國立臺灣師範大學

資訊工程研究所碩士論文

指導教授：陳柏琳 博士

英文連續語音辨識之初步研究

An Initial Study on English Continuous Speech Recognition

研究生：許庭瑋 撰

中華民國 九十六 年 七 月

摘要

本論文為英文連續語音辨識之初步研究。我們實作英文連續語音辨識器，並探討其主要組成，包含語音特徵擷取、聲學模型及語言模型等。首先，針對語音特徵擷取，我們比較傳統式梅爾倒頻譜係數(Mel-frequency Cepstral Coefficients, MFCC)與線性鑑別分析(Linear Discriminant Analysis, LDA)和異質性線性鑑別分析(Heteroscedastic Linear Discriminant Analysis, HLDA)之效能。再者，針對聲學模型，我們探討詞內三連音素模型(Intra-word Triphone Models)、狀態連結(State-Tying)技術、音素模糊矩陣(Phone Confusion Matrix)與非監督式聲學模型訓練(Unsupervised Acoustic Model Training)的使用，以提升語音辨識率。最後，針對語言模型，我們在語音辨識過程中分別利用詞頻數混合法(Count Merging)與模型插補法(Model Interpolation)，結合背景與同領域語言模型訓練語料，以達到較佳之詞發生預測。本論文實驗是以美國之音與台灣腔英文語料為題材，並有一些初步的觀察及發現。

Abstract

This thesis is intended to perform a preliminary study on English continuous speech recognition. An English continous speech recognizer was implemented, while parts of its major constituents, including speech feature extraction, acoustic modeling and language modeling, were extensively investigated as well. First, for speech feature extraction, we compared the performance of linear discriminant analysis (LDA) and heteroscedastic linear discriminant analysis (HLDA) to that of the conventional Mel-frequency cepstral coefficients (MFCC) .Second, for acoustic modeling, we explored the use of the intra-word triphone models, the state-tying scheme and the phone confusion matrix, as well as the unsupervised training of acoustic models, for better speech recognition results. Finally, for language modeling, both count-merging and model-interpolation approaches were respectively expoited to combine the background and in-domain language model training corpora to enable better prediction of word occurrences during the speech recognition process. The experiments were conducted on the Voice of America (VOA) and the English Across Taiwan (EAT) corpora.

誌謝

非常感謝陳柏琳教授在碩士兩年給予我的指導，老師在專業領域上的出類拔萃、教學的用心及嚴以律己的生活態度，是我望塵莫及和永遠的榜樣。每當學生在研究上遇到瓶頸，老師總是耐心的傾聽、並給我機會學習和精闢的指導，讓我從無知的小毛蟲慢慢蛻變，縱使不是最美麗的蝴蝶，但是有能力開始學飛，在未來的人生旅程中謹記老師的諄諄教誨。

感謝口試委員張森嘉博士、高照明教授、陳浩然教授與洪志偉教授對我論文的指正和建議，使我的論文更臻完善。

感謝實驗室人瑋學長和斯涵學妹，很有緣分在不同時間點認識你們，在知識和生活中給予我建議和鼓勵。感謝文鴻、耀民、成韋、志豪及惠銘學長，帶領我認識語音的各項領域。感謝士傑學長為我們帶來消息和打氣。感謝怡婷學姊、燦輝、士弘和士翔學長，你們優秀的實力和口才，是我仰慕與學習的榜樣。特別感謝炫盛學長，一路到最後耐心的指導，每每遇到困難總是能在你身上尋求到方向，最後迎刃而解，讓我的學識在這兩年提升許多。感謝鴻彬和芳輝在修業和實驗研究上激發我的思考，給予我莫大的協助。感謝鴻欣學弟為實驗室帶來理性與感性的活水。語音實驗室是我這輩子永遠不會忘記的第二個家。

感謝摯愛的阿公、阿媽、外婆和在研究所期間逝世的外公，感謝您們給我愛與力量。感謝爸爸、媽媽多年來的養育與栽培之恩，讓我無後顧之憂、順利完成學業。感謝哥哥、軒軒指引我決策方向。感謝漢洲一路的支持和陪伴，讓我的生命更加精采。

這本論文的完成，沒有大家無盡的愛與包容，我想我沒有今日。

謝謝大家 庭瑋 謹誌

目次

表目錄

圖目錄

第1章　緒論

1.1　研究動機

隨著科技的日新月異，電腦處理速度突飛猛進，消費性電子產品的發明帶領人類邁向更便捷的生活模式。與這些電子產品溝通的方法，除了有一般電視、電話、與電腦鍵盤的按鍵模式，還有指紋機、手寫板等觸控模式，聰明的人類思考並尋求其他種更簡易的溝通方法。其中利用「語音」輸入，已成為劃時代的研究議題。因為語言是亙古以來人類仰賴彼此溝通、了解最自然快速的重要工具，目前已知世界上有多達數千種不同的人類語言，如果再加上動物界其他聲音(如海豚發出的聲音、火車經過的聲與下雨聲等)，這些種種聲音，都可當作辨識的圖案(Pattern)或碼(Code) 。如果我們能直接透過語音操作電子設備，且電腦能夠理解我們的要求，做適當的處理，將能節省許多人力和時間。

語音辨識的研究發展，從 1952 年美國貝爾實驗室發展的獨立數字辨識(Isolated-Digit Recognition)，之後隨著演算法、電腦速度的進步，由數字單詞、關鍵詞擷取(Keyword Spotting)[Wilpon *et al.* 1990]演進到口語對話系統(Spoken Dialogue)[Hazen *et al.* 2002]、大詞彙連續語音辨識(Large Vocabulary Continuous Speech Recognition, LVCSR)以及其他語音檢索(Speech-based Information Retrieval)的應用，可見語音辨識研究的蓬勃發展。

目前語音辨識遇到的重要問題是：語音辨識的正確率並沒有達到 100%，這也是多年來語音辨識專家、學者致力研究的核心重點。另一方面，英語是邁入國際化必要學習的重要語言，世界各國爭相將英語加入到國人必要學習的第二外語，然而非英語系國家在學習其他語言，可能因擁有第一語言的發音特性或習

慣，故在學習英語上會產生不同的發音腔調或變異，本論文初步探討台灣腔英語之連續語音辨識的情況與發音變異。

1.2 語音辨識流程

語音辨識流程簡單的說，將輸入的語音訊號，輸出成對應的文字或語音。然而若要達成此目的，則需經多重複雜步驟，如前端處理(Front-End Processing)、聲學比對(Acoustic Matching)與語言解碼(Linguistic Decoding)等細部運算。其中在聲學比對與語言解碼部分，需準備使用聲音語料訓練過的聲學模型(Acoustic Model)、使用文字語料訓練過的語言模型(Language Model)，以及經前端處理轉換過的語音特徵，以產生最相符對應的辨識文句輸出。基本語音辨識流程如圖 1-1 所示：

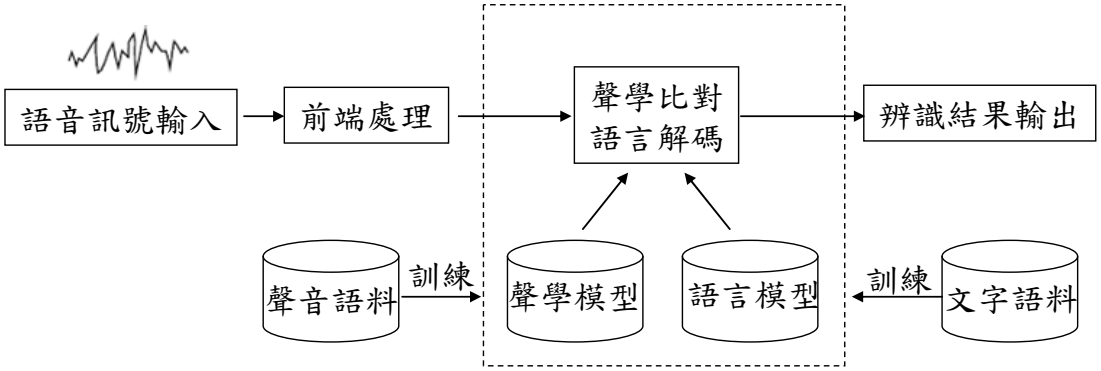


圖 1-1　基本語音辨識流程圖

以數學式來看[Jelinek 1999]，一段輸入的語音訊號段落以 O 表示，其中 O 可表示成語音特徵向量序列 $o_1, o_2, ..., o_T$，經由語音辨識器(Recognizer)辨識成對應的文字詞序列以 $\hat{W}$ 表示，為一連串詞 $w_1, w_2, ..., w_m$ 組成，語音辨識的過程即為找出

具有最大事後(Maximum A Posteriori, MAP)機率的詞序列，也就是代表O最有可能的對應輸出文句$\hat{W}$，可表示成式(1-1)：

$$\begin{aligned}\hat{W} &= \arg\max_{W} P(W|O) \\ &= \arg\max_{W} \frac{P(W)p(O|W)}{p(O)} \\ &= \arg\max_{W} P(W)p(O|W)\end{aligned} \tag{1-1}$$

$p(W|O)$為給定語音段落O時，詞序列W的事後機率。經過貝氏定理(Bayes Theory)轉換，可表示成$P(W)$、$p(O|W)$與$p(O)$。其中$p(O|W)$代表聲學模型(Acoustic Model)產生語音段落O的機率密度函數(Probability Density Function, PDF)，直接估測語音段落O發生在詞序列W對應的聲學模型相似度(Likelihood)；$P(W)$代表語言模型(Language Model, LM)產生詞序列W的機率，用於評估詞序列W於自然語言的合理性，可視為詞序列W的事前機率。語言模型輔助解決聲學上之混淆，使得最後選出的詞序列$\hat{W}$能夠符合該語言特性。另一方面，$p(O)$代表語音O之事前機率密度，因為對某句語音O進行辨識，每條詞序列都同除以$p(O)$，故可忽略。本論文使用連續密度隱藏式馬可夫模型(Continuous Density Hidden Markov Model,CDHMM)[Rabiner *et al.* 1989]作為聲學模型、N-連(N-gram)模型作為語言模型。

在語音辨識過程中，聲學比對負責將音素及語句中每個可能的段落做比對，計算其相似度；語言解碼使用維特比動態規劃搜尋(Viterbi Dynamic Programming Search) [Viterbi 1967]，對聲學相似度和語言機率進行解碼以便找出機率最大的可能詞序列。然而搜尋過程會隨著模型愈複雜，搜尋空間也呈現指數成長，故為了降低搜尋複雜度，本論文利用兩階段的搜尋來完成：第一階段進行聲學比對，並使用較低階的語言模型來搜尋，以產生詞圖(Word Graph)，第二階段在詞圖上使

用較高階的語言模型進行重新搜尋(Rescoring) [Ortmanns *et al.* 1997]。

1.2.1 特徵擷取 (Feature Extraction)

語音訊號前端處理，目的是擷取出合適的語音特徵參數，目前廣為人知的特徵參數有：梅爾倒頻譜係數(Mel Frequency Cepstral Coefficients, MFCC)[Davis *et al*. 1980]，線性預測係數(Linear Prediction Coefficients, LPC)[Gray *et al*. 1973]與感知線性預測係數(Perceptual Linear Prediction Coefficients, PLPC)[Hermansky 1990]等。本論文之特徵擷取實驗以梅爾倒頻譜係數(MFCC)為基礎，其擷取參數流程如圖 1-2：

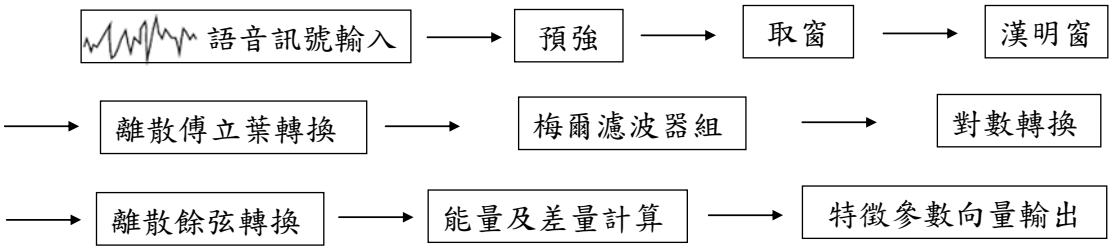


圖 1-2　梅爾倒頻譜係數之前端特徵擷取步驟

以下概述梅爾倒頻譜係數語音前端擷取過程如下：

1. 預強(Pre-emphasis)：

   因語音在空氣中傳送，或於發聲過程中聲門(Glottal)壓抑高頻部分能量，導致高頻的能量會快速遞減，故預強的功能是模擬人耳外聽道的功能，在時域上強調高頻能量。

$$H(z) = 1 - \alpha \cdot z^{-1} \tag{1-2}$$

式(1-2)可用來表示預強，其中 $H(z)$ 為高通濾波器在Z轉換(Z-Transform)的表示。實作上可在時域上處理，如式(1-3)，其中 $s(n)$ 為第 n 個採樣點，$\hat{s}(n)$ 為第 n 個採樣點經預強後的值。α 為預強參數，本論文設定為0.975。

$$\hat{s}(n) = s(n) - \alpha \cdot s(n-1) \tag{1-3}$$

2. 取框(Framing)：

在時域上觀測語音訊號的波形變化為十分迅速且無一定規則，但於頻域上觀察，則可以發現短時間(20ms~40ms)的情況下頻譜具有週期性的改變，所以在語音辨識的前處理，會假設語音訊號為短時間穩定(Short Time Stationary)，所以每隔一小段時間對語音訊號取一音框(Frame)，為了讓音框與音框之間聯繫著關係，所以音框與音框間會重疊(Overlap)一小段時間，此動作稱為取框(Framing)。本論文設定一個音框長為20ms，音框間重複為 10ms。

3. 漢明窗(Hamming Window)：

將時域的每個音框經離散傅立葉轉成頻域的訊號，但由於每個音框是固定時間點切割，所以音框左、右端的邊緣會造成訊號不連續現象，使得頻域上產生摺積效果，故在離散傅立葉轉換前會乘上一個漢明窗，特性在於主辦葉(Main Lobe)較寬，邊葉(Side Lobe)較窄，因此能有效壓抑訊號兩端，聚集中央部份特徵。漢明窗公式如式(1-4)，其中 α 為控制漢明窗之參數，本論文設定為 0.46。

$$w(n) = \begin{cases} (1-\alpha) - \alpha\cos\left(\dfrac{2\pi n}{N-1}\right) & n = 0,1,...,N-1 \\ 0 & otherwise \end{cases} \tag{1-4}$$

4. 離散傅立葉轉換(Discrete Fourier Transform, DFT)：

語音訊號在時域上的變化迅速且會隨著時間不斷改變，故不容易觀察出週期性的變化。為了找出語音訊號特性，故將語音訊號由時域轉成頻域，因為短時間內語音訊號在頻域上的能量分布是有規則性的，故一般可經由離散傅立葉轉換達成。

5. 梅爾頻率濾波器組(Mel-frequency Filter Bank)：

人耳對於聲音的高頻與低頻敏感度不同，在低頻部分人耳感受比較敏銳，而在高頻部分人耳的感受較不敏銳，因在耳蝸中不同位置的感受器連結到不同的反應神經，不同的反應神經代表不同的反應頻率，梅爾濾波器組主要是模擬人耳內部基底膜(Basiler Membrane)傳遞刺激到聽覺神經的過程。

6. 對數轉換(Logarithm)：

因人耳對於音強大小有不同解析度，為了保護人耳不受傷害，對音強小的聲音解析度較高，音強大的聲音有壓抑功能，使解析度較低，為模擬人耳此項功能，此將步驟 5 濾波器輸出取對數轉換。

7. 離散餘弦轉換(Discrete Cosine Transform, DCT)：

對數轉換後的梅爾頻率濾波器組的輸出，再經離散餘弦轉換成為梅爾倒頻譜係數，用意為降低維度間的關係，有助於隱藏式馬可夫模型在儲存共變異矩陣時資料的縮減，並可加快辨識效率，本論文之梅爾倒頻譜係數為 12 維。

8. 能量及差量計算(Log Energy and Time Derivatives)：

不同的音素(Phoneme)在能量(Energy)上的差異很大，由此可知能量為一重要的聲學特徵，一般會把能量與梅爾倒頻譜特徵結合。故於梅爾倒頻譜係數加上能量維後共 13 維，並加入各 13 維的一階與二階的差量計

算後，總共 39 維的 MFCC 語音特徵向量。

1.2.2 聲學模型 (Acoustic Model)

聲學模型的建立，首先對訓練語料中出現的各種語音訊號建立隱藏式馬可夫(Hidden Markov Model)之聲學模型，此模型中定義了四種參數：S、Π、A 與 B，其中 $S=\{s_1,...,s_n\}$ 表示每種模型中存有可能的狀態(State)，且此狀態中包含事件，這些事件可能為離散的事件或是連續事件，在連續事件中通常以高斯分布(Gaussian Distribution)來表示事件的分布情況，即為訓練語料中定義音素的分布情形，分布數目可能大於一，稱為高斯混合模型(Gaussian Mixture Model,GMM)。$\Pi=\{\pi_1,...,\pi_n\}$ 表示進入模型狀態的初始機率(Initial Probability)。$A=\{a_{ij}\}$ 代表任意狀態 s_i 與狀態 s_j 之間的轉換機率(Transition Probability)，$B=b_j(o_t)$ 代表語音特徵向量 o_t 在任意狀態 s_j 的觀測機率(Observation Probability)，如圖 1-3 代表單連(Monophone)音素 ax 對應的隱藏式馬可夫聲學模型：

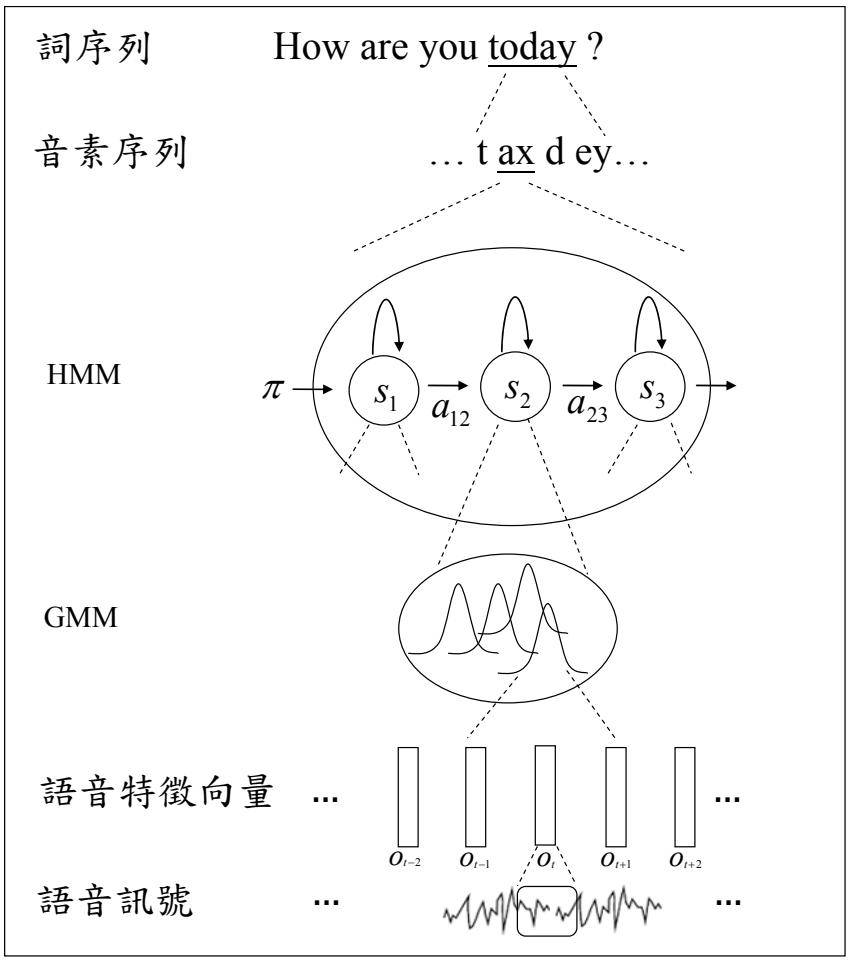


圖 1-3　單連音素 ax 之隱藏式馬可夫聲學模型

式(1-1)中 $p(O|W)$ 代表聲學模型分數，是假設經前端處理的語音訊號特徵向量序列 O（O 是經由一連串語音特徵向量 $o_1,o_2,...o_T$ 組成），特定文句 $W=w_1,w_2,...w_m$ 所對應的聲學模型下出現的機率。本論文使用英文訓練語料來訓練單連音素(Monophone)、二連音素(Biphone)、三連音素(Triphone)之聲學模型。聲學模型是利用最大化相似度訓練法(Maximum Likelihood Estimation, MLE) [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]訓練而得。我們可使用訓練好的單連音素、二連音素、三連音素之模型來構成詞或文句的對應聲學模型。

