國 立 交 通 大 學

電信工程學系碩士班

碩 士 論 文

自發性中文語音基本辨認系統之建立
An Implementation of Spontaneous Mandarin Speech Recognition Baseline System

研 究 生：羅應順

指導教授：陳信宏 博士

中華民國九十四年六月

自發性中文語音基本辨認系統之建立

An Implementation of Spontaneous Mandarin Speech Recognition Baseline System

研 究 生：羅應順　　Student : Ying-Shuen Lo

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

國立交通大學

電信工程學系

碩士論文

A Thesis

Department of Communication Engineering

College of Electrical Engineering and computer Science

National Chiao Tung University

In Partial Fulfillment of Requirements

for the Degree of

Master of Science

in Electrical Engineering

June 2005

Hsinchu,Taiwan,Republic of China

中華民國九十四年六月

自發性中文語音基本辨認系統之建立

研究生：羅應順　　指導教授：陳信宏 博士

國立交通大學電信工程學系碩士班

中文摘要

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，如感嘆詞(particles)、不確定語音發音（uncertain sounds）、非語音聲音(paralinguistic sounds)等。我們使用中研院提供的八段雙人對話語料庫做實驗，最後，獲得的音節辨識率約為 56.4% (引入語音模型)。除此之外，在我們的系統裡，我們使用 KPCA(kernel principal components analysis)方法，去進行基本音節 HMM 模型分裂，來模擬發音變異現象。

關鍵詞：自發性中文對話語音辨識、發音變異

An Implementation of Spontaneous Mandarin Speech Recognition Baseline System

Student： Ying-Shuen Lo Advisor: Dr. Sin-Horng Chen

Department of Communication Engineering
National Chiao Tung University

Abstract

In the thesis，a basic spontaneous Mandarin speech recognition 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 database called MCDC (Mandarin Conversational Dialogue Corpus). Finally, A syllable accuracy rate of 56.4% with adapted language model. In addition，the kernel principal components analysis (KPCA) method is used to split the base-syllable HMM models in order to model the pronunciation variation in our system.

Keywords: Spontaneous Mandarin speech recognition, pronunciation variation

誌謝

首先我要感謝陳信宏及王逸如老師；由於，有他們細心的指導，讓我在這兩年學到了許多作研究的技巧，我想這對我未來應該會有很大的助益；其次，我要感謝中央研究院語言研究所曾淑娟 博士，她提供了我們做實驗的語料庫，讓我們能夠順利進行研究。

還要感謝實驗室的學長文輝、振宇及智合，和同學隆勳、柏暄、金翰、希群及佩穎，有他們專長的分享，使我在各方面學到了很多；除此之外，我們還感謝開朗的學弟們，有了你們，讓我們實驗室迷漫著"快樂時光"的氣氛。

最後我要向關心我的家人及朋友，敬上我最誠致的謝意，因為有了你們的支持，才能使我的學識更邁向一大步。

目錄

表目錄

圖目錄

第一章 緒論

在多媒體充斥的世界裡，語音辨識系統扮演著人機介面的橋樑；在過去 5-10 年間隨著信號處理、演算法、硬體的進步，自動語音辨識也有長足的進步，其中進步的項目，包含了統計式圖形辨識（Statistical Pattern Recognition）的方法、資料驅使（Data Driven）的方法、產生聲學及語言模型的方法及基於動態之搜尋方法[1][2]。

若使用現今的語音辨識技術，去辨識朗讀式語音(Read Speech)，例如閱讀報紙或廣播新聞，則其辨識正確率可以達到 90%以上，但是自發性語音的辨識正確率卻非常低；而辨識正確率低的理由，是由於自發性語音辨識器的聲學模型及語言模型，一般而言，是建構在書面上的語言，因此，若用其所建構出來的模型去描述實際自發性語音，會有一些差異存在。

1.1 研究動機

在朗讀式語音裡，我們無法感受，說話者的情緒及所屬的環境，其最主要的原因，是語者刻意朗讀語音，所以，語者所呈現的語音是比較不自然的；因此，就語音辨識的精神而言，研究自發性語音辨識的價值遠高於朗讀式語音。由以上的觀點，我們即選定自發性語音辨識來作為本論文研究的課題；在自發性語音（Spontaneous Speech）有許多非常難以處理的現象，而這些現象影響的層面可以區分為一聲學層面及語言層面，舉例來說，口吃(stutter)、重覆詞語(repetition)…等，就屬於語言層面，而另一方面，鼻化音(nasalized)、發音偏差（inappropriate pronunciation)、音節合併（syllable contraction）…等，就屬於聲學層面。由於我們所使用的語料庫中，有許多音節合併及發音偏差的現象，且礙於有限的研究時間，因此，我們決定從聲學層次上來研究發音變異的情況，以改善自發性語音之辨識率。

1.2 研究方向

我們的研究方向，主要是朝向聲學層次發音變異的情形來做研究，由於自發性語音裡常常會發生輕微或嚴重的音節合併現象和音素(phoneme)變異之情形，因此，針對這些現象，我

們在本篇論文提出了，以資料驅使的方法為基礎想法，試圖將發生音節合併及未發生音節合併的訓練語料找出來,並且用合理的方式，將其轉換為 HMM（Hidden Markov Model）模型。因此，我們在處理發音變異的程序中，我們使用了 KPCA (Kernel Principle Component Analysis)的分析方式，來協助我們找出發音變異的音節，然後進一步為其建立合適的 CD HMM（Context Dependent Hidden Markov Model）模型，加入到我們的辨識系統內，以提升辨識系統的辨識率。

1.3 章節概要

本論文共分為七大章，而各章的內容概要如下:

第一章　緒論

介紹研究的動機、方向及章節概要。

第二章　現代漢語口語對話語料庫之介紹

介紹現代漢語口語對話語料庫之起源及其使用的音檔和轉寫格式,並且有較為詳細的轉寫標籤（tag）說明，和轉寫內容之文字統計。

第三章　基本語音辨識系統之架構

介紹辨識系統之前級特徵參數設定及其相關設定之物理意義,並且以有系統的方式來說明後級所使用的聲學模型，及其建構之方式。

第四章　基本語音辨識系統效能之分析

比較基本語音辨識系統之聲學模型之優劣，並作一些簡易之分析。

第五章　加入語言模型至基本語音辨識系統

語言模型之建構方式,及加入語言模型後,基本語音辨識系統能在辨識率上提升多少。

第六章　語音辨識之發音變異處理

使用 KPCA 的方析的方式，將 411 音節由原來的空間轉換至另一個空間（Feature Space)，然後，我們根據空間座標軸的聲學特性，找出鑑別度比較大的軸之後，利用合適的分類機制將 411 音節予以擴充，最後，再為被分類的音節建立新的 HMM 模型。

第七章　結論與未來展望

第二章 現代漢語口語對話語料庫之介紹

一個實用的語音辨識系統，必須要能接受自發性的語音輸入（包含口吃、笑聲…）。而這種語音辨識系統，我們稱為自然語音辨識系統，而自發性語音辨識系統的辨識能力會比朗讀式辨識系統來得低得很多，其最主要的原因，是因為自然語音多了許多的口語現象（如呼吸聲、吐氣聲..），因此，對於自然語音方面，的確有很大的研究空間。而國內著名的學術機構－中央研究院語言研究所，有鑑於此，於是開始著手籌備一個較為完整的自然語音語料庫－現代漢語口語對話語料庫（Mandarin Conversational Dialogue Corpus，MCDC）。

在這一章裡，將以較有系統的方式，來詳細介紹 MCDC。而介紹的次序，依序為－MCDC 之簡介、MCDC 口語之轉寫標示、MCDC 之相關統計。

2.1 MCDC 之簡介

MCDC 語料是由中央研究院語言學研究所籌備處於 2000～2002 年間所錄製的[3]，其語料發音人是由台北市民中隨機抽樣，並依據 16-25 歲、26-35 歲以及 36-45 歲三大年齡層，最後找了 60 位語者（37 位女性、23 位男性），共錄製了 30 段對話，但由於有轉寫的對話僅有 8 段對話，因此，我們最後共拿了 8 段對話來當作為語料，其中有 16 位發音者（9 位女性、7 位男性），兩兩互相交談，發音者彼此在錄音前都不認識，為確保不陷入無話可談的窘境，除了一開始的自我介紹外，該單位還提供了一些主題，以供發音人選擇對話主題的參考，但這不代表發音人一定要按照所提供的主題，來當作對話的主題，發音人也可自行選擇所想要講的主題來和對方聊天。最後，我們依照選擇的 8 段對話（每段約 60 分鐘），整理出對話主題總表，如表 2.1 所示。

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

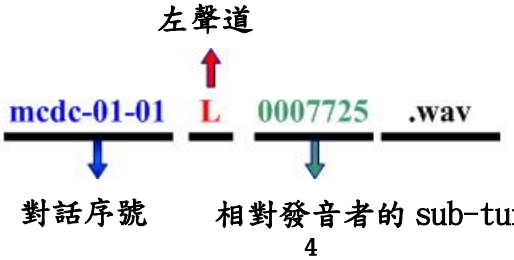
2.1.1 音檔格式說明

1.　原始音檔處理方式

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 分別將兩位發音人的語音錄於左右聲道，因此，每段對話會被儲存在一個音檔裡，再以長度約三分鐘左右，找一個明顯的位置切開，另存為若干個新檔。

2.　音檔後處理方式

由於在語料的轉寫內容中，每位語者的應答(Sub-Turn)皆已標示了時間標籤(time Mark)，所以，為了方便處理語料資訊，我們根據轉寫的時間標籤，將每一段對話的轉寫內容，切割成較小段的應答轉寫內容檔並將原始音檔分割成較小的單聲道音檔，其格式如下



最後，再利用音檔編輯軟體 Cold Editor 將分割好的音檔 Down Sampling 至 16kHz，然後再轉為 PCM 檔。此外，對話序號、左右聲道及語者代號之關係，如附錄一所示。

3. 音檔格式比較

由於我們對原始音檔的格式做了修正，而修正的情形如下圖 2.1，所示

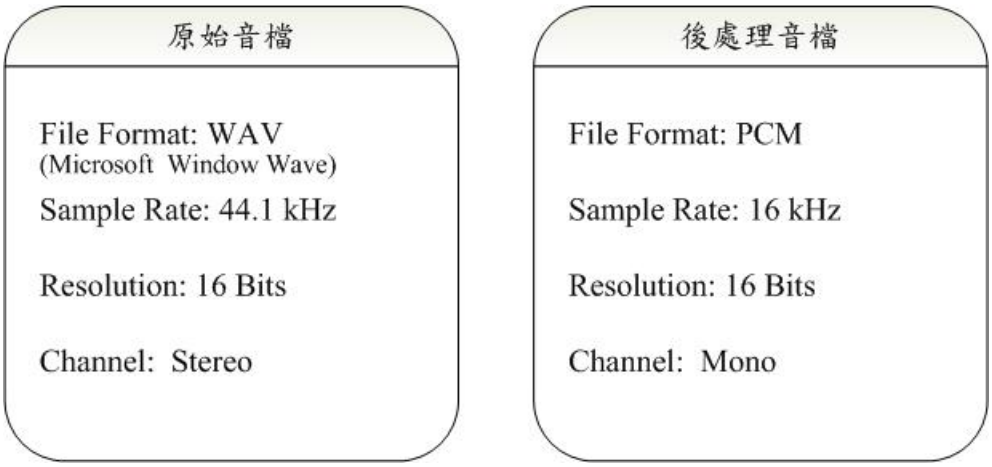


圖 2.1 音檔格式之設定

2.1.2 文字轉寫格式說明

一般的語料庫應該要包含錄製的音檔及音檔內容的文字轉寫(Transcription)這兩個部分；這一節，說明文字轉寫的方式。每一個不同的語料庫對於轉寫的方式會有很大的差異，其主要的差異，通常是文字轉寫的格式，就 MCDC 這個語料庫來講，其所使用的轉寫格式是一種標籤式的語言格式，這種語言格式是有點類似 XML 語法，在結構上大致是以一個 sub-turn 來當作一個單元，如圖 2.1、2.2 所示

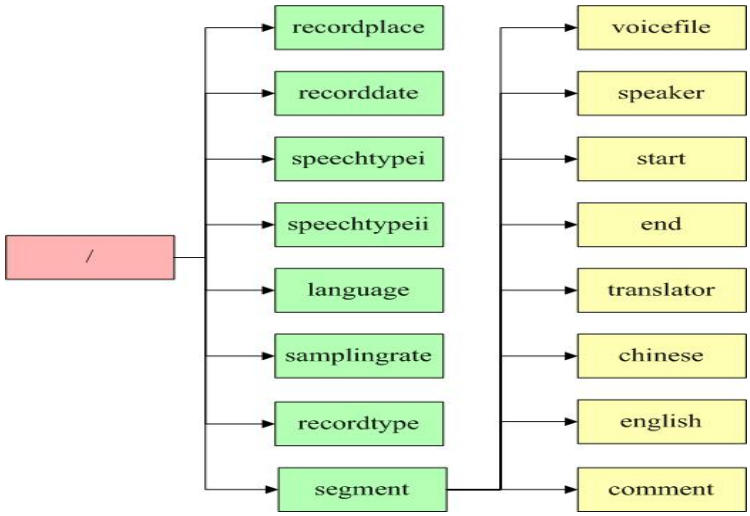


圖 2.2 MCDC 標籤階層圖

One Sub Turn Example

 (sub-turn 開始)
<voicefile>d:\分割完成的檔\stereo_05\mcdc-05-07.wav (音檔名稱)
<speaker>MISC-15-male-40 (發音者)
<start>072172 (音檔啟始時間)
<end>073843 (音檔結束時間)
<translator>Tanya (文字轉寫人)
<chinese> (內容文字標示)
<b particle>O</b particle>內勤<b particle>MHMM </b particle>
</chinese>
<english> (內容漢語併音標示)
O nei4 qin2 MHMM
</english>
<comment> (註解標示)
</comment>
 (sub-turn 結束)

圖 2.3 MCDC 一個 Sub-Turn 標籤範例

2.2 MCDC 口語之轉寫標示

由於 MCDC 語料庫所蒐集的語料是非常貼近日常生活的對話，也就是語音特性是屬於自發性的；因為 MCDC 語料庫的自發特性，這造成了其與朗讀式語料的差異，其中 MCDC 語料庫多出了許多口語現象，像笑聲、咳嗽聲、哈欠聲...等;而在 MCDC 語料庫的轉寫文字資訊上，就標示了許多這方面特性，如，拖長音、音節合併、發音偏差、不確定音...等（如附錄二）。由於，此語料庫有非常多的口語現象之標示[3]，因此，我們無法在這一章節全部說明，以下僅對本次研究，有興趣的口語標示，拿來作一個較為詳細的介紹。

1. 非語音現象(Paralinguistic Phenomena)

我們所處理的非語音現象就性質上,可以分成兩大類: 一、無伴隨語言內容之非語音但確定是由人所發出的聲音，包括吸氣聲、笑聲、吐氣聲、吞口水聲 …… 等，和其他口腔發出無法辨識的聲音等等。二、無伴隨語言內容之非語音且確定非人所發;以下是實際標示於語料庫的兩類例句。

(1)實例一（非語音、人聲）:

<b inhale>@ </b inhale>對我家就住捷運永春站那兒。

（inhale 吸氣聲）

(b)實例二（非語音、非人聲）：

<b noise in room>@ </b noise in room> NHN

（在說話者發出感嘆詞前，有一個敲到麥克風的聲音）

2. 不確定音／字（Uncertain）

我們所處理的不確定音/字現象，就性質上,可以分成兩大類: 一、文字轉寫人可以依據前後語意，可以大略猜測出的語意內容，但無法百分之百確定。二、文字轉寫人無法依據語意猜測出對應字詞，但可用漢語拼音清楚紀錄出其發音;以下是實際標示於語料庫的兩類例句。

(a)實例一（語意猜測字詞）：

至少我對我自己的車子<b uncertain>有 </b uncertain>有一個瞭解程度 BA。

(b)實例二（實際發音）：

我是不曉得在我<b uncertain>[x] </b uncertain>我離開的時候他們訂單都還不錯。

3. 語助詞(Marker)

發音者本身在對話中的慣用插語，這些習慣插語有其基本詞彙意義。但是，在對話中的習慣插語已不保有其原有的完整語意。例如，作用於口語中發音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;以下是實際標示於語料庫的例句。

實例（習慣插語）：

<b marker>NA</b marker>請問怎麼稱呼您。

4. 感嘆詞(Particle)

不具標準語意的感嘆詞，通常，這一類詞是用在表示回應或同意之類的情況。在對話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;四、其他感嘆詞，如嗯哼哼，以下是實際標示於語料庫的兩類例句。

(a)實例一（無相對應國字）:

<b particle>EI </b particle>你好。

(b)實例二（有相對應國字）:

真的<b particle>A </b particle><b particle>0 </b particle>謝謝。

5. 語言轉換(Code Switching)

當發音者使用漢語以外的語言（如閩南語、客家話、英語…等），即稱為語言轉換現象。以下是實際標示於語料庫的例句。

實例（英語）:

它外面沒有一個<b code switching><b English>marker </b English></b code switching>嗎?

