemp 1997; Chase 1997; San-Segundo *et al*. 2001; Sanchis *et al*. 2003; Lane and Kawahara 2005; Benitez *et al*. 2000]。

在本節中我們將以自然貝氏分類器(Naïve Bayes Classifier)為例，介紹如何根據以上這些特徵值來判斷辨識結果的正確性。為了要使用自然貝氏分類器來從事信心度評估，我們必須事先定義兩種類別:C_1 代表語音辨識結果為正確，C_2 代表語音辨識結果為錯誤。另外，需為每一個辨識詞序列中的每一個辨識詞 w 找出一組預估特徵參數向量 $\vec{f}$ (每一維代表一種預估特徵)，之後再利用貝氏定理可為每一個 w 求取屬於 C_1 之信心度，亦即

$$P(C_1 \mid \vec{f}, w) = \frac{P(C_1 \mid w) P(\vec{f} \mid C_1, w)}{\sum_{C'=C_1 \text{or} C_2} P(C' \mid w) P(\vec{f} \mid C', w)} \tag{2-9}$$

為了增加運算速度及減少參數估測量，可假設當給定詞及類別資訊 C_i 時，每一維的特徵 f_d 為互相獨立，所以 $P(\vec{f} \mid C_i, w)$ 可表示成:

$$P(\vec{f} \mid C_i, w) = \prod_{d=1}^{D} P(f_d \mid C_i, w) \tag{2-10}$$

而 $P(C_i \mid w)$ 與 $P(f_d \mid C_i, w)$ 中的每個式子皆可利用最大相似度估測(Maximum Likelihood Estimation, MLE)，由頻率統計(Frequency Count)求得:

$$P(C_i \mid w) = \frac{N(C_i, w)}{N(w)} \tag{2-11}$$

$$P(f_d \mid C_i, w) = \frac{N(f_d, C_i, w)}{N(C_i, w)} \tag{2-12}$$

其中 $N(\bullet)$ 代表某個事件(event)出現的次數，而上述兩式有機率為零的問題，可使用一般在語言模型常使用的絕對折扣平滑(Absolute Discounting Smoothing)技術來解決:

$$P(C_i \mid w) = \begin{cases} \dfrac{N(C_i, w) - b}{N(w)} & \text{if } N(C_i, w) > 0 \\ \dfrac{b}{N(w)} & \text{if } N(C_i, w) = 0 \end{cases} \tag{2-13}$$

$$P(f_d \mid C_i, w) = \begin{cases} \dfrac{N(f_d, C_i, w) - b}{N(C_i, w)} & \text{if } N(f_d, C_i, w) > 0 \\ M \dfrac{P(f_d \mid C_i)}{\sum_{f'_d : N(f'_d, C_i, w) = 0} P(f'_d \mid C_i)} & \text{if } N(f_d, C_i, w) = 0 \end{cases} \tag{2-14}$$

其中式(2-13)成立的原因為 C_i 在這裡的討論假設僅只有分正確及不確定兩類，亦即 $C_i \in \{C_1, C_2\}$。而 b 為介於0~1的數字，式(2-14)中的 M 則代表 $N(f_d, C_i, w)$ 為非零值的事件個數乘上非零值的事件扣掉的值 $\dfrac{b}{N(C_i, w)}$。在最後決定信心度評估時，我們是計算 $P(C_1 \mid \vec{f}, w)$ 的值。當其值大於事先設定的門檻值時，便認為辨識詞為正確之語音辨識結果，否則就當成是錯誤的語音辨識結果。

一般而言，由於每個上述聲學或語言模型等方面的預估特徵都有極高的相關性[Kemp and Schaaf 1997; Schaff and Kemp 1997]。因此，即使將所有較有用的預估特徵都集合起來使用，對效能也不一定會有很大的提昇(與單一最有用的特徵比較)。因此，近來有學者嘗試將單一最有用的聲學預估特徵(如事後機率或*N*-最佳詞序列相關

特徵)，與純粹的語言特徵，如語意剖析(Semantic Parsing)或其它與語意相關資訊相互結合[Zhang and Rudnicky 2001; Guo *et al*. 2004]，的確對效能有些許的提昇。

2.2 事後機率之信心度評估

先前於1.1.4小節提到，傳統自動語音辨識演算法是在給定任何的聲學觀測序列(Acoustic Observation Sequence) X 時，使用最大事後機率決策方式找出一條最有可能的詞序列 $\hat{W}$，使得它有最大的事後機率 $P(\hat{W} \mid X)$:

$$\begin{aligned} \hat{W} &= \arg\max_{W \in \overline{\mathbf{W}}_{\Sigma}} P(W \mid X) \\ &= \arg\max_{W \in \overline{\mathbf{W}}_{\Sigma}} \frac{p(X \mid W)P(W)}{p(X)} \\ &= \arg\max_{W \in \overline{\mathbf{W}}_{\Sigma}} p(X \mid W)P(W) \end{aligned} \tag{2-15}$$

其中 $\overline{\mathbf{W}}_{\Sigma}$ 代表語言中所有詞序列的組合，$P(W)$ 為 W 的語言模型機率，$p(X)$ 為聲學觀測序列 X 的事前機率，而 $p(X \mid W)$ 則是代表假設 W 為辨識結果的情況下，產生 X 的機率(也就是之前提到的聲學模型)。通常來說，此事後機率 $P(W \mid X)$ 對辨識結果的詞序列是一個很好的信心度評估。但是由於實際運用上的考量，在使用式(2-15)當作自動語音辨識的方法時，因為分母項 $p(X)$ 不影響詞序列排序，我們都會忽略 $p(X)$。這也說明了為什麼語音系統辨識結果的分數是不適合用來當作評估辨識結果可靠度的依據。但只要我們將語音系統辨識結果的分數再除以 $p(X)$ 的值，那麼此新的數值便是介於0~1的定量數值，可以用來判斷 X 跟 W 之間匹配的程度。就理論來說，我們可以依式(2-16)計算 $p(X)$ 的值:

$$p(X) = \sum_{W \in \overline{\mathbf{W}}_{\Sigma}} p(X, W) = \sum_{W \in \overline{\mathbf{W}}_{\Sigma}} P(W)p(X \mid W) \tag{2-16}$$

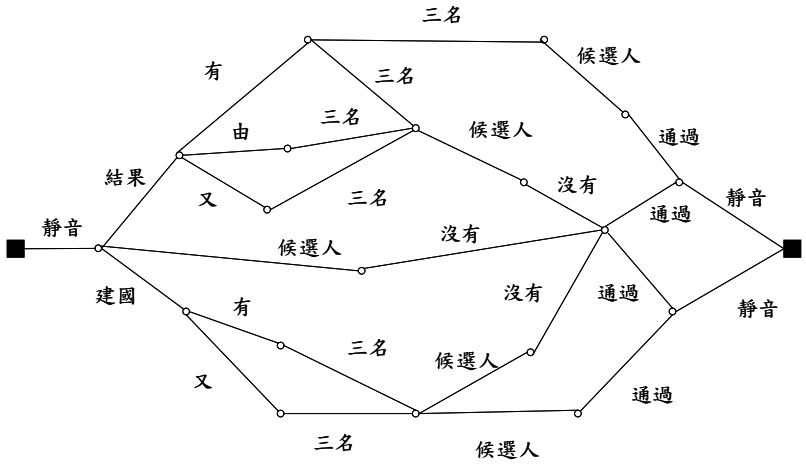

圖 2-4 詞圖 Ψ^X，為所有可能詞序列 $\overline{\mathbf{W}}_\Sigma$ 的近似表示

其中 W 代表 X 一個可能的辨識詞序列。顯而易見的，如果我們沒有對此假設做一些限制，要求得 $p(X)$ 是一件很困難的事，畢竟我們無法加總語言中的所有可能詞序列。所以我們通常會做一些限制或是用近似的方法計算 $p(X)$。如1.2.2小節所提到的，通常會有兩類方法，亦即圖形化基礎法與填充化基礎法，而本論文主要是討論圖形化基礎法。

在圖形化基礎法中，通常會先對每一個 X 產生對應的詞圖 Ψ^X，如圖 2-4所示，由於只有 $p(X|W)P(W)$ 機率較大的詞序列才會有可能留在 Ψ^X 中。因此，詞圖可說是所有可能詞序列的近似表示，以用來求得式(2-16)的近似值。在有了詞圖及每條詞序列的事後機率後，可進而求得每個辨識詞的事後機率。在下面的小節中，本論文將會更詳細的介紹有關於事後機率的計算及其相關變形。

2.2.1 詞圖簡介

詞圖為一有向性，非環狀(Acyclic)的圖形表示法，圖 2-4中的每一個節點代表一個時間點，每個詞段(Word Arc) a 由三個變數表示，$a:[w_a;s_a,e_a]$，其中 w_a 代表其對應的詞編號為何，s_a 代表詞段開始時間，e_a 則代表結束時間。每個詞段通常會給定某種分數，較常見的為此詞段產生語音段落的聲學分數 $p(X_{s_a}^{e_a}|w_a)$。而每個詞圖都會有兩個特殊的節點，分別代表詞圖的開始及結束(如圖 2-4的兩個實心點)。只要是從開

始節點到結束節點的任何路徑都可視為一條完整路徑(Complete Path)，而任一條完整路徑都可以代表聲學觀測序列 X 的一條可能辨識詞序列。

2.2.2 計算一個詞段的事後機率

根據2.2.1小節的說明，我們可以將詞圖 Ψ^X 上某個詞段 $a:[w_a;s_a,e_a]$ 的事後機率，$P(a:[w_a;s_a,e_a]\,|\,\Psi^X)$(在詞圖上計算一個詞段的事後機率，與原本給定一個聲學觀測序列 X 計算事後機率並不相同, 因為 Ψ^X 只為 $\overline{\mathbf{W}}_\Sigma$ 之近似表示)看成是所有通過這個詞段的所有完整路徑的分數總和除以詞圖上所有完整路徑分數總和，如式(2.17)所示:

$$P(a:[w_a;s_a,e_a]\,|\,\Psi^X)=\frac{\sum_{\{\overline{W}:[w^n;s^n,e^n]_{n=1}^{N}\}\in\Psi^X,\,a\subset\overline{W}}\left\{\prod_{n=1}^{N}p(X_{s^n}^{e^n}\,|\,w^n)\cdot P(w^n\,|\,h^n)\right\}}{\sum_{\{\overline{W}':[w^m;s^m,e^m]_{m=1}^{M}\}\in\Psi^X}\left\{\prod_{m=1}^{M}p(X_{s^m}^{e^m}\,|\,w^m)\cdot P(w^m\,|\,h^m)\right\}} \quad (2\text{-}17)$$

其中，$\overline{W}$ 代表在詞圖的一條完整路徑，共有 N 個詞段，$a\subset\overline{W}$ 代表包含詞段 a 的完整路徑 $\overline{W}$，h^x 為 w^x 的詞歷史(Word History)。$p(X_{s^n}^{e^n}\,|\,w^n)$ 代表開始時間 s^n 至結束時間 e^n 此段聲學觀測序列的聲學相似度，而 $P(w^n\,|\,h^n)$ 代表其語言模型分數。$P(a:[w_a;s_a,e_a]\,|\,\Psi^X)$ 可以用前向後向(Forward-backward)演算法有效率地求解。演算法詳細過程如圖 2-5所示，其主要概念為在計算通過某詞段 $a:[w_a,s_a,e_a]$ 的所有完整路徑分數總和時，首先加總由其結束時間為 s_a-1 的所有詞段其轉移至此詞段機率乘上先前每個詞段所累積的分數，最後乘上 $a:[w_a,s_a,e_a]$ 的聲學分數，代表由從 t 為0至此詞段的所有路徑分數總和(也就是所謂的"前向"部份，如圖 2-5中前兩個最外層的for迴圈);接著再加總由其開始時間為 e_a+1 的詞段其轉移至此詞段機率乘上每個開始時間為 e_a+1 的詞段先前所累積的分數及其對應的聲學分數，做為由 $t=T-1$ 至 e_a 的所有路徑分數總和(也就是所謂的"後向"路份，如圖 2-5中第三及第四個最外層的for迴圈)。將此前向及後向的分數相乘之後，便是通過詞段 $a:[w_a,s_a,e_a]$ 的所有完

整路徑分數總和。只要將所有結束時間點為$t = T - 1$的詞段對應之前向分數加總，即可獲得詞圖所有完整路徑總和。

假設 a 與 r 均為詞圖中之詞段，$P(a|r)$ 表示由詞段 $r:[w_r;s_r,e_r]$ 至詞段 $a:[w_a;s_a,e_a]$ 的轉移機率(Transition Probability)，γ_a 則為詞段 a 的事後機率，κ 為聲學分數的權重。詞圖起始時間為 0，結束時間為 T-1

for 開始時間為 0 的詞段 a

$\alpha_a = p(X_{s_a}^{e_a} | w_a)^{\kappa}$

end //前向初始化

for t=1 to T-1

for 開始時間為 t 的詞段 a

$\alpha_a = 0$

for 結束時間為 t-1 的詞段 r

$\alpha_a = \alpha_a + \alpha_r P(a|r)$

end

$\alpha_a = \alpha_a \cdot p(X_{s_a}^{e_a} | w_a)^{\kappa}$

end

end //前向遞迴

for 結束時間為 T-1 的詞段 a

$\beta_a = 1$

end //後向初始化

for t=T-2 to 0

for 結束時間為 t 的詞段 a

$\beta_a = 0$

for 開始時間為 t+1 的詞段 r

$\beta_a = \beta_a + \beta_r P(r|a) p(X_{s_\gamma}^{e_\gamma} | r)^{\kappa}$

end

end

end //後向遞迴

for 每一詞段 a

$$\gamma_a = \frac{\alpha_a \beta_a}{\sum_{\text{所有在時間}T-1\text{結束的詞段}a} \alpha_a}$$

end

圖 2-5 利用前向後向演算法求得詞圖中某一詞段的事後機率

2.2.3 計算一個辨識詞的信心度

基本上，我們可以直接使用某個詞段 $a:[w_a;s_a,e_a]$ 的事後機率，$P(a:[w_a;s_a,e_a]\,|\,\Psi^X)$，當作是辨識詞 w_a 的信心度評估，便能得到一定的效果。但我們知道，在詞圖中，對詞段 $a:[w_a;s_a,e_a]$ 而言，會有許多詞段 γ 與詞段 a 對應到相同的詞編號，只是開始時間及結束時間有些許不同，因此求某個詞 w_a 的信心度時，除了考慮自己本身的事後機率 $P(a:[w_a;s_a,e_a]\,|\,\Psi^X)$ 之外，也可以加入那些開始及結束時間有些微的差距，但是詞編號是一樣的詞段。在前人的研究中，[Wessel *et al.* 2001] 提出了三個計算方法:

$$C_{\text{sec}}(a:[w_a;s_a,e_a]) = \sum_{\substack{r:[w_r;s_r,e_r],w_r=w_a \\ \{s_r,\ldots,e_r\}\cap\{s_a,\ldots,e_a\}\neq\phi}} P(r:[w_r;s_r,e_r]\,|\,\Psi^X) \tag{2-18}$$

$$C_{med}(a:[w_a;s_a,e_a]) = \sum_{\substack{r:[w_r;s_r,e_r],w_r=w_a \\ s_r\le[(s_a+e_a)/2]\le e_r}} P(r:[w_r;s_r,e_r]\,|\,\Psi^X) \tag{2-19}$$

$$C_{\max}(a:[w_a;s_a,e_a]) = \max_{t\in\{s_a,\ldots,e_a\}} \sum_{\substack{r:[w_r;s_r,e_r],w_r=w_a \\ s_r\le t\le e_r}} P(r:[w_r;s_r,e_r]\,|\,\Psi^X) \tag{2-20}$$

其中 $P(r:[w_r;s_r,e_r]\,|\,\Psi^X)$ 的算法可參考式(2-17)。式(2-18)的意思為將時間上與現在這個詞段 $a:[w_a;s_a,e_a]$ 相交，而且對應詞編號相同的所有詞段事後機率相加，當作辨識詞 w_a 的信心度。而式(2-19)則是累加詞編號相同，但是必須與詞段 $a:[w_a;s_a,e_a]$ 的時間中點相交之詞段事後機率。最後，式(2-20)則是不單單只看與 $a:[w_a;s_a,e_a]$ 時間中點相交的相同詞編號詞段之累加事後機率，而是也考慮到相交於 s_a 到 e_a 之中任意時間點相同詞編號詞段之累加的事後機率，然後再取有最大累積值的時間點之分數作為辨識詞 w_a 的信心度。

另外，在[Lo and Soong 2005]中則是提到對於在計算 w_a 的事後機率時，應該注意的幾點事項:

(i) 縮減的搜尋空間(Reduced Search Space):如2.2.2小節所提到的，由於我們無法針對語言中所有可能的詞序列做加總的動作，另外也為了避免搜尋空間太過龐大,所以通常都是在一些較簡化的搜尋空間(如詞圖或*N*-最佳詞序列)來計算辨識結果的事後機率。

(ii) 放寬時間的限制:因為一個詞段是由詞編號、開始時間及結束時間三項要素所構成，而辨識詞其開始及結束時間會在搜尋最佳詞序列時，因其搜尋空間大小而有許多不同的可能。因此，針對有不同的開始及結束時間，但其時間點有重疊(Overlap)且詞編號為相同的詞段，應該視為同樣的詞段。

(iii) 給定聲學及語言模型分數不同的權重:考慮這項因素的原因，主要是因為下列二項特徵:

A. 聲學模型分數區間範圍為 0 到正無窮大，但語言模型的分數卻只介於 0~1 之間

B. 每個音框都會計算其聲學模型分數，而語言模型分數則通常是在經過一段時間後才會計算(於 3.1.4 小節說明此緣由)

綜合了以上各項要點後，在[LO and Soong 2005]中，一個詞段的事後機率計算如下:

$$P(a:[w_a;s_a,e_a]\,|\,\Psi^X)=\frac{\sum_{\substack{[w^n;s^n,e^n]_{n=1}^N\in\Psi^X\\ \exists i,1\le i\le N,w^i=w_a,(s_a,e_a)\cap(s^i,e^i)\neq\phi}}\left\{\prod_{i=1}^{N}\ p^{\alpha}(X_{s^i}^{e^i}\,|\,w^i)\cdot P^{\beta}(w^i\,|\,h^i)\right\}}{\sum_{[w^m;s^m,e^m]_{m=1}^M\in\Psi^X}\left\{\prod_{m=1}^{M}\ p^{\alpha}(X_{s^m}^{e^m}\,|\,w^m)\cdot P^{\beta}(w^m\,|\,h^m)\right\}}\quad(2\text{-}21)$$

其中 α 及 β 分別代表聲學及語言模型的權重。

而在[Sanchis *et al*. 2004]中，作者不再只對單一的詞圖作計算，而是在多個詞圖估算信心度評估，其演算法如圖 2-6所示:

令 M_w 為詞典中所有的詞，Ψ_t^X 為在某個時間點 t 時所建構時間持續從 0 至 t 的詞圖，WPost 則是取得當時詞圖上每個詞段的事後機率的函數，γ 則為詞段的事後機率，$C(w,t)$ 用來儲存每個詞 $w \in M_w$ 的每個時間點 $t \in \{0,\ldots,T-1\}$ 之事後機率總和。

for 對每個詞 $w \in M_w$

  for t=0 to T-1

    $C(w,t) = 0$ //初始化

  end

end

for t=0 to T-1

  $A = \text{WPost}(\Psi_t^X)$

  for 每個詞段 $a:([w_a;s_a,e_a],\gamma) \in \Psi_t^X$

    for $t' = s_a$ to e_a

      $C(w_a,t') = C(w_a,t') + \gamma$

    end

  end

end

令 h_{best} 為詞圖 Ψ_{T-1} 中分數最高的詞序列

for 每一詞段 $a:[w_a;s_a,e_a] \in h_{best}$

$$W_{avg}(w_a;s_a,e_a) = \frac{1}{e_a - s_a + 1}\sum_{t=s_a}^{e_a} \frac{C(w_a,t)}{T-t+1}$$

$$W_{\max}(w_a;s_a,e_a) = \max_{s_a \le t \le e_a} \frac{C(w_a,t)}{T-t+1}$$

圖 2-6 在多重詞圖上求辨識詞 w 信心度

在圖 2-6中，W_{avg} 與 $W_{\max}$ 便可以作為辨識詞 w 兩種的信心度評估。另一方面，詞段上的分數除了可以是聲學分數外，也可以是語言模型分數，或是聲學及語言模型分數的組合。

在[Razik *et al.* 2005]中也提出了幾種詞段信心度評估的作法。首先，最簡單的概念為在計算每個詞段的信心度時，將此詞段的聲學分數乘上單連語言模型的分數，而分母項的部份只考慮有同樣的開始、結束時間但詞編號不相同的其它詞段，如式(2-22)所示:

$$C(a:[w_a;s_a,e_a]) = \frac{p(X_{s_a}^{e_a} \mid w_a)\cdot P(w_a)}{\sum_{\substack{r:[w_r,s_r,e_r] \\ s_r=s_a,e_r=e_a}} p(X_{s_r}^{e_r} \mid w_r)\cdot P(w_r)} \tag{2-22}$$

