國　立　交　通　大　學

電信工程學系碩士班

碩士論文

自發性對話語音辨識之初步研究

Preliminary Study on Spontaneous Speech Recognition

研究生：徐文翰

指導教授：王逸如 博士

中華民國九十三年七月

自發性對話語音辨識之初步研究

Preliminary Study on Spontaneous Speech Recognition

研 究 生：徐文翰　　Student： Wen-Han Hsu

指導教授：王逸如 博士　　Advisor： Dr. Yih-Ru Wang

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

自發性對話語音辨識之初步研究

研究生：徐文翰　　指導教授：王逸如 博士

國立交通大學電信工程學系碩士班

中文摘要

在本論文中，我們建立一個自發性中文對話語音辨識基本系統架構，探討中文語音及自發性語料的特殊語音現象，如感嘆語(particles)、語音發音變異(uncertain sounds)、非語音聲音 (paralinguistic sounds)等，之聲學模型建立方法，使用中研院提供的八個雙人對話語料做實驗，獲得之音節辨識率為 43.33%。為使辨識系統更為完善，我們加入語言模型，並以語言調適的技術，使之更為優化，最後音節辨識率達到 53.93%，較基本系統提升了 10.6%。

關鍵詞：自發性中文對話語音辨識、感嘆語、非語音聲音、聲學模型、語言模型

Preliminary Study on Spontaneous Speech Recognition

Student： Wen-Han Hsu Advisor： Dr. Yih-Ru Wang

Department of Communication Engineering
National Chiao Tung University

Abstract

In the thesis, a basic spontaneous Mandarin speech recognition system is established. The study focuses on the acoustic modeling for 411 Mandarin base-syllables as well as some special phenomena of spontaneous speech such as particles, uncertain sounds, and paralinguistic phenomena. Performance of the proposed system was examined by simulations using a Mandarin dialogue speech database called MCDC (Mandarin Conversational Dialogue Corpus). A syllable accuracy rate of 43.33% was obtained. By adding a bi-gram language model with proper adaptation, the syllable accuracy rate increased to 53.93% which was 10.6% better than the baseline system.

Keywords: Spontaneous Mandarin speech recognition, Particles, Uncertain sounds, Paralinguistic phenomena, Acoustic modeling, Language model, MCDC

致謝

這兩年來，非常感謝王逸如老師帶領我進入了語音辨識的世界裡，並且時常苦口婆心的提點我作事的態度及方法，同時也要感謝陳信宏老師在語音大方向上的指引，在這所學到的東西及老師們的教誨我將銘記於心。

當然，還得感謝實驗室的同學們，阿德、嘉俊、性獸、小 z、小孫、俊良、祺翰和智合，大家各有各的特色，各有各的長處，經過這兩年的朝夕相處，我學到了很多，尤其是俊良，沒有你在 HTK 上的幫助，我想我論文必定無法如期完成。而學弟妹們，以後實驗室就靠你們了，加油。

最後，僅將此篇論文獻給我的家人及我所關心的人，因為有了你們的支持，我才能順利完成學業並邁向人生另一個旅途。

目錄

表目錄

圖目錄

第一章　緒論

科技長足的進步，e 世代的來臨，使得生活更為便捷，也不斷滿足人類無窮無盡的想像。就如同電影裡面的情節，語音辨認的技術的確可以讓機器不再只是冷冰冰的金屬，或者是一排排的按鍵。不論在通訊、電腦甚至於家電方面，只要提供足夠的語音辨識功能，它們都可以成為主人身邊俐落的好幫手。為了達成這樣的夢想，世界各地均不遺餘力的推動語音辨認的研究與發展，朝著更人性化的人機介面邁進。

1.1 研究動機

語音辨識的技術在這二十多年來的發展已有長足的進步，相關的應用方面如：語音控制、聽寫機乃至於資料檢索都已逐步趨向成熟，然而在自發性語音方面，由於其多出了許多特殊的現象，如：音節合併、口吃...等，導致其實用性受到近一步限制，以往在這方面的研究，並沒有一個完整且有系統的語料庫，導致無法進一步做分析，不過現代漢語口語對話語料庫(Mandarin conversation dialogue corpus)的完成，將使得對於自發性語音的研究變的更為方便，因此若能利用這個方便的語料庫，分析並改善其辨識率，未來將可以多加發展這方面的應用，為人們帶來更便利的生活。

1.2 研究方向

一般之中文連續音辨認系統可分成三個層次：(1)語音訊號前處理，(2)聲學解碼(acoustic decoding)，及(3)語言解碼(linguistic decoding)，下圖 1.1 為其方塊圖。

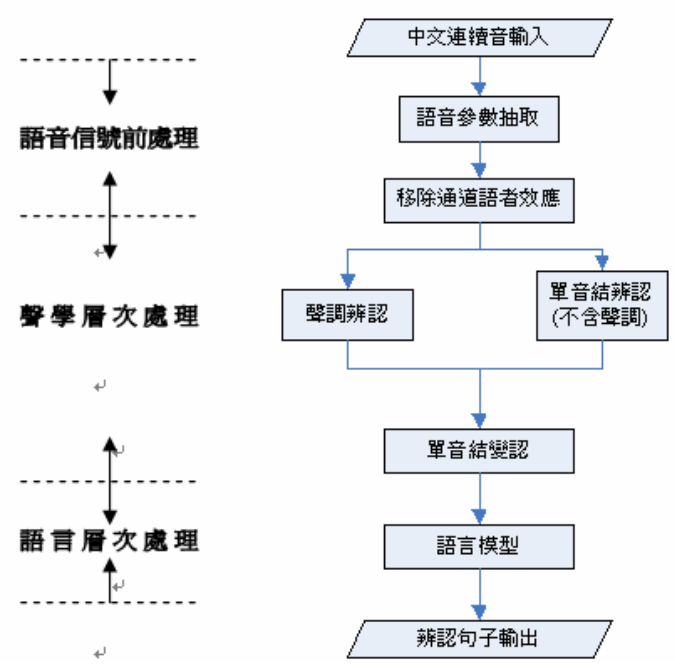


圖 1.1：語音辨認基本系統方塊圖

中文連續音是由字詞串接而成，由於中文的詞彙非常多，不適合作為辨認的基本單元，所以在聲學層次的辨認，一般都以單音節為辨認基本單位，至於將單音節串接成詞或句子的辨認問題則是由語言層次的語言模型來處理

本論文所著重的部份為聲學層次及語言層次的處理，而聲學層次上的處理僅探討中文連續音節的辨認，並未做聲調辨認方面的研究。

1.3 章節概要

本論文共分為六章，各章節編排如下：

第一章　緒論：介紹研究動機、方向及章節概要

第二章　現代漢語口語對話語料庫簡介：簡單介紹現代漢語口語對話語料庫，並根據此語料庫分析自發性語料與朗讀式語料的相異處，再跟據分析做語料庫的後處理

第三章　基本系統架構：說明系統辨識器中聲學模型之建立方法與訓練方式

第四章　自發性語音辨識器之使用及效能探討：利用上一章所建立的基本系統，觀查其效能，並對辨識錯誤的部份做分析，找到改進的方法

第五章　加入語言模型之自發性國語辨識器：加入兩種不同性質的語言模型，探討比較其效能。

第六章　結論與未來展望

第二章　現代漢語口語對話語料庫介紹

做任何語音方面的研究，都需要一套合適的語料庫，否則所產生出來的研究結果將只是象牙塔裡的產物。本論文主要的研究方向為自發性語音的分析，因此必須找一套具有自發性語音特性的語料庫，目前國內符合此種特性且較為完整的即為現代漢語口語對話語料庫(Mandarin conversational dialogue corpus)，以下我們簡稱 MCDC，此語料庫不同於一般朗讀式(Read Speech)資料庫，如 MAT(Mandarin speech data across Taiwan)。MCDC 語料的語音是較自發性、較口語化的對話語料(dialogue)，所以非常的適合用於作自發性語音的初步研究。

本章將會簡單介紹這套語料庫，並以此語料庫為例，說明自發性語料的特性，及其與朗讀式語料間的差異，而對於這些差異上的了解將有助於建立基本辨識的系統。

2.1 Mandarin conversational dialogue corpus 簡介

MCDC 語料是由中央研究院語言學研究所籌備處曾淑娟博士【1】於二 00 一年夏天錄製完成，語料發音人的選取是依據 16-25 歲、26-35 歲以及 36-45 歲三大年齡層，由台北市市民中隨機抽樣選出，最後選取出 16 位參與錄音，共 9 位女性，7 位男性，兩兩互相談話，發音者雙方在錄音前是未曾謀面的，為確保不陷入無話可談的窘境，除了一開始的自我介紹外，錄音前有準備一些主題供發音人參考，發音人可以任意選擇所提供的主題或是任何其他的話題與對方聊天，最後共錄下八個對話，平均每個對話大約是六十分鐘，以下是對話內容的總表：

表 2.1：對話主題總表

對話序號	長度(分)	發音人：性別(年齡)	對話主題
mcdc-01	61	女(29)，男(25)	工作、休閒活動、經濟、開車
mcdc-02	63	女(37)，男(35)	休閒活動、經濟、工作、性別、政治
mcdc-03	61	女(16)，女(17)	家庭、學校、購物、生涯規劃、明星
mcdc-05	63	男(40)，女(46)	工作、家庭、社會階層、保險、歷史、省籍情結、名人
mcdc-09	66	女(30)，女(35)	工作、旅行、生活態度、環保、健康
mcdc-10	54	男(35)，男(23)	電影、政治、軍隊、捷運、學校、經濟
mcdc-25	55	男(43)，女(45)	交通、工作、小孩、旅行、電腦、管理
mcdc-26	46	女(37)，男(24)	工作、求職、家庭、休閒活動、車禍、學英文、婚姻

2.1.1 音檔錄製格式

錄音設備採用 SONY TCD-D10 PRO II DAT 的數位錄音機，使用 Audio-Technica ATM 33a 手持式麥克風。以取樣率 44.1 kHz 將兩位發音人的語料分別錄於左右聲道，共錄製八段對話，每個對話儲存一個聲檔，為了使處理語料內容更有效率，再利用軟體將之分割成更小的雙聲道音檔，方式是在長度約三分鐘左右找到一個清楚可辨的停頓切開，因此 MCDC 的音檔主要是八個對話，每個對話再細分成約每 3 分鐘一個的音檔，其格式列在下面