2.3 MCDC 之相關統計

在 2.1.1 節裡，我們已說明了如何將雙聲道的音檔，轉換為單聲道的音檔，而這個動作就是將雙聲道的音檔切割成一個一個單聲道的音檔，而切割的依據是以一個 sub-turn 為單位。最後，我們可以很容易將整個 MCDC 語料庫中，每段對話 sub-turn 總和統計出來。如表 2.2 所示。

在文字轉寫的部分，我們可以依據 MCDC 語料庫中，文字轉寫的標籤（tag），取出我們想要處理的語音資訊－如 2.2 節所說明；因此，我們可以很容易的在 MCDC 語料庫中，擷取所有語者的對話內容，並且保留我們想要處理的資訊。最後，我們將對話中的語音，分成二大類：

一、正規性語音（411 syllable）

二、非正規性語音

包含了三小類，(1)Particles(含感歎詞及語助詞)、(2) Paralinguistic Phenomena(含

所有非語言現象)、(3) Uncertain (含不確定字／音)。

以下是實際標示於語料庫的例句：

<b inhale>@</b inhale>那用<b exhale>@</b exhale>我走過一次<b particle>A </b particle>。

其中 Paralinguistic Phenomena：@ 。

Particle：A。

411 Syllable: 那用,我走過一次。

最後，我們將這兩大類的語音，統計出來，如表 2.3 所示。

表 2.2 MCDC 語料庫之音檔長度與 Sub-Turn 數量統計

對話序號	音檔長度(分鐘)	Sub-Turn 數量
mcdc-01	61	867
mcdc-02	63	1094
mcdc-03	61	981
mcdc-05	63	865
mcdc-09	66	671
mcdc-10	54	513
mcdc-25	55	669
mcdc-26	46	833
總合	469	6488

表 2.3　MCDC 語料庫，正規性語音及非正規性語音之文字統計表

	正規性語音	非正規性語音		
	411 syllable	Particle	Paralinguistic Phenomena	Uncertain
字數	116,657	10,386	12,199	8,324
百分比	79%	7%	8.2%	5.6%
總字數	147,548			
Sub-Turn 數	6,488			

接下，再進一步針對 sub-turn 與音節數的關係，作一個 syllable 數的分佈圖，如下圖所示。

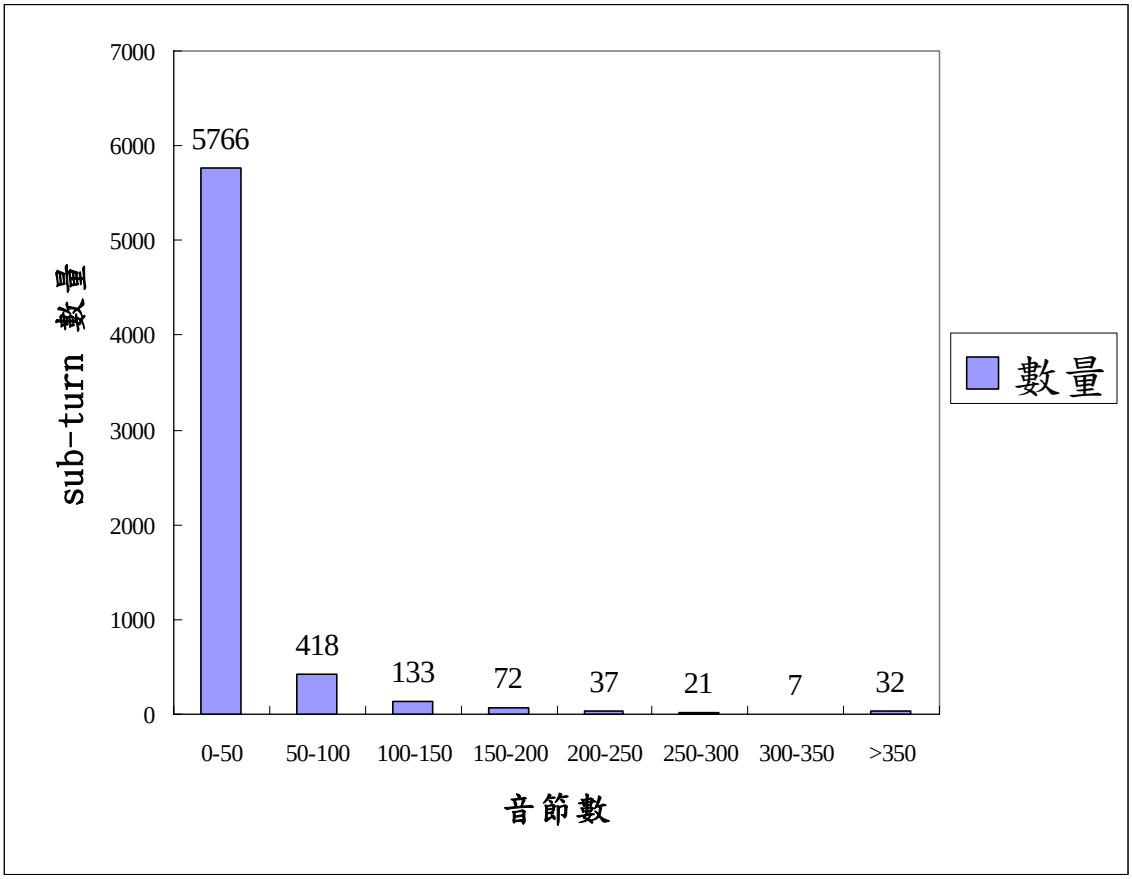


圖 2.3　MCDC 語料庫音節數的分佈圖

第三章 基本語音辨識系統之架構

本論文的基本語音辨識系統架構是建構在 HTK（HMM Tool Kit） v3.2.1[4]上。HTK 是由英國劍橋大學針對語音辨識，所開發出來的工具；這套工具主要是基於 HMM 用來實行語音辨識；在使用 HMM 實行大詞彙語音辨識上，已被證實能夠獲得不錯的效果；而 HMM 最大的特點是它利用訊號的統計特性去描述訊號。

3.1 系統環境參數設定

1. HTK 特徵向量求取原理說明

我們首先做一個假設－聲音訊號在幾毫秒內是 stationary。然後，再將聲音訊號做一連串的處理，處理順序如下[5]：

- 將聲音訊號切成若干的區塊，而每個區塊大小 32 ms 間隔 10 ms，相互重疊。
- 對聲音訊號做預強調的動作（即高頻放大），用來補償從口腔所引起的發散衰減現象。然後，再對每塊區塊取 Hamming Windows 作 smoothing 動作。
- 接下來就是 MFCCs(Mel-Frequency Cepstral Coefficients)參數求取流程，如圖 3.1 所示。

為了要做模型比對，所以我們就必需將語音訊號轉換為一串的聲學向量，而向量的計算方式就是每 10 ms 做一次平滑式對數頻譜。為了要改善模型比對的效能，所以，我們在頻譜上多加了梅爾頻率量度(Mel-frequency Scale)，接著再加上 DCT(Discrete Cosine Transform)。加上 DCT 用途就是將訊號作 de-correlation 動作，因此，可以對先前假設訊號是統計獨立的關係，有所改善。最後，再對參數(MFCC)做一、二次微分後，加到聲學向量以增加聲音訊號的動態資訊。以下是計算微分的公式[4]：

$$\Delta:\ d_t = \frac{\sum_{\theta=1}^{\Theta}\theta(c_{t+\theta}-c_{t-\theta})}{2\sum_{\theta=1}^{\Theta}\theta^2}$$

$$\Delta^2:\ d_t^2 = \frac{\sum_{\theta=1}^{\Theta}\theta(d_{t+\theta}-d_{t-\theta})}{2\sum_{\theta=1}^{\Theta}\theta^2}$$

Δ : delta coefficients
Δ^2 : acceleration coefficients
d_t : the first differential coefficient
d_t^2 : the second differential coefficient
c_t : MFCC coefficient
Θ : delta window size

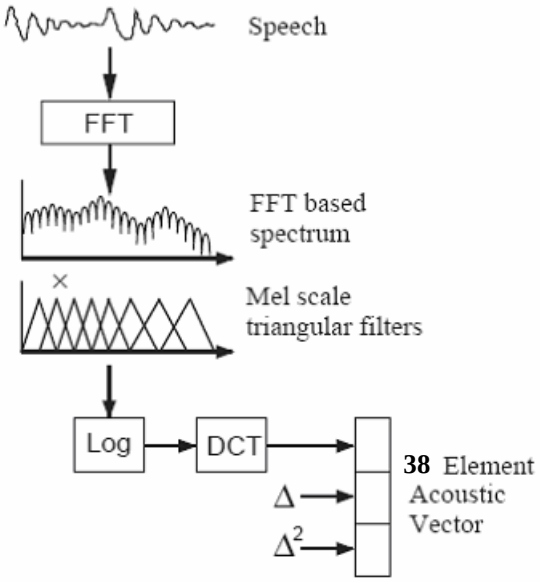


圖 3.1 抽取參數流程圖

2. **特徵向量求取(Feature Extraction)環境設定**

- (語音信號) 預強調： $1-0.97z^{-1}$
- (在做 FFT)視窗類別：漢明視窗 (Hamming Window)
- (在做 FFT)音框長度：32 ms
- (在做 FFT)音框偏移量：10 ms
- (在求 MFCC) Filter Band 頻率範圍：0~8kHz
- (在求 MFCC) 通道數：24
- 38 維之參數向量
  - 12 MFCCs
  - 12 MFCCs 之一階微分項(微分間距:2 音框)
  - 12 MFCCs 之二階微分項(微分間距:2 音框)

- 能量之一階微分項(微分間距:2 音框)
- 能量之二階微分項(微分間距:2 音框)

- 倒頻譜平均值消去法(Cepstrum Mean Subtraction，CMS)

3.2 基本聲學模型的建立

針對 MCDC 語料庫，我們想要為它建立一個聲學模型，我們的作法就是利用一個已經建構好的模型－朗讀式(read speech)語料（TCC300）所訓練的模型，去協助建構 MCDC 語料聲學模型。首先，我們考慮到 MCDC 語料裡有許多特殊模型，在 TCC300 聲學模型並不存在，因此，在 MCDC 裡有相似 411 的特殊模型，我們就將其退化至近似 TCC300 聲學模型，反之，沒有近似 TCC300 聲學模型，則視為 filler 模型（Garbage Model）。可是由於特殊語言現象中，Paralinguistic Phenomena 類別的某些現象，如表 3.1，在語料中出現的次數又非常多，所以我們必須用人工作一個切音的處理；其餘的音節則是利用 TCC300 聲學模型去做切音的動作，即可獲得具有時間資訊之 MCDC 語料轉寫。

表 3.1 人工標記時間資訊的 Paralinguistic phenomena

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

3.2.1 RCD(Right Context Dependent) HMM 模型的建立

由於 MCDC 語料拼音轉寫內容有時間切割資訊，所以，我們就可以利用併音的時間資訊(某一段切割時間)，來針對該段拼音作 Isolated Unit Training [4]的動作，其運算法如附錄三所示，而其整體流程如圖 3.2 所示。

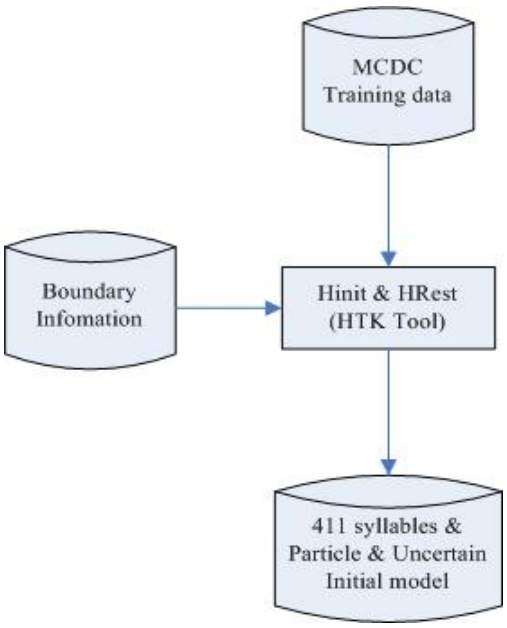


圖 3.2 MCDC 語料庫初始 RCDHMM 模型建構流程

而初始模型的建立原則，如下表 3.2 所示

表 3.2 HMM 模型建立方式

標註類別	狀態數量	Mixture Gaussian 的數量
RCD-Initial	3	依據 3.1 式去計算，若 Mixture 數大於 32 以 32 計之。
Final	5	
Paralinguistic	5	
Particle	3	
Uncertain	3	
Silence	3	64
SP(Tie to silence)	1	64

Mixture Gaussian 的數量計算公式如下式

$$N_{state\_mix} = Min\left(\left[\frac{N_{frame}}{50}\right], 32\right) \quad (3.1)$$

$$\begin{cases} N_{state\_mix}: \text{該狀態的混和高斯數} \\ N_{frame}: \text{該狀態的音框總數} \end{cases}$$

經由上述原則，我們就確立了要如何為拼音及特殊現象，選擇一個適當的 Mixture Gaussian HMM。但到目前為止，尚未說明要如何去選取所要模擬的拼音及特殊現象，基本上，我們以在 MCDC 語料文字轉寫內容出現 20 次當作一個基準，如果某拼音及特殊現象出現次數超過或等於這個基準，就為它們建立 Mixture Gaussian HMM，至於，初始模型，若是有相近 411 音的特殊現象，則以相近的 411 當作其初始模型，反之，若無相近 411 音的特殊現象，則以 filler 模型當作其初始模型。

最後，我們所建立出模型如下表-表 3.3 所示，其中 100 個 RCD Initial 及 40 個 Final 如 附錄四 所示。

表 3.3 完整的 HMM 模型數量

	411 syllable	Paralinguistic Phenomena	Particles	Uncertain
建出初始模型的數量	100 RCD-initial 40 final	11	40	76

然後，我們再根據上表的模型，再更新 MCDC 語料拼音轉寫內容，接著再做 Embedded (Baum-Welch) training（不需要時間切割資訊），訓練直到收斂為止。

3.2.2 CD(Context Dependent) HMM 模型的建立

我們的想法是想要從 Context-Independent HMM 去合成出比較精簡的 CD HMM，如果，我們不用這樣的想法去建構 CD HMM 的話，而直接去為每一個相關拼音建構一個 CD HMM，那麼 CDHMM 的參數數量將會變得非常龐大；因此，我們必須要對所有的 CD HMM 的狀態作一個合併(state tying)動作，對於狀態合併的方法，大致上可以區分為兩種策略－資料驅使的分類（Data-Driven Clustering）方法及樹狀結構基礎的分類(Tree-Based Clustering)方法[4]。

3.2.2.1 資料驅使的分類

我們一開始先假設所有狀態(states)都有屬自己的資料類別，然後在去計算資料類別間的相似度(距離)，如果，類別間的距離較近的話，我們就給予合併成一個類別，如圖 3.3 所示。

其演算法如下：

（1） 找出一對類別，而此對類別之間的距離，是符合小於類別合併距離臨界值之條件。

（2） 將此對類別的所有狀態 tie 在一起。

（3） 合併此對類別成一個新的類別。

（4） 重覆(1)－（3）直到達到所設定的類別數量臨界值。

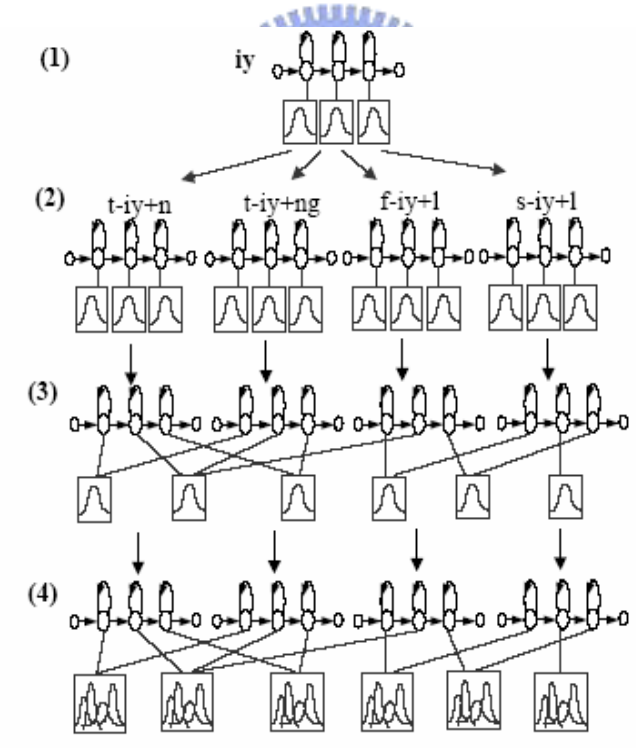


圖 3.3 Data-Driven Clustering 流程圖

3.2.2.2 樹狀結構基礎的分類

這種方法是結合了知識基礎（Knowledge-Based）的方法和資料驅使(Data-Driven)的方法，如圖 3.4 所示;

其演算法如下：

（1） 收集所有的狀態(states)放在樹的根節點。

（2） 找出能最大化資料相似度（Likelihood）的二元問題(Binary Question)

（3） 將資料分成兩個部分，其中一部分，答案為 Yes，另一部分，答案為No。

（4） 根據(3)我們會得到新的兩類，然後這兩類個別再回到(2)去進行分兩類的動作，直到增加的最大相似度低於某一個相似度變化的臨界值。

（5） 最後，我們再判斷所有的葉節點間，合併後的最大相似度，如果，減少的相似度低於某一個相似度變化的臨界值，則合併兩葉節點為一個新的葉節點。

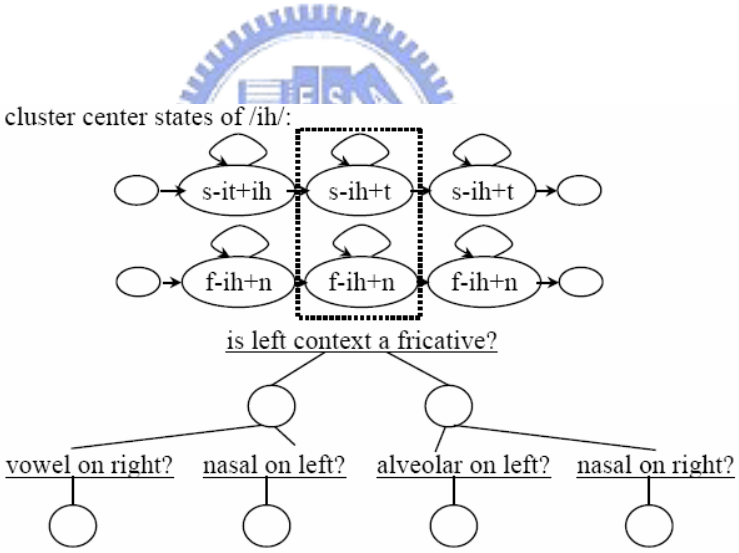


圖 3.4 Tree-Based Clustering 流程圖

1. 決策樹（Decision Tree）建立的概念

在本論文中，建構 CD HMM 是採用樹狀結構基礎的分類方法－決策樹的分類方法，而採用此種方法的理由是決策樹的分類方法,可以很容易的合成出所有可能的 CD HMM/SRU(Speech Recognition Unit)(依據所屬的類別)，即使有某一些 CD HMM/SRU 未包含在訓練語料中，而資料驅使的分類的方法則無法做到這一點。

