國 立 交 通 大 學

電信工程研究所

碩士論文

中文自發性語音辨認系統

Mandarin Spontaneous Speech Recognition

研 究 生：許誌宏

指導教授：陳信宏 博士

中華民國九十九年八月

中文自發性語音辨認系統

Mandarin Spontaneous Speech Recognition

研 究 生：許誌宏　　Student：Chih-Hung Hsu

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

國 立 交 通 大 學

電 信 工 程 研 究 所

碩 士 論 文

A Thesis

Submitted to Institute of Communication Engineering

College of Electrical and Computer Engineering

National Chiao Tung University

in Partial Fulfillment of the Requirements

for the Degree of

Master

in

Communication Engineering

August 2010

Hsinchu, Taiwan, Republic of China

中華民國九十九年八月

中文自發性語音辨認系統

研 究 生：許誌宏　　指導教授：陳信宏 博士

國立交通大學電信工程研究所

中文摘要

近年來朗讀式語音辨認技術已經相當成熟，下一階段的目標將轉往自發性語音辨認，因此本論文將建立一套中文自發性語音辨認系統。在本論文中首先將說明如何建立自發性語音聲學模型；在語言模型方面，自發性語音中有許多常用的口語詞、感嘆詞或語助詞，這些都與傳統語言模型差異甚大，因此本論文以傳統語言模型為基礎，建立一套自發性語音語言模型。

在自發性語音韻律模型的改進方面，本研究將重新定義自發性語音中的特殊韻律現象。除此之外，傳統語音辨認通常只使用聲學模型及語言模型，但是自發性語音中有許多的特殊現象是朗讀式語音中沒有的，因此本論文試著將韻律模型也運用進來，希望能利用韻律資訊來解決這些狀況。

Mandarin Spontaneous Speech Recognition

Student：Chih-Hung Hsu　　Advisor：Dr. Sin-Horng Chen

Institute of Communication Engineering
National Chiao Tung University

Abstract

In recent years, the read-speech recognition technology has been quite mature, and the goal for the next phase will be transferred to spontaneous speech recognition; therefore, this paper will establish a Chinese spontaneous speech recognition system. In this paper, we will describe how to build the spontaneous acoustic model first. In the language model, there are many popular spoken word, particle and expletive in the spontaneous speech, which has great difference with the traditional language model. So, we will adapt a spontaneous speech language model based on the traditional language model in this paper.

In the improvement of the spontaneous speech prosody model, this paper will re-define the special prosodic phenomena of the spontaneous speech. In addition, people usually use the acoustic model and the language model in the traditional speech recognition. However, there are many differences between read-speech and spontaneous speech. So, this paper will try to use some prosodic information to help the speech recognition and wish to resolve these situations.

誌謝

學生時代終於真的要結束了！還真是謝天謝地蟹老闆勒~~哇哈哈！不過在這裡好像應該要感性一點齁…，感謝每個禮拜督促我們進度的陳老師，雖然最後是「功虧一簣」，但是我們雖敗猶榮！再來是要感激每次輪我報告時就會突然出現電爆我的王老師，雖然每次我都只能啞口無言站在台上，但是事後想想都覺得老師說的很有道理，還有就是感謝您在口試時助我一臂之力！

再來就是實驗室的大家，首先是帶著我一路走來的性獸、帶我加入 DMC 行列的合哥、每次都走過來問我有什麼問題的阿德，還有地球防衛小組的希群和準備開公司的巴金叔，你們應該算是 707 的五虎將吧！有著納豆明星臉的輝哥，總是喜歡講一些嚇人的話；這兩個月來一直和我一起住實驗室的承燁，恭喜你最後陶小姐對你露出勝利的微笑；「超屌」的皓翔，我很怕 10 年後只記得你叫暢邱翔；在系辦打工超夯的雲舒，我永遠不會忘記你口試時的樣子；口試時就變成小孬孬的財祿，祝你在晨星能熬過三個月；進實驗室前被我跟大學同學偷偷稱電信大小喬的小喬依玲，我只要兩妹就好了不需要到 20 妹謝謝；還有我們永遠的一哥宥余，什麼時候可以幫我切我的語料啊？哈哈…感謝學長們小宋、杜 Q、小帥哥還有把我騙入 spontaneous 世界的普烏；感謝碩一的學弟們：後來常幫我寫程式的銘傑、大一上邏設第一次期中考就考一百分的人妻殺手胖胖，用合成器造出一堆告白句子卻都沒用過的啟全、舞林高手智障、聯誼咖大胖跟小蝦、健身達人豆腐喵以及神秘人佳緯。

最後感謝我的大學同學們，以電信嘟為首的台北幫以及常常揪團吃飯看電影的新竹幫，讓我碩班生活不至於都在實驗室度過！套句老梗：要感謝的人太多了，那就謝天吧！

目錄

中文摘要 I
Abstract II
目錄 IV
表目錄 VII
圖目錄 IX
第一章 緒論 1
1.1 研究動機 1
1.2 相關研究 2
1.2.1 語言模型之相關研究 2
1.2.2 韻律模型幫助辨認之相關研究 2
1.3 研究方向 3
1.4 章節概要 4
第二章 漢語口語對話語料庫介紹 5
2.1 語料庫介紹 5
2.1.1 音檔格式說明 5
2.1.2 語料標註格式說明 6
2.2 自發性語音之特性 7
2.3 MCDC 語料庫之後處理 9
2.3.1 斷詞與相關統計 9
2.3.2 標記標點符號 11
2.3.3 非流暢現象中斷點 12
2.3.4 修正音節切割位置 13
第三章 自發性語音辨識系統 15

3.1 聲學模型.....15
3.1.1 訓練語料及測試語料.....15
3.1.2 聲學模型之建立.....16
3.1.2.1.特徵參數抽取..... 16
3.1.2.2 聲學模型之建立流程..... 16
3.1.3 實驗結果.....19
3.2 語言模型.....21
3.2.1 初始模型之建立.....21
3.2.1.1 訓練語料..... 21
3.2.1.2 模型訓練..... 21
3.2.2 語言模型之調適.....22
3.2.2.1 自發性語音之訓練單元..... 22
3.2.2.2 調適模型之設計..... 23
3.2.3 語言模型效能分析.....29
第四章 自發性語音韻律模型.....30
4.1 韻律模型設計.....30
4.1.1 中文語音韻律階層式架構.....30
4.1.2 韻律參數之介紹.....33
4.1.3 模型設計.....35
4.2 韻律模型之建立.....39
4.2.1 初始化.....39
4.2.2 重覆疊代.....41
4.3 韻律模型之分析.....42
4.3.1 音節韻律模型.....42
4.3.2 停頓標記聲學模型.....43
4.3.3 停頓標記結果之分析.....45

4.4 以韻律模型協助語音辨認.....47
第五章 實驗結果及討論.....49
5.1 辨認系統加入語言模型實驗.....49
5.1.1 MCDC 辨識結果.....49
5.1.2 中央社對話語料庫之辨識實驗.....51
5.2 辨認系統加入韻律模型實驗.....53
5.2.1 以韻律模型協助辨認之辨識率.....53
5.2.2 實驗結果討論.....54
第六章 結論與未來展望.....59
6.1 結論.....59
6.2 未來展望.....59
參考文獻.....61
附錄一 詞性分類表.....63
附錄二 韻律模型初始停頓標記門檻值之選定.....65
附錄三 停頓標記聲學模型之問題集.....70
附錄四 停頓標記語言模型之問題集.....72

表目錄

表 2.1：MCDC 資料庫的對話主題與語者對照表 6

表 2.2：MCDC 資料庫中 disfluency 與 IP 出現之個數 12

表 3.1：訓練語料統計 16

表 3.2：測試語料統計 16

表 3.3：HMM 模型之設定 18

表 3.4：MCDC 音節辨認率 19

表 3.5：General-LM 訓練語料統計 21

表 3.6：在辨認字典中加入之 27 類 particle 23

表 3.7：在辨認字典中加入之 9 類 paralinguistic 23

表 3.8：MCDC 中各類別之相接機率 28

表 3.9：語言模型混淆度評估 29

表 4.1：韻律結構之停頓標記 32

表 4.2：歸類為基本音節之 particle 個數 33

表 4.3：韻律標記、聲學參數以及語言參數之數學符號 34

表 4.4：基本音節中，不同組合之 AP 下音節韻律模型參數之 TRE 42

表 4.5：特殊音節中，不同組合之 AP 下音節韻律模型參數之 TRE 42

表 5.1：各階段語言模型之辨識率 49

表 5.2：口語對話常用詞之辨識率 50

表 5.3：語言模型不同調適方法之辨識率 51

表 5.4：中央社測試語料統計 51

表 5.5：MCDC 與中央社辨識率比較(Acc(%)) 51

表 5.6：原始句子與切短句後詞之辨認涵蓋率 53

表 5.7：加入韻律模型之辨識率 53

表 5.8：較佳之 Top-N 候選詞串 54

表 5.9：較差之 Top-N 候選詞串 55

表 5.10：表 5.8 範例音段之聲學、語言模型分數 56

表 5.11：表 5.9 範例音段之聲學、語言模型分數 56

表 5.12：辨認句子結構表 58

圖目錄

圖 1.1：訓練模型架構圖 3

圖 2.1：語料庫中詞長分布圖 10

圖 2.2：語料庫中 POS 種類分布圖 10

圖 2.3：語料庫中每個 sub-turn 之音節數分佈圖 11

圖 2.4：發生非流暢現象現象之音節數分佈圖 12

圖 2.5：不同音節數發生非流暢現象現象之機率分佈圖 13

圖 3.1：聲學模型之建立流程 17

圖 3.2：加入狀態轉移機率示意圖 19

圖 3.3：General-LM 訓練流程圖 22

圖 3.4(a)：由 MCDC 估出各類別機率 24

圖 3.4(b)：將 Gerneral-LM 依圖 3-4(a)估算的比例重新分配 24

圖 4.1：中文語音韻律之階層式架構概念 31

圖 4.2：中文自發性韻律階層式架構 31

圖 4.3：音節及音節間韻律參數 34

圖 4.4：初始韻律模型建立流程圖 40

圖 4.5：分類停頓標記之決策樹示意圖 40

圖 4.6：韻律模型重覆疊代流程圖 41

圖 4.7：五個中文聲調之 AP 43

圖 4.8：(a)音節停頓長度 (b)音節間能量低點 (c)正規化音節延長因子與 (d)正規化基頻跳躍

值之分布圖 44

圖 4.9：韻律停頓標記分佈圖 45

圖 4.10：韻律停頓標記範例一之音檔信號圖 46

圖 4.11：韻律停頓標記範例二之音檔信號圖 46

圖 4.12：辨識系統架構圖 48

圖 5.1：解答聲學分數與 Top-N 聲學分數相減之分布圖 57

第一章 緒論

1.1 研究動機

近年來隨著科技的進步，語音辨認(Automatic Speech Recognition, ASR)系統已經有相當成熟的技術，對於朗讀的語音輸入辨認效果極佳，然而要實際應用在生活化的商品上，則必須考慮到更接近於人們日常生活對話的自發性語音(Spontaneous speech) 。但自發性語音之辨認率仍舊與朗讀式語音(Read speech)有一段差距。

造成自發性語音辨認率不如朗讀式語音之原因主要是因為自發性語音常會伴隨著非正規化(ill-formed)以及不流暢語流(disfluency)現象。首先，因為自發性語音有複雜的韻律(prosody)變化及較快的說話速度(speaking rate)，使得音節與音節間會發生嚴重的相互影響，例如：人們在日常對話時，大腦其實是只取一段句子中的關鍵字即可理解對方所要表達的意思，因此造成人類在發音時某些語音會被省略或產生發音變異(pronunciation variation)以及音節合併(syllable contraction)等現象。此外，由於自發性語音未經大腦良好的規劃，使得語流中常會出現遲疑(hesitation)、口吃(stutter)、非流暢現象(disfluency)等不合乎文法結構之語句，以及許多感歎詞(particle)、語者慣用的語助詞(marker)出現。因此自發性語音在辨識上會比一般的朗讀式語音還要困難許多。如果我們能夠有效解決上述之自發性語音問題，相信一定能為人們在將來生活上帶來許多更便利的幫助。

本論文將建立一套中文自發性語音辨認系統。傳統上語音辨認系統皆是以聲學模型為主，語言模型為輔，然而在自發性語音中由於發音問題，較難有極佳的聲學模型，因此我們希望能藉由語言模型以及韻律模型的幫助來提升辨識率。

1.2 相關研究

1.2.1 語言模型之相關研究

在自發性語言模型中所遭遇的問題，主要在於自發性語音與朗讀式語音有很大的文法型態差異，而且自發性語音未經大腦良好的規劃，使得語流中常會出現遲疑、口吃、詞語修補等現象，因此較難估測出詞與詞之間的關聯程度；另外自發性語音中有許多口語化詞彙或不同語者的有不同的慣用語，但是在書寫上的文章並不會有此類文字出現，造成文字資料量不足以致於無法直接建立自發性語音之語言模型。

近年來許多研究者皆使用基於結合之方式調適出自發性語音之語言模型，首先在解決資料尚不足的問題時，Ng and Ostendorf 【1】利用從網路上收集文字資料，再與其基本語言模型作調適結合；Hiroaki and Tatsuya 【2】則使用基礎語言模型對其語料作辨識，再從辨識結果中挑出較佳之文字資料，與原本的基礎語言模型作調適結合，如此反覆結合及訓練出較好的語言模型。接著在調適訓練時，Bacchiani and Roark【3】使用最大事後機率(Maximum A Posteriori, MAP)估算法來估算機率；另外也有許多文獻利用基於分類(Class-based)的方法將有限的文字語料藉由分類(例如：詞性(POS))來增加訓練資料量【4】【5】【6】。

1.2.2 韻律模型幫助辨認之相關研究

除了由語言模型去計算詞與詞之間的關聯程度，若能有效地運用詞與韻律結構之間的關係，對於語音辨認上將是一大幫助。近年來利用韻律資訊協助語音辨認之方法主要分為三種，首先是以事件為基礎(event-based)的方式增加語音辨認之效能【7】，利用韻律參數建立一個偵測事件之模型，例如：類語句邊界(sentence-like unit)或詞語修補中斷點，並利用事件及詞的序列一起建立語言模型，對辨認結果所產生之詞格(word lattice)重新計算分數；第二種為利用韻律參數對初級辨認結果所產生之詞格重新計算分數，直接利用韻律參數來驗證在詞格中不同路徑其對應切割位置之可靠程度【8】；第三類則是利用韻律以及句法的關係建立

一套韻律相關之語言模型(prosody-dependent language model) 【9】，用以描述韻律以及詞之結合機率，並利用韻律邊界的資訊建立韻律相關的聲學模型(prosody-dependent acoustic model) 【10】。

1.3 研究方向

在本篇論文中，將以建立一套自發性語音辨認系統並以韻律模型協助辨認為目標。首先，本研究先建立一套語音辨識系統及韻律模型，架構圖如圖 1.1 所示，在聲學方面，使用隱藏式馬可夫模型(Hidden Markov Model, HMM)以聲母及韻母為單位，採用音節內右相關聲母/韻母模型；在語言學方面，先利用大量文字資料建立合呼朗讀式語音之語言模型，接著使用 MAP 及 Class-based Deleted Interpolation Smoothing 方法調適成合乎自發性語音之語言模型；最後，本研究利用【11】所提出之非監督式中文自發性語音韻律標記及模型為基礎作修改，獲得本研究使用之韻律模型。

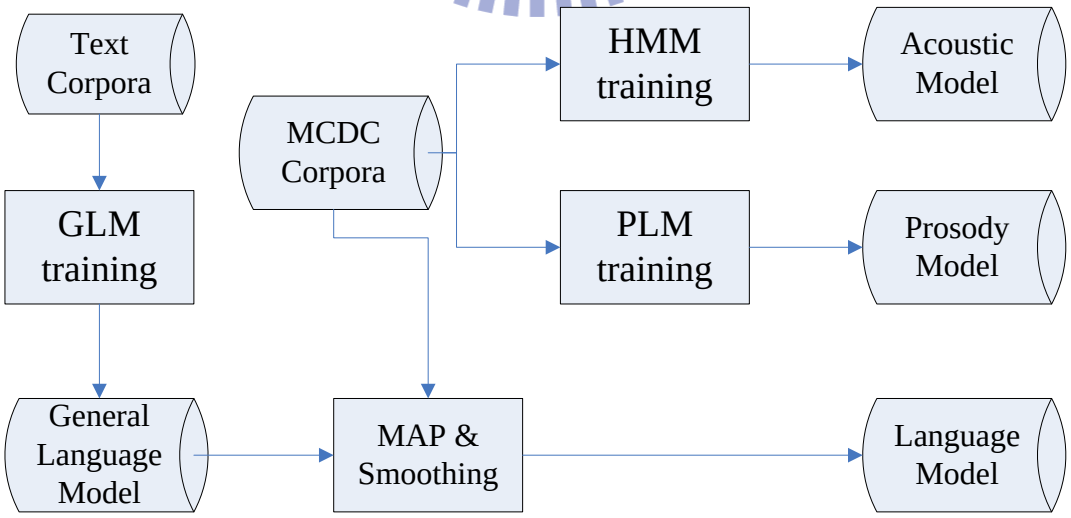


圖 1.1：訓練模型架構圖

在辨認階段本研究將採用兩段式(two pass)語音辨認架構。第一階段利用聲學模型計算語音信號聲學分數，接著利用語言模型計算詞與詞的關聯程度與出現機率，辨認產生最佳 N 條詞串(Top-N word sequence)；由於自發性語音中有許多不合乎文法結構之語句，詞組與詞組間之連接機率較不穩定，因此較難正確估算其關聯程度，但詞組與韻律結構之間的關係則較為明確，因此本研究在第二階段將另外利用韻律參數，給予每個音節邊界一個機率的分數，最後將此三個模型機率分數作權重結合並對每一條路徑重新計算分數，決定出最可能之辨認結果。

1.4 章節概要

本論文共分為六章，各章節編排如下：

第一章　緒論：說明研究動機與研究方向。

第二章　漢語口語對話語料庫介紹：介紹本論文實驗使用之自發性語音語料庫及其特性、統計分析以及後處理。

第三章　自發性語音辨識系統架構：說明自發性語音聲學模型之建立以及語言模型之建立及調適。

第四章　結合韻律模型之辨識系統：說明自發性語音韻律模型之建立、分析及實驗整體架構。

第五章　實驗結果：語言模型及結合韻律模型辨識系統之實驗結果。

第六章　結論與未來展望。

第二章 漢語口語對話語料庫介紹

本研究將使用中央研究院語言學研究所提供一個完整的自發性語音語料庫－現代漢語口語對話語料庫(Mandarin Conversational Dialogue Corpus，MCDC)作為研究素材。現代漢語口語對話語料庫【12】是由中央研究院語言學研究所曾淑娟博士等人於 2000～2002 年間所錄製，其語者是由台北市民隨機抽樣，並依據 16～25 歲、26～35 歲以及 36～45 歲三大年齡層，選出 60 位語者(37 位女性、23 位男性)，共錄製 30 段對話，但其中有轉寫的對話僅有 8 段對話，分別為編號 01、02、03、05、09、10、25 以及 26，其中包含了 16 位語者(9 位女性、7 位男性)。本研究將以此 8 段對話作為實驗之語料。本章節為此語料庫作簡介，包括語料標註(transcription)格式、此語料庫之自發性語音特性、MCDC 語料之後處理以及其他相關統計。

2.1 語料庫介紹

2.1.1 音檔格式說明

MCDC語料庫為兩語者對話之語料，音檔採雙聲道且取樣率為48kHz的方式錄製，並且將兩位發音人之語料分別錄於左右聲道，再利用軟體 Cool Edit Pro將它們分割成小的雙聲道音檔，依長度約三分鐘找到一個清楚可辨的停頓切開，其簡介如表2.1所示。在本研究中將每組對話語料之左右聲道抽取，並轉換為兩個單聲道之音檔，分別為對話中兩位語者之語料，並且將其取樣頻率下降至16kHz，再利用每一段落相對應之開始及結束時間作切割，經由以上處理後產生7,085段音檔，扣除一些未包含語音現象之音檔後剩下之6,570段音檔將作為本研究所使用之語料。

表 2.1：MCDC 資料庫的對話主題與語者對照表

對話序號	長度 (分鐘)	發音人	聲道 (L/R)	語者編號	對話主題
mcdc-01	61	MISC-08-male-25	R	01R	工作、休閒活動
		MISC-07-female-29	L	01L	、經濟、開車
mcdc-02	63	MISC-10-male-35	R	02R	休閒活動、經濟
		MISC-09-female-37	L	02L	、工作、性別、政治
mcdc-03	61	MISC-12-female17	R	03R	家庭、學校、購物
		MISC-11-female16	L	03L	、生涯規劃、明星
mcdc-05	63	MISC-15-male-40	L	05L	工作、家庭、社會階級
		MISC-16-female-46	R	05R	、保險、歷史、省籍情節
mcdc-09	66	MISC-23-female-30	R	09R	工作、旅行、生活態度
		MISC-24-female-35	L	09L	、環保、健康
mcdc-10	54	MISC-26-male-23	R	10R	電影、政治、軍隊
		MISC-25-male-35	L	10L	、捷運、學校、經濟
mcdc-25	55	MISC-57-male-43	L	25L	交通、工作、小孩
		MISC-58-female-45	R	25R	、旅行、電腦、管理
mcdc-26	46	MISC-60-male-24	R	26R	工作、求職、家庭
		MISC-59-female-37	L	26L	、車禍、學英文、婚姻、軍隊

2.1.2 語料標註格式說明

本研究採用中央研究院語言學研究所釋出之版本，在標註時大致以對話中語者轉換處為一個段落作轉寫，轉寫內容主要包括：對應之音檔名稱、語者代號、音檔起始及結束時間、語音之文字轉寫以及其發音相對之漢語拼音，文字以及漢語拼音的轉寫包括語言及非語言部

分，非語言部分主要是標記非人類產生聲音以及人類所產生但不是語音的聲音，例如：咳嗽聲、笑聲、呼吸聲等。以下為一個段落之文字轉寫範例及說明：

```
<segment>                                                    /*語者轉換開始處*/
<voicefile>D:\MCDC\stereo_01\mcdc-01-01.wav                  /*對應之音檔名稱*/
<speaker>MISC-07-female-29                                   /*語者代號*/
<start>077458                                /*此段文字轉寫對應音檔之開始時間*/
<end>080547                                  /*此段文字轉寫對應音檔之結束時間*/
<translator>Fen                                              /*文字轉寫人*/
<chinese>                                                    /*語者發音內容之文字標記*/
 NA 賴先生呢您從事什麼工作 (unrecognizable non-speech sound)
</chinese>
<english>                                                    /*語者發音之漢語拼音標記*/
NA lai4 xian1 sheng1 [nen2] nin2 cong2 shi4 shen2 me5 gong1 zuo4 (unrecognizable non-speech sound)
</english>
</segment>
```

2.2 自發性語音之特性

自發性語音與朗讀式語音最大的差異在於朗讀式語音是經過事先設計好的，人類在說話時常常會伴隨著因大腦思考或情緒變化而產生一些無法預期的聲音或發生在語言學中較詞層次(word level)更為上層之行為，增加許多訓練與辨認時的困難度。以下我們將介紹幾種 MCDC 語料庫中常出現的特性：

- 感歎詞(particle)

不具標準語意的感嘆詞，其語用成份居多如回應或同意。語流中出現的感歎詞有四類：一、有相對應國字的感歎詞，例如：A、BA、LA、MA、O；二、無相對應國字的感歎詞，例如：AI YE、EI、HEN、NE、NEI；三、源於台語的感歎詞，例如：EIN、HEIN、HO；四、其他的感歎詞(Fillers)，例如：UHN、NHN、MHM、MHMHM。

- 無法或難以辨識的語音

無法或難以辨識的語音主要可分為無法辨識的語音(unrecognizable speech sound)以及不確定字/音(uncertain)。無法辨識的語音為標記員確定此為人類所發出之語音但無法辨認何字何意何音；而不確定字/音包括：一、可猜測出大概的語音內容，但無法百分之百確定；二、無法根據語意猜測出對應字詞，但可清楚記錄出其發音。