Sample rate： 44.1kHz

Resolution： 16bits

Channel： Stereo

Format： WAV (Microsoft Windows Wave File)

2.1.2 文字標示格式

一般的語料庫除了音檔之外，免不了都會有文字標記(Transcription)的部份，只是各個語料庫格式都有所不同，在 MCDC 這個語料庫中所使用的是類似 XML 的語法來進行文字標記，結構上是以一個 sub-turn 做為一個單元，下面舉一段例子來做說明

Example：

<voicefile>d:\分割完成的檔\stereo_01\mcdc-01-01.wav
<speaker>MISC-08-male-25
<start>000000
<end>009514
<translator>Fen
<chinese>
EI (clear throat) 你好我姓賴請問一下貴姓 (hiccup) (breathe)
</chinese>
<english>
EI (clear throat) ni3 hao3 wo3 xing4 lai4 qing3 wen4 yi2 xia4 gui4 xing4 (hiccup) (breathe)
</english>
<comment>
</comment>

可看出標註內容包括聲音的檔案<voicefile>、發音者<speaker>、檔案起迄時間<start><end>、檔案標記員<translator>、語料中文內容<chineses>、語料漢語拼音內容<english>。

2.2 自發性語料庫的特性

由上節對 MCDC 語料庫的簡介，我們已知其所蒐集的語料是非常貼近日常對話的，也就是語音是自發性的，因此若要對此語料庫作研究，必須先了解自發

性語料與朗讀式語料間特性的差異，在 MCDC 語料庫中，標註了許多這方面特性的資訊，如：拖長音、音節合併、發音偏差、不確定音...等，而標註種類詳細的介紹，請參考中文詞知識庫小組為此語料庫所出的一本技術報告--現代漢語口語對話語料庫標註系統說明【1】，下面僅對於本論文所使用到的特性做詳細的介紹。

1、 語助詞 (marker)

說話者本身在語流中慣用的插用語，這些習慣插語有其基本詞彙意義。但在語流中習慣插語已不保有其原有的完整語意。而較具語用功能。例如，作用於口語中說話者意欲保有其說話權且又需緩衝時間去思索組織其想說的話的句子，此時習慣插語”那”便常被使用。

原始句例：室內就是一小間一小間嘛那露天就是大家一起

(“那”屬語流中習慣插語，已不保有指涉某物的意義)

2、 感嘆詞 (particle)

不具標準語意的感嘆詞，其與用成分居多如回應或同意。語流中出現的感嘆詞有四類，一、有相對應國字的感嘆詞；二、無相對應的國字的感嘆詞；三、源於台語的感嘆詞，如；四、其他感嘆詞，如嗯哼哼，下面列出一、二類的例句

(1) 有相對應國字的感嘆詞

原始句例：去什麼富基漁港阿那些

(專有名詞”富基漁港”後面跟著一個不具標準語意的感嘆詞”阿”)

(2) 無相對應國字的感嘆詞

原始句例：EI 你好我姓賴請問一下貴姓

(句子一開始，即以一個不具標準語意的感嘆詞”EI”作為起頭)

3、 非語音現象(Paralinguistic Phenomena)

分成兩類，一、凡非語音但確定由人所發出的聲音，包括笑聲、咳嗽聲、呼

吸聲、吐氣聲.....等，和其他口腔發出無法辨識的聲音等等。二、非語音且確定非人所發，下面列出兩類的例句。

(1) 原始例句：大概是我們運氣不好

(講完，即笑)

(2) 原始例句：NHN

(在說話者發出感嘆詞前，有一個敲到麥克風的聲音)

4、 不確定字/音(Uncertain)

一、標記員根據前後語意，可以猜測出大概的語意內容，但無法百分之百確定。二、標記員無法根據語意猜測出對應字詞，但可以漢語拼音清楚紀錄出其發音。下面列出一些的例句。

原始例句： [fa1]因為大概離台北市區比較遠一點

(在明確的語言內容前有一個不確定音[fa1])

5、 外國語言(Foreign)

語料庫中會出現一些外國語言，如英語、日語

原始例句：真正通化街那一條不是有 HANGTEN NA GIORDANO 那一些

(說話者在講到該衣服品牌時適用該品牌的英文名字作為指稱)

以上為本論文於基本辨識系統中所要處理的一些現象，希望能藉由處理這些現象來提升自發性語料中 411 音節的辨識率，處理的方法我們將於第三章做進一步討論。

2.3 MCDC 語料庫的後處理

語料庫中的音檔是雙聲道音檔，兩個語者分別使用左、右兩個聲道錄製，因

此首先必須將之轉換成單聲道，並在論文中我們辨識是以 sub-turn 為單位，因此再根據文字標註檔中的檔案起迄時間，將原本三分鐘的音檔，切割成以 sub-turn 為單位的音檔，做過以上處理後，為減少資料量及配合原本實驗室的辨識器，再將原本 44.1kHz 的音檔，down-sampling 至 16kHz，下表 2.2 列出處理過後之音檔時間統計表。

表 2.2：音檔時間及 Sub-turn 數統計表

對話序號	音檔長度(分鐘)	Sub-turn 數量
mcdc-01	87.1	867
mcdc-02	71.5	1094
mcdc-03	66.3	981
mcdc-05	89	865
mcdc-09	83.8	671
mcdc-10	70.9	513
mcdc-25	65	669
mcdc-26	66.7	833
總合	600	6488

用於研究的語料庫除了須要音檔外，還需要每個音檔的文字標註內容，否則無法訓練聲學模型，也無法做辨識，取得的方式為由文字標註檔中抽取出我們想要的資訊，處理方式和音檔同樣的以 sub-turn 為單位，將語料中中文內容、語料漢語拼音內容及我們所需要之關於自發性語料特性的標註保留下來，有關自發性語料特性的標註，依據 2.2 節所討論之特性，將之分為三大類，其類別名稱，1、Particles(包含感歎詞及語助詞)；2、Paralinguistic phenomena(包含非語言現象)；3、Uncertain(包含不確定字/音)，經過文字抽取處理過後的內容，下面以一個例子做說明

例句： EN NA 請問怎麼稱呼您 @吸氣聲

文字檔：EN NA qing3 wen4 zen3 me5 cheng1 hu1 nin2 @INHALE

其中全大寫之 EN NA 為感嘆詞，@起始的如：@INHALE 為吸氣音屬於非語言現象

對於文字處理過後，其統計表列於下表 2.3

表 2.3：MCDC 語料庫之文字統計

	411 syllable	Particle	Paralinguistic phenomena	Uncertain
字數	116657	10386	12199	8324
百分比	79%	7%	8.2%	5.6%
總字數	147548			
Sub-turn 數	6488			

下圖為觀看 sub-turn 中 syllables 數量的分佈圖

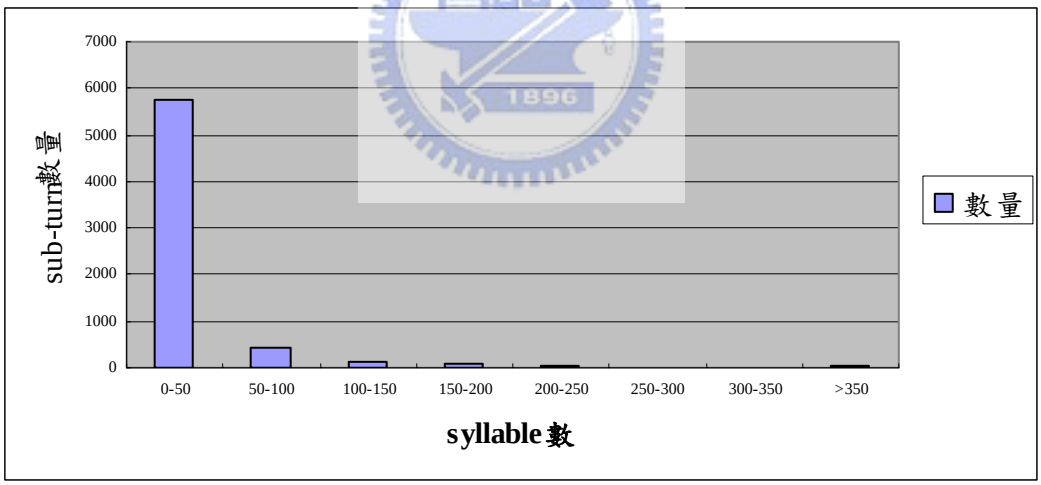


圖 2.1：以 Sub-turn 為單位之音節數量分佈圖

根據以上處理過後的音檔及文字內容後，即為本論文在研究上所使用的語料庫了。

第三章　基本系統架構

本論文的辨識方法，是使用隱藏式馬可夫模型(Hidden Markov Model)，因為 HMM 是目前最適合用於模擬口腔聲道變化的過程，也是目前最受到歡迎的方法。我們基本架構的建立，除了使用實驗室自行開發的 HMM 辨認程式外，還有採用英國劍橋大學開發的工具 HMM Tool Kit(HTK)，目前使用的版本為 HTK3.2.1。【2】

3.1 語音參數設定

我們求取的語音參數是採用梅爾倒頻譜參數(Mel-Frequency Cepstral Coefficents)，而語音參數求取時所使用之系統參數設定如下，1、預強調(pre-emphasis)，語音信號通過一個 $H(z)=1-0.97z^{-1}$；2、音框化中音框的長度設定為 32ms，10ms window shift；3、做 FFT 時使用漢明窗(Hamming window)；4、使用 DC-Bias removal；5、Filter Band 範圍為 0~8KHz，共 24 個 channels；6、求取 Delta-MFCC 及 Delta-Delta-MFCC；7、做 CMN(Cepstral Mean Normalization)，求取的 MFCC 參數為 13 維度，加上一維與二維的變化量，再扣除第零階的能量，因為它並不重要，最後我們得到的是一個 38 維度的語音參數向量。

