國 立 交 通 大 學

電信工程學系

碩 士 論 文

國語廣播新聞語音基本辨認系統之建立

An Implementation of Mandarin Broadcast News Speech Recognition Baseline System

研 究 生：張隆勳

指導教授：陳信宏 博士

中華民國九十四年七月

國語廣播新聞語音基本辨認系統之建立

An Implementation of Mandarin Broadcast News Speech Recognition Baseline System

研 究 生：張隆勳　　Student : Lung-Hsun Chang

指導教授：陳信宏 博士　　Advisor : Dr. Sin-Horng Chen

國立交通大學

電信工程學系

碩士論文

A Thesis

Department of Communication Engineering

College of Electrical Engineering and computer Science

National Chiao Tung University

In Partial Fulfillment of Requirements

for the Degree of

Master of Science

in Electrical Engineering

July 2005

Hsinchu,Taiwan,Republic of China

中華民國九十四年七月

國語廣播新聞語音基本辨認系統之建立

研究生：張隆勳　　指導教授：陳信宏 博士

國立交通大學電信工程學系碩士班

中文摘要

廣播新聞節目在生活中相當普遍，而語音辨識技術在這方面有許多相關應用值得發展。在本論文中，我們首先利用廣播新聞語料庫（MATBN）依據語者環境個別建立聲學模型，並利用其中自發性語音現象，訓練出相對應的聲學模型，作為基本辨識系統。接下來，建立出語言模型並經過調適，使其特性更符合廣播新聞環境中使用，另外，為了將標點符號的訊息也加入辨識系統中，我們在建立語言模型的過程中額外保留了標點符號之相關轉移機率，並且建立出音節間靜音長度模型配合使用提升辨識率。最後，還將更上層的詞類資訊也加入使用，以求辨識器效能的進一步提升。將最終的音節辨識率與基本辨識系統相比，內場主播、外場記者與外場受訪者的辨識率各約提昇 16.04%、20.52%與 22.79%，此外，我們還將觀察標點符號自動標識所得到的結果。

An Implementation of Mandarin Broadcast News Speech Recognition Baseline System

Student： Lung-Hsun Chang Advisor： Dr. Sin-Horng Chen

Department of Communication Engineering
National Chiao Tung University

Abstract

Broadcast news (BN) is very general in our daily life, many researches were investigated into combining speech recognition technique. In the thesis, the Mandarin broadcast news corpus, MATBN, was adopted to train the acoustic model at first. And the models for spontaneous phenomena were built by making use of this corpus, too. Then, the language model (LM) for BN recognition system was built and adapted. Besides, in order to exploit the information of punctuation mark, the related transition probabilities were retained while building LM, and the pause duration models were trained to improve the performance. At last, we tried to use the information on part of speech at the same time. Finally, the syllable recognition rate was increased about 16.04% for anchor, 20.52% for reporter and 22.79% for interviewee. In addition to the recognition rate, the performance of automatic punctuation will be observed and analyzed.

致謝

研究所兩年的時間一轉眼就過去了，尤其是最後這半年真是快的令人感到有點措手不及。

首先，真的很感謝陳信宏老師和王逸如老師，除了在與研究相關的專業知識外，由於他們的細心指導，讓我對於做事情的方法與態度都有所轉變，相信這對於我的未來，不論是在工作上或者是生活中，一定有相當大的幫助，關於您們的教誨我會永遠銘記在心。

至於我們的語音實驗室，真的是讓人印象深刻，之前的學長：俊良、性獸、阿樹、阿德、立彥、嘉俊、小 z、祺翰和智合，每位學長都對我們非常的好，尤其是俊良學長，我的論文得以達到目前的程度真的是因為站在俊良學長這位巨人的肩膀上啊！而且也非常謝謝性獸學長給予我的諸多指導，跟你的討論中實在受益良多。而我們這一屆的好戰友們：順哥、金翰、Lubo、希群、佩穎和智顯，雖然研究過程中大家都經歷了不同的酸甜苦辣，在此先恭喜大家都順利的畢業了！由於你們的陪伴，讓我的研究所生涯留下了許多美好的回憶，其中，順哥非常有耐心的所給予我許多技術上的指導，真是令我感激不盡、佩服不已，若是沒有你的協助，我現在應該還在哭著寫程式無法畢業吧.....。而開朗活潑的學弟們也相當有禮貌，實驗室由於你們的加入，讓苦悶的研究生活中充滿了歡樂的氣息，希望能夠一直保持下去，也先預祝各位在一年後也都可以順利畢業！

最後，家人的鼓勵、還有女友的陪伴，是我在低潮時期的寄託，對你們的感謝之意真的是無法用言語來形容，因為你們的支持，才使得我人生中的這一階段能夠順利完成，因為你們的存在，才使得我的人生得以圓滿。

目錄

表目錄

圖目錄

第一章　緒論

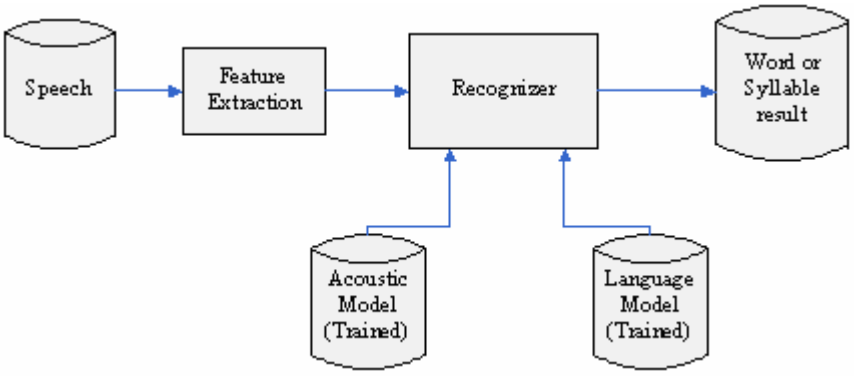
1.1　研究動機

現代科技中的一項重要發展便是用電腦來處理語言問題，而最終的目標，就是利用語音辨識技術建立人與機器之間的溝通橋樑 [1]。由於訊號處理、演算法和電腦硬體設備的進步，語音辨識技術在過去的十到二十年間確實在許多方面均有長足的進展，例如資料驅使（Data-driven）方法、聲學模型和語言模型建立方式，以及基於動態編輯程序（Dynamic Programming-based）之搜尋方法等 [2]。

近年來，廣播新聞非常普及，若能夠在語音辨識技術的配合下，建立出一套良好的廣播新聞資訊檢索系統（Broadcast News Information Retrieval System），那麼將得以對大量的新聞資訊做高效率的擷取，因此，一個高效能的語音辨識系統將是建立新聞資訊檢索系統的第一要素。基於以上理由，本論文將以中文廣播新聞語料為對象，針對此種環境下的語音辨識做深入的研究與探討。

1.2　研究方向

一個基本的語音辨識系統方塊圖如下：



圖一-1　基本辨識系統方塊圖

如圖中所示，一個基本辨識系統包含：求取語音參數、聲學模型（Acoustic Model）的訓練、語言模型（Language Model）的訓練還有進行辨識的方法，根據不同的使用領域，必須對辨識系統做適當的調整。中研院在 2002 至 2004 年的計畫裡錄製了一套廣播新聞資料庫 MATBN（MAT Broadcast News），提供學術研究使用，因此我們將針對中文廣播新聞的特性建立語音辨識系統 [3]。

本論文的研究重點，除了基本的聲學模型和語言模型的建立之外，將會在語言模型上做變化，還會加入標點符號、詞類連接規則以及句中 pause 帶有的訊息，並觀察、比較各種不同條件下的辨識情形。

1.3 章節概要

本篇論文章節內容區分如下：

第一章 緒論：介紹研究動機、研究方向及章節概要。

第二章 廣播新聞資料庫：說明廣播新聞資料庫 MATBN 的語料特性，以及介紹標記軟體 Transcriber 與 XML-Parser，並分析語者環境的差異。

第三章 基本辨識系統：對一個廣播新聞類型的基本辨識系統之建立流程與參數設定做個說明，並檢視其辨識效能。

第四章 加入語言模型之語音辨認器：詳細說明詞典的選擇方法及語言模型的訓練過程，並將語言模型加入基本辨識系統，以提升與辨識效能；另外，還將破音字列入考慮，並且評估加入後改善的結果。

第五章 將標點符號、音節間靜音長度與詞類模型加入口語語音辨識器：評估標點符號、pause duration 以及詞類連接規則對辨識器的影響程度，並對將其中資訊加入辨識系統之構想、流程與方法做個說明，將辨識結果與之前的系統做比較，最後，觀察並分析標點符號自動標示的效能。

第六章 結論與未來發展。

第二章　廣播新聞資料庫

本章首先介紹研究中所使用的一套廣播新聞資料庫，MATBN（Mandarin Speech Data Across Taiwan Broadcast News），並說明廣播新聞語料中一些較常見的特性。由於廣播新聞語料中的語音會隨著內容、思考及情緒等因素變化，所以廣播新聞語料的特性不同於一般的 read speech 而比較傾向自然語音（Spontaneous Speech），又研究中所進行的是連續語音辨識，因此廣播新聞的辨識會比 read speech 或 isolated word 辨識複雜許多。

2.1　MATBN

進行語音辨識相關研究的過程中，首要、也是最困難的條件就是有一套適合的資料庫。在國語資料庫的蒐集上曾經有個成功的經驗，MAT（Mandarin speech data Across Taiwan），過程中所蒐集完成的是一個經過設計的 read speech 資料庫，國內在此資料庫的蒐集錄製完成後，一些關於語音辨識的相關研究便得以方便的進行。

但是，不同類型的語音特性所適用的研究領域也不相同，而且國內目前並沒有關於中文廣播新聞的資料庫存在，而 MAT 中也沒有包含廣播新聞語料的部份，因此，中研院王新民教授又進行了另一個計畫，MATBN（MAT Broadcast News） [4]。這項計畫的目的是錄製一套中文廣播新聞資料庫，其中包括廣播新聞的節目聲音內容，以及針對其聲音所標記（Transcribe）出的文字內容與聲音現象，在有了這套廣播新聞資料庫後，我們便可以進行中文廣播新聞辨識系統的訓練、測試等相關研究。

MATBN 計畫中所錄製的是「公視新聞深度報導」和「公視晚間新聞」的節目內容，每次節目進行長度一個小時，錄製與處理標記共分三年進行，從 2001 年 11 月到 2004 年 7 月，錄製結果如下表所示：

表二-1　MATBN 錄製結果

錄製時間	錄製資料量
第一年（2001～2002）	40 小時
第二年（2002～2003）	80 小時
第三年（2003～2004）	78 小時
總計	198 小時

因為廣播新聞資料庫的蒐集，不同於 read speech 事先有文字的存在，所以在節目錄製完成後，會需要進行標記的動作。MATBN 資料庫的標記工作是由中研院所聘請的兩位標記員進行，由於標記結果的正確性是未來相關研究的基礎，因此每小時的節目內容都先經由一名標記員處理，完成後再由另一名來檢查，並且每週定期開會針對問題進行檢討，以確保標記的品質。

到目前為止三年的資料均已經錄製、標記完成，我們（交通大學語音實驗室）也已將第一年與第二年的部份做好了適當的處理，至於第三年的部份，則因為剛取得不久，處理過程尚未完成，所以目前的辨識系統只利用了第一、二年合計 120 小時的資料，未來當第三年的部份準備好後也將加入系統中使用。

2.1.1　聲音特性

節目中聲音的錄製過程是，直接在電視台利用 DAT（Digital Audio Tape）以 44.1 KHz 的取樣率和 16 bits 的精確度錄製，然後再做 down sampling 的動作，將取樣率降到 16 KHz，並將音檔格式轉換成可由電腦讀取的 WAV（Microsoft Windows Wave）檔。

- Sampling Rate：16 KHz
- Resolution：16 bits

- Channel：Mono
- Format：WAV

2.1.2 資料特性

由於廣播新聞不同於一般 read speech，其處理流程是先有語音後才將它轉成文字，因此必須使用一套適合用來標記廣播新聞語料的軟體，LDC（Linguistic Data Consortium）的 Transcriber [5]，標記員可以在它的輔助下，清楚地將聽到的聲音中所包含的各類資訊標記出來。另外，又因為廣播新聞節目中，並非全部時間都有語音存在，研究中這些部分均不使用，依據標記結果所得到的 MATBN 第一、二年資料統計如下所示：

- 總時間：120 小時
- 有語音時間：86.5 小時
- 總文字數：約 140 萬字

2.2 LDC Transcriber 與 XML-Parser

一般 read speech 資料庫的錄製過程，都是有事先準備好並經過設計的文字之後，再在語者唸出文字內容的同時進行音檔錄製工作，所以錄製的聲音通常發音比較標準且具有一致性；但廣播新聞資料庫的錄製過程中，則是先錄好了並非依據文字稿所唸出的聲音之後，才再去針對錄製的聲音標記（Transcribe）出其中的文字內容，因此便需要一套合適的輔助軟體以方便進行標記的工作，而目前所採用的是一套稱作「Transcriber」的軟體。

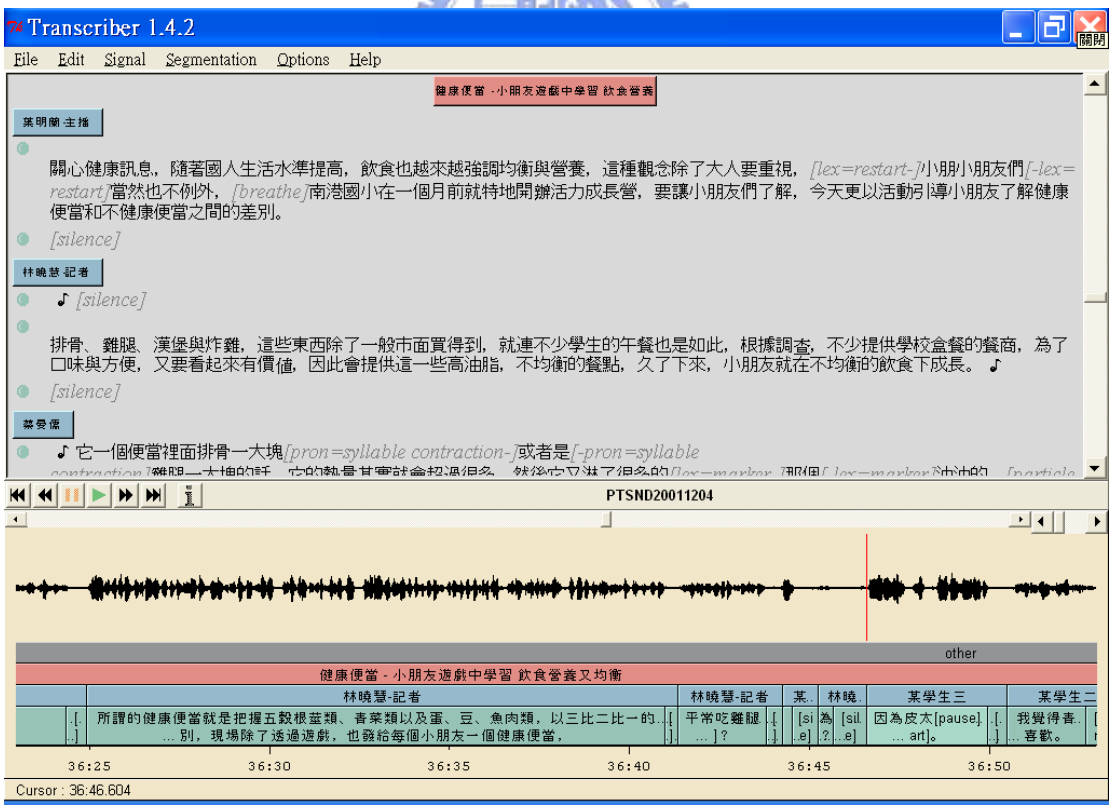
除此之外，又因為標記軟體 Transcriber 的輸出檔案儲存格式為 XML（Extensible Markup Language），所以我們並不能直接將它拿來使用，因此還必須撰寫一個用來處理 XML 格式檔案的解譯器（XML-Parser），將我們所需要的資訊從檔案中取出，並存成適當的形式，以提供給接下來進行中文廣播新聞相關

研究使用。

2.2.1 LDC Transcriber

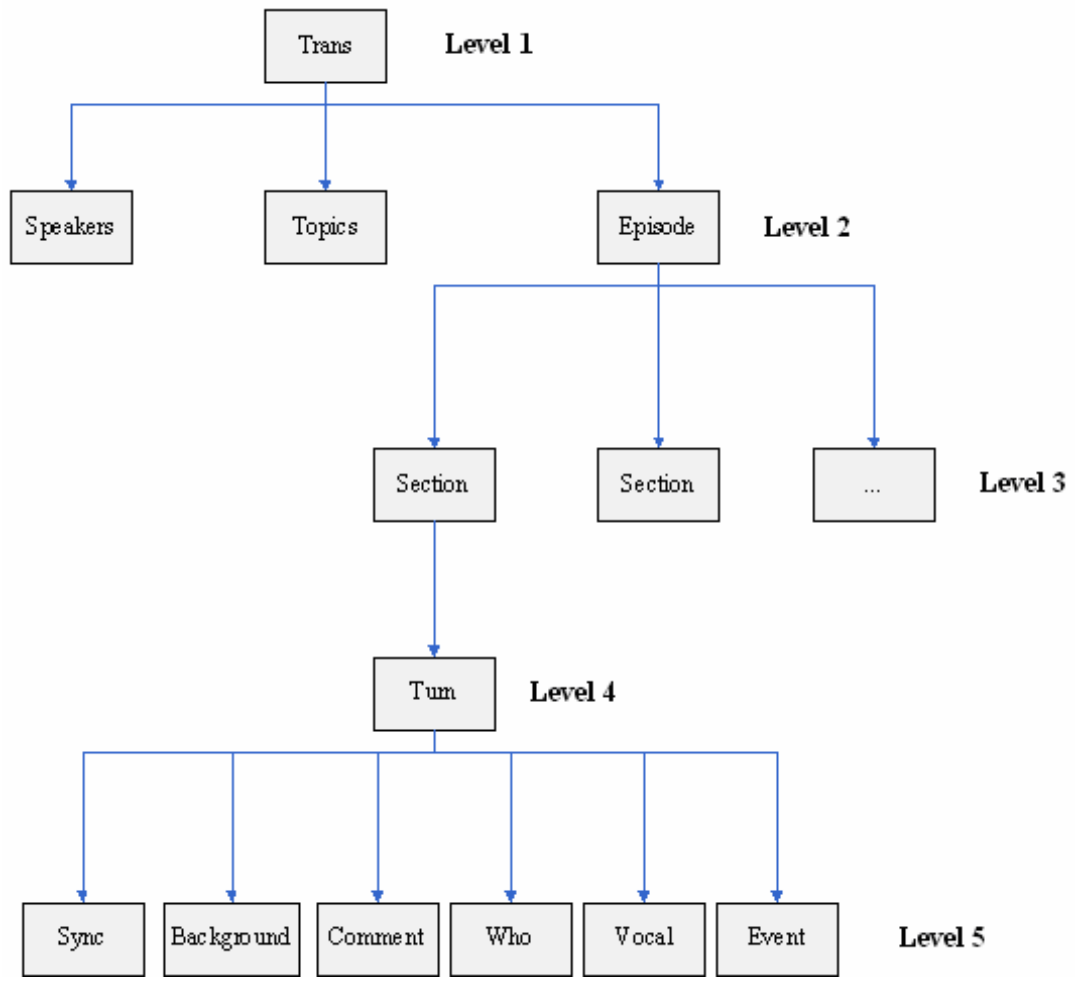
標記軟體 Transcriber，是一套可以顯示聲音波形，並同時提供標記背景聲（Background Sound）、內容主題（Topic）、語者（Speaker）與聲音內容文字等功能的一套軟體，另外，還可以記錄除了一般文字之外的常見自然語音現象，例如：呼吸聲、particle 以及笑聲、嘆氣聲、砸嘴聲等語言學現象（Paralinguistic Phenomena）。詳細標記之聲音現象內容與標記方法請參考附錄一。

這套標記軟體的編輯環境介面如下圖所示，其中任何一段時間的聲音資訊標記均使用四層狀態來記錄，由上而下分別為背景聲、內容主題、語者以及文字內容，Transcriber 之所以被選為這套資料庫的標記軟體，也正因為這四層狀態正好可以完整標記出廣播新聞的內容。



圖二-1　標記軟體 Transcriber 編輯介面

至於 Transcriber 所儲存之檔案格式乃是使用 XML 的語法，一個格式完整的 XML 檔案，需要伴隨一個 DTD（Document Type Definition）檔，DTD 檔案清楚定義了檔案的格式，因此可以避免資料傳送時格式的錯誤。我們可以藉由閱讀 DTD 之檔案內容看出 Transcriber 中定義的 XML 之階層式架構檔案格式，而利用這種階層式架構便可方便的記錄上述的聲音中各層資訊。若將此 DTD 檔圖形化表示則如下所示：



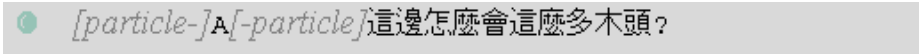
圖二-2　圖形化 DTD 檔階層架構

由上圖可以看出它的架構類似一個樹狀結構，其 root node 稱為 trans（Transcription），而一段語音都會首先標明語者、內容主題，然後真正語音的內容是存放在 episode 內，進入 episode 後，還分成好幾個 section，每個 section 又由數個 turn 所組成，在語言學關於對話的定義中，一個 turn 代表著對話中語者

的轉換，而樹狀結構中最底層便記錄著一個 Turn 中的語音文字內容、起始時間與各種聲音現象等訊息。一個簡短的 XML 原始碼及與其相對應在 Transcriber 中之顯示情形例子如下：

```
<Sync time="2774.734"/>
<Event desc="particle" type="noise" extent="begin"/>
A
<Event desc="particle" type="noise" extent="end"/>
這邊怎麼會這麼多木頭?
```

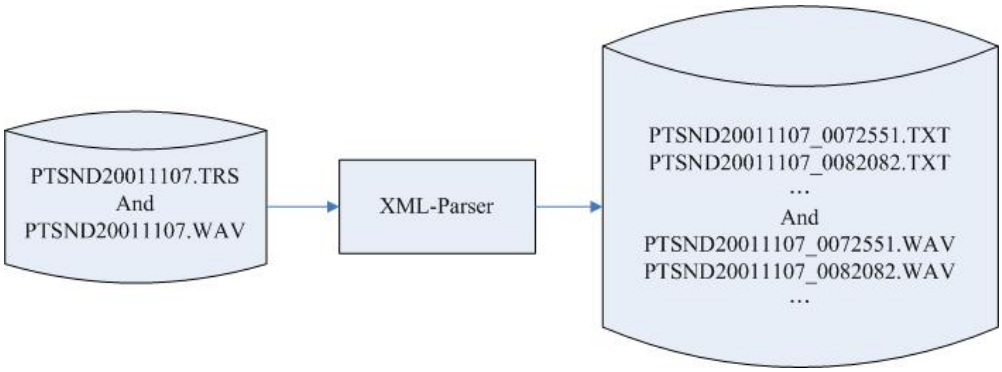
圖二-3　XML 原始碼



圖二-4　Transcriber 中顯示情形

2.2.2　XML-Parser

由於標記軟體 Transcriber 輸出檔案之儲存格式為 XML，因此無法直接拿來使用，但從上方的 XML 原始碼例子中可得知，每句話的起始、結束時間均可以從其中的含有的訊息中推得，且可進一步將需要的資訊挑出，因此我們撰寫了一個程式，XML-Parser，將一個小時節目音檔依據 XML 中的訊息，切割成以 turn 為單位的語料，並加以篩選，只留下其中包含有中文語音的部份，提供給接下來的研究中語音辨認使用。



