國 立 交 通 大 學

電信工程學系

碩士論文

自發性國語語音辨識

Spontaneous Mandarin Speech Recognition

研 究 生：李柏蒼

指導教授：王逸如　博士

中華民國九十六年八月

自發性國語語音辨識

Spontaneous Mandarin Speech Recognition

研 究 生：李柏蒼　　Student：Bo-Cang Li

指導教授：王逸如　　Advisor：Dr. Yih-Ru Wang

國 立 交 通 大 學

電 信 工 程 學 系

碩 士 論 文

A Thesis
Submitted to Department of Communication Engineering
College of Electrical Engineering and Computer Science
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of
Master
in Electrical Engineering

August 2007
Hsinchu, Taiwan, Republic of China

中華民國九十六年八月

自發性國語語音辨識

研究生：李柏蒼　　指導教授：王逸如 博士

國立交通大學電信工程學系碩士班

中文摘要

自發性語音是最接近人們在自然情況下的語音，較能實際應用於日常生活中，因此漸趨於重要。本論文將先用國語 sub-syllable HMM 建立聲學模型，從錯誤分析知道，Uncertain 與 Particle 常與 411 syllable 互相辨識，將 Uncertain 不加入辨識器，可提高辨識率，Particle 在聲學上，十分相似 411 syllable，只差異語意，聲學模型是難以區別。再以 Syllable HMM 建立聲學模型，將可提高辨識率 3.3%，使用 skip state 改善 Deletion 錯誤，改善並不如預期，從錯誤分析知道 Deletion 錯誤大部分由 Syllable Contraction 造成。

Spontaneous Mandarin Speech Recognition

Student : Bo-Cang Li Advisor : Dr. Yih-Ru Wang

Department of Communication Engineering

National Chiao Tung University

Abstract

The spontaneous speech is most close to the speech in natural cases of people and can relatively apply to daily life actually, so become more and more important gradually. The first sets up acoustics model with mandarin sub-syllable HMM. From the error analysis, the recognizing device doesn't often distinguish between Uncertain and 411 syllable, and between Particle and 411 syllable. Do not put Uncertain model into the recognizing device, can improve the recognizing rate. Particle is in acoustics, very similar 411 syllable, and the language purpose of Particle is only different from the language purpose of 411 syllable and this is difficult to distinguish between their acoustics models. The second sets up acoustics model with mandarin sub-syllable HMM, and then can improve 3.3% of recognizing rate, and uses skip state to improve Deletion error, and then that does not improve so well as expectancy. From Deletion error analysis, we know Syllable Contraction causes the most of Deletion error.

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。

還要感謝愛喇賽輝哥學長、常說晚餐吃什麼的性獸學長、一直坐在我旁邊的智合學長、阿德學長、希群學長、巴金叔叔學長、阿勇學長、凡士林學長、總是令人滿意的滿意、常去台北的張友驊、大大、常回家的啟風、哈男專區管理者小肚腩、牛頭不對馬嘴的小鄧、打球很厲害的小傅，及瞇瞇眼的胤賢，有他們專長的分享，使我在各方面學到了很多；除此之外，我們還要感謝人帥真好的小廣、有宅氣的阿宅、衝撞達人小邱毅、姜公子，幫我檢查音檔錯誤。

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

圖目錄

表目錄

第一章　緒論

1.1 研究動機

隨著科技的進步，可以讓我們的生活越來越方便，科技產品誕生的目的，就是要學習如何與人溝通、節省工作時間，因此利用語音辨識技術建立人與機器之間的溝通橋樑【1】。由於訊號處理、演算法和電腦硬體設備的進步及語音辨識技術在過去的十到二十年間確實在許多方面均有長足的進展，例如資料驅使（Data driven）方法、聲學模型和語言模型建立方式以及基於動態編輯程序（Dynamic Programming-based）之搜尋方法等【2】。

近年來語音辨識技術已臻於成熟，但就目前為止，對於更接近於人們日常生活的自發性語音（Spontaneous Speech）的辨識正確率卻非常低。在自發性語音裡，有朗讀式語音（Read Speech）所沒有的口語化現象，例如：口吃（stutter）、重複詞語（repetition）、修正詞語（repair）、重開始詞語（restart）、鼻音化（nasalized）、發音偏差（inappropriate）、音節合併（syllable contraction）...等，這些現象造成比一般朗讀式語音辨識率低的主要原因。本論文將以現代漢語口語對話語料庫為對象建立其基本聲學模型，分析辨識錯誤原因，以改善自發性語音之辨識率。

1.2 研究方向

目前對自發性語音辨識相關研究有使用Decision tree（DT）與maximum entropy model（Maxent）的結合（DT- Maxent）來偵測不流暢現象之中斷點【3、4】、使用條件隨機域來偵測不流暢現象之中斷點【5】、使用Kernel Principle Component Analysis（KPCA）來偵測發音變異的音節【6】、自發性對話語音辨識之初步研究【7】。

本論文，使用國語sub-syllable HMM建立自發性語音基本辨識器，分析其辨

識錯誤原因，試圖改善其辨識率。由於自發性語音受前後文影響比起一般朗讀式語音大，因此採用Syllable HMM可以提高辨識率3.3%，分析其Deletion錯誤，試圖改善其辨識率。

1.3 章節概要

本論文共分為五大章，而各章的內容概要如下:

第一章 緒論

介紹研究動機、方向及章節概要

第二章 現代漢語口語對話語料庫之介紹與統計

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 MCDC 基本語音辨識實驗

介紹使用的特徵參數、國語 sub-syllable HMM 之建立、辨識結果後之錯誤分析，可能的改善方法，以及語料錯誤的檢查。

第四章 Syllable HMM 之建立

使用 syllable HMM，取代國語 sub-syllable HMM，實驗結果發現 Deletion 錯誤還是太多，嘗試使用 skip state syllable HMM，來改善 Deletion 錯誤的問題，但改善不多，分析其 Deletion 錯誤。

第五章 結論與未來展望

第二章 現代漢語口語對話語料庫之介紹與統計

語音辨識系統要實際應用，必須考慮口語化語音，而自發性語音辨識系統也變的越來越重要，自發性語音多了許多口語現象，如呼吸聲、笑聲...等，是朗讀式語音所沒有的。而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）。本章節將介紹此語料庫，包括音檔格式、文字轉寫格式、如何處理 MCDC 語料、如何標示 MCDC 口語現象，以及 MCDC 語料相關統計。

2.1 MCDC 之簡介

MCDC 語料是由中央研究院語言學研究所在 2000～2002 年間所錄製【8】，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，找 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，因此拿 8 段對話當語料，其中有 16 位語者（9 位女性、7 位男性）兩兩互相交談。依照 8 段對話（每段約 60 分鐘）整理出對話主題，如表 2.1 所示。

表 2.1：對話主題

對話序號	長度(分鐘)	發音人：語者編號	聲道(L/R)	對話主題	子音檔(n)範圍
mcdc-01	61	MISC-08-male-25	*R	工作、休閒活動、經濟、開車	01~20
		MISC-07-female-29	L		
mcdc-02	63	MISC-10-male-35	R	休閒活動、經濟、工作、性別、政治	01~22
		MISC-09-female-37	*L		
mcdc-03	61	MISC-12-female17	R	家庭、學校、購物、生涯規劃、明星	01~21
		MISC-11-female16	*L		
mcdc-05	63	MISC-15-male-40	L	工作、家庭、社會階級、保險、歷史、省籍情節、名人	01~20
		MISC-16-female-46	*R		
mcdc-09	66	MISC-23-female-30	R	工作、旅行、生活態度、環保、健康	01~21
		MISC-24-female-35	*L		
mcdc-10	54	MISC-26-male-23	*R	電影、政治、軍隊、捷運、學校、經濟	01~18
		MISC-25-male-35	L		
mcdc-25	55	MISC-57-male-43	L	交通、工作、小孩、旅行、電腦、管理	01~19
		MISC-58-female-45	*R		
mcdc-26	46	MISC-60-male-24	*R	工作、求職、家庭、車禍、休閒活動、學英文、婚姻、軍隊	01~16
		MISC-59-female-37	L		
*代表該音檔首位發音者所使用的聲道					

2.2 音檔格式說明

2.2.1 原始音檔處理方式

數位錄音採用 SONY TCD-D10 PRO II DAT 的數位錄音機，使用 Audio Technica ATM 33a 手持式麥克風，以取樣頻率 44.1 kHz 將兩位發音人的語料分別錄於左右聲道，錄音地點為普通房間。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。

2.2.2 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)，故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，並將原始雙聲道音檔分割成較小單聲道音檔。其檔名格式，如圖 2.1 所示。最後，利用 Unix 上，sox 音檔轉換程式，進行 Down Sampling 至 16kHz。

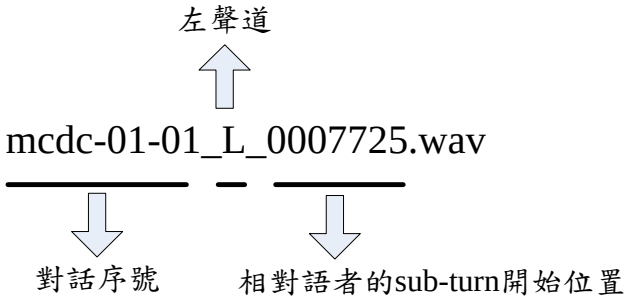


圖 2.1：切割後音檔檔名格式

2.2.3 音檔格式比較

我們對原始音檔格式做了處理，而處理前後音檔的比較如表 2.2 所示。

表 2.2：音檔比較

	原始音檔	處理後音檔
Sampling Rate	44.1kHz	16kHz
Channel	Stereo	Mono

2.3 文字轉寫格式說明

MCDC 所使用的轉寫格式是一種標籤式的語言格式，這種格式是有點類似 XML 語法，結構上大致是以一個 sub-turn 來當作一個單元。如圖 2.2、表 2.3 是其中一個範例。

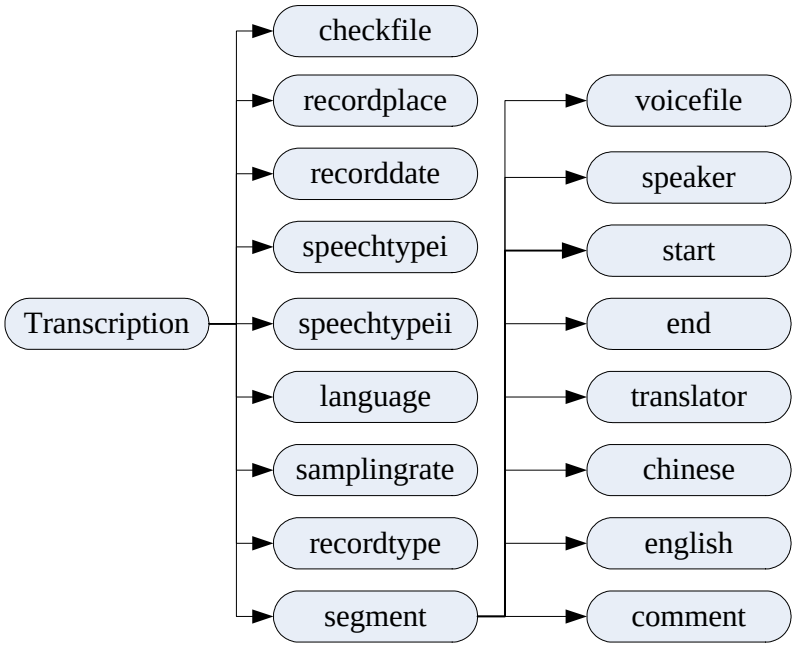


圖 2.2：標籤式的語言格式

表 2.3：文字轉寫範例

(sub-turn 開始)
<voicefile>D:\MCDC\stereo_01\mcdc-01-01.wav (音檔名稱)
<speaker>MISC-07-female-29 (發音者)
<start>045432 (音檔開始時間)
<end>045767 (音檔結束時間)
<translator>Fen (文字轉寫人)
<chinese> (內容文字標示)
對
</chinese>
<english> (內容漢語拼音標示)
dui4
</english>
<comment> (註解標示)
</comment>
(sub-turn 結束)

2.4 MCDC 口語之轉寫標示

自發性口語語音與朗讀式語音的不同在於自發性口語語音中，有許多不流利現象，因此在 MCDC 口語之轉寫標示中，多了許多標記符號，將簡述如下：

2.4.1 非語音現象(Paralinguistic Phenomena)

非語音現象就性質上，可以分兩大類：

(1) 無伴隨語言內容之非語音聲但確定是由人所發出的聲音。

EX：(inhale)我目前是從事…

(inhale)經過處理標示，@inhale。

(2) 無伴隨語言內容非語音聲且確定非人所發聲。

EX：(noise in room) 對對對

(noise in room)經過處理標示，@noise

2.4.2 不確定字/音(Uncertain)

不確定音，就性質上，可以分兩大類：

(1) 文字轉寫人可以依據前後語意，大略猜出。

EX：NA 賴先生呢您從事什麼工作

NA lai4 xian1 sheng1 [nen2] nin2 cong2 shi4 shen2 me5 gong1 zuo4

[nen2]經過處理標示，~nen。

(2) 文字轉寫人無法依據語意猜測出對應字詞，但是可以用漢字拼音紀錄此發音。

EX：(inhale) 那是 (uncertain)…

(inhale) na4 shi4 [f]