- 非語言聲音(Non-Speech sounds)

在口語對話語料庫中常常會有一些非語音的聲音出現，非語言部分可分為人類所產生之副語言現象(para-linguistic)或非語言現象(non-linguistic)。一般的非語音但確定是由人所發出來的即稱為副語言現象，例如：笑聲、咳嗽聲、吞口水聲等等；而非語音且確定不是由人所發出來的則稱為非語言現象，例如：背景的雨聲、敲擊到麥克風聲等等。

- 語流中斷

在本研究中關注之語流中斷主要有沉默(silence)、停頓(pause)或短停頓(short break)，為語者在語流中因話題銜接不上或自身所產生之沉默。

- 非流暢現象(disfluency)

不流暢的語音為自發性語音中一個重要特性，在本研究中關注之詞語修補主要有重覆(repetition)、詞語更正(repair)、部分重覆(restart)以及更正插語(editing term)，重覆是指完整地重覆詞語一次以上；詞語更正為說話者覺得說出的話不適當，立即更正說話內容；而部分重覆則是說話者重新說出這個句子且重覆起頭詞語的片斷，與完整的詞語重覆不同。更正插語是出現在被更正詞語(reparandum)與更正詞語(correction)之間，或是出現在完整重覆或部分重覆中，兩個重覆詞語之間。本研究定義詞語修補中斷點(IP)為被更正詞語與更正後詞語間之停頓點，或完整重覆或部分重覆中的兩個重覆詞語間之停頓點，本研究在文字轉寫中將詞語修補中斷點標記成「*」。以下為幾種非流暢現象範例：

基本型態：　(被更正詞語)*[更正插語](更正詞語)

- 重覆範例：　昨天卡卡表現的(普通)*(普通)
- 詞語更正範例：　今晚世足賽是(烏拉圭)*[EN](巴拉圭)對日本
- 部份重覆範例：　(今)*(今天)晚上是冠軍賽

2.3 MCDC 語料庫之後處理

由於中央研究院語言學研究所釋出之 release 版本中並無標記某些聲學現象及語言資訊，但這些資訊在語音處理中相當重要，因此在本節介紹本研究將對語料庫聲學以及語言學資訊的處理方法及相關統計。

2.3.1 斷詞與相關統計

由於自發性語音目前尚無良好之語言模型，而在調適出好的語言模型之前需要有正確的斷詞標記，因此本研究先利用國立交通大學語音處理實驗室文字處理器以及中研院提供之斷詞器對MCDC之文字轉寫進行斷詞標記，詞性標記以中研院的46類詞性(Part of Speech, POS)(參見附錄一)為標準，接著再將兩者之斷詞標記進行強迫對齊以求出較佳之斷詞結果，最後以人工檢查方式修正出最佳斷詞標記。斷詞標記方式如下：

- 一般詞：以”( )”標記其詞性。例：我(Nh)、公司(Nc)
- 感歎詞：以半形字體表示。例：LA、HON
- 英語：以全形字體表示並標記”(FW)” 。例：VCD(FW)、Why(FW)
- 副語言現象(paralinguistic)：以”{ }”標記。例：{非語音聲}、{呼吸聲}

斷詞標記範例如下：

※原始內容：

ｏｋ NA {重覆開始} 我 {重覆中斷點} 我 {重覆結束} 叫 賴
{吸氣聲} HEN 你好 NA 最近在從事些什麼事情
O 我在一家公關公司上班 {非語音聲}
我目前是從事 {吸氣聲} EN 外貿

※斷詞後內容：

ｏｋ(FW) NA {重覆開始} 我(Nh) {重覆中斷點} 我(Nh) {重覆結束} 叫(VF) 賴(Nb)
{吸氣聲} HEN 你(Nh) 好(VH) NA 最近(Nd) 在(P) 從事(VJ) 些(Nf) 什麼(Nep) 事情(Na)
O 我(Nh) 在(P) 一[Neu]家[Nf](DM) 公關(Na) 公司(Nc) 上班(VA) {非語音聲}
我(Nh) 目前(Nd) 是(SHI) 從事(VJ) {吸氣聲} EN 外貿(Na)

接著統計MCDC語料庫之語言學相關資訊，主要有詞長(word length)以及詞性。MCDC語料庫經由本研究斷詞整理後共斷有79935個詞，圖2.1統計出各個詞長之個數，由圖2.1可明顯看出詞長以一字詞及二字詞居多，共佔總詞數93.66%；圖2.2為詞性之分布圖[1]，其中FW、ParL及Par分別表示foreign word、paralinguistic以及particle。由此可猜測人們在平時對話大多以一字詞或二字詞構成句子，而詞性則以動詞及名詞佔大多數。圖2.3為語料庫中每個sub-turn的音節數分佈圖，由圖可看出此語料庫以短句居多[2]。

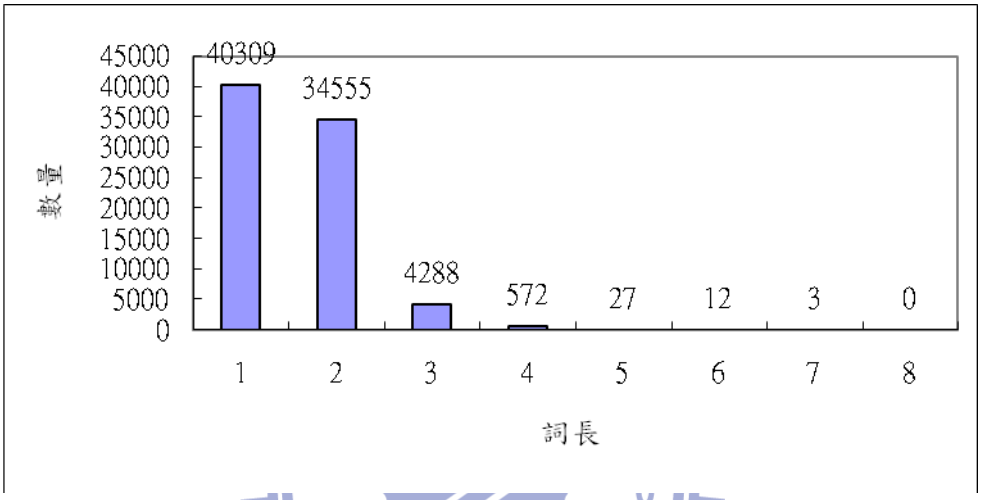


圖 2.1：語料庫中詞長分布圖

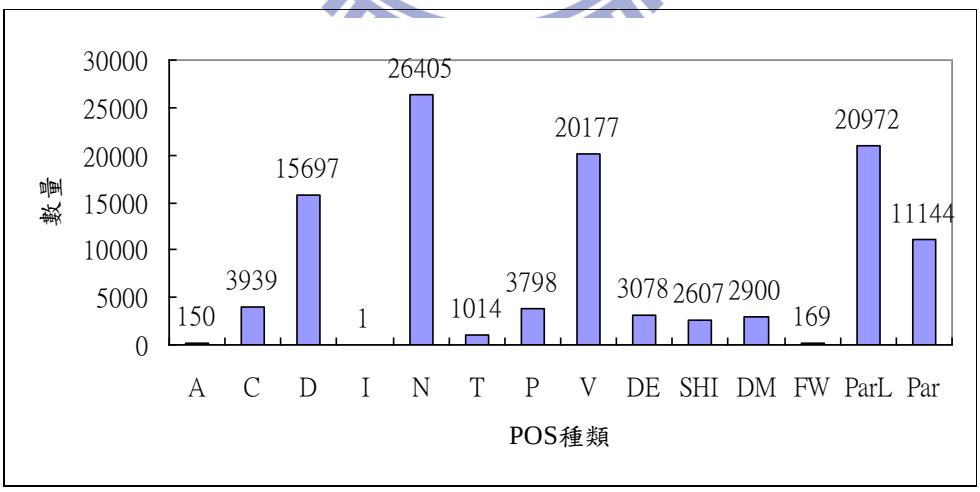


圖 2.2：語料庫中 POS 種類分布圖

[1] 在此為了觀察方便，因此將 46 類 POS 簡化統整為 8 類 POS，其對照表請參考附路一。

[2] 在 MCDC 語料庫中，短句主要以 particle、簡答或回應詞為主。

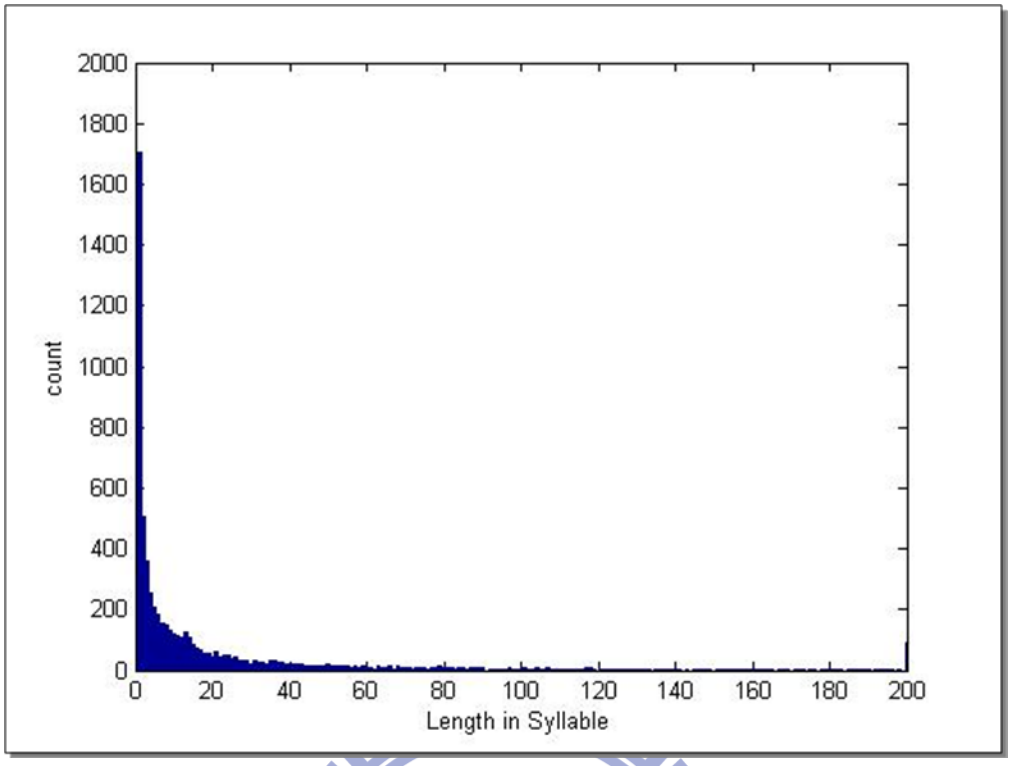


圖 2.3：語料庫中每個 sub-turn 之音節數分佈圖

2.3.2 標記標點符號

在訓練韻律模型時，標點符號(punctuation makes)能提供不少的韻律停頓(prosodic break)資訊，例如在頓號「、」的地方通常會有小停頓，而在逗號「，」之處則會有較長的停頓，句號「。」亦是如此且可能會有句尾拖長音現象發生。因此本研究將對 MCDC 文字語料進行標點符號標記，標記之符號包含「，」、「、」、「。」、「！」、「：」及「？」。標記時主要以一般文法斷法為主，而遇到 particle 時則判斷其與前一個或後一個詞相接念起來是否通順，若不通順則用標點符號將其隔開。以下為幾個標記範例：

E ， 也(D) 不(D) 算(VG) LA ！

我(Nh) 搭(VC) 捷運(Nb) 到(VC) NE GE 捷運忠孝復興站(Nc) 。

NA 我們(Nh) 就是(D) 以(P) 台幣(Na) 在(P) 報(VC) 。

EN ， NA 請問(VE) 怎麼(D) 稱呼(VG) 您(Nh) ？

2.3.3 非流暢現象中斷點

如何處理非流暢現象在自發性語音辨認中是一個重要的議題，本研究對此語料庫之非流暢現象現象做初步統計。首先以 2.1.1 節中音檔切割完產生之 6570 個語句 sub-turn 為基準，經由統計後發現其中有 1225 個 sub-turn 發生非流暢現象現象，圖 2.4 與圖 2.5 為語料庫中發生非流暢現象現象 sub-turn 的音節數和機率分佈圖。另外語料庫中標記「重覆」、「詞語更正」以及「部分重覆」之總個數各為 1011、333 以及 665 如表 2.2 所示，而本研究定義所有語句中每個音節邊界(共 136801 個)都有可能是非流暢現象中斷點的候選邊界(candidate)，經由統計共有 2649 個中斷點出現，由此可知在所有音節邊界中出現非流暢現象中斷點之機率為 1.94%。

表 2.2：MCDC 資料庫中 disfluency 與 IP 出現之個數

	Repetition	Repair	Restart	總數
個數	1011	333	665	2009
IP個數	1518	362	769	2649

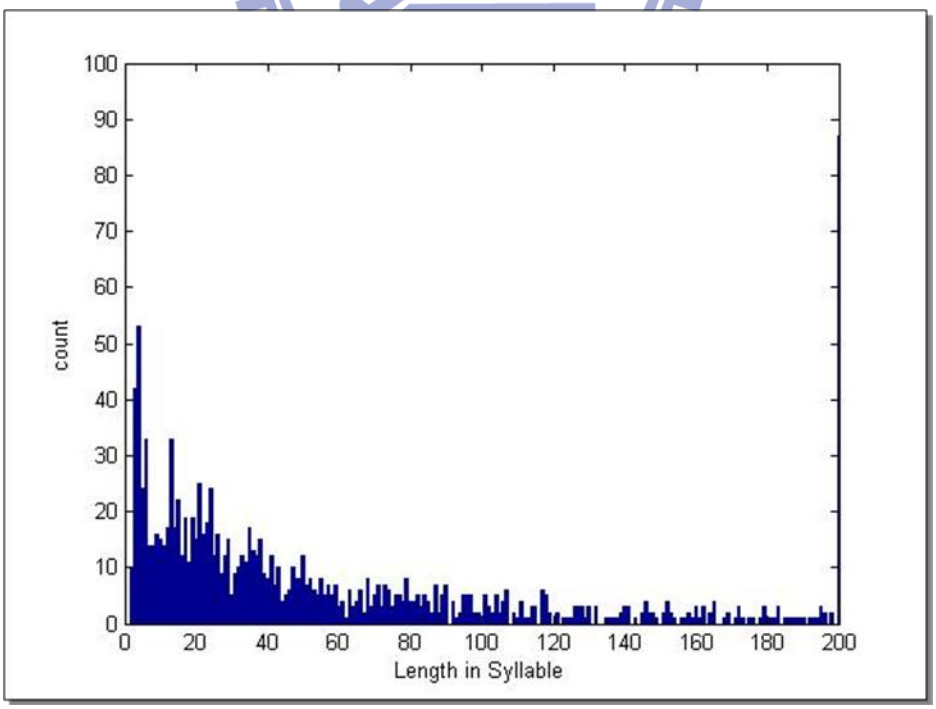


圖 2.4：發生非流暢現象現象之音節數分佈圖

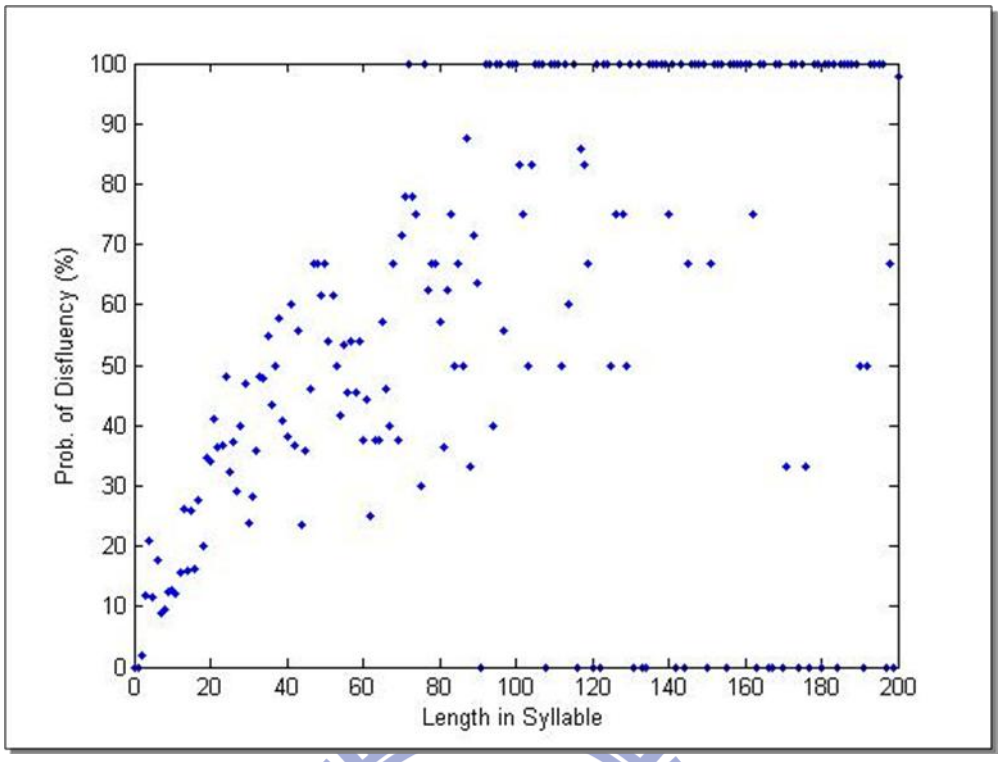
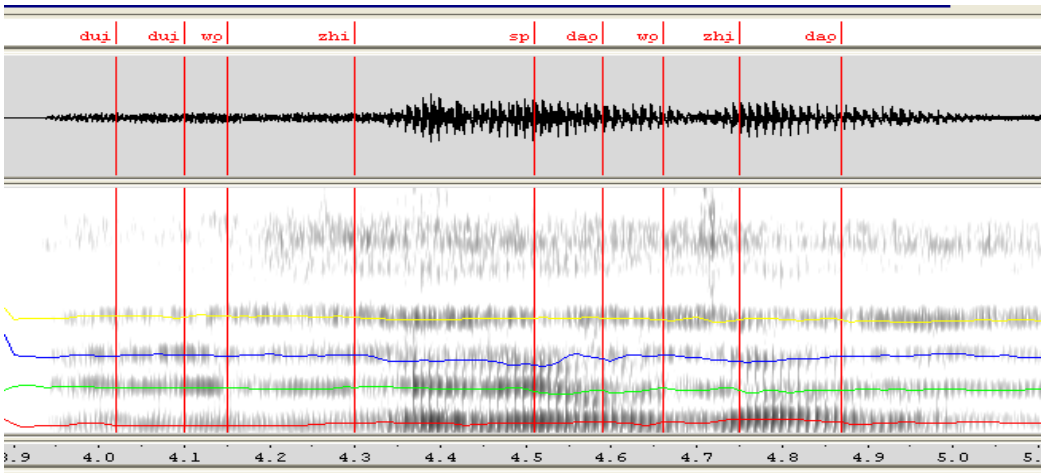


圖 2.5：不同音節數發生非流暢現象現象之機率分佈圖

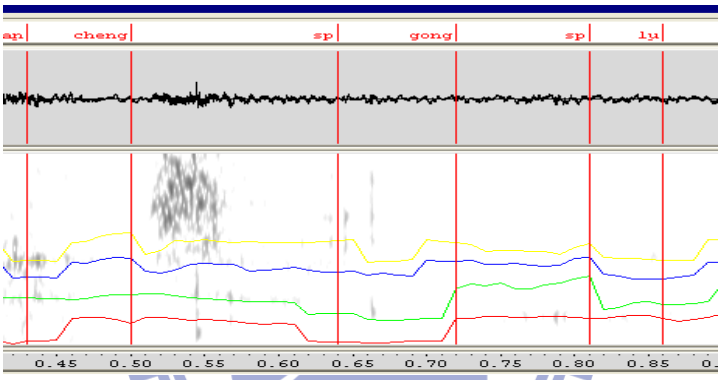
2.3.4 修正音節切割位置

要有好的聲學模型及韻律模型首先要有良好的音節切割位置，但目前自發性語音語料庫尚無準確的音節切割位置，而且由於自發性語音在發音時音量、頻率及時間長度較無法像朗讀式語音穩定，因此無法由自動的方式標示出良好的音節切割位置。如果要採用人工的方式標記方式則是一個浩大的工程，因此本研究首先使用【11】建立之自發性語音聲學模型，利用 HTK 軟體以強迫對齊的方式對每一個音檔作音節之切割，得到每一個音節的音節長度以及停頓長度；接著檢查若在詞內邊界(Intra-word)中出現超過 0.1 秒的 pause，則我們假設其中很有可能有音節切割錯誤的情況發生並修正之。在經由人工修正時，我們發現通常此現象發生在於語者在發音時帶有笑聲、音量太小、音節合併或遭受隔壁聲道干擾等情形，其中受到隔壁聲道污染更可能導致整段句子音節完全切割錯誤，因此將其移除。以下列出兩段例子：

對對我知道我知道-（嚴重連音+隔壁聲道污染）



成功路-（音量太小變氣音）



第三章 自發性語音辨識系統

在本章節中將建立一套自發性語音辨識系統，主要可分為聲學模型及語言模型。本研究利用 MCDC 語料庫以及劍橋大學開發之 HTK(HMM Tool Kit)軟體【13】建立一套自發性語音聲學模型；另外利用 MCDC 之文字資料以及朗讀式語音之文字資料調適出自發性語言模型。3.1 節將介紹聲學模型之建立流程；3.2 節將介紹建立及調適語言模型的方法。

3.1 聲學模型

3.1.1 訓練語料及測試語料

如何從語料庫分配訓練語料及測試語料主要依據作何種語音辨識系統而定。而語音辨識的類型大致上可分為三種：語者相依(Speaker Dependent, SD)辨識、語者獨立(Speaker Independent ,SI) 辨識以及多語者(Multi-Speaker)辨識。對於這三種辨識系統的語料庫分配方式如下：

- 語者相依辨識系統

  測試語料和訓練語料是同一語者。

- 語者獨立辨識系統

  訓練語料與測試語料的語者是不同人的。

- 多語者辨識系統

  訓練語料與測試語料有相同的語者但測試語料與訓練語料之句子是不同的。

本研究目的在於建立一套自發性語音辨認系統且找出韻律能用來幫助辨認的方法，因此是採用多語者辨識系統。本研究中訓練語料包含 16 位語者各自取 9/10 之語音段落所組成，而剩餘之 1/10 語料即為本研究所使用之測試語料，其詳細統計資料如表 3.1 及表 3.2 所示：

表 3.1：訓練語料統計

	411 syllable	Particle	Paralinguistic	Uncertain	Filler	Foreign Word
音節數	104,736	9,688	11,289	3,725	1,743	156
總音節數 總段落數 音檔長度	131,337 6,121 約 8.97 (hours)					

表 3.2：測試語料統計

	411 syllable	Particle	Paralinguistic	Uncertain	Filler	Foreign Word
音節數	12,156	916	1,186	546	253	20
總音節數 總段落數 音檔長度	15,077 417 約 1.09 (hours)					

3.1.2 聲學模型之建立

3.1.2.1.特徵參數抽取

在訓練模型前，首先必須獲得足以充分描述語音特性，且參數量較原語音信號小之特徵參數，而語音處理當中，最廣泛為人使用之特徵參數為梅爾頻率倒頻譜係數(Mel-Frequency Cepstrum Coeiffient, MFCC)，本研究也將使用此特徵參數，以 32 毫秒之漢明窗(Hamming window)且每位移 10 毫秒為一筆資料，求取 12 維 MFCC 並加上一維能量係數，以及這 13 維係數之一階與二階變量(delta and delta-delta)為特徵參數，但單純的能量在參數中較缺乏鑑別性，因此去除能量係數，得到 38 維向量作為本研究語音資料之聲學特徵參數。在本研究也將利用倒頻譜平均值正規化法(Cepstrum Mean Normalization, CMN)藉此消除不同語音信號之通道效應。