3.2 初始聲學模型的建立

MCDC 語料庫在一開始是沒有切割資訊的，若要建立其初始的聲學模型，可使用 Uniform segmentation 來做粗略的切割，但是此方法用在較長的句子時容易產生錯誤蔓延，在語料庫中有許多超過百字的長句，可知此法並不適用。因此為了得到較好的切割資訊，我們的作法為參考其他已有模型之語料庫，利用其模型來做已知字串切割，如此可得到較好的切割資訊，然後再利用此建立本身語料庫的初始模型。

3.2.1 Force alignment

下圖 3.1 為我們做已知字串切割的流程圖

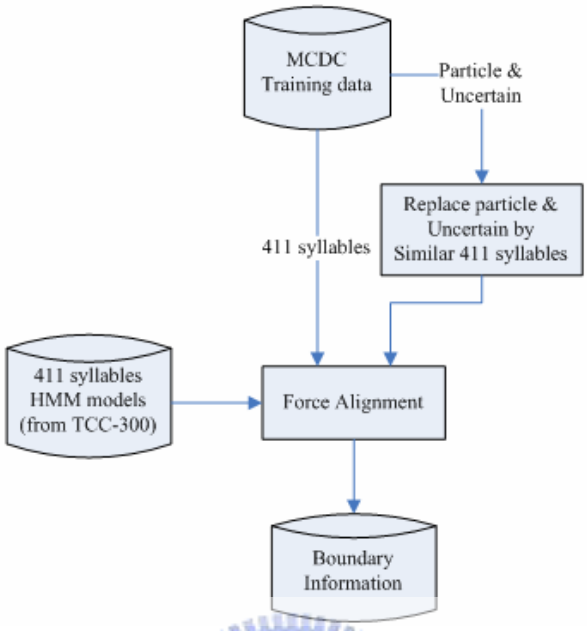


圖 3.1：已知字串切割之流程圖

要做已知字串的切割，必須要有聲學模型及待切割音檔之標示內容，在此我們使用朗讀式語料(TCC-300)所訓練之聲學模型，由於模型是由朗讀式語料所訓練的，因此僅可處理 411 音節，由上一章可知道，MCDC 音檔之標示內容中有我們所留下之許多代表自發性語料中特殊現象的標註，包含 Uncertain、Particle、Paralinguistic phenomena 三大類，下面將介紹本論文對於這些非 411 音的處理方法。

1、Uncertain

對於 Uncertain，可找到與其相近的 411 音，因此先用相近的 411 音取代之，以相近音的聲學模型來切割出其位置，待得到較精準的切割位置後，再建立自己的聲學模型。

2、Particle

與 Uncertain 的處理方式相同，以相近的 411 音當作其參考答案做切割。

3、Paralinguistic Phenomena

由於此種類之標註無法找到與其相近之 411 音，因此若句中含有此類標註，則無法用任何聲學模型來訓練其切割位置，造成整個 sub-turn 無法使用，如此做會造成可得到的切割資訊量非常的少，因此我們的作法是對語料庫中大量出現的 Paralinguitic Phenomena 現象，以人工切割方式得到少量切割位置，每個約切出 50 個音，並進而訓練出其初始模型，如此即可增加不少語料可用於作已知字串切割，表 3.1 列出人工切割之 Paralinguistic Phenomena

表 3.1：使用人工切割位置之 Paralinguistic Phenomena

標註	實際發音
@BREATHE	呼吸音
@INHALE	吸氣音
@EXHALE	呼氣音

根據以上對於非 411 音於標示內容所做的處理，我們統計可用於做已知字串切割的資料，表示於下表 3.2

表 3.2：用於訓練初始模型之資料統計

	411 syllable	Particle	Paralinguistic phenomena	Uncertain
字數	18046	3206	64	862
百分比	80.24%	7.3%	6.12%	2.64%
總字數	22178			
Sub-turn 數	2655			

比較表 2.3 及表 3.2，可以清楚了解到可用於訓練初始模型的資料量是非常少的，這代表著在語料中完全都是 411 音的 sub-turn 的數量是少之又少的，同時也說明了為什麼自發性語料比朗讀式語料難以處理的原因。

3.2.2 初始模型的建立

在我們求得 411 音、Particle、Uncertain 的端點切割資訊之後，我們會進行已知位置的初始模型訓練，HTK 中建立之方法為固定音節端點，對音節中之狀態做 Baum-Welch 參數估計，然後在放開音節切割位置，對整句話做 Baum-Welch 參數重估，下圖 3.2 為初始模型建立方法之方塊流程圖

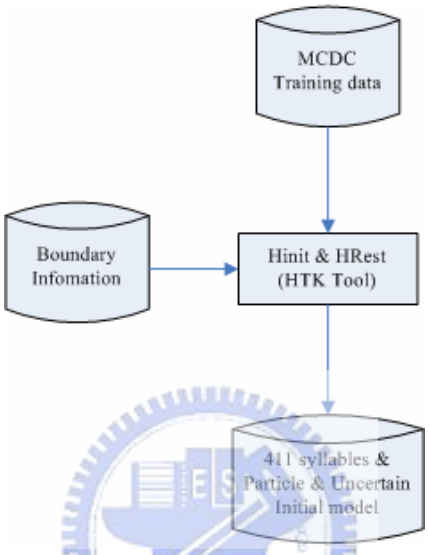


圖 3.2：初始模型建立之流程圖

而模型的初始設定列於下表 3.3

表 3.3：初始模型設定

標註類別	狀態數量	Mixture 數量
聲母	3	根據表 3.2 中的資料量，依據式 3.1 去計算，若 Mixture 數>32 以 32 計
韻母	5	
Particle	3	
Uncertain	3	
Silence	3	64
SP（Tie to silence）	1	64

Mixture 數量計算如下：

$$N_s = \min\left(\lfloor n / 50 \rfloor , 32 \right)$$

其中 N_S ： State Mixture no.

n ： 資料總 Frame 數 (3.1)

由於上述具切割資訊之語料只有整份語料的約 1/7，因此許多聲學模型會因為資料量過少而訓練不出來，這些建不出來的模型，將於 3.2.4 中另做討論，下表 3.4 列出我們所訓練出來的模型總數

表 3.4：初始模型總數統計

	411 syllable	Paralinguistic Phenomena	Particles	Uncertain
可建出初始模型的數量	64 RCD initial 34 final	3	27	3

由上面的介紹我們可建立出具足夠資料之 411 音、Uncertain、Particle 的模型，但在 MCDC 語料庫中，還有許多的現象我們無法以上述之方式建立出初始模型，對於在語料庫中極少出現之語言現象，我們會共同建立一個特殊共用模型來描述尚未建立模型之語言現象【3】，稱為「Filler Model」，下一節將介紹此種模型的建立方法。

3.2.3 Filler Model 的建立

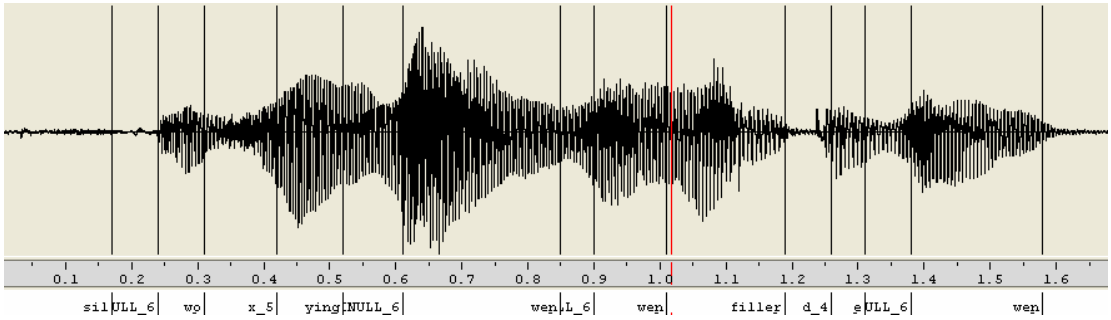
Filler Model 初始模型的訓練是利用我們在 3.2.1 中產生的切割資訊，將所有 non-silence 的資料一起去訓練而得，即可得到一個 Variance 非常大的 Gaussian distribution，模型的設定列於下表 3.5，而之後訓練 Filler model 之語料為那些極少出現的音，而不是初始模型之訓練方式

表 3.5：Filler Model 的設定

模型類別	狀態數量	Mixture 數量
Filler	3	32

當我們在進行語音切割或辨識時，正常的語音在 Filler Model 跟其他的語音模型相比，Filler Model 的辨識分數會顯得較小而不被選擇，但當遇到如咂嘴聲…

等這些在語料庫中極少出現的語音現象時，使用正常語音模型所得到的分數會較小，但會被分佈較廣的 Filler Model 所切割出來。下圖 3.3 為一個我們在做已知字串切割時利用 Filler Model 來取代一個極少見的 Uncertain 的實例



音檔下面那排文字為音檔之文字標註

圖 3.3：已知字串切割實例

由圖中我們可看出 filler 的位置在 wen 及 d_4 的中間，此 filler 代表的是一極少出現的 Uncertain，其切割位置是相當不錯的，由此可知利用 Filler Model 來取代那些少數音，使之不去影響其他模型的訓練，這個方法是可行的。

3.3 聲學模型的訓練

在 3.2.2 中我們介紹到如何建立 411 音、Particle、Uncertain 的初始模型，有了這些足夠的初始模型後，我們將可用所有的訓練語料來訓練出更精確的聲學模型，首先先選取出用於訓練模型之語料，我們在八個對話語料中各平均選取 9/10 的 sub-turn 作為訓練語料，如此較能涵蓋所有 Speaker 的語音特性，所有訓練語料音檔的時間統計列於下表 3.6

表 3.6：訓練語料音檔統計表

句數（sub-turn）	時間（小時）
5765	8.77

訓練語料文字資訊部分，由於 Uncertain、Particle、Paralinguistic phenomena 都具有非常多的種類，尤其 Uncertain 為最，因為它是 transcriber 就其所聽到的

語意直接使用漢語拼音標出其可能之組合情況，所以非 411 的種類，是非常之多的(在整份語料中共出現 1000 種非 411 音節之標音)，而在這麼多的種類中有些在語料庫中出現次數非常的少，對於這種出現次數很少的，我們並不訓練其模型，而將之用於訓練 Filler Model，如此這些音雖然沒有自己的模型也不會去污染別的音，根據此我們統計所有訓練語料之文字資訊，列於下表 3.7

