國　立　交　通　大　學

電信工程學系碩士班

碩士論文

國語廣播新聞語音辨識之研究

A Study on Mandarin Broadcast News Speech Recognition

研 究 生： 陳俊良

指導教授： 陳信宏 博士

中華民國九十三年七月

國語廣播新聞語音辨識之研究

A Study on Mandarin Broadcast News Speech Recognition

研 究 生：陳俊良　　Student： Chun-Liang Chen

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

國立交通大學

電信工程學系碩士班

碩士論文

A Thesis
Submitted to Department of Communication Engineering
College of Electrical Engineering and Computer Science
National Chiao Tung University
In Partial Fulfillment of the Requirements
For the Degree of
Master of Science
In
Electrical Engineering
June, 2004
Hsinchu, Taiwan, Republic of China

中華民國九十三年七月

國語廣播新聞語音辨識之研究

研究生：陳俊良　　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

廣播新聞在現代生活裡非常普遍，結合語音辨識的技術可以應用在許多領域上。在本論文中，我們專注於廣播新聞語音辨識，首先針對廣播新聞類似自發性語音的特性，說明如何建立聲學模型，並且根據內外場環境的不同各自訓練模型，可以提高辨識效能。接著還會說明語言模型的建立，如何將一個語言模型調整成適合廣播新聞，並且使用語言模型調適的技術可以使得語言模型優化，最後的音節辨識率與基本辨識系統相比，內場可提升約 16%、外場可提升 20%以上。而廣播新聞的資料來源是採用中研院錄製的公視新聞資料庫（PTSND），我們也會在論文中對資料庫內容做簡介說明。

A Study on Mandarin Broadcast News Speech Recognition

Student： Chun-Liang Chen Advisor： Dr. Sin-Horng Chen

Department of Communication Engineering
National Chiao Tung University

Abstract

Broadcast news is very general in this day, many researches were applied by combining speech recognition technique. In the thesis, we focus on the Mandarin broadcast news speech recognition, because of the spontaneous speech characteristics, several spontaneous acoustic models were constructed, and individual model training were performed under different environments. Then, the language model construction will be described carefully. Performance will be improved by using the language model adaptation. At last, the syllable recognition rate was increased about 16% for anchor. Above 20% increase was obtained for reporter and interviewee. All related experiments proceeded over the broadcast news database：PTSND. Database content will also be introduced in this paper.

誌謝

研究所兩年的時間轉眼就過，尤其在最後半年內的時間才漸漸進入狀況，而密集的學習是論文能完成的關鍵。很感激陳信宏老師在語音大方向上的指引，還有王逸如老師常常給予的震撼教育，工作的態度決定做事的結果，老師的教誨我銘記在心。

而我們語音實驗室真的是人才輩出，各自都有與眾不同的長處。很感謝有阿德、嘉俊、性獸、小 z、小孫、阿樹、棋翰和揚智合這兩年大家同舟共濟的相處，而且有阿樹哥當我的魔獸戰友，時常凌晨彌戰也不懈怠，是苦悶生活裡的唯一樂趣。而學弟妹們也相當有禮貌，讓我感覺到實驗室裡就像家裡一樣。

父母與家人的鼓勵、還有女友無時給我的支持，是我在低潮時期的支柱，很高興有你們在我身邊陪伴，讓我覺得人生的路上並不孤獨。研究所生活的所做所學，是我在大學以前所沒有的，很幸運能念到畢業，也期許未來的學弟妹們在語音領域上能有所貢獻。

目錄

表目錄

圖目錄

第一章 緒論

1.1 研究動機

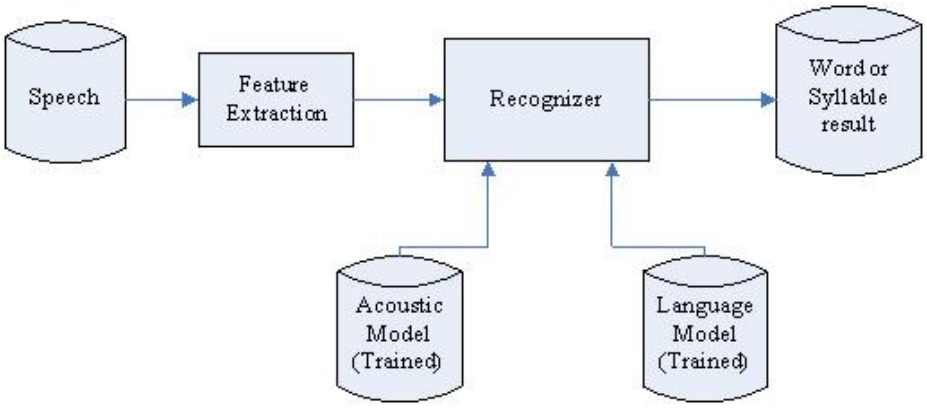
隨著科技的進步，可以讓我們的生活越來越方便，科技產品誕生的目的，就是要學習如何與人溝通、節省工作的時間，因此人機介面的研究是一個非常重要的問題，而利用語音溝通是最自然的介面之一。微軟總裁比爾蓋茲曾在他的演講中說道，未來的十年至二十年，語音辨識的技術將會大量的普及，屆時電腦硬體可能近乎免費，而原本身段高昂的技術，例如語音辨識的功能將會大為普及，深入到我們的日常生活裡。

由此所見，語音辨識尚有很大的研究空間，舉例來說，目前我們處在新聞媒體革命的時代，每天都有大量的新聞資訊產生，而如果能結合語音的技術，建立出一個良好的檢索系統（Information Retrieval），可以讓我們對於大量的新聞資訊能有效率的擷取。但是要建立檢索系統，必須要有準確的語音辨識器分析資料，所以在本論文中，我們將會針對中文廣播新聞，對此環境下的語音辨識做深入淺出的研究。

1.2 研究方向

一個全世界最簡單的辨識系統，可能是把全世界的人都找來，並且將每個人所有的聲音都錄製到系統裡，然後當我們要進行辨識時，只要一句一句比對，找出是或否就可以了。但是，這最簡單的方法也是最不可能、最困難的方法，所以我們會改以有效率的將辨識系統拆成幾個方塊，如將語音有效的編碼，利用幾

個聲學模型來延展出所有的語音，以統計學的方式統計語言規則。所以一個基本的辨識系統我們會設計如下：



圖一-1　基本辨識系統方塊圖

由上圖可看出，一個基本辨識系統包含了：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。但是根據不同的主題，我們還是要對辨識系統做不同的調整，在 2002 至 2004 年的中研院 MATBN（Mandarin Chinese broadcast news corpus）計劃裡，錄製了一套廣播新聞資料庫 PTSND（Public Television Service News Database），提供學術界在中文廣播新聞的研究。因此我們將建立一個廣播新聞的辨識器，針對中文廣播新聞的特性，如內外場環境的差異、不同的語者類型等等，探討如何以聲學模型上來處理這些問題。另外廣播新聞語音有很類似自發語音（Spontaneous Speech）的特性，自發語音在目前語音辨識上仍有很大的研究空間，而其常見的問題，我們也都會逐一去處理、分析。

在聲學模型上，我們以音節（Syllable）當作辨識的基本單元，而且並不考慮中文的音調變化。但是當加入語言模型時，我們可以適當的定義一個詞典，裡面包含中文常出現的詞彙，使得辨識器則在加入語言模型之後，可以直接辨識出詞（Word）或者字（Character），這樣就是一個完整的結果。所以在這裡我們會

說明語言模型建立的方法，還有加入語言模型之後辨識系統的比較。

1.3 章節概要

我們將整篇論文以幾個章節區分如下：

第一章：緒論 — 說明研究動機、研究方向及章節概要。

第二章：廣播新聞資料庫介紹 — 說明 PTSND 廣播新聞資料庫的特色，以及標記軟體與 XML-Parser。

第三章：基本辨識系統 — 說明如何建立一個廣播新聞類型的基本辨識系統。

第四章：依照環境分別建立聲學模型 — 依據環境與語者的差異，分別建立聲學模型，與基本辨識系統效能的比較。

第五章：加入語言模型至辨識系統 — 詞典選擇的方法及如何訓練語言模型，並且將語言模型加入到辨識系統裡。

第六章：結論與未來發展。

第二章 廣播新聞資料庫介紹

本章將介紹一套中文廣播新聞的資料庫（PTSND），並且以此資料庫為例，說明廣播新聞的特性。由於廣播新聞語音不同於一般的 Read speech，其語音是比較自發性口語化（Spontaneous）的，因此語音內容會隨著情緒起伏改變，或者是會受到背景聲音的影響。並且，我們進行的是連續語音辨識，因此新聞語音的辨識將會較 Read speech 及 Isolated word 的辨識複雜許多。

所以接下來，我們將說明如何建立一個適合廣播新聞語音的基本辨識系統，內容包括語音參數的編碼、聲學模型的建立，和我們如何進行模型的訓練，直到在基本辨識系統下進行辨識分析。

2.1 PTSND 介紹

要進行任何語音相關的研究，首先需要有良好的資料庫才能完成，國語語料在資料庫蒐集上過去就有一個成功的經驗：MAT（**M**andarin speech data **a**cross **T**aiwan），MAT 是一個設計好的 Read speech 資料庫，國內在出現此資料庫後，可以開始方便的進行語音相關的研究。

但是，不同的語音類型有不同的研究領域，由於國內目前在中文廣播新聞並沒有可實驗的語料，而 MAT 又沒有廣播新聞的語料，並且大陸地區的語音腔調和用字與台灣不同，因此中研院的王新民教授進行另一個計畫：MATBN（MAT Broadcast News）【1】，此計畫的目的在錄製國內中文廣播新聞的節目內容，並標記（Transcribe）成文字，資料合計 220 個小時，分三年逐步進行，錄製的節目來源為公共電視公司的節目，我們簡稱為 PTSND（Public Television Service News Database）。在有這套資料庫之後，我們可針對中文環境的 Broadcast channel 上提

供訓練與測試語料，進行連續語音辨識的計算。

本計畫錄製的節目內容為「公視新聞深度報導」及「公視晚間新聞」，每次節目為一小時，錄製及處理時間共分三年進行，從 2001/11 至 2004/7。分別為：

表二-1　PTSND 錄製計劃表

錄製時間	錄製資料量
第一年（2001~2002）	40 hours
第二年（2002~2003）	80 hours
第三年（2003~2004）	100 hours
合計	220 hours

其中錄製完成的節目，會再進行標記的動作（Transcribe）。整個 PTSND 資料庫標記的工作，是由中研院聘請兩位專職的標記員來進行。標記後的資料是進行任何相關研究的基礎，因此每個小時的節目都由一名標記員處理，然後再由另一名標記員做檢查，並且每週定時開會檢討。

目前前兩年的資料（40 小時與 80 小時）已經錄製並且標記完成，我們（交通大學語音實驗室）也已經處理好了，合計共 120 小時。而第三年的資料因為仍在錄製當中，等到錄製和標記完成後，未來我們也會將資料加入到辨識系統使用。

2.1.1 聲音格式

聲音錄製時是直接位於電視台，採用 DAT（Digital Audio Tape）以 44.1KHz 取樣率和 16bit 的精確度錄製，然後再作 Down-sampling，將聲音轉成電腦可讀取的 WAV 音檔格式。

Sample rate： 16KHz

Resolution： 16bits

Channel： Mono

Format： WAV（Microsoft Windows Wave File）

2.1.2 資料格式

用來進行標記的工具稱為「*Transcriber*」，這是一套適合廣播新聞使用的標記軟體，它可用來幫助標記員將聽到的聲音輸入成各類資訊，並且儲存的檔案為 XML 的語法格式。而 PTSND 錄製兩年的資料統計如下：