3.1.2.2 聲學模型之建立流程

本研究利用【11】所建立之自發性語音聲學模型，反覆訓練自發性語音之聲學模型及修

改出較正確之文字轉寫。其建立流程如圖 3.1 所示。由於自發性語音尚無良好之語料標註，然而錯誤的標注將可能導致聲學模型混淆以致辨認效能降低，但是要將文字轉寫以人工方式完全修改正確是一個極大的工程，因此在訓練模型前本研究先利用其聲學模型將 MCDC 語料之漢語拼音轉寫和語音信號作強迫對齊，獲得每一個音節切割位置的資訊，接著利用 2.3.4 中提出的方法對切割位置做些微修正。

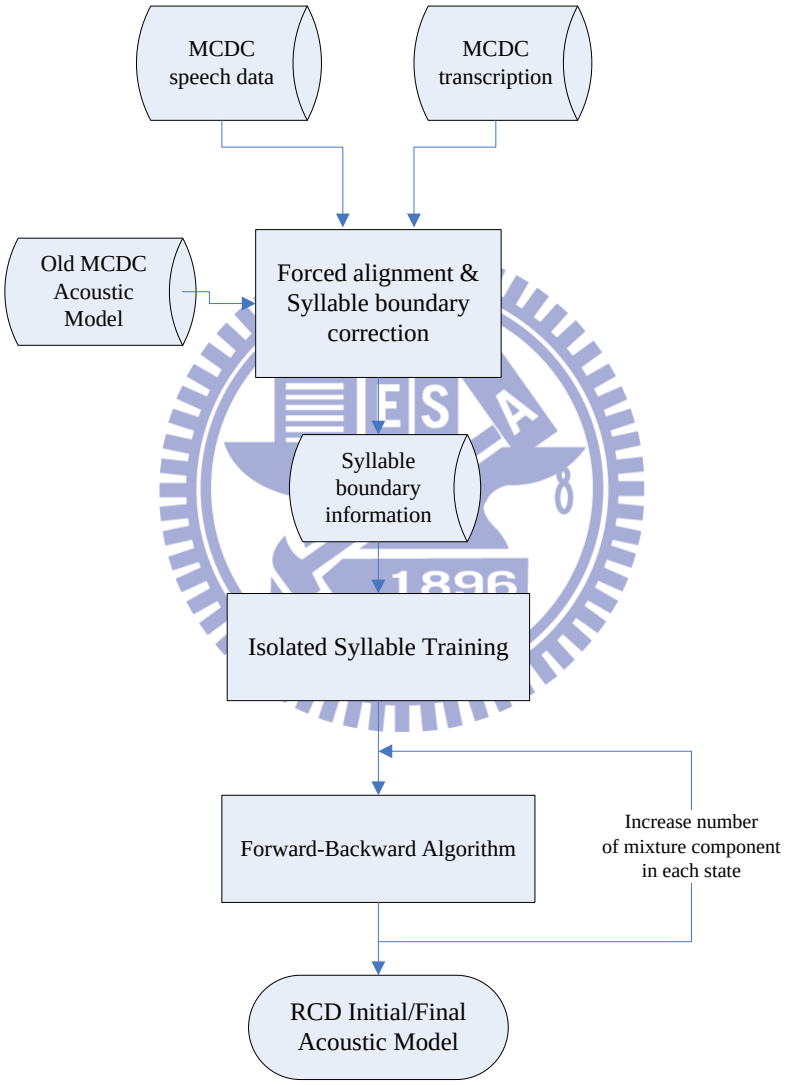


圖 3.1：聲學模型之建立流程

在有了較好的切割資訊後，本研究接著利用此切割資訊訓練各個音節之聲學模型，值得注意的是，在自發性語音當中具有許多基本音節之外的其他音節，例如：語助詞、感歎詞、

不確定字或音以及副語言現象，在本研究中這些特殊音節同樣將利用【11】所建立之自發性語音聲學模型，經由強迫對齊後獲取特殊音節之切割位置，並另外建立其聲學模型，在此數量過少之其他音節將被歸類至填充模型(filler model)。

本研究使用之隱藏式馬可夫模型以聲母及韻母為單位，採用音節內右相關聲/韻母模型(Right-ocntext-dependent Initial/Final Model, RCD)。每一個聲母之 HMM 模型採用 3 個由左至右(left-to-right)的狀態(state)表示；而韻母之 HMM 模型則採用 5 個狀態來表示，另外填充模型以及短靜音將以 1 個狀態來表示。其中每一個狀態以平均 64 個高斯分布之高斯混合模型(Gaussian Mixture Model, GMM)描述其特徵參數之分布，各模型之 HMM 設定如表 3.3 所示。

表 3.3：HMM 模型之設定

HMM 模型類別	狀態個數	模型數量[3]
RCD initial	3	100
Final	5	40
Particle	3	24
Uncertain	3	73
Filler	1	1
Foreign word	3	1
Silence	3	1
Short pause	1	1
Paralinguistic	3	9

值得注意的是，由於人類在口語對話時常因為節省發音力氣而省略音節內的某些發音，因而產生了音節合併(syllable contraction)的現象。在此本研究利用 HTK 軟體計算各個音節之狀態轉移機率，若狀態轉移機率大於 50%，則允許此狀態可以被跳過，如圖 3.2 所示，並重新對模型做訓練至收斂為止。

[3]在 MCDC 語料庫缺少「c_o」、「n_o」以及「s_o」之聲母模型；「eh」、「yai」以及「yo」之韻母模型。

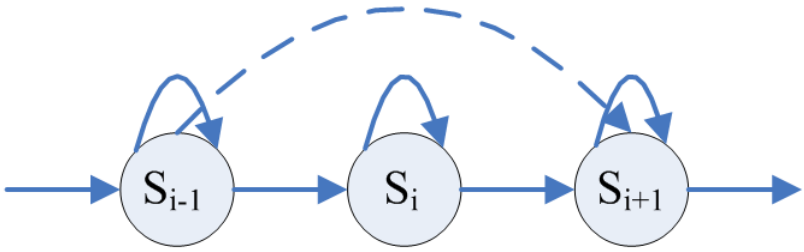


圖 3.2：加入狀態轉移機率示意圖

3.1.3 實驗結果

一般來說聲學模型之效能是由辨認率來評估，因此本研究將使用 3.1.2 建立之 MCDC 聲學模型對測試語料進行音節辨認以評估其效能。在做辨認之前，本研究先建立一個 free-gram 音節語言模型，其中包含 411 音節、Foreign word、filler、Particle 及 Paralinguistic 且各個相接機率相等。接著，本研究使用 HTK 工具對測試語料進行辨認，採用音節作為辨認的基本單元，其中包含 411 音節及 Particle、Paralinguistic、Filler、Foreign word(Eng)。

表 3.4：MCDC 音節辨認率

	Correct	Accuracy	Hit	Deletion	Substitution	Insertion	Total
Syllable	51.74%	48.74%	7,779	2,139	5,117	451	15,035
411only	52.45%	48.55%	6,868	1,684	4,542	511	13,094

本研究之音節辨認率如表 3.4 所示，由表中可以發現，自發性語音中刪除型錯誤(Deletion)數量較多，經由辨認結果之觀察，我們可以發現此乃因為在口語對話中因為語速較快，造成許多詞語會有音節合併現象發生，例如：「因為」、「這樣」、「所以」…等等。當發生音節合併現象時，通常會增加取代型(Substitution)及刪除型(Deletion)錯誤，以下為一種因音節合併造成錯誤之範例：

正確答案(canonical form)	辨認答案(surface form)	錯誤型態(error type)
zhe (這)	jiang (降)	取代型
yang (樣)	NULL	刪除型

本研究另外實驗單純辨識 411 音節之辨認，由表 3.4 可以發現辨識率略微下降，此乃因為自發性語音中有許多感歎詞、語助詞以及副語言現象的情況存在，若我們能有效地將這些現象加入當作辨識單元，使其不干擾正常語音之辨識，辨識率將能有效提升，對於要做語音之自動文字語料標註(transcribing)也將會有很大的幫助。最後，本研究也注意到目前語料庫的音節切割位置還是有許多錯誤，但因為人工修正切割位置是一個浩大的工程，所以目前都只能以半自動的方式修改，若能有效改善此狀況，對於聲學模型的訓練將會有極大的幫助。

3.2 語言模型

由於所有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，一般稱此為語言模型(Language Model, LM)。在語音辨認時，除了聲學模型外，若能加入語言模型的參考，通常能大幅提升辨認系統的效能。由於自發性語音至今尚無良好的語言模型，在此本研究將利用現有的朗讀式語音文字資料以及 MCDC 的文字資料調適出一套適用於自發性語音的語言模型。

3.2.1 初始模型之建立

3.2.1.1 訓練語料

本研究使用光華雜誌(Sinorama)、NTCIR 以及中研院的平衡語料庫(Sinica Corpus)訓練出一套屬於 read speech 的語言模型(以下皆稱此模型為 General-LM, GLM))。所有訓練之文字資料數量如下：

表 3.5：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

3.2.1.2 模型訓練

本研究先利用 General-LM 文字資料，以統計的方式計算詞與詞之間的聯接規則，且利用退化平滑法(back off)及使用 Good-Turing discounting 建立而成一雙連(bi-gram)語言模型。其數學式如(3-1)式所示。若定義訓練語料中詞串出現的次數門檻值 k，則可將詞串分為出現

次數高於門檻值、出現次數低於門檻值及從未出現三種。則參數可表示為下式：

$$P(w_i \mid w_{i-1}) = \begin{cases} \alpha(w_{i-1}) \cdot P(w_i) & C(w_{i-1}, w_i) = 0 \\ d_c(w_{i-1}) \cdot \frac{C(w_{i-1}, w_i)}{C(w_{i-1})} & 1 \le C(w_{i-1}, w_i) \le 5 \\ \frac{C(w_{i-1}, w_i)}{C(w_{i-1})} & C(w_{i-1}, w_i) > 5 \end{cases} \quad (3\text{-}1)$$

有了訓練語料及方法，便可以開始訓練語言模型，訓練過程如圖 3-3 所示：

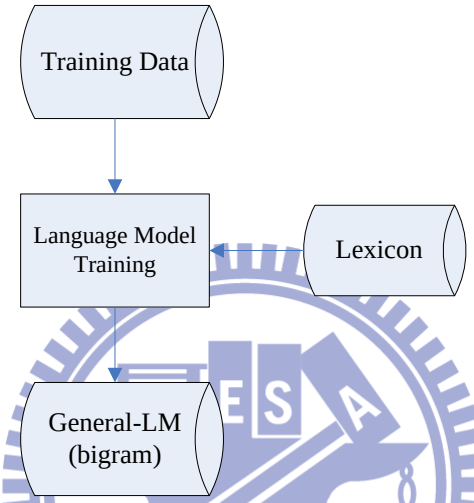


圖 3.3：General-LM 訓練流程圖

其中詞典(Lexicon)的產生是先對大量的文字資料進行斷詞，然後依詞頻排序並挑出詞頻最高的前 6 萬詞作為詞典。經過圖 3.3 之訓練流程後，即完成建立一套雙連 General-LM，以下本研究將利用此語言模型與 MCDC 文字資料進行調適。

3.2.2 語言模型之調適

3.2.2.1 自發性語音之訓練單元

由於 General-LM 中並沒有估算自發性語音中的「particle」及「paralinguistic」的出現機率，因此要調適出一套適用於自發性語音語言模型前，我們勢必要想辦法估算出 particle 及 paralinguistic 的機率並將 particle 及 paralinguistic 的辨認單元加入到辨認字典中。在此，本研

究將挑出 27 類 particle 及 9 類 paralinguistic 當作新的辨認單元，如下表所示：

表 3.6：在辨認字典中加入之 27 類 particle

A	HAN	ME	O
AI	HEIN	MHM	SHEN
BA	HEN	MHMHM	WA
E	HO	NA	YA
EI	LA	NE	YOU
GE	MA	NO	ZHE
NE-GE	SHEN-ME	ZHE-GE	

表 3.7：在辨認字典中加入之 9 類 paralinguistic

呼吸聲	清喉嚨聲	咳嗽聲	吞口水聲	咂嘴聲
語音聲	非語音聲	笑聲	雜訊	

3.2.2.2 調適模型之設計

在本研究中，我們將 MCDC 的文字資料區分成三種資料類別：一般詞(Lexical Word - LWord)、particle(Par)以及 paralinguistic(Para)。首先，我們可以從 MCDC 文字資料中統計出此三種資料類別的出現機率，分別為：$P_{T_S}(G(w_i)=LWord)=0.763$；$P_{T_S}(G(w_i)=Par)=0.110$；$P_{T_S}(G(w_i)=Para)=0.127$，如圖 3.4(a)所示。本研究將計畫把 General-LM 中一般詞、particle 以及 paralinguistic 的機率比例根據由 MCDC 文字資料所估算出來的出現比例做加入及調整，如圖 3.4(b)所示。接下來我們將實驗分成三個階段，以不同的估算方法將「particle」及「paralinguistic」加入到語言模型中，得到一個較佳的模型來幫助辨識。

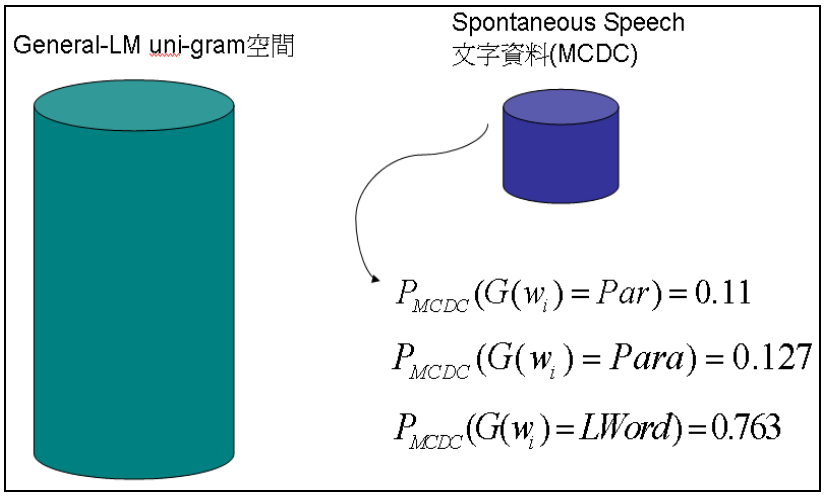


圖 3.4(a)：由 MCDC 估出各類別機率

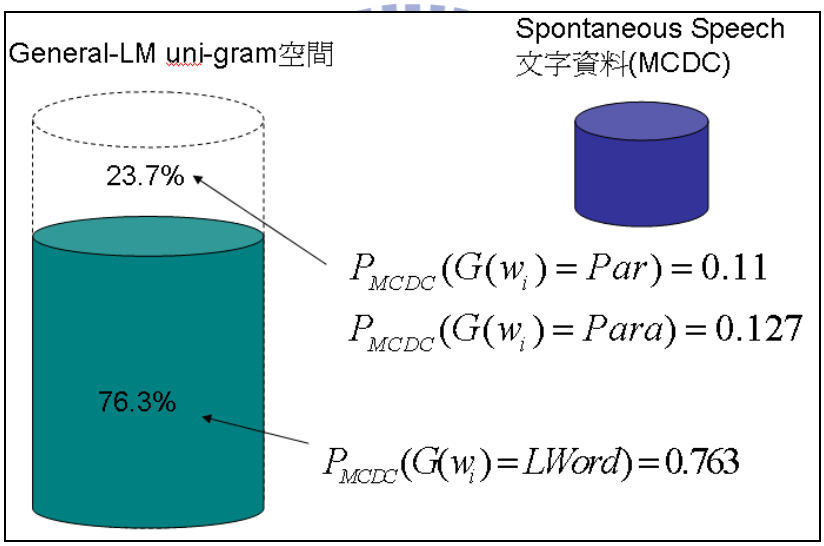


圖 3.4(b)：將 Gerneral-LM 依圖 3-4(a)估算的比例重新分配

➢ Stage 1：給予 particle 及 paralinguistic uni-gram 機率

在此階段我們將根據上述統計的各類別之出現機率來分配 General-LM 中 uni-gram 機率部分。首先我們先從 MCDC 文字資料中估出 particle 及 paralinguistic 的出現機率 $P_M(w_i)$，我們可以發現所有 particle 之機率總和將等於其在 MCDC 中的 particle 類別之出現機率，paralinguistic 亦是如此，如下式：

$$\sum_{w_i} P_M(w_i) = P_M(G(w_i)) \qquad ,w_i \in particle,\ paralinguistic \tag{3-2}$$

接著我們將此機率 $P_M(w_i)$ 乘上一個係數 λ，使得

$$\sum_{w} P'_M(w_i) = \lambda \cdot P_M(w_i) = 1 \qquad ,w_i \in particle,\ paralinguistic \tag{3-3}$$

最後，我們重新分配 General-LM 裡的 uni-gram 空間，其數學式如下：

$$P'(w_i) = \begin{cases} P_M(G(w_i) = LWord) \cdot P_G(w_i) & ,w_i \in lexical\ word \\ P_M(G(w_i) = Par) \cdot P'_M(w_i) & ,w_i \in particle \\ P_M(G(w_i) = Para) \cdot P'_M(w_i) & ,w_i \in paralinguistic \end{cases} \tag{3-4}$$

值得注意的是，其中一般詞是由 General-LM 中原本估算出來的機率乘上在 MCDC 中一般詞的出現機率。在此所有 word 機率總和將滿足：

$$\begin{aligned} \sum_{w_i} P'(w_i) &= \sum_{G}\sum_{w_i} P_M(G(w_i) = LWord) \cdot P_G(w_i) + P_M(G(w_i) = Par) \cdot P'_M(w_i) + P_M(G(w_i) = Para) \cdot P'_M(w_i) \\ &= \sum_{G} P_M(G(w_i) = LWord) + P_M(G(w_i) = Par) + P_M(G(w_i) = Para) \\ &= 1 \end{aligned} \tag{3-5}$$

在此階段因為我們並無估算 particle 及 paralinguistic 的 bi-gram 機率，因此我們將 particle 及 paralinguistic 的 back-off 係數皆設為 1。

➢ Stage2: 估算 particle 及 paralinguistic 的 bi-gram 機率

由於現有的 Read-speech(General-LM)文字資料量與 Spontaneous-speech(MCDC)文字資料量差異懸殊，而且在 General-LM 文字資料中不會有 paralinguistic 出現，particle 也極少，因此若直接將 General-LM 與 MCDC 文字資料合併訓練 bi-gram 語言模型則在估算一般詞與 particle 或 paralinguistic 的 bi-gram 機率時勢必會出現數量過少而導致機率極低的情形。然而，在觀察 MCDC 文字資料後，我們可發現某些詞與 particle 或 paralinguistic 相接的機率甚高，例如：假設我們直接將訓練 General-LM 的文字資料與 MCDC 的文字資料合併訓練則會因為 $Count(對, A) < Count(對, 一個)$[4] 而產生 $P'(A|對) < P'(一個|對)$ 的機率現象，但是我們可以由

[4] $Count(w_{i-1}, w_i)$ 表示詞 w_{i-1} 接詞 w_i 的次數。

MCDC 中觀察出「A」比「一個」的機率要高許多。為了防止此情形發生，我們將使用 Deleted Interpolation Smoothing【14】方法，在估算一般詞與 particle 或 paralinguistic 相接時藉由加入 particle 和 paralinguistic 的 uni-gram 機率來提升其 bi-gram 機率。

首先，傳統的 Deleted Interpolation Smoothing 公式如下：

$$P'(w_i|w_{i-1}) = \alpha \times P(w_i|w_{i-1}) + (1-\alpha) \times P(w_i) \quad (3\text{-}6)$$

其主要概念是：當原bi-gram機率 $P(w_i|w_{i-1})$ 較低時可能造成不可靠性增加，此時可利用與uni-gram機率 $P(w_i)$ 做interploate以達到補強的效果。此方法適用於結合高階(higher-order) n-gram 語言模型與低階(lower-order) n-gram語言模型，因為在一個高階n-gram語言模型中可能有些資料因為出現數量較低因此較不可靠，此時低階之n-gram語言模型將可能可以提供較可靠性的機率資訊。

再來值得注意的是，因為 MCDC 的文字資料量少，我們認為直接由 MCDC 中估出來的 particle 及 paralinguistic 的 uni-gram 機率 $P(w_i)$ 可能不可靠，一種最直接的解決方法就是將種類繁多但數量少的文字資料退化成種類較少的 Class 形式(例如：詞性)，如此一來每個 Class 的數量將會增加，再由此估算出較為可靠的機率。所以在此我們將使用 Class-based 的方法來估算(3-6)式中 $P(w_i)$ 之機率分數。

傳統的 Class-based bi-gram model 數學式定義如下：

$$P_{class}(w_i|w_{i-1}) = P(w_i|G(w_i), G(w_{i-1}), w_{i-1}) \times P(G(w_i)|G(w_{i-1}), w_{i-1}) \quad (3\text{-}7)$$

假設 $P(w_i|G(w_i), G(w_{i-1}), w_{i-1})$ 與 $G(w_{i-1})$ 及 w_{i-1} 獨立且 $P(G(w_i)|G(w_{i-1}), w_{i-1})$ 與 w_{i-1} 獨立，則 (3-7)式可改寫為：

$$P_{class}(w_i|w_{i-1}) = P(w_i|G(w_i)) \times P(G(w_i)|G(w_{i-1})) \quad (3\text{-}8)$$

由(3-8)式得知，欲估算出 $P'(w_i|w_{i-1})$ 前必須先定義好類別($G(w_n)$)，本研究先將分類出三種類別：$G(w_n) = LWord, Par, Para$，分別表示「一般詞」、「Particle」及「Paralinguistic」，

表 3.8 為由 MCDC 訓練語料中所估算的機率。

最後，我們即可將 (3-6)式和(3-8) 式結合改寫成以下形式：

$$P'(w_i|w_{i-1}) = P_M(w_i|w_{i-1}) + (1-\alpha)P_M(w_i|G(w_{i-1}))P_M(w_i|G(w_i)) \quad (3\text{-}9)$$

其中 $\alpha = \sum_{w_i} P_M(w_i|w_{i-1})$。在此值得注意的是， (3-9)式中在使用 Class-based 方法時是估算前一個詞 w_{i-1} 之類別接到某個詞 w_i 之機率 $P_M(w_i|G(w_{i-1}))$，而不是估算 $P_M(G(w_i)|G(w_{i-1}))$，因為經由觀察我們發現某些詞組 $Count(w_{i-1}, w_i)$[5] 當詞 w_i 為某些常出現的 Particle 或 Paralinguistic 時(例如：MHM、A、@呼吸聲)，若單純使用 $P_M(G(w_i)|G(w_{i-1}))$ 去做 Smoothing 時可能因為 $P_M(w_i|G(w_i))$ 機率很大造成估算錯誤。例如：在 MCDC 文字資料中 $Count$(是,O) 並不多，而 $Count$(是,MHM) 更是稀少，因此需要依賴 Smoothing 機率，然而在 Smoothing 時若直接使用 $P_M(G(w_i)|G(w_{i-1}))$ 則會因為 $P_M(O|G(Par)) < P_M(MHM|G(Par))$ 而造成 $P'(O|是) < P'(MHM|是)$ 的狀況發生，但是我們知道在「一般詞」的前提下，「O」的出現機率會比「MHM」高出許多。因此本研究將採用 $P_M(w_i|G(w_{i-1}))$ 來做 Smoothing 動作。

