國 立 交 通 大 學

電信工程學系

碩士論文

中文語音屬性偵測之研究

A Research of Mandarin Speech Attribute Detection

研 究 生：許見偟

指導教授：陳信宏　博士

中華民國九十五年八月

中文語音屬性偵測之研究

A Research of Mandarin Speech Attribute Detection

研 究 生：許見偟　　Student：Chien-Huang Hsu

指導教授：陳信宏　　Advisor：Dr. Sin-Horng Chen

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 2006
Hsinchu, Taiwan

中華民國九十五年八月

中文語音屬性偵測之研究

研究生：許見偟　　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

新世代的自動語音辨識技術架構是一個以知識為基礎（knowledge-based），加上資料驅動（data-driven）的模式，其前端為語音屬性與事件偵測器群，藉由抽取不同的語音特徵參數去偵測某一時段中語音的屬性及事件，尋找任何可以提供語音辨識的線索，提供給後級作語音事件及知識整合後，作證據確認及決策，以其能夠突破目前語音辨識的能力與技術。

本論文基於此概念，首先以英文語料製作以高斯混合模型為基礎的語音屬性貝氏偵測器，包含發音方法偵測器及發音位置偵測器，並結合此兩類偵測器的結果觀察是否能夠使效能提昇。再利用中文語料製作語音屬性偵測器，但是中文語料庫並無精確的音素切割位置，因此我們從中文音節的切割位置起始對語料庫作自動切割以求得音素的切割位置，並以此切割位置製作中文語音屬性偵測器。而在中文音節的切割過程中，我們建立一個背景模型（universal background model，UBM）去描述語料庫中的錄製者所發出的呼吸聲、背景雜訊及背景人聲的分佈，以期望能有效的區分語音、靜音與雜訊的部份，使得能夠訓練出更精確的模型。最後再對中、英文的語音屬性偵測器作效能與錯誤分析。

A Research of Mandarin Speech Attribute Detection

Student：Chien-Huang Hsu Advisor：Dr. Sin-Horng Chen

Institute of Communication Engineering

National Chiao Tung University

Abstract

Next generation ASR system is a knowledge-based and data-driven paradigm. It's front-end is the bank of speech attribute and event detectors, and it's function is to detect the speech attributes and events in the speech signal. By organizing the outputs of front-end and knowledge, it would be sent to next stage to make evidence verified and decision. It would be expected to exceed the current state-of-the-art HMM-based ASR.

Based on the concept, firstly, in this thesis we will use English corpus to make speech attribute detectors, including manner of articulation and place of articulation detectors. And the performance of combination of this two class detectors was examined. Secondly, speech attribute detectors of Mandarin was investigated. Without precise Mandarin phoneme labeling, the auto-labeling results from HMM system was used as the labels of Mandarin speech attribute detectors. In the process of auto-labeling Mandarin syllable boundary, we build a universal background model (UBM) to model the breath and noise in the corpus, and it is expected to improve the performance of Mandarin speech attribute detectors. Finally, we would make error analysis of English and Mandarin speech attribute detectors.

誌謝

這篇論文的完成，首先感謝指導教授陳信宏老師以及王逸如老師，這兩年期間在兩位老師的身旁，學習作研究的方法及態度。跟著老師出席研討會，更讓我增廣見聞，了解到研究領域的廣泛。

這兩年在實驗室的期間，特別要感謝智合學長，給予我研究上的建議、幫我找程式的 bug、教我使用工作站上的軟體、偶爾還找你鬼扯...等，對我而言等於是 Google search。感謝遠在日本的性獸學長，愛看黑澀會阿德學長，看起來很斯文的希群學長，愛講八卦的輝哥學長，金髮的 barking 學長，還有上一屆畢業的學長姐們。

感謝這兩年來，夜晚相處白天睡覺的戰友們，愛看日劇的程式小老師國興，愛打電動的鴻彥，帥氣的世哲，失蹤的世帆，愛唱歌的振豐，有猛男身材的東毅，愛讀書的阿勇，眉飛色舞的 paul，有各位兩年的陪伴，一起完成碩士的學業。

除此之外感謝一群很吵的學弟，不離不棄的宏宇，愛回高雄的啟風，籃球很強的小傅，眼睛咪咪的胤賢，操刀煮水餃的銘彥，不定時上台北的友駿，全村最帥的水球存智，大大的接班人獻文，老是要出發才跑廁所的小迷彩，有了各位的吵鬧，讓我度過最後一年的學生生活。另外感謝冠榮、坤浩，及來交大第一個認識的室友泓毅。

最後感謝辛苦的父母及弟弟們，支持我讓我完成碩士的學業。

感謝我的女朋友慧蘋，能陪伴我鼓勵我度過研究所這兩年。

目錄

表目錄

圖目錄

第一章　緒論

1.1 研究動機

回顧現今的語音辨識技術，由早期的學者研究聲學與語言學的規則建立一個以規則為基礎的（rule-based）語音辨識系統，此種辨識系統可以說是以知識驅動（knowledge-driven）的解決方式，但此種系統無法應付複雜的語音變化，因此語音辨識技術繼續進展至以資料驅動（data-driven）的模式，機器由資料中學習，再進展至大詞彙的連續語音辨識（large vocabulary continuous speech recognition，LVCSR）技術，其所依賴的就是大量的語音資料與語言資料。然而這些方法雖然大大改進機器語音辨識的能力，但漸漸的可以發覺到與人類辨識語音的能力相比，仍然有一段不小的差距。因此為了使語音辨識技術有所突破，近年來國際上不斷有學者主張，應該回頭將語音與語言的知識結合進現今的辨識技術，建立一個以知識為基礎（knowledge-based）加上資料驅動的（data-driven）模式，開放測試平台，共享一個合作的設計與評量機制（如圖 1 所示），邁向下一代自動語音辨認技術。

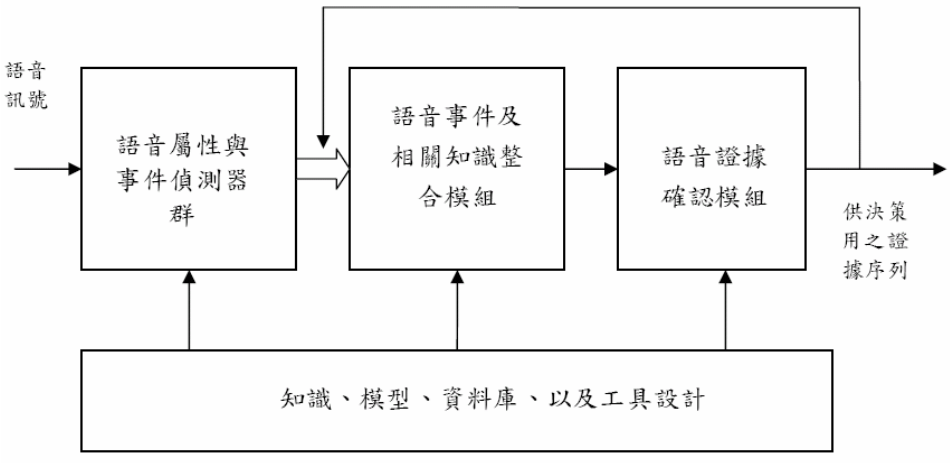


圖 1.1 新世代自動語音辨識技術架構圖

1.2 研究方向

由於新世代的語音辨識系統架構前端為一個偵測器群，而語音屬性偵測器-如發音方法（manner of articulation）及發音位置（place of articulation）等偵測器，為最基本的偵測器，它們可以提供語音特性之資訊，因此我們將致力於製作語音屬性偵測器。而由於國語 TCC300 語料庫並無人工切割位置，而製作偵測器首先需要的是精確的切割位置，因此我們首先將先對中文語料庫作自動切割(forced alignment)以取得好的切割位置[11]，並以此製作語音屬性偵測器，並且對所製作出來的語音屬性偵測器作效能及錯誤分析。

1.3 章節概要

本論文共分為五章：

第一章　緒論：介紹本論文之研究動機與研究方向。

第二章　建立以高斯混合模型的英文語音屬性貝氏偵測器

第三章　中文語音屬性偵測器之製作

第四章　中文語音屬性偵測器的效能分析與討論

第五章　結論與未來展望

第二章 以高斯混合模型為基礎的英文語音屬性貝氏偵測器

本章內容為介紹使用英文語料庫 TIMIT Corpus [2]，建立以音框為基礎（frame-based）的英文語音屬性（speech attribute）偵測器，其中語音屬性包含發音方法（manner of articulation）與發音位置（place of articulation）。此語音屬性偵測器要偵測以音框為單位的語音是否為其所要偵測的語音屬性。而此語音屬性偵測器的架構是以高斯混合模型為基礎的貝氏偵測器（GMM-based Bayesian detectors），其對每一個所要製作的語音屬性偵測器，訓練兩種高斯混合模型，一為標的模型（target model）、另一個為非標的模型（anti-model），再藉由計算每個音框（frame）在此兩種模型上的似然度分數（likelihood score）及考慮事前機率，採用最大似然度法則（Maximum Likelihood Criterion）來決定每個音框是否屬於所要偵測的類別。而本章所得到的偵測器效能結果、以及統計資料，也將對我們在後面章節所要製作的中文語音屬性偵測器提供比較參考。

2.1 語音資料庫、特徵參數抽取

由於我們要製作英文語音屬性的偵測器，因此我們採用的語料庫為英文語料庫：TIMIT 語料庫。在語音特徵參數抽取部分，我們採用傳統的梅爾頻率倒頻譜（MFCC）參數，以 32ms 為單位取一音框，每隔 10ms 重取音框。每一音框，包含 12 維的倒頻譜參數，12 維的一階差量倒頻譜參數，12 維的二階差量倒頻譜參數，1 維的一階差量對數能量以及 1 維的二階差量對數能量，共計 38 維。

而在英文音素的發音方法與發音位置分類表部份,TIMIT 語料庫附有英文發音方法的分類表，因此我們將 TIMIT 原有的 61 個音素（phonemes）依照其分類表(表 2.1)分成六類發音方法,包含塞音(stop)、鼻音(nasal)、摩擦音(fricative)、流音（glide)、塞擦音（affricate)、與母音（vowel)。而在[3]論文中有發音位置的分類表，因此參考其分類表（表 2.1）將 TIMIT 的音素分成十類發音位置，包含雙唇音（bilabial)、唇齒音（labiodental)、齒音（dental)、舌尖前（alveolar)、舌後音(velar)、喉音(glottal)、r 系音(rhotic)、前元音(front)、央元音(central)、後元音（back)，其中 r 系音是涵蓋所有（r）音類型的音[4][5]。另外在訓練語音屬性偵測器時，亦將訓練靜音（silence）的偵測器。

下表為英文 TIMIT 語料庫音素的發音方法、發音位置分類表，橫軸為發音方法類型，縱軸為發音位置類型。

表 2.1 英文 TIMIT 語料庫音素的發音方法、發音位置分類表

	Bilabial	Lab-dent	Dental	Alveolar	Velar	Glottal	Rhotic	Front	Central	Back
Stop	b p			d t dx	g k	q				
Nasal	m em			n en nx	ng eng					
Fricative		f v	th dh	s z	sh zh					
Glide						hh	r	y	l hv el	w
Affricate				jh ch						
Vowel							er axr	iy ih eh ey ae ay ix	aa aw ax ax-h	ah ao oy ow uh uw ux

P.s. Silence : pau , epi , h# , bcl , dcl , gcl , pcl , tcl , kcl

由於 TIMIT 語料庫附有人工切割的音素位置，因此我們可由統計資料中了解最基本的發音方法與發音位置的統計特性，而且也可以當作我們在後面所要做的中文語音屬性偵測器的參考比較。由於 TIMIT 語料庫所附的人工切割位置是以取樣點(sample)當單位，因此在此將其轉為以時間 10ms 為一個音框的單位。下表為 TIMIT 語料庫發音方法統計資料。

表 2.2 TIMIT 語料庫人工切割發音方法的統計資料

TIMIT 語料庫	訓練語料					測試語料				
	總音框數: 1,416,713					總音框數: 513,526				
發音方法	次數	音框總數	最短音框	平均音框	最大音框	次數	音框總數	最短音框	平均音框	最大音框
Vowel	57463	549896	<1	9.57	43	20911	202289	1	9.67	48
Fricative	21424	195416	<1	9.12	38	7724	71036	<1	9.20	33
Stop	25871	106575	<1	4.12	28	9176	37755	<1	4.11	30
Nasal	14157	80454	<1	5.68	26	5104	29043	<1	5.69	22
Glide	20257	129666	<1	6.40	25	7822	51199	1	6.55	24
Affricate	2031	14181	2	6.98	34	631	4470	2	7.08	23
Silence	35877	340525	<1	9.48	300	12777	117734	<1	9.20	464

p.s. 10 毫秒/音框

由表 2.2 可以看出以人工切割的 stop 發音方法其平均切割長度最短，長度僅 4.12 個音框，且統計其小於 1 個音框長度的佔其 0.22%。而 affrcate 發音方法其最短的音框長度有 2 個音框，不過其出現的總次數及音框總數遠低於其他種類。另外可以看出除了 stop 以外，子音的部份，fricative 平均長度長達 9.12 個音框，而其餘皆約為 5～7 個音框長。

下表為 TIMIT 語料庫發音位置統計資料。

表 2.3　TIMIT 語料庫人工切割發音位置的統計資料

TIMIT 語料庫	訓練語料					測試語料				
	總音框數: 1,416,713					總音框數: 513,526				
發音位置	次數	音框總數	最短音框	平均音框	最大音框	次數	音框總數	最短音框	平均音框	最大音框
bilabial	8796	40182	<1	4.57	26	3416	15486	<1	4.53	20
lab-dent	4210	34866	1	8.28	31	1622	13638	1	8.41	30
dental	3577	17536	<1	4.90	31	1320	6373	<1	4.83	22
alveolar	32662	214114	<1	6.56	38	11375	75028	<1	6.60	33
velar	10648	66628	<1	6.26	30	3658	23504	1	6.43	27
glottal	4547	29533	1	6.50	28	1600	10671	<1	6.67	30
rhotic	11992	91398	<1	7.62	34	4708	36827	1	7.82	37
front	34883	316266	1	9.07	43	12503	114284	1	9.14	39
central	15684	119361	<1	7.61	42	5881	45035	1	7.66	48
back	14204	146304	1	10.30	43	5285	54946	<1	10.40	39
silence	35877	340525	<1	9.49	300	12777	117575	<1	9.20	464

p.s. 10 毫秒/音框

發音位置裡面平均音框長度最短的類別是 bilabial，其平均長度為 4.57 個音框，而音框小於 1 的個數約佔 0.41%。dental 的平均長度為 4.90 個音框，是發音位置裡面平均長度最短的兩類。

2.2 英文語音屬性偵測器架構

2.2.1 貝氏偵測器架構

在此我們將建立以音框為基礎的英文語音屬性偵測器，而採用的偵測器架構為以高斯混合模型為基礎的貝氏偵測器架構。此貝氏偵測器架構為製作每一種發音方法以及發音位置偵測器時，我們將訓練兩種高斯混合模型[8]：一個為 target model，另一個為 anti-model。再藉由計算每個音框在此兩種模型上的似然度分數及考慮事前機率，採用最大似然度法則（Maximum Likelihood Criterion）來決定每個音框是否屬於所要偵測的類別。

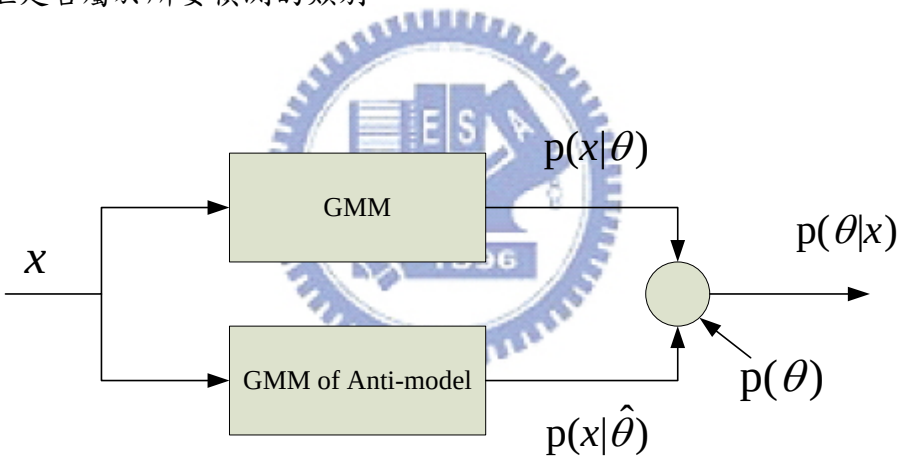


圖 2.1 貝氏偵測器架構圖

其中 X 語料庫中每一個音框的特徵參數向量，θ 為 target model、$\hat{\theta}$ 為 anti-model，$p(x|\theta)$ 為每一個音框在 target model 的近似度(likelihood)，$p(x|\hat{\theta})$ 為每一個音框在 anti-model 的近似度，$p(\theta)$ 為 target model 的事前機率，$p(\theta|x)$ 為每一音框屬於 target model 的事後機率(a posterior probability)。