- 總時間：120 小時（兩年合計）
- 有語音時間：86.5 小時
- 總字數：約一百四十萬字

（註：有語音時間的意思為一句當中至少有中文語音存在才納入計算）

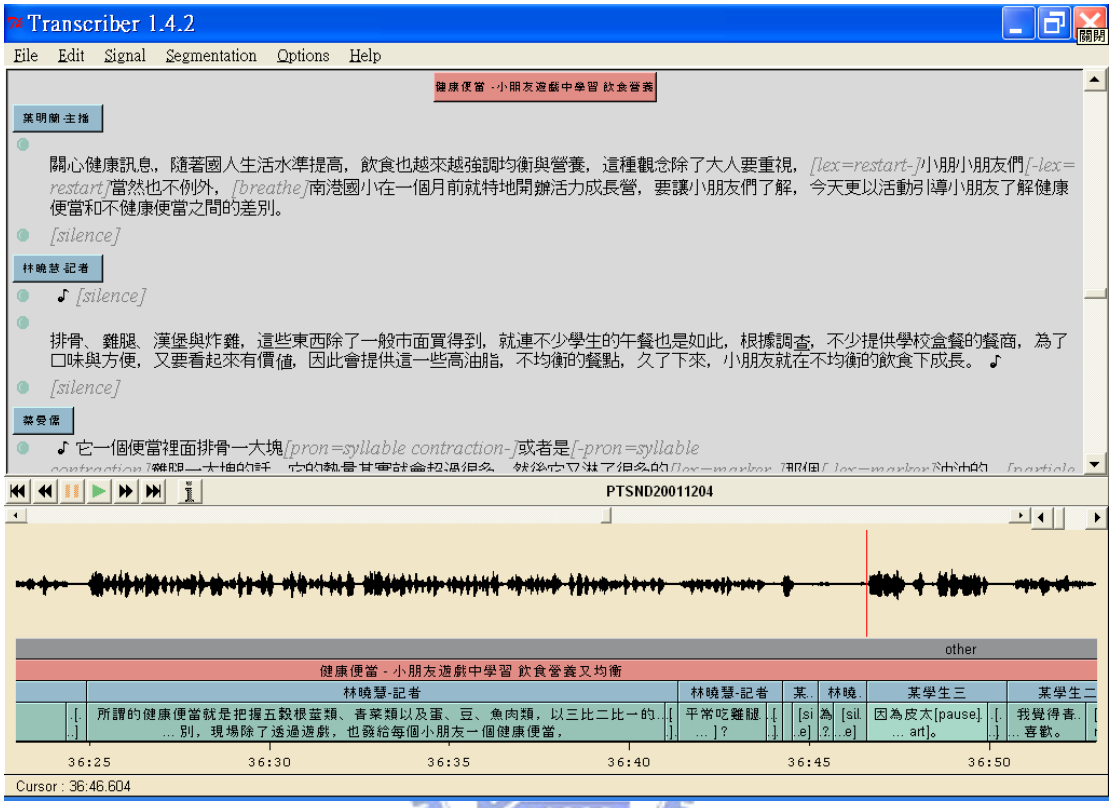
2.2 標記軟體 *Transcriber* 及 XML-Parser

在一般 Read Speech 的語音資料庫（以 MAT 為例），都是先設計好文字再由語者照著唸，並且同時錄製語音，所以發音通常都較為標準且有一致性。但是在廣播新聞裡，語音內容並不是照著稿子念的，因此我們會先有語音，再將錄好的語音標記出其語音內容，所以多了一個標記的動作（Transcribe）。而要進行標記就必須有一套輔助的軟體，目前使用的軟體程式的名稱為「*Transcriber*」。

另外，由於 *Transcriber* 軟體會以 XML 格式儲存檔案，所以我們還必須撰寫一個解譯器（XML-Parser）將我們需要用到的資料轉成可使用的格式，以供之後進行廣播新聞語音的研究時有正確的語料使用。

2.2.1 *Transcriber* 介紹

在標記軟體 *Transcriber* 下，其編輯環境如下所示：



圖二-1 *Transcriber* 標記軟體介面

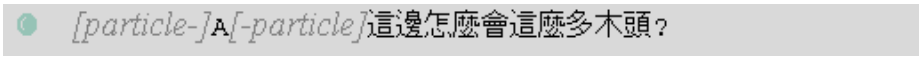
Transcriber 中可以紀錄所有的語者（Speaker）和主題（Topic）的資訊，同時還能詳盡的標記一段語音其文字的內容。標記的內容除了一般文字外，還可以記錄自發語音（Spontaneous speech）中常出現的情況，如 Particle、呼吸聲等等，或者是一些語言學現象（Paralinguistic Phenomena）也可標示出來，如砸嘴聲、漬舌聲、嘆氣聲等等。所有的聲音請參考【附錄一】。

另外，所有的語音標示資料在任何時間下都有四層的狀態，由下往上為：**語音文字、語者名稱、主題名稱、背景聲**（可參考上方圖片），由於四種狀態的特性正好符合廣播新聞的內容，因此 *Transcriber* 才會被選為使用。

Transcriber 儲存的檔案格式是採取 XML 的語法（或稱 SGML），利用 XML 的階層式架構可以容易的描述上述階層式之標示資料。一個格式完整的 XML 檔

案，需要伴隨一個 DTD 檔案（Document Type Definition）*(DTD 檔案的目的為定義文件格式)*，因此我們可以從閱讀 DTD 檔案瞭解 *Transcriber* 中定義的 XML 資料格式。

在語言學有關談話（Conversation）的定義中，一個 Turn 通常代表一來一往的對話，如果只是一方的聲音則稱為 Sub-turn。在 *Transcriber* 中，一個 Sub-turn 的例子如下：



圖二-2　*Transcriber* 中的一個 Sub-turn

則其對應的 XML 原始碼為：

```
<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?
```

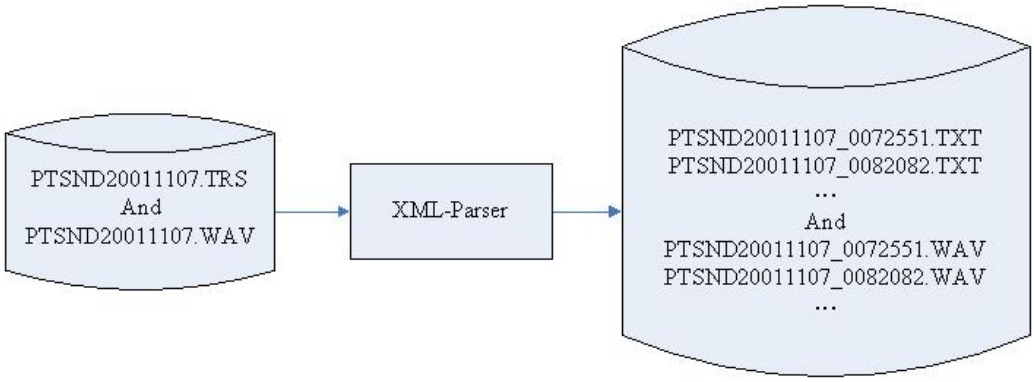
圖二-3　Sub-turn 對應的 XML 原始碼

在 *Transcriber* 中，每個 Sub-turn 的前方都會有一個圓形的符號，表示一段話的開始，而且在 XML 的原始碼中，代表為 Sync 元素（Element），並且有 time 的標籤（Attribution），因此我們可以從標籤的內容知道每個 Sub-turn 的起始時間。

2.2.2 XML-Parser

在上方的例子中， 我們可以從 XML 內容推得每個 Sub-turn 的起始時間與結束時間，因此可以將資料切割出來，而我們便撰寫一個 XML-Parser，可以將一個小時的節目切成以 Sub-turn 為單位的語料，並去除一些我們不要的 Sub-turn；例如若整個句子裡都是英文、或者是只有笑聲，基本上我們只會將有中文語音的

Sub-turn 才會切割出來。



圖二-4　XML-Parser 的功能

而每個 XML-Parser，就是將一個小時的節目讀入，然後依據時間切割成一個個 Sub-turn 的結果，依上節的例子其文字內容為：

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

圖二-5　XML-Parser 產生的 Sub-turn 結果

我們會對每個 Sub-turn 記錄下起始結束時間、語者 ID、語者名稱、語者性別和我們所關心的語音內容，也就是將去除了一些用不到的資訊。而語音內容我們會以固定的方式表示之【附錄一】。

2.3　廣播新聞特性

廣播新聞的語音因為類似於 Spontaneous speech（自發性語音）【2】，不像 Read speech 的語音是經過事先設計的，所以語音內容常常會隨著思考、情緒的變化，出現無法預期的聲音問題，增加了許多訓練與辨識時的困難度，基本上我們列出以下幾個口語語料中常見的特性：

- **Particle and Paralinguistic phenomena**

Particle 是 Spontaneous speech 中最常見的語音特性，在語言學上稱為「感嘆詞」，其可以分成 Grammatical Particle 與 Discourse Particle 兩種。但是在這裡我們並不考慮 Grammatical Particle，因為只針對聲學上的層次，所以之後的 Particle 均代表為 Discourse Particle。下方的一個 Discourse Particle 例子，例如「為什麼這樣 NEI？」，當中的「NEI」就是一個 Particle。

在 Spontaneous speech 中另一個常見的現象，比如笑聲、咳嗽聲、咂嘴聲 … 等，這些都稱為 Paralinguistic Phenomena。這類的聲音傳統語音辨識器是一大問題，而我們也將在基本辨識系統中提出處理這種問題的方法。

- **Background sound**

在廣播新聞當中，我們可以常常見到，除了語者的語音外，時常會出現背景聲音如音樂、人聲等等。不同類型的背景聲會使得 SNR 降低，同時也會干擾語音參數的正確性。

- **Foreign language**

由於國際化的趨勢，因此在廣播新聞我們可以聽到大量的外國語言或者是各地的方言，不同的語言各有複雜的建構方法，但是目前的辨識系統是針對中文語系，因此對於語音中出現這類外國語言的問題我們也會特別處理之。

- **Pronounce error**

在一般 Read speech，由於語音經過事先設計，因此可以發音正確不會有問題。但是在 Spontaneous speech 中，是先錄音再標記，所以發音不正確的情形就會時常發生，例如發音偏差（Inappropriate Pronunciation）、口吃（Stutter）、音節合併（Syllable Contraction）等等。

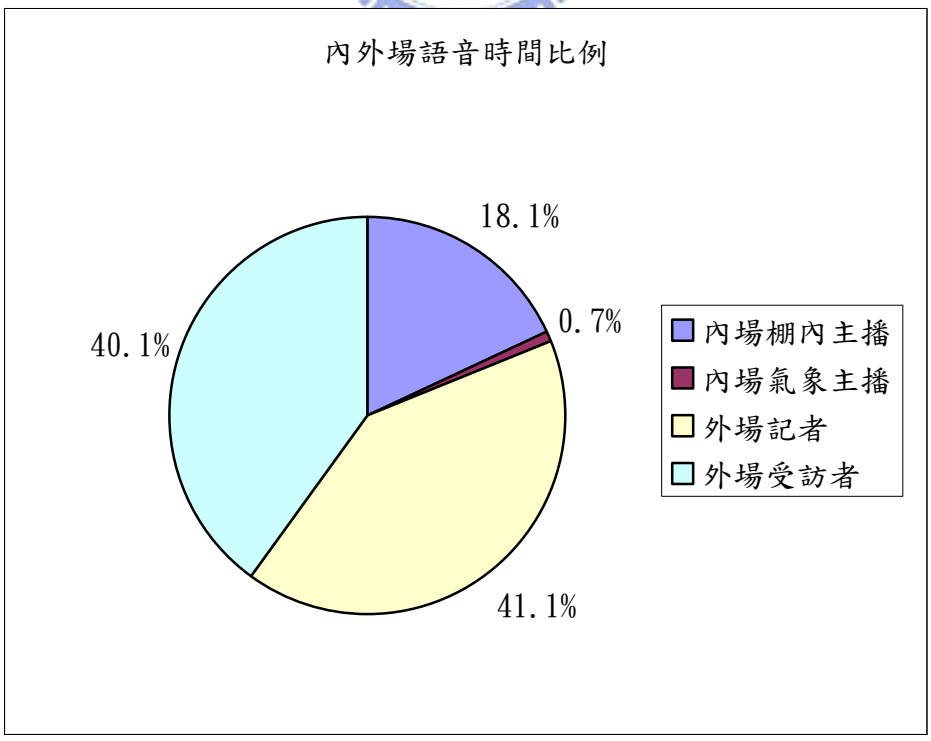
- **內外場環境**

一個廣播新聞節目語音，是由內場的主播（Anchor），和外場的記者（Field Reporter）與受訪者（Interviewee）所構成的，不同的內外場環境，其特性就會各不相同，而不同的語者，其語音正確性也會有所不同，例如主播和記者大多有受過發音訓練，而受訪者為一般民眾，內外場講話會較為口語化。

在 PTSND 第一年與第二年有語音的語料中，各環境的語者人數為：

- 內場棚內主播（Studio Anchor）：4
- 內場氣象主播（Weather Anchor）：2
- 外場記者（Field Reporter）：89（實際上可能小於此數，因為有部份記者身份不明）
- 外場受訪者（Interviewee）：3429

而在所有的語音時間（包含無背景聲與有背景聲的語音）上，內外場環境其各佔的時間比例為：



圖二-6　內外場佔所有語音時間比例

由此可看出整個廣播新聞內，棚內主播佔一個節目裡的語音時間其實大約20%左右，而外場語音才是佔廣播新聞內大多數的時間。

（註：目前氣象主播所佔的部份極小是因為第一年 40 小時的 PTSND 節目只有標記 10 小時，而第二年目前均不標記，因此我們時間無法計算，但實際資料大大於此處統計的）

第三章 基本辨識系統

目前語音辨識的主要方法有兩種，一種是採用 HMM（Hidden Markov Model；隱藏式馬可夫模型）另一種是 RNN（Recursive Neural Network；遞迴式神經網路），因為 HMM 是以口腔發音狀態（state）去模擬語音，因此近年來採用 HMM 於語音辨識的方法大受歡迎，而我們本篇論文也是採用 HMM 做研究。

實際上我們採用的是 left-to-right HMM，因為口腔聲道會隨著時間改變，並且語音信號具備短時間穩定特性，因此我們會假設在同個音框（frame）下，口腔狀態是相同。另外，狀態轉移機率（state transition probability）代表每個音框的信號在狀態下是要轉移或停留，而狀態觀測機率（state observation probability）代表音框與各狀態的相似程度，我們一般使用混合高斯模型（Mixture Gaussian Model）來表示。【3】

訓練模型時，可以進行 Baum-Welch 參數估計方法，依據轉移規則，由已知狀態序列推測出每個音框在哪個口腔狀態下是最佳的，並進而重估模型直到穩定為止。而當進行辨識時，則使用 Viterbi Search 令每個音框都要對所有聲學模型估計，並找出最佳結果，因此其運算量會較訓練時龐大。

在本論文中，我們進行實驗的環境，除了實驗室自行開發的程式外，還有採用英國劍橋大學開發的 HTK 工具（HMM Tool Kit），目前使用的程式版本為 HTK3.2.1。【4】