圖二-5　XML-Parser 功能

除了切割音檔外，XML-Parser 也將各個切割後的音擋所對應到的標記內容由 XML 原始碼中挑選出來，但只留下起始與結束時間、語者 ID、語者名稱、語者性別和語音文字內容等接下來的研究中所需要的資訊，其中語音現象表示方法詳見附錄一。依據上節中的例子，其產生之文字內為：

```
2774734 2776214 spk46 柯金源-記者 M
<PARTICLE> A </PARTICLE> 這邊怎麼會這麼多木頭?
```

圖二-6　XML-Parser 處理後結果

2.3　廣播新聞語料特性

由於廣播新聞語料的錄製不像於 read speech 有準備好的文字稿，因此其聲音特性比較類似自然語音 [6]，所以語料中含有許多因為內容、思考及情緒等因素而產生之無法預期的聲音。接下來，便列出幾個自然語音語料中比較常見的一些口語現象。

- **Particle**

  自然語音中最常見的現象就是 particle，語言學上稱之為「感嘆詞」，particle 又可分為 discourse particle 與 grammatical particle 兩大類，但在此僅將 discourse particle 列入考慮加以處理，因此接下來爲了方便起見，discourse particle 均只以 particle 表示。一個常見的例子是：「為什麼這樣 NEI？」，其中「NEI」便是一個 particle。

- **Paralinguistic Phenomena**

  自然語音中另一個普遍的口語現象便是 paralinguistic phenomena，例如：笑聲、嘆氣聲、砸嘴聲等。語音辨識器中若不針對這些現象去做適當的處理，則辨識結果會受到一定程度的影響，因此我們的辨識器會將這類現象產生的問題列入考慮，並提出合適的解決方法。

- **Pronunciation Error**

不同於一般 read speech，廣播新聞中語者的說話內容並沒有經過設計，因此發音不正確的情形便可能存在，例如發音偏差（Inappropriate Pronunciation）與音節合併（Syllable Contraction）等現象，各舉一個較常見的例子，如「發生」卻唸成「ㄏㄨㄚㄕㄥ」與「這樣子」讀為「降子」，研究中也會對這類情形做特別的處理。

- **Foreign Language**

因為國際化的趨勢，即使是中文廣播新聞中也經常可聽到一些方言或外國語言穿插其中，而不同的語言之建構方法大不相同，因此對於這種非中文的語言所造成的問題也將會對其進行處理。

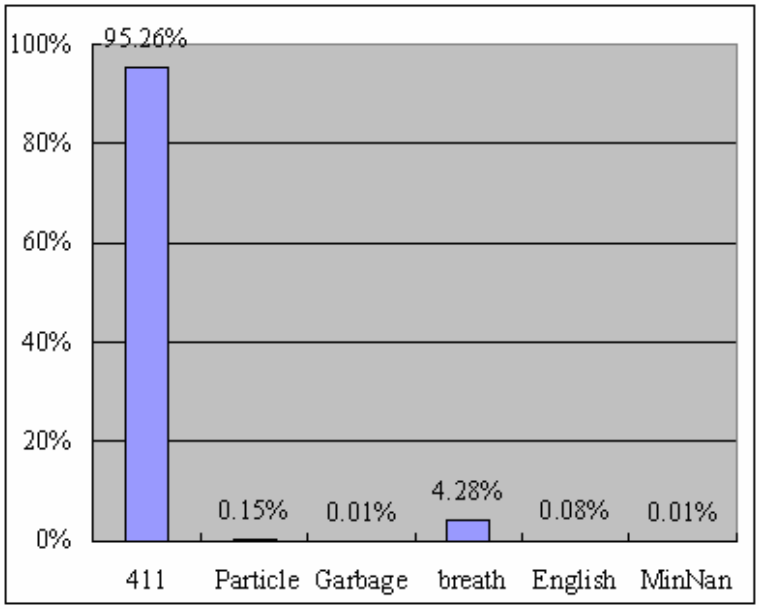
- **Background Sound**

在廣播新聞節目進行中，除了語者說話的部份之外，還常常存在著音樂、汽車聲等背景聲的存在，背景聲的存在將會干擾抽取出的語音參數之正確性而對辨識器效能有嚴重的影響。由於對於有背景聲的語音辨識相關處理並不是我們的重點，所以研究中使用的語音均是不包含背景聲的部份。

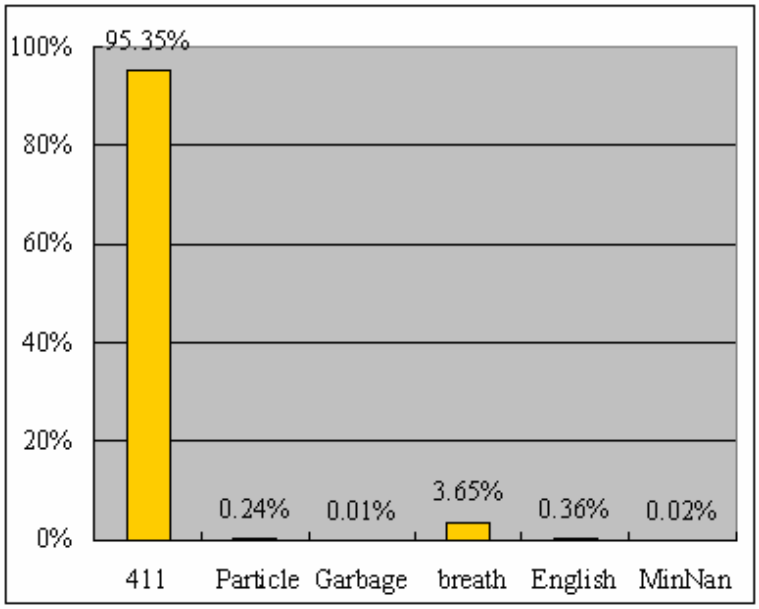
2.4 MATBN 語者環境差異

MATBN 廣播新聞語料中，依據語者環境可以區分為內場主播（Anchor）、外場記者（Reporter）和受訪者（Interviewee）三種，因為不同的語者環境，其發音特性有一定程度的差異，例如：主播與記者大多因受過發音訓練而發音咬字比較正確、清晰，說話文字內容較符合文法規律性；然而受訪者則大多為一般民眾，所以說話必較含糊、情緒化而且含有較多口語現象。

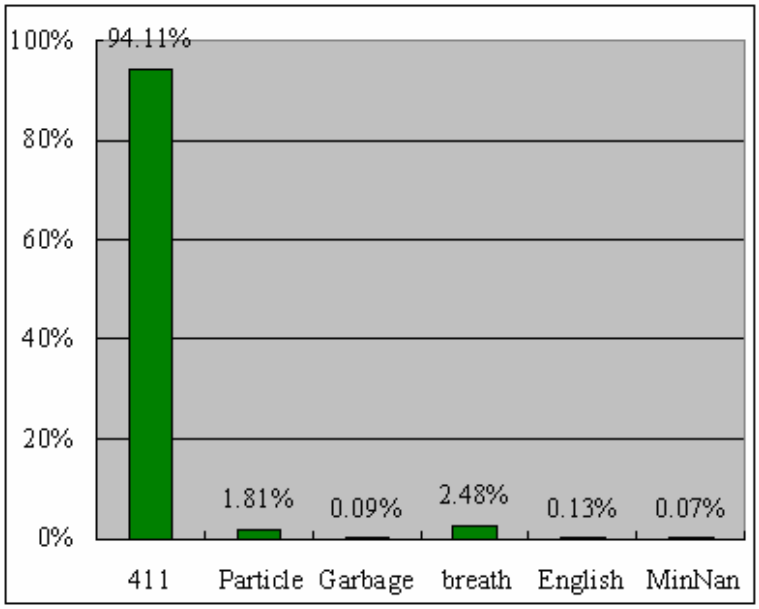
接下來，再由訓練語料（訓練語料選擇方式在接下來的章節中有詳細說明）中，分別統計國語 411 音、particle、garbage（笑聲、砸嘴聲等 paralinguistic phenomena 及一些無法處理的聲音現象）、呼吸聲以及英語、閩南語這兩種較常見的外國語言跟方言等現象，並畫出其比例統計圖，從中觀察比較三種語者環境的特性差異。



圖二-7　內場主播語料現象比例圖



圖二-8　外場記者語料現象比例圖



圖二-9　受訪者語料現象比例圖

從以上三圖中可看出，內場主播的呼吸聲佔的比例最高，這是因為呼吸聲出現的次數通常隨著一句話的長度越長而增加，而三種環境的句子平均長度由長到短為：內場主播、外場記者、受訪者，正好與此呼吸聲比例統計符合；外場記者與受訪者的語料中，非 411 音的比例均比內場主播高，而又以受訪者的 particle 和 garbage 比例最高，也和預期相吻合。

除此之外，我們再觀察發音偏差的現象，在各個語者環境的比例統計如下表所示：

表二-2　各語者環境發音偏差比例

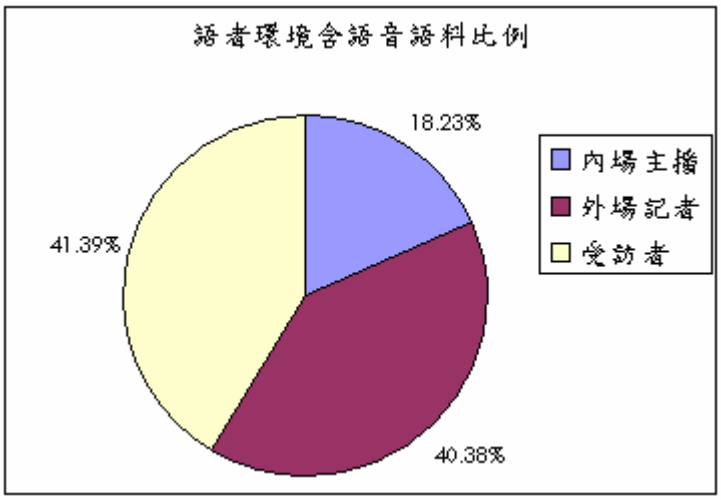
語者環境	內場主播	外場記者	受訪者
發音偏差比例	0.21%	0.36%	1.31%

因為以上的理由，接下來的研究中將針對這三種語者環境的語料分開處理，對每個環境各別進行訓練、建立辨識系統。在 MATBN 第一年與第二年的語料

中，三種語者環境個別的語者個數統計大致如下：

- 內場主播（Anchor）：4 人
- 外場記者（Reporter）：89 人
- 受訪者（Interviewee）：3429 人

另外，在資料庫中三種環境之有語音部份的語料比例如下：



圖二-10　三種語者環境含語音語料比例

由此可看出整個廣播新聞內，內場主播在節目中所佔有比例並不高，大約是總語音時間的 20%左右，而外場語音才是佔廣播新聞內大部份的時間。

第三章　基本辨識系統

近年來進行語音辨識的相關研究中，最常被採用的方法便是利用隱藏式馬可夫模型（Hidden Markov Model，HMM），這種方法是藉由機率模型來描述發音過程中的狀態（State）轉移現象與輸出結果，因為這種方法可以得到不錯的辨識效能，所以本論文中也將採用此種方法進行研究。研究中的實驗環境主要採用英國劍橋大學開發之 HTK（HMM Tool Kit），而目前使用的版本為 version 3.2.1 [7]。

3.1　語音參數設定

進行語音辨識系統之訓練、測試，首先的前處理工作就是將語音參數從輸入語音中抽取出來。因為語音訊號之短時間穩定特性（Short Term Stationary），加上考慮到人耳聽覺效應的補償作用，所以在此所使用的參數為 MFCC（Mel-Frequency Cepstral Coefficients，梅爾倒頻譜參數）。

在做語音參數求取時，有進行 DC 效應的消除、在 FFT（Fast Fourier Transform）之前先通過 Hamming window、最後進行 CMS（Cepstral Mean Subtraction）等動作，此外除了 MFCC 參數，也針對 Delta-MFCC 跟 Delta-Delta-MFCC 參數進行求取，將參數變化中所包含的訊息也提供給辨識器使用，系統參數設定如下：

- Frame Size / Frame Shift：32/10 ms
- Pre-emphasis Filter：First order with coefficient 0.97
- Filter Bank：Pass band 0 ~ 8 KHz with 24 channel
- Dimension of MFCC：12

其中因為一般情況下語者的說話數度是每秒 3 個字，但是 MATBN 語料中語者的說話速度則大約是平均每秒 5 個中文字，比一般速度來的快，所以系統設定

中對於 Delta-window 與 Delta-Delta-window 大小之選取均設定為 2。另外，第 0 階的 MFCC 參數代表著語音的能量，而能量的大小對於語音辨認並不重要，所以會將其省略，因此，針對語音訊號進行參數求取之後，得到的參數便是一維度總共 38 維的語音參數向量。

3.2 聲學模型及其訓練與測試

明確的講，我們採用的是 left-to-right HMM，雖然口腔聲道會隨時間而變，但因為語音訊號具備短時間的穩定特性，因此假設在同一音框（Frame）中，口腔狀態是相同的。此外，對於代表每個音框的信號在狀態下是否要改變之狀態轉移機率（State Transition Probability），以及代表音框與各狀態的相似程度的狀態觀測機率（State Observation Probability），均使用混合高斯模型（Mixture Gaussian Model）來表示 [8]。

另外，訓練模型、估計參數時採用的方法則利用 Baum-Welch 參數估計法，從已知狀態序列，根據轉移規則，推測出每個音框所屬的最佳口腔狀態，並重複估測聲學模型至穩定為止；至於辨識工作的進行則是使用 Viterbi search，讓每個音框均對所有模型進行估計，並找出最佳結果。

3.3 建立不分環境聲學模型

在此不分環境模型建立的資料來源，是使用 MATBN 資料庫中第一年的語料，總計 40 小時。在所使用的工具 HTK 中，HMM 初始模型的建立方法有兩種，第一種稱為 flat start，這種方法首先假設一段語音中的切割位置平均分布，並先用全部的語料訓練出一個初始的模型，將它提供給所有的次音節（Sub-syllable，在此即中文的聲母和韻母）使用，但此種作法在一段語音較長的情況下容易發生切割位置完全錯誤的情形，並且需要花較多的時間訓練出一個正確的模型；因此，在此採用第二種方法，這種方法必須先知道切割位置，然後在已知位置（Fixed

Boundary）下做 Viterbi estimation，這種方法因為每個次音節的位置已知，所以如此估計出的參數會較為準確。

3.3.1 求得切割位置（Force Alignment）

因為廣播新聞語料中，除了國語中的 411 音節（Syllable）、silence 之外，還包含一些其他的口語現象，例如：呼吸聲、particle、笑聲和咳嗽聲等語言學現象，還有外語和地方方言等，以下說明各現象切割位置求得方法：

- 411 音節、silence 與 particle

  因為一般 read speech 語料所訓練出的 HMM 模型中，會包含有國語 411 音節和 silence 的聲學模型，因此在此我們先使用 TCC300 所訓練出的 HMM 模型來對目前要用來建立初始模型的初始訓練語料進行切割，但因為受限於 TCC300 之 HMM 模型中並沒有包含廣播新聞語料中全部現象的聲學模型，所以必需對初始訓練語料做適當的選取。

  由於 read speech 語料並不會訓練出 particle 這類 spontaneous speech 語料中經常出現的聲音模型，所以在進行切割，求取 particle 的切割位置時，會先將 particle 以相近的 411 音節替代。

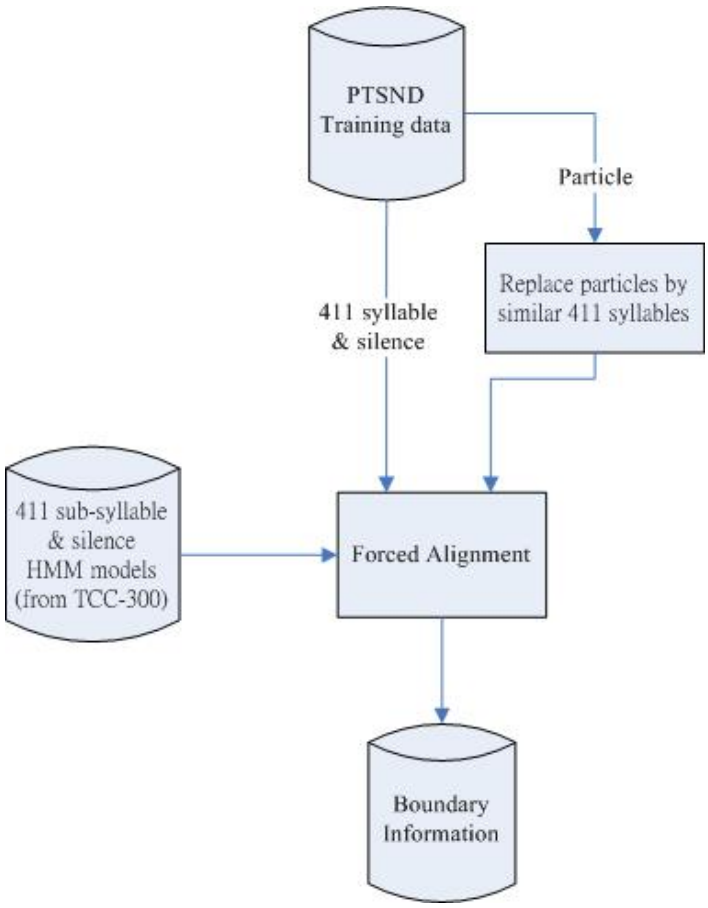
表三-1　替代 particle 之相近 411 音

Particle	相近 411 音節
A	a
AI	ai
AM	an
…	…

其於上述的考量，所選擇當作初始訓練語料的聲音，會先排除掉含有背

景聲的部份（有背景聲情況下的語音辨識並非本研究涵蓋範圍），同時留下的語料內容中只能夠有國語 411 音節、silence 與 particle 這三種資料，資料數量統計為 665 個句子（Turn），共有 35,388 個字，時間長度約 2.05 個小時。

綜合以上內容，藉由 force alignment，便可得到 411 音、silence 和 particle 的切割位置資訊（Boundary Information），切割資訊求取方塊圖如下：



圖三-1　切割位置資訊求取方塊圖

- 呼吸聲

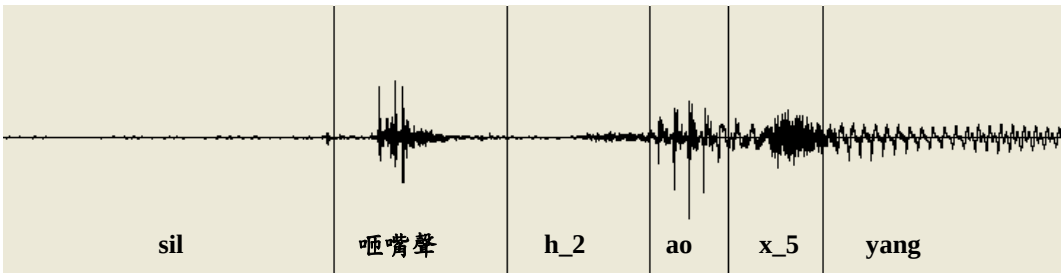
因為呼吸聲（Breath），在廣播語料庫中是一種蠻普遍的現象，但是又無法利用如之前所述，以一般 read speech 之聲學模型進行切割取得切割位置，所以這裡是先用人工的方式對部分訓練語料進行切割，以得到建立初始模型所需的呼吸聲之切割位置。

- Garbage

廣播新聞語料中存在著許多自然語音中經常出現的口語現象，而且其中有許多是不易個別處理的，例如笑聲、砸嘴聲等語言學現象和無法辨識的字詞，因為在語料中存在的資料量並不足夠訓練出個別的聲學模型，所以我們採用的方法是建立一個共同的特殊聲音模型對這些現象進行處理，並將這些現象統稱為「Garbage」，而稱呼用來取代各種特殊聲音的語音模型為「Garbage Model」 [9]。

由於 garbage model 被用來取代許多的特殊聲音，因此它包含的資料範圍相當的廣，從機率模型來看，相當於是一個變異數（Variance）很大的 Gaussian distribution，但為了避免 garbage model 與 silence model 發生混淆，在進行 garbage model 訓練時所使用的資料是所以 non-silence 的聲音語料。

因此在進行語音切割或辨識時，一般正常語音在 garbage model 與在其他語音模型相較之下，garbage model 的分數會顯得小很多而不會被選中，但在這些特殊聲音的情況下，則使用正常語音模型所得到的分數則會比較小，而會被落在變異數很大的 garbage model，因此被切割出來。依照這種處理方法，發現可以得到不錯的切割結果，下面便是一個用 garbage model 切割砸嘴聲的結果：



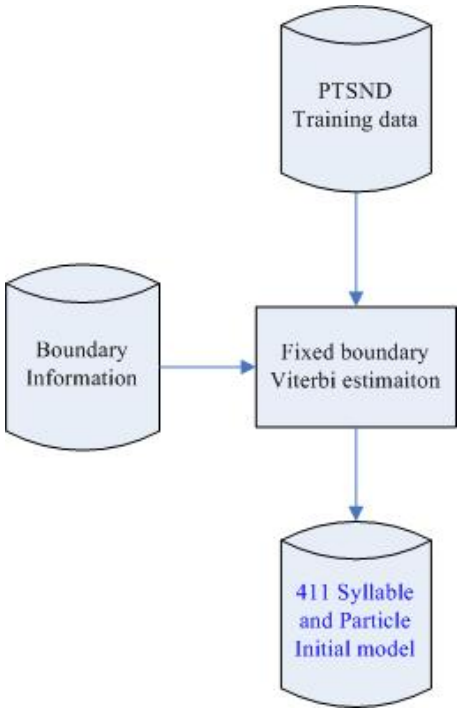
圖三-2　Garbage model 切割砸嘴聲之切割結果

除了各種的特殊聲音現象，另外像是廣播新聞節目中經常穿插在國語中出現的英語和閩南語，也是必須加以處理的對象，仿照之前的方法，也是各建立一個模糊的 garbage model 分別給英文和閩南語使用。如此一來，就也

可以利用這類模型進行切割，得到這些現象的切割位置資訊。

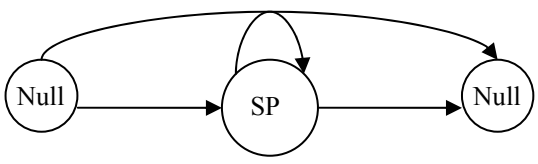
3.3.2 建立初始模型

到目前為止，廣播新聞語料中之國語 411 音、particle、silence、garbage、英文和閩南語的切割資訊（Boundary Information）都準備好了，便可以進行已知切割位置的初始模型訓練，方塊圖如下所示：



圖三-3　已知切割位置模型訓練

另外我們還會建立一個 SP（Short Pause）的 HMM 模型，這是代表音節之間的短暫靜音，SP 只有一個狀態，此狀態允許跳躍（Skip），並且與 silence 的中間狀態合併（Tying）。



圖三-4　SP HMM 模型

3.3.3 訓練程序

具備了初始模型之後，接著必須進行 embedded re-estimation，此時語料的切割資訊已不再需要，在這裡便將第一年語料中，所有無背景聲（Clear Speech）的部份均拿來訓練聲學模型使用，資料數量統計為 2,198 個句子（Turn），共有 168,690 個字，時間長度約 9.5 個小時。在訓練過程中，會依據資料量多寡，針對國語 411 音節和 particle 的狀態 mixture 個數進行調整（每 50 個音框增加一個 mixture），使每個狀態的 mixture 數介在 1 到 16 之間，重複訓練至收斂為止。最後總共訓練出的聲學模型資訊統整如下：

表三-2　基本辨識系統 HMM 參數設定

模型種類	個數	狀態數	Mixture/狀態
聲母	100(RCD)	3	1 ~ 16
韻母	40	5	1 ~ 16
Particle	19	3	1 ~ 16
Breath	1	3	16
Silence	1	3	32
SP（Tie to the middle state of Silence）	1	1	32
Garbage	3	3	32