[f] 經過處理標示，~f。

2.4.3 語助詞(Marker)

語者本身在對話中的慣用性插語，這些習慣性插語有其基本詞彙意義。但是在對話中的習慣性插語已不保有其原本完整語意。例如：作用於口語中語者意欲保留其說話權且又需要緩衝時間去思考組織其想說的話之句子，此時習慣插語“那”便經常被使用。

EX：NA 你們公司在哪…

2.4.4 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是同意之類的情況。在對話中出現的感嘆詞有四種：

(1) 有相對應國字的感嘆詞：

EX：外貿 A…

(2) 無相對應國字的感嘆詞：

EX：EI (clear throat) 你好…

(3) 源於台語的感嘆詞：

EX：EIN 他是賣比較貴…

(4) 其他感嘆詞，如嗯哼哼：

EX：MHM 對 A 你去了那邊…

2.4.5 語言轉換 (Code Switching)

當語者使用漢語以外的語言（如閩南語、客家語、英語...等）

EX：ok (breathe) (pause) 瞭解 (breathe)

2.5 MCDC 之相關統計

在第 2.2.2 節裡，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，將整個 MCDC 語料庫中，每段對話 sub-turn 總和統計出來，如表格 2.4。在文字轉寫的部份，可以依據 MCDC 語料庫中，文字轉寫的標籤(tag)，取出想要處理的語音資訊，如第 2.3 節和第 2.4 節說明。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，有三小類：

(a) Particles (含感嘆詞及語助詞)

(b) Paralinguistic Phenomena (含所有非語言現象)

(c) Uncertain (含不確定字/音)

EX：(inhale) NA 如果從南港過去要怎麼去 (breathe)

其中：Paralinguistic Phenomena：(inhale)、(breathe)

Particle：NA

411 syllable：如果從南港過去要怎麼去

表 2.4：sub-turn 統計

對話序號	音檔長度(分鐘)	sub-turn 數量
mcdc-01	61	871
mcdc-02	63	1397
mcdc-03	61	1109
mcdc-05	63	911
mcdc-09	66	669
mcdc-10	54	569
mcdc-25	55	683
mcdc-26	46	866
總和	469	7075

將這兩大類的語音統計出來如表 2.5，正規性語音僅佔約 **81%**，由此可見自發性口語語音中的不流暢現象，將嚴重影響自發性口語語音之自動辨識系統的效能。

表 2.5：Syllable 及語音現象出現次數統計

	正規性語音	非正規性語音		
	411 syllable	Particle	Paralinguistic Phenomena	Uncertain
字數	127678 (81%)	10620 (7%)	12563 (8%)	6511 (4%)
總字數	157372			
音檔數	6656			
時間長度	10.2 (hours)			

針對 sub-turn 與 syllable 數量的關係，做一個 syllable 數量的分布圖，如圖 2.3 所示。發現 MCDC 語料是長短句分布的語料。

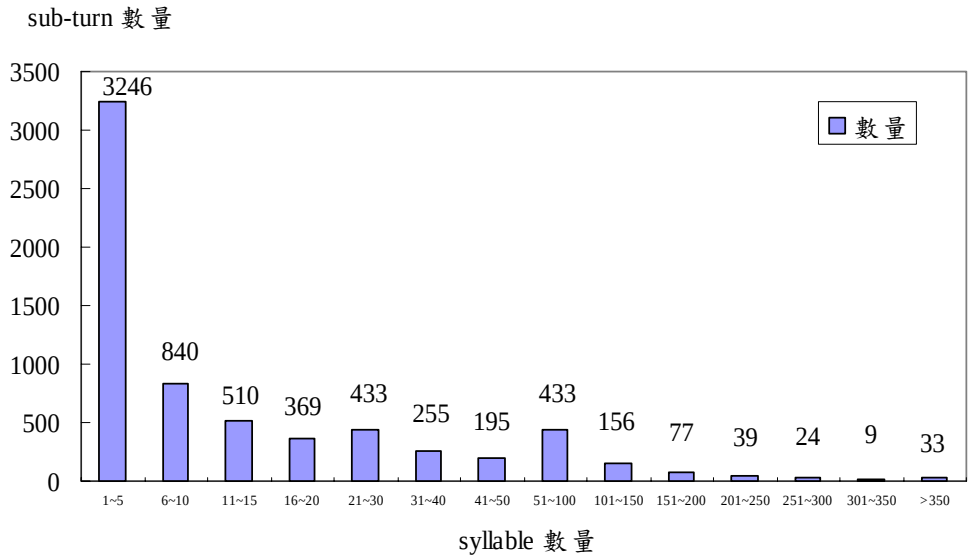


圖 2.3：Syllable 數量分佈圖

第三章　MCDC 基本語音辨識實驗

我們使用的辨識方法是隱藏式馬可夫模型 (Hidden Markov Model，HMM)，這種機率模型用來模擬口腔聲道變化的過程最適合。我們使用的工具是採用英國劍橋大學開發的 HMM Tool Kit (HTK)【9】。本章節依序介紹使用的特徵參數、國語 sub-syllable HMM 模型之建立、辨識結果之錯誤分析，可能的改善方法，以及語料錯誤的檢查。

3.1 訓練語料與測試語料

本實驗採多語者辨識系統，是指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。從 MCDC 語料 8 段對話中，平均各選取 9/10 的 sub-turn，來當作訓練語料，剩餘 1/10 語料去除整句無 411 syllable 的句子，當作測試語料，對音檔的時間、句子、及文字做一個統計，如表 3.1、表 3.2。

表 3.1：訓練語料統計

	正規性語音	非正規性語音			
	411 syllable	Particle	Paralinguistic Phenomena	Uncertain	Filler	Eng
字數	104719 (79.7%)	9690 (7.4%)	11295 (8.6%)	3736 (2.8%)	1745 (1.3%)	157 (0.1%)
總字數	131342					
音檔數	6168					
時間長度	9.08 (hours)					

表 3.2：測試語料統計

	正規性語音	非正規性語音			
	411 syllable	Particle	Paralinguistic Phenomena	Uncertain	Filler	Eng
字數	14702 (83.3%)	912 (5.16%)	1205 (6.82%)	557 (3.15%)	263 (1.49%)	20 (0.1%)
總字數	17659					
音檔數	447					
時間長度	66.16 (minutes)					

由於對於原始 MCDC 的語料中做了調整，所以表 3.1 和表 3.2 的總和與表 2.5 不一致，其中訓練語料移除了音檔錄音品質較差的部份，測試語料將不選應答內未含 411 syllable 的部份。

3.2 聲學模型的建立

3.2.1 特徵參數

我們使用的參數為 MFCC (Mel Frequency Cepstral Coefficients，梅爾倒頻譜參數)，成份包括 12 維 MFCC 加上能量共 13 維，取 Delta 和 Delta-Delta，加起來總共 39 維，而能量的大小對語音辨認不重要，因此省略能量參數，而得到 38 維的語音特徵參數，表 3.3 為參數抽取主要設定檔。

表 3.3：參數抽取設定檔

Sample Rate	16 kHz
Frame Size	32 ms
Frame Shift	10 ms
Filter Bank	24 個三角帶通濾波器

3.2.2 國語 sub-syllable HMM 之建立

利用一個已經建構良好的模型－朗讀式（read speech）語料(TCC300)所訓練的模型，來協助產生 MCDC 語料聲學模型。考慮到 MCDC 語料中，有許多特殊模型(Particle 模型、Paralinguistic Phenomena 模型、Uncertain 模型)，是在 TCC300 語料的聲學模型不存在，所以在 MCDC 中有相似 411 syllable 的 Uncertain 模型，將其退化至近似 TCC300 語料的聲學模型，Particle 模型與 Paralinguistic Phenomena 模型則是使用實驗室先前所建立 MCDC 之 HMM，利用這些聲學模型去做時間切割的動作，將獲得具有時間切割資訊的 MCDC 語料文字轉寫。MCDC 語料文字轉寫有時間切割資訊，所以可以利用文字的時間資訊（某一段切割時間），來對該段文字作 Isolated Unit Training 的動作，其國語 sub-syllable

HMM 之建立流程圖如圖 3.1 所示。表 3.4 列出每一個 Model 的 state 數量，每一類國語 sub-syllable 的 Model 數量。

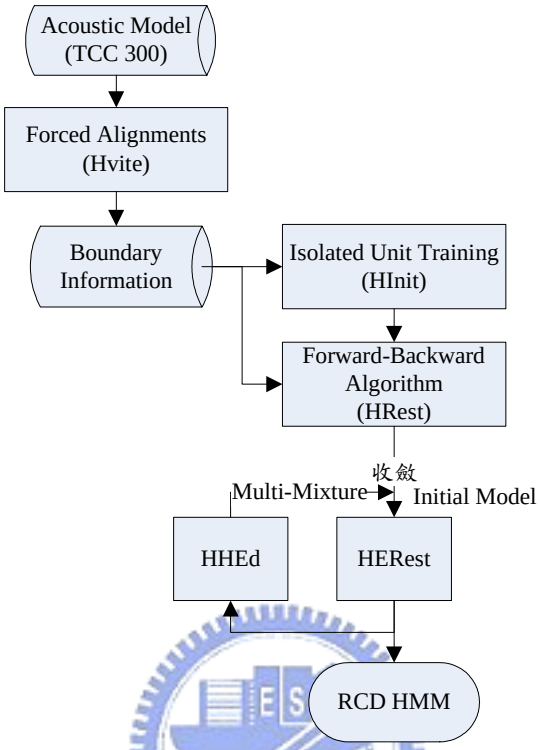


圖 3.1：國語 sub-syllable HMM 流程圖

表 3.4：Model 數量

	state 數量	model 數量
sub-syllable_Initial	3	97
sub-syllable_Final	5	37
Paralinguistic	3	11
Particle	3	37
Uncertain	3	73
Filler	1	1
Eng	3	1
silence	3	1
sp	1	1

- 在 MCDC 語料中，Initial 和 Final model 各缺少 3 種分別是 c_o、n_o、s_o、eh、yai、yo。

3.3 實驗結果

表 3.5 是利用所建立的國語 sub-syllable HMM，得到全部 syllable 的辨識率，表 3.6 是利用所建立的國語 sub-syllable HMM，得到辨識結果，辨識結果與參考答案拿掉 Particle、Paralinguistic Phenomena 及 Uncertain…等 syllable，只有留下 411 syllable 的辨識率。

表 3.5：All Syllable 辨識率

	Correct	Accuracy	Deletion	Substitution	Insertion	Total
syllable	46.66%	43.56%	16.57%	36.77%	3.09%	17659

表 3.6：Only 411 syllable 辨識率

	Correct	Accuracy	Deletion	Substitution	Insertion	Total
syllable	49.67%	43.83%	14.28%	36.05%	5.84%	14685

3.3.1 錯誤分析

實驗列出基本辨識率，可以拿這結果和 read speech 語料庫(TCC-300)的辨識結果做比較，TCC-300 的基本辨識率為 67.5%，比 MCDC 實驗中的辨識率高出約 23.94%，這樣的情況，可以預知，因為自發性的語料中，包含了許多語者口語特殊現象，也就造成 Insertion、Deletion、Substitution 皆高於朗讀式語料，進而使辨識率下降。對於基本辨識之混淆矩陣(Confusion Matrix)做細部的分析，將 syllable 音節分成四類：411 syllable、Uncertain、Particle、Paralinguistic Phenomena。以便分析高錯誤率的原因，列於表 3.7。

表 3.7：國語 sub-syllable HMM Confusion Matrix

Ans \ Rec	411	Uncertain	Particle	Paralinguistic	Deletion	Total	%
411	81.2%	0.44%	1.08%	1.14%	15.83%	14702	83.25%
Uncertain	54.94%	22.08%	3.77%	2.51%	16.7%	557	3.15%
Particle	32.02%	0.11%	44.3%	5.7%	17.87%	912	5.16%
Paralinguistic	8.88%	0.17%	2.82%	64.48%	23.57%	1205	6.82%
Insertion	53.85%	0.37%	13.96%	31.87%		546	

下面我們將就各項錯誤之原因分析如下 ：

3.3.1.1 Uncertain 取代型錯誤分析

由表 3.7 可以看出互相辨識情形嚴重的就是 411 syllable 和 Uncertain，但是 Uncertain 也的確是特性最接近 411 syllable 的一種現象，因為它只是發音錯誤的音而已，而發音錯誤的可能情況有無限多種，很難用幾個模型來精確來描述，因此做一個實驗，不將 Uncertain Model 加入辨識器來辨識，於是將參考的轉寫內容(即是比對答案)中，Uncertain 的字，必須將其用相近 411 syllable 取代。Uncertain 共有 557 個，10 個用相近 Particle 取代，而其他 547 個則用相近 411 syllable 取代。表 3.8 是全部 syllable 的辨識率、表 3.9 是拿掉 Particle、Paralinguistic Phenomena 及 Uncertain…等 syllable，只有留下 411 syllable 的辨識率。

表 3.8：All Syllable 辨識率 (Uncertain 退化後)