3.1 語音參數之設定

在進行訓練或辨識之前，我們均會先將輸入的語音均進行前處理，即求取其語音參數。我們求取的語音參數是梅爾倒頻譜參數（Mel-Frequency Cepstral

Coefficients；MFCC），利用語音在頻譜上具有短時間穩定的特性，並且 MFCC 有考慮到補償人耳的聽覺效應。

而語音參數求取時所使用之系統參數如下所示：音框化，音框長度 32ms、Overlap 22ms；消除 DC 效應；做 FFT 之前使用 Hamming Window；Pre-emphasis filter 為$1-0.97z^{-1}$；通過 Filter Bank：0~8KHz、24 通道；並求取 Delta-MFCC 與 Delta-MFCC，Delta-window=2、Delta-Delta-window=2；最後還會再做 CMN（Cepstral Mean Normalization）。

其中由於在 PTSND 中平均每秒的中文字數約達 5 個字，較一般速度（每秒 3 個字）來的快，因此我們在 Delta-window 與 Delta-Delta-window 均設定為 2。

我們求取的 MFCC 參數為 13 維度，再加上 1 維與 2 維的變化量。但是因為第 0 階的能量並不重要，因此第 0 階的能量會被省略，而第 1、2 階代表的是能量變化，因此會保留下來。所以最後求出的是一個 38 維度的語音參數向量。

3.2 初始模型之建立

在 HTK 中初始 HMM 模型之建立，有兩種方法，一種是 Flat Start，即所有次音節（sub-syllable，如中文的聲母、韻母）都使用相同的初始模型，然後再去做 Re-estimation，這種作法比須花費比較久的時間才能得到正確的模型，在語句中狀態數很多時也容易發生狀態位置完全錯誤之情況。而另一種做法就是已知位置（Fixed Boundary）下去做 Viterbi estimation，由於已經知道每個次音節的位置，所以直接對這些資料估計 HMM 參數會較為準確。

另外，我們建立初始模型的語料來源是採用 PTSND 第一年 40 小時的語料。

3.2.1 Forced Alignment

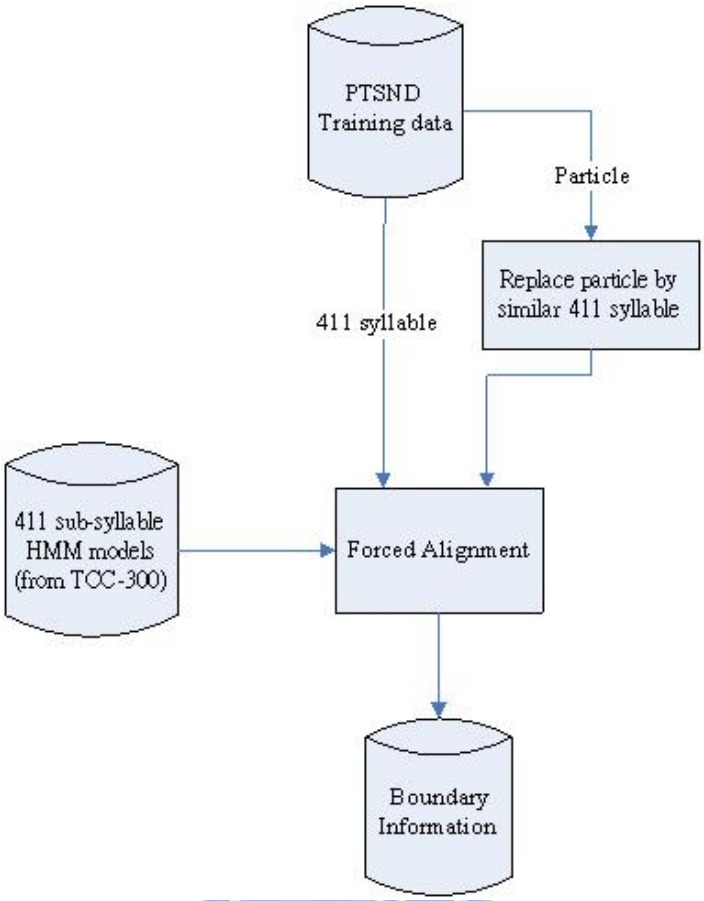
為了能求得切割位置，我們先以 Read speech（TCC300）所訓練之 HMM 模型來切割我們欲建立初始模型的訓練語料，初始訓練語料目前只選取單純的資料使用，也就是語料內容裡沒有背景聲（有背景聲下的語音辨識目前不是我們研究的方向），同時只能有國語 411 音節（Syllable）與 Particle 這兩種資料，資料的數量為 665 個 sub-turn、字數 35,388 字，時間約 2.05 個小時。

因為 Read speech 不會去訓練 Particle 這類在 Spontaneous speech 常出現的聲音模型，因此在進行切割（Forced Alignment），Particle 則使用相近 411 音節替代。

表三-1　Particle 以相近 411 音節替代表

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

經過 Forced Alignment 之後，可以產生 411 音節、Particle 和 Silence 的切割資訊（Boundary Information）。方塊圖如下所示：

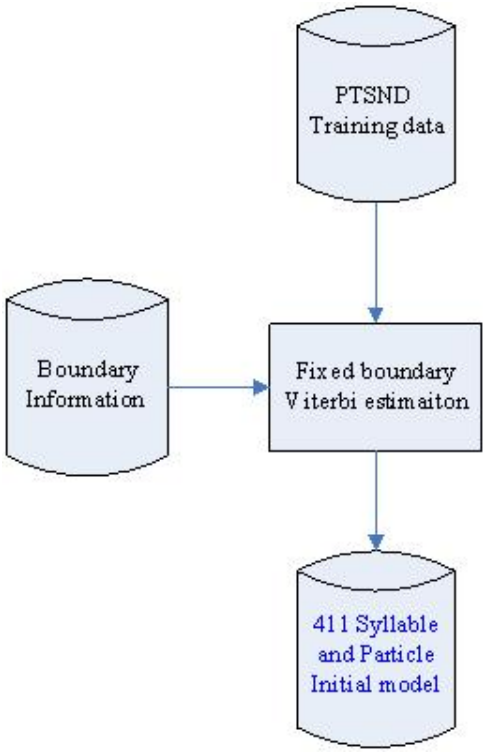


圖三-1　切割資訊求取方法

3.2.2 主要初始模型之建立

- **411 model 及 Particle model 之建立**

在求得 411 與 Particle 的切割資訊（Boundary Information）之後，我們會進行已知位置的初始模型訓練。訓練出的聲母和 Particle 採用 3 個狀態，韻母用 5 個狀態之 HMM 模式，Mixture 個數均為 16。訓練方塊圖如下：

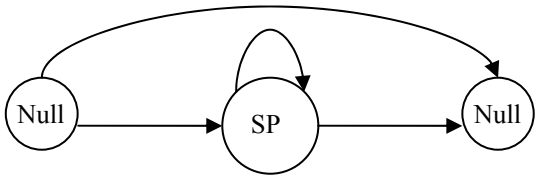


圖三-2　已知切割位置模型訓練

- **Silence model 之建立**

在之前進行切割後，我們亦得到 Silence 的切割資訊，因此 Silence model 的訓練是跟圖三-2 一樣的。訓練出的 Silence model 狀態有 3 個，並且因為資料量多故採用 32 個 Mixture。Silence model 的訓練來源有二，一個是每個 Sub-turn 的頭尾，另一個是當原始標記內容標上「Break」時，我們也會再增加訓練。

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（State），此狀態允許跳躍（Skip），並且與 Silence 的中間狀態合併（Tying）。



圖三-3　SP（Short Pause）HMM 模型

- **Breath model 之建立**

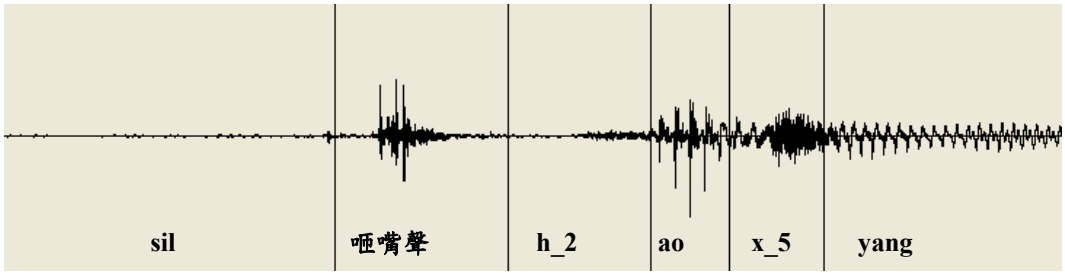
Breath 在廣播新聞或者自語音中常常可見，但是我們缺乏 Breath 的切割位置，因此這裡是先以人工切割出一些訓練語料，再建立其初始模型。建立的 Breath 狀態有 3 個，Mixture 個數為 16。

3.2.3 Garbage Model 之建立

在廣播新聞裡，會出現各種自發性語音（Spontaneous speech）中常見的現象，而且其中有許多是不易處理的，舉例來說，在語言學現象（Paralinguistic Phenomena）中的咳嗽聲、笑聲等這些聲音，這些資料量不一定足以訓練模型，因此對於這種聲音，我們會共同建立一個特殊語音模型來處理，稱為「Garbage model」，這是用來取代各種特殊聲音的語音模型【5】。

由於 Garbage model 被用來取代許多特殊聲音，因此資料分布範圍很大，以機率模型來看，它是一個變異數（variance）很大的 Gaussian distribution，當我們要進行語音切割或辨識時，正常的語音在 Garbage model 跟其他的語音模型相比，Garbage model 的分數會顯得較小而不會被選擇。但是當遇到咳嗽聲、咂嘴聲時，這些聲音使用正常語音模式所得到的辨認分數會較小，但是會被落在分佈較廣的 Garbage model 而可以被切割出來。

以下是一個以 Garbage model 去切割咂嘴聲的結果：



圖三-4　使用 Garbage model 切割咂嘴聲的結果

由此可推廣出，假如在新聞節目裡若穿插有英文、閩南語的語言時，因為

我們只著重在國語語音的部分（Mandarin），不去處理外國語言的問題，因此也會以一個模糊的 Garbage model 來涵蓋這些資料。

Garbage model 之建立的幾個重點如下：

- **Forced alignment**

為了避免 Garbage model 與 Silence model 發生混淆，在利用 3.2.1 的切割資訊時，必須使用所有 non-silence 的資料一起去訓練 Garbage model 的初始模型。因此最後訓練了 3 個狀態的 Garbage model，Mixture 數為 32。

- **使用範圍**
  - 英文
  - 閩南語
  - 語音學現象、無法辨識的字詞等

3.3 訓練程序

初始模型訓練好後，必須再進行整個句子的 Embedded re-estimation，此時已經不需要切割資訊了。另外由於初始訓練語料並不多，模型估計不一定準確，因此在這裡我們會將所有的聲音都拿進來，但是如果資料當中有背景聲的話，目前我們均不處理，因此本論文的語料都是在無背景聲的環境下（Clear speech）。

另外，我們還會調整國語 411 音節與 Particle 的狀態 Mixture 個數，而調整的方式是依據其資料量的多寡，每 50 個音框增加一個 Mixture，每個狀態的 Mixture 數為 1 至 16。

我們會進行訓練，直到模型穩定收斂為止。而判定穩定的方法如下：

$$\text{Convergence Condition: } \frac{P-P'}{P'} < 10^{-4} \tag{3.1}$$

其中 P = Average log Prob. per frame，P'= Last Average log Prob. per frame。

目前語料採用的是 PTSND 第一年 40 小時。我們將所有可用語料的十分之九歸於訓練語料，十分之一歸於測試語料。其中訓練語料的時間大約有 8.5 個小時。而總共可用語料如下：

表三-2　基本辨識系統訓練語料數量

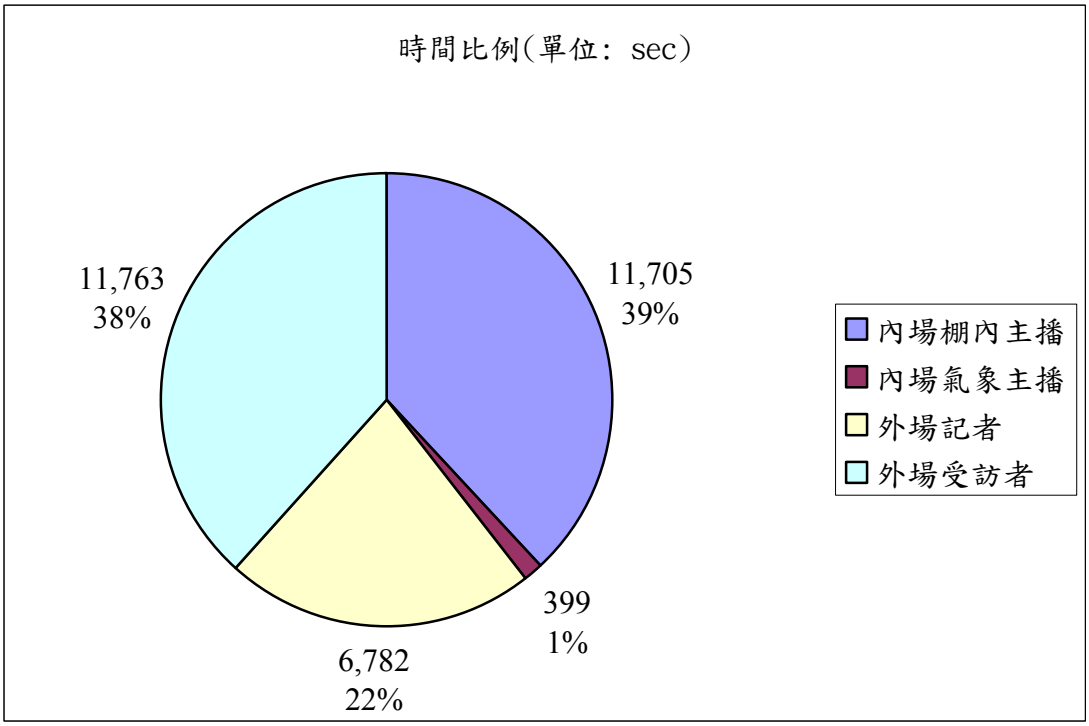
語料種類	Sub-turn 數	字數
訓練語料	1,978	149,548
測試語料	220	19,142
可用語料	2,198	168,690

