第2章 實驗架構

在本章中將先介紹台師大之大詞彙連續語音辨識系統。接著說明本論文所使用的聲學模型訓練語料、語言模型訓練語料、調適語料及語音測試語料。

2.1 台師大之大詞彙連續語音辨識系統

以下將分別介紹台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[Chen *et al.* 2004a]。

2.1.1 前端處理與聲學模型

本系統語音特徵抽取使用梅爾倒頻譜係數(MFCC)或是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]兩種不同語音特徵參數。在本論文中，我們主要使用異質性線性鑑別分析(HLDA)配合最大相似度線性轉換(MLLT)做為語音特徵參數，並使用倒頻譜正規化法(CN)加強語音特徵。聲學模型部分，我們分別為聲母及韻母建立INITIAL與FINAL模型。基本的INITIAL模型為 22 種，FINAL模型為 38 種，因為聲母會被右邊相連韻母影響其發音特性，所以我們將INITIAL模型細分為 112 種，即使用右相關聯模型(Right-Context-Dependent Model, RCD)，最後加上一個靜音(Silence)模型，共有 151 個聲學模型，即 151 個連續密度隱藏式馬可夫模型(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布，其中使用的高斯分布個數分別為 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)。

2.1.2 詞典建立

在中文裡約有 7,000 個單字詞，新詞可由這些單字詞合併產生，本系統根據字詞在語料中的統計特性，以自動化的方式產生新的複合詞(Compound Words)。對於語料中任意相鄰的兩個詞(w_i, w_j)，分別計算它們的前向二連(Forward Bigram)機率$P_f(w_j \mid w_i)$與後向二連(Backward bigram)機率$P_b(w_i \mid w_j)$，再以前後向二連的機率幾何平均$FB(w_i, w_j) = \sqrt{P_f(w_j \mid w_i) P_b(w_i \mid w_j)}$，作為詞$w_i$與詞$w_j$是否合併的依據。文字語料先經由一個含有一至四字詞約六萬六千個詞的原始詞典斷詞，再利用上述的計算方法，經過數次的迭代以及不同的門檻值(Thresholds)設定，產生約五千餘個二至十字詞的複合詞。最後將這五千餘個新詞加入原始詞典中，得到一個含有約七萬兩千個詞的新詞典。

2.1.3 詞彙樹複製搜尋

本系統的大詞彙連續語音辨識方法是採用由左至右(Left-to-right)、音框同步(Frame-synchronous)的詞彙樹複製搜尋方式[Aubert 2002]。在詞彙樹中每個分枝(Arc)代表一個INITIAL或FINAL的隱藏式馬可夫模型，由根節點(Root)到任一個葉節點(Leaf)的路徑代表一個詞或一些發音相同的詞，路徑上的分枝就是代表這個詞或這些詞會使用到的隱藏式馬可夫模型。具體來說，所採用的詞彙樹複製搜尋演算法，在搜尋時每個音框會同時存在數棵詞彙樹複製(Tree Copies)，每個詞彙樹代表不同的語言模型歷史詞序列(Language Model History)。實際上，搜尋時產生的不完全路徑(Partial Path)如果擁有相同的歷史詞序列會被歸類在同一棵詞彙樹複製裡，以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。在每個音框中，若有不完全路徑已到達葉節點時，代表一個完整詞已可被產生；同時，不同棵詞彙樹複製間已抵達葉節點的不完全路徑，若具有相同