1.2.3 語言模型 (Language Model)

式(1-1)中 $P(W)$ 為文句 W 的語言模型分數。一篇文章出現某個詞的機率，可能與過去(History)出現的詞有關，因此式(1-1)的 $P(W)$ 可進一步利用連鎖率(Chain Rule)表示成式(1-5)：

$$\begin{aligned} P(W) &= P(w_1, w_2, ..., w_m) \\ &= P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1, w_2)...P(w_m \mid w_1, w_2, ..., w_{m-1}) \\ &= P(w_1)\prod_{i=2}^{m} P(w_i \mid w_1, w_2, ..., w_{i-1}) \end{aligned} \tag{1-5}$$

若再進一步於式(1-5)中，假設目前的詞與過去出現過的詞無關，稱為單連詞(Unigram)，如式(1-6)所示。

$$P(W) = P(w_1)\prod_{i=2}^{m} P(w_i \mid w_1, w_2, ..., w_{i-1}) \approx \prod_{i=1}^{m} P(w_i) \tag{1-6}$$

如果目前的詞和過去緊鄰出現過的 N-1 個詞有關、和其他詞無關，稱為 N 連詞(N-gram)，如式(1-7)所示。

$$\begin{aligned} P(W) &= P(w_1)\prod_{i=2}^{m} P(w_i \mid w_1, w_2, ..., w_{i-1}) \\ &\approx P(w_1)\prod_{i=2}^{m} P(w_i \mid w_{i-N+1}, ...w_{i-2}, w_{i-1}) \end{aligned} \tag{1-7}$$

式(1-7)中機率模型可利用大量的文字訓練語料，利用最大化相似度估側法(MLE)，訓練而得。而如果有些詞沒有出現在訓練語料中，可用模型平滑化(Smoothing)技術，如 Katz [Katz 1987]、Kneser-Ney [Ney *et al*. 1994]與 Witten-Bell [Witten *et al.* 1991]等語言模型平滑化方法，使得在訓練語料中有出現的 N 連詞折扣(Discount)部分次數給在訓練語料中未出現過的 N 連詞，以解決機率值為 0 的情況[Chen and Goodman 1999]。

圖 1-4 是表示以詞「today」為現在目前出現的詞，往前看零個詞、一個詞(you)及二個(are、you)緊鄰出現過詞的情況：

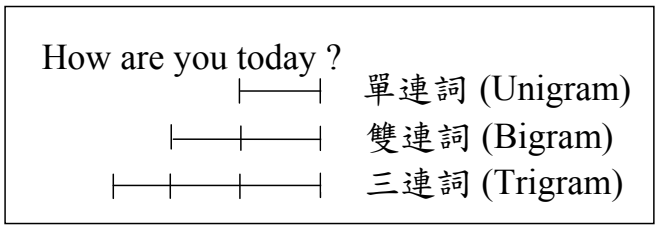


圖 1-4　以詞「today」為例的 *N*-gram 語言模型

1.2.4　語言解碼 (Linguistic Decoding)

依式(1-1)尋找最佳詞序列，備好所需之聲學模型、語言模型以及對應詞典(Lexicon)，使用維特比動態規劃搜尋法(Viterbi Dynamic Programming Search) [Viterbi 1967]找出輸入的語音訊號對應的最佳詞序列。由於搜尋空間會隨詞典與語言模型模型複雜度(例如詞二連、詞三連語言模型)大小呈指數成長，因此在搜尋時，通常會透過路徑裁減(Pruning)技術，停止搜尋機率較低的詞序列路徑，以降低計算複雜度與記憶體使用量。

1.3　研究內容

本論文研究內容，主要為：

1. 建立英文詞內(Intra-word)單連音素、二連音素、與三連音素(Triphone)狀態分享(State-tying)之聲學模型，觀察不同的聲學模型對語音辨識率的影響。

2. 探討台灣腔英語(EAT)、美國之音(VOA)兩套英文語料之連續語音辨識；於辨識器各階段中，分別使用鑑別性特徵擷取法、增加聲學模型之高斯混合數、

調適背景語言模型來提高語音辨識率。

3. 利用模糊矩陣尋找台灣腔英文發音變異，基於觀察之變異情況來修改訓練聲學模型狀態分享規則問題條件，重新修改狀態分享列表。另一方面，於語音辨識搜尋階段修正語音向量在隱藏式馬可夫模型的狀態之觀測機率，以改善系統辨識率。

4. 探討非監督式聲學模型訓練，首先對大量語料進行語音辨識，並使用語料及經辨識後自動轉寫文字(Transcription)資訊重新訓練聲學模型。

1.4 論文大綱

本論文大綱如下：

第二章　回顧現階段國外研究機構發展英文大詞彙連續語音辨識系統概況，以及聲學模型音素單位相似度測量方法。

第三章　介紹本論文實驗所用之音素、詞典與語料設定，以及所用之台師大大詞彙連續語音辨識系統。

第四章　詞內單連、二連、三連音素聲學模型訓練與兩組英文實驗語料之基礎實驗結果。

第五章　改善英文連續語音辨識之討論：使用鑑別式特徵擷取方法、語言模型調適方法、應用模糊矩陣於聲學模型訓練與語音辨識，最後討論非監督式最大化相似度之聲學模型訓練之方法與實驗結果。

第六章　結論與未來展望。

第2章　文獻回顧

2.1　現階段英文語音辨識研究內容

語音辨識系統的發展，從 1952 年美國貝爾實驗室發展的小詞彙獨立數字辨識，之後隨著演算法、電腦速度的進步，由數字單詞演進到任意口語連續語句的大詞彙連續語音辨識(Large Vocabulary Continuous Speech Recognition, LVCSR)系統，其中大詞彙連續語音辨識系統，從 1990 年初開始發展至今已有 15 年以上的歷史，目前較著名國外發展語音辨識器的主要學術單位、科技公司與機構如表 2-1 所示：

表 2-1　國外發展語音辨識器之學術單位、科技公司與機構

1	美國麻薩諸塞州 BBN 科技公司
2	美國 IBM 華生(T.J. Watson)研究中心
3	英國劍橋大學電機系(Cambridge University Engineering Dept.)
4	美國卡內基美隆大學電腦科學學校 (Carnegie Mellon University - School of Computer Science)
5	美國麻薩諸塞州 Dragon Systems 科技公司
6	法國 LIMSI-CNRS (Centre National de la Recherche Scientifique)機構
7	美國加州 SRI 國際機構之語音科技和研究實驗室 (Speech Technology and Research Laboratory)
8	美國 AT&T 實驗室
9	美國密西西比州 MsState – ISIP 學術機構
10	美國微軟(Microsoft)科技公司

語音辨識器，可能因為訓練語料、測試語料、詞典的不同而導致不同的語音辨識率，故美國國家標準技術局(National Institute of Standards and Technology,

NIST) [NIST 2007]自 1996 年起，提供標準的對話電話語料(Conversational Telephone Speech, CTS)，並每年舉辦語者辨識評比(Speaker Recognition Evaluation, SRE)競賽。NIST SRE 是目前國際公認的語者辨識評比基準(Benchmark)。評比項目包含有：

1. 單一語者偵測(Single-Speaker Detection)：判斷一段語音是否為假設語者(Hypothesized Speaker)的語音，相當於語者驗證(Speaker Verification)。

2. 雙語者偵測(Two-Speaker Detection)：從一段二人對話中判別假設語者是否在其中。

3. 語者分段(Speaker Segmentation)：經由找出一段語音中各語者的聲音段落，進而將這些聲音段落依據語者分群。

4. 語者追蹤(Speaker Tracking)：將一段語音中屬於某一假設語者的段落一一標出。

從 2002 年 3 月開始，美國國際電腦科學組織(International Computer Science Institution, ICSI)的語音研究團隊著手進行美國國防部先進研究計畫機構(DARPA)委託的 EARS(Effective Affordable Reusable Speech-to-text Program)計畫[EARS]，計畫目的為在接下來五年內，能讓多語辨識器的辨識率達到 10%以下的詞錯誤率(Word Error Rate, WER)，而主要的辨識語料為廣播新聞(Broadcast News, BN)與人類對話(Conversational Speech)語料。EARS 計畫主要包含有兩個子計畫，分別為大量轉寫文字(Rich Transcription)與創新方法(Novel Approaches)之研究，其中大量轉寫文字之研究主要為設計適當的評比語料，供辨識器研究者做測試，例如 RT03、RT04 等語音評比語料。

語音辨識之訓練語料的來源有很多，美國語言資料協會(Linguistic Data

Consortium, LDC)[LDC]為國際非營利組織之語言相關的教育研究及科技發展機構，提供有關於 Switchboard、Switchboard Cellular 及 Callhome 等語音語料。在 EARS 計畫中，就有幾千小時的語音資料來自於 LDC，這些語料被稱為費雪集合 (Fisher Collection)。

本論文將於 2.1.1 至 2.1.3 擇要簡介其中 BBN、IBM、與劍橋大學三家著名研究機構目前在英文大詞彙連續語音辨識器之發展現狀。

2.1.1 美國 BBN 科技公司

美國 BBN 科技公司於 2005 年，與法國 LIMSI 機構共同發表的「2004 BBN/LIMSI 英文對話電話語料辨識系統」(2004 BBN/LIMSI English Conversational Telephone Speech Recognition System)[Nguyen *et al.* 2005; Prasad *et al.* 2005; Matsoukas *et al.* 2002; Colthurst *et al.* 2000]，此系統需 20 倍時間(Real-time)，評比語料為 RT04，詞錯誤率(WER)為 13.5%。

1. 語料來源：

   訓練聲學模型的聲音語料為 2,300 小時的費雪集合中 Switchboard I and II、CallHome、Cellular 語料；而語言模型的文字語料為詞彙量 27M 同聲學模型的對話電話語料、260.3M 的廣播新聞文字語料、115.9M 的 CNN 文字語料與 525M 經由美國華盛頓大學蒐集的網路文字語料。

2. 前端語音特徵擷取：

   利用 Vocal Track Length Normalization [Molau *et al.* 2001]，解決聲腔長度因人而異的變異性，此為語者正規化(Speaker Normalization)的前端處理技術，目的為調整測試語音的線性頻率尺度，符合原本訓練語料之頻率特性。而此系統利用感知線性預測技術(Perceptual Linear Prediction

Coefficients, PLPC)擷取出 14 維倒頻譜係數、並加入第 15 能量維、音框重疊為 10ms，且用倒頻譜平均消去法(Cepstral Mean Subtraction, CMS)增強語音特性並減少噪音干擾，最後各取一階、二階與三階導數相加而成共 60 維度的語音特徵向量。

3. 聲學模型訓練：

系統之聲學模型分別訓練語者獨立(Speaker-Independent)與語者調適(Speaker-Adaptive)之模型，而此兩種模型都會利用最小音素錯誤(Minimum Phone Error, MPE)[Povey 2004]訓練法做模型訓練。

A. 最大相似度-語者獨立模型訓練(Maximun Likelihood- Speaker Independent model, ML-SI model)：

將前端擷取之 60 維特徵向量，利用異質性線性鑑別分析法(Heteroscedastic Linear Discriminant Analysis, HLDA) [Kumar 1997] [Kumar and Andreou 1998]作投影、降維成 46 維特徵向量，之後利用期望最大(Expectation Maximization, EM) [Dempster *et al.* 1977]演算法來訓練三個語者獨立三連音素聲學模型，分別是：

I. 連結狀態混合(State Tied Mixture, STM)模型。

II. 分群連結狀態混合(State Clustered Tied Mixture, SCTM)模型。

III. 詞間分群連結狀態混合(Cross-word SCTM)模型。

B. 最大相似度-異質性線性鑑別分析-語者調適模型訓練(Maximun Likelihood- Speaker-dependent HLDA Transforms, ML-HLDA-SAT)：

同樣將原本前端擷取之 60 維特徵向量，利用異質性線性鑑別分析

法降維成 46 維特徵向量，並利用條件式最大化相似度性線迴歸(Constrained Maximum Likelihood Linear Regression, CMLLR) [Gunawardana & Byrne 2001]來估測鑑別式線性轉換矩陣，之後利用最大化相似度(Maximum Likelihood)法訓練語者調適模型。

4. 語言模型訓練：

此系統利用各種來源的文字語料，並且利用改良式 Witten-Bell 語言模型[Witten *et al.* 1991]平滑化方法，以解決機率值為 0 的情況。

5. 解碼步驟：

系統辨識器核心為名 Byblos 的辨識器[Colthurst *et al.* 2000]，此辨識器須經過多階段(Multi-pass)的辨識過程，利用多階段不同種類的模型所產生的可能結果，以產生最大可能出現的語音訊號對應詞序列。解碼過程有：

A. 語者獨立(Speaker-Independent)解碼階段：

I. 快速對應(Fast Match)：利用向前(Forward)演算法對 STM 三連音素聲學模型、與詞二連語言模型做統計運算。

II. 利用向後(Backward)演算法對 SCTM 詞內(Within-word)五連音素(Quinphone)聲學模型、詞三連語言模型做統計運算，以找出 *N* 條最有可能的出現詞序列(*N*-best Hypotheses)。

III. 利用 Cross-word SCTM 聲學模型、詞四連(Fourgram)語言模型，對此 *N* 條可能出現的詞序列重算分數(Re-scoring)和重新排列(Re-rank)，最後出現機率最大的(Top 1)即為語者獨立階段的辨識結果。

B. 語者調適(Speaker-Adaptive)解碼階段：

利用語者獨立解碼階段最後產生的假設語句，訓練語者調適聲學模型，再重複相同解碼過程，找出新的假設語句。

C. 最後解碼階段利用大量的統計迴歸組合(Regression Classes) [Leggetter & Woodland 1995; Gales & Woodland 1996]使用對語者調適階段產生的假設語句再來調適聲學模型，產生最後辨識結果。

BBN 結合 LIMSI 的辨識模型，將不同的獨立、調適聲學模型與不同連結的 *N* 連語言模型做結合，最後利用 ROVER 方法[Fiscus 1997]，對每種模型產生最大假設語句中的部份小段(Segmentation)詞句，依據不同權重值做挑選(Voting)，產生組合成一條機率最大的詞序列。

6. 系統架構圖：

如圖 2-1 所示，2004 BBN/LIMSI 英文對話電話語料辨識系統混合比較 BBN、LIMSI 不同解碼階段(B1、B2、B3、B4、L1、L2、L3)產生的不同假設語句，利用 ROVER(圖 2-1 中的 R1 及 R2)挑選出機率最大的詞序列，其中各階段設定如下：

A. B1：使用感知線性預測特徵(PLP)擷取、MPE 的三連音素 STM 聲學模型(365K 個高斯混合分布)與代入詞二連語言模型。

B. B2：使用感知線性預測特徵擷取、MPE 的五連(Quinphone)音素 SCTM 聲學模型(843K 個高斯混合分布)與代入詞三連語言模型。

C. B3：使用感知線性預測特徵擷取、MPE 的詞間五連(Crossword Quinphone)音素 Crossword SCTM 聲學模型(855K 個高斯混合分布)與代入詞四連語言模型。

D. B4：使用梅爾倒頻譜係數特徵擷取、MPE 的詞間五連音素 Crossword SCTM 聲學模型(708K 個高斯混合分布)與代入詞四連語言模型。

E. L1：使用兩個性別相依(Gender-dependent)聲學模型(48 個單連音素、包含 30K 個狀態連結(Tied States)，其中每個狀態有 32 個高斯混合分布) ，並使用詞三連與詞四連插補過的語言模型(Interpolated Language Model)。

F. L2：將單連音素減少成為 38 個，28K 個內文音素共有 30K 個連結狀態 (Tied States)，並代入詞三連與詞四連插補過的語言模型。

G. L3：使用最大事後機率演算法調適性別相依聲學模型(43K 個內文相依(Context-dependent)音素，共有 31K 個連結狀態)。

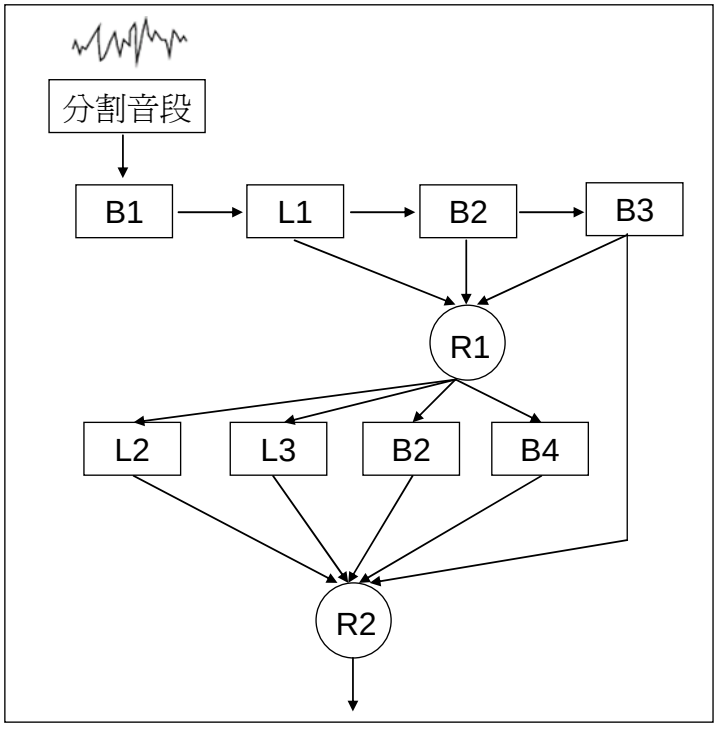


圖 2-1　2004 BBN/LIMSI 英文對話電話語料辨識系統架構圖

2.1.2 美國 IBM 華生研究中心

美國 IBM 華生研究中心曾於 2005 年發表的「IBM 2004 英文對話電話語料辨識系統」(IBM 2004 Conversational Telephony System for Rich Transcription)[Soltau *et al.* 2005]，此系統需 10 倍時間，評比語音語料為 RT04，詞錯誤率為 15.2%。

此系統特別之處是結合特徵最小音素錯誤 (Feature Minimum Phone Error, fMPE)[Povey *et al.* 2005]特徵擷取法，與最小音素錯誤(MPE)聲學模型訓練方法。實驗結果如表 2-2 所示，由表中數據發現，結合 fMPE 與 MPE 法，比單用 ML、fMPE 或 MPE 方法可得較低的詞錯誤率。

表 2-2　RT03 評比語料的詞錯誤率

	ML	MPE	fMPE	fMPE+MPE
詞錯誤率	22.1	20.6	20.2	19.2

1. 語料來源：

   聲學模型的語音訓練語料為共 2,100 小時的 Fisher parts1-7、Switchboard-1、BBN/CTRAN Switchboard-2、Switchboard Cellular、Callhome English 語料。語言模型的文字訓練語料來源有很多，有 SWB (LDC transcripts of Switchboard-1, Switchboard Cellular and Callhome English) 、 BBN (BBN/CTRAN transcripts of Switchboard-2)、BN (廣播新聞語料)、FSH(Fisher Collection Parts1-7)、UW191(由華盛頓大學蒐集的 191M “Switchboard-like”網路語料)、UW175(由華盛頓大學蒐集舊版的 175M “Fisher-like”語料)、UW525(由華盛頓大學蒐集新版的 525M “Fisher-like”語料)。

2. 前端語音特徵擷取與聲學模型訓練：

此系統的聲學模型分三類如下：

A. 語者獨立-對角化共變異矩陣-感知線性預測技術(Speaker Independent- Diagonal Covarience- PLP, SI-DC-PLP)：

語音訊號特徵擷取利用感知線性預測(PLP)技術、線性鑑別分析(Linear Discriminant Analysis, LDA)加上最大相似度線性轉換(Maximum Likelihood Linear Transformation, MLLT)[Gopinath 1998][Saon *et al.* 2000]與倒頻譜正規化法(Cepstral Mean and Variance Normalization,CMVN) [Viikki *et al.*1998]做正規化。語者獨立模型共有 150K 個 40 維的對角化共變異高斯混合分布與利用 MPE 法訓練的 8K 個英文五連音素狀態。

B. 語者調適-共變異矩陣-fMPE (Speaker Adaptation- Full Covarience-fMPE, SA-FC-fMPE)：

語音訊號特徵利用 VTLN 及感知線性預測技術、fMPE 與 LDA 配合 MLLT 及倒頻譜正規化法做擷取。語者調適模型共有 143K 個 39 維高斯混合分布與利用最大交互資訊法(Maximum Mixture Information, MMI)和語音特徵最大化相似度線性迴歸(fMLLR)法訓練的 7.5K 個五連音素狀態。

C. 語者調適-對角化共變異矩陣-fMPE+MPE (Speaker Adaptation-Diagonal Covarience- fMPE+MPE, SA-DC-fMPE+MPE)：

語音訊號特徵利用 VTLN 及感知線性預測技術、fMPE 與 LDA 配合 MLLT 及倒頻譜正規化法做擷取。語者調適模型共有 849K 個 39 維對角化高斯混合分布，與利用 fMPE 和 MPE 及 fMLLR 法和訓練的 22K 個七連(Septaphone)音素狀態。