而總共訓練出的聲學模型數量如下，HMM 參數的設定則與初始模型時相同：

表三-3　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

而我們若再將訓練語料以內外場分成四類的話，則此訓練語料在所佔的比例為：



圖三-5　訓練語料內外場比例

可以看出訓練語料裡，內場和外場所佔的比例為 2：3。由於不同的環境，對辨識效能有不同的影響，所以在之後的辨識實驗中，我們將針對不同的內外場語料進行辨識。

3.4 實驗-基本辨識系統分析

接下來將要進行基本辨識系統的辨識分析，我們辨識系統是以音節（Syllable）當做辨識的輸出單元。

- **測試語料**

依前面廣播新聞的環境特性我們將測試語料分成三種類型，並且都是沒有背景聲音的語料，其中內場主播包含棚內主播與氣象主播。

我們還會對個各環境下進行 Outside 與 Inside 語料的測試，Outside 是指測試語料使用與訓練模型無關的語料，Inside 是指從訓練模型語料抽取出來的語料。因此一般 Inside 語料辨識的結果會比較好，但是以 Outside 語料進行辨識才

是公正的方法。

表三-4　基本辨識系統測試語料數量

辨識種類	環境	Sub-turn 個數	Time(小時)	字數
Outside	內場主播	108	0.52	9,291
	外場記者	55	0.28	5,197
	外場受訪者	57	0.29	4,672
Inside	內場主播	189	0.94	16,711
	外場記者	197	0.73	13,660
	外場受訪者	189	0.78	13,167

文法規則（Grammar rule）上我們是採用無文法（Free Grammar），也就是所有文字出現的機率都一樣。而辨識結果可能會有 Particle、Breath 等出現，為了避免混淆，因此觀察辨識率時，我們只考慮國語 411 音節。

而辨識率計算的方式為：

$$\text{Accuracy} = \left(1 - \frac{\text{Sub+Del+Ins}}{\text{Number}}\right) * 100\% \qquad (3.2)$$

其中 Sub 為替代型錯誤，Del 為刪除型錯誤，而 Ins 為插入型錯誤。Number 為文字的個數。

- **辨識率結果**

在基本辨識系統裡，是採用音節作為辨識的基本單元，因此我們計算辨識率時，就考慮音節的辨識率（Syllable recognition rate），並且只計算國語 411 音節的辨識率。則 Outside 與 Inside 測試語料的辨識率結果如下：

表三-5　Outside 測試語料的 syllable 辨識率

環境	Del	Sub	Ins	Number	Accuracy
內場主播	3.14%	23.35%	0.88%	9,291	72.63%
外場記者	2.12%	33.83%	2.50%	5,197	61.55%
外場受訪者	8.05%	46.94%	5.27%	4,672	39.75%

表三-6　Inside 測試語料的 syllable 辨識率

環境	Del	Sub	Ins	Number	Accuracy
內場主播	3.44%	19.56%	0.80%	16,711	76.20%
外場記者	2.94%	26.35%	1.28%	13,660	69.43%
外場受訪者	8.72%	41.32%	3.30%	13,167	46.66%

- **實驗分析**
  - 內場主播因為人數少、受過專業發音訓練，並且在安靜不受干擾的環境下，因此可以反映到辨識率為最高。
  - 記者雖然發音也較標準，但是身在外場的環境，容易受環境的影響，因此辨識率介於中間。
  - 受訪者的辨識率最差，是拉低辨識效能的主因。這是因為受訪者人數最多，且大多為一般民眾，情緒起伏比較大，發音較不準確且非常口語化所致。因此外場受訪者是廣播新聞研究的主要重點。
  - Inside 的辨識率比 Outside 的辨識率佳，一般大約高 4%至 6%不等。

3.4.1 與 SoVideo 檢索系統的比較

在中研院王新民教授的 SoVideo 檢索系統中【6】，必須使用語音辨識進行新聞語音的前處理，在辨識完成後，才可以知道文字內容以進行檢索。由於此檢索系統也是使用 PTSND 作為測試語料，因此以下將說明與我們基本辨識系統的比較。

SoVideo 的語音辨識部份，在聲學模型（Acoustic Model）上，採用 16 小時的廣播新聞做為訓練語料（但此廣播新聞來源並不是 PTSND），總共訓練出 112 個右相關聲母模型與 38 個韻母模型，並且每個模型狀態會有 4 至 64 不等的 Mixture 個數。而在語言模型(Language Model)方面，訓練語料使用中央社(Central News Agency）的資料，共 6 千 5 百萬的中文字，建立出了 Bigram 與 Trigram 的音節語音模型，並且使用一個 61,521 詞的詞典定義詞（Word）與音節間的關係。

在 SoVideo 的語音辨識中，對於無背景聲的測試語料，同樣也是 Outside 測試語料進行辨識，其音節辨識率如下，而詳細的辨識結果請參考【6】:

表三-7　SoVideo 的辨識率比較

環境	SoVideo（已加 LM）	基本辨識系統
棚內主播	69.87%	72.63%
外場記者	62.78%	61.55%
外場受訪者	30.48%	39.75%

（註：在基本辨識系統 72.63%是指內場主播，是同時包含棚內主播與氣象主播）

由表三-7 發現，在棚內主播與外場受訪者的部分都是基本辨識系統較高，這可能是因為基本辨識系統的訓練與測試語料都在相同環境，因此聲學模型比較準確。而在外場記者部份則為差不多的結果。

因為 SoVideo 的聲學模型訓練語料並不是採用 PTSND，因此聲學模型並不算準確，但是在 SoVideo 的辨識系統裡有加入語言模型（Language Model），本節中之基本辨識系統只是 Free Grammar，加入語言模型能額外提升辨識效能，因此本節的辨識率比較，我們僅作為參考。

第四章 依照環境分別建立聲學模型

由上章的實驗中可以發現，基本辨識系統，是以整個廣播新聞當作訓練語料的來源，訓練出的聲學模型對於各種環境進行辨識時，辨識出的效能差異是相當大的。因此本章將探討如何把廣播新聞依照環境及語者的特性，將訓練語料拆開，分別建立聲學模型，然後在已知或未知環境的情況下，觀察能否提高辨識率。

4.1 內外場環境之探討

廣播新聞資料庫因為不同於一般 Read speech，在不同的環境和不同的語者下，都會產生不同的語音品質。如果以環境來區分的話，廣播新聞可以分成內場與外場，內場就代表在攝影棚或者錄音室這類的環境下，因為比較安靜所以訊號品質都比較佳（SNR 較高）；而外場環境就是指戶外的環境，可能是在街道上，或者是在公共場合裡，通常語者在外場時，往往會因為較大的噪音而使得語音受到影響（SNR 較低），而且不同的環境錄音器材也有所不同，因此外場語音品質會是最差的。

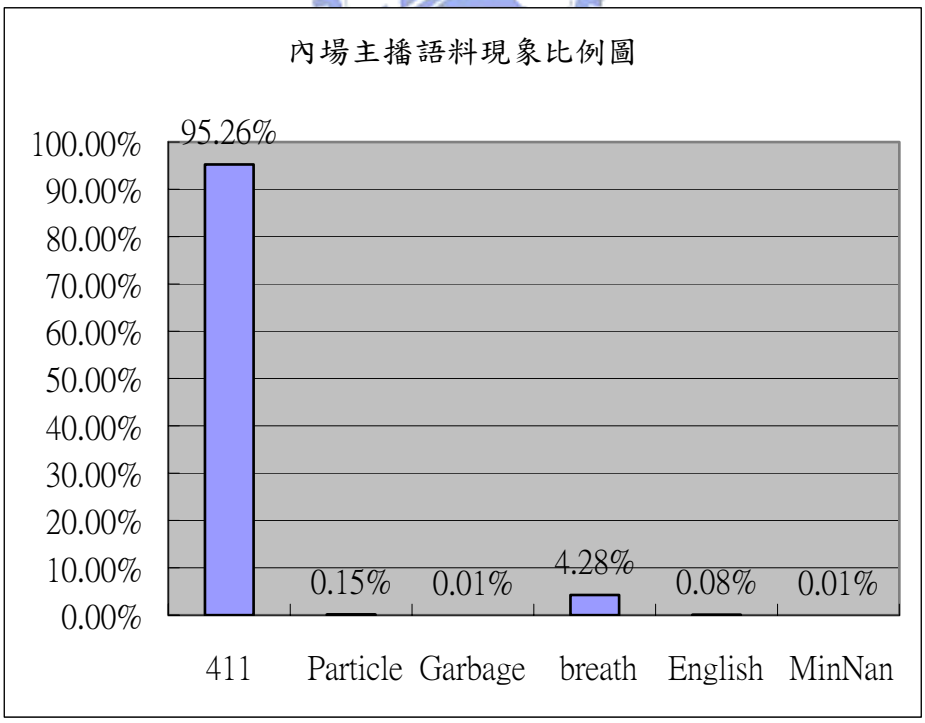
若以語者來區分的話，在一個廣播新聞裡，我們可以區分成四種，分別為棚內主播（Studio Anchor）、氣象主播（Weather Anchor）、記者（Field Reporter）與受訪者（Interviewee）。【1】

由於棚內主播與氣象主播都是屬於內場的聲音，差異性並不大，因此我們在進行實驗時會將兩者合併，只以內場主播（Anchor）來概稱之。而記者的聲音，由觀察新聞內容可得知，大多數記者的聲音不是伴隨著新聞影片的旁白，就是實地去採訪受訪者的語音，而這些聲音都是在外場環境下進行的，因此我們會將記

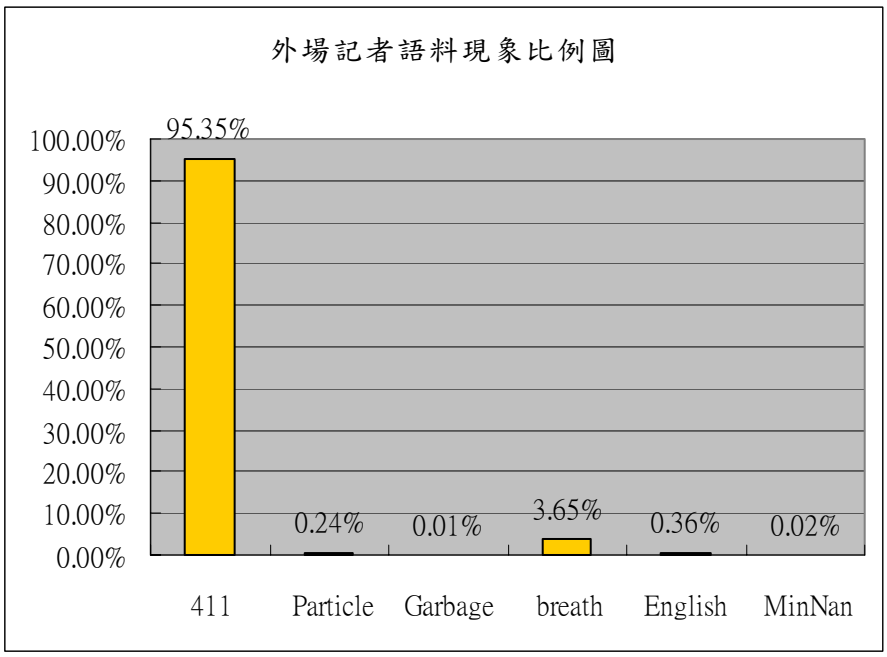
者歸類為外場環境。然後受訪者的部分更是無庸置疑，大部份的聲音都是屬於外場的環境。

而記者與受訪者的語音雖然都是在外場，但是兩者的發音品質各不相同。記者受過發音訓練比較準確、說話也比較規律，而受訪者大多為一般的民眾，說話比較口語化（spontaneous）和情緒化，因此我們不會像棚內主播與氣象主播合併成一種內場主播而已。最後，我們可以將一個廣播新聞節目的所有語音，區分成三種環境，總共為：內場主播、外場記者、外場受訪者。