(3-9)式是估算詞與 particle 及 paralinguistic 的相接機率，而要放入原本 General-LM 的 bi-gram 時，因為對於某個詞 w_{i-1} 而言，其本身的 bi-gram 機率總和滿足 $\sum_{w_i} P_G(w_i|w_{i-1}) = 1$，若直接將我們估算出來的機率放入，則會造成機率總和大於 1 的情況發生，因此我們將使用類似 Stage 1 時的概念，將這個詞的機率空間重新作分配，因此對於某個詞 w_{i-1} 而言，我們的式子如下：

$$P'(w_i|w_{i-1}) = \begin{cases} P_M(G(w_i)|G(w_{i-1})) \cdot P_G(w_i|w_{i-1}) \\ \qquad , w_{i-1}, w_i \in lexical\ word \\ P_M(G(w_i)|G(w_{i-1})) \cdot \left[P_M(w_i|w_{i-1}) + (1-\alpha)P_M(w_i|G(w_{i-1}))P_M(w_i|G(w_i)) \right] \\ \qquad , w_{i-1} \in lexical\ word\ ,\ w_i \in particle, paralinguistic \end{cases} \quad (3\text{-}10)$$

[5] 此處之 $Count(w_{i-1}, w_i)$ 中定義其 w_{i-1} 與 w_i 不會同屬於 lexical word。

表 3.8：MCDC 中各類別之相接機率

$G_M(w_{i-1})$ \ $G_M(w_i)$	Lexical word	Particle	Paralinguistic
Lexical word	0.830	0.068	0.073
Particle	0.426	0.168	0.187
Paralinguistic	0.482	0.119	0.192

- **Stage3: 調適一般詞之機率**

本研究在建立 General-LM 時是利用了大量的文章文字資料訓練且較具有普遍性(General)，而且在前面階段我們已經估算出一些自發性語音才有的特殊單元之機率，但是因為 MCDC 語料庫實質上是屬於不同領域(domain)的資料，使得初始語言模型在估算一般詞的出現機率仍較不準確，若能結合擁有大量資料的 General-LM 和 MCDC 文字資料在同領域下的優點去進行語言模型調適，則可以使得新的語言模型更加合乎自發性語音的特性。

首先，觀察 2.3.1 中圖 2.1 可發現在 MCDC 語料庫中一字詞與二字詞出現機率極高，這些一、二字詞中又以一些特定詞(例：我、你、他、對、有、是、嗎、就是、我們)居多，因此本研究將利用最大事後機率(Maximum A Posteriori, MAP)估算法來建立一套 bi-gram 語言模型以達到較佳的辨識效益。

MAP 估算法可用來當作調適模型的一種方法，其定義是在擁有一個已知模型參數分布(model parameter distribution) X 以及有限的觀察資料(observation) W 的情況下：

$$\begin{aligned} X_{\text{MAP}} &= \arg\max_X P(X|W) \\ &= \arg\max_X P(W|X)P(X) \end{aligned} \tag{3-11}$$

其中 $P(W|X)$ 為觀察資料 W 的概似度(likelihood)。此方法的主要概念是當觀察資料數量越多時，則越不相信原本的模型參數分布，簡而言之，就是改變已知的模型參數分布使其與觀察資料的參數分布相符合。

3.2.3 語言模型效能分析

評估語言模型通常是以計算其混淆度(perplexity, *PP*)來判斷。混淆度是根據消息理論(information theory)而得，如下式：

$$H = -\frac{1}{m}\log P(W = w_1, w_2, \ldots, w_m) \tag{3-12}$$

上式為一個詞串 $W = w_1, w_2, \ldots, w_m$，對於每個新詞提供的平均資訊量(entropy)，經過適當的化簡而得。而混淆度可直接使用 (3-12) 式進一步定義為：

$$PP = \exp(H) \tag{3-13}$$

若 $P(W = w_1, w_2, \ldots, w_m) = \prod_{i=1}^{m} P(w_i \mid w_1, w_2, \ldots, w_{i-1})$ 則可發現，混淆度就是 $P(w_i \mid w_1, w_2, \ldots, w_{i-1})$ 的幾何平均數的倒數。因此混淆度可以解讀為語言模型估測一個歷史詞串後面，平均可能的可接詞數；混淆度越高，表示一個歷史詞串後接詞有較多的選擇，辨認時就越難找到確切的答案；反之，則較易找到正確答案。

利用(3-13)式來評估本研究實驗之各個語言模型的效能，如表 3.9 所示。由表中可觀察到在 Stage1 時我們在語言模型中估算了 Particle 及 Paralinguistic 的機率後，混淆度確實能有效下降；在 Stage3 時我們有效地調適了一般詞的機率後也能使得混淆度大幅下降。此效能評估也將反映在辨識率上，本研究將在第五章予以觀察及討論。

表 3.9：語言模型混淆度評估

語言模型	混淆度(*PP*)
General-LM	748.7
Stage1-LM	598.15
Stage2-LM	547.46
Stage3-LM	394.8

第四章 自發性語音韻律模型

本研究採用【11】所提出之方法做修改，重新定義自發性語音中的特殊韻律現象，並利用語音信號中之聲學參數及文字轉寫(transcription)上的語言參數，以非監督式(unsupervised)的方法訓練出自發性語音韻律模型(PLM)。在語音辨認方面，利用韻律模型對第一級辨認出來的 Top-N 詞串重新計分並產生最佳之詞串辨認結果。本章將在 4.1 節介紹韻律模型之設計；4.2 節將會介紹如何建立韻律模型；4.3 節將針對訓練之韻律模型做簡單分析；4.4 節將介紹以韻律模型協助語音辨認之方法。

4.1 韻律模型設計

在本節中將簡單說明本論文所使用的中文語音韻律階層式架構以及韻律模型之設計，在此定義四個子模型來描述語音信號中聲學參數以及文字上語言參數與韻律標記之關係。

4.1.1 中文語音韻律階層式架構

中文的韻律結構是具有階層性的架構(hierarchy structure)，由底層至上層主要由音節(Syllable, SYL)、韻律詞(Prosody Word, PW)、韻律短句(Prosody Phrase, PPh)以及句調(intonation phrase)所構成。此外，鄭秋豫博士【15】提出將連續的 PPh 組合成一個呼吸群(Breathe Group, BG)來代表大範圍且具有基頻及音節長度高度變化之語句，藉此表示韻律更上層之貢獻，同樣地定義由連續的 BG 所組成的韻律群(Prosody Group, PG)，值得注意的是鄭秋豫博士在流利語音的韻律架構當中，定義 PPh 間存在著某些可插入的篇章提語(Discourse Marker, DM)或韻律填充(Prosody Filler, PF)，以連接鄰近的 PPh，如圖 4.1 所示：

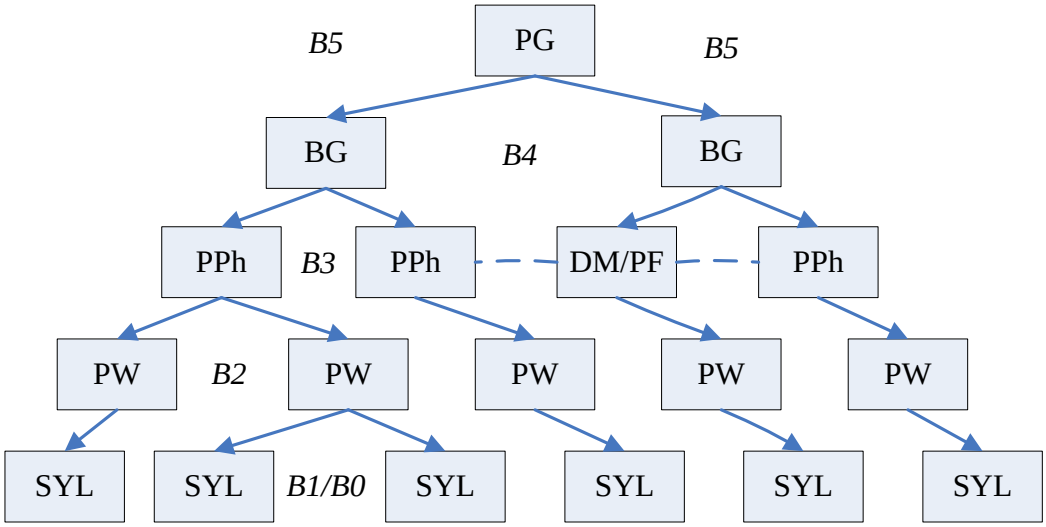


圖 4.1：中文語音韻律之階層式架構概念

在自發性語音韻律結構的研究上【11】提出一個特殊韻律現象(Particular Prosody Phenomena, Par)的單元，藉此隔離正常語流之語句。特殊韻律現象中包含：韻律特性和基本音節差距較大之語助詞或感歎詞[6]、無法或難以辨識的語音、受相鄰音節同化之音節以及發生嚴重拉長之音節[7]，這些嚴重拉長的音節數量雖小但長度較其他基本音節大，會嚴重影響模型之統計特性，如圖 4.2 所示，在本研究中將基於【11】所提出的中文自發性韻律階層式架構為基礎作修改並設計韻律模型。

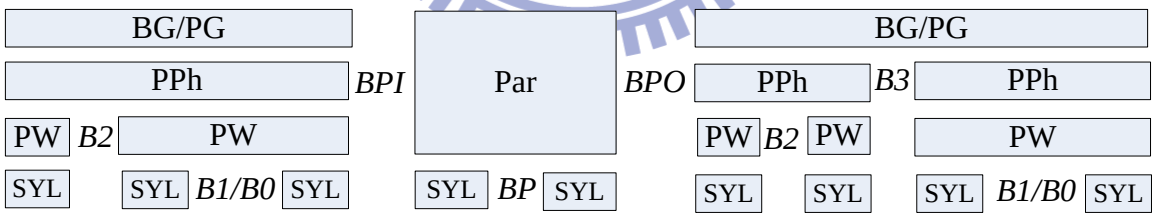


圖 4.2：中文自發性韻律階層式架構

基於圖 4.2 的架構下，本研究將所有音節分為對應正常語流之基本音節(base syllable)以及對應特殊韻律單元之特殊音節(particular syllable)。此外在本研究所使用的四層韻律結構

[6]本研究將第二章所定義之感歎詞或語助詞中，「ZHE GE」、「NA GE」、「NE GE」、以及「SHEN ME」視為兩個音節，其餘皆視為一個音節。

[7]嚴重音節拉長決定方式為一經驗法，本研究定義音節長度前 99%的資料為嚴重音節拉長。

中，主要由十種停頓標記(break type)來區分韻律結構中每一層的韻律單元，對應法則如表 4.1 所示。

表 4.1：韻律結構之停頓標記

韻律結構	停頓標記	意義
韻律群(PG) 或呼吸群(BG)	*B3*	長停頓
	B4	長停頓且含有明顯的基頻跳躍
韻律詞(PW)	*B2-1*	相鄰兩音節具有明顯的基頻跳躍
	B2-2	短停頓
	B2-3	前一音節發生音節拉長
音節(SYL)	*B0*	音節邊界相鄰兩音節是緊密連接(tightly coupling)
	B1	音節邊界相鄰兩音節是普通連接(normal coupling)
特殊韻律現象 (Par)	*BPI*	後一音節為特殊音節
	BP	相鄰之兩音節皆為特殊音節
	BPO	前一音節為特殊音節

值得注意的是，【11】將所有 particle 皆定義為特殊韻律現象，而本研究實地觀察後發現自發性語音中其實有許多 particle 存在於順暢語流中，通常其韻律變化就如同基本音節，因此本研究將把屬於流暢語流中的 particle 視為基本音節作訓練。

首先，本研究先利用【11】訓練之韻律模型計算 *B2-2* 之門檻值(threshold)，檢查所有不為拖長音且後一音節為基本音節的 particle 與前一音節或後一音節間的停頓長度(pause duration)，若停頓長度小於此門檻值則將此 particle 視為流暢語流中的基本音節，並將它的聲調(tone)退化成近似之 411 音節的聲調；若停頓長度大於此門檻值則將此 particle 視為特殊音節，給予一個特別的聲調並當成特殊韻律現象來處理。其中值得注意的是我們將 NE-GE、ZHE-GE 及 SHEN-ME 視為可合併成一個單元的候選 particle 詞組，檢查此類 particle 詞組時除了規定其與前一音節或後一音節間的停頓長度要小於 *B2-2* 門檻值外，其本身兩個 particle 相接的停頓長度也需小於 *B2-2* 門檻值才可被轉換成基本音節。篩選結果如表 4.2 所示：總共有 8554 個 particle，其中有 4462 將被視為流暢語流中的基本音節；有 4092 個 particle 被視為特殊韻律現象之音節。

表 4.2：歸類為基本音節之 particle 個數

particle	個數	particle	個數	particle	個數
A	1024	LA	203	HEN	40
NA	482	MA	156	MHMHM	35
O	425	BA	149	HEIN	25
GE	392	HO	132	WA	22
NE	363	ME	70	HAN	19
E	250	SHEN	70	NO	18
MHM	224	YA	63	AI	15
EI	214	ZHE	61	YOU	10

4.1.2 韻律參數之介紹

接著定義本研究中使用之韻律聲學參數 **A**，以及文字上的語言學參數 **L**，如表 4.3 所示。本研究考慮的韻律聲學參數包含兩大類如圖 4.3 範例：第一類是與韻律狀態有緊密關係的音節韻律參數 **X** (syllable prosodic feature)主要有：音節基頻軌跡 **sp** (syllable pitch contour)、音節長度 *sd* (syllable duration)以及音節能量 *se* (syllable energy level)；第二類是與停頓標記有緊密相關的特徵參數，又細分為兩類分別是音節間韻律參數 **Y** (inter-syllable prosodic feature)以及相鄰兩音節差異之韻律參數 **Z** (differencial prosodic feature)，音節間韻律參數有：音節間停頓長度 *pd* (pause duration)以及音節間能量低點 *ed* (energy-dip level)；相鄰兩音節差異之韻律參數有：相鄰兩音節之正規化基頻跳躍值 *pj* (normalized pitch jump)以及相鄰兩音節之正規化音節延長因子 *dl* (normalized duration lengthening factor)。而文字上的語言學特徵參數 **L** 主要包含語言學中音節以及詞層次上的參數。音節層次上的參數主要包含了音節聲調序列 **t** (tone sequence)、基本音節型態 *s* (base syllable type)或韻母型態 *f* (final type)；其它語言參數 **l**，主要包含音節邊界種類[8](syllable juncture type)、詞長以及詞性。

[8] 音節邊界主要分為詞內音節邊界(intra-word syllable juncture)以及詞間音節邊界(inter-word syllable juncture)

表 4.3：韻律標記、聲學參數以及語言參數之數學符號

T: prosodic tag	**B**: break type ={*B0*, *B1*, *B2-1*, *B2-2*, *B2-3*, *B3*, *B4*, *BPI*, *BP*, *BPO*}	
	PS: prosodic state	**p**: pitch prosodic state **q**: duration prosodic state **r**: energy prosodic state
A: prosodic feature	**X**: syllable prosodic feature	**sp**: syllable pitch contour **sd**: syllable duration **se**: syllable energy level
	Y: inter-syllabic prosodic feature	**pd**: pause duration **ed**: energy-dip level
	Z: differential prosodic features	**pj**: normalized pitch jump **df**: normalized duration lengthening factor
L: linguistic feature	**l**: reduced linguistic feature set **t**: syllable tone sequence **s**: base-syllable type **f**: final type	

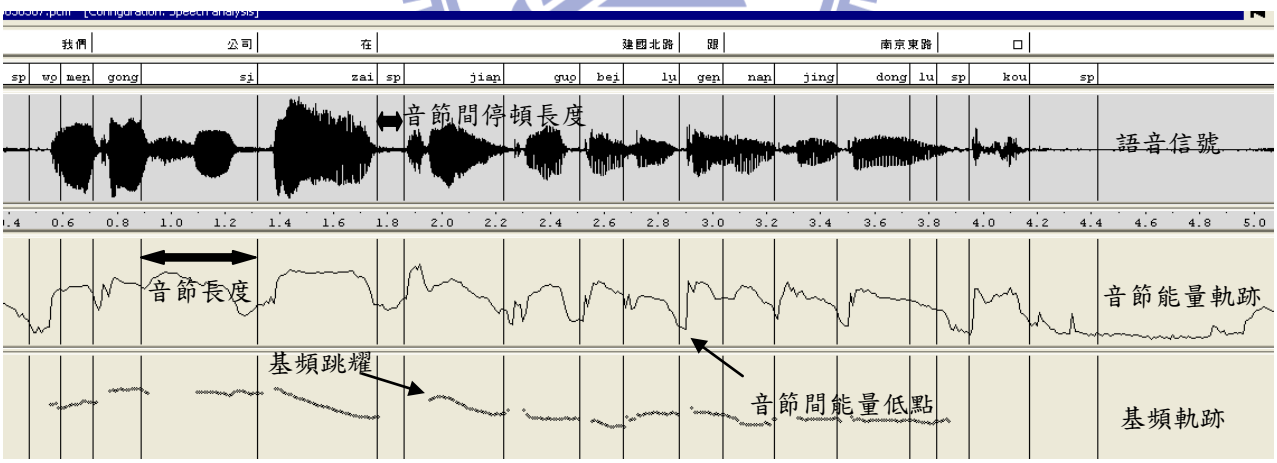


圖 4.3：音節及音節間韻律參數

4.1.3 模型設計

本研究利用語音信號上的聲學參數 **A**，以及文字上語言學的參數 **L**，以模型為基礎(model-based)估計此語句中最有可能的韻律標記序列 $\mathbf{T}^*$，因此可將其看作一個數學估計的問題，其數學式如下：

$$\mathbf{T}^* = \arg\max_{\mathbf{T}} P(\mathbf{T}|\mathbf{A},\mathbf{L}) = \arg\max_{\mathbf{T}} P(\mathbf{T},\mathbf{A}|\mathbf{L}) \tag{4-1}$$

在此定義兩種韻律標記，第一種為 4.1.1 節中所定義之音節停頓標記序列 **B**；第二種則是韻律狀態(prosody state)序列 **PS**，它是經由扣除音節及其相鄰音節對韻律之影響並量化後所得到，以描述韻律上層之變化狀況及其對韻律參數之貢獻值，在本研究中，將所有基本音節量化為 16 個韻律狀態，並且將特殊音節另外分出 4 個韻律狀態量化之。根據 4.1.2 所介紹的韻律參數，我們可以將 4-1 式改寫為：

$$P(\mathbf{T},\mathbf{A}|\mathbf{L}) = P(\mathbf{A}|\mathbf{T},\mathbf{L})P(\mathbf{T}|\mathbf{L}) = P(\mathbf{X},\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{PS},\mathbf{L})P(\mathbf{B},\mathbf{PS}|\mathbf{L}) \tag{4-2}$$

其中 $P(\mathbf{X},\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{PS},\mathbf{L})$ 為廣義韻律參數模型(general prosodic feature model)，其物理意義為下層所得到的韻律聲學參數 **X**、**Y**、**Z**，是由上層的韻律標記 **B**、**PS** 以及語言參數 **L** 所控制。而 $P(\mathbf{B},\mathbf{PS}|\mathbf{L})$ 為廣義韻律語言模型(general prosody-syntax model)，它主要在描述韻律標記 **B**、**PS** 和語言參數 **L** 之間的關係。

由於定義之停頓標記 **B** 已帶有相鄰兩音節間之韻律資訊，因此在已知停頓標記的狀況下，可以假設音節韻律參數 **X** 與音節間韻律參數 **Y** 及相鄰兩音節差異之韻律參數 **Z** 互相獨立，因此可將廣義韻律聲學模型 $P(\mathbf{X},\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{PS},\mathbf{L})$ 一分為二，其數學式如下：

$$P(\mathbf{X},\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{PS},\mathbf{L}) = P(\mathbf{X}|\mathbf{B},\mathbf{PS},\mathbf{L})P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{PS},\mathbf{L}) \tag{4-3}$$

其中 $P(\mathbf{X}|\mathbf{B},\mathbf{PS},\mathbf{L})$ 為音節韻律模型(syllable prosodic model)，其物理意義為音節中的基頻軌跡、音節長度及音節能量，是由上層的韻律標記 **B**、**PS** 以及語言參數 **L** 所控制，其中語言參數又以音節的聲調序列 **t** 之影響最為嚴重。而 $P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{PS},\mathbf{L})$ 為停頓標記聲學模型

(break-acoustic model)，它描述不同韻律標記 **B**、**PS** 以及語言參數 **L** 的狀況之下，音節間韻律參數 **Y** 及相鄰兩音節差異之韻律參數 **Z** 分布的情況。同樣地，經由假設上層的韻律標記 **B**、**PS** 與音節聲調 **t** 互相獨立，我們也可將廣義韻律語言模型 $P(\mathbf{B},\mathbf{PS}|\mathbf{L})$ 一分為二，數學式如下：

$$P(\mathbf{B},\mathbf{PS}|\mathbf{L}) \approx P(\mathbf{B},\mathbf{PS}|\mathbf{l}) = P(\mathbf{PS}|\mathbf{B},\mathbf{l})P(\mathbf{B}|\mathbf{l}) \approx P(\mathbf{PS}|\mathbf{B})P(\mathbf{B}|\mathbf{l}) \tag{4-4}$$

其中 $P(\mathbf{PS}|\mathbf{B})$ 為韻律狀態轉移模型(prosodic state model)，描述在已知停頓標記的狀況之下，韻律狀態轉移之機率。 $P(\mathbf{B}|\mathbf{l})$ 則為停頓標記語言模型(break-syntax model)，主要描述停頓標記 **B** 和語言參數 **l** 之間的關係。因此本研究中設計了四個子模型分別為：音節韻律模型、停頓標記聲學模型、韻律狀態轉移模型以及停頓標記語言模型，來描述韻律狀態、停頓標記與聲學和語言學參數之間的關係。

本研究之音節韻律模型是假設音節基頻軌跡、音節長度以及音節能量可拆解成各個影響因子(affecting factor)之貢獻，這些影響因子包含：音節之韻律狀態 p_n、包含音節左右邊界停頓標記 B_n、B_{n-1} 及鄰近音節聲調 t_{n+1}、t_{n-1} 影響之音節聲調 t_n (i.e.音節連音現象之影響)以及音節基本型態 s_n 或音節韻母型態 f_n，不同的音節韻律模型將視影響程度，對應至不同影響因子之組合，將在之後做更詳細之介紹。因此可將音節韻律模型拆解成三個模型分別為音節基頻軌跡模型、音節長度模型以及音節能量模型，其數學式如下：

$$\begin{aligned} & p(\mathbf{X}|\mathbf{B},\mathbf{PS},\mathbf{L}) \approx p(\mathbf{sp}|\mathbf{B},\mathbf{p},\mathbf{t})\,p(\mathbf{sd}|\mathbf{B},\mathbf{q},\mathbf{t},\mathbf{s})\,p(\mathbf{se}|\mathbf{B},\mathbf{r},\mathbf{t},\mathbf{f}) \\ & \approx \prod_{n=1}^{N} p(\mathbf{sp}_n|B_{n-1}^{n},p_n,t_{n-1}^{n+1}) \prod_{n=1}^{N} p(sd_n|B_{n-1}^{n},q_n,t_{n-1}^{n+1},s_n) \prod_{n=1}^{N} p(se_n|B_{n-1}^{n},r_n,t_{n-1}^{n+1},f_n) \end{aligned} \tag{4-5}$$

其中

$$\mathbf{sp}_n = \begin{cases} \mathbf{sp}_n^r + \boldsymbol{\beta}_{t_{n-1}^{n+1},B_{n-1}^{n}} + \boldsymbol{\beta}_{p_n} + \boldsymbol{\mu} & \text{, if } n\text{th syllable is base syllable} \\ \mathbf{sp}_n^r + \boldsymbol{\beta}_{pr_n} + \boldsymbol{\beta}_{p'_n} + \boldsymbol{\mu}' & \text{, if } n\text{th syllable is particular syllable} \end{cases} \quad \text{for } 1 \le n \le N \tag{4-6}$$