3.4 依語者環境個別建立聲學模型

由第二章得知，雖然同屬廣播新聞語料庫，但是不同的語者環境，在發音特性和語料現象組成比例仍然有所差異，因此依據不同語者環境各別訓練、建立其

聲學模型實屬必須，接下來便要進行依照語者環境建立各自的聲學模型。

3.4.1 訓練語料選擇

要進行依照語者環境區分建立模型，首要條件便是要有足夠的訓練語料，所以在此使用 MATBN 資料庫中第一年、第二年的語料，扣除掉其中不適用的部份語料後，目前可用資料共 24.2 小時，約 42 萬個字，再依三種語者環境將其分開，並取十分之九做為訓練語料，保留十分之一留待測試時使用，各環境所有可用語料及訓練語料資訊如下：

表三-3　各環境下的可用語料數量

訓練語料環境	Turn 數	中文字數	時間（小時）
內場主播	2,261	190,100	10.94
外場記者	2,373	114,239	6.3
外場受訪者	1,849	109,416	7.03

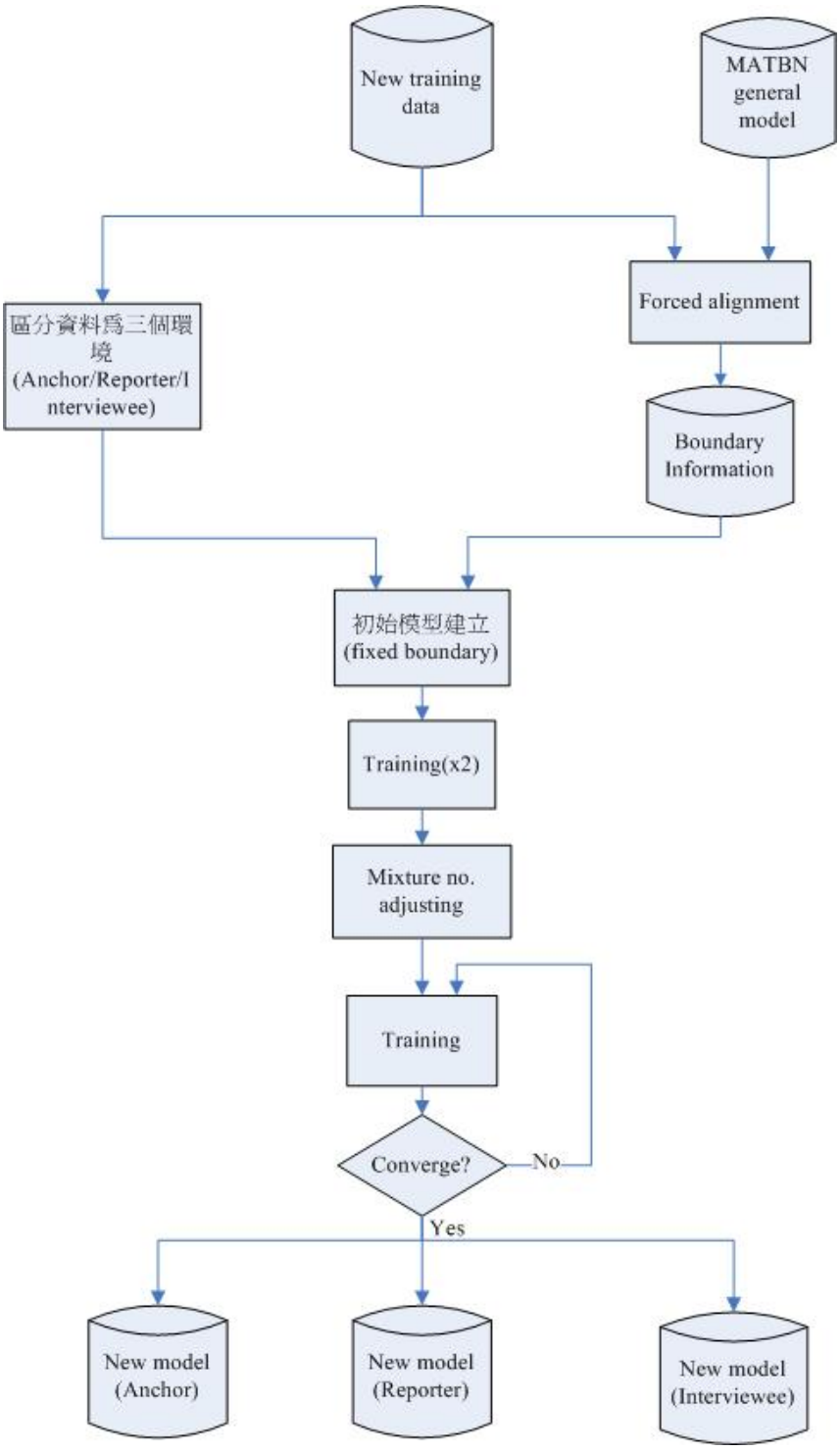
表三-4　各環境下的訓練語料數量

訓練語料環境	Turn 數	中文字數	時間（小時）
內場主播	2,071	175,194	10.1
外場記者	2,167	104,960	5.8
外場受訪者	1,666	99,039	6.4

3.4.2 訓練流程

在上一節中，我們已經利用廣播新聞語料庫 MATBN 建立了一個不分語者環境的 general model，其中包含有廣播新聞語料中各種類型聲音現象的聲學模型，

因此現在便可以用來對目前依據語者環境分為三類的訓練語料進行切割，以得到切割資訊，緊接著便可以分別建立三種環境的初始模型，訓練聲學模型，其中詳細方法與建立 general model 時相同。由下方流程圖可清楚了解依環境訓練聲學模型之流程。



圖三-5　依環境訓練聲學模型流程圖

在訓練的過程中，對於聲學模型中 mixture 個數的調整方式仍與之前相同，每達到 50 個音框增加一個 mixture，但在此我們將 mixture 個數的上限增加到 32 個 mixture，整理最後各個語者環境的聲學模型，其中模型參數如下表所示：

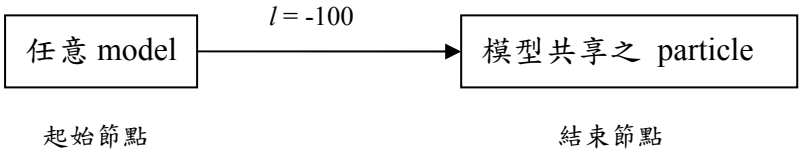
表三-5　各環境下 HMM 參數設定

模型種類	個數	狀態數	Mixture/狀態
聲母	100(RCD)	3	1 ~ 32
韻母	40	5	1 ~ 32
Particle	35(4/7/16)	3	1 ~ 32
Breath	1	3	32
Silence	1	3	64
SP（Tie to the middle state of Silence）	1	1	64
Garbage	3	3	32

如之前提及，建立出一個模型必須要有足夠的資料，又在各個環境中 particle 出現次數不一，因此可建立的初始模型數量也各不相同，但是為了考慮到可用語料中所有出現的 particle 均可能在測試語料中發生，所以我們對每個環境都建立 35 個 particle 的模型，但實際能建出模型的，對三個環境（內場主播、外場記者與受訪者）只有 4 個、7 個與 16 個，其中受訪者因為 particle 出現比例最高，所以可建立的模型數量也最多，而對於資料過少無法建立的 particle 模型，則採用模型合併（Model Tying）的方式去跟相近音共享訓練語料。

但是因為部分 particle 聲學模型的建立利用了模型合併的方法，所以必須要避免因此而造成辨識時採用模型共享的 particle 出現機率超過原 particle 的情形，

為了解決這個問題，我們會在由原本無文法規則（Free Grammar）所轉出之word-net（文字網路）上做適當的處理。Word-net 的功能是定義辨識系統中所有節點之間的轉移關係，而每個轉移都是由一個起始節點（Start Node）和一個結束節點（End Node）組成。所以，我們只要在所有結束節點為這種採用模型共享的 particle 的轉移上減去一個分數（Language Model Score），使其被辨識出之機率降低，便可避免上述問題發生。處理方法圖示如下：



圖三-6　模型共享之 particle 在 word-net 的處理

3.5　實驗—不同環境之聲學模型辨識效能

聲學模型建立完成後，便可以利用之前保留的測試語料來進行辨識的實驗，觀察各個環境聲學模型的辨識效能。

3.5.1　測試語料

如之前所述，我們排除掉有背景聲之部分剩餘的語料作為可用語料，並隨機保留十分之一作為系統 outside 測試所需的語料（其中均無外國語言），又因為從第二章中語料庫不同環境語者總數統計得知，內場主播和外場受訪者的語者數都不算多，留下來的測試語料跟之前使用的訓練語料會有語者重複的情形，如此一來實驗所得到的只能算是多語者（Multi-speaker）辨識系統的結果，但是受訪者的語者數量則有將近三千五百人，即使測試語料是隨機保留，會遇到跟訓練時的使用的語料相同語者的機率並不大，所以可以算是語者獨立（Speaker Independent）辨識系統的實驗，保留的測試語料數量統計結果如下：

表三-6　各環境下的測試語料數量

訓練語料環境	Turn 數	中文字數	時間（小時）
內場主播	190	14,906	0.84
外場記者	210	9,279	0.5
受訪者	186	10,382	0.63

3.5.2　實驗結果

在基本辨識系統中，是採用音節作為辨識單元，並且在此用無文法（Free Grammar）之文法規則，因此辨識率的計算便只考慮音節辨識率（Syllable Recognition Rate），又辨識結果中可能包含有 particle、呼吸聲等現象，為了避免混淆，在計算辨識率時均只將國語 411 音節列入考慮，計算方式如下：

$$Accuracy = (1 - \frac{Sub + Del + Ins}{Number}) \times 100\% \qquad (3.1)$$

其中 *Sub* 為替代型錯誤，*Del* 為刪除型錯誤，*Ins* 為插入型錯誤，而 *Number* 則是音節的總數。最後計算得到之 outside 測試辨識率如下表所示：

表三-7　Outside 測試語料 syllable 辨識率

環境	Sub	Del	Ins	Accuracy
內場主播	18.51%	2.88%	0.85%	77.76%
外場記者	27.88%	2.56%	1.15%	68.40%
受訪者	45.73%	6.87%	5.08%	42.32%

3.5.3 實驗分析

根據以上辨識結果，可以得到以下幾個結論：

- 內場主播因為受過專業發音訓練，而且人數少、大多數語料均為同一位語者的聲音，此外，主播所在的環境安靜無打擾，所以有較高的辨識率。

- 外場記者雖然咬字也很清晰，又因為處在吵雜的環境，容易受到影響，所以辨識率不如內場主播。

- 受訪者因為人數眾多，說話比較口語化、發音比較不精確，而且處在易受打擾的環境，所以辨識率與之前兩者相較之下有一段不小的差距。

- 由於內場主播與外場記者的語者數都不算多，因此建立的系統只能算是多語者辨識系統，但是受訪者的辨識器則是由上千名語者的聲音建立，真正是語者獨立辨識系統，基於如此的差異，必定將造成受訪者的辨識系統會得到較低的辨識率。

第四章　加入語言模型之語音辨認器

語言模型（Language Model，LM），可區分為兩種，第一種是依據語言的文法、字詞的詞性，訂定文章出現一定要符合的規則之語言模型（Rule-Based LM）；另一種則是藉由處理很大量的文字資料，從中利用統計的方法，以得到詞與詞之間的聯接規則而建立的語言模型（Statistic-Based LM）。在此我們使用一般辨識器較常使用的基於統計之語言模型。

4.1　語言模型簡介

事實上所有的語言都有一些文法規則，而利用這類文法規則所建立出的機率模型，則稱之為語言模型。若在進行語音辨識時，能夠將 LM 的資訊，配合聲學模型（Acoustic Model）共同使用，通常能夠大幅提升辨識系統的效能。

在漢語中文（Mandarin）的情況下，建立語言模型時，一般是以詞（Word）作為基本單位。理由是在中文語言中，以「詞」為基本單位建構成句子比較符合語言規則，以「藝術」這個二字詞為例，若將它拆成「藝」和「術」這兩個字元（Character），則並不如原本的詞那樣具有意義。因此，在正式開始建立 LM 的工作前，必須先準備一個詞典（Lexicon），其中定義著全部所要使用的詞。

4.1.1　n-gram 語言模型

假設有一個句子（Sentence），其內容以詞（Word）為單位所組成，總共有 m 個詞，也就是「$w_1, w_2, ..., w_m$」，其中「w_i」代表句子中的第 i 個詞，則產生這個句子所對應的機率，可以拆解為一連串的條件機率（Conditional Probability）之連乘 [10]：

$$P(w_1, w_2, ..., w_m) = P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1, w_2) \cdots P(w_n \mid w_1, w_2, ..., w_{n-1}) = \prod_{i=1}^{m} P(w_i \mid w_1, ..., w_{i-1}) \quad (4.1)$$

但是因為記憶體的大小有限，而且要求得所有詞的條件機率是不可能的，所以若是給予適當的假設，則可以使用 n-gram 的機率去趨近（4.1）式。

$$P(w_1, w_2, ..., w_m) \simeq \prod_{i=1}^{m} P(w_i \mid w_{i-n+1}, ..., w_{i-1}) \quad (4.2)$$

其中每個 n-gram 的機率，可藉由在大量文章中詞串（Word Sequence）所累積的出現次數計算而得，如下式所示：

$$P(w_i \mid w_{i-n+1}, ..., w_{i-1}) = \frac{Count(w_{i-n+1}, ..., w_i)}{Count(w_{i-n+1}, ..., w_{i-1})} \quad (4.3)$$

其中，*Count*(•) 表示詞串的出現次數。而語言模型也就是由許許多多的 n-gram 之機率所組合而成。

4.1.2 機率的 smoothing

由（4.3）式可以知道，如果在分子的*Count*(•) 值為 0 時，則此 n-gram 的機率會等於 0，但是一個詞串在部分文章中沒有出現過，並不代表辨識結果中絕不會有這種組合出現，因此這種情況下機率的給定是不合理的，而且在消息理論（Information Theorem）上來看機率 0 會使得資訊量無窮大，而造成錯誤的估計。此外，當*Count*(•) 的值很小的時候，所計算出的 n-gram 機率也是不準確、信心度不足的，所以還必須對利用（4.3）式所計算出的機率做 smoothing 的動作 [11]，使所有的 n-gram 機率均能夠被良好的估計，一種常見的 smoothing 方式如下：

$$P(w_i \mid w_{i-n+1},...,.w_{i-1})$$

$$= \begin{cases} a(w_{i-n+1},...,w_{i-1}).P(w_i \mid w_{i-n+2},...,w_{i-1}) & : Count(w_{i-n+1},...,w_i) = 0 \\ d_a.\dfrac{Count(w_{i-n+1},...,w_i)}{Count(w_{i-n+1},...,w_{i-1})} & : 1 \le Count(w_{i-n+1},...,w_i) \le k \\ \dfrac{Count(w_{i-n+1},...,w_i)}{Count(w_{i-n+1},...,w_{i-1})} & : Count(w_{i-n+1},...,w_i) > k \end{cases} \quad (4.4)$$

式中 $a(w_{i-n+1}, ..., w_{i-1})$ 表示為 back-off 係數，也就是當計算 n-gram 機率所用的詞串出現次數為 0 時，則利用其 $(n-1)$-gram 的機率，再乘上 back-off 係數，這樣便可避免機率 0 的出現，並分配給它一個適當的機率值。而 $a(w_{i-n+1}, ..., w_{i-1})$ 的選定，還會經過 normalization，令其滿足：

$$\sum_{w \in V} P(w_i = w \mid w_{i-n+1},...,w_{i-1}) = 1 \quad (4.5)$$

而關於 $Count(\bullet)$ 的值很小時所造成的 n-gram 機率不準確的問題，解決方法是當詞串的出現次數小於 k 次時，則將這個 n-gram 機率乘上一個依據 Good-Turning discounting 所計算出之小於 1 的值 d_a（Discount Coefficient Factor）[11]，以減低其機率（相當於對它的值較沒信心），並將扣除的這些機率分給詞串沒有出現（Unseen Event）的 n-gram 機率使用。

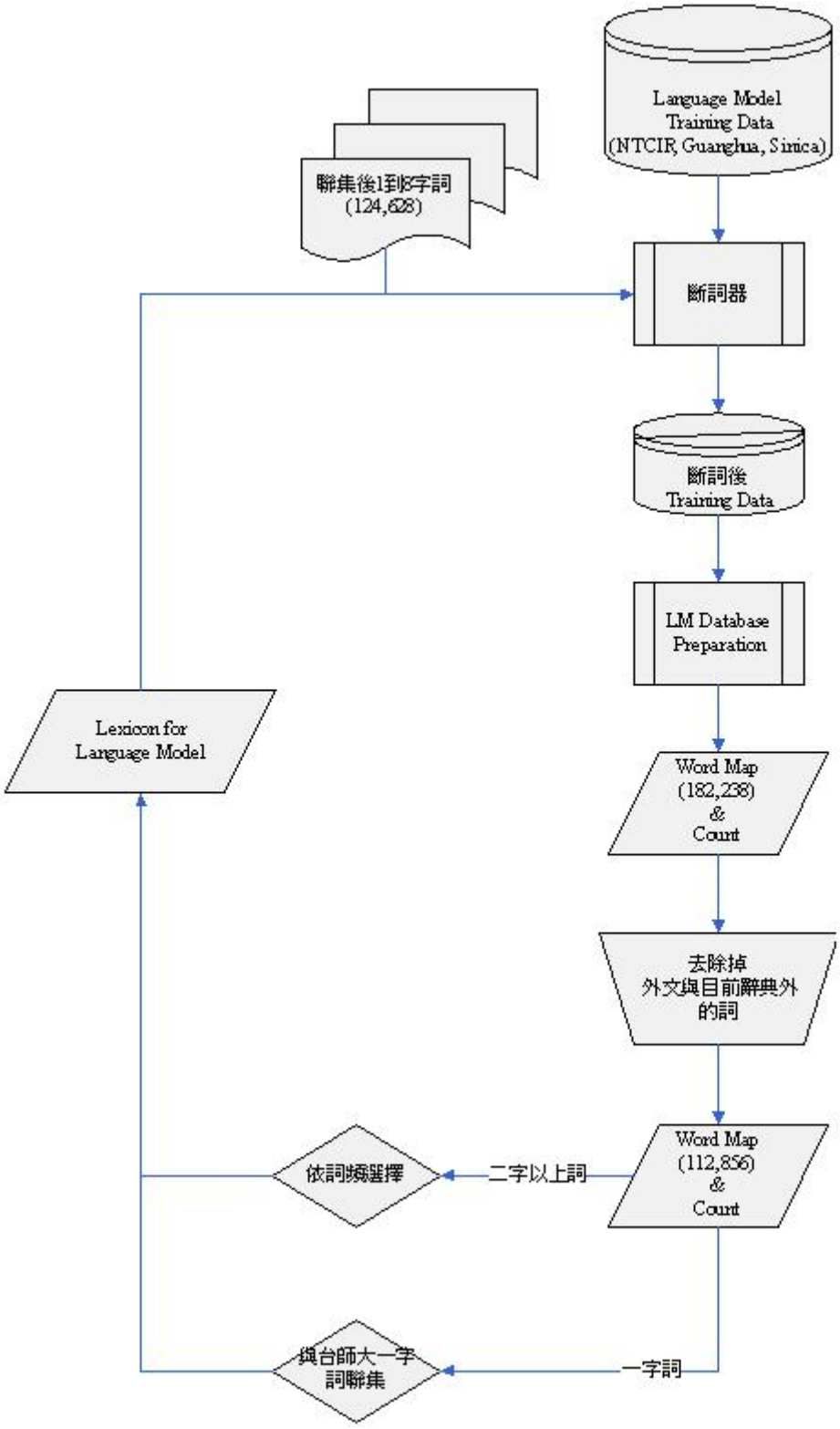
4.2 詞典的選擇

想要建立出一個完善的語言模型，其中一項關鍵的因素便是詞典（Lexicon）的選擇。詞典中的內容決定著辨識結果中能夠出現的詞，所以其最終目的是將一種語言中所有存在的詞都納入其中，但受限於記憶體的大小，我們僅能夠將較常出現、較重要的詞整理出來，包含在詞典之中提供給建立 LM 時使用。

4.2.1 標準詞典的建立

語言模型的建立過程中，其中首要物件就是一個包含有語言模型中所要認識的所有詞的詞典。實驗室目前所擁有的詞典來源共有三個，分別是中研院八萬詞詞庫、交通大學語音實驗室自訂詞條與台師大詞典，取三者的聯集作為一開始所使用的詞典，其中共計有十二萬四千多詞，但因為記憶體容量等因素限制，並不能將目前所擁有的全部詞都加入最後要使用的詞典中。關於詞典大小的選擇，可以藉由觀察其包含率（Inclusion Rate）來決定，因此便先利用實驗室的中文斷詞器 [12]，依據十二萬多詞的詞典對所擁有的文字資料作斷詞，而文字資料庫來源為光華雜誌（Sinorama）、中文資訊檢索標竿測試集（NACSIS Test Collections For IR，NTCIR）以及中研院平衡語料庫（Sinica Corpus），這三個資料庫除了在選擇詞典時使用外，也將作為 LM 建立的主要訓練資料來源。

而接下來的工作便是由目前全部的十二萬四千多詞中，挑選出較重要、較具代表性的詞，作為接下來要建立語言模型時要使用的詞典，整體詞典處理選擇流程圖如下所示：

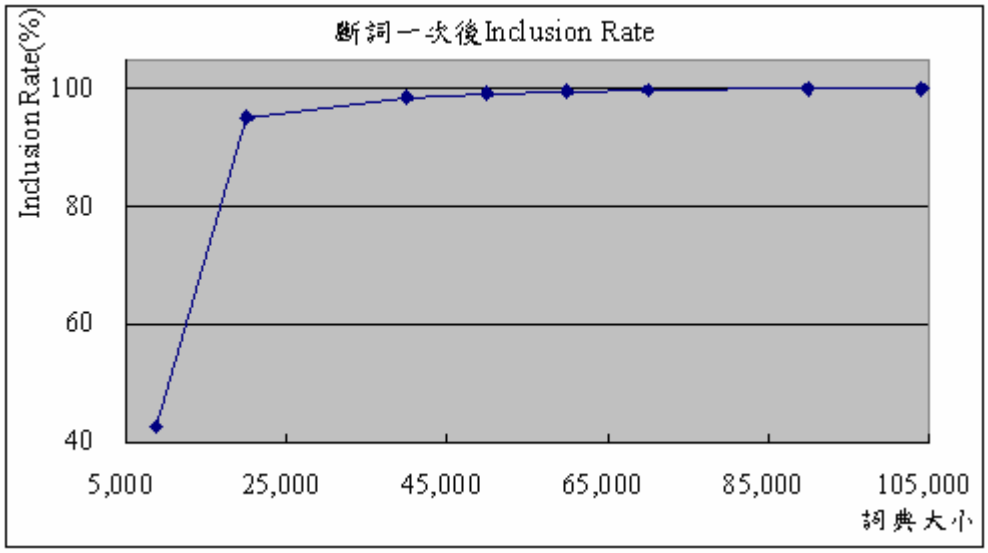


圖四-1　詞典處理選擇流程方塊圖

在此詞典的選擇方式，是用最直覺、也是最簡單的選詞方法，便是由斷詞結果中統計出各詞的詞頻，並依據詞頻大小來決定詞的重要性，但又因為考慮到被

刪除的長詞會變為 OOV（Out of Vocabulary），所以對於一字詞部分的做法稍有不同，為了避免 OOV 的問題，必須將全部的一字詞都保留在詞典中，如此一來即使長詞被刪去，它也能夠被斷成數個短詞的連接，而不會造成 OOV 的出現，所以我們將全部斷詞結果中有出現的一字詞與台師大的一字詞作聯集，得到的結果當作是詞典中一字詞的部份，其它二字詞與二字詞以上的詞則依據詞頻出現次數來選擇，使最後詞典的只需要適度的大小便能夠有足夠的包含率。