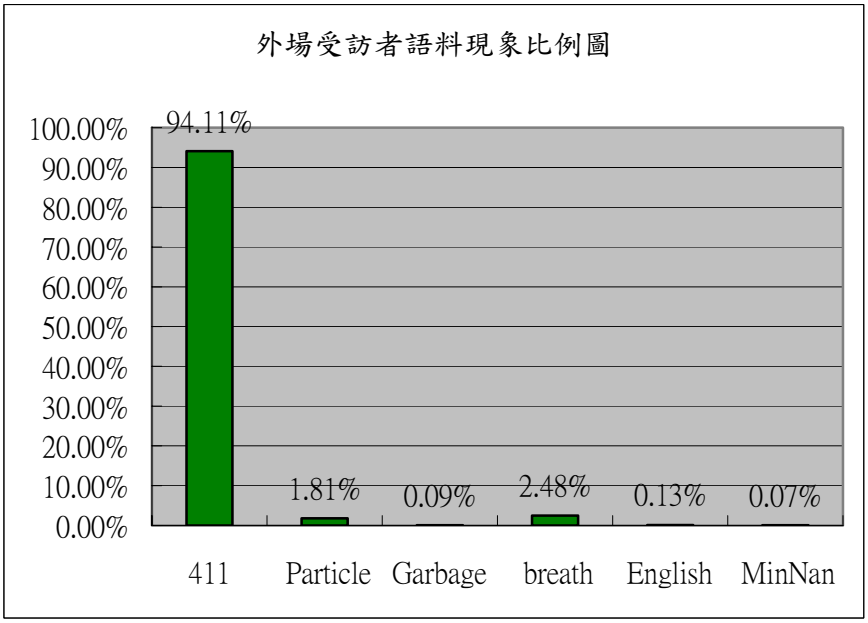
以下我們再從訓練語料中觀察（訓練語料內容請參考 4.2 說明），國語 411 音節、Particle、呼吸聲等聲音在三個環境所佔的比例統計。411 即代表中文字的數量，Garbage 表示咳嗽聲、笑聲及一些無法處理的聲音，而呼吸聲因為出現次數特別多所以另外統計出來，還有英文與閩南語這兩個是最常見的外國語言，因此也一併統計。



圖四-1　內場主播（Anchor）語料現象比例圖



圖四-2　外場記者（Reporter）語料現象比例圖



圖四-3　外場受訪者（Interviewee）語料現象比例圖

很明顯可以看出來，三個環境的比較下，外場在非國語 411 語音外的比例比較多。以外場受訪者來看，其 Particle 與以 Garbage 這類的聲音佔的比例是最高的；而內場主播在呼吸聲比例上為最高，這是因為呼吸聲隨著句子長短而出現，每個環境的句子由長至短為：內場主播、外場記者、外場受訪者，而呼吸聲

的比例也與此符合。另外我們再觀察發音偏差（Inappropriate Pronunciation）的比例，其例子如，當「發生」唸成「ㄏㄨㄚ生」時，此時會標記發音偏差與其唸錯的音，這也是在自發性語音中常見的現象。其出現的比例統計如下：

表四-1　各環境下發音偏差的比例

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

其中又以外場受訪者在發音偏差上的問題最嚴重，當發音偏差比例過高時，會增加模型訓練的困難度。由此可知，不同環境會有不同的特性，因此我們不用將所有資料共同訓練一種模型，而可以考慮對每個環境個別訓練，所以接下來就是要進行分開環境去建立各自的聲學模型。

4.2 訓練程序

- **訓練語料重選**

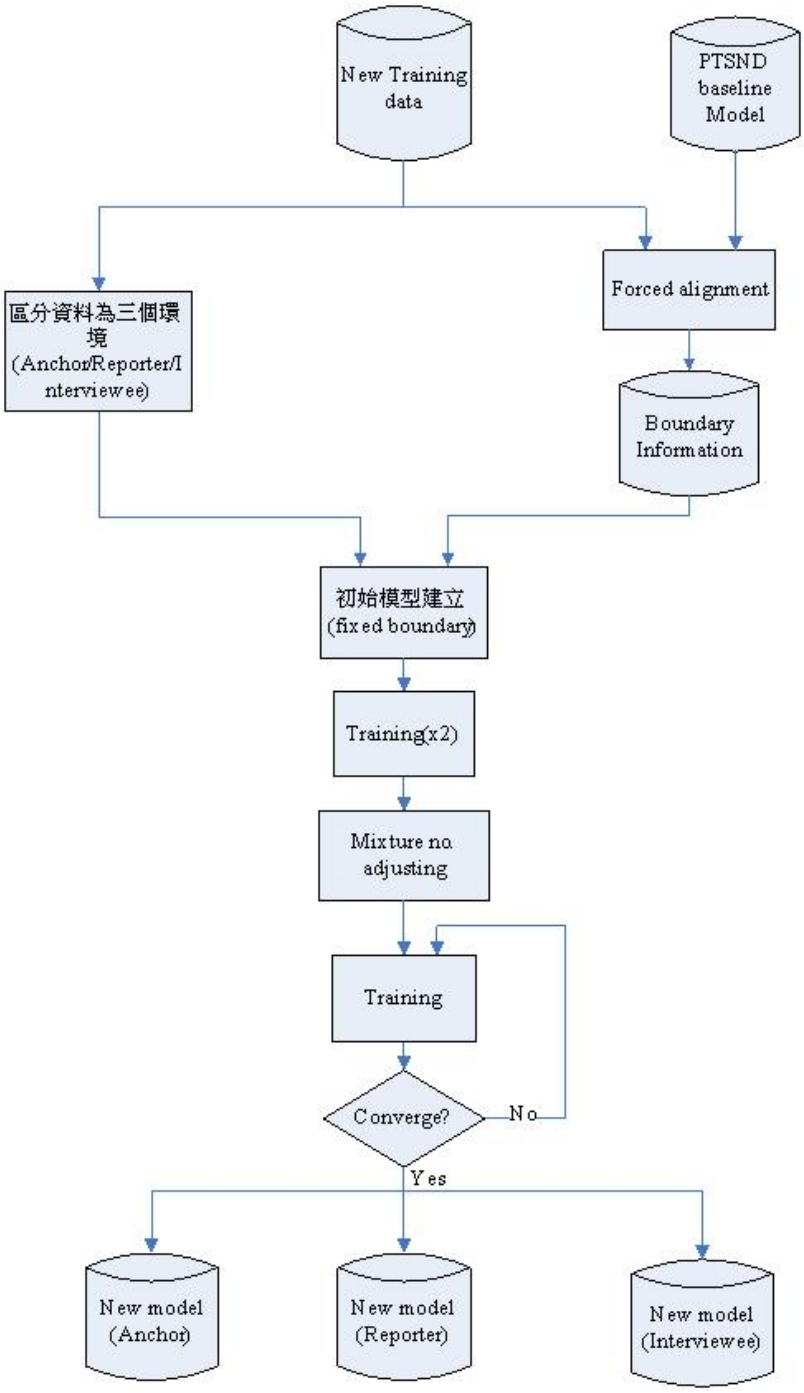
由於若要進行個別環境聲學模型的訓練，首先遇到的問題就是訓練語料是否足夠，因此在這裡我們同時加入第 1 年與第 2 年的語料，並且修改我們的 XML-Parser，使切出來的可用資料更多。所以目前的可用資料共有 24.2 個小時，約 42 萬個字。再拆開成三個環境後，我們也是選取十分之九做為訓練語料，十分之一為測試語料。則各環境訓練語料有：

表四-2　各環境下的訓練語料數量

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

- **訓練流程圖**

我們使用在基本辨識系統訓練的模型（PTSND baseline Model）對新的訓練語料進行切割，這樣就可以得到每個次音節（Sub-syllable）的切割資訊。再把訓練語料分成三種環境，各別建立初始模型，而建立初始模型的方法與基本辨識系統均相同。



圖四-4　依環境訓練聲學模型流程圖

同時，我們也會進行每個聲學模型狀態中 Mixture 個數的調整，而調整的方式與基本辨識系統相同，Mixture 的個數為 1 至 16。

最後我們就可以得到各別環境的聲學模型。對每個環境下我們都會得到模型數如下：

表四-3　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 ~ 16
韻母	40	5	1 ~ 16
Particle	35(4/7/16)	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

在之前我們已經說明過，由於在各個環境裡 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，所以我們會對每個環境都建立 35 個 Particle，但實際能建出模型的，對三個環境（內場主播、外場記者與外場受訪者）來看只有 4 個、7 個與 16 個。外場受訪者因為 Particle 出現比例最高，所以可建立的模型數也最高。而對於資料過少無法建立的 Particle 模型時，我們會採用模型共享（Model Tying）的方式去跟相近音共享訓練語料。

4.3 實驗-不同環境聲學模型的辨識分析

當我們完成聲學模型的訓練後，就開始進行辨識的實驗，觀察分開環境訓練模型之後的結果。而我們主要分成兩種方法實驗，分別為已知環境與未知環境。已知環境下就是代表以相同環境的聲學模型去對測試語料辨識，而未知環境，就是由輸入語音去自己去比對找出最高分的環境模型當作其辨識結果。

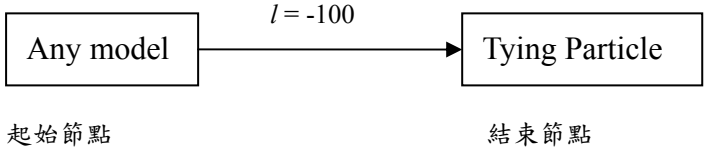
4.3.1 已知環境辨識

由於在每個環境下，我們實際建立的 Particle 模型數量較少，而無法建立的會以模型合併（Model Tying）的方式處理，但是為了避免辨識時採用模型共享的 Particle 出現機率會超過原 Particle，因此我們會在 Word-net（文字網路）上做前處理。

方法就是將原本的無文法規則（Free Grammar）轉到 Word-net 後，Word-net 定義了辨識系統裡所有節點之間的轉移（transition）關係，而每個轉移都是由一個起始節點（Start node）與結束節點（End node）所組成，我們就令所有結束節點為這種 Particle 的轉移上加一個分數（Language Model Score），使其被辨識出的機率較低，如下：

$$\log(P(w_i = Tying\ Particle \mid w_{i-1})) = -100 \qquad (4.1)$$

在 Word-net 上的處理則如下圖：



圖四-5　Tying Particle 在 Word-net 的處理

- **測試語料**

在這裡測試語料選取無背景聲，以及無外國語言的語料，並且均為 Outside 測試語料的辨識。

表四-4　各環境下的測試語料數量

測試語料	Sub-turn 個數	Time(小時)	中文字數
內場主播	190	0.84	14,906
外場記者	210	0.5	9,279
外場受訪者	186	0.63	10,382

- **辨識率結果**

依模型環境與測試語料的關係，進行了各環境間的交叉辨識，因此可以列出所有的音節（Syllable）辨識率結果。其辨識率的計算方法與基本辨識系統相同，可參考 3.4。

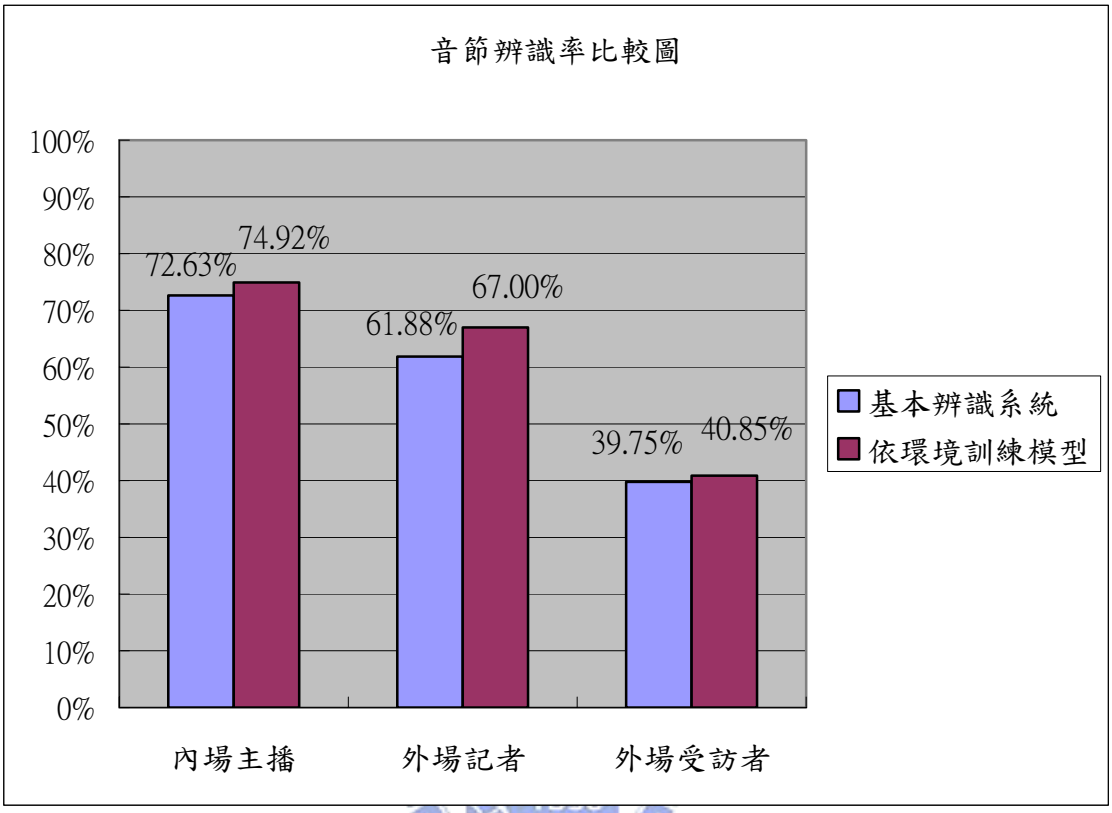
表四-5　已知環境交叉辨識的 Syllable 辨識率

模型環境 / 測試語料	內場主播	外場記者	外場受訪者
內場主播	**74.92%**	55.2%	42.94%
外場記者	48.14%	**67%**	52.32%
外場受訪者	20.62%	29.74%	**40.85%**