為第 n 個音節之音節基頻軌跡，$\mathbf{sp}_n^r$ 為 $\mathbf{sp}_n$ 正規化(normalization)後之基頻殘存值(residual)；$\boldsymbol{\beta}_x$ 則為某一影響因子 x 之影響型態(Affecting Pattern, AP)；$\boldsymbol{\mu}$ 為所有 AP 之總體平均值(global

mean)。同樣地，第 n 個音節之音節長度及音節能量可表示如下：

$$sd_n = \begin{cases} sd_n^r + \gamma_{t_{n-1}^{n+1}, B_{n-1}^n} + \gamma_{q_n} + \gamma_{s_n} + \mu_d & \text{, if } n\text{th syllable is base syllable} \\ sd_n^r + \gamma_{pr_n} + \gamma_{q_n'} + \mu_d' & \text{, if } n\text{th syllable is particular syllable} \end{cases} \quad \text{for } 1 \le n \le N \quad (4\text{-}7)$$

$$se_n = \begin{cases} se_n^r + \alpha_{t_{n-1}^{n+1}, B_{n-1}^n} + \alpha_{r_n} + \alpha_{f_n} + \mu_e & \text{, if } n\text{th syllable is base syllable} \\ se_n^r + \alpha_{pr_n} + \alpha_{r_n'} + \mu_e' & \text{, if } n\text{th syllable is particular syllable} \end{cases} \quad \text{for } 1 \le n \le N \quad (4\text{-}8)$$

如同以上數學式所描述，本研究將特殊音節另行賦予其特殊音節型態之影響因子 pr_n 以及特殊音節之韻律狀態 p_n'。接著經由假設正規化後之殘存值為一平均值為零之高斯分佈，可將音節基頻軌跡模型、音節長度模型以及音節能量模型改寫為：

$$P(\mathbf{sp}_n | p_n, B_{n-1}^n, t_{n-1}^{n+1}) = \begin{cases} N(\mathbf{sp}_n; \boldsymbol{\beta}_{t_{n-1}^{n+1}, B_{n-1}^n} + \boldsymbol{\beta}_{p_n} + \boldsymbol{\mu}, \mathbf{R}) \\ \qquad \text{, if } n\text{th syllable is base syllable} \\ N(\mathbf{sp}_n; \boldsymbol{\beta}_{pr_n} + \boldsymbol{\beta}_{p_n'} + \boldsymbol{\mu}', \mathbf{R}') \\ \qquad \text{, if } n\text{th syllable is particular syllable} \end{cases} \quad \text{for } 1 \le n \le N \quad (4\text{-}9)$$

$$P(sd_n | q_n, B_{n-1}^n, t_{n-1}^{n+1}, s_n) = \begin{cases} N(sd_n; \gamma_{t_{n-1}^{n+1}, B_{n-1}^n} + \gamma_{q_n} + \gamma_{s_n} + \mu_d, R_d) \\ \qquad \text{, if } n\text{th syllable is base syllable} \\ N(sd_n; \gamma_{pr_n} + \gamma_{q_n'} + \mu_d', R_d') \\ \qquad \text{, if } n\text{th syllable is particular syllable} \end{cases} \quad \text{for } 1 \le n \le N \quad (4\text{-}10)$$

$$P(se_n | r_n, B_{n-1}^n, t_{n-1}^{n+1}, f_n) = \begin{cases} N(se_n; \alpha_{t_{n-1}^{n+1}, B_{n-1}^n} + \alpha_{r_n} + \gamma_{f_n} + \mu_e, R_e) \\ \qquad \text{, if } n\text{th syllable is base syllable} \\ N(se_n; \alpha_{pr_n} + \alpha_{r_n'} + \mu_e', R_e') \\ \qquad \text{, if } n\text{th syllable is particular syllable} \end{cases} \quad \text{for } 1 \le n \le N \quad (4\text{-}11)$$

值得注意的是，本研究將利用決策樹(decision tree)以資料驅動(data-driven)的方式，自動分類 $\boldsymbol{\beta}_{t_{n-1}^{n+1}, B_{n-1}^n}$、$\gamma_{t_{n-1}^{n+1}, B_{n-1}^n}$ 以及 $\alpha_{t_{n-1}^{n+1}, B_{n-1}^n}$ 之 AP，藉此同時描述音節聲調及前後音節連音現象(coarticulation)對韻律參數之影響。

接著經由假設音節間韻律參數 **Y** 及相鄰兩音節差異之韻律參數 **Z** 與韻律狀態 **PS** 及聲調

序列 **t** 之間互相獨立，可將停頓標記聲學模型之數學式改寫如下：

$$P(\mathbf{Y,Z|B,PS,L}) \approx P(\mathbf{Y,Z|B,l}) \approx \prod_{n=1}^{N-1} P(pd_n, ed_n, pj_n, dl_n | B_n, \mathbf{l}_n) \tag{4-12}$$

為了數學上容易處理，我們將 $P(pd_n, ed_n, pj_n, dl_n | B_n, \mathbf{l}_n)$ 簡化成停頓長度之伽瑪分布與其他參數之高斯分布相乘，其數學式如下：

$$\begin{aligned} P(pd_n, ed_n, pj_n, dl_n | B_n, \mathbf{l}_n) = & g(pd_n; \alpha_{B_n,\mathbf{l}_n}, \beta_{B_n,\mathbf{l}_n}) N(ed_n; \mu_{B_n,\mathbf{l}_n}, \sigma^2_{B_n,\mathbf{l}_n}) \\ & \cdot N(pj_n; \mu_{B_n,\mathbf{l}_n}, \sigma^2_{B_n,\mathbf{l}_n}) N(dl_n; \mu_{B_n,\mathbf{l}_n}, \sigma^2_{B_n,\mathbf{l}_n}) \end{aligned} \tag{4-13}$$

在每種停頓標記狀況下，各個參數之機率分布，將以最大概似度增益為分裂準則(splitting criterion of maximum likelihood gain)之決策樹實現，其問題集(question set)將由詢問語言參數 **l** 之問題所產生。

此外韻律狀態轉移模型將利用馬可夫模型(Markov Model)來實現，其數學式如下：

$$P(\mathbf{p|B}) \approx P(p_1) \left[\prod_{n=2}^{N} P(p_n | p_{n-1}, B_{n-1}) \right] \tag{4-14}$$

其中 $P(p_1)$ 為各語句中第一個音節韻律狀態之機率；$P(p_n | p_{n-1}, B_{n-1})$ 為已知音節前邊界之停頓標記 B_{n-1} 之下，前一音節韻律狀態 p_{n-1} 轉移到現在音節韻律狀態 p_n 之機率。

最後我們簡化停頓標記語言模型為：

$$P(\mathbf{B|l}) = \prod_{n=1}^{N-1} P(B_n | \mathbf{l}_n) \tag{4-15}$$

並且以最大概似度增益為分裂準則之決策樹來實現它，每一節點中將產生每一種停頓標記之機率，其問題集將由詢問語言參數 **l** 之問題所產生。

4.2 韻律模型之建立

本研究在訓練韻律模型部分主要分為兩個步驟：初始化(initialization)以及重覆疊代(iteration)，分別會在 4.2.1 節以及 4.2.2 節作介紹。而在訓練韻律模型之四個韻律子模型時，將基於最大概似度準則(maximum likelihood criterion)採用逐項最佳化程序(sequencial optimization procedure)來訓練並更新模型參數，其標的函數(objective function)如下：.

$$Q=\left(\prod_{n=1}^{N}P(\mathbf{sp}_n\mid p_n,B_{n-1}^{n},t_{n-1}^{n+1})\,p(sd_n\mid q_n,B_{n-1}^{n},t_{n-1}^{n+1},s_n)\,p(se_n\mid r_n,B_{n-1}^{n},t_{n-1}^{n+1},f_n)\right)$$
$$\left(P(p_1)P(q_1)P(r_1)\prod_{n=2}^{N}P(p_n\mid p_{n-1},B_{n-1})P(q_n\mid q_{n-1},B_{n-1})P(r_n\mid r_{n-1},B_{n-1})\right) \qquad (4\text{-}16)$$
$$\left(\prod_{n=1}^{N-1}\left(p(pd_n,ed_n,pj_n,dl_n\mid B_n,\mathbf{l}_n)P(B_n\mid\mathbf{l}_n)\right)\right)$$

4.2.1 初始化

為了幫助模型的訓練能快速收斂，我們首先將建立一個初始韻律模型，其流程圖如圖 4.4 所示，而在訓練過程中將以音節為單元。

首先判斷其音節屬性，分為基本音節以及特殊音節，並依據不同屬性音節計算其各個影響因子之影響型態；接著使用音節停頓長度、音節間能量低點、正規化基頻跳躍值以及正規化音節延長因子，並利用決策樹的方式，對所有音節邊界處標記初始的停頓型態，如圖 4.5 所示。而為了得到一個具有物理意義且較合理的決策樹門檻值，在此我們使用附錄二之定義方法，以一個系統化的統計方式得到這些決策樹之門檻值(Th1~Th7)。

有了初始停頓標記之後，扣掉各個影響因子 AP，再以向量量化的方式計算基頻、音節長度以及能量之每一個韻律狀態的代表值 $\boldsymbol{\beta}_{p_n}$、γ_{q_n} 以及 α_{r_n}，並針對每一個音節標記其擁有之韻律狀態。完成停頓標記及韻律狀態後，將這些標記的結果以相對計數率(relative count)的方

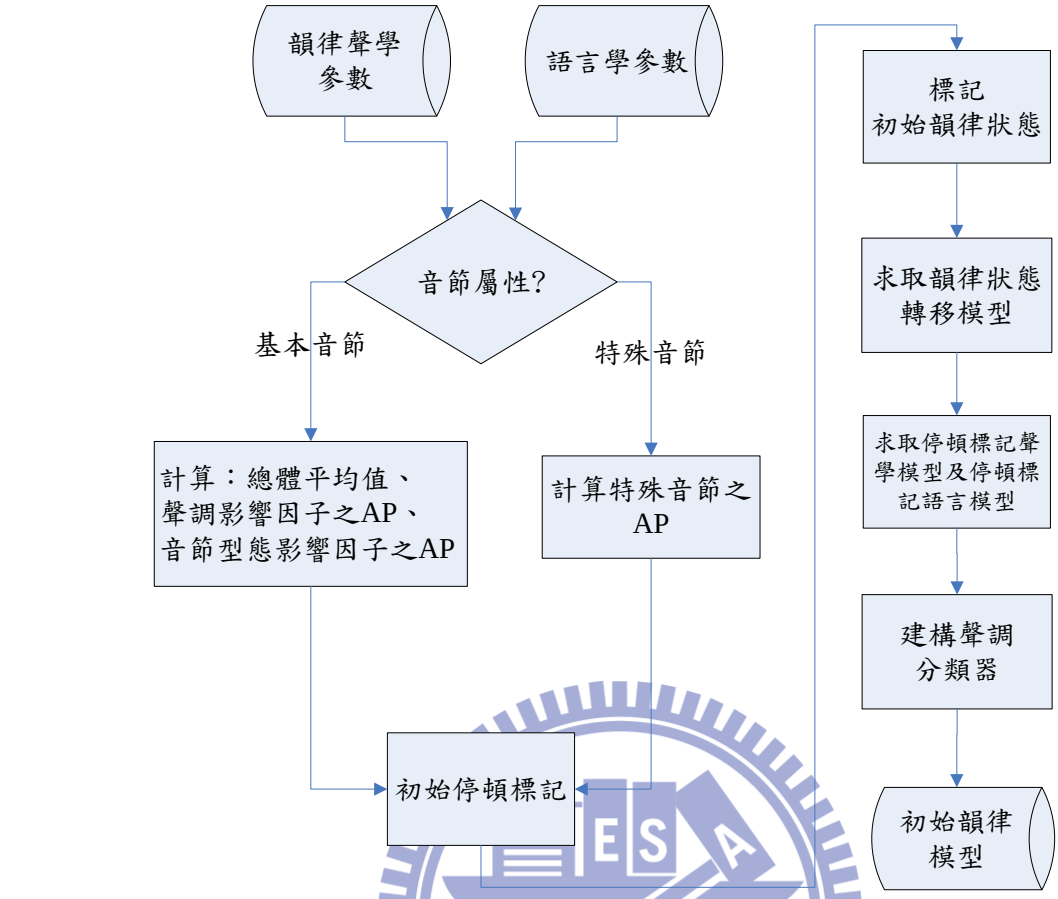


圖 4.4：初始韻律模型建立流程圖

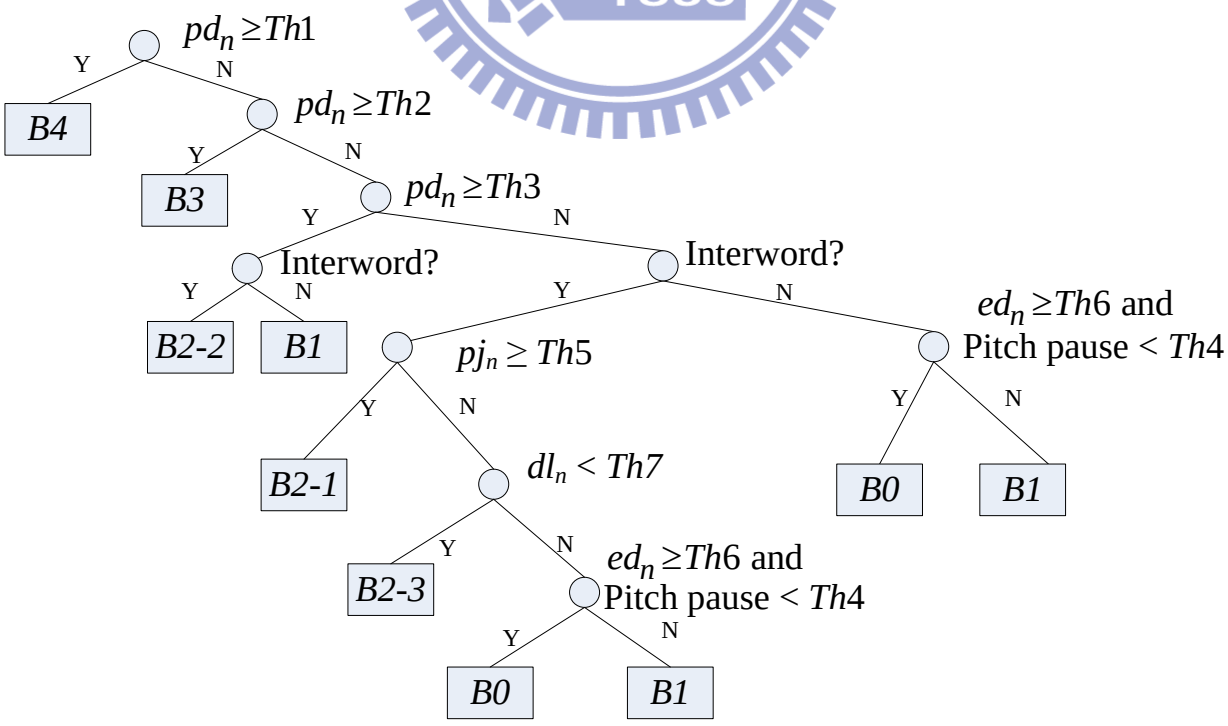


圖 4.5：分類停頓標記之決策樹示意圖

式計算每種停頓標記的情況之下韻律狀態轉移之機率 $P(\mathbf{P}|\mathbf{B})$。

在停頓標記聲學模型以及停頓標記語言模型之決策樹方面本研究將以CART(Classification And Regression Trees)演算法實現，其問題集分別參見附錄三、四。最後，在利用音節聲調之影響因子求得較為準確之韻律狀態貢獻值後，利用扣除總體平均值以及韻律狀態貢獻之基頻殘存值，以 CART 演算法及最小歐幾里得距離(minimum Euclidean distance)為分裂準則區分包含連音現象聲調之 AP，實現一個包含連音現象之聲調 AP 分類器(classifier)，供重覆疊代時使用。

4.2.2 重覆疊代

有好的初始化模型後，即可進行重覆疊代訓練模型直至收斂為止，流程圖如圖 4.6 所示。首先，使用聲調分類器計算聲調影響因子對基頻、音節長度以及能量產生之 AP 以及更新所有音節型態影響因子之 AP；接著利用維特比搜尋演算法(Viterbi search algorithm)重新標記韻律狀態及重新標記停頓標記，使(4-16)式之標的函數 Q 之概似度(likelihood)最大化；最後再以 CART 演算法更新停頓標記聲學模型以及停頓標記語言模型，重覆疊代訓練圖 4.6 之流程直至收斂為止。其中值得注意的是每次更新模型時，其共變異矩陣也將重新計算。

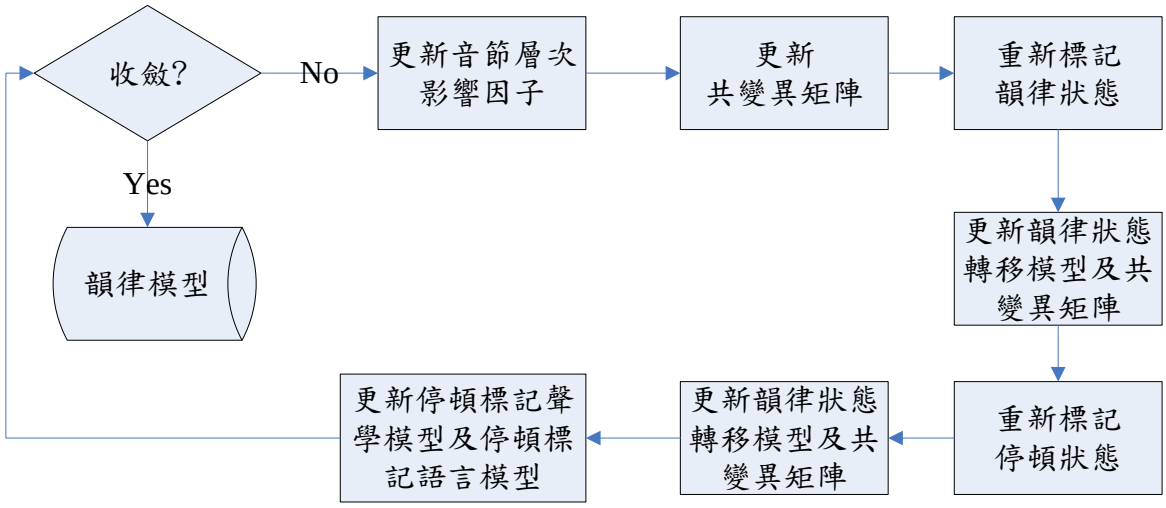


圖 4.6：韻律模型重覆疊代流程圖

4.3 韻律模型之分析

本研究在訓練韻律模型時採逐項最佳化程序重覆疊代共 55 次收斂。本節中將會簡單分析訓練之音節韻律模型、停頓標記聲學模型以及韻律標記結果之觀察結果。

4.3.1 音節韻律模型

音節韻律模型中包含音節基頻軌跡模型 $p(\mathbf{sp}|\mathbf{B},\mathbf{p},\mathbf{t})$、音節長度模型 $p(\mathbf{sd}|\mathbf{B},\mathbf{q},\mathbf{t},\mathbf{s})$ 以及音節能量模型 $p(\mathbf{se}|\mathbf{B},\mathbf{r},\mathbf{t},\mathbf{f})$ 三個子模型，主要用來描述音節中各個影響因子對音節基頻軌跡、長度以及能量之貢獻。表 4.4 及表 4.5 中列出在基本音節及特殊音節中，扣除不同影響因子之下，各韻律參數之總殘餘誤差值(Total Residual Error, TRE)，即為扣掉各種組合 AP 後殘存值之變異數與原始資料之變異數的比值，藉此觀察各個影響因子對音節中韻律參數變化之貢獻大小。

表 4.4：基本音節中，不同組合之 AP 下音節韻律模型參數之 TRE

Pitch		Duration		Energy	
APs	TRE	APs	TRE	APs	TRE
+Tone with Coarticulation	91.21%	+Tone with Coarticulation	95.88%	+Tone with Coarticulation	95.98%
		+Base Syllable	86.51%	+Final	89.16%
+Prsodic State	14.18%	+Prsodic State	1.92%	+Prsodic State	2.76%

表 4.5：特殊音節中，不同組合之 AP 下音節韻律模型參數之 TRE

Pitch		Duration		Energy	
APs	TRE	APs	TRE	APs	TRE
+Particular Syllable Class	86.23%	+Particular Syllable Class	94.03%	+Particular Syllable Class	88.41%
+Prsodic State	40.21%	+Prsodic State	13.82%	+Prsodic State	38.20%

基本音節中基頻軌跡、長度及能量的變化將受到包含連音現象之聲調所影響，其中音節長度又會受到音節型態的影響，而音節能量則是受韻母形態所影響，由表 4.4 可以得知在扣除這些影響因子後，總殘餘誤差值確實是有降低，且在加入韻律狀態影響因子後都會有較大的變化貢獻。

接著觀察考慮連音現象之聲調影響因子，如圖 4.7 所示。由此圖 4.7 與朗讀式語音之聲調【16】相比，我們發現在自發性語音中聲調之 AP 之動態範圍(dynamic range)及其基頻軌跡之曲度都較小，在此我們推測其原因可能為自發性語音之語速較快，造成音節會產生緊密連接的狀況較多，因此音節基頻軌跡會受鄰近音節的嚴重干擾，以及自發性語音中發音經常不完全而使聲調之基頻軌跡相當凌亂；另外，由於本語料庫尚有切割位置不準確之處，此狀況將影響音節中基頻軌跡之求取。

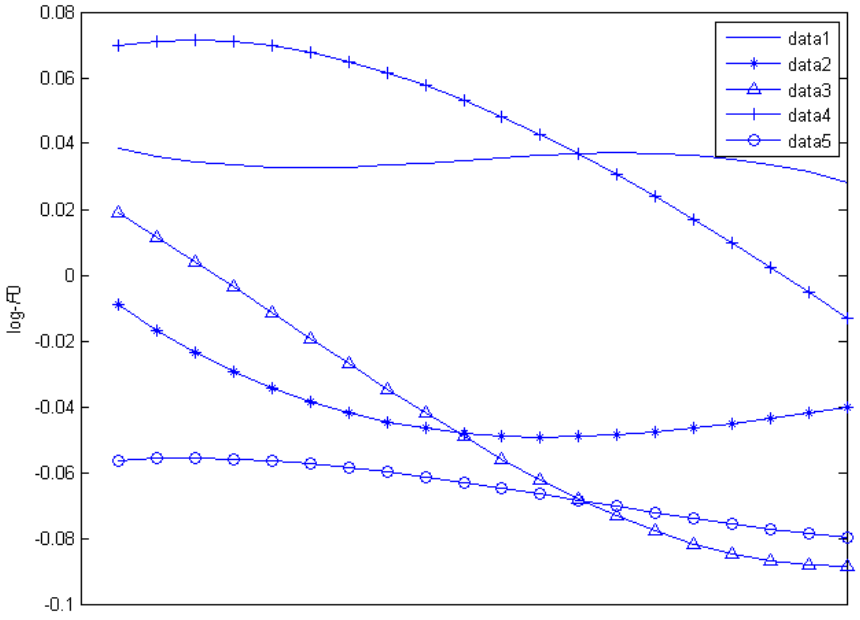


圖 4.7：五個中文聲調之 AP

4.3.2 停頓標記聲學模型

停頓標記聲學模型描述了音節邊界之停頓標記**B**、語言參數**l**與音節間韻律參數 **Y** 及相鄰兩音節差異之韻律參數 **Z** 之間的關係，主要可分為四個子模型 $g(pd_n;\alpha_{B_n,\mathbf{l}_n},\beta_{B_n,\mathbf{l}_n})$、

$N(ed_n;\mu_{B_n,\mathbf{l}_n},\sigma^2_{B_n,\mathbf{l}_n})$、$N(pj_n;\mu_{B_n,\mathbf{l}_n},\sigma^2_{B_n,\mathbf{l}_n})$ 以及 $N(dl_n;\mu_{B_n,\mathbf{l}_n},\sigma^2_{B_n,\mathbf{l}_n})$。圖 4.8 顯示不考慮語言參數**l**下，*B0~B4* 停頓標記決策樹根節點中各個參數之分布圖。觀察圖 4.8(a)，*B0* 與 *B1* 的停頓長度相當短，此表示 *B0* 與 *B1* 通常出現在相鄰兩音節為緊密連接時，而 *B3*、*B4* 會有較長的停頓長度，因為其主要用來隔離上層韻律單元 PG/BG 以及 PPh 之停頓標記。在觀察圖 4.8(b)音節間能量低點分布圖亦是如此，*B0* 因為表示相鄰兩音節為緊密連接而有較大的能量低點，*B3/B4* 則因為相鄰兩音節停頓大而有較低的音節間能量低點。接著觀察圖 4.8(c)與(d)可發現 *B2-1* 擁有較大之基頻跳躍值及 *B2-3* 有較大之正規化音節延長因子，由此可知人類不但使用明顯的停頓來表示韻律詞的邊界，另外也使用基頻的跳躍以及音節的拉長來表示之。