	Correct	Accuracy	Deletion	Substitution	Insertion	Total
syllable	47.73%	44.25%	15.93%	36.34%	3.78%	17659

表 3.9：Only 411 Syllable 辨識率 (Uncertain 退化後)

	Correct	Accuracy	Deletion	Substitution	Insertion	Total
syllable	49.69%	45.25%	13.86%	36.45%	4.44%	15234

由表 3.5 和表 3.8 比較，我們可以知道用相近 411 syllable 取代 Uncertain 後，辨識率大約提昇 0.69%。由表 3.6 和表 3.9 比較，我們可以知道用相近 411 syllable

取代 Uncertain 後，411 syllable 辨識率大約提昇 1.42%，411 syllable 辨識率提升的幅度是比較令人驚訝的，仔細觀察發現，Insertion 錯誤突然少 1.4%，這代表有可能計算辨識率時答案與辨識結果對齊不一。進一步分析 Confusion Matrix，如表 3.10。

表 3.10：國語 sub-syllable HMM Confusion Matrix (Uncertain 退化後)

Ans \ Rec	411	Particle	Paralinguistic	Deletion	Total	%
411	82.04%	1.29%	1.38%	15.23%	15249	86.35%
Particle	31.24%	45.77%	5.86%	17.34%	922	5.22%
Paralinguistic	8.96%	3.07%	65.45%	22.74%	1205	6.82%
Insertion	52.12%	14.66%	33.22%		614	

由表 3.10 得知，將相近 411 syllable 取代 Uncertain 之後，411 syllable 被辨識成 411 syllable 為 82.04%，相較表 2.7 的 81.2%提升了 0.84%，而且 Uncertain 只佔了所有 syllable 總數的 3.15%，因此由表 3.5 和表 3.8 比較，用相近 411 syllable 取代 Uncertain 之辨識率大約提昇 0.69%，由表 3.6 和表 3.9，用相近 411 syllable 取代 Uncertain 之 411 syllable 辨識率大約提昇 1.42%。我們在訓練時加入 Uncertain 模型，事實上可將不可靠之 syllable 去除不用來訓練 411 syllable model，可獲得較佳之 411 syllable model，但在辨識時，Uncertain model 應不加入辨識器，因為 Uncertain 基本上只是不可靠 411 syllable，因此之後實驗都不將 Uncertain model 加入辨識器。

3.3.1.2 Particle 取代型錯誤

由表 3.10 發現 411 syllable 與 Particle 的互相辯識情形也是很嚴重，尤其是 Particle 常常會辨識錯誤成 411 syllable，其實有很多 Particle 其發音很像相近 411 syllable，幾乎是一模一樣。因此尋找出 Particle 出現次數較多次與對應 411 syllable 也出現較多次，來分析 Particle 與同音 411 syllable 的特性，圖 3.2 是 Particle 出現次數前八名與同音 411 syllable 出現次數的分布圖。由圖 3.2 先分析出現次數相

似的 Particle 與相對應的 411 syllable，EX：NA 與 na、GE 與 ge、NE 與 ne。

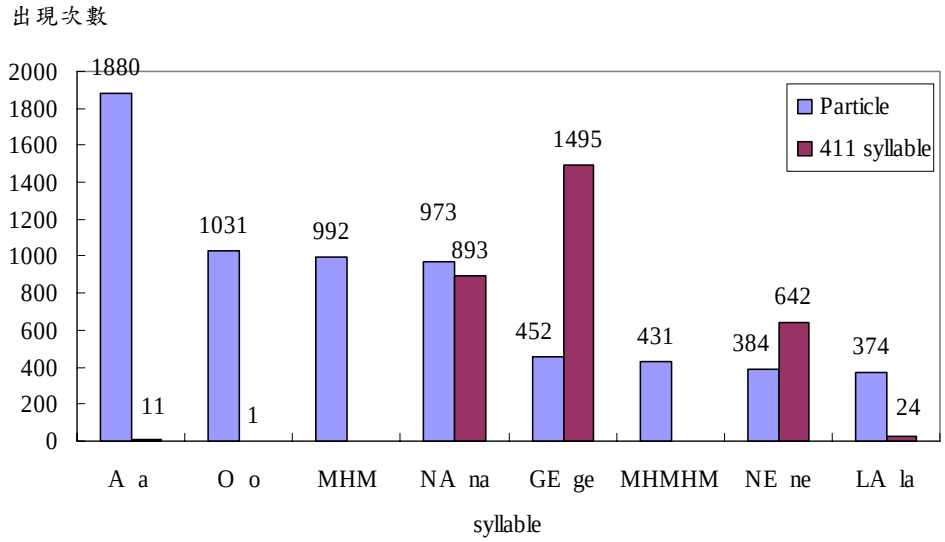


圖 3.2：Particle 出現次數與同音 411 Syllable 出現次數的分布圖

3.3.1.3 Particle model 與同音 411 syllable model 之分析

在 Particle 中有兩類是有同音 411 syllable 分別為慣用插入語和有相對應國字，以下只分析這兩類 Particle 之中 Particle 與同音 411 syllable 出現較多的 Particle 與同音 411 syllable。為了更精確確認，Particle 與同音 411 syllable 是否會混淆，在本節將對訓練語料分析。我們假設已知 Particle 與同音 411 syllable 之時間切割位置，對它們做辨識，觀察它們的辨識結果是否會混淆。首先利用 HMM 對訓練語料做 Forced Alignments 取得單一 syllable 的時間切割位置，對該時間位置抽取單一 syllable 的 38 維 MFCC 參數，然後使用辨識器辨識單一 syllable 的 MFCC，如圖 3.3 示意圖。

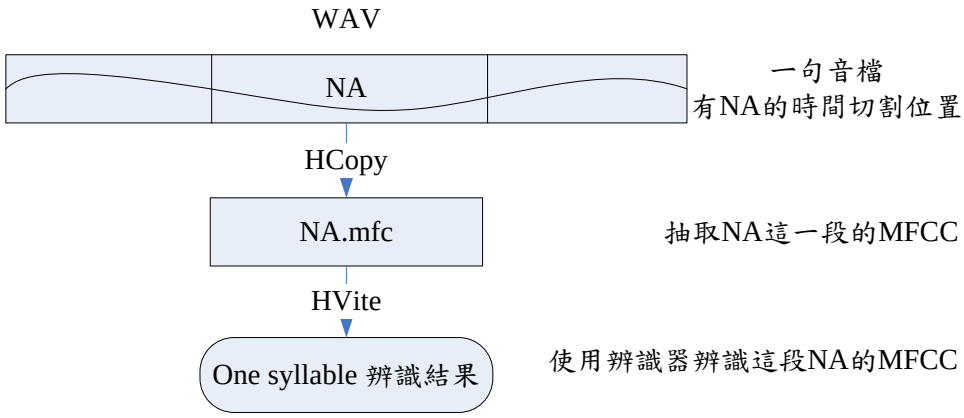


圖 3.3：已知 syllable 時間切割位置辨識分析示意圖

NA 在訓練語料出現 871 次，做 NA 已知 syllable 時間切割位置辨識，比較 NA 與 na 的辨識分數。圖 3.4 左邊是 NA 已知 syllable 時間切割位置辨識，參考答案是 NA，橫軸是 NA 的辨識分數，縱軸是 na 的辨識分數。na 在訓練語料出現 795 次，做 na 已知 syllable 時間切割位置辨識，比較 na 與 NA 的辨識分數。圖 3.4 右邊是 na 已知 syllable 時間切割位置辨識，參考答案是 na，橫軸是 na 的辨識分數，縱軸是 NA 的辨識分數。表 3.11 是 NA 與 na 的 Confusion Matrix，表示已知 syllable 時間切割位置互相辨識的情況，other 是辨識錯誤成其他 syllable，由於種類太多，且單一出現次數太少所以以 other 表示。

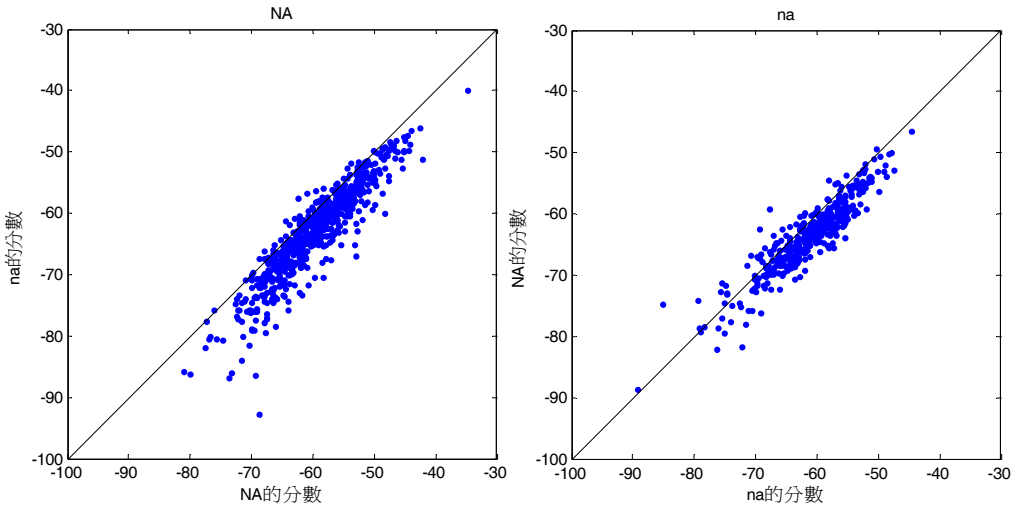


圖 3.4：已知 syllable 時間切割位置辨識，左邊是 NA，右邊是 na

表 3.11：NA 與 na 的 Confusion Matrix

Ans \ Rec	NA	na	other	total
NA	691 (79.33%)	35 (4.02%)	145 (16.65%)	871
na	42 (5.28%)	373 (46.92%)	380 (47.8%)	795

GE 在訓練語料出現 374 次，做 GE 已知 syllable 時間切割位置辨識，比較 GE 與 ge 的辨識分數。圖 3.5 左邊是 GE 已知 syllable 時間切割位置辨識，參考答案是 GE，橫軸是 GE 的辨識分數，縱軸是 ge 的辨識分數。ge 在訓練語料出現 1303 次，做 ge 已知 syllable 時間切割位置辨識，比較 GE 與 ge 的辨識分數。圖 3.5 右邊是 ge 已知 syllable 時間切割位置辨識，參考答案是 ge，橫軸是 ge 的辨識分數，縱軸是 GE 的辨識分數。

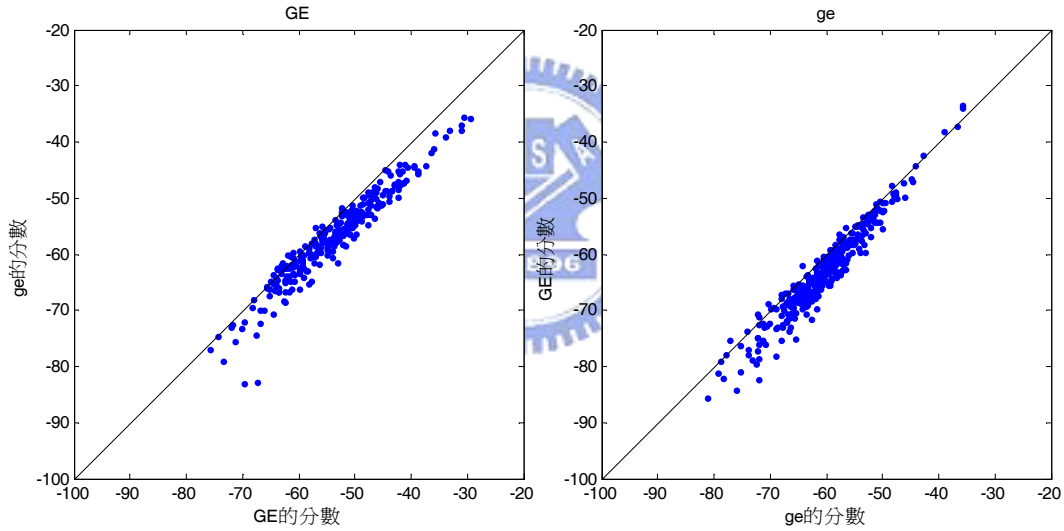


圖 3.5：已知 syllable 時間切割位置辨識，左邊是 GE，右邊是 ge

表 3.12 是 GE 與 ge 的 Confusion Matrix，表示已知 syllable 時間切割位置互相辨識的情況。other 是辨識錯誤成其他 syllable，由於種類太多，單一出現次數太少所以以 other 表示。

表 3.12：GE 與 ge 的 Confusion Matrix

Ans \ Rec	GE	ge	other	total
GE	302 (80.75%)	13 (3.48%)	59 (15.78%)	374
ge	21 (9.29%)	726 (55.72%)	380 (29.16%)	1303