藉由以上詞典處理流程，以使用十二萬詞斷詞之結果為正確答案，經由觀察如下圖所示之使用不同詞數之詞典包含率，發現取六萬詞左右便能夠使包含率超過 99.6%，所以便將最後詞典的大小設定為六萬詞。

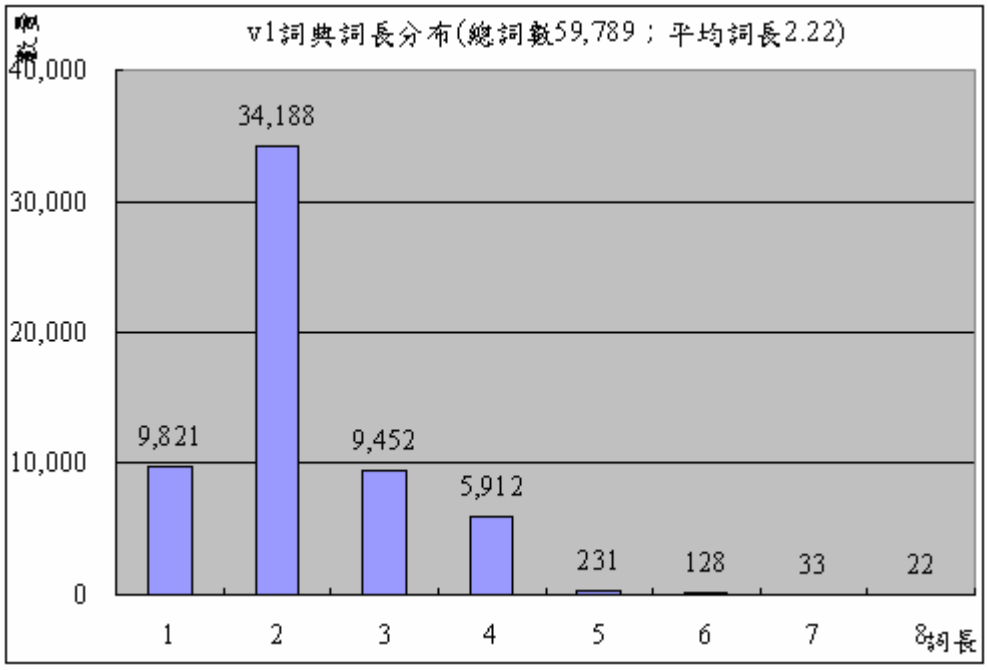


圖四-2　詞典包含率

我們共選擇了較重要的 59,787 個詞作為接下來建立語言模型時所需的詞典，而此詞典之平均詞長為 2.22(字 / 詞)，其中各詞長比例表與數量分布圖如下所示：

表四-1　六萬詞詞典詞長比例表

詞長	1	2	3	4	5	6	7	8
百分比	16.35%	56.92%	15.74%	9.84%	0.38%	0.21%	0.05%	0.04%



圖四-3　六萬詞詞典詞長分布圖

而在詞典中，我們對每個中文詞的表示方式均採用「Big5 碼_漢語拼音」，例如「表示」這一個二字詞，在詞典中其格式便是「AAEDA5DC_biao3shi4」。

4.2.2　加入廣播語料新詞

因為我們用來選擇詞典所使用的文字資料庫內容均為文字資料，其中內容並不包含口語語音中才會出現的呼吸聲、particle 等聲音現象，如此建立出來的語言模型中便不會含有這類現象的 n-gram 機率，所以為了使訓練出的語言模型能夠較接近廣播新聞語料的特性，我們會將廣播新聞中最常出現的兩種口語現象，呼吸聲和 particle，加入到詞典中，並在最後再利用 MATBN 的語料文字，建立出含有這類 n-gram 機率的語言模型，提升辨識器的辨識效能。

4.3 語言模型建立過程

在此首先利用大量的文字資料訓練出一個涵蓋範圍廣泛、通用於各個領域的語言模型，基於此種語言模型的普遍性，稱之為「General LM」。當 general LM 建立完成後，為了使語言模型的特性比較接近廣播新聞語料的特性，接著利用語言模型調適（Language Model Adaptation）的方法，利用 MATBN 的語料之轉記文字內容用去訓練另外一個語言模型（Language Model For Adaptation），在此我們使用除了測試語料之外的 MATBN 全部文字資料，如此建立的語言模型不但能夠產生含有 particle 和呼吸聲的 n-gram 機率，而且更為符合研究中廣播新聞領域的應用，最後再將兩個模型的訊息加以組合，產生一個調適過後的語言模型（Adapted Language Model）配合之前建立的聲學模型合併使用。

4.3.1 Bigram 語言模型的建立

- 訓練文字資料

  要建立一個好的語言模型，除了適當的詞典選擇之外，另外就是必須要有大量的文字資料庫。在這裡用來建立 general LM 的文字資料庫總共有三個，分別介紹如下。

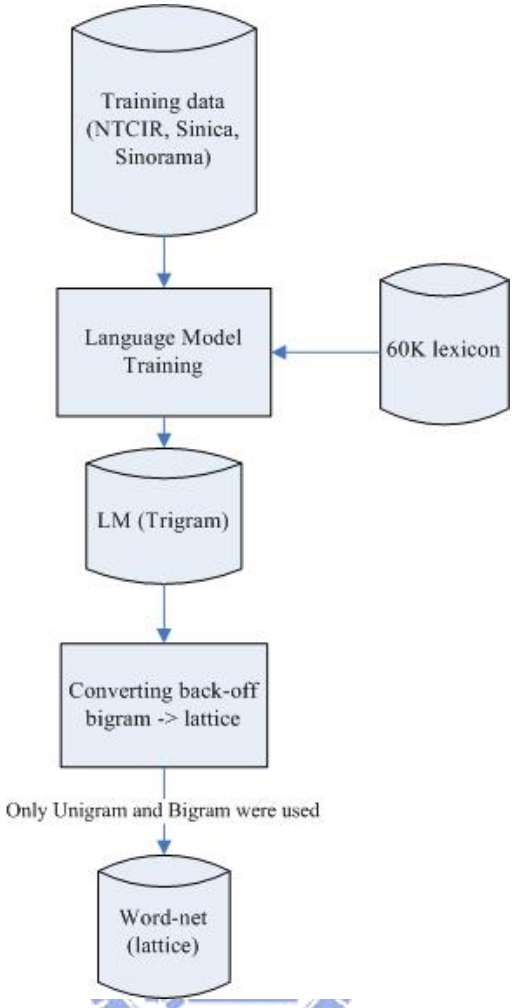
  第一個是光華雜誌，其內容為一般雜誌的文章，蒐集的資料年代範圍介在 1976 年到 2000 年之間；其次是 NTCIR 是一個建立資訊檢索系統的標竿測試集，其內容由數種不同學科領域的文章構成；最後是中研院平衡語料庫，是一套由中研院錄製，內容包含多種主題，以語言分析研究為目的的資料庫。利用之前選定的六萬詞詞典，去對這三個資料庫利用中文斷詞器斷詞後的斷詞結果之詞數與字數數量統計見下表：

表四-2　General LM 訓練語料統計

訓練語料	詞數（Word）	字數（Character）
光華雜誌	9,870,430	16,406,485
NTCIR	124,442,861	206,847,107
平衡語料庫	4,796,163	7,972,113
合計	139,109,455	231,225,705

- 語言模型產生流程

藉由輸入大量的文字資料，統計出各種詞串在文章中累計的出現次數後，便可以利用（4.3）式並配合（4.4）式的 smoothing 方法，計算出建立 LM 所需的 n-gram 機率，在此為了接下來的研究所需，我們分別求出 unigram、bigram 和 trigram 機率。雖然到此為止已算是將 LM 建立完成，但若要將語言模型加入辨識系統中和聲學模型共同使用，則仍須將 LM 轉為 word-net 的形式，其中清楚描述著詞與詞之間的連接關係與轉移機率，又因為在此的目標是建立 bigram 語言模型，所以目前只使用了 unigram 以及 bigram 的機率。整體語言模型建立流程可參考下方流程圖：



圖四-4　Bigram 語言模型建立流程圖

4.3.2　語言模型的調適

- 調適文字資料

上述方法所建立的語言模型的長處是利用了大量資料的普遍性（General），不過會存在著因為文字資料和辨識器應用領域不同而造成語言模型不準確的缺點，為了讓語言模型能夠更接近廣播新聞語料的特性，在此採用語言模型調適（LM Adaptation）的方法 [13]，利用調適語料另外建立一個特性相同的語言模型，再用來對 general LM 進行優化（Refinement）的動作，使調適後的語言模型兼具資料量大和特性相近的好處。其中除了中文字之外的詞，我們將全部的 particle 共同作為一個 class 進行訓練，另外也對呼吸聲獨立訓練。最後，調適語料 MATBN 斷詞後現象數量統計於下表中：

表四-3　MATBN 調適資料統計

MATBN 文字資料	中文詞數	中文字數	Particle	呼吸聲
數量	1,309,020	2,249,724	23,314	90,052

- 模型調適流程

模型調適的過程中對於條件機率所造成的影響，以一個 trigram 的例子來看，條件機率變化如下式所示：

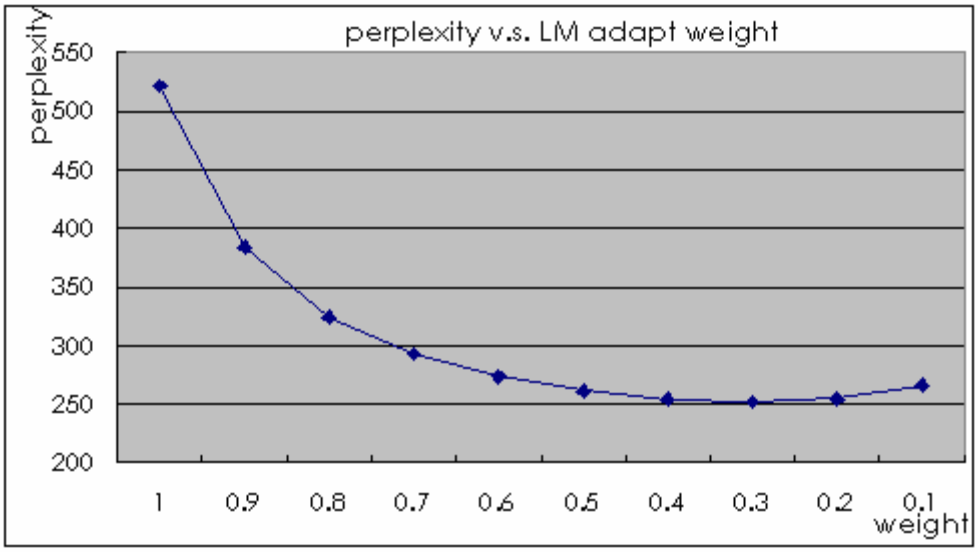
$$P_{adap}(w_i \mid w_{i-1,} w_{i-2}) = \lambda P_{gen}(w_i \mid w_{i-1}, w_{i-2}) + (1-\lambda) P_{MATBN}(w_i \mid w_{i-1}, w_{i-2}) \quad (4.5)$$

式中，P_{adap} 是調適後語言模型中的 trigram 條件機率，P_{gen} 是原本 general LM 中的 trigram 機率而 P_{MATBN} 是在 MATBN 訓練語料所訓練出的 LM 的 trigram 機率，另外 λ 則代表所給定的調適比重（Adaptation Weight），其數值介於 0 到 1 之間，而數值的大小則會依據調適文字資料的多少而不同。

關於 λ 的選擇，目的就是希望調適後的 LM 在進行辨識時能夠有比較好的效能，而 LM 的好壞則可以用 perplexity（PP）來量測，定義如下：

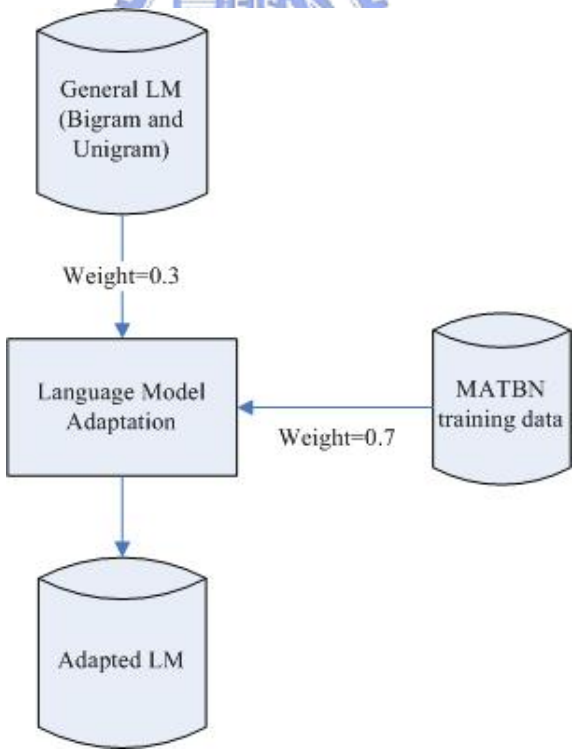
$$PP = 2^{\hat{H}} \quad , \quad \hat{H} = -\frac{1}{m}\log_2 P(w_1, w_2, ..., w_m) \quad (4.6)$$

上式是一個句子（Sentence）之內容共由 m 個詞所組成的例子，其中並對於每個新詞提供的平均資訊量，熵（Entropy，H），經過了 ergodic 的假設與適當的化簡，最後以（4.6）式來對 H 做近似。所以，在此將用來作為辨識系統之測試語料的撰寫文字做為輸入，用來計算給予不同調適比重進行調適後的語言模型之 perplexity，便可以依據得到的 perplexity 大小來進行 λ 的選擇，結果如下圖所示：



圖四-5　不同調適比重之語言模型 perplexity

由上圖中發現，當 λ 為 0.3 時，perplexity 值等於 251.9 最小，所以接下來進行語言模型調適時的調適比重便給定其值為 0.3，而調適流程如下：

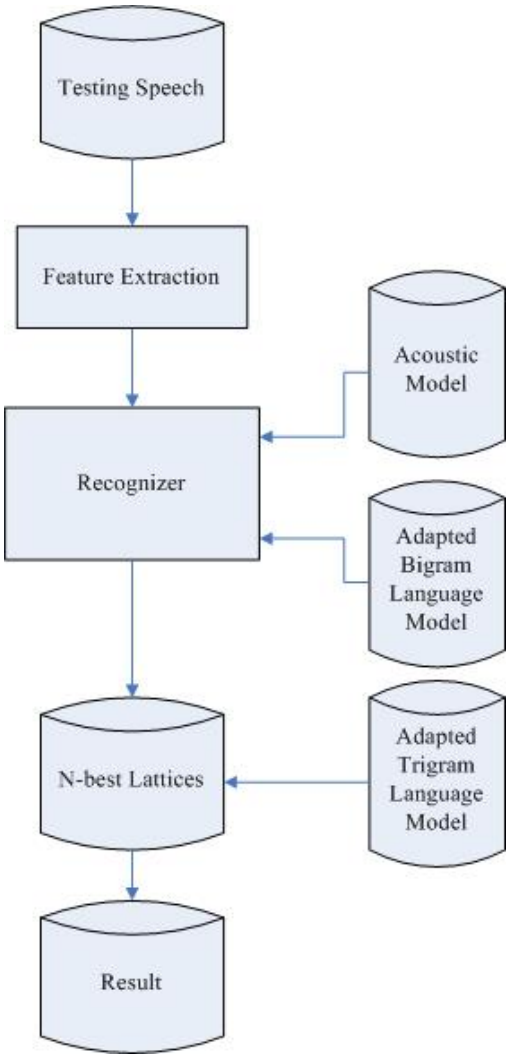


圖四-6　語言模型調適流程圖

4.3.3 Trigram 語言模型的使用

- 模型使用流程

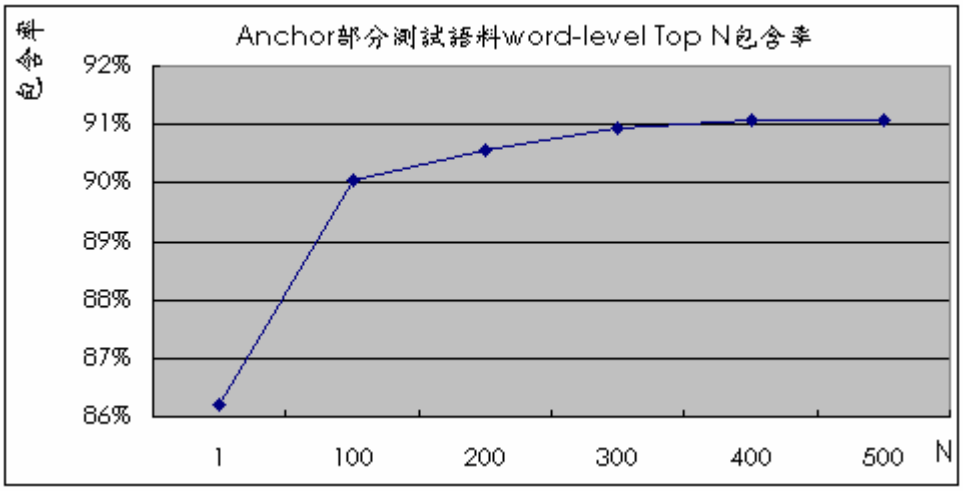
上述過程中，我們計算了 unigram、bigram 和 trigram 條件機率，但因工具限制，辨識系統使用的 word-net 只能夠利用到 unigram 與 bigram 的資訊。可以預期的，若能夠將 trigram 的資訊也加入到辨識系統使用的語言模型中，勢必能夠提升辨識效能，所以採用一種變通的方法利用 trigram 的機率：首先用 bigram LM 之 word-net 配合聲學模型進行辨識，但過程中並非只找出最好的一組答案，而是在每個 state 均保留 N 個 token（N-best），建立 Top-N lattice 並且可以從中產生很多組較好的辨識結果的組合，最後再利用 trigram 語言模型從中選出最好的一組，作為真正的辨識系統輸出結果。



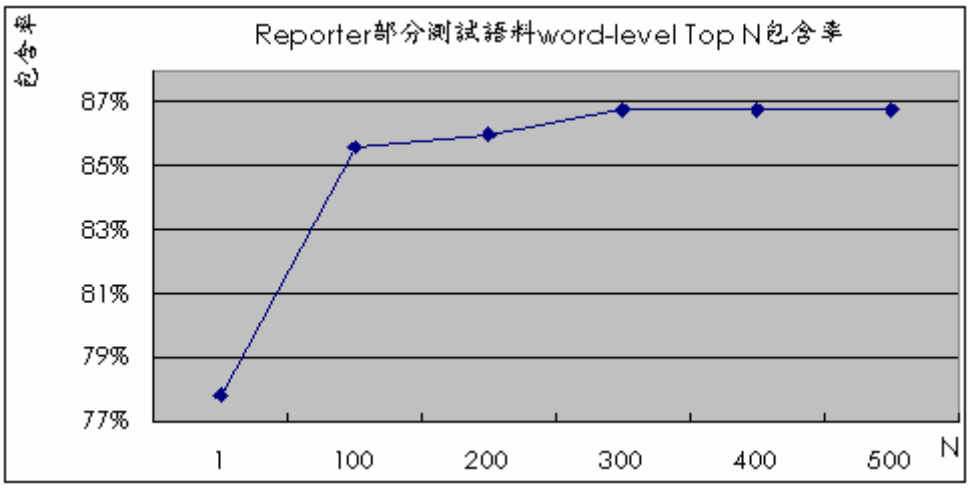
圖四-7　Trigram 語言模型使用方式流程圖

- State 之 token 數的選擇

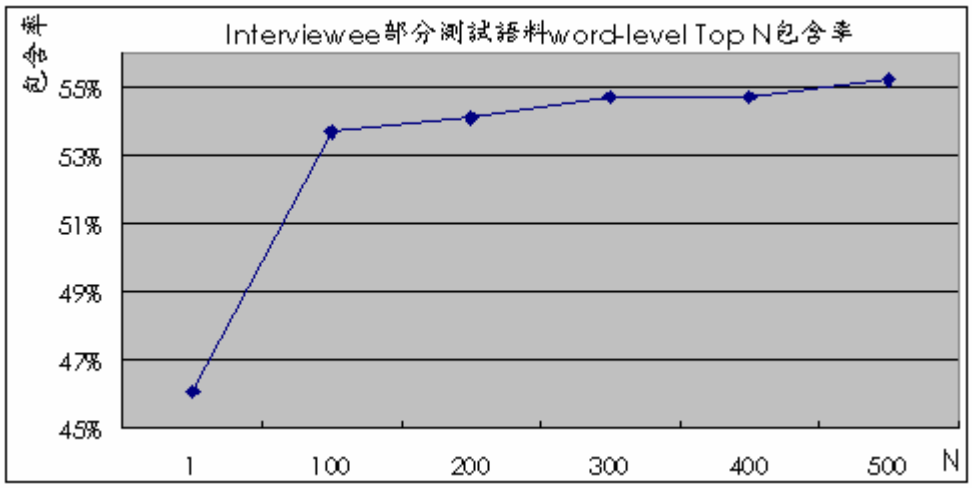
State 要保留的 token 數的選擇目標，是希望能夠找到一個越小越好的值，但又能夠包含越高越好的辨識率在其所有的輸出答案組合中（包含率），以提供足夠的進步空間再配合使用 trigram LM 進行辨識，但可預期的是這兩者對於 token 數的選取所造成的影響，將呈反比例關係。在此我們採用測試語料的十分之一，作為選擇 token 數的實驗語料，並觀察所選擇的 token 數是否能夠提供足夠的辨識率成長空間。三種語者環境在 token 數設為 10 的情況下得到的辨識包含率為：



圖四-8　內場主播 10-best 詞辨識包含率



圖四-9　外場記者 10-best 詞辨識包含率



圖四-10　受訪者 10-best 詞辨識包含率

由以上實驗結果圖形中可以看出，三種環境之曲線均有類似的情形，當 token 設定為 10，辨識結果組合數到達 500 時，便已經提供了一段不小的辨識率進步空間給配合 trigram LM 辨識時使用。

4.4　考慮破音字效應

如第二章所述，研究中使用的資料庫屬於廣播新聞語料，音檔的正確答案（Transcription）均是經由標記員去聽錄製完成的新聞節目，同時對節目內容進行標記，而過程中對於文字內容的標記則記錄到文字層級（Word-level），但是在辨識系統建立流程中，必須使用到音節層級（Syllable-level）的正確答案才能對聲學模型進行訓練，因此，之前使用的音節層級正確答案都是直接採用每個字的最常見音，對於測試語料的正確答案亦是如此。在此將探討破音字的存在對辨識器造成的影響，以及如何加入破音字於辨識系統中。

4.4.1　破音字的影響

若是依照之前的作法，直接對於每個標記的文字都選用其最常見音作為這個字在音檔中的發音標記，並且當作是訓練語料的正確答案提供給 HMM 模型訓練

時使用，這樣便會存在有標音錯誤的情形，造成拿了不完全正確的聲音去訓練 411 音模型的問題，使得訓練出來的模型受到污染而不夠精確；另外，在辨識過程中也會發生測試語料的正確答案與其真實音檔之發音不相吻合的現象，以上兩者均會影響辨識系統的效能以及結果的正確性。