的語言模型歷史詞序列，則會進行再結合(Recombination)，保留最大分數者，並以它們的歷史詞序列為標註，產生一棵新的詞彙樹複製，或加入到一棵已存在且具有相同歷史詞序列的詞彙樹複製中。值得注意的是，在實作時並不需要真的建立如此多的詞彙樹複製，僅需建立一棵詞彙樹作為搜尋時路徑展開參考之用即可，並分別記錄搜尋時存活下來的隱藏式馬可夫模型狀態節點(也就是不完全路徑目前拜訪到的節點)的相關資訊。另一方面，由於存活的隱藏式馬可夫模型狀態節點可能會隨音框數呈指數倍增加，因此採用光束搜尋(Beam Search)技術，適當地裁減分數較低的狀態節點或不完全路徑。在執行裁減動作時會同時考量每一個詞彙樹複製之內部節點(Internal Node)下涵蓋的可能拜訪葉節點所代表之所有詞對應的語言模型機率，並以其中最大者當作每一個詞彙樹複製內部狀態節點的語言模型前看分數(Language Model Look-ahead Score)[Aubert 2002]，再加上內部狀態節點本身搜尋時所累積的解碼分數(Decoding Score)及聲學模型前看分數[Chen *et al.* 2004a]來當成裁減比較的依據。本系統採用單連語言模型前看(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之最大一連語言模型機率，作為該內部狀態節點的語言模型前看分數。此外，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(這些葉節點本身代表著可能的候選詞)，諸如它們的語言模型歷史詞序列、候選詞所對應的開始與結束的音框以及搜尋時聲學模型解碼的分數，然後再依此資訊建立詞圖(Word Graph)，並在詞圖上使用更高階的語言模型，重新進行一次詞圖動態規劃搜尋(Word Graph Rescoring)[Ortmanns *et al.* 1997]，找出最佳的辨識詞序列。在本系統中，詞彙樹複製搜尋階段是使用二連詞語言模型，而在詞圖搜尋階段是使用三連詞語言模型。

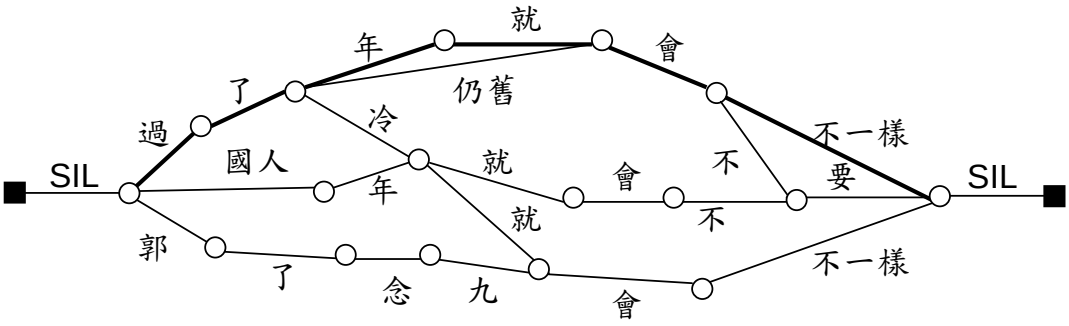

圖 2-1 詞圖範例

2.1.4 詞圖搜尋

經過詞彙樹搜尋後，可以產生詞圖(Word Graph or Lattice)，詞圖範例如 圖 2-1 所示。每一個分支代表經過裁減所保留的詞段，且每一個詞段會記錄其聲學分數。接著我們針對每一個詞段進行維特比搜尋，記錄與其相連(結束時間與目前詞段開始時間相同)且最可能(維特比分數最高)的詞段。詞圖所保留下來的詞段，在聲學上大多是混淆的，所以需要透過語言模型的輔助。由於詞圖已經簡化，搜尋時，可以使用較複雜的語言模型而不至於花費太多時間，例如三連詞模型。而要使用更長的歷史詞序列資訊時，可以於某詞段的開始時間後徹(Backtrace)最有可能的詞序列，再根據不同的模型進一步利用。而完成整個詞圖的重計分之後，同樣從語音結尾的詞段後徹最可能詞序列當作最後辨識結果。

2.2 實驗語料

聲學模型訓練語料來自公視廣播新聞語料(Mandarin Across Taiwan - Broadcast News, MATBN)。公視新聞語料[Wang *et al.* 2005]是由中央研究院資訊所口語小組[SLG]與公共電視台[PTS]，在 2001 年到 2003 年之間合作錄製的；其中 2001 年包含了 30 小時，2002 年包含了 146 小時，2003 年包含了 24 小時，總共是 200 小時的新聞語料。這 200 小時的語料包含了 28,913 個句子(Sub-term)，每個句子都經過人工轉寫且標註了額外的資訊，如各種背景雜訊、語者性別、停頓、呼吸、語助詞等。