3. 語言模型訓練：

此系統利用各種來源的語言資料，在實作時會依據與語音相關的程度不同權重(Weight)值做詞四連語言模型插補調適，並且利用改良式 Kneser-Ney 語言模型平滑化方法[Ney *et al*. 1994]，以解決語言模型機率值為 0 的情況。

4. 解碼步驟：

A. 對語音訊號做前端切刻(Segmentation)，區分出語音段(Speech)與非語音段(Non-speech) 。

B. 將語音訊號經由第一階段 SI.DC.PLP 所使用的語者獨立聲學模型做解碼。

C. 利用 VTLN 技術對與音訊號做語者正規化，將辨識結果再次經由 SA.FC.fMPE 調適模型做解碼。

D. 將辨識結果經由 SA.DC.fMPE+MPE 辨識模型，產生最後詞圖(Lattice) 。

E. 再經由語言模型分數做重新評分(Rescoring)，產生模糊網路(Confusion Network) [Mangu *et al*. 2000]，產生事後機率最大詞序列。

5. 系統架構圖：

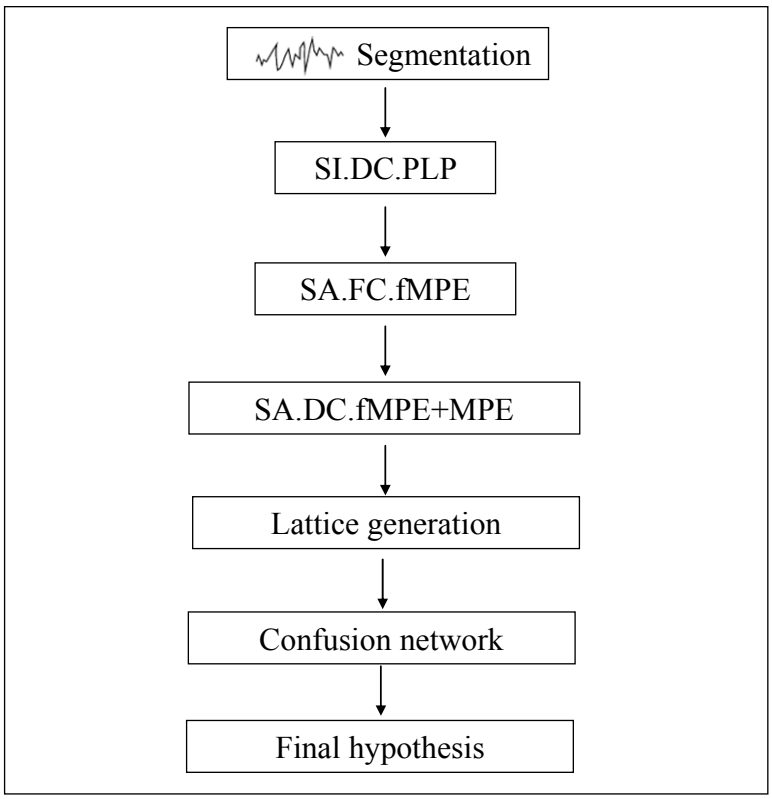


圖 2-2　IBM 2004 英文對話電話語料辨識系統架構圖

2.1.3　英國劍橋大學

英國劍橋大學曾於 2003 年發表的「2003 CU-HTK 英文對話電話語料辨識系統」(2003 CU-HTK Conversational Telephone Speech Transcription System)[Evermann *et al*. 2003]，此系統需 190 倍時間，評比語料為 RT03，詞錯誤率為 20.7%。而在 2004 年加入上達數千小時的訓練語料，發表新版的「2004 CU-HTK 英文對話電話語料辨識系統」(2004 CU-HTK Conversational Telephone Speech Transcription System)[Evermann *et al*. 2004]，系統時間為 10 倍時間，評比語料為 RT03，詞錯誤率為 17%。

1.　語料來源：

2003 CU-HTK 所用聲學模型語音訓練語料有 363 個小時的 Switchboard I、

Call Home English、Switchboard Cellular 語料，而語言模型文字訓練語料為 427M 詞彙量的廣播新聞文字語料、對話文字語料與從網路蒐集的 62M 詞彙量文字語料。

2004 CU-HTK 用到更多訓練語料，聲學模型部份為 2,180 小時、語言模型是 1,044M 詞彙量的 Fisher data。

2. 前端語音特徵擷取：

前端語音訊號利用 HLDA 做降維，並加入 VTLN 做語者正規化及倒頻譜正規化(CMVN)。

3. 聲學模型訓練：

利用最大相似度與 MPE 三連音素聲學模型，其中有 6K 個連結狀態，每個狀態中有 28 個高斯混合分布。

4. 語言模型訓練：

此系統利用各種來源的語言資料，在實作時會依據與語音相關的程度不同權重(Weight)值做詞四連語言模型插補調適，並且利用 Kneser-Ney 與 Good-Turing 語言模型平滑化方法，來找出沒有出現詞的語言模型機率。

5. 解碼步驟：

此系統需三個主要階段的解碼過程。

A. 第一階段(Pass 1,P1)使用 MPE 三連音素聲學模型、前端語音訊號利用 HLDA 做降維、及使用詞四連語言模型分數產生一條詞序列。

B. 將此詞序列利用 VTLN 找出語者正規化與倒頻譜正規化找出語音特徵向量。

C. 第二階段(Pass 2,P2)使用同第一階段的 MPE 三連音素聲學模型、VTLN 與 HLDA 前端特徵向量與語言模型產生較小的詞圖(Lattices)。

D. 第三階段(Pass 3,P3)使用詞圖最大相似度線性迴歸法(Lattice MLLR)[Uebel *et al*.2001]做調適，並使用詞四連插補法語言調適模型產生新詞圖。

E. 將第三階段產生的詞圖於 P4.1~P4.n(圖 2-3)與 P5.1~P5.n(圖 2-3)各階段，使用三連音素、五連音素聲學模型與語言模型重新評分，尋找機率最大(Top 1)詞序列。

F. 最後利用模糊網路結合(Confusion Network Combination, CNC) [Mangu *et al*. 2000]技術找出最終機率最大的辨識詞序列。此多重解碼步驟讓辨識結果更準確。

6. 系統架構圖：

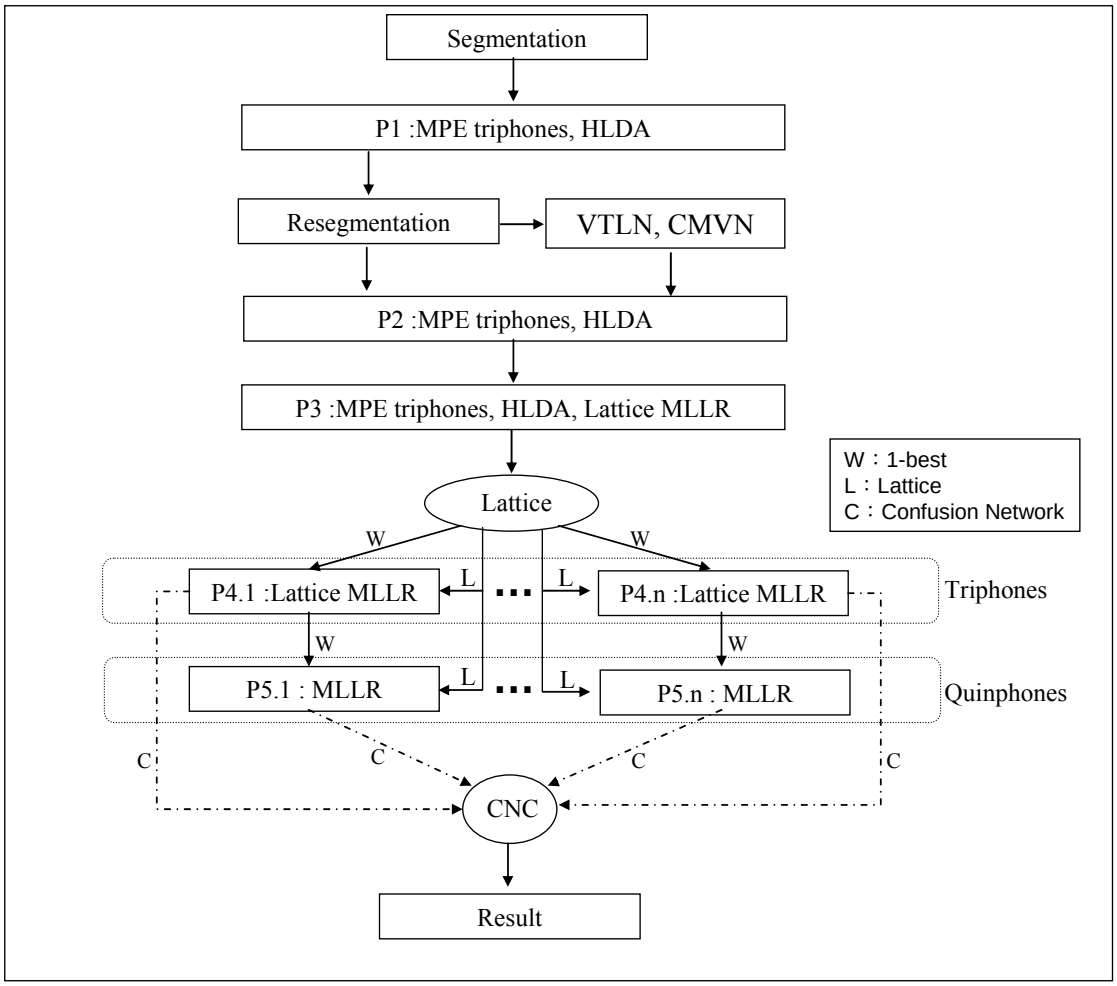


圖 2-3　2003 CU-HTK 英文對話電話語料辨識系統架構圖

2.1.4　綜合討論

此節綜合討論國外三家機構現階段大詞彙連續語音辨識器之內容特色，如下表 2-3 所示。

表 2-3　國外三家現階段大詞彙連續語音辨識器之內容特色

	BBN	IBM	CU
系統名稱	2004 BBN/LIMSI 英文對話電話語料辨識系統	IBM 2004 英文對話電話語料辨識系統	2004 CU-HTK 英文對話電話語料辨識系統
執行時間	20RT	10RT	10RT
評比語料	RT04	RT04	RT03
詞錯誤率	13.5%	15.2%	17%
聲學語料	2,300(時)	2,100(時)	2,180(時)
前端特徵	VTLN PLP + CMS HLDA+MLLT	VTLN PLP + CMVN +LDA fMPE + LDA+MLLT	VTLN HLDA+ CMVN
聲學模型	1. ML-SI (+HLDA) I. STM II. SCTM III. Cross-word SCTM 2. ML-HLDA-SAT (+MLLT)	1.SI.DC.PLP 2.SA.FC.fMPE 3.SA.DC.fMPE+MPE	MPE + Triphone Quinphone
語言模型	Witten-Bell+ Interpolated LM	Kneser-Ney + Interpolated LM	Kneser-Ney + Good-Turing+ Interpolated LM
解碼步驟	1. ML-SI : I.Triphone + Bigram II.Within-word Quinphone + Trigram III.Cross-word Quinphone+Fourgram 2. ML-HLDA-SAT 3. Regression Classes	1. SI.DC.PLP: Quinphone + Fourgram 2. SA.FC.fMPE: Quinphone + Fourgram 3. SA.DC.fMPE+MPE: Septaphone + Fourgram	1.Triphone+Fourgram 2.Quinphone+Fourgram 3.Lattice MLLR

觀察表 2-3 可得知，三家研究機構利用上千小時之聲學模型訓練語料，與豐富的大量語言模型訓練語料訓練模型，及前端語音特徵擷取技術、模型調適技術，增加語音辨識率。在辨識過程中為多階段(Multi-pass)之辨識，並分別訓練語者獨立

與語者調適之模型。三家機構之辨識器之詞錯誤率介於 13.5%至 17%之間。

2.2 聲學模型音素單位相似度測量

語言是亙古以來人類仰賴彼此溝通、了解最自然快速的重要工具，目前世界上有多達數千種的不同語言。本國人(Native)在學習非本國人(Non-Native)之語文時，可能因擁有本國人或第一語言之發音特性或習慣，故在學習非本國語言時會產生不同的發音腔調或變異(Variability)。如本國人說非本國語言，容易將某二個音混淆，則在本國人所說非本國語之聲學模型上，測量兩個音素於聲學模型中的音素相似度(Phoneme Similarity)，如果距離愈近代表相似程度愈高，可以合併來思考。

本節介紹在聲學模型尋找發音變異特性。主要有兩種方向，分別為資料導向方法(Data Driven Methods)與以知識為基準之方法(Knowledge Based Methods)。

2.2.1 資料導向方法

此方法為由上而下(Top-down)運算每個音素模型間的距離(Distance)，利用距離值當作相似度的比較，如果距離愈近代表相似程度愈高，可以合併來思考，以彌補語音訓練資料量的不足，運算此距離的運算方式有如下幾種[Le *et al.* 2006]：

1. HMM 距離 [Kohler *et al.* 1996]：

   如假設兩個模型 w_i 與 w_j ，與分別對應的語音特徵向量序列為 O_i 與 O_j ，則 HMM 距離算法為式(2-1)所示，即分別算出觀察值在模型中的 log 機率值(Log Likelihood)，再依對稱(Symmetric)取平均的方法運算其距離。

$$D(w_i, w_j) = \log p(O_i \mid w_i) - \log p(O_i \mid w_j)$$
$$D(w_j, w_i) = \log p(O_j \mid w_j) - \log p(O_j \mid w_i) \quad (2\text{-}1)$$
$$D(w_i, w_j) = \frac{1}{2}\left(D(w_i, w_j) + D(w_j, w_i)\right)$$

2. Kullback-Leibler 距離：

兩個機率密度函數 $p(o)$ 和 $q(o)$ 的相關熵值(Relative Entropy)可以式(2-2)所表示，用來表示兩個機率分布的差異程度，其中 o 代表語音特徵向量：

$$D(p \| q) = \int p(o) \log \frac{p(o)}{q(o)} do \quad (2\text{-}2)$$

3. Bhattacharyya 距離[Brian Mak *et al.*1996]：

如式(2-3)來測量兩個機率密度函數 $p(o)$ 和 $q(o)$ 的距離，其中 o 代表語音特徵向量：

$$D(p, q) = \int \sqrt{p(o)q(o)} do \quad (2\text{-}3)$$

4. Euclidean 距離[Sooful *et al.* 2001]：

如式(2-4)來測量兩個機率密度函數分布 i、j 的距離，其中 μ_i、μ_j 與 σ_i、σ_j 分別代表平均值與標準差，V 表向量串列的維度(Dimension) [Young *et al.* 2006]：

$$D(i, j) = \frac{1}{V} \sum_{k=1}^{V} \left[\frac{(\mu_{ik} - \mu_{jk})^2}{\sigma_{ik}\sigma_{jk}} \right]^{1/2} \quad (2\text{-}4)$$

5. 模糊矩陣(Confusion Matrix) [Beyerlein *et al.* 1999] [Bayeh *et al.* 2004]：

建立模糊矩陣來表示兩個音素之間的模糊機率(Likelihood)。計算語料之正確文字標記與辨識結果之最小編輯距離(Levenshtein Distance) ，找出每個音素 M「取代」(Substitution)成 $N_1 ... N_k$ 的次數正規化值，以 A_{MN_i} 表示，其中 $i = 1...k$

且$\sum_{i=1}^{k} A_{MN_i} = 1$。本論文欲利用此法，尋找語料辨識結果與正確解答之間的音素差異性，修改訓練聲學模型前的音素狀態連結規則，與將變異性加入於語音辨識階段，觀測語音辨識率。以圖 2-4 為例，代表統計辨識結果中，正確標記之單連音素「eh」易被取代為「ae」與「aa」的次數經正規化後分別為 0.3 與 0.7。

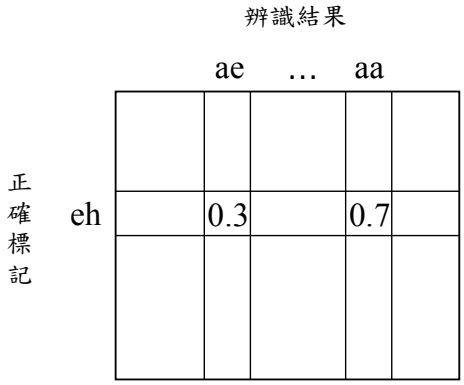


圖 2-4　模糊矩陣示意圖

2.2.2　以知識為基準之方法

傳統上以知識為基準的方法用來尋找原始語言(Source Language)對應目標語言(Target Language)的最佳音素[Beyerlein *et al*. 1999；Schultz *et al.* 2001]，但是並無比較兩個音素之間的相似度，在[Le *et al.* 2006]論文中提出一個新的以知識為基準方法來計算兩個音素之間的相似度，此法為一由下而上(Bottom-up)演算法，由兩個步驟組成：

1. 使用階層圖(Hierarchical Graph)做由上而下的分類：

   利用國際音素標準(International Phonetic Alphabet, IPA)所訂定的音素規則，如子音(Consonant)、母音(Vowel)分類，子音分類中又可分為破裂音

(Plosive)、雙唇音(Bilabial)，使用者自訂不同規則所占訓練資料量應有的比例，產生 k 層(Layer)分配，將每層訂定比例值，以 G_i 表示，代表使用者自訂第 i 層(i=0…k-1)的比例值，愈下層代表分配比例愈細緻，但內容所佔比重愈小，故 G 值隨之遞減。

2. 由下而上音素距離估測：

使用步驟 1 的定義，建立好規則階層圖型後，如果想找音素 s 與 t 之音素距離，首先觀察音素 s 與 t 是否有出現在階層的葉節點(Leaf Node)，如果有則從階層圖回朔、由下往上尋找兩音素最近的相交母節點(Parent Node)，觀察母節點所在的階層 G 值，當作兩個音素的距離，如式(2-6)：

$$d(s,t) = G_i \tag{2-6}$$

如圖 2-5 為例，某使用者自訂分層規則，將階層圖分五層，分別為層 0 至層 4，G_0 至 G_4 值分別為 0.9、0.45、0.25、0.1、0.0。此時欲找葉節點音素 p 與 m 之相似度，由下往上回朔觀察兩音素之相交母節點於層 2 之「Bilabial」(爆破音)，故兩音素之相似度設定為 0.25。

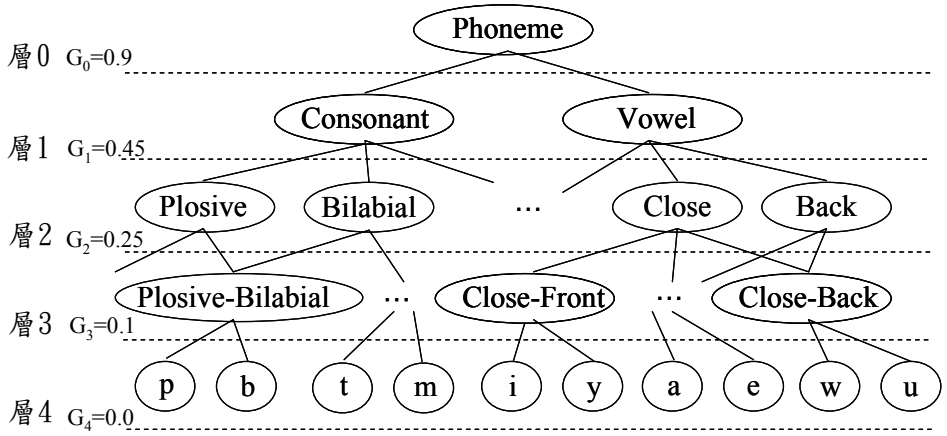


圖 2-5　音素相似度階層圖範例

第3章　實驗語料與設定說明

3.1　實驗詞典與英文音素定義

本實驗所用之英文詞典(Lexicon)為選自美國發音(American English)之 Festlex CMU 詞典(Lexicon for Festival Speech System Festlex CMU)[Festlex CMU]，此詞典有 105,626 個英文詞彙，如圖 3-1 所示：

```
…
("begin" nil (((b ih g) 0) ((ih n) 1)))
…
("coffee" nil (((k aa f) 1) ((iy) 0)))
…
("hello" nil (((hh ax l) 0) ((ow) 1)))
…
("yes" nil (((y eh s) 1)))
…
```

圖 3-1　原 Festlex CMU 詞典

```
…
begin b ih g ih n
…
coffee k aa f iy
…
hello hh ax l ow
…
yes y eh s
…
```

