國 立 交 通 大 學

電信工程研究所

碩士論文

國語語音屬性偵測器之製作及其應用

Implementation and applications of Mandarin Pronunciation manner detector

研 究 生：林彥邦

指導教授：王逸如 博士

中華民國一百零二年六月

國語語音屬性偵測器之製作及其應用

Implementation and applications of Mandarin Pronunciation manner detector

研 究 生：林彥邦　　Student：Yen-Bang Lin

指導教授：王逸如 博士　　Advisor：Dr. Yih-Ru Wang

國 立 交 通 大 學

電信工程研究所

碩 士 論 文

A Thesis
Submitted to Institute of Communication Engineering
College of Electrical and Computer Engineering
National Chiao Tung University
in Partial Fulfillment of the Requirements
for the Degree of
Master
In

Communication Engineering

June 2013

Hsinchu, Taiwan, Republic of China

中華民國一百零二年六月

國語語音屬性偵測器之製作及其應用

研 究 生：林彥邦　　指導教授：王逸如 博士

國立交通大學電信工程研究所碩士班

中文摘要

使用語音屬性偵測架構之語音辨認器在近十年來已重新受到學者之重視，並且有許多的相關之研究正在進行。然而，一般語音屬性偵測器之建立都需要使用監督式的方式訓練，但當訓練語料在缺乏一個已標示正確之答案時，其所製作的偵測器之效能結果多是不佳。本論文使用 HMM 辨認器之音素標示及英文語音屬性偵測器之結果來得到國語語料之自動音素端點校正，並以此製作一組可靠的國語語音屬性偵測器。

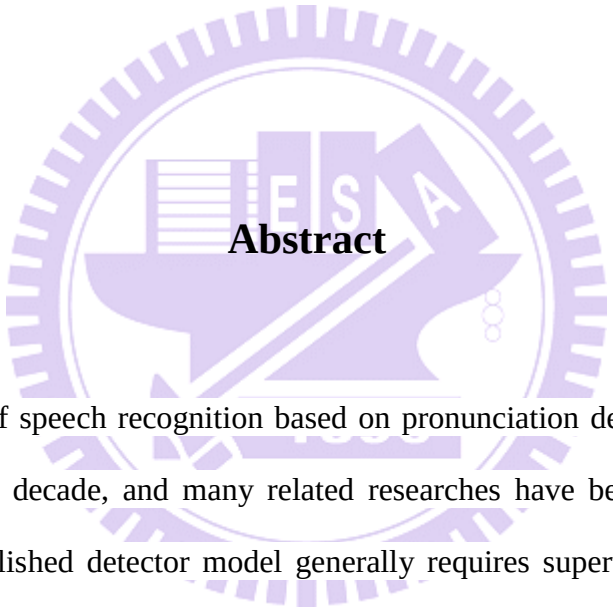
同時，論文中將所製作之國語語音屬性偵測端點偵測器運用於自發性語料。在自發性語料中有著許多音素省略或是同化之現象，以至於HMM語音辨認器之效能結果不佳。本論文中，使用訓練而得的國語語音屬性偵測器來觀察口語中常見詞彙之音素省略現象。

Implementation and applications of Mandarin Pronunciation manner detector

Student：Yen-bang Lin Advisor：Dr. Yih-Ru Wang

Institute of Communication Engineering

National Chiao Tung University



Abstract

The importance of speech recognition based on pronunciation detector has been highly recognized in the past decade, and many related researches have been conducted as well. However, a well-established detector model generally requires supervised training, and the effectiveness of detected results have mostly been poor for the lack of correct segmentations during corpus training. In this paper, we use the given results of the segmentations from HMM recognizer as well as the output from the English pronunciation manner detector established by TIMIT corpus to acquire a reliable Mandarin pronunciation manner detector. Meanwhile, this paper also utilizes this Mandarin pronunciation manner detector system in the tests of spontaneous speech. There are many linguistic phenomena, like phone reduction or assimilation, in spontaneous speech, making the HMM recognizer really difficult to achieve the reliable results. In this paper, we use the acquired manner detector system for to observe the common phone-reduction phenomena in spontaneous speech.

致謝

其他人只需兩年的時間就結束這段碩士的求學生涯，我卻整整花了四年才完成，但即便如此，我並不認為多了這兩年是白費的。人生總是需要挫折後，才會有所成長，在這短短的四年裡遇到的事情不少，不論是一改再改的研究題目或者是錯綜的情感事件等，都曾經讓我心力交瘁，而我也在這些事情中得到了很多人生的體悟以及心靈上之成長，就如毛蟲總要被困住後才能破繭擁有美麗的蛻變一般。

這四年來要感謝的人太多，無論是在課業上或者是在生活中。謝謝陳信宏老師，讓我還能在實驗室繼續努力；謝謝王逸如老師，即便是在實驗上遇到了挫折，你仍是當著我研究上明燈；俏秘書靜觀，實驗室有了你的管理後一切順遂了許多；一樣非常鄉民化的奕勳，希望我們反社會的血液繼續流動，而不是隨著社會洪流飄動；熱衷於水族的子睿，做事高效率，會玩會唸書就是說你；脾氣極好的良基，跟你相處真的非常輕鬆無壓力；蒙主感召的婉君，雖然跟你的交流不多，但你的單純和善良仍是映在我的眼簾；嘴巴沒停過的仲銘，你努力的表現一定會換來好的結果；鬼靈精怪的仲堯；棒球好手柏鈞；想當王仁甫的阿璋；籃球強爆了的阿峻；網路殺手茂隆；來無影去無蹤的佩樺；大胖等 98 級以及喬華等 99 級的成員，謝謝你們，讓實驗室貧乏的生活變得多采多姿。

另外，也感謝我大學的同學，不管發生了什麼事情，我們的友情依然維持不變。說曹操曹操就到的曹操，去當兵千萬別彎下腰；拒吃泡麵的小香，國際人力派遣顧問公司就靠你了；街頭藝人 MAN 哥，感謝你三年的收留讓我們天天都在表演輸。其餘尚有阿祖、薛帥、阿肚…等族繁不及備載。

最後，謝謝我的父母，雖然我多花了兩年才完成這次過程，但您們仍是在我背後無怨無悔地支持著我，對於不擅表達情感的我，那感動是隱隱的藏在心中，久久不能自已，我好想對您們說：「爸媽，我愛您們」。

謹以此論文獻給所有我愛的你們。

目錄

表目錄

圖目錄

第一章 緒論

1.1 研究動機

現今語言技術相關研究方面都和語料庫的資訊密不可分。在語音研究中，無論是在模型訓練或者是實驗結果之比對時，皆是以語料庫裡的文本內容作為答案以進行所有研究之程序。通常在國語語料庫之文本，朗讀式語音（Read speech）是已經經過詳細的規劃後產生，而在自發性語音（Spontaneous speech）中則是根據語者說話發音內容之正確字詞選定後再加以轉寫，然而實際上語者在發音時並不一定會完整地將字詞讀出，其在發音時很容易因為語速的變化或者是其餘不可測之因素導致語音產生了改變，如：字詞的轉調，甚至是有音素弱化或脫落之現象，語言學家將這些變化之現象稱為音變（pronunciation variation），因此在文本和實際發音彼此的比對上是很有可能會有不一致的情況發生，此種現象在自發性語音中發生的次數更是多於朗讀式語音之結果。

因為文本內容和實際語音比對上的誤差，這就會造成模型系統在訓練上會產生了不穩定的現象，進而導致整個研究之成效不佳。因此如果能夠解決文本和實際發音不一致的問題，這除了對於模型訓練上的穩定有極大的幫助之外，在此同時我們也可以簡單地觀察語音學家所討論的語言上之現象。

為了讓文本能夠符合語者實際之發音，研究中先行對幾個目標字詞進行實驗以了解其可行性，在本論文中提出了依語言屬性偵測結果為基礎的(detection-based)系統架構，並且經由動態時間校正，除了能夠將語言學家們所提到的語言現象以統計之結果清楚地表達之外，也希望藉此程序能夠將文本內容修正為符合實際發音的結果，以利於其餘語音相關之研究。

1.2 研究方向

在本篇論文中，將以建立一套優良的國語語言屬性偵測系統作為最終之目標。首先以目標音框為中心抽取參數，利用這些參數特性的不同便可將音框進行分類。

在本研究皆是使用類神經網路多層感知器結構（Multi-Layer Perceptron，MLP）來建立語言屬性偵測器之系統模型，其自我調適能力、非線性運算、自我學習的特性等特性有助於目標的分類。實驗中採用兩段式（two-stage）的偵測架構，在第一階段時採取左相關及右相關的偵測模型，其先行對目標音框及其左或右相關參數各自進行偵測並輸出結果；在第二階段時則是將第一階段時的左和右兩個偵測器其輸出合併當作新的參數，並建造一個合併偵測模型。最後，一個偵測器之系統架構即完成，如圖 1.1。

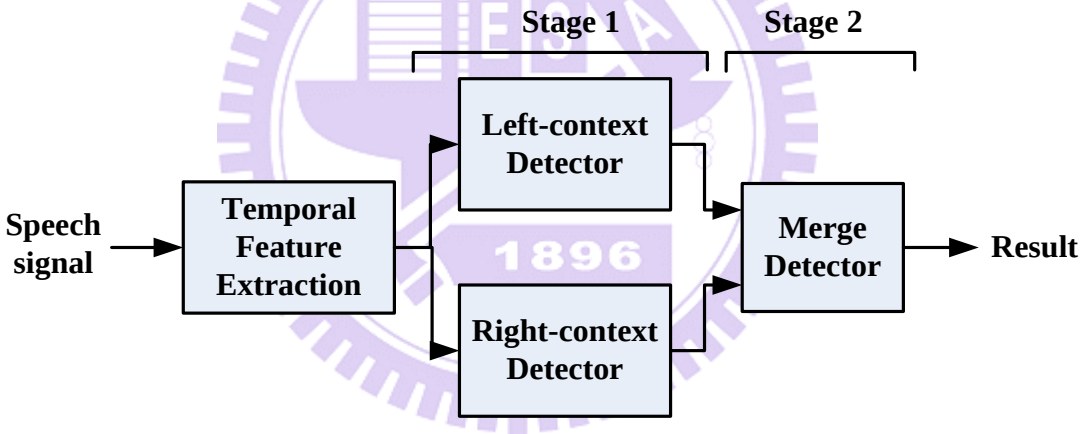


圖 1.1：偵測器系統架構圖

在語音經過偵測模型系統輸出後，研究中使用動態時間校正（Dynamic time warping，DTW）對於目標字詞進行調整，接著計算信心度量值以判斷目標字詞之正確性，最後根據統計結果觀察語音在自發性語言中的語音現象，並且將文本進行簡單的修正，以期許其能更符合實際語音。

1.3 相關研究

過去提出的語言屬性偵測系統中，通常都是運用於語音辨識的目的。由 C-H Lee[1] 提出的新時代自動語音辨識系統，其目的是希望能夠克服近年發展近於趨緩的隱藏式馬可夫模型（Hidden Markov Model，HMM）架構，運用前級的語言屬性及事件偵測器群組，經由抽取語音特徵參數來偵測不同時序之語音屬性及事件輸出概率值，並提供後級系統進行語言事件和語言學知識彙整的決策，以此達到最終語音辨識的目的。

而在 S. Siniscalchi 的研究中[2]，其對於語言屬性偵測器進行了跨語言之測試，實驗中使用了六國之語音共同訓練了一個通用語言屬性偵測器組，並運用於第七國語言進行了測試以證明語言屬性偵測器確實可以使用在跨於言上，且其實驗結果之差距並沒有劣化許多。然而本論文中要求一組最佳的語言屬性偵測器，因此在實驗中仍會使用目標語音重新訓練模型，藉此達到最佳化的效果。

1.4 章節概要說明

本論文內容總共分為五個章節，分別為：

第一章：緒論：介紹本論文之研究動機以及研究方向。

第二章：語料庫介紹：介紹在本論文中所使用之語料庫的統計分析。

第三章：語言屬性偵測器：介紹實驗中所抽取之參數以及說明建立國語語言屬性偵測系統的詳細步驟。

第四章：實驗結果：將語言屬性偵測器使用於不同的國語語料庫，統計其結果以觀察語音的現象並分析在不同語料庫中之差別。

第五章：結論與未來展望。

第二章 語料庫介紹

本論文將以不同語言的語料庫進行語言屬性偵測器上的訓練以及實驗，在此包括兩種語言：英文和中文，下列我們將會對於我們所使用的三種語料庫進行簡單的介紹。在 2.1 節中我們將介紹 TIMIT 語料庫的資料格式；在 2.2 節則會介紹國語的 TCC-300 語料庫其資料格式；在 2-3 節中則是介紹國語 MCDC 語料庫的格式。

2.1 英語 TIMIT 語料庫簡介

2.1.1 語音資料

在本論文中，TIMIT（The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus，TIMIT）語料庫為我們實驗中的前導性語料，因其中具有人工手動標記的時間資訊的特性以及高度的準確性可供我們使用。而 TIMIT 語料庫是屬於朗讀句子的語音（read speech），這些語句皆是由德州儀器（Texas Instruments，TI）、麻省理工學院（Massachusetts Institute of Technology，MIT）以及史丹佛研究機構（Stanford Research Institute，SRI）共同設計。而語料庫中的語句是經由德國儀器請美國不同區域的居民朗讀後並錄製成音檔，爾後由麻省理工學院對於每句語句進行人工轉寫而成。

TIMIT 語料庫中共有 6300 個語句，而這些語句來自 630 位不同的語者，此 630 位語者又可以分成美國主要的八種不同口音（Dialect）地區，每位語者各朗讀 10 個語句錄製而成，其在收錄時是以 16kHz 的取樣頻率經過 16 位元量化後錄製的單聲道音檔，音檔的檔頭為 1024 位元組（byte）。其餘的語料庫的詳細資料包括男女、地區分布就如下表 2.1。

另外，每位語者朗讀的 10 句語句中的類型如表 2.2，其中包括 2 句為了顯現不同區域的語者口音差異的方言（SA）語句；其中 5 句是為了讓每個音素出現頻率相近的

phonetically-compact（SX）語句；剩下的 3 句則是 phonetically-diverse（SI）語句，此語句乃從當時現存的文字語料庫挑選而出，如布朗文字語料庫（Brown Corpus，Kuchera and Francis，1967）等。

表 2.1：TIMIT 語料庫其方言之人數分布

方言的地區人數分布							
編號	區域	男性		女性		總計	
		人數	百分比	人數	百分比	人數	百分比
1	New England	31	63%	18	31%	49	8%
2	Northern	71	70%	31	30%	102	16%
3	North Midland	79	77%	23	23%	102	16%
4	South Midland	69	69%	31	31%	100	16%
5	Southern	62	63%	36	37%	98	16%
6	New York City	30	65%	16	35%	46	7%
7	Western	74	74%	26	26%	100	16%
8	Army Brat (moved around)	22	67%	11	33%	33	5%
	總計	438	70%	192	30%	630	100%

表 2.2：TIMIT 語料庫不同語句類型之分布

語句類型	語句數	每一語句的語者數目	總計	每位語者之語句數
Dialect(SA)	2	630	1260	2
Compact(SX)	450	7	3150	5
Diverse(SI)	1890	1	1890	3
總計	2342		6300	10

2.1.2 文字轉寫之人為時間標記

TIMIT 語料廣泛的用於語音上面的研究，皆因此語料庫完整地囊括不同對應層級的人為時間轉寫標記，其無論是文字層級（word level）或者是音素層級（phone level）皆有清楚的人為時間轉寫標記，這使得 TIMIT 語料庫成為一個廣泛使用的語言相關研究

的平台，並由此語料庫驗證其理論、方法以及研究結果的成效優劣。

而無論何種層級的文字轉寫，皆是由標音員根據語音中的語音信號給予正確的標音符號，並且依其起始和結束的位置的取樣點給予時間標記。雖然此標記受標音員的主觀影響結果，在信號起始和結束的特性上會有偏差不一致的問題存在，話雖如此，但其目前對語言準確分段來說仍是最佳的方式。也因此，在本論文中，我們是以 TIMIT 語料庫當作我們的前導性語料。

2.2 國語 TCC-300 語料庫簡介

本語料庫乃是由國立交通大學、國立台灣大學以及國立成功大學錄製集結而成，並且由中華民國計算語言學學會發行，屬於麥克風朗讀語音，其主要目的是提供語言辨認研究。台灣大學所錄製的音檔主要是包含詞以及短句，其文字經過設計，並有考慮音節與其相連出現之機率，語者共 100 人，每人錄製一句；成功大學以及交通大學所錄製的音檔則為長文語料，其語句內容是由中研院所提供的 500 萬詞詞類標示語料庫中選取，每篇文章內含數百個字，再將其分割成三至四段音檔，每段至多 231 個字，兩校的語者各 100 人，每人各朗讀一篇長文，文章內容不重複。此語料庫的錄音環境為 16kHz 取樣頻率，取樣位元為 16 位元，檔頭資訊為 4096 位元組（byte），其副檔名為*.vat。其檔案詳細資料如下表 2.3。

表 2.3：TCC-300 語料庫之檔案資料

學校	語音檔案(*.vat)	文字檔案(*.tab)	群集(group)
台灣大學	6509	6509	1
交通大學	1238	1238	5
成功大學	1170	1170	5

在屬於聲調語言之國語音節中，其結構可以粗略分成聲母以及韻母，而韻母又可以再細分為介音和韻腳，韻腳中包含主要元音以及韻尾，而本論文所使用的 TCC-300 國語

語料庫乃是使用類音素為單元做為語言屬性偵測器的基本語言單元，也因此國語中的複韻韻母必須拆解成單韻母，所以國語音節的結構就分成聲母、韻母（不包含鼻音韻尾）、鼻音韻尾等三個部分以依照語言之特性簡化結構。

而在實驗語料的選取中，我們從 TCC-300 語料庫中選取了共 6858 句，其中隨機挑選 6000 句為訓練語料，剩餘的 858 句為測試語料。在本論文中是根據每個音素的語言屬性去訓練一個語言屬性偵測器，爾後再比對答案證明模型的穩定以及評斷實驗的結果，因此我們需要音素的時間位置。由於 TCC-300 國語語料庫並沒有人工標記音素分段模型，所以我們使用 HTK（Hidden Markov toolkit）訓練隱藏式馬可夫模型（Hidden Markov Model，HMM），接著由此模型對音檔進行強迫對齊（Forced alignment）後取得自動分段結果，並作為 TCC-300 語料庫的音素初始分段位置。

2.3MCDC 語料庫簡介

2.3.1 語音資料

MCDC 語料庫為兩語者對話之語料，其採用雙聲道且取樣頻率為 48kHz 的方式錄製，兩語者發音之語料則分別錄於左右聲道，再利用 Cool Edit Pro 軟體將其分割成較小的雙聲道音檔，並依長度約三分鐘處找到一處明顯清楚的停頓切開，其語者對話介紹如表 2.4 所示。在本論文中將左右聲道語料分開，並將其轉換成兩個單聲道的音檔，且同時將取樣頻率下降至 16kHz，接著再利用每一段落相對應之開始以及結束位置做切割，經由上述處理後共產生了 7085 段音檔，扣除未包含語言現象的音檔後剩下 6570 段音檔。

表 2.4：MCDC 語料庫之檔案資料及對話主題簡介

對話序號	長度(分)		聲道	語者編號	對話主題
mcdc-01	61	MISC-08-male-25	R	01R	工作、休閒活動、經濟、開車
		MISC-07-female-29	L	01L	
mcdc-02	63	MISC-10-male-35	R	02R	休閒活動、經濟、工作、性別、政治
		MISC-09-female-37	L	02L	
mcdc-03	61	MISC-12-female-17	R	03R	家庭、學校、購物、生涯規劃、明星
		MISC-11-female-16	L	03L	
mcdc-05	63	MISC-15-male-40	R	05R	工作、家庭、社會階級、保險、歷史、省籍情結
		MISC-16-female-46	L	05L	
mcdc-09	66	MISC-23-female-30	R	09R	工作、旅行、生活態度、環保、健康
		MISC-24-female-35	L	09L	
mcdc-10	54	MISC-26-male-23	R	10R	電影、政治、軍隊、捷運、學校、經濟
		MISC-25-male-35	L	10L	
mcdc-25	55	MISC-57-male-43	R	25R	交通、工作、小孩、旅行、電腦、管理
		MISC-58-female-45	L	25L	
mcdc-26	46	MISC-60-male-24	R	26R	工作、求職、家庭、車禍、學英文、婚姻、軍隊
		MISC-59-female-37	L	26L	

2.3.2 自發性語音之特性

自發性語音和朗讀式語音之間最大的差異在於朗讀式語音是事先經過設計的文本，但人類實際在說話時常常會伴隨著大腦思考以及情緒的變化而產生一些無法預期的聲音或發生在語言學中較詞層級（word level）更為上層的行為，此些現象就會造成訓練和研究上的困難。以下我們介紹幾種 MCDC 語料庫中常見的特性：