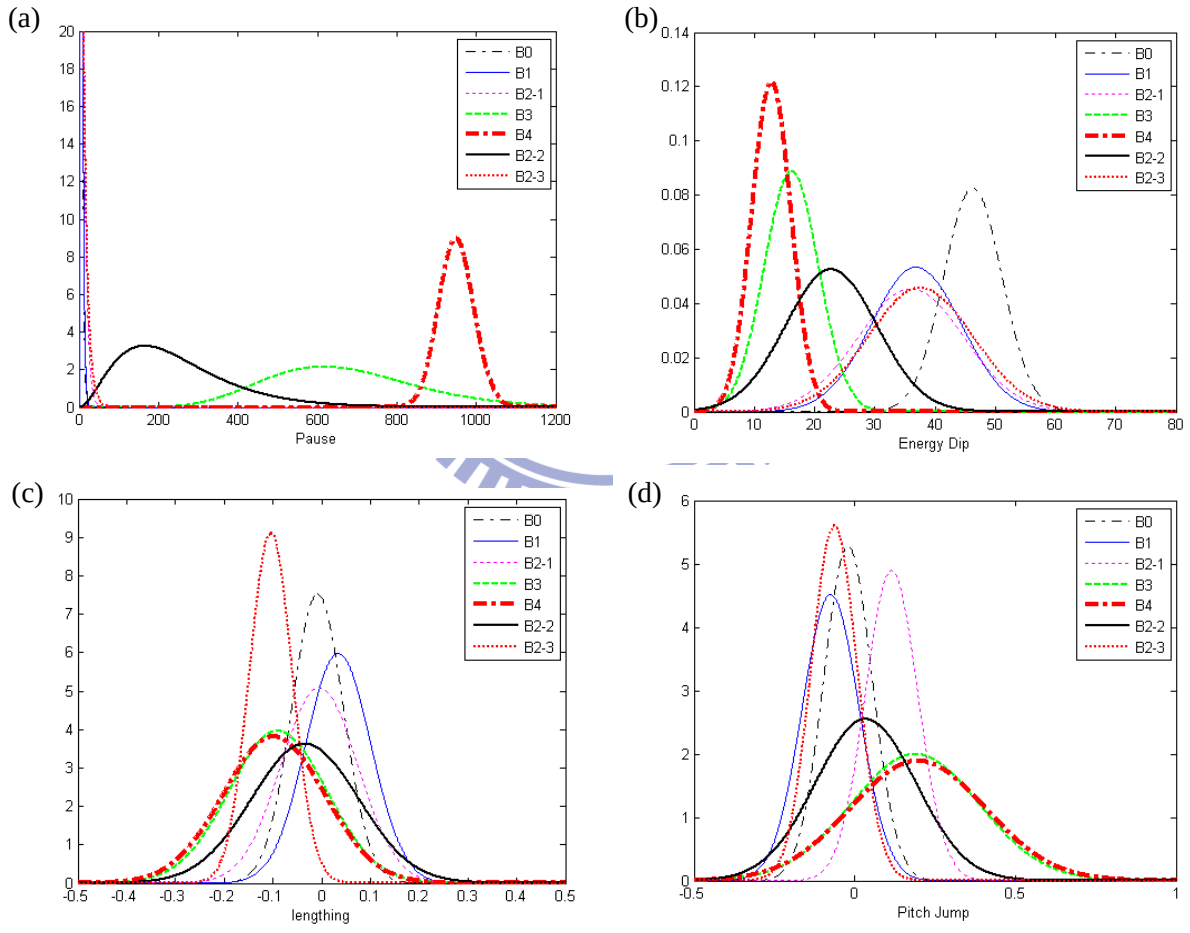


圖 4.8：(a)音節停頓長度 (b)音節間能量低點 (c)正規化音節延長因子與 (d)正規化基頻跳躍值之分布圖

4.3.3 停頓標記結果之分析

圖 4.9 為所有音節邊界之停頓標記的分布圖，由圖中可發現 *B0* 和 *B1* 所佔的比率很高，此表示人類在自發性語音中因為說話速度較快因此有音節較常有緊密相連的情況；而 *B2-1*、*B2-2* 以及 *B2-3* 與朗讀式語音【16】相比出現的比率差不多，但 B2-1 所佔之比率上升且 B2-2 所佔之比率下降，表示人類在語速較快的自發性語音中，在一韻律詞邊界時較少產生音節間的停頓，而是較常使用基頻的跳躍。最後與【11】相比，*BPI*、*BPO* 以及 *BP* 數量少了許多，而 *B0* 和 *B1* 數量增加，這表示人類在聊天時常出現的「particle」通常都是與前後詞緊連，流暢性就如同一般基本音節，因此我們還是可以用 *B0*~ *B4* 來標記其韻律停頓。

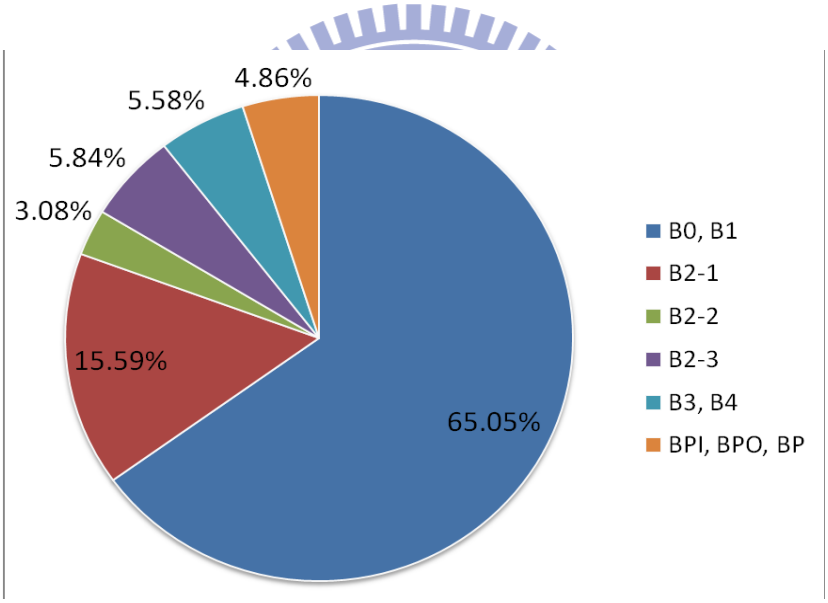


圖 4.9：韻律停頓標記分佈圖

接著觀察實際的韻律停頓標記結果，以下將列出兩個範例及圖 4.10 和圖 4.11 為其相對應之音檔信號圖。由範例一對照圖 4.10 可看出 *B2-1* 和 *B2-2* 確實能反映出基頻跳躍及音節間短停頓的現象。而範例二對照圖 4.11 中顯示 *B2-3* 之處會有音節拖長音的現象發生，值得注意的是 *B2-3* 也可能因為前一音節長度太短而造成相較此音節有較長的音節長度，如範例二中的「一些」,「一」的音節數明顯比「些」要來的短，造成在訓練模型時「些」會被判斷為有拖長音的現象發生。此外，在「particle」部分，範例一中的「O」及範例二中的「LA」與

【11】相比，本研究訓練之模型已將他們歸類為順暢語流中的音節，因此不會出現特殊韻律現象停頓標記。

範例一：O(Par) 我(Nh) **{B2-1}** 在(P) 一家(DM) **{B2-2}** 公關(Na) 公司(Nc) 上班(VA)

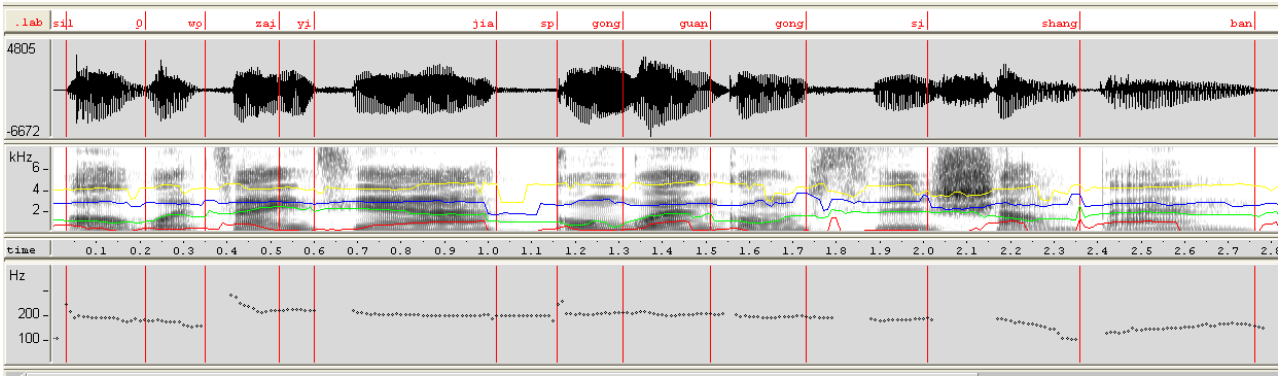


圖 4.10：韻律停頓標記範例一之音檔信號圖

範例二： 在(P) 一些(Neqa) **{B2-2}** 飯**{B2-2}** 店(Nc) LA(Par) **{B2-3}** 或是(Caa) 說(VE) **{B2-3}** 一些(Neqa) **{B2-3}** 大型(Na) 的(DE) **{B2-2}** 購物(VA) 中心(Nc) 辦(VC) 活動(Na)

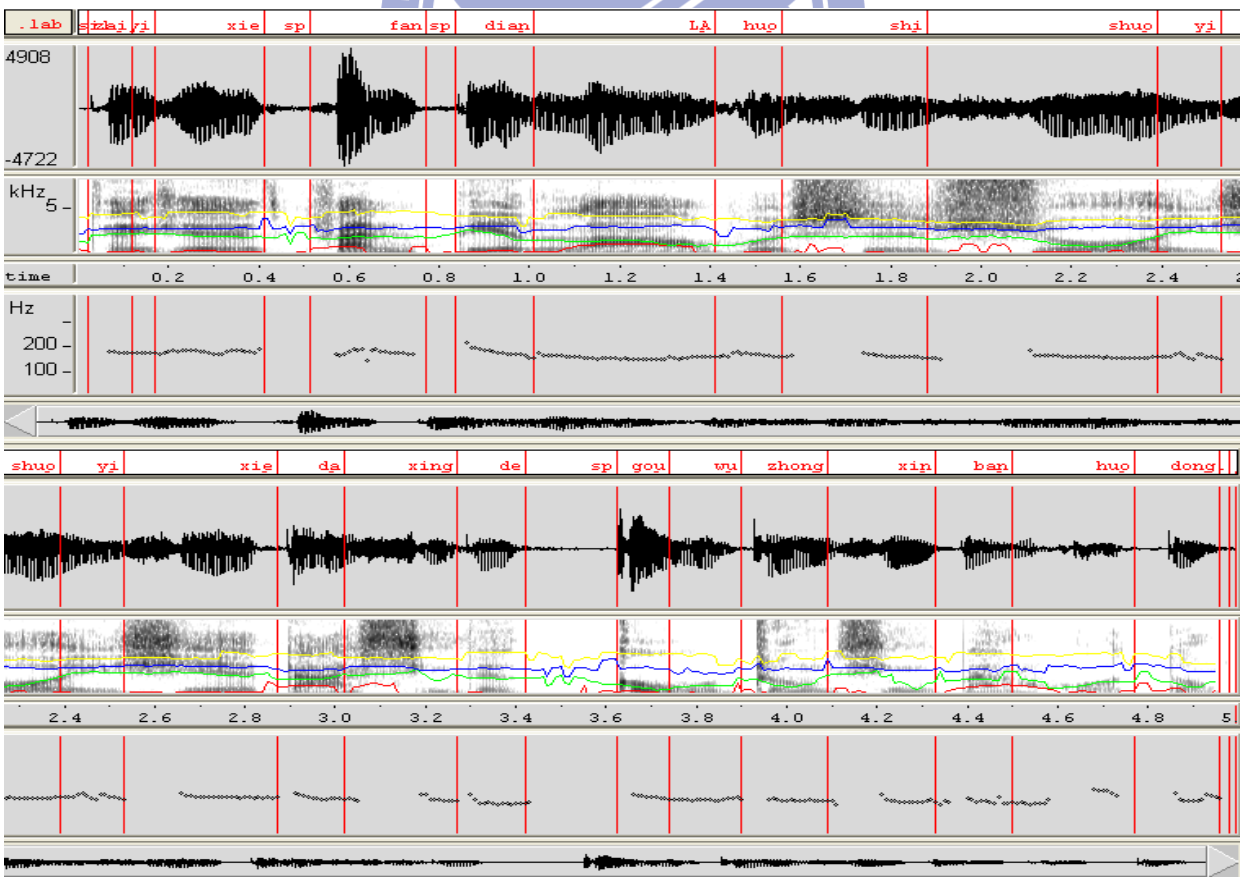


圖 4.11：韻律停頓標記範例二之音檔信號圖

4.4 以韻律模型協助語音辨認

在有了自發性語音韻律模型之後，本研究將利用韻律參數對初級辨認結果所產生之詞串重新計算分數，挑選出最佳之辨識結果。傳統的語音辨識數學式如下：

$$W^* = \arg\max_W P(W|A) \tag{4-17}$$

其中 W 為 n 個詞所組成的詞串 $\{w_1, w_2, \ldots, w_n\}$；$P(W|A)$ 為給定一組聲學參數向量序列 A 下，詞串 W 之事後機率(Posterior Probability)，(4-17)式表示找尋一組詞串 W 使得 $P(W|A)$ 得到最大值。除了聲學模型之外，我們可將韻律模型加入(4-17)式中，在此我們定義詞串 W 的對應韻律參數為 $F=\{f_1, f_2, \ldots, f_n\}$，因此(4-17)式可改寫為：

$$W^* = \arg\max_W P(W|A,F) \tag{4-18}$$

根據貝式定理(Bayes' Theorem)，我們可將(4-18)式展開成：

$$\begin{aligned} W^* &= \arg\max_W P(W|A,F) \\ &= \arg\max_W P(A,F|W)P(W) \\ &= \arg\max_W P(A|W)P(F|W)P(W) \end{aligned} \tag{4-19}$$

在此為了簡化實驗，我們假設聲學參數序列 A 與韻律參數序列 F 互相獨立。其中 $P(A|W)$ 由聲學模型求得，$P(W)$ 則由語言模型求得，而 $P(F|W)$ 將由韻律模型求得。我們可將 $P(F|W)$ 展開如下：

$$\begin{aligned} P(F|W) &= \sum_{POS,B} P(F,POS,B|W) \\ &= \sum_{POS,B} P(F|POS,B,W)P(B,POS|W) \\ &= \sum_{POS,B} P(F|POS,B,W)P(B|POS,W)P(POS|W) \end{aligned} \tag{4-20}$$

其中 $P(F|POS,B,W)$ 及 $P(B|POS,W)$ 分別為停頓標記聲學模型和停頓標記語言模型，將由韻律模型求得；$P(POS|W)$ 為詞對應的詞性之模型。而為了簡化實驗，我們將(4-20)式簡化為此三個模型分數機率相乘之最大值，如下：

$$P(F|W)\approx\max_{POS,B} P(F|POS,B,W)P(B|POS,W)P(POS|W) \quad (4\text{-}21)$$

本研究採用兩段式語音辨認架構，如圖 4.12 所示。第一階段利用聲學模型計算語音信號的聲學分數，接著利用語言模型計算詞與詞的關聯程度與出現機率，辨認產生最佳 N 條詞串。為了產生較多正確的候選詞串供第二階段選擇，通常會將較長的測試音檔切成短句再做最佳 N 條詞串之辨認。在此，我們將先由音節辨認結果去對音檔進行強迫對齊，求得每個音節的時間資訊，接著根據短靜音(short pause, sp)之長度，將 sp 大於 350ms[9]之處再切開成較小的音段去做辨認，並產生最好的前 100 名詞串辨認結果，使得辨認涵蓋率(coverge rate)較高；第二階段將利用韻律參數，給予每個音節邊界一個機率的分數，最後將聲學模型、語言模型及韻律模型分數作權重結合並對每一條路徑重新計算分數，決定出最可能之辨認結果。重新計分公式如下：

$$W^{*}=\lambda_1\log P(A|W)+\lambda_2\log P(W)+\lambda_3\log P(F|W) \quad (4\text{-}22)$$

其中 λ_1、λ_2 及 λ_3 分別為各個模型之權重；而 $P(F|W)$ 將分為停頓標記聲學模型、停頓標記語言模型及詞對應詞性模型等三個子模型如(4-21)式所示。

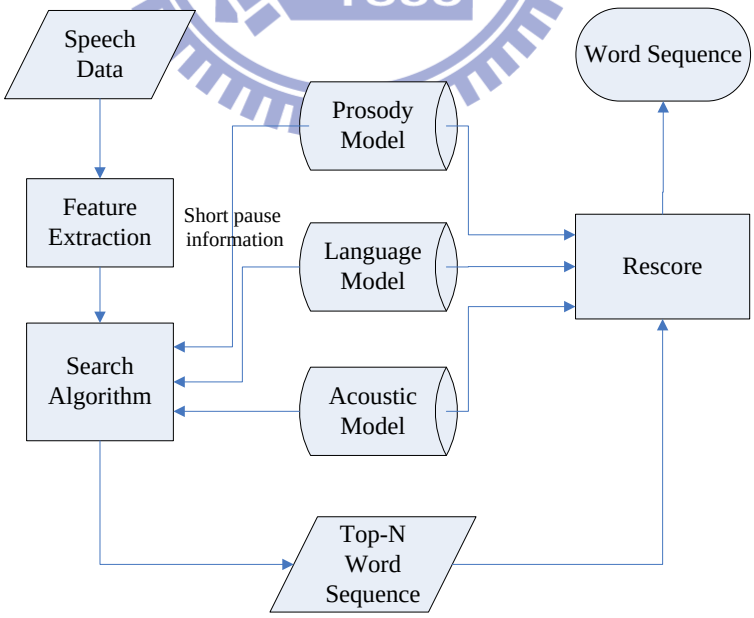


圖 4.12：辨識系統架構圖

[9] 350ms 約為自發性語音韻律模型中 *B3* 之門檻值。

第五章 實驗結果及討論

本章節將討論使用本研究調適之語言模型及韻律模型之辨認狀況。在 5.1 節將分析語言模型之實驗結果，本研究亦將實驗延伸到不同的口語對話語料庫；5.2 節將分析以韻律模型協助自發性語音辨認之結果並討論。

5.1 辨認系統加入語言模型實驗

5.1.1 MCDC 辨識結果

利用聲學模型對語音作辨識將產生音節辨認資訊，若加入語言模型，辨識單元即可由音節變為詞，但為了與聲學模型比較，我們亦會將詞轉為字元及音節來做比較。表 5.1 為本研究由建立 General-LM 後，階段性調適語言模型後之辨識結果。其中 Free-gram-syllable-LM 即為 3.1 節聲學模型之辨認實驗，只有音節辨識率；在加入 General-LM 之後的實驗開始有字元及詞的辨識率，在此值得注意的是，由於 MCDC 的測試語料中的辨識單元含有「particle」及「paralinguistic」，但是 General-LM 中並無這兩類的機率，為了比較之公平性，本研究也針對辨認結果另外計算「一般詞」的辨識率，以下各辨識率表中的(*)皆為「一般詞」之辨識率。而本實驗之最大辨認率為 99.64%。

表 5.1：各階段語言模型之辨識率

Acc(%) / case	Word	Character	Syllable
Free-gram-syllable-LM			48.74(48.55)
General-LM	29.30(31.80)	38.45(40.40)	46.78(49.51)
Stage1-LM	36.35(36.15)	45.03(45.07)	53.75(54.49)
Stage2-LM	37.91(37.61)	45.93(46.26)	54.26(55.75)
Stage3-LM	48.65(48.90)	55.01(55.69)	61.28(62.95)

註: (*)中為只計算一般詞之辨識率。

觀察表 5.1 可注意到在加入 General-LM 後音節辨識率下降了 1.96%，此乃因為 General-LM 中並無「particle」及「paralinguistic」之機率，若觀察「一般詞」的辨識率則可發現辨識率上升了 0.96%，代表在加入語言模型後，因為有了文法的資訊，確實能協助中文的辨識。在 Stage1 本研究先簡單估計「particle」及「paralinguistic」之 uni-gram 機率，接著在 Stage2 時以較嚴謹的方法估算其 bi-gram 機率，值得注意的是，在此二階段「一般詞」的辨識率會比含「particle」及「paralinguistic」的辨識率低，由此可知本研究提出的估算方法是能有效辨認出「particle」及「paralinguistic」。最後在 Stage3 再以 MAP 方法調適「一般詞」的機率分布，由表中可看出各個階段皆能有效提升辨識率。

此外，在口語對話中存在許多常用詞，這些詞以連接詞及副詞居多，在此我們列出幾個測試語料中常出現的詞，並比較在使用初始模型及本研究之 Stage3 語言模型之辨識率，如表 5.2 所示。由表可看出在本研究調適後的語言模型對於口語中的常用詞確實能比合乎文法的朗讀式語音語言模型有效地辨認出來。值得注意的是，表 5.2 中對於「就是」、「沒有」及「所以」的辨識率較其他詞要來的低，其原因可能是因為這三個詞常被說話者拿來當說話時句首的慣用插入詞，其本身已不保有其原有的完整語意，因此在語音信號中這類詞通常會有音節合併的現象，造成無法正確建立其聲學模型，若能有效處理音節合併的問題，辨識率將能有效地改善。

表 5.2：口語對話常用詞之辨識率

詞(數量) / 模型	就是(114)	我們(94)	然後(78)	因為(64)	沒有(60)	所以(44)
General-LM	58.77%	78.72%	53.85%	67.19%	58.33%	36.36%
Stage3-LM	71.05%	85.11%	75.64%	78.13%	70%	68.18%

本研究也另外實驗將 General-LM 與 MCDC 訓練語料直接使用 MAP 估算法做調適，表 5.3 為實驗結果之辨識率，結果也顯示本研究提出的估算「particle」及「paralinguistic」方法將使得整體辨識率有較佳的效益。

表 5.3：語言模型不同調適方法之辨識率

Acc(%) / case	Word	Character	Syllable
General-LM	29.30(31.80)	38.45(40.40)	46.78(49.51)
MAP 估算法	48.17(48.09)	54.81(55.39)	60.91(62.45)
Stage3-LM	48.65(48.90)	55.01(55.69)	61.30(62.95)

註: (*)中為只計算一般詞之辨識率。

5.1.2 中央社對話語料庫之辨識實驗

本研究建立之語言模型在 MCDC 語料庫上有效地改善後，我們也將辨識實驗延伸到中央社的對話語料庫中。中央社語料庫為一男一女之口語對話語料庫，對話內容分為工作、休閒運動、旅遊...等主題。本研究挑選其中兩種與 MCDC 語料庫差異較大的主題進行語音辨認實驗，一個是談「命運」，另一個則是聊「居家生活」。表 5.4 為挑選測試語料統計，在此值得注意的是此語料庫並不像 MCDC 語料庫一樣有標記各種副語言現象，因此「Paralinguistic」的數量很低。

表 5.4：中央社測試語料統計

	Lexical word	Particle	Paralinguistic	Total word	Total syllable
數量	6,336	721	52	7,109	10,006

本研究使用 MCDC 之聲學模型當作基本辨識系統，並針對五種語言模型對中央社語料庫作語音辨認實驗，其辨識結果如下表：

表 5.5：MCDC 與中央社辨識率比較(Acc(%))