4.4.2 辨識系統的對應處理

由上節中可知，破音字的存在對於辨識器的確會有一定程度的影響，不過仍然有破音字存在的數量和受到影響之音節分布情形等因素需要評估，以決定考慮破音字於辨識系統中的必要性，接下來的過程中將對這個疑問進行驗證，並詳細說明加入破音字後，辨識系統所需要進行的調整。

- 正確答案與聲學模型的修正

  研究中的破音字（目前僅考慮一字詞破音字）來源有二，分別是交通大學語音實驗室常見一字詞破音字表、以及台師大詞典中一字詞破音字[14]，在此取兩者的聯集，又因為包含音調的辨識並非研究中重點，所以再將聯集的結果中只有音調不同的破音字排除，總計共有 510 個一字詞破音字，詳列於附錄二。

  要列入考慮的破音字決定以後，對於破音字在音檔中正確讀音的選擇，最精準的方式便是由人去聽音檔後判定，但在音檔數量龐大的情況下，這種方法並不可行；幸運的，先前所建立的辨識系統即擁有尚可接受的效能，所以利用它來對訓練語料進行 re-alignment，並對破音字的不同發音自動選擇較正確的讀音，如此一來便可以得到正確的訓練語料音節層級正確答案，並可從中統計出若是忽略破音字，對於訓練語料會有何種程度的影響，數量統計如下表所示：

表四-4　考慮破音字後訓練語料變化

條件	Anchor	Reporter	Interviewee
總 sub-syllable 數量	353,682	213,820	208,550
改變的 sub-syllable 數量	2,926	2,082	2,396

由上表中可發現，三種語者環境多考慮了破音字後，對於 sub-syllable 的造成的改變大概僅佔總 sub-syllable 數量的百分之一左右，乍看之下似乎影響範圍有限，但進一步統計後發現，這些改變的 sub-syllable 分布極不平均，例如「那」、「長」、「行」、「和」等破音字均會造成其對應的 sub-syllable 超過一成以上的改變，如此看來，在辨識系統中加入破音字，絕對有其必要性存在。

因此，使用這個較正確的訓練語料音節層級答案去對 HMM 模型進行再訓練，便可得到較精確的聲學模型，接著利用新的聲學模型去對測試語料的正確答案做相同的處理，以減少其中的錯誤標音，避免影響接下來的實驗中辨識結果的正確性。

- 重要破音字的選擇

在進行辨識工作時，並沒有必要將大量的破音字列入考慮，而只需挑選出較重要的破音字即可，在此用來衡量破音字重要性的指標有兩個，首先是每個破音字在斷詞過後的 MATBN 文字資料中的出現次數，其次是由訓練語料中統計出的它們的亂度（Perplexity），也就是當亂度越大的同時，便表示這個破音字會被唸成不同音的機率越高，在這兩個參數的協助下，總共挑選了 27 個較重要的一字詞破音字列入辨識系統中，如下表所示：

表四-5　辨識時加入的破音字

MATBN 中較重要一字詞破音字
了、地、行、佛、沒、那、和、的、長、重、哪、差、參、得、從、都、曾、朝、給、著、說、彈、樂、調、親、還、露

- 語言模型的修正

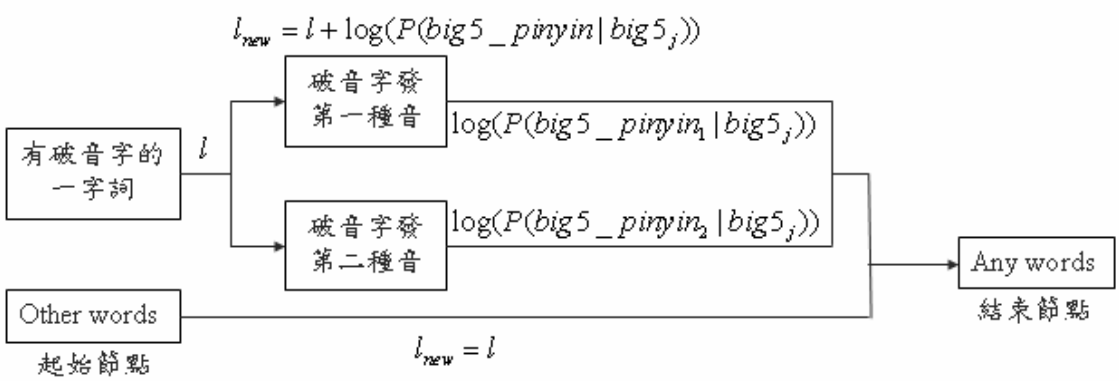
因為將破音字列入了考慮，所以在語言模型中這部份的轉移機率必須要做些適當的修正，我們會將 Big5 碼相同但漢語拼音不同的這些破音字視為同屬於一個 class，以「還」這個破音字為例，則這一個 class 內包含有{還_ㄏㄞˊ，還_ㄏㄨㄢˊ}兩者，因此牽涉到破音字的轉移機率將改為：

$$P_{new}(w_i \mid w_{i-1} = big5_j) = P(w_i \mid w_{i-1} = big5_j) \times P(big5_j \_ pinyin \mid big5_j) \qquad (4.7)$$

其中 $P(w_i \mid w_{i-1} = big5_j)$ 即為原本建立的語言模型中之轉移機率，而 $P(big5_j \_ pinyin \mid big5_j)$ 則可以從訓練語料中統計得到。將（4.7）取對數之後，相乘的關係則改變為相加：

$$\begin{aligned} &\log\left(P_{new}(w_i \mid w_{i-1} = big5_j)\right) \\ &\quad = \log\left(P(w_i \mid w_{i-1} = big5_j)\right) + \log\left(P(big5_j \_ pinyin \mid big5_j)\right) \end{aligned} \qquad (4.8)$$

所以在語言模型的 word-net 的處理上，為了達到（4.8）的方式，我們必須引入 sub-network 的作法，建立出 multi-level 的 word-net，也就是 word-net 內還包含有小的 word-net，下圖中我們將這個概念以圖形化表示：



圖四-11　破音字在 word-net 之轉移機率處理方式

4.5　實驗一──加入語言模型後辨識效能

本實驗在聲學模型部分，仍採用之前依據不同語者環境所建立的三個聲學模型（內場主播、外場記者及受訪者），所使用的測試語料也和 3.5.1 中相同，如此便可觀察到各個語言模型對系統辨識效能的影響。

接下來的實驗中，將分別採用之前所建立的 bigram LM、adapted bigram LM 和 adapted trigram LM 三種不同的語言模型，配合聲學模型進行辨識，過程中為了加快 Viterbi search 以提升辨識速度，都有使用 beam search。

另外，因為聲學模型的使用，辨識結果的基本單元將不再僅止於音節（Syllable），而辨識器將會輸出以詞（Word）為單位的辨識結果，因此我們將可以計算詞的辨識率以及音節辨識率，此外還會將詞的結果轉換到字元（Character），已進行字元辨識率的計算。其中音節辨識率可以用來與先前無文法規則（Free Grammar）時所得到的辨識結果做比較，觀察 LM 效能。

當辨識器中除了聲學模型之外，還另外加入了語言模型時，進行辨識時所產生的結果便不再只考慮聲學模型的分數，而會再配合著語言模型分數的影響，因此會需要給定一個語言模型的分數比重（LM Weight），以決定語言模型對最後辨識結果造成的影響程度，經過在實驗中的測試發現，在使用 trigram LM 時除了主播比重為 11，其它語言模型都是在比重的值為 9 的時候會有較好的辨識效能。

4.5.1 使用 Bigram LM

首先將利用大量文字資料所建立的 general bigram LM 配合聲學模型來進行辨識，三種語者環境下，個別的辨識結果詳列於下表中：

表四-6　Outside 測試語料 word 辨識率

環境	Sub	Del	Ins	Accuracy
內場主播	18.56%	3.09%	3.46%	74.89%
外場記者	26.06%	3.61%	4.33%	66.01%
受訪者	46.50%	8.30%	7.01%	38.19%

表四-7　Outside 測試語料 character 辨識率

環境	Sub	Del	Ins	Accuracy
內場主播	12.72%	2.97%	0.25%	84.06%
外場記者	19.68%	2.58%	0.35%	77.39%
受訪者	38.81%	8.20%	2.34%	49.35%

表四-8　Outside 測試語料 syllable 辨識率

環境	Sub	Del	Ins	Accuracy
內場主播	7.44%	2.99%	0.27%	89.30%
外場記者	12.37%	2.60%	0.37%	84.66%
受訪者	29.38%	8.25%	2.39%	59.98%

4.5.2 使用 Adapted Bigram LM

接下來的實驗中，採用的語言模型是經過利用 MATBN 的文字資料進行調適過後，結合了文字資料量大和模型特性接近廣播新聞兩項好處的語言模型，得到如下的辨識效能：

表四-9　Outside 測試語料 word 辨識率

環境	Sub	Del	Ins	Accuracy
內場主播	9.85%	3.14%	1.31%	85.70%
外場記者	17.46%	4.84%	1.81%	75.89%
受訪者	36.39%	11.94%	3.63%	48.04%

表四-10　Outside 測試語料 character 辨識率

環境	Sub	Del	Ins	Accuracy
內場主播	6.64%	2.62%	0.15%	90.59%
外場記者	13.40%	2.74%	0.37%	83.49%
受訪者	30.88%	8.92%	1.91%	58.29%

表四-11　Outside 測試語料 syllable 辨識率

環境	Sub	Del	Ins	Accuracy
內場主播	4.78%	2.63%	0.16%	92.43%
外場記者	9.53%	2.72%	0.35%	87.41%
受訪者	24.71%	9.06%	2.05%	64.18%

4.5.3 使用 Adapted Trigram LM

最後，進行的實驗中為了能夠利用 trigram 的機率，利用 rescoring 的方法使用了調適過後的 trigram 語言模型，如此所產生的辨識結果如下表所示：

表四-12 Outside 測試語料 word 辨識率

環境	Sub	Del	Ins	Accuracy
內場主播	8.52%	2.99%	0.96%	87.53%
外場記者	15.47%	4.22%	1.76%	78.56%
受訪者	34.81%	12.24%	3.40%	49.55%

表四-13 Outside 測試語料 character 辨識率

環境	Sub	Del	Ins	Accuracy
內場主播	5.51%	2.48%	0.14%	91.86%
外場記者	12.04%	2.55%	0.30%	85.11%
受訪者	30.04%	8.91%	1.86%	59.18%

表四-14 Outside 測試語料 syllable 辨識率

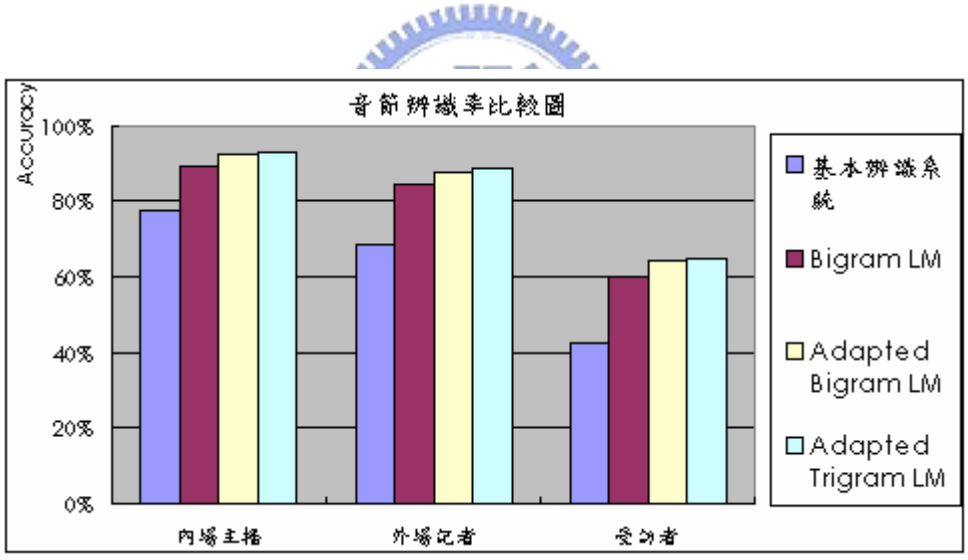
環境	Sub	Del	Ins	Accuracy
內場主播	4.27%	2.50%	0.15%	93.08%
外場記者	8.63%	2.55%	0.30%	88.52%
受訪者	24.24%	9.09%	2.04%	64.63%

4.5.4 實驗分析

在此將第三章基本辨識系統之辨識結果與本章之三種辨識結果之音節（Syllable）辨識率統整於下表中，並以圖形化長條圖表示如下：

表四-15　各種方法的音節辨識率比較表

語者環境	基本辨識系統	Bigram LM	Adapted Bigram LM	Adapted Trigram LM
內場主播	77.76%	89.30%	92.43%	93.08%
外場記者	68.40%	84.66%	87.41%	88.52%
受訪者	42.32%	59.98%	64.18%	64.63%



圖四-12　不同條件下音節辨識率比較圖

- 由上圖中可看出，在辨識系統中加入語言模型之後，三種語者環境下的辨識率均有大幅度的提升，這是因為一般正常情況下講話大多會符合文法規則，若能夠在辨識系統中加入文法規則，而對於 HMM 聲學模型之判斷結果加以限制、修正，則可以產生較符合文法的辨識結果而增加辨識正確的機率。

- 藉由比較圖中的辨識率進步幅度，可發現從完全不使用語言模型的基本辨識系統到使用 bigram 語言模型，辨識率的成長幅度很大，但是 bigram 和 trigram 語言模型之間的辨識率提升幅度則非常有限，因此可預期即使在語言模型的使用上再加入更多的參數量，也無法使辨識器的效能再有大幅度的提升。

接著來看詞辨識率在經過調適後的變化：

表四-16　詞（Word）辨識率比較

語言模型	內場主播	外場記者	受訪者
Bigram LM	74.89%	66.01%	38.19%
Adapted Bigram LM	85.70%	75.89%	48.04%

- 由上表中可發現，三種環境下的詞辨識率均因為調適而有很大幅度的改善，其中又以內場主播的辨識率提升幅度最明顯（error reduction rate 高達 43%），是由於主播說話最符合廣播新聞的文法規則，因此經過調適之後，能夠有很高的 error reduction rate，同理，外場記者的 error reduction rate 也會比說話最口語化的受訪者要來的高。

- 綜合比較三種 level（詞、字元和音節）的辨識率，可以發現不論在哪種環境或條件下，都是音節的辨識率最高、字元其次，而詞辨識率最低，這是因為有些情況下，雖然 word boundary 的選擇而使詞的辨識結果並不正確，但將其轉到字元或是音節時卻有可能是正確的，同理，即使辨識出的字元是錯誤的，但音節也有可能正確，所以會有以上的結果。

4.6 實驗二—考慮破音字後辨識效能

在此，修正前的 word-net 建立方法仍與建立 bigram 語言模型時經過的流程相同，但是因為語言模型過於龐大，為了使加入 sub-network 的工作得以完成，我們首先必須在當時的建立流程中做些改變，也就是將由建立 general LM 時使用的文字資料得到的大量 bigram 中只出現一次的忽略掉，因為這些出現次數極少的 bigram 重要性原本就很低，所以即使對於建立完成的語言模型之效能會造成影響，但是其程度應該有限，接著再對它做調適的動作使它的特性較接近廣播新聞語料的說話特性，最後再加入破音字的 sub-network，便得到了加入破音字的修正後 bigram 語言模型。

另外，接下來的實驗中所使用的測試語料依舊和先前章節實驗中所使用的相同，不過 syllable-level 測試語料的正確答案，則會因為多考慮的破音字而重新經過選擇而和之前有所不同。最後，過程中為了加快 Viterbi search 以提升辨識速度，都有使用 beam search。

4.6.1 實驗結果

在此仍是依據三種語者環境，個別使用將破音字列入考慮後經過再訓練的聲學模型進行辨識工作，在修正後的語言模型配合使用下（三種語者環境之語言模型比重均與之前實驗中相同，設定為 9），我們將可以得到 word、character 以及 syllable 三種不同層級的辨識結果。三種不同語者環境的各個層級辨識結果詳列如下：

表四-17　Outside 測試語料 word 辨識率

環境	Sub	Del	Ins	Accuracy
內場主播	9.44%	3.09%	1.18%	86.29%
外場記者	17.72%	4.76%	1.93%	75.59%
受訪者	37.26%	11.33%	3.57%	47.84%

表四-18　Outside 測試語料 character 辨識率

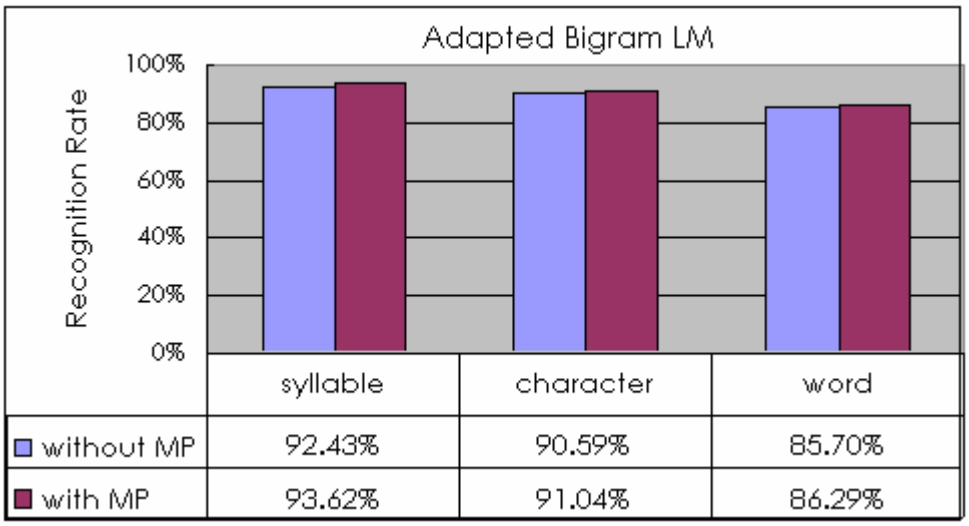
環境	Sub	Del	Ins	Accuracy
內場主播	6.28%	2.52%	0.17%	91.04%
外場記者	13.61%	2.71%	0.30%	83.37%
受訪者	31.49%	8.71%	1.82%	57.99%

表四-19　Outside 測試語料 syllable 辨識率

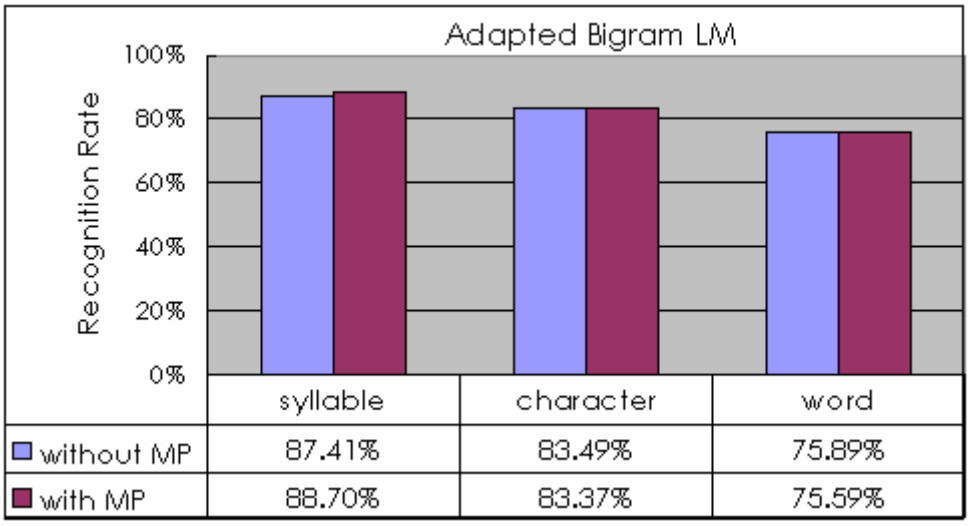
環境	Sub	Del	Ins	Accuracy
內場主播	3.67%	2.53%	0.18%	93.62%
外場記者	8.29%	2.71%	0.30%	88.70%
受訪者	24.21%	8.93%	2.04%	64.82%

4.6.2　實驗分析

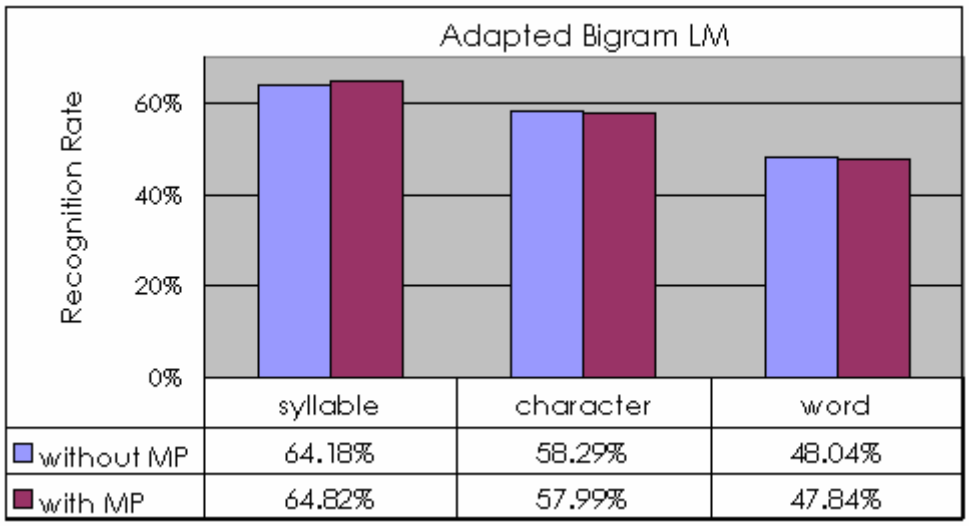
接下來，我們分別將三個環境的三種層級加入破音字後得到的辨識結果，與在第四章使用 adapted bigram 語言模型沒有考慮破音字時所得到的結果做個統整，並畫出長條圖方便觀察。



圖四-13　加入破音字前後內場主播辨識率比較圖



圖四-14　加入破音字前後外場記者辨識率比較圖



圖四-15　加入破音字前後受訪者識率比較圖

- 從以上三圖中可以明顯看出，三種語者環境在加入考慮破音字之後，音節辨識率均有所提升（主播、記者與受訪者三種環境之 error reduction rate 分別為 15.7%、10.2%與 1.8%），這是因為由於破音字的加入，提升了訓練語料和測試語料的音檔發音與標記之間的一致性，所以減少了訓練 HMM 模型時標音錯誤所造成的污染，建立出了較精確的聲學模型。

- 破音字的唸法，事實上會與前後文有關，所以未來應該探討破音字發音與前後文之間的關係。

第五章　將標點符號、音節間靜音長度與詞類模型加入口語語音辨認器

前人在做語音辨認時，大多以一個句子為單位，所以都不會考慮句子中標點符號對辨識系統的影響，而在進行語音辨識工作，建立語言模型的過程中，一般並不將標點符號（Punctuation Mark，PM）列入考慮，但語言模型究竟該如何對待標點符號？將它們忽略是否真的適當？以及更高層的訊息，詞類標記（Part of Speech，POS），是否也能夠對辨識結果的文法規則正確性有所提升？另外，詞內（Intra-word）和詞間（Inter-word）所存在的 pause、以及詞間的 pause 是否有對應到 PM，在如此分類所建立的 pause duration 模型之協助之下，對於 word-level 辨識結果的 word boundary 判斷選擇上，或許能夠有所幫助。

5.1　標點符號特性與分類

標點符號的主要功能是使文章閱讀更生動、意義表達更明確，若希望利用標點符號提供的資訊，並能夠發揮應有的功能，那麼依據標點符號的功能與特性給予適當的分類，將是一個不可或缺的過程。