在本論文裡，建立決策樹的工具是採用 HTK v3.2.1，以下先來介紹決策樹的基本架構：在 HTK 裡 CDHMM 的標準命名方式-“SRU1-SRU2+SRU3”指 SRU2 左邊接 SRU1，SRU2 右邊接 SRU3，以基於漢語拼音的中文音節為例：

q-yue+INULL_w

中文音節(Syl)通常會用 Initial(I)、Final(F)兩大類的 SRU 來組成，因此，可能出現組合，Syl+Syl+Syl=I。F+I。F+I。F（I-F+I 或 F-I+F），以上面這個例子來說，”yue”F-SRU 左邊接”q”I-SRU 而右邊接”INULL_w”I-SRU。接下來我們再針對相同 Central SRU 的 CDHMM 依據所建立的問題集，來作分群的動作，如下圖 3.5 所示，

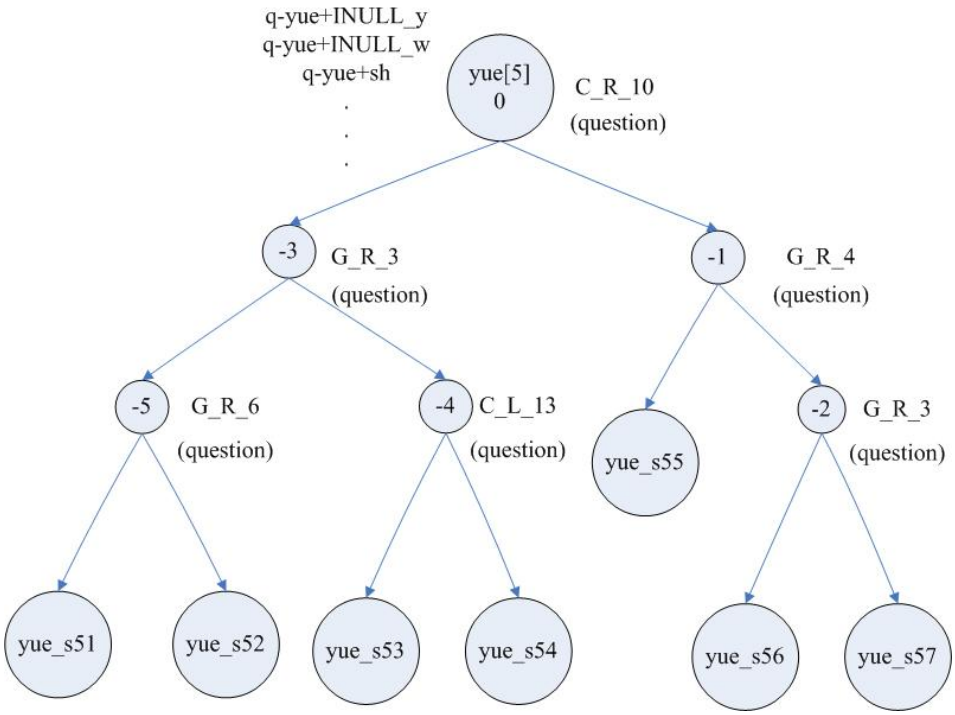


圖 3.5 Decision Tree 結構範例

2. 建立決策樹之原則

- 建立問題集－依據如附錄五 所示

例如： HTK 問題的標準格式

QS "V_R_1" {*+a, *+eh, *+o, *+e, *+er, *+FNULL1, *+FNULL2}

QS "V_L_1" {a-*, eh-*, o-*, e-*, er-*, FNULL1-*, FNULL2-*}

- 節點分裂之限制

  a. 葉節點的資料筆數。

  b. 最大相似度的改變量

- 節點分裂之標準（基於高斯分配）

  在分類的過程中，最大相似度的變化量是用來分裂節點的標準[6]，而最大相似度的定義如下式，

$$L(\mathbf{S}) = -\frac{1}{2}(\log[(2\pi)^n \left|\sum(\mathbf{S})\right|] + n)\sum_{s\in\mathbf{S}}\sum_{f\in\mathbf{F}}\gamma_s(\mathbf{o}_f) \qquad (3.2)$$

  γ_s : the a posteriori probability of the observed frame o_f being generated by state *s*.

  n : the dimension of the data

  $\sum S$: the pooled state variance

節點分裂（如圖 3.6 所示）的變化增益計算公式如下，

$$\Delta L_q = L(S_y(q)) + L(S_n(q)) - L(S) \qquad (3.3)$$

其中

L : log likehood

q : the question

S : the pooled state

S_y : fitted question sub-class

S_n : non-fitted question sub-class

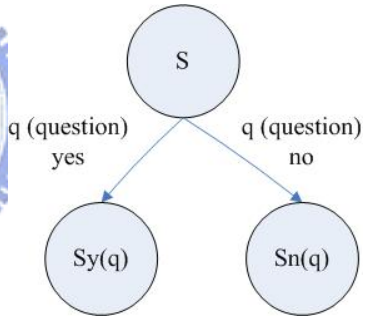


圖 3.6 節點分裂示意圖

第四章 基本語音辨識系統效能之分析

在這一章裡，我們要開始測試我們的基本系統的效能，並且對效能上做一個評估，而評估的方向，主要是針對音長（Duration）及 411 音節與特殊發音的混淆情形，作一個分析，以期待能找出改進辨識率的方法。

4.1 MCDC 語料庫使用之分配

在作語音辨識之前，對整個語料庫作分配是一個重要課題；要如何分配，主要是要看辨識系統要作何種辨識而定。而辨識的類型，大致上可分為語者獨立(Speaker Independent , SI)辨識、語者相依(Speaker Dependent, SD)辨識、多語者(Multi-Speaker)辨識。對於這些辨識系統的語料分配方式如下：

- 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。

- 語者相依辨識系統

  測試語料和訓練語料是同一語者。

- 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的句子是不一樣的。

在本論是採用多語者辨識系統，以下是語料庫的分配情形，

4.1.1 訓練語料

本論文是在 MCDC 8 段對話語料中，平均各選取 9/10 的 sub-turn 來當作訓練的語料，如此較能涵蓋所有語者的語音特性。然後，根據所有分配得到的訓練語料，來對音檔的時間、句數及文字作一個統計，列於表 4.1 及表 4.2。

表 4.1 訓練語料時間統計

句數（sub-turn）	時間（小時）
5,701	8.65

表 4.2 訓練語料文字資訊統計

	正規性語音	非正規性語音			
	411 syllable	Particles	Paralinguistic phenomena	Filler	Uncertain
字數	104,176	9,130	10,752	314	3,309
百分比	81.59	7.15	8.42	0.25	2.59
總字數	127,681(不含 silence)				
Sub-turn 數	5,701				

4.1.2 測試語料

我們已將每一段對話 9/10 的語料用於訓練模型，而其餘剩下來的約 1/10，而後再將 1/10 sub-turn 中，整句話裡無 411 音的語料除去，除去後，剩餘的語料即為實驗用的測試語料。然後，依據測試的語料，來對音檔的時間、句數及文字作一個統計，分別列於表 3.3 及表 3.4。

表 4.3 測試語料時間統計

句數（sub-turn）	時間（分鐘）
447	65.18

表 4.4 測試語料文字資訊統計

	正規性語音	非正規性語音			
	411 syllable	Particles	Paralinguistic phenomena	Filler	Uncertain
字數	14,997	936	1,205	47	517
百分比	84.72	5.29	6.81	0.27	2.92
總字數	17,702(不含 silence)				
Sub-turn 數	447				

由於我們對於原始 MCDC 的語料中做了調整，所以，表 4.2 及表 4.3 的總和，與表 2.3 不一致，其中，訓練語料－我們藉著強迫對齊的方法，移除了音檔錄製品質較差的部分；測試語料－我們將應答內未含 411 音節的部分移除。

4.2 辨識率的計算方法

對連續語音而言，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，因此對於辨認的結果，就產生三大類的錯誤－「替代型」(Substitution)錯誤、「插入型」(Insertion) 錯誤以及「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(Dynamic Programming，DP)，將已知的正確音節與辨認出來的音節做比對的動作，同時進行錯誤類型的認定，即可找到一條可得到最佳辨認率的路徑；至於其辨識情形可以區分為兩個測度－辨識率(Accuracy)及包含率(Correct)，以下是它們的計算公式：

$$\text{辨識率}=\frac{\text{正確音節數} - (\#\text{Sub} + \#\text{Ins} + \#\text{Del})}{\text{正確音節數}} \quad (4.1)$$

$$\text{包含率}=\frac{\text{正確音節數} - (\#\text{Sub} + \#\text{Del})}{\text{正確音節數}} \quad (4.2)$$

- #Sub: 替代的音節數
- #Ins: 插入的音節數
- #Del: 刪除的音節數

4.3 基本實驗

在上一章裡，我們已提出兩種 HMM 模型－ Final-Dependent Initial/Final 及 CD HMM，在這一章裡，我們需要比較這兩類模型辨識情形。由於，我們的語料庫是自發性語料，所以它有別朗讀式語料，多了特殊的語音現象（第二章提及）。對於模擬特殊現象，我們是採用 CI(Context Independent)HMM，其餘的 411 音節 IF 模型則依實設定採用 RCD 或 CD HMM。最後，我們將會做兩個實驗來探討 RCD、CD HMM 的效能。

4.3.1 實驗 1(RCD)

本實驗的目的，最主要是要針對使用 RCD HMM，去描述此種模型對正規性語音(411 音節)及非正規性語音（特殊現象發音）的鑑別能力。

4.3.1.1 使用 RCD-Intial+Final 之辨識結果

RCD HMM 的模型建構方式是根據 3.2.1 節及而其辨識的環境是依據 3.1 節，而最後，其辨識能力，如下表 4.5 所示，

表 4.5 RCD HMM 辨識結果

	Corr(%)	Acc(%)	Sub(%)	Del(%)	Ins(%)
Sum/Avg	47.18	41.97	40.27	12.5	5.20

4.3.1.2 實驗 1 之錯誤分析

實驗一已列出其基本的辨識率，所以，我們可以拿這個結果和 Read Speech 語料庫(TCC-300)的辨識結果[7]做比較，TCC-300 的基本辨識率為 67.5%，比實驗一的基本辨識率高出約 25.63%，這樣的情形，我們是可以預知的，因為自發性的語料中，包含了許多語者口語特殊的現象，這也就造成 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 4.6，也就是對於基本系統之混淆矩陣（Confusion Matrix）做細部的分析，我們依 HMM 模型的特性分成 Paralinguistic Phenomena、Particle、Uncertain、411 syllables 這四大類，以便我們分析高錯誤率的原因，我們將整過後的結果，列於下表 4.6。

表 4.6 RCD HMM 辨識結果的 Confusion Matrix 分析

辨識結果 / 音節類別	Paralinguistic	Particle	Uncertain	411
Paralinguistic	60.54%	3.6%	2.55%	13.74%
Particle	4.17%	49.46%	2.24%	32.26%
Uncertain	2.71%	5.43%	33.91%	48.25%
411	1.24%	1.51%	1.94%	83.20%

1. 替代型錯誤分析

在這一節中，我們將對非 411 音節型態被辨認為 411 音節部分（替代型錯誤）來做一個探討；以下將以順序為 Paralinguistic Phenomena、Uncertain、Particle 取代型錯誤來探討分析。

(a) Paralinguistic Phenomena 取代型錯誤分析

由表 4.4 可看出 Paralinguistic Phenomena 大約有 13%的音是被辨識為 411 音，直覺上，Paralinguistic Phenomena 與 411 音的特性應該是有蠻大的差距的，可是在誤辨為 411 音的比例竟然有 13.74% 之多，於是我們就想要了解為何會這種現象，因此，我們在表 4.7 列出 Paralinguistic Phenomena 中較容易辨識為 411 音的前三名。

表 4.7 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 4.7 中三種音節所代表的意義，(1)Unrecognizable Speech Sound 代表的確是屬人所發出來的語音，但標記員無法辨認其文字意義；(2)Unrecognizable Non-Speech Sound 代表其他由人發出來的非語音，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，而這三種聲音總共佔了 Paralinguistic Phenomena 辨識為 411 音錯誤的 54%，它們都有一個特徵－無法確定的聲音；所以它們並沒有固定特性，因此訓練出來的模型，其變異量必定大，這應該就是造成比起其他 Paralinguistic Phenomena 易造成錯誤的原因。

(b) Uncertain 取代型錯誤分析

由表 4.6 可看出互相辨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，於是我們將參考的轉寫內容（即比對的答案）中是 Uncertain 的字，必須將其退化為相近的 411 音，其結果列於表 4.8

表 4.8 Uncertain 退化至相近 411 音之辨識率

	Corr(%)	Acc(%)	Sub(%)	Del(%)	Ins(%)
Sum/Avg	47.41	42.55	39.48	13.1	4.86

由表 4.5 和表 4.8 比較，我們可知道 Uncertain 退化至相近的 411 後，辨識率大約提升 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（如表 4.9）來作比較。

表 4.9 Uncertain 退化至相近 411 音 Confusion Matrix 分析

辨識結果 / 音節類別	Paralinguistic	Particle	411
paralinguistic	61.2%	4.07%	14.22%
Particle	4.27%	50.10%	33.33%
411	1.42%	1.76%	84.27%

由表 4.9 得知，將 Uncertain 退化至相近的 411 音節後，411 音節被辨認成 411 音節為 84.27%，相較於表 4.6 的 83.20%提升了 1.07%，而且 Uncertain 只佔了所有音節總數的 2.92%，因此我們由表 4.5 和表 4.8 比較，Uncertain 退化至相近的 411 之辨識率大約只提升 0.5%，推論出，的確 Uncertain 基本上與 411 音節是具有相同性質的。

(c) **Particle 取代型錯誤分析**

根據表 4.6，我們可以發現 411 音與 Particle 的互相辨識情形也是很嚴重，尤其是 Particle 的音常常會辨識為 411 音，這種情形其實是蠻正常的，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。

為了解決 411 音與 Particle 容易混淆的情況，我們希望藉由音長(Duration)上的差異來區別 411 音及 Particle;但是，由於我們目前的辨識工具 HTK，無法建立音長模型，因此，我們必須用其它的方法求出 Particle 及與其相近的 411 音之間的差異；最後，我們決定由 Particle 及與其最相近的 411 音的 Duration 分佈，定義一個臨界值。藉由此臨界值，來對參考轉寫內容(辨識答案)做一個適當的修正。

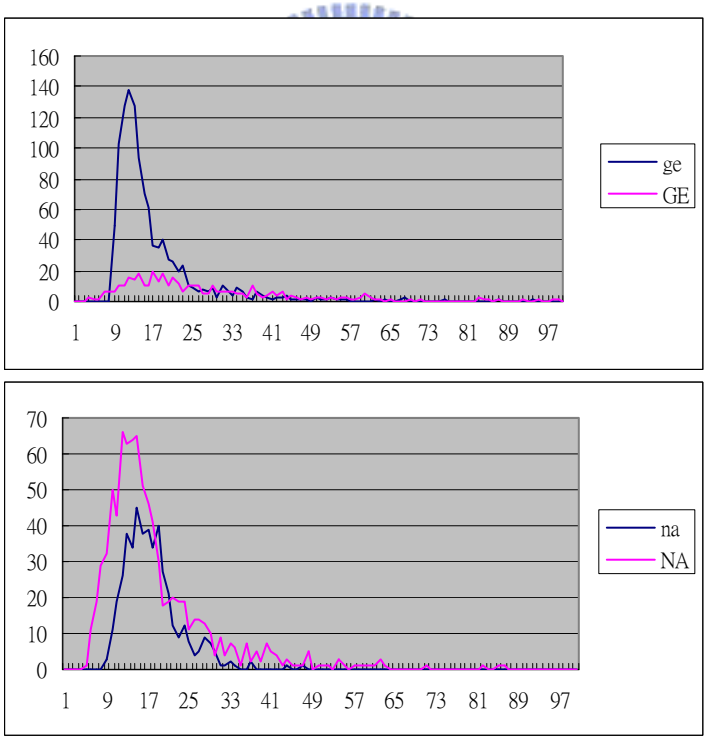


圖 4.1 Particle 及其相近 411 音之音長分佈圖

由上圖可以看得出來，原本期望藉由音長來區分出 Particle 及 411 音，是件非常困難的事情，因為其音長分佈非常的像，我們很難以一個臨界值來決定參考的轉寫內容（辨識答案）是否要更改。因此，想利用 Particle 及其相近 411 音，在音長的特性，去區別它們是一件非

常難的事情，且因為具有相近 411 音的 Particle，大都是虛詞(Function Word)，所以，在語音信號上的表現應該較為相似。

由 2.2 節 MCDC 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 4.10：Particle 分類後之辨識率分析