NE 在訓練語料出現 325 次，做 NE 已知 syllable 時間切割位置辨識，比較 NE 與 ne 的辨識分數。圖 3.6 左邊是 NE 已知 syllable 時間切割位置辨識，答案 NE，橫軸是 NE 的辨識分數，縱軸是 ne 的辨識分數。ne 在訓練語料出現 557 次，做 ne 已知 syllable 時間切割位置辨識，比較 NE 與 ne 的辨識分數。圖 3.6 右邊是 ne 已知 syllable 時間切割位置辨識，答案 ne，橫軸是 ne 的辨識分數，縱軸是 NE 的辨識分數。

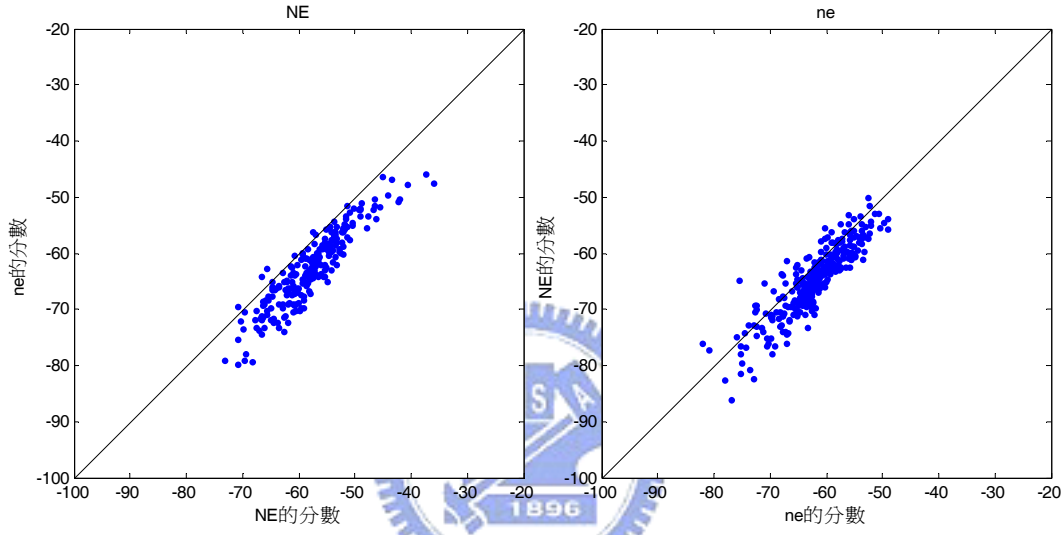


圖 3.6：已知 syllable 時間切割位置辨識，左邊是 NE，右邊是 ne

表 3.13 是 NE 與 ne 的 Confusion Matrix，表示已知 syllable 時間切割位置互相辨識的情況。other 是辨識錯誤成其他 syllable，由於種類太多，單一出現次數太少所以以 other 表示。

表 3.13：NE 與 ne 的 Confusion Matrix

Ans \ Rec	NE	ne	other	total
NE	287(88.31%)	4(1.23%)	34 (10.46%)	325
ne	39(7%)	269(48.29%)	249 (44.7%)	557

由上面分析三個最多的例子，可以得到對於已知 syllable 時間切割位置辨識是可以區分 Particle 與同音的 411 syllable，它們互相辯識不超過 10%，但是對於整個句子辨識時卻還有 31%，我們還是期待找出 Particle 與同音 411 syllable 相異之處。

3.3.1.4 Particle 與同音 411 syllable 的 duration 分佈

NA 與 na 的 duration 分析，如圖 3.7，橫軸為 frames 數量，縱軸為出現次數。

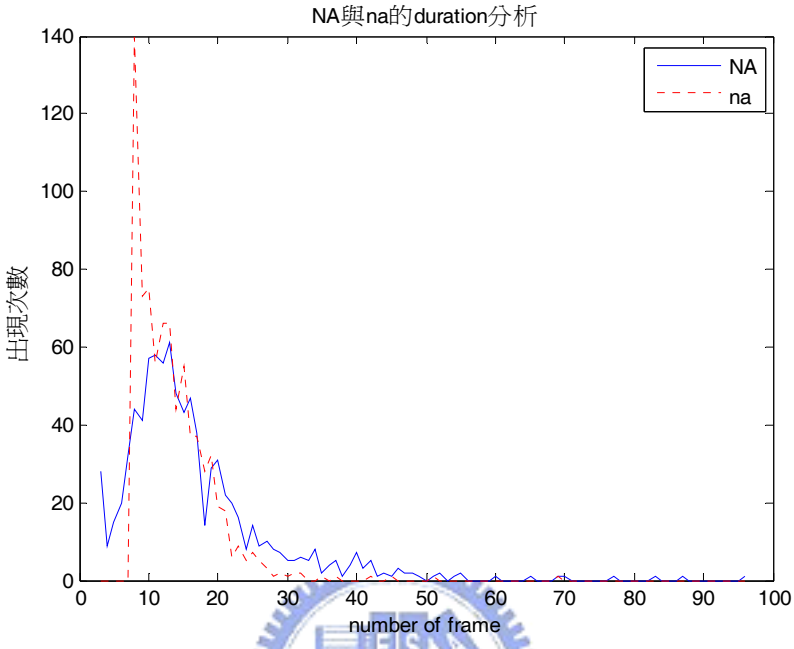


圖 3.7：NA 與 na 的 duration

GE 與 ge 的 duration 分析，如圖表 3.8，橫軸為 frames 數量，縱軸為出現次數。

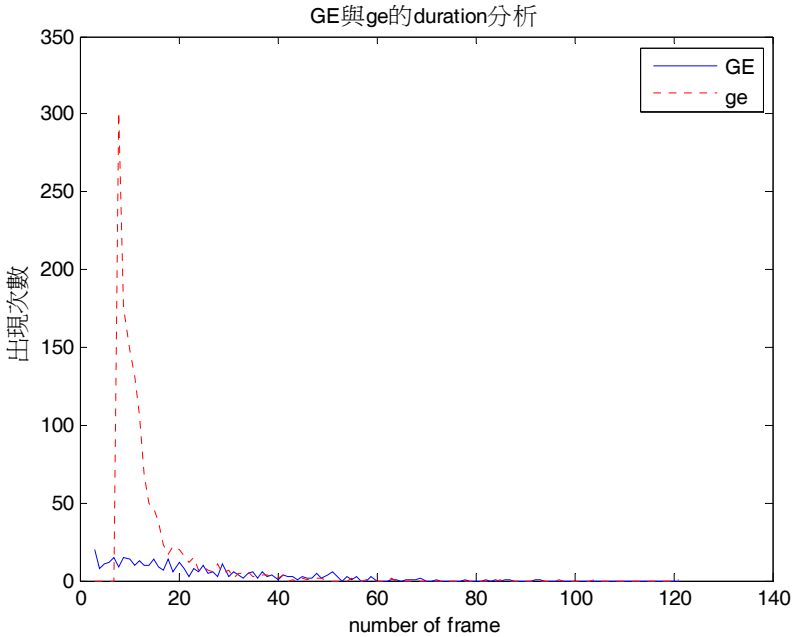


圖 3.8：GE 與 ge 的 duration

NE 與 ne 的 duration 分析，如圖 3.9，橫軸為 frames 數量，縱軸為出現次數。

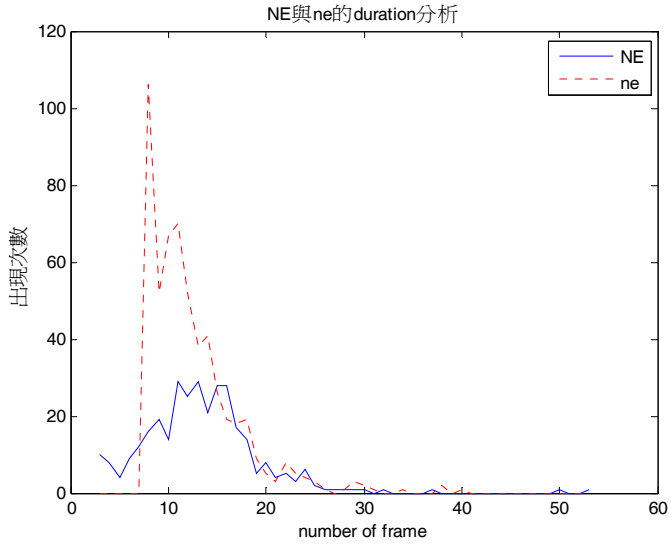


圖 3.9：NE 與 ne 的 duration

由上面分析三個最多的例子，可以得到 Particle 與同音 411syllable 的 duration 分佈是幾乎無法分辨 Particle 與同音 411syllable，分佈太相似了。

3.3.1.5 Particle 前後 silence 的分析

我們假設語者說 Particle 時會停頓的話，也就是如果 Particle 前後有 silence，辨識器可以依照這特性，加以區別分開 Particle 與同音的 411 syllable。觀察 Particle 前後 short pause，如果 short pause 的 duration 有超過 1 個 frame，就認定 Particle 前後有 silence，表 3.14 列出 Particle 前後 silence 的分析。

表 3.14：Particle 前後 silence 分析

	sil_particle	particle_sil	sil_partilce_sil	particle	total
Particle 次數	33.15%	12.86%	10.28%	43.71%	9701
	56.29%				

- sil_particle：代表 Particle 前面有 silence。
- particle_sil：代表 Particle 後面有 silence。
- sil_particle_sil：代表 Particle 前後都有 silence。
- particle：代表 Particle 前後都沒有 silence。

由表 3.14 並發現 Particle 前後 silence 並不明顯，只佔了 56.29%。所以根據 Particle 的特性分類，再觀察其特性。Particle 分成五類：

一、 語者本身在語流中慣用的插入語，這些習慣插語有其基本詞彙意義。例如：NA、NE…

二、 只是回應或同意，有相對應國字。例如：A、AI…

三、 只是回應或同意，無相對應國字。例如：HEN、EI…

四、 只是回應或同意，源於台語。例如：EIN、HAN...

五、 其他感嘆詞。例如：UHN、UHNN…

對於每一類分析 Particle 前後 silence，如表 3.15。

表 3.15：分類 Particle 前後 silence 分析

	sil_particle	particle_sil	sil_partilce_sil	particle	total	%
一、慣用插入語	25.36%	12.79%	3.61%	58.25%	1830	18.86%
	41.75%					
二、有相對應國字	18.57%	18.6%	6.42%	56.42%	4006	41.29%
	43.58%					
三、無相對應國字	35.4%	11.65%	12.01%	40.94%	1082	11.15%
	59.06%					
四、源於台語	42.96%	9.22%	15.53%	32.28%	412	4.25%
	67.72%					
五、其他感歎詞	61.11%	4.43%	20.24%	14.26%	2371	24.44%
	85.74%					

從表 3.15 看起來，只是其他感嘆詞，silence 接 Particle 接 silence 比例比較高，有 81.35% (61.11%+20.24%)，Particle 前面會有 silence，我們可以針對這種 Particle 處理，更改辨識器的其他感嘆詞字典。將在辨識器的字典上，在其他感嘆詞前面加上「sp」，使得辨識器更容易區分出其他感嘆詞。這樣更改影響了 77 個 syllable【0.44%】。辨識率列於表 3.16。

表 3.16：All Syllable 辨識率 (更改字典後)

	Correct	Accuracy	Deletion	Substitution	Insertion	Total
syllable	47.30%	44.31%	16.75%	35.95%	3%	17659

從表 3.8 與表 3.16 的比較辨識率提高了 0.06%，對於這 77 個 syllable 的辨識率是提升了 13.64%，但對於 Particle 與同音 411 syllable 互相辯識情況是沒有幫助，因為想改善的是 Particle 之第一類與第二類有相對應國字。我們發現有一半 Particle 前後有 silence，所以期待考慮加上前後 silence 後，Particle 與相似 411 syllable 的 duration 有相異之處。