中文所使用的 PM 共有十六種，並可區分為標號與點號兩大類，其中標號常用的有書名號、破折號、省略號、括號、引號等九種，而點號則有逗號、頓號、句號、冒號、分號、問號及驚嘆號共七種，這兩大類中又以點號跟閱讀時的停頓與否及文法規則有較大的關連性，與目前研究中所需的資訊較符合，所以接下來提及標點符號時均是指點號為主，分類和處理的過程也僅以點號為對象。

首先統計 MATBN 語料之標記內容，利用第四章所建立的六萬詞詞典斷詞，統計過後得到的總詞數、字數以及所標記的點號數量，如下所示：

- 總詞數：1,309,020

- 總字數：2,249,724
- 總點號數：178,267

由以上的統計結果發現，PM 所佔的比例並不算小，其數量約為總詞數的13.6%，所以若加入 PM 的資訊，確實有機會能夠對辨識率有所影響。

接著同時考慮標點符號可能造成的語句停頓和代表的語意文法，對七種點號進行數量統計並分成三類，結果如下表所示：

表五-1　MATBN 語料標記之 PM 數量統計與分類

MATBN	，	、	。	：	；	？	！
數量	124,520	4,318	46,320	56	79	2,950	24
分為三類	COM	DOT	OTH				
數量	124,520	4,318	49,429				

5.2　加入標點符號、POS 和 pause duration 資訊之構想

若在進行辨識工作時，除了詞和聲音參數（Acoustic Feature）之外，也同時將標點符號與 pause duration feature 列入考慮，則式子如下 [15，16]：

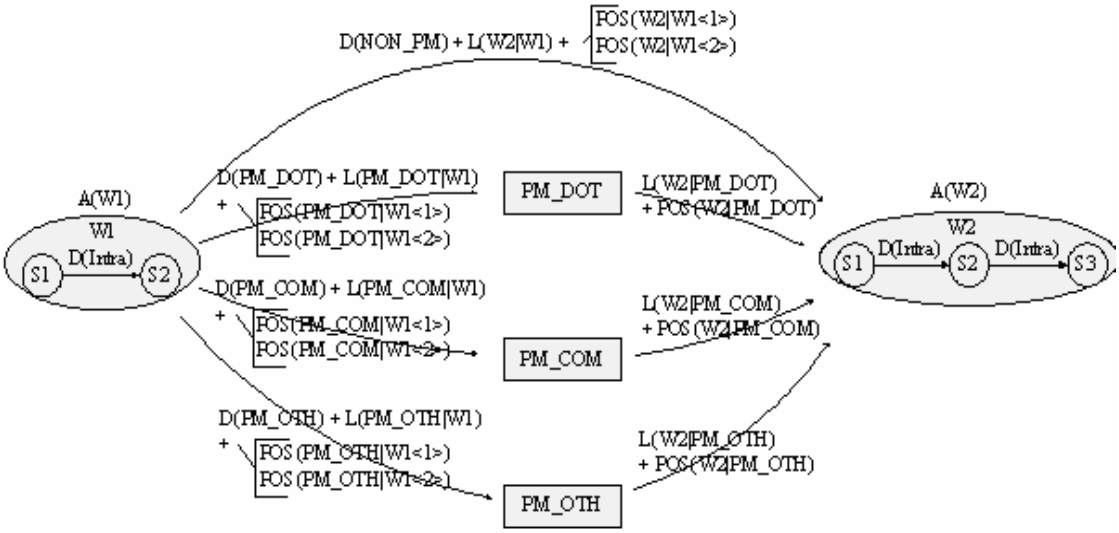
$$
\begin{aligned}
&P(X, X_d, W, P, POS) \\
&\quad = P(X, X_d \mid W, P, POS) \times P(W, P, POS) \\
&\quad = P(X \mid W, P, POS) \times P(X_d \mid W, P, POS) \times P(W, P, POS) \\
&\quad = P(X \mid W, P) \times P(X_d \mid W, P) \times P(W, P, POS) \\
&\quad = P(X \mid W, P) \times P(X_d \mid W, P) \times P(W, P) \times P(POS \mid W, P)
\end{aligned}
\qquad (5.1)
$$

其中W代表詞、X為聲音參數、POS是詞的詞類，而P與X_d則分別代表標點符號以及pause duration，並且假設聲音參數和詞類無關，而pause duration則是與詞類無

關。將（5.1）取對數之後，相乘的關係則改變為相加：

$$\log P(X, X_d, W, P, POS) = \log P(X \mid W, P) + \log P(X_d \mid W, P) + \log P(W, P) + \log P(POS \mid W, P) \quad (5.2)$$

從上式可看出在進行辨識工作時，一句測試語料的分數將由四項分數相加組成，分別為acoustic model分數、pause duration model分數、包含PM的language model分數和POS model的分數。若將四種分數同時列入考慮，以概念圖的方式表示辨識工作的進行，則任意兩個詞（此圖中假設第一個詞為一個擁有兩種詞類的二字詞，第二個詞為三字詞且只有一種詞類）之間所存在的分數如下圖所示：



圖五-1　辨識路徑概念圖

上圖的例子中，W 和 S 分別表示詞（Word）和音節（Syllable），另外 A(・)、D(・)、L(・)、POS(・)分別為 acoustic、pause duration、language、POS model 分數。所以，在詞本身的分數則除了 AM 分數之外，還另外加入了 intra-word pause DM 分數，而原本詞 W1 到詞 W2 的唯一連接法則因為 PM 與 POS 的加入而增加為八種可能，並且現今哪一種連接法最為適當，兩詞間是否有標點符號存在，將變成由 inter-word pause DM 分數配合加入了 PM 的語言模型之 LM 分數和 POS

的模型分數共同來決定。

5.3 標點符號、詞類與 pause duration 模型使用過程

配合工具 HTK 的使用，為了實現以上之構想，在此必須對於第四章建立語言模型的流程稍作修改，使其中 PM 的資訊得以適當的保留，訓練出包含 PM 的語言模型，至於 POS 模型的部份，則需要建立出一個相當於是 class-based 的語言模型配合使用，另外，還需要統計 MATBN 語料中的 pause 長度分佈，建立符合廣播新聞語料的 pause duration 模型提供給計算 DM 分數時使用，最後則是要有一套正確的 rescoring 流程，使需要的資訊得以合理的應用在辨識過程之中。

5.3.1 包含標點符號語言模型的建立

語言模型的建立過程，與第四章中介紹過的流程大致上仍然相同，首先利用大量的文字資料訓練出一個適用領域範圍廣泛的的語言模型（General Language Model），在此使用的文字資料來源有光華雜誌、NTCIR 以及中研院平衡語料庫，採用之前選定的六萬詞詞典斷詞過後，統計得到三者的總詞數近一億四千萬詞，接下來，為了使語言模型的特性能夠比較適合廣播新聞語料辨識系統的使用，利用 MATBN 的語料文字內容（約一百三十萬詞）訓練另外一個符合廣播新聞領域應用的語言模型，最後再將兩個模型的訊息加以組合，產生一個調適過後的語言模型（Adapted Language Model），並在轉換為 word-net 之後提供給辨識器使用。

和之前不同的是，一般用來訓練語言模型、計算 n-gram 機率的文章，都會以 PM 為依據斷開成句子（Sentence），並於頭尾加註 sentence-boundary（Sentence-start 與 Sentence-end）的標記，以此為單位進行下一步的機率統計；如今希望能夠在建立出的 LM 中保留 PM 的機率，所以會將文章中存在且屬於決定要保留的幾種 PM，將其視為文字並給予適當的分類與標記（PM_COM、PM_DOT、PM_OTH）。但是，若只是將 sentence-end 前的 PM 視為文字加以保

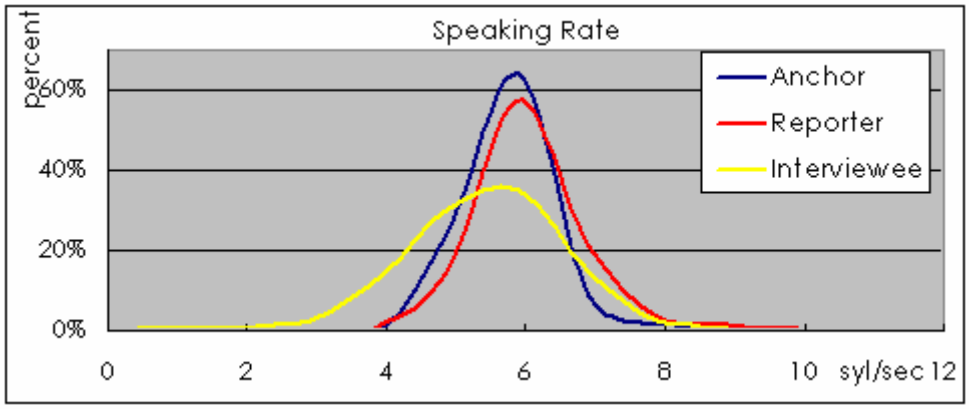
留，而仍然以句子為單位進行語言模型的訓練，如此一來在統計 n-gram 機率的過程中，所有的 PM 都將是伴隨著 sentence-end 而存在，那麼辨識系統想必幾乎不會產生句中的標點符號，因此，除了必須保留 PM、給予標記外，還需要將訓練語言模型的文字資料改為以一段或一節（Paragraph）為單位 [17]，並在頭尾加註 sentence-boundary 的標記，而廣播新聞語料的部份則是以 speaker-turn 為單位進行語言模型的調適，這樣建立出的包含 PM 語言模型，才能夠對 PM 的標記與辨識系統效能的提升有適當的貢獻跟幫助。

第四章中曾經提及 LM 的好壞，可以藉由輸入一段文章所得到的 perplexity 來衡量，所以，在此將辨識系統之測試語料的文字內容經過斷詞並留下 PM 之標記做為輸入，用來計算考慮標點符號之調適後語言模型 perplexity，結果發現 perplexity 從原本不含 PM 時的 255.0 下降為 249.2，由此可發現加入 PM 後確實能夠令語言模型的效能有所提升。

5.3.2 音節間靜音長度模型的建立

- 說話速度與停頓特性

  有鑑於之前的特性統計，三種語者環境的說話聲音特性確實有所差異，所以在此也將根據語者環境的不同，個別建立符合其特性的 pause duration 模型。為了建立適合廣播新聞語料使用的 pause duration 模型，首先將 MATBN 訓練語料區分為三種語者環境，個別計算出每句訓練語料的 speaking rate，且將結果以分布圖表示，另外，並統計詞內的 syllable-boundary 以及詞間的 word-boundary 有 pause 存在所佔有的比例，統計結果如表五-2 所示。



圖五-2　三種語者環境 speaking rate

上圖中可清楚看到三種語者環境的平均說話速度（Speaking Rate）由快到慢依序是，外場記者（5.55 syl/sec）、內場主播（5.27 syl/sec）與受訪者（4.93 syl/sec），其中又以受訪者的語者說話速度差異較大而分布範圍最廣。

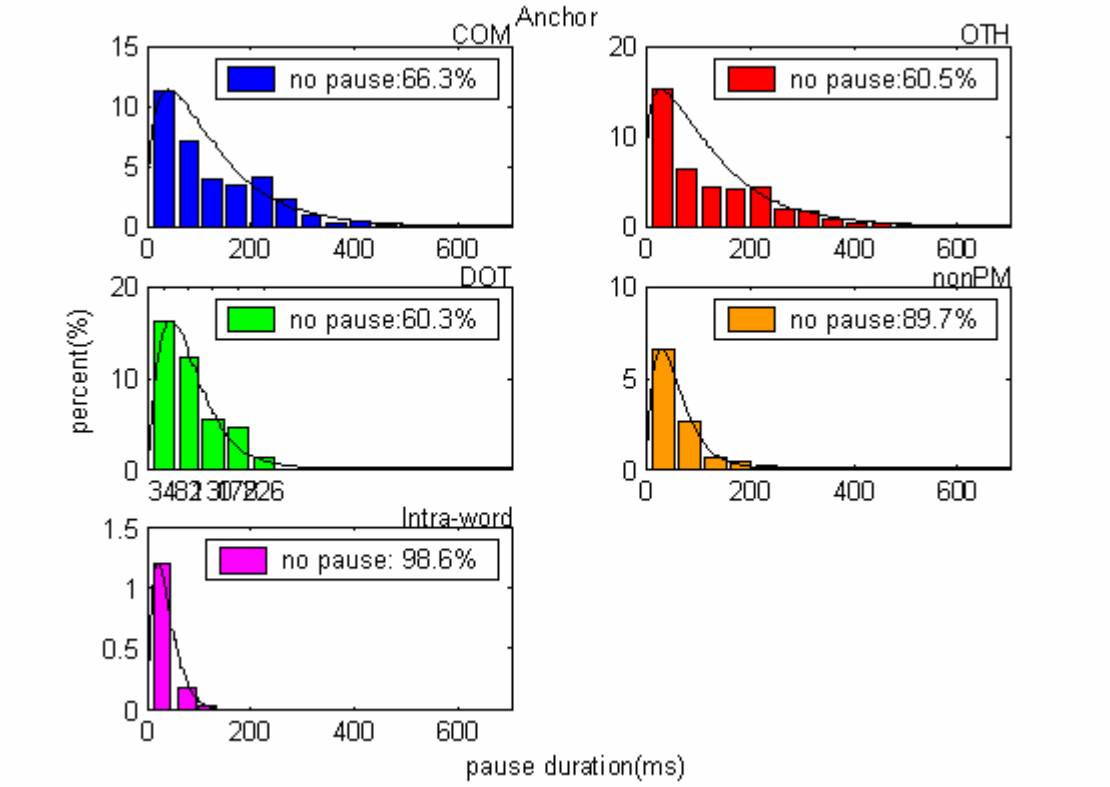
表五-2　MATBN 訓練語料詞內詞間 pause 存在情形

Environment	MATBN	Inter-word				Intra-word
		PM_COM	PM_OTH	PM_DOT	NON_PM	
Anchor	Total number	8,676	1,763	239	94,956	77,706
	With pause	33.7%	39.5%	39.7%	10.3%	1.5%
	Without pause	66.3%	60.5%	60.3%	89.7%	98.5%
Reporter	Total number	5,309	629	373	59,011	44,414
	With pause	48.0%	60.1%	46.9%	7.8%	1.2%
	Without pause	52.0%	39.9%	53.1%	92.2%	98.8%
Interviewee	Total number	5,611	277	211	58,609	42,386
	With pause	60.2%	66.4%	49.3%	13.2%	2.0%
	Without pause	39.8%	33.6%	50.7%	86.8%	98.0%

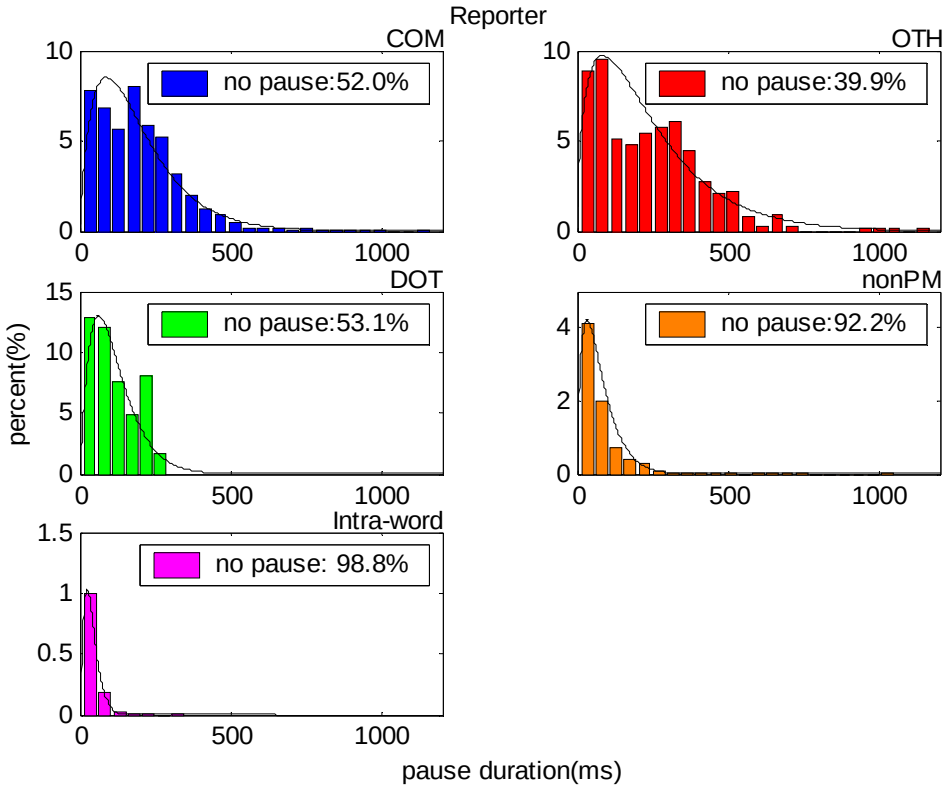
上表之統計結果中，首先觀察三種語者環境共同的特性，由於 MATBN 屬於廣播新聞語料，聲音特性偏向於自然流利語音，而且說話速度較快，所以發現所要標記的三類標點符號，有停頓現象發生所佔的比例都不算太高，即使是說話速度最慢的外場受訪者也只有六成左右的標點符號有停頓，而內場主播 PM 停頓的比例甚至不到四成，另外，和預期相同的是，有 PM 存在的 word-boundary 有停頓現象之比例均較沒有 PM 存在的部份要高出許多，而三種 PM 之中又以 PM_OTH 這類語句結束點號發生停頓的機率最高，最後，詞內的 syllable-boundary 發生停頓現象的比例都非常的低。接著來看三種環境間的差異之處，在沒有 PM 存在的 word-boundary 以及詞內的 syllable-boundary 部分可明顯看到，依據語者環境平均說話速度的快慢，停頓發生的比例也有相對應的結果，停頓比例由高到低為受訪者、內場主播接著是外場記者，但是在三種標點符號的部分，雖然受訪者仍然因為說話速度最慢而有最多的停頓現象發生，不過主播卻因為說話較具連慣性，而造成雖然說話速度較慢，但標點符號不停頓的比例卻比記者要來得多的現象發生。

- Pause duration model 產生流程

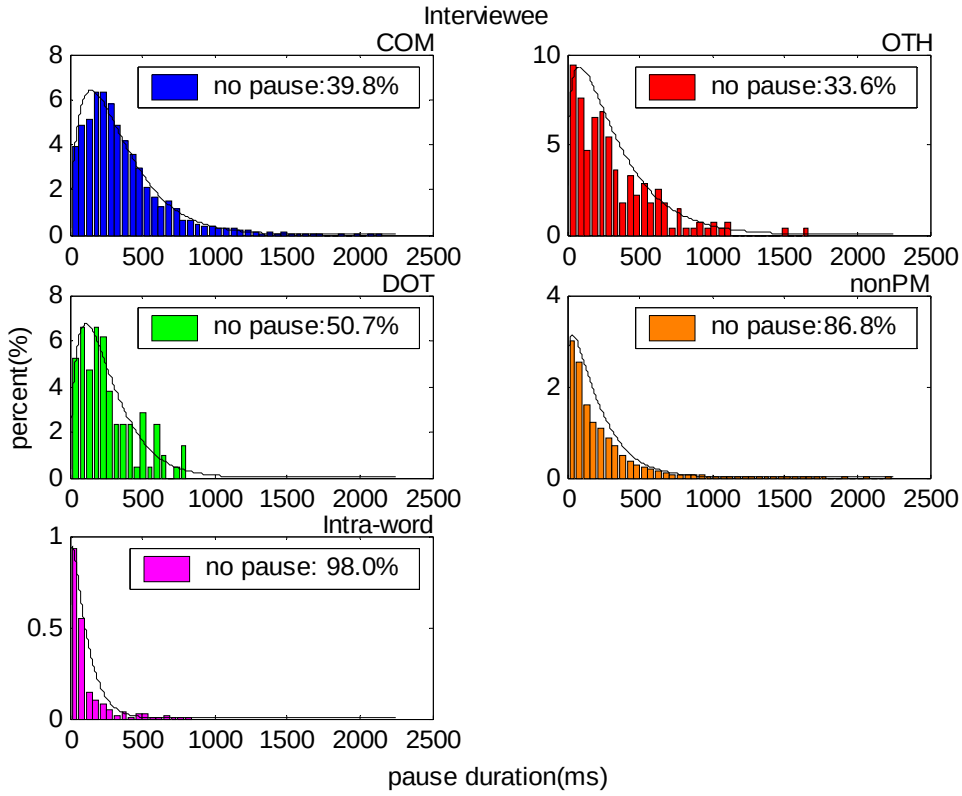
使用第四章所建立之考慮破音字後聲學模型，去對建立 HMM 模型時使用的訓練語料進行切割位置的求取，從 force alignment 所得到的結果中可以得到產生 pause duration 模型所需的 pause 長度，並依據之前決定的種類加以區分為 inter-word 與 intra-word 兩大類，而 inter-word 的部分又可以分為 PM_COM、PM_DOT、PM_OTH 和 NON_PM 四種類型，從各個不同情況下的非零部份之 pause duration 分布情形長條圖看來，我們可以利用 Gamma distribution 去對各種情況之分布圖進行近似與模擬，三種環境之各長條圖與 Gamma distribution 分布近似曲線如下所示：