表 3.7：訓練語料文字統計表

	411 syllable	**Particles**	**Paralinguistic phenomena**	**Filler**	**Uncertain**
字數	101464	9235	7744	4665	3342
百分比	80.24%	7.3%	6.12%	3.69%	2.64%
總字數	126450				
Sub-turn 數	6011				

有了更多的語料，我們就可以訓練出更多資料足夠的模型，之前由於在建立初始模型時訓練語料的不足，造成一些模型建立不出來，使得在利用較大的語料來訓練時，很多音是沒有其初始模型的，下面表 3.8 我們列出此種音建立其初始模型的方法。

表 3.8：無法建立初始模型之處理方法

所缺模型種類	**作法**
聲母/韻母	Copy 其相近之聲母/韻母的模型
Particle	Copy 相近 Particle 的音之模型，若找不到相近音則以 Filler Model 當做其訓練時的初始模型
Uncertain	方法同於 Particle 之處理方法
Paralinguistic Phenomena	以 Filler Model 當做其初始 Model

下表 3.9 列出所有我們建立出來的模型

表 3.9：所有建立出之模型

	411 syllable	Paralinguistic Phenomena	Particles	Uncertain
可建出初始模型的數量	100 RCD initial 40 final	15	40	76

模型之 State 數的設定同於初始模型，參考表 3.3，但 Mixture 數量的計算，就有所不同了，一個是依據表 3.2 的資料量來求得，一個是依據表 3.7 的資料量來求得。

當所有的模型建立完成後，即可對於所有的訓練語料用於訓練，我們將會一直訓練至穩定為止，而判定穩定的方法如下：

P = Average log Prob. per frame

P’ = Last Average log Prob. per frame

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

第四章　自發性語音辨識器之使用及效能探討

上一章我們已介紹了基本系統架構的建立，本章將探討在此基本系統下的辨識效能，並對於辨識錯誤的部份做分析，以期能找到改進的方法。

4.1 測試語料

上一章中我們使用了每一個對話之 9/10 的語料用於訓練模型，而剩下的約 1/10，再將 1/10 Sub-turn 中，整句話裡無 411 音的語料去除，其餘的即為實驗用的測試語料，由於測試語料中包含每一個 Speaker 的語料，因此不是辨識單一語者，而是多個語者(multi-speaker)的辨識，最後統計測試語料其總時間，列於下表 4.1

表 4.1：測試語料時間統計

句數（sub-turn）	時間（分鐘)
447	65.18

測試語料中的文字資訊，亦為計算辨識結果的答案，在訓練語料中我們可將過少無法獨自建立聲學模型的音，用於訓練 Filler Model，但在辨識時，若將過少的音以 Filler 當作其答案，由於 Filler Model 為一個 Variance 非常的大的模型，可預期的是這些音的辨識率將非常的差，因此我們必須決定無模型的這些音，其答案應該是什麼，在本論文中的處理方法列於下表 4.2

表 4.2：無模型之音節於參考答案之處理方式

種類	處理方式
Uncertain	由於 Uncertain 其性質近似 411 音，因此以相近 411 音當作其答案
Particle	以近似 Particle 的音當作其答案
Paralinguistic Phenomena	以 Filler 當作其答案

經由上面的處理方式，我們將文字資訊的統計，列於下表 4.3

表 4.3：訓練語料文字資訊統計

	411 syllable	Particles	Paralinguistic phenomena	Uncertain pronunciation
字數	14996	936	1254	516
百分比	80.24%	7.3%	6.12%	2.64%
總字數	17702			
Sub-turn 數	447			

4.2 辨識率計算方法

對連續音而言，由於辨認結果所得的音節總數，未必會等於正確的音節總數，所以辨認的結果除了「替代型」(Substitution)錯誤以外，還會包含一些「插入型」(Insertion)以及「刪除型」(Deletion)錯誤。我們對於替代型、插入型、刪除型錯誤的認定方式，即是已得到最佳辨認率為準則，其具體作法，則是利用動態規劃法，將正確音節字串與辨認結果做一對應，進行錯誤類型的認定，找到一條可得到最佳辨認率的路徑，在以下列計算辨識率及包含率

辨識率 = (正確音節數 - (Sub + Ins + Del)) / 正確音節數

包含率 = (正確音節數 - (Sub + Del)) / 正確音節數

本論文中還分成兩種情況來探討辨識率，一種是將非 411 音及 411 音的辨識情況都考慮在內的辨識率，下面我們稱之為整體辨識率，一種是將辨識結果及參考答案中的非 411 音都去除，只考慮 411 音辨識結果的辨認率，下面我們稱之 411 only 的辨識率。

4.3 基本實驗

首先我們需要決定的是該使用那一些聲學模型用於辨識，這對於朗讀式語料而言，是較為簡單的，只須使用 411 音節模型即可，然而對自發性語料而言，由前面的介紹，可知它多出許多了朗讀式語料中所沒有的現象，因此必須決定是否使用 411 音節外的模型加以辨識，為此我們利用一個實驗來驗證並分析。

實驗分兩種情況討論：

1、在辨識系統中只使用 411 音節及 Filler model。

2、將所有用於訓練的模型，也就是除了包含上種情況的模型外，再加入 Particle、Uncertain、Paralinguistic Phenomena 這些為了自發性語料中的特殊現象所建立的模型。

下兩節將分別為此兩種情況做討論

4.3.1 實驗一

本實驗我們使用了 411 音節、Filler Model 加入辨識器，加入 Filler Model 的目的，在於期望能以一個 Variance 非常大的模型，來描述所有非 411 音節，由於辨識結果只可能是 411 音節和 Filler，因此必須將參考答案中非 411 音都改為 Filler，以此觀看之實驗辨識結果，以及 4.2 節所描述之辨識率計算方式，分成整體及 411 only 的情況，列於下表 4.4

表 4.4：使用 411、Filler Model 於辨識器中之辨識結果

	包含率	辨識率	Del(%)	Sub(%)	Ins(%)
整體	41.84	39.96	18.2	39.85	1.98
411 only	49.15	39.36	11.52	39.3	9.8

為了能更清楚了解表 4.4 的結果，分析 Confusion Matrix 是必須的，在此我們分為兩個類別對表 4.4 中整體辨識率來做分析，非 411 音為一個類別，411 音為一個類別，表 4.5 列出統整過後的情況

表 4.5：表 4.4 中辨識結果之 Confusion Matrix 分析

辨識結果 / 正確答案	非 411	411	Deletion	類別總字數
非 411	1.75%	59.17%	39.07%	2706
411	0.14%	85.39%	14.45%	14996

由表中可以看出，表 4.4 中 411 Only 辨識率的 Insertion 偏高是由於非 411 音被辨識成具有較精確模型的 411 音所造成，而 Deletion 的降低是因為非 411 音的 Deletion 較 411 音為高之故，由表 4.5 可看出其高達將近 40%，大約是整份語料的 6%左右。

4.3.2 實驗二

本實驗的考慮，是將所有用於訓練的模型，也就是為了自發性語料中特殊現象所建立的模型，皆用於辨識系統中，因此與實驗一的不同點，只在於我們利用較為精確且較多的模型來描述非 411 音節，期望以此來可得到更好的效果，實驗的辨識結果，列於下表 4.6

表 4.6：加入自發性語料中特殊模型於辨識器之辨識結果

	包含率	辨識率	Del(%)	Sub(%)	Ins(%)
整體	47.18	41.97	12.5	40.27	5.20
411 Only	48.28	42.83	14.1	37.6	5.4

為了能更清楚了解表 4.6 的結果，進一步對於 Confusion Matrix 做分析，和實驗一同樣的，我們分成兩種類別，非 411 音節與 411 音節兩類，分析統整的結果，列於下表 4.7

表 4.7：表 4.6 中辨識結果之 Confusion Matrix 分析

辨識結果 / 正確答案	非 411	411	Deletion	類別總字數
非 411	55.4%	33.85%	12.8%	2706
411	5.35%	84.05%	10.5%	14996

對於實驗一及實驗二的各項結果之分析我們將細列於下：

1、首先分析整體的辨識率，也就是將非 411 音及 411 音的辨識結果皆考慮進來，比較表 4.4 和表 4.6，可以發現實驗二的辨識率是較高的，大約比實驗一的結果提升了約 2%，我們進一步比較表 4.5 和表 4.7，表 4.5 中非 411 音的辨識結果不是被辨識為 411 音，就是造成刪除型錯誤，若是造成刪除型錯誤這是較無關係的，因為這些音原本就不希望被辨識出來，但被辨識為 411 音，這就較為嚴重了，而在表 4.7 中我們可以明顯感覺到這種情況降低了許多，由 59.17%降至 33.85%。

2、對於只考慮 411 音辨識的分析，我們由表 4.4 及表 4.6 可看出， Insertion 明顯的降低了，由 9.8%下降至 5.4%，這代表著由於實驗二中對非 411 音描述的較為精確，可使非 411 音辨識為 411 音的情形大為減少，進而使的我們在只考慮 411 音的辨識時 Insertion 的降低，不過由於為了提高非 411 音自己的辨認率，相對的將會犧牲 411 音的辨識率，由表可看出包含率由 49.15%下降至 48.28%。

藉由上面的分析，可以得到一個結論，加入為了 Spontaneous Speech 中特殊現象所建立的這些模型，我們可以提升非 411 音的辨識率，進而使 Insertion 錯誤降低，而且只需付出一點點 411 音包含率下降不利因素。因此本論文所採用的基本辨識器為實驗二的系統。

4.4 錯誤分析

上一節的實驗二已列出本論文基本辨識率的結果了，我們可以將之和 Read Speech 語料庫(TCC-300)【4】做比較，TCC-300 的基本辨識率為 70.6%，比本論文之基本辨識高出約 27%，我們可以藉此得知自發性的語料，其中必有許多特殊的特性，造成 Insertion、Deletion、Substitution 皆高於朗讀式語料，進而使辨識率下降，本節將對於表 4.7，也就是對於基本系統之 Confusion Matrix 做更細部的分析，此處不再粗略的分成 411 音及非 411 音，而是不同性質的分為一類，也就是分成 Paralinguistic Phenomena、Particle、Uncertain、411 syllables 這四組做更細部的討論，希望能找出造成高錯誤率的原因，我們將統整過後的結果，列於下表 4.8