公視新聞語料包含了內場以及外場兩個部分，內場為主播(Studio Anchors)語料，外場又可分成採訪記者(Field Reporters)與受訪者(Interviewees)語料。但是由於內場主播語料大部分為同一個主播所錄製的因素，為了避免語料的偏差性讓實驗偏向語者相依(Speaker Dependent)，故不採用內場主播語料；又發現外場的受訪者語料，包含了太多的語助詞，故論文的實驗初步採用外場記者語料。我們從外場記者語料中挑選了約 28.5 個小時的語料，其中 25.5 個小時(包含了 5,774 個句子)拿來當作聲學模型訓練語料，1.5 個小時(包含了 292 個句子)拿來當作發展語料，1.5 個小時(包含了 307 個句子)拿來當作評估語料。聲學模型訓練語料是從 2001 年及 2002 年挑選出來的，其中男女生的語料各半，且都為不包含語助詞(Particles)的語料。發展與測試語料是由中央研究院所選定的評估集中，挑選出採訪記者語料並過濾掉含有語助詞的句子而來。發展與測試語料涵蓋的時間是2003年，分別來自 0128、0129、0211、0307、0403 與 0124、0127、0207、0305、0306 兩組不同的日期。語音語料的相關統計資訊如 表 2-1 所示。

語言模型的背景訓練文字語料蒐集自中央通訊社(Central News Agency, CNA)2001 至 2002 年，類型為報導(Story)的新聞，大約包含一億五千萬(150M)個中文字，經斷詞後約八千萬詞[CNA News]。調適語料則分為兩部分，同時期(Contemporary)語料與同領域(In-Domain)語料。因為測試語料是來自 2003 年的廣播新聞，所以我們收集 2003 年 1 月到 4 月的東森新聞當作同時期語料[ETtoday

表 2-1 外場記者訓練、發展及評估語音語料之統計資訊

	訓練集語料	發展集語料	評估集語料
語音長度(男)	766.68 分鐘	21.69 分鐘	23.38 分鐘
語音長度(女)	766.78 分鐘	65.23 分鐘	60.86 分鐘
總字數	477,098 字	26,219 字	26,767 字
總詞數	289,513 詞	16,106 詞	16,494 詞

News]。經過去除非中文字元及斷詞後，最後保留約四萬篇新聞，包含大約兩千萬(20M)個中文字，斷詞後約一千兩百萬(12M)個詞，以SetET表示。我們亦從公視廣播新聞語料2001與2002年的人工轉寫文件選擇出一部分當作同領域調適語料。我們篩選出約三千五百篇報導，約兩百萬(2M)個中文字，斷詞後約一百萬(1M)個詞，以SetMAT表示。使用的文字語料詳細統計資訊如 表 2-2 所示。我們採用SRI Language Modeling Toolkit訓練辨識所需要的*N*連詞語言模型[SRILM]。

2.3 語言模型評估

本小節介紹評估語言模型品質的兩種主要方法，語言複雜度與字錯誤率。

2.3.1 語言複雜度

一般評估語言模型的方式為計算語言複雜度(Perplexity)[Manning and Schutze 1999]。使用欲評估的語言模型 M ，計算測試語料正確語句 $w_1, w_2, ..., w_n$ 的機率，再取其倒數的幾何平均即為此語言模型的複雜度：

$$PP = \sqrt[n]{\frac{1}{P_M(w_1, w_2, ..., w_n)}} \tag{2-1}$$

n 為測試語料總詞數。語言複雜度亦可視為語言模型預測詞的平均分支數(Average Branching Factor)。

表 2-2 訓練及調適文字語料之統計資訊

	CNA	SetET	SetMAT
類型	中央社新聞文字語料	東森新聞文字語料	公視廣播新聞人工轉寫語料
文件篇數	314,592	41,345	3,551
總字數	142,519,874	19,436,419	1,922,713
總詞數	85,769,244	12,132,356	1,178,076
文件平均長度(字)	約 453 字	約 470 字	約 541 字
文件平均長度(詞)	約 272 詞	約 293 詞	約 331 詞