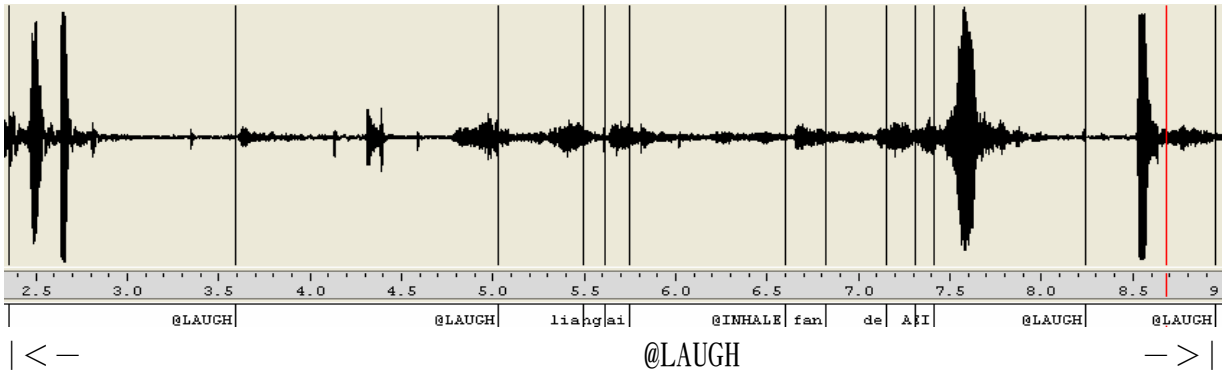
Particle 類別	被辨識為 411 音	辨識為 411 音且為相近 411 音	辨識為同類 Particle	Testing Data 中之音節數
有相對應國字	32.71%	11.56%	47.35%	813
無相對應國字	18.6%	4.85%	69.8%	123

由表 4.10 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。

2. 插入型錯誤分析

由於 MCDC 語料庫所採用的錄音方式是用兩支麥克風錄製，因此當發生雙方同時間講話的情況時，彼此的聲音很容易會錄進彼此的麥克風中，造成串音（Crosstalk）的現象。因此，如果沒有有效的隔離語者錄音時的通道，將會造成嚴重的串音；很不幸地，MCDC 語料庫的串音現象相當的嚴重，所以，這也就容易造成文字轉寫內容的錯誤，進而造成插入型錯誤。下面列出幾種串音造成插入型錯誤的例子：

- 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，因此造成原音檔的文字轉寫內容只有笑聲，但實際上，中間卻穿插了許多人耳可辨識的微弱語音，而這些將造成辨識的錯誤，如圖 4.2 所示。



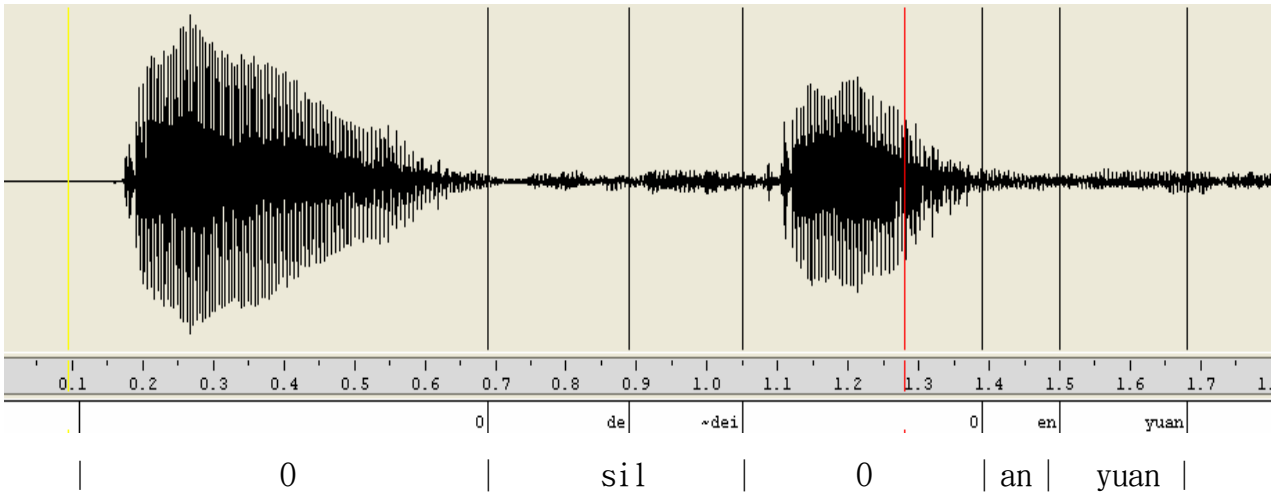
音檔下面第一排文字為辨識出來的結果。

音檔下面第二排文字為實際音檔之文字轉寫內容。

圖 4.2　串音現象範例(一)

上圖中表示辨識出了哪些音節及其切割位置，我們可看出此音檔的辨識結果有非常多的音節，但是事實上其文字轉寫內只有一個音節－笑聲，所以，這一句話將會造非常多個插入型錯誤。

- 此檔的情況為說話者和對方同時講話時的串音現象



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實際音檔之文字轉寫內容。

圖 4.3　串音現象範例(二)

上圖所顯示的 0 與 0 之間辨識出來的 de 和~dei 這兩個音節，於文字轉寫內容中是沒有的，因為這兩個音節是由另一個語者所發出的，這幾個音節聽起來是蠻清楚的，因此這也是由串音所造成之刪除型錯誤。

3. 刪除型錯誤分析

對所有訓練語料 Force Alignment，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，快速語料之錯誤率相對於正常速度是較高的，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音節合併現象，合併現象有三：

(a) 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。

(b) 音節雖無短少，但卻都連在一起，難以切割。

(c) 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音節耦合（Syllable 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 4.11 易發生音節耦合之音節及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	13.9%	31.7%
一	一個	92	525	17.5%	42.0%
的	是的	87	453	19.2%	42.3%
我	我們	55	465	11.8%	34.5%
他	他們	43	276	15.5%	32.0%
對	對阿	40	308	12.9%	41.4%
這	這樣	36	143	25.1%	34.5%
有	沒有	35	338	10.3%	34.0%

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。

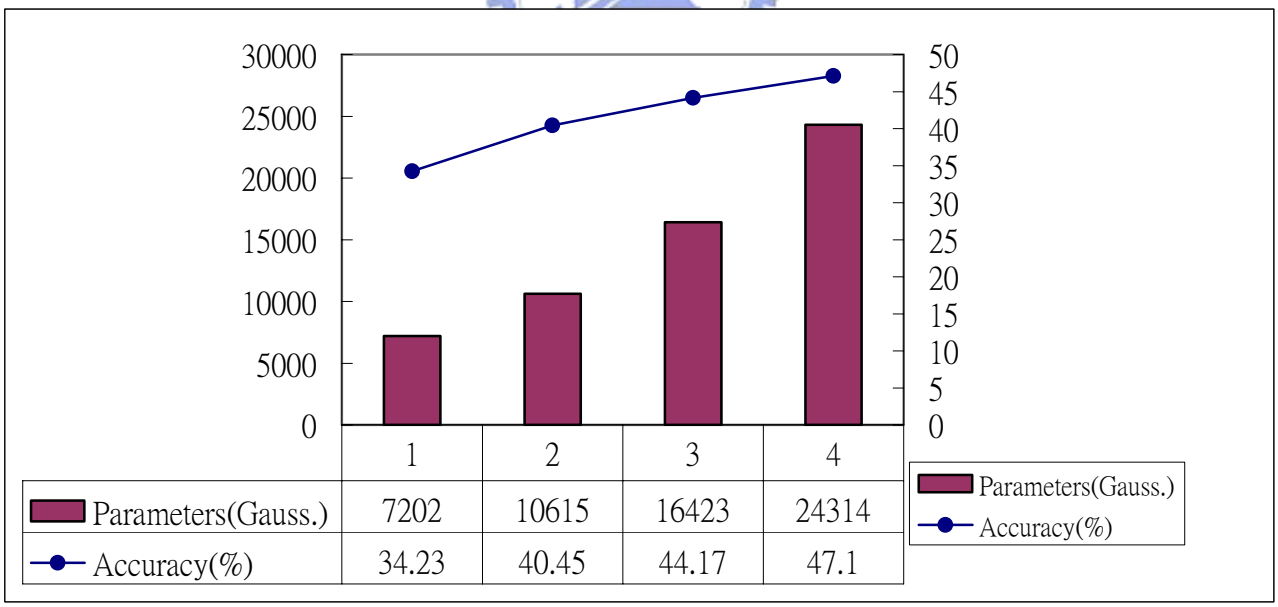
4.3.2 實驗 2(CD HMM)

在這一節裡，我們想要以 CD HMM 模型來當作辨識系統的核心，試圖想要改善辨識系統對 411 音節的辨識率，設定及建構方式如 3.2.2 節所示，其中分裂節最大相似度變量設為 200，Word Penalty 設為 20，而下圖為其辨識結果，

表 4.12　CD HMM 參數量 v.s.正確率

參數量（Gauss.）	7202	10615	16423	24314
正確率（Acc.）	34.23%	40.45%	44.17%	47.1%
基本音節參數量（Gauss.）	3227	6640	12488	20339
基本音節狀態最大混合高斯數(Gauss.)	1	2	4	8
Note ： (1)非基本音節參數量（Gaussians）:3975 (2)所有相同 Central SRU 的轉移機率皆 tying 在一起				

圖 4.4　CD HMM 參數量 v.s.正確率趨勢圖



上圖參數量在 16423 時的辨識率已超越，RCD HMM（參數量為 15532）的辨識率(如表 4.5)41.97%。

4.3.2.1 增進 CD HMM 辨識率的方法

底下我們將以修改原始 CD HMM 模型(Prototype)，來改善辨識率，

1. Untying Same Central SRU CD HMM 轉移機率的方法

如果，我們考慮不要將所有相同 Central SRU 的轉移機率皆 tying 在一起，所得到的辨識率為下表所示

表 4.13 Untying 轉移機率 CD HMM (參數量 24958)的辨識結果

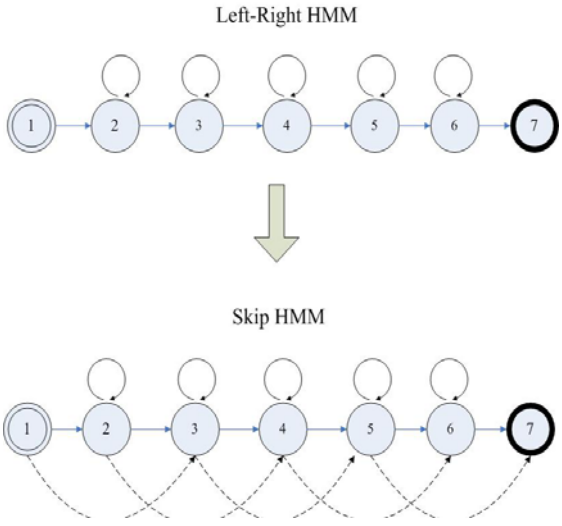
	Corr(%)	Acc(%)	Sub(%)	Del(%)	Ins(%)
Sum/Avg	53.90	48.63	34.09	12.02	5.26

比較表 4.10 與表 4.11 得知轉移機率 Untying 的 CD HMM 辨識率高於 Tying CD HMM 1.53 %。

2. 更改原始 HMM 模型的方法

接下來我們考慮改變 HMM 原始模型，想法如下(如圖 4.5)，

圖 4.5 更改原始 5-State HMM 模型(左至右)至新 HMM 模型(Skip HMM)



至於何時該跳躍狀態，我們在此定出一個標準－以被跳躍狀態的平均音長做為量度。我們假設某 5 個狀態的轉移矩陣為

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} \\ a_{51} & a_{52} & a_{53} & a_{54} & a_{55} \end{pmatrix}$$

計算每個狀態平均音長的公式[8]如下

$$\bar{d}_i = \frac{1}{a_{ii}} \quad , for\ i = 1 \cdots 5$$

$\bar{d}_i$:第 i 狀態平均音長

我們首先定下一個被跳躍狀態臨界值，根據每個狀態的平均音長是否小於這個臨界值來決定要不要跳躍此狀態，如果小於這個臨界值，則跳躍，反之，則不跳躍此狀態。然後，我們再利用設定完之後的 skip one state HMM 來做一個 Re-train 的動作，再來分析被 skip 的狀態數量來作一個比較，如 圖 4.6 所示，由 圖 4.6 得知實際有發生跳躍狀態是相當少的，所以，我們可以期待辨識率改善的程度將會非常少。

利用臨界值為 1.75 的 Skip CD HMM 去做辨識，而得到辨識率如下表所示，

表 4.14　Skip One State CD HMM 的辨識結果

	Corr(%)	Acc(%)	Sub(%)	Del(%)	Ins(%)
Sum/Avg	55.28	48.99	33.88	10.84	6.29

所以，比較表 4.11 與表 4.12，發現 Skip CD HMM 辨識率僅上升 0.36%。

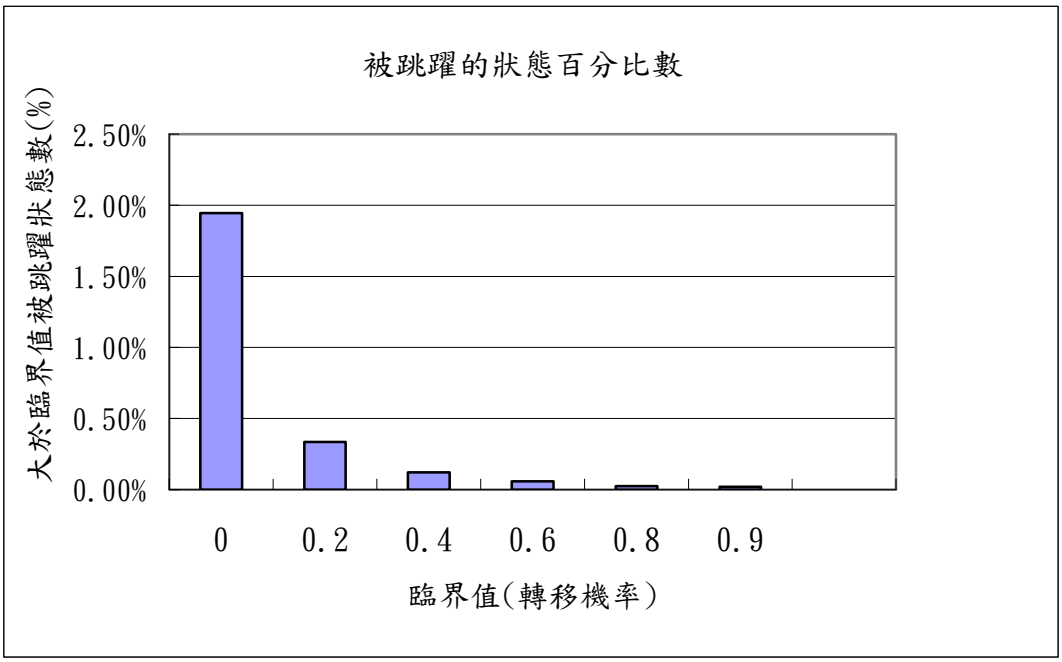


圖 4.6　被跳躍 Skip One State CD HMM 的狀態數統計圖

座標軸說明－
縱軸：(Re-train 後跳躍的狀態數(>臨界值))／被設定跳躍的狀態數
橫軸：某一臨界值〔HMM Transition Prob.〕

4.3.2.2 實驗 2 的錯誤分析

我們以實驗 2，最後所得到的（Skip）CD HMM，去建構混淆矩陣（如表 4.13）來做錯誤的分析。

表 4.15　CD HMM 辨識結果之混淆矩陣

辨識結果 / 音節類別	411	Paralinguistic	Particles	Uncertain	Filler
411	96.60%	1.30%	1.03%	1.03%	0.01%
Paralinguistic	8.28%	87.72%	2.69%	1.10%	0.20%
Particles	35.84%	0.03%	59.32%	0.02%	0.00%
Uncertain	75.80%	3.61%	3.61%	16.99%	0.00%
Filler	80.00%	0.18%	2.50%	0.00%	0.00%

由上表和表 4.2，我們可以發現 411 音節被辨識為 411 音節的百比率已由 83.20 增加至 96.60，由此可知 CD HMM，的確對 411 音節的辨識有很大的助益，而像 filler 被辨認為自己的機率為零，事實上，這是不可靠的數據，因為在測試語料中，它僅占約 0.25%，數量非常少。其餘的音節模型的錯誤分析與 RCD HMM 錯誤分析（如 4.3.1.2）雷同。

接下來我們再來看音節的音長對最大相似度的關係，在這裏我們僅舉出幾個在訓練語料出現機率的音節來看，下圖 4.7、4.8 所示

圖4.7 /shi/ (是)之音長及其最大相似度分佈圖

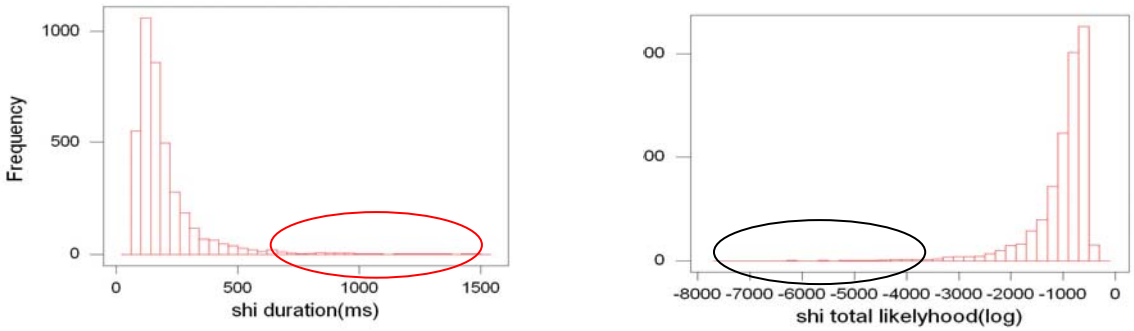
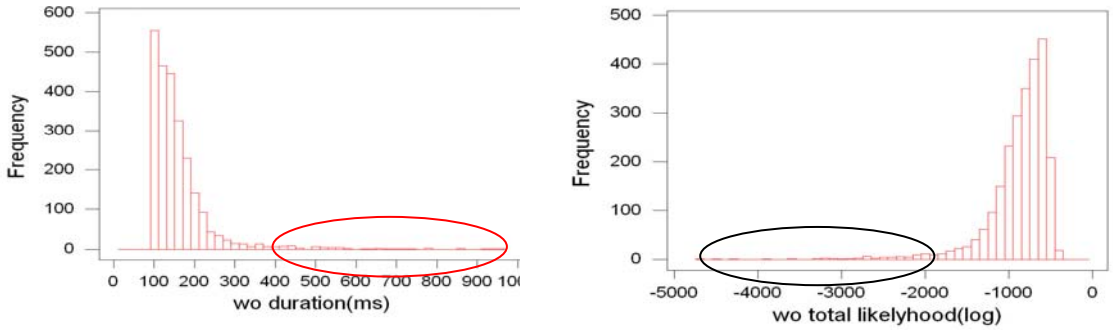