圖 3-2　處理後之 Festlex CMU 詞典

圖 3-1 Festlex CMU 詞典代表每個英文詞的音素(Phones)與重音音節(Syllables)之發音狀態，以正規表示式(Regular Expression)表示，在實驗前先對詞典做前處理動作，例如拿掉「(」、「)」、「”」、與代表發音重音音節的「0」、「1」和代表詞結尾「nil」等符號，留下英文詞與對應的單連音素，如圖 3-2 所示。統計所有相異的音素個數，總共有 40 個相異英文音素，並加入代表靜音(Silence)的「sil」與代表英文字與字之間暫停(Pause)的「sp」到音素列表中，故列表中共有 42 個相異單連音素，我們稱此 42 個不同的音素為單連音素(Monophone)。

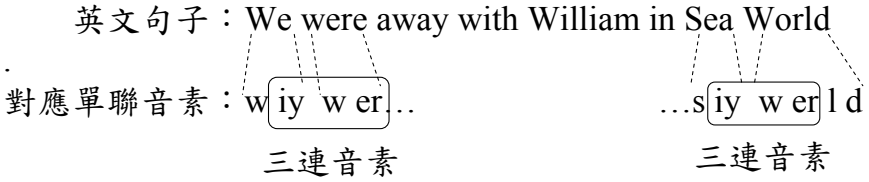
表 3-1-1 代表 Festlex CMU 詞典用到的音素列表與對應的國際音素標準(International Phonetic Alphabet, IPA)。

表 3-1-1　Festlex CMU 詞典所用之英文音素列表

Phone	Example	IPA	Phone	Example	IPA
母音(Vowels)			破裂音(Plosives)		
aa	bott	/ɑ/	**b**	bet	/b/
ae	bat	/æ/	**d**	debt	/d/
ah	but	/ʌ/	**g**	get	/g/
ao	bought	/ɔ/	**k**	cat	/k/
aw	bout	/ɑʊ/	**p**	pet	/p/
ax	about	/ə/	**t**	tat	/t/
ay	bite	/aɪ/	摩擦音(Fricatives)		
eh	bet	/ɛ/	**dh**	that	/ð/
er	bird	/ɝ/	**th**	thin	/θ/
ey	bait	/e/	**f**	fan	/f/
ih	bit	/ɪ/	**v**	van	/v/
iy	beet	/i/	**s**	sue	/s/
ow	boat	/o/	**sh**	shoe	/ʃ/
oy	boy	/ɔɪ/	**z**	zoo	/z/
uh	book	/ʊ/	**zh**	measure	/ʒ/
uw	boot	/u/	破擦音(Affricates)		
滑音(Glides)			**ch**	cheap	/tʃ/
l	led	/l/	**jh**	Jeep	/dʒ/
r	red	/r/	鼻音(Nasals)		
w	wed	/w/	**m**	met	/m/
y	yet	/j/	**n**	net	/n/
hh	hat	/h/	**ng**	thing	/ŋ/

人的說話聲中，有內文相依(Context Dependency)[Odell, 1995]的變異存在，因而影響發音的差異，主要分為外在的影響(Sessional Effects)與內在的影響(Local Effects)。外在的影響又分為語者影響(Speaker Effects)與環境影響(Environmental Effects)，語者影響為不同語者所發出的不同語言與本身的性別、

年齡、說話方式等差異，此為最主要影響發音變異差異的來源；環境影響為當語者發音時所受背景環境噪音(Background Noises)與通道效應(Channel Effects)的影響。內在的影響針對單一語者在發音時連音(Co-articulation)、強調(Stress)、與重音(Emphasis)的差異。如果能掌握以上這些發音變異重點，將可找出語者變異所在，並能試著調整其聲學模型中的參數或其他方法，讓模型辨識率提高。其中內在影響中的連音現象，為欲觀察發音文句的差異時，如果從詞(Word)的變化看至音素的變化，將可得更精確的觀察效果，如下列文句「We were away with William in Sea World」所示：



可發現文句中的「We were」與「Sea World」是兩個不同的英文字相連結，但是如果看到英文字所對應音素，是相同的「iy-w-er」相連結的三連音素，如果觀察其音素 w 的聲譜圖(Spectrogram)，可發現具有相似的曲線變化。又此跨出二詞(We, were)與(Sea、World)之間界線(Word Boundary)的音素相連結看法，為詞間(Inter-word)內文相依(Context Dependency)。倘若只看單一詞之間音素相連結的作法，則稱為詞內(Intra-word)內文相依，文獻[Odell 1995]實驗說明若能適當處理，詞間內文相依法可增加語音辨識的正確率。

圖 3-3 表示二連與三連音素詞內與詞間內文相依情況，在二連音素中，「_」代表音素兩兩連結狀況；在三連音素中，「-」代表左方連接(Left context)的音素、「+」代表右方(Right context)連接的音素。本論文實驗作法為主要為三連音素詞內內文相依。表 3-1-2 代表統計本論文所用 Festlex CMU 詞典之單連音素、詞內內文相依之二連音素、三連音素統計個數。

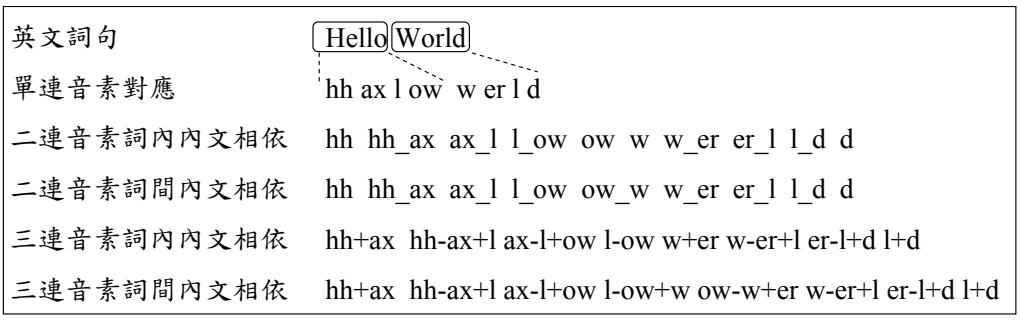


圖 3-3　二連、三連音素詞內與詞間內文相依示意圖

表 3-1-2　Festlex CMU 詞典的不同音素個數

個數	單連音素	二連音素	三連音素
Festlex CMU	42	1,364	19,339

3.2　實驗語料

3.2.1　台灣腔英語(English Across Taiwan, EAT)

本論文所用的台灣腔英語語音語料(English Across Taiwan, EAT)[EAT corpus]是從 2004 年 5 月開始收集，至 2005 年 1 月初步完成收集，經由台大、清大、交大、成大與師大等學校共同分配錄音，回收之語料經由工研院電通所彙整，並請專人整理語料庫，依據語料音檔之品質及發音內容之正確性分為可用(Usable)及不可用(Unusable)兩大類，可用語料再依英語系、非英語系與男、女性別做分類。這些語料又可分為麥克風語料(MIC)、電話語料(PSTN)、及手機語料(GSM)分類。本論文採用麥克風可用語料，其統計資訊如下表 3-2 所示。

表 3-2　EAT 麥克風語料音檔資料統計

	英語系		非英語系	
	男性	女性	男性	女性
原有句數	11,977	30,094	25,432	15,540
非含 OOV 句數	9,230	23,075	19,580	12,032
非含 OOV 時間(hr)	7.5	14.0	12.4	8.0
人數	166	406	368	224

麥克風語料錄製取樣頻率為 16 KHz(取樣點大小為 16 Bits)，是由個人電腦及麥克風經個人電腦之音效卡錄製的聲音訊號，最後將所有取樣點以 wav 格式音檔儲存。

EAT 語料內容中，有英文單字、片語、縮寫、數字與單字連續語音和中英文混合句，如表 3-3 所示：

表 3-3　EAT 語料中不同句型範例

	句型種類	標記檔範例
1	英文單字	grandpa
2	英文片語	for instance
3	英文縮寫	R. S. R. T. E. K.
4	英文數字連續語音	six five seven seven four five seven
5	英文連續語音	Green Mountain Energy
6	中英文混合句	我要查 Fuji
7	中英文混合句加縮寫	這家的刀削麵吃起來很 Q.

本論文主要探討英文單字、片語與連續語音的辨識(如表 3-3 之 1、2、4 和 5)，並將語料人工轉寫 Transcription)，含有沒出現在 Festlex CMU 詞典中的詞以 OOV(Out of Vocabulary Words, OOV)刪除之。

3.2.2 美國之音(The Voice of America, VOA)

本論文利用美國之音(The Voice of America, VOA)語音語料[VOA corpus]，美國之音成立於 1942 年，屬於美國政府的對外廣播電臺與電視台，每天以五十三種語言、每星期超過一千三百小時向世界各地廣播。內容包括新聞時事、專題節目、英語教學節目、美國流行音樂，與反映美國政府立場的社論，是最為大眾所熟知的國際廣播。

本論文收集美國之音語料，為 2004 至 2006 年 VOA 語料，取樣頻率為 16 KHz(取樣點音檔為 16 Bits)，最後將所有取樣點以 pcm 格式音檔儲存，包含男、女聲混合，經由事前處理對 3 分鐘以上的語音檔案與文字對應檔做切割(Segmentation)，利用 Transcriber 軟體將 5 至 10 字切成一句，去除雜音與音樂音段，並將文字檔中的阿拉伯數字轉換為詞典中之英文詞彙，且標點符號暫不考慮，剩下人聲的部份，統計共有 4.5 小時的語料。表 3-4 為 VOA 部份語句：

表 3-4　VOA 實驗語句

	句型範例
1	their workshops were long ago damaged
2	an internet message taking responsibility for their deaths
3	it is one of those things that i dreaded the entire time

3.2.3 英國國家文字語料庫(British National Corpus, BNC)

英國國家文字語料庫(British National Corpus, BNC)[BNC Corpus]，為一蒐集上達約一億個詞(102M)有關說、寫的文字語料庫，資料來源為近百年來各種領域的說、寫資料。寫的資料占 90%，資料來源包含各領域新聞、期刊、學術書籍、小

說、出版品、信件、章程、備忘錄等文字資料；而說的資料占 10%，資料來源為經由不同種族、年齡、職業及商業或政府開會或是廣播新聞等對話資料。本論文用此BNC語料庫,訓練語言模型,當成背景語言模型(Background Language Model)之訓練語料。

3.3 台師大大詞彙連續語音辨識系統

本論文所用之辨識器為台師大大詞彙連續語音辨識系統[Chen *et al.* 2004, 2005]，以下分別介紹系統實驗用到的語音特徵擷取、聲學模型建立、語言模型建立、詞典建立與語言解碼等部分。

3.3.1 語音特徵擷取

語音辨識系統可以看成一種圖樣辨識(Pattern Recognition)系統，此概念屬於分類(Classification)的問題，如果擷取出的特徵向量可以有較高的鑑別力(Discrimination)，分類的結果也會較準確。而特徵的維度如果增加，代表後端分類器的參數或複雜度也會隨之增加，但是訓練資料有限，如此可能造成分類器參數估測不準確，進而影響辨識率，故如果能夠降低語音特徵維度，並擷取出具有鑑別力的語音資訊，勢必可以增加語音辨識效能。

語音特徵擷取技術經由模擬人耳聽覺感知特性，利用 1.2.2 節敘述之梅爾倒頻譜係數(MFCC)、線性預測係數(LPC)與感知線性預測係數(PLPC)等基礎方法做降維、並增強語音訊號和壓抑非語音訊號，但不保證對語音辨識或分類有較高的鑑別力。故再利用資料相關線性特徵轉換(Data-Driven Linear Feature Transform)，進一步降低維度並找出較具有代表性與鑑別力的特徵。例如線性鑑

別分析(Linear Discriminant Analysis, LDA)[Campbell 1984]、異質性線性鑑別分析(Heteroscedastic Linear Discriminant Analysis, HLDA)[Kumar 1997] [Kumar and Andreou 1998]與最大相似度線性轉換(Maximum Likelihood Linear Transformation, MLLT)[Gopinath 1998][Saon *et al.* 2000]等方法。最後可再經由強健性方法，增強語音特性並減少噪音干擾。例如倒頻譜平均消去法(Cepstral Mean Subtraction, CMS) [Furui 1981]與倒頻譜正規化法(Cepstral Mean and Variance Normalization, CMVN)[Viikki *et al.*1998]。

本論文結合不同種類的語音特徵擷取方式，訓練不同的聲學模型，觀測語音辨識率的變化，並使用較好的特徵值與聲學模型做相關實驗，運用到的不同特徵擷取法有：

1. 梅爾倒頻譜係數(MFCC)。
2. 梅爾倒頻譜係數配合倒頻譜平均消去法(MFCC+CMS)。
3. 梅爾倒頻譜係數配合倒頻譜正規化法(MFCC+CMVN)。
4. 線性鑑別分析配合最大相似度線性轉換，加上倒頻譜正規化法(LDA+MLLT+CMVN)。
5. 異質性線性鑑別分析配合最大相似度線性轉換，加上倒頻譜正規化法(HLDA+MLLT+CMVN)。

3.3.2 聲學模型建立

本實驗利用微軟公司(Microsoft Corporation)與英國劍橋大學電機研究所(Cambridge University Engineering Department)研發之隱藏式馬可夫模型訓練工

具 3.4 版(Hidden Markov Model Toolkit, HTK Version 3.4)，分別訓練 EAT 與 VOA 兩套語料之單連音素(Monophone)、二連音素(Biphone)與三連音素(Triphone)之聲學模型。在語音辨識的評估方面，首先對測試語料(Testing Set)進行語者辨識，之後將辨識結果與正確標記問文字進行編輯距離(Levenshtein Distance)比對，找出詞的取代(Substitution)錯誤、插入(Insertion)錯誤、與刪除(Deletion)錯誤與相符(Match)的個數，詞錯誤率(Word Error Rate, WER)運算如式(3-1)所示，為 1 減去詞正確率(Accuracy)：

$$\text{詞錯誤率} = 1 - \text{詞正確率} \quad = 1 - \left(\frac{\text{相符詞數} - \text{插入詞數}}{\text{正確句子中的所有詞數}} * 100\% \right) \qquad (3\text{-}1)$$

而在 HTK 中設定詞的取代錯誤、刪除錯誤、與插入錯誤的懲罰值(Penalty)，分別為 10、7 與 7。

HTK 執行流程主要分為四部份，事前資料整理(Data preparation)、訓練(Training)、測試(Testing)、與分析(Analysis)，如圖 3-4 所示[Young *et al.* 2006]。

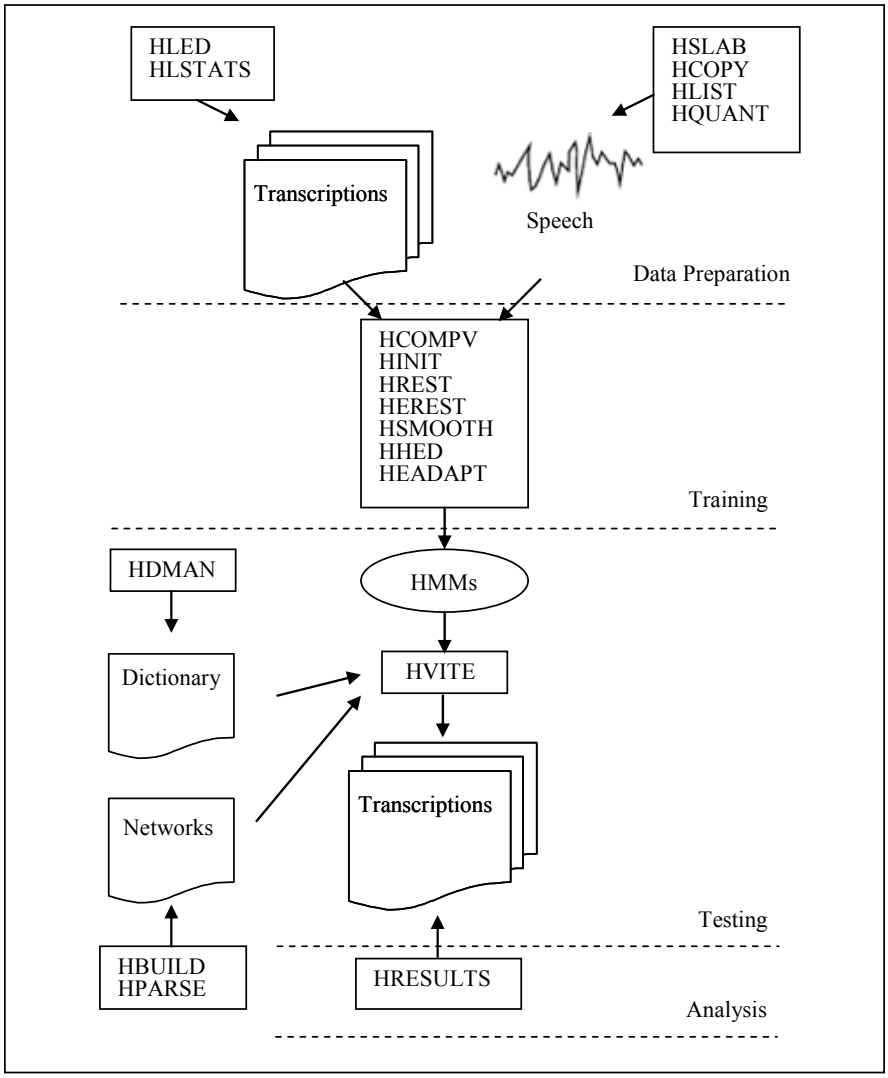


圖 3-4　HTK 處理流程

本論文利用 HTK 訓練實驗所需之英文單連音素、二連音素與三連音素狀態分享(Tied State Triphones)聲學模型。三連音素狀態分享之聲學模型建立過程主要有四個步驟[Young 1994]：

1. 建立單連音素聲學模型：

   首先將訓練語料做前端語音特徵擷取，並且備好語音語料對應的人工文字轉寫(Transcription)及語音檔內容，計算訓練語料中所有單連音素的高

斯分布，先以全域平均值(Global Mean)與共變異數(Covariance)來初始化基本單連音素聲學模型中的高斯分布，本論文設定每個聲學模型中有3至5個狀態，每個相異的初始單連音素都擁有相同的平均值、共變異數及轉換機率，此初始化訓練法稱為「單調開始」(Flat Start)。因為每個單連音素的參數都相同，故利用波式重估演算法，對訓練語料做參數估測更新(Re-estimate)。如圖 3-5 建立單連音素聲學模型所示：

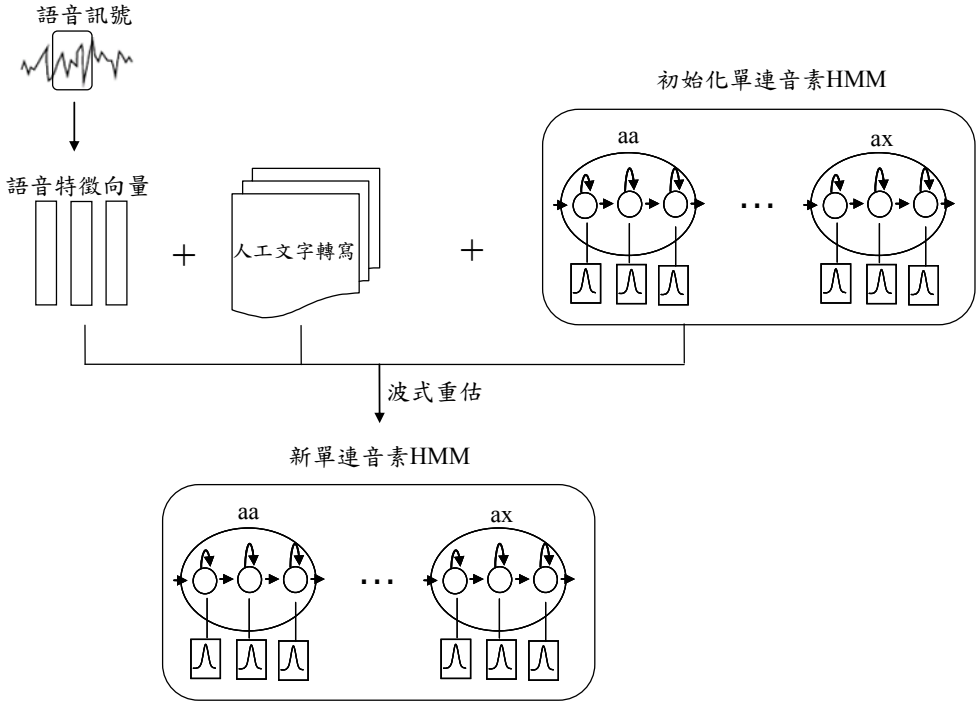


圖 3-5　建立單連音素聲學模型

2. 由單連音素模型建立三連音素模型：

利用第一階段產生的新單連音素模型，與統計訓練語料中三連音素列表，重新初始化三連音素聲學模型，此時三連音素模型的分布，是利用中央音素(Central Phone)的單連音素模型分布值，如圖 3-6 中初始化三連音素 HMM 中之「k-aa+b」、「k-aa+g」、「l-ax+m」為以中央音素為「aa」