圖五-3　內場主播 pause duration 分布圖



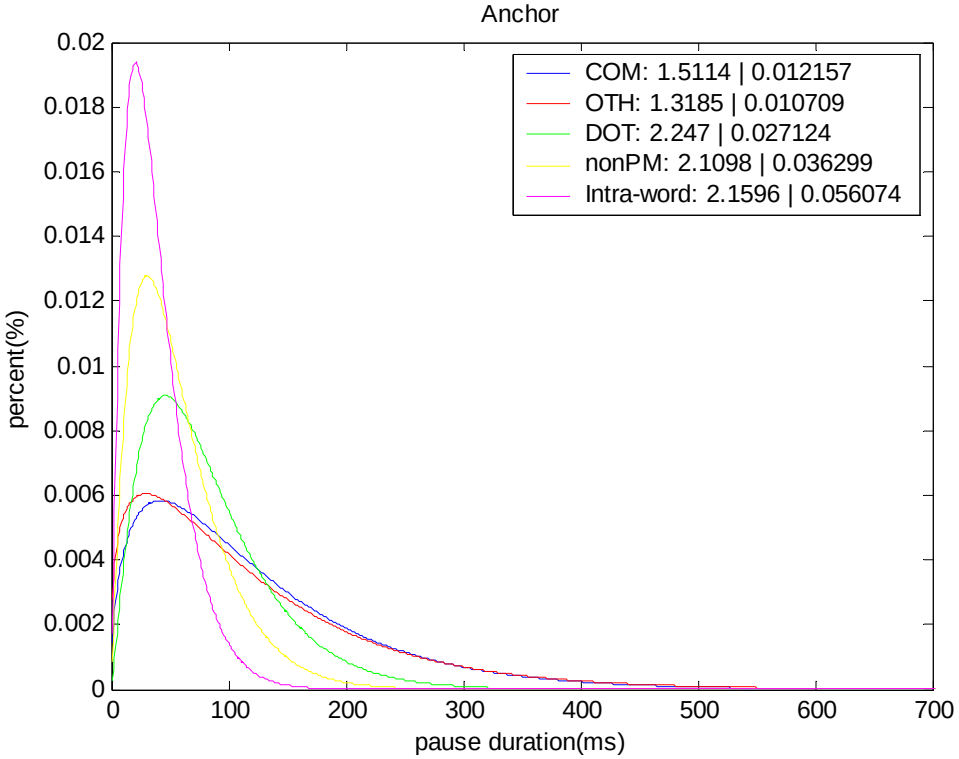
圖五-4　外場記者 pause duration 分布圖



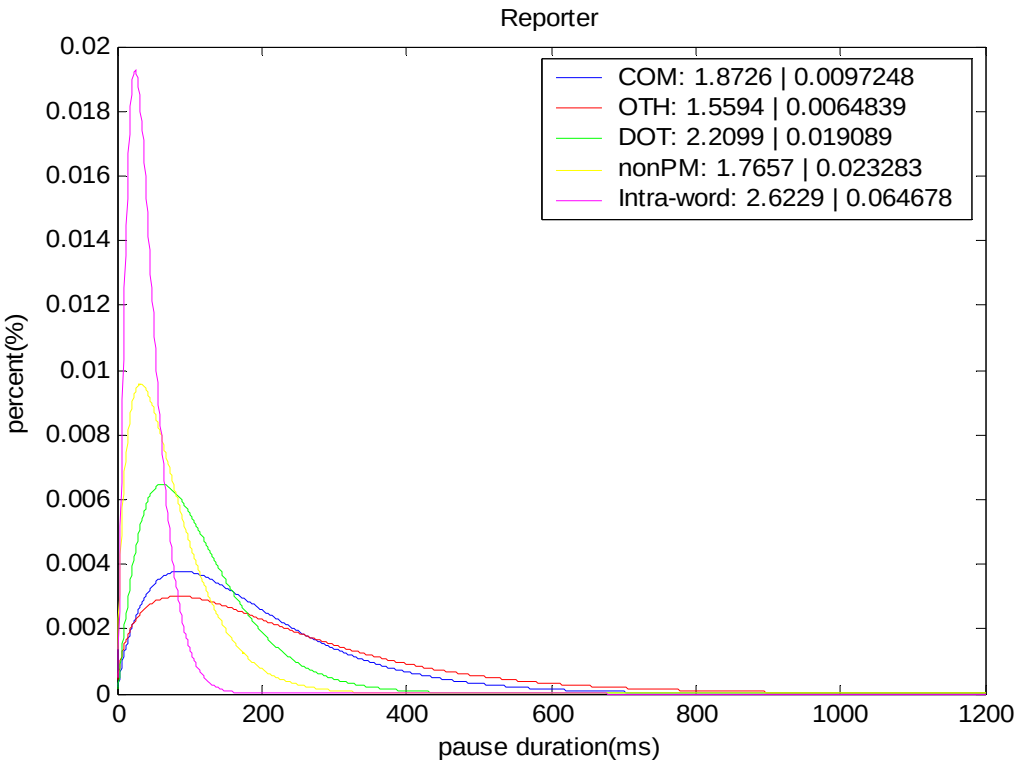
圖五-5　受訪者 pause duration 分布圖

依據以上的 pause duration 分布長條圖近似所得到的 Gamma distribution 圖形，各個語者環境均可從每個分布圖中得到一組參數（α 與 λ），利用這些參數，便可以在得到一段 pause 的長度之後，進一步計算出此段 pause 屬於各個 pause duration 模型的機率與 DM 分數（機率的詳細計算過程於下節中說明），如此便完成了 pause duration 模型的建立。

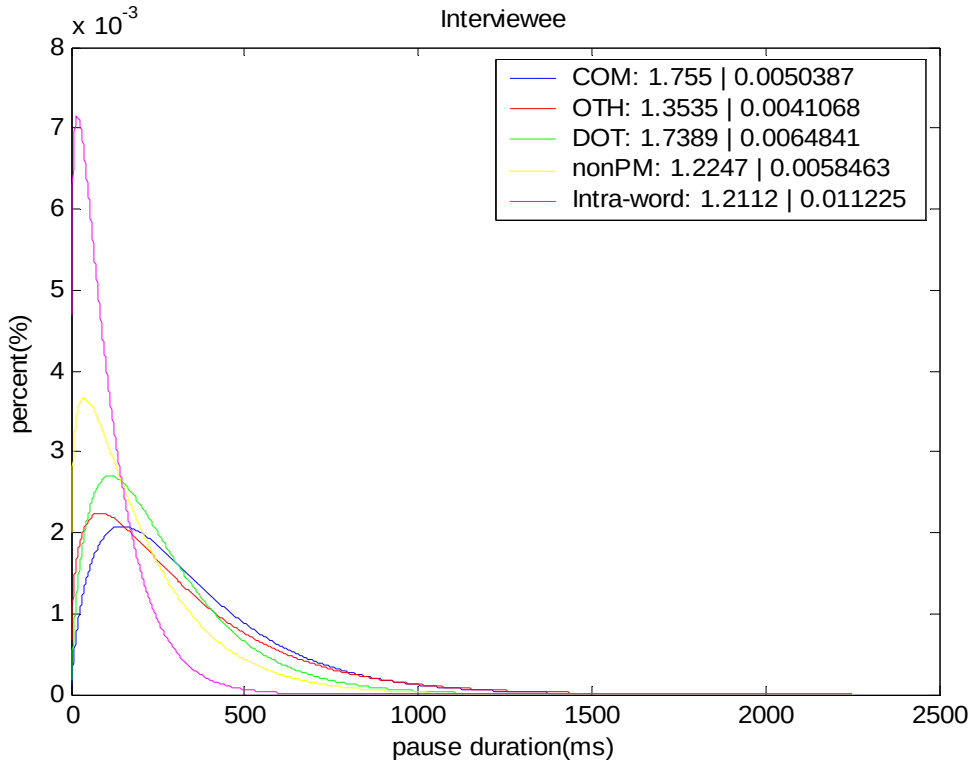
為了進一步觀察所要區分的五種情況之 pause duration 模型是否有鑑別度，接下來根據統計得到的各組參數，以三種語者環境區分，將五個分布圖畫在一起，並將各組音節間靜音長度模型參數（α|λ）同時顯示於圖型右上方，結果如下：



圖五-6　內場主播 pause Gamma duration 圖



圖五-7　外場記者 pause Gamma duration 圖



圖五-8 受訪者 pause Gamma duration 圖

- Pause duration model score 的計算

從統計的結果圖形中發現，pause duration 的分布可以用 Gamma distribution 近似，論文中所用 Gamma distribution [18]為：

$$f(x)=\frac{\lambda(\lambda x)^{\alpha-1}e^{-\lambda x}}{\Gamma(\alpha)} \qquad x>0,\ \alpha>0,\ \lambda>0 \tag{5.3}$$

式中 Γ(α)為 Gamma funciton：

$$\Gamma(\alpha)=\int_0^{\infty}x^{\alpha-1}e^{-x}dx \qquad \alpha>0 \tag{5.4}$$

且變數的 mean 跟 variance 和 α 與 λ 之間存在著如下關係：

$$E[X]=\alpha/\lambda \qquad Var[X]=\alpha/\lambda^2 \tag{5.5}$$

又因為 Gamma distribution 所模擬的均是 pause 長度大於零的部份，但從之前統計結果得知，每種情形均有 pause 長度為 0 的部份，而且佔有相當程度的比重，所以最後用來計算 pause duration model 分數的機率公式如下：

$$f_D(d)=\begin{cases} w & , d=0 \\ (1-w)\cdot f(d), & d>0 \end{cases} \tag{5.6}$$

上式中 w 為各個情形中 pause 長度為 0 所佔的比例，將（5.6）之機率公式取對數之後，便可以得到計算辨識率時各個情況下的靜音長度分數。

5.3.3 POS 模型的建立

- POS model score 公式推導

在此將對（5.2）公式中 POS model 分數部份的機率作進一步說明，因為過程中也將一類 PM 視為一種 POS，所以定義變數 *W'*表示詞與標點符號穿插組合而成的字串，因此

$$W'=w'^{N}_{1}=w'_1\ w'_2\ .....\ w'_N \tag{5.7}$$

對應的詞類串則為

$$POS=pos^{N}_{1}=pos_1\ pos_2\ .....\ pos_N \tag{5.8}$$

接下來，（5.2）式中 POS 模型分數進一步推導過程如下：

$$
\begin{aligned}
P(POS \mid W, P) &= P(POS \mid W') \\
&= P(pos_1 \mid w'^{N}_{1}) \prod_{i=2}^{N} P(pos_i \mid pos_1^{i-1}, w'^{N}_{1}) \\
&= P(pos_1 \mid w'_1) \prod_{i=2}^{N} P(pos_i \mid pos_{i-1}, w'_i)
\end{aligned}
\tag{5.9}
$$

在式（5.9）化簡過程中，必須經過合理的簡化來降低複雜度，我們假設詞串符合穩態（Stationary）馬可夫序列以有效減少需要的參數量。

接著再繼續對 $P(pos_i \mid pos_{i-1}, w'_i)$ 進行整理：

$$
\begin{aligned}
P(pos_i \mid pos_{i-1}, w'_i) &= \frac{P(pos_i, pos_{i-1}, w'_i)}{P(pos_{i-1}, w'_i)} = \frac{P(pos_i, pos_{i-1}, w'_i)}{P(pos_{i-1})P(w'_i)} \\
&= \frac{P(pos_i, w'_i \mid pos_{i-1})}{P(w'_i)} \\
&= \frac{P(w'_i \mid pos_{i-1}, pos_i)P(pos_i \mid pos_{i-1})}{P(w'_i)} \\
&= \frac{P(w'_i \mid pos_i)P(pos_i \mid pos_{i-1})}{P(w'_i)} \\
&= \frac{P(pos_i \mid w'_i)P(w'_i)P(pos_i \mid pos_{i-1})}{P(pos_i)P(w'_i)} \\
&= \frac{P(pos_i \mid w'_i)}{P(pos_i)} P(pos_i \mid pos_{i-1})
\end{aligned}
\tag{5.10}
$$

將以上結果代回（5.9）式，則 POS model 之機率可表示為

$$
P(POS \mid W, P) = P(pos_1 \mid w'_1) \prod_{i=2}^{N} \left[\frac{P(pos_i \mid w'_i)}{P(pos_i)} P(pos_i \mid pos_{i-1}) \right] \tag{5.11}
$$

從（5.11）之推導結果發現，計算 POS 模型分數所需要的資訊和建立 class-based 語言模型所能夠得到的結果資訊相同，分別是以 POS 為單元的 unigram、bigram 機率以及每個詞可能擁有不只一種的詞類，而它屬於可能

的各個詞類的機率。

- POS 語言模型的建立

為了得到計算 POS 模型分數所需的資訊，必須建立一個 POS 語言模型，在此，文字資料將採用中研院平衡語料庫五百萬詞之標有詞類斷詞結果，其中不只記錄著文章的斷詞結果，還額外保留有斷出的每個詞所屬的詞類，經過處理後便可以得到以詞類為單元所組成的文章句子，並進一步從中統計語言模型所需要的 unigram 和 bigram 機率。詞庫小組標記使用的標記種類詳見附錄三。

如今的辨識環境語料來自於廣播新聞，從第二章的語料特性統計結果得知，其中存在著許多一般 read speech 所沒有的現象，例如：particle、呼吸聲以及 garbage，又因為中研院平衡語料庫的文章中，並不會有這類型的特殊標記存在，所以，這部份的 unigram 與 bigram 機率必須經由其他的方式得到；在此，我們將 particle 視為一類，而它的 unigram 與 bigram 機率則個別採用其他所有的機率之平均值使用，至於呼吸聲和 garbage 則與外文標記歸為同一類，使用相同的機率值，又因為本章所進行的研究，希望加入標點符號所包含的訊息，所以也將三種標點符號各視為一種詞類使用；因此，除了中文原有的 46 種詞類，又另外加入了上述的詞類共 4 種，合計 50 種。

另外，我們還需要有六萬詞詞典內的每一個詞屬於某幾種詞類的機率，因為研究中使用的六萬詞詞典是從中研院八萬詞詞庫、交通大學語音實驗室自訂詞條與台師大詞典三者的聯集中取得，其中，中研院詞庫部份有正確的詞類標記之相關資訊，而自訂詞條也有給予標記，其中標記內容格式如下：

中文之 Big5 碼_漢語拼音	出現次數	詞類總數	第一種詞類	第二種詞類	第一種次數	第二種次數
A4C9A4D1_SHENG1TIAN1	19	2	36	40	19	0

圖五-9　詞類標記之相關資訊範例

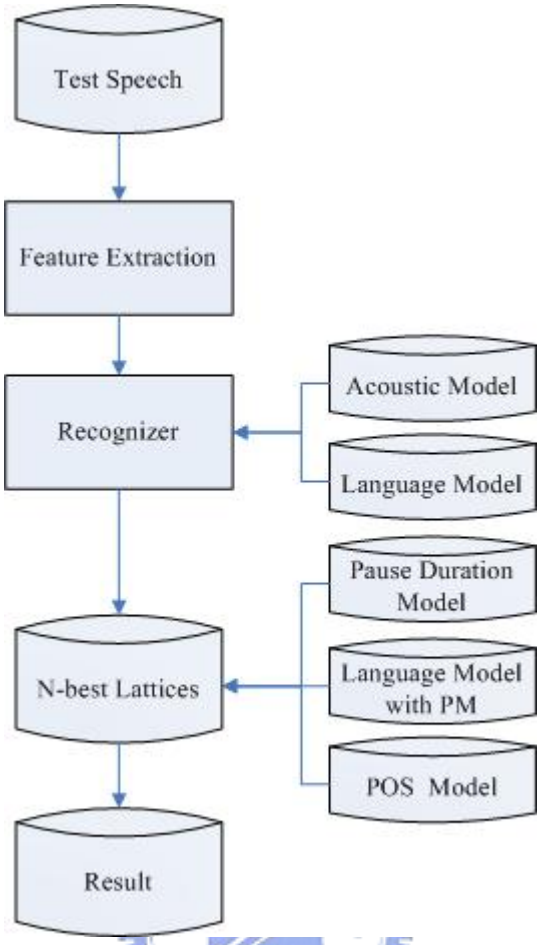
從上圖的例子便可以發現，因為其中的出現次數是由平衡語料庫五百萬詞之斷詞結果中統計而來，但是因為語料庫的內容詞數並不算多，所以會有出現次數為 0 的現象發生，會造成那個詞的其中一種詞類之機率為 0 的情況，而且並無法肯定辨識結果中這個詞不會有屬於這個詞類的情形發生，因此，我們利用 zero-order interpolation 的方法，將每個詞的每種可能詞類之出現次數都加上 1，以解決機率為 0 所造成的問題。

經由統計後發現，六萬詞詞典中完全沒有詞類標記資訊的部份所佔的比例不到百分之三，由於所佔的比例不多，這部份便均統一給予詞類標記為普通名詞使用。

5.3.4 Rescore 方法與流程

- 模型使用流程

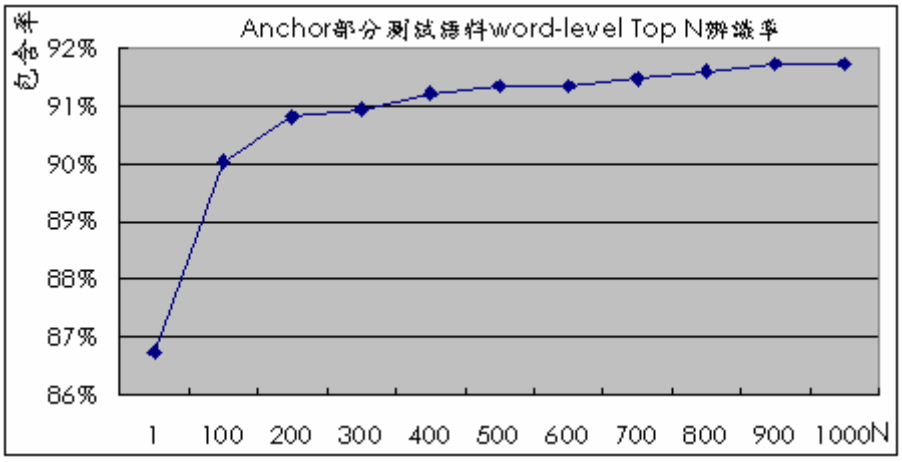
配合工具HTK的使用，我們將採用如同第四章中利用trigram語言模型時的two-pass rescoring方式，使辨識系統能夠利用到（5.2）中的四項分數進行辨識，也就是一開始先利用之前章節所建立的聲學模型和語言模型讓HTK對測試語料進行辨識，而在辨識的同時保留數組比較好的結果提供給下一階段，再配合加入PM的語言模型、inter-word（又細分為PM_COM、PM_DOT、PM_OTH與NON_PM）和intra-word pause duration模型以及POS模型，找出最好的一組辨識結果。



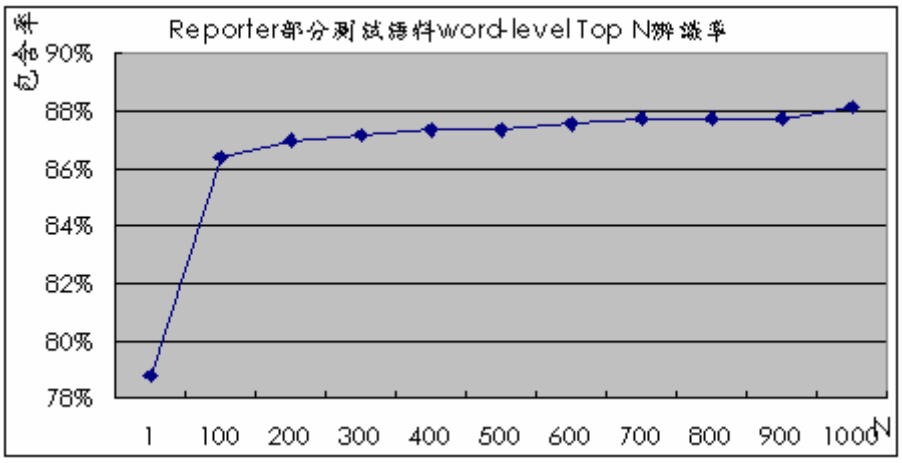
圖五-10　Two-pass rescore 流程方塊圖

- State 之 token 數的選擇

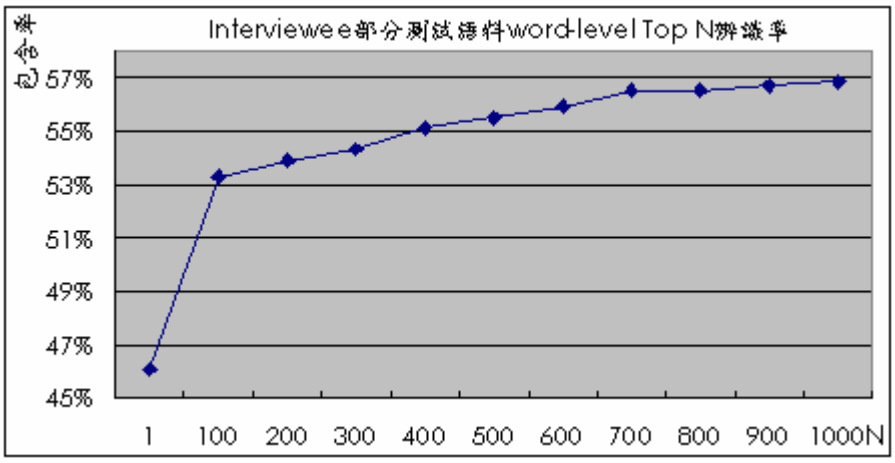
在此也將遇到利用 trigram 語言模型時相同的問題，就是 state 所需保留 token 數的選擇，期望達到的目標依舊是希望能夠在每個 state 只保留越少越好的 token 數量，但又能夠在其所有的輸出答案組合中包含越高越好的辨識率（包含率）。在此仍採用測試語料的十分之一，作為選擇 token 數的實驗語料，並觀察所選擇的 token 數是否能夠提供足夠的辨識率成長空間給 rescoring 時使用。三種語者環境在 token 數設為 10 的情況下辨識包含率為：



圖五-11　內場主播 10-best 詞辨識包含率



圖五-12　外場記者 10-best 詞辨識包含率



圖五-13　受訪者 10-best 詞辨識包含率

由以上實驗結果之圖形中可以看出，三種環境之包含率曲線均有類似的上升情形，當 token 設定為 10，辨識結果組合數到達 1000 時，便已經提供了一段不小的辨識率進步空間給下一階段 rescoring 時使用。

5.4 實驗一—利用標點符號、pause duration 資訊辨識效能

本實驗中所使用的聲學模型，將利用第四章之考慮破音字後依據不同語者環境所建立的三個聲學模型，所使用的測試語料則仍和之前章節相同，而 syllable-level 測試語料的正確答案，則是採用和第四章中實驗時相同，因考慮了破音字而重新選擇過後較為正確的答案內容。

接下來的實驗過程中，首先將利用第四章所建立的 adapted bigram 語言模型，配合聲學模型進行第一階段的辨識，並保留 10-best lattice 的辨識結果，過程中為了加快 Viterbi search 以提升辨識速度，都有使用 beam search。接下來先同時採用加入 PM 後的語言模型與 pause duration 模型的資訊，進行 rescoring 的工作，並產生一組最好的辨識結果，然後再加入詞類模型，最後，我們還將顯示標點符號的標示結果，並對於標點符號自動標示的效能做個分析與評估。

5.4.1 實驗結果

在此依據之前所述 two-pass rescoring 流程進行辨識，第二階段 rescoring 過程中，除了語言模型之外，還另外加入了 pause duration 模型，進行辨識時所產生的結果便將同時考慮聲學模型、語言模型以及 pause duration 模型的分數，因此除了語言模型的分數比重外，還必須選擇一個 pause duration 模型分數比重（DM Weight），以決定 pause duration 模型對最後辨識結果造成的影響程度，經過實驗中測試發現，三種語者環境之 inter-syllable pause duration model weight 均是在選擇 0.4 時能夠有最好的辨識結果，個別的三個層級 411 音節辨識結果詳列如下：

表五-3　Outside 測試語料 word 辨識率

環境	Sub	Del	Ins	Accuracy
內場主播	8.96%	3.21%	0.97%	86.86%
外場記者	16.98%	4.95%	1.70%	76.36%
受訪者	36.61%	11.57%	3.37%	48.44%

表五-4　Outside 測試語料 character 辨識率

環境	Sub	Del	Ins	Accuracy
內場主播	5.96%	2.52%	0.15%	91.37%
外場記者	13.14%	2.67%	0.31%	83.88%
受訪者	31.44%	8.51%	1.72%	58.33%

表五-5　Outside 測試語料 syllable 辨識率

環境	Sub	Del	Ins	Accuracy
內場主播	3.53%	2.52%	0.15%	93.80%
外場記者	8.19%	2.62%	0.27%	88.92%
受訪者	24.14%	8.77%	1.98%	65.11%

5.4.2　標點符號自動標識結果

在標點符號自動標識的效能評估過程中，所有的錯誤類型有三種，第一個是正確答案中有 PM 存在的位置，但是卻沒有 PM 被辨識出來的刪除型錯誤（*Del*），第二種則是答案中並沒有 PM 出現的位置，而辨識結果有 PM 存在的插入型錯誤（*Ins*），另外，還有一種則是在答案中有 PM 存在的位置辨識器也辨識出了 PM 的存在，不過卻發生了類型上的辨識錯誤，這類型錯誤則稱之為替代型錯誤

（*Sub*），但是因為標點符號標記的類型原本就可能有不只一種的選擇，因此，接下來所定義的項目將根據上述三類錯誤為基礎進行計算，作為辨識器標記標點符號效能優劣的衡量標準，各項所代表的意義與計算公式接下來有進一步的說明。