2.2.2 高斯混合模型

高斯混合模型是以高斯機率分佈為主體，包含多個高斯機率分佈，因此模型參數包含平均值向量（mean vector）、變異數向量（variance vector）以及混合數權重（mixture weight）。

下列式子為 n 個基本高斯機率分佈加權和(weighted summation)之高斯混合模型。

$$p(x|\theta) = \sum_{i=1}^{n} C_i \cdot N(\mu_i, \Sigma_i) \tag{2.1}$$

$$N(\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} \exp[-\frac{1}{2}(x-\mu_i)^T (\Sigma_i)^{-1}(x-\mu_i)] \tag{2.2}$$

$$\theta = \{(C_i, \mu_i, \Sigma_i), 1 \le i \le n\} \tag{2.3}$$

其中 x 為一 D 維度大小之特徵參數向量，θ 為高斯混合模型，$N(\mu_i, \Sigma_i)$ 為高斯混合模型中各高斯分佈之機率密度函數，μ_i 為平均向量，Σ_i 為共變異矩陣(Covariance Matrix)，C_i 為混合權重，且須滿足 $\sum_{i=1}^{N} C_i = 1$。而在此實驗，我們假設共變異矩陣為一對角矩陣(Diagonal Matrix)。

在高斯混合模型的訓練中，可以利用最大似然度法則(Maximum Likelihood Criterion)來求得最佳模型，假設 $\bar{\theta}$ 為更新之模型、θ 為初始模型，使用預估最大值演算法（EM algorithm）去重新估算模型參數，使其滿足 $p(X|\bar{\theta}) \ge p(X|\theta)$ 之條件。亦即根據所有資料來估計統計特性，因此我們可以估算所有的平均向量，共變異矩陣，及各混合高斯模型之混合加權值，並將該統計出來的資料結果，根據最大似然度方法達到最大化 $p(X|\bar{\theta})$ 的要求，如此即可找到模型參數，重估公式

如下[10]：

$$C_i = \frac{1}{K}\sum_{k=1}^{K} p(i \mid x_{k,}\, \theta) = \frac{1}{K}\sum_{k=1}^{K} \frac{C_i\ p(x_k \mid i\,,\, \theta)}{\sum_{i=1}^{n} C_i\ p(x_k \mid i\,,\, \theta)} \qquad (2.4)$$

$$\mu_i = \frac{1}{K}\sum_{k=1}^{K} \frac{C_i\ p(x_k \mid i\,,\, \theta)\, x_k}{\sum_{i=1}^{n} C_i\ p(x_k \mid i\,,\, \theta)} \qquad (2.5)$$

$$[\Sigma_i]_{dd} = \frac{1}{K}\sum_{k=1}^{K} \frac{C_i\ p(x_k \mid i\,,\, \theta)\, [x_k - \mu_i]_d^2}{\sum_{i=1}^{n} C_i\ p(x_k \mid i\,,\, \theta)} \quad ; \quad 1 \le d \le D \qquad (2.6)$$

其中$[x]_d$是指向量中的第 d 個元素，$[\Sigma]_{ij}$則是指矩陣Σ的第 i 行第 j 列之元素。

2.3 英文語音屬性偵測器之效能

由於 TIMIT 語料庫已經附有人工切割的音素位置，且音素與發音方法以及發音位置的分類表也以具備，因此可以利用分類表將音素的切割位置轉換成發音方法或發音位置的切割位置，去求取我們所要製作的語音屬性偵測器的二種高斯混合模型的參數。在訓練完每一個發音方法、發音位置偵測器的兩種高斯混合模型後，我們將對於測試語料，求得各類語音屬性偵測器的偵測效果。在此每一個發音方法、發音位置偵測器皆將利用最大似然度法則，去偵測測試語料的每一個音框是否為所要偵測的種類(人工所標示音素的分類)。

下式為最大似然度法則：

$$\frac{p(x|\theta)}{p(x|\hat{\theta})} \geq \frac{p(\hat{\theta})}{p(\theta)} \times \zeta \quad , \quad x \rightarrow \theta \qquad (2.7)$$

$$p(\theta) + p(\hat{\theta}) = 1$$

其中 θ 為 target model，而 $\hat{\theta}$ 為 anti-model， x 為每一個音框的特徵參數向量，

ζ 為臨界值（threshold）。 $p(\theta)$ 與 $p(\hat{\theta})$ 為各個種類的 target 與 anti-target 的事先機率。若 2.7 式成立，則將此音框判定為 target，反之則否。

我們利用 TIMIT 訓練語料中每個語音屬性種類的 anti-target 與 target 的音框數比值求得事先機率比值 $p(\theta) / p(\hat{\theta})$ ，並藉著微調臨界值（threshold），可以得到偵測器對測試語料的錯誤警戒率（false alarm rate，FA）以及錯誤拒絕率（false reject rate，FR）的值。當錯誤警戒率等於錯誤拒絕率時，此時的音框錯誤率（frame

error rate）即是等錯誤率(Equal Error Rate，EER)。

以下為錯誤警戒率、錯誤拒絕率、音框錯誤率的定義：

FA Rate = # of FAs / total # of non-target (2.8)

FR Rate = # of FRs / total # of targets (2.9)

Frame Error Rate = (# of FAs + # of FRs)/ total # of labels (2.10)

2.3.1 發音方法偵測器之效能

下表為以高斯混合模型為基礎的英文語音發音方法貝氏偵測器，其發音方法偵測器與國外學者用不同偵測器架構所做出來的性能[6]比較：

表 2.4 以高斯混合模型為基礎的貝氏偵測架構與其它偵測器架構的發音方法偵測效能比較

	Frame-based detector		Segment-based detector	
EER(%)	Baseline(GMM)	ANN*	HMM	SEG_MCE
Vowel	12.3	9.0	1.7	1.8
Fricative	10.0	11.3	6.4	3.6
Stop	16.7	14.5	9.9	5.4
Nasal	8.7	12.2	11.2	5.4
Glide (Approximant)	16.3	15.9	8.0	6.1
Affricate	7.2			
Silence	9.7	3.7	2.1	0.8

*Frame-base ANN with a detection threshold of 0.5

在表 2.4 的類神經網路（ANN)部份的作法，各個偵測器其網路輸入部分有 9 個音框，每個音框有 13 維的特徵向量(12MFCCs+energy)，因此共有 117 個輸入節點(input nodes)，而音框間距（frame shift）為 10ms。且有一個隱藏層（hidden layer）其中有 100 個節點。輸出（output node）部份僅有一個節點。

由表 2.4 可以看出，以以高斯混合模型為基礎的 fricative 與 nasal 偵測器，其效能較 ANN 為佳，尤其是 nasal 偵測器改善了約 3%，其他的發音方法偵測器均較差於 ANN，尤其是 silence 偵測器差了 6%。另外由表 2.4 可以看出，以 HMM Segment-Based 做發音方法偵測器普遍比 GMM 以及 ANN 架構好，這提供了我們未來在做其他偵測器一個參考的依據。

2.3.2 發音位置偵測器之效能

下表為以高斯混合模型為基礎的英文語音發音位置貝氏偵測器的 EER

表 2.5 以高斯混合模型為基礎的的英文發音位置貝氏偵測效能比較

EER(%)	GMM-based Bayesian detector
Bilabial	12.2
Lab-dent	11.0
Dental	12.7
Alveolar	12.0
Velar	12.4
Glottal	18.3
Rhotic	9.4
Front	13.5
Central	17.7
Back	17.8

由表 2.5 可以看出幾乎全部的發音位置偵測器的 EER 均大於 10%以上、除了 Rhotic 偵測器、但也很接近 10%。其中以 Glottal、Central、Back 偵測器錯誤率皆大於 17%以上為最差。

2.4 英文發音方法結合發音位置偵測器之效能

我們已經製作出發音方法與發音位置偵測器，並也已知這兩類偵測器的效能，那我們能藉由結合這兩類偵測器而改進效能嗎？在此我們將有組合可能(參考表 2.1)的發音方法偵測器與發音位置偵測器結果結合起來，並假設所有音素其所屬的發音方法與發音位置是獨立的關係。在新一代語音辨識系統（NG-ASR）第一級的各個偵測器，其輸出為語音屬性或者事件的機率值。因此，藉由已訓練出來的各個發音方法與發音位置偵測器的 target model 以及 anti-model，我們將可以求出每一個音框是否為 target 的機率。

下式為透過每一音框在 target model 與 anti-model 的似然度分數（likelihood score），再考慮事先機率比值後，所求得的每一個音框是否為 target 的機率。

$$p(\theta \mid x) = \frac{p(x|\theta)\, p(\theta)}{p(x)} = \frac{p(x|\theta)\, p(\theta)}{p(\theta)\, p(x|\theta) + p(\hat{\theta})\, p(x|\hat{\theta})} \tag{2.11}$$

其中 θ 為 target model，而 $\hat{\theta}$ 為 anti-model，x 為每一個音框的特徵參數向量。而 $p(x|\theta)$ 與 $p(x|\hat{\theta})$ 為音框在 target model 與 anti-model 上的似然度分數， $p(\theta)$ 與 $p(\hat{\theta})$ 為 target 與 anti-target 的事先機率。如此便能得到每一個音框為 target 的機率。

由於我們假設所有音素（phoneme）其所屬的發音方法與發音位置是獨立的關係，因此發音方法及發音位置結合機率密度函數（joint probability density function）可以拆開為兩個機率密度函數相乘。由此我們可以得到每一個音框其發音方法結合發音位置的機率值，並且藉由調整臨界值，可以得到偵測器對測試語料的錯誤警戒率以及錯誤拒絕率的值。最後將所有錯誤警戒率與錯誤拒絕率的值畫出一個 FA-FR 的曲線圖。下式為其數學表示式，在此以 manner 表示發音方法，以 position 表示發音位置。

$$\mathrm{p}(\mathrm{manner} \cap \mathrm{position} \mid x) \overset{indep.}{=} \mathrm{p}(\mathrm{manner} \mid x) \times \mathrm{p}(\mathrm{position} \mid x) \tag{2.12}$$

$$\frac{\mathrm{p}(\mathrm{manner} \cap \mathrm{position} \mid x)}{\mathrm{p}(\overline{\mathrm{manner} \cap \mathrm{position}} \mid x)} > \zeta \quad , \quad x \to \mathrm{manner} \cap \mathrm{position} \tag{2.13}$$

其中 $\mathrm{p}(\mathrm{class} \mid x)$ 為每一個音框屬於何種類別（class）的機率值，$\mathrm{p}(\overline{\mathrm{class}} \mid x)$ 則為每一音框不為類別（class）的機率值。

而由 TIMIT 語料發音方法與發音位置分類表（表 2.1）可以得知，6 種發音方法與 10 種發音位置共有 21 種組合，每一種組合可能包含多個音素集合，亦有少量是單一個音素。下面我們將觀察將發音位置以及發音方法偵測器結合後，效能的改善變化。所有 21 種的發音方法結合發音位置 FA-FR 的曲線圖請參考附錄 2。

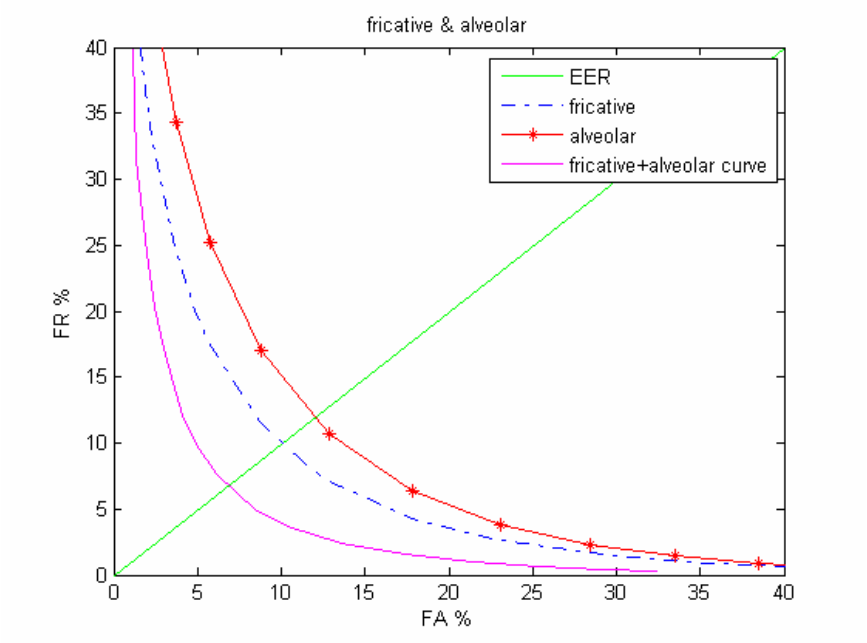


圖 2.2 發音方法結合發音位置 FA-FR 曲線圖

由圖 2.2 可以看出原先 fricative 偵測器其 EER 為 10.0%，而 alveolar 偵測器其 EER 為 12.0%，將兩種偵測器結合在一起後所得到的 EER 約為 7~8%，很明顯的可以看出藉由結合 fricative 偵測器與 alveolar 偵測器可以得到較佳的偵測效能，而不用特別去訓練一個 fricative 與 alveolar 交集的偵測器。而觀察所有的發音方法結合發音位置的 FA-FR 曲線圖，可以得出 stop + velar、nasal+ bilabial 、fricative+ alveolar 、 fricative + velar 、 glide + central 、glide + back 此六種的發音方法結合發音位置的組合，其 EER 比其原先結合的兩種較佳，而 stop + glottal 、 fricative + lab-dental 、 fricative + dental、glide + glottal 、 vowel + rhotic 、vowel + front 、 vowel +central 、 vowel +back 共八種的發音方法與發音位置的組合其 EER 比其原先結合的兩種較差，而其餘的七種組合則其結合的 EER 則介於其原先結合的兩種之間。可以明顯看出與任何發音位置偵測器與 vowel 偵測器結合的偵測效果，並不會較原先的佳，而發音位置偵測器 EER 最高的 glottal 偵測器，其兩種組合方式的結合偵測效果亦較原先的差。

下面我們將表 2.1 著上顏色以明顯區分發音方法與發音位置偵測器的結合偵測效果，粉紅色表示結合偵測效果較原先獨立的佳，灰色表示結合偵測效果較原先獨立的差，而其餘的淡綠色表示結合偵測效果介於原先兩偵測器之間。

表 2.6 發音方法結合發音位置偵測器效能改善圖示

	Bilabial	Lab-dent	Dental	Alveolar	Velar	Glottal	Rhotic	Front	Central	Back
Stop	B p			d t dx	g k	q				
Nasal	m em			n en nx	ng eng					
Fricative		f v	th dh	s z	sh zh					
Glide						hh	r	y	l hv el	w
Affricate				jh ch						
Vowel							er axr	iy ih eh ey ae ay ix	aa aw ax ax-h	ah ao oy ow uh uw ux

第三章 中文語音屬性偵測器

製作語音屬性偵測器的第一步，就是需要有語音資料庫及詳細的標示資料，而此標示資料必須標示到語音屬性階層並包含其時間訊息，不像從前以隱藏式馬可夫模型為基礎(HMM-Based)所作的語音辨認研究，其所需之語料庫都只需標示到音節階層即可，也不需要詳細的時間資訊。在第二章中，由於英文 TIMIT 語料庫附有音素的人工切割位置，因此很容易的可以利用音素的切割位置以及英文語音屬性分類表去製作英文屬性偵測器，然而由於中文 TCC300 語料庫缺少正確的標音位置，因此首先我們將以隱藏式馬可夫模型對中文 TCC300 語料庫的訓練語料與測試語料求取音素切割位置[7]，再以此切割位置製作中文語音屬性偵測器。而本章主要重點放在發音方法偵測器上的製作，因此僅考慮製作發音方法偵測器。在製作出中文發音方法偵測器之後也將與英文發音方法偵測器作效能比較。