- **感嘆詞（Partical）**

不具標準語意的感嘆詞，其在對話中占有極大的成分，如回應或者是同意。語流中常出現的感嘆詞可分成下列四類：

一、有相對應的國字感嘆詞，如：A、BA、LA、MA、O。

二、無相對應的國語感嘆詞，如：AI YE、EI、HEN、NE、NEI。

三、源於台語的感嘆詞，如：EIN、HEIN、HO。

四、其他感嘆詞，如：UHN、NHN、MHM、MHMHM。

◆ **無法或難以辨認語音**

此部分主要分為無法辨識的語音（unrecognizable speech sound）以及不確定字詞（uncertain）。無法辨識的語音乃為標音員確定此聲音為人類發出的語音，但卻無法辨認其為何字何義；而不確定字詞則可分成兩種，一種為標音員大致上可以猜出語音內容，但卻並非百分百確定，另一種則是可以清楚記錄其發音，但卻無法根據語義猜出相對字詞。

◆ **非語言聲音（Non-speech sounds）**

在人類自發性語音中，常常會有一些非語言現象出現，此部分可以分做副語言現象（para-linguistic）以及非語言現象（non-linguistic）。一般非語音但確定是由人發出的聲音即稱為副語言現象，如：笑聲、咳嗽聲、呼吸聲、吞口水聲…等；而非語言現象則為確定不為人類發出的非語音，如：背景聲、麥克風的敲擊聲。

◆ **語流中斷**

此語料庫中的語流中斷現象主要有沉默（silence）、停頓（pause）以及短停頓（short break），這些現象乃語者在語流中因話題銜接不上或自身所產生的沉默造成。

◆ **非流暢現象**

不流暢的語音為自發性語音裡的一個重要特性，其主要有重覆（repetition）、詞語更正（repair）、部分重覆（restart）以及更正插語（editing term）…等。在此，重覆指的是完整地重覆詞語一次以上；詞語更正則是語者覺得說出的話語不適當，當下立即做出更改；部分重覆則是語者重覆句子的起頭詞片段，與重覆完整的詞句不同；更正插語則是出現在被更正詞語（reparandum）以及更正詞語（correction）之間，或者是出現在完整重覆或部分重覆中兩個重覆語詞中間。下面為非流暢現象的範例

基本型態　　(被更正詞語)*[更正插語](更正詞語)

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

如同之前所述，在本論文中必須要有音素的時間位置當作正確答案，如此才能對於區段內的語言屬性進行比對的步驟，因此，我們對於 MCDC 語料庫一樣使用 HTK 訓練 HMM 模型。在此，因為自發性語言很容易會產生音素遺失的現象，因此我們所訓練的模型只分成聲母模型以及韻母模型，接著進行強迫對齊對音檔自動分段，最後將此結果當作聲母韻母的分段結果以進行後續的研究。

第三章 語言屬性偵測器

本論文乃使用音框式聲學參數的語言屬性偵測系統，由參數在不同屬性中之變化特性做為依據，並在最初以英文 TIMIT 語料庫其音素層級之文字轉寫的人工標記時間做為模型之訓練目標，爾後再將此模型跨語言運用於中文上，並由國語 TCC300 語料庫反覆訓練語言屬性偵測模型，以達到使用於國語語言屬性偵測上的目的。3.1 節將介紹英文語言屬性偵測器的架構以及概念；3.2 節則介紹在此使用的音框式聲學參數之萃取；3.3 節介紹的是語言屬性偵測模型的訓練；3.4 節則是介紹語言屬性偵測器在跨語言後之測試以及由目標語言重新調整訓練之架構。

3.1 語言屬性偵測器

3.1.1 語言屬性偵測器之架構

儘管是在不同的語言中，人類發音系統的構造對語音特性的影響在一段語句中仍舊能夠清楚的展現，而不論是發音方法或者是發音位置都和語音中的音素有著相當大的關聯性，因此偵測語言中不同的屬性就等於可以猜測語者所發音之音素。在此處，本論文關注於發音的方法，藉由偵測音框式聲學參數以及其左相關（left-context）和右相關（right-context）的數個音框內的聲學參數特性的變化，由此判斷目標音框的發音方式以達到語言屬性偵測的過程。

語言屬性偵測器以音素層級文字轉寫之標記時間來訂定目標函數狀態種類以做為訓練模型時的依據，在語言屬性偵測的系統中，我們將其分為屬於此語言屬性（Y）、非此語言屬性（N）以及其他（other）共三種分類。在所有目標音框及其左相關或右相關的數個音框內的聲學特性參數以及目標函數狀態標記對應之後，經由多層感知器的學習特性，反覆疊代訓練將音框的語音特性做發音方法之語言屬性分類，之後再將左相關及

右相關的語言屬性偵測器輸出合併成為新的輸入，並再次經過多層感知器以及反覆疊代訓練得到一個語言屬性偵測之系統模型，藉此模型達到語言屬性偵測的目的。

在此，語言屬性偵測器是以英語 TIMIT 語料庫中的音素層級文字轉寫之人工標記時間經過目標函數狀態轉換後成為模型訓練之目標，而本論文中所使用的語言屬性偵測器乃由發音方式分類，共分成：母音（vowel）、鼻音（nasal）、摩擦音（fricative）、流音（approximant）、爆破音（stop）以及靜音（silence）六種，而此六種語言屬性偵測器乃是平行的偵測模型。而此處訓練所得的模型，會使用於跨語言的語料庫如國語的 TCC-300 中進行測試，並由 TCC300 語料庫再對模型繼續反覆的訓練和調整，圖 3.1 為語言屬性偵測器的架構流程圖。

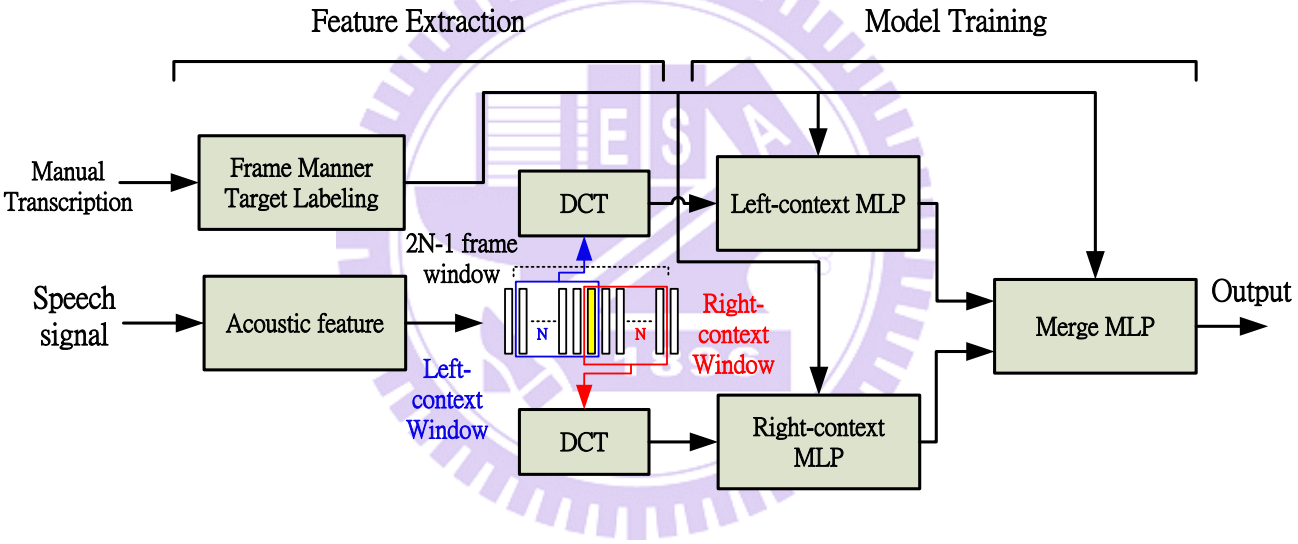


圖 3.1：使用多層感知器之語言屬性偵測器

3.1.2 語言屬性之分類

在人類發音的過程中，會藉由各個發聲器官如：口腔、舌頭、牙齒...等的配合以及調節，再經由肺部發送氣流通過發聲器官所形成之通道，如此達到目標之發聲。以下所有發音器官透視圖皆來自[11]，圖 3.2 為人類之發聲器官透視示意圖。

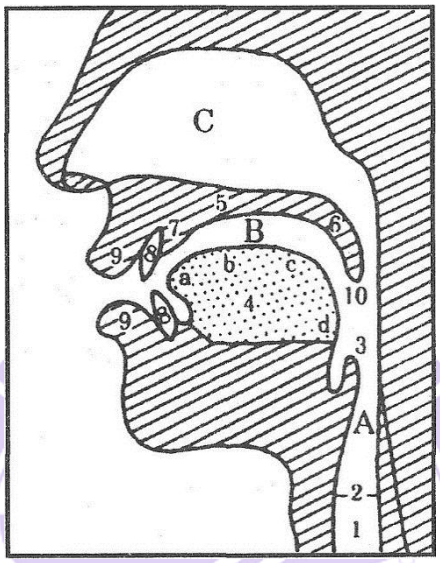


圖 3.2：人類發聲器官透視示意圖

在人類語音中，有些音素間的特徵相距甚多，而有些音素間的特徵則是相距不遠，如何將這些音素做分類是實作語言屬性偵測器的第一步，因此實驗前必須先制定一個音素分類標準。根據 S. Siniscalchi 的分類中[3]，其對於音素共分成了發音位置以及發音方法，在此，我們只制定了發音方式中之六個種類，介紹如下：

■ **母音**

母音又稱做元音，而因為發母音時聲帶必會震動，所以母音皆為濁音，其在發聲的過程是由氣流通過口腔而不受阻礙所發出的聲音。母音在發聲時，氣流從肺部通過聲門直接衝擊聲帶，致使聲帶產生均勻地震動，而震動的氣流並不受口腔或鼻腔的阻礙，而後經過舌頭和嘴唇的調節產生不同的聲音。圖 3.3 為母音發聲的例子，此處為母音發聲時通道不受阻之發音器官透視圖。

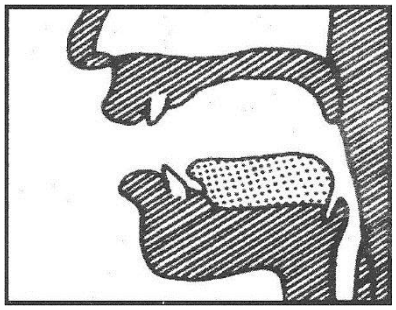


圖 3.3：母音 a 發聲之發音器官透視圖

■ **鼻音**

鼻音是一種輔音，其在發音時，軟顎下垂，氣流在口腔的通路被阻塞而轉往鼻腔通過。圖 3.4 為鼻音發聲的例子，此處為鼻音 m 發聲時的透視圖，發聲時雙唇緊閉，軟顎下垂，以致氣流由鼻腔通過。

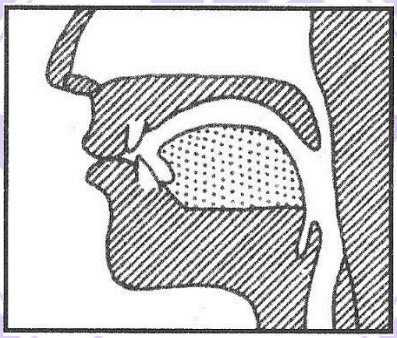


圖 3.4：鼻音範例，m 的發聲之發音器官透視圖

而鼻音在頻譜上的能量是集中於低頻部分，如圖 3.5 所示。

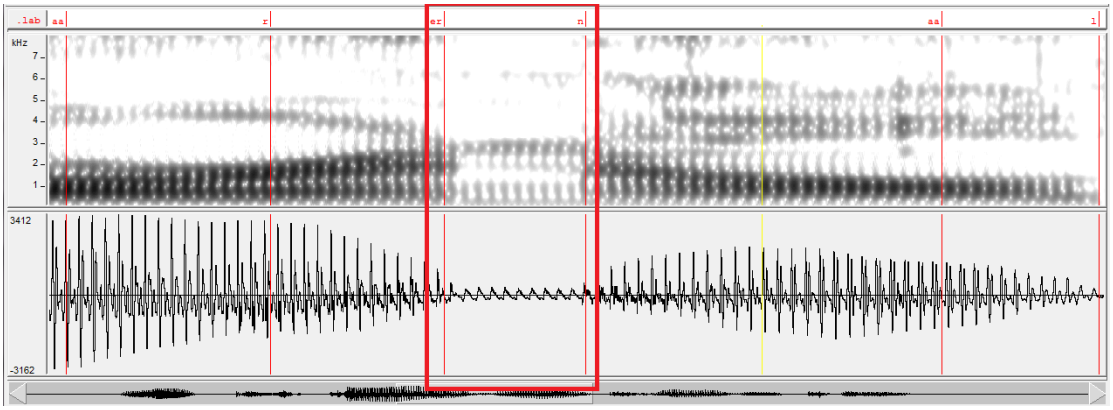


圖 3.5：鼻音之能量頻譜分佈示意圖。由上而下分別是類音素層級文字轉寫時間標記、聲譜圖、原始語音信號

■ **摩擦音**

摩擦音也是輔音的一種，其在發音時，由兩個發聲器官靠攏形成一個狹窄的通道，氣流經過通道時就會產生摩擦而發出聲音。圖 3.6 為摩擦音發聲之範例，此處為摩擦音 f 或 v 之透視圖，發聲時下唇輕碰上齒形成一個狹窄之通道，氣流由此通過造成湍流摩擦因而發出聲音。

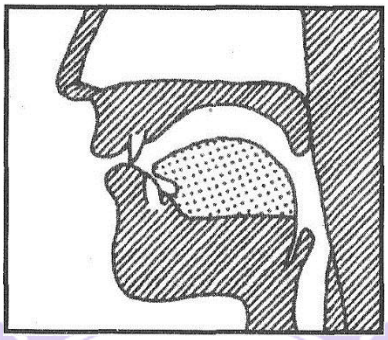


圖 3.6：摩擦音範例，f 或 v 的發聲之發音器官透視圖

摩擦音的發聲可以持續一段時間，且其在頻譜上的能量是集中於高頻的部分，圖 3.7 充分地展示了此項特性。

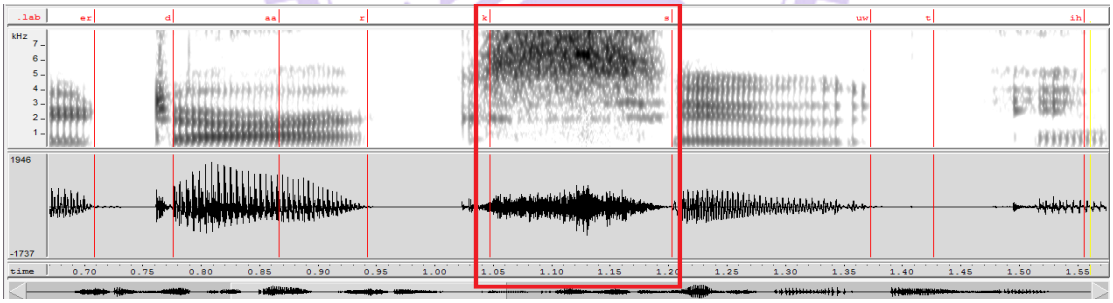


圖 3.7：摩擦音之能量頻譜分佈示意圖。由上而下分別是類音素層級文字轉寫時間標記、聲譜圖、原始語音信號

■ **流音**

流音是介於母音和輔音的聲音，語言學稱之為近音，其在發聲時，由兩個發聲器官靠攏形成一個較窄的通道，但此通道因不夠狹窄所以氣流可以較無礙的流動，產生的湍流就較弱，因此形成了介於母音和摩擦音間的聲音。圖 3.8 為流音之範例，此處為流音 l 的透視圖，發聲時舌尖抵住上齒齦，氣流從舌頭兩側通過以發出聲響。

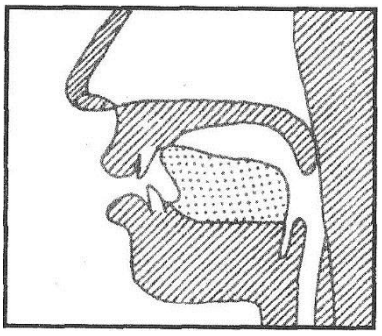


圖 3.8：流音範例，l 的發聲之發音器官透視圖

■ **爆破音**

爆破音是輔音的一種，其在發聲時是先將氣流通路閉塞，然後突然打開讓氣流通過，藉此產生聲音，因此爆破音前面會有一個短停頓產生，在語言學上稱為嗓音起始時間（voice onset time，VOT）。圖 3.9 為爆破音發聲的例子，此處為爆破音 k 或 g 的發聲之透視圖，發聲時舌後向上頂住軟顎形成阻隔，而後快速打開讓氣流通過口腔，以產生聲音。

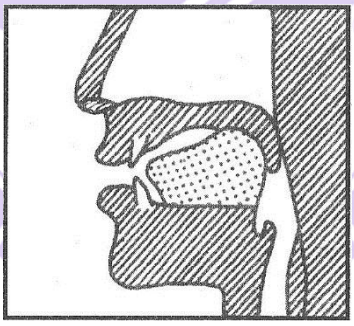


圖 3.9：爆破音範例，k 或 g 的發聲之發音器官透視圖

爆破音相較於其他的發音方式來說，因為其發聲過程是靠突然的氣流通過而產生的噪音，所以發聲的時間長度較其他的發音方式短。但在英語 TIMIT 語料庫中，我們使用的人為時間轉寫標記之層級並未將嗓音起始時間另外標記，因此爆破音的平均時間長度較短之現象無法從統計中求得，但從頻譜中可以發現爆破音前確實是有短暫的停頓產生，而有聲音的部分就非常的短，如圖 3.10 中所示。

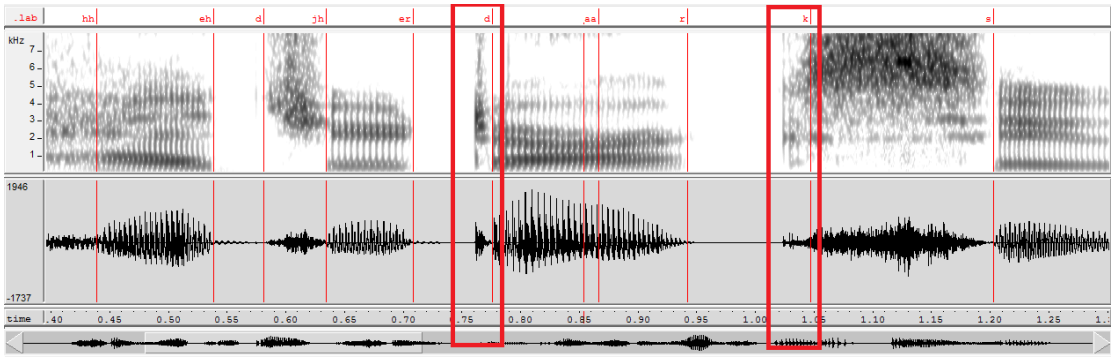


圖 3.10：爆破音之能量頻譜分佈示意圖。由上而下分別是類音素層級文字轉寫時間標記、聲譜圖、原始語音信號

■ **靜音**

沒有聲音的停頓情況，或者是爆破音前面的短停頓，皆屬於靜音。

由上面介紹的發音方法之語音屬性分類方式，我們根據[3]對英文 TIMIT 語料庫的 39 個音素分類到六個語言屬性中，分類結果如下表：

表 3.1：英文 TIMIT 語料庫之 39 個音素其發音方法語言屬性分類

Manner	Phone
Vowel	aa ae ah aw ay eh ey ih iy ow oy uh uw
Approximant	l r er w y
Fricative	ch dh f hh jh s sh th v z
Nasal	m n ng
Stop	b d dx g k p t
Silence	pau

根據上面發音方法之語言屬性分類，每個發音方法都有各自的語言屬性偵測模型，而語音信號平行的輸入此六個模型進行語言屬性之偵測以獲得結果。而若將六個模型輸出的結果合併成為一個新的輸入參數，並將其對應至目標函數狀態後訓練，即可得到一組發音方法之語言屬性辨認器，其架構如圖 3.11。