3.3.1.6 加上前後 silence 後，Particle 與同音 411 syllable 的 duration 分佈

sp_NA_sp 與 sp_na_sp 的 duration 分析，如圖 3.10 橫軸為 frames 數量，縱軸為出現次數。

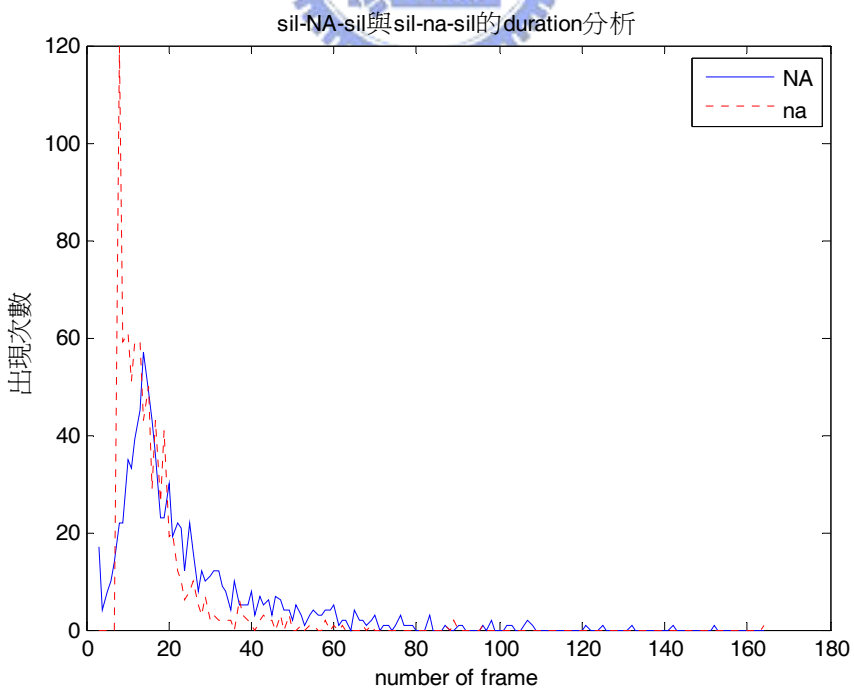


圖 3.10：sp_NA_sp 與 sp_na_sp 的 duration

sp_GE_sp 與 sp_ge_sp 的 duration 分析，如圖 3.11 橫軸為 frames 數量，縱軸為出現次數。

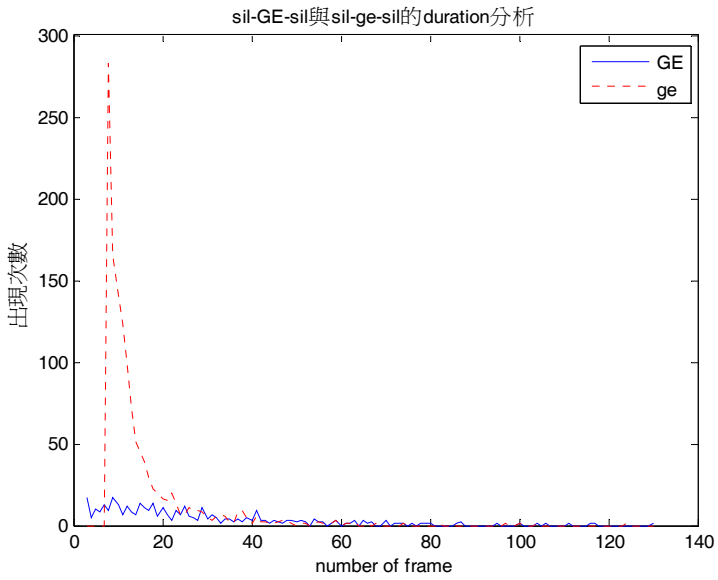


圖 3.11：sp_GE_sp 與 sp_ge_sp 的 duration 分析

sp_NE_sp 與 sp_ne_sp 的 duration 分析，如圖 3.12 橫軸為 frames 數量，縱軸為出現次數。

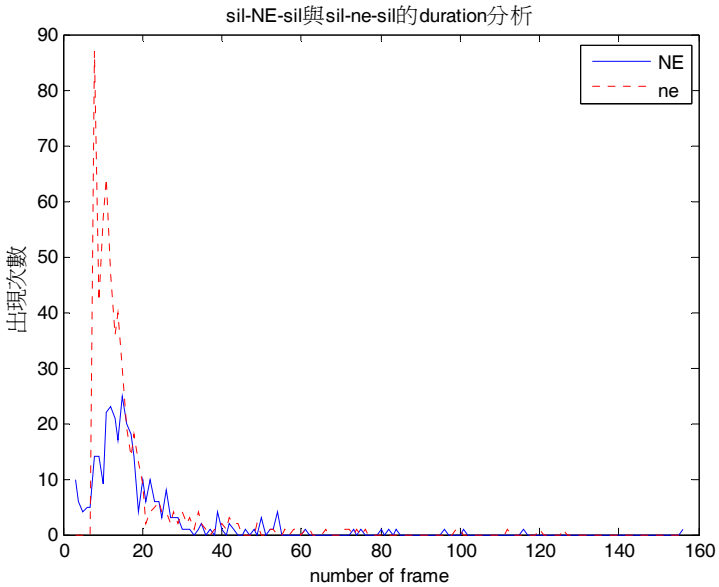


圖 3.12：sp_NE_sp 與 sp_ne_sp 的 duration

由上面分析三個最多的例子，得到加上前後 silence 的 Particle 與加上前後 silence 的相似 411 syllable 的 duration 分佈是也是無法分辨 Particle 與同音 411 syllable，分佈太相似了。

3.3.1.7 調整 Particle 與 411 syllable 發生的機率

由表 3.10 得知 Particle 辨識錯誤成 411 syllable 的數量較 411 syllable 辨識錯誤成 Particle 多，所以調整 Particle 與 411 syllable 發生的機率，提高 Particle 發生的機率，降低 411 syllable 發生的機率，來減少 Particle 辨識錯誤成 411 syllable 的可能性。應該能降低一些錯誤，提高一些辨識率。但是發現 HTK 在答案與辨識結果對齊時，有一些瑕疵，它只依照文字對齊，並沒有依照時間切割位置對齊，也是說它有可能會對齊錯誤，造成 Confusion Matrix 跟實際上有些誤差。圖 3.13 為其中一個例子。圖中 yi 是辨識錯誤成 deletion，下面參考答案沒標示任何 syllable，只是 silence,卻辨識出@INHALE。HTK 它認為 yi 辨識成@INHALE，這樣大大影響 Confusion Matrix 的準確度。這樣造成調整 Partile 與 411 syllable 發生機率的困難度。首先加上文字的時間切割位置資訊，來幫助對齊正確，參考答案之文字使用 Forced Alignments 的時間切割位置，辨識結果之文字使用辨識的時間位置。依照時間位置，調整對齊，以便得到較精準的 Confusion Matrix，如表 3.17。

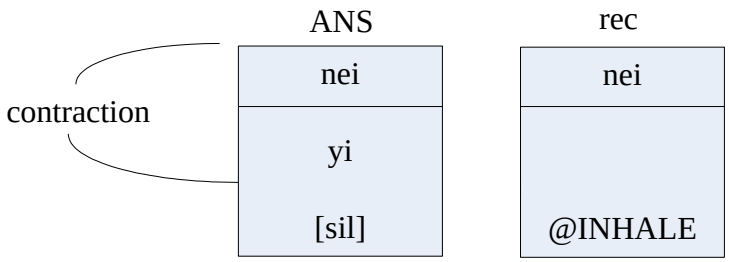


圖 3.13：對齊錯誤示意圖

表 3.17：國語 sub-syllable HMM Confusion Matrix (調整對齊後)

Ans \ Rec	411	Particle	Paralinguistic	Deletion	Total	%
411	82.12%	1.03%	0.75%	16.04%	15249	86.35%
Particle	28.74%	47.94%	3.15%	20.17%	922	5.22%
Paralinguistic	2.57%	1.83%	65.89%	29.63%	1205	6.82%
Insertion	43.56%	14.93%	36.51%		871	

- 其中 411 syllable 辨識成 411 syllable 的辨識率

	Accuracy	total
411	59.22%	12522

- 其中 Particle 辨識成 Particle 的辨識率

	Accuracy	total
Particle	75.57%	442

- 其中 Paralinguistic Phenomena 辨識成 Paralinguistic Phenomena 的辨識率

	Accuracy	total
Paralinguistic	85.14%	794

從表 3.10 與表 3.17 比較我們可以發現 Particle 與 411 syllable 和 Paralinguistic Phenomena 與 411 syllable 的互相辯識錯誤都降低了，但是 Particle 辨識錯誤成 411 syllable 還是高達 28.74%。所以調整 Particle 與 411 syllable 發生的機率。調高 Particle 發生的機率，調低 411 syllable 發生的機率。Particle 發生的機率為 1.0 ，411 syllable 發生的機率為 0.9，表 3.18 是調整 Particle 與 411 syllable 發生機率後辨識率。

表 3.18：All Syllable 辨識率 (調整發生機率)

	Correct	Accuracy	Deletion	Substitution	Insertion	Total
syllable	47.19%	44.32%	16.75%	35.95%	3%	17659

從表 3.8 與表 3.18 的比較辨識率提高了 0.07%，對這提升辨識率並不是很顯著。表 3.19 其 Confusion Matrix，可以發現雖然降低 Particle 辨識錯誤成 411 syllable，但也提高 411 syllable 辨識錯誤成 Particle，所以辨識率上升幅度不高。

表 3.19：國語 sub-syllable HMM Confusion Matrix (調整發生機率)

Ans \ Rec	411	Particle	Paralinguistic	Deletion	Total	%
411	80.46%	1.78%	0.64%	17.08%	15249	86.35%
Particle	23.1%	54.34%	2.49%	20.07%	922	5.22%
Paralinguistic	1.99%	2.57%	62.66%	32.7%	1205	6.82%
Insertion	45.84%	22.19%	31.97%		757	

- 其中 411 syllable 辨識成 411 syllable 的辨識率

	Accuracy	total
411	59.53%	12270

- 其中 Particle 辨識成 Particle 的辨識率

	Accuracy	total
Particle	75.65%	501

- 其中 Paralinguistic Phenomena 辨識成 Paralinguistic Phenomena 的辨識率

	Accuracy	total
Paralinguistic	85.96%	755

由表 3.10 得知 Particle 辨識錯誤成 411 syllable 一直高居不下，如果像之前調整 Particle 與 411 syllable 的發生機率，也只是挖東牆補西牆，並沒有具體的好方法。我們看更仔細表 3.10，Particle 與同音 411 syllable 互相辯識錯誤佔 Particle 與 411 syllable 互相辯識錯誤的 16.29%，其他 83.71%錯誤有其他原因。因為 Particle 與同音 411 syllable 不論聲音或 duration 都很相似，只能靠語意區別，假如對 Particle 與同音 411 syllable 互相辯識，不算錯誤的話，辨識率可從 44.25%提升至 44.69%。

3.4 檢查語料錯誤

根據之前的分析，我們開始懷疑 MCDC 語料的文字轉寫，是否有誤，期待檢查語料對辨識率有幫助。對每一個 syllable 做已知切割位置之辨識，方法類似第 3.3.2.1 節如果參考答案沒在辨識前 N 名出現，懷疑有可能文字轉寫錯誤、Forced Alignment 切割錯誤，也期待這樣的觀察能找出 Contraction 現象特性。表 3.20 是(TOP N)參考答案沒在辨識前 N 名出現的統計，表 3.21 是(TOP N)已人工觀察錯誤分類統計。

表 3.20：參考答案沒在辨識前 N 名出現的統計

訓練語料	TOP 20	TOP 30	TOP 40	TOP 50
411 syllable	4530 (97.36%)	3399 (98.38%)	2751 (98.89%)	2302 (99.18%)
Uncertain	23 (0.49%)	13 (0.38%)	6 (0.22%)	3 (0.13%)
Particle	58 (1.25%)	22 (0.64%)	12 (0.43%)	6 (0.25%)
Paralinguistic	42 (0.9%)	21 (0.61%)	13 (0.47%)	10 (0.43%)
syllable 錯誤數量	4653 (3.53%)	3455 (2.62%)	2782 (2.11%)	2321 (1.76%)
File 數量	1589	1343	1184	1060
syllable total	131714			

表 3.21：觀察錯誤分類統計

訓練語料	TOP 50	Total
syllable 錯誤數量	2321 (1.76%)	131714
已觀察 syllable 錯誤數量	996 (0.76%)	
文字轉寫錯誤	569 (57.13%)	996
contraction 現象	216 (21.69%)	
辨識不出來	190 (19.08%)	
時間切割不好	11 (1.1%)	
念不清楚	10 (1%)	

syllable 錯誤分佈太分散，目前只看有超過 5 個 syllable 錯誤的音檔，其中訓練語料有 2 個音檔完全錯誤，文字與音檔完全對不起來，不列入統計直接踢除，修正語料文字轉寫後，重新得到辨識率於表 3.22。

表 3.22：All Syllable 辨識率 (修正語料後)

	Correct	Accuracy	Deletion	Substitution	Insertion	Total
syllable	48.33%	44.83%	15.77%	35.9%	3.5%	17659

從表 3.8 與表 3.22 的比較辨識率提高了 0.58%，可以提升辨識率，代表訓練出來 HMM 比之前更為好，只有 569 個 syllable 的修正(0.43%)，能提高 0.58 辨識率，代表語料錯誤會影響辨識率。表 3.23 其修正語料文字轉寫後 Confusion Matrix。各個類別 syllable 互相辯識的情形也有降低。

表 3.23：國語 sub-syllable HMM Confusion Matrix (修正語料後)

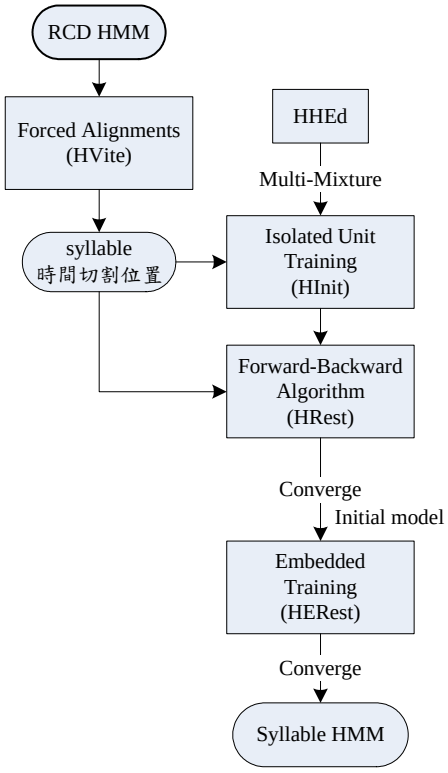
Ans \ Rec	411	Particle	Paralinguistic	Deletion	Total	%
411	82.49%	1.17%	1.27%	15.01%	15249	86.35%
Particle	32.21%	45.12%	5.21%	17.46%	922	5.22%
Paralinguistic	9.38%	2.32%	64.81%	23.4%	1205	6.82%
Insertion	50%	15.7%	34.3%		618	