表 4.8：基本辨識系統的 Confusion Matrix 分析

辨識結果 / 答案	Paralinguistic	Particle	Uncertain	411	Del	類別字數
paralinguistic	60.54%	3.6%	2.55%	13.74%	19.6%	1254
Particle	4.17%	49.46%	2.24%	32.26%	11.8%	936
Uncertain	2.71%	5.43%	33.91%	48.25%	9.7%	516
411	1.24%	1.51%	1.94%	83.20%	12.1%	14996

下面各小節將對表 4.6 中插入型錯誤、取代型錯誤、刪除型錯誤個別做進一步的分析。

4.4.1 音節插入型錯誤分析

由表 4.6 只考慮 411 音的辨識結果看出，插入型錯誤是偏高的，約 5.4%，我們發現有兩種原因會造成插入型的錯誤，下面將對這兩種原因分別做說明。

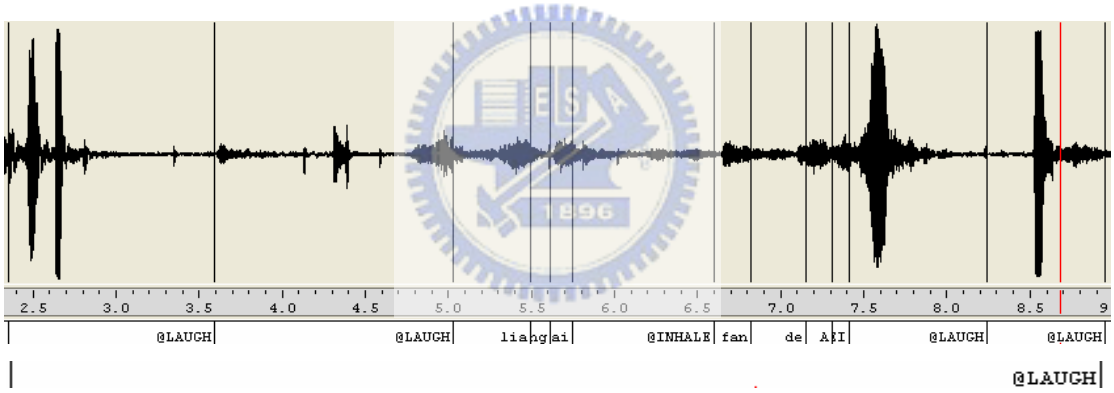
第一種原因是由非 411 音辨識為 411 音所造成的，因為此種狀況的發生將導致在只考慮 411 音辨識時會多辨識出字來，我們將在後面 4.4.3 節插入型錯誤分

析時對於非 411 音與 411 音間互相辯識情形做討論並提出改善。

第二種原因是由於在語料庫中的 Crosstalk 現象造成，此種現象會影響並增加插入型的錯誤。造成 Crosstalk 的原因是語料庫的錄製是用兩支麥克風並於房間內錄製，因此當雙方發生同時間講話的情況時，彼此的聲音難免會錄進互相的麥克風中，造成 Crosstalk 的現象，此種現象嚴重時，音檔中可清楚聽出對方在說些什麼，但 transcribe 標示中這段語音之內容是被標示在另一個 Sub-turn 中，這將造成辨識器多辨識出對方的聲音，進而造成插入型錯誤的產生。

下面列出幾種 Crosstalk 造成插入型錯誤的例子：

1、此音檔的情況為其中一人不斷的在笑(2.5~9sec)，而另一個人同時間在說話，因此造成音檔的文字標記只有笑聲，但實際聽起來，中間穿插了許多人耳可辨識的語音，而這些將造成辨識的錯誤

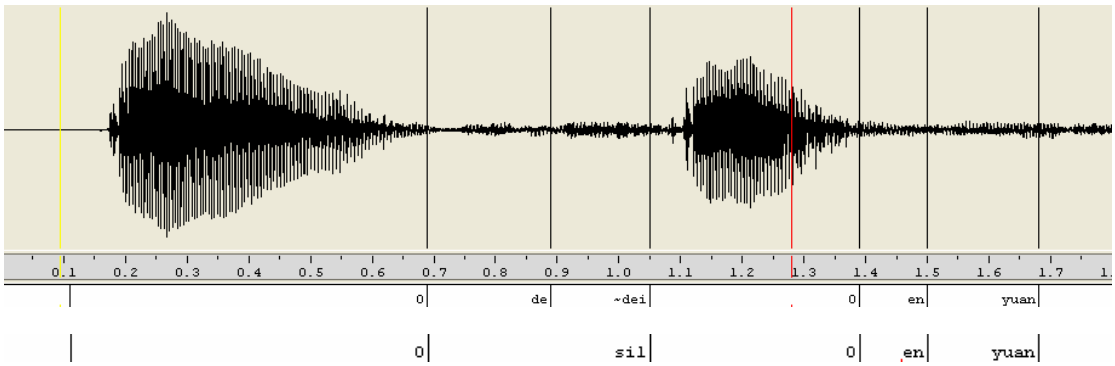


音檔下面第一排文字為辨識出來的結果
音檔下面第二排文字為音檔之文字標註

圖 4.1：Crosstalk 情況一之辨識結果及切割位置示意圖

圖中表示辨識出了哪些字及其切割位置，我們可看出此音檔的辨識結果有非常多的字，但是其實其文字標註只有一個字，就是笑聲，可以知道此句將會造非常多的插入型錯誤

2、此檔的情況為說話者和對方同時講話時的 Crosstalk 現象



音檔下面第一排文字為辨識出來的結果

音檔下面第二排文字為音檔之文字標註

圖 4.2：Crosstalk 情況二之辨識結果及切割位置示意圖

上圖所顯示的 O 與 O 之間辨識出來的 de 和~dei 這兩個字，於文字標註中是沒有的，因為這兩個字是由另一個語者所發出的，其音是清晰可辨的，因此這也是由 Crosstalk 所造成之刪除型錯誤。

此種狀況所造成的 Insertion，無法正確的估計有多少，只大略估計約佔插入型錯誤的 10%，雖並不是最主要的原因，卻也是一個不可忽視的現象。

4.4.2 音節刪除型錯誤分析

對於所有訓練語料作已知答案切割，可以得到所有音節的切割位置，我們可以藉此進一步的統計出，MCDC 語料庫的平均說話速度大約在 5~5.5 syllables/sec，可以知道 Spontaneous Speech 中的講話速度是偏快的，快速語料之錯誤率是較高的【5】，由於說話速度較快，因此有些音的狀態很有可能被省略，譬如在某些習慣用語上，因為太常使用而使某些音節被合併或省略，如「這樣子」會發出近似「醬子」的音，這種現象較嚴重者，我們稱之音節合併(syllable contraction)，所謂的音節合併，是當說話者說得太快或不清楚時出現的現象，其共有三種：

1、清楚可辨的音節短少，像是從原本正常的三個字三個音節變成三個字兩個音節，或者是兩個字兩個音節變成兩個字一個音節。

2、音節雖無短少，但卻都連在一起，難以切割。

3、音節無短少且音節可切割，只是音節結構有變。

在 MCDC 語料庫的標示資料中已標示所有 Contraction 發生的地方，根據統計約佔所有音節總數的 20%，下表為列出容易發生 Contraction 的音節其發生刪除型錯誤的情況。

表 4.9：易發生 Contraction 的音節其發生插入型錯誤之情況

音節	常見合併之詞	Deletion 數量	出現次數	Deletion 發生率	發生 Contraction 比率
是	是阿	106	762	13.91%	31.7%
一	一個	92	525	17.52%	42%
的	是的	87	453	19.205%	42.3%
我	我們	55	465	11.825%	34.5%
他	他們	43	276	15.575%	32%
對	對阿	40	308	12.985%	41.4%
這	這樣	36	143	25.17%	34.5%
有	沒有	35	338	10.355%	34%

由上表八個拼音所佔的 Deletion 數量為 494 個，而 411 所有的 Deletion 為 1527，大約佔了 1/3 左右，且這些音的 Deletion 發生率大多高於整體平均值，可知 Syllable Contraction 是造成 Deletion 發生的重要原因。

4.4.3 音節取代型錯誤分析

4.4.1 節分析曾提到，對於非 411 音辨識為 411 音的這種取代型錯誤，將造成當只考慮 411 音辨識時的插入型錯誤，因此本節將對於 Paralinguistic Phenomena、Uncertain、Particle 這些非 411 音做分析，以期能找到降低取代型錯誤的方法，對於這些類別在下面分節做探討。

4.4.3.1 Paralinguistic Phenomena 取代型錯誤分析

由表 4.8 可看出 Paralinguistic Phenomena 大約有 13%的音是被辨識為 411 音，按照道理來說 Paralinguistic Phenomena 與 411 音的特性應該是有蠻大的差距的，因此會有這麼高的錯誤是值得探討的，表 4.10 列出 Paralinguistic Phenomena 中較容易辨識為 411 音的前三名

表 4.10：易辨識為 411 音之 Paralinguistic Phenomena

音節	該音節被辨識為 411 音之百分比	該音節總數
unrecognizable speech sound	30%	30
unrecognizable non speech sound	26%	281
noise	15.4%	65

首先說明表 4.11 中三種音節所代表的意義，(1)unrecognizable speech sound 代表確屬人所發出之語音，但標記員無法辨認何字何義何音；(2)unrecognizable non speech sound 代表其他由人發出非語音，而且無法辨識的聲音；(3)noise 代表非語音且確定非人所發出的聲音，包括如雨聲、手機聲...等，這三種聲音辨識為 411 音的錯誤，佔了所有 Paralinguistic Phenomena 辨識為 411 音錯誤之 54%，它們都有一個特徵，就是都是無法確定的聲音，因此訓練出來的模型 Variance 必定非常的大，而無法有效的描述其特性，這應該就是造成比起其他 Paralinguistic Phenomena 易造成錯誤的原因。