3.1 中文發音方法貝氏偵測器之製作

3.1.1 以 flat-start 方式求取語料庫初始切割位置

在此節將以中文 TCC300 語料庫來製作中文語音屬性的偵測器，由於中文 TCC300 語料庫並沒有正確的標音位置，因此我們在製作中文語音屬性偵測器之前便是需要一個初始切割位置。在此我們對中文語料庫 TCC300 求取 38 維梅爾頻率倒頻譜參數，訓練與前後文無關（context independent）的音素的隱藏式馬可夫模型（Phone HMM），而其每個音素的隱藏式馬可夫模型將其狀態數設為 3 個，且以 flat start 的方法訓練隱藏式馬可夫模型，在此所訓練的音素分為子音、介音、

母音、以及鼻音韻尾共四類，詳細的中文音素分類請參考附錄二。之後再將訓練好的音素的隱藏式馬可夫模型拿來對 TCC300 的訓練語料以及測試語料作強迫切割（forced-alignment），因此可以得到一個有粗略位置資訊的音素切割位置，最後利用中文音素的發音方法分類表(表 3.1)，將訓練語料以及測試語料的音素切割位置轉為發音方法的切割位置。而此訓練語料的發音方法切割位置便作為我們在製作中文發音方法偵測器的初始切割位置，最後再將製作出來的中文發音方法偵測器對測試語料作偵測求取偵測效能。

表 3.1 中文發音方法分類表

p.s.括弧中為 IPA 表示

1	爆破音（Stop）	ㄅ (p)	ㄆ (p^h)	ㄉ (t)	ㄊ (t^h)	ㄍ(k)	ㄎ (k^h)
2	鼻音　（Nasal）	ㄇ (m)	ㄋ (n)	n_n , ng			
3	摩擦音（Fricative）	ㄖ (ʐ)	ㄈ (f)	ㄙ (s)	ㄒ (ɕ)	ㄏ (x)	ㄕ (ʂ)
4	塞擦音（Affricate）	ㄓ (t)	ㄔ ($t^hʂ$)	ㄑ ($t^hɕ$)	ㄐ (tɕ)	ㄘ（t^hs）	ㄗ（ts）
5	流音　（Liquid）	ㄌ (l)					
6	母音　（Vowel）	others					

P.S. n_n , ng 為ㄤㄣㄢㄥ的鼻音韻尾

下圖為以 flat start 的方法訓練中文音素的隱藏式馬可夫模型後，對語料庫切割的流程圖。

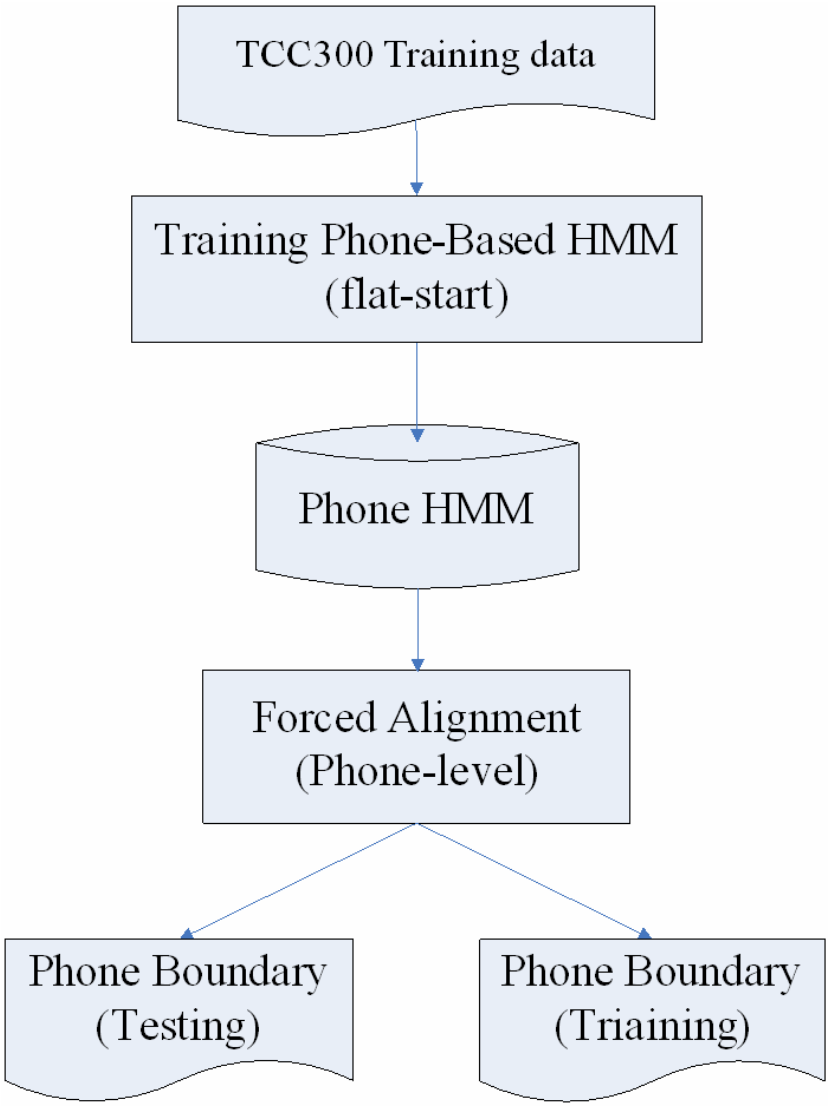


圖 3.1 以 flat start 方式訓練 HMM 的流程圖

下表為以 flat-start 的方式對 TCC300 訓練語料所作的初始音素切割位置轉成發音方法類別的統計資料

表 3.2 以 flat-start 的方式對 TCC300 訓練語料所作的初始音素切割位置的發音方法類別統計資料

發音類別	次數	音框總數	最短音框	平均音框	最長音框
Vowel	418343	4086526	3	9.77	67
Fricative	74276	844507	3	11.37	45
Stop	76291	642205	3	8.42	29
Nasal	119535	707626	3	5.92	50
Liquid	14653	103489	3	7.06	33
Affricate	75889	784833	3	10.34	47
Silence	16074	406290	2	25.28	2314
Sp	36582	1028709	1	28.12	304

p.s. 10 毫秒/音框

其中音節跟音節間的短暫靜音（short pause ,sp）共有 300836 個，實際切出的次數約為 36582，約佔 12.16%，而上表的 sp 統計出至少切出長度為 1 個音框的資料，由表可以看出由於 TCC300 語料庫包含長句錄音，因此句中會出現錄音者停頓很久的情況，導致 sp 的平均長度會很長，且 stop 的平均切割長度為 8.42 個音框，與英文人工切割的平均長度比較似乎過長。

3.1.2 中文發音方法偵測器架構

中文發音方法偵測器的架構，同於我們在第二章製作英文語音屬性偵測器所使用的以高斯混合模型為基礎的貝氏偵測器。我們依照 3.1.1 節所得到的中文 TCC300 訓練語料的初始切割位置，訓練各個發音方法的 target model 與 anti-model，訓練好高斯混合模型後，再利用最大概似法則(Maximum likelihood Criterion)，去決定測試語料的每一個音框是否為所要偵測的種類(人工所標示音素的分類)，最後再藉著微調臨界值（threshold），可以得到偵測器對測試語料的錯誤警戒率（false alarm rate，FA）以及錯誤拒絕率（false reject rate，FR）的值。最後將所有錯誤警戒率與錯誤拒絕率的值畫出一個 FA-FR 的曲線圖。將可得到當錯誤警戒率等於錯誤拒絕率時的等錯誤率(Equal Error Rate，EER)，相關數學式請參考第二章第三節。

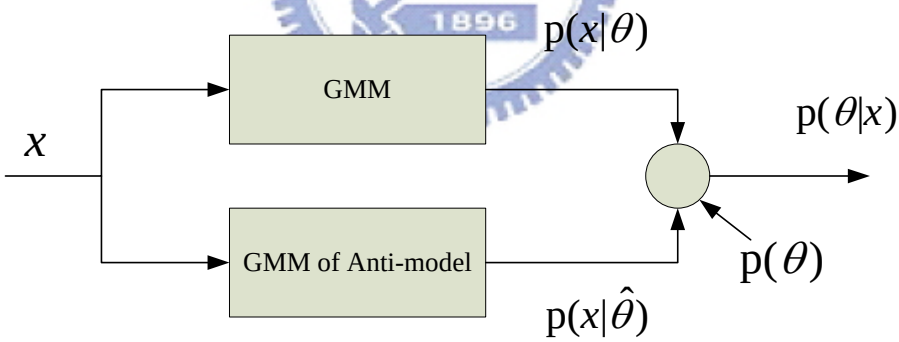


圖 3.2 貝氏偵測器架構圖

3.2 中文發音方法偵測器之偵測效能

以 flat start 的方式訓練隱藏式馬可夫模型後，對中文 TCC300 的測試語料作強迫切割（forced alignment），所得到的測試語料初始切割位置拿來求取以高斯混合模型為基礎的中文發音方法貝式偵測器的偵測效能。

3.2.1 偵測器的 EER 比較

下表比較以不同架構所作出來的中文與英文發音方法偵測器的效能比較。

1. 英文 TIMIT 語料庫(MFCC12+energy)加上人工標記的切割位置以類神經網路（ANN）的偵測架構所作出來的英文發音方法偵測器的偵測效果。

2. 英文 TIMIT 語料庫(MFCC38)加上人工標記的切割位置訓練高斯混合模型並以貝氏偵測架構所作出來的英文發音方法偵測器的偵測效果。

3. 中文 TCC300 語料庫(MFCC38)以 flat-start 的方式訓練狀態數為 3 的中文音素隱藏式馬可夫模型（3-states phone HMM）後，對中文語料庫 TCC300 作強迫切割所得到的音素切割位置，再以訓練語料的音素切割位置製作中文發音方法偵測器的偵測效果。

表 3.3 發音方法偵測器效果比較

	Frame-based		
	ANN	Bayesian detector	
EER(%)	**English**	**English**	**Mandarin**
vowel	9.0	12.3	10.70
fricative	11.3	10.0	15.7
stop	14.5	16.7	11.5
nasal	12.2	8.7	11.5
glide/liquid	15.9	16.3	9.2
affricate	-	7.2	11.5
silence	3.7	9.7	8.0

在第二章所作的以已有人工切割位置的 TIMIT 英文語料庫製作以高斯混合模型為基礎的英文發音方法貝式偵測器所得到等錯誤率（EER）約在 10%上下。而由表 3.3 可以看出以高斯混合模型為基礎的中文發音方法貝式偵測器的等錯誤率（EER）大概也約 10%上下。但是可以明顯看出以類神經網路（ANN）所製作出來的 silence 偵測器的 EER 僅僅只有 3.7%，而中文的 silence 偵測器的 EER 高達 8.0% 。另外中文 fricative 偵測器的 EER 高達 15.7%，且中文 affricate 偵測器的 EER 為 11.5%，均高於英文的 4~6%。雖然對於 stop、glide/liquid 而言，中文偵測器的效能明顯比英文偵測器的好多了。除此之外我們知道以隱藏式馬可夫模型對語料庫作強迫切割的初始切割位置並非十分準確，一個不準確的切割位置可能會影響偵測器的偵測效能。是否因為如此，造成 fricative、affricate 偵測器的效能較差。因此我們將檢查以 flat-start 方式訓練隱藏式馬可夫模型後對語料庫作強迫切割的初始切割位置的效果。

3.2.2 切割位置統計資料比較

下表比較中文自動切割與英文人工標記的發音方法平均長度以及最大長度的統計資料。

1. 中文 TCC300 語料庫以 flat-start 的方式訓練狀態數為 3 的中文音素隱藏式馬可夫模型（3-states phone HMM），並以此模型對中文 TCC300 訓練語料做強迫切割後，將切割位置轉為所得到的自動切割統計資料。

2. 英文 TIMIT 語料庫人工標記的訓練語料統計資料。

表 3.4 中文自動切割與英文人工標記的發音方法統計資料

	TIMIT training Data by manual labeling			TCC300 training Data by flat-start phone HMM Forced-Alignment		
class	Average frame	> TIMIT 2*average frame	Max frame	Average frame	> TIMIT 2*average frame	Max frame
Vowel	9.57	3.9%	43	9.77	5.5%	67
Fricative	9.12	2.2%	38	11.37	2.9%	45
Stop	4.12	7.1%	28	8.42	43.9%	29
Nasal	5.68	2.6%	26	5.92	5.4%	50
Glide/Liquid	6.40	3.1%	25	7.06	7.0%	33
Affricate	6.98	2.8%	34	10.30	16.4%	47
silence	19.28	-	300	25.28	-	2314
Sp	6.07	-	123	28.12	-	304

p.s. 10 毫秒/音框

$$\text{兩倍長度比率} = \frac{\text{語料庫發音方法長度大於兩倍TIMIT發音方法平均長度的個數}}{\text{語料庫發音方法總個數}} * 100\%$$

在表 3.4 中，我們除了統計兩邊切割位置發音方法基本的平均長度與最大長度，還統計了對中文 TCC300 語料庫作自動切割後其發音方法長度大於兩倍人工切割的英文發音方法平均長度的比率。因為過長的切割長度可能是因為雜訊干擾、背景人聲所造成的，尤其是對子音這種平均長度較短的音素，受到雜訊干擾導致過長的切割位置會嚴重影響到製作出來的偵測器效能，因此在這裡我們採用兩倍的英文人工切割的發音方法平均長度當作標準，觀察各個發音方法過長切割的比例。由於中文 TCC300 語料庫，裡面包含長句錄音，因此句頭及句尾的 silence 的長度會遠高於 TIMIT 的 silence，而且也因為錄製長句的關係，使得在句中可能會出現停頓很長的短暫靜音（short pause，sp），因此在這我們不考慮中、英文 silence 與 sp 長度的比較。

1. 人工標記的英文發音方法 stop 的平均長度為 4.12 個音框，而以自動切割的中文 stop 的平均長度卻長達 8.42 個音框，整整多出了一倍，且自動切割出來的中文 stop 長度大於 2 倍英文的 stop 長度佔中文全部 stop 的次數比例有 43.9%，而英文 stop 的長度大於 2 倍英文 stop 平均長度的比例只佔 7.1%。
2. 自動切割的中文 affricate 的平均長度長於英文人工標記的 affricate 3 個音框，且兩邊大於人工標記 affricate 長度兩倍的比例是 2.8%與 16.4%。而自動切割的中文 fricative 的平均長度長於英文人工標記的 fricative 2 個音框。
3. 對於 vowel、nasal、liquid 而言，自動切割的中文發音方法平均長度與人工標記的英文發音方法平均長度是差不多的，不過對於大於兩倍人工標記的英文發音方法長度而言，自動切割的中文發音方法所佔的比例均較高。
4. 由以上幾點結果可以得知以 flat-start 訓練音素的隱藏式馬可夫模型後，對中文 TCC300 的訓練語料作強迫切割後的結果，對於所有的發音方法而言，比人工標記的英文發音方法長度還長，這也意味著 silence 的切割長度將較短，甚至沒切出來。