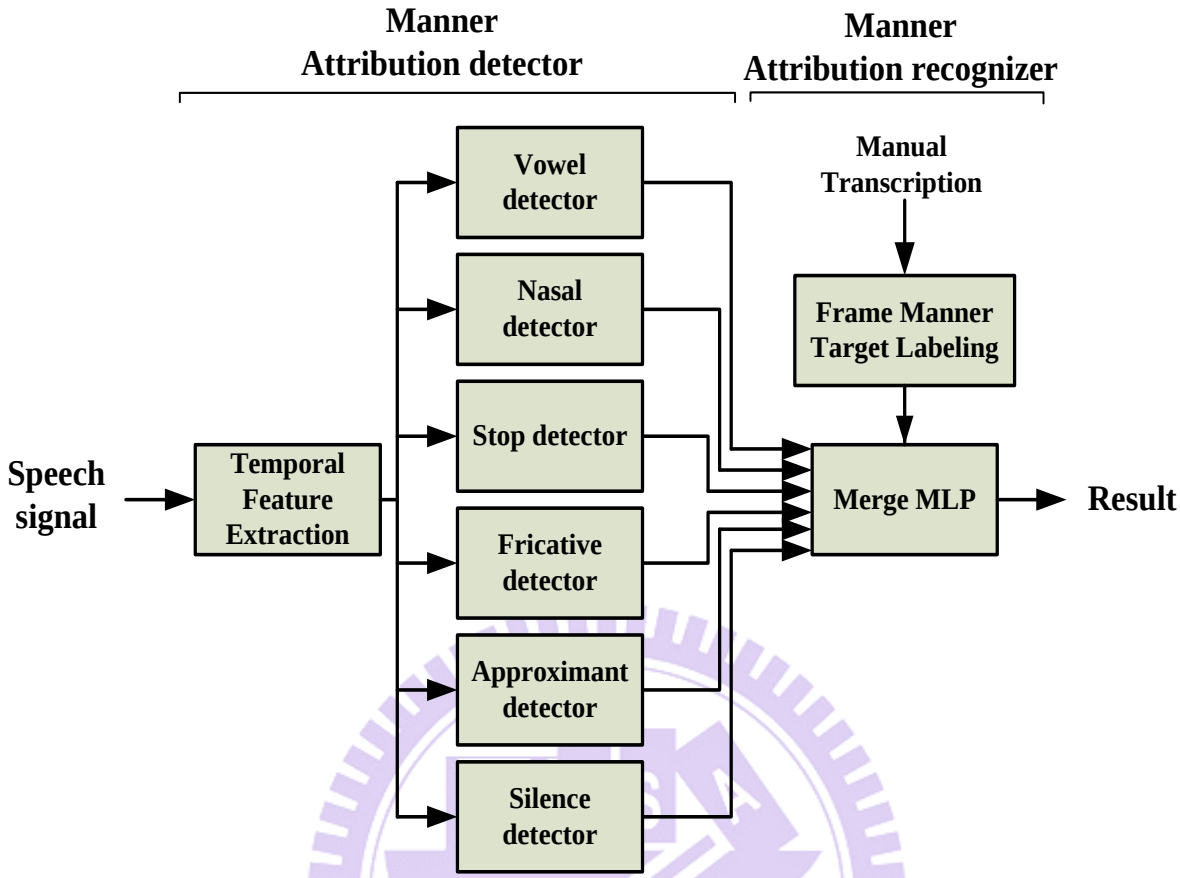


圖 3.11：語言屬性辨認器之架構圖

3.2 參數萃取

聲學參數抽取的目的是為了能夠有數值資訊可以呈現一段語音信號中的特性和現象，在語音信號中不同的音素就會有不同的特徵，而相似的音素其特徵也不會相差太多。然而，不同的語者以及朗讀語句中不同的內容皆會讓數值上起了一定的變化、差異甚至是音位變體（Allophone），但是大致上來說，其語言屬性是不會有改變的，因此可以藉由抽取語音信號的聲學參數，藉此將語音信號之聲學資訊萃取出來，並對此資訊進行分析和分類（Classification）以提供下一步研究或程序處理。

人耳對於各個頻域上的聲音之感知，並非在所有的頻域上皆有相同的敏感度。事實上，人耳在對於低頻時的聲音解析度比較高，亦即較能分辨低頻時的頻率差異性；反之，在高頻的時候人類聽覺對聲音之解析度就降低了。所以人耳對於頻率的對應投射並非全是線性的，其在頻域 1 千赫茲（kHz）以下維持線性的，但在 1 千赫茲（kHz）以上則是對數投射。為了模仿人耳內的基底膜（Basement membrane）其對語音信號聲音的臨界頻帶（Critical band）之刺激反應，在此我們使用的是一個經過聽覺感知對應後的聲學參數，信號經過梅爾刻度頻率曲線之三角帶通濾波器（Mel-scale filter bank）組進行對數能量之求取後，再進行離散餘弦轉換。圖 3.12 為梅爾刻度頻率曲線之三角帶通濾波器組。

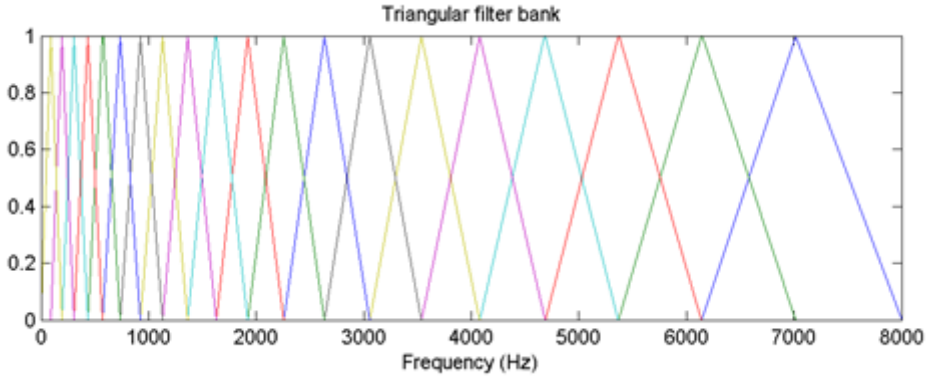


圖 3.12：梅爾刻度頻率曲線之三角帶通濾波器組

在本研究中，根據 P. SCHWARZ 所述[4]，對於數個參數向量窗框化並將其經過離散餘弦轉換，再把此輸出值輸入類神經網路後的結果會優於未進行窗框化和離散餘弦轉換的結果。其原因乃是窗框化可以將目標音框的影響放大，而越邊緣的音框其影響則會越小，因此焦點就可以放於目標音框，周圍的音框則可以視為輔助辨識結果；離散餘弦轉換則是可以將參數向量之維度降低，且可以將快速變化的部份剔除進而致使參數較平滑，這意味著類神經網路在訓練的過程中會有較少的機會只收斂於局部最佳（Local optimal），而經過離弦轉換後的各參數間是獨立不相關的關係。

在此，我們根據[4]之架構設定研究之模型。首先，對目標音框抽取之聲學參數是所有三角帶通濾波器的輸出對數能量，此處設定之濾波器組共有 23 個三角帶通濾波器，亦即每個音框皆有 23 維的對數能量輸出。此外，在本論文中的語言屬性偵測器是使用目標音框以及其左相關和右相關的聲學資訊當作偵測器的輸入以進行偵測，於是我們以目標音框當作中心，並往左和右各加了數個音框合組成一個新的窗框，實驗中由目標窗框往左和右各延伸了 15 個音框，因此左窗框和右窗框在包含著目標音框的情況下皆含了 16 個音框，最後形成了一個 31 個音框之新窗框。而在偵測器的架構上，我們分成了左相關語言屬性多層感知器以及右相關語言屬性多層感知器，所以在輸入的聲學參數方面，其左窗框和右窗框內的輸入聲學參數是不相同的。在左相關語言屬性多層感知器中，由左窗框之 16 個音框內之各個濾波器的對數能量聲學參數皆各自經過 11 階的離散餘弦轉換得到輸出係數，並當作多層感知器的輸入；同理，右相關語言多層感知器的輸入則由右窗框內之 16 個音框中的每個頻帶之濾波器其對數能量經過 11 階的離散餘弦轉換而得，所以左和右的語言屬性多層感知器各是 253 維之輸入。其離散時間餘弦轉換公式如下：

$$F(m,l)=\sum_{r=1}^{16}E_r(m)\cos\left[\frac{\pi}{16}\left(r+\frac{1}{2}l\right)\right]\text{，}l=0,\cdots,10\text{ ，}m=1,\cdots,23 \tag{3-1}$$

其中，$E_r(m)$ 為第 r 個音框中的第 m 個頻帶濾波器經過窗框化後之對數能量，l 為窗框中的每一個頻帶濾波器所求取的係數階數。

而最後的合併左右相關之多層感知器輸入，則是將這左和右相關語言屬性感知器的輸出合併，並再訓練一個新的感知器，此三個多層感知器就組成了一個語言屬性偵測器。圖 3.13 清楚地展現了左右相關之聲學參數萃取方式的整體架構。

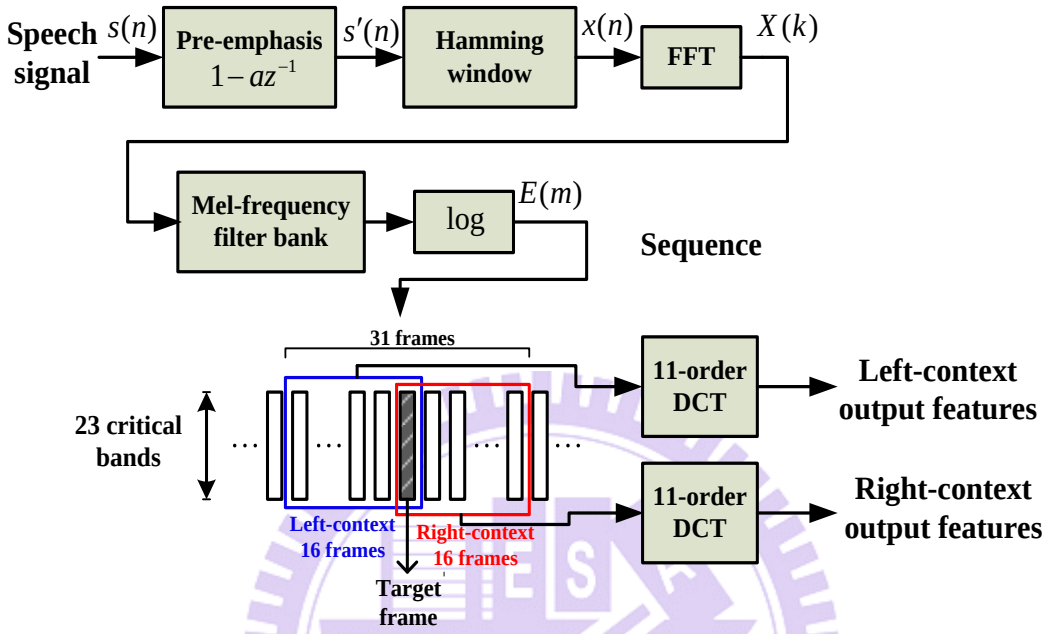


圖 3.13：本論文所使用之左和右相關聲學參數之萃取流程圖

3.3 模型訓練

在聲學參數的萃取完成後，我們使用 Quicknet 多層感知器架構建構系統，其使用倒傳遞演算法（Back-propagation algorithm）對於聲學參數向量做目標的分類，並以達到系統最小平方差（Minimum least square）為目標。下面將會依序介紹模型的訓練步驟：3.3.1 節將會介紹類神經網路的基本單位以及架構；3.3.2 節則是介紹模型訓練之步驟。

3.3.1 多層感知器之類神經網路架構

3.3.1.1 神經元

神經元乃類神經網路中最基本的單位，其主要是負責處理和計算資料之間的關係。而每一個神經元所接受到的輸入參數資料量從一個至數個不等，端看不同的網路連結線路而決定輸入參數資料的數目。另外，神經元對每一個輸入參數 x_j 經過處理後，便會依照參數的重要性而給予一個加權權重（weight，w）做相乘，接著，神經元將所有經過加權權重相乘後的輸入參數做累加，並加上神經元中的偏移量（bias，b）後產生一個淨值（net），最後，將此淨值輸入激發函數（activation function），藉此函式將其轉換而成模擬人類神經元的輸出訊號。下式即為神經元內的計算數值：

$$net = \sum_j w_j x_j + b \tag{3-2}$$

$$a = f(net) = f\left(\sum_j w_j x_j + b\right) \tag{3-3}$$

其中 a 表示為輸入信號經由激發函數所輸出的神經元訊號。下圖 3.14 為神經元輸入輸出的關係圖：

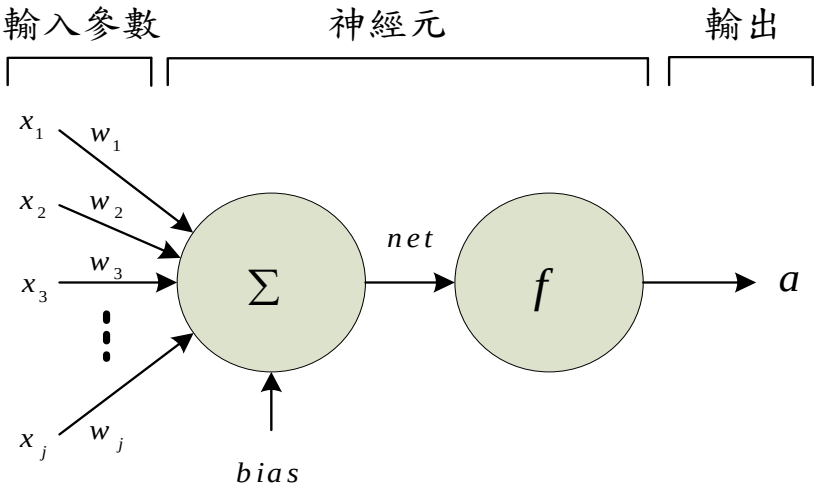


圖 3.14：神經元與輸入參數和輸出數值的關係圖

神經元中的激發函數可以依據面臨問題的不同或者是想要達成的目的而挑選其所需要的特性，其主要上可以分成兩種：線性函數以及非線性函數，而為了使倒傳遞演算法進行有效的學習和訓練，激發函數就必須限定為可微函數。事實上，激發函數的類型其實有很多種，在此，我們選擇使用最基本的邏輯乙形函數（Logic sigmoid function）做為實驗用的激發函數，其數學表示式子如下：

$$f(x) = \frac{1}{1+e^{-x}} \tag{3-4}$$

因為函數本身的限制，所以不管輸入的數值大小為何，其輸出的數值範圍只會介於 0 至 +1 之間，此特性由下圖 3.15 可以清楚地表現。

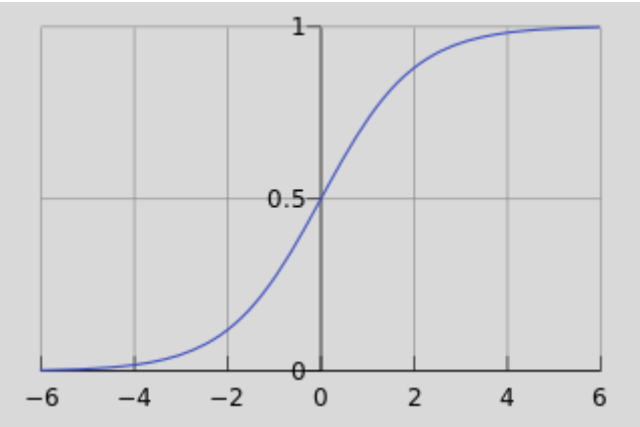


圖 3.15：函數之輸入輸出關係曲線圖

3.3.1.2 網路架構

本論文使用的是前饋式類神經網路，其網路結構如圖 3.18 所示，其由數個平行輸入參數之神經元並聯組成一個結構層（layer），再經由數個結構層間的神經元彼此相互鏈結串聯而形成一個網路。在類神經網路中，結構層其特性和功能作用皆有所不同，由這些性質上的差異，其大致上可以將結構層分成三類：輸入層、隱藏層以及輸出層。

1. 輸入層：此層的每一個神經元都只有一個輸入參數以當作其輸入值，並提供下一層的每個神經元之輸入以進行訓練，因此此層神經元的作用只有類似暫存器而已，並無任何運算之功能。另外，在此層的神經元數量取決於輸入參數向量的個數，本論文的語言屬性偵測實驗中，在左相關和右相關發音方法語言屬性偵測器中使用的皆是 253 維之參數向量；而合併左右相關之多層感知器其輸入則是由左相關和右相關各輸出之 3 維參數向量組合，因此，在語言屬性偵測器方面的輸入層之神經元數目就只會有 253 個或 6 個。而在發音方法語言屬性辨識器中，其輸入的參數向量則是由每個語言屬性偵測器之 3 維輸出合併，因此在此的輸入神經元數目則是 18 個。

2. 隱藏層：隱藏層是介於輸入層和輸出層之間的結構層，其階層數和神經元數目並沒有一個硬性的規定，而是必須視資料的複雜度做測試後再做選擇。通常隱藏層之階層數越多，其分類結果也會較佳，但相對來說模型的運作和訓練也就較於費時，且其結果並不一定有著明顯之差距。在本實驗裡，隱藏層的階層數只設為一層，而使用的神經元的數目為 500 個。

3. 輸出層：輸出層的神經元個數取決於偵測任務（task）的目標種類，且前後兩者的數目必須且絕對要相同。而輸出層神經元最後輸出的結果，即為類神經網路經由訓練後所輸出的目標函數值。

下圖 3.16 即為前饋式類神經網路的表示圖。

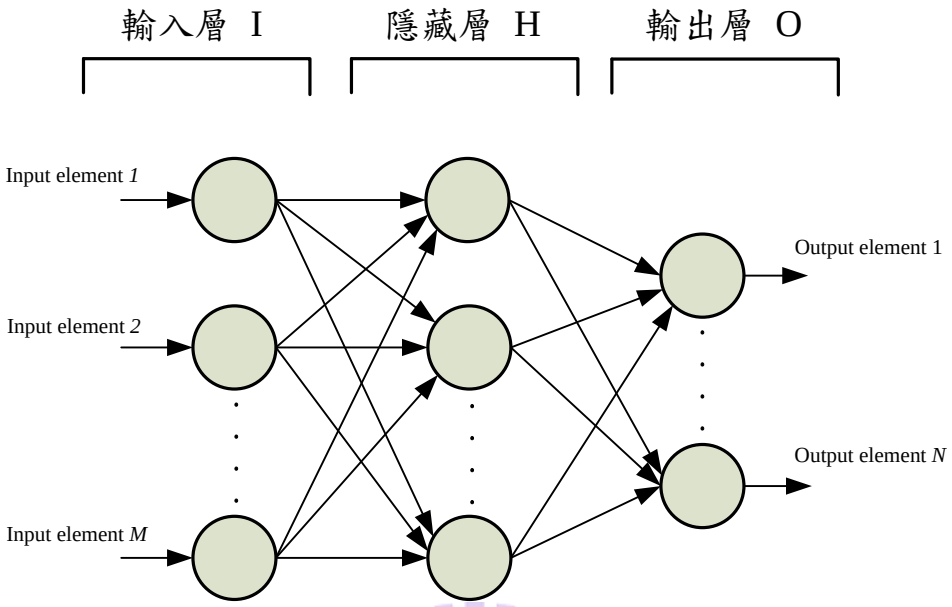


圖 3.16：多層前饋式類神經網路架構範例圖

3.3.2 網路學習方法

類神經網路在最後的輸出層輸出後，我們另外使用了一個 softmax 激發函數對輸出值做一個簡易的轉換，因為 softmax 函數可以保證輸出之總合為 1，這會讓每個輸出節點的值類似事後機率一樣，softmax 激發函數如下式：

$$f(k) = \frac{e^{-o_k}}{\sum_{i=1}^{N_3} e^{-o_i}} \tag{3-5}$$

此外，根據 R. Dunne 的研究[10]，經由 softmax 激發函數後，使用最小平方差之訓練結果優劣近似於使用最小交叉熵值（Minimum cross-entropy），且此系統優於未經過 softmax 激發函數的結果，在此我們使用最小平方差的方法。下面為網路訓練的步驟：

- Step1：參數向量經過類神經網路後輸出。
- Step2：計算輸出和目標函數的誤差，在此使用最小平方差。
- Step3：更新網路類神經網路之權重等參數。

重複以上 Step1 至 Step3，直至模型收斂。

3.4 國語語言屬性偵測器

在英文 TIMIT 語料庫中，因為其為人工時間轉寫標記，雖然標記時會因標音員主觀的意識而產生誤差和不一致的現象，但其仍為目前最佳的標音結果，而從最佳標音結果開始訓練一個語言偵測模型對於研究上來講絕對是較優的，表 3.2 為 TIMIT 語料庫所訓練之語言屬性偵測辨識系統之測試統計結果。

表 3.2：由英文 TIMIT 語料庫訓練之語言屬性偵測辨識系統模型其運用於英文 TIMIT 語料庫之測試結果

Corpus	**Training**	TIMIT
	Testing	TIMIT
Manner detection	**Vowel**	90.43%
	Approximant	91.94%
	Nasal	96.44%
	Fricative	94.65%
	Stop	94.45%
	Silence	96.89%
Manner recognition	85.06%	

但在國語語料庫中，並沒有如 TIMIT 語料庫一般的人為時間轉寫，而因為人為時間轉寫所需耗費的時間和金錢甚鉅，因此使用隱藏式馬可夫模型對文本強迫對齊以取而代之是較為可行之方法。在此，我們使用的是國語 TCC-300 語料庫，因為其屬於朗讀式語音，相較於自發性語音來講其信號較穩定，且少了副語言現象，對於分析或訓練模型來講會更有利。

在有了中文強迫對齊後的時間結果後，我們使用由 TIMIT 語料庫訓練而得的語言屬性偵測系統對其進行測試，表 3.3 為測試結果之統計。

表 3.3：由英文 TIMIT 語料庫訓練之語言屬性偵測辨識系統模型其運用於國語 TCC-300 語料庫之測試結果

Corpus	**Training**	TIMIT
	Testing	TCC-300
Manner detection	**Vowel**	81.87%
	Approximant	88.61%
	Nasal	91.33%
	Fricative	88.02%
	Stop	92.27%
	Silence	91.08%
Manner recognition	65.53%	

其正確率在跨語言偵測辨識後大幅降低了近 20%。在此情形下，我們另行使用 TCC-300 語料庫訓練中文的語言屬性偵測系統，其架構和訓練方式就和英文使用 TIMIT 語料庫如出一轍，回顧圖 3.1。而最後使用新訓練而得的系統測試國語語音上的應用，其統計結果如下表 3.4。