4.4.3.2 Uncertain 取代型錯誤分析

由表 4.8 可看出互相辯識情形最嚴重的就是 411 音和 Uncertain，而 Uncertain 也的確是特性最接近 411 音的一種現象，因為它只是發音錯誤的音而已，而發錯音的可能情況有無限多種，原本就較難以只用幾個模型來精確描述，因此下面作了一個實驗，就是不將 Uncertain Model 加入辨識器來辨識，看看是否可以藉此

改善辨識率，由於不使用 Uncertain Model，那麼在原本答案中是 Uncertain 的字，必須將其改成相近的 411 音，否則無法求取辨識率，因為那些答案將會是全錯，其結果列於表 4.11

表 4.11：辨識器中去除 Uncertain Model 後之辨識率

	包含率	辨識率	Del(%)	Sub(%)	Ins(%)
整體	47.41	42.55	13.1	39.48	4.86
411 Only	47.69	43.33	13.68	38.6	4.36

由表 4.11 和表 4.6 比較，我們可知道辯識率大約提升 0.5%左右，而這 0.5%的提升是怎麼來的，必需再進一步由 Confusion Matrix 的比較，較可看出，其統整的結果列於表 4.12

表 4.12：去除 Uncertain 影響後之 Confusion Matrix 分析

辨識結果 / 答案	Paralinguistic	Particle	411	Del	總數
paralinguistic	61.2%	4.07%	14.22%	20.52%	1254
Particle	4.27%	50.10%	33.33%	12.28%	936
411	1.42%	1.76%	84.27%	12.55%	15512

由表 4.8 中，Uncertain 辨識為 Uncertain 的機率為 33.91%，而進一步統計其完全正確的機率為 25.58%，在答案中總共有 516 個音節，也就是完全正確的音節有 132 個，而不將 Uncertain Model 用於辨識後，我們比較 411 音正確字的數目，可進一步統計出去掉前比去掉後 411 音多正確了 176 個字，Paralinguistic Phenomena、Particle 合算多正確了 10 個字，這些多正確的字，同等的就是取代型錯誤的降低，0.5%的提升大多就是由這些錯誤的降低所得到的，我們再根據 411 Only 之辨識率分析，Uncertain 容易辨識為 411 音，是由於其 Model state 數量少，且其訓練語料也少，因此模型較為粗糙，比較表 4.6 與表 4.11，可知去除 Uncertain 易與 411 音混淆的情形後，Insertion 的情況的確是降低了，顯示在 4.4.1 節中所討論的造成 Insertion 錯誤之第一種原因是確實存在的。

我們根據實驗結果可得到了一個結論，系統中不加入 Uncertain Model 去辨識是較為正確的作法，況且若要將辨識結果轉為字元輸出時，一些 Uncertain 也會造成無字元可以對應的情況，而必須以拼音型式輸出。

雖然 Uncertain model 不適用於加入辨識器中，但在訓練模型時我們仍使用較粗糙之 Uncertain model 的原因是在於 (1)能使切割更為正確 (2)不將 uncertain 用來訓練 411 音節模型。

4.4.3.3 Particle 取代型錯誤分析

根據表 4.8，我們還可發現 411 與 Particle 的互相辯識情形也是很嚴重，尤其是 Particle 的音常常會辨識為 411 音，這個結果其實不令人意外，因為有很多 Particle 其發音聽起來就和其相近 411 音幾乎是一模一樣的，但 Particle 的模型僅使用三個狀態來描述，因此結果是合理的。

為了解決 411 音與 Particle 互相辨識的不平衡狀況，我們希望藉由 411 音及 Particle 在 Duration 上的差異來解決，由於在 HTK 中無法建立 Duration Model，因此我們的方法為求出 Particle 及與其最相近的 411 音在 Duration 分布上的 PDF，再定義一個臨界值，當得到的辨識答案為其中一個音，便根據此臨界值來決定該辨識結果是否正確，是否需要更改。

圖 4.1 舉出兩個較常出現的 Particle，及與其最相近的 411 音之 Duration 的 PDF

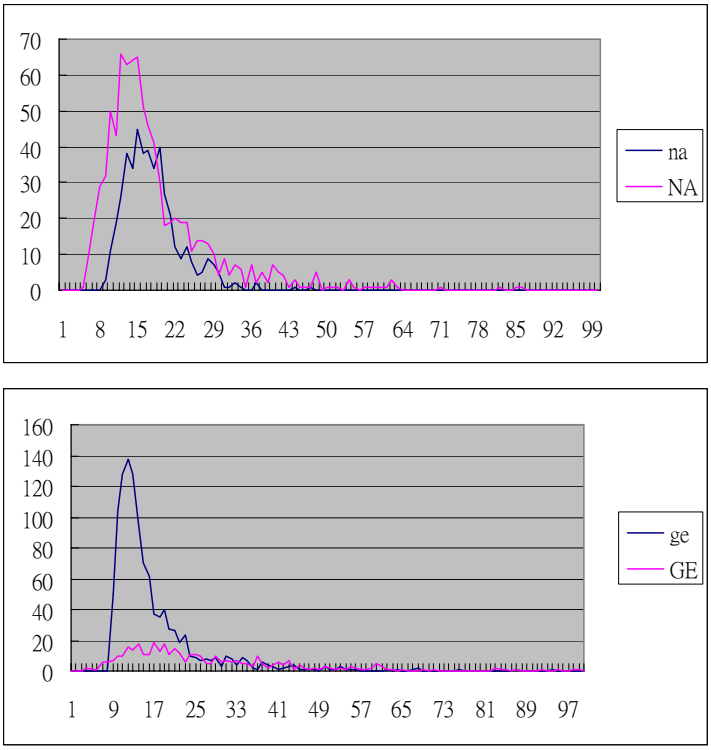


圖 4.3：Particle 及其相近 411 音之 Duration 分佈

由圖即可知道，原本期望藉由 Duration 來區分出 Particle 及 411 音，是件非常困難的事情，因為其分佈非常的相像，我們難以界定一個臨界值來決定辨識答案是否更改，我們可對此下一個結論，Particle 及其相近 411 音在 Duration 的特性上差距是很小的。

由 2.3 節自發性語料庫特性分析中，可知感嘆詞分為四類，(1)有相對應國字的感嘆詞；(2)無相對應國字的感嘆詞；(3)源於台語的感嘆詞；(4)其他感嘆詞，在此我們將之重新歸類合併為兩類，一為有相對應國字的感嘆詞，一為無相對應國字的感嘆詞

表 4.13：Particle 分類後之辨識率分析

Particle 類別	**被辨識為 411 音**	**辨識為 411 音且為相近 411 音**	**辨識為同類 Particle**	**Testing Data 中之音節數**
有相對應國字	32.71%	11.56%	47.35%	813
無相對應國字	18.6%	4.85%	69.8%	123

由表 4.13 我們可看出有相對應國字的 Particle，其模型是較容易會與 411 音節混淆的，而無相對應國字的辨識結果是較好的，也較不會辨識為 411 音，這個結果並不令人意外，因為有相對應國字的 Particle 其發音與 411 音的特性非常相近，它與 411 音的最大的差別只在於是否含有語意而已，而 Particle 對於一整句話是無意義的，在辨識器中我們並無判斷有無語意的機制，因此對於答案中硬將兩種答案分開是不太恰當的，不過聲學模型的辨識結果亦達到 50%以上，因此對於上層語意之分析時，仍具參考價值，也由於此，我們的處理方式不像處理 Uncertain 般，還是將之保留於辨識器中。

經由上面的討論，我們稍加改變一下觀察答案比對的方法，也就是當答案屬於有相對應國字的 Particle，則辨識結果除了辨識為自己是正確外，若辨識為相近 411 音也算是正確，反之亦為正確，如：411 的 a 辨識為 particle 的 A 及 particle 的 A 辨識為 411 的 a 皆為正確，以如此看法，將結果列於表 4.14

表 4.14：Particle 辨為相近 411 音修正後之辨識率

	包含率	辨識率	Del(%)	Sub(%)	Ins(%)
基本辨識系統	**47.69**	**43.33**	**13.68**	**38.6**	**4.36**
Particle 修正後	**47.90**	**43.87**	**13.2**	**38.8**	**4**

由表 4.11 與表 4.14 我們比較辨識率，發現由 43.33%升至 43.87%，這約 0.5%的上昇，絕大部份都是有相對應國字的 Particle，與相近 411 音的音節，互相辨識錯誤修正後的結果。

由以上觀察及分析，我們發現自發性語料辨識率較朗讀式語料下降 20-30%的主要原因在於：

(1) 音節合併現象造成許多刪除型及取代型錯誤的增加。

(2) Particle、Paralinguistic 辨識為 411 音及 411 音被辨識為 Particle、Paralinguistic 而造成插入型及刪除型錯誤的增加。

(3) 由於 Crosstalk 現象造成了插入型錯誤的增加。

以上三點皆是造成錯誤率增加的主要原因，而這也都是在自發性語料才會出現的現象，所以找到更好的方法來解決這些現象，將是提升辨識率的有效方式。

第五章　加入語言模型之自發性國語辨識器

一個國語的辨識器其辨識結果應該是文字而不是音節結果，但是中文字的音轉字，是一個非常複雜的處理，首先中文字有一音多字的情形，而且加上辨識器中並無對於聲調的討論，這會使音轉字是更加的困難，但若能加入語言模型，我們將可得到文字的辨識，不但可使辨識器更加的完整，更可提高它的效能。

由於所有的語言都有其獨特的文法規則，因此我們可針對此規則性，來求得一個機率模型，對於此種機率模型，一般我們稱為語言模型，簡稱 LM(Language Model)，在辨識時，除了聲學模型外，若能加入語言模型的參考，通常能大幅提高辨識系統的效能。

在本章將建立兩種不同性質的語言模型，一種是以文字性質的語料所訓練而得，一種是由對話性質的語料訓練而得，我們將會討論加入不同性質的語言模型，對於 MCDC 語料庫辨識情況的影響，並以一個適當比重來調適兩種語言模型，以增加辨識系統之效能。