不過，由於這樣的限制太嚴格，會導致$C(a:[w_a;s_a,e_a])$容易趨近於1。因此，在[Razik *et al.* 2005]中試著放寬關於開始、結束時間或詞段時間長度的限制。但由於詞圖上有許多詞段都符合現在所規定的限制，因此在決定分子或分母項的聲學分數$p(X_s^e \mid w)$時，便決定取相同的詞編號且符合目前的限制，具有最大聲學分數的詞段:

$$C(a:[w_a;s_a,e_a]) = \frac{\max_{\substack{a':[w_{a'};s_{a'},e_{a'}] \\ w_{a'}=w_a,a'\in E}} p(X_{s_{a'}}^{e_{a'}} \mid w_{a'})^{\alpha}\cdot P(w_a)^{\beta}}{\sum_{r:[w_r;s_r,e_r]\in E} \max_{\substack{r':[w_{r'};s_{r'},e_{r'}] \\ w_{r'}=w_r,r'\in E}} p(X_{s_{r'}}^{e_{r'}} \mid w_{r'})^{\alpha}\cdot P(w_r)^{\beta}} \tag{2-23}$$

其中α及β分別代表聲學及語言模型的權重，而E代表那些符合放寬時間及詞段長度後但為相同詞編號的其它詞段，如開始時間及結束時間不一定要相同，只要有交集就可以或其它任意放寬的限制。值得注意的是式(2-22)及式(2-23)都只考慮到單連語言模型，這樣的資訊可能太過於粗糙，所以作者便進一步考慮藉由使用雙連語言模型來包含前後詞的資訊。首先，只多看前一個詞的話，便可將式(2-23)改為:

$$C(a:[w_a;s_a,e_a]) = \frac{\max_{\substack{a':[w_{a'};s_{a'},e_{a'}] \\ w_{a'}=w_a,a'\in E}} p(X_{s_{a'}}^{e_{a'}} \mid w_{a'})^{\alpha}\cdot \sum_{\substack{a_p:[w_{a_p};s_{a_p},e_{a_p}] \\ e_{a_p}=s_{a'}-1}} P(w_a \mid w_{a_p})^{\beta}}{\sum_{r:[w_r;s_r,e_r]\in E} \max_{\substack{r':[w_{r'};s_{r'},e_{r'}] \\ w_{r'}=w_r,r'\in E}} p(X_{s_{r'}}^{e_{r'}} \mid w_{r'})^{\alpha}\cdot \sum_{\substack{r_p:[w_{r_p};s_{r_p},e_{r_p}] \\ e_{r_p}=s_{r'}-1}} P(w_r \mid w_{r_p})^{\beta}} \tag{2-24}$$

若再多考慮後面一個詞，最後的式子便可以寫成:

$$C(a:[w_a;s_a,e_a]) = \frac{\max\limits_{\substack{a':[w_{a'};s_{a'},e_{a'}]\\ w_{a'}=w_a, a'\in E}} (p(X_{s_{a'}}^{e_{a'}} \mid w_{a'}))^{\alpha} \cdot \Gamma_{a':[w_{a'};s_{a'},e_{a'}]}}{\sum\limits_{r:[w_r;s_r,e_r]\in E} \max\limits_{\substack{r':[w_{r'};s_{r'},e_{r'}]\\ w_{r'}=w_r, r'\in E}} (p(X_{s_{r'}}^{e_{r'}} \mid w_{r'}))^{\alpha} \cdot \Gamma_{r':[w_{r'};s_{r'},e_{r'}]}} \tag{2-25}$$

其中

$$\Gamma_{a:[w_a;s_a,e_a]} = \sum_{\substack{a_p:[w_{a_p};s_{a_p},e_{a_p}]\\ e_{a_p}=s_a-1}} \sum_{\substack{a_n:[w_{a_n};s_{a_n},e_{a_n}]\\ s_{a_n}=e_a+1}} \left\{P(w_a \mid w_{a_p})P(w_{a_n} \mid w_a)\right\}^{\beta} \tag{2-26}$$

其中 w_{a_p} 及 w_{a_n} 分別代表 w_a 前一個及後一個連接詞。

2.3 引用高階資訊(High Level Information)之信心度評估

除了上述所提到關於聲學、語言及語法方面的信心度評估外，[Cox and Dasmahapatra 2002]認為人們通常還可以藉由語意資訊(Semantic Information)來辨別辨識結果的正確性。因此，他們便提出利用潛藏語意分析(Latent Semantic Analysis, LSA)來判別辨識結果的正確與否，另外[Guo *et al*. 2004]也提出利用詞與詞之間交互訊息(Inter-word Mutual Information)來與事後機率做適當地結合，進而提昇信心度評估的正確率。以下兩小節將會分別介紹此兩種方法。

2.3.1 潛藏語意分析(Latent Semantic Analysis, LSA)

潛藏語意分析近年被廣泛運於資訊檢索(Information Retrieval)[Furnas *et al*. 1988]及語音辨識的語言模型[Belllegarda 1998]等領域。主要是利用線性代數的奇異值分解(Singular Value Decomposition, SVD)來將原本高維度且不相關的詞向量(Word Vector)與文件向量(Document Vector)投影到較低維度的潛藏語意空間(Latent Semantic Space)。如果兩個詞向量在此潛藏語意空間利用餘弦估測(Cosine Measure)的值愈大(在空間上愈接近)，則此兩個詞也有比較接近的語意。

在進行奇異值分解之前，我們必須先建立一個詞-文件矩陣(Word-document Matrix) A，此矩陣可經由事先收集大量文件資料求得。假設我們詞典有 m 個詞，而文件有 n 篇，則此矩陣 A 的維度大小便是 $m \times n$，而每個元素的算法通常可表示為

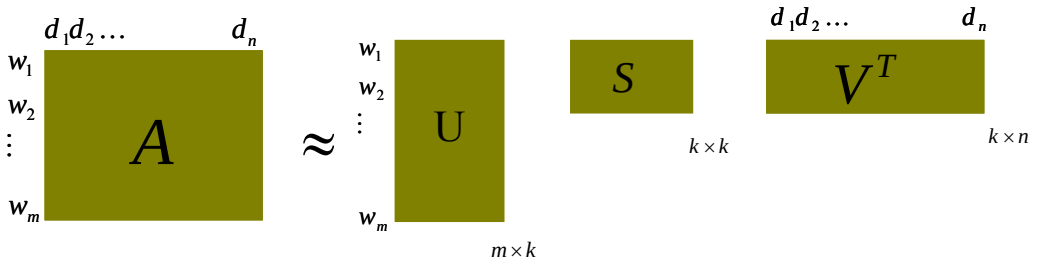

圖 2-7 奇異值分解

$$a_{ij} = (1 - E_i) \times \log(1 + \frac{c_{ij}}{n_j}) \tag{2-27}$$

c_{ij} 是詞 w_i 出現在第 j 篇文件 d_j 的次數;n_j 則是代表 d_j 的大小。而 E_i 則可視為 w_i 的正規化熵值(Normalized Entropy) :

$$E_i = -\frac{1}{\log_2(N)} \sum_{j=1}^{N} f_{ij} \log_2 f_{ij} \tag{2-28}$$

其中

$$f_{ij} = \frac{c_{ij}}{t_i} \tag{2-29}$$

t_i 代表詞 w_i 出現在 n 篇文件的總次數。而 $0 \le E_i \le 1$(在有 $f_{ij} = 1$ 以及所有 $f_{ij} = \frac{1}{N}$ 時分別有最小和最大值),因此當 E_i 的值愈大,則代表 w_i 在各個文件中出現的次數趨近相同,其重要性便比較低;反之當 E_i 的值愈小,w_i 出現次數集中於某幾篇文件,其重要性因而提高。另外 $\frac{c_{ij}}{n_i}$ 可視為 w_i 在 d_j 的重要性,其值愈大,則代表其重要性愈高;反之,其重要性愈低。雖然有了此詞-文件矩陣 A,但由於每篇文件不可能包含所有的詞,通常此矩陣 A 裡面的元素會有許多的值為0,且詞與文件的維度不同,其代表的意義也不同。因此,可以利用進行奇異值分解來做降維的動作。而奇異值分解的公式如下:

$$A \approx USV^T \tag{2-30}$$

其中U代表$m \times k$維的左奇異矩陣;S為$k \times k$維的對角矩陣;V是$n \times k$維的右奇異矩陣;V^T則是代表V的轉置矩陣(Transposition Matrix);k為小於等於矩陣A的秩R的一個整數值。奇異值分解的概念可以用圖 2-7來表示。經過奇異值分解後，詞和文件就都被投影到維度較低的潛藏語意空間，而原本在矩陣A的列向量便可用U的列向量$\vec{u}_i$來表示，而A的行向量可以改用V^T的行向量$\vec{v}_j^T$來代表。其中$\vec{u}_i$與$\vec{v}_j^T$的每一維度有一對一的對應關係，代表某一種潛藏的語意空間[Bellegarda 2000; 2005]。因此，如果我們想計算兩個詞w_i及w_j在語意上是否有相關聯，便可以藉著$\vec{u}_i$及$\vec{u}_j$的餘弦估測值來決定。

在實作上，當我們將一個聲學觀測序列辨識成一詞序列$W = w_1, w_2, \ldots, w_N$時，一個詞$w_i$的信心度評估可用下式計算:

$$MSS_i = \frac{1}{N} \sum_{j=1}^{N} \text{Cos}(U(w_i), U(w_j)) \tag{2-31}$$

MSS_i代表w_i的平均語意相似度(Mean Semantic Similarity)，而$U(w_i)$代表做完奇異值分解後的$\vec{u}_i$向量。Cos(·,·)則是兩個向量的餘弦估測函式。但由於功能詞(Function Word)對其它的詞在語意上幾乎都很接近，加上功能詞常常會一直出現在辨識詞序列中，使得MSS_i的值容易變大，而影響(2-31)式的正確性。為了避免這樣的情況，在求w_i的平均語意相似度時，通常都不考慮功能詞，而採用式(2-32):

$$MSS_i = \frac{1}{N - N_{w_j \in W^f}} \sum_{j=1, w_j \notin W^f}^{N} \text{Cos}(U(w_i), U(w_j)) \tag{2-32}$$

其中W^f代表所有的功能詞集合，$N_{w_j \in W^f}$則代表詞序列中功能詞的個數，除了MSS_i之外，更進一步關於利用潛藏語意分析計算某詞w_i的信心度，請參考[Cox and Dasmahapatra 2002]。

2.3.2 交互資訊(Mutual Information, MI)

交互資訊可以視為是兩個變數(Variables)相依(Dependence)程度。而當給定兩個詞 w_i 及 w_j 時，其交互訊息的計算如式(2-33)所示:

$$MI(w_i, w_j) = \log\left(\frac{P(w_i, w_j)}{P(w_i)P(w_j)}\right) \tag{2-33}$$

其中

$$P(w_i) = \sum_{w_j} P(w_i, w_j) \tag{2-34}$$

而

$$P(w_i, w_j) = \frac{N(w_i, w_j)}{\sum_{w_l, w_k} N(w_l, w_k)} \tag{2-35}$$

其中 $N(w_i, w_j)$ 代表 w_i 及 w_j 同時在訓練資料出現的次數，而式(2-35)的分母項則是代表語料庫中所有詞對(Word Pair)個數。因此，一個辨識詞序列 $W = w_1, w_2, \ldots, w_N$ 中某個詞 w_i 的信心度可以表示成 w_i 與辨識詞序列其它詞的平均交互資訊(Average Mutual Information, AMI) :

$$AMI_i = \frac{1}{N}\sum_{j=1}^{N} MI(w_i, w_j) \tag{2-36}$$

雖然單獨使用上述的兩種較高階資訊的信心度評估沒有辦法比聲學方面的信心度評估獲得較好的效用[Jiang 2005]。但如果將此兩種高階資訊與事後機率相關的信心度評估做適當的結合，如線性插補法(Linear Interpolation)，可以獲得單獨只用事後機率相關的信心度評估更佳的效果[Guo *et al*. 2004]。

目前信心度評估除了用於驗證語音辨識結果的可信度之外，在進行詞彙樹複製搜尋的往前觀測(Look-ahead)，或是降低詞錯誤率(Word Error Rate, WER)也有相關的研究，在稍後的2.4及2.5兩小節將會一一說明。

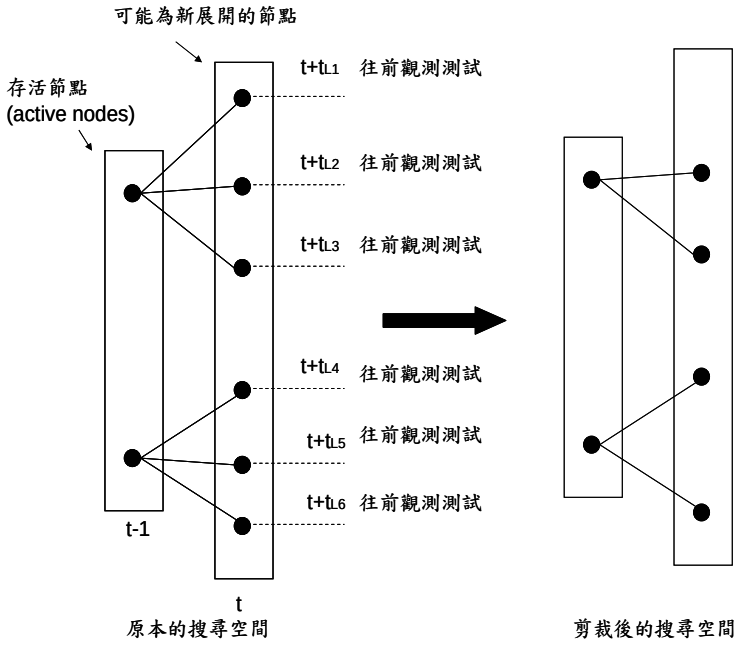

圖 2-8 往前觀測基本概念

2.4 信心度評估於詞彙樹複製搜尋之研究

在大詞彙連續語音辨識的研究領域中，要如何降低詞彙樹複製搜尋的搜尋空間(Search Space)及運算時間(Computation Time)是極為重要的課題。而往前觀測(Look-ahead)便是希望能降低運算時間卻又不影響語音辨識系統的正確率，其基本概念為在進行詞彙樹複製搜尋時，往前多看幾個音框或是事先看完整個語句，只有通過往前觀測測試的節點才會存活下來，如圖 2-8所示。在[Afify *et al*. 2005]中，作者提出了以驗證為基礎(Verification-based)的往前觀測法。作者將往前觀測當作是一個假設檢定(Hypothesis Testing)問題，針對某個音素 α 提出虛無假設(Null Hypothesis) H_0 及對立假設(Alternative Hypothesis) H_1:

$$H_0 \ : \ \alpha \ \text{ starts at time } \ t$$
$$H_1 \ : \ \alpha \ \text{ does not start at time } \ t$$

接下來便是採用兩元假設檢定(Binray Hypothesis Testing)，以決定要接受兩個假設中那一個。其檢定式如式(2-37)所示:

$$LRT = \frac{p(X \mid H_0)}{p(X \mid H_1)} \underset{H_1}{\overset{H_0}{\gtrless}} \tau \tag{2-37}$$

其中τ代表我們預先設定的門檻值(Threshold)，如果LRT的值大於τ，則採用虛無假設，也就是將音素α留下來。反之，則裁剪音素α。實作上，$p(X \mid H_0)$及$p(X \mid H_1)$的機率估測可以分別寫成$p(x_t^{t+d_\alpha} \mid \alpha)$及$p(x_t^{t+d_\alpha} \mid \bar{\alpha})$，其中$d_\alpha$代表要往前看多少時間(在這邊設定虛無假設及對立假設往前看的時間是一樣的)，而α及$\bar{\alpha}$分別代表音素對應的隱藏式馬可夫模型及反對應模型(Anti-model)，更詳細的資訊請參照[Afify *et al*. 2005]，另外在[Fabian *et al*. 2005]及[Abdou and Scordilis 2003]也都有探討信心度評估運用於往前觀測的相關探討。

2.5 信心度評估於降低詞錯誤率之研究

語音辨識最終的夢想便是希望能讓電腦可以像人一樣，在辨識人們所說的每一句話時，詞或字正確率能達到百分之百。因此，不論是研究聲學、語言模型或是強健性語音辨識等方面的研究學者，其實都是為了達到這個目標而努力。近年來，開始有研究學者也試著將信心度評估應用於詞圖搜尋或是*N*-最佳詞序列重新排序，以增進語音辨識的正確率。接下來的幾個小節將分別介紹搜尋最佳詞序列的基礎觀念，以及探討應用信心度評估於提昇語音辨識系統的正確率之相關研究。

2.5.1 最小化貝氏風險(Minimum Bayes Risk)

當我們將一段聲學觀測序列X辨識成某個詞序列W時，有時難免會產生辨識錯誤的情況(如2.1小節所提到會有詞的刪除、插入及替代等錯誤)。而一個好的語音辨識系統對X的辨識錯誤率當然是愈小愈好。更進一步而言，我們可以將辨識系統視為一個將X對應到詞序列的一個映射函式(Mapping Function)，或稱為分類器(Classifier):

$$F(X): X \rightarrow W_h^X \tag{2-38}$$

其中W_h^X代表一個假設空間(Hypothesis Space)，為詞典中所有詞的可能組合$\overline{\mathbf{W}}_\Sigma$的子集合(Subset)。另外，我們先定義一個成本函式(Cost Function)$\ell(W,W')$(W屬於$\overline{\mathbf{W}}_\Sigma$，而$W'$屬於$W_h^X$)，代表當將一個詞序列$W$的聲學觀測序列辨識成$W'$時的成本(為一個實數值)。此成本函式的定義通常與任務相關(Task-dependent)，如對語音辨識的領域來說，此函式通常定義成Levenshtein距離(此為語音辨識系統中判別辨識率好壞的評估標準，將會於3.2.3小節介紹)，或是與Levenshtein距離有關的函式。在估計此映射函式的期望錯誤時，通常都是利用貝氏風險來估算:

$$\sum_{W\in\overline{\mathbf{W}}_\Sigma} \ell(W,W')P(W\mid X) \tag{2-39}$$

有了風險評估函式之後，此映射函式$F(X)$對X的最佳辨識結果便是相當於在W_h^X此假設空間中選擇一條貝氏風險為最低的詞序列:

$$F^*(X)=\arg\min_{W'\in W_h^X}\sum_{W\in\overline{\mathbf{W}}_\Sigma} \ell(W,W')P(W\mid X) \tag{2-40}$$

2.5.2 最大事後機率與最小化貝氏風險尋找最佳詞序列之關聯

如果將$\ell(W,W')$定義成一個簡單的0/1對稱(Zero-one and Symmetric)成本函式:

$$\ell_{0/1}(W,W')=\begin{cases}1 & \text{if } W'\neq W\\ 0 & \text{otherwise}\end{cases} \tag{2-41}$$

則式(2-40)可以改寫成

$$\begin{aligned}F(X)&=\arg\min_{W'\in W_h^X}\sum_{W\in\overline{\mathbf{W}}_\Sigma}\ell_{0/1}(W,W')P(W\mid X)\\&=\arg\min_{W'\in W_h^X}\sum_{W\in\overline{\mathbf{W}}_\Sigma,W\neq W'}\ell_{0/1}(W,W')P(W\mid X)\\&=\arg\min_{W'\in W_h^X}1-P(W'\mid X)\\&=\arg\max_{W'\in W_h^X}P(W'\mid X)\end{aligned} \tag{2-42}$$

也就變成我們目前一般使用的最大事後機率尋找最佳詞序列的方法。

2.5.3 事後機率詞圖搜尋

在[Wessel *et al*. 2000]中，作者提出了使用詞圖計算出來的事後機率來增加辨識結果的正確率。假設我們有了詞圖 Ψ^X 之後，在實作最大事後機率辨識法則時，會以 Ψ^X 取代假設空間 W_h^X 。另外，在產生詞圖之後，我們便有了每個詞段的詞編號及對應的開始和結束時間。因此，(2-42)式便可以改寫成式(2-43):

$$\begin{aligned} F(X) &= \arg \max_{[w^n;s^n,e^n]_{n=1}^N \in \Psi^X} P([w^n;s^n,e^n]_{n=1}^N \mid \Psi^X) \\ &= \arg \max_{[w^n;s^n,e^n]_{n=1}^N \in \Psi^X} \prod_{n=1}^M P([w^n;s^n,e^n] \mid [w^n;s^n,e^n]_{n=1}^{n-1}, \Psi^X) \qquad (2\text{-}43) \\ &\approx \arg \max_{[w^n;s^n,e^n]_{n=1}^N \in \Psi^X} \prod_{n=1}^M P([w^n;s^n,e^n] \mid \Psi^X) \end{aligned}$$