表 3.4：由國語 TCC-300 語料庫訓練之語言屬性偵測辨識系統模型其運用於國語 TCC-300 語料庫之測試結果

Corpus	**Training**	TCC-300
	Testing	TCC-300
Manner detection	**Vowel**	91.21%
	Approximant	92.94%
	Nasal	95.25%
	Fricative	94.98%
	Stop	97.24%
	Silence	97.96%
Manner recognition	85.41%	

由表 3.2 和表 3.4 可以發現，由國語語料庫訓練國語語言屬性偵測之系統，其在國語測試語料的表現和由 TIMIT 所訓練而得的偵測器測試英文語料的結果差不多，而辨識器之正確率也是大致一樣。雖然如此，但由偵測的結果去比對音檔，我們可以發現偵

測系統大致上是和強迫對齊結果相符，但事實上強迫對齊的結果和正確的語音信號標記之差距仍是很大，很多音素區段會產生前移的現象，進而導致某音素區段中涵蓋了前音素的資訊，而本該屬於自己之區段卻劃分給後面的音素。圖 3.17 充分表示了此種現象。

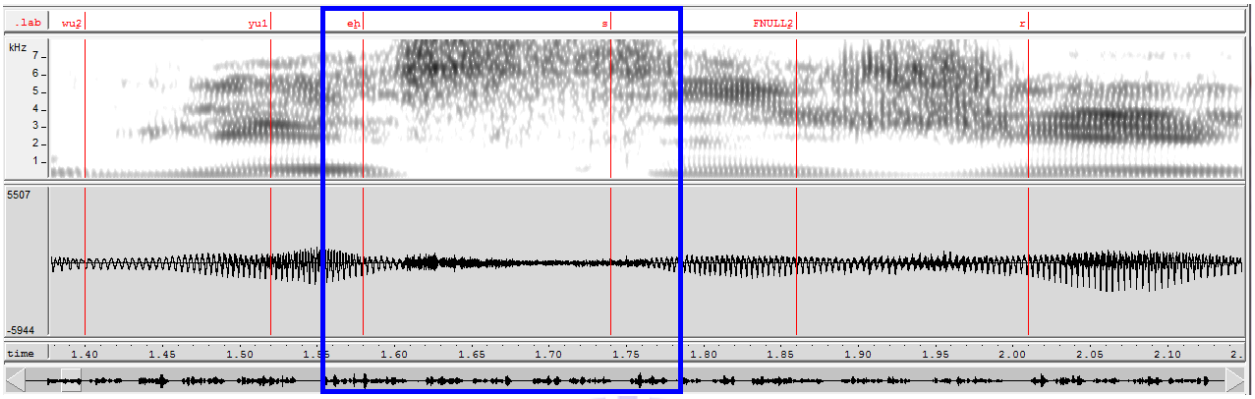


圖 3.17：文本強迫對齊時間與實際語音之誤差範例圖。由上而下分別是文字時間轉寫標記、聲譜圖以及原始語音信號

於此因素，我們可以推測由 TIMIT 訓練而得的語言屬性偵測系統跨語言到國語進行偵測其測試結果劣化如此多乃是因為中文的強迫對齊不佳導致答案不準確所致。而另一方面，因為系統可以準確學習轉寫時間標記，我們也就可以認定如果有一個良好的轉寫時間標記當作初始，則模型的結果絕對會較優良。在此種條件下，本論文提出了利用跨語言之屬性偵測器訓練一組新的偵測系統，3.4.1 節介紹國語語言屬性偵測系統之架構以及注音之屬性分類；3.4.2 節則是介紹音素時間標記之重製方法；3.4.3 節介紹新的時間標記所訓練出的語言屬性偵測系統其測試統計以及系統再次重新建立後的結果。

3.4.1 國語語言屬性偵測系統

3.4.1.1 國語語言屬性偵測系統之架構

在先前的實驗中，我們已從英文 TIMIT 語料庫獲得了語言屬性偵測以及辨識系統，然而實際上不同的語言間其在發音方式的接續上會有所差異存在，如在中文中爆破音後不會接摩擦音，但在英文中卻會有此種發音排序之發生，而偵測系統的架構是藉由長窗

框裡的參數向量判別，這也代表著語言屬性的偵測跟前後發音方式是有些許的相關性存在；另外，不同語言間其音素不一定能夠完全對應，而且其在發音上仍舊是有些許之不同。以上所述之差別可能會影響所萃取之參數，進而在跨語言上的測試導致結果劣化。於此因素，跨語言的偵測系統勢必得經過目標語言重新調整訓練模型，如此才能得以獲得較佳的目標語言屬性偵測系統以及實驗測試結果。

當然，我們可以從 TCC-300 語料庫的資訊從頭訓練一個國語語言屬性偵測系統，然而即便是使用的文字轉寫時間經過重新對齊調整後仍不夠精準，這也代表著由此訓練而得的模型其偵測能力是較差的。而 TIMIT 語料庫的文字轉寫乃人為時間標記，其時間標記之準確性最佳，因此訓練而得的偵測模型系統的效能和穩定性都較優。如果從一個已存在且較優良的模型開始調整和更新模型，其結果應該是會優於一個從無到有之模型。於此因素，我們將由 TIMIT 語料庫訓練而得的模型當作國語語言屬性偵測模型訓練中的初始模型，接著由 TCC-300 語料庫的參數向量以及重新製作的音素切割分段位置調整並訓練系統，如此即獲得了一個國語語言屬性偵測系統。圖 3.18 清楚展示了國語屬性偵測系統之訓練流程。

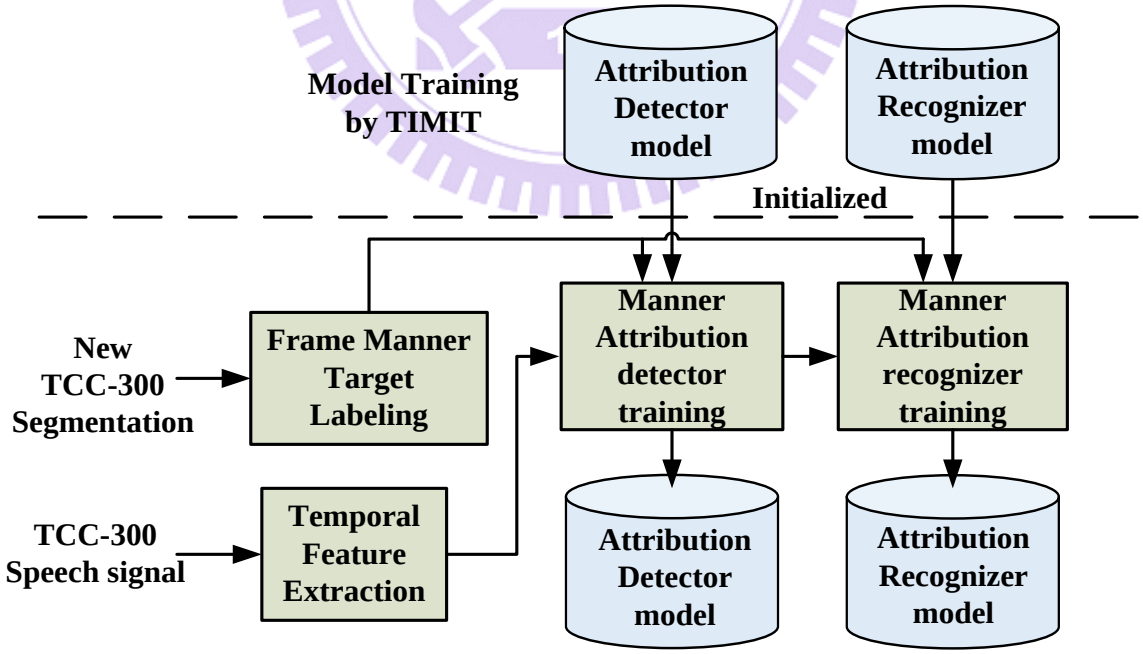


圖 3.18：國語語言屬性偵測系統之訓練流程

在模型完成調整和訓練後，我們即可其依圖 3.11 的流程進行偵測以及辨識。

3.4.1.2 國語語言屬性之分類

文字轉寫在標記時可使用不同的層級表示之，而在中文裡可用的層級有聲母韻母層級或者是類音素層級...等。然而，在聲母韻母層級中，有些韻母是屬於複韻韻母類型，這表示著韻母乃是由兩個單韻母組成，如：ㄠ是由ㄚ和ㄨ所組成；又或者是包含介音所構成的韻母，如：一ㄡ，這些韻母其實都含有數個不同的發音，因此若使用聲母韻母層級對照發音方式進行分類就會顯得不夠精細。於此因素，在此研究中使用的文字轉寫是屬於類音素層級，其將注音簡單轉換對照成幾個類音素所組成，表 3.5 顯示了注音轉換成類音素之標記。

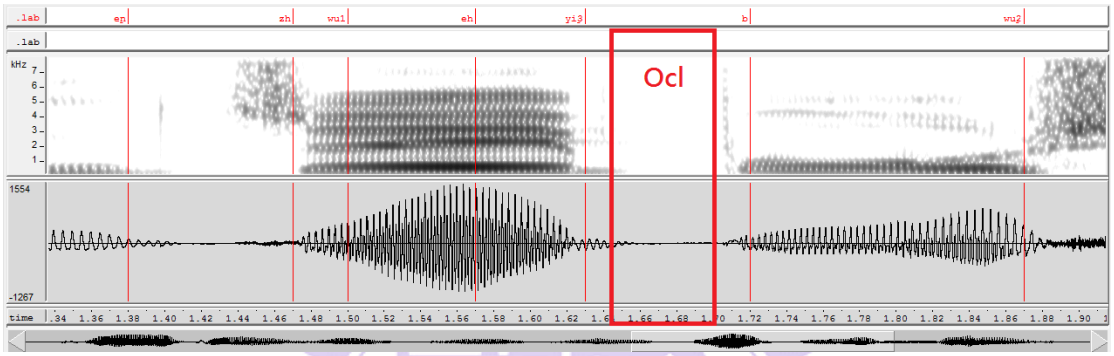
表 3.5：國語注音轉換類音素層級之對照表

注音	類音素標記	注音	類音素標記	注音	類音素標記
ㄅ	b	ㄒ	x	ㄜ	e
ㄆ	p	ㄓ	zh	ㄝ	eh
ㄇ	m	ㄔ	ch	ㄞ	a yi3
ㄈ	f	ㄕ	sh	ㄟ	eh yi3
ㄉ	d	ㄖ	r	ㄠ	a wu3
ㄊ	t	ㄗ	z	ㄡ	o wu3
ㄋ	n	ㄘ	c	ㄢ	a en
ㄌ	l	ㄙ	s	ㄣ	e en
ㄍ	g	ㄧ	yi1/yi2	ㄤ	a ng
ㄎ	k	ㄨ	wu1/wu2	ㄥ	e ng
ㄏ	h	ㄩ	yu1/yu2	ㄦ	er
ㄐ	j	ㄚ	a	ㆠ	FNULL1/FNULL2
ㄑ	q	ㄛ	o		

在表 3.5 中值得注意的是介詞部分，介詞的類音素發音除了單獨成為韻尾以及置中連接聲母和韻母之外，其發音也有可能成為複韻母之韻尾。而介詞在字裡的位置不同，也會影響到其表現和特性，因此我們將其分為三個不同之類別：連接聲母和韻母的在類音素標記後加上數字 1；單獨成為韻尾的則在類音素標記後加上 2 的數字；複韻母之韻尾則在類音素後標記 3。

另外，我們先前提到爆破音的發聲前會有著嗓音起始時間，此現象在中文裡也是存在的。然而中文文本在過往的文字轉寫時間標記中，在爆破音前都不會給予標記此停頓，而實際上爆破音本身的長度較短，此種不含嗓音起始時間的標記在隱藏式馬可夫模型訓練時很容易導致爆破音是運用大量的短停頓訓練，這也導致了文本強迫對齊的同時，爆破音區段內其實是涵蓋了大部分的短停頓，圖 3.19 充分顯示了此種情形。

(a)



(b)

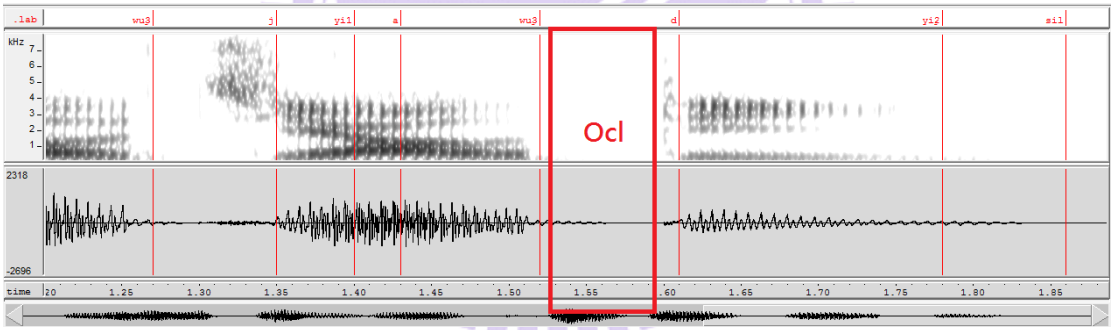


圖 3.19：中文強迫對齊之文字時間轉寫標記中爆破音前的嗓音起始時間之範例圖(a)和(b)，紅色四方框裡表示的是嗓音起始時間。由上而下分別是類音素層級文字轉寫時間標記、聲譜圖、原始語音信號

於此因素，若本研究中希望能夠獲得一個更精準的文字標記以及轉寫時間，勢必得在爆破音上有所行動。在此，我們對於所有的爆破音前面都加上了「Ocl」以表示嗓音起始時間。因此，所有國語類音素的發音方式分類皆清楚地列於表 3.6 中。

表 3.6：國語 TCC-300 語料庫之類音素對應發音方法語言屬性之分類

Manner	Phone-like
Vowel	a、o、e、eh、yi1、yi2、yi3、wu2、FNULL1、FNULL2、yu2、er
Fricative	zh、ch、sh、z、c、s、f、j、q、x、h
Approximant	l、r、wu1、wu3、yu1
Stop	b、p、d、t、g、k
Nasal	m、n、en、ng
Sil	sp、sil、Ocl

3.4.2 利用跨語言屬性偵測器之資料校正國語分段位置

一個良好的文字時間轉寫標記對於系統模型的訓練和穩定是有極大的幫助，然而在國語 TCC-300 語料庫中並沒有人為時間文字轉寫標記，且其經由抽取梅爾倒頻譜參數所訓練的隱藏式馬可夫模型之強迫對齊結果也不佳，因此，如何獲得一個較佳的文字轉寫時間標記是我們必需面對的問題。

在此，本論文中提出了一個想法：既然語言屬性偵測系統是對語音信號進行發音方式之分類，而音素正是其所使用之層級，因此其辨識結果輸出應能在文本強迫對齊時提供一個絕佳的參數向量，幫助文本在時間轉寫標記之結果。又我們先前提到語言屬性偵測系統在跨語言運作時，其語言中前後之發音方式接續可能會有所不同，另外還有不同語言間標記相同之音素其發音上卻存在著些許差異，這些因素都極有可能導致偵測結果不佳，因此我們需要一個由目標語言重新訓練的語言屬性偵測系統；TCC-300 原始的強迫對齊之音素切割分段位置極差，由此訓練而得之系統其穩定性以及偵測能力必定不佳，因此我們需要一個標記較佳的音素切割分段位置來訓練偵測系統模型。於上之兩個因素，我們將擁有最佳標記的 TIMIT 語料庫以及屬於目標語言之 TCC-300 語料庫合併，並將其共同訓練一套通用語言屬性偵測系統以供後續運用。在此，本論文使用 80 個 TIMIT 音檔以及 750 個 TCC-300 音檔進行訓練。

在語言屬性偵測和辨識模型訓練完成後，即可對 TCC-300 所有的音檔進行辨識測試，而感知器之輸出包含六個發音方式以及一個其他項（Other）之概似率，共 7 維參數。

接著，研究中另行對音檔抽取 38 維之梅爾倒頻譜係數，並將每個音框之語言屬性辨識系統輸出對應每個音框的倒頻譜係數，而後將兩組參數合併成一個新的參數向量，藉此來表達音框內的資訊。新的參數向量除了梅爾倒頻譜係數之 38 維外，另有不含其他項的辨識器輸出 6 維參數，共 44 維。

接著，研究中使用 HTK 訓練聲學模型，由新的 44 維參數向量訓練國語的音素層級之馬可夫模型，並依序提升其模型複雜度（Mixture），由 1、2、4、8、…直至目標複雜度後停止，在此，我們最終設定之音素模型複雜度為 64。最後，由新的音素層級之馬可夫模型對文本進行強迫對齊，並獲得最新之音素切割分段位置。文字時間轉寫標記之重新製作的流程清楚顯示於圖 3.20。

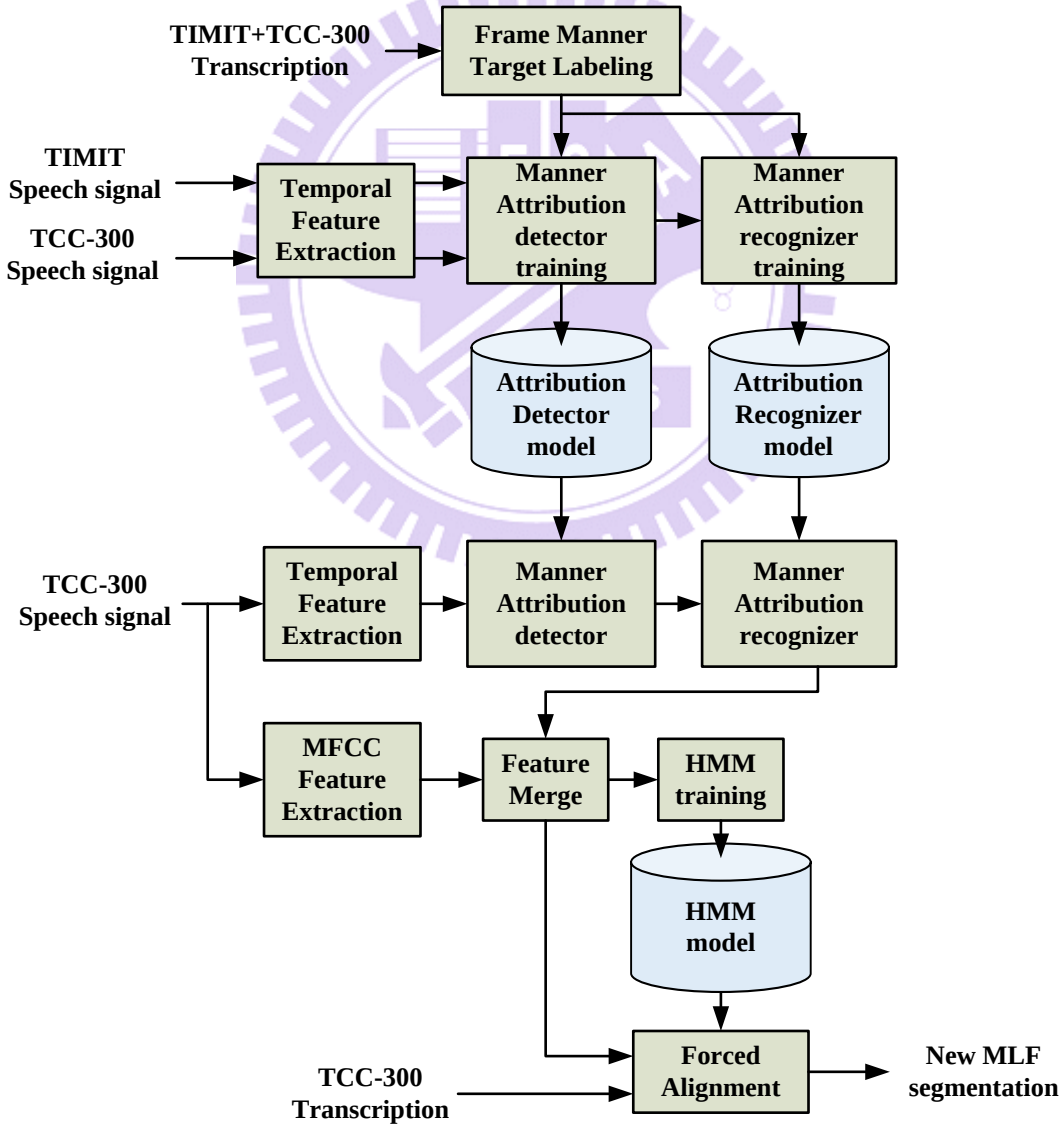


圖 3.20：文字時間轉寫標記之重新製作流程

在文本重新經過強迫對齊後，我們比較不同複雜度之文字時間轉寫標結果後，發現在較低的複雜度下之摩擦音的時間對齊優於複雜度高時的結果。然而在複雜度較低時，因為其在參數向量空間上的分類少，其強迫對齊結果較有可能存在時間嚴重偏差的情形發生，導致音素之時間標記錯誤；反之，複雜度較高時，強迫對齊的結果則較不易出現時間嚴重偏差的現象。圖 3.21 顯示不同複雜度下對時間標記的影響，其中低複雜度下強迫對齊產生錯誤造成極大的偏差。