- **實驗分析**

由辨識結果可以明顯發現，在每個測試語料中，都是在相同環境模型下的辨識率為最高，以表四-5 來看就是對角線中的辨識率為最高。

並且，與基本辨識系統相比，分開環境個別訓練聲學模型確實可以提升辨識效能，如下圖所示：

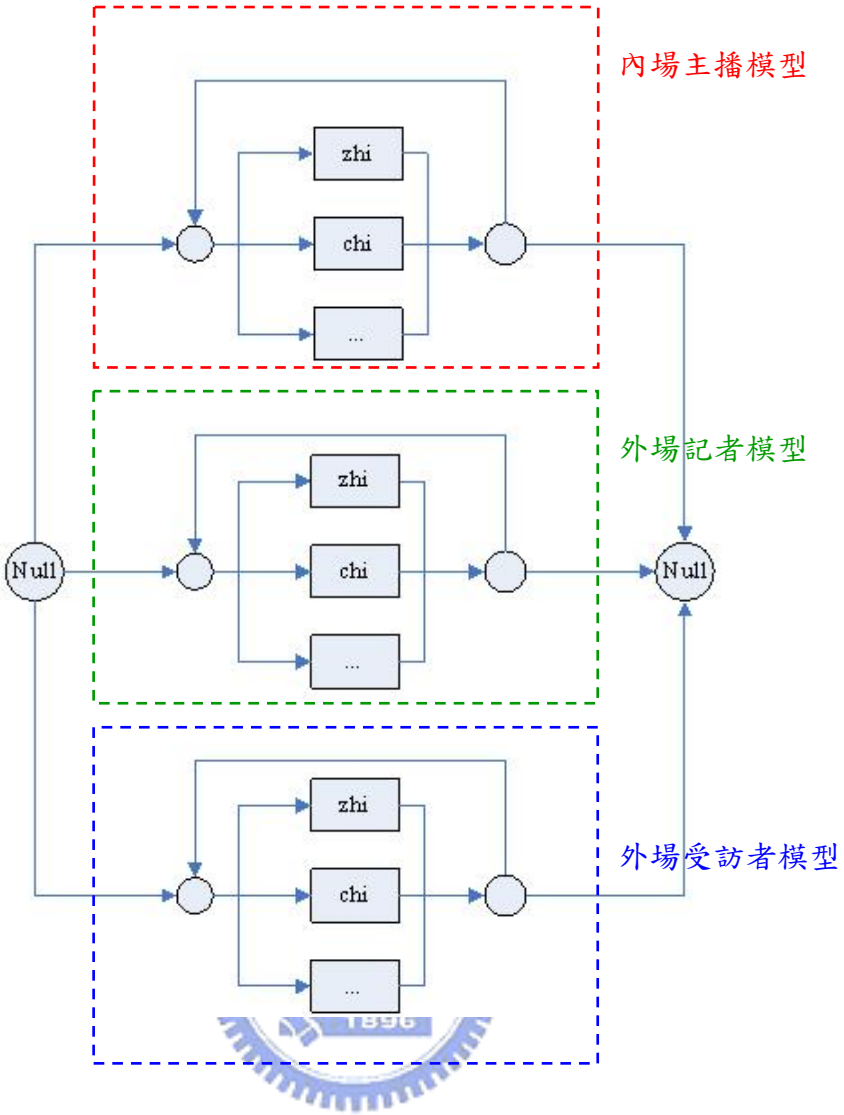


圖四-6　依環境訓練模型與基本辨識系統的辨識率比較

4.3.2 未知環境辨識

在實際的廣播新聞辨識系統裡，我們並不會知道此時輸入的聲音會是在哪個環境下，因此為了符合實際的系統，我們應該整合各個環境下的語音模型，由輸入語音去比對找出一個最適合的環境。這也類似一般有考慮性別的辨識器裡，我們會對男、女個別訓練語音模型，在測試時就由語音自己選擇較高分的模型當作其辨識結果。

而我們的做法就是，修改原本的無文法規則（Free Grammar），使得進行辨識時可選擇環境，並且在同個句子的每個音框都是在相同環境模型下。則我們將原有的 Free Grammar 修改如下：



圖四-7　Free Grammar 於未知環境辨識下的修改

並且，對於進行模型合併的 Particle 之處理與測試語料的選取，與已知環境的辨識均相同。

- **辨識率結果**

表四-6　未知環境的 Syllable 辨識率

環境	Del	Sub	Ins	Number	Accuracy
內場主播	3.05%	21.19%	0.94%	14,906	74.82%
外場記者	2.61%	29.11%	1.31%	9,279	66.97%
外場受訪者	6.63%	45.71%	4.87%	10,377	42.80%

- **實驗分析**

由未知環境辨識的結果可發現，在內場主播與外場記者的辨識率均較已知環境的辨識來得低，不過差距並不大，這是因為有部分的 Sub-turn 被辨識成別的環境下，辨識率因此而下降。

但是在外場受訪者，其辨識率竟然反而較為上升，似乎與正常結果背道而馳，這可以由下表來解釋之。下表是一個測試語料原本環境與辨識出環境的比例關係：

表四-7　未知環境下各環境 Sub-turn 相互辨識比例

原本環境 / 辨識環境	內場主播	外場記者	外場受訪者
內場主播	95.26%	0.48%	0.54%
外場記者	4.21%	93.33%	16.67%
外場受訪者	0.53%	6.19%	82.80%

這當中的原因是，因為辨識時的環境可以自動選擇，有部分的外場受訪者的環境比較偏向外場記者，比如在辦公室的訪問新聞中，兩者的環境是相同的，因此外場受訪者有部分 Sub-turn 會被選到外場記者的環境下，所以才造成未知環境下的模型辨識反而效能會比較好。

而從上方的辨識環境比例中發現，在進行外場受訪者測試語料的辨識時，有 16.67%的 Sub-turn 環境確實被選擇在外場記者環境下，因此我們上方的推論是合理的。

第五章　加入語言模型至辨識系統

5.1　語言模型（Language Model）簡介

由於所有的語言都有其文法規則，因此我們若能針對此規則性，求得一個機率模型，則我們稱此為語言模型，簡稱 LM（Language Model）。在進行語音辨識時，除了考慮聲音模型（Acoustic Model）外，若還能加入 LM 的參考，通常可以大幅提高辨識系統的效能。

一般在 LM 下，我們建立模型是以 Word 為基本單位，而在漢語中文（Mandarin）下，就是以「詞」為單位。因為在自然語言中以詞來建構會比較符合語言規則。例如有一個詞（Word）為「蘋果」，我們不會將這個詞拆成兩個字「蘋」和「果」（Character），而是以一個詞來看它才有意義。因此，在建立 LM 之前，我們會先建立一個詞典（lexicon），詞典裡面定義了所有我們要使用的詞。

5.1.1 n-gram 語言模型

假設有一個詞串（Word sequence）或句子（Sentence），其內容以詞（Word）為單位為「$w_1, w_2, ..., w_m$」，則此詞串對應的機率為【7】：

$$\begin{aligned} P(w_1, w_2, ..., w_m) &= P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1, w_2)\cdots P(w_n \mid w_1, w_2, ..., w_{n-1}) \\ &= \prod_{i=1}^{m} P(w_i \mid w_1, ..., w_{i-1}) \end{aligned} \tag{5.1}$$

由於要求得所有詞的條件機率是不可能的，所以我們可以使用 n-gram 的機率去趨近。

$$P(w_1, w_2, ..., w_m) \simeq \prod_{i=1}^{m} P(w_i \mid w_{i-n+1}, ..., w_{i-1}) \tag{5.2}$$

其中每個 n-gram 的機率如下式所示：

$$P(w_i \mid w_{i-n+1}, ..., w_{i-1}) = \frac{Count(w_{i-n+1}, ..., w_i)}{Count(w_{i-n+1}, ..., w_{i-1})} \tag{5.3}$$

其中，$Count(.)$ 表示為詞串出現的次數。在求得所有詞串 n-gram 的機率後，我們就得到 LM 了。

5.1.2 Smoothing 機率

但是，當 $Count(.)$ 為 0 次時，則此 n-gram 機率就為 0，這樣在資訊理論（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，而且我們也無法肯定辨識結果不會有這個組合，因此這是不合理的。另外當 $Count(.)$ 次數太小時，也會使得機率估計並不準確。所以我們會進行 Smoothing【8】，使得所有 n-gram 機率能被良好的估計。因此一個詳細的估計方法如下所示：

$$P(w_i \mid w_{i-n+1}, ..., .w_{i-1}) = \begin{cases} a(w_{i-n+1}, ..., w_{i-1}).P(w_i \mid w_{i-n+2}, ..., w_{i-1}) & : Count(w_{i-n+1}, ..., w_i) = 0 \\ d_a.\dfrac{Count(w_{i-n+1}, ..., w_i)}{Count(w_{i-n+1}, ..., w_{i-1})} & : 1 \le Count(w_{i-n+1}, ..., w_i) \le k \\ \dfrac{Count(w_{i-n+1}, ..., w_i)}{Count(w_{i-n+1}, ..., w_{i-1})} & : Count(w_{i-n+1}, ..., w_i) > k \end{cases} \tag{5.4}$$

其中 $a(w_{i-n+1}, ..., w_{i-1})$ 表示為 Back-off 係數，當 n-gram 詞串出現次數為 0，則取其 $(n-1)$-gram 的機率做趨近，並乘上 Back-off 係數，這樣可避免機率 0 的出現。而 $a(w_{i-n+1}, ..., w_{i-1})$ 的選定為令其滿足：

$$\sum_{w \in V} P(w_i = w \mid w_{i-n+1}, ..., w_{i-1}) = 1 \tag{5.5}$$

另外當 n-gram 詞串出現次數在 k 以下時，我們會再乘上一個 d_a（discount coefficient factor）【8】，這是令部分詞串的機率少算一點，然後將多出的機率分給未出現的 n-gram 詞串（Unseen Event）機率。

上方這是語言模型在理論上的推導，而實際建立時我們也是依此建立我們的 LM。而且我也將與一般辨識器相同，利用大量的文字資料庫求取 Unigram、Bigram 與 Trigram 的機率。

5.2 詞典（lexicon）選擇

建立語言模型之前，我們必須先要有詞典（Lexicon）。詞典的目的是將所有常見的詞（Word）整理出來，而語言模型負責定義詞與詞之間的機率關係。在詞典的建立上，參考至 SoVideo 的詞典大小【6】，我們也以大約 6 萬詞的詞數當作標準詞典。

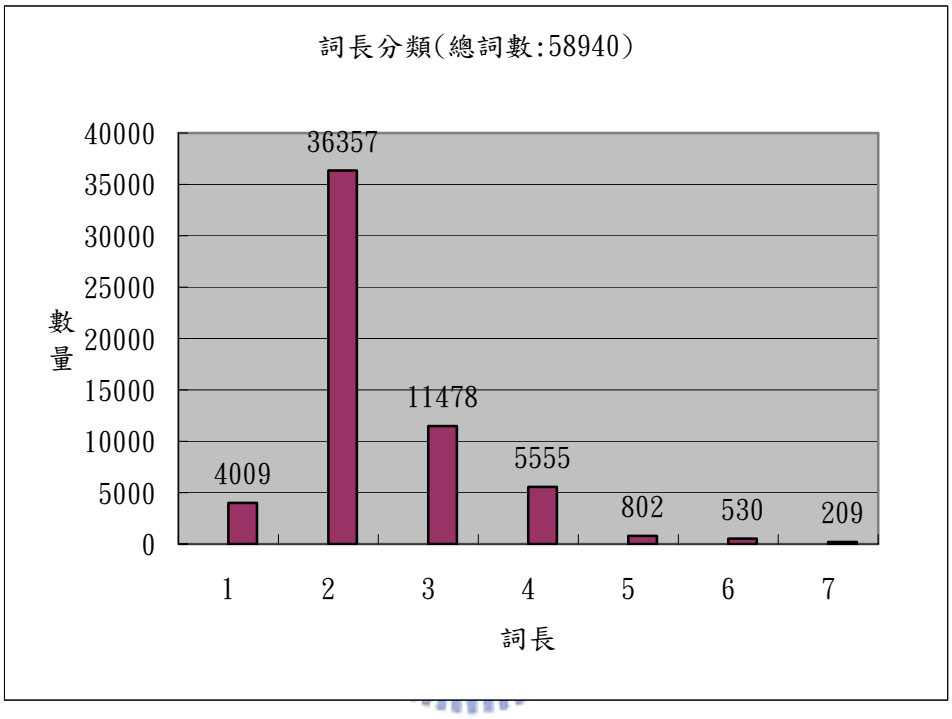
5.2.1 標準詞典建立

詞典的來源，我們是與清大資工所的張俊盛教授合作，請他們幫我們產生詞典，然後我們再針對需求做後處理。而產生詞典的方法是，先對大量的文字資料庫進行斷詞，然後將詞頻最高的詞排列，挑選最高的前 6 萬詞作為詞典。在這裡我們很感謝張教授的幫忙。

我們的文字資料庫來源有三個，分別為光華雜誌（Sinorama）、NTCIR 和中研院的平衡語料庫（Sinica Corpus），這些資料除了求得詞典外，也是作為 LM 的