其中 $P([w^n;s^n,e^n] \mid \Psi^X)$ 也就是2.2.2小節中所介紹的詞段之事後機率。

2.5.4 最小化音框錯誤率詞圖搜尋

當我們將貝氏風險中的成本函式訂為0/1函式後，便可以將最大事後機率的辨識方法視為找出一條正確率為最高的詞序列(因為詞序列的事後機率也可想像為詞序列正確的機率)。但是由於一般是以詞錯誤率(Word Error Rate)為語音辨識系統好壞的評估標準，因此會有成本函式與估評標準不匹配的情況出現，在[Mangu *et al*. 2000]中便有提出最小化詞序列錯誤率並不等於同時最小化詞錯誤率的例子。要解決這個不匹配的問題，最簡單的方法便是將成本函式改為與評估標準一致。雖然這樣本質上解決了不匹配的問題，但是由於計算詞錯誤率時，必須計算兩個詞序列的Levenshtein距離(也就是統計兩個詞序列中的取代、刪除及插入的次數)。如果要在詞圖中眾多可能的詞序列中，針對任兩個詞序列都要計算Levenshtein距離的話，其組合有太多種，計算複雜度便成了這個方法的主要缺點。為了解決此計算複雜度所造成的缺點，[Stolcke *et al*. 1997]中將此兩兩比對(Pairwise Alignment)限制在*N*-最佳詞序列範圍內。幾年後，[Mangu *et al*. 2000]將原本比對的對象再度從*N*-最佳詞序列改成詞圖。但是考量前面所提到的高計算複雜度缺點，[Mangu *et al*. 2000]試著將詞圖原本包含

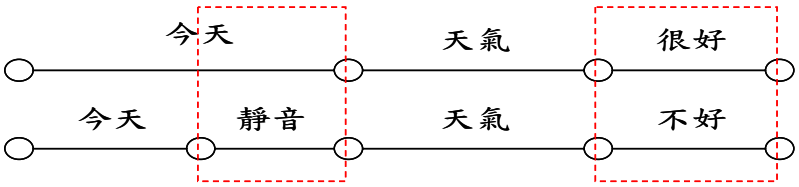

圖 2-9 音框錯誤率圖解

不同長度的詞序列改成長度一模一樣，一般稱為一致性網路(Consensus Network or Sausage)。因此只要執行複雜度較低的多重字串比對(Multiple String Alignment)。

字串比對中的刪除及插入錯誤是造成兩兩比對的複雜度會如此高的原因。當字串比對只剩下取代錯誤時，原本動態規畫(Dynamic Programming)比對法的Levenshtein距離成本函式之計算複雜度將大幅下降。[Wessel *et al.* 2001]便是以此想法為出發點，提出了新的成本函式-音框錯誤率(Time Frame Error)。

由於詞圖擁有每個詞的開始及結束時間資訊，當我們比對詞圖上任意的兩條詞序列時，如圖 2-9所示，虛線框框的地方便是代表有音框錯誤。從圖 2-9可以看出，原本字串比對時會存在的詞刪除、插入或替代錯誤，現在都已可以用音框錯誤來取代。基於這個觀念，便可以定義一個新的不對稱成本函式，稱為音框錯誤率[Wessel *et al.* 2001] :

$$\ell([w^n;s^n,e^n]_{n=1}^{N},[w^m;s^m,e^m]_{m=1}^{M})=\sum_{n=1}^{N}\frac{\sum_{t=s^n}^{e^n}1-\delta(w^n,w_t^m)}{1+\alpha(e^n-s^n)} \tag{2-44}$$

其中$[w^n;s^n,e^n]_{n=1}^{N}$及$[w^m;s^m,e^m]_{m=1}^{M}$分別代表詞圖上的兩條完整路徑，其詞的個數分別為N及M，而

$$\delta(w^n,w_t^m)=\begin{cases}1, & \text{if } w^n=w_t^m\\ 0, & \text{if } w^n\neq w_t^m\end{cases} \tag{2-45}$$

w_t^m代表在t這個音框時w^m的詞編號。α則是決定是否要做正規化(Normalization)的參數。當α為1時，代表採取正規化的動作，換言之此辨識方法有長詞優先的傾向。現在，將音框錯誤率成本函式代入最小化貝氏風險的準則，則式(2-40)可以改寫為

$$
\begin{aligned}
F^{*}(X) &= \arg\min_{[w^n;s^n,e^n]_{n=1}^N} \left\{ \sum_{[w^m;s^m,e^m]_{m=1}^M} \ell([w^n;s^n,e^n]_{n=1}^N,[v^m;s^m,e^m]_{m=1}^M) P([v^m;s^m,e^m]_{m=1}^M \mid \Psi^X) \right\} \\
&= \arg\min_{[w^n;s^n,e^n]_{n=1}^N} \left\{ \sum_{[w^m;s^m,e^m]_{m=1}^M} \sum_{n=1}^N \frac{1-\delta(w^n,w_t^m)}{1+\alpha(e^n-s^n)} P([v^m;s^m,e^m]_{m=1}^M \mid \Psi^X) \right\} \\
&= \arg\min_{[w^n;s^n,e^n]_{n=1}^N} \left\{ \sum_{n=1}^N \frac{\sum_{t=s_n}^{t=e_n} \left[1 - \sum_{[w^m;s^m,e^m]_{m=1}^M} \delta(w^n,w_t^m) P([v^m;s^m,e^m]_{m=1}^M \mid \Psi^X \right])}{1+\alpha(e^n-s^n)} \right\} \\
&= \arg\min_{[w^n;s^n,e^n]_{n=1}^N} \left\{ \sum_{n=1}^N \frac{\sum_{t=s_n}^{t=e_n} \left[1 - \sum_{[w^m;s^m,s^m]_{m=1}^M} \sum_{w^m:s^m\le t\le e^m} \delta(w^n,w_t^m) P([v^m;s^m,e^m]_{m=1}^M \mid \Psi^X) \right]}{1+\alpha(e^n-s^n)} \right\} \\
&= \arg\min_{[w^n;s^n,e^n]_{n=1}^N} \left\{ \sum_{n=1}^N \frac{\sum_{t=s_n}^{t=e_n} \left[1 - \sum_{[w^m;s^m,s^m]:s^m\le t\le e^m} \delta(w^n,w_t^m) P([v^m;s^m,e^m] \mid \Psi^X) \right]}{1+\alpha(e^n-s^n)} \right\}
\end{aligned}
\tag{2-46}
$$

其中我們將 $\sum_{[w^m;s^m,e^m]:s^m\le t\le e^m} \delta(w^n,w_t^m) P([w^m;s^m,e^m] \mid \Psi^X)$ 表示成 $P(w^n \mid t,\Psi^X)$，可以視為在音框 t 時，通過 w^n 的機率為多少，也就相當於 w^n 在音框 t 時正確的機率。

第3章 實驗架構

在本章中將先介紹臺師大大詞彙連續語音辨識系統[Chen *et al*. 2004; 2005]。接著介紹及分析本論文所使用的公視晚間新聞(MATBN)外場記者及外場受訪者語料。最後則是介紹實驗的評估方式。

3.1 臺師大大詞彙連續語音辨識系統

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

3.1.1 前端處理

本系統支援梅爾倒頻譜係數(Mel-frequency Cepstral Coefficients, MFCC)以及異質性線性鑑別分析(Heteroscedastic Linear Discriminant Analysis, HLDA)加上最大相似度線性轉換(Maximum Likelihood Linear Transformation, MLLT)[Gopinath 1998; Saon *et al*. 2000]兩個不同的語音特徵參數。在本論文中，我們先比較兩種語音特徵參數對語音辨識系統的正確率影響，根據實驗結果，論文後半部主要使用異質性線性鑑別分析配合最大相似度線性轉換做為語音特徵參數。

3.1.2 聲學模型

本系統使用1.1.2小節所介紹的112個聲母(INITIALs)，38個韻母(FINALs)及1個靜音(Silence)共151個連續密度隱藏式馬可夫模型(Continuous Density Hidden Markov Models,CDHMMs)。每個模型的狀態有3至6個不等，每個狀態皆為高斯混合分佈，其中每個高斯混合分佈的個數分別為1至128個不等。此外，這些聲母和韻母共組合成403個不同的基本音節(Base Syllables)。

3.1.3 詞典建立及語音模型訓練

本系統所使用的詞典是先將大量的文字語料經由一個含有一至四字詞約六萬八千個詞的詞典來斷詞，配合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)。新增複合詞的方式則如下所述:對於語料中任意相鄰的兩個詞(w_i, w_j)，分別計算它們的前雙連(Forward Bigram)機率$P_f(w_j \mid w_i)$與後雙連(Backward Bigram)機率$P_b(w_i \mid w_j)$，再以前後雙連(Forward and Backward Bigrams)的機率，求其幾何平均(Geometric Average)$FB(w_i, w_j) = \sqrt{P_f(w_j \mid w_i) P_b(w_i \mid w_j)}$，作為$(w_i, w_j)$是否合併的依據。根據上述的公式，經數次迭代(Iteration)以及不同的門檻值(Threshold)設定，產生約五千個二至十字詞的複合詞，使得最後的語音辨識詞典約有七萬二千個一至十字詞。

本系統使用詞雙連(Bigram)及詞三連(Trigram)語言模型，外場記者語料的部份是從中央通訊社(Central News Agency, CNA)在2001與2002年間收集到的約一億七千萬個中文字語料作為背景語言模型(Background Language Model)的訓練資料[LDC]。而在外場受訪者的部份，由於此語料具有偏口語對話(Spontaneous Speech)的特性，較容易有不流暢(Disfluency)或有語助詞的情況發生，因此，除了上述中央通訊社語料之外，我們另外從曾淑娟博士的漢語連續口語對話語音語料庫(Mandarin Conversational Dialogue Corpus, MCDC)[Tseng and Liu 2001]擷取一些可用的語句文本加上外場受訪者聲學模型訓練語料的文字檔，作為相同領域(In-domain)的語言模型訓練語料。本論文中的語言模型使用Katz語言模型平滑技術，語言模型訓練工具採用SRI Language Modeling Toolkit (SRILM) [SRILM 2000]。

3.1.4 詞彙樹複製搜尋

本系統是採用由左至右(Left-to-right)且音框同步(Frame Synchronous)的詞彙樹複製搜尋方式[Aubert 2002]。詞彙樹的架構如圖 3-1所示，樹中的每個分枝(Arc)代表一個聲母(INITIAL)、韻母(FINAL)或靜音(Silence)模型。由樹的根節點(Root Node，圖 3-1的方型實心點)走到樹的葉節點(Leaf Node，圖 3-1的圓形實心點)的某一條完整路徑代表走完一個或一組發音相同的詞。而路徑上的每一個分枝正好對應到這些詞的一組聲學模型。詞彙樹複製搜尋在執行時，每個音框會同時存在數棵詞彙樹複製(Tree Copies)，而每棵詞彙樹代表來自不同的語言歷史或限制(Language Model History or Constraint)。在同一棵詞彙樹裡，會進行隱藏式馬可夫模型狀態層次(State Level)維特比(Viterbi)動態規劃搜尋。在詞彙樹搜尋中，只有在走到葉節點時，才能確定所搜尋的一個完整詞為何。另外，當具有相同語言模型歷史之不同詞彙樹分別都已經走到自己所屬那棵樹的葉節點時，則會進行結合(Recombination)，只保留其中分數最大者，並針對留下來的詞彙樹繼續執行詞彙樹複製搜尋。然而，真正在實作時，並不需要產生如此多的詞彙樹，僅需建立一棵詞彙樹作為參考之用，並分別記錄搜尋時存活下來之隱藏式馬可夫模型狀態節點的相關資訊(如到目前為此所累積的分數及前一狀態為何)。另外一方面，由於存活的狀態節點通常會隨著音框數呈指數倍成長，因而必須以光束剪裁(Beam Pruning)技術將分數較低的狀態節點做剪裁的動作。在對每個狀態節點執行光束剪裁時，會依此節點所有可拜訪的葉節點之最大單連語言模型往前觀測分數(Unigram Language Model Look-ahead Score)[Aubert 2002]及聲學往前觀測分數(Acoustic Look-ahead Score)[Chen *et al*. 2004; 2005]做為剪裁與否的依據。此外，在每個音框，利用存活的詞彙樹複製樹其葉節點(代表可能的候選詞)所儲存的語言模型歷史、開始音框、結束音框及其聲學解碼的分數等資訊，建立如2.2.1小節所提到的詞圖。而後使用更高階的語言模型，如詞三連或詞四連(Fourgram)語言模型，抑或採用更複雜的聲學模型，如三連音素(Triphone)，進行詞圖搜尋[Ortmanns *et al*. 1997]，找出最佳的詞序列。在本論文中，詞彙樹複製搜尋階段是採用詞雙連語言模型，詞圖搜尋階段則是使用詞三連語言模型。

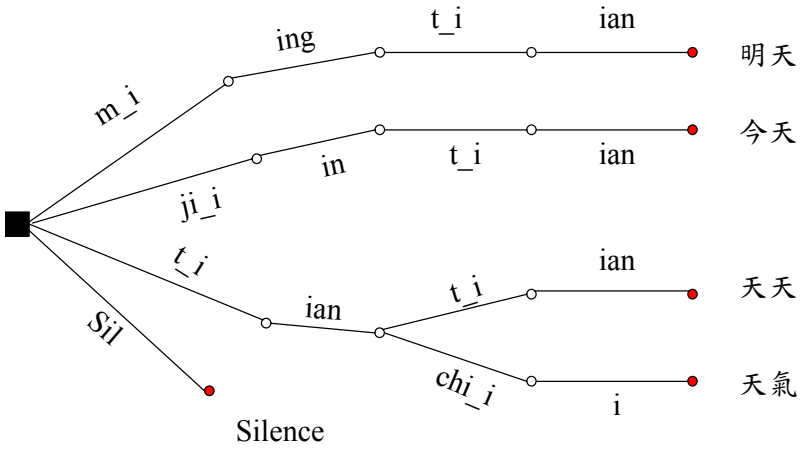

圖 3-1 詞彙樹範例

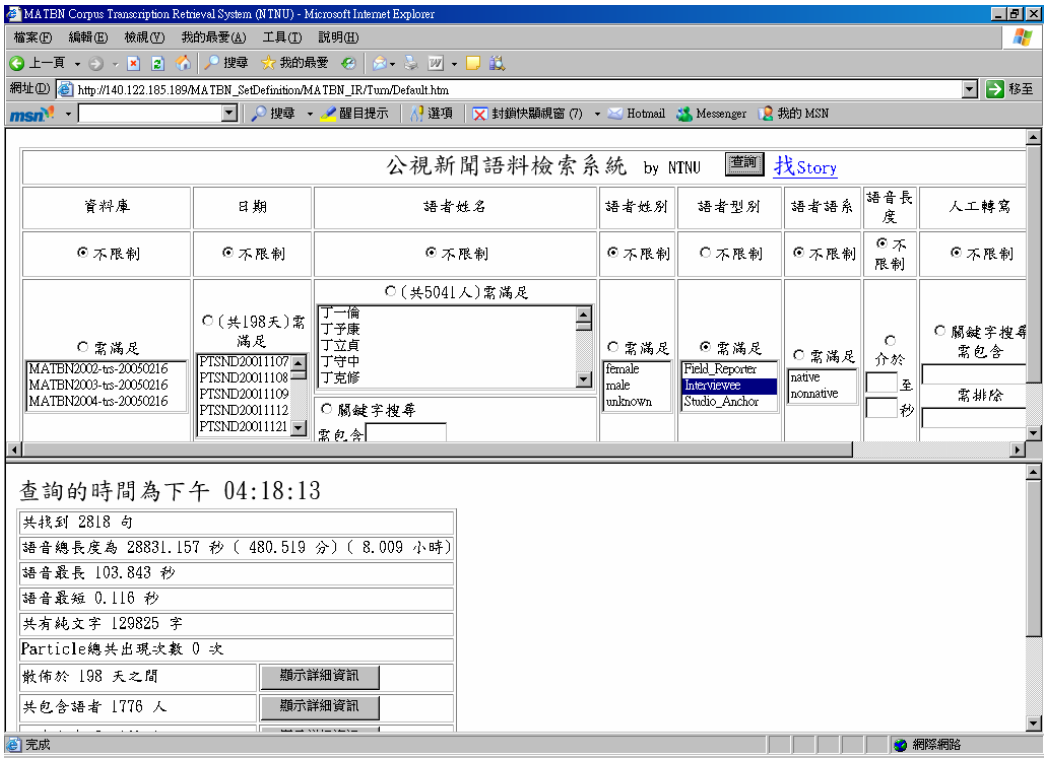

圖 3-2 臺師大資工所公視新聞語料檢索系統，檢索語句(Sub-term)的統計資訊

3.2 實驗語料

本論文所使用的兩套語料庫皆取材自公視新聞語料庫(MATBN)[Wang *et al*. 2005]，此語料庫為中央研究院資訊所中文資訊處理實驗室口語小組[SLG]耗時三年與公共電視台[PTS]合作錄製完成。錄製的對象為公視晚間新聞，其每天的長度皆為一個小

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

表 3-1 主播語料分佈表

時，收錄了 198 天的新聞語料，其中包含 2001 年的新聞 30 小時、2002 年 146 小時及 2003 年 24 小時。所有的新聞語料都有詳細的人工轉寫以及其它的標註資訊(如：音樂、背景雜訊、停頓、語助詞、呼吸、強調語氣、反覆及不適當的發音等)，所有的人工轉寫與標註均使用 DGA&LDC 的轉寫器(Transcriber)[Barras *et al.* 1986]來完成。

每天的公視晚間新聞約含有二十多則報導，每則報導為一完整主題。除了語音資料，文字語料在其它應用，如資訊檢索(Information Retrieval)、文件摘要(Document Summarization)也提供了很好的實驗平台。此新聞語料大致上可分內場及外場兩個部份，內場部分主要為攝影棚內場主播(Studio Anchors)的語料，外場部分則可分為採訪記者(Field Reporters)與受訪者(Interviewees)的語料。在篩選實驗語料時，考量新聞的特性，主播多為同一人所擔任，如表 3-1所示，葉明蘭主播的語料在本語料庫中約佔了所有主播語料的84%，這將使得實驗偏向語者相依(Speaker-dependent)的環境，加上女性主播約佔了所有主播語料的94%，也造成了性別相依(Gender-dependent)的問題，如果使用主播語料的話，可能無法提供聲學模型良好的訓練與客觀的評估。故本實驗不採用主播語料，而是採用外場記者與受訪者做為實驗的語料。在選取語

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

表 3-2 外場記者訓練與評估語料分佈表

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

表 3-3 語助詞出現次數統計表

料的工具選擇方面，我們是採用臺師大資工所語音實驗室針對MATBN電視新聞語料所開發的語料資訊檢索系統[NTNU 2004]，如圖 3-2所示。此系統可檢索語句的統計資訊，如語者資訊、語音長度、所含背景雜訊、說話速度及正確轉譯文句等資訊，適合用來分析且定義出實驗的訓練集(Training Set)與評估集(Evaluation Set)。

3.2.1 外場記者語料

外場記者語料指的是採訪記者的語料，共包含25.5小時的訓練集(5774句，再切成34,964個短句供聲學模型訓練之用)和1.5小時的評估集(292句，供評估語音辨識系統正確率之用)。其中男女語料大約各半，詳細的資訊如表 3-2所示。訓練集選自2001和2002兩年的新聞語料，而評估集選自該語料庫設定的評估語料[Wang *et al.* 2005]，但濾掉了含有語助詞的語句。更詳細的資訊可參考[郭人瑋 2005]。

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

表 3-4 外場受訪者訓練與評估語料分佈表

3.2.2 外場受訪者語料

由於受訪者語料跟內場主播及外場記者比較起來，如表 3-3 所示，其語音資料包含了許多的語助詞。如果不做一些前處理的動作，而直接將所有包含語助詞的語音資訊濾除，再加上考慮男女語料平衡的因素，堪用的訓練語料大概共只有 235 分鐘，要用來訓練聲學模型可能有所不足。因此，為解決此項問題，我們將人工轉寫的文字檔中一般常見的語助詞符號轉為中文字(例如"MA"轉為"嗎")，進而獲得更多的訓練及評估語料。我們最後收集了約 530 分鐘的訓練集(2,002 句，後來切割 9,764 個較短的語句)以及約 36 分鐘(196 句)的評估集。其詳細資料如表 3-4 所示。外場受訪者的訓練語料也是選自 2001 及 2002 年的語料，而評估集則選自該語料庫設定的評估語料[Wang *et al.* 2005]。

3.2.3 實驗評估方式

本論文針對信心度評估及辨識系統正確率各有一套評估標準，以下將會分別介紹。

(i) 信心度評估:

當估算出每個詞的信心度後，每個詞將會根據其信心度的值是否大於或小於事先設定好的門檻值而標注為正確(Correct)或錯誤(False)。在進行信心度評估時，通常會發生兩種錯誤，一類是錯誤接受(False Acceptance);也就是辨識錯誤的詞被標注為正確。另一類是錯誤拒絕(False Rejection);也就是辨識正確的詞被標注為錯誤。如果事先設定的門檻值太高的話，通常會降低錯誤接受的次數，但錯誤拒絕的次數反而會增加;而事先設定的門檻值太低的話，則相反。因此，這兩類的錯誤會因事先設定的

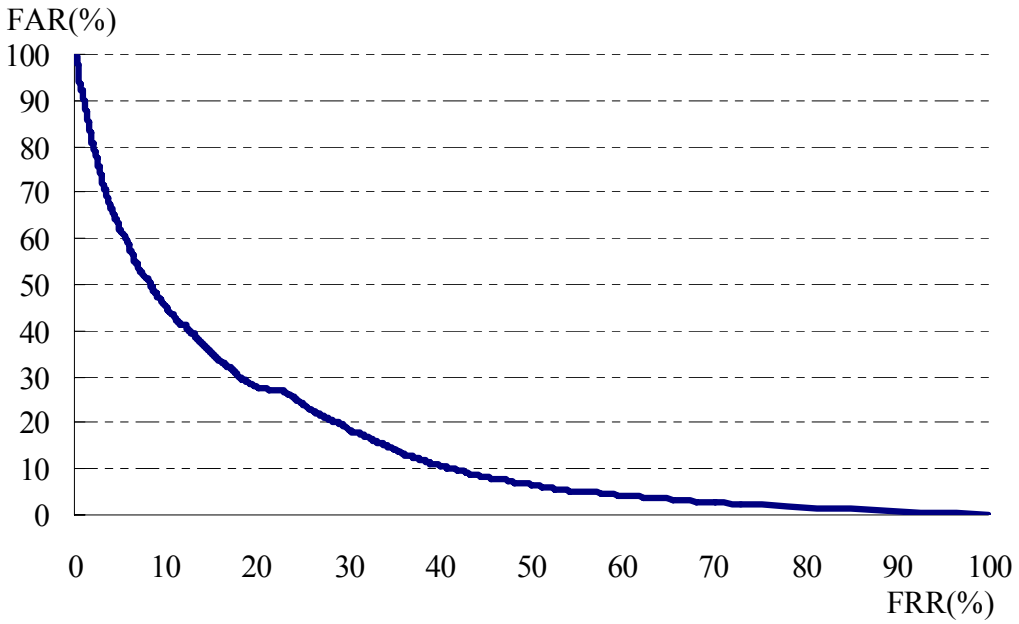

圖 3-3 偵測錯誤交易曲線圖範例

門檻值高低不同而有所權衡取捨(Trade-off)。

本論文採用的第一個的評估標準使用信心度錯誤率(Confidence Error Rate)，其定義如式(3-1)所示:

$$\text{信心度錯誤率} = \frac{\text{錯誤拒絕(False rejection)個數} + \text{錯誤接受(False Acceptance)個數}}{\text{辨識詞總個數}} \tag{3-1}$$

而信心度錯誤率的基礎實驗結果(Baseline)則是定義如下 ：

$$\frac{\text{插入(Insertion)個數} + \text{替代(Substitution)個數}}{\text{辨識詞總個數}} \tag{3-2}$$

從式(3-1)的定義可以發現，信心度錯誤率的大小會直接受到事先設定好的門檻值影響。因此，在實作時，此事先設定的門檻值通常是額外使用一套驗證語料。在接下來的實驗中，我們各從外場記者及受訪者聲學模型訓練語料中隨機抽取1,000句(約0.74小時)及500句(約0.45小時)當作驗證語料，使得事先設定好的門檻值在此驗證語料有最低的信心度錯誤率。

本論文採用的第二項評估標準則是偵測錯誤交易曲線圖(Detection-error-tradeoff Curve, DET Curve)，偵測錯誤交易曲線圖是針對不同的門檻值而可以劃出相對應的錯誤接受率(False Acception Rate, FAR)及錯誤拒絕率(False Rejection Rate, FRR)(縱軸為錯誤接受率;橫軸為錯誤拒絕率)，如圖 3-3所示。而錯誤接受率及錯誤拒絕率的算法分別如式(3-3)及式(3-4)所示:

$$\text{錯誤接受率} = \frac{\text{錯誤接受(False Acceptance)個數}}{\text{辨識錯誤的詞個數}} \tag{3-3}$$

$$\text{錯誤拒絕率} = \frac{\text{錯誤拒絕(False Rejection)個數}}{\text{辨識正確的詞個數}} \tag{3-4}$$

(ii) 辨識系統正確率:

此評估法則是採用美國國家標準與技術中心(National Institute of Standards and Technology, NIST)[NIST]所訂立的評估標準來進行正確答案的詞序列與辨識詞序列的比較。此評估標準需要使用動態規畫(Dynamic Programming)來做詞序列比對(也就是2.5.4小節所提到的Levenshtein距離)。由於在中文會有斷詞不一致的問題，因此在本論文的實驗中主要是以字為比對單位。令 H 為正確答案詞序列與辨識詞序列比對後相同(Match)的字的個數、I 為辨識詞序列多餘插入(Insertion)的字的個數、N 為正確答案詞序列的字的個數，則語音辨識系統的正確率(Accuracy)的計算方式為 $\frac{H-I}{N}\times 100\%$ ，而錯誤率(Error Rate)則為1-正確率。在進行動態規畫比對時，替代(Substituion)錯誤的懲罰權重(Penalty Weight)為10分，插入及刪除的權重則皆為7分。

第4章 基礎實驗討論

本章實驗包含了三小節。4.1小節為建立外場受訪者聲學模型之實驗，此小節除了比較梅爾倒頻譜係數(MFCC)以及異質性線性鑑別分析加上最大相似度線性轉換(HLDA+MLLT)兩種語音特徵參數對語音辨識系統正確率的好壞之外，同時也探討最大化相似度訓練以及最小化音素錯誤訓練對於聲學模型訓練的影響。最後，則是嘗試在語言模型方面加入相同領域的語言模型訓練語料(請參照3.1.3小節)，來降低大詞彙連續語音辨識系統的錯誤率。4.2小節針對傳統信心度評估方法進行實驗討論;4.3小節則是討論關於前人應用信心度評估於降低詞圖搜尋錯誤率之實驗。

4.1 外場受訪者基礎實驗

4.1.1 最大化相似度(Maximum Likelihood, ML)訓練之實驗

此實驗的初始聲學模型之狀態高斯混合機率分佈均視為平均值等於0、標準差為1的標準常態分佈(Standard Normal Distribution)，利用HTK Tookit[Young *et al*. 2002]內建函數，根據MFCC和HLDA+MLLT兩種不同的語音特徵參數，各進行30次最大化相似度訓練。每間隔5次最大化相似度訓練後之聲學模型對於MATBN外場受訪者測試語料的自由音節辨識(Free Syllable Decoding)之錯誤率結果請參考表 4-1。而30次訓練之自由音節辨識錯誤率曲線請參考圖 4-1。接著，我們採用第30次的聲學模型作為大詞彙連續語音辨識系統的初始聲學模型，在執行詞彙樹複製搜尋及詞圖搜尋時，分別調整語言模型分數的權重，如式(4-1)中的β:

$$p(X \mid W)P(W)^{\beta} \tag{4-1}$$

觀察其對語音辨識系統錯誤率的影響。詞彙樹複製搜尋的結果可參考表 4-2及圖 4-2;而詞圖搜尋的結果可參考表 4-3及圖 4-3。

訓練次數	MFCC	HLDA+MLLT
5	67.89	66.03
10	67.40	65.56
15	67.02	65.44
20	67.11	65.34
25	67.04	65.23
30	66.80	65.27

表 4-1 外場受訪者:30 次最大化相似度訓練，每間隔 5 次之自由音節辨識錯誤率(%)

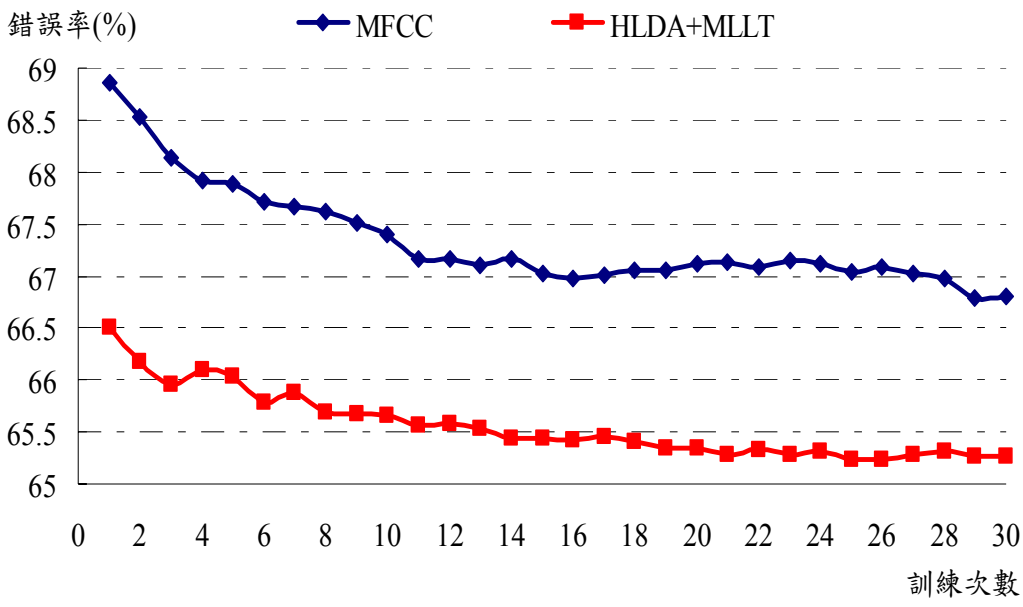

圖 4-1 外場受訪者:30 次最大化相似度訓練之自由音節辨識音節錯誤率曲線圖

語言模型權重	MFCC	HLDA+MLLT
5	62.60	57.70
6	61.65	57.11
7	61.60	56.74
8	62.00	56.85
9	62.17	57.30
10	62.42	57.52
11	63.20	58.13
12	63.47	58.70

表 4-2 外場受訪者:不同的語言模型權重，經詞彙樹複製搜尋後之字錯誤率(%)

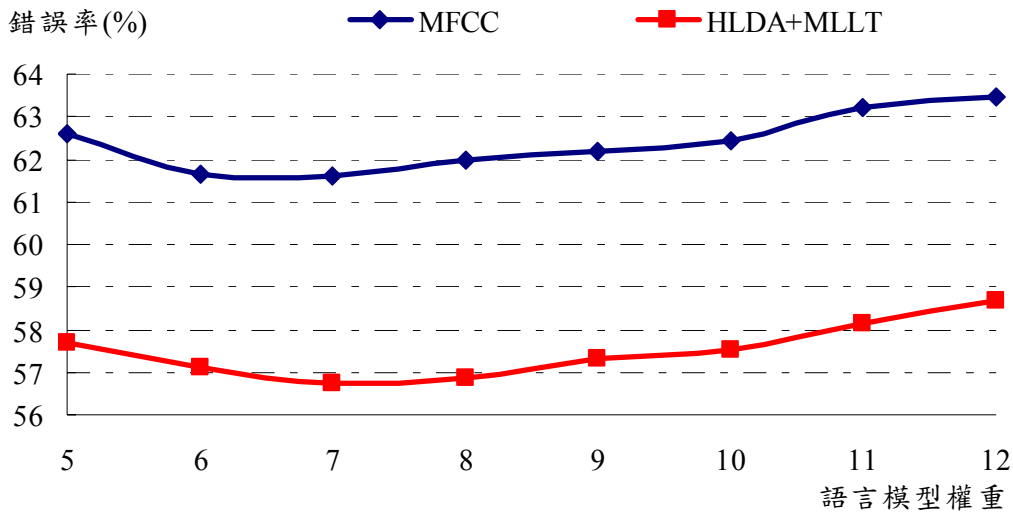

圖 4-2 外場受訪者:不同的語言模型權重，經詞彙樹複製搜尋後之字錯誤率曲線圖

語言模型權重	MFCC	HLDA+MLLT
5	60.73	56.11
6	59.71	55.21
7	59.57	55.50
8	60.14	55.75
9	60.35	55.69
10	60.73	55.77
11	61.09	56.40
12	61.20	57.00

表 4-3 外場受訪者:不同的語言模型權重，經詞圖搜尋後字錯誤率(%)

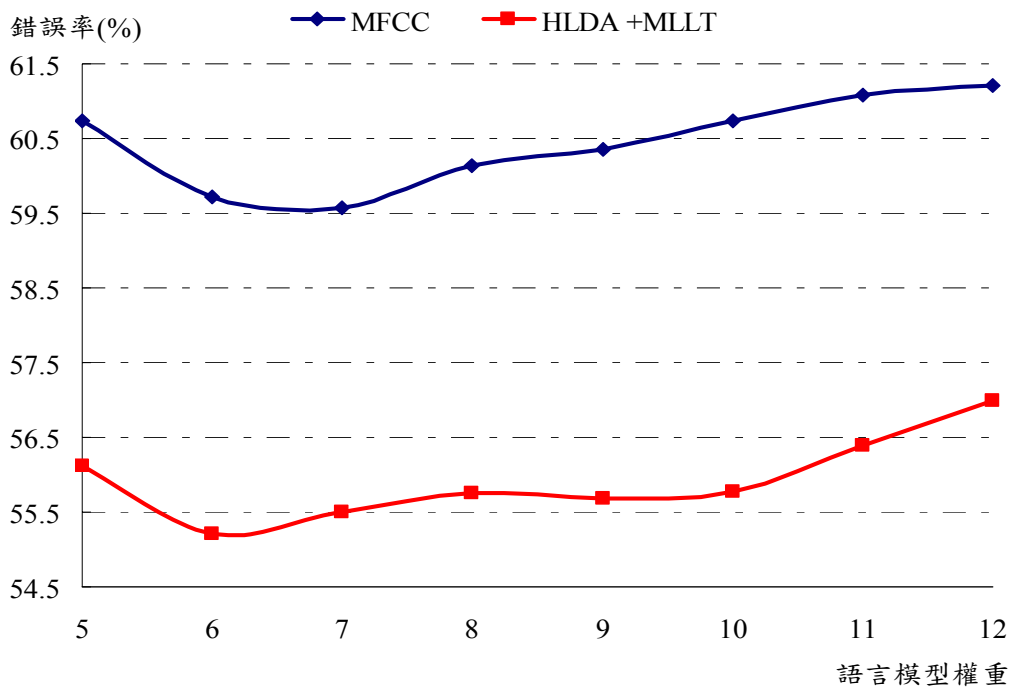

圖 4-3 外場受訪者:不同的語言模型權重，經詞圖搜尋後之字錯誤率曲線圖

【實驗討論】

在自由音節辨識實驗中，不論採用是何種語音特徵參數，最大化相似度訓練大概都在 15 次訓練次數之後呈現飽和(Saturation)。15 次之後的訓練，其錯誤率曲線會呈現上下振盪的情況，辨識錯誤率已無法有進一步明顯的下降。

在調整語言模型權重的實驗中，不論是採用詞彙樹複製搜尋或詞圖搜尋、在何種語音特徵參數，其錯誤率最低的語言模型權重皆約為 6 或 7 左右，顯示背景語言模型跟外場受訪者測試語料的領域(Domain)似乎有些不同。此外，詞圖搜尋由於用到更高階的語言模型，因此其錯誤率能比詞彙樹複製搜尋更降低一些。在往後的實驗，外場受訪者的詞彙樹複製搜尋語言模型權重將固定設 7，詞圖搜尋的權重則將設為 6。

4.1.2 最小化音素錯誤(Minimum Phone Error, MPE)訓練之實驗

在上一小節的實驗中，由於最大化相似度訓練很快就達到了飽和狀度。因此，接下來我們便再對經過30次最大化相似度訓練後的聲學模型進行10次的最小化音素錯誤訓練[Povey 2004]，觀察能否對聲學模型有所幫助。實驗結果可參照表 4-4，而對應的自由音節辨識錯誤率及字錯誤率曲線請參照圖 4-4及圖 4-5。

	自由音節辨識錯誤率		詞彙樹複製搜尋字錯誤率	
訓練次數	MFCC	HLDA+MLLT	MFCC	HLDA+MLLT
1	66.31	64.64	60.57	55.45
2	66.30	64.56	59.29	54.98
3	66.69	64.31	58.55	54.86
4	66.90	64.45	58.32	54.55
5	67.02	64.54	58.19	54.26
6	67.27	64.56	57.89	54.03
7	67.33	64.65	58.02	53.90
8	67.33	64.81	57.97	53.87
9	67.33	65.05	58.17	53.89
10	67.58	65.10	58.09	54.24

表 4-4 外場受訪者:10 次最小化音素錯誤訓練之音節與字錯誤率(%)

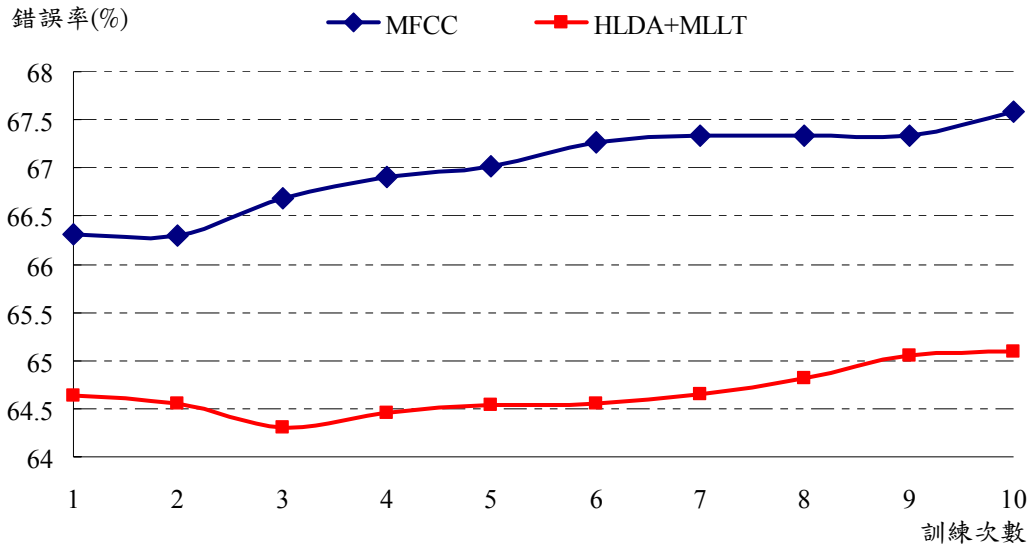

圖 4-4 外場受訪者:10 次最小化音素錯誤訓練之音節錯誤率曲線圖

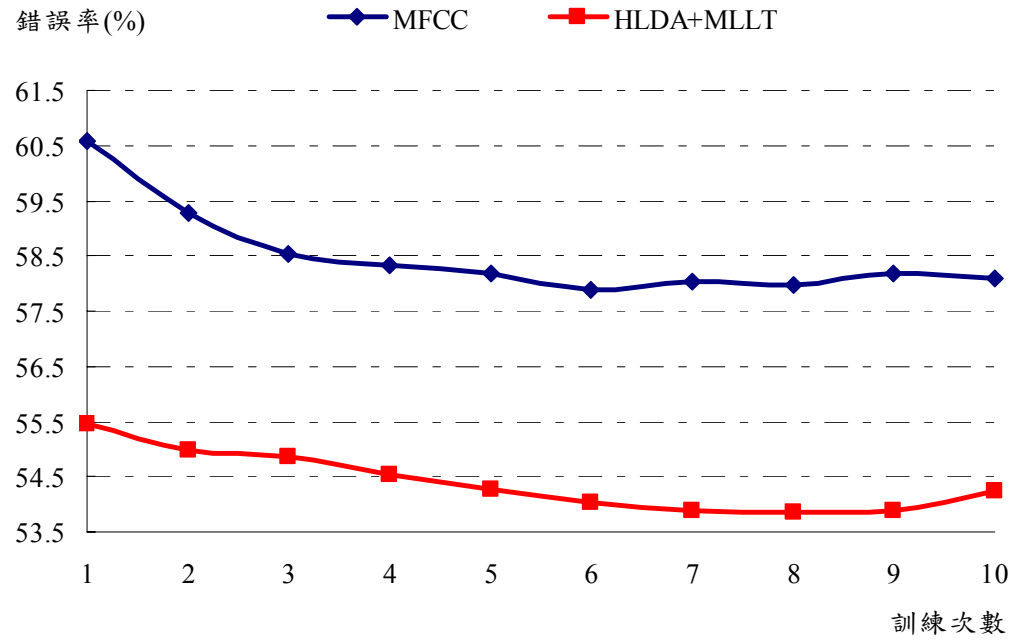

圖 4-5 外場受訪者:10 次最小化音素錯誤訓練之詞彙樹複製搜尋之字錯誤率曲線圖

【實驗討論】

在自由音節辨識實驗中，最小化音素錯誤訓練法則其作用似乎比較不明顯(其音節錯誤率約降低0.5%~1%左右)，而且很快就因為過度訓練(Over-training)而帶來錯誤率的上昇，對於辨識率較低的初始聲學模型尤其顯著，其原因可能出自辨識率較低的聲學模型無法提供較多的鑑別資訊。