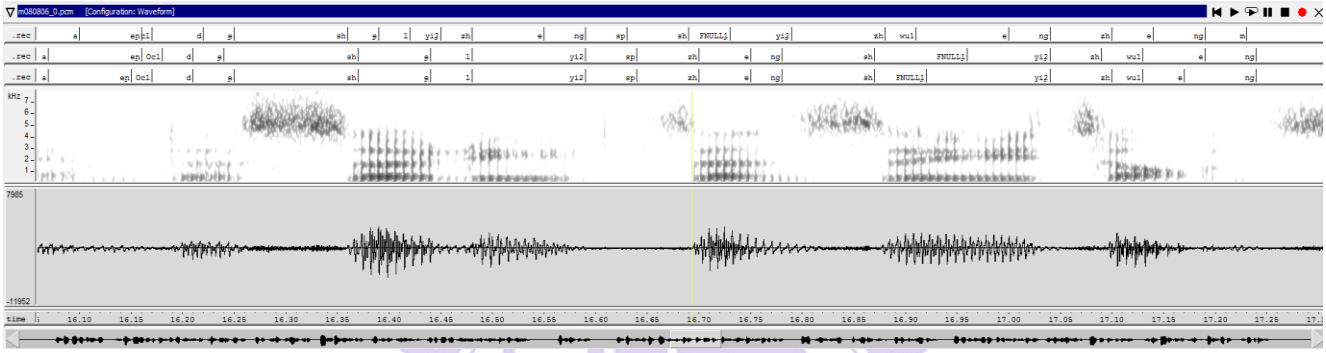


圖 3.21：不同複雜度之音素模型對文本強迫對齊後之結果。由上而下分別是複雜度 1、複雜度 8、以及複雜度 32 之強迫對齊時間、頻譜圖以及原始語音信號

另外，圖 3.22 顯示摩擦音在不同的複雜度下的強迫對齊結果，在低複雜度下的摩擦音之對齊時間普遍是優於高複雜度的結果，紅色方框內的摩擦音「zh」和「h」皆是如此。

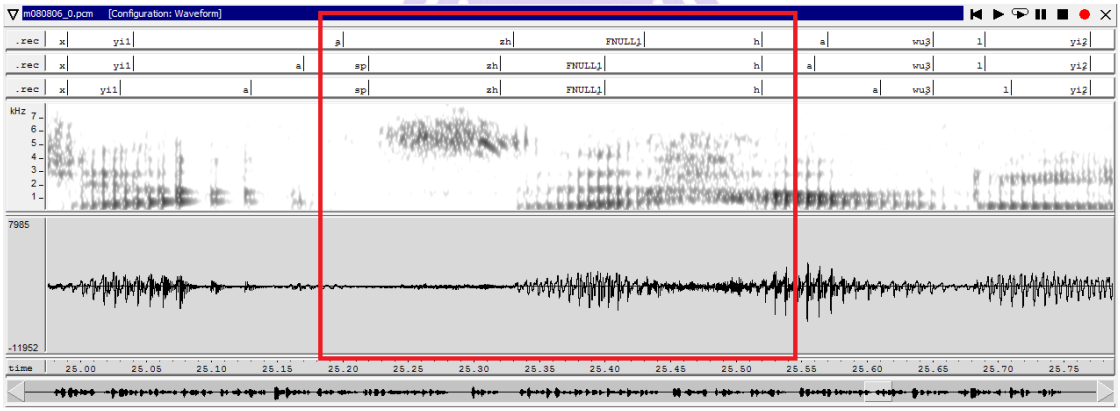


圖 3.22：不同複雜度之音素模型對文本強迫對齊後之結果。由上而下分別是複雜度 1、複雜度 8、以及複雜度 32 之強迫對齊時間、頻譜圖以及原始語音信號

由上面所述之討論，我們知道高複雜度和低複雜度之文本強迫對齊結果各有其優缺點，而本研究又以取得一個良好的時間標記為目標，因此我們結合兩種結果之優點並改善其缺點以得到一個較佳之文字時間轉寫標記。實驗中依較高複雜度的強迫對齊結果為主，而低複雜度的結果則當作輔助，其用來調整高複雜度之文本對齊時間中之摩擦音時間起始和結束位置，如此，不僅可以避免強迫對齊時的嚴重偏差，更可以調整摩擦音至較佳的位置。在此，我們觀察文本強迫對齊在複雜度為 8 時就已經有足夠能力可以避免音素之時間標記嚴重偏差，因此，本實驗中以複雜度為 8 之對齊結果為主，複雜度為 1 的對齊結果為輔，重新製作後的文字轉寫表現如圖 3.23。

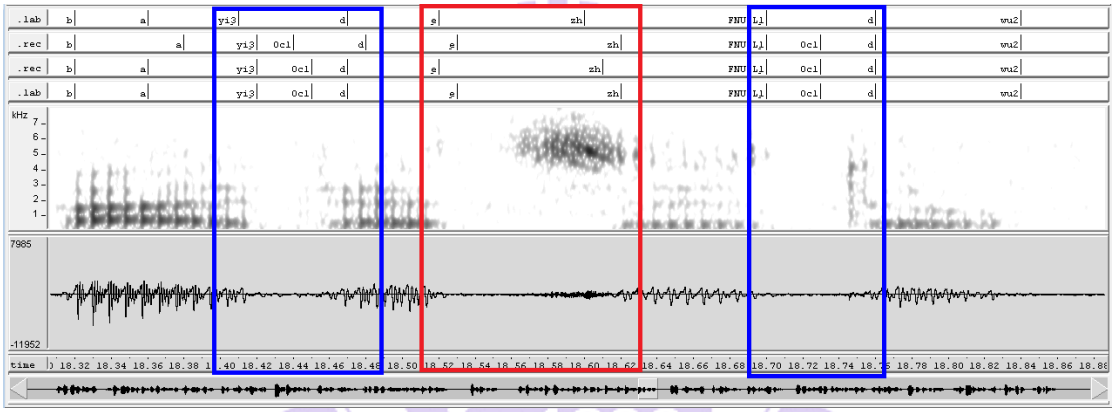


圖 3.23：文字時間轉寫標記重製後之結果比較圖。由上而下分別是原始強迫對齊時間標記、複雜度 1、複雜度 8 之強迫對齊時間標記、重製後之文字時間轉寫標記、聲譜圖以及原始語音信號

由圖中紅色方框處可以清楚看見摩擦音在調整之後的時間標記，比對上聲譜圖之能量分布確實是移到了較佳之位置；藍色方框處則是顯示了爆破音前的嗓音起始時間被獨立標記了出來，由聲譜圖裡的能量分布可以清楚得知嗓音起始時間確實是該獨立並歸於靜音，而不是將其歸於爆破音內，也因此在重新製作後之文字時間轉寫標記中其爆破音平均時間較短，較符合實際爆破音之特性。雖然重製之結果和原始強迫對齊相比仍是有優有劣，但是相較起來對齊時間變好的的比例是大於劣化的，也因此後續的研究是使用重製後之文字時間轉寫標記。

3.4.3 模型測試及再次重製

有了重製後之文字時間轉寫標記之後，我們由此開始訓練語言屬性偵測模型系統，訓練的流程就如 3.4.1 節所介紹，使用 TIMIT 之模型當初始值開始建立系統，並經過測試語料進行測試，表 3.7 為使用重製之音素時間標記所訓練之系統模型測試之統計結果。

表 3.7：由 TIMIT 訓練之模型作為初始值之國語語言屬性偵測及辨識系統其運用於國語 TCC-300 語料庫測試結果

Corpus	**Training**	TCC-300
	Testing	TCC-300
Manner detection	**Vowel**	92.43%
	Approximant	95.27%
	Nasal	96.42%
	Fricative	97.19%
	Stop	98.45%
	Silence	97.14%
Manner recognition	89.53%	

雖然文字時間轉寫標記在經過重製之處理程序後，其結果相較於之前的強迫對齊已經好了很多，然而實際上其仍是存在著時間之誤差；此外，人類在講話時常常會節省發音力氣因而導致字詞內的音節弱化甚至是消失，這種情形尤其是自發性語音時更是明顯。然而，我們在製作文本時並非是依照實際發音之情形而標記，因為此法必須耗費大量之人力與時間，轉而代之的是將文本中的中文直接轉換成類音素之標記使用，因此，音檔中實際之發音有可能和轉寫標記不同，這些因素就會影響整個系統的建立。

為了獲得一個更佳的模型，在 3.4.1 節中語言屬性偵測和辨識系統建立好後，研究中將辨識器輸出對發音方式進行對齊之程序，藉此獲得一個新的音框目標狀態函數。此次的對齊中，狀態是有可能跳過分數太低的類型輸出，也就是會將一些弱化或不存在之

發音方式移除，因此對齊後之音框目標狀態函數是一個較符合實際發音的目標狀態序列。由對齊後之結果再次對模型進行調整以及更新，如此便可以獲得一個更佳之語言屬性偵測及辨識系統，圖 3.24 則為模型再度重製訓練的流程圖。

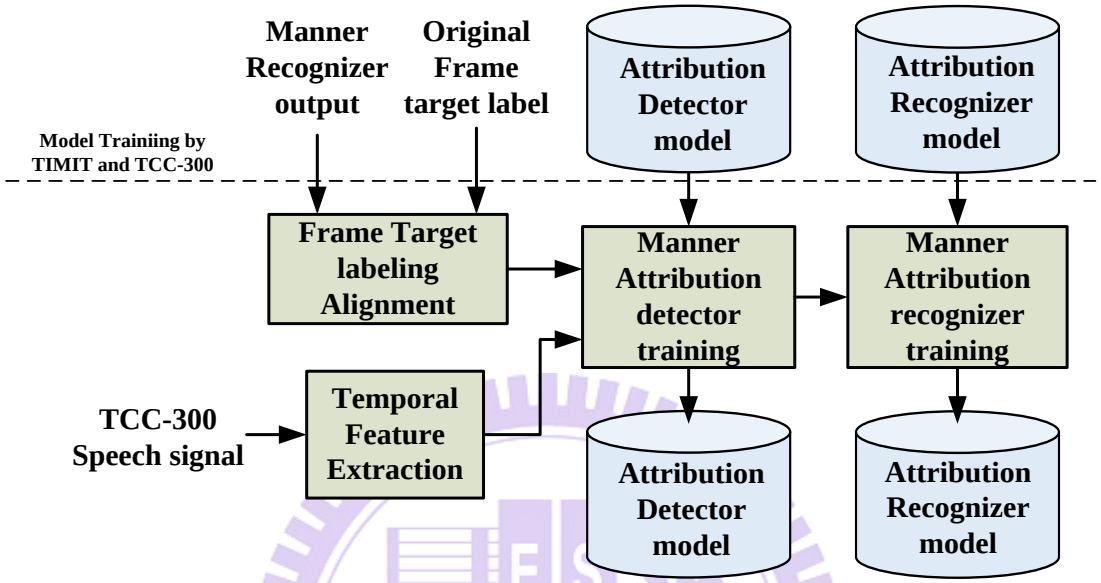


圖 3.24：經由音框目標狀態函數對齊之模型重製流程圖

在系統重製之後，我們一樣使用測試語料判別系統的優劣，而在此的測試語料音框目標狀態函數也是將經過強迫對齊後的結果當作實驗用的答案，表 3.8 為測試結果統計。

表 3.8：經由目標狀態函數對齊後再重製之國語語言屬性偵測及辨識系統其運用於國語 TCC-300 語料庫測試結果

Corpus	**Training**	TCC-300
	Testing	TCC-300
Manner detection	**Vowel**	95.57%
	Approximant	97.68%
	Nasal	98.22%
	Fricative	97.99%
	Stop	99.11%
	Silence	97.98%
Manner recognition	94.88%	

第四章 實驗結果

人類在講話時，語者會不自覺地將發音中的音節改變或是弱化，更甚至是刪除，此現象在語音學中稱為音變。而語言學上將音變現象分成了歷史音變以及語流音變，歷史音變乃語言經過時間發展而產生之變化，如古代的平、上、去、入四聲變成普通話之陰平、陽平、上聲、去聲；語流音變則是語流中音節和音節間產生之影響導致發音之變化。然而語言學家其探討的內容太過廣泛，譬如考慮不同年紀之語者其在發音上的音素變化差別，或者是不同地區的語者其發音的差別...等，在本研究中只先選擇較簡單的部分進行探討。

4.1 常見字詞之觀察分析

首先，人類在朗讀式語音時會因為文本已經完整規劃，語者只需照著文本一字一句的朗讀即可，因此郎讀式語音的發音相對來說是完整的；而自發性語音則是因為語者一邊產生語音一邊經由大腦規劃構詞，在此，通常語者所使用的詞彙會集中於常用的兩千個字詞內，另外，語者在自發性語音時經常會有習慣用詞或是無意義字詞，而在這些無意義以及常用字詞之發聲時，語者通常會快速地帶過或變調，也因此這些字詞相對地容易造成發音上之變異以及發音方式的弱化甚至是脫落。

為了比較兩種語音間的差別，研究中分析在兩種語音裡皆常出現的字詞。根據曾淑娟博士對於朗讀式以及自發性語音的探討和整理[5]，我們選擇了幾個在兩種語音中都常出現的字詞進行觀察和分析。研究中使用的是本論文 3.4.3 節所訓練而得的模型系統並對語音信號進行發音方法語言屬性的偵測，且由 HTK 對齊之目標字詞的時間結果當作偵測區段，接著將偵測器之輸出進行動態時間校正，在此處我們設定發音方法狀態序列不允許跳過，這代表的是目標字詞裡的每個發音方式至少會擁有一個音框長度，而後我們再將目標字詞內的每個發音方式計算其信心度量值（confidence measure）。信心度量

值 c_i 的計算如下：

$$c_i = \left(\prod_{x_j \in \text{Manner } i} x_j \right)^{\frac{1}{n_i}} \tag{4-1}$$

$$n_i = \sum_{\substack{x_j \in \text{Manner } i \\ j=1}}^{N} 1 \tag{4-2}$$

其中，N 為目標偵測區段的總音框數，c_i 為經過動態時間校正後第 i 個發音方法的信心度量值，x_j 則是第 j 個音框的偵測器輸出，n_i 則是發音方法 i 在動態時間校正後佔有的音框數。

在所有常出現之字詞經過動態時間校正以及信心度量值之計算後，我們比對同一個字在字詞裡面不同位置其發音方式的信心度量值分布，並比對在不同語音中信心度量值的差別。圖 4.1 為「因」的鼻音信心度量值分布圖，由圖可以發現在朗讀式語音中「因為」的「因」以及所有語料庫中的「因」其鼻音的信心度量值大多是聚集於較高的值，而「原因」的「因」相對來說則不是那麼集中，此外，這些類別在低值時也有一個集中的情形；自發性語音中「因」的鼻音信心度量值則是集中於較低的地方。

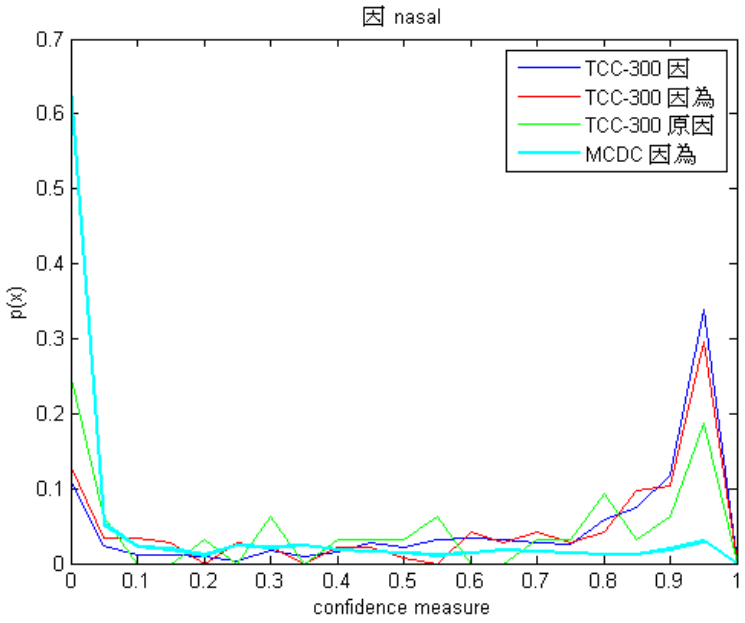


圖 4.1：「因」之鼻音發音其信心度量值分布圖

除了觀察信心度量值的分布圖之外，我們也對發音方法經過動態時間校正後所佔有之時間長度進行統計。圖 4.2 顯示了在朗讀式語音中其鼻音之時間分布是較平均的，但其在一個音框長度仍有一個高值，在比對信心度量分布之結果後，我們可以推論經過動態時間校正後發音方法之長度有很短的情況時，這些部分很可能會有脫落的情況發生，而其餘發音長度較長的部分則較不易產生發音方式脫落的現象，因此整體來說，「因」的鼻音在朗讀式語音中是不至於有嚴重的脫落；；在自發性語音中，除了經過動態時間校正之結果其發音方式佔有之音框大多集中於一個之外，信心度量值分布也集中於低值，由此兩樣資訊我們可以推論在自發性語音中，「因」的鼻音發音應該是有著嚴重的脫落現象。

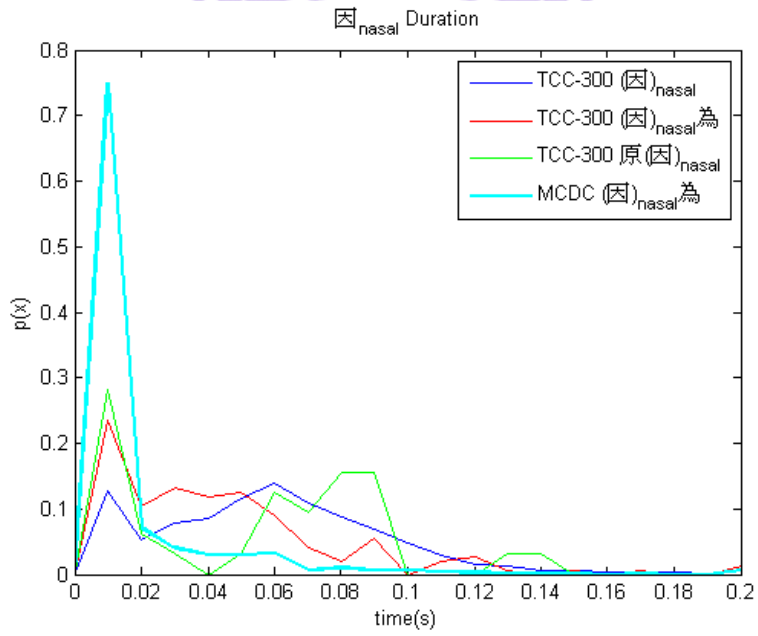


圖 4.2：「因」經過動態時間校正後其鼻音所佔之時間長度

圖 4.3 則是「為」其流音之信心度量值分布。在朗讀式語音中不論是統計所有語料中的「為」或者是含有「為」的字詞，其流音之信心度量值也集中於高處；而在自發性語音中則是聚集於低處。

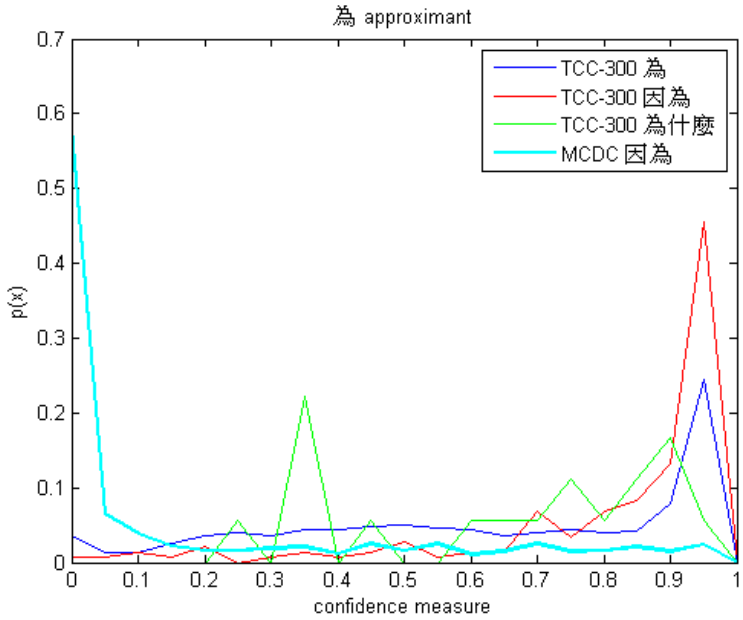


圖 4.3：「為」之流音發音其信心度量值分布圖

圖 4.4 則是「為」經過動態時間校正後所佔有之時間長度統計。圖中顯示了在朗讀式語音中，「為」的流音長度大多是佔了數個音框長度，比對了信心度量值分布後，我們推論其發音方式脫落的現象也不嚴重；而在自發性語音中，不論時間長度或者是信心度量值分布都是集中於低值，因此我們也推論了在自發性語音中「為」的流音也會產生嚴重的脫落。