第四章　Syllable HMM 之建立

自發性語音辨識中，每一個 syllable 會受前後 syllable 影響比一般朗讀式語音辨識大，因此使用國語 sub-syllable HMM 可能較不能描述 syllable 特性。本章節採用 syllable HMM，取代國語 sub-syllable HMM，期待能描述 syllable 更好，實驗結果發現 Deletion 錯誤還是太多，嘗試使用 skip state syllable HMM，來期待改善 Deletion 錯誤的問題，但改善不多，分析其 Deletion 錯誤。

4.1 Syllable HMM

利用國語 sub-syllable HMM 做 Forced Alignments，取得每一個 syllable 時間切割位置，再對其中每段時間切割位置做 Isolated Unit Training 的動作，來建初始每一個 syllable HMM，再利用 Forward-Backward Algorithm 訓練每一個 syllable HMM 至收斂，最後用 Embedded Training 訓練 HMM 至收斂，其流程圖，如圖 4.1 左邊。我們發現語料有些 syllable 出現次過少，無法建立 syllable HMM，這些 syllable 將以 Initial 和 Final phone model 合成 syllable HMM，其流程圖，如圖 4.1 左邊。表 4.1 列出每一類 syllable model 數量，每一種 model state 的數量，表 4.2 列 411 syllable HMM 分佈之統計，表中 syllable HMM 是指能依圖 4.1 左邊流程圖建出 model 的 syllable，Initial-Final 是指 syllable 出現太少只能利用 Initial 和 Final phone model 合成 syllable HMM 的 syllable，no model 是指連 Initial 和 Final phone model 都缺少的 syllable，因此也無法建出 model，在 MCDC 語料中，有 6 個 syllable 是我們常用之 411 syllable 表中未出現之音節。



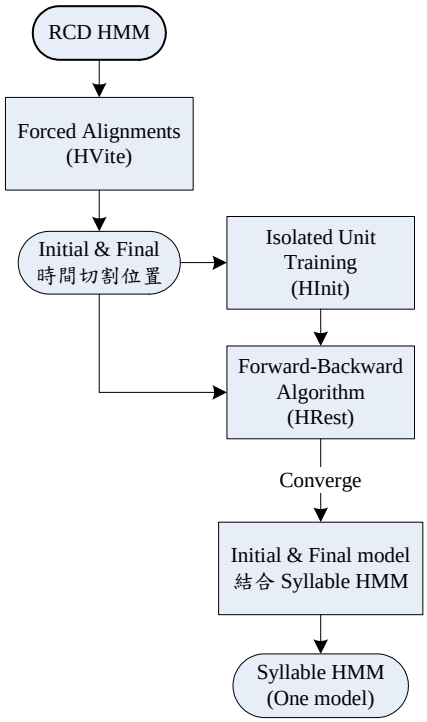


圖 4.1：左邊為 Syllable HMM 流程圖，右邊為 Syllable HMM 流程圖 (出現次數太少)

表 4.1：Syllable HMM 之統計

	state 數量	model 數量
Syllable	8	417
Paralinguistic	3	11
Particle	3	37
Uncertain	3	73
Filler	1	1
Eng	3	1
Silence	3	1
sp	1	1

表 4.2：411 Syllable HMM 分佈之統計

	Model 數量	411 syllable
Syllable HMM	366	411
Initial-Final	45	
No Model	6 (未建 model)	

實驗結果：表 4.3 是 Syllable HMM 的 all syllable 的辨識率。

表 4.3：All Syllable 的辨識率 (Syllable HMM)

	Correct	Accuracy	Deletion	Substitution	Insertion	Total
syllable	51.24%	48.13%	16.22%	32.53%	3.11%	17659

從表 3.21 與表 4.3 的比較辨識率提高了 3.3%，取代型錯誤少 3.3%。這代表 syllable HMM 比國語 sub-syllable HMM 更能精確描述 syllable 特性，表 4.4 其 Confusion Matrix。

表 4.4：Confusion Matrix (Syllable HMM)

Ans \ Rec	411	Particle	Paralinguistic	Deletion	Total	%
411	82.15%	1.7%	1%	14.94%	15249	86.35%
Particle	26.68%	49.89%	3.47%	19.96%	922	5.22%
Paralinguistic	9.46%	2.9%	58.84%	28.55%	1205	6.82%
Insertion	59.74%	12.39%	27.32%		549	

由表 3.22 與表 4.4 比較可以知道 Particle 辨識 Particle 為 49.89%，相較表 3.22 的 45.12%提升了 4.77%，這可以知道取代型錯誤變少了，這相呼應表 3.21 和表 4.3 取代型錯誤下降了 2.81%。但由表 4.3 發現 Deletion 錯誤還是高達 16.22%，對語音辨識來說 16.22%的 Deletion 錯誤太多。因此將統計 MCDC 語料所有語者的說話速度，所有語者平均說話速度為 5.64 (syllable/sec)，可以知道 MCDC 語料語者說話速度是比一般朗讀式語音快，因此 Deletion 錯誤也會比一般朗讀式語音多，我們嘗試 skip state Syllable HMM，期待改善 Deletion 錯誤太多的問題。

4.2 Skip State Syllable HMM

原本我們的Syllable HMM，只准許每一個state跳回自己的state和跳下一個

state，現在允許能跳回自己的state、跳下一個state，以及跳下下一個state，圖4.2為skip state示意圖。如何決定某一個state是否能被skip，我們是依照Forced Alignments切出每一個state的duration來決定。假設某一state的duration 有50%以上是1個frame (10ms)，表示這state可能可以被skip，將設一個轉移機率0.01，從上一個state跳至下一個state，圖4.3為轉移機率示意圖。我們遇到一個問題，假如skip某個state，它擁有syllable不可缺少的音素，將造成它無法跟其他syllable區別。例如：dan (ㄉㄢ) 與dian (ㄉㄧㄢ)，假如介音(ㄧ) 的state被skip，將無法區分dan與dian，會造成取代型錯誤上升，辨識率下降。因此我們將使用KULLBACK-LEIBLER (K-L) distance【10】來判斷相鄰state是否相似，假如某個state與下一個state相似，又它的duration 有50%以上是1個frame (10ms)，才允許這個state可以被skip。

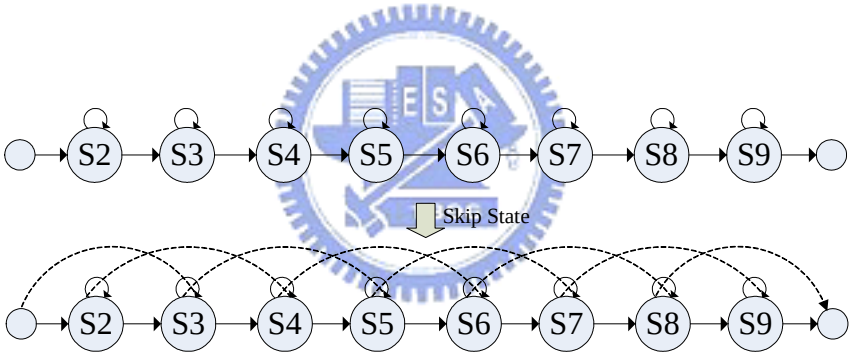


圖 4.2：Skip Sate 示意圖

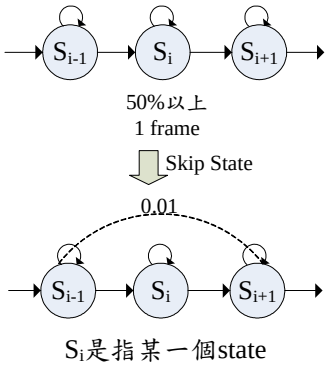


圖 4.3：轉移機率示意圖

4.2.1 K-L Distance

K-L distance 是用來估計兩個機率分佈相差值，我們使用它來估計相鄰兩個 state 是否相似，假設 $p_1(x)$ 和 $p_2(x)$ 為相鄰兩個 state 的機率分佈，從 $p_1(x)$ 來看 K-L distance 為 $D(p_1, p_2)$ 如下數學式子（4.1）、（4.2），從 $p_2(x)$ 來看 K-L distance 為 $D(p_2, p_1)$，平均 K-L distance 為 $D_{avg} = \frac{\{ D(p_1, p_2) + D(p_2, p_1) \}}{2}$ 。

$$D(p_1, p_2) = \int p_1(x) \cdot \log \frac{p_1(x)}{p_2(x)} \cdot dx \tag{4.1}$$

$$D(p_1, p_2) = E\left[\log \frac{p_1}{p_2} \right] \tag{4.2}$$

我們所建 Syllable HMM，每一個 state 是用 Multi-mixtures 高斯混合機率分佈來描述模型 $b(X) = \sum_i c_i \cdot N(\mu_i, \Sigma_i)$，將其簡化為單一高斯機率分佈 $p(X) = N(\mu, \Sigma)$，化簡如下數學式子（4.3）、（4.4）以便於使用 K-L distance 估計相鄰兩個 state 是否相似。例子如圖 4.4、4.5 為一個 syllable（de）（shi），每兩個相鄰 state 之 K-L distance 曲線，橫軸為相鄰 state，縱軸為相鄰 state 之 K-L distance 值。

$$b(X) = \sum_i c_i \cdot N(\mu_i, \Sigma_i) \tag{4.3}$$

$$\Rightarrow p(X) = N(\mu, \Sigma),$$

$$\mu = \sum_i c_i \cdot \mu_i \tag{4.4}$$

$$[\Sigma]_{mn} = \sum_i c_i \cdot ([\Sigma_i]_{mn} + [\mu_i]_m [\mu_i]_n) - [\mu]_m [\mu]_n$$