圖 4.8 /wo/(我)的音長及其最大似相度分佈圖



由上圖得知，如果音節的音長比大部分的音長來得長的話，將造成其最大相似度的降低，這可視為某程度的群聚效應。

由於，我們不針對音長的部分來做深入的研究，因此，在此我們僅說明在我們語料中有如此的現象而已。

第五章 加入語言模型至基本語音辨識系統

在第三、四章我們已經介紹了聲學模型的比對方式，而聲學模型的比對是屬於較低層次的作法，因為其未包含任何有關的語音資訊，所以，一般而言，一部較佳的中文辨識器能接受一連串的聲音訊號輸入，並輸出較為合理的口語句子。而其實際上的作法，就是它會根據一部有限詞彙的詞典去猜測聲音訊號，比較可能是什麼詞彙，最後就輸出有可能出現的句子。

上述所說的猜測的動作，即是依據語言的統計模型；所謂的語言模型－具有其獨特的文法規則，及語言特性，所求得一個機率模型，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，通常能提高辨識系統的辨識率。

在本章將建立兩種不同性質的語言模型，一種是內容較為廣泛及文字語料較多的，所訓練而得的語言模型（通用語言模型），另一種是由特定領域的語料訓練而得的（MCDC 語言模型）；由於 MCDC 語料庫文字語料太少，所以必須用通用語言模型來協助 MCDC 語言模型已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，

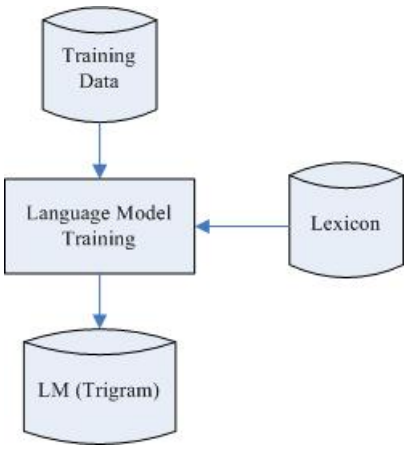


圖 5.1 LM 訓練流程圖

5.1.1 訓練語料及詞典(lexicon)

建立語言模型必須要準備的兩樣資料－訓練語料及詞典，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，下面將稱之為通用語料庫。

(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，是由交大電信所語音實驗室的詞典和台灣師大資工所做聯集及後處理動作而得到新的詞典，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 詞典中之詞長分佈

詞長	1	2	3	4	5	6	7	8	總合
數量	9,821	34,188	9,452	5,912	231	128	33	22	59,787
百分比	16.43	57.18	15.81	9.89	0.39	0.21	0.06	0.04	100%

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 通用語料庫之詞數表

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

表 5.3 MCDC 語料庫之詞數表

訓練語料	詞數（Word）	字數(Character)
MCDC 語料庫	64,720	102,217

5.1.2 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、Bigram、Trigram 的機率，分別為 $P(w_i \mid w_{i-1})$、$P(w_i \mid w_{i-1,} w_{i-2})$ 及 $P(w_i \mid w_{i-1}, w_{i-2}, w_{i-3})$，下面將介紹，求取 n-gram 機率的方法，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。

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

其中每個 n-gram 的機率如下式所示：

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

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

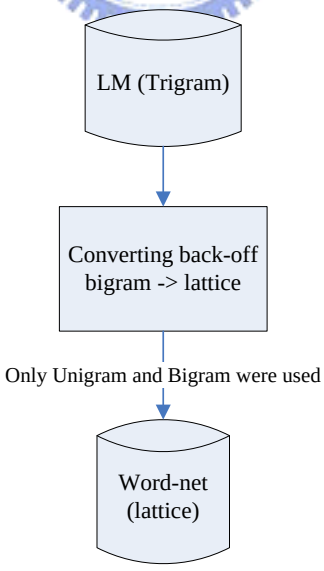


圖 5.2 LM 轉 Word-Net 之流程圖

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六(此為經驗值)，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，本節將做二個實驗，以討論加入二種 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通用語料庫訓練而得的，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

	Corr(%)	Acc(%)	Sub(%)	Del(%)	Ins(%)
Sum/Avg	58.08	53.83	28.56	13.37	4.25

5.2.2 實驗二

由於 MCDC 語料庫的文字語料太少，所以，我們可以預期到，本實驗必須採用語料較多的語音模型(General LM)來協助 MCDC 的語言模型而這種協助的行為，稱作語言模型調適（Language Model Adaptation），以下將介紹如何作語言模型調適，

以一個 Trigram 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 Bigram 條件機率，P_{Gen} 是原本 General LM 的 Trigram 機率以及 P_{MCDC} 是在 MCDC 訓練語料中的 Trigram 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

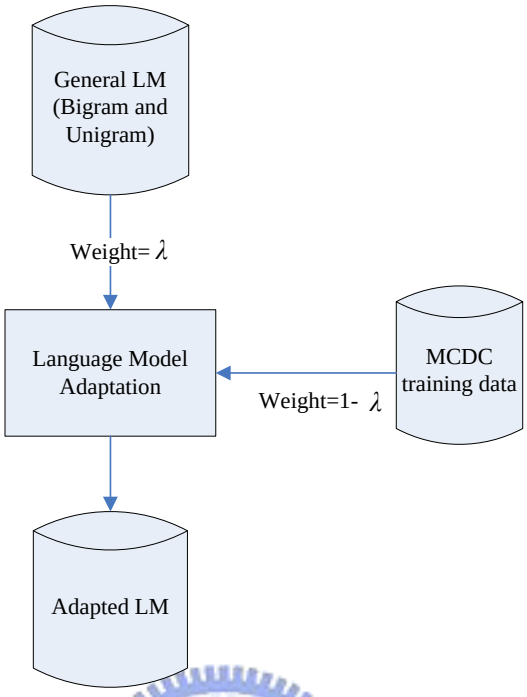


圖 5.3 語言模型調適流程圖

在這個實驗中，我們將調整 weight 為 0.9，0.8，0.6，0.4，0.2，試圖找到最佳的語言模型。而其相對的辨識結果列於下表 5.5 及圖 5.4

表 5.5　不同 weight 調適語言模型之辨識結果

λ	0.9	0.8	0.6	0.4	0.2
Acc.(%)	55.12	56.40	55.72	55.96	55.49

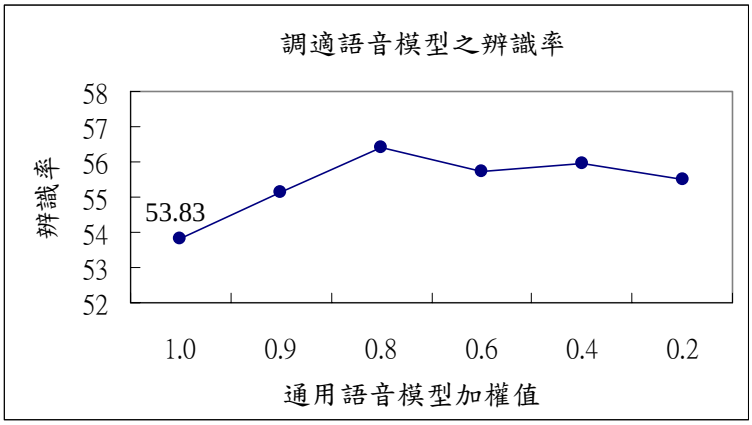


圖 5.4　不同 weight 調適語言模型不同之辨識結果之趨勢圖

5.2.3 實驗分析

對於實驗一、實驗二及結果之分析我們將細列於下：

1、相較於表 4.11 的辨識率，實驗一（General LM）為 53.83%，增加了 4.84%，增加的部分應該只有 411 音節，因為通用語言模型，並未包含特殊音節(Particles、Uncertain..等)。

2、實驗二(MCDC LM+ General LM)為 56.4%，相較於實驗一，增加了 2.57%，增加的部分，大部分的貢獻是因為多了特殊音節，和語者口語風格。

由上述分析，我們可以得到一個就結論－口語性語料與文字資料的語言模型還是有明顯的差異存在。

第六章 語音辨識之發音變異處理

6.1 發音變異之處理方式

相較於朗讀式自動語音辨識器而言，其音節辨識能力，通常會比自發性語音辨識器來得高。而會有這種情形的主要原因就是在於自發性語音，有許多不同發音類型，大致可以分為兩大層面－聲學層面、語言層面。

在聲學層面上，通常包含了音素結構的變化（插入式、刪除式、取代式）和發音方式的變化(鼻音化、央化、兒化…)，這些變化現象發生的原因－可能是語者不同的說話速度、心情、韻律…等。其它的現象還有拖長音、呼吸聲、笑聲、咳嗽聲、吞口水聲、清喉嚨聲、非語音聲、雜音…等[3]，這些都是造成 ASR 辨識上的困難。

在語言層面上，包含了許多口語現象－重覆詞語、更正詞語、口吃…等。這些現象發生的原因－通常是在日常生活對話中，語者正在思考所引起的。因此，在這些口語現象的影響之下，造成了語言模型(N-Gram)統計上的困難。

本篇論文僅針對聲學模型部分來研究發音變異的情況;由於中研院曾淑娟 博士的研究[9]發現音節合併與 Phone Structure 並無絕對的關連性存在，反而，是在口語中常見的虛詞影響較大，因此，即使使用 Context-Dependent 音學模型也無法有效的描述此種語音變異現象；基於這樣的想法，我們必須研究新的聲學變異模型，來處理此種語音變異現象；而至於如何有效的捕捉發音變異的特性及建構聲學變異模型，其方法不外乎有兩種，大概可以區分為兩大類－知識基礎的方法(Knowledge-Based Method)和資料驅使的方法(Data-Driven Method)[10]；（1）知識為基礎的方法－發音資訊來自於語言學知識或者手寫式字典，例如，在文字轉寫(Transcription)上，對於標音，會有兩個不同層級的標音方式[11]－標準式(Canonical-Form)標音，依照標準字典，所標的拼音；表層式（Surface-Form）標音，依據語言學家實際上所聽到的聲音，所標的拼音。有了這兩層標音，我們就能依據它們之間的差異，去建構發音偏差模型。(2)資料驅使的方式－發音資訊是以某些方法自資料中綷取出來；然後再根據資料間的關連性，找出單筆資料的發音特性，然後再進一步去建構發音偏差模型。

由於，知識為基礎的方法，需要語言學知識及大量的人力，所以我們在有限的資源下，採用了資料驅使的方式來作為研究發音變異的方式。在這裡我們所提出來的想法，是針對中文 411 音節來當作建構發音偏差的對象；對於中文音節而言，每個音節可以被分成兩個語音辨識單元（Speech Recognition Unit, SRU）Initial 和 Final，在本論文 Initial 使用 3 個狀態 HMM 去模擬它，而 Final 使用 5 個狀態 HMM 去模擬它，因此，一個中文音節，就使用 8 個狀態。然後，我們在就音節每個狀態，去計算其觀測向量(Observation Vector)的平均值（向量），然後，用每個狀態的平均值，構成一個超向量(Super-Vector)。

我們依據上述的概念，為訓練語料中的每個中文音節，建構出個別音節的超向量；然後再將相同的 411 音節，收集為一類，進而得到若干個超向量；我們在依據這些超向量的關連性，來進行資料的分羣的動作。最後，就可以將音節分成若干類，然後，我們在根據各類，去為它們建立發音變異的 HMM 模型。

6.2 使用 KPCA 建構發音變異模型

前一節我們已經大略了說明了發音變異模型的建構方式。建構發音變異模型的目的簡單來說就是找出標準式標音以外的發音（表層式標音），在這裡我們提出了使用 KPCA 方析的方式，來協助我們將被標示為相同節音的訓練語料，再做進一步的分類，找出表層式標音。在這一節我們將要說明如何將我們的訓練語料，應用到 KPCA 架構上，去進行分類的動作，而我們將分以下幾個部分來說明－KPCA 資料點之計算方式、KPCA 之基本原理介紹、音節分羣之方式。

6.2.1 KPCA 資料點之計算方式

對於 KPCA 架構之輸入資料點，我們試圖以音節為單位作為 KPCA 之輸入資料點。而選擇音節最主要的原因－由於自發性語料中，常常會發生嚴重或輕微的音節合併（Syllable Contraction）現象，因此，為了觀察上及分析上的方便，採取了以音節為單位的方式。接下來，我們來大略說明獲得資料點的作法，

（1） 抽取音節特徵參數－從訓練語料中抽取固定大小之音節特徵向量（Feature Vector），而實際的作法就是使用強迫對齊(Force-Alignment)的方法找出相對應的音節位置。

（2） 針對每個音節（Initial＆Final）去計算其每個狀態的平均值（12MFCC 之向量），最後，依序將每個狀態的平均值合併為一個比較大的向量，我們稱之為超向量(96維，8 個狀態的平均值)。

現在，我們就實際取得超向量的作法，來作一個詳細的說明－

假設我們在句子觀察到的特徵向量序列為 $O=\left[o_1,o_2,\cdots\right]$，而其第 k 個音節之超向量為 $x_k=\left[m_{k,1},m_{k,2},\cdots,m_{k,8}\right]^T$，

$$m_{k,i}=\frac{\sum_t P(s_t=k,q_t=i|O,\lambda)\cdot o_t}{\sum_t P(s_t=k,q_t=i|O,\lambda)} \qquad (6.1)$$

其中

s_t 和 q_t：分別代表音節及狀態在第 t 個音框的標註。

λ：為所有音節的 HMM 模型。

6.2.2 KPCA 之基本原理介紹

在上一節中，我們已知如何獲得輸入的資料點，接下來本節將就 KPCA 方析[12]的實際作法來做介紹－

首先，先介紹 PCA 的作法，

假設我們從訓練語料中，獲得 K 個音節 $\{x_k;k=1,\cdots,K\}$，來當作輸入的資料點，接下來我們就計算其 Covariance Matrix，

$$C=\frac{1}{K}\sum_k x_k x_k^T \qquad (6.2)$$

其中

我們假設 $\sum_k x_k=0$

然後，我們再去解它的特徵方程式，找出以下的關係式，

$$\lambda \mathbf{v} = C\mathbf{v} \qquad (6.3)$$

其中

λ 和 $\mathbf{v}$ 分別代表 C 的 eigenvalue 及相對應的 eigenvector

接下來，我們依據 PCA 的作法，延伸出 KPCA 的作法

假設現在有另一個內積空間 $\mathbb{F}$，它具有以下非線性的映射關係，

$$\Phi : \mathbb{R}^N \to \mathbb{F}, x \mapsto \mathbf{X} \qquad (6.4)$$

其中

我們假設 $\sum_k \Phi(x_k) = 0$

然後，我們在另一個內積空間 $\mathbb{F}$，找出其 Covariance Matrix

$$\overline{C} = \frac{1}{K}\sum_k \Phi(x_k)\Phi(x_k)^T \qquad (6.5)$$

相似於 PCA 的情形，我們可以找出 $\overline{C}$ 在內積空間 $\mathbb{F}$ 的 eigenvalue 及相對應的 eigenvector。

在在內積空間 $\mathbb{F}$ 內，有一個 eigenvector V，且存在一組係數 $\alpha_k, k = 1, \cdots, K$，使得

$$\mathbf{V} = \sum_k \alpha_k \cdot \Phi(x_k) \qquad (6.6)$$

由方程式 (6.5)、(6.6) 我們可以整理出以下的式子，

$$\lambda \alpha = P\alpha \qquad (6.7.1)$$

其中

$P_{i,j} = \left(\Phi(x_i), \Phi(x_j)\right)$ 和 $\alpha = [\alpha_1, \cdots, \alpha_K]^T$

假如 $\sum_j \Phi(x_j) \neq 0$，我們仍然可以有類似的方程式

$$\tilde{\lambda}\tilde{\alpha} = \tilde{P}\tilde{\alpha} \qquad (6.7.2)$$

其中 $\tilde{P}_{i,j} = (P - 1_K P - P 1_K + 1_K P 1_K)$ 和 $\left(1_K\right) = 1/K$

找到 P 的 eigenvector α 後，為了將在 $\mathbb{F}$ 空間內的向量 V 正規化至 V'，我們必需對 α 作一個標準化至 α' 也就是說，

$$\mathbf{V}' = \sum_k \alpha'_k \cdot \Phi(x_k) \quad \text{and} \quad |\mathbf{V}'| = 1 \tag{6.8}$$

最後，我們可以根據 (6.6) 或 (6.8) 式找出向量x在內積空間$\mathbb{F}$內，Principal Components（PCs），

$$(\mathbf{V}', \Phi(x)) = \sum_k \alpha'_k \cdot (\Phi(x_k), \Phi(x)) \tag{6.9}$$

因此，我們藉由本節的 KPCA 的方式，可將我們音節的超向量轉換到內積空間$\mathbb{F}$內，以 PCs 為分析對象的向量，最後，我們依據不同 PCs 所代表的聲學特性，來將音節作一個分羣的動作。

6.2.3 音節分羣之方式

針對音節分羣而言，分羣的目的在於音節的合理分裂，進而能夠更精細的描述音節特性，如有發生音節合併或未發生音節合併，因此，我們憑藉著 KPCA 分析的方式，來將音節裡的雜訊去除，進而找出音節中重要的聲學特性。