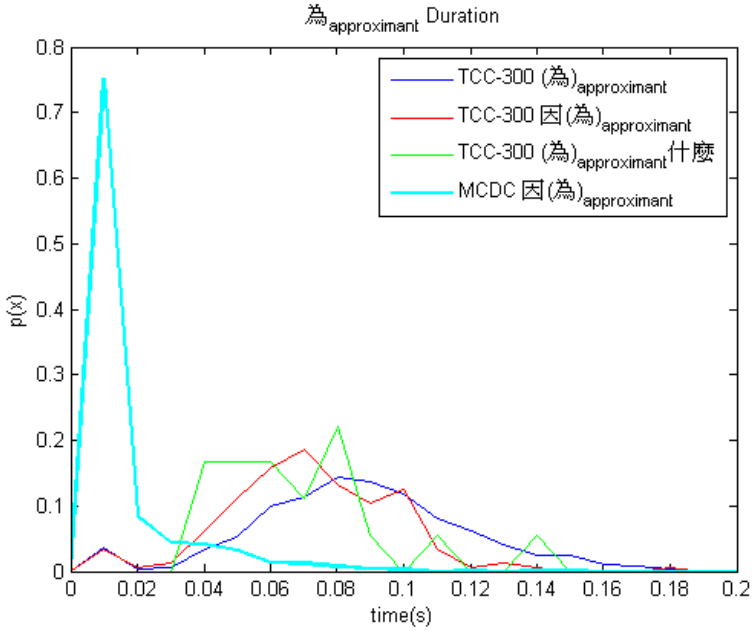


圖 4.4：「因」經過動態時間校正後其鼻音所佔之時間長度

由上述兩個例子我們可以清楚地發現，不論是「因」或者是「為」，其發音方式之信心度量在朗讀語音裡大都聚集於高值，雖然經過動態時間校正的長度並不一定較長，但由此兩筆資料進行分析，我們可以推論發音方式是較不易產生脫落的，這也代表著人類在朗讀式語音時確實是會較完整的發音，也因此偵測系統的結果比對文本來說正確率是較高的；自發性語音的發音方式信心度量值則是聚集在低值，且其經由動態時間校正的長度分布也極度集中於一個音框，這代表著在動態時間校正時因為並沒有設定可以跳過發音狀態，而語者在發音時也並沒有此種發音方式，此音框在程序中是強迫給予的，這結果也表示著人類在自發性語音時發音方式會很容易弱化或脫落，因為在聲音上產生了變異，也因此偵測系統的結果比對文本的正確率是較低的。

下面我們繼續對於常出現之字詞進行分析。圖 4.5 為「的」之爆破音信心度量分布圖，在朗讀式語音的部分，「真的」的「的」和所有的「的」經過統計後，其爆破音之信心度量值分布是集中在較高的範圍，而「真的」的「的」更是極端地集中於此；在自發性語言中，其爆破音之信心度量值分布則是集中於低值。

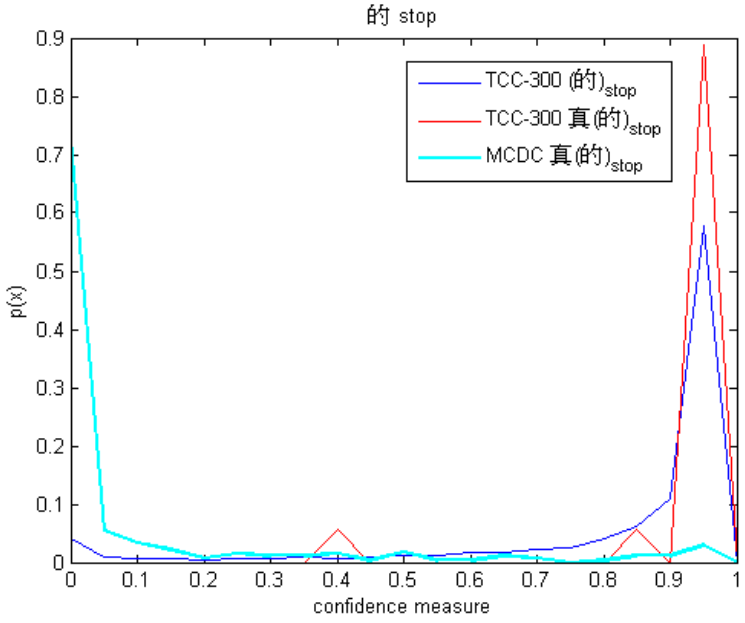


圖 4.5：「的」之爆破音發音其信心度量值分布圖

圖 4.6 則是「的」經過動態時間校正後所佔有之時間長度統計。由圖我們可以發現在自發性語音中，「的」之爆破音時間長度有近九成只有一個音框長度，比對其信心度量值集中在低值範圍，我們可以由此推斷「真的」的「的」在自發性語言中其爆破音的弱化或脫落的現象是非常嚴重的；在朗讀式語音中，「的」之爆破音平均時間長度大約佔了四個音框，這符合先前提到爆破音長度較短的結果，此外其信心度量值則都是在極高之值，代表語者在朗讀式語音中確實是有產生此種發音方法，也代表著字詞「真的」在不同方式的語音中其發音是有差異存在的。

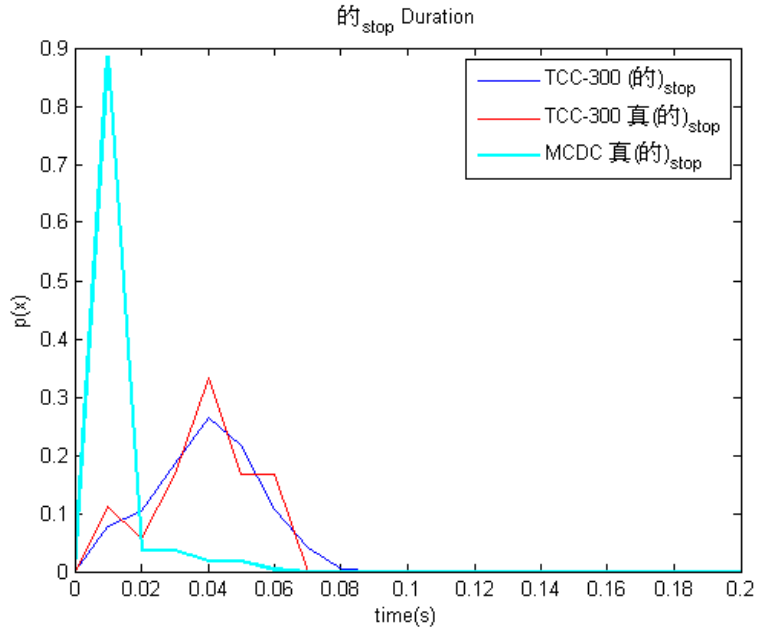


圖 4.6：「的」經過動態時間校正後其爆破音所佔之時間長度

圖 4.7 則是「所」的流音信心度量值分布圖。圖中顯示，在朗讀式語音中不論是語音中所有「所」之統計或者是字詞「所以」的「所」之統計其分布雖然沒有極端的集中於某個範圍，但仍是有較多的比例在高值部分，此外這兩種的統計分布是近似的，亦即「所」這個詞在朗讀式語音中並不會因為所在的字詞位置不同而有太大的差異性；而在自發性語音中，「所以」的「所」其流音之信心度量值分布則集中於低值的範圍。

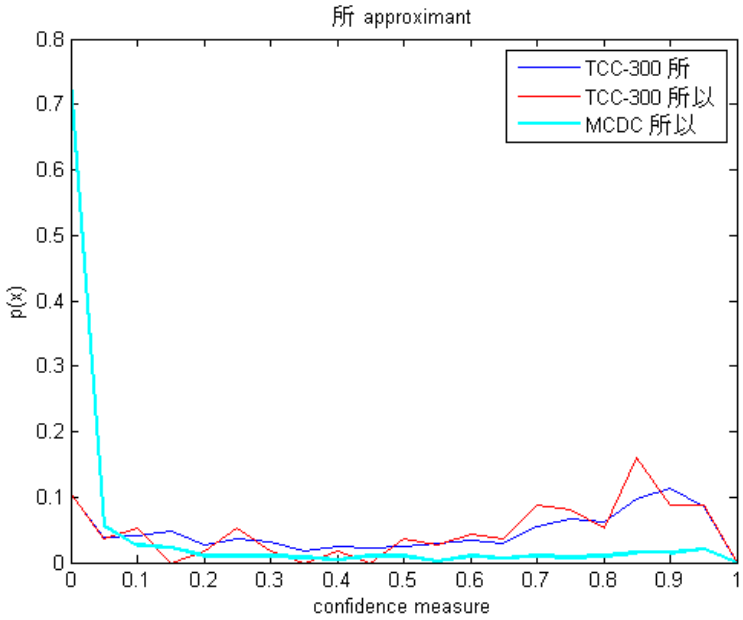


圖 4.7：「所」之流音發音其信心度量值分布圖

圖 4.8 則是「所」經過動態時間校正後所佔有之時間長度統計。由圖中可以發現在朗讀式語音時，其分布最多的是在一個音框長度，在比對的信心度量值分布後，我們可以推斷「所」的流音發聲在朗讀式語音裡雖然存在，但其長度很短；在自發性語音中，有超過九成的「所」其流音長度只有一個音框，而其信心度量值也集中於低值，因此，在自發性語音中，其流音的發音弱化或者是脫落的現象是非常嚴重的。

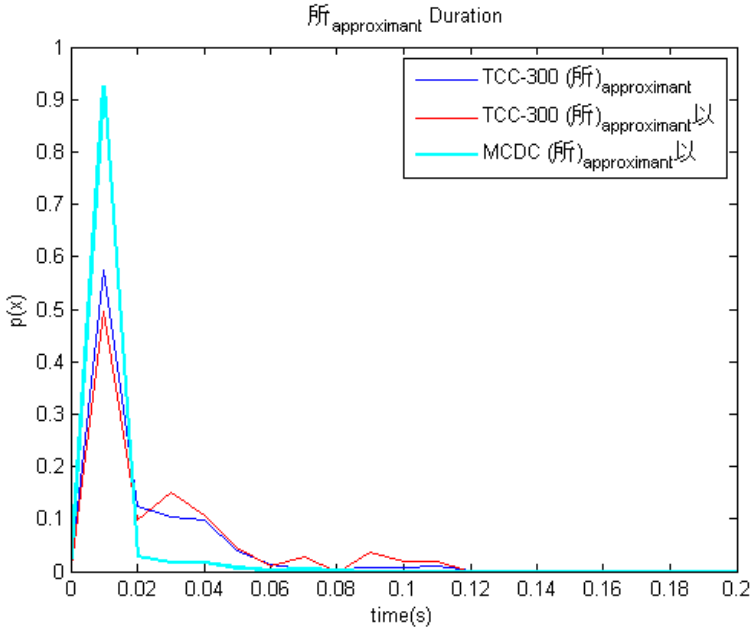


圖 4.8：「所」經過動態時間校正後其流音所佔之時間長度

下面我們觀察的是「沒有」的「有」其流音發聲的現象。圖 4.9 顯示朗讀式語音中「有」的流音信心度量值分布是較平均的，而在較高和最低部分則分布略有上升；在自發性語音中，「有」的流音信心度量值分布則是集中於低值，這代表著有很多的「有」其流音在經過偵測後是非常不確定其是否存在的。

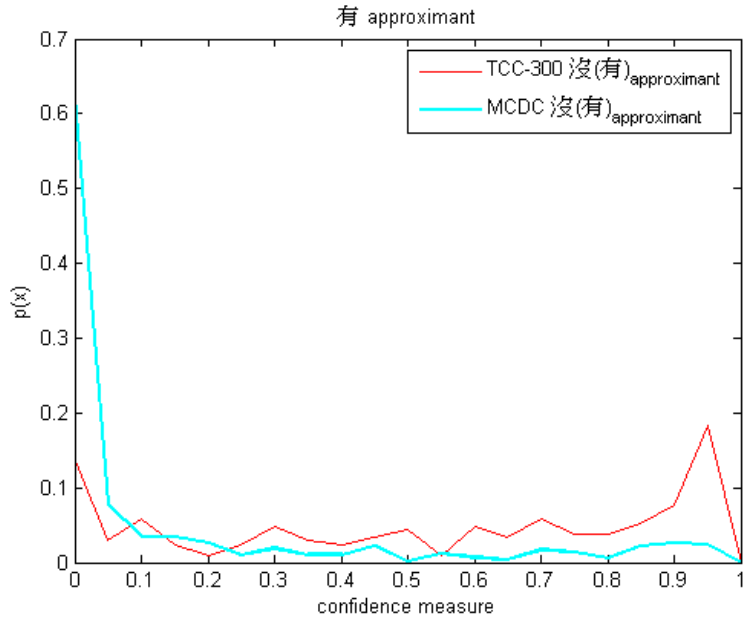


圖 4.9：「有」之流音發音其信心度量值分布圖

圖 4.10 則是「沒有」的「有」其流音部分經過動態時間校正後之時間長度分布結果。由圖中可以看見在朗讀式語音中「有」的流音長度集中於一個音框以及四個音框附近，比對其信心度量分布後，我們可以推測長度只有一個音框的結果應該大部分是發音方法產生了脫落的現象，而其餘非一個音框之結果則是有些許部分因為弱化而不是如此確定其存在，就整體來說在朗讀式語音中「有」的流音部分的脫落現象是輕微的；在自發性語音中，「有」的流音部分不論是信心度量值或者是時間長度分布都是聚集於最低值，如同於之前其他字詞的推測方法，我們一樣是斷定「有」在自發性語音中其流音確實是產生了嚴重脫落現象。

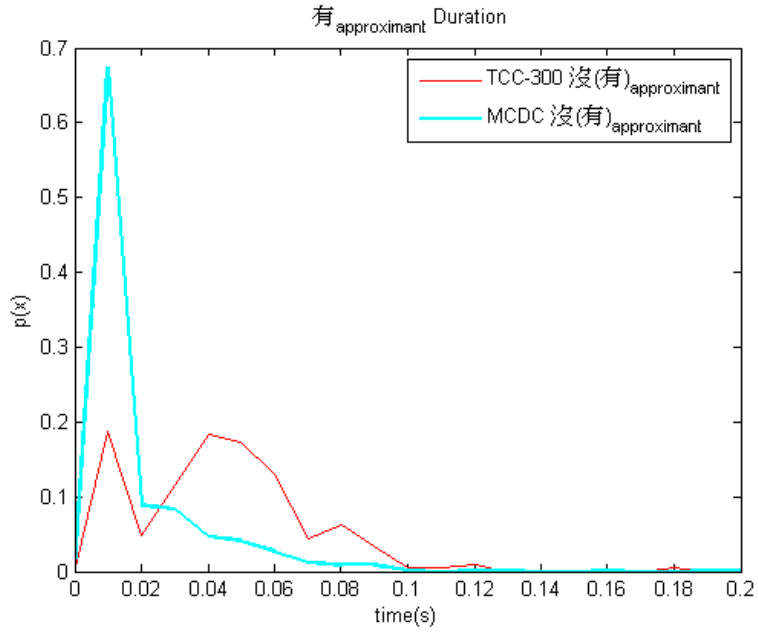


圖 4.10：「有」經過動態時間校正後其流音所佔之時間長度

接著研究對於字詞「可以」進行觀察，在此我們觀察的部分是「可」的爆破音。圖 4.11 顯示著其信心度量值的分布，由圖我們可以發現，不論是在自發性語音或者是朗讀式語音中，「可」之爆破音信心度量值在最高值時都有集中的現象，另外，自發性語音的信心值分布在最低值時也有集中的情形發生。

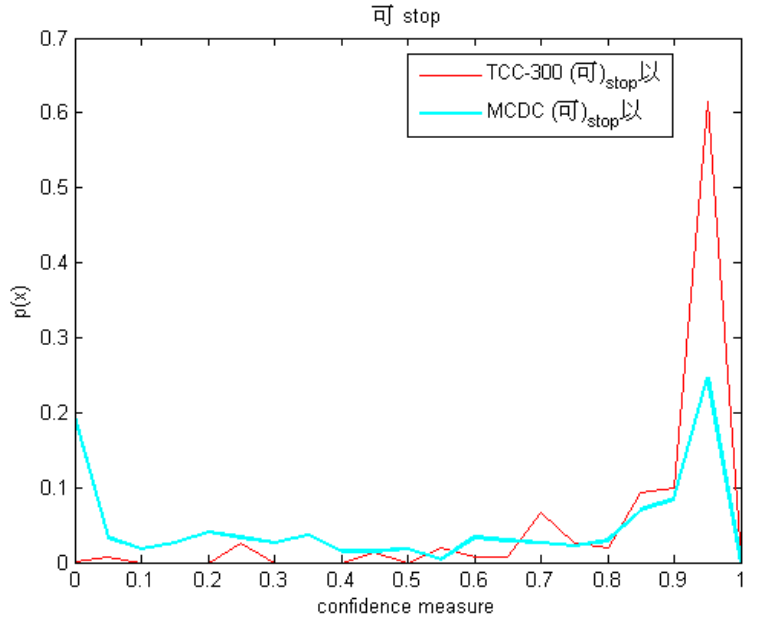


圖 4.11：「可」之爆破音發音其信心度量值分布圖

而圖 4.12 則是「可」在經過動態時間校正後其爆破音的時間長度統計。圖中顯示了自發性語音中「可」的爆破音集中於一個音框以及五個音框附近，比對其信心度量的分布我們可以推斷一個音框長度的結果大多是爆破音產生了脫落，而其餘非一個音框長度的爆破音則是存在的，因此整體來說在自發性語言中「可」的爆破音雖然會有脫落的現象，但仍屬於較輕微的；在朗讀式語音中，「可」之爆破音不管是信心度量分布或者是時間長度分布都顯示了其並不易產生弱化或者脫落之現象。最後，「可」雖然在自發性語音中可能會因語者說話而產生變異，但由兩種不同方式的語音觀察後，我們可以推論「可」的爆破音是較不容易有脫落的現象發生的。

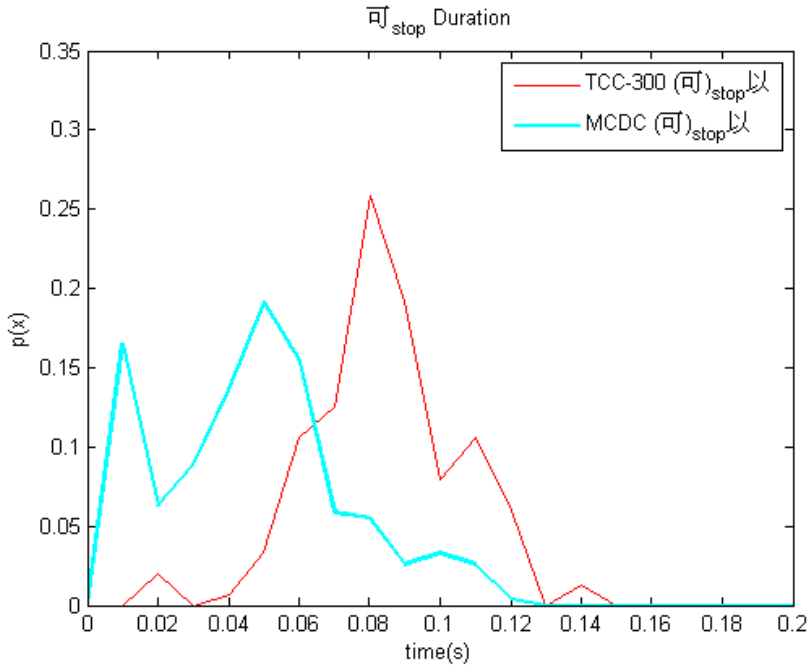


圖 4.12：「可」經過動態時間校正後其爆破音所佔之時間長度

圖 4.13 則是「什麼」的「麼」其母音的信心度量值分布。由圖可以看見自發性語音以及朗讀式語音其集中的範圍不同，在自發性語音中「麼」之母音其信心度量值是集中於最低值部分，相反地朗讀式語音則集中於最高值的範圍，這顯示著字詞「什麼」在兩個不同的語音中其發音是有差異的。