與「ax」之單連音素分出來之初始化模型，同樣的，因為每個中央音素相同的三連音素之模型參數都相同，故利用波式重估演算法，對訓練語料做參數估測更新。圖 3-6 為由單連音素模型建立三連音素模型：

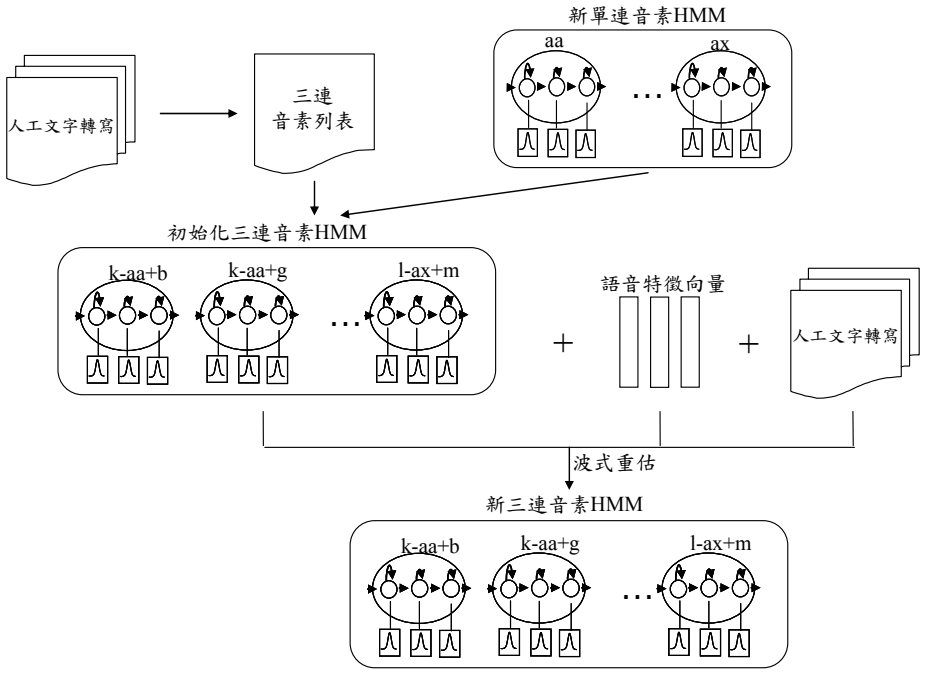


圖 3-6　由單連音素模型建立三連音素模型

3. 建立狀態分享之三連音素模型：

單連音素的模型有 41 種，而三連音素的模型至多有 41 的三次方種，然而除非訓練語料量夠大夠多，使得每種三連音素組合都有出現，不然建立內文相依(Conetxt Dependence)的聲學模型，經常有資料稀疏(Data Sparseness)的問題產生，即有些三連音素組合沒有足夠的訓練語料亦或是測試語料中的三連音素無法找到適當的三連音素模型對應，為了解決這個問題，過去學者[Lee 1989]嘗試以模型為基礎之分享法(Model-based Sharing)，將無法找到對應的三連音素模型機率，以相似模型的單連、

二連音素模型的機率分布做平滑化(Smoothing)，但是此法所得的三連音素機率值並不可靠，故考慮利用模型間的狀態(State)分布做連結(Tying)分享，因為看的更細，所得結果能更有鑑別性。建立狀態連結分享之三連音素模型方法有兩種，分別為資料導向之分群法(Data-driven Clustering)及以樹為基礎之分群法(Tree-based Clustering)。

資料導向之分群法為一種由下而上(Bottom-up)法，利用馬氏距離(Mahalanobis Distance)，如式(3-2)所示：

$$d(i,j)=\frac{1}{V}\sum_{k=1}^{V}\left[\frac{(\mu_{ik}-\mu_{jk})^2}{\sigma_{ik}\sigma_{jk}}\right]^{1/2} \tag{3-2}$$

計算三連音素模型中兩兩狀態之高斯分布群集 i 與 j 之 V 維向量平均值 μ 與標準差 σ 之間的距離，距離小代表分布較相近，故可合併成同一種群集，利用此法達到狀態分享。此法可解決模型訓練語料的不足，但是如果詞典中之三連音素沒出現在訓練語料中，則無法找到適當的三連音素模型做連結對應。

以樹為基礎之分群法為一種由上而下(Top-down)法，先將所有訓練語料的三連音素模型的每個狀態依據條件置於根(Root)群集中，如圖 3-7 第一列表示將訓練語料中三連音素的組合中，中央音素為「aa」音素之狀態 2(State 2)蒐集起來為一群集，「*」代表任意音素。

```
…
"aa_s2" {(aa, *-aa, *-aa+*, aa+*).state[2]}
"aa_s3" {(aa, *-aa, *-aa+*, aa+*).state[3]}
"aa_s4" {(aa, *-aa, *-aa+*, aa+*).state[4]}
"ax_s2" {(ax, *-ax, *-ax+*, ax+*).state[2]}
"ax_s3" {(ax, *-ax, *-ax+*, ax+*).state[3]}
"ax_s4" {(ax, *-ax, *-ax+*, ax+*).state[4]}
…
```

圖 3-7　狀態群集設定範例

```
...
QS 'L_ae' { "ae-*" }
QS 'R_ae' { "*+ae" }
QS 'L_Liguid' { “hh-*“,”l-*”,”r-*”,”w-*”,”y-*” }
QS 'R_Liguid' {“*+hh“,” *+l”,” *+r”,” *+w”,” *+y” }
...
QS 'L_Class-Stop' { "b-*","d-*","g-*","k-*","p-*","t-*" }
QS 'R_Class-Stop' { "*+b","*+d","*+g","*+k","*+p","*+t" }
QS 'L_Nasal' { "m-*","n-*","ng-*" }
QS 'R_Nasal' { "*+m","*+n","*+ng" }
...
```

圖 3-8　決策樹問題集範例

表 3-5　以樹為基礎之分群法之分類問題條件

音素本身	每個單連音素
停頓音(Stop)	b　d　g　k　p　t
鼻音(Nasal)	m　n　ng
摩擦音(Fricative)	ch　dh　f　jh　s　sh　th　v　z　zh
流音(Liquid)	hh　l　r　w　y
母音(Vowel)	aa　ae　ah　ao　aw　ax　ay　eh　er　ey　ih　iy　ow　oy　uh　uw
舌前音(Front)	ae　b　eh　f　ih　iy　m　p　v　w
舌中音(Central)	ah　ao　d　dh　er　l　n　r　s　t　th　z　zh
舌後音(Back)	aa　ax　ch　g　hh　jh　k　ng　ow　sh　uh　uw　y

接著開始建立決策樹(Decision Tree)，自定分裂決策樹之問題條件，如圖 3-8 代表問題集(Question Sets, QS)範例，表 3-5 代表本論文實驗中所用之分類問題條件。以圖 3-8 第一列為例，「*」代表目前觀察之三連音素狀態，「ae-*」代表目前觀察之三連音素狀態之左方是否為音素「ae」。用訂定的每個問題條件，運算對數機率值(Log Likelihood)，作為分裂群集的依據，如果每個問題條件之最大機率值中，有大於目前所在群集之機率值，則做分裂群集的動作，一直到最後問題條

件都使用完畢或到達門檻值(Threshold)則停止分群，最後落在同個葉節點(Leaf Node)的狀態可做狀態連結。因此詞典中每種三連音素組合都可以找到對應。如圖 3-9 代表以中央音素為「aa」之第三狀態決策樹，分群之問題條件為：中央音素左方是否為滑音(Liquid)?中央音素左、右方所接音素是否鼻音(Nasal)與停頓音(Stop)，最後三連音素「k-aa+b」與「k-aa+g」之第三個狀態(state[4])落於同個葉節點，則兩狀態可做連結。

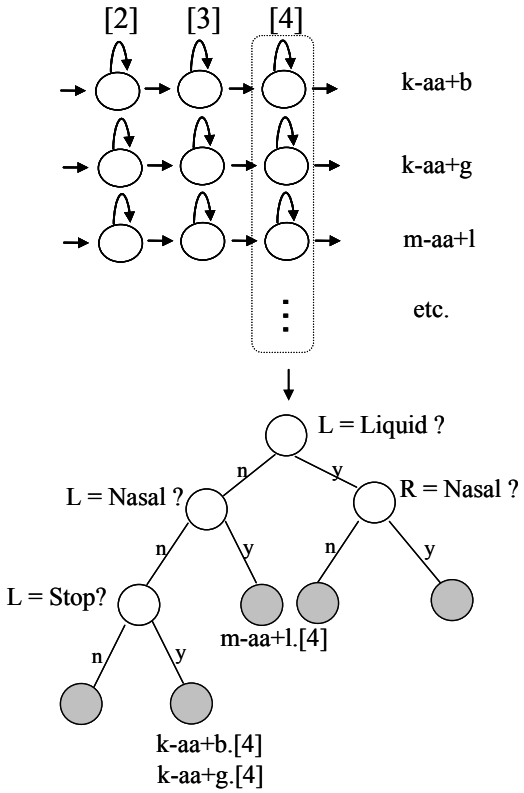


圖 3-9　以中央音素為/aa/之第三狀態決策樹

利用狀態連結法，產生狀態連結之三連音素聲學模型，如圖 3-10 建立狀態分享之三連音素模型所示：

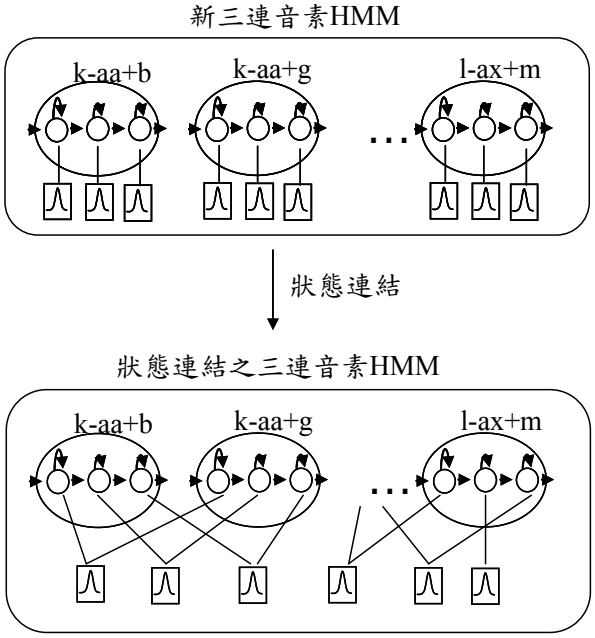


圖 3-10　建立狀態分享之三連音素模型

4. 增加三連音素模型之高斯混合數目：

最後一個步驟為增加高斯混合數目(Mixture Splitting)，以提升模型的效能，增加高斯混合數目的方法可以有很多，可以同時增加某個值，或是依據訓練語料的相異資料量，做不同比重的高斯混合數增加，如圖 3-11 增加三連音素模型之高斯混合數目所示，以產生初始化狀態連結之三連音素模型，最後再利用波式重估法，產生新狀態連結之三連音素模型，讓模型與訓練語料更為匹配。

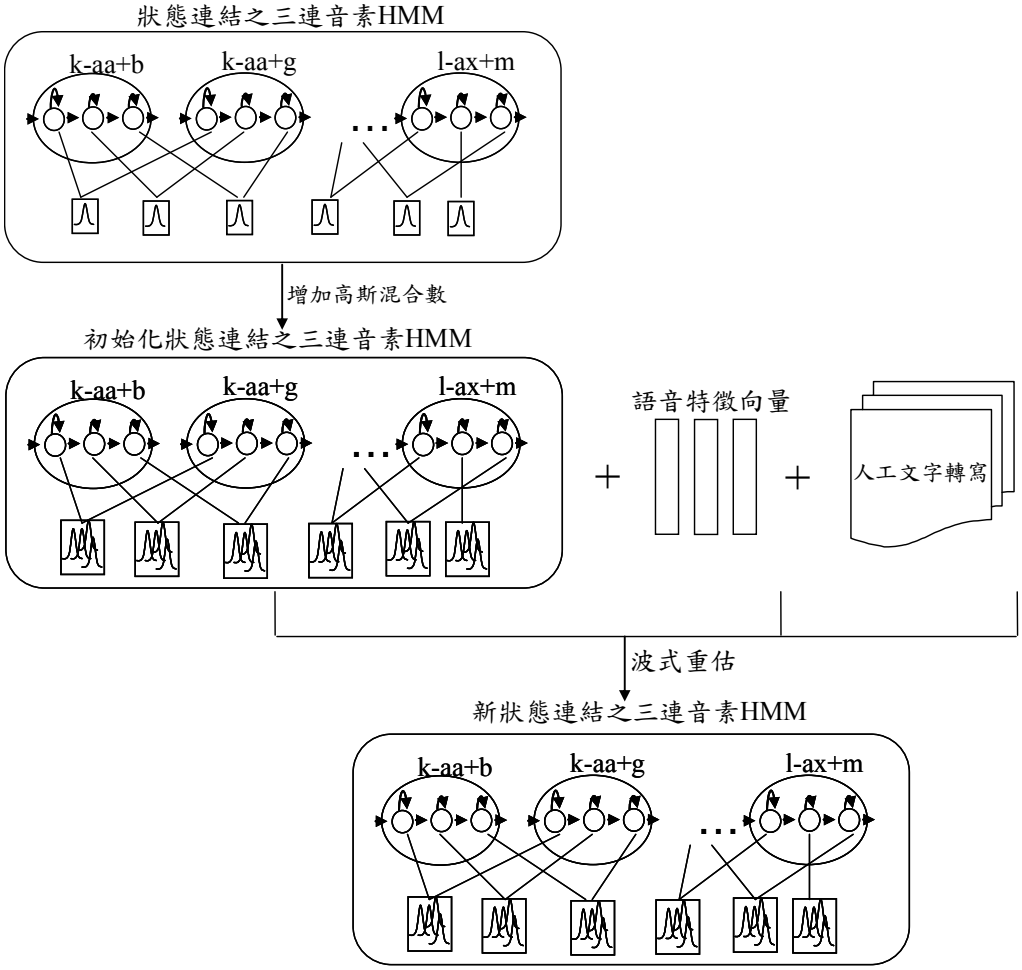


圖 3-11　增加三連音素模型之高斯混合數目

3.3.3　語言模型建立

本論文系統利用詞二連與詞三連之語言模型，語料來源利用同聲學模型訓練語料之美國之音(VOA)、台灣腔英語(EAT)當作相同領域(In-domain)之語言模型訓練語料，而英國國家語料庫(BNC)作為背景語言模型(Background Language Model)之訓練語料。本論文中的語言模型使用 Katz 語言模型平滑化技術，語言模型訓練工具為 SRI Language Modeling Toolkit(SRILM)[SRILM]。表 3-6 代表本論文第 4、5 章用到實驗語料之詞彙數統計，詳細說明於 4.1.1、4.2.1、5.4.2 節之實驗設定。

表 3-6　本論文所用語料之詞彙數統計

語料	語料時間(hr)	詞彙數(個)
VOA 訓練語句	3.33	30,637
VOA 測試語句	0.56	4,373
EAT 監督式訓練語句	7.02	53,922
EAT 非監督式訓練語句	33.4	108,323
EAT 測試語句	0.65	2,781
BNC	(純文字檔)	102,350,485

3.3.4　詞典建立

本論文所用之辯識器詞典，是將原本 105,626 個 Festlex CMU 做縮減，僅保留在實驗訓練語料與測試語料中有出現的詞，以此詞典來當作語言模型訓練詞典及語音辨識之詞典，所用詞典個數如表 3-7 所示：

表 3-7　本論文所用語料之詞典個數

語料	語料內容(hr)	詞典個數
VOA 訓練	3.89 (3.33 + 0.56)	5,178
EAT 監督式訓練	7.67 (7.02 + 0.65)	2,370
EAT 非監督式訓練	41.07 (7.02 + 33.4 +0.65)	4,229

3.3.5　語言解碼

本論文在語言解碼階段，利用二階段(Two-Pass)解碼過程，找出語音訊號對應的最佳詞序列，第一階段為詞彙樹複製搜尋(Tree-Copy Search)，第二階段為詞圖重新評分(Word Graph Rescoring)。

本論文之系統於第一階段採用由左至右(Left-to-right)且音框同步(Frame

Synchronous)[Aubert 2002]的詞彙樹複製搜尋方式，在詞彙樹樹中每一個分枝(Arc)代表一個三連音素隱藏式馬可夫聲學模型，如圖 3-12 詞彙樹範例所示，於樹中利用維特比動態規劃搜尋法，對三連音聲學模型之狀態做搜尋，並檢查是否走到樹的葉節點(Leaf Node，圖 3-12 的菱形實心點)，如果有，則代表一個完整詞或一些完整詞的產生，然而由於詞彙樹中存活的隱藏式馬可夫模型狀態節點可能會隨著音框數呈指數倍增加，因此必須利用光束剪裁(Beam Pruning)技術適當的剪裁掉分數較低的詞彙樹內部狀態節點(Internal Nodes)或不完全路徑(Partial Paths)。

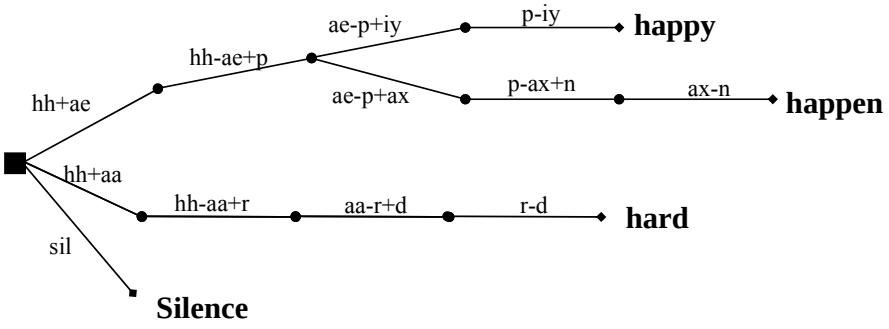


圖 3-12　詞彙樹範例

詞彙樹複製搜尋演算法於搜尋時，每個音框會同時存在數棵詞彙樹複製(Tree Copies)，每棵詞彙樹代表不同的語言模型歷史(Language Model History)，然而實際上，搜尋時產生的不完全路徑(Partial Paths)如果擁有相同的語言模型歷史，則會被歸類在同一棵詞彙樹裡，再進行隱藏式馬可夫模型狀態層次(State-level)維特比動態規劃搜尋。如果每個音框中，有不完全路徑已達葉節點，若具有相同的語言模型歷史，則會進行再結合(Recombination)，只保留其中分數最大者，並以它們的語言模型歷史為標註，產生新的一棵詞彙樹複製，或加入到一棵已存在且具有相同語言模型歷史的詞彙樹複製中。

為了降低搜尋空間，近年來學者提出語言模型前看(Language Model Look-ahead)[Ney *et al.* 1999]與聲學模型前看分數[Chen *et al.* 2004, 2005]等技術預先估計尚未搜尋到的語音段落的語言模型分數及聲學分數，當成剪裁比較的依據。本系統只採用單連詞語言模型前看(Word Unigram Language Model Look-ahead)技術，對每一個詞彙樹複製內部狀態節點，會以其所在分枝(或隱藏式馬可夫模型)之可能拜訪葉節點中具最大單連詞語言模型機率，做為該內部狀態節點的語言模型前看分數。

此外在每個音框會記錄存活的詞彙樹複製葉節點中分數較高者的相關資訊，例如語言模型歷史、對應候選詞開始與結束的音框以及搜尋時聲學解碼的分數，然後再依此資訊建立詞圖(Word Graph)，如圖 3-13 所示，並在詞圖上使用更高階的語言模型，如詞三連、詞四連之語言模型分數，重新進行一次詞圖動態搜尋重新評分(Word Graph Rescoring)，找出最佳的文句。本論文於第一階段詞彙樹複製搜尋使用詞二連(Bigram)語言模型分數，之後於第二階段詞圖動態搜尋重新評分使用詞三連(Trigram)語言模型分數。

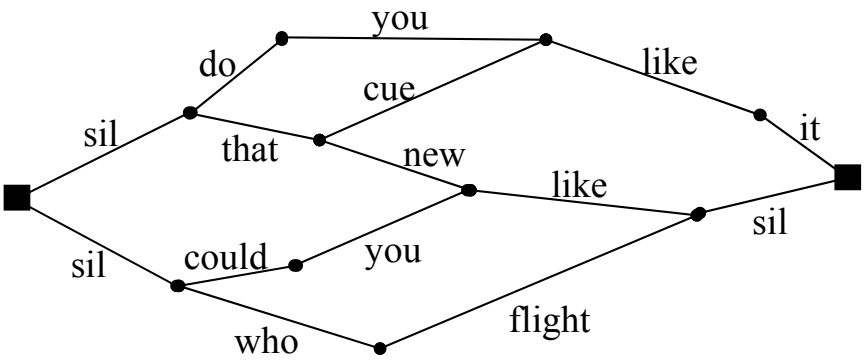


圖 3-13　詞圖示意圖

第4章　英文語音辨識之基礎實驗

4.1　VOA 語料之基礎實驗

4.1.1　實驗設定

本論文所用 VOA 實驗語料設定如表 4-1-1 所示。訓練語料為 3.33 小時廣播新聞語料，內容有男、女主播及男、女受訪者語料。測試語料為 0.56 小時。詞典個數為訓練語料與測試語料中相異之詞組成，共有 5,178 個詞。