		MCDC	中央社
Word	General-LM	29.30(31.80)	25.78(26.51)
	MCDC-LM	48.02(49.09)	29.20(30.69)
	Stage1-LM	36.35(36.15)	26.34(26.98)
	Stage2-LM	37.91(37.61)	27.67(28.46)
	Stage3-LM	**48.65(48.90)**	**33.22(34.87)**

Character	General-LM	38.45(40.40)	32.51(33.14)
	MCDC-LM	56.08(56.89)	37.34(38.87)
	Stage1-LM	45.03(45.07)	34.52(35.23)
	Stage2-LM	45.93(46.26)	35.78(37.12)
	Stage3-LM	**55.01(55.69)**	**40.33(41.76)**
Syllable	General-LM	46.78(49.51)	40.09(41.19)
	MCDC-LM	60.21(61.66)	42.39(44.35)
	Stage1-LM	53.75(54.49)	41.12(42.30)
	Stage2-LM	54.26(55.75)	42.23(43.84)
	Stage3-LM	**61.30(62.95)**	**47.03(48.92)**

註: (*)中為只計算一般詞之辨識率。

觀察表 5.5 可以發現中央社對話語料庫的辨識率普遍比 MCDC 語料庫的辨識率要來的低，在此我們推測原因有以下幾點：

- 因為基本辨認系統中最重要的聲學模型是由 MCDC 訓練來的，因此在 MCDC 測試語音信號上的預估自然比中央社的好。
- 再來中央社語料庫中有許多 cross talk 的狀況，這些在語音辨識上有極大的困難度，為了實驗方便，我們將發生 cross talk 的地方切斷不拿來做語音辨認，因此測試語料可能有一些文法很亂的短句出現。
- 本研究調適的語言模型與 MCDC 本身的關係較密切，且本研究對於「particle」及「paralinguistic」有較好的預估能力，但是中央社的測試語料中較少標記此現象，因此較無法發揮本模型之能力。

雖然在中央社的辨識率不盡理想，若是能使用中央社語料庫本身建立之聲學模型，相信辨識率應該能提升許多，但是本研究主要在於探討調適語言模型的方法是否有效，因此目前暫不另外訓練聲學模型，且實驗結果證明本研究提出之方法確實能有效相對提升語言模型的辨識率。

5.2 辨認系統加入韻律模型實驗

5.2.1 以韻律模型協助辨認之辨識率

在 4.4 節我們提到在把韻律模型加入幫助辨認前為了有較多較好的候選詞串，我們會先將測試語料切短句以提高辨認涵蓋率。原始句子中平均每個音段約有 27 個詞，而切短句後每個音段平均約只有 14 個詞。表 5.6 列出測試語料未切短句以及切完短句後的辨認涵蓋率。

表 5.6：原始句子與切短句後詞之辨認涵蓋率

	Top-1 (%)	Coverge Rate (%)
原始句子	48.65	51.25
切短句	47.45	52.18

由表中可以發現在切完短句後辨認涵蓋率雖然有提高，但是提升幅度並不大，無法像朗讀式語音一樣有較高的涵蓋率；而觀察 Top-1 辨識率時，切完短句後的辨識率比原始句子要來的低，在此我們檢討發現主要原因為在自發性語音中停頓點較不像朗讀式語音叫有語法性，人們在聊天時常常因說錯話或被打斷而造成非流利語音現象發生，而我們在實驗時是以停頓長度為依據來切短句，因此造成切出來的句子有可能只有一、兩個詞，或甚至是詞被切開成兩個字落在兩個不同的音段裡，此時語言模型較無法發揮其效能。

在辨認系統加入韻律模型後之辨識率如表 5.7 所示，在未加入韻律模型前的詞辨識率是 48.65%，加入韻律模型後我們做了兩種實驗：原始句子辨認以及切短句辨認。

表 5.7：加入韻律模型之辨識率

		Word (%)
未加入 PLM (Top-1)	原始句子	48.65
	切短句	47.45
加入 PLM	原始句子	48.7
	切短句	47.89

觀察表 5.7 可發現對原始句子做辨認時，在加入韻律模型後辨識率提升 0.05%，算是持平狀態；對切短句做辨認時提升 0.44%，算是微小幅提升，但是並未超越原始句子 Top-1 的辨識率，因此無法達到我們切短句以提升辨識率的目的。由表 5.7 可推論以韻律模型協助自發性語音辨認似乎並無太大成效，我們將在下一小節探討其原因。

5.2.2 實驗結果討論

本研究在 4.4 節提出的方法主要概念是藉由韻律模型的幫助挑出最符合正確解答的候選詞串，而在此因為切短句後有較高之辨認涵蓋率，因此我們首先觀察切完短句在第一級辨認所產生之 Top-N 候選詞串。表 5.8 及表 5.9 共舉出 5 個音段例子[10]，因為空間關係，我們只列出每段辨認結果最好之之前五名詞串，表 5.8 為較好之例子，表示在加入韻律模型幫助後較有機會選中正確解答；表 5.9 則顯示出候選詞串本身就與解答差異甚遠，表示不管如何選擇皆無法與正確解答相符。觀察表 5.8(b)與表 5.9(a)可發現在切完短句後前半段是有機會選中正確解答，但後半段則無法救回。

表 5.8：較佳之 Top-N 候選詞串

(a) mcdc-25-04_R0100172_seg1	
解答	**O 在 承德路 這邊**
Top-1	O 在 承德路 事變
Top-2	O 在 承德路 是 變
Top-3	O 在 承德路 是 邊
Top-4	**O 在 承德路 這邊**
Top-5	O 的 承德路 事變
(b) mcdc-03-10_R0141143_seg1	
解答	**可是 我 覺得**
Top-1	可是 我 絕對
Top-2	可是 我 絕對 的
Top-3	**可是 我 覺得**
Top-4	可是 我 絕對 @非語音聲
Top-5	可是 我 就 對 對

[10] 我們觀察到當 Top-5 的候選詞串狀況不佳時，Top-100 的候選詞串幾乎大同小異，無法有符合解答的詞串。

表 5.9：較差之 Top-N 候選詞串

(a) mcdc-03-10_R0141143_seg2	
解答	**就是 我 覺得 我們 學校 的 主編 並 不 是**
Top-1	@非語音聲 就是 我 就 我們 學校 的 朱延平 不 是 的 @呼吸聲
Top-2	就是 我 就 我們 學校 的 朱延平 不 是 的 @呼吸聲
Top-3	@非語音聲 就是 我 就 我們 學校 的 朱延平 不 是 了 @呼吸聲
Top-4	就是 我 就 我們 學校 的 總編輯 不 是 的 @呼吸聲
Top-5	就是 我 就 我們 學校 的 朱延平 不 是 了 @呼吸聲
(b) mcdc-01-11_L0031915_seg1	
解答	**因為 造成 很多 人 會 誤解 那個 印象 說 改 車 都 是 拿去 飆車**
Top-1	因為 造成 很多 人 會 物件 那個 影響 說 改 車 都 是 拿去 標
Top-2	因為 造成 很多 人 會 五件 那個 影響 說 改 車 都 是 拿去 標
Top-3	因為 造成 很多 人 會 武檢 那個 影響 說 改 車 都 是 拿去 標
Top-4	因為 造成 很多 人 會 午間 那個 影響 說 改 車 都 是 拿去 標
Top-5	因為 造成 很多 人 會 五間 那個 影響 說 改 車 都 是 拿去 標
(c) mcdc-25-07_L0087755_seg2	
解答	**我 之前 把 我 的 舊 鋼琴 賣掉**
Top-1	我 之前 包含 我 就 港區 賣掉
Top-2	我 之前 保養 我 就 港區 賣掉
Top-3	我 之前 褒揚 我 就 港區 賣掉
Top-4	我 之前 保單 我 就 港區 賣掉
Top-5	我 之前 包含 我 就 鋼琴 賣掉

在第一級做 Top-N 辨認時若無法產生與解答相符的候選詞串，則第二級時就幾乎無法救回來了！而根據表 5.9 可以得知，我們的實驗在第一級辨認時所產生的 Top-N 候選詞串有許多音段根本無法產生與解答相同的詞串，所以在第二級加入韻律模型重新計分後效果也無法呈現出來。

基於此狀況，我們必須檢討第一級的辨認系統，因此接著觀察在做第一級辨認時聲學模型以及語言模型的辨認分數狀況。在此我們同樣以上面 5 個音段為例，觀察正確解答及 Top-5 候選詞串的聲學模型及語言模型分數，如表 5.10 及表 5.11 所示，其中分數為取對數(Log_{10})之結果。觀察表 5.10，我們發現 Top-1 到 Top-3 辨認結果的模型分數總合皆贏過解答的分數，直到 Top-4 的分數總合才等於解答的分數，而根據表 5.8，Top-4 也正好是辨認正確解答。

表 5.10：表 5.8 範例音段之聲學、語言模型分數

(a) mcdc-25-04_R0100172 _seg1	音段內容：O 在承德路這邊	
	聲學模型分數	語言模型分數
解答	**-8915.80**	**-28.25**
Top-1	-8873.22	-31.32
Top-2	-8873.22	-31.64
Top-3	-8873.22	-32.22
Top-4	**-8915.80**	**-28.25**
Top-5	-8873.22	-34.15
(b) mcdc-03-10_R0141143_seg1	音段內容：可是我覺得	
	聲學模型分數	語言模型分數
解答	**-6942.16**	**-15.19**
Top-1	-6858.85	-20.65
Top-2	-6847.16	-23.90
Top-3	**-6942.16**	**-15.19**
Top-4	-6859.33	-22.69
Top-5	-6905.95	-18.71

接著觀察表 5.11，此時可以發現辨認結果的模型分數總合幾乎皆贏過正確解答的分數，其中又以聲學模型最為明顯，圖 5.1 為解答之聲學分數減掉 Top-N 聲學分數之分布圖，由圖可知大部份 Top-N 的聲學分數皆贏過解答的聲學分數，此現象表示辨認器在做辨認時會認為此音段某區的聲學參數較符合某個詞而不是正確解答的詞，而聲學模型又正好是目前做語音辨認時最重要的參考依據，此部分的分數若有偏差，較難由其他模型救回來。根據以上的觀察結果，我們發現並不是韻律模型無法幫助自發性語音辨認，而是目前的自發性語音聲學模型的部分還不夠完善。

表 5.11：表 5.9 範例音段之聲學、語言模型分數

(a) mcdc-03-10_R0141143_seg2	音段內容：就是我覺得我們學校的主編並不是	
	聲學模型分數	語言模型分數
解答	**-17552.22**	**-49.29**
Top-1	-17240.64	-55.76
Top-2	-17262.98	-54.58
Top-3	-17239.40	-56.06
Top-4	-17249.06	-55.75
Top-5	-17261.74	-54.87

(b) mcdc-01-11_L0031915_seg1	音段內容：因為造成很多人會誤解那個印象說改車都是拿去飆車	
	聲學模型分數	語言模型分數
解答	**-23748.90**	**-106.79**
Top-1	-23654.78	-105.24
Top-2	-23654.78	-105.26
Top-3	-23654.78	-105.31
Top-4	-23654.78	-105.34
Top-5	-23654.78	-105.35
(c) mcdc-25-07_L0087755_seg2	音段內容：我之前把我的舊鋼琴賣掉	
	聲學模型分數	語言模型分數
解答	**-16099.65**	**-55.71**
Top-1	-16015.44	-53.72
Top-2	-16008.15	-54.72
Top-3	-16008.15	-55.00
Top-4	-16011.64	-54.73
Top-5	-16030.61	-53.09

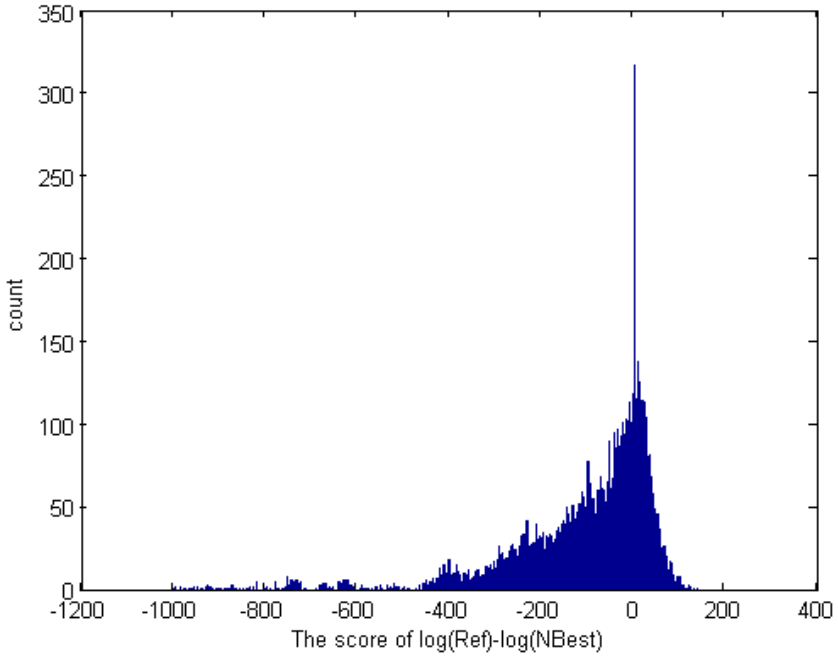


圖 5.1：解答聲學分數與 Top-N 聲學分數相減之分布圖

最後，我們將分析韻律模型目前對於自發性語音辨認所能提供的貢獻。在第一章我們提到過自發性語音中有許多不合乎文法結構之語句，詞組與詞組間之連接機率較不穩定，但是

詞組與韻律結構之間的關係還是存在一定的規則架構，例如：在詞間邊界(Inter-word)的停頓長度通常比詞內邊界長、詞內邊界較少發生基頻跳躍。在做語音辨認時，這些韻律資訊將可以用刪除一些較不可能的候選詞，我們也實地去觀察辨認結果以探討韻律模型的功效。表 5.12 展示五個在未加入韻律模型以及加入後的辨認句子，表中的「句法結構」表示句子中「詞」接「詞」的對應「字數」之連接關係，例如：3→2 表示三字詞接二字詞；而「句子內容」中灰底的部分表示此處結構與解答有差異。在不看辨認出來的詞是否正確的前提下，單看句法結構，由表中可以發現在加入韻律模型的幫助後，辨認器比較能挑出與正確解答句法結構相近的詞串，這表示韻律資訊跟句法結構之間確實帶有許多有用的訊息。

表 5.12：辨認句子結構表

	句子內容	句法結構
(a)		
解答	不至於→這樣→ＮＨＮ→ＨＥＮ→ＨＥＮ	3→2→1→1→1
未加 PLM	不→是→以→以→我→覺得→ＨＥＮ→ＨＥＮ	1→1→1→1→1→2→1→1
加入 PLM	不至於→這樣→ＭＨＭ→ＨＥＮ→ＨＥＮ	3→2→1→1→1
(b)		
解答	對→Ａ→根本→就→沒→什麼→垃圾	1→1→2→1→1→2→2
未加 PLM	對→Ａ→可能→就→沒有→一個→車隊	1→1→2→1→2→2→2
加入 PLM	對→Ａ→可能→就→在→這個→車隊	1→1→2→1→1→2→2
(c)		
解答	ＥＮ→我→家→是→住→在→象山	1→1→1→1→1→1→2
未加 PLM	那個→我→家→是→住→在→香山	2→1→1→1→1→1→2
加入 PLM	也→我→家→是→住→在→香山	1→1→1→1→1→1→2
(d)		
解答	摸→ＯＮ	1→1
未加 PLM	ＭＨＭ	1
加入 PLM	Ｏ→Ｏ	1→1
(e)		
解答	我→住→政治→大學→那裡	1→1→2→2→2
未加 PLM	我→就→真的→再接再厲	1→1→2→4
加入 PLM	我→就→真的→他→現在	1→1→2→1→2

第六章 結論與未來展望

6.1 結論

本研究主要是建立一套中文自發性語音之辨認系統，主要有三個模型分別為：聲學模型、語言模型及韻律模型。在自發性語音中有許多常用的口語詞、感嘆詞或語助詞，這些都與傳統語言模型差異甚大，所以本研究以傳統語言模型為基礎，建立一套自發性語音語言模型，實驗證明此模型可有效運用在自發性語音辨認上。

在自發性語音韻律模型的改進方面，本研究認為口語中的許多感嘆詞的韻律特性與一般基本音節並無差異，因此將把應屬於基本音節的感嘆詞歸為正常語流韻律現象。除此之外，傳統語音辨認通常只使用到聲學模型及語言模型，但是自發性語音中有許多的特殊現象是朗讀式語音中沒有的，因此本研究試著將韻律模型也運用進來，希望能解決這些狀況。而實驗結果雖然不盡理想，但也顯示韻律在自發性語音中句法結構的辨認上帶有一些有用的訊息。

6.2 未來展望

自發性語音辨認要達到商業運用的地步還有很長一段距離，在自發性語音語言模型的研究上通常會面臨語料短缺的問題，若能持續收集人類日常生活聊天中大量的口語詞、新世代用語、及感嘆詞的文字記錄，或學習出自發性語音凌亂的文法結構，對於自發性語言模型的訓練將會有很大的幫助。

另外一方面，本研究實驗結果也注意到自發性語音聲學模型還有很大的改善空間，而聲學模型的訓練遇到的問題主要是音檔的品質及音節切割位置的正確性，音檔的品質或許也反映出自發性語音中的特有現象：常用詞的音節合併現象、忽快忽慢的語速、時大時小的音量、突如其來的背景雜訊、或同時存在兩個人的聲音等情況，都是未來自發性語音辨認在應用上會面臨的問題。若能利用韻律資訊來協助協助聲學模型的建立，或許能夠改善一些問題。待

有了較強健的聲學模型後，能產生更好的 Top-N 候選詞串，再加入韻律模型的幫助，挑選出句法結構正確的詞串，相信將能大幅改善現在的自發性語音辨認系統，若能使自發性語音的辨認率達到如朗讀式語音的辨認率，則電影中人類與機器人對話的情景將不再只是幻想。

參考文獻

【1】 T. Ng, M. Ostendorf, M.Y. Hwang, M. Siu, I. Bulyko, X. Lei, "Web-Data Augmented Language Models for Mandarin Conversational Speech Recognition," in *Proceedings of ICASSP*, 2005, pp. 589–592.

【2】 Hiroaki Nanjo, Tatsuya Kawahara, "Language Model and Speaking Rate Adaptation for Spontaneous Presentation Speech Recognition," *in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), vol. 12, 2004*

【3】 M. Bacchiani, B. Roark, and M. Saraclar, "Language model adaptation with MAP estimation and the perceptron algorithm," in *Proceedings of the HLTNAACL*, Boston, MA, May 2004.

【4】 Samuelsson, C., Reichl, W., 1999. A class-based language model for large-vocabulary speech recognition extracted from part-of-speech statistics. In: Proceedings of ICASSP. pp. 537–540.

【5】 T. Yokoyama, T. Shinozaki, K. Iwano and S. Furui, "Unsupervised class-based language model adaptation for spontaneous speech recognition," Proc ICASSP, vol.1, pp. 236–239, 2003.

【6】 G. Moore and S. Young. Class-based Language Model Adaptation using Mixtures of Weights. In *Proc. ICSLP*, 2000.

【7】 A. Stolcke, E. Shriberg, D. Hakkani-Tür ,and G. Tür, "Modeling the prosody of hidden events for improved word recognition," in *Proc. of Eurospeech 1999*, pp. 311-314.

【8】 M. Ostendorf, I. Shafran, and R. Bates, "Prosody models for conversational speech recognition," in *Proc. of the 2nd Plenary Meeting and Symposium on Prosody and Speech Processing 2003*, pp. 147–154.

【9】 K. Chen and M. Hasegawa-Johnson. "Improving the robustness of prosody dependent

language models based on prosody syntax dependence". In *Proceedings IEEE Workshop on Speech Recognition and Understanding*, pp. 435–440, St. Thomas, U. S. Virgin Islands, 2003.

【10】 K. Chen, M. Hasegawa-Johnson, A. Cohen, S. Borys, S.S. Kim, J. Cole, and J.Y. Choi, "Prosody dependent speech recognition on radio news corpus of American english," *IEEE Transactions on Speech, Audio, Language Processing*, Vol. 14, No. 1, pp. 232–245, 2006.

【11】 周裕倫，"中文自發性語音之韻律標記及韻律模式"，國立交通大學碩士論文，民國九十八年七月。

【12】 S.C. Tseng, "Processing Spoken Mandarin Corpora," *Traitement automatique des langues, Special Issue: Spoken Corpus Processing*, Vol. 45, No. 2, pp. 89-108, 2004.

【13】 S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P.Woodland, The HTK Book Version 3.0. Cambridge, U.K.: Cambridge Univ. Press, 2000.

【14】 X. Huang, A. Acero, HW. Hon, "Spoken Language Processing, " pp.558-559, 2001

【15】 C.Y. Tseng and Z.Y. Su, "Corpus approach to phonetic investigation - methods, quanitative evidence and findings of Mandarin speech prosody," in *Proc. of Oriental COCOSDA Workshop 2006*, pp. 123-138.