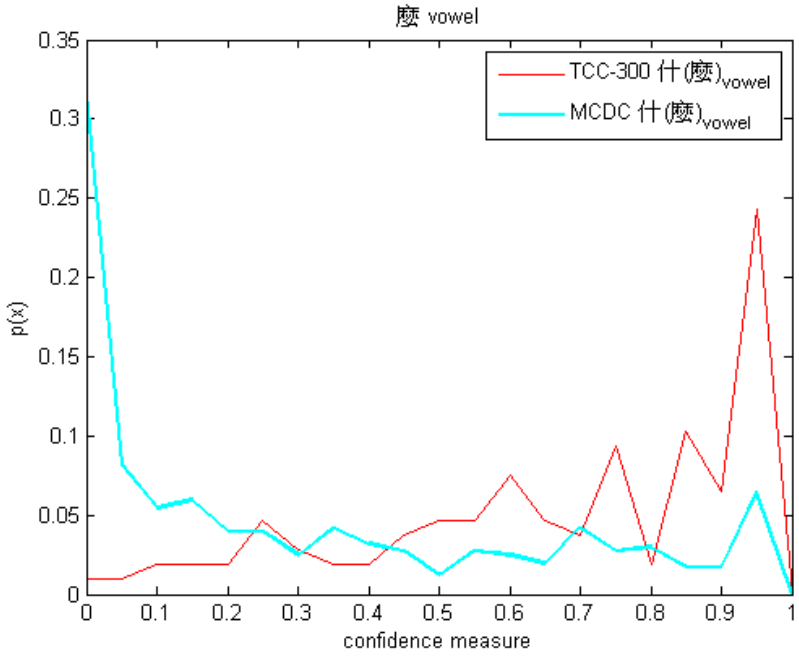


圖 4.13：「麼」之母音發音其信心度量值分布圖

圖 4.14 則是顯示「麼」在經過動態時間校正後其母音的時間長度統計。圖中顯示在朗讀式語音中，「麼」之母音長度並沒有極端集中於某一個值，反而是較平均的散落於各個長度分布，而其平均長度大約是 0.1 秒左右，跟先前討論過的字詞中其他發音方式之長度相比後是較長的，這符合了母音發音相較於其他發音方式其長度會較大的結果，而在比對了其信心度量值分布後，我們可以推斷「麼」的母音在朗讀式語音中大多是存在的；在自發性語音中，「麼」之母音長度則是集中於一個音框長度，但不同於先前觀察之發音方式是極度集中於此，其只有近四成左右，而比對信心度量值分布，我們也可以發現雖然其雖然集中於最低值，但也只有三成左右而已，因此我們可以堆斷其雖然仍舊會有發音方式脫落現象之產生，但其只能算稍微嚴重，而並非會極端的造成脫落的發生。比較了「麼」在兩種語音中的統計分布結果以及經過簡單的推論，我們可以知道「什麼」在不同的語音中其發音的差異是存在的，但是其產生發音方式脫落的機會並不如其他觀察之目標字詞來個高。

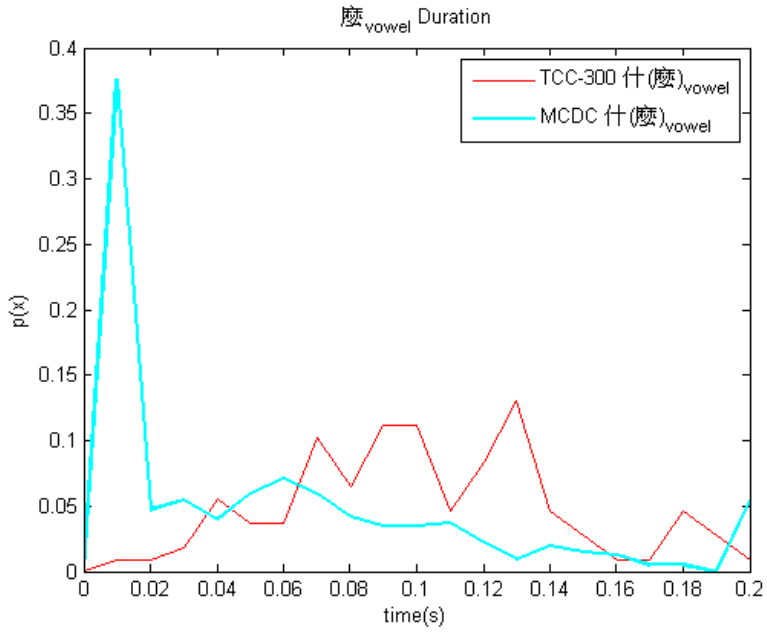


圖 4.14：「麼」經過動態時間校正後其母音所佔之時間長度

圖 4.15 則是「我們」的「我」其流音之信心度量值分布。圖中顯示了在自發性語音中，「我」之流音部分是非常的集中於最低值，而在朗讀式語音中雖然其也是集中於最低值，但其信心分布相較於自發性語音的分布是較平均的分散於各個值。

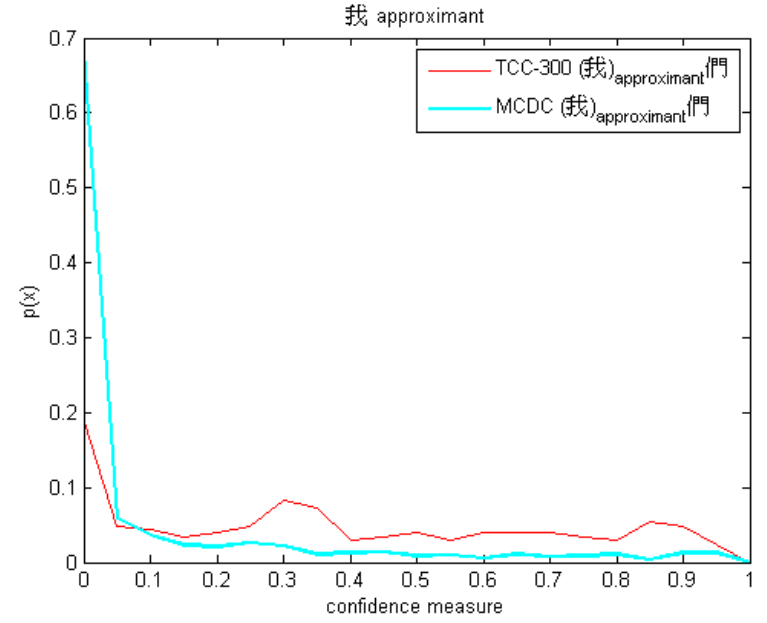


圖 4.15：「我」之流音發音其信心度量值分布圖

接著研究觀察其時間長度，圖 4.16 為「我」在經過動態時間校正後其母音的時間長度統計。圖中顯示自發性語音中「我」的流音部分極端的集中於一個音框長度，比對其信心度量的分布，我們可以推斷「我」在自發性語音中，其流音之部分的脫落現象是很嚴重的；在朗讀式語音中，「我」的流音則是較平均地分布於各個時間長度，在比對其信心度量之分布後，我們可以推測在朗讀式語音中「我」的流音部分雖然其信心度量值並非都很高，亦即並非所有流音都是非常確定，但其大多仍是存在語音中，因此在朗讀式語音中「我」的發音絕對是優於在自發性語音中。

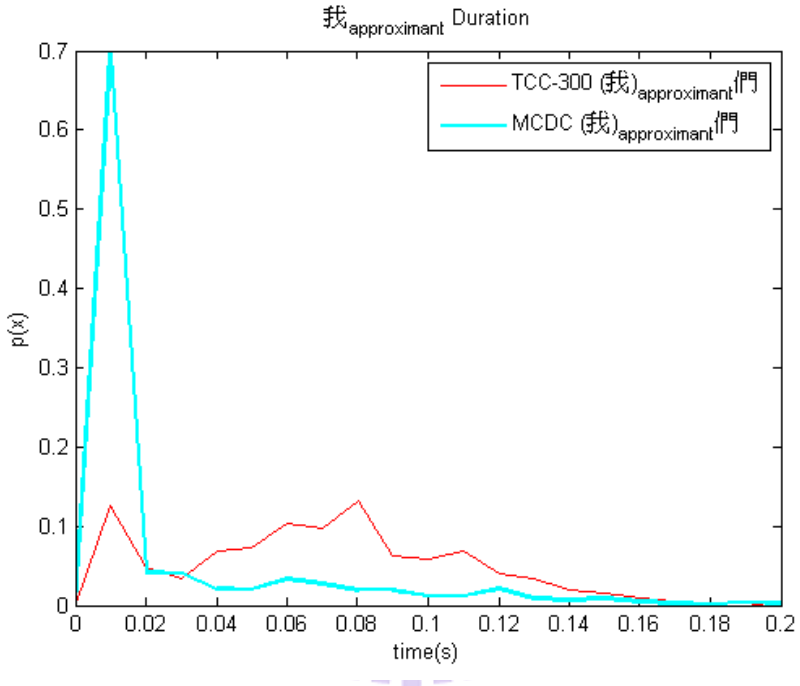


圖 4.16：「我」經過動態時間校正後其流音所佔之時間長度

圖 4.17 則是「果」其爆破音之信心度量值分布。由圖可以看見在朗讀式語音中，「果」的爆破音之信心度量值不論在字詞「如果」中或者是所有語料中「果」的統計都是非常地集中於最高值的部分，這代表著在動態時間校正後我們非常確定爆破音確實是存在；在自發性語音中，「果」的爆破音信心度量值則是相反，其近八成是極度的集中於最低值，這意味著很多的爆破音是不確定甚至是不存在的。由這張圖我們可以初步斷定在不同的語料中，「果」的發音差別是很大的。

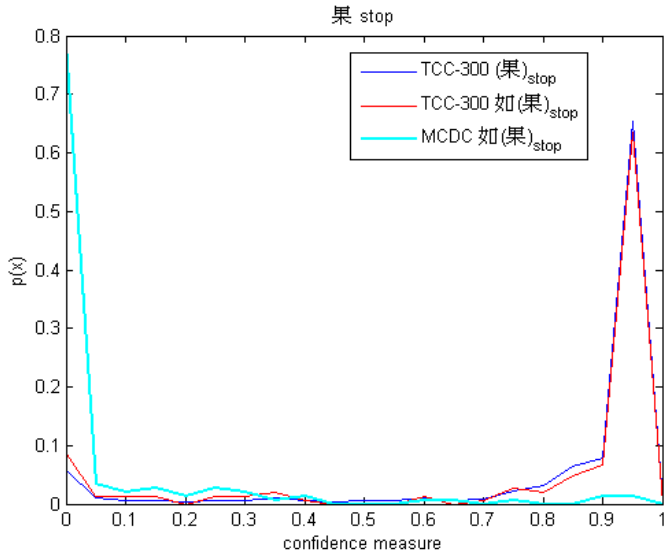


圖 4.17：「果」之爆破音發音其信心度量值分布圖

而圖 4.18 則是「果」在經過動態時間校正後其爆破音的時間長度統計。由圖可以清楚了解在朗讀式語音中「果」的爆破音大約集中於 0.04 秒左右，這長度符合我們先前提到爆破音長度較短的事實，比對其信心度量值的分布可以推測「果」在朗讀式語音中是幾乎是清楚完整的發音；在自發性語音中，其時間長度超過九成只有一個音框，而其信心度量值又集中於最低值，因此我們可以推斷在自發性語音中，「果」的爆破音是會有很嚴重的脫落現象。比對兩個語音的結果，我們可以說字詞「如果」在不同的語音中其發音方式是會有很大的差異的。

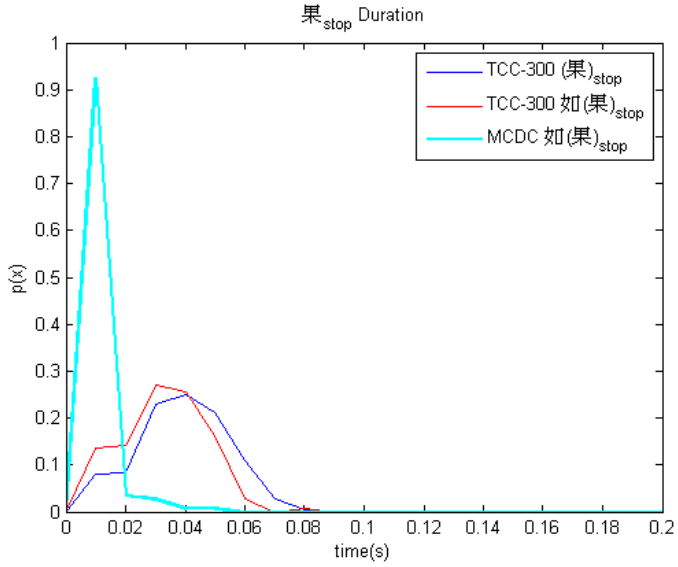


圖 4.18：「果」經過動態時間校正後其爆破音所佔之時間長度

除了由[5]挑選兩種語音皆常用的字詞外，研究中另外挑了在自發性語音中常出現的字詞。在此，我們選擇觀察「覺得」的「覺」的流音方法部分，圖 4.19 中顯示在朗讀式語音時，「覺」不管是單一字或字詞的統計其信心度量分布近似，皆是有集中於高值的範圍的現象，而在低值部分也有集中的現象發生；在自發性語音中，「覺」的流音其信心度量分布則是集中於低值，相較於朗讀式語音，其發音是較有問題的。

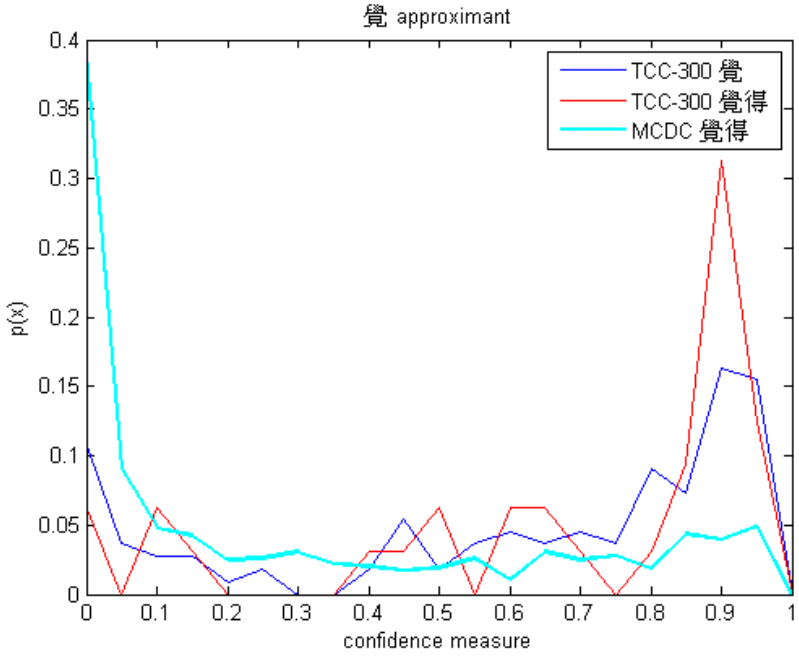


圖 4.19：「覺」之流音發音其信心度量值分布圖

圖 4.20 則是「覺」在經過動態時間校正後其流音的時間長度統計。圖中顯示了在朗讀式語音中「覺」的流音集中於一個音框以及四個音框長度左右，根據信心度量分布的圖，我們可以推斷一個音框長度的結果有比較多脫落的現象，而其餘非一個音框之結果的則脫落的機會較低；在自發性語音中，其「覺」之流音有超過八成的結果是集中於單一個音框，在比對其信心度量值分布後，我們推斷「覺」之流音的脫落機率是非常的高。

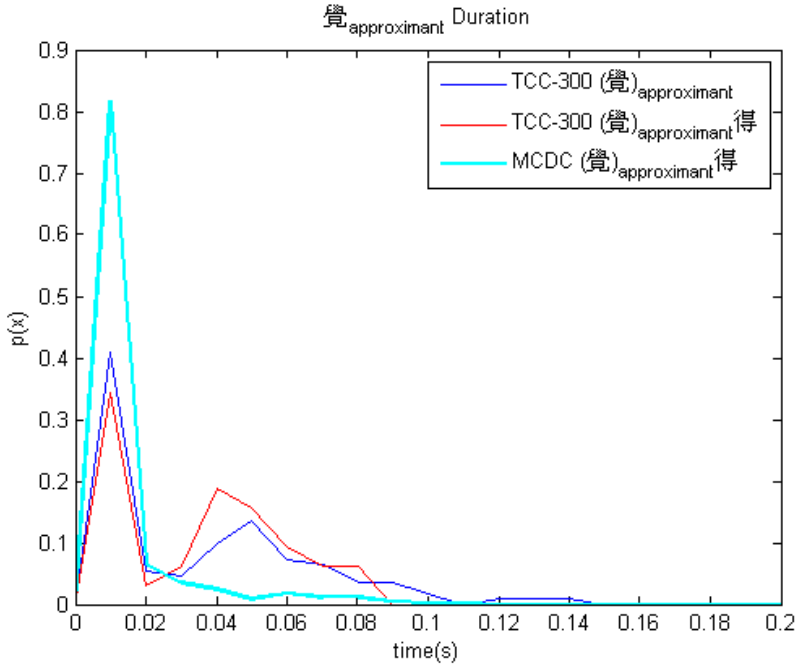


圖 4.20：「覺」經過動態時間校正後其流音所佔之時間長度

由以上的觀察，我們可以推斷在朗讀式語音中，因為語者會將字詞唸的較完整，因而語音之語速較慢，所以不論是何種發音方式其信心度量值之分布依舊是能維持在高值範圍，且動態時間校正之結果也不會極端的集中於一個音框長度；但當在自發性語音中，因為語者此時沒有文本的完整規劃，且除了語速會加快外，其在產生語音時會為了節省發音力氣而將某些發音弱化或脫落，造成發音方式偵測的難度上升，因此信心度量值容易集中於最低值外，而其長度也集中於一個音框長度。

因為在自發性語音中音素脫落之現象是非常容易產生的，然而先前的動態時間校正之設定中並不允許跳過目標字詞之發音方式，因此每個發音方式在經過校正後無論如何都至少會有一個音框長度，這也是為何先前觀察的字詞其發音方式很容易集中於一個音框長度的原因，因此研究中重新設定動態時間校正的條件，允許其可以跳過發音方式，並且對於上面觀察過的字詞其發音方式的脫落進行統計。在此，我們只統計自發性語音的結果，而其統計結果如表 4.1。

表 4.1：中文常用字詞在自發性語音中脫落現象之統計

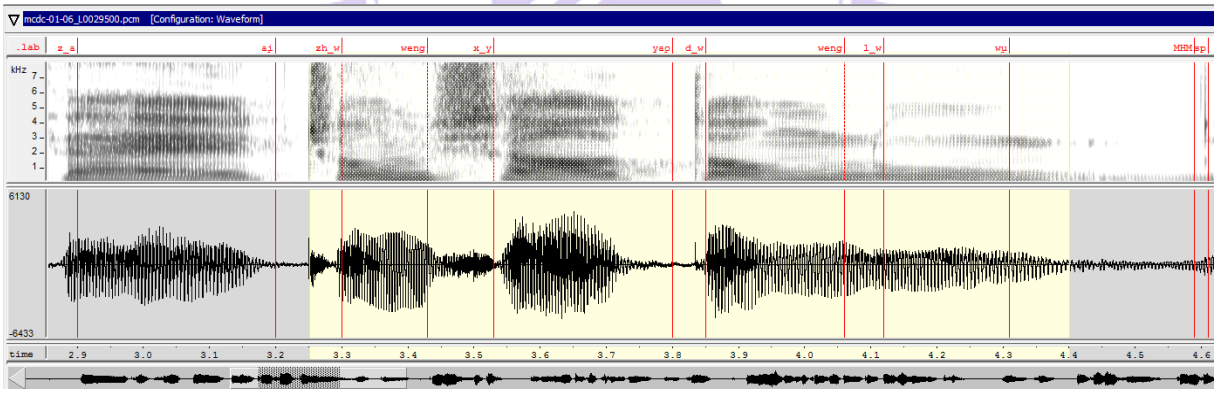
字詞	發音方式	總數	個數	遺失率
因為	因 nasal	675	429	63.6%
因為	為 approximant	675	445	65.9%
如果	果 stop	146	119	81.5%
所以	所 approximant	491	435	88.6%
真的	的 stop	273	238	87.2%
覺得	覺 approximant	688	497	88.7%
我們	我 approximant	912	515	56.5%
什麼	麼 vowel	404	128	31.7%
可以	可 stop	272	33	12.1%
沒有	有 approximant	586	396	67.6%

由表 4.1 可以看見除了字詞「可以」以及「什麼」的其目標發音方式遺失率較低外，其餘字詞的目標發音方式的遺失率都是超過五成以上，這跟我們先前觀察信心度量值分布以及校正後之長度的推論是一樣的。這個結果也代表著在動態時間校正時設定發音方式允許跳過是可行的，而且此結果更加地符合語者實際上的發音方式，其根據偵測器輸出值將較不可能存在之發音方式予以刪除，故不須經由統計信心度量值以及音框長度後再進行觀察判斷發音方式是否脫落或存在，此為一個較佳的作法，因此之後的觀察統計其設定都為動態時間校正時可允許跳過發音方式。