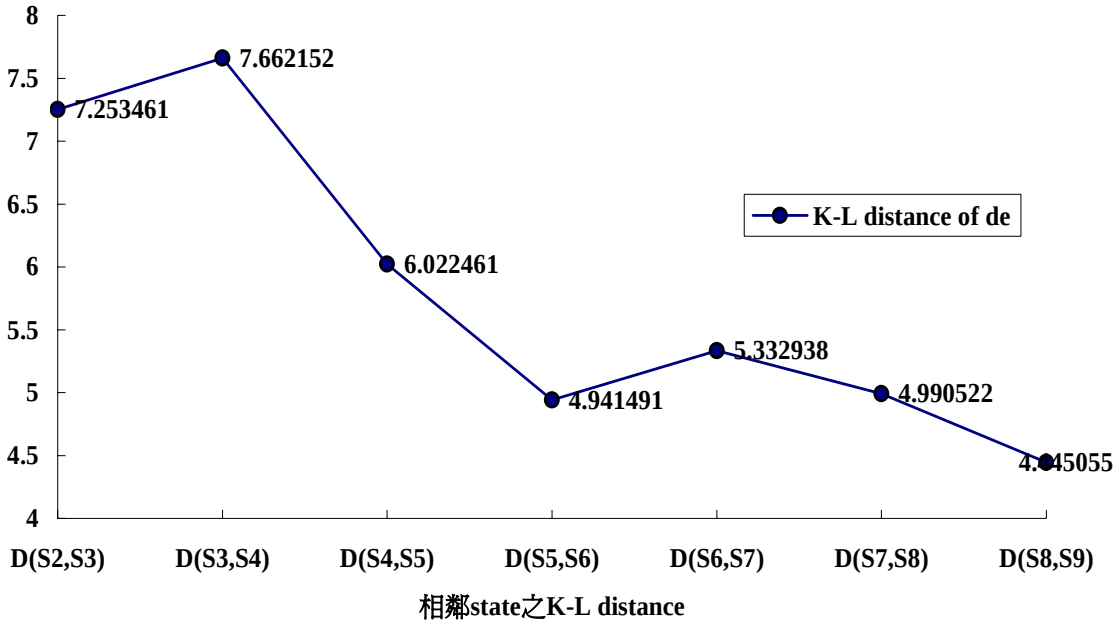


圖 4.4：de，每兩個相鄰 state 之 K-L distance 曲線

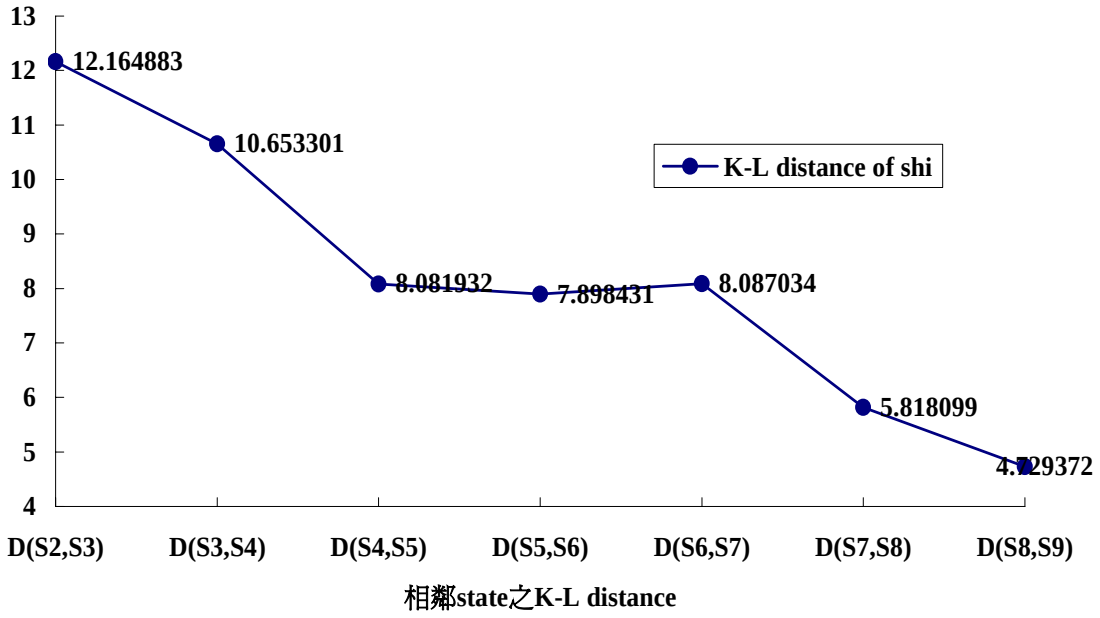


圖 4.5：shi，每兩個相鄰 state 之 K-L distance 曲線

每一 frame 的 log likelihood 值約-54.8，我們認為相鄰兩個 state 的 log likelihood 相差超過-5.48（10%）將認為兩個 state 不相似，因此 K-L distance threshold 值設 5。由圖 4.4、4.5 得知 de 的 state5 與 state6 相似、state7 與 state8 相似、state8 與 state9 相似及 shi 的 state8 與 state9 相似。因此 state 與相鄰下一個 state 之 K-L distance 值小於 5 且 duration 有 50%以上是 1 個 frame (10ms) ，允許這個 state

可以被 skip。

實驗結果：表 4.5 為 skip state syllable HMM 的辨識率。

表 4.5：All Syllable 辨識率 (skip state syllable HMM)

	Correct	Accuracy	Deletion	Substitution	Insertion	Total
syllable	51.27%	48.32%	16.24%	32.49%	2.95%	17659

由表 4.3 與表 4.5 比較，辨識率約提高 0.19%，但 Deletion 錯誤並沒有改善，而是 Substitution 錯誤降低 0.04%、Insertion 錯誤降低 0.16%，表示 skip state 可能無法改善 Deletion 錯誤，表 4.6 其 Confusion Matrix。

表 4.6：Confusion Matrix (skip state syllable HMM)

Ans \ Rec	411	Particle	Paralinguistic	Deletion	Total	%
411	82.43%	1.55%	0.93%	14.87%	15249	86.35%
Particle	25.81%	50%	3.25%	20.93%	922	5.22%
Paralinguistic	8.8%	2.99%	58.67%	29.38%	1205	6.82%
Insertion	60.65%	11.32%	27.64%		521	

由表 4.4 與表 4.6 比較，411 syllable 辨識成 411 syllable 約提高 0.28%，表示 Substitution 錯誤降低，411 syllable 的 Deletion 錯誤下降 0.07%，而所 skip state 的 HMM 幾乎都是 411 syllable 的 HMM，表示 skip state 有改善一點點 Deletion 錯誤，但成效不大。

4.3 Deletion 錯誤分析

我們知道 Deletion 錯誤下降 0.07%，再分析哪些 skip syllable 的 Deletion 錯誤下降了，表 4.7 為 skip state 數量統計。發現會 skip syllable 並不多，分析其 skip state 前與 skip state 後之 Deletion 錯誤與辨識率，如表 4.8。

表 4.7：skip state 統計

	Skip	Total
state 數量	39	3659
model 數量	17	536

表 4.8：skip state 前與 skip state 後之 Deletion 錯誤與辨識率

Skip state 前				Skip 數量	Skip state 後			
Syllable	Correct	Deletion	Total		Syllable	Correct	Deletion	Total
shi	68.66%	16.47%	868	2	shi	69.01%	16.59%	868
de	43.56%	24.16%	567	4	de	44.62%	24.34%	567
yi	54.63%	21.55%	529	5	yi	53.50%	21.93%	529
dui	67.08%	17.70%	322	1	dui	67.70%	17.39%	322
zhe	32.64%	30.56%	144	3	zhe	34.03%	31.94%	144
le	26.23%	27.05%	122	8	le	25.41%	24.59%	122
ni	69.58%	12.92%	240	2	ni	67.50%	15.42%	240
na	31.82%	25.45%	110	1	na	34.55%	24.55%	110
ye	50.96%	23.08%	104	1	ye	50.96%	24.04%	104
ran	51.38%	18.35%	109	2	ran	52.29%	17.43%	109
ren	54.24%	15.25%	118	1	ren	53.39%	16.10%	118
yang	59.83%	13.68%	117	2	yang	57.26%	15.38%	117
er	30%	17.50%	80	1	er	30%	16.25%	80
me	46.07%	14.61%	89	2	me	49.44%	13.48%	89
yin	61.33%	17.33%	75	2	yin	61.33%	20%	75
yu	22.41%	20.69%	58	1	yu	22.41%	16%	58
Total	55.21%	19.76%	3652	38	Total	55.35%	20%	3652
@NOISE 之 skip state 未列入統計								

由表 4.8 得知加入 skip state 轉移機率的 syllable 大部分是語者唸得比較短或是語者較為習慣用詞，表中 skip state 前與 skip state 後，Deletion 錯誤、辨識率大約差不多，有些 syllable 變好有些 syllable 變差。Spontaneous Speech 中，語者說話速度偏快，在某些常用慣用詞，常會有些音節被省略，或被合併。例如「這樣子」語者發出類似「醬子」的聲音，我們稱之為 syllable contraction【11、12】。syllable contraction 現象有三種：

一、清楚可以分辯的 syllable 缺少，像是從原本正常三個字三個 syllable 變成三個字兩個 syllable 或是從原本正常兩個字兩個 syllable 變成兩個字一個 syllable。例如「這樣子」語者發出類似「醬子」的聲音。

二、syllable 雖無缺少，但卻都連在一起，難以切割。

三、syllable 雖無缺少，但 syllable 的音素結構有變。

MCDC 語料標示 syllable contraction 發生的地方約 16239 次，411 syllable 總數為 127678 個，syllable contraction 約佔 12.72%。表 4.9 是列出常發生 syllable contraction 之 syllable 其 Deletion 錯誤的情況。表中第一欄為容易發生 Deletion 錯誤之 syllable，其測試語料中的 Deletion 數量列於第二欄，而在測試語料中出現次數列於第三欄。表中第四欄列出常見 syllable contraction 之詞的例子，其在測試語料 Deletion 錯誤列於第五欄，例如：時候的「時」Deletion 錯誤為 40.85%，其在 MCDC 語料 syllable contraction 的發生率列於第六欄，其在 MCDC 之出現率列於第七欄，MCDC 之出現率為 Syllable Contraction 之詞出現次數乘於 Syllable Contraction 之詞的 syllable 數目除以 MCDC syllable 總數。

表 4.9：常發生 syllable contraction 之 syllable 其 Deletion 錯誤的情況

1	2	3	4	5	6	7
Syllable	測試語料 Deletion	測試語料出現次數	常見 Syllable Contraction 之詞	測試語料 Deletion	Contraction 發生率	MCDC 出現率
shi	144 (16.59%)	868	時候	40.85%	60.43%	0.59%
			就是	20.49%	35.77%	1.13%
			可是	30.65%	46.47%	0.47%
			其實	30.61%	58.03%	0.39%
			是不是	26.92%	70.21%	0.09%
de	138 (24.34%)	567	覺得	44.90%	69.37%	0.72%
			真的	19.57%	33.63%	0.28%
			我(你、他、它)(們)的	36.59%	35.78%	0.26%
			的時候	33.33%	33.33%	0.05%
			好的	37.50%	33.33%	0.08%

yi	116 (21.93%)	529	一個	28.77%	24.62%	0.41%
			所以	36.54%	88.20%	0.59%
			可以	25.58%	61.37%	0.35%
bu	59 (21.61%)	273	不會	41.38%	45.69%	0.25%
			會不會	75%	65.38%	0.05%
			是不是	76.92%	70.21%	0.09%
A	59 (26.94%)	219	對啊	38.95%	62.02%	0.97%
dui	56 (17.39%)	322	對啊	23.16%	62.02%	0.97%
			對對對	20.22%	46.78%	0.33%
			對不對	25%	75.76%	0.13%
you	43 (12.68%)	339	沒有	17.65%	50.25%	0.75%
			有沒有	27.78%	62.16%	0.07%
			有一*	14.89%	45.19%	0.26%
zhe	46 (31.94%)	144	這樣	45.24%	78.22%	0.67%
			這樣子	35.48%	31.94%	0.14%
wo	43 (9.19%)	468	我們	10.91%	55.08%	1.01%
			我覺得	10.64%	22.45%	0.19%
ran	19 (17.43%)	109	然後	18.18%	67.96%	0.79%
hou	34 (16.19%)	210	然後	11.36%	67.96%	0.79%
yin	15 (20%)	75	因為	21.21%	66.88%	0.80%
wei	32 (27.12%)	118	因為	42.42%	66.88%	0.80%
xian	8 (8.99%)	89	現在	16%	65.79%	0.43%
zai	24 (11.71%)	205	現在	34%	65.79%	0.43%
Total	836			27.25%	58.38%	

由表 4.9 觀察的 syllable Deletion 錯誤 836 個，411 syllable 之 Deletion 錯誤 2268 個(表 4.6)，大約佔 36.86%，其 syllable 在常見 syllable contraction 之詞的 Deletion 發生率高於平常，且其 syllable 的 Deletion 發生率大多高於整體平均值，可知 syllable contraction 是造成 Deletion 發生重要原因。下面是實際的例子，如圖 4.6、4.7、4.8、4.9、4.10。