而在詞彙樹複製搜尋方面的錯誤率似乎就較無自由音節辨識的問題。大致上來說，不論是何種語音特徵參數，都有2%~3%的字錯誤率絕對下降(Absolute Reduction)。可能是因為最小化音素錯誤訓練法則除了在訓練聲學模型本身會有效果之外，也會讓聲學模型在訓練時同時考慮到與語言模型結合的影響，而使得字錯誤率有較佳的結果。

在以上兩小節的實驗中，不論對最大化相似度或最小化音素錯誤訓練來說，HLDA+MLLT的語音特徵參數在自由音節辨識的音節錯誤率或詞彙樹複製搜尋的字錯誤率都較MFCC的語音特徵參數有明顯的下降，因此有關於外場受訪者實驗中，我們將採用HLDA+MLLT的語音特徵參數。另外，使用HLDA+MLLT語音特徵參數的最小化音素錯誤訓練聲學模型於詞圖搜尋的錯誤率(WG:CHAR表示)如表 4-5所示。由於在經第8次訓練後的字錯誤率最小，因此本論文最後採用第8次最小化音素錯誤訓練的聲學模型為外場受訪者的聲學模型，而在外場記者語料部份，則是之前的實驗，採用經由150次最大化相似度訓練及10次最小化音素錯誤訓練後的聲學模型做為記者語料的聲學模型, 而語音特徵參數也是使用HLDA+MLLT。

訓練次數	1	2	3	4	5
WG:CHAR	54.92	54.37	53.85	53.21	52.79
訓練次數	6	7	8	9	10
WG:CHAR	52.38	52.30	52.29	52.47	52.55

表 4-5 外場受訪者:10 次最小化音素錯誤率訓練詞圖搜尋之字錯誤率(%，語音特徵參數為 HLDA+MLLT)

4.1.3 相同領域與背景語言模型線性插補實驗

由於外場受訪者是屬於偏即性口語對話語料，單靠中央通訊社的新聞背景語料似乎有所不足，因此，本論文額外使用了外場受訪者聲學訓練語料的人工轉寫文字檔共2,002句及從漢語連續口語對話語音語料庫抽出的1,791句，合併成3,793句的相同領域語言模型訓練資料。將得到的相同領域語言模型與中央通訊社背景語言模型做線性插補，而線性插補的公式如式(4-2)所示:

$$P(W) = \alpha \cdot P_{BG}(W) + (1-\alpha)P_{InDmain}(W) \tag{4-2}$$

其中α代表背景語言模型$P_{BG}(W)$的權重(其值介於0~1之間)，$P_{InDomain}(W)$則是代表相同領域語言模型的分數。觀察對詞圖搜尋錯誤率的的影響，其實驗結果可參考表4-6。其中MATBN_IV_LM代表單單只使用外場受訪者聲學訓練語料的文字檔進行線性插補，而MCDC_MATBN_IV_LM則代表將外場受訪者聲學模型訓練語料的文字檔與漢語連續口語對話語音語料庫合併後進行線性插補。而詞圖搜尋字錯誤率曲線圖可參考圖 4-6。

插補權重α	MABN_IV_LM	MCDC_MATBN_IV_LM
0.1	50.91	50.85
0.2	50.41	50.36
0.3	50.35	50.16
0.4	49.72	50.71
0.5	49.72	49.60
0.6	49.73	49.69
0.7	49.75	49.56
0.8	50.06	50.21
0.9	50.48	50.38
1	52.40	52.19

表 4-6 相同領域語言模型與背景語言模型做線性搜補之詞圖搜尋字錯誤率(%)

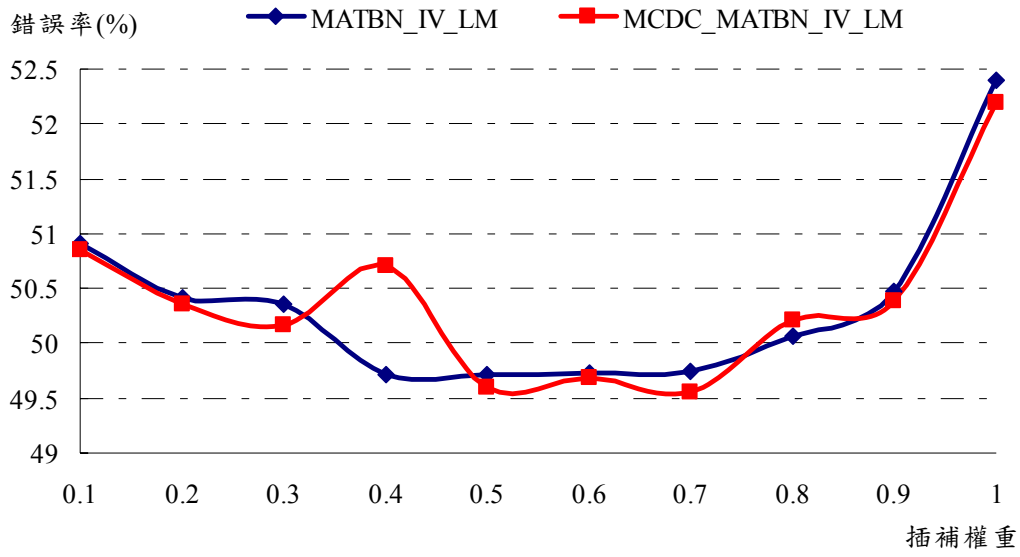

圖 4-6 相同領域語言模型與背景語言模型做線性插補之詞圖搜尋字錯誤率曲線圖

【實驗討論】

將相同領域的語言模型與背景語言模型做線性插補後，使得新的語言模型能更符合外場受訪者的情況，不論是單單使用外場受訪者聲學訓練語料的文字檔或是將外場受訪者聲學訓練語料的文字檔與漢語連續口語對話語音語料庫合併後進行線性插補，對減低詞圖搜尋的字錯誤率皆有一定的效果，根據表 4-6可得知最好的結果為使用兩套文字語料合併的效果最好，當插補權重α設為0.7時，字錯誤率為49.56%。

4.2 傳統信心度評估之實驗

本小節分為兩部份:4.2.1小節討論以特徵為基礎之信心度評估;4.2.2小節則為討論事後機率之信心度評估。而信心度評估的對象分別為外場記者與受訪者兩套語料庫其語音辨識系統字錯誤率基礎實驗結果中個別的最佳辨識詞序列中之每一個詞。其中MATBN_R代表外場記者語料，而MATBN_IV則代表受訪者語料。這兩套語料其信心度評估之基礎實驗結果(Baseline)可參照表 4-7。

MATBN_R	MATBN_IV
24.52	51.97

表 4-7 外場記者與受訪者信心度評估之信心度錯誤率(%)基礎實驗結果

4.2.1 以特徵為基礎之信心度評估實驗

在此實驗中，我們主要分別使用2.1小節所提到的聲學穩定度(AS)及候選詞假設密度(HD)兩種傳統較常使用的特徵，或是將此兩種特徵合併使用(HD+AS)，搭配2.1小節所提到的自然貝氏分類器來對外場記者與受訪者兩套語料庫做信心度評估。其實驗結果可參照表 4-8。

	MATBN_R	MATBN_IV
HD	23.34	33.08
AS	21.25	31.74
HD+AS	21.08	30.50

表 4-8 以特徵為基礎之信心度評估之信心度錯誤率(%)

【實驗討論】

由實驗結果可得知，不論是聲學穩定度或候選詞假設密度在兩套語料皆能有效的降低信心度錯誤率，當進一步採用自然貝氏分類器的假設合併此兩項預估特徵時，更能降低信心度錯誤率，相較於表 4-7，表 4-8的最佳的結果對外場記者及受訪者測試語料分別有14.03%，41.31%的信心度錯誤率相對下降(Relative Reduction)。

4.2.2 事後機率之信心度評估實驗

我們針對 2.2.2 小節中式(2-17)，也就是以一般傳統的事後機率為辨識詞的信心度，以及 2.2.3 小節中的式(2-18)至(2-20)三種不同的信心度評估進行實驗比較。在計算事後機率的時候，由於語言模型的分數為介於 0~1 之間的值，但聲學分數的區間則是 0 至無窮大。所以我們通常在計算傳統的事後機率時，會使用一個權重$\kappa(\kappa>1)$來拉近聲學與語言模型之間分數的比例，如將式(2-17)修改為:

$$P(a:[w_a;s_a,e_a]\,|\,\Psi^X)=\frac{\sum_{\{\overline{W}:[w^n;s^n,e^n]_{n=1}^N\}\in\Psi^X,a\subset\overline{W}}\left\{\prod_{n=1}^{N}p(x_{s^n}^{e^n}\,|\,w^n)^{1/\kappa}\cdot P(w^n\,|\,h^n)\right\}}{\sum_{\{\overline{W}':[w^m;s^m,e^m]_{m=1}^M\}\in\Psi^X}\left\{\prod_{m=1}^{M}p(x_{s^m}^{e^m}\,|\,w^m)^{1/\kappa}\cdot P(w^m\,|\,h^m)\right\}} \tag{4-3}$$

$1/\kappa$代表壓縮聲學模型分數，使之分數區間能較接近語言模型的分數區間。在本實驗中，首先試著調整κ的值，使得式(4-3)在測試語料的信心度錯誤率為最低。實驗結果可以參照表 4-9，對應的信心度錯誤率曲線圖可見圖 4-7及圖 4-8。

κ	MATBN_R	MATBN_IV
5	24.13	37.46
6	24.31	36.38
7	24.18	34.18
8	23.65	32.92
9	22.80	32.45
10	22.32	32.55
11	22.18	31.31
12	22.43	31.31

表 4-9 使用不同κ值計算事後機率之信心度錯誤率(%)

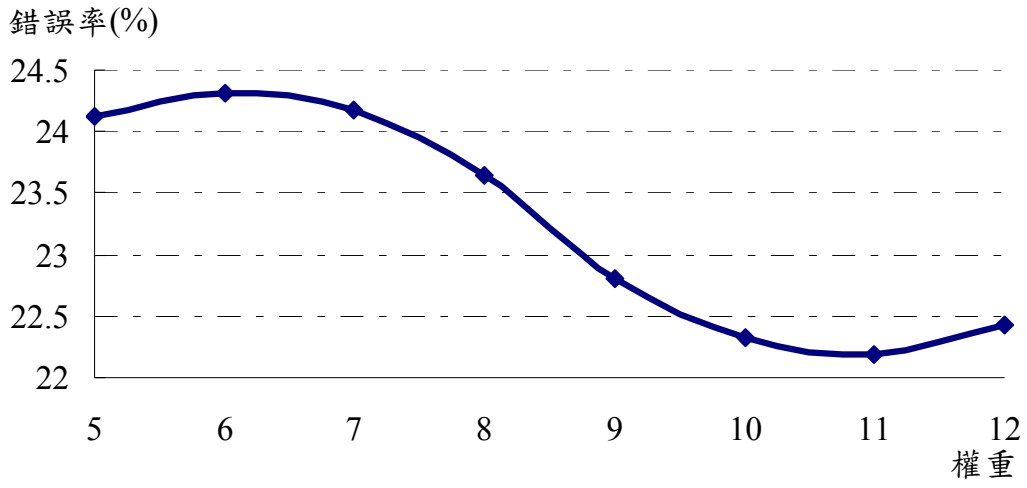

圖 4-7 外場記者:使用不同κ值計算事後機率所獲得的信心度錯誤率曲線圖

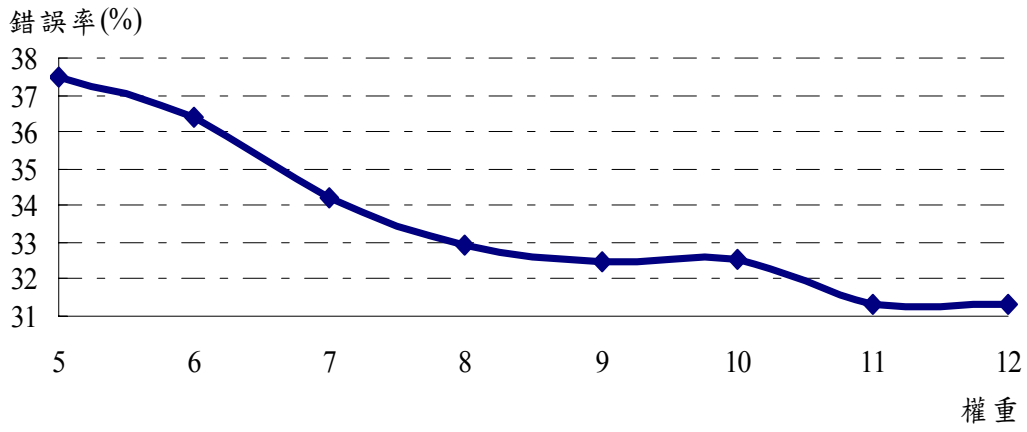

圖 4-8 外場受訪者:使用不同κ值計算事後機率所獲得的信心度錯誤率曲線圖

根據表 4-9的結果，我們進一步比較式(2-17)至(2-20)四種信心度評估之效果。實驗結果可參考表 4-10。其中Cnormal代表式(2-17)，也就是傳統事後機率的算法，而外場記者與受訪者在計算式(2-17)的事後機率時聲學分數權重κ根據表 4-9皆設為11，使其在評估語料有最低的信心度錯誤率。

	MATBN_R	MATBN_IV
Cnormal	22.18	31.31
Csec	21.47	32.32
Cmed	21.47	32.32
Cmax	21.47	32.32

表 4-10 不同的辨識詞事後機率方法之信心度錯誤率(%)

【實驗討論】

由實驗結果可得知，雖然Cmed、Csec及Cmax三種信心度評估雖然跟基礎實驗結果比較有相當的進步，但是與傳統的事後機率方法相比，並不一定會有較佳的效果。其主要原因可能在於外場受訪者測試語料本身的辨識率較低，造成當使用Cmed、Csec及Cmax時，大部份都是加大錯誤的詞信心度，反而造成信心度錯誤率因此升高。

4.3 信心度評估應用於降低詞圖搜尋錯誤率之實驗

此實驗主要分為兩個部份: 4.3.1小節討論關於運用傳統的事後機率於降低詞圖搜尋之錯誤率;4.3.2小節則是探討最小化音框錯誤率對於詞圖搜尋的影響。外場記者與受訪者字錯誤率的基礎實驗結果可參考表 4-11。

MATBN_R	MATBN_IV
20.79	49.56

表 4-11 外場記者與受訪者語料經詞圖搜尋後之字錯誤率(%)，此為基礎實驗結果

4.3.1 運用事後機率降低詞圖搜尋錯誤率之實驗

根據2.5.3小節中的式(2-43)，必須先求得詞圖中每個詞段的事後機率。我們根據4.2.2小節的實驗，聲學分數權重κ設為11去計算每個詞段的事後機率，再進行式(2-43)的詞圖搜尋，其實驗結果可參照表 4-12。

MATBN_R	MATBN_IV
20.68	47.60

表 4-12 外場記者與受訪者語料經事後機率詞圖搜尋後之字錯誤率(%)

【實驗討論】

由實驗結果可得知，如果事後機率的信心度估評在某個語料其信心度錯誤率下降越明顯，其運用在降低詞圖插尋錯誤率有較佳的效果。如在本實驗中，外場受訪者測試語料就有較佳的進步。

4.3.2 最小化音框錯誤詞圖搜尋之實驗

本實驗主要是探討2.5.4小節所討論的式(2-46)，其中α的部份我們嘗試設0~0.1的區間，間隔為0.01，來代表是否要強調長詞。實驗結果可參考表 4-13，而對應的錯誤率曲線請參考圖 4-9及圖 4-10。

α	MATBN_R	MATBN_IV
0	20.81	47.35
0.01	20.68	47.45
0.02	20.67	47.74
0.03	20.61	47.87
0.04	20.62	48.01
0.05	20.60	48.16
0.06	20.57	48.19
0.07	20.56	48.37
0.08	20.58	48.44
0.09	20.59	48.57
0.1	20.57	48.59

表 4-13 運用最小化音框錯誤率於詞圖搜尋之字錯誤率(%)

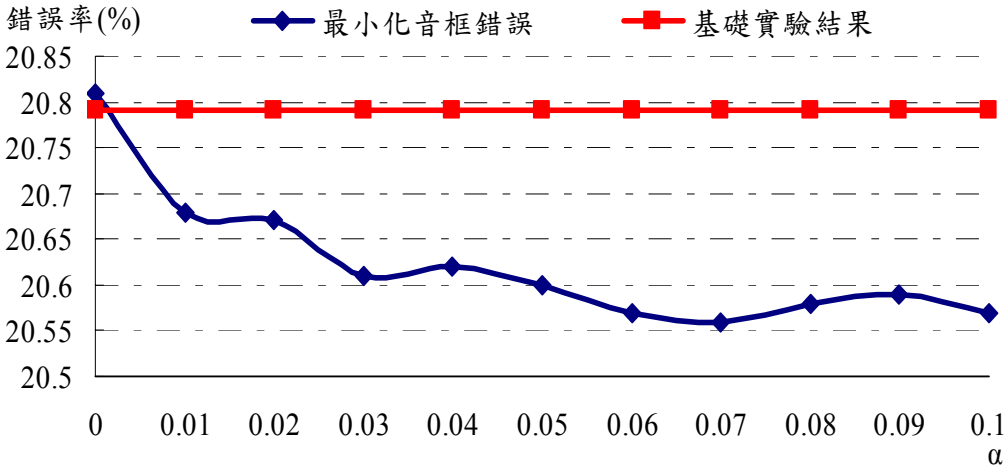

圖 4-9 外場記者:運用最小化音框錯誤率於詞圖搜尋之字錯誤率曲線圖

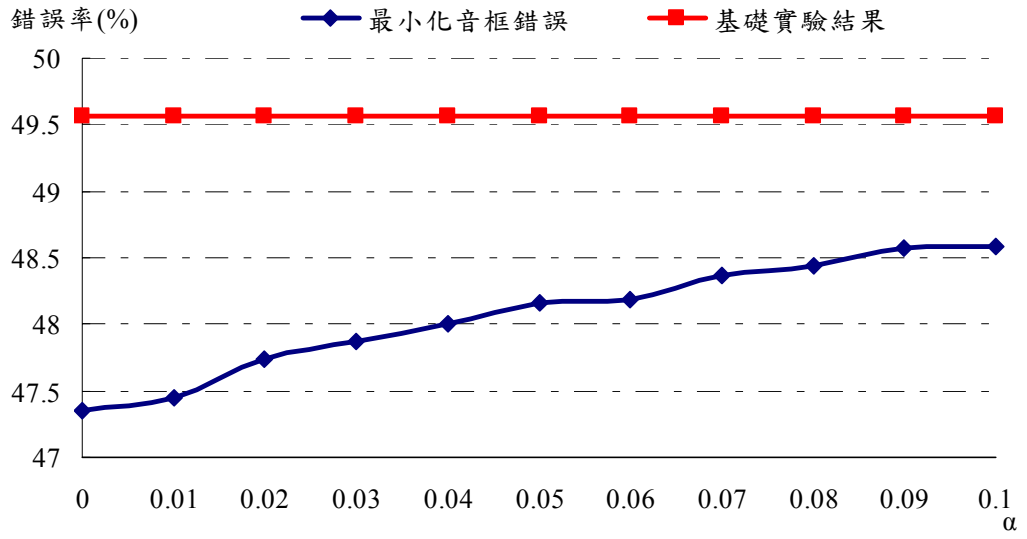

圖 4-10 外場受訪者:運用最小化音框錯誤率於詞圖搜尋之字錯誤率曲線圖

【實驗討論】

由實驗結果可得知，在外場記者的評估語料中，最好的結果相較於基礎實驗結果可以有 1.11%的相對進步。而在外場受訪者的部份則更可以有 4.56%的相對進步。此方法似乎在語音辨識系統的正確率較低的情況之下，有較大的進步的空間。另外，從實驗可觀察出外場記者的語料似乎有出現少許的長詞。而受訪者的部份，因為α越大，則錯誤率越高，有偏向於短詞較多的情形。

第5章 改良信心度評估及運用於降低語音辨識系統錯誤率實驗

此章分為4小節，5.1小節為本論文提出結合熵值與現有的信心度估評方法，以降低信心度錯誤率之實驗;5.2及5.3小節則是有關最小化音框錯誤率之改進;5.4小節則是探討有關信心度評估如何運用於最小化貝氏風險法則。