表 4-1-1　VOA 實驗語料設定

語料分配	種類	句數	時間(hr)	詞彙數
	訓練語料	5,340	3.33	30,637
	測試語料	500	0.56	4,373
詞典個數	5,178 個			

4.1.2　基礎語音特徵擷取

影響語音辨識率重要因素之一為前端語音特徵擷取方式，本論文使用不同種類的語音特徵擷取法，分別為梅爾倒頻譜係數(MFCC)法、梅爾倒頻譜係數配合倒頻譜平均消去法(MFCC+CMS)與梅爾倒頻譜係數配合倒頻譜正規化法(MFCC+CMVN)，作為基礎語音特徵擷取法之實驗。

梅爾倒頻譜係數的擷取步驟如論文 1.2.1 圖 1-2 所示。倒頻譜平均消去法(CMS)與倒頻譜正規化法(CMVN)此兩種方法，原先是用來消除通道雜訊

(Channel Effects)，然而經過實驗證實，對於語音特徵強健性處理也同樣有幫助，且計算量很小，目前已廣泛應用於語音辨識上。

倒頻譜平均消去法是假設語者發音期間，通道效應對語音訊號的影響在倒頻譜上的表現為常數(Constant)，因此在倒頻譜上經過倒頻譜平均消去法處理後，可適度移除由不同的通道所造成的不匹配現象，假設一段語句經過特徵擷取後為一連串倒頻譜特徵向量 $C=\{C_1,C_2,...,C_t,C_T\},\ t=1,...,T$，其中 C_t 代表語句中第 t 個語音特徵向量，T 代表語句的總特徵向量個數。則倒頻譜平均消去法(CMS)所得新特徵向量為 $\widetilde{C}_t$，可表示成式(4-1)：

$$\widetilde{C}_t = C_t - \overline{C},\ t=1,...,T \tag{4-1}$$

其中 $\overline{C}=\frac{1}{N}\sum_{t=1}^{N}C_t$，而 $\widetilde{C}_t$ 為原本語句中第 t 個特徵向量 C_t 減去此語句的平均向量 $\overline{C}$，故新的倒頻譜特徵向量的平均值向量被正規化為零向量(Zero Vector)。

另一方面，經倒頻譜正規化法(CMVN)所得新特徵向量可表示成式(4-2)：

$$\widetilde{C}_t = \frac{C_t[n]-\overline{C}[n]}{S[n]}, t=1,...,T \tag{4-2}$$

其中 $S[n]=\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(C_t[n]-\overline{C}[n]\right)^2}$，代表語音特徵向量的標準變異向量。因此 CMVN 除了將 C_t 減去語句的平均向量，再對特徵向量的變異量做正規化。除了移除通道效應的影響，在對變異數進行正規化的過程中，降低不同維度間的語音特徵機率分布的差異程度，故相對於倒頻譜平均消去法，文獻證明此法能更進一步降低環境不匹配對特徵參數所造成的不良影響。

HMM 狀態中，依據每個 HMM 模型所分配到訓練語料段落數，分配 1 至 128 個不等的高斯混合數目，分配規則如表 4-1-2 所示。實驗嘗試將高斯混合數目依

據規則倍數做調整，例如表 4-1-2 第三列代表 HMM 分配到之語音段落數數目小於等於 10 之高斯混合數給定高斯混合數為 2，於實驗中嘗試將高斯混合數依倍數放大，觀察辨識率的改變。如前例依據 2 倍放大，混合數設定為 4；依據 3 倍放大，混合數設定為 6。

表 4-1-2　高斯混合數依音素出現比例之分配規則表

HMM 分配到之語音段落數	高斯混合數
≤ 5	1
≤ 10	2
≤ 100	4
≤ 500	8
≤ 2500	16
≤ 12500	32
≤ 62500	64
> 62500	128

表 4-1-3 代表三種不同基礎特徵擷取法於 VOA 語料的辨識結果，其中「TC」代表辨識器第一階段(One Pass)詞彙數複製搜尋(Tree Copy Search)後辨識之結果；「WG」代表第二階段詞圖重新評分(Word Graph Rescoring)後辨識結果。此實驗之語言模型為使用 VOA 訓練語料訓練而成。

表 4-1-3　VOA 不同基礎特徵擷取法之辨識結果

實驗	語音特徵	混合數	詞正確率(%)	
			TC	WG
1	MFCC	78,412	46.95	48.75
2	MFCC_CMS	76,073	47.95	49.92
3	MFCC_CMVN	73,083	47.12	49.10

本論文之 VOA 語料之三種基礎特徵擷取法(MFCC、MFCC 配合 CMS、MFCC 配合 CMVN)，觀察辨識結果，以 MFCC 配合 CMS 之前端特徵擷取法，可得較高之辨識率為 49.92%。表中之高斯混合數計算有些微差異，因在多次訓練單連音素模型階段中，為了讓訓練語料對模型間的效能更匹配，會對訓練做一次重新校準(Realign)，捨去辨識過程中小於門檻值的訓練語句，故導致最後 HMM 模型高斯混合數稍微減少。

4.1.3 基礎三連音素聲學模型

利用表 4-1-3 實驗 2 之 MFCC 配合 CMS 擷取出的語音特徵來訓練聲學模型，並增加高斯混合數，依表 4-1-2 之高斯混合數分配規則(表 4-1-4 之實驗 1)、規則之 2 倍(表 4-1-4 之實驗 2)、規則之 3 倍(表 4-1-4 之實驗 3)、規則之 4 倍(表 4-1-4 之實驗 4)訓練聲學模型，語言模型仍使用 VOA 訓練語料訓練而成，實驗結果如表 4-1-4 所示。

表 4-1-4　VOA 不同高斯混合數之辨識結果

實驗	語音特徵	混合數	詞正確率(%)	
			TC	WG
1	MFCC_CMS	76,073	47.95	49.92
2	MFCC_CMS	145,318	46.89	48.25
3	MFCC_CMS	217,744	45.23	47.01
4	MFCC_CMS	290,505	43.42	45.77

觀察發現，高斯混合數增加，對辨識率的改變並不大，可能原因有：高斯混合數的分配比例是依據訓練語料量增加，又因為訓練語料量不足，故模型中存在資料稀疏(Data Sparseness)問題，使辨識率下降，接著吾人將針對其他可能影響辨識率之因素做分析。

4.1.4 基礎語言模型

本小節觀察是否因為語言模型對辨識效果產生不同影響。首先，已知辨識器於搜尋第一階段使用詞二連語言模型分數、第二階段使用詞三連語言分數。如果以表 4-1-3 之實驗 2 於第二階段使用詞二連語言模型分數，其辨識率如表 4-1-5 實驗 1 所示，辨識第二階段利用詞三連語言分數可讓辨識率相對地提高 4.3%。

表 4-1-5　第二階段使用詞二連語言模型分數

實驗	語音特徵	第二階段 LM	混合數	詞正確率(%)
				WG
1	MFCC_CMS	詞二連	76,073	47.86
2	MFCC_CMS	詞三連	76,073	49.92

再者，吾人觀察僅使用 BNC 文字語料或 VOA 訓練語料之正確轉寫文字當語言模型之訓練語料，觀察語音辨識的影響。實驗結果如表 4-1-6 所示，所使用之語音特徵向量為 MFCC 配合 CMS。

表 4-1-6　VOA 不同語言模型之辨識結果

實驗	語言模型	混合數	詞正確率(%)	
			TC	WG
1	BNC	76,073	45.90	51.43
2	VOA	76,073	47.95	49.92

實驗觀察，使用 BNC 語料訓練語言模型比使用 VOA 語料之詞正確率相對地提升 3.02%，因 BNC 語料不僅包含與 VOA 統計特性較相關的會議或廣播新聞等文字語料，且 BNC 語料的量較大。

4.2 EAT 語料之基礎實驗

4.2.1 實驗設定

本論文所用 EAT 實驗語料設定如下表 4-2-1 所示，各使用 5,000 句英語系男生、女生，非英語系之男生、女生之四種語料組合成共 20,000 句(7.02 小時)當訓練語料。再另外選用英語系男生、女生，非英語系之男生、女生之四種語料組合各 250 句組合成共 1,000 句(0.65 小時)當測試語料。詞典個數為訓練語料與測試語料中的所有相異詞，共有 2,370 個詞。

表 4-2-1　EAT 實驗語料設定

語料分配	種類	句數	時間(hr)	詞彙數
	訓練語料	20,000	7.02	53,922
	測試語料	1,000	0.65	2,781
詞典個數	2,370 個			

4.2.2 基礎語音特徵擷取

表 4-2-2 代表使用三種不同基礎語音特徵擷取方式的辨識結果，此實驗之語言模型使用 EAT 訓練語料之正確轉寫文字訓練而成，高斯混合數依表 4-1-2 之分配規則分配並訓練聲學模型。

表 4-2-2　EAT 不同基礎特徵擷取法之辨識結果

實驗	語音特徵	混合數	詞正確率(%)	
			TC	WG
1	MFCC	145,319	29.69	40.04
2	MFCC_CMS	143,735	36.41	49.53
3	MFCC_CMVN	138,713	33.93	47.02

觀察發現以 MFCC 結合 CMS 之基礎特徵擷取法，得到較佳之詞正確率 49.53%，而 MFCC 較 MFCC_CMS 與 MFCC_CMVN 詞正確率低很多，表示 EAT 語料受通道效應(Channel Effects)非常嚴重。

4.2.3　基礎三連音素聲學模型

利用表 4-2-2 實驗 2 之 MFCC 配合 CMS 所擷取出的語音特徵來訓練聲學模型，觀察高斯混合數之差異是否影響辨識率結果。實驗設定高斯數目全部設定為 1(表 4-2-3 之實驗 1)，或依表 4-1-2 之分配規則(表 4-2-3 之實驗 2)、規則之 4 倍(表 4-2-3 之實驗 3)來訓練聲學模型，語言模型為 EAT 訓練語料之正確轉寫文字訓練而成，實驗結果如表 4-2-3 所示：

表 4-2-3　EAT 不同高斯混合數之辨識結果

實驗	語音特徵	混合數	詞正確率(%)	
			TC	WG
1	MFCC_CMS	25,375	30.12	40.55
2	MFCC_CMS	143,735	36.41	49.53
3	MFCC_CMS	549,953	36.45	49.35

觀察發現，高斯混合數依規則分配時，詞正確率由 40.55%提升至 49.53%，約提

升 22.14%，然而增加至規則的 4 倍時，辨識率卻些微下降，原因可能為每個高斯混合數被分配到的訓練語料不夠多，無法隨高斯混合數目增加而提高辨識率。

4.2.4 基礎語言模型

觀察僅使用 BNC 文字語料或 EAT 訓練語料之正確轉寫文字當語言模型之訓練語料，辨識結果如表 4-2-4 所示。高斯混合數依表 4-1-2 之規則分配、特徵擷取為 MFCC 配合 CMS。

表 4-2-4　EAT 不同語言模型之辨識結果

實驗	語言模型	混合數	詞正確率(%)	
			TC	WG
1	BNC	143,735	28.40	31.02
2	EAT	143,735	36.41	49.53

由實驗觀察，使用 EAT 訓練語料比使用 BNC 語料之語言模型辨識率提高 59.67%。可能原因為 EAT 與 BNC 語料的統計特性差異很大，EAT 語料中大多為英文單字、片語或數字連續語音，而 BNC 為開會或是廣播新聞等對話資料。

4.3 實驗討論

本節討論 4.1 與 4.2 實驗如下：

1. 基礎特徵擷取法：

   A. VOA 語料之特徵擷取以 MFCC 配合 CMS、高斯混合數依表 4-1-2 規則分配、語言模型為使用 VOA 訓練語料、可得到較佳詞正確率

(49.92%)。

B. EAT 語料之特徵擷取以 MFCC 配合 CMS、高斯混合數依表 4-1-2 規則分配、語言模型為 EAT 訓練語料、可得到較佳詞正確率(49.53%)。

2. 基礎三連音素聲學模型：

A. 於VOA語料使用MFCC配合CMS之特徵,觀察高斯混合數表4-1-2規則分配，可得到較佳的詞正確率(49.92%)。但是詞正確率沒有隨高斯混合數目增加而提高。

B. 於 EAT 語料使用 MFCC 配合 CMS 之特徵，觀察高斯混合數依表 4-1-2 規則分配，可得到較佳的詞正確率(49.53%)。詞正確率未隨高斯混合數目增加至規則的 4 倍而提高。因訓練語料量少，每個高斯混合數被分配到的訓練語料不夠多,無法隨高斯混合數目增加而提高辨識率。

3. 基礎語言模型：

A. VOA 語料之特徵擷取以 MFCC 配合 CMS、高斯混合數依表 4-1-2 規則分配，使用 BNC 語料訓練之語言模型，因統計特性較相關且內容更豐富，對詞正確率提升有幫助。

B. EAT 語料之特徵擷取僅以以 MFCC 配合 CMS、高斯混合數依表 4-1-2 規則分配、語言模型 EAT 訓練語料之人工轉寫語料訓練而成，比僅以 BNC 語料或結合訓練之語料，有較佳之詞正確率(49.53%)。可能原因為 EAT 語料中大多為英文單字、片語或數字連

續語音，與 BNC 語料為會議或是廣播新聞等對話資料的統計特性差異較大。

第5章　改進英文辨識之各項實驗

5.1　鑑別性特徵擷取

資料相關線性特徵轉換(Data-Driven Linear Feature Transform)近幾年在語音特徵擷取的研究上佔有相當重要的地位，因為資料相關線性特徵轉換可以藉由統計訓練資料來自動地找出特徵空間中重要的基底向量(Basis Vectors)，使得經轉換後的特徵能保有重要的成份或具有較高的鑑別力，並且可以進一步去除多餘的維度，由於基底向量是根據訓練資料而來，所以找出的基底向量將較能代表語音訊號的特徵[Hung *et al.* 2001]。

本論文嘗試使用資料相關線性特徵轉換中的線性鑑別分析(LDA)、異質性線性鑑別分析(HLDA)，並結合最大相似度線性轉換(MLLT)方法於語音特徵擷取。

線性鑑別分析也可以以最大相似度(Maximum-likelihood)估測法[Campbell 1984]來解釋，作法為使用訓練資料的類別資訊來統計各類別的分布，求取一個轉換矩陣再藉此作線性轉換與特徵降維，其目的在於使得轉換後特徵之間可以保有最大的分類鑑別資訊。期望轉換後類別內(Within-Class)的分布越凝聚越好，而類別間(Between-Class)的分布距離越遠越好；也就是說轉換後，類別內的變異量越小越好，而類別間的變異量越大越好。

線性鑑別分析假設各類別分布的變異量相同[Campbell 1984]，然而現實上大多的訊號特徵分布的變異量皆為異質性。異質性線性鑑別分析[Kumar 1997][Kumar and Andreou 1998]假設各類別分布的變異量為異質性(Heteroscedastic)，去除鑑別性特徵係數法中各類別分布變異量相同的限制，同樣再以最大相似度估

測目標函式(Objective Function)，以求得較具鑑別性的語音特徵向量，詳細異質性線性鑑別分析推導可參見[張志豪 2005]。

本論文使用線性鑑別分析或異質性線性鑑別分析之特徵擷取，再加入最大化相似度線性轉換[Gopinath 1998；Saon *et al.* 2000]，因為目前我們使用的隱藏式馬可夫模型為對角化(Diagonal)之共變異矩陣，最大相似度線性轉換目的為保留矩陣維度，並使轉換後類別的共變異矩陣對角化。

本論文使用不同特徵擷取方式與語音強健性技術。VOA 語料的辨識結果如表 5-1-1 所示。觀察可知，VOA 語料之前端擷取利用 LDA 配合 MLLT 與 CMVN、高斯混合數目依規則分配、語言模型使用 VOA 訓練語料訓練，可得詞正確率(57.21%)。

表 5-1-1　VOA 不同特徵擷取法之辨識結果

實驗	語音特徵	混合數	詞正確率(%)	
			TC	WG
1	MFCC	78,412	46.95	48.75
2	MFCC_CMS	76,073	47.95	49.92
3	MFCC_CMVN	73,083	47.12	49.10
4	LDA+MLLT_ CMVN	70,672	51.82	57.21
5	HLDA+MLLT_ CMVN	71,627	49.72	52.95

另一方面，EAT 語料之語音特徵利用基礎與線性鑑別式擷取，結果如表 5-1-2 所示，觀察可知，EAT 語料之語音擷取利用 HLDA 配合 MLLT 與 CMVN、高斯混合數目依規則分配、語言模型僅為 EAT 訓練語料訓練，可得最佳詞正確率(59.71%)。

表 5-1-2　EAT 不同特徵擷取法之辨識結果

實驗	語音特徵	混合數	詞正確率(%)	
			TC	WG
1	MFCC	145,319	29.69	40.04
2	MFCC_CMS	143,735	36.41	49.53
3	MFCC_CMVN	138,713	33.93	47.02
4	LDA+MLLT_CMVN	138,289	47.30	59.53
5	HLDA+MLLT_CMVN	141,333	46.48	59.71

5.2　語言模型調適

統計式語言模型調適技術，為利用大量語料訓練背景(Background)語言模型，這些語料包含許多領域和主題，可以從中求得一般性(Genaral)的自然語言規則。另外再準備調適語料，此為包含較少量的語料，並與欲辨識語料相關，利用調適語料中所取得的資訊來調適背景語言模型。本論文利用 BNC 文字語料當成背景語言模型之訓練語料，並以訓練聲學模型之正確人工轉寫當作調適語料。

經由調適語料中取得的 N 連詞頻(N-gram Count)：N 連詞出現於訓練語料的次數，可透過不同方式來調適背景語言模型，以求得語言模型分數 $P(w_i \mid h_i)$ 之值，其中 h_i 是詞 w_i 的歷史詞序列。常見的語言模型調適法有詞頻數混合法(Count Merging)與模型插補法(Interpolation)。此兩種方法可視為最大事後機率(Maximun A Posteriori, MAP)[Bacchiani *et al.* 003]調適法的一種。

詞頻數混合法作用在詞頻數階級(Frequency Count Level)，算法如式(5-1)所示，其中 $C_B(\bullet)$ 表某個詞或詞序列在背景語料中出現的次數、$C_A(\bullet)$ 表在調適語料中出現的次數，α 與 β 可藉由期望值最大化(Eapectation-Maximization, EM)[Dempster *et al.* 1977]演算法求得。

$$P(w_i \mid h_k) = \frac{\alpha \cdot C_B(h_k w_i) + \beta \cdot C_A(h_k w_i)}{\alpha \cdot C_B(h_k) + \beta \cdot C_A(h_k)} \tag{5-1}$$

線性插補法作用在模型階段(Model Level)，算法如式(5-2)所示，其中λ代表線性插補法的係數，也可視為各個機率分布的權重。

$$P(w_i \mid h_k) = \lambda P_A(w_i \mid h_k) + (1-\lambda) P_B(w_i \mid h_k) \tag{5-2}$$

5.2.1 詞頻數混合法

表 5-2-1 為 VOA 詞頻數混合法之語言模型之辨識結果，所用之語音特徵為 LDA 加上 MLLT 配合 CMVN，高斯混合數目為 76,672 個，而語言模型調適法中之 α 與 β 值設定如表所示，觀察背景語料與調適語料混合，能比原先單獨使用調適語料的辨識率高，因 BNC 語料不僅包含與 VOA 統計特性較相關的會議或廣播新聞等文字語料，且 BNC 語料的量較大。1:100 59.14

表 5-2-1　VOA 詞頻數混合法之辨識結果

實驗	α	β	語言模型	詞正確率(%)	
				TC	WG
1	1	0	BNC	47.68	56.60
2	0	1	VOA	51.82	57.21
3	1	1	BNC+VOA	49.51	59.04
4	1	50	BNC+VOA*50	48.96	59.07
5	1	100	BNC+VOA*100	48.91	59.14

表 5-2-2 為 EAT 詞頻數混合法之語言模型對之辨識結果，所用之語音特徵為 HLDA 配合 CMVN，高斯混合數為依規則分配，共 141,333 個，語言模型調適法中之α與β值設定如表所示，觀察實驗結果發現，BNC 語料對 EAT 之辨識率提

升效果非常小，可能原因為 EAT 與 BNC 語料的統計特性差異很大，EAT 語料中大多為英文單字、片語或數字連續語音，而 BNC 為開會或是廣播新聞等對話資料。