在分群的作法上，我們想以最簡易且合理的方式來進行，在進行的過程中我們必須考慮到三點：(1)音節分群之標準、(2)音節分類後資料集中的程度、(3)音節資料的數量。

(1) 音節分群之標準

在音節分群上，我們主要是在內積空間$\mathbb{F}$內進行分群別類的動作，而我們所使用的分類策略為 $(\mathbf{V}^k, \Phi(x)) \gtrless 0 \ \forall k$，也就是音節的 PC 是大於 0 或小於 0 來作為分類的標準。

(2) 音節分類後資料集中的程度

首先我們假設音節分裂前及分裂後的資料量，兩者皆有相當多的資料量，足以讓我們認為其分佈大略為高斯分佈，接下來我們就使用以下的量度，來作為音節分類後，資料集中的程度；其數學表示式[13]如下（6.10）式而其概念示意圖，如圖 6.1 所示。

$$\Delta L = \frac{(n_1 + n_2)\log|\Sigma| - n_1 \log|\Sigma_1| - n_2 \log|\Sigma_2|}{(n_1 + n_2)} \tag{6.10}$$

其中

$$\begin{cases} \Delta L: \text{最大相似度增益} \\ \Sigma: \text{父節點之變異量} \\ n: \text{父節點之資料數量} \\ \Sigma_i: \text{子節點之變異量,i} \in 1,2 \\ n_i: \text{子節點之資料數量,i} \in 1,2 \end{cases}$$

$$n = n_1 + n_2$$

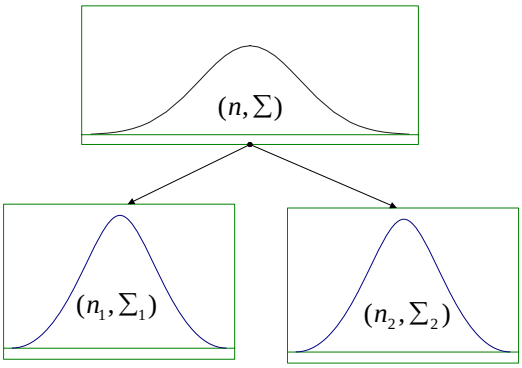


圖 6.1 資料分裂示意圖

由上式（6.10）得知，如果被分裂的兩子節點的資料分佈若非常集中的話，則Σ_1 和 Σ_2的數值將會變得較小，此時最大相似度增益，將會變得比較大，而在另外一種情形，兩子節點的資料數差不多（$n_1 \approx n_2$）時，$-n_1 \log|\Sigma_1| - n_2 \log|\Sigma_2|$ 的量會變得比較小，在此時可以獲得較大的最大相似度增益。因此，(6.10)式正符合我們測試資料集中程度的需求。

(3) 音節資料的數量

為了考量能夠使用 KPCA 分析的方式，來呈現音節完整的聲學特性，我們必須選擇資料量較大的音節，來做我們音節分類的候選人，所以，我們在這裡採用了相對比較的方法，即是我們選擇資料量較大的前一百名，來當作我們分類的候選人。

接下來，我們就直接來說明我們分裂中文 411 音節的流程，如圖 6.2 所示，

(1) 收集在訓練語料中所有的 411 音節，在此我必須強調此時中文音節只有 411 類。

(2) 我們針對在訓練語料中，標示為相同的 411 音節之超向量，拿來做 KPCA 的分析並且使用上述的量度來分析的音節分群的成效，在此時我們就有 411 組音節分析的結果。

(3) 在音節資料量的考量下，我們只選出（2）411 組音節中，資料量較大的前 100 名。

(4) 在 100 組音節裡，裡找出獲得最大相似度增益的一組音節。

(5) 將此組音節的所有超向量，分為兩大類（ex. bu->bu1、bu2）。

(6) 回朔至訓練語料上，重新標示音節，此時在訓練語料中的總音節類別數會增加一。

(7) 檢查總音節數量是否達到我們的臨界值，不是的話，則回到（2）－（7），是的話則

達到目標停止分裂。

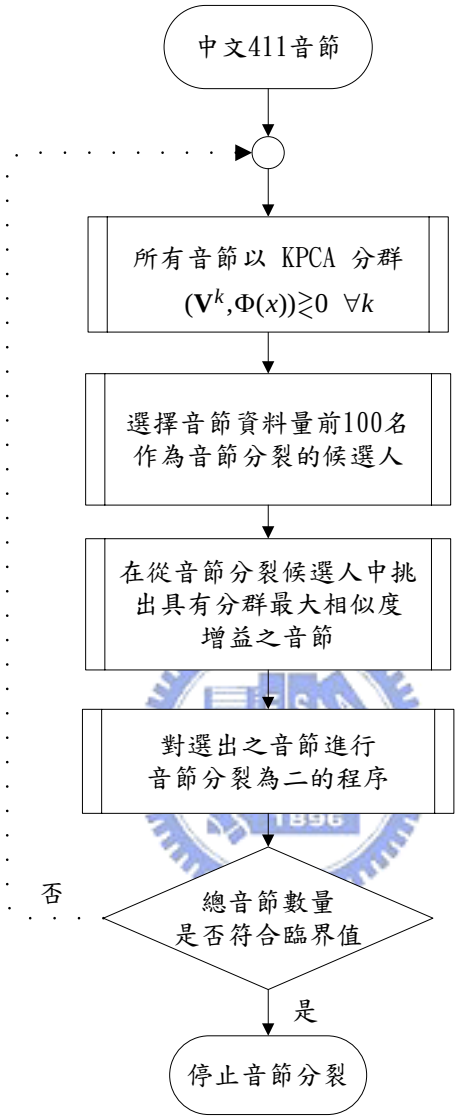


圖 6.2 中文音節分裂程序

6.3 實驗

在這一節裡，我們的實驗可以被劃分為兩大部分－KPCA 之基礎向量（Principal Component，PC）聲學特性、分裂音節之 HMM 成效分析。在此，定義 KPCA 之 kernel function 為 $ker(x,y) = (\Phi(x), \Phi(y))$ ，而我們所使用的 kernel function 是 RBF(radian Bayesian Function)，

$$ker(x,y) = e^{-(x-y)^2/(200\cdot\sigma^2)} \tag{6.11}$$

其中 σ : *the standard deviation of data*

6.3.1 KPCA 之基礎向量所示聲學差異之觀察

在這一節，我們將就所觀測到基礎向量之聲學特徵，來作一個簡單的說明，並藉此說明來佐證我們使用 KPCA 方法來做音節分類的合理性。

由於篇幅有限，所以我們無法將所有 411 音在 eigen-sapce 的分佈呈現出來，故我們在此僅舉音節 “/xiang/(相)” 做為 411 音在 eigen-sapce 的分佈的範例；如圖 6.3 所示，PC2 所代表的應該是性別；就我們觀察，若我們以該軸的 0 點為性別的分界點的話，我們可以蠻精確的區分出男性（╳）或女性（○）

接下來，我們繼續觀察 PC1（相對特徵值最大之成分），我們將其區分為三個區域，然後再個別來觀察其聲學特性；首先，我們觀察 class2，我們發現其音節之結構並未發生發音變異之情形，如 圖 6.4 所示，在此我們可以發現一個事實，就是在我們語料中，多數人對於此音節的唸法，較集中於 0 附近，這代表多數人對此音節的唸法上，並未有太大的差異，這通常應該是標準式的音節結構；我們再觀察 class1，我們發現其音節之結構發生變異，如 圖 6.5 所示，我們發現其音節的 INITIAL 的部分已經不見了，這表示此音節之結構發生短少現象，也就是發生 syllable contraction，故這應該屬於表層式的音節結構; 我們再觀察 class1，我們發現其音節之結構並未發生太大的變異，如 圖 6.6 所示，但卻發生了音節的振動(vibration)現象。

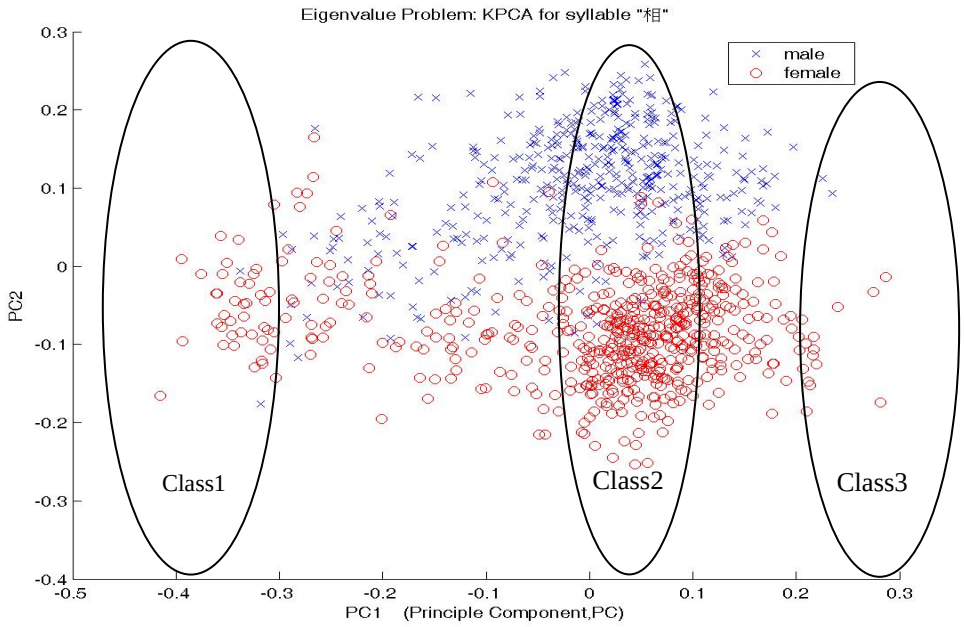


圖 6.3 /xiang/(相)在 eigen-space 的分佈

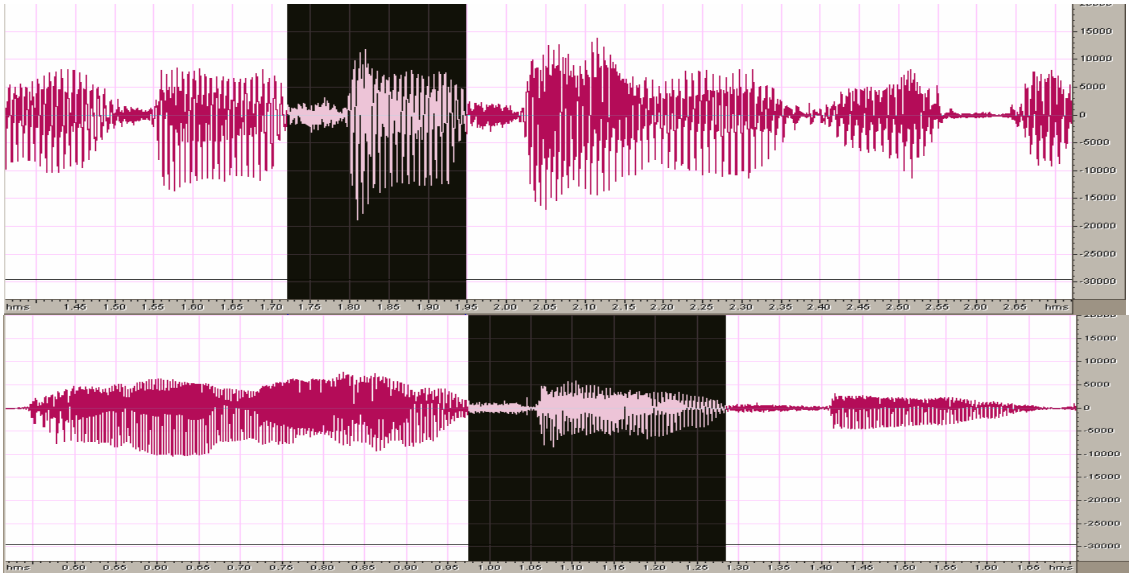


圖 6.4 在 class2（正常情形）中/xiang/(相)的波形例子

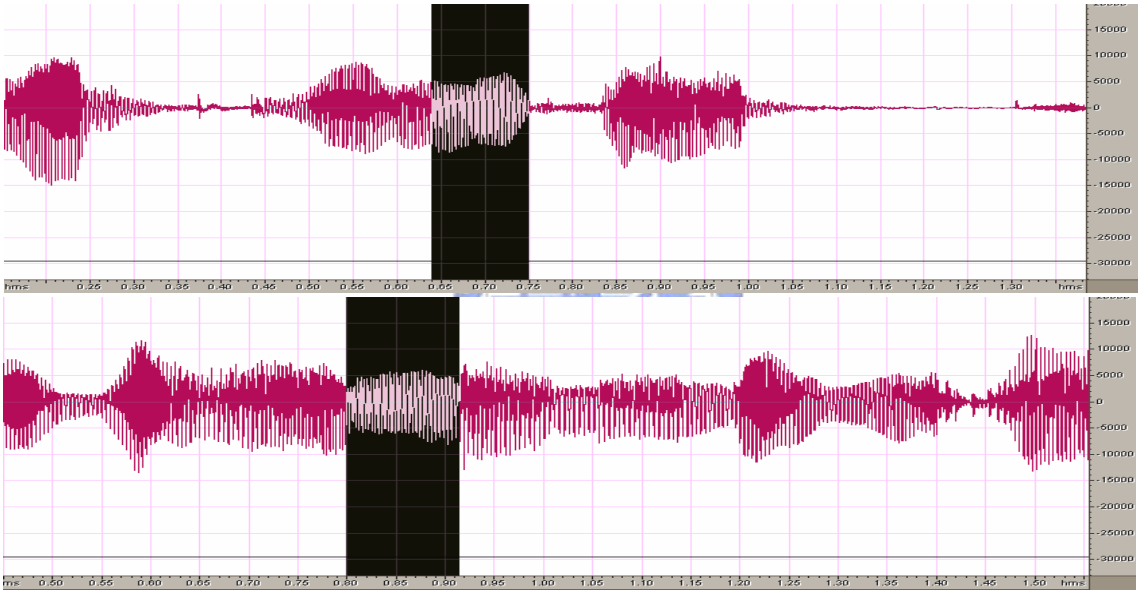


圖 6.5 在 class1(變異情形)中/xiang/(相)的波形例子

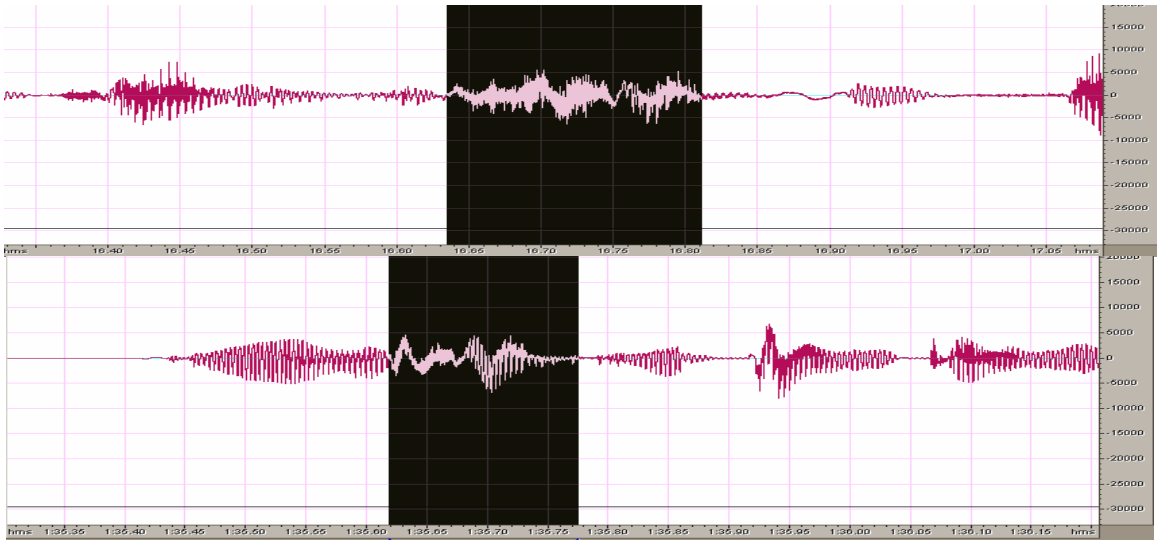


圖 6.6 在 class3(變異情形)中/xiang/（相）的波形例子

由以上所觀察的現象得知，以 PC1 的正，負值來看，正值及負值分別代表不同的變異情形及程度，若我們以正負值來區分音節的變異情形，應該可以獲得某種程度的成效。

6.3.2 分裂音節之 HMM 成效分析

我們根據 6.2.3 節，所介紹的音節分類方法，來進行 411 音節分裂程序(不考慮男女分群)，最後我們將 411 音節中的 26 個音節，分裂成 57 個音節，如表 6.1 所示；根據我們的統計，這 57 個音節佔了測試語料音節總數的 26.22%，因此，對於分裂的音節數量，應該是足夠去判別我們所使用的方法是否具有顯著的成效。

6.3.2.1 使用模型分裂（Model Splitting）建立分裂音節模型

針對被分裂出來的音節部分，我們視分裂之音節為相異之音節，然後再重新建立系統之聲學模型（CD HMM)，然後，再進行辨識實驗；而實驗的環境設定與基本系統環境設定一致，最後，我們得到以下結果，如 表 6.2 所示，

表 6.1 被分裂音節一覽表

編號	分裂音節名稱	類別總數	訓練語料音節總數	編號	分裂音節名稱	類別總數	訓練語料音節總數
1	/bu/（不）	2	1015	15	/chang/（昌）	2	189
			1136				180
2	/duo/（多）	2	263	16	/di/（滴）	2	197
			213				158
3	/fu/（夫）	2	161	17	/hen/（很）	2	521
			122				512
4	/hui/（輝）	2	529	18	/jian/（間）	2	151
			625				137
5	/jue/（絕）	2	376	19	/ke/（柯）	2	582
			310				513
6	/lu/（魯）	2	185	20	/men/（悶）	2	772
			151				634