訓練語料。

由於剛取得 6 萬詞的詞典時，裡面還夾帶英文、注音符號這些詞，所以我們會將這些詞剔除，並進行一些前處理的動作。最後實際的詞典大小為 58,940，在各詞長下的數量如下所示：



圖五-1　6 萬詞詞典詞長分布圖

表五-1　6 萬詞詞典詞長比例表

詞長	1	2	3	4	5	6	7
百分比	6.8%	61.7%	19.5%	9.4%	1.4%	0.9%	0.4%

詞典的平均詞長為 2.409（字 / 詞），二字詞的數量是最多的，而最長的詞為七字詞。

而在這裡我們對每個中文詞的表示採用「Big5 碼_漢語拼音」，例如若有一個詞稱為「表示」，則格式就為「AAEDA5DC_biao3shi4」。

5.2.2 廣播新聞新詞加入

- **Unknown Word**

當我們在建立 LM 時，訓練語料中一定會有不在標準詞典裡的詞，對於所有這類的詞，我們統稱為 OOV（Out Of Vocabulary）。在訓練 LM 時，這些 OOV 會被視為一個詞，或是一個 Class，稱為 Unknown Word，所以在詞典外我們也加上了一些 Unknown Word，如國語 411 音節、一般 OOV。

對於 Unknown Word 的訓練，假如 OOV 是由全中文組成時，我們會把詞（Word）退化成音節（syllable），於是 OOV 就會變成數個的國語 411 音節，因此就會多出 411 種中文單音節的詞，這樣做的目的是為了讓辨識系統當辨不出原詞典的詞時，就會退化到音節當成輸出結果，詳細做法我們會在 5.3.3 說明。

- **Particle 與呼吸聲**

在訓練語料當中，因為文章內容裡並不會標記 Particle、呼吸聲等其他語言學的現象，因此我們就無法統計這種詞的 Language model，為了使得我們訓練出的 LM 能接近廣播新聞的特性，因此我們會考慮廣播新聞中最常出現的 Particle 與呼吸聲，利用 PTSND 的訓練語料，建立出這兩個新詞的 Language Model。

5.3 語言模型訓練程序

在這裡我們訓練的是 Tri-gram 的 LM，由於使用大量的訓練語料，因此涵蓋的範圍比較廣泛，因此語言模型就比較普遍性（General），我們稱為「General LM」。在完成標準詞典 LM 訓練後，會再利用語言模型調適（Language Model Adaptation）的方法將 PTSND 中的訓練語料加入到 General LM，這樣可以產生

Particle 與呼吸聲的機率。最後我們在進行 Unknown word 的後處理。

5.3.1 Trigram 語言模型產生

- **訓練語料**

要建立 General LM，必須要有大量的文字資料庫。因此我們使用以下的資料庫當作訓練語料的來源為：光華雜誌（Sinorama）、NTCIR 和中研院的平衡語料庫（Sinica Corpus）。

光華雜誌內容為一般雜誌文章，總共蒐集了 1976 年至 2000 年的資料。而 NTCIR（NACSIS Test Collections for IR）是一個建立檢索系統的標竿測試集，內容包含數種不同的學科領域。平衡語料庫是由中研院所錄製的，內容包含多種主題，目的在於研究語言分析。

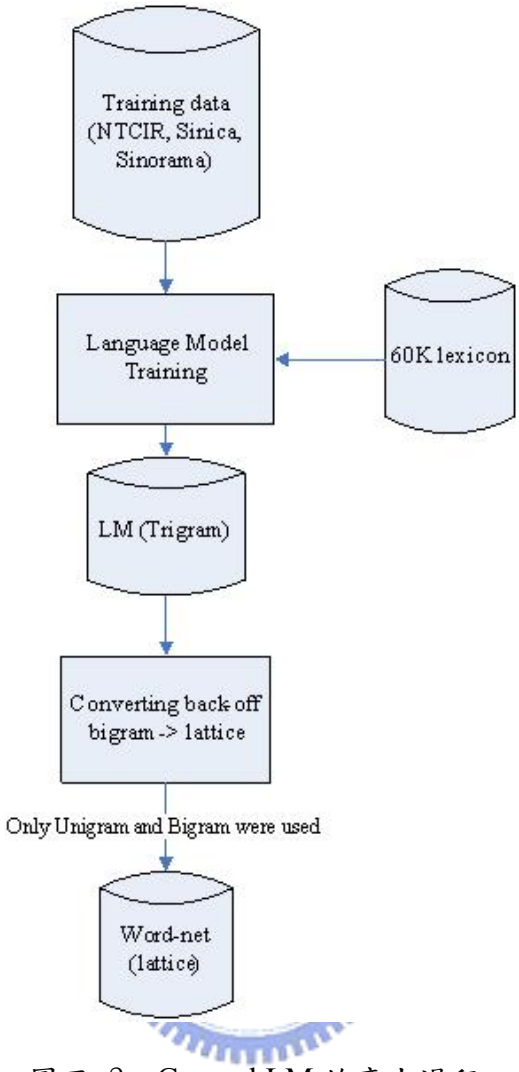
因此所有的訓練語料數量如下：

表五-2　General LM 訓練語料統計

訓練語料	詞數（Word）	字數（Character）
光華雜誌	11,348,465	15,669,241
NTCIR	59,862,541	83,116,970
平衡語料庫	5,816,309	8,078,119
合計	77,027,315	106,864,330

- **訓練過程**

而一個 Tri-gram LM 的訓練過程如圖五-2 所示：



圖五-2　General LM 的產生過程

藉由訓練語料與詞典，我們要求出 Bigram、Trigram 的機率，分別為 $P(w_i \mid w_{i-1})$ 和 $P(w_i \mid w_{i-1,} w_{i-2})$。

我們使用(5.4)的方法計算，總共得到約 9.07 百萬的 Bigrams 與 29.6 百萬的 Trigram。雖然已經求出 LM 了，但若要加入到辨識系統，還必須將 LM 轉成 Word-net，因為 Word-net 才是清楚的描述詞跟詞的轉移關係，因此最後一個步驟是將 LM 轉換 Word-net，目前只有使用到 Bigram 和 Unigram 機率。

5.3.2 語言模型調適

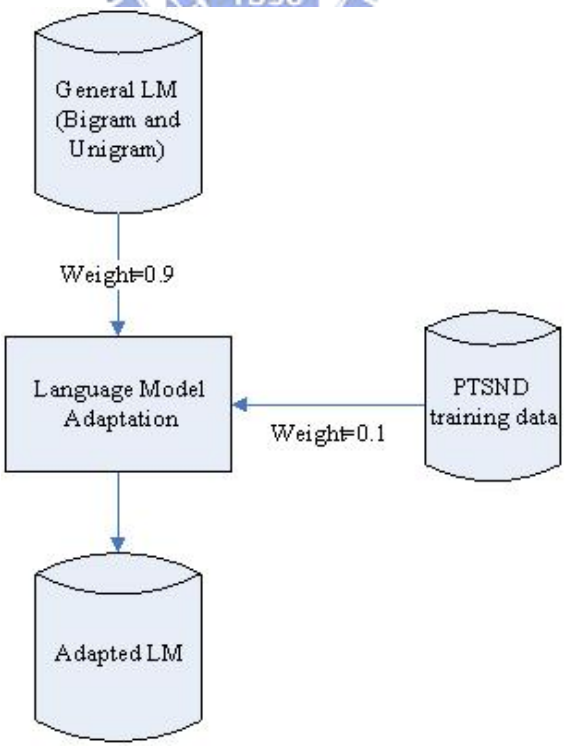
語言模型調適（Language Model Adaptation）是一個可以優化 LM 的方法（Refinement）。雖然我們現在訓練出的 LM 因為資料很大而具有普遍性（General），但是由於領域（domain）不同仍會使得 LM 並不準確，所以我們可以結合 General LM 資料多和 PTSND 訓練語料在同領域下的優點，去進行語言模型調適，則可以使得產生出的新 LM，更合乎廣播新聞的特性。

以一個 Tri-gram 的條件機率來看，我們進行調適後會變成：

$$P_{adap}(w_i \mid w_{i-1,} w_{i-2}) = \lambda P_{Gen}(w_i \mid w_{i-1}, w_{i-2}) + (1-\lambda) P_{PTSND}(w_i \mid w_{i-1}, w_{i-2}) \qquad (5.6)$$

其中，P_{adap} 是調適後的 Tri-gram 條件機率，P_{Gen} 是原本 General LM 的 Tri-gram 機率以及 P_{PTSND} 是在 PTSND 訓練語料中的 Tri-gram 機率。而 λ 是代表調適比重（Adaptation weight）。

我們進行語言模型調適的方塊圖如圖五-3：



圖五-3　General LM 的調適方法

實際的建立方法，使用 PTSND 中的訓練語料，其中原本的 LM 的調適比重為 0.9，而 PTSND 的調適比重為 0.1。語言模型調適的作法可以參考至【9】。PTSND 訓練語料的總詞數約為 24 萬，詳細統計如下：

表五-3　PTSND 訓練語料進行 LM Adaptation 資料統計

PTSND 訓練語料	中文詞數	Particle	呼吸聲	其他	合計
數量	221,390	3,480	14,673	858	240,401

對於 PTSND 訓練語料中，除了 6 萬詞典以外的詞，我們會把所有的 Particle，共同以一個 Class 來訓練，然後 Breath 也獨立訓練，剩下的詞若是純中文 OOV 則退化成音節，若是一般 OOV，則直接以 Unknown word 去訓練。

5.3.3 Unknown Word 處理

- **一般 OOV**

在詞典以外的詞我們統稱為 OOV，但是一般的 OOV 是指文章中出現一些非中文的詞類，如英文、注音符號等，或者是中英夾雜的詞，如「卡拉 OK」等，由於目前我們並未訓練出此類文字的聲學模式，因此對於這類的 OOV，我們均歸類為一般 OOV，並且也會用這些資料訓練出 Unknown Word 的 LM。而這類 Unknown Word 的 LM 我們也會使用於廣播新聞中的外國語文或者是 Garbage 的辨識上。

- **純中文 OOV**

當辨識系統找不到詞典的結果時，理論上這就是 Unknown Word，而我們希望這些 Unknown Word 能退化到音節（Syllable）表示，因此在訓練時，就必須

訓練出這些中文 411 個音節的這些 LM，並且也要 411 個音節加到詞典後方。

所以在訓練語料中，如果遇到這種純中文 OOV 的話，就會將其退化到音節。如下方的例子所示，「葉明蘭」這個人名不在詞典內，因此我們就會改以音節去進行 LM 的訓練：

Origin ： 我 是 **葉明蘭** 我們 明天 晚上九點 見
Training：我 是 **?ye ?ming ?lan** 我們 明天 晚上九點 見

圖五-4　純中文 OOV 在訓練語料退化成音節

但是，將詞（Word）退化成音節（Syllable）後，從統計中發現，在所有訓練語料當中，有大約 5%的詞不在詞典內，只有 95%的詞能在詞典中找到。這樣子就代表由 6 萬詞去共用 95%的資料，而 411 個音節就用使用了 5%的資料，這樣是太多的，容易使得辨識器都是 411 的音節。

因此我們會考慮當所有的字音節 Class 的影響，舉例來說，以音節「ye」為例，在這個音節 Class 內可以有{業, 葉, 夜, 頁}，但這些字（Character）全部都屬於同一種 Class，因此我們會再乘上一層機率來抵銷這個影響。

$$P_{new}(w_i \mid w_{i-1} = s_j) = P(w_i \mid w_{i-1} = s_j) * P(char \mid s_j) \tag{5.7}$$

其中，s_j 為 411 音節，j=1~411。$P(char \mid s_j)$ 為在 s_j 下出現一種字（Character）的機率，令：

$$P(char \mid s_j) = \frac{1}{c_j} \tag{5.8}$$

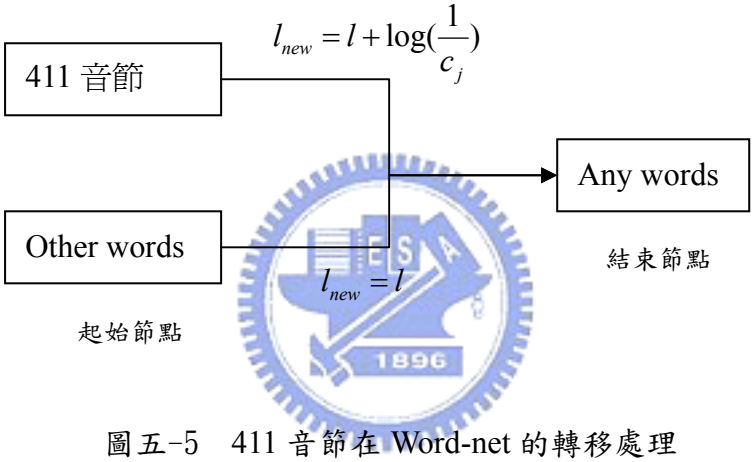
其中 c_j 是在 s_j 下的字之種類數。將(5.7)取對數後，相乘的關係就由相加取

代：

$$\log(P_{new}(w_i \mid w_{i-1} = s_j)) = \log(P(w_i \mid w_{i-1} = s_j)) + \log(\frac{1}{c_j}) \tag{5.9}$$