【16】 江振宇，"非監督式中文語音韻律標記及韻律模式"，國立交通大學博士論文，民國九十八年三月。

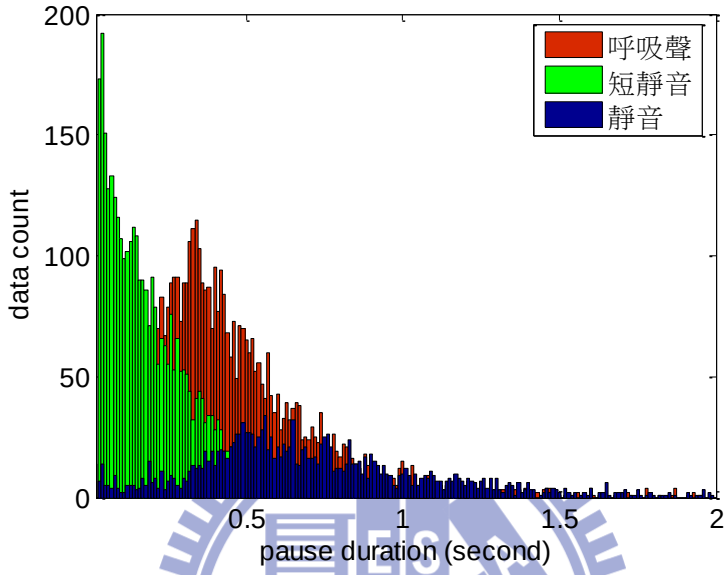
附錄一 詞性分類表

2 類詞性			8 類詞性			23 類詞性			46 類詞性		
編號	中文詞性	代號	編號	中文詞性	代號	編號	中文詞性	代號	編號	中文詞性	代號
1	實詞	S	1	非謂形容詞	A	1	非謂形容詞	A	1	非謂形容詞	A
2	功能詞	F	2	連接詞	C	2	連接詞	C	2	對等連接詞	Caa
2	功能詞	F	2	連接詞	C	2	連接詞	C	3	連接詞，如：等等	Cab
2	功能詞	F	2	連接詞	C	2	連接詞	C	4	連接詞，如：的話	Cba
2	功能詞	F	2	連接詞	C	2	連接詞	C	5	關聯連接詞	Cbb
2	實詞	S	3	副詞	D	5	副詞	D	6	數量副詞	Da
2	實詞	S	3	副詞	D	3	動詞前程度副詞	Dfa	7	動詞前程度副詞	Dfa
2	實詞	S	3	副詞	D	4	動詞後程度副詞	Dfb	8	動詞後程度副詞	Dfb
2	實詞	S	3	副詞	D	5	副詞	D	9	時態標記	Di
2	實詞	S	3	副詞	D	5	副詞	D	10	句副詞	Dk
2	實詞	S	3	副詞	D	5	副詞	D	11	副詞	D
1	實詞	S	4	體詞	N	6	普通名詞	N	12	普通名詞	Na
1	實詞	S	4	體詞	N	6	普通名詞	N	13	專有名詞	Nb
1	實詞	S	4	體詞	N	6	普通名詞	N	14	地方詞	Nc
1	實詞	S	4	體詞	N	9	後置詞,位置詞	Ng	15	位置詞	Ncd
1	實詞	S	4	體詞	N	7	時間詞	Nd	16	時間詞	Nd
1	實詞	S	4	體詞	N	8	定詞,量詞	Ne	17	數詞定詞	Neu
1	實詞	S	4	體詞	N	8	定詞,量詞	Ne	18	特指定詞	Nes
1	實詞	S	4	體詞	N	8	定詞,量詞	Ne	19	指代定詞	Nep
1	實詞	S	4	體詞	N	8	定詞,量詞	Ne	20	數量定詞	Neqa
1	實詞	S	4	體詞	N	8	定詞,量詞	Ne	21	後置數量定詞	Neqb
1	實詞	S	4	體詞	N	8	定詞,量詞	Ne	22	量詞	Nf
1	實詞	S	4	體詞	N	9	後置詞,位置詞	Ng	23	後置詞	Ng
1	實詞	S	4	體詞	N	6	普通名詞	N	47		Nv
1	實詞	S	4	體詞	N	10	代名詞	Nh	24	代名詞	Nh

2 類詞性			8 類詞性			23 類詞性			46 類詞性		
編號	中文詞性	代號	編號	中文詞性	代號	編號	中文詞性	代號	編號	中文詞性	代號
2	功能詞	F	5	感嘆、語助詞	T	12	感嘆、語助詞	T	25	感嘆詞	I
2	功能詞	F	6	介詞	P	11	介詞	P	26	介詞	P
2	功能詞	F	5	感嘆、語助詞	T	12	感嘆、語助詞	T	27	語助詞	T
1	實詞	S	7	動詞	V	13	不及物動詞	VA	28	動作不及物動詞	VA
1	實詞	S	7	動詞	V	13	不及物動詞	VA	29	動作使動動詞	VAC
1	實詞	S	7	動詞	V	14	及物動詞	VC	30	動作類及物動詞	VB
1	實詞	S	7	動詞	V	14	及物動詞	VC	31	動作及物動詞	VC
1	實詞	S	7	動詞	V	14	及物動詞	VC	32	動作接地方賓語動詞	VCL
1	實詞	S	7	動詞	V	14	及物動詞	VC	33	雙賓動詞	VD
1	實詞	S	7	動詞	V	14	及物動詞	VC	34	動作句賓動詞	VE
1	實詞	S	7	動詞	V	14	及物動詞	VC	35	動作謂賓動詞	VF
1	實詞	S	7	動詞	V	13	不及物動詞	VA	36	分類動詞	VG
1	實詞	S	7	動詞	V	15	狀態不及物動詞	VH	37	狀態不及物動詞	VH
1	實詞	S	7	動詞	V	15	狀態不及物動詞	VH	38	狀態使動動詞	VHC
1	實詞	S	7	動詞	V	15	狀態不及物動詞	VH	39	狀態類及物動詞	VI
1	實詞	S	7	動詞	V	14	及物動詞	VC	40	狀態及物動詞	VJ
1	實詞	S	7	動詞	V	14	及物動詞	VC	41	狀態句賓動詞	VK
1	實詞	S	7	動詞	V	14	及物動詞	VC	42	狀態謂賓動詞	VL
1	實詞	S	7	動詞	V	16	有	V_2	43	有	V_2
2	功能詞	F	8	的	DE	17	的，之，得	DE	44	的，之，得	DE
2	功能詞	F	9	是	SHI	18	是	SHI	45	是	SHI
1	實詞	S	11	外文	FW	20	外文標記	FW	46	外文標記	FW
1	實詞	S	10	定量複合詞	DM	19	定量複合詞	DM	58	定量複合詞	DM
4	Paralinguistic	ParaL	13	Paralinguistic	ParaL	24	Paralinguistic	ParaL	59	Paralinguistic	ParaL
5	Particle	Par	14	Particle	Par	25	Particle	Par	60	Particle	Par
5	Particle	Par	14	Particle	Par	25	Particle	Par	61	Marker	Marker

附錄二 韻律模型初始停頓標記門檻值之選定

A. *Th*1, *Th*2 和 *Th*3 之定義:



圖二.1：已標記音節邊界之音節停頓長度分佈

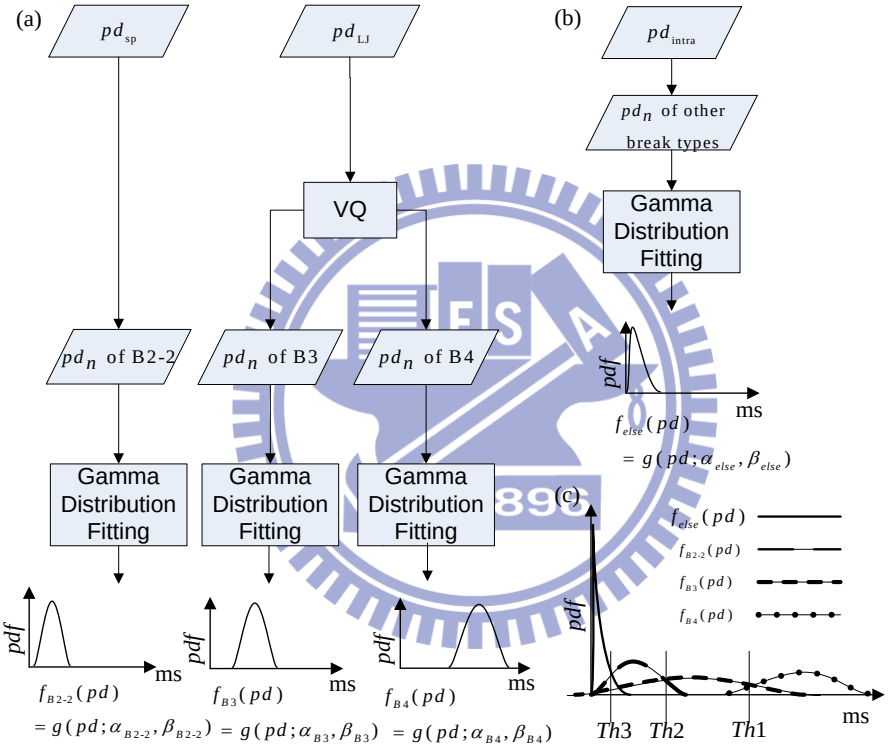
Th1、*Th2* 和 *Th3* 為區分 *B4*、*B3*、*B2-2* 以及其它停頓標記之停頓長度門檻值，由於 MCDC 語料庫當中已具有一些語流中斷之標記（Labeled Juncture, LJ），例如：靜音（silence）、短靜音（short pause, sp）或呼吸聲（breathe），表示這些標記所對應之音節邊界，都有人類可明顯觀察到的停頓現象，其資料分布如圖二.1 所示，因此在本研究使用一個半監督式（semi-supervised）的方法得到 *B4*、*B3*、*B2-2* 以及其它停頓標記對應之停頓長度機率分佈 $f_{B3}(pd)$、$f_{B4}(pd)$、$f_{B2\text{-}2}(pd)$和 $f_{else}(pd)$ 以決定 *Th1*、*Th2* 和 *Th3* 這些門檻值，如圖 4.10 所示。但是由於人類的標記會有不一致性以及自動切割的方式會造成切割位置不準確的狀況，因此先計算標記為「短靜音」以及其他已標記之停頓長度平均值 μ_{sp} 以及 μ_{LJ}，以收集較為可靠之停頓長度資料，其中詞內音節邊界、短靜音以及其他已標記之停頓長度資料，如下：

$$pd_{\text{intra}} = \{pd_n : pd_n \in \text{intra-word syllable jucture}\}$$

$$pd_{\text{sp}} = \{pd_n : pd_n \in \text{short pause}, pd_n - \mu_{sp} < pd_n - \mu_{LJ}, pd_n \geq 0.03\}$$

$$pd_{\text{LJ}} = \{pd_n : pd_n \in \text{silence or breathe}, pd_n - \mu_{\text{LJ}} < pd_n - \mu_{sp}, pd_n \geq 0.03\}$$

接著以向量量化（Vector Quantization, VQ）的方式將 pd_{LJ} 中的資料分為兩群，pd_{sp} 以及 pd_{intra} 的資料視為第三群和第四群，使用伽瑪分佈建構這四群資料的機率分佈，如圖二.2 (a)和(b)所示，平均值由高至低排列之分布分別為 $f_{B4}(pd)$、$f_{B3}(pd)$、$f_{B2\text{-}2}(pd)$ 及 $f_{else}(pd)$，並且設定等機率點對應之停頓長度值即為 *Th*1、*Th*2 和 *Th*3，如圖二.2 (c) 所示。



圖二.2：*Th1, Th2 和 Th3* 之定義方法：計算 (a) *B4*、*B3 和 B2-2* 與 (b)其他停頓標記之停頓長度的機率分佈以及 (c) *B4*、*B3*、*B2-2* 和其他停頓標記之門檻值

B. *Th*5 之定義：

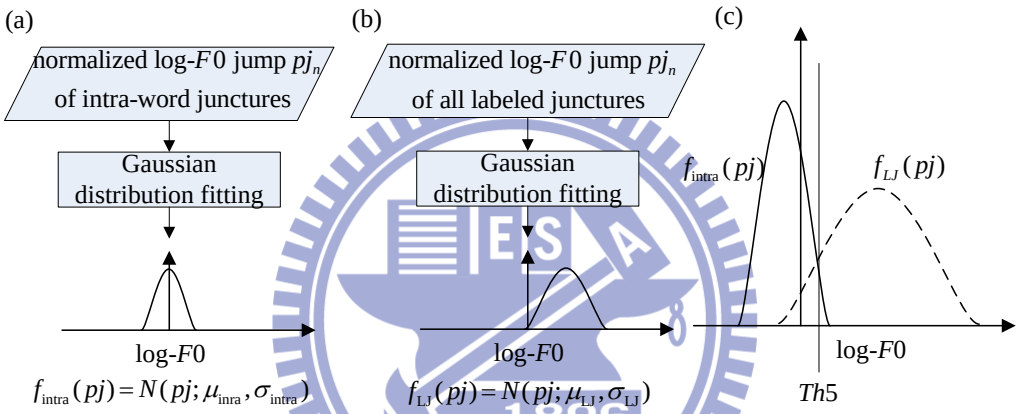
Th5 是將小於 *Th3* 之詞間音節邊界，再分為 *B2-1* 以及 *B0/B1/B2-3* 之正規化基頻差值。首先定義第 *n* 個音節的正規化基頻跳躍值 pj_n 為扣掉總體平均值以及聲調 AP 後，基頻殘餘量之第一維差值，如下：

$$pj_n = \mathbf{sp2}_{n+1}(1) - \mathbf{sp2}_n(1)$$

其中扣掉總體平均值以及聲調 AP 後之基頻殘餘量如下：

$$\mathbf{sp2}_n = \begin{cases} \mathbf{sp1}_n - \boldsymbol{\beta}_{pr_n}, & \text{if } n\text{th syllable is particular syllable } pr \\ \mathbf{sp1}_n - \boldsymbol{\beta}_{t_n}, & \text{if } n\text{th syllable is base syllable } s \text{ corresponding tone } i \end{cases}$$

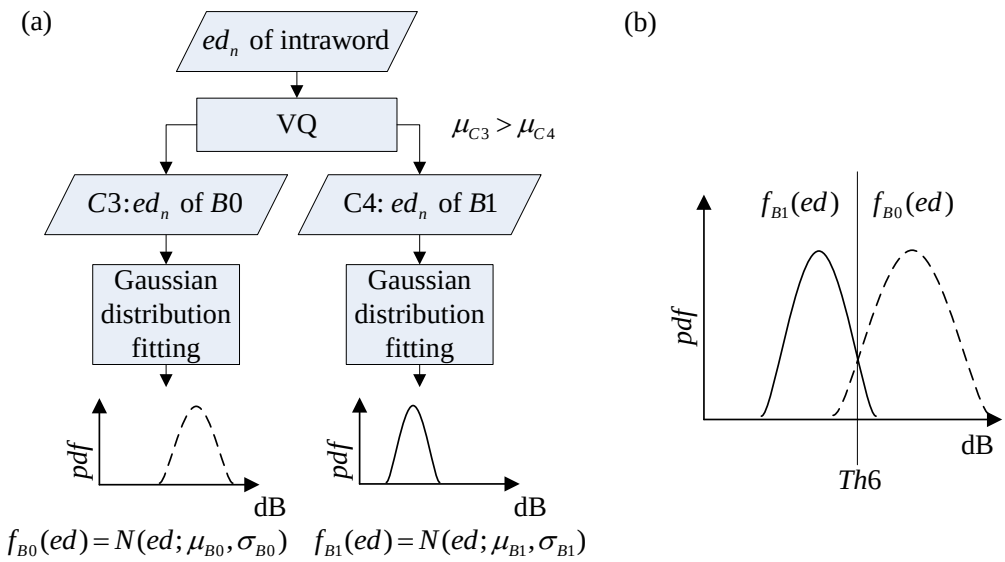
接著將詞內音節邊界和所有已標記之音節邊界，所對應到之相鄰兩音節正規化基頻跳躍值，各自以高斯分佈建構這兩群資料之機率分布，並設定兩個高斯機率分佈的等機率點所對應到之正規化基頻差值為 *Th5*。



圖二.3：*Th5* 之定義方法：計算 (a)詞內與 (b)已標記之音節邊界正規化基頻差之機率分佈以及 (c)門檻值之定義

C. *Th4* 和 *Th6* 之定義：

Th4 和 *Th6* 是將小於 *Th3* 之詞內音節邊界，再分為 *B0* 以及 *B*1 之基頻停頓長度及音節間能量低點，其中基頻停頓長度即為兩音節間基頻軌跡之距離。一般來說 *B0* 表示兩音節有非常緊密的連接，所以相較於 *B1* 會有較短的基頻停頓長度，以及較大的音節能量低點，所以我們定義 *Th4* 的值為一個音框長度（=10ms）。接著將所有詞內音節邊界所對應之能量低點資料，以向量量化的方式分為兩群，並且用兩個高斯分佈建構這兩群的機率分布，平均值由高至低排列分別為 $f_{B0}(pd)$ 及 $f_{B1}(pd)$，並且設定等機率點對應之能量低點值即為 *Th6*，如圖二.4 所示。



圖二.4：*Th6* 之定義方法：計算 (a)*B0* 和 *B1* 音節能量低點之機率分佈以及 (b)門檻值之定義

D. *Th7* 之定義：

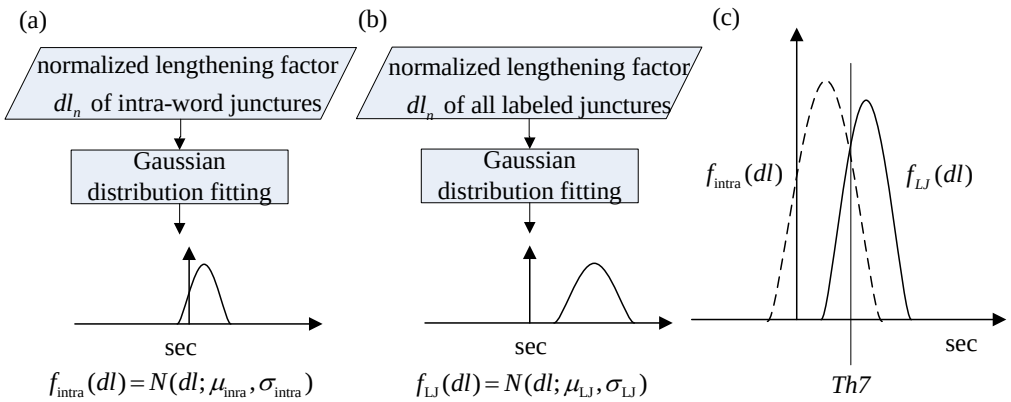
Th7 是將小於 *Th3* 且小於 *Th5* 的詞間音節邊界，再分為 *B2-3* 以及 *B0/B1* 之相鄰兩音節正規化音節延長因子門檻值。首先定義相鄰兩音節之正規化音節延長因子 dl_n，如下：

$$dl_n = sd3_n - sd3_{n+1}$$

其中 $sd3_n$ 為第 n 個音節扣掉總體平均值、音調以及基本音節型態 AP 之音節長度殘存值，如下：

$$sd3_n = \begin{cases} sd1_n - \gamma_{pr}, & \text{if } n\text{th syllable is particular syllable} \\ sd2_n - \gamma_s, & \text{if } n\text{th syllable is base syllable} \end{cases}$$

本研究以正規化音節延長因子來表示在上層的韻律當中，音節邊界之前一音節是否有發生音節延長的現象，若數值越大表示前一音節延長之現象越嚴重。接著我們將詞內音節邊界和所有已標記之音節邊界，所對應到之相鄰兩音節正規化音節延長因子數值，各自以高斯機率建構這兩群之機率分佈，並設定兩個高斯機率分佈等機率點所對應到之正規化音節延長因子值為 *Th7*，如圖二.5 所示。



圖二.5：*Th7* 之定義方法：計算 (a)詞內音節邊界和 (b)已標記之音節邊界的相鄰兩音節正規化音節延長因子之機率分佈以及 (b)門檻值之定義

附錄三 停頓標記聲學模型之問題集

The question set Θ_1 used to construct the decision trees for building the break-acoustics model $p(pd_n, ed_n, pj_n, dl_n, df_n | B_n, \mathbf{l}_n)$ is listed below:

1. Syllable Level

$Q_1$1.1: Is the initial of the following syllable a null one or in {*m, n, l, r*}?

$Q_1$1.2: Is the initial of the following syllable a null one?

$Q_1$1.3: Is the initial of the following syllable in {*b, d, g*}?

$Q_1$1.4: Is the initial of the following syllable in {*f, s, sh, shi, h*}?

$Q_1$1.5: Is the initial of the following syllable in {*m, n, l, r*}?

$Q_1$1.6: Is the initial of the following syllable in {*ts, ch, chi*}?

$Q_1$1.7: Is the initial of the following syllable in {*p, t, k*}?

$Q_1$1.8: Is the initial of the following syllable in {*tz, j, ji*}?

$Q_1$1.9: Is the inter-syllable location an inter-word?

$Q_1$1.10: Is the inter-syllable location a intra-word?

2. Questions related to sentence level features

All the following questions are subject to a prerequisite condition that the current inter-syllable location is an inter-word.

2.1 Word length

$Q_1$2.1.1~4: Is the preceding word an $n \in$ {1, 2, 3, 4}-syllable word?

$Q_1$2.1.5~8: Is the following word an $n \in$ {1, 2, 3, 4}-syllable word?

$Q_1$2.1.9: Is the length of the preceding word in syllable greater than 4?

$Q_1$2.1.10: Is the length of the following word in syllable greater than 4?

2.2 Level-1 POS and special tags

$Q_1$2.2.1~11: Is the POS of the preceding word A/C/D/N/I/P/T/V/DE/SHI/DM?

Q_1 2.2.12 ~ 22 : IS the POS of the following word A/C/D/N/I/P/T/V/DE/SHI/DM?

2.3 Level-2 POS

Q_1 2.3.1 ~ 33 : Is the POS of the preceding word Ca/Cb/Da/Db/Dc/Dd/Df/Dg/Dh/Di/Dj/Dk/Na/Nb/Nc/Nd/Ne/Nf/Ng/Nh/VA/VB/VC/VD/VE/VF/VG/VH/VI/VJ/VK/VL/V_2?

Q_1 2.3.34 ~ 66 : Is the POS of the following word Ca/Cb/Da/Db/Dc/Dd/Df/Dg/Dh/Di/Dj/Dk/Na/Nb/Nc/Nd/Ne/Nf/Ng/Nh/VA/VB/VC/VD/VE/VF/VG/VH/VI/VJ/VK/VL/V_2?

2.4 Level-3 POS

Q_1 2.4.1 ~ 15 : Is the POS of the preceding word Caa/Cab/Cba/Cbb/Dfa/Dfb/Ncd/Neu/Nes/Nep/Neq/VA2/VC1/VH16/VH22?

Q_1 2.4.16 ~ 30 : Is the POS of the following word Caa/Cab/Cba/Cbb/Dfa/Dfb/Ncd/Neu/Nes/Nep/Neq/VA2/VC1/VH16/VH22?

2.5 Combination of POS

Q_1 2.5.1 ~ 7 : Does the POS of the preceding word belong to {Da, Db, Dc, Dd, Dg, Dh, Di, Dj, Dk}/{Na, Nb, Nc}/{Ncd, Ng}/{I, T}/{VA, VG}/{VB, VC, VD, VE, VF, VJ, VK, VL}/{VH, VI}?

Q_1 2.5.8 ~ 14 : Does the POS of the following word belong to {Da, Db, Dc, Dd, Dg, Dh, Di, Dj, Dk}/{Na, Nb, Nc}/{Ncd, Ng}/{I, T}/{VA, VG}/{VB, VC, VD, VE, VF, VJ, VK, VL}/{VH, VI}?

附錄四 停頓標記語言模型之問題集

The question set used to construct the decision trees for building the break-syntax model $P(B_n|\mathbf{l}_n)$ is listed below:

1. Syllable Level

$Q_2$1.1: Is the initial of the following syllable a null one or in {*m*, *n*, *l*, *r*}?

$Q_2$1.2: Is the inter-syllable location an inter-word?

$Q_2$1.3: Is the inter-syllable location a intra-word?

2. Word Level

All the following questions are subject to a prerequisite condition that the current inter-syllable location is an inter-word.

2.1 Word length

$Q_2$2.1.1~4: Is the preceding word an $n \in \{1, 2, 3, 4\}$-syllable word?

$Q_2$2.1.5~8: Is the following word an $n \in \{1, 2, 3, 4\}$-syllable word?

$Q_2$2.1.9: Is the length of the preceding word in syllable greater than 4?

$Q_2$2.1.10: Is the length of the following word in syllable greater than 4?

2.2 Level-1 POS and special tags

$Q_2$2.2.1~11: Is the POS of the preceding word A/C/D/N/I/P/T/V/DE/SHI/DM?

$Q_2$2.2.12~22: IS the POS of the following word A/C/D/N/I/P/T/V/DE/SHI/DM?

2.3 Level-2 POS

$Q_2$2.3.1~33 : Is the POS of the preceding word Ca/Cb/Da/Db/Dc/Dd/Df/Dg/Dh/Di/Dj/Dk/Na/Nb/Nc/Nd/Ne/Nf/Ng/Nh/VA/VB/VC/VD/VE/VF/VG/VH/VI/VJ/VK/VL/V_2?

Q_2 2.3.34 ~ 66 : Is the POS of the following word Ca/Cb/Da/Db/Dc/Dd/Df/Dg/Dh/Di/Dj/Dk/Na/Nb/Nc/Nd/Ne/Nf/Ng/Nh/VA/VB/VC/VD/VE/VF/VG/VH/VI/VJ/VK/VL/V_2?

2.4 Level-3 POS

Q_2 2.4.1 ~ 15 : Is the POS of the preceding word Caa/Cab/Cba/Cbb/Dfa/Dfb/Ncd/Neu/Nes/Nep/Neq/VA2/VC1/VH16/VH22?

Q_2 2.4.16 ~ 30 : Is the POS of the following word Caa/Cab/Cba/Cbb/Dfa/Dfb/Ncd/Neu/Nes/Nep/Neq/VA2/VC1/VH16/VH22?

2.5 Combination of POS

Q_2 2.5.1 ~ 7 : Does the POS of the preceding word belong to {Da, Db, Dc, Dd, Dg, Dh, Di, Dj, Dk}/{Na, Nb, Nc}/{Ncd, Ng}/{I, T}/{VA, VG}/{VB, VC, VD, VE, VF, VJ, VK, VL}/{VH, VI}?

Q_2 2.5.8 ~ 14 : Does the POS of the following word belong to {Da, Db, Dc, Dd, Dg, Dh, Di, Dj, Dk}/{Na, Nb, Nc}/{Ncd, Ng}/{I, T}/{VA, VG}/{VB, VC, VD, VE, VF, VJ, VK, VL}/{VH, VI}?