5.1 結合熵值(Entropy)與信心度估評之實驗

在資訊理論中，熵值代表一個隨機變數(Random Variable)所含有的資訊，通常是以位元(Bits)來表示其資訊量，值越大代表資訊越混亂。其表示法如式(5-1)所示:

$$H(R) = -\sum_{r \in R} P(r) \log_2 P(r) \tag{5-1}$$

其中 $P(r)$ 代表一個隨機變數 R 在值為 r 時的機率質量函數(Probability Mass Function)。當有了熵值這個概念，我們進而考慮如圖 5-1的情形:在圖的左右邊各有五個詞段，每個小括號中的數字代表其詞段對應的信心度值。就此圖來說，雖然”今天”這個詞的信心度在兩個詞圖中皆為2/3。但是在左邊的詞圖中，其它的詞段信心度也皆為2/3，就常理而言，在某段時間區間，不應該同時會有許多詞的信心度都

圖 5-1 兩個詞圖中，某一段時間區段的所有詞段的信心度分佈情形

很高或很接近(造成熵值會較高)。因此，右邊詞圖的"今天"其信心度應該比左邊詞圖的"今天"更值得信任。而在本論文中，便嘗試加入熵值的概念，觀察是否對信心度評估有所幫助。首先，本論文提出一個新的信心度評估算式:

$$CM_{new}([w,s,e]) = CM([w,s,e]) * (1 - E_{avg}([w,s,e])) \tag{5-2}$$

其中$[w,s,e]$代表詞圖或N-最化詞序列中的某一個詞段，其開始及結束時間分別為s及e，而$CM([w,s,e])$代表此詞段原本的信心度(可用第2章所提到之任何信心度評估方法求取)，而$E_{avg}([w,s,e])$代表此詞段的平均正規化熵值:

$$E_{avg}([w,s,e]) = \frac{1}{e-s+1}\sum_{t=s}^{e} E_f(t) \tag{5-3}$$

$$E_f(t) = -\frac{1}{\log_2 N}\sum_{[w,s,e],s\le t\le e} P_{CM}([w,s,e],t)\log_2 P_{CM}([w,s,e],t) \tag{5-4}$$

其中，我們先將每個音框的信心度值正規化為0~1的機率分佈值，如式(5-5)及(5-6)

$$P_{CM}([w,s,e],t) = \frac{CM_{sum}([w,s,e],t)}{\sum_{[w',s',e'],e'\le t\le e'} CM_{sum}([w',s',e'],t)} \tag{5-5}$$

$$CM_{sum}([w,s,e],t) = \sum_{\substack{[w',s',e'],w'=w \\ s'\le t\le e'}} CM([w',s,e],t) \tag{5-6}$$

在考慮信心度的熵值時，由於可能會有詞編號相同的詞，但是開始及結束時間不完全一樣的詞段，在計算熵值時，我們只考慮不同詞編號的詞之間其信心度分佈情形。因此，在式(5-6)，會先將相同詞編號之信心度加總。而在式(5-5)到(5-6)中的加總條件w'，代表在t此時間點時，所有不同詞編號的詞。在此實驗中，我們先將熵值求取對象設定為以事後機率為基礎的信心度評估，也就是將式(2-17)至式(2-20)的信心度做為式(5-2)中的$CM([w,s,e])$，並直接在詞圖計算熵值，其實驗結果可參考表5-1(其中 MATBN_R 代表外場記者語料，MATBN_IV 代表外場受訪者語料)，而相對之偵測錯誤交易曲線圖(與傳統事後機率為基礎的信心度評估最好結果相比較)，則可

參考圖 5-2 及圖 5-3，其中 FAR 代表錯誤接受率(False Acceptance Rate)，FRR 代表錯誤拒絕率(False Rejection Rate)。接著，我們再試著將熵值資訊應用於以特徵為基礎的信心度評估法則。根據 4.2.1 小節的實驗，我們直接使用最好的結果，也就是結合聲學穩定度及候選詞假設密度(AS+HD)此兩項預估特徵所求得之信心度做為式(5-2)中的 $CM([w,s,e])$。此外，由於聲學穩定度無法直接於詞圖上計算熵值，因此，我們改在 *N*-最佳詞序列上計算熵值。其實驗結果可參考表 5-2。

	MATBN_R	MATBN_IV
基礎實驗	24.52	51.97
Cnormal	22.18	31.31
Csec	21.47	32.32
Cmed	21.47	32.32
Cmax	21.47	32.32
Centropy(Cnornal)	18.55	31.08
Centropy(Csec)	18.44	28.50
Centropy(Cmed)	18.44	28.44
Centropy(Cmax)	18.45	28.51

表 5-1 結合詞圖上的熵值與事後機率相關信心度評估之信心度錯誤率(%)

	MATBN_R	MATBN_IV
HD+AS	21.08	30.50
HD+AS(entropy)	21.02	30.20

表 5-2 結合 100-最佳詞序列熵值與以特徵為基礎信心度評估之信心度錯誤率(%)

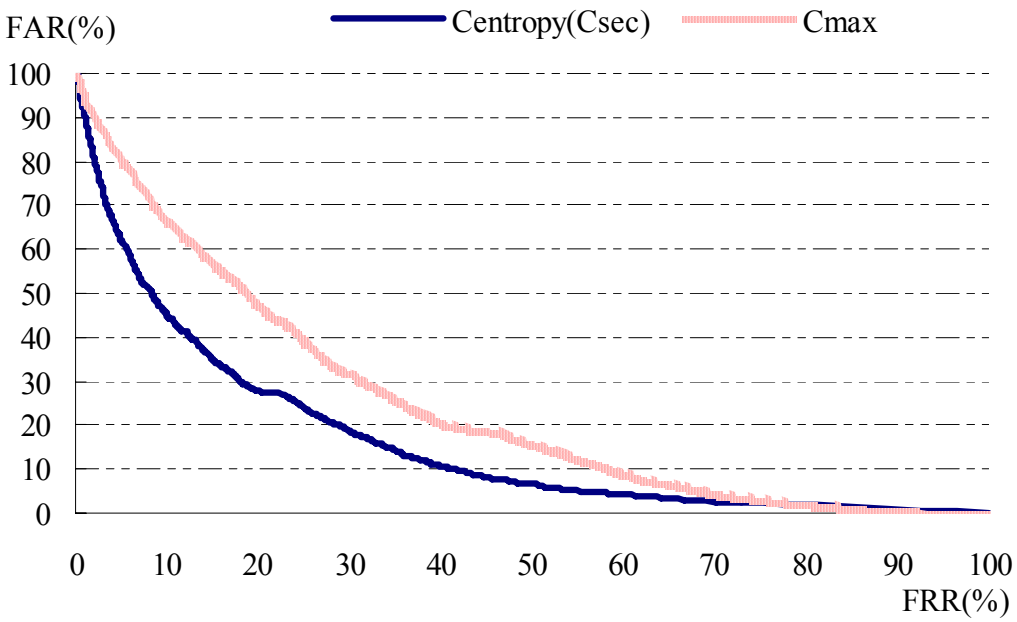

圖 5-3 外場受訪者語料:結合詞圖上的熵值資訊與事後機率信心度評估對應之偵測錯誤交易曲線圖

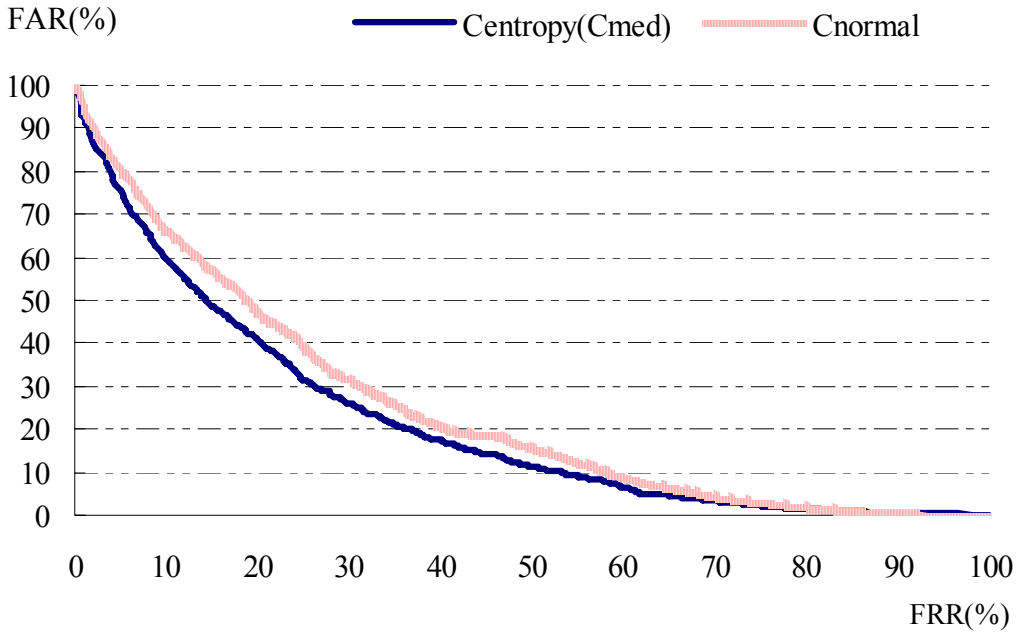

圖 5-2 外場記者語料:結合詞圖上的熵值資訊與事後機率信心度評估對應之偵測錯誤交易曲線圖

【實驗討論】

由實驗結果可得知，不論是對外場記者或受訪者的語料，本論文所提出的方法都較過去所提出之任何信心度評估有更佳的表現。在外場記者的語料部份，本論文所提出的方法與傳統最好的事後機率相關之信心度評估(以式(2-18)Cmed 為比較對象)可以有高達 16.37%的進步，而在外場受訪者部份則有 12.00%的進步(與式(2-17)一般事後機率計算方法相比)。此外，以事後機率為基礎的信心度評估結合熵值似乎有較佳的結果，其原因可能是因為 *N*-最佳詞序列中，由於每個時間點不同的詞資訊較詞圖少，因此較無法表現信心度的混淆度，造成以特徵為基礎之信心度評估的效果較不明顯。為了驗證此想法，本論文另外嘗試同樣在計算 *N*-最佳詞序列的熵值資訊與事後機率為基礎的信心度評估結合，其實驗結果如表 5-3 所示。相較於表 5-1，在 *N*-最佳詞序列中計算熵值的效果的確較在詞圖上求取而言來得差。接著，我們再進一步分析表 5-1 的實驗結果，觀察事後機率為基礎之信心度評估加入詞圖熵值資訊前後，其錯誤接受率(FAR)及錯誤拒絕率(FRR)的變化，其結果如表 5-4 所示。由表 5-4 可發現，適當地結合熵值資訊，能有效地使錯誤接受率下降。其原因應在於此方法能降低大部份辨識錯誤的詞信心度，不過，也可能同時降低小部份辨識正確結果信心度的值，造成錯誤拒絕率也勢必會有所增加。最後，我們再試著探討時間複雜度的問題，由於計算熵值的步驟必須是在計算事後機率相關的信心度評估之後，所以，其運算量勢必較為龐大。由於一般事後機率(也就是式(2-17))的計算量較其它三種方法來得少，因此我們接下來只針對式(2-17)與在詞圖上計算熵值的運算量進行比較。

	MATBN_R	MATBN_IV
Centropy(Cnornal)	20.34	34.06
Centropy(Csec)	20.03	33.66
Centropy(Cmed)	20.03	34.20
Centropy(Cmax)	20.03	33.67

表 5-3 結合 100-最佳詞序列熵值與以事後機率為基礎信心度評估之信心度錯誤率(%)

	MATBN_R		MATBN_IV	
	FAR	FRR	FAR	FRR
Cnormal	80.07	3.38	35.41	26.87
Csec	77.18	3.38	42.24	21.59
Cmed	77.18	3.38	42.21	21.63
Cmax	77.18	3.38	42.21	21.63
Centropy(Cnornal)	56.54	6.21	44.46	16.60
Centropy(Csec)	51.98	7.54	31.66	25.08
Centropy(Cmed)	52.01	7.54	31.60	25.02
Centropy(Cmax)	51.98	7.54	31.66	25.11

表 5-4 結合詞圖熵值與以事後機率為基礎信心度評估之錯誤接受率及錯誤拒絕率(%)

	MATBN_R		MATBN_IV	
	Cnormal	熵值	Cnormal	熵值
第一次	1.59	0.94	15.94	5.89
第二次	1.58	0.93	15.91	5.85
第三次	1.58	0.94	15.98	5.84
第四次	1.58	0.93	15.93	5.84
第五次	1.58	0.94	15.93	5.84
平均	1.58	0.94	15.94	5.85

表 5-5 測試語句平均每一秒計算詞圖熵值或計算一般事後機率(Cnormal)所需時間(百分之一秒)之比較

實驗結果如表 5-5 所示，其中每一欄代表每一秒測試語句平均所需計算時間(Real-time Factor, RT)。由實驗結果可得知，即使加入計算熵值的時間，仍然能即時地計算出每一個辨識詞的信心度。

5.2 融合不同詞圖並配合最小化音框錯誤率詞圖搜尋方法之實驗

在2.5.4小節所介紹的最小化音框錯誤率(Time Frame Error)詞圖搜尋方法中，是運用於單一語音特徵參數所構成之詞圖上。而在本論文中，則嘗試結合兩種由不同語音特徵參數所形成的詞圖。希望藉由兩種不同的詞圖，能帶來更多的資訊，進而找出更正確的詞序列。而由於兩種詞圖是分別由不同的語音特徵參數所形成，所以可能造成兩個詞圖的詞段分數區間不同。舉例來說，在本論文中由梅爾倒頻譜係數(Mel-frequency Cepstral Coefficients, MFCC)語音特徵參數所形成的詞圖，其中的詞段聲學對數相似度趨為正值。而由異質性線性鑑別分析(Heteroscedastic Linear Discriminant Analysis, HLDA)搭配最大相似度線性轉換(Maximum Likelihood Linear Transformation, MLLT)語音特徵參數所產生的詞圖，其詞段聲學對數相似度趨為負值。該如何融合此兩種詞圖，便成為一個值得思考的方向。在本論文中，我們試著以最小化音框錯誤率的方式合併兩種詞圖，如式(5-7)所示:

$$F^{*}(X)=\arg\min_{[w^{n};s^{n},e^{n}]_{n=1}^{N}}\left\{\begin{array}{l}\displaystyle\sum_{[v^{m};s^{m},e^{m}]_{m=1}^{M}\in\Psi_{A}^{X}}\sum_{n=1}^{N}\frac{\displaystyle\sum_{t=s^{n}}^{t=e^{n}}1-\delta(w^{n},w_{t}^{m})}{1+1*\alpha(e_{n}-s_{n}-1)}P(v^{m};s^{m},e^{m}]_{m=1}^{M}\mid\Psi_{A}^{X})\cdot P(\Psi_{A}^{X}\mid X)\\ +\displaystyle\sum_{[r^{y};s^{y},e^{y}]_{y=1}^{Y}\in\Psi_{B}^{X}}\sum_{n=1}^{N}\frac{\displaystyle\sum_{t=s^{n}}^{t=e^{n}}1-\delta(w^{n},r_{t}^{y})}{1+1*\alpha(e_{n}-s_{n}-1)}P([r^{y};s^{y},e^{y}]_{y=1}^{Y}\mid\Psi_{B}^{X})\cdot P(\Psi_{B}^{X}\mid X)\end{array}\right\} \tag{5-7}$$

其中，Ψ_{A}^{X} 及 Ψ_{B}^{X} 分別代表由同一個聲學觀測序列 X ，兩種不同的語音特徵參數 A 及 B 所形成的詞圖。而 $[v^{m};s^{m},e^{m}]_{m=1}^{M}$ 及 $[r^{y};s^{y},e^{y}]_{y=1}^{Y}$ 則分別代表 Ψ_{A}^{X} 及 Ψ_{B}^{X} 兩個詞圖上的詞序列，其詞的個數為 M 及 Y ，其它數學符號可參照2.5.4小節。接著假設每個詞圖的事後機率 $P(\Psi^{X}\mid X)$ 為均等分配(Uniform Distribution)，再經由類似式(2-46)的推導之後，融合兩種詞圖的最小化音框錯誤率詞圖搜尋最後便可表示成式(5-8) :

$$F^*(X)=\arg\min_{[w^n;s^n,e^n]_{n=1}^N}\left\{\sum_{n=1}^{N}\frac{\sum_{t=s_n}^{t=e_n}\left[1-\sum_{\substack{[w^m;s^m,s^m]\in\in\Psi_A^X\\ s^m\le t\le s^m}}\delta(w^n,w_t^m)P([v^m;s^m,e^m]\,|\,\Psi_A^X)\right]}{2+2\alpha(e^n-s^n)}+\sum_{n=1}^{N}\frac{\sum_{t=s_n}^{t=e_n}\left[1-\sum_{\substack{[r^y;s^y,e^y]\in\Psi_B^X\\ s^y\le t\le s^y}}\delta(w^n,r_t^y)P([r^y;s^y,e^y]\,|\,\Psi_B^X)\right]}{2+2\alpha(e^n-s^n)}\right\} \quad (5\text{-}8)$$

其中，$P([v^m;s^m,e^m]\,|\,\Psi_A^X)$代表詞段$[v^m;s^m,e^m]$在詞圖$\Psi_A^X$的事後機率，$P([r^y;s^y,e^y]\,|\,\Psi_B^X)$則是詞段$[r^y;s^y,e^y]$於詞圖$\Psi_B^X$的事後機率。實驗結果可參考表5-6，兩套測試語料對應的字錯誤率曲線圖則可參照圖 5-4及圖 5-5。圖 5-4及圖 5-5中的基礎實驗結果為傳統最大事後機率解碼(MAP Decoding)法則，其語音特徵參數為異質性線性鑑別分析搭配最大相似度線性轉換(HLDA+MLLT)。而HLDA+MLLT及MFCC則代表最小化音框錯誤率單獨配合使用其中任一語音特徵參數的結果。

α	MATBN_R	MATBN_IV
0.00	19.83	47.14
0.01	20.08	49.26
0.02	20.73	51.93
0.03	21.35	54.12
0.04	21.68	55.33
0.05	22.06	56.65
0.06	22.38	57.54
0.07	22.74	58.35
0.08	23.00	58.98
0.09	23.23	59.37
0.10	23.44	59.82

表 5-6 運用最小化音框錯誤率融合不同詞圖於詞圖搜尋之字錯誤率(%)

