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。

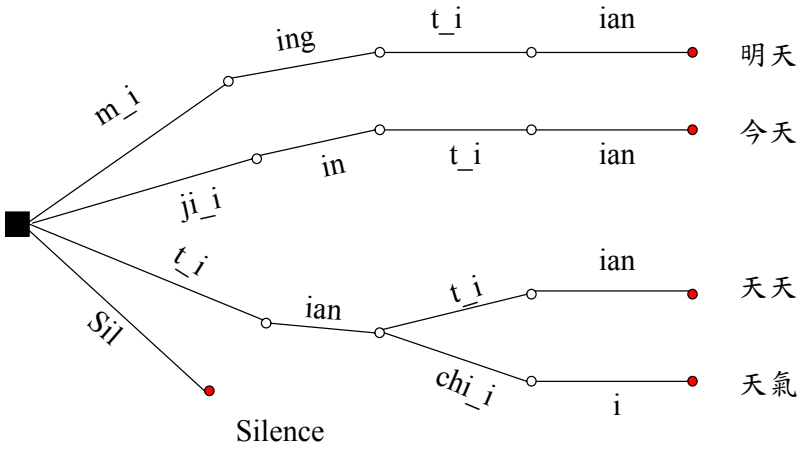

圖 3-1 詞彙樹範例

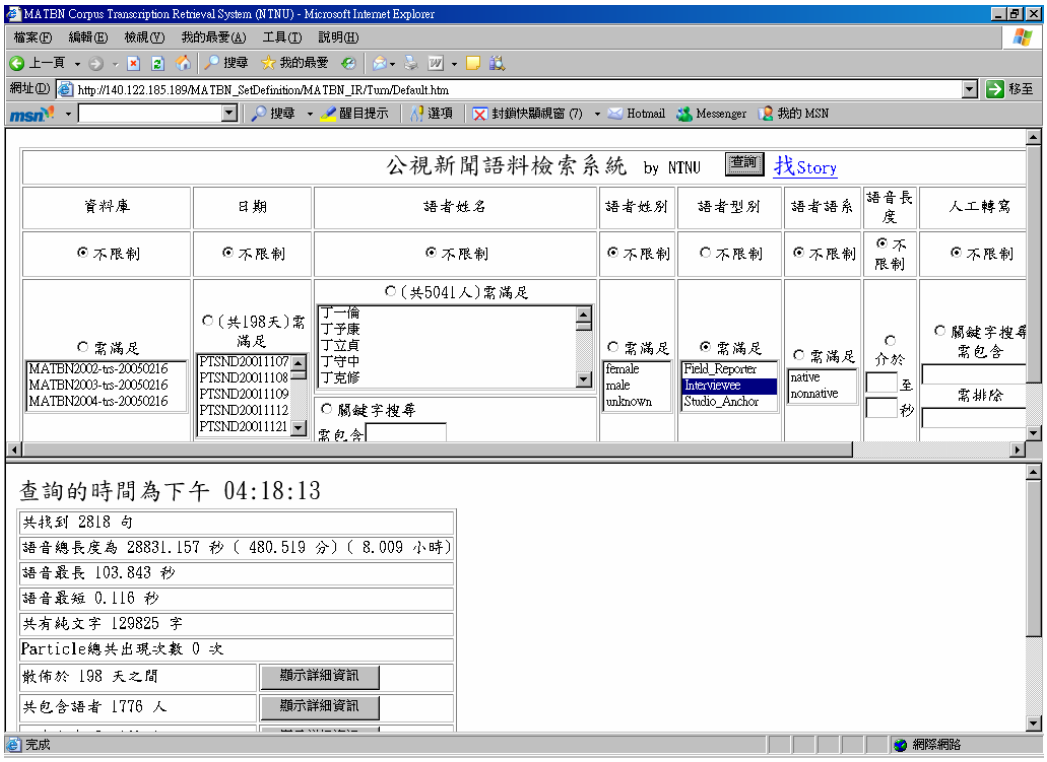

圖 3-2 臺師大資工所公視新聞語料檢索系統，檢索語句(Sub-term)的統計資訊

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

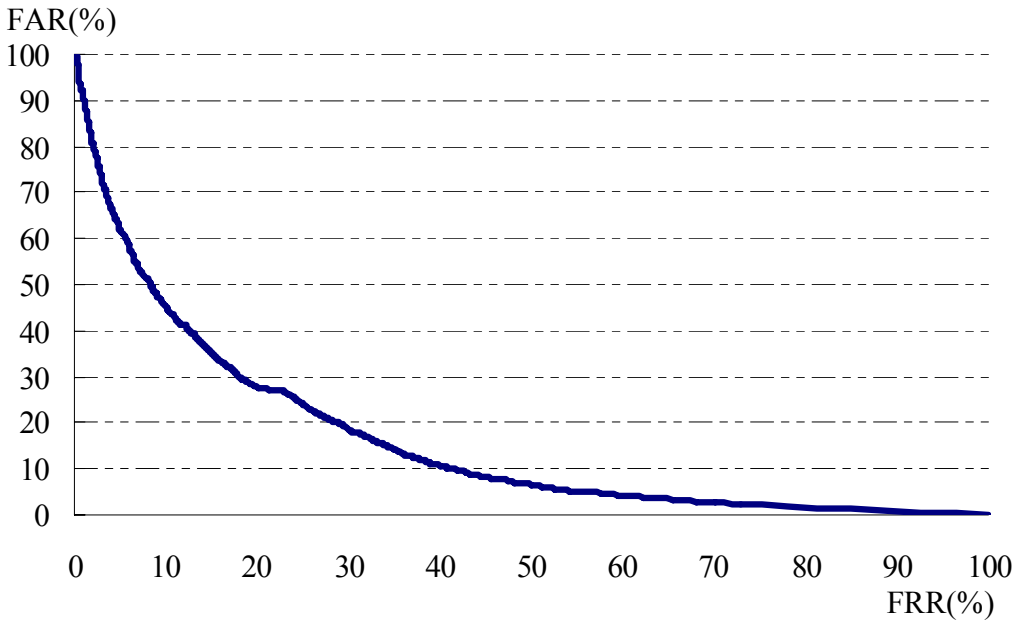

圖 3-3 偵測錯誤交易曲線圖範例

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。