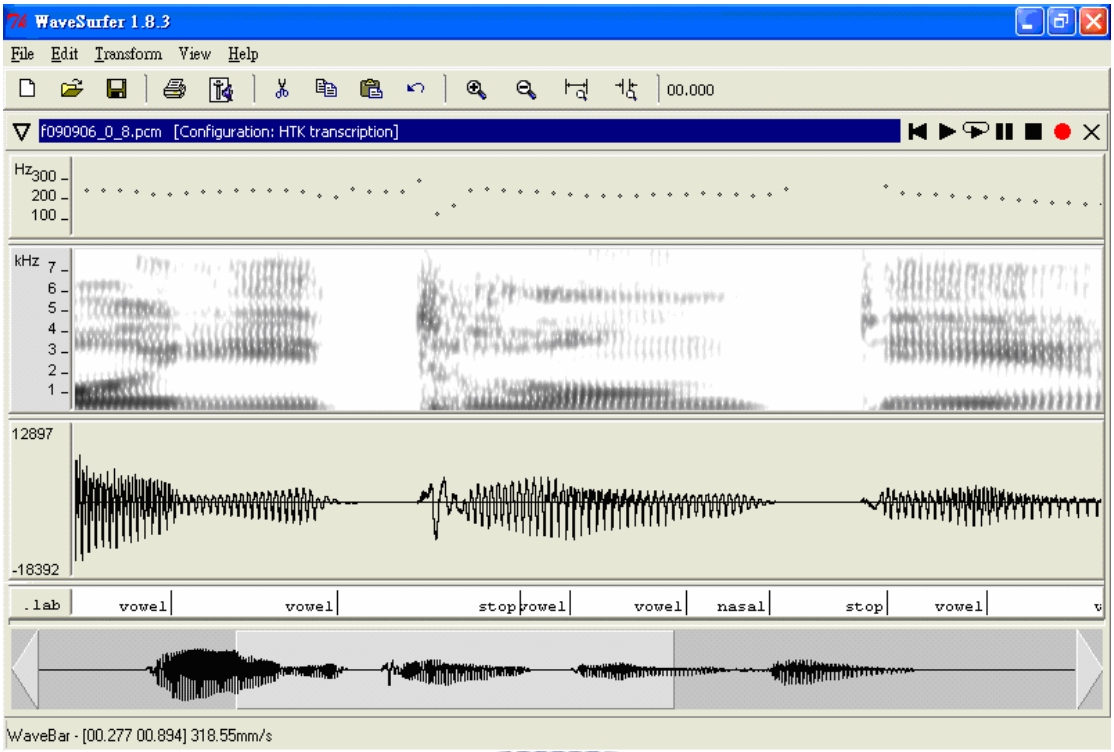


圖 3.3 以 flat start 方式訓練音素的馬可夫模型後對中文語句切割的實例

因此由前面所列的切割位置統計資料以及偵測器的 EER 結果來看，可以知道以 flat start 的方式訓練狀態數為 3 的中文音素隱藏式馬可夫模型後，對中文語料庫 TCC300 作強迫切割後所得到的音素切割位置，可能並不夠精確來當作訓練發音方法模型的初始切割位置，因為一個不夠好的切割位置將會影響後面所製作偵測器的效能，因此針對這一點我們將加以改善。而下一節我們將要改進以 flat start 方式訓練音素 HMM 後對語料庫切割所造成的誤差。

3.3 初始切割位置之改進

在前一節中我們利用 flat start 的方式訓練中文音素的隱藏式馬可夫模型後，對中文語料庫 TCC300 作強迫切割得到音素的切割位置，再以訓練語料的音素切割位置製作中文發音方法偵測器，但對於製作出來的發音方法偵測效果並不是非常理想，尤其是 silence 的偵測效能與使用 TIMIT 語料庫及類神經網路所製作的 silence 偵測器有一段很大的差距。因此我們將改變方法去求得較佳的中文語料庫初始切割位置，使所作出來的偵測器效能能夠提昇，並且對製作出來的中文發音方法偵測器作錯誤分析。

3.3.1 由音節切割位置起始求得音素的初始切割位置

由 4.1 節可以得知以 flat-start 的方式訓練狀態數為 3 的中文音素隱藏式馬可夫模型後，對中文語料庫 TCC300 作強迫切割所得到的音素切割位置，而這個切割位置可能不太精確，尤其對於各個發音方法的長度都切的偏長，如此一來也壓抑了音節間的短暫靜音（short pause，sp），這樣導致了訓練出來的 silence 偵測器效果不佳，也會影響到其他發音方法偵測器。因此我們要去改進中文 TCC300 語料庫的初始切割位置，以求得更好的偵測器效能。我們捨棄之前由 flat start 的方式訓練狀態數為 3 的中文音素隱藏式馬可夫模型後，再以訓練出來的 HMM 對語料庫作強迫切割，以求得中文 TCC300 語料庫的初始切割位置，取而代之的方法是從中文音節的切割位置起始著手（with syllable boundary constraint start）。

在這中文音節的切割裡面，我們將錄製者所發出的呼吸聲、背景雜訊及背景人聲考慮進去，因為這些不屬於語音的雜訊將會影響我們所要製作的偵測器模型，因此我們將對這些雜訊建立一個模型去模擬其分佈，以期望與語音及靜音的部份區隔開來。而此模型建立的方式為將所有中文的訓練語料拿來訓練一個狀態數為 3 的隱藏式馬可夫模型當作是背景模型（Universal Background Model，UBM）。然後在做中文音節切割的時候，將語料在其他音節模型與此背景模型的分數（likelihood）作比較，比背景模型的分數還低的音段（segment）當作是雜訊，而且以”breath ”標記之。breath 發生的位置有可能在句頭句尾的 silence 或者發生在句中的 sp，藉由標記出 breath 的位置跟 silence 與 sp 作區分，如下圖 3.4，以期望能切出較乾淨的 silence 以及 sp。

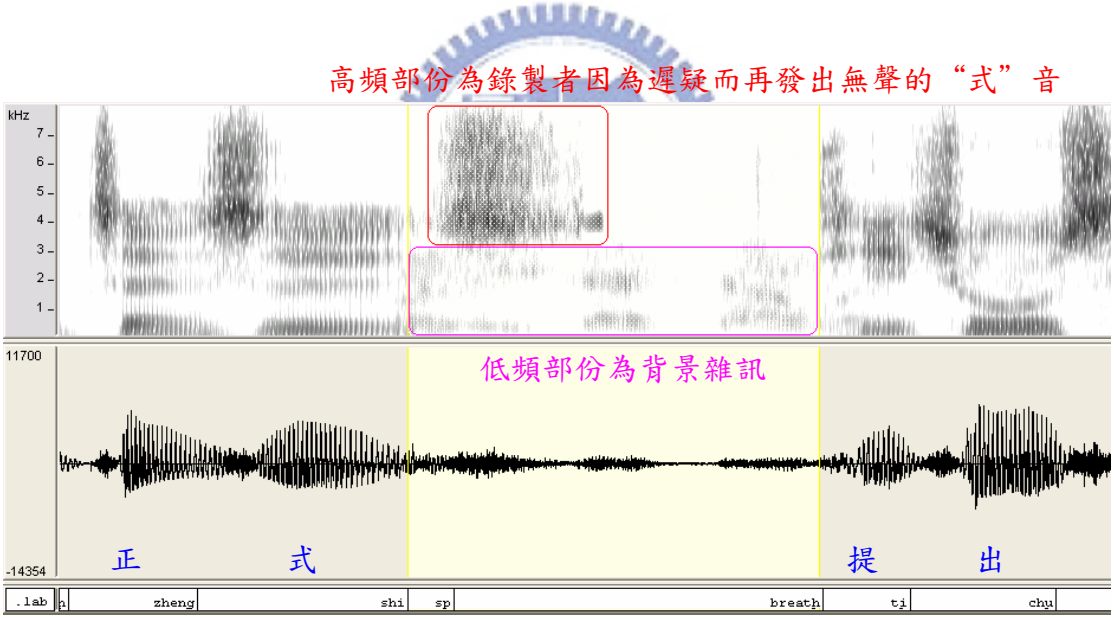


圖 3.4 加入 breath 模型的中文音節切割實例

而有了中文的音節切割位置，再將各個音節依據其組成的音素個數，均勻切割長度給其組成的音素，再由音素的切割位置去訓練個別狀態數為 3 的中文音素隱藏式馬可夫模型(包含 sp 與 breath model)，最後再利用全部音素的隱藏式馬可夫模型對中文 TCC300 語料庫作強迫切割，如此一來便得到語料庫音素的切割位

置。下圖 3.5 是以音節切割位置去訓練音素的隱藏式馬可夫模型後，對中文語料庫作強迫切割的流程圖。

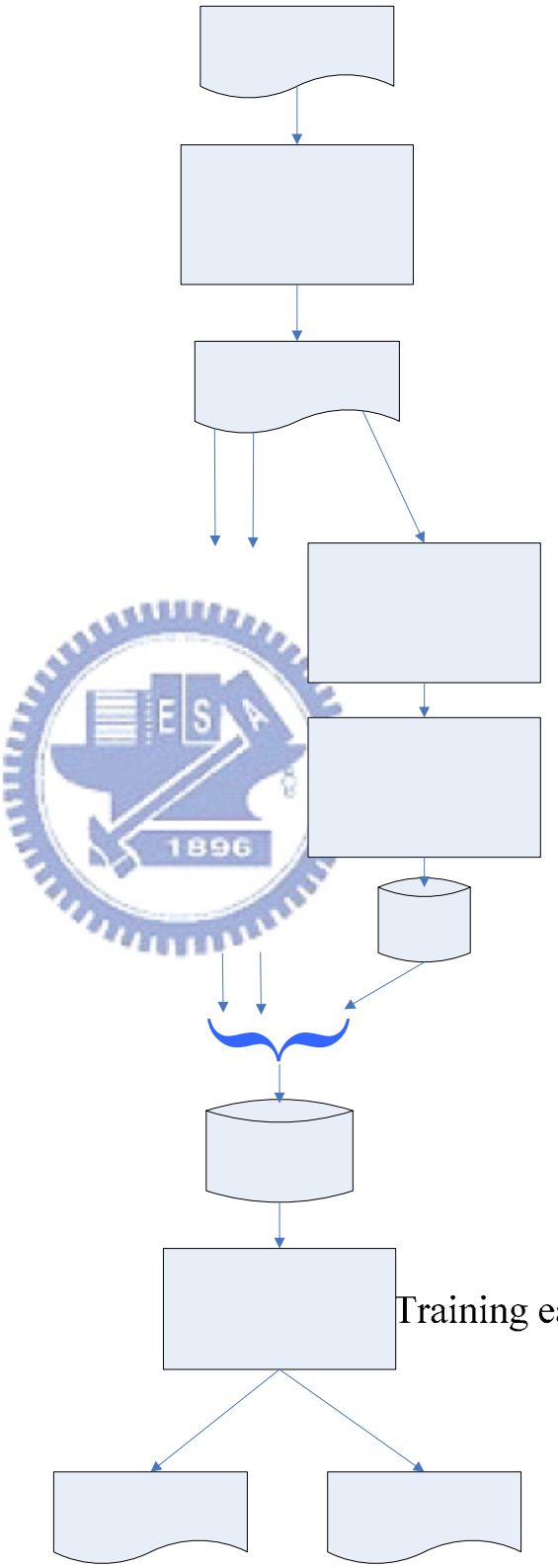


圖 3.5 由音節切割位置訓練音素的隱藏式馬可夫模型到切割語料庫的流程圖

3.3.2 切割位置的統計資料比

下表比較從音節的切割位置起始著手去訓練音素的隱藏式馬可夫模型以及以 flat start 的方式去訓練音素的隱藏式馬可夫模型後，兩者對中文訓練語料作強迫切割，其中非語音部份的切割統計資料。

表 3.5 非語音部份的中文訓練語料切割統計資料比較

	by syllable boundary constraint start phone HMM forced-alignment			by flat start phone HMM forced-alignment		
class	counts	frames	average frame	counts	frames	average frame
silence	16,881	387,858	22.98	16,074	406,290	25.28
short pause (sp)	50,018	1,026,704	20.53	36,582	1,028,709	28.12
breath	19,515	146,429 (9.4%)	7.50	-	-	-

P.S. 10ms/frame

由表 3.5 可以看出原來以 flat start 的方式訓練隱藏式馬可夫模型去對語料作強迫切割後，其非語音部份共約 143 萬個音框，而以音節的切割位置起始著手訓練音素的隱藏式馬可夫模型去對語料作強迫切割後，其非語音的部份為 silence + sp 約 141 萬個音框以及 breath 的 14 萬個音框，其中 breath 所切出的總量約佔非語音部份的 9.4%。很明顯的非語音的音框數由音節切割位置起始的方式比以 flat start 的方式約增加了 12 萬個，如此一來，不但降低之前切割長度過長的語音部份的音框數，又可以使切出來的 silence 與 sp 較不受雜訊影響。

下表比較從音節的切割位置起始著手去訓練音素的隱藏式馬可夫模型以及以 flat start 的方式去訓練音素的隱藏式馬可夫模型後，對 TCC300 訓練語料作強迫切割，其中語音部份的切割統計資料，以及兩倍平均長度比率。

表 3.6 語音部份的中文訓練語料切割統計資料比較

	by syllable boundary constraint start phone HMM forced-alignment			by flat-start phone HMM forced-alignment		
class	average frame	> TIMIT 2*average frame	max frame	average frame	> TIMIT 2*average frame	max frame
vowel	10.28	6.9%	75	9.77	5.5%	67
fricative	9.25	0.76%	41	11.37	2.9%	45
stop	6.74	19.9%	24	8.42	43.9%	29
nasal	6.81	7.7%	49	5.92	5.4%	50
liquid	5.59	1.8%	35	7.06	7.0%	33
affricate	8.51	4.9%	37	10.30	16.4%	47

由表 3.6 可以看出對於子音的部份，從音節切割位置起始著手訓練音素的隱藏式馬可夫模型，再將訓練好的模型對語料庫作切割的方法，其所切出來的音素所對應至的發音方法平均長度都有明顯的下降。而以兩倍 TIMIT 人工切割的平均長度為參考而言，更可以明顯的看出從音節切割位置起始的方法，其所切出來的發音方法含高長度的切割數目大大的降低。

下圖為各個發音方法在不同切割方法所切出來的發音方法長度分佈圖

1. 以 flat-start 的方式訓練音素的隱藏式馬可夫模型，再對語料作強迫切割。
2. 從音節切割位置起始經過均勻切割後，訓練音素的隱藏式馬可夫模型，再對語料作強迫切割。
3. TIMIT 語料庫人工切割。

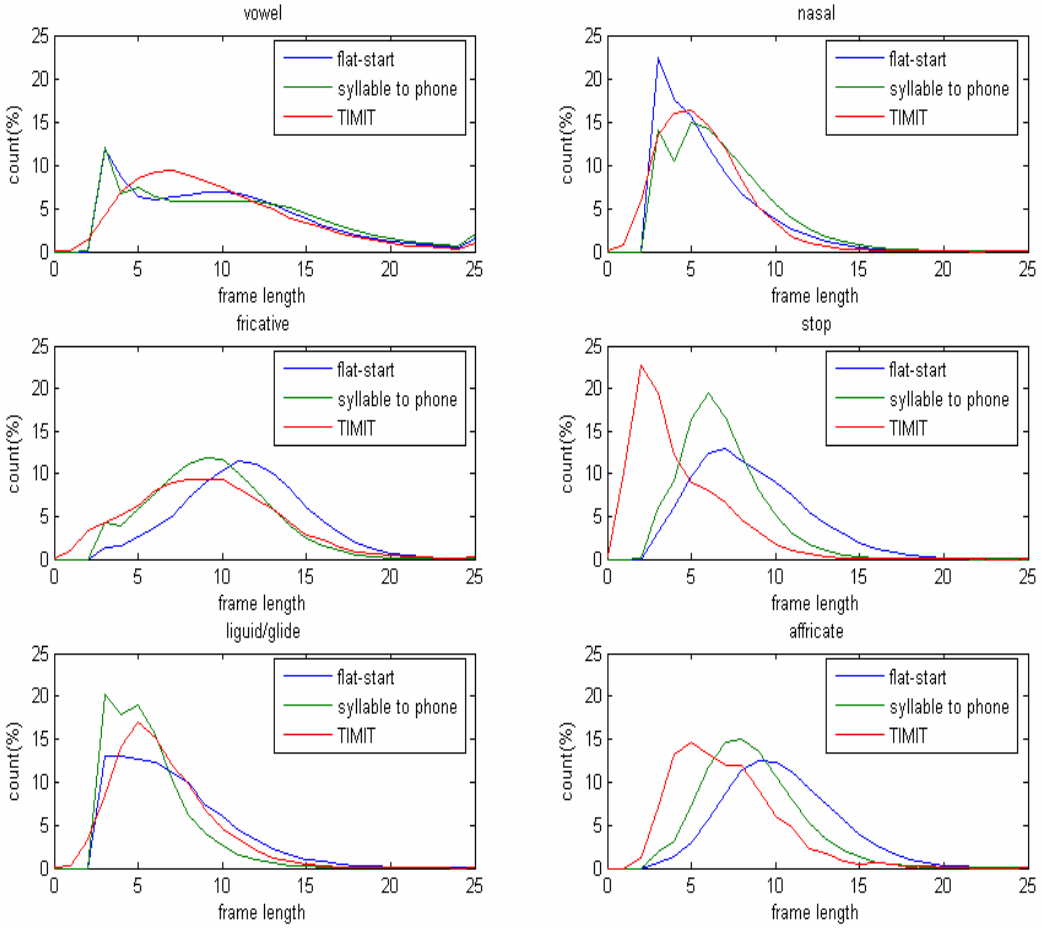


圖 3.6 發音方法切割長度比較圖

由圖 3.6 可以看出，從音節切割位置起始著手訓練音素的馬可夫模型後，對語料庫作切割的方法，除了母音與鼻音外，其餘子音的切割長度分佈曲線均向英文人工切割的長度分佈曲線偏移，這也表示中文 TCC300 語料庫中在子音的部份，切出長的長度能夠有效的降低，如此一來，能使從音節切割位置起始著手訓練音素的馬可夫模型對語料庫作強迫切割後所得到的音素切割位置比以 flat start 的方式所得到的音素切割位置改進許多，而我們也將以此較佳的切割位置當作是中文 TCC300 訓練語料的初始切割位置，重新再訓練各個發音方法的高斯混合模型後製作偵測器。