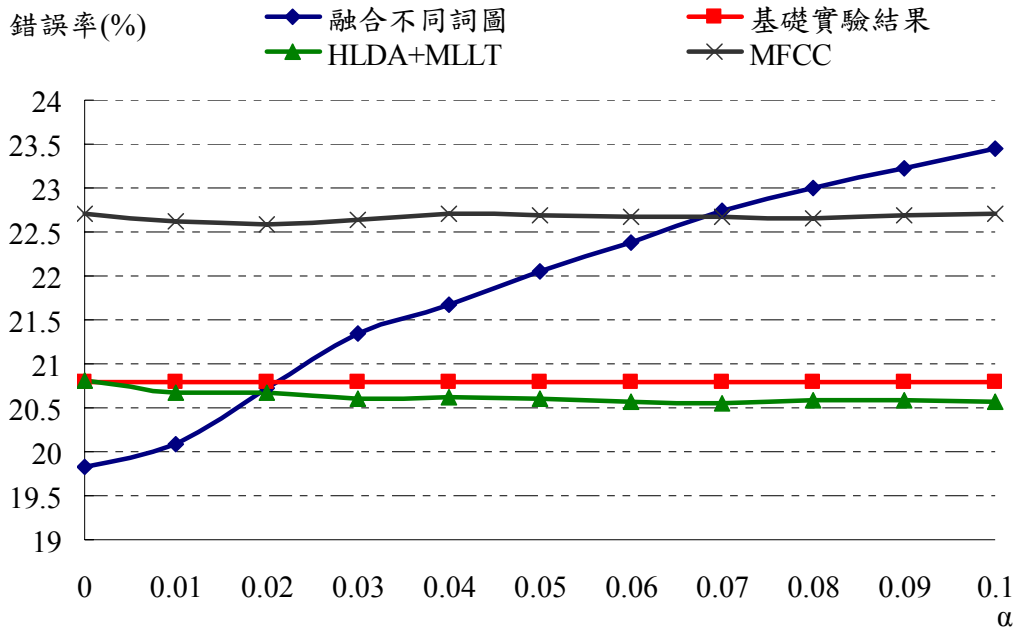

圖 5-4 外場記者:運用最小化音框錯誤率融合不同詞圖於詞圖搜尋之字錯誤率曲線圖

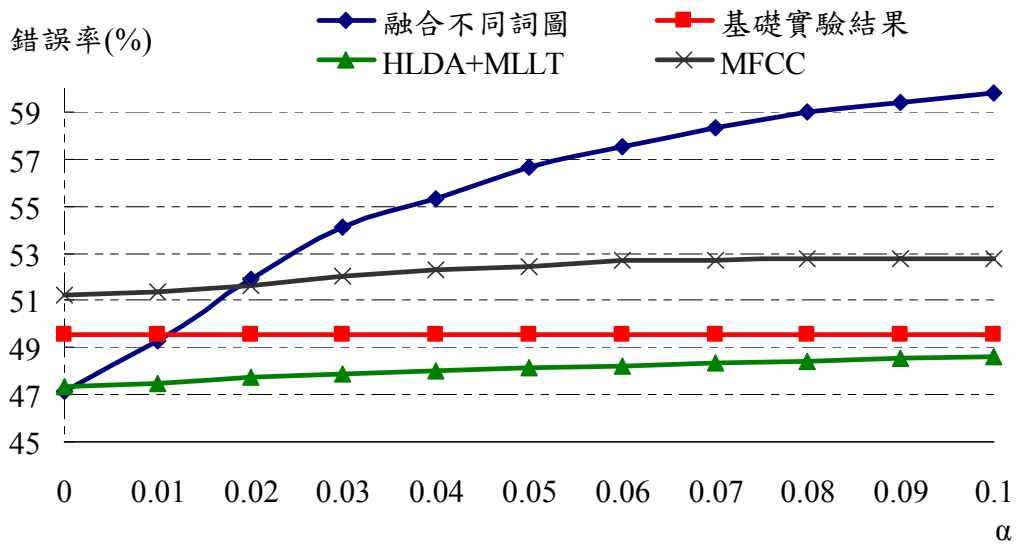

圖 5-5 外場受訪者:運用最小化音框錯誤率融合不同詞圖於詞圖搜尋之字錯誤率曲線圖

【實驗討論】

在融合了兩種不同語音特徵參數的詞圖後，不論是在外場記者測試語料或是外場受者測試語料，皆有不錯的效果。在外場記者測試語料較基礎實驗結果有4.6%的字錯誤率相對下降，而在外場受訪測試語料的部份則有4.8%的字錯誤率相對下降(皆於α為0時)。而這個方法除了可以降低語音辨識系統的錯誤率之外，另外一個優點便是可以系統化的方式加入兩個以上的詞圖。其通式如式(5-9)：

$$F^*(X)=\arg\min_{[w^n;s^n,e^n]_{n=1}^N}\left\{\begin{array}{l}\displaystyle\sum_{n=1}^{N}\frac{\displaystyle\sum_{t=s_n}^{t=e_n}\left[1-\sum_{\substack{[w^m;s^m,s^m]\in\Psi_A^X\\ s^m\le t\le s^m}}\delta(w^n,w_t^m)P([v^m;s^m,e^m]\,|\,\Psi_A^X)\cdot P(\Psi_A^X\,|\,X)\right]}{1+\alpha(e^n-s^n)}\\ \displaystyle+\sum_{n=1}^{N}\frac{\displaystyle\sum_{t=s_n}^{t=e_n}\left[1-\sum_{\substack{[r^y;s^y,e^y]\in\Psi_B^X\\ s^y\le t\le s^y}}\delta(w^n,r_t^y)P([r^y;s^y,e^y]\,|\,\Psi_B^X)P(\Psi_B^X\,|\,X)\right]}{+\alpha(e^n-s^n)}\\ +\ldots\\ \displaystyle+\sum_{n=1}^{N}\frac{\displaystyle\sum_{t=s_n}^{t=e_n}\left[1-\sum_{\substack{[\omega^y;s^\Omega,e^\Omega]\in\Psi_\Omega^X\\ s^\Omega\le t\le s^\Omega}}\delta(w^n,r_t^y)P([\omega^y;s^y,e^y]\,|\,\Psi_\Omega^X)P(\Psi_\Omega^X\,|\,X)\right]}{1+\alpha(e^n-s^n)}\end{array}\right\}$$

(5-9)

其中Ω代表詞圖的個數，而Ψ_Ω^X則代表第Ω個詞圖，而

$$\delta(w^n,w_t^m)=\begin{cases}1,\ \text{if } w^n=w_t^m\\ 0,\ \text{if } w^n\neq w_t^m\end{cases}\qquad(5\text{-}10)$$

5.3 以字(Character)為單位之最小化音框錯誤率詞圖搜尋

在2.5.4及5.2兩小節中，當比較詞圖中兩個詞段時，皆以詞為單位。但對中文而言，通常會有斷詞不一致的問題(例如甲系統辨識成"今天天氣"這個詞，但乙系統則是辨識成"今天"及"天氣"兩個詞，在計算語音辨識系統的字錯誤率時，其辨識正確率應為相同)。如我們考慮到圖 5-6的情況時，假設每個詞段的事後機率皆相同。如以詞為比對單位，使用最小化音框錯誤率詞圖搜尋時，"今天天氣"與"今天"此兩詞段應是不同，但如果我們以"字"為單位時，卻會有可能對應到"今"、"天"兩個字。由此可知，不同比對單位，可能會有不同的結果。因中文的語音辨識系統正確率的評估單位為字，所以在進行中文語音辨識系統的最小化音框錯誤率詞圖搜尋時，比較單位也應以字為基準。其實作的式子只需修改式(2-46)為:

$$F^*(X) = \arg\min_{[c^n;s^n,e^n]_{n=1}^N} \left\{ \sum_{n=1}^{N} \frac{\sum_{t=s_n}^{t=e_n} \left[1 - \sum_{\substack{[c_i^m;s_i^m,e_i^m]\in[v^m;s^m,e^m], \\ s_i^m \le t \le e_i^m}} \delta(c^n, c_t^m) P([c_i^m; s_i^m, e_i^m] \mid \Psi^X) \right]}{1+\alpha(e^n - s^n)} \right\} \tag{5-11}$$

其中$[c^n, s^n, e^n]_{n=1}^N$代表詞圖上一條完整的字序列，但在本論文中假設此字序列必由詞圖上原本的詞序列組成，本論文僅將此詞序列切割成以字為單位的序列，並不額外產生所謂的字圖(Character Graph)。$[c_{ni}^m, s_{ni}^m, e_{ni}^m] \in [v^m, s^m, e^m]$代表組成$[v^m, s^m, e^m]$此詞段的第$i$個字，其開始及結束時間為$s_{ni}^m$及$e_{ni}^m$。在此小節的實驗中，取得字的開始及結束時間分為兩種方法，一種為組成詞段的所有字平均分配詞段的持續時間;另一種則為執行詞段中以字為單位的強制切齊(Force Alignment)。實驗結果可參考表5-7，而對應的字錯誤率曲線圖則可參考圖 5-7及圖 5-8。這裡所使用的語音特徵參數仍然為HLDA+MLLT。

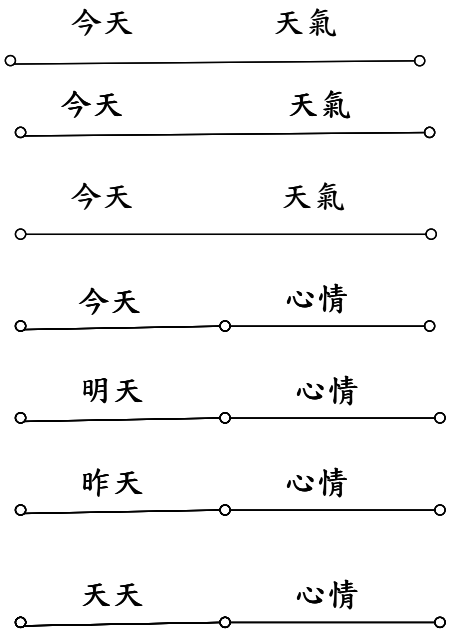

圖 5-6 詞圖中某段音框區間其詞段分佈

	MATBN_R		MATBN_IV	
α	平均分配	強制切齊	平均分配	強制切齊
0.00	20.73	20.73	47.32	47.31
0.01	20.63	20.63	47.33	47.35
0.02	20.66	20.64	47.76	47.73
0.03	20.60	20.60	47.87	47.88
0.04	20.58	20.59	48.05	48.08
0.05	20.62	20.61	48.28	48.17
0.06	20.62	20.61	48.35	48.32
0.07	20.62	20.61	48.44	48.38
0.08	20.61	20.58	48.50	48.46
0.09	20.61	20.60	48.60	48.57
0.10	20.60	20.61	48.62	48.65

表 5-7 以字為比對單位之最小化音框錯誤率詞圖搜尋之字錯誤率(%)

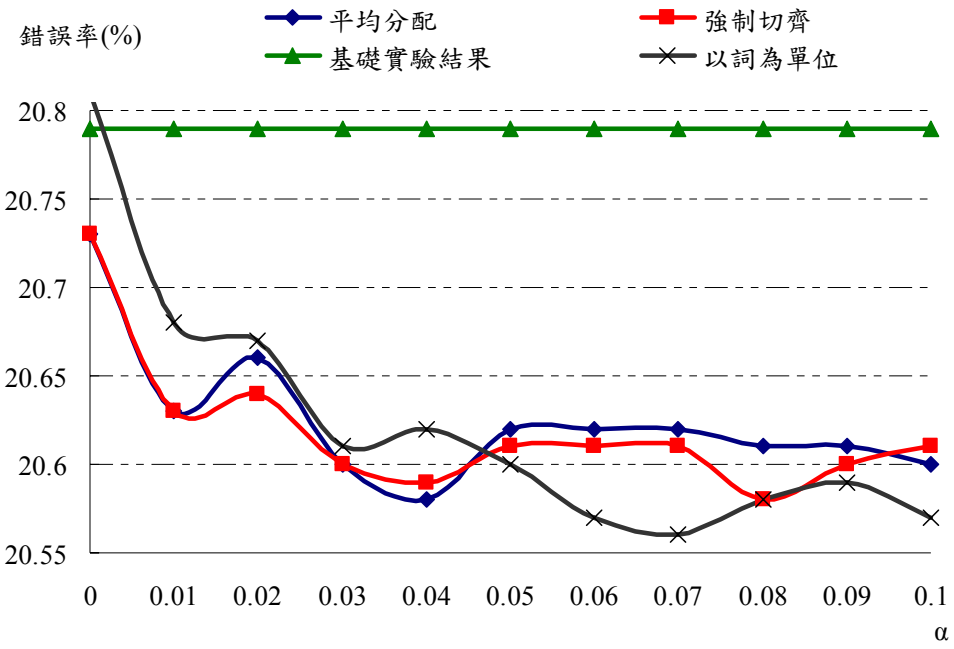

圖 5-7 外場記者:以字為比對單位之最小化音框錯誤率詞圖搜尋之字錯誤率曲線圖

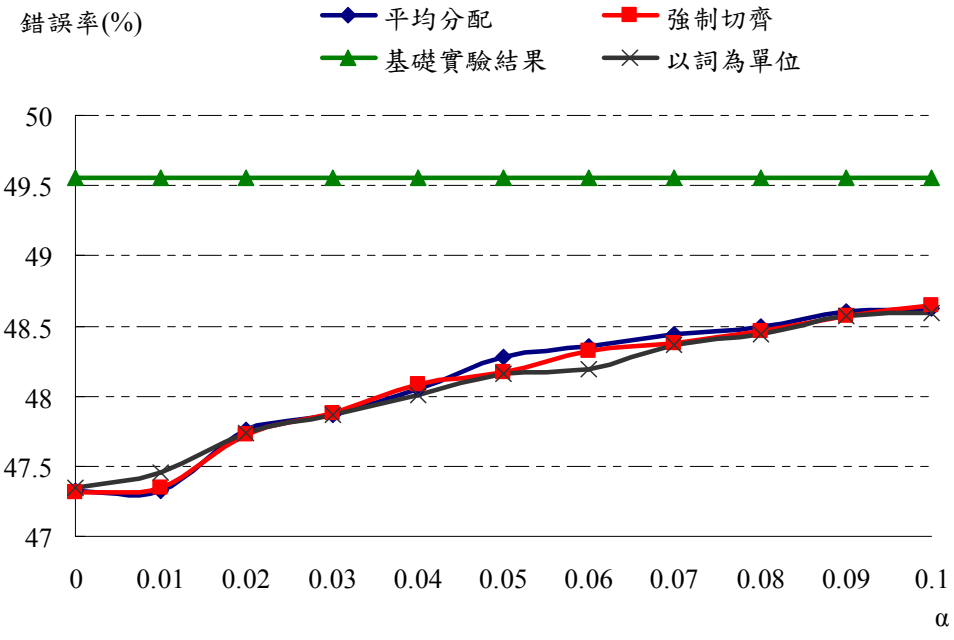

圖 5-8 外場受訪者:以字為比對單位之最小化音框錯誤率詞圖搜尋之字錯誤率曲線圖

【實驗討論】

由實驗結果可得知，以字為比對單位之最小化音框錯誤率詞圖搜尋較基礎實驗結果而言，在外場記者的評估語料中，最好的結果有 1.01%的相對進步，而在外場受訪者的部份則更可以有 4.46 %的相對進步。而在決定字的開始及結束時間時，不論是使用平均分配其對應詞段的持續時間抑或採用強制切齊的方式，其結果似乎不會有太大的差距。我們推測是因為不論是用平均分配或強制切齊的方式，對於詞段上每個字的正確開始及結束時間皆會有錯誤的關係。才會造成實驗數據相差不太。

5.4 結合信心度評估與以 Levenshtein 距離為成本函式之最小化貝氏法則

如2.5.1小節中所提到的，傳統以Levenshtein距離為成本函式之最小化貝氏法則常用於*N*-最佳化詞序列的重新排序(Reranking)，其數學式如式(5-12)所示:

$$F^*(X) = \arg \min_{W' \in \mathbf{W}_{N-\text{best}}} \sum_{W \in \mathbf{W}_{N-\text{best}}} \ell(W,W') \cdot P(W \mid X) \qquad (5\text{-}12)$$

其中 $\mathbf{W}_{N-\text{Best}}$ 是由詞圖中產生*N*條分數最高的詞序列，而 $\ell(W,W')$ 則是兩條詞序列的Levenshtein距離。式(5-12)其意義在於找出一條與其它條詞序列之Levenshtein距離最低(或者說與其它詞序列期望最相近)之詞序列。本論文進一步提出一種想法:當目前此條詞序列與其它詞序列較不相近時，需要多考慮較不相近詞序列的信心度，如果不相近的詞序列信心度較低，則應降低其成本函式。反之，則應保持其原本之成本函式值。此觀念可用式(5-13)表示:

$$F^*(X) = \arg \min_{W' \in \mathbf{W}_{N-\text{best}}} \sum_{W \in \mathbf{W}_{N-\text{best}}} \ell(W,W') \cdot CM(W) \cdot P(W \mid X) \qquad (5\text{-}13)$$

其中 $CM(W)$ 代表經與 W' 計算Levenshtein距離後，和 W' 不同的詞對應之信心度評估。稍後的實驗，我們首先分析在外場記者與外場受訪者兩套語料中，其*N*-最佳化詞序列最小字錯誤率(在每句測試句之*N*-最佳化詞序列中皆挑選字錯誤率最低的詞序列，統計其平均字錯誤率)。其值為實驗結果的上限(Upper Bound)，外場記者及受訪者的*N*-最佳化詞序列最小字錯誤率曲線圖可參考圖 5-9及圖 5-10。

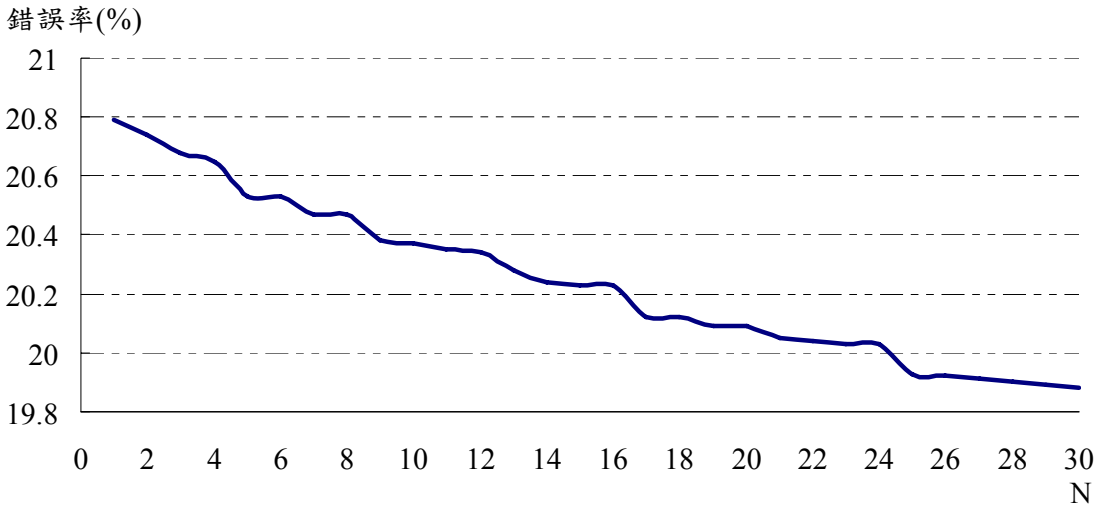

圖 5-9 外場記者: 1 至 30-最佳詞序列之最小字錯誤率曲線圖

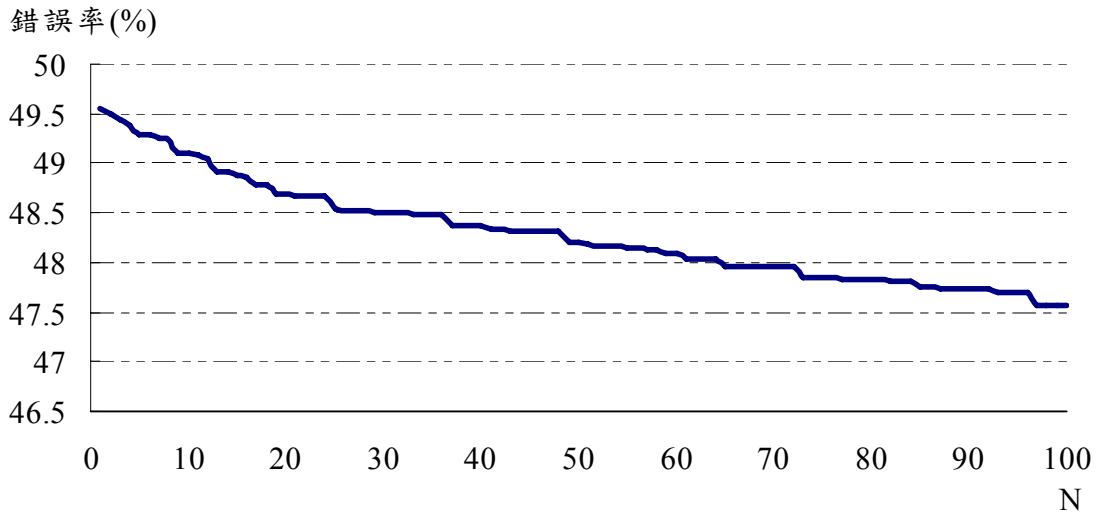

圖 5-10 外場受訪者: 1 至 100-最佳詞序列之最小字錯誤率曲線圖

	MATBN_R	MATBN_IV
基礎實驗結果	20.79	49.54
MBR	21.02	49.73

表 5-8 外場記者與受訪者之基礎實驗結果及 MBR 之字錯誤率(%)

由圖 5-9及圖 5-10可以發現，不論是在外場記者或受訪者測試語料當*N*-最佳詞序列的N值愈大時，語音辨識系統的進步空間也就愈大。在本實驗中，外場記者測試語料的N取30，而外場受訪者測試語料的N取100。而兩套測試語料所使用的語音特徵參數為HLDA+MLLT，其最大事後機率解碼基礎實驗結果及傳統以Levenshtein距離為成本函式之最小化貝氏風險法則(以MBR表示)之錯誤率可參考表 5-8。由實驗結果得知，過去以Levenshtein距離為成本函式的方法，也就是找一條跟其它詞序列期望最相似的詞序列，並不一定對語音辨識系統的正確率有所幫助。

接著，本論文將討論當加入信心度評估時，是否對傳統的最小化貝氏風險法則有所助益。由於在最小化貝氏風險法則中，已經有使用到詞序列的事後機率 $P(W \mid X)$ 。為了避免資訊的重複使用，這裡所使用的信心度評估主要是使用以特徵為基礎的信心度評估法(合併聲學穩定度及候選詞假設密度此兩項預估特徵)，而不採用事後機率相關的信心度評估。此外，在實作上，我們將以式(5-14)為主:

$$F^*(X) = \arg \min_{W' \in \mathbf{W}_{N-\text{best}}} \sum_{W \in \mathbf{W}_{N-\text{best}}} \ell(W, W') \cdot CM(W')^{\beta} \cdot P(W \mid X) \qquad (5\text{-}14)$$

式(5-14)與式(5-13)主要的差別在於式(5-14)加入一個權重β，用來調整對信心度的信賴程度。其實驗結果可參考表 5-9、圖 5-11及圖 5-12。

α	MATBN_R	MATBN_IV
0.01	21.01	49.69
0.02	21.02	49.70
0.03	21.01	49.67
0.04	21.01	49.70
0.05	21.00	49.67
0.06	21.02	49.65
0.07	21.01	49.67
0.08	21.01	49.67
0.09	21.00	49.66
0.1	21.00	49.66