5.1 建立語言模型

本節將介紹我們是如何訓練語言模型，其流程如圖 5.1：

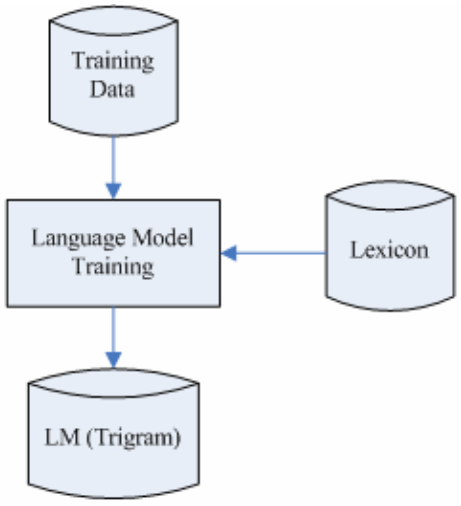


圖 5.1：LM 訓練流程圖

5.1.1 訓練語料及詞典(lexicon)

訓練語料及詞典是建立語言模型所必須要準備的兩樣資料，下面將介紹其用途，及本論文中所使用之訓練語料、詞典為何。

5.1.1.1 訓練語料

建立語言模型必須要有大量的文字資料庫，才可分析其語言規則，對於不同種類的訓練語料所分析出的語言規則也必定不同，本論文中所採用的訓練語料有兩種，(1)包含光華雜誌(Sinorama)、NTCIR 和中研院的平衡語料庫，下面將稱之為 General 語料庫；(2)MCDC 語料庫之訓練語料的部份。

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科學領域。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，這三種語料庫的內容皆是文字性質，我們可藉此訓練出具有文字性質語言規則的語言模型。

MCDC 語料庫是一個內容為對話性質的語料，利用此語料庫將可建立出具對話性質語言規則的 LM，不過由於本論文基本架構中所用於測試的語料即為 MCDC 語料庫中的一部分，因此只可將論文中用於訓練聲學模型的語料來建立，否則將產生不公平的現象。

5.1.1.2 詞典

上一節介紹了訓練語言模型所需的兩種語料庫，有了語料庫我們即可做其語言上的分析，在漢語中文(Mandarin)下，以詞為單元來做分析是較符合語言規則的，所以必須將語料庫由原本以音節為單位轉換成以詞為單位，這時便需要詞典來做轉換，下面將對於本論文所使用之詞典其來源做介紹。

詞典的來源，我們是由清大資工所的張俊盛教授所提供之詞典(此詞典只有

詞且無 POS(part of speech))，利用此詞典，對 General 語料庫作斷詞，然後依詞頻做排列，挑選出最高的前 6 萬詞，由於這六萬詞的詞典中還夾帶著英文、注音符號這些詞，將之剔除後，最後實際的詞典大小為 58,940 個，此即為本論文中所使用之詞典，對於詞典中詞長分佈統計於表 5.1

表 5.1：詞典中之詞長分布

詞長	1	2	3	4	5	6	7
數量	4009	36357	11478	5555	802	530	209
百分比	6.8%	61.7%	19.5%	9.4%	1.4%	0.9%	0.4%

根據論文中所使用的詞典，對兩種語料庫作斷詞後的結果，其資料分別統計於表 5.2 及表 5.3

表 5.2：General 語料庫之詞數及字數統計表

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

表 5.3：MCDC 語料庫之詞數及字數統計表

訓練語料	**詞數（Word）**	**字數（Character）**
MCDC 語料庫	96,816	126,450

5.1.2 語言模型的訓練

有了訓練語料及詞典，便可以開始訓練語言模型，在訓練過程中必定會有一個問題發生，那就是訓練語料中會出現詞典中所沒有的詞，對於這類的詞，統稱為 OOV(Out Of Vocabulary)，下面將討論對於兩種語料庫在 OOV 上的處理及語言模型訓練的方法。

5.1.2.1 OOV 的處理

對於 General 平衡語料，我們在辭典中加入一個特殊的詞，稱之 Unknown Word，在訓練語言模型時將發生 OOV 的所有詞皆用於訓練 Unknown Word。而對於 MCDC 語料庫，由於語料庫中包含 Particle、Paralinguistic Phenomena 這兩類語言學的現象，為了保留這兩種現象之語言規則，我們並不將之用於訓練 Unknown Word，而是於詞典中增加兩個新詞，分別代表這兩種現象，來特別訓練這兩類的語言規則。

5.1.2.2 訓練語言模型的方法

藉由訓練語料與詞典，本論文中我們是要訓練出 Trigram 的語言模型，因此要求出 Unigram、Bi-gram、Trigram 的機率，Bi-gram 及 Trigram 之機率分別為 $P(w_i \mid w_{i-1})$、$P(w_i \mid w_{i-1,} w_{i-2})$，下面將介紹，求取 n-gram 機率的方法【6】，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：

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

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。

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

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，我們即可得到所需求的語言模型了。

5.2 基本辨識器加入語言模型之辨識分析

要將語言模型加入辨識系統中，我們還需將之轉換為 Word-net，因為 Word-net 才是清楚的描述詞跟詞的轉移關係，由於 HTK 中轉換上的問題，我們只使用到 Bigram 和 Unigram 的機率，其轉換流程如圖 5.2

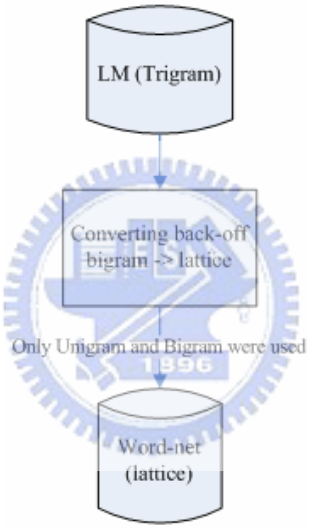


圖 5.2：LM 轉 Word-net 之流程圖

有了 Word-net，相當於文法規則，之後便可將此文法加入基本的辨識系統中，而加入了語言模型，在辨識時我們除了會得到聲學模型的分數外，還會再得到語言模型的分數，本論文中我們較為重視語言模型，因此將其所得之分數乘五，以提高其影響，實驗中我們的基本辨識系統為 4.4.2.3 中只使用 411 音節、Particle、Paralinguistic Phenomena 這些聲學模型的系統，本節將做三個實驗，以討論加入三種 LM 所產生的 Word-net，對於基本辨識系統的改善。

5.2.1 實驗一

本實驗所使用的語言模型是由 General 語料庫訓練而得的，以下我們稱之為 General LM，加入語言模型後的辨識單位由音節變為詞，但是為了能與未加入語言模型的系統比較，我們還是會將詞轉成音節來做辨識，加入 General LM 後的辨識結果列於表 5.4

表 5.4：加入 General LM 之辨識結果

	包含率	辨識率	Del(%)	Sub(%)	Ins(%)
整體	52.36	49.56	15.2	32.4	2.8
411 only	53.88	51.23	16.1	30	2.6

5.2.2 實驗二

本實驗所採用的語言模型是由 MCDC 語料庫所訓練而得的，在此我們稱之 MCDC LM，與實驗一同樣的將詞轉為音節來辨識，並多觀察只考慮 411 音的辨識情況，加入 MCDC LM 後的辨識結果列於表 5.5

表 5.5：加入 MCDC LM 的辨識結果

	包含率	辨識率	Del(%)	Sub(%)	Ins(%)
整體	53.99	51.16	15.2	30.7	2.8
411 only	55.33	52.65	14.7	30	2.68

5.2.3 實驗三

本實驗所採用的語言模型是經由實驗一及實驗二的語言模型所調適而得，所謂語言模型調適（Language Model Adaptation），以一個 Bigram 的條件機率來看，我們進行調適後會變成：

$$P_{adap}(w_i \mid w_{i-1}, w_{i-2}) = \lambda P_{Gen}(w_i \mid w_{i-1}, w_{i-2}) + (1-\lambda) P_{MCDC}(w_i \mid w_{i-1}, w_{i-2}) \qquad (5.1)$$

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_{MCDC} 是在 MCDC 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流程如圖 5.3