3.4 改進初始切割位置後的中文發音方法偵測器效能

我們利用由中文音節的切割位置起始經過均勻切割後所得到的音素切割位置，再以此音素切割位置訓練音素的隱藏式馬可夫模型後，對中文 TCC300 訓練語料作強迫切割所得到的音素切割位置當作初始切割位置，並以此切割位置去製作中文發音方法偵測器，底下我們將求取所製作出來的中文發音方法偵測器對中文測試語料作偵測的效能。

偵測器的架構如同 3.1.3 節所述，製作中文發音方法的貝式偵測器，其中每一個發音方法均訓練兩種高斯混合模型，一為 target model 、另一為 anti model，藉由計算每個音框在這兩種模型的似然度（likelihood），且考慮事先機率後，調整臨界值（threshold），以求得偵測器對中文測試語料的等錯誤率。

下表為發音方法偵測器的偵測效能比較。

表 3.7 發音方法偵測器效能比較

	Frame-based			
	ANN	Bayesian detector		
EER(%)	English	English	Mandarin (flat start)	Mandarin (syllable boundary constraint start)
Vowel	9.0	12.3	10.70	10.6
Fricative	11.3	10.0	15.7	11.80
Stop	14.5	16.7	11.5	8.14
Nasal	12.2	8.7	11.5	10.77
Glide/Liquid	15.9	16.3	9.2	8.33
Affricate		7.2	11.5	9.22
Silence	3.7	9.7	8.0	3.96

由表 3.7 可以明顯看出

1. silence 偵測器效能大大提昇，其 EER 由原本的 8.0%降至 3.96%。stop 偵測器的 EER 也從 11.5%降至 8.14%。fricative 偵測器的 EER 從原本最高 15.7%降至 11.80%。affricate 偵測器的 EER 從 11.5%降至 9.22%，可以看出對於子音的部份，偵測器的改善效能相當的大。
2. 其餘偵測器仍然有小幅改善，並且由上述結果可以看出，有效的將子音切割長度縮短後，其改善的偵測器效能相當明顯，並且可以看出從音節切割位置起始製作偵測器的方法比以 flat-start 的方式要好。

第四章 中文語音屬性偵測器的效能分析與討論

本章將探討在第三章製作出來的各個中文發音方法偵測器，其對中文測試語料的偵測效能及錯誤分析，分析內容包含中文發音方法偵測器對測試語料作偵測後，偵測錯誤的發生位置、以及是否對於某一類發音方法或特定的音素容易發生偵測錯誤，成為容易混淆的類別。再比較以英文發音方法偵測器，而最後我們將利用中、英文所製作出來的發音方法偵測器作跨語言的發音方法偵測，以比較發音方法在跨語言實驗中的性能。

4.1 中文發音方法偵測器之錯誤分析

下圖為中文音素的分類表，將中文音素分類為發音方式、發音方法及有聲無聲等，底下將利用此音素分類表的資訊幫助進行效能與錯誤分析。

大類	小類	發音方式		音素	
子音 (廣義) 25	子音 (狹義) 22	塞音 6	不送氣	ㄅ、ㄉ、ㄍ	無聲
			送氣	ㄆ、ㄊ、ㄎ	
		塞擦音 6	不送氣	ㄗ、ㄓ、ㄐ	
			送氣	ㄘ、ㄔ、ㄑ	
		擦音 6	不上提	ㄈ、ㄏ	
			舌葉上提	ㄙ、ㄕ、ㄒ	
				ㄖ	有聲
		鼻音 3		ㄇ、ㄋ、ㄥ	
		接近音 4	流音 1	ㄌ	
	半母音 4		介音 3	ㄧ、ㄨ、ㄩ	
母音 (廣義) 13	母音 (狹義) 9	單母音 9	中低音	ㄚ、ㄛ、ㄜ、ㄝ、ㄦ	
			高母音	ㄧ、ㄨ、ㄩ、ㄭ	
	雙母音 4	雙母音 4		ㄞ、ㄟ、ㄠ、ㄡ	

圖 4.1 中文音素分類表

4.1.1 中文發音方法偵測器容易偵測錯誤的發音方法類別分析

首先先觀察各個中文發音方法偵測器是否容易對某一發音方法發生偵測錯誤。下表為各個中文發音方法偵測器對測試語料中標記為某一個發音方法作偵測，且將其偵測為 target 的比例統計。

說明 1：

對測試語料中所有的 fricative 作偵測，vowel 偵測器將其偵測為 target 的比例此為偵測錯誤率。

$$\text{error rate} = \frac{\text{detect "fricative" as "vowel"}}{\text{total "fricative" frame}}$$

說明 2：

對測試語料中所有的 vowel 作偵測，vowel 偵測器將其偵測為 target 的比例此為偵測正確率（為表 4.1 的斜對角線上），而以 1 減去此偵測正確率的值即為 vowel 偵測器對 vowel 作偵測的錯誤拒絕率。

$$\text{accuracy rate} = \frac{\text{detect "vowel" as "vowel"}}{\text{total "vowel" frame}}$$

表 4.1 中文發音方法偵測器容易偵測錯誤的發音方法類別統計

Desired / Detector detected as target	vowel	fricative	stop	nasal	liquid	affricate	silence
vowel	89.40	11.24	8.33	26.33	36.66	9.41	1.13
fricative	7.91	88.21	25.93	7.15	38.67	36.49	6.75
stop	5.28	17.42	91.86	4.74	23.66	19.13	6.31
nasal	15.33	6.95	5.16	89.23	39.58	2.40	2.68
liquid	10.45	7.62	11.93	16.22	91.67	1.06	0.39
affricate	4.62	39.02	26.82	2.06	2.34	90.78	5.98
silence	0.63	6.44	8.75	3.53	0.71	6.27	96.04

由表 4.1 可以看出，silence 偵測器對 vowel 作偵測，僅僅將 0.63%的 vowel，誤偵測為 silence，又由於語料庫中 vowel 的總量是最多的，因此如此低的錯誤偵測，使得 silence 偵測器對中文測試語料的等錯誤率能降到僅僅只有 3.96%。除此之外 affricate 與 nasal 以及 affricate 與 liquid 此兩類均有極低的偵測錯誤率，顯示出偵測器能夠有效的區分出此兩類。

affricate 偵測器以及 fricative 偵測器在互相偵測對方時，其所造成的錯誤率，均高達 36%以上，此兩類在中文裡均包含六個音素，且在中文語料庫中所佔的音框數非常接近，fricative 約佔語料庫所有音框數的 7.98%，affricative 則為 7.51%，顯示此兩種語料容易發生混淆的情況。下圖 4.2 為中文發音方法偵測器偵測中文測試語料的實際例子，我們秀出 silence、vowel、fricative、affricate 偵測器對測試語料中的每一個音框作偵測後，偵測為 target 的事後機率圖[9]。

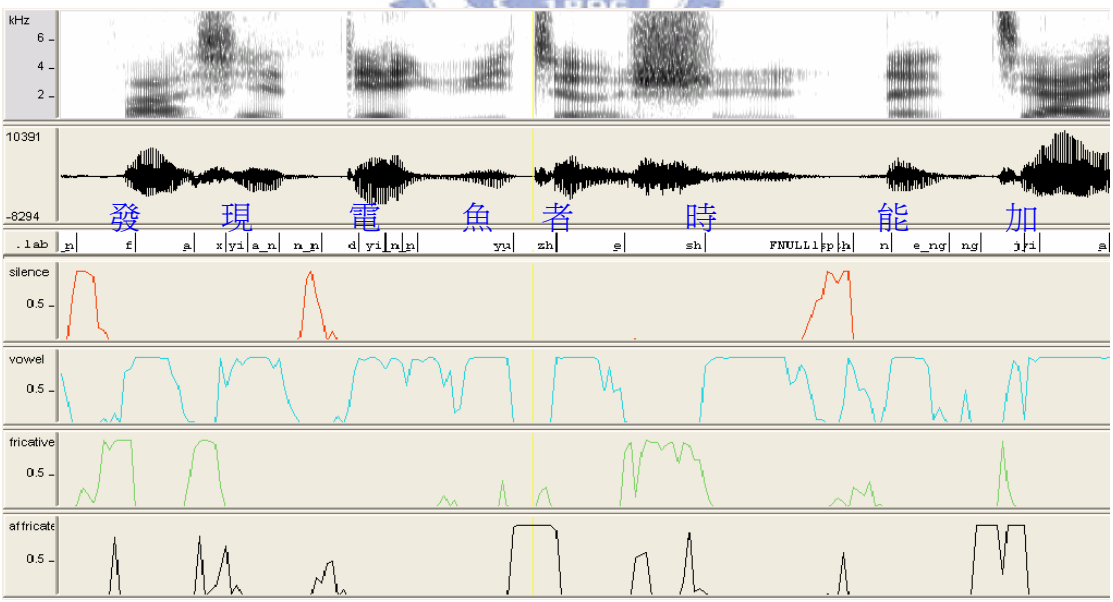


圖 4.2 fricative 與 affricate 相互偵測混淆實例

由圖 4.2 可以看出，fricative 與 affricate 兩者的確有互相混淆的情況，而由下圖 4.3 可以看出 fricative 與 affricate 其兩個高斯混合模型的平均值 C1,C2 分佈相當接近，可以得知此兩種發音方法類別的聲學特性相當的近似。除此之外，stop 與這兩種發音方法類型互作偵測，錯誤率亦明顯的高，因此 affricate 以及 fricative 此兩類與 stop 亦有部份混淆的情況發生。

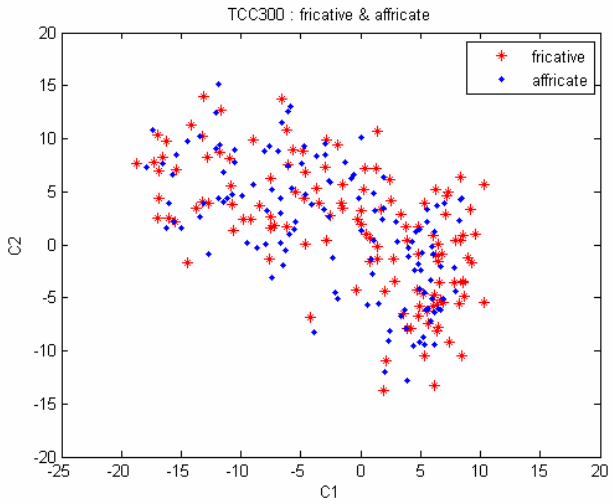


圖 4.3　fricative 與 affricate 兩個高斯混合模型的平均值 c1,c2 分佈

對於 vowel 與 nasal 而言，其交互偵測的錯誤率亦很高，尤其是以 nasal 偵測器偵測 vowel 的錯誤率高達 15.33%，是所有偵測器偵測 vowel 錯誤率最高的一個，而這也使得 nasal 偵測器的等錯誤率達到 10.77%的原因。而 nasal 與 vowel 互相偵測錯誤率偏高，應該是發生在屬於子音的 nasal 與 vowel 交界處，或是 vowel 與鼻音韻尾的交界處這兩處之一，這部份將在下面繼續探討。

另外在表 4.1 的部份，以 vowel、fricative、stop、nasal 這四類發音方法偵測器，對測試語料中屬於 liquid 的語料作偵測，除了 stop 偵測器錯誤率為 23.66%，其餘三類的錯誤率均高達 36%以上，但以 liquid 偵測器對上述的四種發音方法類別的語料作偵測，其錯誤率分別為 10.45%、7.62%、11.93%、16.22，聲學性質

若近似的話，其交互偵測錯誤率應該會接近對稱，但由 liquid 與上述四類的發音方法類別明顯有不相稱的錯誤率，因此我們認為這四類發音類別裡，應該會有特別一個或某一部份的音素與 liquid 的聲學性質相近，因此導致這四類發音方法偵測器對 liquid 的偵測錯誤率特別高。因此接下來我們將對表 4.1 作更細的分類，統計各個發音方法偵測器偵測各個中文音素的錯誤率排序。

4.1.2 中文發音方法偵測器容易偵測錯誤的音素類別分析

前一小節列出各個發音方法偵測器對語料庫中的各個發音方法作偵測後，觀察各個發音方法偵測器容易發生偵測錯誤地方，以及對何種發音方法容易偵測錯誤，在本節中，我們將細分為各個發音方法偵測器對各個中文音素作偵測，觀察發音方法偵測器是否對某一發音方法的偵測錯誤是來自於所偵測的發音方法其所屬的某一或某群音素所造成。下表為各個中文發音方法偵測器偵測各個中文音素的錯誤率排序。

說明：

對測試語料中所有的ㄖ /r/ 作偵測，vowel 偵測器將其偵測為 target 的比例此為偵測錯誤率。

$$\text{error rate} = \frac{\text{detect "ㄖ /r/ " as "vowel"}}{\text{total "ㄖ /r/" frame}}$$

表 4.2 中文發音方法偵測器容易偵測錯誤的音素類別統計

Rank / Detector	1	2	3	4	5	6	7	
vowel	ㄖ /r/ 49.56	/ng/ 31.79	/n_n/ 24.22	ㄏ /h/ 21.35	ㄋ /n/ 20.30	ㄇ /m/ 14.60		
fricative	ㄑ /q/ 51.3	ㄘ /c/ 44.53	ㄗ /z/ 42.05	ㄔ /ch/ 37.27	ㄓ/zh/ 36.95	ㄅ /b/ 32.82	other stop 22~26	ㄐ /j/ 21.36
stop	ㄈ /f/ 64.66	ㄏ /h/ 47.73	ㄔ /ch/ 31.41	ㄘ /c/ 28.87	ㄗ /z/ 26.63	ㄓ /zh/ 21.62		
nasal	ㄥ /e_ng/ 41.97	ㄖ /r/ 28.07	ㄣ /e_n/ 25.84	FNULL2 24.96	ㄤ /a_ng/ 21.28	ㄟ /ei/ 17.68		
liquid	ㄖ /r/ 76.61	ㄋ /n/ 68.13	ㄇ /m/ 64.30	ㄉ /d/ 22.26	ㄩ /yu/ 20.19	ㄧ /yi / 17.67		
affricate	ㄙ /s/ 57.5	ㄊ /t/ 53.11	ㄒ /x/ 53.66	ㄕ /sh/ 46.52	ㄆ /p/ 28.14	ㄉ /d/ 28.37	ㄎ /k/ 26.52	
silence	ㄈ /f/ 18.34	ㄆ /p/ 16.51	ㄅ /b/ 13.55	ㄊ /t/ 11.71	ㄏ /h/ 11.63	ㄎ /k/ 10.31		

P.S.　ng、n_n 為鼻音韻尾、FNULL2 為空韻母

表 4.2 列出偵測器對中文音素作偵測的錯誤率排序，而由於在表 4.1 中已經有列出各個發音方法偵測器對 liquid 的偵測，而 liquid 只包含一個ㄌ/ l /音，因此在表 4.2 中我們略掉對音素ㄌ/ l /的偵測項目。