表 5-2-2　EAT 不同語言模型之辨識結果

實驗	α	β	語言模型	詞正確率(%)	
				TC	WG
1	1	0	BNC	37.46	34.72
2	0	1	EAT	46.48	59.71
3	1	1	BNC+EAT	37.20	38.50
4	1	100	BNC+EAT*100	44.28	48.24

5.2.2　線性插補法

表 5-2-3 與圖 5-2-1 代表利用表 5-2-1 實驗 1 之以 BNC 為背景語料實驗之第一階段產生之詞圖資訊，代入詞三連機率之線性插補法，其中 λ 值為調適語料所佔比重。當調適模型與背景模型比重為 0.15 與 0.85，可得較佳的詞正確率 59.04%。

表 5-2-3　VOA 語言模型線性插補法之辨識結果

調適模型比重(%)	詞正確率(%)	調適模型比重(%)	詞正確率(%)
0.00	56.60	0.55	58.72
0.05	58.70	0.60	58.61
0.10	59.07	0.65	58.36
0.15	59.04	0.70	58.40
0.20	59.04	0.75	58.47
0.25	59.18	0.80	58.22
0.30	58.95	0.85	57.92
0.35	58.70	0.90	57.58
0.40	58.66	0.95	56.92
0.45	58.56	1.00	51.82

0.50	58.66	-	-

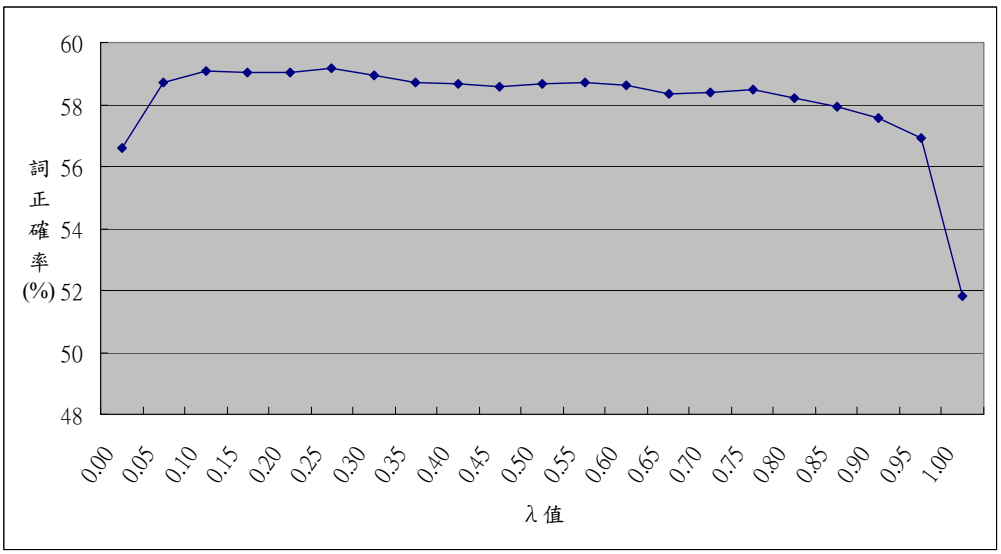


圖 5-2-1　VOA 語言模型線性插補法辨識結果示意圖

5.3　模糊矩陣之使用

本論文利用英文辨識器之第二階段辨識結果，與正確轉寫文字做單連音素、三連音素與詞之比對，統計發生「取代」的個數，利用模糊矩陣(Confusion Matrix)法統計並正規化(Normalized)容易辨識錯誤的個數。取出大於門檻值之辨識錯誤情況，將單連音素辨識錯誤統計結果，加入訓練聲學模型階段之決策樹問題條件，觀測辨識率之變化。以及將三連音素辨識錯誤統計結果，應用於辨識器搜尋階段，觀測辨識率之變化。

5.3.1　聲學模型訓練階段使用

此節實驗利用表 5-2-2 實驗 4 之 VOA 內測試(Inside test)語料，詞正確率為 78.05%之實驗做模糊矩陣之實驗，門檻值設定為 0.5，取出大於門檻值之辨識錯誤單連

音素統計結果，結果如表 5-3-1 所示：

表 5-3-1　VOA 內測試語料之模糊矩陣之單連音素辨識錯誤統計表

正確音素	辨識音素	次數正規化	正確音素	辨識音素	次數正規化
ng	ae	1.00	ih	ae	0.50
oy	ay	1.00	jh	b	0.50
s	z	0.67	jh	uw	0.50
z	s	0.60	l	w	0.50
er	r	0.57	p	t	0.50
ey	l	0.50	p	uw	0.50
ey	t	0.50	-	-	-

將表 5-3-1 之辨識音素變化應用於訓練聲學模型階段分裂決策樹之問題條件，重新訓練聲學模型與辨識，當門檻值設為 0.5，得到新的詞正確率為 78.10%；門檻值降低為 0 時，新的詞正確率為 78.07%，此規則對模型辨識率之提升影響力較小，可能原因為額外加入的問題條件占訓練語料之機率值過小。

5.3.2　辨識器搜尋階段使用

本論文統計語料之正確人工轉寫與辨識結果，找出編輯距離(Levenshtein Distance)中每個三連音素 M 「取代」(Substitution)成 $N_1...N_k$ 的次數正規化值，以 A_{MN_i} 表示，其中 $i=1...k$ 且 $\sum_{i=1}^{k} A_{MN_i}=1$。再將此模糊矩陣挑選門檻值大於 α 以上的結果，代入英文辨識器，重新計算每個時間點每個狀態的機率值，以 $\widetilde{B}_M$ 表示，$\widetilde{B}_M$ 計算方式如式(5-3)所示，其中 λ 代表原本三連音素 M 之狀態機率值所佔比列：

$$\widetilde{B}_M=\lambda B_M+(1-\lambda)\sum_{i=1}^{k}\left(A_{MN_i}\times B_{N_i}\right) \tag{5-3}$$

圖 5-3-1 第一列代表編號 10 號的三連音素容易被取代成編號 12 號，正規化次數為 0.5，其他變異如表所示。當 α 值設定為 0.4，則取出編號 10 的所有變異狀況於辨識器搜尋階段。

M	N	A_{MN}
10	12	0.5
10	15	0.5
12	16	0.4
102	140	0.4
:	:	:

$\alpha = *$

圖 5-3-1　模糊矩陣示意圖

本節實驗利用論文 5.4.2 之實驗設定，訓練聲學模型，利用辨識結果建立模糊矩陣。表 5-3-2 為 EAT 測試語料之單連音素辨識錯誤統計表，取出 0.2 以上的單連音素變化。

表 5-3-2　EAT 測試語料之單連音素辨識錯誤統計

正確音素	辨識音素	次數正規化	正確音素	辨識音素	次數正規化
z	s	0.38	ay	ax	0.25
sh	s	0.38	ay	t	0.25
jh	r	0.33	k	t	0.23
jh	t	0.33	uh	ax	0.23
zh	ax	0.33	m	n	0.23
zh	l	0.33	ao	ow	0.23
zh	sh	0.33	ch	n	0.22
aw	l	0.30	th	s	0.22
ng	n	0.29	b	f	0.21
d	t	0.27	l	r	0.20
aw	aa	0.25	iy	ih	0.20

表 5-3-3 代表將 EAT 測試語料之三連音素模糊矩陣變化，應用於辨識器階段

之詞正確率，α 全為 0，代表模糊矩陣找出之變異全部加入辨識器搜尋階段，λ 為 0.97 時有最佳詞正確率 62.98%。

表 5-3-3　EAT 測試語料之模糊矩陣應用於辨識器階段之詞正確率

		詞正確率(%)	
實驗	λ	TC	WG
1	-	50.14	57.84
2	0.20	39.25	48.15
3	0.50	40.84	52.36
4	0.80	52.14	62.85
5	0.97	52.87	62.98

如果使用大量 EAT 語料進行辨識，將其辨識結果與正確轉寫文字比對，建立一般化(General)模糊矩陣，再將此矩陣應用於辨識階段，詞正確率如表 5-3-4 所示。

表 5-3-4　EAT 一般化模糊矩陣應用於辨識器階段之詞正確率

			詞正確率(%)	
實驗	λ	α	TC	WG
	-		50.14	57.84
1	0.80	0.0	48.13	55.07
2	0.97	0.0	50.18	57.40
3	0.97	0.1	50.32	57.40
4	0.97	0.3	50.54	57.94

觀察實驗結果，當 α 為 0.3、λ 為 0.97 時有最佳詞正確率 57.94%，比原本 57.84 之辨識率提高 0.17%。

如果利用信心度評估(詳見論文 5.4.1、5.4.2)，選出辨識結果中的詞信心度為 1 的轉寫文字(26,810 句)，由此建立一般化模糊矩陣，再將矩陣加入辨識階段，詞正確率如表 5-3-5 所示。觀測當 α 為 0.3、λ 為 0.97 時有最佳詞正確率 58.41%，比原本 57.84 之辨識率提高 0.98%。並與表 5-3-4 比較，使用信心度挑選出的語句對詞正確率有提升效果。可知使用信心度評估法可挑選出適當語句且其變異狀況較有代表性。

表 5-3-5　EAT 一般化模糊矩陣應用於辨識器階段之詞正確率

			詞正確率(%)	
實驗	λ	α	TC	WG
	-		50.14	57.84
1	0.80	0.0	49.09	56.06
2	0.97	0.0	50.19	57.34
3	0.97	0.1	51.44	57.98
4	0.97	0.3	51.10	58.41

5.4　非監督式聲學模型訓練

過去研究語音辨識，語料的來源取得較為困難，因需人工特別去錄製，且需大量人力利用聽寫的方式，找出訓練語料對應的正確轉寫文字(True Transcription)及詞與音素之邊界。如圖 5-4-1(a)所示，稱其為監督式(Supervised)聲學模型訓練，此法須人工的介入，相當費力耗時。

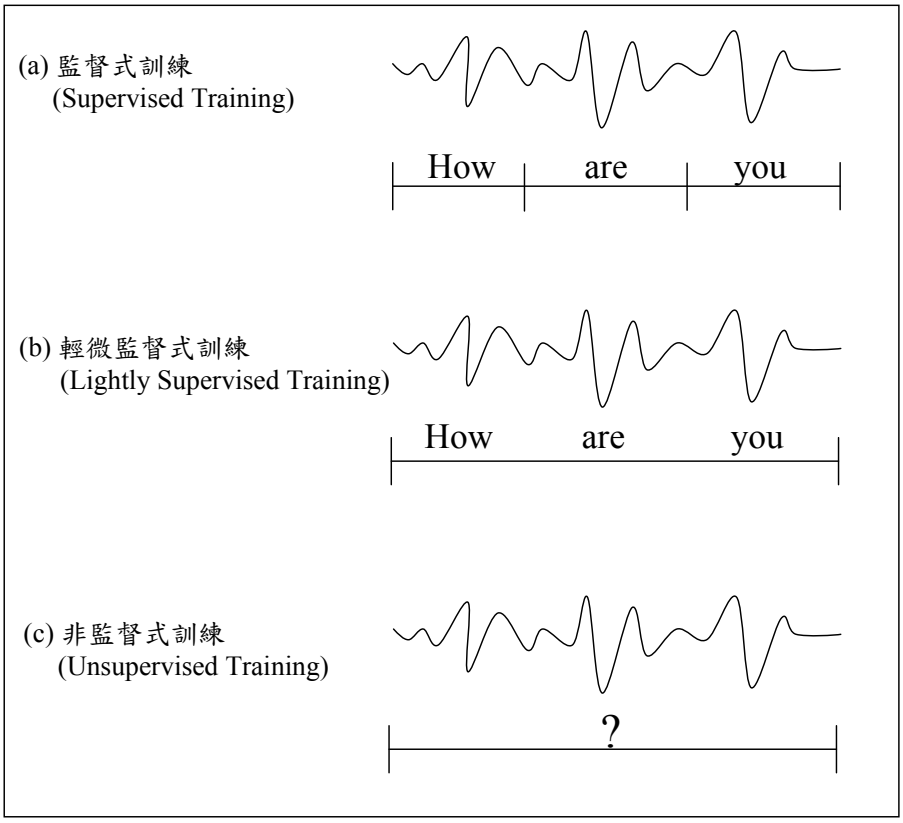


圖 5-4-1　三種訓練聲學模型方式示意圖

隨著科技發達，語料的來源可以有很多，使得大量的多媒體資料存在於網路中，例如：電視新聞、廣播新聞、多媒體影音等，使得收集語料不再是困難的工作。但是收集到的語料通常沒有正確的轉寫文字(如廣播新聞)，或者僅有近似的正確文字，如現場直播(Live)的電視新聞有人工即時(Real-time)將字幕(Closed-caption)打出，然而這些字幕可能並不視為完整的人工轉寫。利用字幕或是近似且沒有正確詞或音素邊界之語料來訓練聲學模型的方法，稱為輕微監督式(Lightly Supervised)聲學模型訓練[Lamel *et al.* 2002]，如圖 5-4-1(b)所示。如果只有語料而沒有正確轉寫文字和詞或音素邊界，稱其為非監督式(Unsupervised)聲學模型訓練[Wessel *et al.* 2001, Chen *et al.* 2004b]，如圖 5-4-1(c)所示。

對於大詞彙連續語音辨識研究中，訓練語料的多寡對模型的強健度佔了很重要的因素，例如三連音素聲學模型訓練實驗中，如果訓練語料量夠多，三連音素

平均出現的機會也增多，辨識率相對有提升之效果，但是在語料隨手可得的今日，我們仍無法有效地提升自動語音辨識器(ASR)效能，這是因為取得的大量語料可能不具有正確轉寫文字，且字幕取得也不容易，故如果想要有正確的轉寫文字或詞與音素的邊界，必須投入大量人力去標註，非常耗時。如果不想花費大量人力去轉寫正確的文字，可用現有的自動語音辨識器去辨識大量未轉寫之語料，再進行非監督式模型訓練，在大詞彙連續語音辨識研究中，非監督式模型訓練為一重要議題。

非監督式之聲學模型訓練通常利用最大化相似度(Maximum Likelihood)估測法來達成模型參數最佳化，作法為利用現有的人工轉寫過少量的語料，訓練初始聲學模型，利用此聲學模型對大量未經人工轉寫的語料做辨識，利用辨識後第一名(Top 1)辨識結果當成正確轉寫文字，再利用第一名之辨識結果的大量語料與現有之人工語料重新利用最大化相似度訓練法訓練聲學模型。

5.4.1 信心度評估法

非監督式聲學模型訓練中，利用初始之聲學模型對大量未經人工轉寫的語料做一次辨識，利用辨識後之詞圖產生第一名之辨識結果當成正確轉寫文字，再重新訓練初始之聲學模型。但是語音辨識總有辨識錯誤產生，如果拿錯誤的轉寫文字去訓練模型，會使訓練出的聲學模型不正確，降低辨識效能。信心度評估[Wessel *et al* 2001]法為判斷辨識結果的可靠度，給定辨識結果一個分數，如 0 至 1 之間的實數值，再設定門檻值，選出大於門檻值之語料與原本語料重新訓練模型。

本論文利用的信心度評估為事後機率法中的圖形化基礎(Graph-based)法來求詞圖中每個詞段(Word Arc)的信心度值。假設某一段語音特徵序列為O，詞圖Ψ^O中每個節點代表一個時間點，每個詞段(Word Arc)a由三個變數組成，

$a:[w_a;s_a,e_a]$，其中 w_a 代表為詞編號、s_a 代表詞段開始時間、e_a 代表詞段結束時間，每個詞段產生這個語音段落的聲學分數 $P(O_{s_a}^{e_a} \mid w_a)$，且每個詞圖有兩個特殊節點，分別為詞圖的開始與結束(如圖 3-13 之方形節點)，只要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)，而任一條完整路徑代表某一條聲學觀測值之辨識詞序列。在詞圖 Ψ^O 上利用前向後向演算法計算某詞段 $a:[w_a;s_a,e_a]$ 的事後機率如式(5-4)所示：

$$P(a:[w_a;s_a,e_a] \mid \Psi^O) = \frac{\sum_{\{\overline{W}:[w^n;s^n,e^n]_{n=1}^{N}\}\in\Psi^O, a\subset\overline{W}} \left\{\prod_{n=1}^{N} p\left(O_{s^n}^{e^n} \mid w^n\right)\cdot P\left(w^n \mid h^n\right)\right\}}{\sum_{\{\overline{W}:[w^m;s^m,e^m]_{m=1}^{M}\}\in\Psi^O} \left\{\prod_{n=1}^{M} p\left(O_{s^m}^{e^m} \mid w^m\right)\cdot P\left(w^m \mid h^m\right)\right\}} \tag{5-4}$$

其中，$\overline{W}$ 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^n 為 w^n 的詞歷史(Word History)，$p\left(O_{s^n}^{e^n} \mid w^n\right)$ 代表開始時間 s^n 至結束時間 e^n 此段語音特徵序列的聲學相似度、$P\left(w^n \mid h^n\right)$ 代表語言模型分數。於實作時，求得每個詞段的信心度，再利用維特比(Viterbi)動態搜尋解碼得第一名之詞序列，此詞序列中的每個詞都有一個信心度值，再利用事先訂好的門檻值(Threshold)，來決定第一名詞序列中的某個詞是否拿來作聲學模型訓練，本論文挑選詞序列其信心度值全為 1 之句子。

研究指出[Wessel *et al* 2001b]，非監督式之模型訓練應以迭代方式實現，即以現有人工轉寫語料訓練之聲學模型，對未轉寫的語料做一次辨識，再將第一名辨識結果與現有人工轉寫語料再次訓練聲學模型，迭代(Iterative)過程可以有多次，且信心度評估於迭代過程中，應隨迭代次數而門檻值降低，因一開始之聲學模型之辨識率較低，所以門檻值設高，可過濾許多辨識錯誤語句，當幾次迭代後，可得到較佳的聲學模型。

5.4.2 實驗設定與結果

本論文利用 EAT 語料做非監督式最大化相似度聲學模型訓練，所使用的語音特徵、語料、詞典個數、語言模型設定如下表 5-4-1 所示。首先監督式訓練語料與測試語料同 4.2.1 節之實驗設定，另外再使用其他更多的英語系男生、女生，與非英語系男生、女生之語料混合，共 33.4 小時當非監督式訓練語料。而詞典個數為出現在監督式與非監督式訓練語料與測試語料中所有相異詞，共 4,229 個詞。

表 5-4-1　EAT 語料之非監督式最大化相似度聲學模型訓練實驗設定

語音特徵	HLDA+MLLT+CMVN			
實驗語料	種類	句數	時間(hr)	詞彙數
	監督式訓練語料	20,000	7.02	53,922
	非監督式訓練語料	42,960	33.4	108,323
	測試語料	1,000	0.65	2,781
詞典個數	4,229 個			
語言模型	監督式訓練語料之轉寫文字			

首先尋找非監督式之聲學模型上界，實驗流程如圖 5-4-2 所示。首先針對 7.02 小時之監督式語料訓練三連音素狀態分享模型、高斯混合數依據規則分配(表 4-1-2)，訓練之聲學模型為 HMM(1)(圖 5-4-2 中)。再將 7.02 與 33.4 小時監督式語料混合，依據訓練語料量重新分配並依規則分配高斯混合數目，與聲學模型 HMM(1)重新訓練成聲學模型 HMM(2)(圖 5-4-2 中)。HMM(2)為本節實驗詞正確率上界之聲學模型。詞正確率為 64.74%，如表 5-4-2 所示。

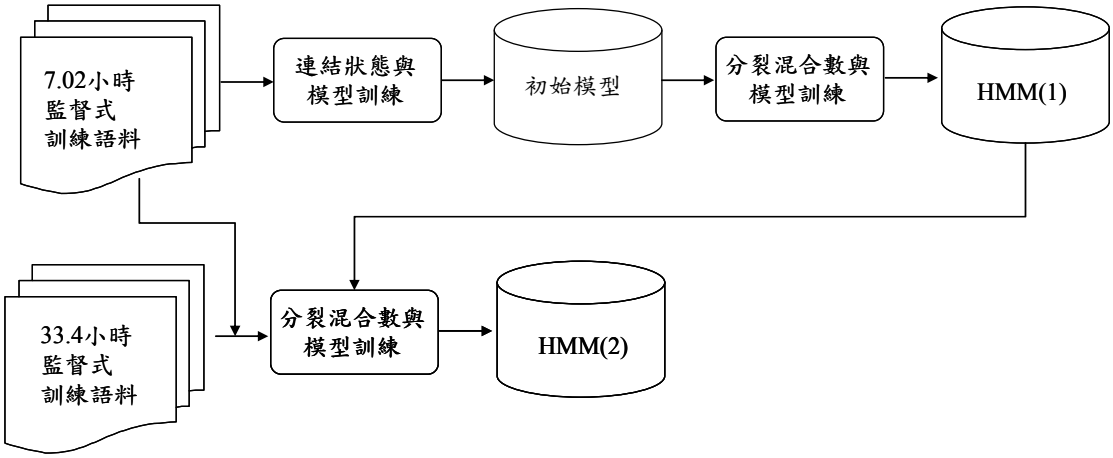


圖 5-4-2　非監督式之聲學模型上界

表 5-4-2　非監督式之聲學模型上界之詞正確率

實驗	聲學模型	混合數	詞正確率(%)	
-	-	-	TC	WG
1	HMM(1)	141,333	50.14	57.84
2	HMM(2)	216,318	56.29	64.74