7	/na/ (哪)	2	446	21	/ne/ (呢)	2	309
			356				258
8	/neng/ (能)	2	186	22	/ran/ (冉)	4	260
			199				174
9	/shi/ (師)	2	2459				131
			3058				199
10	/shuo/ (說)	2	418	23	/wan/ (彎)	2	192
			396				161
11	/wei/ (威)	2	496	24	/xiao/ (消)	2	429
			484				214
12	/suo/ (縮)	3	175	25	/xue/ (穴)	3	130
			165				178
			196				192
13	/ye/ (也)	2	477	26	/zhong/ (中)	3	338
			439				175
14	/zuo/ (左)	2	270				165
			285				

表 6.2 音節分裂成效比較表

HMM Models	Acc(%)	Sub(%)	Del(%)	Ins(%)
Baseline (CD HMMs)	45.23	37.70	11.91	5.16
Baseline with proun. Variation (音節分裂重建模型)	46.05	37.10	12.47	4.38
包含所有模型(411 音節+特殊音節)之辨識率				

我們從上表得知，相較於基本系統，使用聲學變異的模型將得額外的辨識率為 0.82%；接下來再來分析音節混淆表，如表 6.3、6.4 ，我們得知 411 對 411 的辨識率升高了 1.71%，但是其它的特殊模型被辨為 411 的百分比也增加，而由這個結果看來 411 的聲學模型變得較為精細後，會使得較多特殊模型被辨識為 411，但在一升一降中，我們仍然能獲得額外的 0.82%，最主要的原因是 411 音節佔了測試語料的總音節數，約 8 成多。

表 6.3 Confusion Matrix for Baseline with Proun. Variation

辨識結果 / 音節類別	411	Paralinguistic	Particles	Uncertain
411	96.67%	1.00%	1.27%	0.83%
Paralinguistic	15.15%	77.16%	5.09%	1.84%
Particles	46.20%	4.16%	46.20%	3.16%
Uncertain	80.17%	3.88%	2.59%	12.93%

表 6.4 Confusion Matrix for Baseline

辨識結果 / 音節類別	411	Paralinguistic	Particles	Uncertain
411	94.96%	1.43%	1.36%	2.21%
Paralinguistic	8.60%	85.08%	4.46%	1.76%
Particles	39.67%	4.46%	53.44%	2.42%
Uncertain	77.14%	3.30%	4.40%	15.16%

接下來，若我們只考慮 411 的辨識率，所得到的辨識率如 表 6.5 所示，的確，411 的辨識率確實是提高了 0.5%。

表 6.5 411 音節分裂成效比較表

HMM Models	Acc(%)	Sub(%)	Del(%)	Ins(%)
Baseline (CD HMMs)	48.48	35.58	11.95	3.99
Baseline with proun. Variation (音節分裂重建模型)	48.98	34.47	12.44	4.10
僅包含 411 音節模型之辨識率				

最後，分析分裂音節之辨識率，如表 6.6 所示，26 個音節其中有 9 個音節辨識率下降，其可能發生的原因，我們可以從音節的資料量及其資料在 eigen-space 的分佈來做研究，研究發現音節若僅被分一次且資料量（如表 6.1）在 200 以下，其音節辨識正確率會比較差，

反之音節若被分兩次以上則代表此音節變異情形較為明顯，因此，其音節辨識正確率會比較佳；在考慮另外一種情形，就是資料量夠，但其資料點在 eigen-space 的分佈不適合僅用一條垂直線去實行二分法來分類，如圖 6.7、6.8 。

表 6.6 分裂音節之音節辨識正確率比較表

編號	分裂音節名稱	音節辨識正確率		編號	分裂音節名稱	音節辨識正確率	
		分裂前	分裂後			分裂前	分裂後
1	/bu/	76.7%	75.4%	14	/zuo/	50.7%	72.7%
2	/duo/	75.0%	76.4%	15	/chang/	58.8%	69.8%
3	/fu/	94.6%	94.4%	16	/di/	80.6%	73.3%
4	/hui/	59.7%	74.4%	17	/hen/	71.4%	72.3%
5	/jue/	73.7%	82.7%	18	/jian	72.2%	54.1%
6	/lu/	70.2%	62.5%	19	/ke/	74.3%	79.0%
7	/na/	33.3%	33.8%	20	/men/	67.2%	74.9%
8	/neng/	34.2%	31.4%	21	/ne/	52.0%	42.9%
9	/shi/	68.7%	73.2%	22	/ran/	61.6%	71.9%
10	/shuo/	71.7%	79.7%	23	/wan/	78.7%	78.3%
11	/wei/	46.7%	52.9%	24	/xiao/	76.2%	72.8%
12	/suo/	52.4%	73.3%	25	/xue/	70.2%	82.0%
13	/ye/	51.2%	60.9%	26	/zhong/	54.5%	71.0%

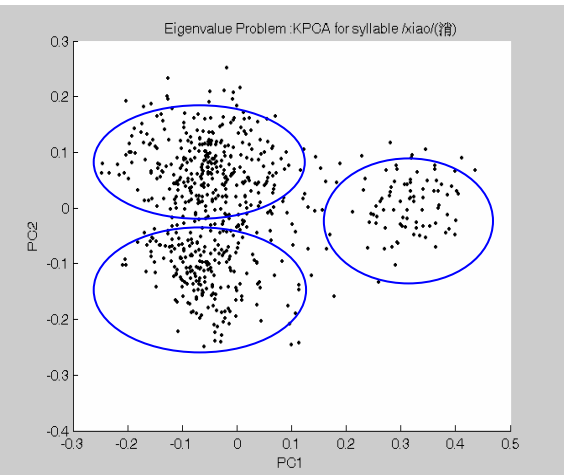


圖 6.7 /xiao/(消)在 eigen-space 的分佈

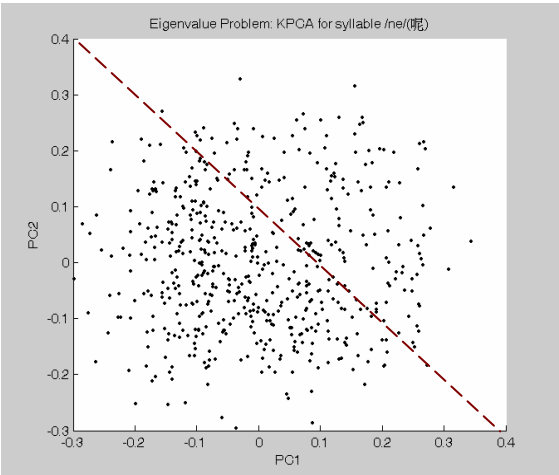


圖 6.8 /ne/(呢)在 eigen-space 的分佈

6.3.2.2 使用模型調適（Model Adaptation）建立分裂音節模型

在這一節裡，我們的做法是直接使用音節原有的模型，再利用分類後的音節資料點當作調適語料，將音節模型，調適到不同類別的分裂音節模型，做法如 圖 6.9 所示；因為，我們所使用的模型是 CD HMM 的緣故，所以，音節相關的模型，也比較多了，這也造成了實質的調適語料大幅的減少，因此，在做調適時，必須對調適語料量作限制，以避免將分裂音節模型調適成特殊的音節模型。

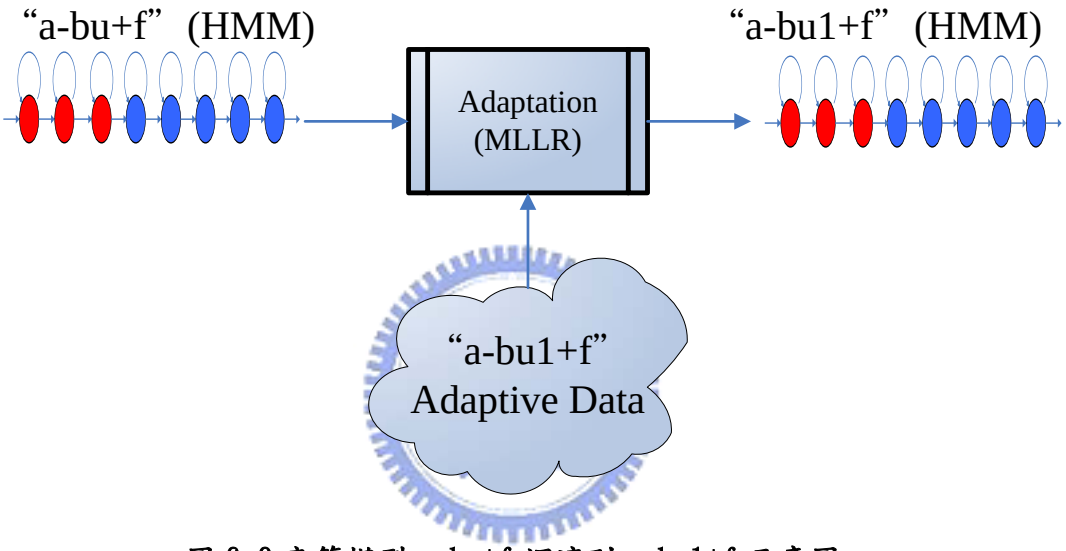


圖 6.9 音節模型 a-bu+f 調適到 a-bu1+f 示意圖

接下來，將調適後的模型添加到辨識系統的聲學模型，然後，再進行辨識實驗，其實驗的環境設定與基本系統環境設定一致;除此之外，我們還對音節的調適語料量做了限制，如下所示：

1. 應調適 HMM 模型數：7827 個模型。
2. 有調適語料的模型數 1874，但通過所設定的調適語料量門檻值僅有 403 個模型。

最後得到以下結果，如表 6.7 及 6.8 所示，

表 6.7 調適後音節分裂成效比較表

HMM Models	Acc(%)	Sub(%)	Del(%)	Ins(%)
Basline (CD HMMs)	45.23	37.70	11.91	5.16
Basline with proun. variation	43.79	39.05	11.64	5.52
包含所有模型(411 音節+特殊音節)之辨識率				

表 6.8 調適後分裂音節之音節辨識正確率比較表

編號	分裂音節名稱	音節辨識正確率		編號	分裂音節名稱	音節辨識正確率	
		分裂前	調適後			分裂前	調適後
1	/bu/	76.7%	73.4%	14	/zuo/	50.7%	44.3%
2	/duo/	75.0%	70.9%	15	/chang/	58.8%	58.8%
3	/fu/	94.6%	91.9%	16	/di/	80.6%	76.7%
4	/hui/	59.7%	58.4%	17	/hen/	71.4%	67.6%
5	/jue/	73.7%	68.8%	18	/jian	72.2%	73.0%
6	/lu/	70.2%	70.7%	19	/ke/	74.3%	69.0%
7	/na/	33.3%	29.2%	20	/men/	67.2%	57.6%
8	/neng/	34.2%	25.0%	21	/ne/	52.0%	51.4%
9	/shi/	68.7%	54.3%	22	/ran/	61.6%	53.7%
10	/shuo/	71.7%	67.7%	23	/wan/	78.7%	78.7%
11	/wei/	46.7%	46.2%	24	/xiao/	76.2%	71.2%
12	/suo/	52.4%	48.9%	25	/xue/	70.2%	64.4%
13	/ye/	51.2%	41.4%	26	/zhong/	54.5%	45.2%

由表 6.7、6.8 得知使用適調分裂音節模型，反而，使得辨識系統的辨識率變差，其中最大的原因可能是調適語料過少，經統計結果得知在我們調適的語料裡，只有 39 個相關分裂音

節模型資料筆數超過 20 筆，如表 6.9，因此，在作調適時，很容易將分裂音節模型調適成獨特的模型，使得辨識率變差。

表 6.9 調適語料筆數超過 20 之 CD HMM 一覽表

CD HMM	調適語料筆數	CD HMM	調適語料筆數
FNULL1-shuo2	26	e-shi1+n	25
FNULL1-shuo1	28	e-shi1+h	136
a-bu1+f	34	an-shi1+INULL_w	20
FNULL1-ke2+INULL_y	20	wu-shi2+INULL_y	23
wu-ran2+j	21	e-shi2+n	26
wei1+sh	28	e-shi2	35
wu-xiao1+d	30	yi-shi2	26
yi-men1+x	29	a-shi2+h	20
a-men1+d	27	e-shi2+q	27
a-men1+j	36	e-shi2+j	27
a-men2+j	24	yi-shi2+n	20
e-di2+f	25	e-shi2+INULL_y	27
e-shi1	39	wu-shi2+h	25
e-shi1+t	30	ai-shi2+INULL_y	41
yi-shi1+sh	21	yi-shi2+INULL_w	23
a-shi1+h	29	wu-shi2	21
e-shi1+INULL_w	63	e-shi2+h	125
ai-shi1+sh	20	e-di1+f	26
FNULL1-shi1+sh	21	FNULL1-bu2+sh	59
yi-shi1+INULL_w	55		

第七章 結論與未來展望

7.1 結論

在本論文裡，我們所使用的基本系統聲學模型是左右相關的 HMM 模型（非共用轉移機率），其辨識率高於 RCD Intial+Final HMM 模型 6.66%；若我們進一步修改成 Left-Right HMM 的模型，准許其跳過一個狀態，便可將辨識率再提升 0.36%；由分析辨識混淆表，我們可以發現 Particle 及 Uncertain 容易與 411 相互辨識，其最主要的原因是因為它們的聲音訊號過於相似，只是語意上有所區別而已。

因此，對於這種情形，我們引用了調適的語言模型，以補償語言及口語上的文字缺陷，最後，我們可以再提升約 7%，但辨識率只有 56.40，我們推斷其原因，可能是 411 聲學模型的辨識能力不夠好，而不夠好的理由是因為語料庫裡有太多的音節合併現象及音檔裡的串音現象；在本論文，我們試圖對音節合併現象來做一個處理，以提升系統聲學模型的辨識率。

基於此目標，我們提出了使用 KPCA 將音節的變異現象（包含了音節合併現象）找出來，並且使用二分法，將音節予以分類，再根據音節類別，建立其 HMM 模型，最後獲得額外的 0.8% 辨識率。或許是因為分類的方式，導致改善辨識率幅度不大，但對於音節訊號之分類上，的確是有一定的成效存在。

7.2 未來展望

由於我們使用的音檔裡有蠻嚴重的串音存在，因此，若將來能夠在建構聲學模型時決解串音的問題，勢必可以大幅度的提升系統的辨識率。

此外，針對使用 KPCA 將音節訊號變異現象做分類的部分，我們未來可以改進的地方，有下列幾點以供參考：

1. 分類之類別邊界
2. 分類之類別數設定
3. 發現更多基礎向量之聲學特徵(例如：語者說話速度、音長、聲調…等)。

參考文獻

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

[2] Rabiner, L.R. and Juang, B.H.,Fundamentals of speech Recognition, New Jersey, Prentice-Hall,Inc.,1993.

[3] 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.

[4] S.Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, The HTK Book（for HTK Version 3.2.1）, 2002.

[5] S.Young. Large Vocabulary Continuous Speech Recognition:a Review. IEEE Workshop on Automatic Speech Recognition, 1996.

[6] S. Young, and P. Woodland. State clustering in HMM-based continuous speech recognition, Computer Speech and Language,vol.8,no.4, pages 369-384, 1994.

[7] 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.

[8] Lawrence Rabiner and Biing-Hwang Juang. Fundamentals of Speech Recognition, Prentice Halll International, pages 324-325, 1993.

[9] Shu-Chuan Tseng. Feature of Contraction Syllable of Spontaneous Mandarin, EUROSPEECH 2003, pages77-80, 2003.

[10] Mirjam Wester and Eric Fosler-Lussier. A Comparison of Data-Driven and Knowledge-Based Modeling of Pronunciation Variation, ICSLP '00, volume I, pages 270-273, Beijing, 2000.

[11] Yi Liu and Pascale Fung. Pronunciation Modeling for Spontaneous Mandarin Speech Recognition,INTERNATIONAL JOURNAL OF SPEECH TECHNOLOGY 7, pages155－172, 2004.

[12] Bernhard Schölkopf and Alexander Smola. Nonlinear Component Analysis as Kernel

Eigenvalue Problem, MIT Press, pages 1299-1319, 1998

[13] Mei-Yuh Huang, Xue-Dong Huang. Dynamically Configurable Acoustic Models for Speech Recognition, ISCAP 1998, Vol. 2.