表 5-9 不同權重β結合信心度評估之最小化貝氏風險其字錯誤率(%)

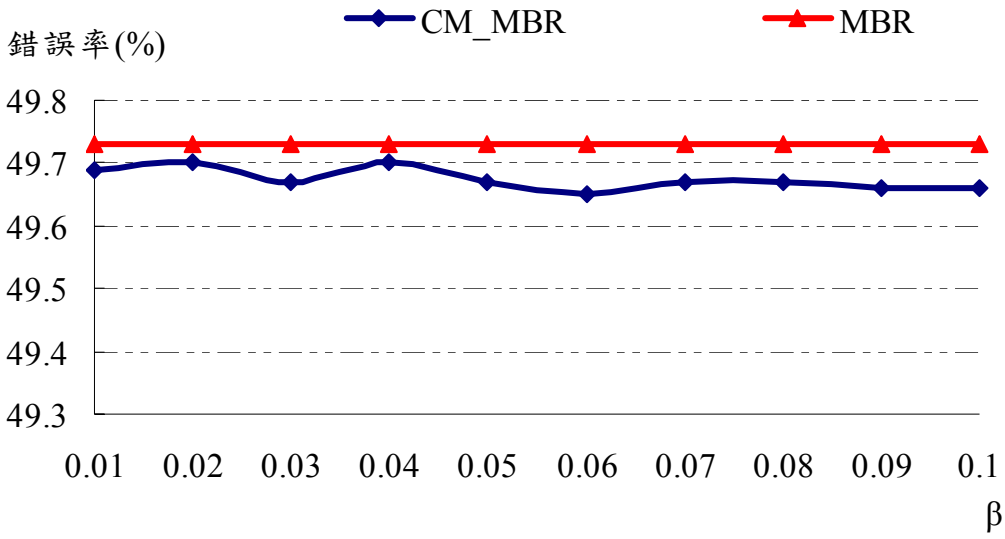

圖 5-11 外場記者:不同權重β對於結合信心度評估之最小化貝氏風險之字錯誤率曲線圖

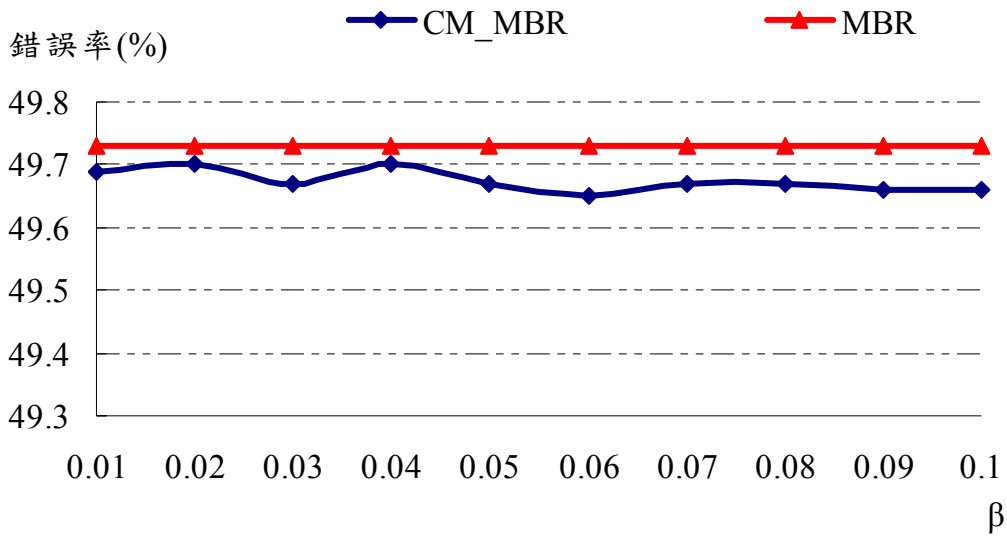

圖 5-12 外場受訪者:不同權重β對於結合信心度評估之最小化貝氏風險之字錯誤率曲線圖

【實驗討論】

由實驗結果可得知，傳統以 Levenshtein 距離為成本函式的最小化貝氏風險法則，不一定能降低語音辨識系統的錯誤率。會有如此的結果可能的原因仍在於傳統的做法主要是希望能找出一條與其它詞序列期望最相近的詞序列，然而這樣的假設並不代表一定是正確的。而在加入了信心度評估後，不論是在外場記者或外場受訪者的測試語料中，結合信心度評估的最小化貝氏風險法則都能較傳統的最小化貝氏風險法則能有些許的進步。

第6章 結論與未來展望

由於目前語音辨識的正確率仍然無法達到百分之百，該如何驗證辨識結果的正確性一直是個很重要的課題。因此，過去這十幾年來，不斷有學者在進行信心度估評的研究。更甚者，除了傳統應用於判斷辨識結果的正確與否之外，近幾年來，信心度評估的應用已越趨廣泛，舉凡非監督式聲學模型訓練、大詞彙連續語音辨識的往前觀測以及增進語音辨識系統的正確率等，都是信心度評估新興的研究方向。此外，現今信心度評估所運用的資訊除了第2章所提到的聲學、語言及語法相關特徵之外，近年來有學者嘗試運用如聲韻(Prosodic)等語言學(Linguistic)的相關知識，來降低信心度評估的信心度錯誤率及提昇語音辨識系統的正確率[Qian *et al*. 2004]。

本論文是以改進傳統的信心度評估及應用信心度評估於提昇語音辨識系統的正確率為主軸。在傳統的信心度評估應用中，提出結合熵值與傳統的信心度評估，在MATBN外場記者與外場受訪者語料中，使用詞圖熵值結合傳統事後機率之信心度評估，相對於傳統事後機率的方法，信心度錯誤率各有16.37%及12.00%的相對下降。此外，與以特徵為基礎之信心度評估比較，也有較佳的結果。這是因為熵值資訊考慮到在同一段發音區間，不應該同時有兩個以上的詞都擁有相當高的信心度。當有這種情形發生時，就代表此信心度評估較不可靠，應當降低對此信心度評估結果的可信度。而在降低語音辨識系統錯誤率的實驗中，本論文使用了最小化音框錯誤率法則，結合梅爾倒頻譜係數(MFCC)，以及異質性線性鑑別分析搭配最大相似度線性轉換(HLDA+MLLT)兩種不同語音特徵參數所形成的詞圖。由於本論文適當的結合了兩種不同語音特徵參數所形成的詞圖，使得其帶有的資訊，相較於任一單一的詞圖都來得多，在降低外場記者及外場受訪者兩套測試語料的字錯誤率實驗中，較傳統最大事後機率解碼能各有4.6%及4.8%的字錯誤率相對下降。與過去只使用單一詞圖的最小化音框錯率搜尋法相比，也都有較好的結果。最後，本論文也另外提出了在傳統利用Levenshtein距離為成本函式的最小化貝氏風險(Minimum Bayes Risk)辨識法則中，適當的加入以特徵為基礎的信心度評估，使得成本函式除了考慮兩條詞

序列的相似程度之外，也觀察詞序列的信心度高低。經由實驗得知，雖然在外場記者以及外場受訪者的語料中，對於辨識錯誤率並沒有很明顯的進步或退步，但相較於傳統利用Levenshtein距離為成本函式的最小化貝氏風險辨識法則而言，已有較佳的結果。

目前事後機率的信心度估評除了應用於詞圖之外，最近開始也有學者試著探討信心度評估於一致性網路(Consensus Network or Sausage)的情況[Fabian *et al*. 2003]。根據[Mangu *et al*. 2000]，一致性網路除了在降低語音辨識系統錯誤率有一定的效果之外，應用於信心度評估也有相當的效能。因此吾人未來也將嘗試利用本論文所提出熵值資訊配合傳統信心度評估的方法作用於一致性網路，同時討論以特徵為基礎的信心度評估於一致性網路的效果。另外，在這幾年中，也開始有研究學者進行以單連或雙連詞出現的次數等特徵，採用鑑別式訓練(Discriminative Training)的方式於*N*-最佳化詞序列重新排序(Reranking)之研究[Zhang and Rudnicky 2004; Zhou *et al*. 2006]。未來，吾人會試著以信心度評估取代上述之特徵，並結合鑑別式訓練應用於*N*-最佳化詞序列的重新排序。

參考文獻

[Abdou and Scordilis 2003] S. Abdou and M. S. Scordilis, "An Efficient Fast Matching Approach Using Posterior Probability Estimates in Speech Recognition," Proc. of European Conference on Speech Communication Technology, 2003.

[Afify *et al*. 2005] M. Afify, F. Liu, H. Jiang and O. Siohan, "A New Verification-based Fast-match for Large Vocabulary Continuous Speech Recognition," IEEE Trans. Speech and Audio processing, Vol . 13,No.4,pp 546-553, 2005.

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

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

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

[Bahl *et al*. 1986] L. R. Bahl, P. F. Brown, P. V. de Souza and R. L. Mercer, "Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition," Proc. of International Conference on Acoustic, Speech and Signal Processing, 1986.

[Barras *et al*. 1986] C. Barras, E. Geoffrois, Z. B. Wu, and M. Liberman, "Transcriber: Development and Use of a Tool for Assisting Speech Corpora Production," Speech Communication, Vol. 33, pp.5-22, 2001.

[Baum 1972] L. E. Baum, “An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes,” Inequalities, Vol. 3, No. 1, pp.1-8, 1972.

[Belllegarda 1998] J. R. Bellegarda, “A Multispan Language modeling Framework for Large Vocabulary Speech Recognition,“ IEEE Trans. on Acoustic, Speech and Signal Processing, Vol. 6, No. 5, pp. 456-467, 1998.

[Belllegarda 2000] J. R. Bellegarda, “Exploiting Latent Semantic Information in Statistical Language Modeling,” Proceedings of the IEEE, Vol. 88, pp.1279-1296, 2000.

[Bellegarda 2005] J. R. Bellegarda, “Latent Semantic Mapping,” IEEE Signal Processing Magazine, Vol.22, pp70-80, 2005.

[Benitez *et al*. 2000] M .C. Benitez, A. Rubio, and A. Torre, “Different Confidence Measures for Word Verification in Speech Recognition,” Speech Communication, Vol. 32, pp. 79–94, 2000.

[Boll 1979] S. F. Boll, “Suppression of Acoustic Noise in Speech Using Spectral Subtraction,” IEEE Trans. on Acoustic, Speech and Signal Processing, Vol. 27, No. 2, pp. 113-120, 1979.

[Chase 1997] L. Chase, “Word and Acoustic Confidence Annotation for Large Vocabulary Speech Recognition,” Proc. of European Conference on Speech Communication Technology, 1997.

[Chen and Goodman 1999] S. F. Chen and J. Goodman, “An Empirical Study of Smoothing Techniques for Language Modeling,” Computer Speech and Language, Vol. 13, pp. 359-393, 1999.

[Chen *et al.* 2004] B. Chen, J.-W. Kuo and W.-H. Tsai, “Lightly Supervised and Data-driven Approaches to Mandarin Broadcast News Transcription,” Proc. of International Conference on Acoustic, Speech and Signal Processing, 2004.

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

[Cox and Dasmahapatra 2002] S. Cox and S. Dasmahapatra, “High-level Approaches to Confidence Estimation in Speech Recognition,” IEEE Trans. Acoustic, Speech, and Signal Processing, Vol. 10, No. 7, pp.460-471, 2002.

[Cox and Rose 1996] S. Cox and R. Rose, “Confidence Measures for the Switchboard Database,” Proc. of International Conference on Acoustic, Speech and Signal Processing, 1996.

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

[Eide *et al*. 1995] E. Eide, H. Gish, P. Jeanrenaud and A. Mielke, “Understanding and Improving Speech Recognition Performance Through the Use of Diagnostic Tools,” Proc. of International Conference on Acoustic, Speech and Signal Processing, 1995.

[Fabian *et al*. 2003] T. Fabian, R. Lieb, G. Ruske and T. Thomae, “Impact of Word Graph Density on Quality of Posterior Probability Based Confidence Measures,” Proc. of European Conference on Speech Communication Technology, 2003.

[Fabian *et al*. 2005] T. Fabian, R. Lieb, G. Ruske and T. Thomae, “A Confidence-guided Dynamic Pruning Approach – Utilization of Confidence Measurement in Speech Recognition,” Proc. of European Conference on Speech Communication Technology, 2005.

[Furnas *et al*. 1988] G. W. Furnas, S. Deerwester, S. T. Dumais, T. K. Landauer, R. A. Harshman, L. A. Streeter and L. E. Lochbaum, ”Information Retrieval Using a Singular Value Decomposition Model of Latent Semantic Structure,” Proc. of International Conference on Research and Development in Information Retrieval, pp 465-480, 1988.

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

[Goel and Byrne 2000] V. Goel and W. Byrne, “Minimum Bayes-risk Automatic Speech Recognition,” Computer Speech and Language, Vol. 14, pp.115-135, 2000.

[Gopinath 1998] R. A. Gopinath, “Maximum Likelihood Modeling with Gaussian Distributions,” Proc. of International Conference on Acoustic, Speech and Signal Processing, 1998.

[Guo *et al*. 2004] G. Guo, C. Huang, H. Jiang and R.-H. Wang, “A Somparative Study on Various Confidence Measures in Large Vocabulary Speech Recognition,” Proc. of International Conference on Spoken Language Processing, 2004.

[Hazen *et al*. 2002] T. J. Hazen, S. Seneff, and J. Polifroni, “Recognition Confidence Scoring and Its Use in Speech Understanding Systems,” Computer Speech and Language, Vol. 16, pp.49-67, 2002.

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

[Jelinek 1999] F. Jelinek, “Statistical Methods for Speech Recognition,” the MIT press, 1999.

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

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

[Katz 1987] 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.

[Kamppari and Hazen 2000] S. O. Kamppari and T. J. Hazen, “Word and Phone Level Acoustic Confidence Scoring,” Proc. of International Conference on Acoustic, Speech and Signal Processing, 2000.

[Kemp and Schaaf 1997] T. Kemp and T. Schaaf, “Estimating Confidence Using Word Lattice,” Proc of European Conference on Speech Communication Technology, 1997.

[Korkmazsky 2004] F. Korkmazsky, D. Fohr and I. Illina, “Using Linear Interpolation to Improve Histogram Equalization for Speech Recognition,” Proc. of International Conference on Spoken Language Processing, 2004.

[Lane and Kawahara 2005] I. R. Lane and T. Kawahara, “Utterance Verification Incorporating In-domain Confidence and Discourse Coherence Measures,” Proc. of European Conference on Speech Communication Technology, 2005.

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

[Lo and Soong 2005] W. K. LO and F. K. Soong, “Generalized Posterior Probability for Minimum Error Verification of Recognized Sentences,” Proc. of International Conference on Acoustic, Speech and Signal Processing, 2005.

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

[Neti *et al*. 1997] C. V. Neti, S. Roukos, E. Eide, “Word-based Confidence Measures as a Guide for Stack Search in Speech Recognition,” Proc. of International Conference on Acoustics, Speech and Signal Processing, 1997.

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

[NIST] National Institute of Standards and Technology. http://www.nist.gov/.

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

[Ortmanns *et al*. 1997] S. Ortmanns, H. Ney and X. L. Aubert, “A Word Graph Algorithm for Large Vocabulary Continuous Speech Recognition,“ Computer Speech and Language, Vol. 11, pp.43-72, 1997.

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

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

[Qian *et al*. 2004] Y. Qian, T. Lee and F. K. Soong, "Tone Information as a Confidence Measure for Improving Cantonese LVCSR,” Proc. of International Conference on Acoustic, Speech and Signal Processing, 2004.

[Rabiner 1989] L. R. Rabiner, “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,” Procedings of the IEEE, Vol. 77, No. 2, 1989.

[Razik *et al*. 2005] J. Razik, O. Mella, D. Fohr and J. P. Haton, “Local Word Confidence Measure Using Word Graph and N-best List,” Proc. of European Conference on Speech Communication Technology, 2005.

[Rose *et al*. 1995] R. C. Rose, B. H. Juang and C.-H. Lee, “A Training Procedure for Verifying String Hypothesis in Continuous Speech Recogniton,” Proc. of International Conference on Acoustic, Speech and Signal Processing, 1995.

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

[Sanchis *et al*. 2003] A. Sanchis, A. Juan and E. Vidal, “Improving Utterance Verification Using a Soomthed Naïve Bayes Model,” Proc. of International Conference on Acoustic, Speech and Signal Processing, 2003.

[Sanchis *et al*. 2004] A. Sanchis, A. Juan, and E. Vidal, “New Features Based on Multiple Word Graphs For Utterance Verification,” Proc. of International Conference on Spoken Language Processing, 2004.

[San-Segundo *et al.* 2001] R. San-Segundo, B. Pellom, K. Hacioglu and W. Ward, “Confidence Measures for Spoken Dialogue System,” Proc. of International Conference on Acoustic, Speech and Signal Processing, 2001.

[Saon *et al*. 2000] G. Saon, M. Padmanabhan, R. Gopinath and S. Chen, “Maximum Likelihood Discriminant Feature Spaces,” Proc. of International Conference on Acoustic, Speech and Signal Processing, 2000.

[Schaaf and Kemp 1997] T. Schaaf and T. Kemp, “Confidence Measure for Spontaneous Speech Recognition,” Proc. of International Conference on Acoustic, Speech and Signal Processing, 1997.

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

[SRILM 2000] A. Stolcke, “SRI language Modeling Toolkit,” version 1.3.3, http://www.speech.sri.com/projects/srilm/

[Stolcke *et al*. 1997] A. Stolcke, Y. Konig and M. Weintraub, “Explicit Word Error Rate Minimization in N-Best List Rescoring,” Proc of European Conference on Speech Communication Technology. 1997.

[Tseng and Liu 2001] S.-C. Tseng and Y.-F. Liu, ”Mandarin Conversational Dialogue Corpus. MCDC,” Technical Note 2001-01. Institute of Linguistics, Academia Sinica, Taipei.

[Uhrik and Ward 1997] C. Uhrik and W. Ward, “Confidence Metrics Based on N-gram Language model Backoff Behaviors,” Proc of European Conference on Speech Communication Technology. 1997.

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

[Viterbi 1967] A. J. Viterbi, “Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm,” IEEE Trans. Information Theory, Vol. 13, No. 2, 1967.

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

[Wessel *et al*. 2000] F. Wessel, R. Schlüter and H. Ney, “Using Posterior Word ProBabilities for Improved Speech Recognition,” Proc. of International Conference on Acoustics, Speech and Signal Processing, 2000.

[Wessel *et al*. 2001] F. Wessel, R. Schlüter, K. Macherey and H. Ney, “Confidence Measure for Large Vocabulary Continuous Speech Recognition,” IEEE Trans. Speech and Audio Processing, Vol.9, No. 3, pp.288-298, 2001.

[Wessel *et al*. 2001b] F. Wessel, R. Schlüter, K. Macherey and H. Ney, “Explicit Word Error Minimization Using Word Hypothesis Posterior Probabilities,” Proc. of International Conference on Acoustic, Speech and Signal Processing, 2001.

[Wessel and Ney 2005] F. Wessel and H. Ney, “Unsupervised Training of Acoustic Models for Large Vocabulary Continuous Speech Recognition,” IEEE Trans. Speech and Audio Processing, Vol.13, No. 1, pp.23-31, 2005.

[Wilpon *et al*. 1990] J. G. Wilpon, L. R. Rabiner, C.-H. Lee and R. Goldman, “Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models,” IEEE Trans. Acoustics Speech Signal Process, Vol.38, No.11,pp.1870-1878, 1990.

[Young 1994] S. R. Young, “Detecting Misrecognition and Out-of-vocabulary Words,” Proc. of International Conference on Acoustic, Speech and Signal Processing, 1995.

[Young *et al.* 2002] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. C. Woodland. *The HTK Book.* Version 3.2, 2002. http://htk.eng.cam.ac.uk/.

[Zhang and Rudnicky 2001] R. Zhang and A. I. Rundicky, “Word Level Confidence Annotation Using Combination of Features,” Proc of European Conference on Speech Communication Technology, 2001.

[Zhang and Rudnicky 2004] R. Zhang and A. I. Rundicky, “Apply N-Best List Re-ranking to Acoustic Model Combinations of Boosting Training,” Proc. of International Conference on Spoken Language Processing, 2004.

[Zhou *et al*. 2006] Z. Zhou, J. Gao, F. K. Soong and H. Meng, “A Comparative Study of Discriminative Methods for Reranking LVCSR N-best Hypotheses in Domain Adaptation and Generalization,” Proc. of International Conference on Acoustic, Speech and Signal Processing, 2006.

[郭人瑋 2005] 郭人瑋, ”最小化音素錯誤鑑別式聲學模型學習於中文大詞彙連續語音辨識之初步研究,” 國立台灣師範大學資訊工程所碩士論文, 2005.