從表 4.2 可以看出，以 vowel 偵測器偵測中文音素錯誤率排序中，除了已知在表 4.1 所列出的對 liquid 的/ l /音具有高偵測錯誤率外，其餘前六項中有四項音素是屬於 nasal，符合在表 4.1 所得到結果，且對於 nasal 的高偵測錯誤率分佈在 /n_n/、/ng/ 此兩鼻音韻尾的音素，這兩類鼻音韻尾的音框數佔測試語料中 nasal 的音框數的 87.14%，因此可知 vowel 偵測器偵測 nasal 語料發生偵測錯誤的地方在於 vowel 與鼻音韻尾的交接處，而從 nasal 偵測器來看也是如此，其偵測錯誤率排序的前幾項中有三項是屬於有鼻音韻尾的 vowel。

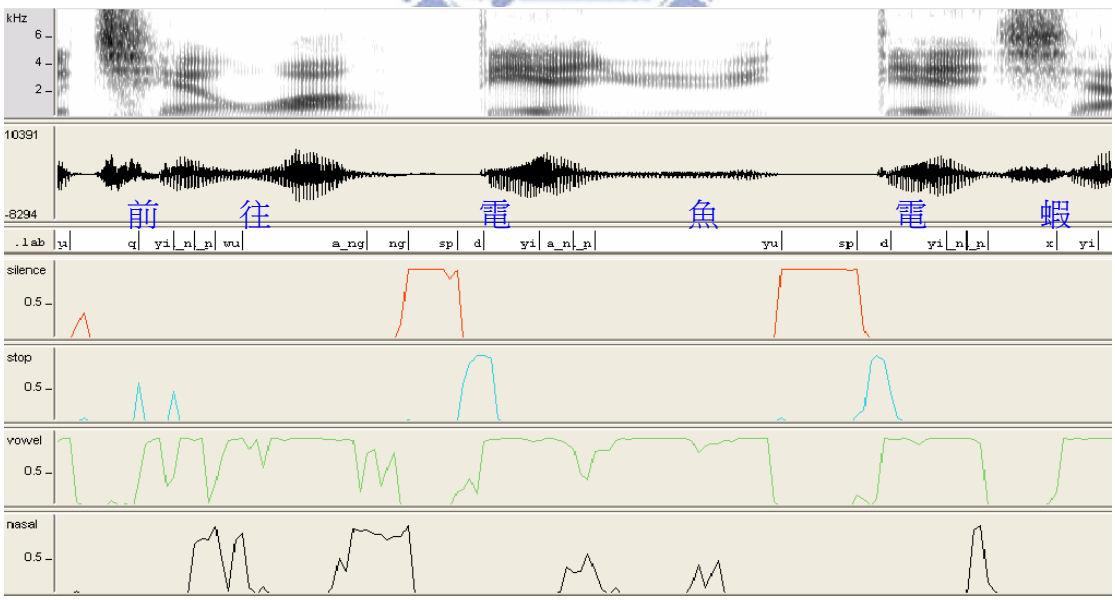


圖 4.4　vowel 與 nasal 相互偵測混淆實例

除此之外 vowel 偵測器偵測到最高錯誤率的音素是ㄖ/r/，而這個音素是屬於 fricative，由圖 4.1 可以得知，ㄖ/r/音是唯一一個有聲（voiced）的 fricative，而 vowel 與 nasal 偵測器均對ㄖ/r/這個音素有高的錯誤偵測率，應該是因為均同屬有聲的音。

對於 fricative 偵測器而言，對 affricative 的音素偵測的效果均較差，對不送氣的ㄐ/j/音作偵測的錯誤率為 21.36%是裡面最低的，而錯誤率為最高的 51.3%為對送氣的ㄑ/q/音作偵測，其餘皆在 36~44%之間。而 affricate 偵測器偵測屬於無聲的且舌葉上提的 fricative 音素，其偵測效能明顯很差，錯誤率在 46%~57%之間，但對於偵測唯一一個屬於有聲的 fricative ㄖ/r/音，錯誤率僅僅只有 5.92%。而這兩個偵測器互相偵測對方效能不好的原因在前面有敘述過，是因為 fricative 與 affricate 的聲學特性近似所導致，但 fricative 的偵測器的偵測效能似乎還比 affricate 偵測器來的差。另外 fricative 偵測器對 stop 的音素偵測，以不送氣的ㄅ/b/錯誤率為最高，達到 32.82%，其餘皆在 22~26%之間。而 affricate 偵測器偵測屬於送氣的 stop 的ㄊ/t/音，其錯誤率高達 53.11%，但對於偵測不送氣的 stop 的ㄅ/b/與ㄍ/g/，錯誤率僅有 15.02%及 12.21%。

對於 stop 偵測器而言，偵測錯誤率最高的是對屬於舌葉不上提的 fricative：ㄈ/f/與ㄏ/h/音素，錯誤率為 64.66%及 47.73%。而對於屬於舌尖前的ㄗ/z/及ㄘ/c/與屬於舌尖後的ㄔ/ch/及之ㄓ/zh/的 affricate 的音素，偵測錯誤率為 21~31%，效果亦較差。

前面曾提及過的在表 4.1 中以 vowel、fricative、stop、nasal 這四類發音方法偵測器，對測試語料中屬於 liquid 的語料作偵測，除了 stop 偵測器錯誤率為 23.66%，其餘三類的錯誤率均高達 36%以上。而從表 4.2 中可以看出偵測錯誤率最高的 76.61%的ㄖ/r/是屬於 fricative，其次為 68.13%的ㄋ/n/與 64.30%的ㄇ/m/，

再來為ㄉ/d/，而最後為ㄩ/yu/及ㄧ/yi /，而後面三個的錯誤率約在 20%上下。從圖 4.1 可以得知 liquid 的ㄌ/l/與 fricative 的ㄖ/r/、nasal 的ㄋ/n/及ㄇ/m/同屬於有聲音，liquid 的ㄌ/l/與 vowel 的ㄩ/yu/及ㄧ/yi /同屬於接近音(approximate)。然而 liquid 偵測器為何對ㄖ/r/及ㄋ/n/的偵測錯誤率高達 68%以上？在此我們認為可能的原因是錄製者在發這三類音時，容易因為發音的不精確導致混淆而發出其他的音，例如中文字中的“熱”與“樂”，ㄖ與ㄌ容易發音混淆、或者在發“難”或者“能”的音，ㄌ與ㄋ亦容易發音混淆。這些發音混淆的情況均有可能導致這三類發音方法偵測器的偵測錯誤。

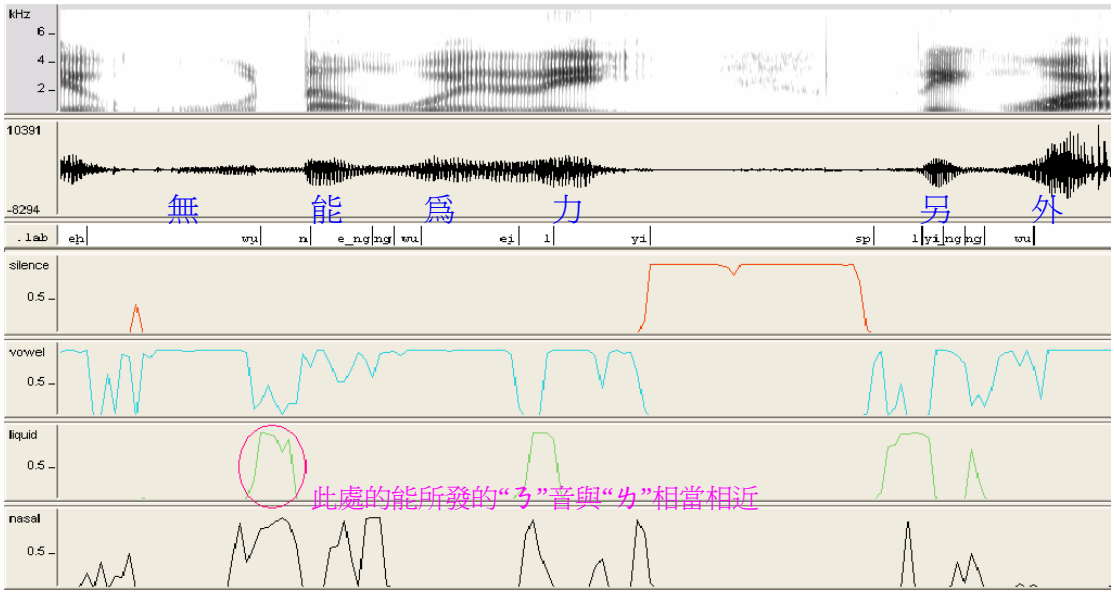


圖 4.5 發音誤差(ㄌ與ㄋ)所導致的偵測錯誤實例

最後對於 silence 偵測器而言，從表 4.1 可以看出其偵測錯誤最高的音素類別為 stop，又由於 stop 後面一定是接 vowel，而在前面的討論中亦提到 silence 偵測器對 vowel 語料偵測的錯誤率相當的低，僅有 0.63%，因此 silence 偵測器偵測 stop 語料所發生的錯誤，應是發生在 stop 音的起始點。這也符合在第三章所作的以 HMM 對中文語料庫作強迫切割的初始切割位置，stop 的平均長度為 6.74 個音框，而人工切割英文的 stop 平均長度為 4.12 個音框，兩者仍存著 2 個音框長度的差異。

4.2 英文發音方法偵測器之錯誤分析

由於在第二章曾製作過英文的發音方法偵測器，再分析了中文發音方法偵測器的效能與錯誤後，我們亦對英文發音方法偵測器分析其容易發生偵測錯誤的發音方法類別，觀察是否和中文發音方法偵測器有相近的錯誤。下表為各個英文發音方法偵測器對測試語料中標記為某一個發音方法作偵測，且將其偵測為 target 的比例統計，此處相關的說明請參考 4.1.1 節。

表 4.3 英文發音方法偵測器容易偵測錯誤的發音方法類別統計

Desired / Detector detected as target	vowel	fricative	stop	nasal	glide	affricate	silence
vowel	87.70	4.78	15.88	17.79	39.92	3.38	3.76
fricative	4.58	90.05	27.40	6.83	5.65	68.03	15.05
stop	12.64	23.43	83.34	14.17	22.30	52.01	19.23
nasal	10.13	4.95	9.18	92.30	11.78	1.61	9.22
glide	27.27	3.59	24.07	13.62	83.66	3.22	5.40
affricate	1.41	27.35	16.99	1.13	1.96	92.75	5.13
silence	3.29	16.87	25.03	13.31	5.67	13.13	90.29

由表 4.3 可以看出，在英文發音方法偵測器中，vowel 與 glide 混淆的情況最為明顯，根據表 2.1 的英文音素的發音方法及發音位置的分類表可以看出，glide 與 vowel 其所組成的音素在發音位置的分類上有高度的重疊，此應為兩者混淆情況嚴重的原因。而 affricate、fricative 與 stop，vowel 與 nasal 亦有混淆偵測的情況，此情況與中文偵測器的錯誤為相似的地方。而 affricate 與 nasal 以及 affricate 與 glide 的互相偵測錯誤率為最低，顯示出此種情況，偵測器能夠有效的區分此兩類的分別，這與分析中文偵測器所觀察到的錯誤有相同的情形。

4.3 跨語言的發音方法偵測器效能比較

由 4.1 與 4.2 節的中英文發音方法偵測器的錯誤分析中，可以得知利用此兩種語言所製作出來的發音方法偵測器，其偵測器的偵測錯誤有類似的情況發生，又因為我們知道發音方法應該是與語言無關的（language independent），因此我們欲作跨語言的偵測實驗。由於英文 TIMIT 語料庫已有人工標記好的音素切割位置，因此我們嘗試利用這人工標記的英文音素位置求取中文發音方法偵測器對其的偵測效能。再比較以英文語料所訓練出來的發音方法的高斯混合模型，以及在第三章所得到的從音節切割位置著手所訓練出來的音素高斯混合模型，兩者對中文語料庫作強迫切割後，所得到的發音方法切割位置的統計比較。

4.3.1 中文發音方法偵測器偵測英文測試語料之效能

我們以中文發音方法偵測器去偵測已有人工標好音素位置的英文 TIMIT 語料庫的測試語料來求取的偵測器效能，並且與英文發音方法偵測器的偵測效能作比較。

表 4.4 英文與中文發音方法偵測器對英文 TIMIT 語料庫的測試語料所做的偵測結果

Test data : TIMIT	**Frame-based detector**	
EER(%)	**detector trained from English**	**detector trained from Mandarin**
vowel	12.3	21.3
vricative	10.0	26.1
stop	16.7	31.0
nasal	8.7	15.6
glide (liquid)	16.3	44.5 /l/
affricate	7.2	18.5
silence	9.7	24.0

由表 4.4 我們可以很清楚的看出，各個中文發音方法偵測器偵測英文測試語料的 EER，幾乎是英文發音方法偵測器所偵測得到的兩倍上下。而對於 glide (liquid)而言，以中文的偵測器去偵測英文的語料，其 EER 高達 44.5%，其中的原因有可能與中文的 liquid 只有ㄌ/l/一個音素，而英文的 glide 有七個音素相關。

4.3.2 中、英文發音方法模型對語料庫的切割位置比較

利用英文語料所訓練出來的發音方法的高斯混合模型，將其視為狀態數為 1 的英文發音方法隱藏式馬可夫模型（1-state HMM）去對中文訓練語料作強迫切割，並且觀察其結果與以狀態數為 3 的由音節切割位置著手所訓練出來的中文音素的隱藏式馬可夫模型（3-states phone HMM）對中文訓練語料作強迫切割的音素初始切割位置轉為發音方法切割位置後之間的差異。

表 4.5　英文發音方法高斯混合模型對中文訓練語料作強迫切割的統計資料

	1-state HMM（English）	HMM（Mandarin）
發音方法	平均長度	平均長度
vowel	8.80	10.28
fricative	8.71	9.25
stop	4.31	6.74
nasal	7.26	6.81
liquid	8.18	5.59
affricate	3.88	8.51

由表 4.5 可以得知兩種切出來的各個發音方法的平均長度差異頗大，尤其對於 stop、affricate 而言，更是明顯，雖然以英文的發音方法 stop 的高斯混合模型對中文語料作強迫切割後，其所切出來的 stop 平均長度為 4.31 個音框，極為接近人工切割的英文 stop 的平均長度，但是對於 affricate 而言，英文所訓練出來的 affricate 高斯混合模型所切的平均長度卻僅有 3.88 個音框數，這是比較不合理的地方。

接下來我們觀察以英文的發音方法高斯混合模型當作是狀態數為 1 的隱藏式馬可夫模型以及狀態數為 3 的中文音素的隱藏式馬可夫模型，兩者對中文語料作強迫切割後的切割位置的差異。