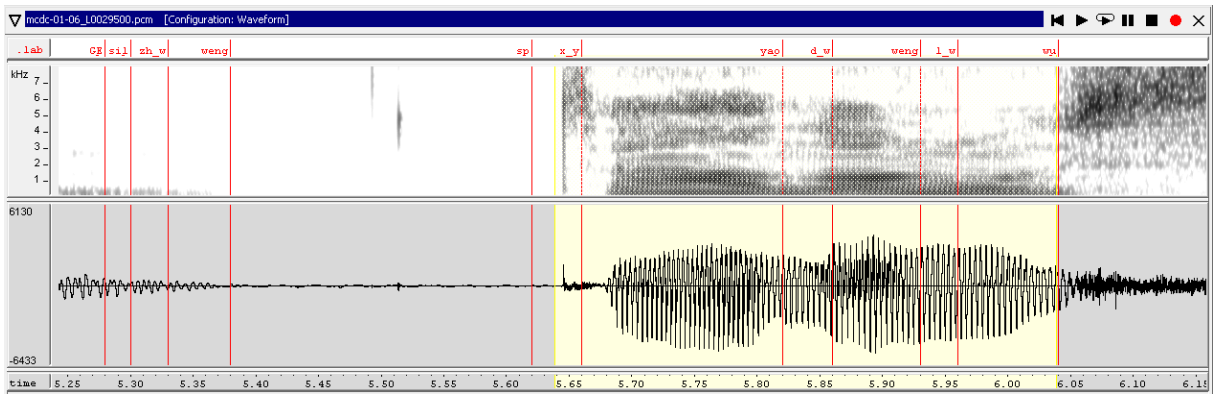
4.2 重複字詞之觀察

此外，語者在講話時其對話中同一個名詞第一次出現時應該是會將語速放慢且發音清楚完整以表示語者之意思，而後續出現之名詞就會因為先前語者已經強調過而產生些許變化，如發音間距縮短且語速加快導致的音節弱化或者脫落。在此，我們聚焦於路名上，因其相較於其他的名詞來說變化性較少所以較容易統計觀察。圖 4.21(a)顯示語者在語句中第一次講到「忠孝東路」時的對齊結果，其強迫對齊時間區段為 1.11 秒，而黃色區段為人工辨識的區段結果，其長度為 1.15 秒；圖 4.21(b)為語者第二次提到「忠孝東路」的對齊結果，其強迫對齊之區段時間長度為 0.74 秒，然而實際上此對齊是有誤差存在的，而黃色區段之人工辨識結果相對來說準確許多，其區段的時間長度為 0.4 秒；圖 4.21(c)則是語者第三次提到「忠孝東路」的結果，對齊區段以及人工標記長度皆為 0.74 秒，唯一的差別在於有一些小小的偏移。

(a)



(b)



(c)

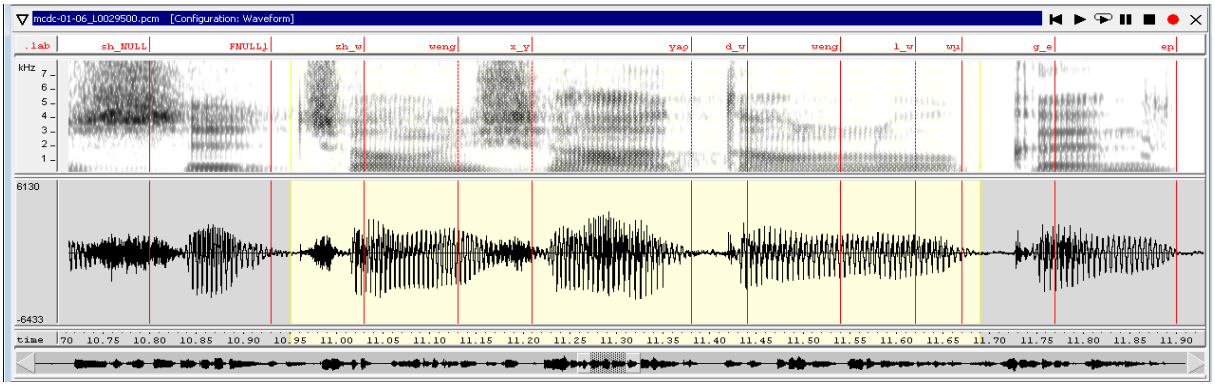


圖 4.21：語者在語句中講同一個名詞「忠孝東路」其在不同時間順序上出現之對齊結果，依據名詞出現先後順序分為(a)、(b)、(c)。圖中由上往下分別為強迫對齊時間文字轉寫、聲譜圖以及原始語音信號

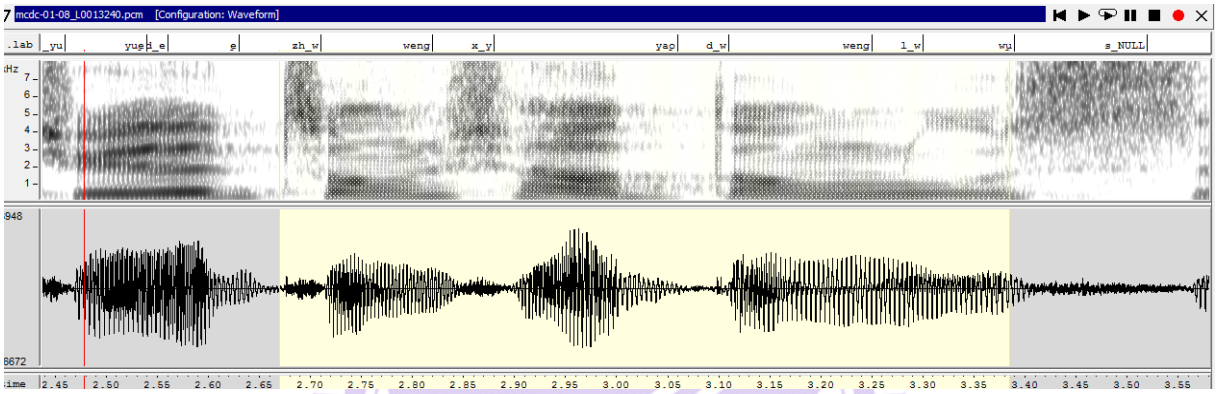
由上圖可以證實在第一次出現之名詞其時間區段對比於其後出現之區段時間長度確實是較長。為了將此三段同一語者在同一語句中不同時間出現之相同名詞做觀察，研究中一樣使用本論文 3.4.3 節所訓練而得的模型系統對語音信號進行語言屬性發音方法的偵測，並一樣由 HTK 對齊之區段結果當作偵測區段進行允許跳過發音方式之動態時間校正，因此輸出之結果更能代表語音實際上之發音。表 4.2 為上述語句中不同位置之三個名詞「忠孝東路」其偵測輸出經過動態時間校正後之結果以及文字之發音方式序列，在發音方法序列中頓號「、」表示隔開每個獨立中文字之發音方式。

表 4.2：在同一語句不同位置之「忠孝東路」經過偵測器和動態時間校正之結果

文字之正確發音方式		
中文字詞	**Manner sequence**	
忠孝東路(correct)	FAVN、FVA、OAVN、AV	
偵測結果		
中文字詞	**Manner sequence**	**Miss rate(%)**
忠孝東路(1st)	FAVN、FVA、OAVN、AV	0
忠孝東路(2nd)	FVN、A、AVN、V	38.5
忠孝東路(3th)	FAVN、FV、OVN、V	23.1

由表中可以看見第一次出現的名詞其偵測之結果並沒有造成發音方法脫落之現象，而後續出現的目標名詞其發音方法偵測結果就開始有脫落的情形發生，這例子符合我們先前提出的理論。研究繼續觀察其他例子，圖 4.22 為同一語者在另外一個語句中講到「忠孝東路」的情況，圖 4.22(a)為第一次出現的目標名詞，其對齊區段長度為 0.76 秒，人工標記區段長度為 0.72 秒；圖 4.22(b)為第二次出現之目標名詞，其對齊區段長度為 0.46 秒，人工標記區段長度為 0.43 秒。

(a)



(b)

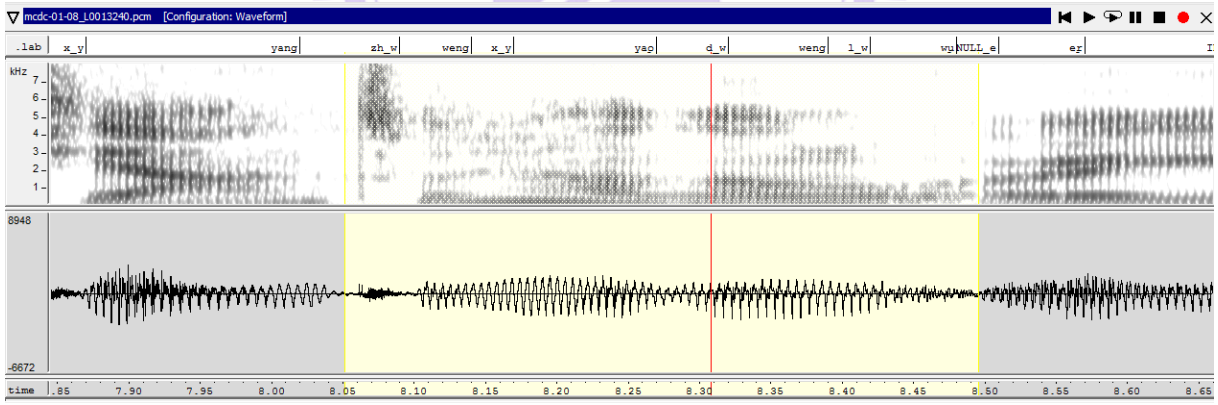


圖 4.22：語者在語句中講同一個名詞「忠孝東路」其在不同時間順序上出現之對齊結果，依據名詞出現先後順序分為(a)、(b)。圖中由上往下分別為強迫對齊時間文字轉寫、聲譜圖以及原始語音信號

而經過偵測器輸出以及動態時間調整後，我們一樣將其結果進行簡單的統計，統計之結果清楚地列於表 4.3。

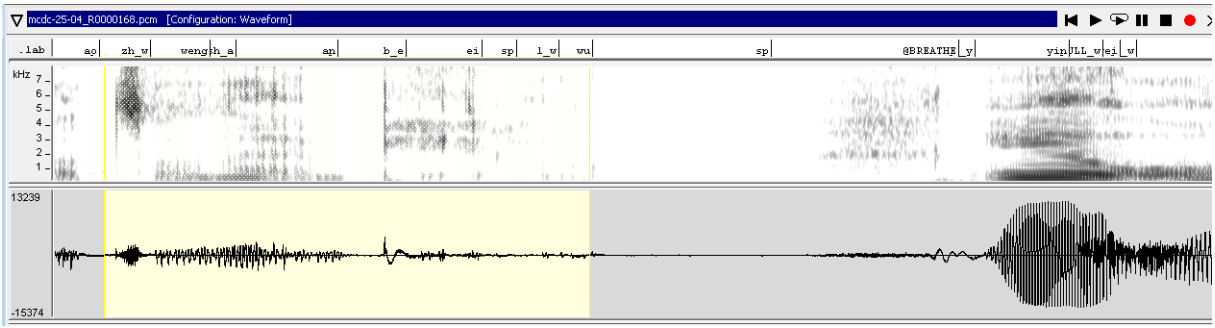
表 4.3：在同一語句不同位置之「忠孝東路」經過偵測器和動態時間校正之結果

文字之正確發音方式	
中文字詞	**Manner sequence**
忠孝東路(correct)	FAVN、FVA、OAVN、AV

偵測結果		
中文字詞	**Manner sequence**	**Miss rate(%)**
忠孝東路(1st)	FAVN、FVA、OAVN、V	7.7
忠孝東路(2nd)	FVN、FV、OVN、AV	23.1

表中一樣顯示了句子中首次出現之目標名詞其脫落的比例是小於其後出現之名詞，因此先前的推論是正確的。然而，此種現象不一定適用於所有人類的語音對話中，人類在自發性語音時會一邊產生語音一邊在大腦中規劃思考構詞，但因為並非是一個良好規劃後之結果，所以語流容易出現遲疑、口吃、非流暢現象...等。而在非流暢現象中，有些語者在補語或句末會將音量降低，即便其大腦已有構詞但發音的表現卻不完整且音量微弱，此種情形下就會造成此段區段偵測上的錯誤率上升，因此若是目標名詞其首次出現是座落於此區段，則後續出現之目標名詞偵測結果就有極大的可能優於首次出現的偵測結果。圖 4.23 為同一語者在同一個語句中講到「中山北路」的情況，圖 4.23(a)為語者在語句中首次說出目標名詞，而此名詞的區段剛好是在句末且音量微弱，對齊區段和人工標記一樣區段完全一致，長度為 0.58 秒；圖 4.23(b)為語者第二次說出目標名詞，此次的發音的音量正常，而對齊區段和人工辨識的區段一致，時間長度為 0.70 秒。

(a)



(b)

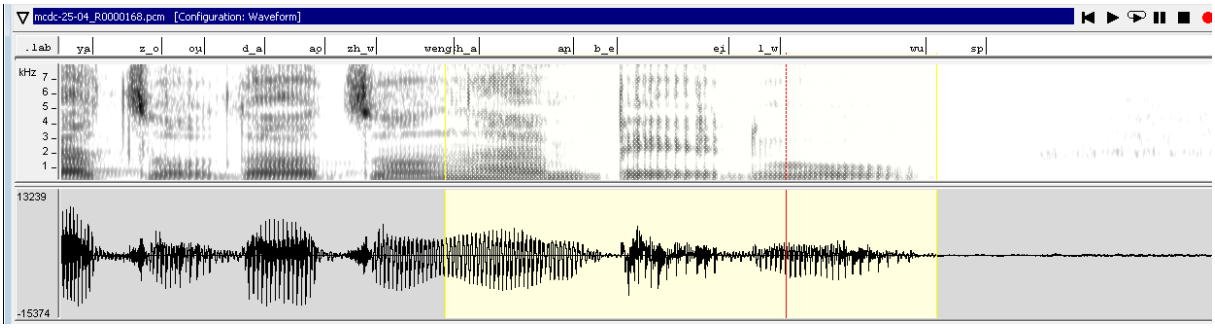


圖 4.23：語者在語句中講同一個名詞「中山北路」其在不同時間順序上出現之對齊結果，依據名詞出現先後順序分為(a)、(b)。圖中由上往下分別為強迫對齊時間文字轉寫、聲譜圖以及原始語音信號

在觀察兩個目標名詞的區段以及其聲音大小後，我們可以推斷在第二次出現的發音結果會優於第一次出現之名詞，表 4.4 顯示了此段語音經過動態時間校正後之結果。

表 4.4：在同一語句不同位置之「中山北路」經過偵測器和動態時間校正之結果

文字之正確發音方式		
中文字詞	**Manner sequence**	
中山北路(correct)	FAVN、FVN、OV、AV	
偵測結果		
中文字詞	**Manner sequence**	**Miss rate(%)**
中山北路(1st)	FVN、VN、OV、AV	18.2
中山北路(2nd)	FVN、FVN、OV、AV	9.1

由結果可以發現兩者的發音方式脫落的比例都不高，然而第二次所出現的目標名詞其發音還是比第一次的完整。最後我們可以簡單的下個結論，在正常的情況來說，通常第一次出現之名詞其發音會比後面出現的完整，這在語料中之比例是佔大多數的，然而在自發性語音中有太多的不可測之因素導致此結果不一定會出現，因此會有少數情況是後面出現之目標名詞其發音較完整的現象發生。

第五章 結論與未來展望

5.1 結論

本論文中以獲得一個精準的國語語言屬性偵測器系統為目標，將模型使用於國語自發性語言中並觀察語音之現象，且期望系統能修改目標字詞之文本讓其能夠確實對應至實際語音發音，如此對於語言方面之模型建立如聲學模型是大有幫助的。

本實驗中，先由一些英文語料庫及國語朗讀式語料庫經由類神經網路之多層感知器架構訓練一組通用語言屬性偵測系統，接著將系統運用於國語朗讀式語料中測試，並將輸出結果合併梅爾倒頻譜係數重新訓練一組國語類音素聲學模型，且對國語朗讀式語料進行語音自動分段的程序。實驗中顯示了新的分段結果對比於未加入偵測輸出所訓練之模型之分段結果準確率上提升了許多。而後實驗由新的分段結果調整語言屬性偵測系統使其對於國語語音之效果能夠達到最佳，模型經過測試後，每一樣語言屬性的偵測正確率都至少達到了 95%以上。

除了獲得了一個良好之國語語言屬性偵測系統外，實驗中將其使用於國語自發性語音中進行偵測，並將輸出經由動態時間校正調整以觀察人類在自發性語音中所產生的語言現象。實驗數據中顯示自發性語音中常常會因為語者將音素省略或弱化，因此常用的目標字詞中很容易會有發音方式脫落之現象，藉由偵測每個目標字詞之語言屬性是否有脫落就可以將文本修正為較符合實際發音之對應。另外研究中也觀察了同一目標字詞在同一語句中的前後其發音之不同，在實驗中顯示了通常目標字詞在首次出現時發音會較完整，後續出現的則會產生了較嚴重的語言屬性脫落之現象，這些數據之結果可以驗證了語言學家提出的語言現象確實是存在的。

5.2 未來展望

本研究中運用了國語語言屬性偵測系統觀察自發性語音中的語言現象，並且可以根據偵測結果有效地修改了目標字詞之文本。然而實驗仍有很多部分可以改進，例如國語類音素其語言屬性之分類，尤其是流音之部分仍尚有討論之空間；另外研究中忽略了韻母加上介音後本來就會產生的發音變異，將這些發音直接由注音轉換成類音素，如「ㄩㄣ」或「一ㄥ」實際上並不等同於「yu2 e en」和「yi2 e ng」，因此在最初的文本標記上也有改進之空間；此外，本論文中只使用了發音方式之語言屬性，然而每個音素其實皆由發音方式以及發音位置組成，因此若是另外建立一組發音位置之語言屬性偵測系統共同使用於研究中的話，應會提高偵測的準確性。除了上述之部分仍有可以進步的地方外，在未來的實驗能夠藉由此偵測系統探討更多的語音現象，並將修正文本之能力從目標字詞擴大範圍至整段語音，如此不僅能夠使文本和實際語音更精準的對應，也能夠使語音相關之研究能夠因為有一個更精確之文本而得到一個更佳的結果。

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*, vol. 4, 2004.

【2】 S. M. Siniscalchi, D-C Lyu, T. Svendsen, and C-H Lee, “Experiments on cross-language attribute detection and phone recognition with minimal target-specific training data,” *IEEE Trans. Audio Speech and Language Processing*, vol. 20(3), pp. 875-887, 2012.

【3】 S. M. Siniscalchi, and C-H Lee, “A study on integrating acoustic-phonetic information into lattice rescoring for automatic speech recognition,” *Speech Communication*, vol. 51(11), pp. 1139-1153, 2009.

【4】 P. Schwarz, “Phoneme recognition based on long temporal context,” Ph.D. dissertation, Faculty of Information Technology BUT, 2009.

【5】 S-C Tseng, “Lexical coverage in Taiwan mandarin conversation,” *Computational Linguistics and Chinese Language Processing*, vol. 18(1), pp. 1-18, March 2013.

【6】 P. Färber, “Quicknet on multispert: fast parallel neural network training,” Technical report, Tech. Rep. TR-97-047, ICSI, 1997, 1997.

【7】 V. H. Do, “Hybrid architectures for speech recognition,” Ph.D. dissertation, NTU, 2011.

【8】 S. M. Siniscalchi, P. Schwarz, C-H Lee, “High-accuracy phone recognition by combining high-performance lattice generation and knowledge based rescoring,” *Proc. ICASSP2007*, vol. 4, pp. 869-872, 2007.

【9】 C-H Lee, Mark A. Clements, S. Dusan, Eric F-C, “An overview on automatic speech attribute transcription(ASAT),” *Proc. Interspeech*, pp. 1825-1825, 2007.

【10】 R. A. Dunne, N. A. Campbell, “On the pairing of the softmax activation and cross-entropy penalty functions and the derivation of the softmax activation function,” *Proc*

conf. on the Neural networks, vol. 185 ,1997.

【11】 http://www.wudisk.com/v-08fd6058e7e27bf2dd3667f78a3b2d92.html

【12】V. H. Do, X Xiao, V. Hautamäki, E. S. Chng, "Speech attribute recognition using context-dependent modeling," *APSIPA ASC*, 2011.

【13】S. M. Siniscalchi, T. Svendsen and C-H Lee, "Toward a detector-based universal phone recognizer," *ICASSP 2008*, pp 4261-4264, 2008.

【14】P. Matějka, P. Schwarz, J. Černocký, "Phonotactic language identification using high quality phoneme recognition," *Proc. Interspeech*, pp 2237-2240, 2005.

【15】D-C Lyu, S. M. Siniscalchi and C-H Lee, "An experimental study on continuous phone recognition with little or no language-specific training data," *Proc. ITRW*, 2008.

【16】Hermansky, H. and Sharma and Sangita, "Temporal Patterns (TRAPS) in ASR of Noisy Speech," *Proc. ICASSP*, vol. 1, pp 289-292, 1999.

【17】C-C Kuo, "Feature Systems in Linguistics," 2005.

附錄一

國語類音素屬性之分類

表一、類音素之語言屬性對照表

類音素	發音方式	發音位置	類音素	發音方式	發音位置
a	Vowel	Low	f	Fricative	Labial
o	Vowel	Mid	j	Fricative	Coronal
e	Vowel	Mid	q	Fricative	Coronal
eh	Vowel	Mid	x	Fricative	Coronal
yi1	Vowel	High	h	Fricative	Velar
yi2	Vowel	High	l	Approximant	Coronal
yi3	Vowel	High	r	Approximant	Retroflex
FNULL1	Vowel	High	b	Stop	Labial
FNULL2	Vowel	High	p	Stop	Labial
wu1	Approximant	High	d	Stop	Coronal
wu2	Vowel	High	t	Stop	Coronal
wu3	Approximant	High	g	Stop	Velar
yu1	Approximant	High	k	Stop	Velar
yu2	Vowel	High	m	Nasal	Labial
er	Vowel	Mid	n	Nasal	Coronal
zh	Fricative	Retroflex	en	Nasal	Coronal
ch	Fricative	Retroflex	ng	Nasal	Velar
sh	Fricative	Retroflex	sil	Silence	Silence
z	Fricative	Coronal	Ocl	Silence	Silence
c	Fricative	Coronal	sp	Silence	Silence
s	Fricative	Coronal			

E S A
1896