2.3.2 字錯誤率

在語音辨識中，常用的評估的方法是計算錯誤率(Error Rate)。透過美國標準與科技組織所訂立的評估標準(U.S. NIST F.O.M. Metric)，對正確參照轉寫與辨識結果進行字串比對，藉由動態規畫(Dynamic Programming)方式及不同層次的對齊單元(Alignment Unit)，求得最佳的字串對齊(Alignment)。對齊過程中，兩字串單元可能發生替代(Substitution)、插入(Insertion)及刪除(Deletion)等錯誤情況。對齊結束後，可以計算辨識正確率(Accuracy)：

$$Acc = \frac{H - I}{N} \times 100\% \tag{2-2}$$

H 代表兩字串相同(Hit)的單元數量，I 代表辨識字串插入的單元數量，N 代表正確參照轉寫的單元數量。錯誤率則為 1-辨識正確率，在英文辨識中，以詞(Word)為對齊單元，計算詞錯誤率(Word Error Rate, WER)，而中文的詞由字(Character)組成，且有新詞定義及斷詞問題，故以字為單元，計算字錯誤率(Charater Error Rate, CER)。

2.4 基礎實驗結果

首先是僅使用背景語言模型的實驗結果。背景語言模型訓練語料收集自中央通訊社 2001 年及 2002 年，並透過Katz平滑化技術訓練三連語言模型，其字錯誤率及語言複雜度如 表 2-3 所示。我們亦分別對同時期調適語料SetET與同領域調適語料SetMAT訓練N連詞語言模型，其中分為平滑化及未平滑化模型，平滑化技術亦是採用Katz平滑化法。接著使用模型插補法(Model Interpolation)與背景N連語言模型結合：

$$\hat{P}\left(w_i \mid h_{w_i}\right) = (1-\lambda) P_{BG}\left(w_i \mid h_{w_i}\right) + \lambda P_{ADP}\left(w_i \mid h_{w_i}\right) \tag{2-3}$$

$P_{BG}\left(w_i \mid h_{w_i}\right)$是背景三連詞語言模型，$P_{ADP}\left(w_i \mid h_{w_i}\right)$是從調適語料訓練出的$N$連詞模

型，λ是可調整的比重。實驗結果如 表 2-4 與 表 2-5 所示。我們觀察到，*N*連詞語言模型經過平滑化後，的確能改善資料稀疏問題，且越高階的情況影響越大，如評估集裡，使用SetET語料，平滑化與未平滑化之三連語言模型的字錯誤率分別為 19.65%與 20.28%，相對改進為 3.1%。平滑化與未平滑化之二連語言模型字錯誤率分別為 19.89%與 20.26%，相對改進為 1.82%，與三連模型比較，二連模型改進幅度較少，單連模型部分則沒有什麼影響。在調適語料SetMAT亦有如此現象。評估集結果大抵相同，惟使用SetMAT語料及平滑化三連模型並無改善，可能是因為模型插補參數是發展集的最佳設定，於評估集未必是最佳。此外，我們亦觀察到，同領域語料SetMAT的效果會比同時期語料SetET來的好。這是因為測試語料與同領域語料皆是廣播新聞語料，用詞較接近，同時期語料SetET雖然包含其新聞內容，但是屬於文字新聞語料，文件風格較不相似。

表 2-3 背景三連詞語言模型基礎實驗結果

基準	字錯誤率(%)	語言複雜度
發展集	20.79	667.23
評估集	20.32	682.10

表 2-4 *N* 連詞模型插補法於發展集結果

字錯誤率(%)			
調適語料	單連詞	二連詞	三連詞
SetET (未平滑化)	20.32	20.26	20.28
SetET (平滑化)	20.32	19.89	19.65
SetMAT (未平滑化)	20.30	19.53	19.89
SetMAT (平滑化)	20.31	19.50	19.46
語言複雜度			
調適語料	單連詞	二連詞	三連詞
SetET (未平滑化)	606.32	573.86	581.40
SetET (平滑化)	606.31	540.10	507.30
SetMAT (未平滑化)	575.37	440.17	487.00
SetMAT (平滑化)	574.31	439.16	426.59

表 2-5 *N* 連詞模型插補法於評估集結果

字錯誤率(%)			
調適語料	單連詞	二連詞	三連詞
SetET (未平滑化)	20.09	19.82	19.93
SetET (平滑化)	20.09	19.65	19.29
SetMAT (未平滑化)	19.85	19.23	19.48
SetMAT (平滑化)	19.87	19.23	19.23
語言複雜度			
調適語料	單連詞	二連詞	三連詞
SetET (未平滑化)	626.33	608.03	614.75
SetET (平滑化)	626.33	576.04	544.04
SetMAT (未平滑化)	587.92	449.10	493.72
SetMAT (平滑化)	586.32	447.55	434.46