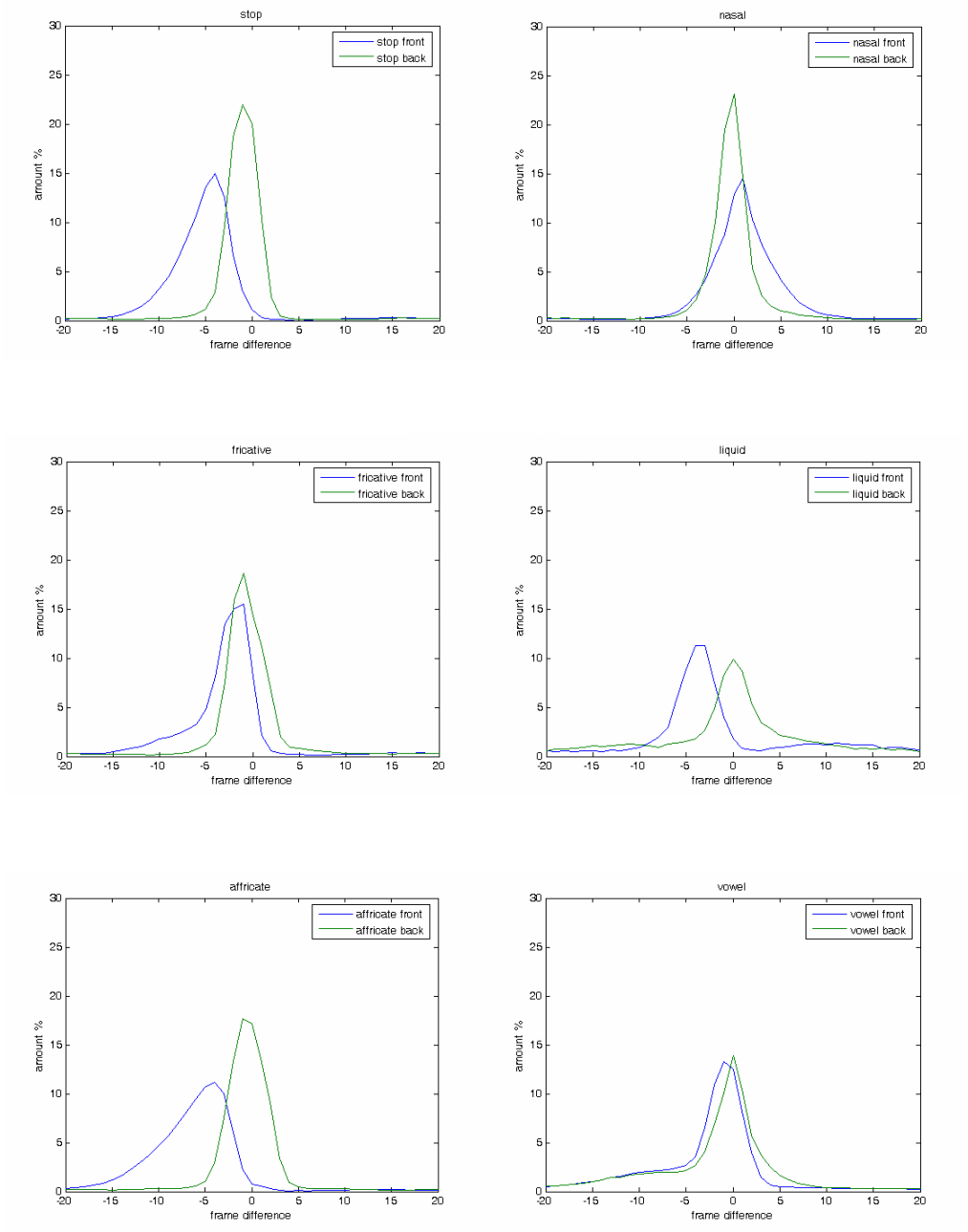


圖 4.6 以英文 GMM 與以中文 HMM 對中文訓練語料

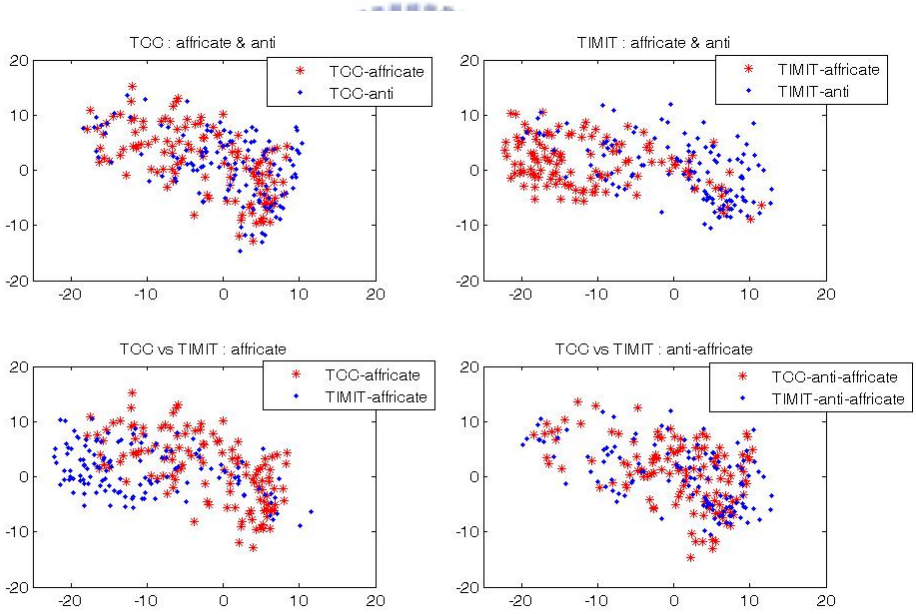
作強迫切割的切割置差異比較

對於各個發音方法觀察兩個切割點，一為起始切割位置（圖中標 front），二為結束切割位置（圖中標 back）。以 stop 的圖而言，藍色曲線代表以中文 HMM 與英文 GMM 所作的 forced-alignment 起始切割差異。藍色曲線代表起始切割位置的差距，而藍色曲線大部分分佈在 x 軸的負值，代表以中文 HMM 所切的起始切割位置在時間上比以英文 GMM 所切的時間點還早。由上圖可以看出兩類的切割結果差距頗大。

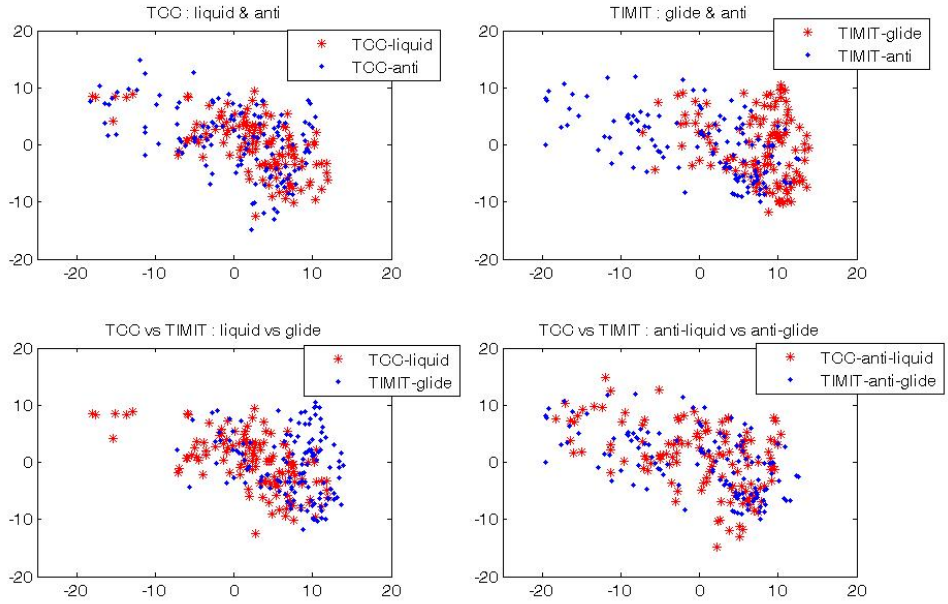
4.3.3 跨語言發音方法偵測器之效能討論

根據以上所得到的資料，對於以中英文發音方法偵測器交互偵測中英文語料的偵測效果不佳的結果，我們認為可能的原因應該是兩種語料的錄音環境所造成的不匹配。對於這點，我們觀察以兩種語言所訓練出來的偵測器模型其參數值的分佈，以及兩種語料所求出來的 MFCC 分佈。下圖為中文發音方法的高斯混合模型與英文發音方法的高斯混合模型，兩者 128 個高斯混合(Gaussian mixture)中，每一個高斯混合模型平均值的 C1 和 C2 分佈圖

1. affricate



2. liquid or glide



4. silence

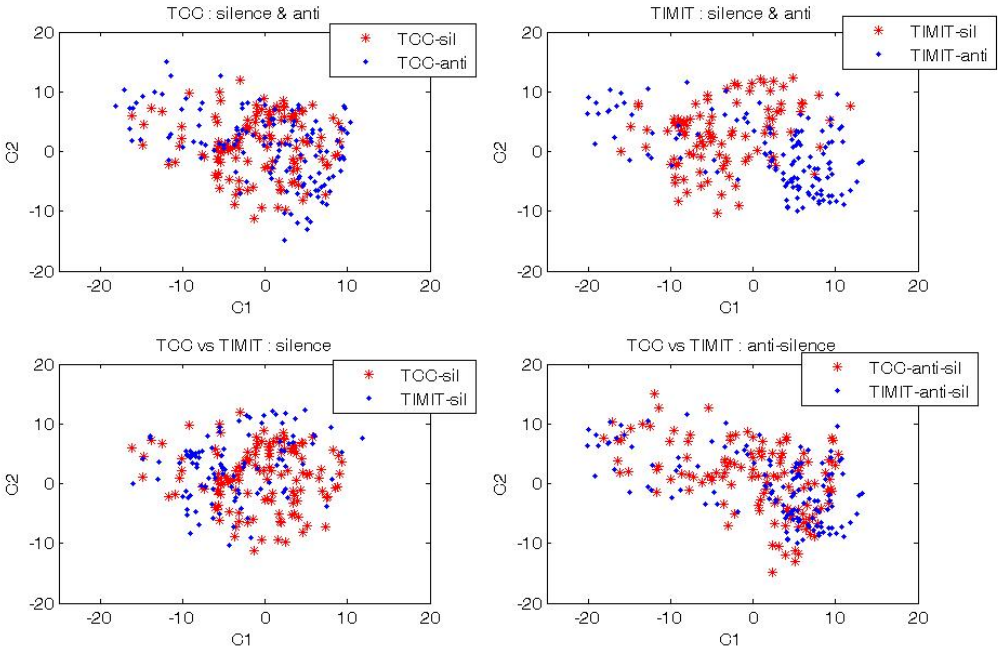


圖 4.7 中、英文發音方法高斯混合模型的 C1 與 C2 平均值分佈

下圖為 TIMIT 語料庫與 TCC300 語料庫全部語料 MFCC 39 維各維平均值分佈圖

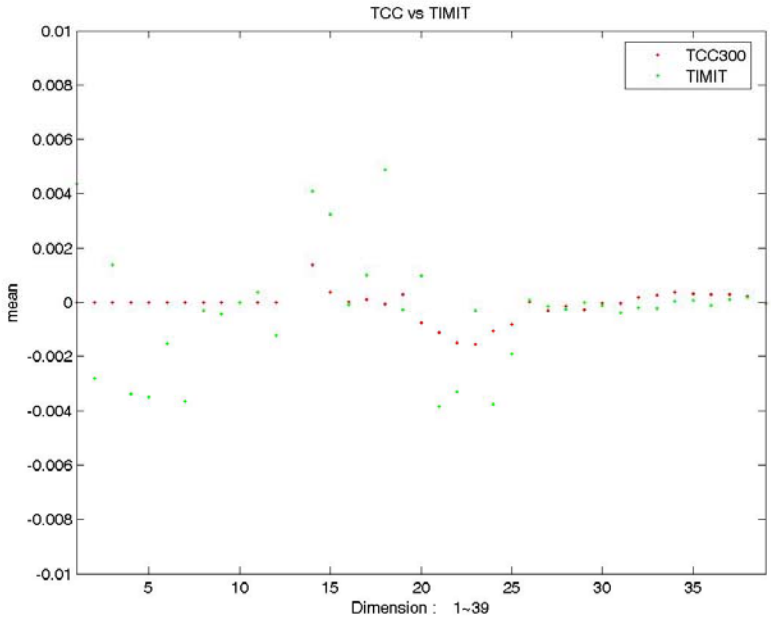


圖 4.8　TIMIT 語料庫與 TCC300 語料庫全部語料 MFCC39 維各維平均值分佈圖

由圖 4.8 及圖 4.7 可以看出兩個語料庫的語料分佈以及所訓練出來的發音方法高斯混合模型的參數分佈差異頗大，因此對於未來要作中、英文跨語言的語音屬性偵測可能要先解決語料庫錄製時的環境不匹配問題。

第五章　結論與未來展望

5.1 結論

在本論文中，首先我們以高斯混合模型為基礎製作英文語音屬性貝氏偵測器，並將其效能當作是後面所要製作的中文語音屬性偵測器的對照組，並且將製作出來的發音方法與發音位置偵測器的偵測結果結合，對某些發音方法及發音位置組合的情況下，能降低偵測此兩者所對應的音素或音素集合的等錯誤率。

而在中文語料庫沒有良好的切割位置情況下，我們考慮了以中文音節的切割位置起始，經過均勻切割長度給其所組成的音素，並以此切割位置重新訓練隱藏式馬可夫模型來對語料庫求取音素的切割位置，以製作中文語音屬性偵測器。其中中文的音節切割位置裡，我們考慮加入一個背景模型來模擬語料庫中背景雜訊及背景人聲，以期望切出較好的靜音位置來訓練語音屬性偵測器模型。而以此方法所製作出來的中文語音屬性偵測器其偵測效能在與基本對照系統相比後的確有明顯的改善。

最後在中文語音屬性偵測器的效能與錯誤分析中，發音方法的摩擦音(fricative)與塞擦音(affricate)，因為其兩者聲學特性相近，而導致在互相偵測時會產生混淆，而爆破音(stop)也與此兩類有部份混淆的情況。除此之外，中文音素中的ㄖ/r/是唯一一個有聲的摩擦音，容易使得母音(vowel)、鼻音(nasal)、流音(liquid)的偵測器對其產生錯誤偵測。而以母音與鼻音偵測器偵測帶有鼻音韻尾的母音時，其交界為一個模糊的地帶，容易造成此兩種偵測器的錯誤偵測。

5.2 未來展望

在中文語音屬性偵測器的部份，仍然有很多路線可以去探索及改進，無論是針對所要偵測的對象，求取具有鑑別性的語音特徵參數，抑或是採用不同的偵測架構。本論文提供一個方法去製作基本的中文語音屬性偵測器以及偵測效能參考及評比，並且以實驗的方式，獲得在偵測中文語音中音素、發音方法或發音位置之間交互的影響，以提供後人在對中文語音偵測時的參考。希望藉由這些經驗、知識的累積，建立一個以知識為基礎(knowledge-based)加上資料驅動(dara-driven)的新一代語音辨識系統架構，以推進語音辨識能力的突破。

參考文獻

【1】 C.-H. Lee, “From knowledge-ignorant to knowledge-rich modeling：A new speech research paradigm for next generation automatic speech recognition” Proc. ICSLP2004, Keynote speech, 2004

【2】 J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D.S. Pallett, and N. L.Dahlgren, “DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus,” U.S.Dept. of Commerce,NIST, Gaithersburg, MD, February 1993.

【3】 Chang, S., Greenberg, S. and Wester, M. “An elitist approach to articulatory-acoustic feature extraction” Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech-2001), 2001

【4】 王小川，“語音訊號處理”，全華科技圖書，中華民國九十三年三月。

【5】 洪惟仁，“聲韻學講義教材”，元智大學中國語文學系。

【6】 C.-H. Lee, “A Study on Separation between Acoustic Models and Its Applications,” Proc. ICASSP2005

【7】 S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey,V. Valtchev, P. Woodland, ”The HTK Book (for HTK Version 3.3)”, Cambridge University, 2005

【8】 R. P. Lippmann, L C. Kukolich, and E. Singer, “LNKnet: Neural Network, Machine Learning, and Statistical Software for Pattern Classification”, Lincoln Laboratory Journal, vol. 6, pp. 249-268, 1993.