首先，定義只考慮刪除型與插入型錯誤，而並不將 PM 的種類選擇上的錯誤列入計算之錯誤總數量，記為 *Err_notype*，而定義 *Total* 為正確答案中有標記之標點符號的總量，個別公式如下 [17]：

$$Err\_notype = Del + Ins \quad ; \quad Total = Corr + Del + Sub \tag{5.7}$$

接著以 *Place Corr, total* 表示 PM 標記位置正確，而忽略辨識器所標記的 PM 種類是否正確，這類型的結果在正確標記之 PM 中所佔的比例：

$$Place\ Corr,\ total = \left((Corr + Sub)/Total\right) \times 100\% \tag{5.8}$$

另外，*Place Corr, type Corr* 則是計算 PM 的標記位置與種類均正確的比例：

$$Place\ Corr,\ type\ Corr = \left(Corr/Total\right) \times 100\% \tag{5.9}$$

接下來的 *Place Corr, type Err* 是標記的 PM 位置正確但種類錯誤的部份：

$$Place\ Corr,\ type\ Err = \left(Sub/Total\right) \times 100\% \tag{5.10}$$

最後，則是刪除型錯誤與插入型錯誤這兩大主要錯誤，在正確答案中所有的 PM 數量中所佔有的比例，公式如下：

$$Place\ Err = \left(Err\_notype/Total\right) \times 100\% \tag{5.11}$$

下列表格，分別為三種語者環境之測試語料標點符號辨識結果，以及為了觀

察 PM 間的相互辨識情形，建立之個別 confusion table：

表五-6　Outside 測試語料標點符號辨識率

環境	Place Corr , total	Place Corr , type Corr	Place Corr , type Err	Place Err
內場主播	78.93%	67.88%	11.05%	35.02%
外場記者	83.99%	77.11%	6.88%	35.95%
受訪者	67.91%	66.18%	1.73%	56.46%

表五-7　內場主播標點符號標記之 confusion table

辨識結果 / 正確答案	PM_COM	PM_OTH	PM_DOT
PM_COM	96.14%	3.31%	0.55%
PM_OTH	31.99%	68.01%	0.00%
PM_DOT	100.00%	0.00%	0.00%

表五-8　外場記者標點符號標記之 confusion table

辨識結果 / 正確答案	PM_COM	PM_OTH	PM_DOT
PM_COM	98.21%	1.49%	0.30%
PM_OTH	11.34%	88.66%	0.00%
PM_DOT	93.75%	0.00%	6.25%

表五-9　受訪者標點符號標記之 confusion table

辨識結果 / 正確答案	PM_COM	PM_OTH	PM_DOT
PM_COM	99.37%	0.63%	0.00%
PM_OTH	4.66%	95.34%	0.00%
PM_DOT	50.00%	50.00%	0.00%

5.4.3　實驗分析

在此，一開始先觀察加入標點符號與 pause duration 模型後對辨識結果的影響，以下將第四章考慮破音字並使用調適後語言模型，以及本章中又加入標點符號與 pause duration 模型的三種語者環境各層級辨識率，統整結果於下表中，同時計算出個別的 error reduction rate（ERR）：

表五-10　三種語者環境各層級辨識結果比較表

語者環境	辨識條件	Syllable-level	Character-level	Word-level
內場主播	同第四章設定	93.62%	91.04%	86.29%
	加入 PM & DM	93.80%	91.37%	86.86%
	ERR	2.82%	3.68%	4.16%
外場記者	同第四章設定	88.70%	83.37%	75.59%
	加入 PM & DM	88.92%	83.88%	76.36%
	ERR	1.95%	3.07%	3.15%
受訪者	同第四章設定	64.82%	57.99%	47.84%
	加入 PM & DM	65.11%	58.33%	48.44%
	ERR	0.82%	0.81%	1.15%

- 從表中可看出，在辨識系統中加入 PM 與 DM 的資訊之後，三種語者環境的各層級辨識率均有小幅度的提升，由此可知，這兩者含有一般使用的聲學模型和語言模型以外的訊息，而確實能夠對辨識器的效能有所助益。

- 藉由比較上表中的 ERR 數值，可發現三種語者環境共有的一個現象，便是 word-level 的 error reduction rate 均較另外兩個層級的 ERR 值要高出一些，這是因為標點符號和 pause duration 模型的主要功用是，改變詞的連接機率以及提升 word boundary 位置判斷的正確性，所以雖然也能夠使另外兩個層級的辨識率有所的改善，但是對於 word-level 辨識結果將有較為顯著的影響。

- 由表五-6 中可知，標點符號 *Place OK, total* 的標示結果從高到低為，外場記者、內場主播、最後是受訪者。在主播和記者的部份，雖然加入 PM 的語言模型對於兩者都有幫助，但主播的 PM 部分 pause 停頓現象不明顯、比例較低，而且根據圖五-6 和圖五-7 的 pause duration model 圖形，相較之下主播的各個 DM 間鑑別度較差，所以 PM 的標示結果不如外場記者優良；另外，受訪者則因為一開始的詞辨識率欠佳，所以標示 PM 的整體效能也最差。

接下來，進一步觀察標點符號自動標示的結果，首先將 *Place Err* 的結果再細分為 miss detection 和 false alarm 兩種，結果如下：

表五-11　標點符號標記之 miss detection 與 false alarm

語者環境	Place Err		
	Miss Detection	False Alarm	Total
內場主播	21.07%	13.95%	35.02%
外場記者	16.01%	19.94%	35.95%
受訪者	32.09%	24.37%	56.46%

- 上表中內場主播和外場記者的部份，可以發現記者的 miss detection 發生機率比主播低，這是因為表五-2 的結果指出，這兩種語者環境之三種標點符號的停頓機率都是記者的比例較高，所以較不會發生 miss detection 的現象。另外，在 false alarm 的部份，因為兩種環境在沒有 PM 的位置有停頓發生的情況都很少，所以主要會由加入 PM 的語言模型分數來決定是否有 PM 存在，而從之前章節的實驗結果中發現，調適後的語言模型對於主播的辨識效能提升最為顯著，所以 false alarm 的出現機率是內場主播較小。至於受訪者的部份，因為一開始的詞辨識結果正確率不足，無法發揮 LM 應有的效能，所以即使在三種 PM 停頓的比例是三個環境中最高的，但是兩種錯誤發生的機率仍然最高。

5.5　實驗二—再加入詞類資訊後辨識效能

在此，實驗設定和所使用的測試語料均與實驗一中相同，但是除了使用加入 PM 後的語言模型與 pause duration 模型外，再加入 POS 模型的資訊，進行 rescoring 得到辨識結果，並觀察標點符號自動標示的效能變化。

5.5.1　實驗結果

在此辨識流程仍需利用 two-pass rescoring 流程，進行辨識時所產生的結果是

同時考慮聲學模型、語言模型、pause duration 模型和 POS 模型分數的產物，下表中記錄著三種不同語者環境在給定不同的 POS 模型比重時， word-level 辨識結果之變化情形：

表五-12　Outside 測試語料 word 辨識率隨詞類模型比重變化情形

環境	0.0	0.2	0.4	0.6	0.8
內場主播	88.86%	86.68%	86.60%	86.57%	86.47%
外場記者	76.36%	76.30%	76.08%	76.00%	75.83%
受訪者	48.44%	47.48%	47.36%	47.30%	47.15%

5.5.2　標點符號自動標識結果

下列表格中，分別記錄了三種語者環境在詞類模型比重等於 0.8 時，測試語料標點符號自動標記所得到的結果：

表五-13　Outside 測試語料標點符號辨識率

環境	Place Corr , total	Place Corr , type OK	Place Corr , type Err	Place Err
內場主播	80.71%	64.61%	16.10%	48.13%
外場記者	85.53%	75.56%	9.97%	44.24%
受訪者	69.77%	65.38%	4.39%	73.24%

表五-14　標點符號標記之 miss detection 與 false alarm

語者環境	Place Err		
	Miss Detection	False Alarm	Total
內場主播	19.28%	28.85%	48.13%
外場記者	14.47%	29.78%	44.24%
受訪者	30.23%	43.01%	73.24%

5.5.3　實驗分析

根據以上實驗結果，可以得到以下幾個結論：

- 從表五-12 中可發現，隨著辨識系統中給予詞類模型的比重越大時，辨識率反而有逐漸下降的趨勢，而造成 POS 對辨識器效能有危害可能的原因有三，首先可能是因為在（5.1）式之推導化簡過程中，pause duration 與詞類無關的假設並不恰當；其次是實驗中建立語言模型和 POS 模型時，所使用的文字資料庫並非同一個來源，而這種資料庫不匹配（Miss Match）的情形，可能會造成新加入的資訊無法發揮應有的效能；第三個原因則是決定每個詞所屬的詞類時，雖然六萬詞詞典中大多數都可以從中研院詞典中找到正確的詞類，以及計算相關機率時所需的資訊，但是其他部分的詞類給定則不盡正確，而對辨識結果也將造成一定程度的影響。

- 藉由比較表五-6、表五-13 以及表五-11 和表五-14 可發現，雖然標點符號自動標示所得到的結果正確率從高到低仍然為，外場記者、內場主播、最後是受訪者，但是可以發現 POS 模型對於標點符號標示所造成的影響，會使 PM 種類的判定較不正確，此外，即使 miss detection 的發生機率有小幅度的下降，但是 false alarm 的機率在三種語者環境下都有大幅度的上昇，而這也是造成標點符號標示效能降低的最主要原因。

第六章　結論與未來發展

6.1 結論

在本論文中，我們使用廣播新聞資料庫 MATBN 進行廣播新聞語音辨識的相關研究，從語料庫的聲音的特性、建立基本辨識系統，到語言模型以及標點符號、pause duration 模型和 POS 模型的加入，有一個循序漸進的說明。在此，我們將幾個主要重點分列如下：

（1） 廣播新聞語音與一般 read speech 語音不同，特性比較傾向於自然語音（Spontaneous Speech），其中存在著一般文字之外的某些自然語音現象，因此除了中文 411 音節模型外，我們還另外建立了 particle、呼吸聲和 garbage model 等聲音模型，使得基本辨識系統更完整。

（2） 廣播新聞語料的特色之一是擁有不同的內外場語者環境，因此依照不同的環境個別建立聲學模型是提高辨識率的必要條件，研究中依據不同環境的語料各自訓練聲學模型，確實可以使辨識器有較好的辨識效能。

（3） 語言模型是完整的語音辨識系統不可或缺的一部分，若能加入語言模型，而非僅採用無文法規則（Free grammar），可使得辨識器更完善，我們在此針對廣播新聞特性，進行語言模型的調適，從音節辨識結果的變化情形來看，加入語言模型真正能有效的提高辨識率；另外，若能夠將破音字加入考慮、給予適當的處理，也確實能夠得到較正確的結果而使得音節辨識效能有進一步的改善。

（4） 一般的辨識器，並不會將文章中的標點符號對辨識系統的影響列入考慮，但是標點符號在句中所扮演的腳色，對於語句的文法結構應有其重要性存

在，若能夠再配合音節間靜音長度模型共同使用，根據研究中詞辨識結果的改變，以上兩者所包含的資訊確實能夠使詞辨識率有所提升；此外，例如詞類這種更高層的文法相關訊息，雖然實驗中無法得到預期的結果，但假使有完整、精確的相關資料，應該也可以對辨識結果有所助益。

6.2 未來展望

國內外的許多辨識系統，均已經採用前後文相關模型（Context Dependent Model）列為基本的條件，但是實驗中由於語言模型模型的加入，已經使得運算量變的相當龐大，辨識工作的進行因而相當費時，因此就沒有再嘗試前後文相關模型的實驗。假若在未來的研究中，我們希望可以再把部分加入辨識系統。

廣播新聞節目中，語者講話的同時經常會伴隨著背景聲（Background Sound）的存在，而本論文所進行的研究均僅針對無背景聲的部份，至於有背景聲的語音辨識至今仍是一大挑戰，但是考慮背景聲的語音辨識是必要的，也可以針對語者語音參數調適和聲學模型的調適去加強辨識系統在有背景聲存在時的效能，這也是未來一個相當不錯的研究主題。

研究中，我們採用了語料中音節間的 pause duration 來幫助辨識系統在標點符號標示與 word boundary 位置的判定，未來的研究中若能夠再將音調（Tone）的影響也加入辨識系統，建立出 411 音節的 duration model，相信對於辨識系統的效能也能有相當程度的幫助。

參考文獻

[1] B. H. Juang and S. Furui, “Automatic recognition and understanding of spoken language – A first step towards natural human-machine communication,” in Proc. IEEE, 88, 8, pp. 1142-1165, 2000

[2] L. R. Rabiner and B. H. Juang, “Fundamental of Speech Recognition,” New Jersey, Prentice-Hall, Inc., 1993

[3] 陳俊良，“國語廣播新聞語音辨識之研究”， 國立交通大學電信工程學系碩士論文，民國九十三年七月

[4] Hsin-Min Wang, “MATBN 2002: A Mandarin Chinese Broadcast News Corpus” ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003)

[5] C. Barras, E. Geoffrois, Z. B. Wu, M. Liberman, “Transcriber: Development and Use of S tool for Assisting Speech Corpora Production,” Speech Communication, 33, pp. 5-22, 2001

[6] Liu, D., L. Nguyen, S. Matsoukas, J. Davenport, F. Kubala, R. Schwartz, “Improvements in Spontaneous Speech Recognition,” DARPA 1998 Broadcast News Transcription and Understanding Workshop, Leesburg VA, Feb. 1998

[7] S. Young, G.. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, “The HTK Book（for HTK Version 3.2.1）”

[8] Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, “Spoken Language Processing, Aguide to Theory, Algorithm, and System Development,” Prentice-Hall, Inc

[9] Kazuyuki TAKAGI, Shuichi ITAHASHI, “Segmentation of Spoken Dialogue by Interjections, Disfluent Utterances and Pauses.” In Proceedings of the ICSLP-96, pp. 697-700

[10] G. Riccardi, E. Bocchieri, and R. Pieraccini. “Non-deterministic stochastic language models for speech recognition.” In Proceedings IEE International Conference on Acoustics, Speech and Signal Processing, volume 1, pages 237-240. IEEE, 1995

[11] Slava M. Katz, “Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer,” IEEE Transactions on Acoustic, Speech and Signal Processing, Vol. ASSP-35, NO. 3, MARCH 1987

[12] 江振宇，“中文段詞器之改進”，國立交通大學電信工程學系碩士論文，民國九十三年七月

[13] H. Meinedo, N. Souto, and J. Neto, “Speech recognition of broadcast news for the european portuguese language,” in Proc. ASRU '2001

[14] 吳季芳，“表列國語一字多音”，文化出版社，民國九十二年三月

[15] Jachym Kolar, Jan Svec, and Josef Psutka, “Automatic Punctuation Annotation in Czech Broadcast News Speech.” In SPECOM-2004, pp. 319-325

[16] Ji-Hwan Kim and P. C. Woodland, “The use of prosody in a combined system for punctuation generation and speech recognition,” in Proc. Eurospeech, page 2757-2760, 2001

[17] C. Chen, “Speech Recognition with Automatic Punctuation,” in Proc. Eurospeech, page 447-450, 1999

[18] Alberto Leon-Garcia, “Probability and Random Processes for Electrical Engineering,” Addison-Wesley Publishing Company, no. 2, pp. 102-119, 1994

附錄一

Background sound 標記方法		
種類	**標記**	**說明**
Music	<BACK_Music> … </BACK_Music>	純音樂
Speech	<BACK_Speech> … </BACK_Speech>	可以聽清楚的人聲
Shh	<BACK_Shh> … </BACK_Shh>	機器聲
Other	<BACK_Other> … </BACK_Other>	噪音，如交通工具的聲音（呼嘯聲、喇叭聲、 警鈴聲）、喧雜的人聲、電子通訊器材發出的干擾聲，以及任何沒有意義的聲響

Noise 標記方法		
種類	**標記**	**說明**
advertisement	<ADV/>	廣告
breath	<BRE/>	喘息聲（含呼吸聲、呼氣聲、吐氣聲）
clear throat	<NOISE/>	清喉嚨聲
click	<NOISE/>	漬舌聲
cough	<NOISE/>	咳嗽聲
cry	<NOISE/>	哭聲
empty	不處理	DAT 轉錄製 PC 時，因無法同時作業而產生的 0 值 Sample，前後各約有數秒的時間（一般只會出現在每個音檔的頭尾而已）
hiccup	<NOISE/>	打嗝聲
laugh	<NOISE/>	笑聲
particle	<PARTICLE> … </PARTICLE>	沒有標準語意的語氣詞
pause	<PAUSE/>	停頓
sign	<NOISE/>	嘆氣聲
silence	<SILENCE/>	沉默
smack	<NOISE/>	砸嘴聲

sneeze	<NOISE/>	噴嚏聲
sniffle	<NOISE/>	吸鼻音
swallow	<NOISE/>	吞口水聲
trill	<NOISE/>	顫音
unrecognizable non-speech sound	<UNRECOGNIZED> … </UNRECOGNIZED>	由人發出非語音且無法辨識的聲音
weather broadcast	<WEATHER/>	氣象預報
yawn	<NOISE/>	哈欠聲
noise	<NOISE/>	其他無法判定的雜音（補充）
inhale	<NOISE/>	吸氣聲（補充）
langthening	不處理	拉長聲（補充）
short break	<PAUSE/>	pause（補充）

Pronounce error 標記方法		
種類	**標記**	**說明**
Inappropriate Pronunciation	發(hua1)生	發音有偏差但仍能辨識的字詞（常見），判斷其拼音是否存在於漢語拼音中，若存在則使用新的拼音
Stutter	<STUTTER> … </STUTTER>	口吃，一直重複某個字或其部分的音，如「對對對」
Syllable contraction	<SYLLABLE_CONTRACTION> … </SYLLABLE_CONTRACTION>	說話太快而出現音節合併的現象（常見），如「這樣子」變成「降子」
Uncertain	<UNCERTAIN> … </UNCERTAIN>	無法確定的字詞，但是當一連串念了一句以後就可以辨別是什麼的字詞
Unrecognizable Speech sound	<UNRECOGNIZED> … </UNRECOGNIZED>	無法辨識的字詞，如方言
Alternative	不處理	尚未被收錄在辭典但被廣為使用之讀音
Zhuyin	不處理	注音符號（非常少用）

Foreign Language 標記方法		
種類	**標記**	**說明**
English	<ENG> … </ENG>	英文
Min-Nan	<MinNan> … </MinNan>	閩南語
Japanese	<JPN> … </JPN>	日語
Formosan	<Formosan> … </Formosan>	原住民語
Hakka	<Hakka> … </Hakka>	客家語
Cantonese	<Foreign> … </Foreign>	廣東語
Other	<Foreign> … </Foreign>	其他所有語言，如拉丁語，法語，阿拉伯語等

附錄二

中文 510 個一字詞破音字												
了	杓	咽	射	得	著	膀	壎	蘗	絜	鯷	免	泊
亡	沁	咯	射	從	著	蓋	戲	躍	腊	鶣	刨	玟
勺	沈	咱	屑	殺	著	裳	檐	露	苛	鶖	吭	的
大	沒	咻	差	液	蛤	裨	濘	鶴	棻	儹	否	芾
已	角	契	差	涸	貴	裨	禪	讀	嗄	鶇	告	軋
什	谷	契	挾	率	軸	說	縮	攪	骰	攢	吱	長
仇	足	姥	挾	畦	閒	說	繆	鑰	稗	蠶	囤	阿
屯	身	屏	旁	盛	馮	閤	繆	鑿	筴	饡	坊	陂
氏	車	度	晁	祭	傳	麼	臂	乜	腄	饡	圾	亟
爪	那	怎	校	紮	剿	麼	薄	匸	萎	碁	坏	便
仔	那	恫	桔	脯	勦	價	薜	万	萹	橦	坏	俊
卡	其	恪	桁	莎	嗑	噓	蟀	卬	蜎	毯	尾	冠
召	卒	扁	殷	莞	塞	墮	螯	爿	跩	膴	彷	削
句	卷	拽	烙	莘	廈	彈	褶	王	鉈	蜛	忸	勁
扒	呱	括	畜	莆	搽	徵	豁	伶	鉈	褟	忪	咳
白	呱	拾	砟	蛇	會	數	谿	价	嗂	鴟	更	咳
石	和	查	祕	蚵	楛	暫	賺	圴	嵺	櫸	疴	繫
互	和	洽	秤	蚵	滑	暴	賺	旮	摽	謏	茎	藥
吃	呢	洸	秘	蛆	溪	模	輾	体	撂	謕	蛉	藷
囝	咋	洮	耙	被	溴	樂	還	吽	滊	謚	埻	識
地	奇	甚	脈	訥	煖	樂	擷	妡	綮	麉	滬	鵲
宅	姊	省	臭	許	瑁	潰	曜	尨	跟	瀧	碗	嚼
弛	宛	矜	般	都	稟	熟	檻	侅	噁	矉	第	礦

扛	宛	衹	衰	釭	粳	熨	璿	坻	敶	羆	祀	耀
朴	居	紅	郝	陸	腳	磅	癘	炔	夔	騔	袚	蘋
艮	怯	耑	釙	雀	落	稽	瞿	吡	瀟	蓋	袚	蠕
艾	拓	胖	釙	傀	落	調	藏	呴	瞉	遍	酞	覺
虫	拗	茄	鬲	喀	葉	誰	藉	枹	蓼	酪	堨	躂
血	拗	郃	乾	喳	蛻	賜	蹣	炵	蝷	鈸	堨	馨
行	杳	郃	偺	單	解	輥	轍	罘	蝎	雋	椑	鰓
佛	杷	重	偕	單	貉	頡	釐	凊	禈	馱	猢	屬
伽	枝	降	勒	孱	賈	頫	騎	悢	遫	僥	猲	攜
伺	泌	食	匙	尋	賃	噱	鵠	挐	圜	屜	給	綸
余	波	食	區	廁	辟	噢	瀕	揤	嬛	強	腌	翟
乘	參	員	啪	哦	圈	夏	埤	娠	尉	家	曾	窘
倆	參	哪	啞	埔	垔	奘	婁	娠	將	學	期	絡
倘	參	哪	啁	埋	堆	娜	宿	娩	惡	彊	朝	榦
龜	漸	嗾	碩	親	湮	穌	殼	歙	椎	橇	棹	漆
嚇	漯	摘	稱	遺	湯	蕈	氯	澠	殼	橈	棲	龜
駭	靦	錯										

附錄三

中文的 46 類詞類標記					
編號	標記	詞類	編號	標記	詞類
1	A	非謂形容詞	24	Nh	代名詞
2	Caa	對等連接詞	25	I	感嘆詞
3	Cab	連接詞，如：等等	26	P	介詞
4	Cba	連接詞，如：的話	27	T	語助詞
5	Cbb	關聯連接詞	28	VA	動作不及物動詞
6	Da	數量副詞	29	VAC	動作使動動詞
7	Dfa	動詞前程度副詞	30	VB	動作類及物動詞
8	Dfb	動詞後程度副詞	31	VC	動作及物動詞
9	Di	時態標記	32	VCL	動作接地方賓語動詞
10	Dk	句副詞	33	VD	雙賓動詞
11	D	副詞	34	VE	動作句賓動詞
12	Na	普通名詞	35	VF	動作謂賓動詞
13	Nb	專有名詞	36	VG	分類動詞
14	Nc	地方詞	37	VH	狀態不及物動詞
15	Ncd	位置詞	38	VHC	狀態使動動詞
16	Nd	時間詞	39	VI	狀態類及物動詞
17	Neu	數詞定詞	40	VJ	狀態及物動詞
18	Nes	特指定詞	41	VK	狀態句賓動詞
19	Nep	指代定詞	42	VL	狀態謂賓動詞
20	Neqa	數量定詞	43	V_2	有

21	Neqb	後置數量定詞	44	DE	的，之，得，地
22	Nf	量詞	45	SHI	是
23	Ng	後置詞	46	FW	外文標記

標點符號及其他詞類標記		
編號	**標記**	**說明**
47	DASHCATEGORY	—
48	ETCCATEGORY	…
49	COMMACATEGORY	，
50	PERIODCATEGORY	。
51	QUESTIONCATEGORY	？
52	COLONCATEGORY	：
53	SEMICOLONCATEGORY	；
54	EXCLANATIONCATEGORY	！
55	PARENTHESISCATEGORY	「」( )【】
56	PAUSECATEGORY	、
57	SPCHANGECATEGORY	‖
58	DM	定量複合詞
100	BM	附著語素