非監督式模型訓練實驗流程如圖 5-4-3 所示。首先對 7.02 小時之監督式語料訓練三連音素狀態分享模型、高斯混合數依據規則分配，訓練聲學模型(圖 5-4-3 中 HMM(1))。再對 33.4 小時非監督式訓練語料做辨識，將辨識結果當成標記文字，與原本 7.02 小時訓練語料混合，重新分配高斯混合數並訓練聲學模型(如圖 5-4-3 中 HMM(3))，詞正確率為 51.73%(表 5-4-3)。如利用信心度評估法，尋找 33.4 小時辨識結果中詞序列之信心度全為 1 的轉寫文字，與原本 7.02 小時訓練語料結合，重新分配高斯混合數並訓練聲學模型(如圖 5-4-3 中 HMM(4))，詞正確率為 58.20%(表 5-4-3)。

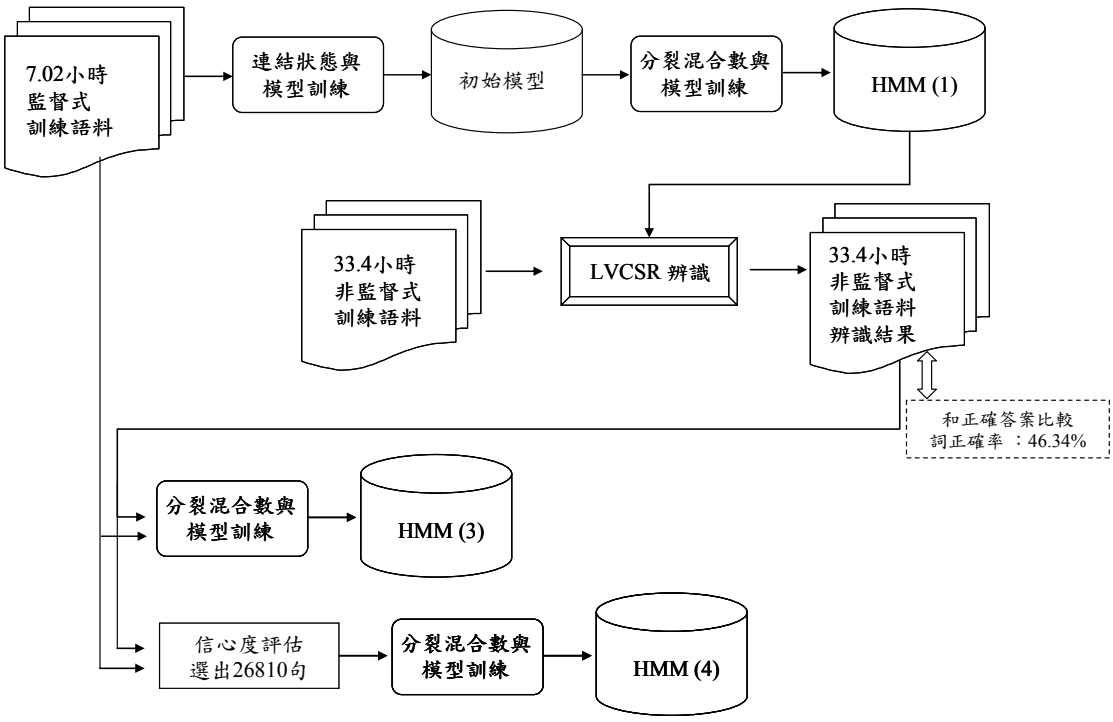


圖 5-4-3　非監督式之訓練示意圖

表 5-4-3　非監督式之訓練之詞正確率

實驗	聲學模型	混合數	詞正確率(%)	
-	-	-	TC	WG
1	HMM(1)	141,333	50.14	57.84
2	HMM(3)	221,820	49.78	51.73
3	HMM(4)	191,314	50.86	58.20

由實驗得知，33.4 小時之大量辨識結果與其正確轉寫文字比對，詞正確率為 46.34%，代表辨識結果含有 53.66%的錯誤比重，如果將全部辨識結果與 7.02 小時監督式語料之正確轉寫文字混合訓練聲學模型，將使詞正確率下降，因其中含有許多錯誤之轉寫文字，讓訓練的聲學模型不正確，降低辨識效能。如果以信心度評估法選出符合條件之語句，將使詞正確率提升 0.62%，代表信心度評估法能選出較正確之辨識語句，重新與原本訓練語料結合訓練模型，讓聲學模型更強健。

5.5 實驗討論

本節討論 5.1 至 5.4 實驗如下：

1. 鑑別性語音特徵擷取：

   A. VOA 語料之特徵擷取以 LDA 配合 MLLT 與 CMVN、高斯混合數依規則分配、語言模型使用 VOA 訓練語料訓練，可得較佳詞正確率(57.21%)。

   B. EAT 語料之特徵擷取以 HLDA 配合 MLLT 與 CMVN、高斯混合數依規則分配、語言模型為 EAT 訓練語料訓練，可得較佳詞正確率(59.71%)。

2. 語言模型調適：

   A. VOA 語料實驗，利用語言模型調適中的詞頻數混合法，以式(5-1)之α與β值，即 BNC 背景語料與 VOA 調適語料設定比重為 1 比 100 的情況下，語音特徵為 LDA 加上 MLLT 配合 CMVN，得到詞辨識率為 59.14%，觀察背景語料與調適語料混合，能比原先單獨使用調適語料的辨識率高，因 BNC 語料不僅包含與 VOA 統計特性較相關的會議或廣播新聞等文字語料，且 BNC 語料的量較大。

   B. VOA 語料之語言模型線性插補法實驗，將表 5-2-1 實驗 1 以 BNC 文字語料為背景語料，於第一階段辨識後詞圖產生的資訊，代入詞三連機率之線性插補，當調適模型與背景模型比重為 0.15 與 0.85 時，能讓詞正確率相對地提高 4.31%。

C. EAT 語料利用語言模型調適法中的詞頻數混合法，以式(5-1)之α與β值，即 BNC 背景語料與 EAT 調適語料設定比重為 0 與 1 的情況下，語音特徵為 HLDA 配合 MLLT 與 CMVN，得到詞辨識率為 59.71%，觀察 BNC 背景語料對 EAT 實驗影響非常小，因 EAT 語料中大多為英文單字、片語或數字連續語音，與 BNC 的統計特性差異很大。

3. 音素模糊矩陣之應用：

A. 尋找辨識結果與正確解答之音素取代情況，製作模糊矩陣並應用於訓練聲學模型階段之決策樹問題條件，觀測對辨識率之影響不大，可能原因為額外加入的問題條件占訓練語料之機率值過小，故不影響辨識率。

B. 應用音素模糊矩陣於辨識器搜尋階段，調整語音特徵向量之觀測機率，如果將辨識結果與正確解答比對建立後之模糊矩陣加入搜尋階段，可讓詞正確率提高。

4. 非監督式聲學模型訓練：

非監督式之聲學模型訓練結合信心度評估法，能選出非監督式訓練語料中較具正確性之語句，重新與原本監督式訓練語料混合訓練，讓聲學模型更強健。

第6章　結論與未來展望

本論文為英文連續語音辨識之初步研究，我們實作英文連續語音辨識器，並探討其主要組成，包含語音特徵擷取、聲學模型及語言模型等之改進方法：

1. 前端語音特徵擷取，分別使用基礎特徵(MFCC)與鑑別性特徵(LDA、HLDA)配合強健性語音特徵技術(CMS、CMVN)，觀察其在語音辨識之表現。

2. 針對聲學模型，探討詞內三連音素模型、狀態連結技術，並增加高斯混合數目以提升語音辨識率。

3. 針對語言模型，在語音辨識過程中利用詞頻數混合法與模型插補法等語言模型調適方法來結合背景與同領域語言模型訓練語料，以達到較佳之詞發生預測。

4. 利用模糊矩陣尋找台灣腔英語發音變異，基於所觀察之變異情況來修改訓練聲學模型之狀態分享規則問題條件。此外，亦根據發音變異情況於語音辨識搜尋階段修正語音向量在隱藏式馬可夫狀態之觀測機率。

5. 探討非監督式聲學模型訓練，首先對大量語料進行語音辨識，並使用語料及經辨識後自動轉寫文字資訊重新訓練聲學模型。

表 6 為本論文使用 VOA 與 EAT 兩套語料，語音辨識系統在使用較好之語音特徵擷取法、三連音素狀態分享之聲學模型、以及搭配語言模型調適法下所得到之最佳辨識結果。其中對於 VOA 語料而言，在語音特徵擷取使用 LDA 配合 MLLT 與 CMVN 語音強健方法、語言模型使用 BNC 文字語料為背景語料加上以 VOA

語料之人工轉寫為調適語料(以 1 比 1 之比重)之混合訓練下可得 59.89%之詞正確率。另一方面，對於 EAT 語料而言，在特徵擷取使用 LDA 配合 MLLT 與 CMVN 語音強健方法、語言模型僅使用 EAT 訓練語料之人工轉寫下可得 65.71%之詞正確率。

表 6　VOA 與 EAT 實驗語料最佳設定與詞正確率

	VOA	EAT
前端特徵種類	LDA+MLLT+CMVN	HLDA+MLLT+CMVN
訓練語料	3.33 小時(5340 句)	40.42 小時(62906 句)
測試語料	0.56 小時(500 句)	0.65 小時(1000 句)
詞典個數	5,178 個	4,229 個
高斯混合數	70,672 個(依規則)	216,310 個(依規則)
模型數目	4,373 個	8,850 個
語言模型	BNC+VOA(1:1)人工轉寫訓練語料	EAT 人工轉寫訓練語料
詞正確率	59.89 %	65.71 %

本論文的研究發現，語音特徵擷取方式、聲學模型與語言模型之訓練、以及語音辨識之語言解碼階段之技術是影響連續語音辨識的重要因素。因此計畫於未來針對下列幾項變因做改進：

1. 增加聲學模型之訓練語料量，提高三連音素之訓練資料出現次數，以減少資料稀疏問題。

2. 豐富語言模型訓練語料，並使用其他層次的語言資訊，如詞類別、語意等。

3. 增加系統辨識速度。

4. 使用鑑別式聲學模型訓練，如最小化音素錯誤(Minimum Phone Error, MPE)訓練，以提高模型辨識率。

參考文獻

[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.

[Bacchiani *et al.* 2003] M. Bacchiani and B. Roark.”Unsupervised Language Model Adaptation, “In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 2003.

[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

[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.

[Bayeh *et al.* 2004] R. Bayeh et al., “Towards multilingual speech recognition using data driven source/target acoustical units association”, ICASSP’04, vol. I, pp. 521-524, Montreal, Canada, May 2004.

[Beyerlein *et al.* 1999] P. Beyerlein et al., “Towards language independent acoustic modeling”, ASRU’99, Keystone, CO, USA, December 1999.

[BNC corpus] British National Corpus：http://www.natcorp.ox.ac.uk/

[Brian Mak *et al.*1996] Brian Mak, E. Barnard, “Phone Clustering Using the Bhattacharyya Distance,” ICSLP ‘96, volume 4, pages 2005-2008, 1996

[Campbell 1984] N. Campbell, “Canonical Variate Analysis – a general formulation,”

Australian Journal of Statistics, 1984.

[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.

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

[Chen and Goodman 1999] S. F. Chen, J. Goodman. An Empirical Study of Smoothing Techniques for Language Modeling. Computer Speech and Language, 13, 1999.

[Colthurst *et al.* 2000] Thomas Colthurst, Owen Kimball, Fred Richardson, Han Shu, Chuck Wooters, Rukmini Iyer, Herbert Gish,"The 2000 BBN Byblos LVCSR System,"In ICSLP-2000, vol.2, 1011-1014.

[Davis *et al.* 1980] DAVIS, S. et MERMELSTEIN, P. "Comparison of parametric representation for mononsyllabic word recognition in continuously spoken sentences." IEEE International Conference on Acoustics, Speech and Signal Processing, 28(4):357–366.1980.

[Dempster et al. 1997] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, Volume 39, no. 1, pages 1-38, 1977.

[EAT corpus] English Across Taiwan：http://www.aclclp.org.tw/

[EARS] EARS at ICSI , http://www.icsi.berkeley.edu/Speech/EARS/index.html

[Evermann *et al*. 2004] G. Evermann, H.Y. Chan, M.J.F. Gales, T. Hain, X. Liu, D. Mrva, L. Wang, P.C. Woodland, " DEVELOPMENT OF THE 2003 CU-HTK CONVERSATIONAL TELEPHONE SPEECH TRANSCRIPTION SYSTEM,"in Proc. ICASSP 2004

[Evermann *et al.* 2003] G. Evermann & P.C. Woodland," DESIGN OF FAST LVCSR SYSTEMS," in Proc. ASRU,2003

[Festlex CMU] Festlex CMU：http://linux.maruhn.com/sec/festlex_cmu.html

[Fiscus 1997] J. G. Fiscus, "A post-processing system to yield reduced word error rates: recognizer output voting error reduction (ROVER)," IEEE ASRU Workshop,

1997.

[Furui 1981] S. Furui, “Cepstral Analysis Technique for Automatic Speaker Verification,” IEEE Trans. Acoust. Speech Signal Process, 1981.

[Gales & Woodland 1996] M. J. F. Gales and P. C. Woodland (1996). “Mean and Variance Adaptation within the MLLR Framework,” *Computer Speech and Language,* Vol. 10, pp.249-264, 1996.

[Gopinath 1998] R. A. Gopinath, “Maximum likelihood modeling with Gaussian distributions,” In Proceedings of ICASSP, Seattle, 1998.

[Gray *et al.* 1973] J.D. Markel, A.H. Gray, and H. Wakita,”Linear Prediction of Speech-Theory and Practice”, SCRL Monograph No. 10, Speech Communications Research Laboratory, Santa Barbara, California, 1973.

[Gunawardana & Byrne 2001] A. Gunawardana and W. Byrne (2001).“Discriminative Speaker Adaptation with Conditional Maximum Likelihood Linear Regression,” in *Proc. Eurospeech’01*.

[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.

[Hermansky 1990] Hermansky, H. “Perceptual linear predictive (PLP) analysis of speech”, J. Acoust. Soc. Am., 87(4), pp. 1738-1752.1990.

[Huo *et al.*1995] Qiang Huo, Chorkin Chan and Chin-Hui Lee,”Bayesian Adaptive Learning of the Parameters of Hidden Markov Model for Speech Recognition,” IEEE Trans. on Speech and Audio Processing, Vol. 3, No. 5, pp.334-345, 1995.

[Hung *et al.* 2001] J-W Hung, H-M Wang and L-S Lee, “Comparative Analysis for Data-Driven Temporal Filters Obtained Via Principal Component Analysis(PCA) and Linear Discriminant Analysis(LDA) in Speech Recognition,” *Eurospeech*, 2001.

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

[Katz 1987] S. M. Katz. Estimation of Probabilities from Sparse Data for the Language Model Component of A Speech Recognizer. IEEE Trans. On Acoustics,

Speech and Signal Processing, Volume 35 (3), pages 400-401, March 1987.

[Kohler *et al.* 1996] J. Kohler,"Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds", ICSLP'96, pp. 2195- 2198, Philadelphia, PA, USA, October 1996.

[Kumar 1997] N. Kumar, "Investigation of Silicon-Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition," Ph.D. thesis, John Hopkins University, Baltimore, 1997.

[Kumar and Andreou 1998] N. Kumar and A. G. Andreou, "Heteroscedastic Discriminant Analysis and Reduced Rank HMMs for Improved Speech Recognition," *Speech Communication*, vol.26 no.4, pp.283-297, Dec. 1998.

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

[Leggetter & Woodland 1995] C. J. Leggetter, P. C. Woodland (1995). "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer Speech and Language*, Vol. 9, pp.171-185, 1995.

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

[Le *et al.* 2006] Viet Bac Le, Laurent Besacier, Tanja Schultz, "Acoustic Phonetic Unit Similarities for Context Dependent Acoustic Model Portability," ICASSP, 2006.

[Lee 1989] Lee K-F. "Automatic Speech Recognition:The Development of the SPHINX System". Kluwer Academic Publishers, Boston. 1989

[Leggetter *et al.* 1995] C. Leggetter, P. Woodland,"Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density HMMs," Computer Speech and Language, Vol. 9, pp. 171-185, 1995.

[Matsoukas *et al.* 2002] Spyros Matsoukas, Thomas Colthurst, Owen Kimball,Alex Solomonoff, Fred Richardson, Carl Quillen, Herbert Gish, Pierre Dognin," THE 2001 BYBLOS ENGLISH LARGE VOCABULARY CONVERSATIONAL SPEECH RECOGNITION SYSTEM," in Proc. ICASSP,2002

[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.

[Molau *et al.* 2001] Sirko Molau,Michael Pitz,Hermann Ney. “Histogram Based Normalization in the Acoustic Feature Space”. ICSLP 2001

[Ney *et al.* 1999] Ney, H., Ortmanns, S., “Dynamic Programming Search for Continuous Speech Recognition,” *IEEE Signal Processing Magazine*, vol. 16, no. 5, 1999, pp. 64-83.

[Nguyen *et al.* 2005] Long Nguyen, Bing Xiang, Mohamed Afify, Sherif Abdou, Spyros Matsoukas, Richard Schwartz, and John Makhoul,” The BBN RT04 English Broadcast News Transcription System,” in Proc. INTERSPEECH, 2005

[NIST 2007] National Institute of Standards and Technology， http://www.nist.gov/speech/participants/index.htm

[Odell 1995] Julian James Odell, The use of context in large vocabulary speech recognition,” Ph.D. dissertation, Univ. Cambridge, Cambridge, U.K., 1995.

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

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

[Povey *et al.* 2005] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau and G. Zweig (2005). “fMPE: Discriminatively Trained Features for Speech Recognition,” in *Proc. ICASSP’05*.

[Prasad *et al.* 2005] R. Prasad, S. Matsoukas, C.-L. Kao, J. Ma, D.-X. Xu, T. Colthurst, O. Kimball, R. Schwartz ,” The 2004 BBN/LIMSI 20xRT English Conversational Telephone Speech Recognition System,” in Proc. INTERSPEECH,2005

[Rabiner *et al*. 1989] Rabiner, L. R., A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 1989

[Saon et al. 2000] G. Saon, M. Padmanabhan, R. Gopinath and S. Chen, “Maximum

Likelihood Discriminant Feature Spaces," ICASSP, 2000.

[Schultz *et al*. 2001] T. Schultz, A. Waibel, "Language independent and language adaptive acoustic modeling for speech recognition", Speech Communication, vol. 35, no. 1-2, pp. 31-51, August 2001.

[Soltau *et al.* 2005] Hagen Soltau, Brian Kingsbury, Lidia Mangu, Daniel Povey, George Saon and Geoffrey Zweig," THE IBM 2004 CONVERSATIONAL TELEPHONY SYSTEM FOR RICH TRANSCRIPTION," in Proc. ICASSP,2005

[Sooful *et al.* 2001] J. J. Sooful, E. C. Botha, "An acoustic distance measure for automatic cross-language phoneme mapping", PRASA'01, pp. 99-102, South Africa, November 2001.

[SRILM] A. Stolcke. SRI Language Modeling Toolkit. version 1.5.2, http://www.speech.sri.com/projects/srilm/ .

[Uebel *et al.* 2001] L.F. Uebel and P.C. Woodland. "Speaker Adaptation Using Lattice-based MLLR." In Proc. ISCA ITRW on Adaptation Methods in Speech Recognition, 2001.

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

[VOA corpus] The Voice of America, VOA：http://www.voanews.com/

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

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

[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.

[Witten *et al.* 1991] Witten, Ian H. and Timothy C. Bell. 1991. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. IEEE Transactions on Information Theory, 37(4):1085~1094, July.

[Young *et al.* 1994] J Young, JJ Odell, PC Woodland,"Tree-Based State Tying for High Accuracy Acoustic Modelling",Proceedings of the workshop on Human Language Technology, 1994

[Young *et al.* 2006] Steve Young, Gunnar Evermann, Mark Gales, Tomas Hain, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, Phil Woodland, The HTK Book (for HTK Version 3.4)

[張志豪 2005] 張志豪,”強健性和鑑別力語音特徵擷取技術於大詞彙連續語音辨識之研究,” 國立台灣師範大學資訊工程所碩士論文, 2005.