【9】 Wavesufer Homepage : http://www.speech.kth.se/wavesurfer/

【10】Bilmes J.A., "A Gentle Tutorial of the EM algorithm and its application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models",

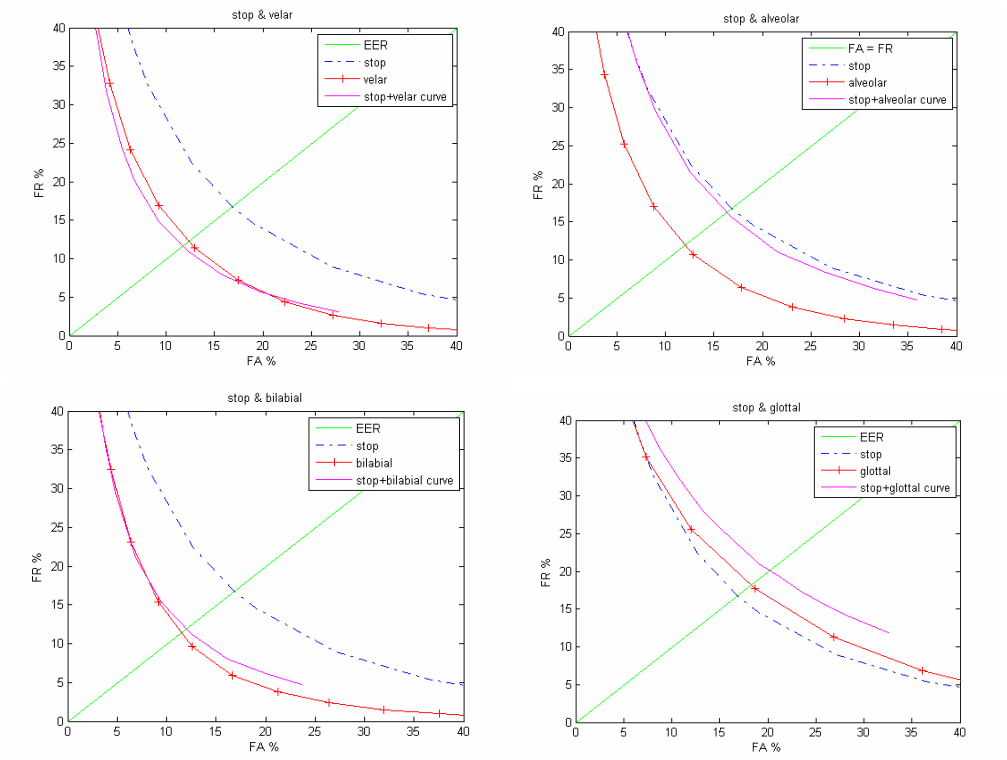
ICSI-Technical Report-97-021, 1997.

【11】Sérgio Paulo，Luís C. Oliveira，“Automatic Phonetic Alignment and Its Confidence Measures”, Advances in Natural Language Processing,Vol.3230, pages 36-44,2004

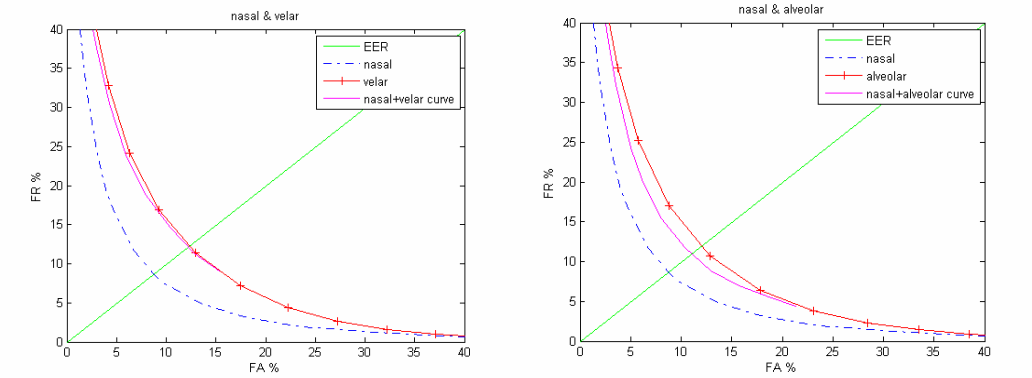
附錄一

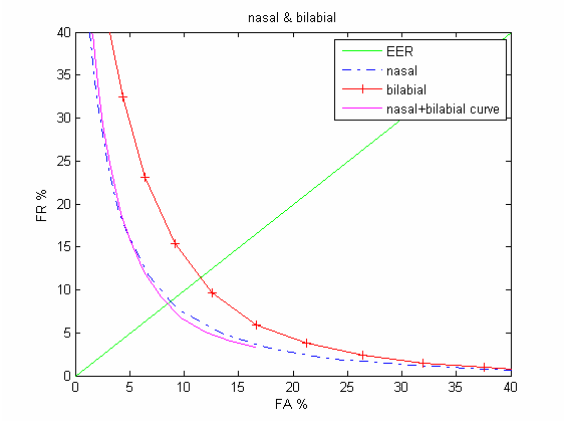
英文發音方法結合發音位置的 FA-FR 曲線圖

1. 發音方法 stop ＋ poison 共四種 FA-FR 的曲線圖

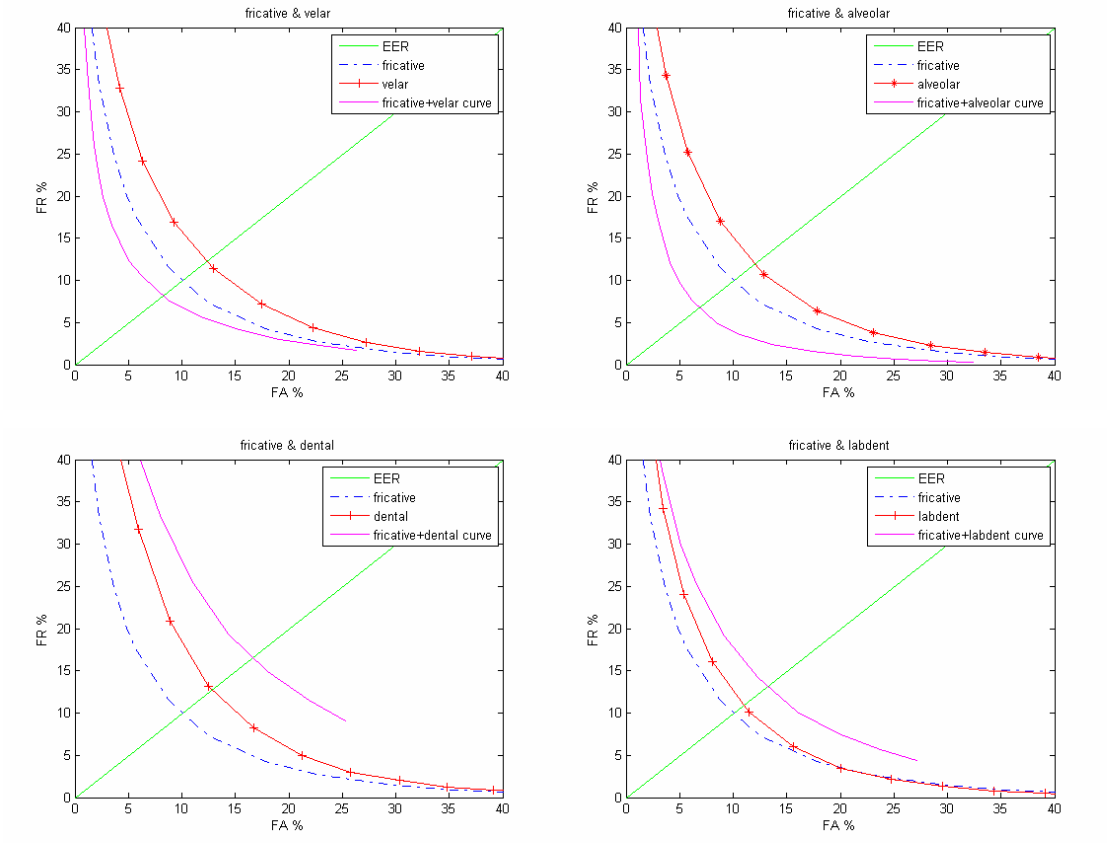


2. 發音方法 nasal ＋ poison 共三種 FA-FR 的曲線圖

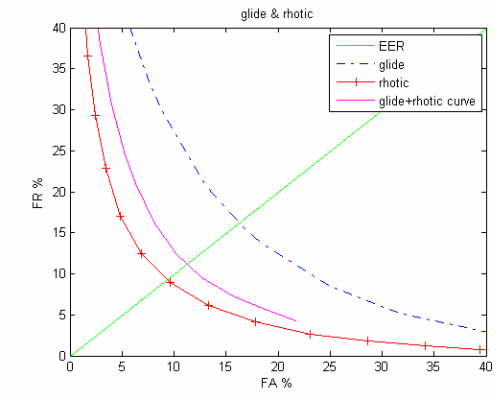


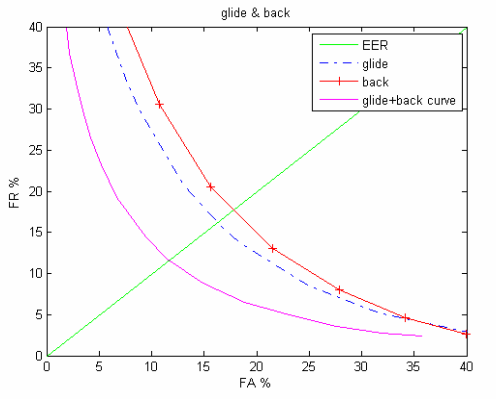


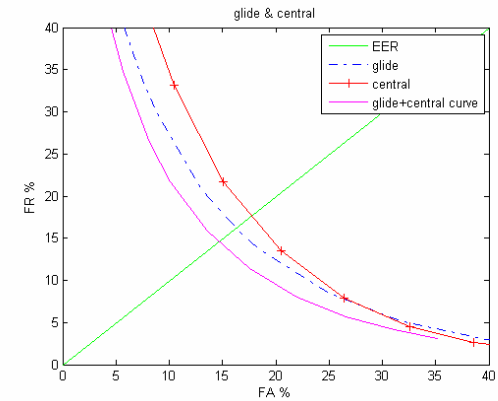
3. 發音方法 fricative ＋ poison 共四種 FA-FR 的曲線圖

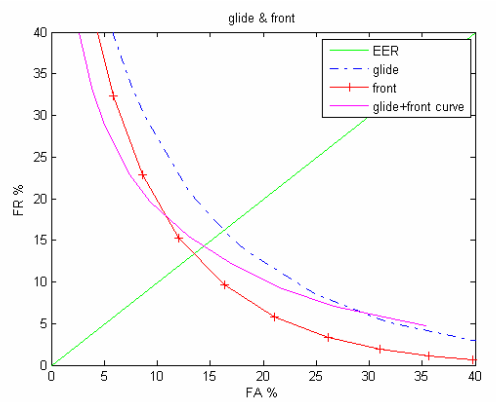


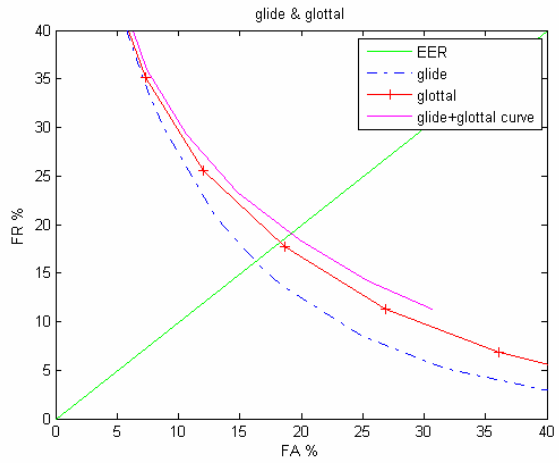
4. 發音方法 glide ＋ poison 共五種 FA-FR 的曲線圖



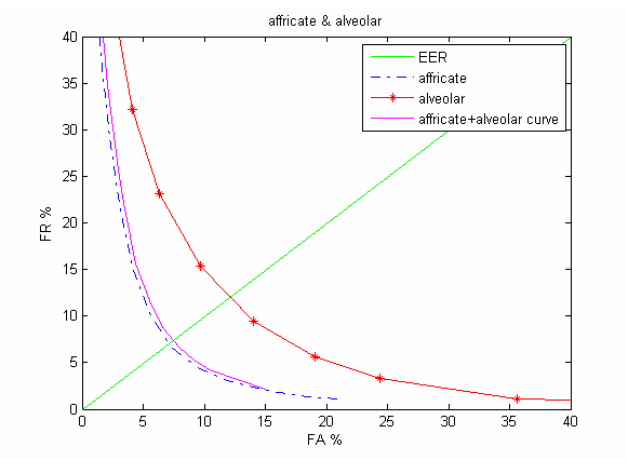




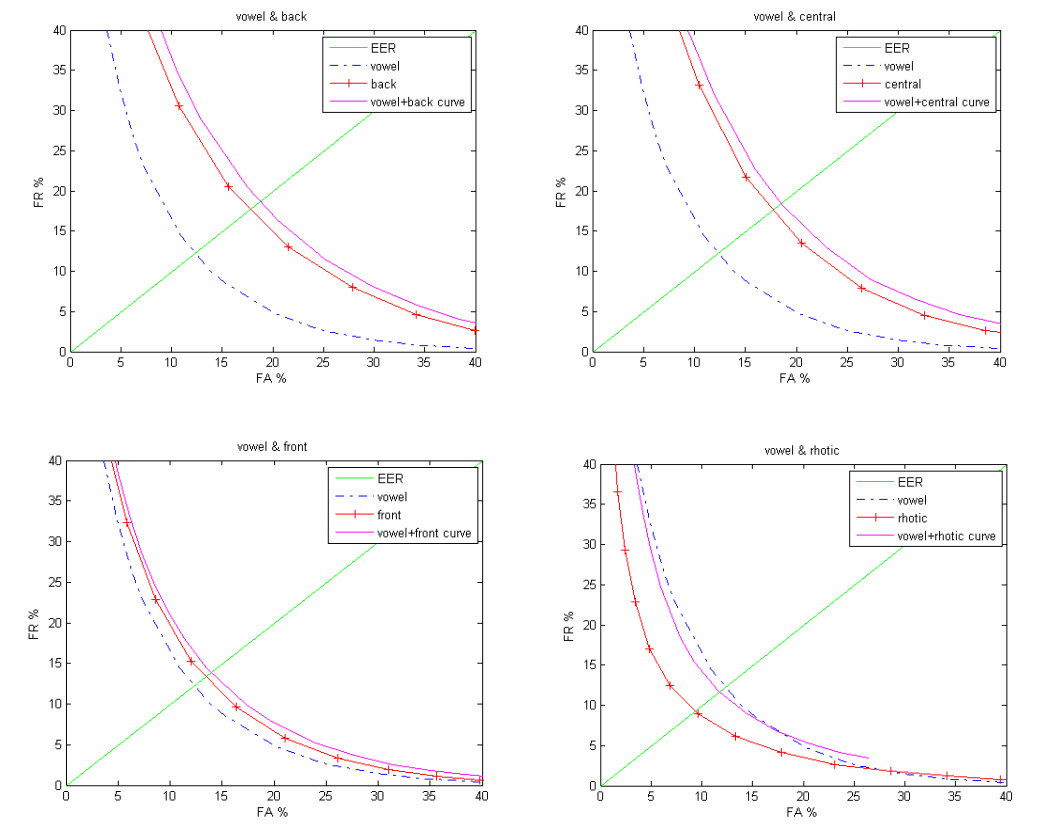




5. 發音方法 affricate ＋ poison 共一種 FA-FR 的曲線圖



6. 發音方法 vowel ＋ poison 共四種 FA-FR 的曲線圖



附錄二

中文音素分類及漢拼、注音對照表

表一　21 類聲母表

編號	注音	漢拼		編號	注音	漢拼		編號	注音	漢拼
1	ㄅ	b		9	ㄍ	g		17	ㄕ	sh
2	ㄆ	p		10	ㄎ	k		18	ㄖ	r
3	ㄇ	m		11	ㄏ	h		19	ㄗ	z
4	ㄈ	f		12	ㄐ	j		20	ㄘ	c
5	ㄉ	d		13	ㄑ	q		21	ㄙ	s
6	ㄊ	t		14	ㄒ	x				
7	ㄋ	n		15	ㄓ	zh				
8	ㄌ	l		16	ㄔ	ch				

表二　16 類韻母表

編號	注音	漢拼		編號	注音	漢拼
1	ㄚ	a		9	ㄢ	a_n
2	ㄛ	o		10	ㄣ	e_n
3	ㄜ	e		11	ㄤ	a_ng
4	ㄝ	eh		12	ㄥ	e_ng
5	ㄞ	ai		13	ㄧ	yi
6	ㄟ	ei		14	ㄨ	wu
7	ㄠ	ao		15	ㄩ	yu
8	ㄡ	ou		16	ㄦ	er

Ps.實際“ㄢ”　“ㄣ”　“ㄤ”　“ㄥ”的漢拼分別為

“an”　“en”　“ang”　“eng”

在此我們將細分至鼻音韻尾，因此做些改變

表三　　空母音　　與　　鼻音韻尾

編號	符號		編號	符號
1	FNULL1		1	n_n
2	FNULL2		2	ng

Ps.其中 n_n 為”ㄢ”與”ㄣ”的鼻音韻尾，ng 為”ㄤ”與”ㄥ”的鼻音韻尾