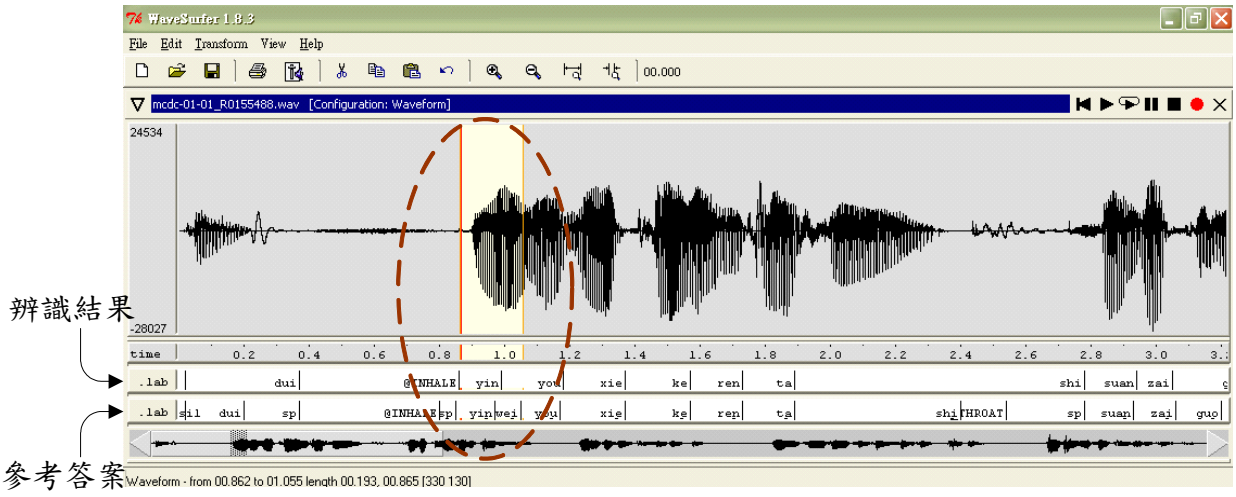


圖 4.6：「因為」Syllable Contraction 而造成 Deletion 錯誤

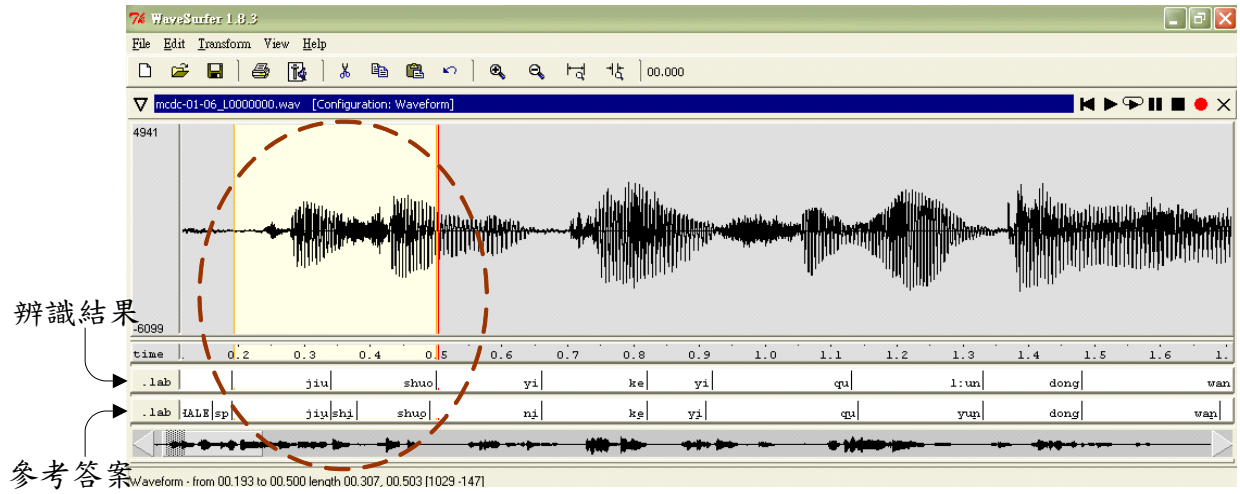


圖 4.7：「就是」Syllable Contraction 而造成 Deletion 錯誤

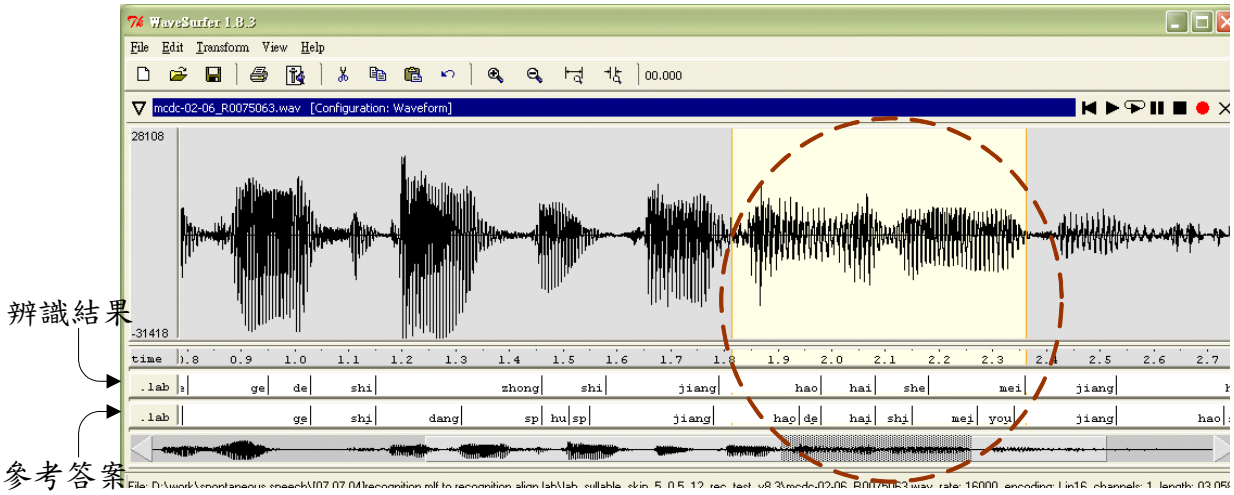


圖 4.8：「好的」「沒有」Syllable Contraction 而造成 Deletion 錯誤

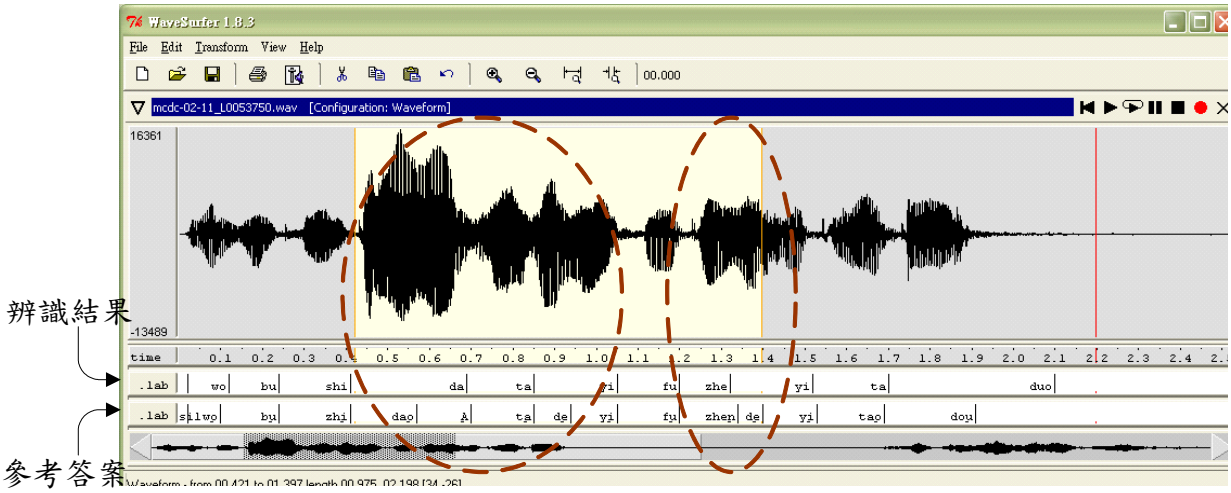


圖 4.9：「知道啊」、「他的」、「真的」Syllable Contraction 而造成 Deletion 錯誤

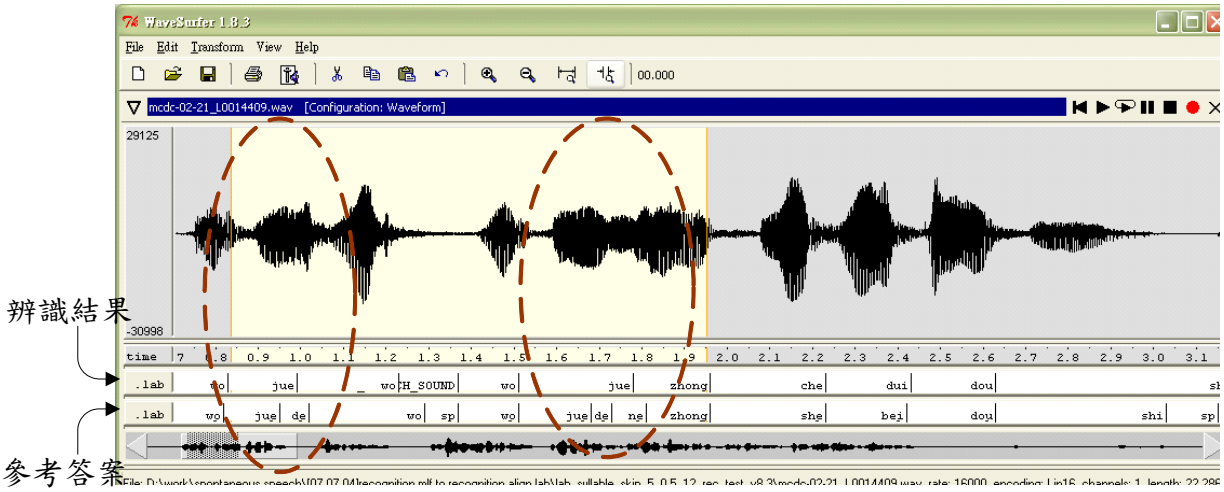


圖 4.10：「覺得」Syllable Contraction 而造成 Deletion 錯誤

第五章 結論與未來展望

5.1 結論

在本論文裡，我們建立國語 sub-syllable Initial + Final HMM 模型，辨識率為 43.56%，將不把 Uncertain HMM 加入辨識器可提高至 44.25%，分析辨識 Confusion Matrix 我們可以得知 Particle 與同音 411 syllable 容易互相辯識，在已知 syllable 時間切割位置辨識得知即使給已知時間切割位置還是會互相辨識約 10%，其辨識分數極為相近，分析 Particle 與同音 411 syllable 之 duration 也相似。我們以語者說 Particle 時，前後會停頓、遲疑的特徵，統計 Particle 前後 silence，此現象約佔所有 Particle 56.29%，並不是特別顯著，再分類 Particle 統計 Particle 前後 silence，發現有同音 411 syllable 之 Particle，此現象佔不到 50%，這樣無法幫助分離 Particle 與同音 411 syllable，再分析加入前後 silence 之 Particle 與加入前後 silence 之 411 syllable 的 duration，也十分相似。分析至此可以說 Particle 與同音 411 syllable 在聲學上是非常相似的，它們只差異於語意。如果 Particle 與同音 411 syllable 互相辨識不算辨識錯誤，辨識率還可以向上提升至 44.69%。使用 TOP N 來偵測語料錯誤，更改 MCDC 語料錯誤將可以其辨識率向上提昇 44.83%，再使用 Syllable HMM，其辨識率高於國語 sub-syllable Initial + Final HMM 模型 3.3%，嘗試使用 skip state 解決 Deletion 錯誤，辨識率提高 0.19%，再分析 Deletion 錯誤，得知 Deletion 錯誤大部分是 Syllable Contraction 造成，我們知道一個 Syllable Contraction 在辨識時，會造成一個 Deletion 錯誤與 Substitution 錯誤。

5.2 未來展望

我們知道 Syllable Contraction 會大大影響辨識率，如果能尋找其特徵或其有用地參數，來偵測 Syllable Contraction 之處，將會對建立聲學模型與辨識大大幫

助，更進一步辨識出 Syllable Contraction 之詞的 syllable，這將能大幅度提高自發性語音辨識率，目前為止只能知道 Syllable Contraction 發生在語者常慣用詞，因為常常使用某種詞，而說話速度變快，因而 Syllable Contraction，但並未找到其他共通特性。

參考文獻

【1】B.H. Juang and S. Furui, Automatic Recognition and Understanding of Spoken Language - A First Step Toward 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】Lin, C.-K., et al., “Important and New Features with Analysis for Disfluency Interruption Point (IP) Detection in Spontaneous Mandarin Speech”, in Proc. of DiSS, 2005.

【4】Lin, C.-K. & Lee, L.-S. Improved Spontaneous Mandarin Speech Recognition by Disfluency Interruption Point (IP) Detection Using Prosodic Features. Proc. Eurospeech’05.

【5】吳維彥，「應用不定長度特徵之條件隨機域於口語不流暢語流修正模型」，國立成功大學資訊工程學系碩士論文，民國九十五年六月。

【6】羅應順，「自發性中文語音基本辨認系統之建立」，國立交通大學電信工程學系碩士論文，民國九十四年六月。

【7】徐文翰，「自發性對話語音辨認之初步研究」，國立交通大學電信工程學系碩士論文，民國九十三年七月。

【8】曾淑娟，劉怡芬，現代漢語口語對話語料庫標註系統說明，中央研究院語言學研究所籌備處，民國九十一年九月。

【9】S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollan, D. Povey, V. Valtchev, P. Wooland, 「The HTK Book（for HTK version 3.4）」.

【10】S. Kullback and R. Leibler, On information and sufficiency, Ann. Math. Statist. , vol. 22, pp. 79–86, 1951.

【11】Shu-Chuan Tseng, Contracted Syllables in Mandarin: Evidence from Spontaneous Conversations, Academia Sinica

【12】Shu-Chuan Tseng, Syllable Contractions in a Mandarin Conversational Dialogue Corpus, Institute of Linguistics, Academia Sinica, Taiwan

附錄

附錄表一

411 syllable 表沒有出現 syllable	
nia	ㄋㄧㄚ
den	ㄉㄣ
dia	ㄉㄧㄚ
kei	ㄎㄟ
lyu	*
nyu	*

附錄表二：Particle 表

Particle	同音 411 syllable	Particle	同音 411 syllable
A	a（ㄚ）	MHMHMHMHM	*
AI	ai（ㄞ）	MHMM	*
BA	ba（ㄅㄚ）	NA	na（ㄋㄚ）
E	e（ㄜ）	NE	ne（ㄋㄜ）
EI	ei（ㄟ）	NEI	nei（ㄋㄟ）
EIN	*	NHN	*
EN	en（ㄣ）	NHNHN	*
GE	ge（ㄍㄜ）	NHNN	*
HAN	han（ㄏㄢ）	NO	*
HEIN	*	O	o（ㄛ）
HEN	hen（ㄏㄣ）	ON	*
HO	*	SHEN	shen（ㄕㄣ）
HON	*	UHN	*
LA	la（ㄌㄚ）	UHNHN	*
MA	ma（ㄇㄚ）	WA	wa（ㄨㄚ）
ME	me（ㄇㄜ）	YA	ya（ㄧㄚ）
MHM	*	YOU	you（ㄧㄡ）
MHMHM	*	ZHE	zhe（ㄓㄜ）
MHMHMHM	*		

附錄表三：Paralinguistic Phenomena 表

@BREATHE	呼吸聲
@CLEAR_THROAT	清喉嚨聲
@COUGH	咳嗽聲
@EXHALE	呼氣聲
@INHALE	吸氣聲
@LAUGH	笑聲
@NOISE	雜訊
@NON_SPEECH_SOUND	非語音聲
@SMACK	咂嘴聲
@SPEECH_SOUND	語音聲
@SWALLOW	吞口水聲