所以我們會在原本 Word-net 的部份，將 Class 機率的消除，方法就是在所有起始節點為 411 音節的轉移均乘上一個條件機率。

在 Word-net 上的處理，就是依照(5.9)的方式，在所有起始節點為 411 音節的轉移（transition），再加上一個 log 機率。



圖五-5　411 音節在 Word-net 的轉移處理

- **Particle 與呼吸聲**

在我們之前的詞典裡，因為沒有自發性語音的紀錄，因此無法統計呼吸聲、Particle 這類的資料，就無法計算 n-gram 的機率。但是在廣播新聞中，自發性現象很普遍，其中 Particle 與呼吸聲是最常見的，因此我們會利用語言模型調適的方法(Language Model Adaptation)來建立之。其中建立時， General LM 與 PTSND 的調適比重為 9：1。

5.4 實驗-加入語言模型的辨識分析

本實驗是採取已知環境的辨識，就是假設測試語料與訓練模型都在相同環境下（內場主播、外場記者或外場受訪者），測試語料也與 4.3.1 相同，這樣就可以觀察效能上的比較，而訓練的聲學模型也是直接使用依照環境分別建立聲學模型的版本。

而 LM 方面，我們會進行兩種實驗，一種是將 5.3 節中訓練出的 General LM 加入到辨識器的結果，另外一個是經過與 PTSND 訓練語料調適後的 Adapted LM，使用 Adapted LM 加入到辨識器的結果。所有的實驗都會加上 Beam Search，使用 Beam Search 可以加快 Viterbi Search 的速度，提高辨識的速度。

另外，因為加入 LM 後，辨識結果的基本單元由原本的音節（Syllable）變成詞（Word），因此我們的辨識結果都是以詞表示。由詞（Word）的結果，可以再轉成字元（Character）或者是音節（Syllable），例如以「蘋果」這個詞為例，假如以字元來看就是「蘋」、「果」；若以音節來看，則為「ping」、「guo」。所以除了觀察詞辨識率外，還可觀察字元辨識率與音節辨識率，並且最後再與無文法規則（Free grammar）時的音節辨識結果做比較。

當 LM 加入辨識器時，就不會像 Free grammar 一樣，進行辨識時只考慮聲學模型分數的最佳化，而是會在加上語言模型的分數影響。所以在本章的實驗，經過我們測試後發現，語言模型的分數統一乘上一個比重（Language Model Weighting），令其為 5，這樣辨識的結果會比較好，並且對每個加上 LM 的結果也比較公正。

5.4.1 使用 General LM 於辨識系統

我們使用大量的文字資料庫所建立的語言模型 General LM，詞典大小為

58,940 個詞，令語言模型提高比重（LM weighting=5），測試語料同 4.3.1。而我們的辨識結果為詞，但是為了能與未加入 LM 的系統比較，我們會將詞轉成音節，並且只觀察中文 411 音節的辨識率（Syllable recognition rate），辨識率如表五-4 所示。

表五-4　加入 General LM 的音節辨識率

環境	Del	Sub	Ins	Number	Accuracy
內場主播	2.68%	9.07%	0.44%	14,694	87.80%
外場記者	2.82%	13.39%	0.36%	9,225	83.44%
外場受訪者	8.96%	28.65%	2.55%	10,376	59.83%

使用語言模型後的辨識系統，與未使用時的辨識效能相差很多，由表五-4 與表四-5 的結果比較，音節辨識率在內場主播部份可以提高約 13%，外場記者部份可以提高 16.4%，而外場受訪者甚至可以提高到 19%左右。

5.4.2 使用 Adapted LM 於辨識系統

Adapted LM 是指將 General LM 與目前 PTSND 的訓練語料做調適（Adaptation）後的語言模型，我們令 General LM 調適的比重為 0.9，PTSND 訓練語料為 0.1，調適後可以使得 LM 比較接近廣播新聞的特性。語言模型比重也設為 5，只考慮中文 411 音節的辨識率，如表五-5。

表五-5　加入 Adapted LM 的音節辨識率

Outside 辨識	Del	Sub	Ins	Number	Accuracy
內場主播	2.48%	7.96%	0.45%	14,694	89.11%
外場記者	2.73%	12.37%	0.34%	9,225	84.56%
外場受訪者	8.25%	28.28%	2.67%	10,376	60.80%

兩種實驗都進行後，除了音節辨識率以外，我們還可以比較使用 General LM 與 Adapted LM 的詞辨識率（Word recognition rate）和字元辨識率（Character recognition rate）。此處辨識率的計算做法是，先以 6 萬詞典對測試語料的正確答案斷詞，若是不在詞典則以一字詞表示，再將辨識結果剔除 Particle 這類的答案，與正確答案做比對。則詞辨識率與字辨識率如表五-6、表五-7。

表五-6　詞（Word）辨識率比較

Outside 辨識	內場主播	外場記者	外場受訪者
General LM	68.48%	58.97%	35.46%
Adapted LM	75.21%	63.85%	38.83%

表五-7　字元（Character）辨識率比較

Outside 辨識	內場主播	外場記者	外場受訪者
General LM	81.66%	76.16%	50.65%
Adapted LM	85.31%	78.43%	52.40%

由上方結果可以發現，進行調適後的詞辨識率增加相當多，尤其以內場主播，外場記者特別明顯。這是因為主播與記者的發言比較符合文法規則，而外場受訪者屬於自發性語音，加入 LM 能提升的詞辨識率就沒那麼高。字元的（Character）辨識結果就是一般的文字，而字元的辨識率大小剛好界於音節與詞辨識率之間，並且在計算時，已包含從詞退化到音節的這些結果（如?wu），但也可發現使用調適後 LM 辨識結果較好。

- **實驗分析**
  - 加入 LM 後的辨識單元由音節變為詞，而我們希望當辨識器找不到詞典裡面的詞時，可以退化成音節作為結果。由辨識結果發現，當辨識答案

的文法不在一個 LM 計算時，就會辨識成音節輸出。如下方圖五-6 採用 General LM 辨識的例子，LAB 為原標示內容，REC 為辨識結果，由於「航速」這個詞不在詞典裡，因此辨識結果確實退化成音節來表示。

LAB： 飛機 航向 **航 速** 等 十 多項 資料
REC： 飛機 航向 **?hang ?su** 等 十多 項 資料

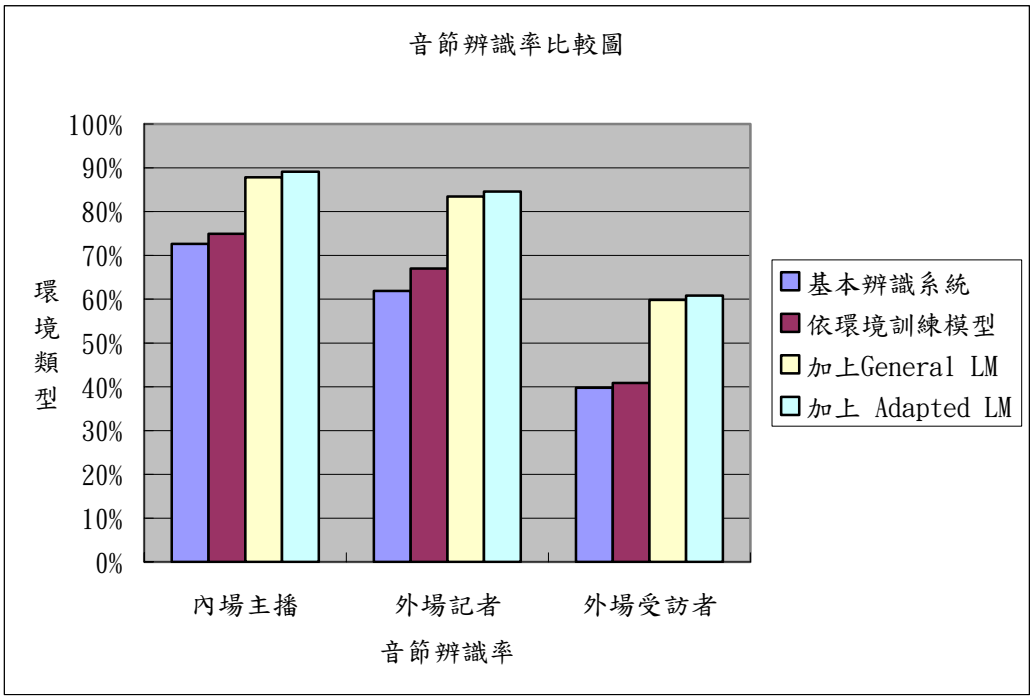
圖五-6 辨識結果退化到音節的範例

- 將本章節的實驗，與第二章的基本辨識系統和第三章依照環境重新訓練聲學模型的辨識結果比較，國語 411 音節（Syllable）的辨識率如下所示：

表五-8 各種方法的音節辨識率比較表

辨識率	內場主播	外場記者	外場受訪者
基本辨識系統	72.63%	61.88%	39.75%
依環境訓練模型	74.92%	67.00%	40.85%
加上 General LM	87.80%	83.44%	59.83%
加上 Adapted LM	89.11%	84.58%	60.80%

而以清楚的條列圖形化表示，如圖五-7：



圖五-7　各種方法的音節辨識率比較圖

- 我們可以發現，將聲學模型分開訓練後，辨識率約略可以上昇，但是幅度有限，這是因為使用 HMM 聲學模型模擬口腔有一定的限制。而加上語言模型以後，無論是在哪個環境辨識率都大幅提升，因為辨識系統加入文法的規則，辨識的結果也會比較符合文法，而將詞的辨識結果再轉成音節時，音節辨識率就可以大為提高了。

- 使用調適後的語言模型，詞辨識率提升較多，而音節辨識率提升較少，這是因為領域（Domain）的不同，若找不到文法規則，詞會退化到音節作為結果。在廣播新聞領域有其常見的名詞與文法，使用 Adapted LM 能加入了廣播新聞領域的特性，使得辨識系統比較合乎測試語料特性，故能提高詞的辨識率。

- 外場受訪者，在各別環境建立聲學模型後，音節辨識率提高並不明顯，這是因為外場受訪者語者很多，可以算是語者不相關（Speaker

Independent）的模型。而內場主播與外場記者因為人數有限，因此建立聲學模型後，辨識率會比較好。

第六章 結論與未來發展

6.1 結論

在本論文中，我們使用 PTSND（公共電視新聞資料庫），進行廣播新聞語音辨識的研究，從語音的特性、建立基本辨識系統，到系統改進以及語言模型的加入，有一個循序漸進的說明。我們將幾個主要重點分列如下：

（1） 廣播新聞語音不比一般 Read Speech 語音，而是比較近似於自發性語音（Spontaneous speech）特性，因此除了國語 411 語音以外，我們還建立了 Particle、呼吸聲和 Garbage model 等語音模型，這是為了能輔助其他聲學模型的訓練，使得基本辨識系統能更完整。

（2） 內外場環境的不同是廣播新聞的特色之一，因此依照各個環境建立聲學模型是必要的，我們針對不同的環境，分開訓練語料各自訓練模型，可以確實的提高辨識率。

（3） 加入語言模型（Language Model）至辨識器是語音辨識基本的需求，在前面的辨識系統裡都只是採用無文法規則（Free grammar），若能加入語言模型，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和 Unknown Word 的處理，從最後音節辨識的結果來看，加入語言模型真正能有效的提高辨識率。

6.2 未來展望

現時國內跟國外的辨識系統，均已經將語言模型（Language Model）、前後文相關模型（Context Dependent Model）均納入基本的條件。由於在我們進行實驗時，加入語言模型已經使得運算量會相當龐大，進行辨識時所費的時間是 Free grammar 的好幾倍，因此就沒有再進行前後文相關模型的實驗。假若在未來的辨

識系統裡，我們希望可以再把這方面加入。

廣播新聞裡，語者在講話的同時，時常會出現背景聲（Background Sound），而本論文的方向只在於無背景聲下的研究。而有背景聲的語音辨識效能一般比較低落【6】，同時也是一大挑戰，考慮背景聲的語音辨識是必要的，我們也可以針對 SNR 的大小、語者語音參數調適和聲學模型的調適去加強辨識系統的效能，這也是未來一個適合研究的主題。

良好的廣播新聞語音辨識器是建立正確檢索系統的關鍵，參考至 SoVideo 的檢索系統中【6】，若是沒有正確的語音辨識，容易使得使用者在搜尋到所需的新聞之後發現是錯的，因此在未來媒體爆炸的時代，為了有效的管理、擷取資訊，我們必須把基本辨識系統作好，也就是建立一個良好的語音辨識器。

參考文獻

【1】 Hsin-min Wang, *"MATBN 2002: A Mandarin Chinese Broadcast News Corpus"* ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003).

【2】 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

【3】 Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, *"Spoken Language Processing, Aguide to Theory, Algorithm, and System Development"*, Prentice-Hall, Inc.

【4】 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)"*

【5】 Kazuyuki TAKAGI, Shuichi ITAHASHI, *"Segmentation of Spoken Dialogue by Interjections, Disfluent Utterances and Pauses"*, In Proceedings of the ICSLP-96, pp. 697--700

【6】 Hsin-min Wang, Shi-sian Cheng and Yong-cheng Chen, *"The SoVideo Chinese Broadcast News Retrieval System"* International Journal of Speech Technology 7, 189-202,2004

【7】 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.

【8】 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

【9】 H. Meinedo, N. Souto, and J. Neto, "*Speech recognition of broadcast news for the european portuguese language*" in Proc. ASRU '2001.

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
breathe	<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