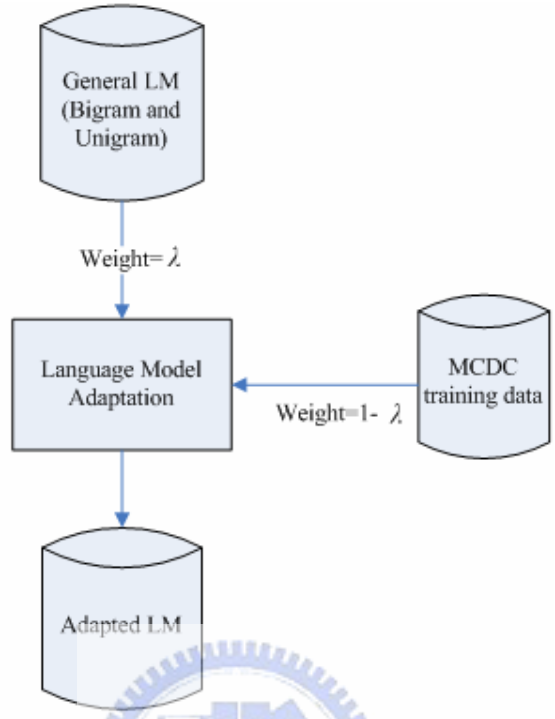


圖 5.3：語言模型調適流程圖

在這個實驗中，我們會調整兩種 weight，分別是 0.8 及 0.2，希望能藉由比例上的分配，辨別哪一個語言模型對於辨識是較有幫助的。

當 Weigh 調整為 0.8 時之辨識結果列於下表 5.6

表 5.6：Adapted LM 之調適 weight 為 0.8 之辨識結果

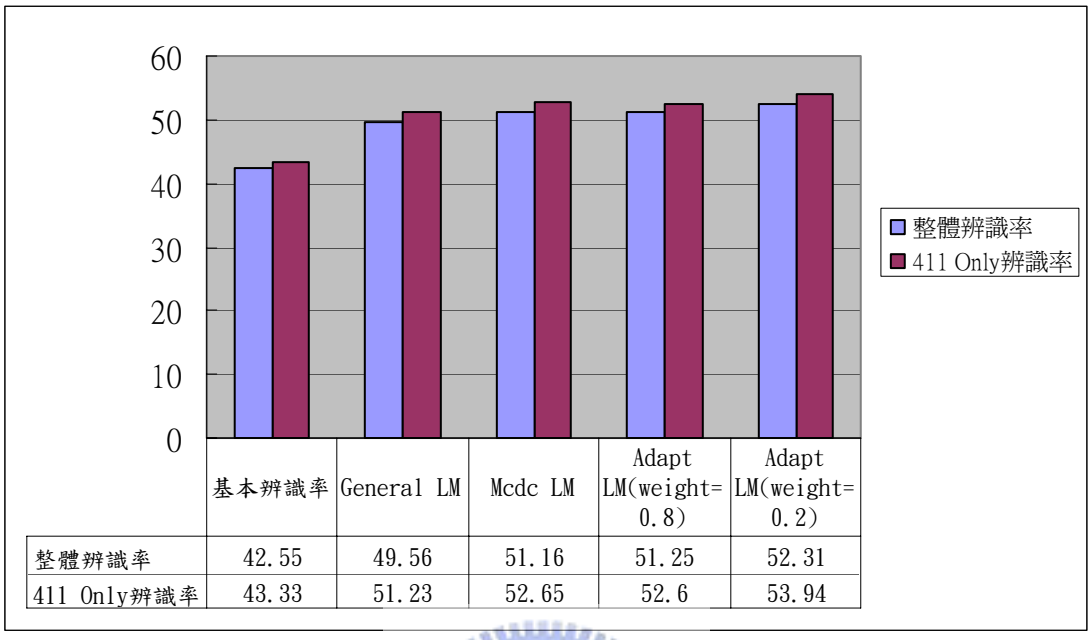
	包含率	辨識率	Del(%)	Sub(%)	Ins(%)
整體	53.62	51.25	16	30.3	2.37
411 only	55.31	52.60	14.8	29.8	2.7

當 Weigh 調整為 0.2 時之辨識結果列於下表 5.7

表 5.7：Adapted LM 之調適 weight 為 0.2 之辨識結果

	包含率	辨識率	Del(%)	Sub(%)	Ins(%)
整體	54.93	52.31	15.3	29.7	2.61
411 only	56.46	53.94	14.8	28.7	2.52

我們將實驗一、二、三的結果與基本辨識系統做比較，最後結果將之整理如下圖 5.4



	基本辨識率	General LM	Mcdc LM	Adapt LM(weight=0.8)	Adapt LM(weight=0.2)
整體辨識率	42.55	49.56	51.16	51.25	52.31
411 Only辨識率	43.33	51.23	52.65	52.6	53.94

圖 5.4：加入語言模型後與基本辨識系統之比較圖

對於實驗一、實驗二及實驗三各項結果之分析我們將細列於下：

1、由圖 5.4 看到加入 General LM 後，在整體的辨識率上比基本辨識器高了約 7%，而在只考慮 411 音辨識率上提高了約 8%，可知由於 General LM 並沒有對於 Particle、Paralinguistic 的文法描述，造成對於非 411 音的辨識率是較無幫助的。

2、對於加入 General LM 與 MCDC 本身語料庫訓練的語言模型，這兩種對於辨識率的幫助，明顯的以使用對話性質語料所建立的語言模型幫助較大，辨識率大約在提升約 1.5%，因此我們可以下一個結論，文字性質與對話性質的語言規則是有其差距的。

3、由表 5.3 可知 MCDC 訓練語料共有 96816 個詞，是否能完整涵蓋對話性質是值得懷疑的，因此試著 Adaptation 一個具有大量詞的 General 語料庫，看看是否能夠補起不足的地方，藉由圖 5.4 的結果發現 Adaptation 的確是有用的，

而對於調試比重上，發現若比重以 General 語料庫為重的話幾乎是沒有什麼幫助的，若比重以對話性語料為重的話，對辨識率的幫助就比較大了，可以知道可藉由加上一點比重小的 General 語料，來補足 MCDC 語料詞太少的缺點，進而提高辨識率。

上面比較了基本系統與加入語言模型前後，音節辨識率的差別，現在我們將觀查加入語言模型後的字元辨識率與詞的辨識率，同樣的比較實驗一、二、三加入不同語言模型的結果，在此只討論 411 Only 的辨識率，將詞與字元之辨識結果列於表 5.8、表 5.9

表 5.8：詞辨識率比較

	General LM	**MCDC LM**	**Adapted LM (weight=0.2)**	**Adapted LM (weight=0.8)**
辨識率	30.38%	33.31%	36.13%	33.31%

表 5.9：字元辨識率比較

	General LM	**MCDC LM**	**Adapted LM (weight=0.2)**	**Adapted LM (weight=0.8)**
辨識率	38.24%	41.03%	43.71%	41.28%

同樣的不論是字元或詞的辨識率，使用 Adapted LM (weight=0.2)的依然是最高的，而且其提高的比率還超過了以音節考慮之辨識率。

5.3 與 Mandarin Call Home 系統比較

IBM T.J Watson Research Center 在 1996 年提出了的一篇對於 Mandarin Call Home 所做的辨識研究【7】，由於 Mandarin Call Home 亦為關於自發性語音的語料庫，因此我們將於此節討論比較兩個辨識系統。

首先先對 Mandarin Call Home 做一個簡單的介紹，Mandarin Call Home 是 Call Home corpus 中的一部份，其音檔為利用電話線，錄製雙方之談話內容，對談的人是認識的，因此內容沒有一定的主題，在【7】中有兩個 baseline system，是對兩個不同時間所錄製之語料庫(Nov 94、Apr95)所做的實驗，在此只討論其中 Apr95 這個語料，因為其訓練語料與我們較相近，下表 5.10 列出兩語料的比較

表 5.10： MCDC 與 Apr95 訓練語料之比較

訓練語料	時間(hr)	詞數
MCDC	8.77	96K
Apr95	7.4	170K

在建立語言模型上的方法相同，同樣是利用語料庫的訓練語料部份做訓練，再利用一份較大且較 General 的語料來調適，此份語料共包含八千七百萬個詞，比我們這邊的七千七百萬個詞稍多，下表 5.11、表 5.12 列出辨識結果比較

表 5.11：兩系統字元辨認率結果比較

LM 種類	本論文	【7】
使用本身訓練語料	41.03%	34.6%
General LM	38.24%	29.2%
Adapted LM (weight=0.2)	43.71%	35.1%

表 5.12：兩系統詞辨認率結果比較

LM 種類	本論文	【7】
使用本身訓練語料	33.31%	25.8%
General LM	30.38%	20.5%
Adapted LM (weight=0.2)	36.13%	25.9%

由上表可知，本論文的辨識率是比較高的，但這應屬正常，因為 MCDC 語

料是麥克風錄音，品質比 Apr95 的電話線錄音好，且 Mandarin Call Home 是普通話，普通話的音節種類是多於 411 音的，因此實驗的結果是蠻合理的。

第六章　結論與未來展望

6.1 結論

在本論文中，我們分析了自發性語料中有別於朗讀式語料的一些特性，並針對其特性建立出屬於自己的聲學模型，藉由實驗發現辨識器中加入這些聲學模型對於辨識系統是有好處的，加入前後約可提升 3.5%，主要的提升是在於插入型錯誤的降低，這是因為加入這些聲學模型，可較清楚的描述非 411 音的現象，使非 411 音辨識為 411 音的機率下降，進而降低插入型錯誤的發生。

接下來探討 Uncertain 這種特性極近似 411 音的現象，發現辨識器中不應加入對於此種現象的描述，這是因為 Uncertain 的種類是無限的，無法以少量的模型來涵蓋，藉由實驗可知去除了 Uncertain 與 411 音易混淆情況後，辨識結果可提升約 0.5%。在論文中還討論到 Particle 與 411 音之間的關係，由於許多 Particle 的特性與 411 音的差別只在於語意上的有無而已，由於辨識器並不包含對於語意上的判斷，因此若 Particle 辨識為相近 411 音應該不可算錯，對此情況做修正後辨識率可再提升約 0.6%。

處理過聲學模型及觀看辨識率的方法後，最後辨識率可達到 43.10%，為了使辨識器更為完整，我們加入了兩種不同性質的語言模型，實驗結果可知使用對話性質的語言模型比使用文字性質的可得到更好的效果，並可藉由兩種性質語料間的調適，補足對話性質語言模型缺少的部份，得到更完整的語言模型，使辨識率再為提升。

6.2 未來展望

未來研究方向可朝下列幾點進行：

1、在論文中有提到 Deletion 的降低，Syllable Contraction 是一個重要的原因，未來可對於這方面特性做更深入的探討，以降低 Deletion 的發生率，進而提升

辨識器的效能。

2、音檔中 Crosstalk 的現象，也是一個造成辨識率下降的一個原因，未來可對於這方面在進一步研究。

3、在語言模型調適中，我們使用了兩種不同的 weight，目地在於提升辨識率及比較何種性質的語言模型較適合 MCDC 語料庫，未來還可進行對語言模型複雜度(perplexity)的分析，找出可使語言模型複雜度最低的 weight 來做調適以得到最佳的辨識結果。

參考文獻

[1] 曾淑娟、劉怡芬, ”現代漢語口語對話語料庫標註系統說明”, 中文詞知識庫小組，民國九十一年一月

[2] 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）”

[3] Kazuyuki TAKAGI, Shuichi ITAHASHI, “Segmentation of Spoken Dialogue by Interjections, Disfluent Utterances and Pauses”, *In Processing of the ICSLP-96*, pp.670—700

[4] 林政賢，”以可靠度量測引導之通道效應及頻寬不匹配補償於牆漸行語音辨認”, *國立台北科技大學電腦通訊與控制研究所*, 民國九十二年六月

[5] H. Nanjo, and T. Kawahara, “Speaking Rate Dependent Acoustic Modeling for Spontaneous lecture Speech Recognition”, *Proc. Eurospeech 2001*, pp2531-2534.

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

[7] F. H. Liu, M. Picheny, P. Srinivasa, M. Monkowski, and J. Chen, “Speech Recognition on Mandarin Call Home: A Large-Vocabulary, Conversational, and Telephone Speech Corpus”, *IBM T.J. Watson Research Center*, P.O. Box 704, Yorktown Height, NY 10598