附錄一

編號	mcdc 對話編號	mcdc 子音檔編號範圍	語者編號	聲道(L/R)	sample rate
1	MCDC-01	MCDC-01-01~20	MISC-08-male-25	*R	44.1kHz
	MCDC-01	MCDC-01-01~20	MISC-07-female-29	L	44.1kHz
2	MCDC-02	MCDC-02-01~22	MISC-10-male-35	R	44.1kHz
	MCDC-02	MCDC-02-01~22	MISC-09-female-37	*L	44.1kHz
3	MCDC-03	MCDC-03-01~21	MISC-12-female-17	R	44.1kHz
	MCDC-03	MCDC-03-01~21	MISC-11-female-16	*L	44.1kHz
4	MCDC-05	MCDC-05-01~20	MISC-15-male-40	L	44.1kHz
	MCDC-05	MCDC-05-01~20	MISC-16-female-46	*R	44.1kHz
5	MCDC-09	MCDC-09-01~21	MISC-23-female-30	R	44.1kHz
	MCDC-09	MCDC-09-01~21	MISC-24-female-35	*L	44.1kHz
6	MCDC-10	MCDC-10-01~18	MISC-26-male-23	*R	44.1kHz
	MCDC-10	MCDC-10-01~18	MISC-25-male-35	L	44.1kHz
7	MCDC-25	MCDC-25-01~19	MISC-57-male-43	L	44.1kHz
	MCDC-25	MCDC-25-01~19	MISC-58-female-45	*R	44.1kHz
8	MCDC-26	MCDC-26-01~16	MISC-60-male-24	*R	44.1kHz
	MCDC-26	MCDC-26-01~16	MISC-59-female-37	L	44.1kHz

備註：＊代表該音檔首位發音者所使用的聲道（Left/Right）。

附錄二

- **特殊音韻現象(pronunciation variation)**

1. 拖長音(lengthening)

   標記實例:

   我目前是從<b lengthening>**事**</b lengthening>外貿

2. 音的同化(assimilation)

   標記實例:

   賴先生<b assimilation>**呢您**<b assimilation>從事什麼工作

3. 音節合併(syllable contraction)

   1).三個字三個音節變成三個字兩個音節 或 兩個字變成一個音節。

   "**我們**"實際發音[om]

   2).音節連在一起,難以分割(音節無短少)。

   3).音節結構有變(音節無短少)。

   標記實例:

   但是相對於跟淡水 A<b syllable contraction>**那種**</b syllable contraction>什麼木柵那邊比就少很多了

4. 鼻化音(nasalized)

   標記實例:

   室內就是一小間一小間 MA NA 露天就是大<b nasalized>**家**</b nasalized>一起 A

5. 發音偏差(inappropriate pronunciation)

   標記實例:

   我<b inappropriate pronunciation >比</b inappropriate pronunciation >較喜歡從事一些球類運動 LA

- **無法或難以辨識的語音(unintelligible speech sound)**

1. 喃喃自語(mumble)

   標記實例:

   都在賺錢 O<b mumble>賺錢</b mumble>

2. 無法辨識的語音(unrecognizable speech sound)

   標記實例:

   因為<b unrecognizable speech sound >@</b unrecognizable speech sound >太貴了

- **不確定字/音(uncertain)**

1. 標記員根據前後語意,可以猜出大概的語意內容,但無法百分之百確定。

   標記實例:

   至少我對我自己的車子<b uncertain>有</b uncertain>有一個瞭解程度 BA

2. 標記員無法根據語意猜測出對應字詞,但可辨識出清楚的發音時,漢字與併音的標記內容都記為[實際發音]。若聲調亦可辨識出,也一併標記。

   標記實例:

   <b uncertain> [fal] </b uncertain>因為大概離台北市區比較遠一點所以人不會那麼多

- **語流中斷(prosodic disfluency)**

1. 沉默(silence)

   標記實例:

   <b silence >@</b silence > (1570 ms)

2. 停頓(pause) (約 600 ms 以上)

   標記實例:

   然後黃線好像是九百<b pause>@</b pause>然後有的開到一千二

3. 短停頓(short break)(約 200-400ms)

   標記實例:

   那邊你要是熟<b short pause>@</b short pause>就要鑽到吳興街那邊算近的了

4. 字詞片段(word fragment)

   標記實例:

   外貿 A 是進<b word fragment >口</b word fragment >EN 出口嗎

5. 口吃(stutter)

   標記實例:

   其實沒什麼影響因為那個價格跟<b stutter>[u5]外國人</b stutter>的那些商人都已經講好了

- **不完整句法結構(lexico-syntactic disfluency)**

1. 不適當用法(inappropriate usage)

   標記實例:

   可是烏來也很塞 EI<b inappropriate usage >上次**是**我們去也**是**一路塞上去然後再塞下來</b inappropriate usage >

2. 被對方打斷(interrupted)

   標記實例:

   Speaker MISC-07-femal-29:0 去山上繞一繞<b interrupted >是</b interrupted >

   Speaker MISC-08-male-25:像譬如說會去烏來 A

3. 句子中斷(abridged)

   標記實例:

   <b abridged>**它有一個天**</b abridged>E 那邊有個天籟渡假村 MA

4. 語誤(error)

   標記實例:

   你也不知道是誰開車的 A 對不對你就開<b error>**這張車子**</b error>而已

- **詞語修補(repair)**

1. 重覆(repetition)

   標記實例:

   A 要處理可是<b repetition>**又有又有**</b repetition>ZHE GE 情理法法理情

2. 部分重覆(restart)

   標記實例:

   真的是稍微動用一下就覺得<b restart>**很很不夠用**</b restart>這樣子

3. 詞語更正(repair)

   1).語意更正

2).語音更正

3).聲調更正

4).詞語更正

標記實例:

<b repair>**你您**的住處</b repair>就是在永春站那附近就對了

標記實例:

當時<b repair>**我才反應到我才意識到**</b repair>說其實愛是需要填補的

4. 更正插語(editing term)

標記實例:

外貿 A 是進口<b editing term>EN</b editing term>出口嗎

- **受外語或方言影響(socio-linguisic phenomena)**

1. 語言轉換(code-switching)

1).閩南語 2).英語

標記實例:

它有一個很大的<b code switching ><b Min-Nan >看板</b Min-Nan ></b code switching >會

標記實例:

真正通化街那一條不是有<b code switching ><b English>HANGTEN</b English></b code switching >NA<b code switching ><b English> GIORDANO </b English></b code switching >那一些

2. 受閩南語影響的發音(Min-Nan-influenced pronunciation)

標記實例:

不然泡在室內不會很溫暖 E 只是很悶而已因為真的很<b Taiwanese-influenced pronunciation > <b r-l>熱</b r-l></b Taiwanese-influenced pronunciation >

3. alternative-約定俗成讀音

標記實例:

真的那邊車子開個八九十沒有關<b alternative-xil>係</b alternative-xil>A

- **其它(Others)**

1. 語助詞(marker)

標記實例:

室內就是一小間一小間 MA <b marker> NA</b marker>露天就是大家一起 A

2. 感歎詞(particle)

1).有相對應國字的感歎詞 ex. A、BA.....

2).無相對應國字的感歎詞 ex. NE、NA….

3).源於台語的感歎詞 ex. HO、HAN....

4).其他的感歎詞(Fillers) ex UHN、UHNN.....

標記實例:

去什麼富基港<b particle >A</b particle >那些

標記實例:

<b particle >EI</b particle >你好我姓賴請問一下貴姓

- **非語音部份口語標註(Non-Speech Sounds)**

1. 人聲(human voice)

笑聲、咳嗽聲、吐氣聲.....

1).伴隨語言內容之人聲

標記實例:

我覺得今天我少一點花個三百塊跟直接投資三萬塊這<b laugh >A</b laugh >

2).無伴隨語言內容之人聲

標記實例:

大概是我們的運氣不好<b laugh >@</b laugh >

2. 非人聲(non human sound)

1).室內雜音(noise in room)

a).伴隨語言內容之非人聲

標記實例:

<b noise in room >像我工作就是在那邊去看的 </b noise in room >(下雨聲)

b).無伴隨語內容之非人聲

標記實例:

<b noise in room>@</b noise in room>NHN

- **同一輪標記(same of the turn)**

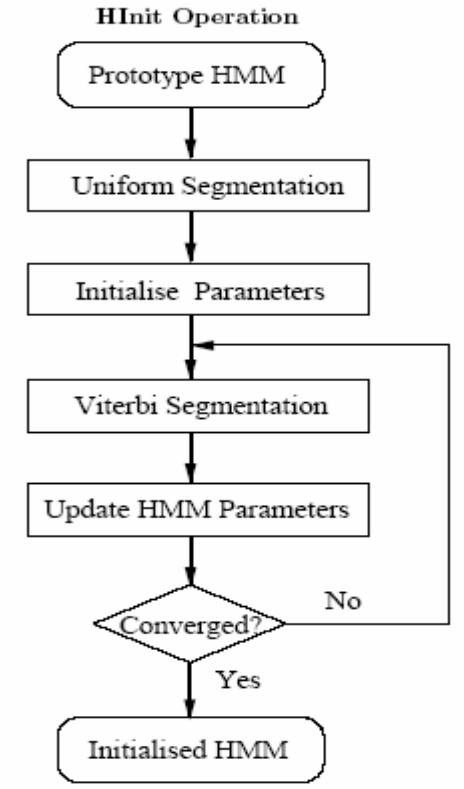
標記實例:

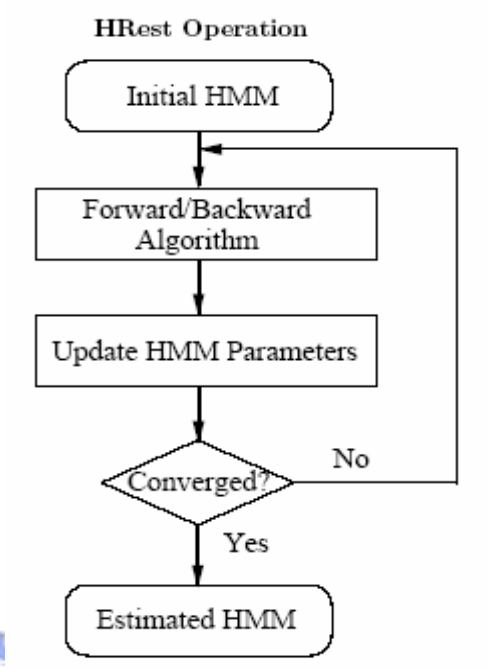
.....看法<b short break>@</b short break>像<b pause>@</b pause>(mcdc-01-09.wav)

><b syllable contraction>像我</b syllable contraction>自己有玩.....(mcdc-01-10.wav)

附錄三

- Hinit & Hrest Alogrithm





附錄四

子音編號(100 類)	子音拼音(100 類)	注音	子音編號(100 類)	子音拼音(100 類)	注音	子音編號(100 類)	子音拼音(100 類)	注音
1	zh_1	ㄓ_1	35	c_4	ㄘ_4	69	b_5	ㄅ_5
2	ch_1	ㄔ_1	36	s_4	ㄙ_4	70	p_5	ㄆ_5
3	sh_1	ㄕ_1	37	g_4	ㄍ_4	71	m_5	ㄇ_5
4	r_1	ㄖ_1	38	k_4	ㄎ_4	72	zh_6	ㄓ_6
5	z_1	ㄗ_1	39	h_4	ㄏ_4	73	ch_6	ㄔ_6
6	c_1	ㄘ_1	40	d_4	ㄉ_4	74	sh_6	ㄕ_6
7	s_1	ㄙ_1	41	t_4	ㄊ_4	75	r_6	ㄖ_6
8	zh_2	ㄓ_2	42	n_4	ㄋ_4	76	z_6	ㄗ_6
9	ch_2	ㄔ_2	43	l_4	ㄌ_4	77	c_6	ㄘ_6
10	sh_2	ㄕ_2	44	b_4	ㄅ_4	78	s_6	ㄙ_6
11	z_2	ㄗ_2	45	p_4	ㄆ_4	79	g_6	ㄍ_6
12	c_2	ㄘ_2	46	m_4	ㄇ_4	80	k_6	ㄎ_6
13	s_2	ㄙ_2	47	f_4	ㄈ_4	81	h_6	ㄏ_6
14	g_2	ㄍ_2	48	r_2	ㄖ_2	82	d_6	ㄉ_6
15	k_2	ㄎ_2	49	zh_3	ㄓ_3	83	t_6	ㄊ_6
16	h_2	ㄏ_2	50	ch_3	ㄔ_3	84	n_6	ㄋ_6
17	d_2	ㄉ_2	51	sh_3	ㄕ_3	85	l_6	ㄌ_6
18	t_2	ㄊ_2	52	r_3	ㄖ_3	86	b_6	ㄅ_6
19	n_2	ㄋ_2	53	z_3	ㄗ_3	87	p_6	ㄆ_6
20	l_2	ㄌ_2	54	c_3	ㄘ_3	88	m_6	ㄇ_6
21	b_2	ㄅ_2	55	s_3	ㄙ_3	89	f_6	ㄈ_6
22	p_2	ㄆ_2	56	g_3	ㄍ_3	90	j_7	ㄐ_7
23	m_2	ㄇ_2	57	k_3	ㄎ_3	91	q_7	ㄑ_7
24	f_2	ㄈ_2	58	h_3	ㄏ_3	92	x_7	ㄒ_7
25	l_3	ㄌ_3	59	d_3	ㄉ_3	93	n_7	ㄋ_7
26	b_3	ㄅ_3	60	t_3	ㄊ_3	94	l_7	ㄌ_7
27	p_3	ㄆ_3	61	n_3	ㄋ_3	95	INULL_2	Φ2
28	m_3	ㄇ_3	62	j_5	ㄐ_5	96	INULL_3	Φ3
29	f_3	ㄈ_3	63	q_5	ㄑ_5	97	INULL_4	Φ4
30	zh_4	ㄓ_4	64	x_5	ㄒ_5	98	INULL_5	Φ5
31	ch_4	ㄔ_4	65	d_5	ㄉ_5	99	INULL_6	Φ6
32	sh_4	ㄕ_4	66	t_5	ㄊ_5	100	INULL_7	Φ7
33	r_4	ㄖ_4	67	n_5	ㄋ_5			
34	z_4	ㄗ_4	68	l_5	ㄌ_5			

母音編號	母音符號(40 類)	注音	母音編號	母音符號(40 類)	注音
1	FNULL1	Φ1	36	yun	ㄩㄣ
2	a	ㄚ	37	yung	ㄩㄥ
3	o	ㄛ	38	er	ㄦ
4	e	ㄜ	39	yo	ㄧㄛ
5	eh	ㄝ	40	FNULL2	Φ2
6	ai	ㄞ			
7	ei	ㄟ			
8	ao	ㄠ			
9	ou	ㄡ			
10	an	ㄢ			
11	en	ㄣ			
12	ang	ㄤ			
13	eng	ㄥ			
14	yi	ㄧ			
15	wu	ㄨ			
16	yu	ㄩ			
17	ya	ㄧㄚ			
18	ye	ㄧㄝ			
19	yai	ㄧㄞ			
20	yao	ㄧㄠ			
21	you	ㄧㄡ			
22	yan	ㄧㄢ			
23	yin	ㄧㄣ			
24	yang	ㄧㄤ			
25	ying	ㄧㄥ			
26	wa	ㄨㄚ			
27	wo	ㄨㄛ			
28	wai	ㄨㄞ			
29	wei	ㄨㄟ			
30	wan	ㄨㄢ			
31	wen	ㄨㄣ			
32	wang	ㄨㄤ			
33	weng	ㄨㄥ			
34	yue	ㄩㄝ			
35	yuan	ㄩㄢ			

附錄五

General Questions (phones in extended SAMPA notations)

Feature	phones
元音性	a o @ E i u y U U’ @’ m n l n# N# Z’ #(a) #(i) #(u) #(E) #(o) #(y) #(@)
輔音性	b p m f d t n l g k h dz ts s dz’ ts’ s’ Z’ dz\ ts\ s\ N# n# #(i) #(u)
突發性	b p d t g k dz ts dz’ ts’ dz\ ts\
聚集性	a E o @ @’ U U’ g k h N# dz\ ts\ s\ #(a) #(E) #(o) #(@)
發散性	i u E o y @ U U’ @’ b p m f d t n l dz ts s Z’ n# #(i) #(u) #(E) #(o) #(y) #(@)
沉鈍性	a u E o @ U U’ @’ b p m f n n# N# h Z’ #(a) #(u) #(E) #(o) #(@)
尖銳性	a @ E i y U U’ @’ d t dz ts s dz\ ts\ s\ #(a) #(i) #(E) #(y) #(@)
降音性	o u y U’ @’ m l dz’ts’ s’ Z’ #(o) #(u)
平音性	a @ E i U b p f d t n h dz ts s dz\ ts\ s\ #(a) #(i) #(E) #(@)
鼻音性	m n n# N#
口音性	a o @ E i u y U U’ @’ #(a) #(i) #(u) #(E) #(o) #(y) #(@)

Vowel Questions (phones in extended SAMPA notations)

Feature	phones
聚集性	a E o @ U U’ @’
發散性	i u E o y @ U U’ @’
沉鈍性	a u E o @ U U’ @’
尖銳性	a E @ I y U U’ @’
降音性	o u y U’ @’
圓音性	o u y
齊口呼	i
合口呼	o u y
撮口呼	y
開口呼	A E @ I U U’ @’
高音性	@ E u y U U’ @’
低音性	a o @ @’
洪音性	a o @ u @’
細音性	E I y U U’

Consonant Questions (phones in extended SAMPA notations)

Feature	phones
元音性	m n l N# n# Z' #(a) #(i) #(u) #(E) #(o) #(y) #(@)
延續性	m f n l N# n# s s' Z' #(a) #(i) #(u) #(E) #(o) #(y) #(@)
聚集性	g k h N# dz\ ts\ s\ #(a) #(i) #(o) #(@)
發散性	b p m f d t n l dz ts s Z' n# #(i) #(u) #(E) #(o) #(y) #(@)
沉鈍性	b p m f n n# N# h Z' #(a) #(u) #(E) #(o) #(@)
尖銳性	d t dz ts s dz\ s\ #(a) #(i) #(E) #(y) #(@)
降音性	m l dz' ts' s' Z' #(o) #(u)
平音性	b p f d t n h dz ts s dz\ ts\ s\ #(a) #(i) #(E) #(@)
口音性	#(a) #(i) #(u) #(E) #(o) #(y) #(@)
粗糙性	dz ts s dz' ts' s' dz\ ts\ s\
柔潤性	b p m f d t n l g k h Z' n# N#
送氣	p t k ts ts' ts\
不送氣	b d g dz dz' dz\
雙唇音	b p m
鼻音性	dz ts s
舌尖前音	d t n l
舌尖音	dz' ts' s' Z'
舌尖後音	dz\ ts\ s\
舌根音	g k h
塞音	b p d t g k
塞擦音	dz ts dz' ts' dz\ ts\
舌尖前塞擦音	dz ts
舌尖後塞擦音	dz' ts'
舌面前塞擦音	dz\ ts\
雙唇塞音	b p
舌尖塞音	d t
舌根塞音	g k
送氣塞擦音	ts ts' ts\
不送氣塞擦音	dz dz' dz\
送氣塞音	P t k
不送氣塞音	b d g